SIMPLE LINEAR REGRESSIONN AND CORRELATION

CONTENTS

- Introduction
- The regression model
- The sample regression equation
- Evaluate the regression equation
- Using the regression equation
- The correlation model
- The correlation coefficient
- Some precaution

Introduction

Bivariate relationships

- **Chi-square tests of independence to determine whether a statistical relationship existed between two variables. It does not tell us what the relationship is.**
- **Regression and correlation analyses will show how to determine both the nature and the strength of a relationship between two variables.**
- **The term "regression " was first used as a statistical concept by Sir Francis Galton. He designed the word regression as the name of the general process of predicting one variable (the height of the children) from another (the height of the parent). Later, statisticians coined the term multiple regression to describe the process by which several variables are used to predict another.**

- In regression analysis an estimating equation – that is a mathematical formula that relates the known variables to the unknown variable.
- Correlation analysis determine the degree to which the variables are related and tell us how well the estimating equation actually describes the relationship.

Bivariate frequency distribution

- In order to summarize the bi-variate data, for each variable a suitable number of classes are taken, keeping in view the same consideration as in the uni-variate case.
- If there are 'K' classes for X and 'L' classes for Y. Then, the data can be summarized by using a two-way frequency table having KL cells.

- By going through the pairs of values of X and Y the number of individuals in each cells can be found out by using the system of tally marks. The whole set of cell frequencies will now define a frequency distribution called bi-variate frequency distribution.

Correlation

- Definition:

  Correlation analysis is a statistical technique of measuring the degree, direction, causation and significance of co-variation between two or more variables.

Correlation analysis consists of three steps

(i) Determining whether relation exists and, if it does, measuring it.

(ii) Testing whether it is significant.

(iii) Establishing the cause & effect relation if any.

Significant correlation between increase in smoking and increase in long cancer does not prove that smoking causes cancer. The proof of a cause & effect relationship can be developed only by means of an exhaustive study of the operative elements themselves.

Significance of the study of correlation

- Most variables are correlated in some way or other & correlation analysis helps to measure the degree & direction of correlation in one figure.

  Ex:- Age & Height, Blood pressure and pulse rate

- Given the close correlation between two or more variables, the value of one can be estimated for known value of other variable.
- The effect of correlation is to reduce the range of uncertainty. The prediction based on correlation analysis is likely to be more reliable and near to reality.

Correlation may be due to following reasons.

- May be due to pure chance, especially in a small sample

ii. Both the correlated variables may be influenced by one or more variables.

Example: Correlation between yield per acre of rice & yield per acre of tea may be due to the fact that both are depended upon amount of rainfall.

iii. Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect.

Example: Correlation" between demand and supply, price & production. The variables in that case mutually interact.

iv. Correlation may be due to the fact that one variable is the cause & other variable the effect.

Types of Correlation

i. positive or negative
ii. Simple, partial & multiple
iii. Linear and non-linear

i. Positive and negative correlation

- Correlation is said to be positive (direct) if both the variables vary in the same direction i.e. if one is increasing (decreasing) the other on an average, is also increasing (decreasing).
- It is said to be negative (inverse) if the variables are varying in opposite direction i.e. if one is increasing, the other is decreasing or vice versa.

Example : In recent years, physicians have used the so-called diving reflex to reduce abnormally rapid heartbeats in humans by submerging the patient's face in cold water.

Simple, Partial and Multiple Correlation

- Simple correlation is a study between two variables
- When three or more variables are studied it is a problem of either multiple or partial correlation. In multiple correlation, three or more variables are studies simultaneously.
- In partial correlation, three or more variables are recognized, but the correlation between two variables are studied keeping the effect of other influencing variable (s) constant.

iii) Liner and Non-Linear (Curvi-linear) Correlation

- If the amount of change in one variable tends to bear a constant ratio to the amount of change in other variable then the correlation is said to be linear.
- If the amount of change in one variable does not bear a constant ratio to the amount of change in other variable then the correlation would be non-linear.
- However, since the method of analyzing non-linear correlation is very complicated, generally, we assume linear relationship between the variables.

Scatter diagram method of studying correlation

- In this method, the given pairs of observations (x_i, y_i) are plotted on a graph paper. We get as many dots in the graph paper as the pairs of observations. By studying the scatteredness of the dots we can form an idea about the direction of the correlation and whether it is higher low. If dots are more scattered, the correlation is less and vice versa.

Karl Pearsons' Co-efficient of Correlation

This is one of the mathematical methods of finding out degree and direction of Correlation and is popularly known as Pearson's Co-efficient of Correlation. This is denoted by the symbol 'r'.

$$r = \frac{{}^{n}\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{\sqrt{{}^{n}\sum x_i^2 - \left(\sum x_i\right)^2}\sqrt{{}^{n}\sum y_i^2 - \left(\sum y_i\right)^2}}$$

Assumptions:

- For each value of X there is a normally distributed subpopulation of Y values.
- For each value of Y there is a normally distributed subpopulation of X values.
- The joint distribution of X and Y is a normal distribution called the bivariate normal distribution.
- The subpopulations of Y values have the same variance.
- The subpopulations of X values have the same variance.

Significance test for correlation:

Hypothesis: $H_0: \rho = 0$

$H_A: \rho \neq 0$

Test Statistics:

Under the null hypothesis (when null hypothesis is true)

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

follows a Students 't' distribution with n-2 degrees of freedom.

Decision rule:

If we let $\alpha = 0.05$, the critical value of t is ± 2.0639

If the computed value of 't' ≥ 2.0639 or ≤ -2.0639, we reject H_0 or it may be accepted.

Hypothesis: $H_0: \rho = \rho_0$, where ρ_0 some value other than 0.

$H_A: \rho \neq \rho_0$

Test statistics:

$$z = \frac{Z_r - Z_p}{1/\sqrt{n-3}}$$

Which follows approximately the standard normal distribution and decision rule of z test is followed.

Sample size should be ≥25

$$Z_r = \frac{1}{2} n\left[\frac{1+r}{1-r}\right]$$

$$Z_p = \frac{1}{2}\frac{1}{n}\left[\frac{1+\rho}{1-\rho}\right]$$

$\frac{1}{\sqrt{n-3}}$ estimated standard deviation.

If sample size ≥10 but <25, Hotelling procedure may be used.

Hoteling procedure:

Test statistics:

$$z^* = \frac{z^* -}{1\sqrt{n-1}}$$

Where

$$z^* = z_r - \frac{3z_r + r}{4n}$$

$$\rho^* = z_\rho - \frac{3z_\rho + \rho}{4n}$$

When the assumptions are not made, spearman's rank correlation co-efficient is used.

Co-efficient of Determination

- The Co-efficient of determination = r^2. This interpretes that r^2 percent of the variation in the dependent variable has been explained by the independent variable. Co-efficient of determination may assume the maximum value 1. The co-efficient of determination r^2 is defined as the ratio of the explained variance to the total variance.
- The ratio of un-explained variance to total variance is frequently called co-efficient of non-determination, denoted by K^2 and its square root is called the co-efficient of alienation, or K.

Merits and limitations of the Pearsons co-efficient

This is the widely used for measuring the degree of relationship. This summarize in one figure the degree and direction of correlation.

Limitations

- Karl Pearson's co-efficient of correlation assumes linear relationship regardless of the fact whether that assumption is correct or not.
- The value of the co-efficient is unduly affected by the extreme items.
- Very often there is risk of misinterpreting the co-efficient hence great care must be exercised, while interpreting the co-efficient of correlation.

Interpretation of Co-efficient of Correlation

The general rules for interpreting the value of r is given below.

- r = +1 implies perfect positive relationship between the variables.
- The closeness of relationship is not proportional to r.

- r = -1 implies perfect negative relationship between the variables.
- r = 0 implies no relationship between the variables.
- The closer r is to = 1 or –1 the closer the relationship between the variables and closer r is to 0, the less close the relationship.

The Simple Regression Model

- **Regression: Helpful in ascertaining the probable form of the relationship between variables.**
- **Main objective of regression is to predict or estimate the value of one variable corresponding to a given value of another variable**
- **One type of probabilistic model, a simple linear regression model, makes assumption that the mean value of y for a given value of x graphs as straight line and that points deviate about this line of means by a random amount equal to e, i.e. $y = A + B\,x + e$**

where A and B are unknown parameters of the deterministic (nonrandom) portion of the model.

- If we suppose that the points deviate above or below the line of means and with expected value $E(e) = 0$ then the mean value of y is $y = A + B\,x$.
- Therefore, the mean value of y for a given value of x, represented by the symbol $E(y)$ graphs as straight line with y-intercept A and slope B.

A SIMPLE LINEAR REGRESSION MODEL

$y = A + B\,x + e,$

where

y = dependent or response variable

x = independent or predictor variable

e = random error

A = y-intercept of the line

B = slope of the line

ASSUMPTIONS REQUIRED FOR A LINEAR REGRESSION MODEL

- The mean of the probability distribution of the random error is 0, $E(e) = 0$. that is, the average of the errors over an infinitely long series of experiments is 0 for each setting of the independent variable x. this assumption states that the mean value of y, $E(y)$ for a given value of x is $y = A + B\,x$.
- The variance of the random error is equal to a constant, say σ^2, for all value of x.
- The probability distribution of the random error is normal.
- The errors associated with any two different observations are independent. That is, the error associated with one value of y has no effect on the errors associated with other values.

Estimating A and B: the method of least squares

To find estimators of A and B of the regression model based on a sample data .

- **Suppose we have a sample of n data points (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). The straight-line model for the response y in terms x is $y = A + B\,x + e$.**
- **The line of means is $E(y) = A + B\,x$ and the line fitted to the sample data $\hat{y} = a + bx$**
- **Thus, $\hat{y}$ is an estimator of the mean value of y and a predictor of some future value of y; and a, b are estimators of A and B, respectively.**
- **For a given data point, say the point (x_i, y_i), the observed value of y is y_i and the predicted value of y would be $\hat{y}_i = a + b\,x_i$**

and the deviation of the ith value of y from its predicted value is $SSE=\sum_{i=1}^{n}\left[y_i-(a+bx_i)\right]^2$

The values of a and b that make the SSE minimum is called the least squares estimators of the population parameters A and B and the prediction equation is called the least squares line.

FORMULAS FOR THE LEAST SQUARES ESTIMATORS

Slope: $b=\frac{SS_{xy}}{SS_{xx}}$ y intercept: $a=\bar{y}-b\bar{x}$

Where

$$SS_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \qquad SS_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \quad , \quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$$

n = sample size

Estimating σ^2

- In most practical situations, the variance σ^2 of the random error e will be unknown and must be estimated from the sample data. Since σ^2 measures the variation of the y values about the regression line, it seems intuitively reasonable to estimate σ^2 by dividing the total error SSE by an appropriate number.

ESTIMATION OF σ^2

$$S^2 = \frac{SSE}{Degreeoffr\ eedom} = \frac{SSE}{n-2}$$

where

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

INTERPRETATION OF *s*, THE ESTIMATED STANDARD DEVIATION OF *e*

- We expect most of the observed y values to lie within 2s of their respective least squares predicted value $\hat{y}$

Making inferences about the slope, B

- The probabilistic model $y = A + B\,x + e$ for the relationship between two random variables x and y, where x is independent variable and y is dependent variable, A and B are unknown parameters, and e is a random error.
- Under the assumptions made on the random error e $E(y) = A + B\,x$. This is the *population regression line*.
- If we are given a sample of n data points (x_i, y_i), $i = 1,\ldots,n$, then by the least squares method the straight line fitted to these sample data. This line is the *sample regression line*.

- It is an estimate for the population regression line.
- The theoretical background for making inferences about the slope *B* lies in the following properties of the least squares estimator *b:*

PROPERTIES OF THE LEAST SQUARES ESTIMATOR *b:*

- *b* will possess sampling distribution that is normally distributed.
- The mean of the least squares estimator *b* is *B*, $E(b) = B$, that is, *b* is an unbiased estimator for *B*.
- The standard deviation of the sampling distribution of *b* is

$$\sigma_b = \frac{\sigma}{\sqrt{SS_{xx}}}$$

where σ is the standard deviation of the random error *e*,

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Since σ is usually unknown, we use its estimator s and instead of $\sigma_b = \frac{\sigma}{\sqrt{SS_{xx}}}$ we use its estimate $S_b = \frac{s}{\sqrt{SS_{xx}}}$

For testing hypotheses about *B* first we state null and alternative hypotheses:

$$H_0 : B = B_0$$

$$H_a : B \neq B_0 \; (or B < B_0 \; or B > B_0)$$

where B_0 is the hypothesized value of *B*.

Often, one tests the hypothesis if $B = 0$ or not, that is, if *x* does or does not contribute information for the prediction of *y*. The setup of our test of utility of the model is summarized in the box.

A TEST OF MODEL UTILITY

ONE-TAILED TEST

$$H_0 : B = 0$$
$$H_a : B < 0 \quad (\text{or} B > 0)$$

Test statistic:

$$t = \frac{b}{s_b} = \frac{b}{s / \sqrt{SS_{xx}}}$$

Rejection region

$$t < -t_\alpha$$

(or $t > t_\alpha$),

Where $t\alpha$ is based on $(n - 2)$ df.

TWO-TAILED TEST

$$H_0 : B = 0$$
$$H_a : B \neq 0$$

Test statistic:

$$t = \frac{b}{s_b} = \frac{b}{s / \sqrt{SS_{xx}}}$$

Rejection region

$$t < -t_{\alpha/2} \text{ or } t > t_{\alpha/2}$$

where $t\alpha/2$ is based on $(n-2)$ df.

Example: To model the relationship between the CO (Carbon Monoxide) ranking, y, and the nicotine content, x, of an american-made cigarette the Federal Trade commission tested a random sample of 5 cigarettes. The CO ranking and nicotine content values are given in Table

Cigarette	Nicotine Content, x, mgs	CO ranking, y, mgs
1	0.2	2
2	0.4	10
3	0.6	13
4	0.8	15
5	1	20

Solution: Testing the usefulness of the model requires testing the hypothesis

$$H_0 : B = 0$$
$$H_a : B \neq 0$$

with $n = 5$ and , the critical value based on (5 -2) = 3 df.

$$t_{\alpha/2} = t_{0.025} = 3.182$$

Thus, we will reject H_0 if $t < -3.182$ or $t > 3.182$.

In order to compute the test statistic we need the values of b, S and SS_{xx}. b =20.5. S = 1.82 and SS_{xx} = 0.4. Hence, the test statistic is

$$t = \frac{b}{s/\sqrt{ss_{xx}}} = \frac{20.5}{1.82/\sqrt{0.4}} = 7.12$$

Since the calculated t-value is greater than the critical value $t_{0.025}$ = 3.182, we reject the null hypothesis and conclude that the slope .B≠0 At the significance level α = 0.05, the sample data provide sufficient evidence to conclude that nicotine content does contribute useful information for prediction of carbon-monoxide ranking using the linear model.

THANK YOU