



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A Study on the Academic Performance of the Students by Applying Multiple Linear Regression Analysis using the method of Least Squares

G. Narasinga Rao, Ch.Sreenu Babu

Computer Science Engineering Department,

GMR Institute of Technology,

GMR Nagar, Rajam, Srikakulam District, Andhra Pradesh, INDIA.

Abstract: Many Educational Institutions want their students to be placed in reputed companies and hence they focus on the academic performance of their students. If they could able to know about the student's academic performance, then they can focus more on those students whose academic performance is poor. They can take special care on those students whose academic performance is poor and make them prepare for the placement. In this paper, we present a study on the academic performance of the students of a reputed college by applying multiple linear regression models on the data set consisting of 23 students and evaluate the dependent variable for two given predictor variables.

Keywords: Data mining, Data Set, Data objects, Classification, Multiple linear regressions.

I. Introduction

Data Mining is the process of analyzing data available in large data sets and obtains knowledge out of it. The knowledge so obtained is known as the Knowledge Discovery in Databases (KDD). This knowledge helps us in different ways. A data set is a collection of data objects. Each data object represents a record. This data object consists of any number of attributes which specifies the characteristics of the data object. Any Educational Institution consists of students with different categories. We can categorize students based on several properties like attendance, percentage of marks secured, attitude, intelligence etc., to judge their academic performance. In this paper, we present a study on the academic performance of the students based on their aggregate percentage of marks, attendance percentage and number of backlogs. Hence we use multiple linear regression models on the data set of 23 students and evaluate the dependent variable for two given predictor variables.

II. Methodology[7]

A linear regression model that contains more than one predictor variable is called a multiple linear regression model. This is of the form $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

The above model is linear in the parameters $\beta_0, \beta_1, \beta_2$ that describes a plane in three dimensional space of Y, X1 and X2. β_0 is the intercept of the plane and β_1, β_2 are considered as partial regression coefficients.

Parameter β_1 represents the change in the mean response corresponding to a unit change in x_1 when x_2 is held constant. Parameter β_2 represents the change in the mean response corresponding to a unit change in x_2 when x_1 is held constant. The above model that we consider here is a first order multiple linear regression models. The true regression model is generally never known. It can be estimated by calculating the parameters of the model for an observed data set.

The first order regression model applicable to the data set having two predictor variables is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Where the dependent variable Y represents the aggregate percentage of marks and the predictor variables X1 and X2 represents the attendance% and number of backlogs.

The least square estimates, $\hat{\beta}$, can now be obtained:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\text{Thus } \hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$$

Then the estimated regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ are calculated. Thus the final regression model is $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

III. Results

We applied the multiple linear regression model on the data set of a reputed Institution consisting of 60 students by considering the aggregate percentage of marks (Y) secured and attendance % (X1), number of backlogs (X2) as attributes of each of the student data object. Then we identified those students with similar attribute values and reduced the total data set consisting of 23 students having unique attribute values. The reduced data set consisting of 23 students with their respective attribute values is shown below in Table 1.

Table 1

Serial No	Aggregate Percentage of Marks (y)	Attendance% (x1)	No. of backlogs (x2)
1	62	84.2	0
2	55.3	80.7	3
3	71.8	91.7	0
4	69.8	93	0
5	46.5	74.5	9
6	66.6	88.4	0
7	53.2	81.2	2
8	56.6	82.1	1
9	50.2	87.6	4
10	59.1	93	0
11	54.6	93.2	2
12	45.8	69.8	14
13	63.2	91.9	0
14	49.1	79.1	8
15	47.2	74.9	7
16	59.6	90.2	1
17	53	75.4	8
18	43.8	81.2	8
19	51.8	75.8	4
20	74.8	88.4	0
21	45.8	68.9	13
22	77.4	96.6	0
23	45.1	79.6	18

Thus the above student data set is expressed in the matrix form as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Now compute $\hat{\beta} = (X^T X)^{-1} X^T Y$

$$\text{Clearly } X^T X = \begin{bmatrix} 23 & 1921.4 & 102 \\ 1921.4 & 161941.92 & 7802.9 \\ 102 & 7802.9 & 1062 \end{bmatrix}$$

$$\text{And } X^T Y = \begin{bmatrix} 1302.3 \\ 4870.2 \\ 110173.37 \end{bmatrix}$$

$$\text{Also } (X^T X)^{-1} = \begin{bmatrix} 0.76 & -0.008 & -0.01 \\ -0.008 & 0.000096203 & 0.000113314 \\ -0.01 & 0.000113314 & 0.000225627 \end{bmatrix}$$

$$\text{Now } \hat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 59.65904 \\ 0.73247055691 \\ 0.56003386358 \end{bmatrix}$$

Hence the estimated regression coefficients are calculated and the final regression model is obtained as

$$Y = 59.65904 + 0.73247055691 X_1 + 0.56003386358 X_2$$

Now after the final results, we can easily evaluate the dependent variable Y if we are given the predictor variables X1 and X2.

IV. Conclusion

In this paper, we present a study on the academic performance of students by considering the aggregate percentage of marks secured by the students, attendance percentage and number of backlogs of the students as attributes of each of the student data object to monitor the performance of the students. We demonstrated this by applying multiple linear regressions on a student data set of a reputed institution for total number of 23 students.

After applying the multiple linear regression model, we can easily predict the dependent variable Y for two given predictor variables X1 and X2.

References

- [1] Larose, D. T., "Discovering Knowledge in Data : An Introduction to Data Mining", *Hoboken, NJ, USA: Wiley*, 2005.
- [2] S.Sujit Sansgiry, M.Bhosle and K. Sail, "Factors that affect academic performance among pharmacy students," *American Journal of Pharmaceutical Education*, 2006
- [3] N. V. Anand Kumar and G. V. Uma, "Improving Academic Performance of Students by Applying Data Mining Technique,"
- [4] R. Baker, "Data mining for education," in *International Encyclopedia of Education*, B.McGaw, P. Peterson, and E. Baker, Eds., 3rd ed. Oxford, U.K.: Elsevier, 2010.
- [5] Garcia, E. P. I., & Mora, P. M., "Model Prediction of Academic Performance for First Year Students", *Paper presented at the Artificial Intelligence (MICAI)*, 10th Mexican International Conference on, Nov. 26 2011-Dec. 4 2011.
- [6] Sachin, R. B., & Vijay, M. S, "A Survey and Future Vision of Data Mining in Educational Field", Paper presented at the *Advanced Computing & Communication Technologies (ACCT)*, Second International Conference on 7-8 Jan. 2012.
- [7] http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis