Unit VI
Unit Name: Linear Statistical Models

Overview:
This unit Correlation and regression, Rank correlation.

Outcome:
After completion of this unit, students would be able to:

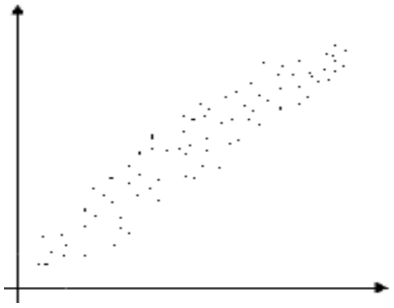
1. apply knowledge of correlation and regression to solve real life problems.

Correlation

Correlation: Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. Two variables are said to be correlated if change in one variable affects the change in other variable, and the relation between them is known as correlation.

Positive Correlation:

Two variables are said to be positively correlated if they deviates in the same direction.
e.g. height & weight, income & expenditure



Negative Correlation

Two variables are said to be negatively correlated if they deviates in the opposite directions.
e.g. volume and pressure of a perfect gas, price and demand

Un-correlation

Two variables are said to be uncorrelated or statistically independent if there is no relation between them.

Karl Pearson's Product Moment coefficient of correlation: Correlation coefficient between two random variables X and Y, usually denoted by $r(X,Y)$ or r_{xy} and defined as

$$r(X,Y)=\frac{\frac{1}{n}\sum x_i y_i-\bar{x}\,\bar{y}}{\sqrt{\left(\frac{1}{n}\sum x_i^2-\bar{x}^2\right).\left(\frac{1}{n}\sum y_i^2-\bar{y}^2\right)}}$$

OR

$$r(X,Y)=\frac{n\sum xy-\sum x.\sum y}{\sqrt{\left[n\sum x^2-\left(\sum x\right)^2\right]\left[n\sum y^2-\left(\sum y\right)^2\right]}}$$

Note:

If $r=+1$ then correlation is perfectly positive,

If $r=-1$ then correlation is perfectly negative,

If $r=0$ then variables are uncorrelated.

Spearman's Rank Correlation:

The method develop by Spearman is simpler than Karl Pearson's method since, it depends upon ranks of the items and actual values of the items are not required. Hence this can be used to study correlation even when actual values are not known.

For instance, we can study correlation between intelligence and honesty by this method.

Spearman's Rank Correlation coefficient is defined by

$$R=1-\frac{6\sum d_i^2}{n^3-n}.$$

Where $d_i=R_1-R_2$,

$R_1=$ Rank of X,

$R_2=$ Rank of Y.

If ranks are repeated then the above formula becomes

$$R=1-\frac{6\left[\sum d_i^2+\sum\left(\frac{m^3-m}{12}\right)\right]}{n^3-n}$$

Where m is the number of times an item is repeated.

Regression:

Regression can be defined as a method to estimate the value of one variable when that of other is known, when the variables are correlated. Regression analysis is a mathematical measure of average relationship between two or more correlated values.

Equations of Lines of regression:

1) Line of regression of y on x is :

$y - \bar{y} = b_{yx}(x - \bar{x})$

where regression coefficient of y on x is given by $b_{yx} = \frac{Cov(x,y)}{\sigma_x^2} = r\frac{\sigma_y}{\sigma_x}$

2) Line of regression of x on y is :

$x - \bar{x} = b_{xy}\left(y - \bar{y}\right)$

where regression coefficient of y on x is given by $b_{xy} = \frac{Cov(x,y)}{\sigma_y^2} = r\frac{\sigma_x}{\sigma_y}$

Properties: i) Lines of regression are passes through the point $(\bar{x}, \bar{y})$

ii) $b_{yx}\, b_{xy} = r^2 \quad \therefore r = \sqrt{b_{yx}\, b_{xy}}$

iii) $b_{yx}\, b_{xy}$ have same sign.

Regression using method of least squares:

1) **Least Squares Straight Line**

For a given set of N data points $(X_1, y_1), (X_2, Y_2), \ldots\ldots, (X_N, Y_N)$ assume that the straight line

$$Y = a_0 + a_1X = f(X) \qquad \ldots (1)$$

fits to the data in the least squares sense.

Normalized equation are given by

$$\sum Y_i = Na_0 + a_1 \sum X_i$$

$$\sum X_i Y_i = a_0 \sum X_i + a_1 \sum X_i^2$$

known as "**Normal equations**". In such a case Equation (1) represents a least squares straight line.

Multiple Regression:

So far, we have seen the concept of simple linear regression where a single predictor variable X was used to model the response variable Y . In many applications, there is more than one factor that influences the response. Multiple regression models thus describe how a single response variable Y depends linearly on a number of predictor variables.

Example: A multiple linear regression model with k predictor variables $X_1, X_2, ..., X_k$ and a response Y, can be written as

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

As before, the ϵ are the residual terms of the model and the distribution assumption we place on the residuals will allow us later to do inference on the remaining model parameters. Interpret the meaning of the REGRESSION COEFFICIENTS $\beta_0, \beta_1, \beta_2, ..., \beta_k$ in this model.

The simplest multiple regression model is one constructed with two independent variables, where the highest power of either variable is 1 (first-order regression model). The regression model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

The constant and coefficients are estimated from sample information, resulting in the following model.

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

For multiple regression models with two independent variables, the result is three simultaneous equations with three unknowns (b_0, b_1, and b_2).

$$b_0 n + b_1 \Sigma x_1 + b_2 \Sigma x_2 = \Sigma y$$
$$b_0 \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2 = \Sigma x_1 y$$
$$b_0 \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2 = \Sigma x_2 y$$

References of the entire unit

1: Probability, Statistics and Random Processes, T. Veerarajan, Tata McGraw Hill, 3rd edition.

2: Applied Mathematics, G.V. Khumbhojkar, C. Jamnadas & Co.

Correlation

Session 1

1. Calculate the correlation coefficient for the following heights (in inches) of fathers (X) and their sons(Y)

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

[**Ans :** r = 0.603]

2. Calculate the Karl Person's correlation coefficient from the following data

X	28	45	40	38	35	33	40	32	36	33
Y	23	34	33	34	30	26	28	31	36	35

[**Ans :** r = 0.5185]

3. Calculate the correlation coefficient from the following data

X	30	33	25	10	33	75	40	85	90	95	65	55

Y	68	65	80	85	70	30	55	18	15	10	35	45

[**Ans :** r = - 0.9935**]**

4. A computer while calculating correction coefficient between two variables X and Y from 25 pairs of observations obtained the following results

$n = 25, \sum X = 125, \sum X^2 = 650, \sum Y = 100, \sum Y^2 = 460, \sum XY = 508$

[**Ans :** r = 0.67**]**

5. The following table gives the distribution of items of production and also the relatively defective items among them according to size-groups. Is there any correlation between size and defect in quantity?

Size – groups	15–16	16–17	17–18	18–19	19–20	20–21
No. of items	200	270	340	360	400	300
No. of defective items	150	162	170	180	180	120

[**Ans :** r = 0.94**]**

6. Calculate the rank correlation coefficient from the following data.

X	1	3	7	5	4	6	2	10	9	8
Y	3	1	4	5	6	9	7	8	10	2

[**Ans :** R= 0.42**]**

7. The following table shows the marks obtained by 10 students in Accountancy and Statistics. Find the Spearman's coefficient of rank correlation.

Student No.	1	2	3	4	5	6	7	8	9	10
Accountancy	45	70	65	30	90	40	50	57	85	60
Statistics	35	90	70	40	95	40	60	80	80	50

[**Ans :** R= 0.90**]**

8. Find the coefficient of correlation between height of father and height of son from the following data.

Height of father	65	66	67	67	68	69	71	73
Height of son	67	68	64	68	72	70	69	70

[**Ans :** R= 0.47**]**

Regression

Session 2

1. The following are the marks in Statistics (X) and Mathematics (Y) of ten students

X	56	55	58	57	56	60	54	59	57	58
Y	68	67	67	65	68	70	66	68	66	70

Calculate the coefficient of correlation and estimate marks in Mathematics of a student who scored 62 marks in Statistics.
[**Ans :** r = 0.44 , Y = 69.5 **]**

2. It is given that the means of x and y are 5 and 10. If the line of regression of y on x is parallel to the line 20y = 9x + 40 , estimate the value of y at x = 30

[**Ans :** 20y = 9x + 155 and y = 175**]**

3. Find the two lines of regression from the following data

X	65	66	67	67	68	69	70	72
Y	67	68	65	68	72	72	69	71

[**Ans :** x = 30.364 + 0.545 y and y = 23.667 + 0.667 x**]**

4. In partially destroyed laboratory record of an analysis of correlation data, the following results only are legible-
   Variance of X = 9, regression equations are: 8X–10Y + 66 = 0 & 40X–18Y = 214
   What was i) the mean of X and Y
   ii) the correlation between X and Y
iii) the S.D. of Y

[**Ans:** $\overline{x} = 13, \overline{y} = 17, \sigma_y = 4, r = 0.6$]

5. You are given the following data

	X	y
Mean	30.1	47.8
Standard Deviation	6.2	9.5

6. Obtain the equation of the line of regression of **cost on age** from the following table giving the age of a car of certain make and the annual maintenance cost.

Age of car (in years)	2	4	6	8
Maintenance (in thousands of Rs.)	5	7	8.5	11

Also find maintenance cost of the car if its age is 9 years
[**Ans :** y = 3 + 0.975 x and y = Rs. 11775**]**

7. Develop the equation of the regression model for the following data. Comment on the regression coefficients. Determine the predicted value of *y* for $x1 = 33$, $x2 = 29$, and $x3 = 13$.

y	x_1	x_2	x_3
114	21	6	5
94	43	25	8
87	56	42	25
98	19	27	9
101	29	20	12
85	34	45	21
94	40	33	14
107	32	14	11
119	16	4	7
93	18	31	16
108	27	12	10
117	31	3	8