Statistical Methods in the Usage of Correlation and Regression of the Machine Learning Models

V. Manikyala Rao, K. Uma Pavan Kumar, B. Priyanka

Abstract— The predictive analytics is the most commonly used methodology in the usage of Machine Learning class of algorithms. Based on the values generated at the time of running the algorithm the significance of the model can be estimated. The current work gives a complete focus on P value and the significance levels of the P value in the correlation analysis of the algorithms. Based on the P value the impact of the model can be notified and the interpretation of the results can be done in the efficient way. The other dimension of the work is the usage of statistical functionalities in the regression analysis, most of the researchers are focusing on the shallow usage of regression analysis in the classification of the tasks. The current work explains the complete internals of the regression models available and the usage of the statistical functionalities utilized in the implementation of the corresponding variants of the algorithms.

We believe that the current work exclusively helps the upcoming researcher in the areas of regression in the context of the statistical functionalities which are vital in the implementation of the tasks. The outcome of the work is to exploit the correlation analysis with various significance levels and the issues in the processing of the analysis. The another point here is the regression internals with the focus of statistical methods available in the processing of regression variants. The regression analysis involves various types like linear regression, multiple regressions and logistic regression. The current work gives an overview of all these three types of regressions and also the significance of P value in the prediction of outcome. In the examples such as house rate prediction based on the given area, salary of an employee based on the experience level, profit of the start-up companies based on the spending on research, admin marketing and state of the country are best suitable in the explanation of regression.

Keywords: Regression, statistical methods, P-Value, Correlation.

I. INTRODUCTION

The regression is very much helpful to study the relationship among the variables so as to predict the best approximations. There exist various regression techniques such as Linear Regression, Multiple Regression and Logistic Regression. In statistical modeling simple linear regression, a form of relationship between two variables, the basic formula is [1]

$$Y=A+BX+U \quad \text{----------(1)}$$

Here dependent variable is Y which we have to predict the value. Here independent variable is X which we are using for predicting Y. If the independent variable is changing then accordingly is there any change exist with dependent variable, which can be observed with regression. The multiple regression handles multiple independent variables with a dependent variable. [2][3]

$$Y=A_0+A_1X_{i1}+A_2X_{i2}+\text{------}A_pX_{ip}+€ \quad \text{----------(2)}$$

The linear regression can be used if dependent and independent values were continuous in nature. In case the independent variables are not having high correlation then the suitable type of regression is multiple regression. Here € is the error term which represents the context such that if the model could not represent the actual relationship between dependent and independent variables. [4][5]

There are certain conditions that the dependent variable is binary and the independent variable might be nominal, ordinal or interval in nature. The scenarios like rate of heart attack based on the weight and calories intake. The logistic regression predicts the probability of occurrence of an event by fitting a model. [6]

The logistic regression general equation is

$$P=e^{\wedge}Y/1+e^{\wedge}Y \quad \text{----------(3)}$$

P is the probability and (3) is Logic Function.

$$\text{Log } (p/1-p)=Y \quad \text{----------(4)}$$

II. IMPLEMENTATION OF LINEAR REGRESSION WITH SALARY.CSV DATA SET

The data set contains the information about the salaries, the model implements the relation between salary and the experience of the employees in the data. To implement the model, caTools package has been used in R, and lm() linear model is used to establish a relationship between Salary and Experience.

Revised Version Manuscript Received on 30 May, 2018.

V. Manikyala Rao, Assistant professor, CSE, Malla Reddy Institute of Technology, Telangana, India.
(Email: manikyalarao.v92@gmail.com)
Dr. K. Uma Pavan Kumar, Associate Professor, CSE, , Malla Reddy Institute of technology, Telangana, India.
(Email: Dr.kethavarapu@gmail.com)
B. Priyanka, Assistant professor, CSE, Malla Reddy Institute of Technology, Telangana, India.
(Email: Priyanka.biradar511@gmail.com)

```
dataset = read.csv('Salary_Data.csv')

# Division dataset into the Training set and Test set
 install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$Salary, SplitRatio = 2/3)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
# Fitting Simple Linear Regression to the Training set
regressor = lm(formula = Salary ~ YearsExperience,
         data = training_set)

y_pred = predict(regressor, newdata = test_set)
# Visualizing the Training set results
install.packages("ggplot2")
library(ggplot2)
ggplot() +
 geom_point(aes(x = training_set$YearsExperience, y = training_set$Salary),
        colour = 'red') +
 geom_line(aes(x = training_set$YearsExperience, y = predict(regressor, newdata = training_set)),
        colour = 'blue') +
 ggtitle('Salary vs Experience (Training set)') +
 xlab('Years of experience') +
 ylab('Salary')

# Visualizing the Test set results
library(ggplot2)
ggplot() +

 geom_point(aes(x = test_set$YearsExperience, y = test_set$Salary),
        colour = 'red') +
 geom_line(aes(x = training_set$YearsExperience, y = predict(regressor, newdata = training_set)),
        colour = 'blue') +
 ggtitle('Salary vs Experience (Test set)') +
 xlab('Years of experience') +
 ylab('Salary')
```

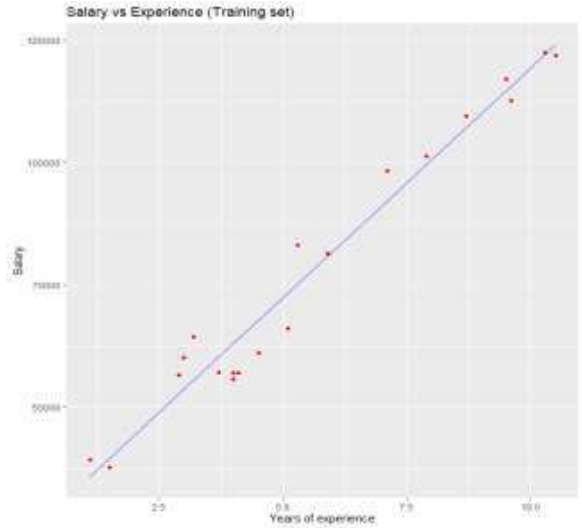

Figure 1: Linear Regression Test Set Results

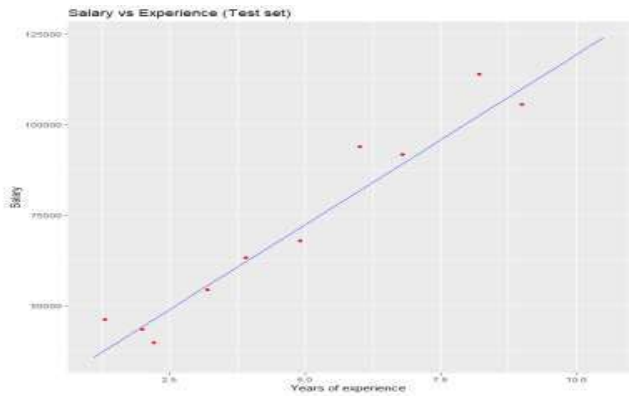

Figure 2: Linear Regression Training Set Results

III. IMPLEMENTATION OF MULTIPLE REGRESSION WITH STARTUP DATASET

The data set considered to implement the multiple linear regressions is start up data which gives the information about the expenditure on various parameters such as R and D, Admin and Marketing and to predict the state information on profit based.

```
# Multiple Linear  Regression
dataset = read.csv('50_Startups.csv')
dataset
dataset$State = factor(dataset$State,

                levels = c('New York', 'California', 'Florida'),
                labels = c(1, 2, 3))
# Splitting the dataset into the Training set and Test set
# install.packages('caTools')
library(caTools)
set.seed(123)
split = sample.split(dataset$Profit, SplitRatio = 0.8)
training_set = subset(dataset, split == TRUE)
test_set = subset(dataset, split == FALSE)
# Fitting Multiple Linear Regression to the Training set
regressor = lm(formula = Profit ~ ., data = training_set)
# Predicting the Test set results
y_pred = predict(regressor, newdata = test_set)
summary(y_pred)
plot(y_pred)
```

Minimum 1st Qtr. Median Mean 3rd Qtr. Maximum. 96867 111541 126298 132621 156702 173981

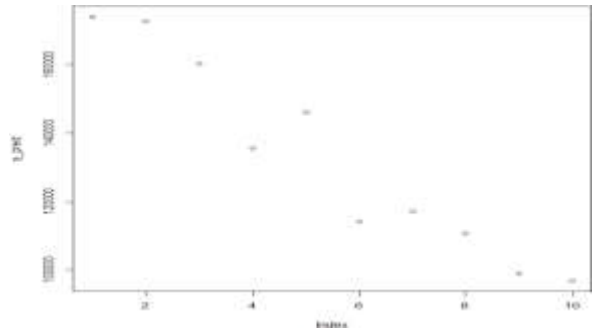

Figure 3: Multiple Regression Outcome

IV. ISSUES USAGE OF P- VALUE IN THE PREDICTIONS & RESULTS

The significance of the hypothesis indicates p-value; there is no effect with null Hypothesis if the coefficient value is zero. A significance of the hypothesis indicates a low p-value, so if there is any changes in the independent variable it will effects the dependent variable and if any changes are there in the predictor's value not associated with the response variable. [7]

Residual standard error: 4.712 on 8 degrees of freedom Multiple R-squared: 0.9548, Adjusted R-squared: 0.9491

F-statistic: 168.9 on 1 and 8 DF, p-value: 1.164e-06

The above outcome gives the various values along with the p-value as the value is low so there is a significant relation between predictor's value and response variable.

V. CONCLUSION

The work describes the usage of regression and the corresponding models available such as linear, multiple and logistic regression. The implementation of the linear and multiple regressions in various data sets and the corresponding plots were presented. In the application of machine learning especially in case of establishing a kind of the relationship between the dependent and independent variables we can make use of regression analysis. The outcome of the work is to present the regression basics and implementation of the regression with R programming and the presentation of the results.

REFERENCES

1. Qiu qiufen et al. Journal of Hunan Institute of Humanities, Sciences and Technology. 2013(2):102-105.
2. Abrams , L. M. (2004). Teachers' views on high-stakes testing: Implications for the classroom. Tempe, AZ: Arizona State University, Education Policy Research Unit.
3. Roles and relationships: School board and superintendents. (Report No. EA 025907). Arlington, VA. (ERIC Document Reproduction Service No. ED37465). American Association of School Adminstrators (ASSA). (2007). (1 993). Pp. 1 1-1 2.
4. A nation at risk. (1 983). The imperative for educational rejbrrn. Vol. 26 issue 7 July 1983. Acm, New York, NY USA.
5. Brown, R. (2011). Research note: Estimates of college football player rents. Journal of Sports Economics, 12(2), 200-212.
6. Tunaru, R., Clark, E., & Viney, H. (2005). An option pricing framework for valuation of football players. Review of financial economics, 14(3-4), 281- 295.