DSDL

Assignment NO:- 07

Title:-

Correlation & linear Regression in R.

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

Pre-Lab:- A basic understanding of the correlation & regression concept is required.

Theory:-

Linear Regression:-

In data analysis we come across the term "Regression" very frequently. e.g. if we say that.

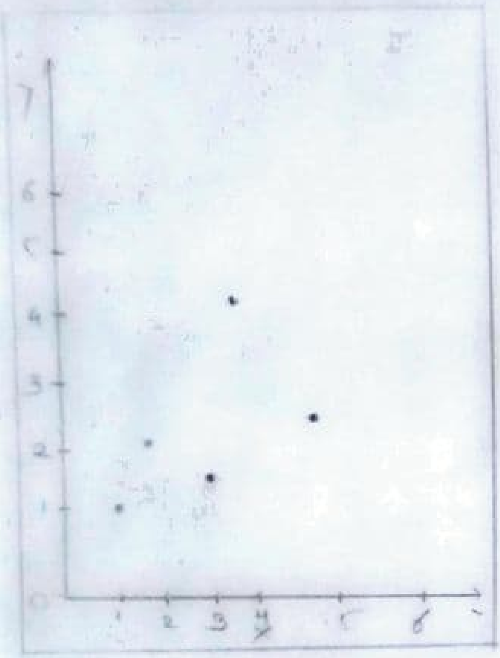
Age = 5 + Height * 10 + Weight * 13

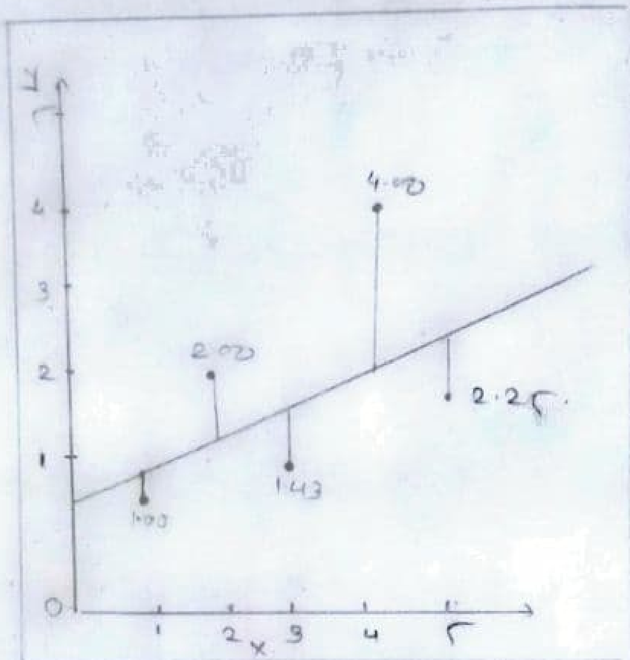
Simple Linear Regression:-

"Linear Regression" is a statistical method to regress the data with depend variable having contineous values. E.g. Predicting traffic in a retail store, predicting a user's dwell time or number of pages visited on Dezyre.com etc.

Prerequisites:-

- Correlation (r) - Explains the relationship betn two variables, posible 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}


x


y
4.00
2.00
1.43
1.00
2.25
x

Assumption of Linear Regression:—

i) Linearity & Additive :- There should be a linear relationship betn dependent & independent variables & the impact of change in independent variable values should have additive impact on dependent variable.

ii) Normality of error distribution :- Distribution of differences betn actual & predicted values (Residuals) should be normally distribution.

iii) Homoscedastisity :-

a) Time

b) The predictions.

c) Independent variable values.

iv) Statistical independance of errors :- The error terms (residuals) 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 (error term) of data points from the fitted line.

following equations.

$Y = B_0 + B_1 X$.

where,

Y = Dependent variable.

X = Independent variable.

B_0 = Constant term / Intercept.

B_1 = coefficient of relationship betn 'x' & 'y'

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".
- B_1 explains the change in Y with a change in x by one unit.

Finding a linear Regression Line:-

Using a statistical tool e.g., Excel, R, SAS etc. For example, let say we want to predict 'y' from 'x' given in following table & let's assume that our regression eqn will look like "$y = B_0 + B_1 * x$".

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	$B_0 + B_1 * 10$

where,
Table 1,

std. Dev of x	3.02765
std. Dev. of y	6617317
Mean of x	5.5
Mean of y	9.7
correlation between x & y	.989938.

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 eqns.

$B_1 = 2.64$

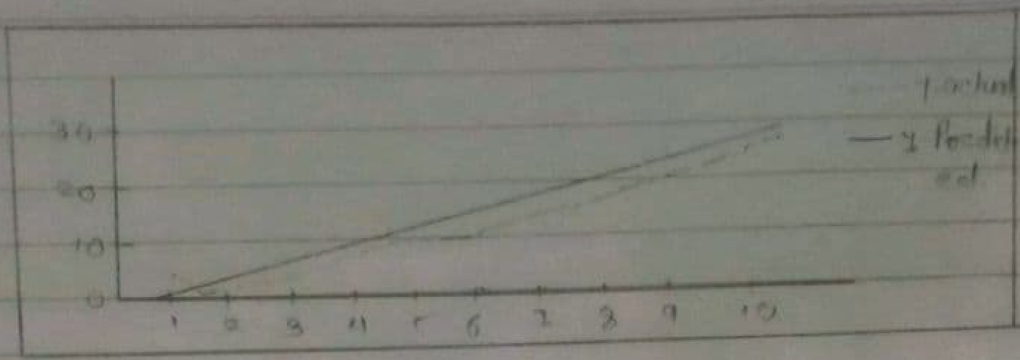
$B_2 = -2.2$.

Hence, the least regression eqn will become -

$Y = -2.2 + 2.64 * X$.

lets see,

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.



Linear Regression in R using lm() function :—
It is easiest way to find regression using lm() fun.
The syntax is.

lm (formula, data).

- formula is a symbol presenting the relⁿ betⁿ x & y.
- data is the vector on which the formula will be applied.

predict() function:
The basic syntax for predict() in in linear regression is -

predict(object, newdata).

- object is formula which is already created using lm() fun.
- newdata is the vector containing the new data value for predictor variable.

Multiple Regression :-
The general mathematical eqⁿ for multiple regression is —

$y = a + b_1x_1 + b_2x_2 + \dots b_nx_n.$

following is discription of parameters used -

- y is the response variable.
- $a, b_1, b_2, \dots$ the the coefficients.
- $x_1, x_2, \dots x_n$ are pridictor variables.

we create the regression model using the lm() function in R.
The ml() funⁿ creates the relationship model betⁿ the predictor & the response variable.
The basic syntax for lm() fun in multiple regression is-
lm(y~x1+x2+x3.....,data).
- formula is a symbol presenting the relan betn the response variable & predictor variables.
- 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 eqn.

Apply Equation for predicting New values:-
we can use the regression eqn created above to predict the new value of dependent variable for the given set of independent variables.

Logistic Regression:-
The general mathematical eqn for logistic regression is
$y = 1/(1+e^{-(a+b_1x_1+b_2x_2+b_3x_3+\cdots)})$
- y is a responce variable.
- x is a predictor variable.
- a & b are the coefficient which are numeric constant
The fun used to create the regression model is a glm fun.

The basic syntax for glm() funⁿ in logistic regression is -
glm(formula, data, family)
Following is the description of the parameters used -
- formula is the symbol presenting the relationship betⁿ the variables.
- data is the data set giving the values of these variables.
- family is R object to specify the details of the model. It's value is binomial for logistic regression

Post-Lab:-
students will be able to find relation betⁿ dependent & independent variables using training dataset & can predict values for the test dataset given.

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