⑦ Assignment 3 II

Date : / /20

Title :- Correlation & Linear Regression in R

Problem statement - use of R for correlation & regression analysis.

Theory :-

Linear regression -

In data analysis we come accross the term "Regression" very frequently. Regression is a statistical way to estabilish a relationship bet" a dependent variable & a set of independent variable (s) e.g if we say Age = 5+ height * 10 + weight * 13.

Simple Linear Regression :-

Linear Regression is a statical method to regress having continuous values whereas independent variables can have their continuous or categorical values. In other words "Linear Regression" is a method to predict dependent Variable (Y) base on values of independent variable.

E.g - predicting traffic in retail store, predicting users dwell time or number of pages visited.

Prerequisties -

To start with linear Regression, few basic concepts are required.

Date : / /20

Give only 10 data points to fit a line our prediction are not pretty accurate but if we see correlation betⁿ 'y-actual' & 'y-Predicted' it will turn out very high.

Linear Regression in R using lm() function
It is easiest way to find regression using lm() the syntax is
Int (formula, data)
- formula is a symbol presenting relation betⁿ x & y.
- data is the vector on which formula will applied.

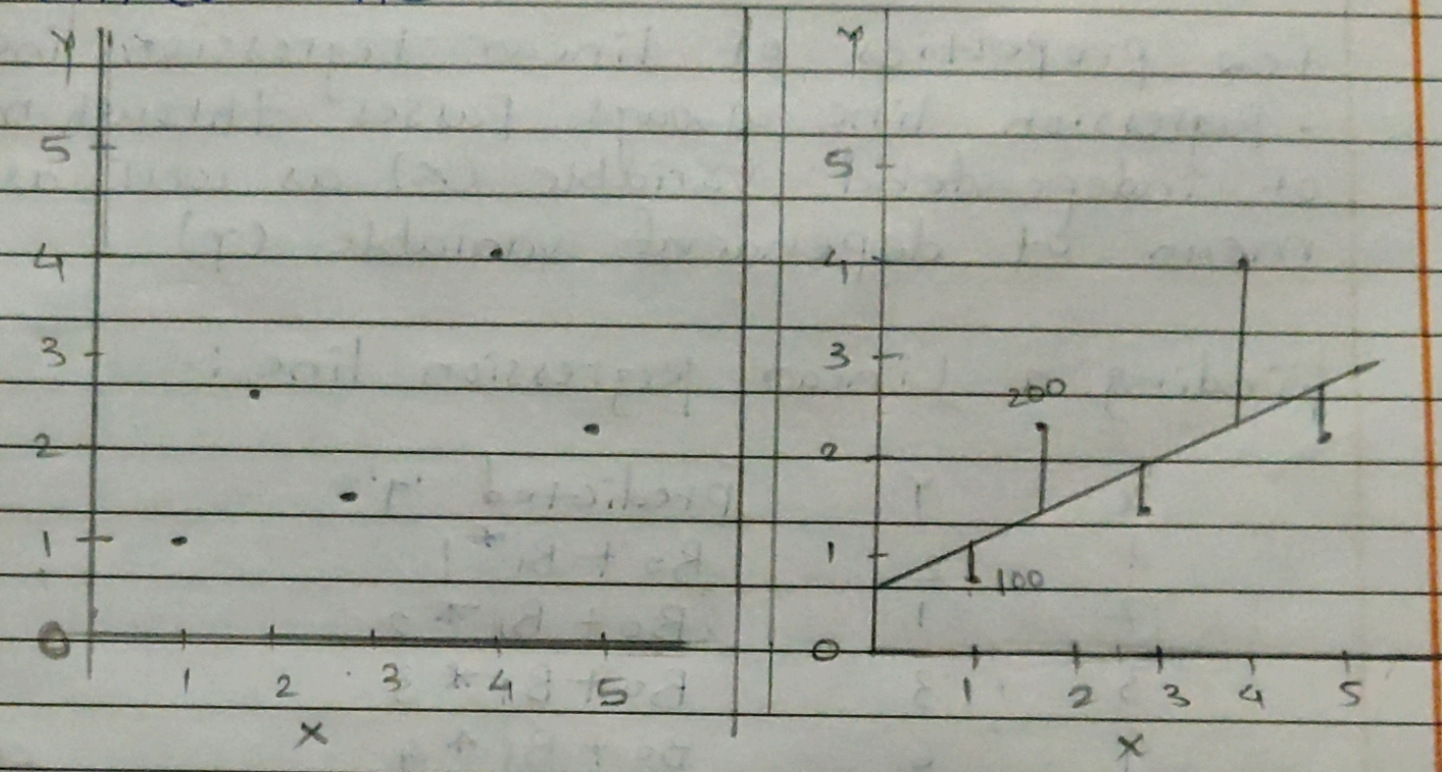
* predict() function :-
The basic syntax for predict() in linear regression predict (object, newdata)
- object is formula which is already created using the lm() function.
- newdata is formula which is containing new value for predictor variable.
value found using lm() function.

Multiple Regression -
multiple regression is an extension of linear regression into relationship betⁿ more than two variables. In Simple linear relation we have one predictor & one

Date : / / 20

Linear regression line -

While doing linear regression our objective is to fit a line through the distribution which is near set to most of the points. Hence reducing the distance (error term) of data points fitted line.



For example in above figure (left) dots represent various data points & line (right) represent an approximate line which can explain relationship betⁿ 'x' & 'y' arces. Through linear regression we try to find out search a line. for. eg if we have re dependent variable 'y' & an independent variable 'x'. relation betⁿ 'x' & 'y' can be represented in a form of following.

Date : / / 20

$Y = B_0 + B_1 X$

Where Y = Dependent variable
X = Independent variable.
B_0 = constant term / Intercept
B_1 = coefficient of relationship bet'n 'X' & 'Y'

few properties of linear regression line :-

- Regression line always passes through mean of independend varidble (x) as well as mean of dependent variable (y)

Finding a Linear Regression line :-

x	Y	predicted 'y'
1	2	$B_0 + B_1 * 1$
2	1	$B_0 + B_1 * 2$
3	3	$B_0 + B_1 * 3$
4	6	$B_0 + B_1 * 4$
5	9	$B_0 + B_1 * 5$
6	11	$B_0 + B_1 * 6$
7	13	$B_0 + B_1 * 7$
8	15	$B_0 + B_1 * 8$
9	17	$B_0 + B_1 * 9$
10	20	$B_0 + B_1 * 10$

Date : / /20

Table 1 :

Std. Dev of x	3.02765
Std. Dev of y	6.6137317
Mean of x	5.5
Mean of y	9.7
correlation bet^n x & y	.989938

If we differentiate the residual sum of square (RSS) wrt B_0 & B_1 & equate results into zero

B_1 = correlation * (std. Dev of y / std. dev. of x)

B_0 = Mean (y) − B_1 * mean(x)

Putting values from table 1

$B_1 = 2.64$

$B_0 = -2.2$

∴ least regression eq^n will be.

$Y = -2.2 + 2.64 * X$

now our prediction are looking like equation

X	Y-actual	Y-predicted.
1	2	0.44
2	1	3.08
3	3	5.72
4	6	8.36
5	9	11
6	11	13.64
7	13	16.28
8	15	18.92
9	17	21.56
10	20	24.2

Date : / / 20

- Correlation (r) - Explain relationship between two variables. Possible values -1 to +1
- variance (σ^2) - Measure of spread in your data
- Standard deviation (σ) - Measure of spread in your data (square root of variance)
- Normal distribution.
- Residual (error term) - { Actual value - predicted val}

Assumption of Linear Regression.

Not a single size fits for all, the same is true for linear regression as well as misleading

i) Lincarity & Additive :- There should be linear relationship between dependent & independent variable. should have additive impact on dependent variable.

ii) Normality of error distribution. Distribution or differences between Actual & predicted values.

iii) Homoscedasticity : variance of errors should be constant versus

a) Time.

b) The predictions

c) Independent variable values.

iv) Statical independence of errors : The error terms should not have any correlation among themselves.

Pooja

Date : / /20

one response variable.
The general mathematical eqn for multiple regression

$$y = a + b1\,X1 + b2\,X2$$

- y is response variable
- a, b1, b2 ... bn are coeficients.

Create Equation for Regression model -
Based on the above intercept & coeficient values we create the mathematical eqn.

Logistic Regression :-
The logistic regression is a regression model in which response variable has categorical values such as True / false or 0 or 1. It actualy measure the probability of a binary response as value of based on mathematical eqn.
The general mathematical eqn for logistic regression is -

$$y = 1/(1 + e^{-(a + b2\,x1 + b2\,X2 + b3\,X3 + \cdots)})$$

following is description of parameter used.
- y is response variable
- x is predictor variable.
- a & b are the coefficient which are numeric constants.

Pooja

Date : / / 20

The basic syntax for glm() function is logistic regression glm (formula, data, family)

- formula is symbol presenting relationship bet^n the variables
- data is data set giving the values of these variables.
- family is R object to specify the details of model It's value is binomial for logistic regression.

Conclusion :- Thus exercised various commands related to linear regression in R.