|  |  |  |
| --- | --- | --- |
| Related image | **KONERU LAKSHMAIAH EDUCATION FOUNDATION**  (Deemed to be University estd, u/s, 3 of the UGC Act, 1956) (NAAC Accredited “A++” Grade University)  Green Fields, Guntur District, A.P., India – 522502  **Department of Computer Science and Engineering**  (DST - FIST Sponsored Department) |  |

**B.Tech. II CSE(H) PROGRAM**

**A.Y. 2023-24 ODD, Semester-II**

**Course Code: 22MT2005**

**PROBABILITY, STATISTICS AND QUEUING THEORY**

**Course Outcome-2**

**Session 18:** **Regression and its Properties**

1. **Course Description (Description about the subject)**

Regression analysis is a statistical method used to examine the relationship between one or more independent variables (also known as predictor or input variables) and a dependent variable (the outcome or response variable). It is widely employed in various fields, including economics, social sciences, finance, medicine, and data science, to understand the effect of different variables on the outcome of interest and make predictions based on the observed data.

1. **Aim**

To explain identify the linear relationship between two variables using regression.

1. **Instructional** **Objectives (Course Objectives)**

To Calculate the linear relationship between two variables using different measures of regression

1. **Learning** **Outcomes (Course Outcome)**

**CO2**: Students will be able to Apply continuous probability distributions tothe real world problems also predict the relationship between variables

1. **Module** **Description** **(CO-2 Description)**

Types of Scatter diagrams, Correlation and Regression

1. **Session** **Introduction**

The main objective of many statistical investigations is to make predictions, preferably on the basis of mathematical equations. For example, in an industrial situation it may be known that the tar content in the outlet stream in a chemical process is related to the inlet temperature. It may be of interest to develop a method of prediction, that is, a procedure for estimating the tar content for various fuels of the inlet temperature form experimental information. If we study several automobiles with the same engine volume, they will not all have the same gas mileage. If we consider houses in the same part of the country that have the same part of the country that have the same square footage of living space, all will not be sold for the same price. Tar content, gas mileage, and the price of houses are natural dependent variables or responses. Inlet temperature, engine volume, and square feet of living space are respectively, independent variables or regressors.

1. **Session description**

A reasonable form of a relationship between the dependent variable and the regressors x is the linear relationship

Y=α+βx

Where, α is the intercept and β is the slope.

If the relationship is exact, then it is a **deterministic** relationship between the two variables. However, in the examples listed above, as well as countless other scientific and engineering phenomena, the relationship is not deterministic and there will be random component in it. The concept of regression analysis deals with finding the best relationship between Y and x, and using methods that allow for prediction of the response values for given values of the regressor x.

In many applications there will be more than one regressor. For example, in the case where the dependent variable is the price of house, one would expect the age of the house to contribute to the explanation of the price so in this case the **multiple regression** structure might be written as

Y=α+β1X1+β2X2

where Y is price, X1 is square footage and X2 is age in years. The resulting analysis is termed as multiple regressions while the analysis of the single regressor case is called simple regression.

**Simple Linear regression model:**

The dependent variable Y is related to the independent variable x through the equation.

Y=α+βx+ε

Where α and β are unknown intercept and slope parameters respectively, and ε is a random variable that is assumed to be distributed with E(ε)=0 and Var(ε)=σ2. Since ε is random the quantity Y is a random variable. The value x of the regressor variable is not random and measured with negligible error. Ε is called **random error or random disturbance**, has constant variance. E(ε)=0 implies that at a specific x and y values are distributed around the **true** or population **regression line** Y=α+βx.

**The method of least squares:**

An aspect of regression analysis is to estimate the parameters α and β. We denote the estimates a for α and b for β. Then the estimated or fitted regression line is given by

where is the predicted or fitted value. We expect that the fitted line should be closer to the true regression line. When a large amount of data is available.

**Residual:**

A residual is essentially an error in the fit of the model

Given a set of regression data {(xi, yi), i=1,2,...,n} and a fitted model

, the ith residual εi is given by εi=yi­-, i=1,2,...,n.

We shall find a and b, the estimates of α and β, so that the sum of the squares of the residuals is a minimum. The residual sum of squares is also called the sum of squares of the errors about the regression line and is denoted by SSE. This minimization procedure for estimating the parameters is called the method of least squares. Hence, we shall find a and b so as to minimize,

Differentiating SSE with respect to a and b, equating the partial derivatives to zero and rearranging the terms to obtain the equations (called the normal equations)

Which may solved simultaneously to yield the computing formulas for a and b.

**One scatter plot but two fitted lines:**

The fitted regression line is used to predict the value of Y when the values of regressor x are given. The estimates so obtain will be best in the sense that it will have the minimum possible error has defined by the principle of least squares. We can also obtain an estimate of the x for any given Y by using (i) but estimates so obtain will not be best since (i) is obtained on minimizing the sum of squares of errors of estimates in Y not in x. Hence to predict the X for any given value of y use the regression equation of x on y.

which is derived on minimizing the sum of squares of errors of estimates in X. Here X is dependent variable and y is an independent variable, where

and c=

The two regression equations are not reversible or interchangeable because of the simple reason that the basis and assumptions for deriving these equations are quite different.

The regression equation of Y on x is obtained on minimizing the sum of squares of the errors parallel to the Y-axis. While the regression equation of x on y is obtained on minimizing the sum of squares of the errors parallel to X-axis.

**Note**: In case of perfect correlation, r=±1, both the lines of regression coincide. Therefore, in general we always have two lines of regression except in the particular case of perfect correlation.

**Regression coefficients:**

1. The slope of the regression line of Y on X is called regression coefficient of Y on X. It is denoted by bYX.

The regression coefficient bYX represents the increment in the value of dependent variable Y for a unit change in the value of dependent variable X.

1. The slope of the regression line of X on Y is called regression coefficient of X on Y. It is denoted by bXY.

The regression coefficient bXY represents the increment in the value of dependent variable X for a unit change in the value of independent variable Y.

**Properties of regression coefficients:**

1. The geometric mean between regression coefficients is the correlation coefficient
2. In one of the regression coefficient is greater than unity, then the other must be less than unity.
3. Regression coefficients are independent of change of origin but not scale
4. The Arithmetic mean of the regression coefficients is greater than the correlation coefficient, provided r>0.
5. **Activities/ Case studies/related to the session.**

**NA**

1. **Examples & contemporary extracts of articles/ practices to convey the idea of the Session**

**Example 1:** The following are measurements of the air velocity and evaporation coefficient of burning fuel droplets in an impulse engine:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Airvelocity (cm/s) x | 20 | 60 | 100 | 140 | 180 | 220 | 260 | 300 | 340 | 380 |
| Evaporation coefficient (mm2)y | 0.18 | 0.37 | 0.35 | 0.78 | 0.56 | 0.75 | 1.18 | 1.36 | 1.17 | 1.65 |

Fit a straight line to these data by the method of least squares, and use it to estimate the evaporation coefficient of a droplet when the air velocity is 190 cm/s.

**Solution:**  For these n=10 pairs (xi, yi) we first calculate

and then we obtain

Sxx=532,000-(2,000)2/10=132,000 Sxy=2,175.40-(2,000)(8.35)/10=505.40

Syy=9.1097-(8.35)2/10=2.13745

Consequently, b=Sxy/Sxx=505.40/132,000=0.00383

and then

Thus, the equation of the straight line that best fits the given data in the sense of least squares is

And for x=190 we predict that the evaporation coefficient will be

Finally, the residual sum of squares is

**Example 2:** Engineers fabricating a new transmission-type electron multiplier created an array of silicon nanopillars on a flat silicon membrane. The precise structure can influence the electrical properties so, subsequently, the height and widths of 50 nanopillars were measured in nanometres or 10-9 meters. The summary statistics, with x=width and y=height, are N=50, Sxx=7239.22, Sxy=17840.1, Syy=66957.2

a) Find the least squares line for predicting height from width

b) Find the least squares line for predicting width from height.

c) Make a scatter plot and show both lines. Comment.

**Solution:**

a) Here y=height and the least squares estimates are

slope=b=Sxy/Sxx=17840.1/7239.22=2.464 and

The fitted line is height =87.88+2.464 width.

b) Width is now the response variable and height the predictor, so x and y must be interchanged.

Slope b= 17,840.1/66976.2=0.266 and

The fitted line is width=6.944+0.266 height.

c) Here we construct the scatter plot and include the two lines of regression. The line from part (b) is written as

Height =-(6.944/0.266)+(1/0.266)width=-26.11+3.759width

Note that both pass through the mean point (

The choice of fitted line depends on which variable you wish to predict.

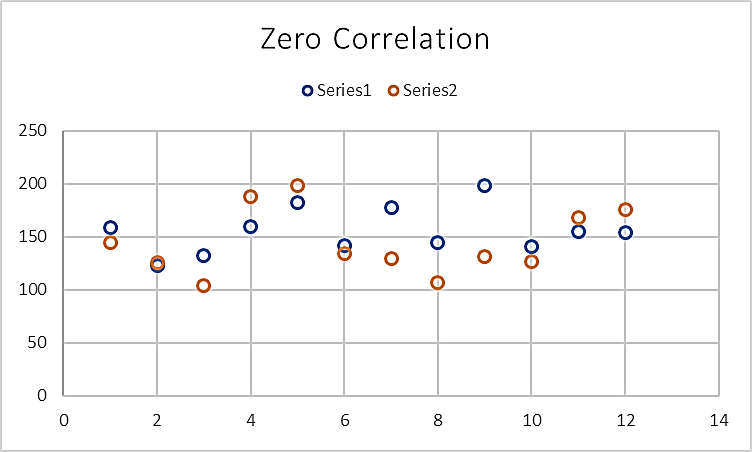
**Example** 3. a) Visualize the Zero, Positive and negative correlations.

b) Verify the type of correlation from the following data.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age (*x*) | 36 | 38 | 42 | 42 | 47 | 49 | 55 | 56 | 60 | 63 | 68 | 72 |
| Blood Pressure (*y*) | 118 | 115 | 125 | 140 | 128 | 145 | 150 | 147 | 155 | 149 | 152 | 160 |

**Answer**:

a) Zero correlation: no association or correlation between the two variables



Positive correlation. If the two variables tend to move in the same direction

A graph with numbers and dots

Description automatically generated

Negative correlation. If the two variables tend to move in the different direction

A graph with a line graph

Description automatically generated

b) Visualization: It is a positive correlation.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age (*x*) | 36 | 38 | 42 | 42 | 47 | 49 | 55 | 56 | 60 | 63 | 68 | 72 |
| Blood Pressure (*y*) | 118 | 115 | 125 | 140 | 128 | 145 | 150 | 147 | 155 | 149 | 152 | 160 |

1. **SAQ's-Self Assessment Questions**
2. In regression, the equation that describes how the response variable (y) is related to the
   1. explanatory variable (x) is:
   2. the correlation model
   3. the regression model
   4. used to compute the correlation coefficient
   5. None of these alternatives is correct.
3. Regression modeling is a statistical framework for developing a mathematical equation that describes how
   1. one explanatory and one or more response variables are related
   2. several explanatory and several response variables response are related
   3. one response and one or more explanatory variables are related
   4. All of these are correct.
4. In regression analysis, the variable that is being predicted is the
   1. response, or dependent, variable
   2. independent variable
   3. intervening variable
   4. is usually x
5. In least squares regression, which of the following is not a required assumption about the error term ε?
   1. The expected value of the error term is one.
   2. The variance of the error term is the same for all values of *x*.
   3. The values of the error term are independent.
   4. The error term is normally distributed.
6. In the case of an algebraic model for a straight line, if a value for the *x* variable is specified, then
   1. the computed response to the independent value will always give a minimal residual
   2. the exact value of the response variable can be computed
   3. the computed value of *y* will always be the best estimate of the mean response
   4. none of these alternatives is correct.

**Answers**:

1.b) 2.c) 3.a) 4.a) 5.b)

1. **Summary**

The students will understand the types of Regression and its properties and also the use of Simple Linear Regression and Multiple Linear Regressions to fit a hyperplane to the data.

1. **Terminal Questions**

1. The failure rate of a certain electronic device is suspected to increase linearly with its temperature. Fit a linear line through the data in following Table (two measurements were taken for each given temperature, and hence we have twelve pairs of measurements) and find a failure rate at ¯x = 80◦F.

**The failure rate versus temperature**

|  |  |
| --- | --- |
| T (*◦*F) | 55 65 75 85 95 105 |
| Failure rate *·*106 | 1.90 1.93 1.97 2.00 2.01 2.01 |

2. Consider a computer system that is subject to periodic diagnosis and maintenance every 1000 h. The diagnosis–maintenance service is assumed not to be perfect, and the probability of its being able to correctly diagnose and correct the fault (if it exists) is c. The expected life y of the system is to be fitted as a power function of the coverage factor c:

y = a + bec.

1. Using the following data (adapted from Ingle and Siewiorek [INGL 1976]), estimate the parameters a, b, and compute the coefficient of determination.
2. Plot a scatter diagram.

***ci yi* (h)**

0.2 11,960

0.4 15,950

0.6 23,920

0.8 47,830

0.9 95,670

0.92 120,000

0.94 159,500

0.96 239,200

0.98 478,340

3. The data in the table are the 2017 team batting averages and the number of runs fora sample of 10 MLB teams.

**Team Team-Batting Average No. of Runs**

Atlanta Braves 0.263 732

Boston Red Sox 0.258 785

Cincinnati Reds 0.253 753

Detroit Tigers 0.258 735

Houston Astros 0.282 896

Kansas City Royals 0.259 702

Miami Marlins 0.267 778

Minnesota Twins 0.260 815

Pittsburgh Pirates 0.244 668

Texas Rangers 0.244 799

a. Plot the points using a scatterplot. Does it appear that there is any relationship between the number of runs and the team batting average?

b. Is there a significant positive correlation between the number of runs and the team batting average?

c. Model a regression model of the above data, if suitable.

4. From the following table find the regression equation and compare it with the scattering plot.

**Student Mathematics Achievement Final Calculus**

**Test Score (x) Grade (y)**

1 39 65

2 43 78

3 21 52

4 64 82

5 57 92

6 47 89

7 28 73

8 75 98

9 34 56

10 52 75

1. **Case Studies (CO Wise)**

**NA**

1. **Answer Key**

**NA**

1. **Glossary**

SSE – Sum of Squared Error

1. **References of books, sites, links Textbooks:**

**Textbooks:**

1. Probability and Statistics Rukmangad Achari E. and E. Keshava Reddy
2. Probability and Statistics for Engineers and Scientists” Ronald E. Walpole, Sharon L. Myers and Keying Ye 8th Edition Pearson pub
3. Probability & Statistics for Engineers Dr. J. Ravichandran first Edition Wiley-India

**Reference books:**

1. Hossein Pishro-Nik, Introduction to Probability, Statistics, and Random Processes, 2014, by Kappa Research LLC; ISBN-13: 978-0990637202

**Web Resources**

1. https://ncert.nic.in/textbook.php?kemh1=0- 16
2. https://ncert.nic.in/textbook.php?jemh1=ps-15
3. **Keywords**

Regression analysis, prediction, Simple Linear Regression, multiple Linear Regression, SSE