IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

PREDICTIVE DATA ANALYTICS USING R LANGUAGE

Dr.R.Jamuna, Associate Professor in Computer Science, S.R.College, Trichy-2,

1. Introduction

Data analysis is the process of analyzing data to increase the inferences and make accurate predictions to promote growth in any field. It involves steps like data cleansing, data warehousing, transformation, inspection, and mathematical modeling to perform statistical analysis .It could gather hidden data insights and generate reports based on the analysis. Today many data analyzing tools like Python, Tableau, Power BI, R language, Apache Spark , Rapid Miner Studio, KNIME,SAP Predictive Analysis etc are available. In this paper R language is chosen for analysis and predictions. Statistical technique called predictive analysis or predictive analytics makes use of machine learning and neural network algorithms to find patterns in the data and forecasts future actions. It is now better to go beyond descriptive analytics in order to learn whether training initiatives are effective and how they may be enhanced. Data from the past as well as the present can be used in predictive analysis to make predictions about what might occur in the future. Here the data taken is the working hours of scholars to reach out the scores of the examinations as goal is taken as trial data.

2. Predictive data analysis

Predictive analytics is also referred to as predictive analysis which is a subset of data analysis that focuses on creating future predictions from existing trial data which is crucial in making decisions in all fields. In any case, predictive analysis typically uses a variety of statistical models, techniques, and tools that helps in understanding the patterns in datasets and making predictions. To predict future trends, these models evaluate previous and present data using algorithms and machine learning techniques. Prediction is a vital component of data mining. Predictive analysis is a method for forecasting future patterns from current or historical data. It can take many different forms, but some of the most advanced models make use of machine learning and artificial intelligence [1] algorithms. Predictive analysis encompasses several different types of data analysis models. Most of these are regression models, which aim to determine the connections between two or more variables. They can help in predicting the value of an unknown variable as the value of a known variable changes by recognizing the relationship between these variables. Many models exist for such predictions such as

- *Regression techniques:* Help in deciphering the connections between variables.
- *Decision trees:* Use branching to illustrate the potential outcomes of each option or course of action.
- *Neural networks:* Make use of algorithms to discover potential connections between data sets.

Usually Generalized Linear Model - Linear Regression can be taken for analysis. The linear regression model [2] is the most basic predictive analysis approach. In this approach, it is presumed that an unknown variable value will scale linearly with a known variable value. Linear regression R consists of two variables that are related by an equation where exponent of both variables equals one, which forms a mathematical straight line. In nonlinear relationships the exponent is not one that leads to a graph. Here in linear regressions y=ax+b is an equation where y is the response variable and x is the predictor variable where a and b are constants or coefficients. Mathematically a linear relationship represents a straight line when plotted as a graph. In case the variables are nonlinear then the exponents of the variables are not equal to one and it creates a curve on the graph. Correlation measures the strength of a linear relationship between two variables, whereas regression describes the relationship as an equation.

3. Method of Analysis

Lm () function of R with which we can

- Create a relationship model using the **lm()** functions in R.
- Find the coefficients from the model created and create the mathematical equations.
- Get a summary of the relationship model to know the average error in prediction, which is also called **residuals**.
- To predict the weight of new persons, use the **predict()** function in R

```
lm(formula, data, subset, weights, na.action,
  method = "qr", model = TRUE, x = FALSE, y = FALSE, qr = TRUE,
singular.ok = TRUE, contrasts = NULL, offset, …)
```

Arguments of the lm() Function

formula: It represents relation between x and y.
It is an object of a class "formula" (or one that can be coerced to that class): a symbolic description of the model to be fitted.

data

It is an optional data frame, list or environment (or object coercible by as.data.frame to a data frame) containing the variables in the model. If not found in data, the variables are taken from environment(formula), typically the environment from which lm is called.

subset

It is an optional vector specifying a subset of observations to be used in the fitting process.

weights

It is an optional vector of weights to be used in the fitting process. It should be NULL or a numeric vector. If non-NULL, weighted least squares is used with weights (that is, minimizing sum(w*e^2)); otherwise ordinary least squares is used.

na.action

It is a function which indicates what should happen when the data contain NAs. The default is set by the na.action setting of options, and is na.fail if that is unset. Another possible value is NULL, no action. Value na.exclude can be useful.

method

It is the method to be used for fitting, currently only method = "qr" is supported; method = "model.frame" returns the model frame.

model, x, y, qr

these are logicals. If TRUE the corresponding components of the fit like model frame, the model matrix, the response and the QR decomposition are returned.

singular.ok

This is also a logical parameter. If FALSE then a singular fit is an error.

contrasts

It is an optional list offset used to specify prior known components that are to be included in the linear predictor.

offset

This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. This should be NULL or a numeric vector or matrix of extents matching those of the response. One or more offset terms can be included in the formula instead or as well, and if more than one are specified their sum is used.

Let us determine the relationship model between the predictor and response variables for a student dataset. The predictor value stores the **number of hours of study** put in by the students, where as the response vector stores the **entrance exam score** for getting into professional courses.

Consider the sample dataset given below:-

nohrs	entrancescore
2	55
2.5	62
3	65
3.5	70
4	77
4.5	82
5	75
5.5	83
6	85
6.5	88

Let us read the data from Data set into a Data Frame using read .table function of R language.

```
> Hs<-read.table("e:/data.txt",sep="",header=TRUE)
> Hs
```

Sl.no	nohrs	entrancescore
1	2	55
2	2.5	62
3	3	65
4	3.5	70
5	4	77
6	4.5	82
7	5	75
8	5.5	83
9	6	85
10	6.5	88

Use of summary() function produces statistical summaries. Here minimum, first quartile ,median,mean,third quartile ,maximum are obtained.

```
>summary (Hs)
Nohrs- entrancescore
 Min.   :2.000   Min.   :55.00
 1st Qu.:3.125   1st Qu.:66.25
 Median :4.250   Median :76.00
 Mean   :4.250   Mean   :74.20
 3rd Qu.:5.375   3rd Qu.:82.75
 Max.   :6.500   Max.   :88.00
```

Next we can check the internal structure of the Data Frame. It shows the ten observations of two variables

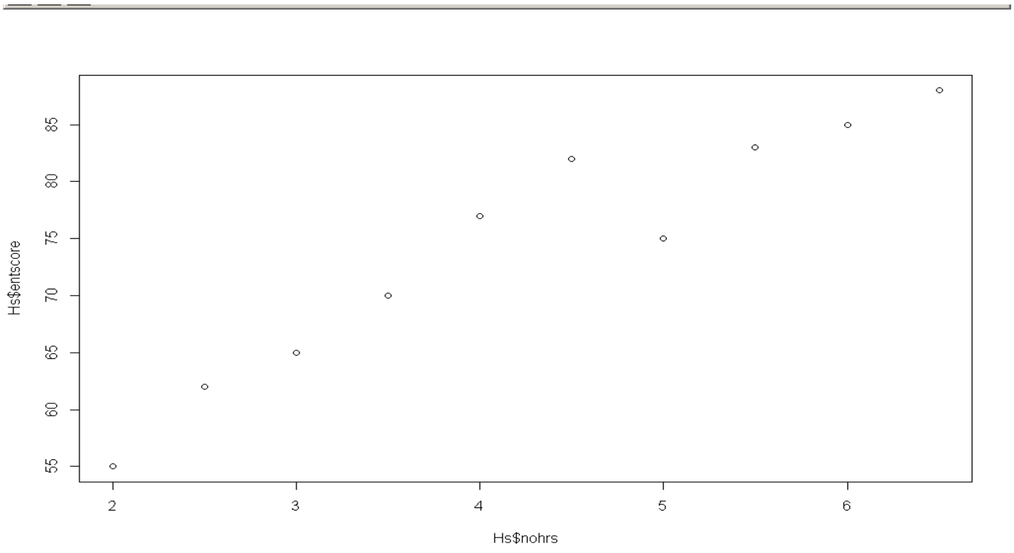
```
> str(Hs)
'data.frame':   10 observations of  2 variables:
```

nohrs	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5
entrancescore	55	62	65	70	77	82	75	83	85	88

```
 $ nohrs   : num  2 2.5 3 3.5 4 4.555.5 66.5
 $ entscore: int  55 62 65 70 77 82 75 83 85 88
```

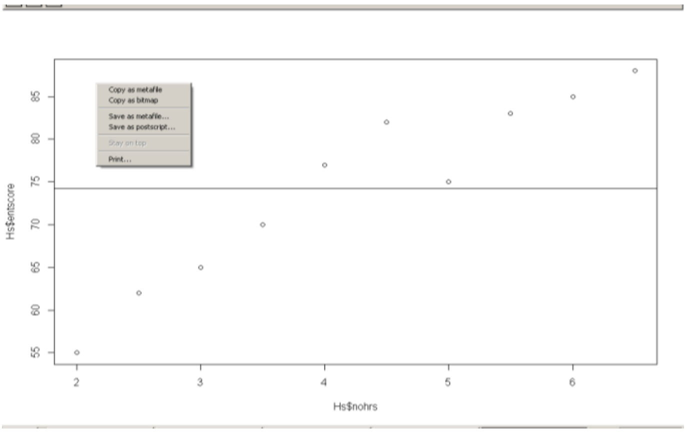
4. Statistical analysis: Plot and Analysis of R object

W ca plot the no of hours worked by the student in X axis with the entrance examination score on the Y axis[3] using the function >plot(Hs$nohrs,Hs$entraancescore).



To draw a horizontal line across the plot at the mean (mean for entrance is 74.20), we can use the abline function.

```
>abline(h=mean(Hs$entscore))
```



Correlation coefficient

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}$$

Solving for this given student data set ,

R=10*(3295.5)-(42.5)*(742)/square root of (10*201.25-1806.25]{10*56130-550564])

=0.947548805

=0.95

5. Results and Discussions

We can use the mean to predict the entrance score at some instances by finding the significant differences between the actual or the observed value and the predicted value .For example the first student has the entrance score of 55 and if we use the mean to predict the score, we would have predicted it as 74.20 and here the observed score is less than the predicted score . For the tenth sample student the observed score is 88 and the value is greater than the predicted value of 74.20.This indicates that we have to consider other factors as well. The factors like concentration, deep insight into the concepts, regular efforts, improving the problem solving aptitudes of the student also play an important and significant role in achieving the desired scores in any examination pattern by formulating multivariate analysis. In data mining [5] prediction plays a vital role where the degree of association and direction of association are key factors to determine the strong or weak association between variables. Multiple variables as stated above as other factors can also be analyzed to reach the desired goal. Next we can also use the cor () function in R language to determine degree of association and direction of association.

```
>cor(Hs$nohrs,Hs$entscore)
[1] 0.9542675
```

6. Conclusions

The correlation value here suggests that there is strong association between the number of hours studied and the entrance score. With more information, future outcomes can be predicted with relative accuracy. This makes it possible for organizations, educational institutes and businesses to make more accurate predictions and decisions to increase the production. Learning the methods of predictive analysis and applying the same has become essential for optimizing the goal for any field in data science. Data analytics has numerous applications and tools to promote the growth and progress of any field like technological, educational, business or industrial world. However there are few cautions to be observed while using correlation analysis. [6]

1. For non linear relationships correlation is not an appropriate measure of association.
2. To test if two variables are linearly related a scatter plot can be used.
3. Pearson coefficient can be affected by outliers. A box plot can be used to show the outliers. Spearman’s coefficient is preferred ,since it minimizes the outlier effect.
4. A correlation coefficient of 0 or nearly zero indicates that the variables are not linearly related but they are related by some way.
5. Correlation does not indicate causation that is one variable causing the other. But helps in finding the degree of association.
6. Since correlation coefficient is inappropriate in determining causation we can use regression techniques to quantify the nature of relationship between the variables.

References

[1**]** Abbott, D. (2014) Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst, New Jersey: John Wiley & Sons

[2] Bafna J. Predictive Analysis Using Linear Regression With SAS.Big Data Zone-DZone;2017

[3] Chambers, J. M. (1992) *Linear models.* Chapter 4 of *Statistical Models in S* eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.

[4] Data Mining Techniques: Algorithm, Methods & top Data Mining Tools. Software Testing Help; March 2020. Available from: https://www.softwaretestinghelp.com/data-mining-techniques/

[5] Han, J., Kamber, M., and Pei, J. (2011) Data Mining Concepts and Techniques (Third ed). Elsevier

[6]Wilkinson, G. N. and Rogers, C. E. (1973). Symbolic descriptions of factorial models for analysis of variance. *Applied Statistics*, **22**, 392--399. 10.2307/2346786.