Title : Correlation & Linear Regression R.

Problem statement : Use of R for Correlation & Regression analysis.

Pre-Lab : A basic understanding of the Correlation & regression concepts is required

Theory :

Linear Regression :

In data analytics we come across the term "Regression" very frequently. Regression is a statical way to establish a relationship between a dependent variable & a set of independent variable(s). eg. if we say that Age = 5 + Height * 10 + Weight * 13

Here we establishing a relationship between Height & Weight of a person with his/her Age.

Simple Linear Regression

"Linear Regression" is a statistical method to regress the data with dependent variable having continuous values whereas independent variables can have either continuous or categorical values. In other words "Linear Regression" is a method to predict dependent variable (Y) based on values of independent variable

EXPT NO | NAME

Page No.

Prerequisites

To start with Linear Regression, few basic concepts of statistics are required:

- Correlation (r) – Explains 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}

Assumptions of Linear Regression:

Not a single size fits or all, the same is true for Linear Regression as well. In order to fit a linear regression line data should satisfy few basic but important assumptions. If your data doesn't follow the assumptions, your results may be wrong as well as misleading.

i. Linearity & Additive: There should be a linear relationship between dependent & independent variables & the impact of change in independent variable value should have establish additive impact on dependent variable.

ii. Normality of error distribution: Distribution of differences between Actual & predicted values (Residuals) should be normally distributed.

Teacher's Signature

EXPT. NO. NAME

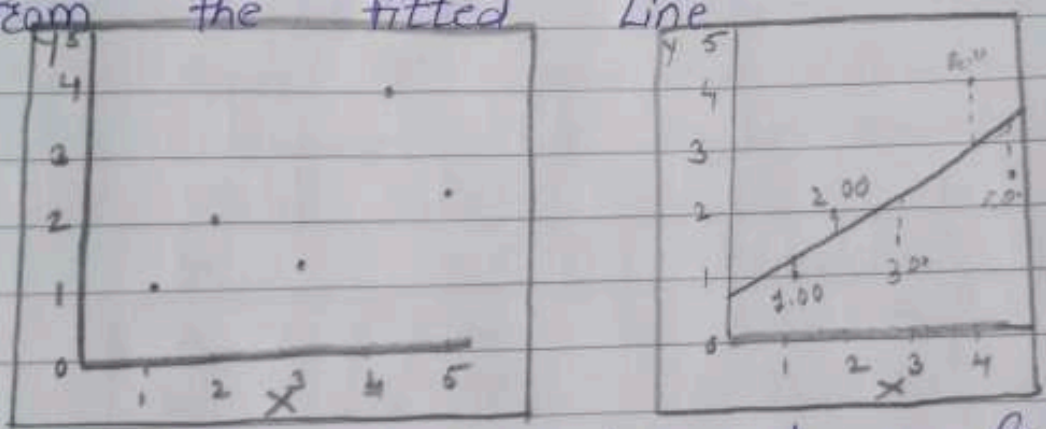
Page No.

iii. Homoscedasticity: Variance of errors shold be constant versus, a. Time
b. The predictions
c. Independent variable values.

iv. Statistical independent 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 nearest to most of the points. Hence reducing the distance of data points from the fitted Line



For example, in above figure dots represent various data points & line represents an approximate line which can explain the relationship between 'x' & 'y' axes. If we have one dependent variable 'y' & one independent variable 'x' - Relationship between 'x' & 'y' can be represented in a form of following equation: $Y = B_0 + B_1 X$

Teacher's Signature:

EXPT. NO. NAME:

Page No.:

Few properties of linear regression line

- Regression line always passes through mean of independent variable (x) as well as mean of dependent variable (y)
- Regression line minimizes the sum of "Square of Residuals". That's why the method of Linear Regression is known as "Ordinary least Square (OLS).

Finding Linear Regression Line.

Using a stastical tool e.g., Excel, R, SAS you will directly find constants as a result of linear Regression function.
For example, let us we want to predict 'y' from 'x' given in following table & let's assume that our regression equation will look like "$y = B_0 + B1 * x$"

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	$B_0 + B_1 * 10$

Teacher's Signature:

Table 1 :

Std. Dev. of x	3.02765
Std. Dev. of y	6.617317
Mean of x	5.5
Mean of y	9.7
Correlation between x & y	.989938

If we differentiate the Residual Sum of Square (RSS) wrt. B_0 & B_1 & equate the results to Zero, we get the following equations as a result :

B_1 = Correlation * (Std. Dev. of y / Std. Dev of x)

B_0 = Mean (Y) - B_1 * Mean (X)

Putting values from table 1 into the above eqution

$B_1 = 2.64$ $B_0 = -2.2$

Hence, the least Regression equation will become — $y = -2.2 + 2.64 * x$

Let See, how our prediction are looking like using this 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

Teacher's Signature

Linear Regression in R using lm() function:
It is the easiest way to find regression using lm() function.
The syntax is:
lm(formula, data)
Following is the description of the parameters used -
- formula is a symbol presenting the relationship between x & y
- data is vector on which the formula will be applied

Predict function:
The syntax is:
predict (object, newdata)
Following is the description of the parameters used -
- Object is the formula which is already created using the lm() function
- newdata is the vector containg the new value for predictor variable.

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

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

Logistic Regression:

The Logistic Regression is Regression Model in which the response variable (dependent variable) has categorised value such as True/false or 0/1.

The general mathematical equation for logistic regression is -

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

glm( ) function.

The basic syntax for glm( ) function in logistic regression is -

glm(formula, data, family).

Post-Lab: Students will be able to find relation between dependent & independent variables using training dataset and can predict values for the new dataset given

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