Department of AI & DS
CSE and CS&IT

COURSE NAME: PROBABILITY, STATISTICS AND QUEUING THEORY

COURSE CODE: 23MT2005

Topic

Regression

Session - 13

AIM OF THE SESSION

To familiarize students with the concept of regression analysis

INSTRUCTIONAL OBJECTIVES

This Session is designed to:

1. Demonstrate Linear regression
2. Describe Linear and Non linear regression in real life applications
3. List out the two lines of regression

LEARNING OUTCOMES

At the end of this session, you should be able to:

1. Define liner regression
2. Describe the method of least squares to fit a linear and non linear association between two variables
3. Summarize the difference between linear and non linear regression.

SESSION INTRODUCTION

CONTENTS

- Linear Regression

- Nonlinear Regression

Regression analysis

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.

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.**

Regression analysis

A reasonable form of a relationship between the dependent variable and the regressors x is the linear relationship $Y=\alpha+\beta x$

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

$$Y=\alpha+\beta_1X_1+\beta_2X_2$$

Where Y is price, X_1 is square footage and X_2 is age in years. The resulting analysis is termed as multiple regressions while the analysis of the single regressor case is called simple regression.

Regression analysis

Simple Linear regression model: The dependent variable Y is related to the independent variable x through the equation

$$Y=\alpha+\beta x+\varepsilon$$

Where α and β are unknown intercept and slope parameters respectively, and ε is a random variable that is assumed to be distributed with $E(\varepsilon)=0$ and $Var(\varepsilon)=\sigma^2$. 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. E is called **random error or random disturbance**, has constant variance. $E(\varepsilon)=0$ implies that at a specific x and y values are distributed around the **true** or population **regression line** $Y=\alpha+\beta x$.

Regression analysis

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

$$\hat{y} = a + bx$$

where $\hat{y}$ 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

$$\hat{y} = a + bx$$

Given a set of regression data $\{(x_i, y_i), i=1,2,\ldots,n\}$ and a fitted model $\hat{y}_i = a + bx_i$, the i^{th} residual ε_i is given by $\varepsilon_i = y_i - \hat{y}_i$, i=1,2,...,n.

ACTIVITIES/ CASE STUDIES/ IMPORTANT FACTS RELATED TO THE SESSION

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

$$\mathrm{SSE} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}_{\mathrm{i}}^{2} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_{\mathrm{i}} - \hat{\mathrm{y}}_{\mathrm{i}})^{2} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{y}_{\mathrm{i}} - \mathrm{a} - \mathrm{bx}_{\mathrm{i}})^{2}$$

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)

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i\,,$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i$$

ACTIVITIES/ CASE STUDIES/ IMPORTANT FACTS RELATED TO THE SESSION

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

$$\mathrm{b} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})(\mathrm{y_i}-\bar{\mathrm{y}})}{\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})^2} = \frac{\mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{x_i y_i} - \sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{x_i}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{y_i}}{\mathrm{n}\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{x_i}^2 - (\sum_{\mathrm{i}=1}^{\mathrm{n}}\mathrm{x_i})^2} = \frac{\mathrm{S_{xy}}}{\mathrm{S_{xx}}}$$

and $\mathrm{a} = \bar{\mathrm{y}} - \mathrm{b}\bar{\mathrm{x}} = \left(\frac{1}{\mathrm{n}}\right)[\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} y_i]$, where $\mathrm{S_{xy}} = \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})(\mathrm{y_i}-\bar{\mathrm{y}})$,

$$\mathrm{S_{xx}} = \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{x_i}-\bar{\mathrm{x}})^2 \text{ and } \mathrm{S_{yy}} = \sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathrm{y_i}-\bar{\mathrm{y}})^2 .$$

EXAMPLES

Example: 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, $\bar{x} = 88.34, \bar{y} = 305.58,$ S_{xx}=7239.22, S_{xy}=17840.1, S_{yy}=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=S_{xy}/S_{xx}=17840.1/7239.22=2.464 and

EXAMPLES

$$a = \bar{y} - b\bar{x} = 305.58 - \frac{17840.1}{7239.22} \times 88.34 = 87.88$$

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

$$a = 88.34 - 0.2664 \times 88.34 = 6.944$$

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 $(\bar{x}, \bar{y}) = (88.34, 305.58)$.

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

SUMMARY

In this session,

1. Define Regression analysis and how it is related with correlation discussed
2. Differentiate the linear and nonlinear regressions.
3. Method of least squares in determining the coefficient have described

SELF-ASSESSMENT QUESTIONS

1. In regression analysis, the variable that is being predicted is the

a) response, or dependent, variable
b) independent variable …
c) intervening variable
d) is usually x

In regression, the equation that describes how the response variable (y) is related to the explanatory variable (x) is:

a) the correlation model
b) the regression model
c) used to compute the correlation coefficient
d) None of these alternatives is correct.

TERMINAL QUESTIONS

1. Describe the linear and non linear regression

2. List out the properties of regression coefficients

3. Analyze the regression analysis and its importance in practical experiment

4. In the accompanying table, x is the tensile force applied to a steel specimen in thousands of pounds, and y is the resulting elongation in thousandths of an inch:

X:	1	2	3	4	5	6
Y:	14	33	40	63	76	85

a) Graph the data to verify that it is reasonable to assume that the regression of Y on x is linear.

b) Find the equation of the least squares line, and use it to predict the elongation when the tensile force is 3.5 thousand pounds.

5) A professor in the school of business in a university polled a dozen colleagues about the number of professional meetings professors attended in the past five years (x) and the number of papers submitted by those to refereed journals (y) during the same period. The summary data are given as follows:

n=12, $\bar{x} = 4, \bar{y} = 12, \sum_{i=1}^{n} x_i^2 = 232, \sum_{i=1}^{n} x_i y_i = 318.$

Fit a straight line to the given data.

REFERENCES FOR FURTHER LEARNING OF THE SESSION

Reference Books:

1. Chapter 1 of TP1:William Feller, An Introduction to Probability Theory and Its Applications: Volume 1, Third Edition, 1968 by John Wiley & Sons,Inc.

2. Richard A Johnson, Miller& Freund's Probability and statistics for Engineers, PHI, New Delhi, 11th Edition (2011).

Sites and Web links:

1. https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/

2.https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/regression-library/v/introduction-to-residuals-and-least-squares

3. https://nptel.ac.in/courses/105105150/24

THANK YOU

Team – PSQT EVEN SEMESTER 2024-25