Part V

Lecture: Regression Analysis and Correlation Analysis

When observations in respect of two variables are available, very often a relation is found to exist between them

When observations in respect of two variables are available, very often a relation is found to exist between them

Example

1. Height and Weight of persons.

When observations in respect of two variables are available, very often a relation is found to exist between them

Example

1. Height and Weight of persons.
2. Expenditure depends on income.

When observations in respect of two variables are available, very often a relation is found to exist between them

Example

1. Height and Weight of persons.
2. Expenditure depends on income.
3. Yield of a crop depends on the amount of rainfall.

When observations in respect of two variables are available, very often a relation is found to exist between them

Example

1. Height and Weight of persons.
2. Expenditure depends on income.
3. Yield of a crop depends on the amount of rainfall.
4. Production depends on price.

When observations in respect of two variables are available, very often a relation is found to exist between them

Example

1. Height and Weight of persons.
2. Expenditure depends on income.
3. Yield of a crop depends on the amount of rainfall.
4. Production depends on price.

When observations in respect of two variables are available, very often a relation is found to exist between them

Example

1. Height and Weight of persons.
2. Expenditure depends on income.
3. Yield of a crop depends on the amount of rainfall.
4. Production depends on price.

Definition

Frequently, it is desirable to express this relationship between variables by means of some mathematical equation, representing a certain geometrical curve. The process finding such a curve or its equation on the basis of a given set of observations is called curve fitting.

Some Common types of Curves

1. Straight Line: $y = a + bx$, Parabola: $y = a + bx + cx^2$.

Some Common types of Curves

1. Straight Line: $y = a + bx$, Parabola: $y = a + bx + cx^2$.
2. Cubic Curve: $y = a + bx + cx^2 + dx^3$, Logistic Curve: $\frac{1}{y} = a + bc^x$.

Some Common types of Curves

1. Straight Line: $y = a + bx$, Parabola: $y = a + bx + cx^2$.
2. Cubic Curve: $y = a + bx + cx^2 + dx^3$, Logistic Curve: $\frac{1}{y} = a + bc^x$.
3. Exponential Curve: $y = ab^x$, Geometrical Curve: $y = ax^b$.

Available Methods for Curve Fitting:

Available Methods for Curve Fitting:

1. Free-hand Method.

Available Methods for Curve Fitting:

1. Free-hand Method.
2. Method of Least Squares.

Available Methods for Curve Fitting:

1. Free-hand Method.
2. Method of Least Squares.

Available Methods for Curve Fitting:

1. Free-hand Method.
2. Method of Least Squares.

Method of Least Squares

1. Method of Least Squares is a device for finding the equation of a specified type of curve, which best fits a given set of observations.

Available Methods for Curve Fitting:

1. Free-hand Method.
2. Method of Least Squares.

Method of Least Squares

1. Method of Least Squares is a device for finding the equation of a specified type of curve, which best fits a given set of observations.
2. The method depends upon the *Principle of Least Squares*, which suggests that for the *best-fitting curve*, "the sum of the squares of differences" between the observed and the corresponding estimated values should be the minimum possible, $i.e.$, $\sum_{i=1}^{n}(y_i - \hat{y})^2$ is minimum.

Fitting of Straight Line

Let $y = a + bx$ be the equation of a straight line to be fitted to a given set of n pairs of observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$. In order to determine a and b we apply method of least squares, $i.e.$,

$$Minimize \sum_{i=1}^{n} (y_i - a - bx_i)^2,$$

which leads the normal equations:

$$\sum y = na + b\sum x, \quad \sum xy = a\sum x + b\sum x^2$$

Fitting of Straight Line

Let $y = a + bx$ be the equation of a straight line to be fitted to a given set of n pairs of observations $(x_1, y_1), (x_2, y_2) \ldots (x_n, y_n)$. In order to determine a and b we apply method of least squares, $i.e.$,

$$Minimize \sum_{i=1}^{n} (y_i - a - bx_i)^2,$$

which leads the normal equations:

$$\sum y = na + b\sum x, \quad \sum xy = a\sum x + b\sum x^2$$

Example

Determine the equation of a straight line which best fits the following data:

x:	10	12	13	16	17	20	25
y:	19	22	24	27	29	33	37

Why Regression Analysis. . .

In many business situations, it has been observed that decision making is based upon the understanding of the relationship between two or more variables.

Why Regression Analysis. . .

In many business situations, it has been observed that decision making is based upon the understanding of the relationship between two or more variables.

Example

A sales manager might be interested in knowing the impact of advertising on sales. Here, advertising can be considered as an independent variable x and sales can be considered as the dependent variable y.

Why Regression Analysis. . .

In many business situations, it has been observed that decision making is based upon the understanding of the relationship between two or more variables.

Example

A sales manager might be interested in knowing the impact of advertising on sales. Here, advertising can be considered as an independent variable x and sales can be considered as the dependent variable y.

This is an example of simple linear regression where a single independent variable is used to predict a single numerical dependent variable.

Why Regression Analysis. . .

In many business situations, it has been observed that decision making is based upon the understanding of the relationship between two or more variables.

Example

A sales manager might be interested in knowing the impact of advertising on sales. Here, advertising can be considered as an independent variable x and sales can be considered as the dependent variable y.

This is an example of simple linear regression where a single independent variable is used to predict a single numerical dependent variable.

The word "Regression" is used to denote *prediction* of the average value of one variable for a specified value of the other variable. The prediction is done by means of suitable equations, derived on the basis of available bivariate data. Such an equation is known as a Regression Equation.

Continued...

1. Regression Analysis is the process of developing a statistical model, which is used to predict the value of a dependent variable by at least one independent variable.

Continued...

1. Regression Analysis is the process of developing a statistical model, which is used to predict the value of a dependent variable by at least one independent variable.
2. In a simple linear regression analysis, only a straight line relationship between two variables is examined.

Continued...

1. Regression Analysis is the process of developing a statistical model, which is used to predict the value of a dependent variable by at least one independent variable.
2. In a simple linear regression analysis, only a straight line relationship between two variables is examined.
3. In fact, simple linear regression analysis is focused on developing a regression model by which the value of the dependent variable can be predicted with the help of the independent variable, based on the linear relationship between these two.

Example

Cadbury India Ltd, incorporated in 1948, is the wholly owned Indian subsidiary of the UK-based Cadbury Schweppes Plc., which is a global confectionary and beverages company. Cadbury India Ltd. operates in India in the segments of chocolates, sugar confectionary, and food drinks. The company has spent heavily on advertisements. The sales and advertisement expenses (in thousand rupees) for the 12 randomly selected months are given in the following table. Develop a regression model to predict the impact of advertisement on sales.

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
Advt.	92	94	97	98	100	102	104	105	105	107	107	110
Sales	930	900	1020	990	1100	1050	1150	1120	1130	1200	1250	1220

In regression analysis, we shall develop an estimating equation, ie., a mathematical formula that relates the known variable to the unknown variable. The main focus of simple linear regression analysis is on finding the straight line that fits the data best.

Equation of simple linear regression line

$\hat{y} = a + bx$, where a is the sample y intercept which represent the average value of the dependent variable when $x = 0$ and b the slope of the sample regression line, which indicates expected change in the value of y for per unit change in the value of x.

In regression analysis, we shall develop an estimating equation, ie., a mathematical formula that relates the known variable to the unknown variable. The main focus of simple linear regression analysis is on finding the straight line that fits the data best.

Equation of simple linear regression line

$\hat{y} = a + bx$, where a is the sample y intercept which represent the average value of the dependent variable when $x = 0$ and b the slope of the sample regression line, which indicates expected change in the value of y for per unit change in the value of x.

Determination of a and b

For determining the equation of the simple regression line, values of a and b are obtained from (using Least-Square method)

$$b = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sum x^2 - \frac{1}{n}(\sum x)^2}, \quad a = \frac{1}{n}\sum y - b.\frac{1}{n}\sum x.$$

Solution of the Problem

For the problem under consideration:
$\sum x = 1221, \sum y = 13060, \sum x^2 = 124581, \sum xy = 1335420,$
$b = 6565/344.25 = 19.0704, a = -852.08$

$$\hat{y} = (-852.08) + (19.07)x$$

This result indicates that for each unit increase in x (advertisement), y (sales) is predicted to increase by 19.07 units. When there is no expenditure on advertisement ($x = 0$), sales is predicted to decrease by 852.08 thousand rupees.

Standard Error Estimate

Though the regression line fits the data best, all the observed data point do not fall exactly on the regression line. There is an obvious variation of the observed data points around the regression line. So, there is a need to develop a statistic which can measure the differences between the actual values (y) and the regressed values $(\hat{y})$. Standard error fulfil this need which measures the amount by which the regressed values $(\hat{y})$ are away from the actual values (y). It represent the variability, or scattar of the observed values around the regression line. Standard error of the estimate is defined as

$$S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}.$$

Standard Error Estimate

Though the regression line fits the data best, all the observed data point do not fall exactly on the regression line. There is an obvious variation of the observed data points around the regression line. So, there is a need to develop a statistic which can measure the differences between the actual values (y) and the regressed values $(\hat{y})$. Standard error fulfil this need which measures the amount by which the regressed values $(\hat{y})$ are away from the actual values (y). It represent the variability, or scattar of the observed values around the regression line. Standard error of the estimate is defined as

$$S_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}} = \sqrt{\frac{\sum y^2 - a\sum y - b\sum xy}{n-2}}.$$

For the problem under consideration: $S_e = 37.1068$.

Why Correlation Analysis. . .

1. Correlation is the statistical tool we can use to describe the degree to which one variable is linearly related to another.

Why Correlation Analysis...

1. Correlation is the statistical tool we can use to describe the degree to which one variable is linearly related to another.
2. It is used in conjunction with regression analysis to measure how well the regression line explains the variation of the dependent variable, y.

Why Correlation Analysis. . .

1. Correlation is the statistical tool we can use to describe the degree to which one variable is linearly related to another.
2. It is used in conjunction with regression analysis to measure how well the regression line explains the variation of the dependent variable, y.
3. Regression and Correlation analysis show us how to determine both the nature and the strength of a relationship between two variables.

Why Correlation Analysis...

1. Correlation is the statistical tool we can use to describe the degree to which one variable is linearly related to another.
2. It is used in conjunction with regression analysis to measure how well the regression line explains the variation of the dependent variable, y.
3. Regression and Correlation analysis show us how to determine both the nature and the strength of a relationship between two variables.
4. In short, regression analysis is concerned with the prediction of the most likely value of one variable when the value of the other variable is known; while correlation is concerned with the measurement of the strength of association between two variables.

There are two measures for describing the correlation between two variables: the coefficient of determination and the coefficient of correlation.

Coefficient of Determination

$$r^2 = \frac{a\sum y + b\sum xy - \frac{1}{n}(\sum y)^2}{\sum y^2 - \frac{1}{n}(\sum y)^2}$$

It lies between 0 and 1.

There are two measures for describing the correlation between two variables: the coefficient of determination and the coefficient of correlation.

Coefficient of Determination

$$r^2 = \frac{a\sum y + b\sum xy - \frac{1}{n}(\sum y)^2}{\sum y^2 - \frac{1}{n}(\sum y)^2}$$

It lies between 0 and 1.

Coefficient of Correlation

$$r = \pm\sqrt{r^2}$$

It takes any value between -1 and $+1$. $r \overset{\text{sign}}{=} b$. The sign of r indicates the direction of the relationship between the two variables x and y.

$r \in (0, 1) \Rightarrow$ direct relationship (if x increases then y increases).

$r \in (-1, 0) \Rightarrow$ inverse relationship (if x increases then y decreases).

Solution of the Previous Problem

For the problem under consideration:

$$r^2 = 0.9009$$

This indicates that 90.09 percent of the variation in sales can be explained by the independent variable, i.e., advertisement. This result also explains that 9.91 percent variation in sales is explained by factors other than advertisement.

Multiple Linear Regression Model

Previously, we discussed the concept of simple linear regression model to analyze how the dependent variable (y) is affected by independent variable (x). This idea can also be generalized to any number of independent variables.

Multiple Linear Regression Model

Previously, we discussed the concept of simple linear regression model to analyze how the dependent variable (y) is affected by independent variable (x). This idea can also be generalized to any number of independent variables.

In so doing, we expect to develop models that fit the data better than would a simple linear regression model. Because, limiting the number of independent variables also limit the usefulness of the model.

Multiple Linear Regression Model

Previously, we discussed the concept of simple linear regression model to analyze how the dependent variable (y) is affected by independent variable (x). This idea can also be generalized to any number of independent variables.

In so doing, we expect to develop models that fit the data better than would a simple linear regression model. Because, limiting the number of independent variables also limit the usefulness of the model.

Although, there are a number of applications where we purposely develop a model with only one independent variable, in general we prefer to include as many independent variables as can be shown to significantly affect the dependent variable.

Multiple Linear Regression Model

Previously, we discussed the concept of simple linear regression model to analyze how the dependent variable (y) is affected by independent variable (x). This idea can also be generalized to any number of independent variables.

In so doing, we expect to develop models that fit the data better than would a simple linear regression model. Because, limiting the number of independent variables also limit the usefulness of the model.

Although, there are a number of applications where we purposely develop a model with only one independent variable, in general we prefer to include as many independent variables as can be shown to significantly affect the dependent variable.

A regression model that contains more than one independent variable is called a multiple regression model.

Many situations occur frequently in real life that have more than one independent variables. For example:

Many situations occur frequently in real life that have more than one independent variables. For example:

1. Weight of a person depends on his height and age.

Many situations occur frequently in real life that have more than one independent variables. For example:

1. Weight of a person depends on his height and age.
2. The life of a cutting tool is related to the cutting speed and the tool angle.

Many situations occur frequently in real life that have more than one independent variables. For example:

1. Weight of a person depends on his height and age.
2. The life of a cutting tool is related to the cutting speed and the tool angle.
3. Patient satisfaction in a hospital is related to patient age, type of procedure performed, and length of stay.

Many situations occur frequently in real life that have more than one independent variables. For example:

1. Weight of a person depends on his height and age.
2. The life of a cutting tool is related to the cutting speed and the tool angle.
3. Patient satisfaction in a hospital is related to patient age, type of procedure performed, and length of stay.
4. Fuel economy of a vehicle depends on engine displacement, horse power, type of transmission and weight of the vehicle.

Many situations occur frequently in real life that have more than one independent variables. For example:

1. Weight of a person depends on his height and age.
2. The life of a cutting tool is related to the cutting speed and the tool angle.
3. Patient satisfaction in a hospital is related to patient age, type of procedure performed, and length of stay.
4. Fuel economy of a vehicle depends on engine displacement, horse power, type of transmission and weight of the vehicle.

Many situations occur frequently in real life that have more than one independent variables. For example:

1. Weight of a person depends on his height and age.
2. The life of a cutting tool is related to the cutting speed and the tool angle.
3. Patient satisfaction in a hospital is related to patient age, type of procedure performed, and length of stay.
4. Fuel economy of a vehicle depends on engine displacement, horse power, type of transmission and weight of the vehicle.

Multiple regression models give insight into the relationships between these variables that can have important practical implications.

We now assume that k-independent variables are potentially related to the dependent variable. A multiple regression model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

where y is the dependent variable, $x_1, x_2, \ldots, x_k$ are the independent variables, $\beta_0, \beta_1, \ldots, \beta_k$ are called the regression coefficients and ϵ is the error term.

We now assume that k-independent variables are potentially related to the dependent variable. A multiple regression model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

where y is the dependent variable, $x_1, x_2, \ldots, x_k$ are the independent variables, $\beta_0, \beta_1, \ldots, \beta_k$ are called the regression coefficients and ϵ is the error term.
The independent variables may actually be functions of other variables. For example:
$x_2 = x_1^2,\ x_3 = x_1 x_2,\ x_4 = x_2^2,\ x_5 = x_3 x_4,\ x_7 = \ln(x_6)$ etc.

We now assume that k-independent variables are potentially related to the dependent variable. A multiple regression model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

where y is the dependent variable, $x_1, x_2, \ldots, x_k$ are the independent variables, $\beta_0, \beta_1, \ldots, \beta_k$ are called the regression coefficients and ϵ is the error term.
The independent variables may actually be functions of other variables. For example:
$x_2 = x_1^2,\ x_3 = x_1 x_2,\ x_4 = x_2^2,\ x_5 = x_3 x_4,\ x_7 = \ln(x_6)$ etc.
When $k = 2$, the regression equation represents a plane where β_0 is the intercept of the plane. We call $\beta_1,\ \beta_2$ partial regression coefficients as β_1 (resp. β_2) measures the expected change in y for per unit change in x_1 (resp. x_2) keeping the other fixed.

We now assume that k-independent variables are potentially related to the dependent variable. A multiple regression model that might describe this relationship is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon$$

where y is the dependent variable, $x_1, x_2, \ldots, x_k$ are the independent variables, $\beta_0, \beta_1, \ldots, \beta_k$ are called the regression coefficients and ϵ is the error term.
The independent variables may actually be functions of other variables. For example:
$x_2 = x_1^2,\ x_3 = x_1 x_2,\ x_4 = x_2^2,\ x_5 = x_3 x_4,\ x_7 = \ln(x_6)$ etc.
When $k = 2$, the regression equation represents a plane where β_0 is the intercept of the plane. We call $\beta_1,\ \beta_2$ partial regression coefficients as β_1 (resp. β_2) measures the expected change in y for per unit change in x_1 (resp. x_2) keeping the other fixed.
Whenever k is greater than 2, we can only imagine the response surface– we can't draw it.

Estimating the Coefficients. . .

For estimating the regression coefficients and assessing the model, let us consider the corresponding sample regression model as:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k.$$

We use the method of least squares which yields the normal equations:

$$\sum y = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_1 + \hat{\beta}_2 \sum x_2 + \ldots + \hat{\beta}_k \sum x_k$$

$$\sum x_1 y = \hat{\beta}_0 \sum x_1 + \hat{\beta}_1 \sum x_1^2 + \hat{\beta}_2 \sum x_1 x_2 + \ldots + \hat{\beta}_k \sum x_1 x_k$$

$$\sum x_2 y = \hat{\beta}_0 \sum x_2 + \hat{\beta}_1 \sum x_1 x_2 + \hat{\beta}_2 \sum x_2^2 + \ldots + \hat{\beta}_k \sum x_2 x_k$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\sum x_k y = \hat{\beta}_0 \sum x_k + \hat{\beta}_1 \sum x_1 x_k + \hat{\beta}_2 \sum x_2 x_k + \ldots + \hat{\beta}_k \sum x_k^2$$

The normal equations can be solved by any method appropriate for solving a system of linear equations to get the estimates of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$ and the required model.

Example-1

The following table shows the weights (to the nearest pound), heights (to the nearest inch), and ages (to the nearest year), of twelve boys. Obtain the linear regression model that fit the data and estimate the parameters. Estimate the weight of a boy who is 9 years old and 54 inches tall.

Weight	64	71	53	67	55	58	77	57	56	51	76	68
Height	57	59	49	62	51	50	55	48	52	42	61	57
Age	8	10	6	11	8	7	10	9	10	6	12	9

Example-1

The following table shows the weights (to the nearest pound), heights (to the nearest inch), and ages (to the nearest year), of twelve boys. Obtain the linear regression model that fit the data and estimate the parameters. Estimate the weight of a boy who is 9 years old and 54 inches tall.

Weight	64	71	53	67	55	58	77	57	56	51	76	68
Height	57	59	49	62	51	50	55	48	52	42	61	57
Age	8	10	6	11	8	7	10	9	10	6	12	9

Example-2 (Home work)

A study was performed to investigate the shear strength of soil (y) as it is related to depth in meters (x_1) and percentage of moisture content (x_2). Ten observations were collected, and the following summary quantities obtained: $n = 10$, $\sum x_1 = 223$, $\sum x_2 = 553$, $\sum y = 1916$, $\sum x_1^2 = 5200.9$, $\sum x_2^2 = 31729$, $\sum x_1 x_2 = 12352$, $\sum x_1 y = 43550.8$, $\sum x_2 y = 104736.8$ and $\sum y^2 = 371595.6$. Fit a multiple linear regression model to the data and estimates the parameters. What is the predicted strength when $x_1 = 18$ meters and $x_2 = 43$ percent?

Solution of Example-1

Here y: weight, x_1: height and x_2: age. Consider the multiple linear regression model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2.$$

Then the normal equations are

$$\sum y = n\hat{\beta}_0 + \hat{\beta}_1 \sum x_1 + \hat{\beta}_2 \sum x_2$$

$$\sum x_1 y = \hat{\beta}_0 \sum x_1 + \hat{\beta}_1 \sum x_1^2 + \hat{\beta}_2 \sum x_1 x_2$$

$$\sum x_2 y = \hat{\beta}_0 \sum x_2 + \hat{\beta}_1 \sum x_1 x_2 + \hat{\beta}_2 \sum x_2^2.$$

For the given data:
$\sum y = 753,\ \sum x_1 = 643,\ \sum x_2 = 106,\ \sum x_1^2 = 34,843,\ \sum x_2^2 = 976,\ \sum x_1 y = 40,830,\ \sum x_2 y = 6796,\ \sum x_1 x_2 = 5779.$ Then

$$12\hat{\beta}_0 + 643\hat{\beta}_1 + 106\hat{\beta}_2 = 753$$

$$643\hat{\beta}_0 + 34843\hat{\beta}_1 + 5779\hat{\beta}_2 = 40830$$

$$106\hat{\beta}_0 + 5779\hat{\beta}_1 + 976\hat{\beta}_2 = 6796$$

which after solving gives $\hat{\beta}_0 = 3.6512$, $\hat{\beta}_1 = 0.8546$ and $\hat{\beta}_2 = 1.5063$.
Hence the regression model is:

$$\hat{y} = 3.65 + 0.855x_1 + 1.506x_2$$

For $x_1 = 54$ and $x_2 = 9$ we have $y = 63.356$ pound.

?

THANK YOU


राजीव गाँधी पेट्रोलियम प्रौद्योगिकी संस्थान, जायस, अमेठी
Rajiv Gandhi Institute of Petroleum Technology, Jais, Amethi
विद्यारत्नम् महद्धनम्