Assignment 8 II

Date : / /20

Title - Correlation and Linear Regression in R

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

Prelab - A basic understanding of correlation & regression concepts

Theory :-

Linear Regression -

In data analytics we come across the term "Regression" very frequently. Regression is a statistical way to establish a relationship between a dependent variable & a set of independent variable (s) e.g if we say that,

Age = 5 + Height * 10 + Weight * 18.

Simple Linear Regression -

"Linear Regression" is a stastical 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) based on values of independent variable.

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

Prerequisites -

To start with Linear Regression, few basic concepts are required -

Pooja

Date : / /20

- Correlation (r) - Explain the 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 value}

Assumption of Linear Regression -

Not a single size fits for all, the same is true for Linear Regression as well as misleading

i) Linearity & Additive : there should be linear relationship between dependent & independent variable & the impact of change in independent variable should have additive impact on dependent variable.

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

iii) Homoscedasticity : variance of errors should be constant versus
   a) Time
   b) The predictions
   c) Independent variable values.

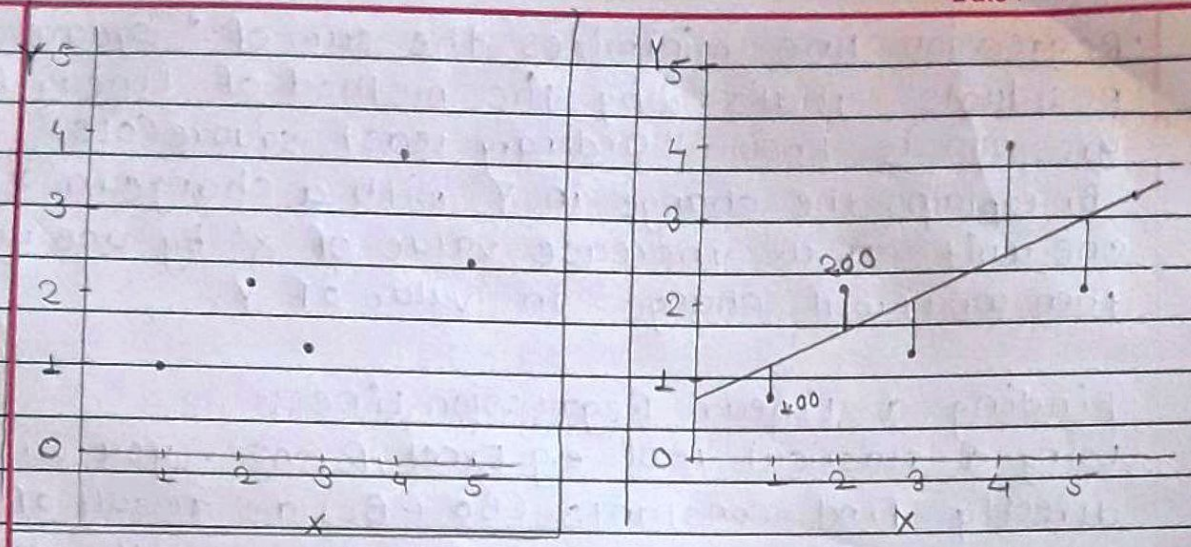
iv) stastical independence of errors : the error terms should not have any correlation among themselves

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

Pooja

Date : / / 20



For examples in above figure (left) dots represent various data points & line (right) represent an approximate line which can explain relationship betw' 'x' & 'y' axes. Through, linear regression we try to find out search a line. For e.g if we have re dependent variable 'Y' and one independent variable 'X' - relation between 'x' & 'Y' can be represented in a form of foll'.

$$Y = B_0 + B_1 X$$

where
- Y = Dependent variable
- X = Independent variable.
- B_0 = Constant term / Intercept
- B_1 = Coefficient of relationship bet' 'x' & 'Y'

Few Properties of linear regression line :-

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

Pooja

Date : / / 20

- Regression line minimizes the sum of "square of Residuals". That's why the method of Linear Regression is known "Ordinary Least square (OLS)".
- B_1 explains the change in Y with a change in x by one unit. If we increase value of 'x' by one unit then what will change in value of Y.

Finding a Linear Regression Line :-
Using a stastical tool e.g Excel, R, SAS. you will directly find constants (B_0 & B_L) as result of linear regression function. But conceptually as discussed it works on OLS concept & tries to reduce square root of errors, using very concept software package.

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

Pooja

Date : / / 20

Table 1 :

Std. Dev of x	3.02765
Std. Dev. of y	6.613747
Mean of x	5.5
Mean of y	9.7
Corelation bet^n x & y	.989938

If we differentiate the Residual sum of square (RSS) wrt B_0 & B_1 & equate results in to 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 equation will become

$Y = -2.2 + 2.64 * x$

how 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

Give only 10 data points to fit a line our prediction are not pretty accurate but if see correlation bet^n 'Y-Actual' & 'Y-Predicted' it will turn out very high

Pooja

Date : / / 20

Linear Regression in R using lm() function.
It is easiest way to find regression using lm()
The syntax is
lm(formula, data)
- formula is a symbol presenting the relation between 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 the vector containing the new value for predictor variable

This function will be used to predict the new value of dependent variable using the newdataset & values found using lm() function.

Multiple Regression.
Multiple regression is an extension of linear regression into relationship between more than two variables. In simple linear relation we have one predictor & one response variable, but in multiple regression we have more than one predictor variable & one response variable
The general mathematical eq^n for multiple regression

$$y = a + b_1x_1 + b_2x_2$$

- y is response variable.
- a, b_1, b_2 ... b_n are coefficients.

Pooja

Date : / /20

We create the regression model using the lm() function in R. The model determines the value of coefficients using the data. We can predict the value of response for given set of predictor variables using these coefficient.

The lm() function create the relation model between the predictor & the response variable.
The basic syntax for lm() function in multiple regression is -

lm (y ~ x1 + x2 + x3 ...., data)

- Formula is a symbol presenting the relation bet^n the response variable & predictor.
- data is the vector on which the formula will be applied.

Create Equation for Regression Model
Based on the above intercept & coefficient values, we create the mathematical equation.

Apply Equation for predicting New values -
We can use regression equation created above to predict the new values of dependent variables for the given set of independent variables.

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 actually measure the probability of a binary response as value of

Pooja

Date : / / 20

based on mathematical equation.
The general mathematical eq^n for logistic regression is -

$y = 1 / (1 + e^{-(a + b1x1 + b2x2 + b3x3 + \ldots)})$

Foll^g is description of parameter used -

- y is the response variable.
- x is the predictor variable.
- a & b are the coefficient which are numeric constants

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

- Formula is symbol presenting the relationship between 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.

Post-Lab :- Students will able to find relation bet^n dependent & independent variables using training dataset & can predict value for new dataset given.

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

Pooja