IARS' International Research Journal
ISSN: 2202-2821
ISSN: 1839-6518
iars.research@gmail.com
International Association of Research Scholars
Organismo Internacional

A Study of Relationship to Absentees and Score Using Machine Learning Method: A Case Study of Linear Regression Analysis

Singh Yadav, Ramjeet

A Study of Relationship to Absentees and Score Using Machine Learning Method: A Case Study of Linear Regression Analysis
IARS' International Research Journal, vol. 12, núm. 1, 2022
International Association of Research Scholars, Organismo Internacional
Disponible en: https://www.redalyc.org/articulo.oa?id=663872727005

Peer Reviewed Research Manuscript

A Study of Relationship to Absentees and Score Using Machine Learning Method: A Case Study of Linear Regression Analysis

Ramjeet Singh Yadav
Dr. Rammanohar Lohia Avadh University, India

IARS' International Research Journal, vol. 12, núm. 1, 2022

International Association of Research Scholars, Organismo Internacional

Revisado: 15 Febrero 2022
Aprobación: 27 Febrero 2022
Publicación: 28 Febrero 2022

Redalyc: https://www.redalyc.org/articulo.oa?id=663872727005

Abstract: Absenteeism from classrooms amongst students is an international problem that does not only affect Indian students. This research is focuses on absentees of student in class and score and has been carried out by using linear regression analysis. Linear regression analysis is one of excellent method of machine learning. The descriptive, student's t-test, Pearson correlation, and regression models were used in this study's statistical analysis. According to the results of this study, there are considerable variations between absentees and score (t-test=-4.06075,p<0.05). The study also discovered that absenteeism from class had a negative link with the score (r=-0.6088). To investigate the impact of class absentees on student score, a regression model was created. This study will benefit both the college administration and the students by raising awareness of the disadvantages of not attending classes.

Keywords: Linear Regression model, Machine Learning, Descriptive Statistical Analysis, Statistical Analysis.

I. Introduction

I In today's technologically advanced world, a major difficulty in primary and middle schools, intermediate colleges, post graduate colleges and university is the rising tendency of student absenteeism. At the present time India also facing these type student absentees' problems. Absenteeism is described as the habit of failing to show up for class or an event without a valid justification, and the term absentee is used to characterize someone who does this regularly. It has been observed via different studies and blind observation that students who do not attend classes receive worse grades, but individuals who have a greater attendance rate than their peers receive higher grades. A student's grade is determined by his/her attendance.

Various studies have shown that successful students have a better attendance record as well as a higher grade. According to Ahmat and Zahari's research, there is a negative relationship between absenteeism and grades, implying that more absence equals a lower grade [1]. The study looks into the link between students' attendance in class and their total marks in a variety of computer science programmes [2].

Zhang and Wang's research reveals that a positive correlation exists between the desired variables using a linear regression model [3]. On the basis of test score and classroom attendance data over two years,

regression analysis is introduced, and correlation curves between test score and classroom attendance are plotted [3]. The regression model revealed a robust link between semester GPA and attendance %, as well as overall GPA [4]. Similarly, Narula and Nagar's research demonstrates a high link between attendance and grades using machine learning methods [5]. In a Finnish University, the research examines the relationship between attendance and performance using the clustering method [6]. The research examines how absenteeism affects student outcomes using administrative panel data from California to estimate the pandemic's impact [7].

The outcomes of this study demonstrated that greater performance in professional assessment tests by medical undergraduate students has a negative link with absenteeism and a positive correlation with high attendance percentage [8]. All of this evidence demonstrates that attendance and grades have a substantial beneficial relationship using machine learning based regression analysis.

The assumptions have been presented as hypotheses in this study, and a proper statistical model has been utilised to prove them, as well as relationships proposed between them to show how they will change with the changing of each and every factor.

In this research, we performed several statistical tests to draw conclusions about the relationships, and we also used a regression model to predict the least square best equation to show how grade varies depending on each element. The Pearson correlation coefficient is a measure that is used to determine how closely the components are related to each other assuming they are related at all. This research provides a machine learning based method of linear regression model that depicts how a student's score changes as a result of factor absenteeism.

II. Machine Learning

The ML is the subclass of artificial intelligence and one of its applications. Machine learning provides the ability to automatically improve and learn from systems without having to be explicitly programmed. The method of machine learning is focused on the development of computer programs and it accesses the data and uses it to learn itself. The machine learning process starts with data and experimental observations like classification information of training data and testing data, examples, direct experience, instruction and pattern in the data. With the help of learning data, machine learning algorithms make deter decisions in the future based on the example. The main purpose of machine learning methods is to permit the computers learn routinely without human interference or support and regulate activities consequently. We have presented two machine learning-based linear regression and statistical methods to the data analysis purpose of in this proposed research work.

III. Linear Regression Analysis

Linear regression is the relationship between independent variable (d) and dependent variable (s). The equation of linear regression is given below:

$$s = \beta_1 d + \beta_0$$

Where are the unknown parameters β_1. and β_0. known as gradient or slope of line and intercept on y axis and is a normal random variable with a mean of zero and an unknown standard deviation. Note that this model is being offered for the entire population of students enrolled in this course, not just those enrolled this semester and especially not just those in the sample. The parameters β_1. and β_0. are all related to this enormous population. The β_1. and β_0. parameters are also referred to as regression model parameters. These parameters can be learned using the least square approach, artificial neural networks, evolutionary algorithms, and other applicable learning approaches. Least square estimation and the gradient descent algorithm can be used to learn the parameters β_1. and β_0.

IV. Experimental Results and Discussion

In general, educators believe that class attendance has a considerable impact on course achievement, all other conditions being equal. An education researcher chooses a multiple part basic computer science course at a large university to evaluate the relationship between attendance and performance. Throughout the semester, the course instructors agree to keep an accurate record of attendance. In this proposed study, we have taken data from Dr. Rammanohar Lohia Avadh University of BCA final year students for experimental purpose. The sample size of data is 30. Here, we have taken 30 students randomly out of 60 students at the end of the semester. Two measurements are taken for each student in the 30 sample which are given below:

1. The number of days (d) the student was unable to attend class.
2. End semester score (s).

Table 1 shows the dataset of 30 students and figure 1 depicts scatter plot of 30 students' data.

Table 1:
Data Set of 30 Students

Student	Absentee in Days (d)	End Semester Score (s)
1	3	77
2	6	28
3	7	64
4	2	97
5	3	81
6	8	72
7	1	90
8	0	91
9	5	52
10	6	75
11	1	85
12	2	76
13	1	64
14	3	40
15	5	66
16	4	89
17	1	97
18	1	99
19	0	90
20	0	97
21	2	92
22	0	89
23	2	69
24	1	86
25	3	79
26	1	79
27	0	90
28	6	65
29	1	80
30	2	98

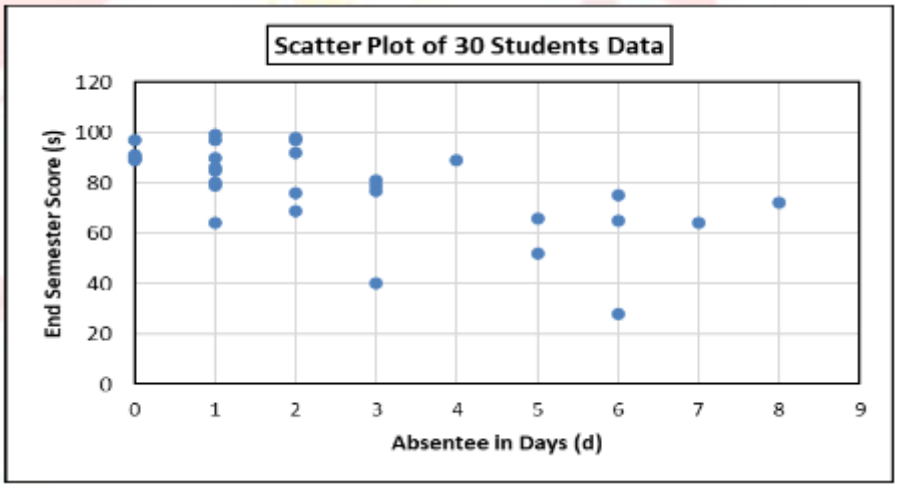

Figure 1:
Scatter plot of 30 Students Data

Table 2 shows the calculation of Calculation of d^2,*ds*, and s^2 In table 2 we have added the numbers in each column which will use in generating the trendline or predicted line.

Table 2:
Calculation of and

Student	Absentee in Days (d)	End Semester Score (s)	d^2	ds	s^2
1	3	77	9	231	5929
2	6	28	36	168	784
3	7	64	49	448	4096
4	2	97	4	194	9409
5	3	81	9	243	6561
6	8	72	64	576	5184
7	1	90	1	90	8100
8	0	91	0	0	8281
9	5	52	25	260	2704
10	6	75	36	450	5625
11	1	85	1	85	7225
12	2	76	4	152	5776
13	1	64	1	64	4096
14	3	40	9	120	1600
15	5	66	25	330	4356
16	4	89	16	356	7921
17	1	97	1	97	9409
18	1	99	1	99	9801
19	0	90	0	0	8100
20	0	97	0	0	9409
21	2	92	4	184	8464
22	0	89	0	0	7921
23	2	69	4	138	4761
24	1	86	1	86	7396
25	3	79	9	237	6241
26	1	79	1	79	6241
27	0	90	0	0	8100
28	6	65	36	390	4225
29	1	80	1	80	6400
30	2	98	4	196	9604
Sum	**77**	**2357**	**351**	**5353**	**193719**

The above data table 2 gives the following expression:

$$\sum d = 77, \sum s = 2357, \sum d^2 = 351, \sum ds = 5353, \sum s^2 = 193719$$

$$SS_{dd} = \sum d^2 - \frac{1}{n}\left(\sum d\right)^2 = (351)^2 - \frac{1}{30} \times (77)^2 = 153.3667$$

$$SS_{ss} = \sum s^2 - \frac{1}{n}\left(\sum s\right)^2 = 193719 - \frac{1}{30} \times 2357^2 = 8537.367$$

$$SS_{ds} = \sum ds - \frac{1}{n}\left(\sum d\right)\left(\sum s\right) = 5353 - \frac{1}{30} \times 77 \times 2357 = -693.633$$

$$\bar{d} = \frac{\sum d}{n} = \frac{77}{30} = 2.566667$$

$$\bar{s} = \frac{\sum s}{n} = \frac{2357}{30} = 78.56667$$

We start by determining the least squares regression line or predicted line, which is the line that best fits for the data. Its y- intercept and slope are given below:

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{-693.633}{153.3667} = -4.54227$$

$$\hat{\beta}_0 = \bar{s} - \hat{\beta}_1 \bar{d} = 78.56667 - (-4.54227) \times (2.566667) = 90.22517$$

The least squares regression line for this data, rounded to two decimal places, is:

$$\hat{s} = -4.54s + 90.23$$

Figure 2 depicts the fitted regression or predicted line of 30 students' data.

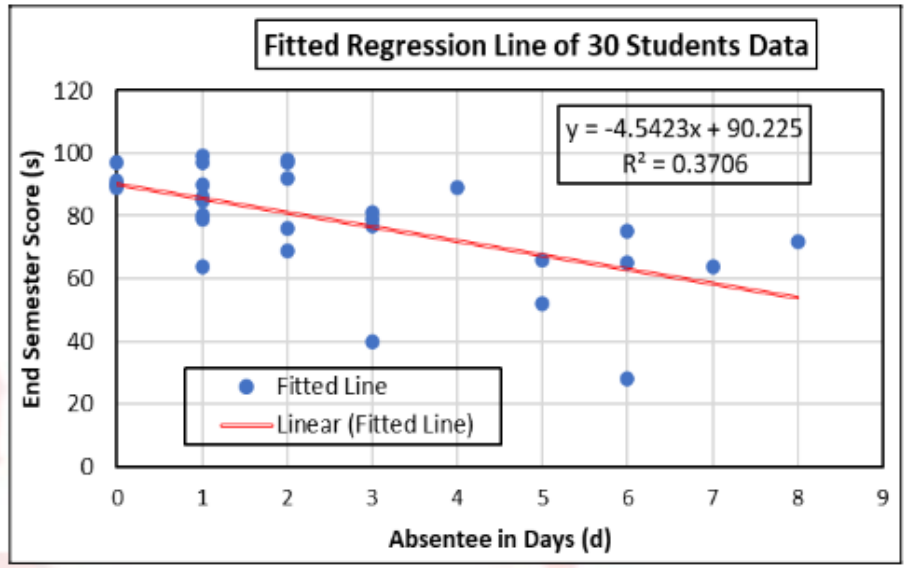

Figure 2:
Fitted regression or predicted line of 30 students' data

The figure 2 also shows that a decreasing trend, indicating that students with more absences perform worse on the final exam on average. The total of the squared errors (Sum of Square Error) of this line's goodness of fit to the scatter plot is:

5373.068

$$SSE = SS_{ss} - \hat{\beta}_1 SS_{ds} = 8537.367 - (-4.54227)(-693.633) = 5373.068$$

This is a huge amount. As a result, it isn't particularly useful in and of itself, but we utilised it to calculate a crucial statistic:

$$S_\epsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{5373.068}{30-2}} = 13.85263$$

The statistic S€ calculates the standard deviation (σ) of the model's normal random variable (€). It means that the standard deviation of final test grades for all students with the same number of absences is around 13.85 points. Because the number of absentees has such a huge impact on a 100-point exam, the final exam scores of each sub-population of students are very diverse. The size and sign of the slope of the predicted line $(\hat{\beta}_1 = -4.54227)$ imply that, on average, students score 4.54 points lower on the final test for each class missed. Similarly, students tend to score 2 x 4.54=9.08 points less on the final test for every two classes missed, or around a letter grade lower on average. The s-intercept has importance in this situation because 0 is inside the range of d- values in the data set. It's a guess about the average final test grade for all students that have perfect attendance. The intercept$(\hat{\beta}_0 = 90.22517)$ is the anticipated average for such pupils/students.

Before we go any farther with the regression equation or run any other analysis, it's a good idea to look at how useful the linear regression model is. This can be accomplished in two ways:

1. By calculating the correlation coefficient r, you can discover how closely the number of absentees (d) and the final exam score (s) are related.
2. By putting the null hypothesis H0: β1=0 to the test (The slope of the population regression line is zero, indicating that d is not a reliable predictor of s) vs the natural alternative H0: β1 <0 (Because the population regression line has a negative slope, final exam scores (s) decrease as absentees (d) increase).

The correlation coefficient *r* is:

$$r = \frac{SS_{ds}}{\sqrt{SS_{dd}SS_{ss}}} = \frac{-696.633}{\sqrt{(153.3667)(8537.367)}} = -0.6088$$

There is a moderately negative connection between the two variables. We can observe that there is a negative relationship between absentees and student scores. It means absentees of the student is increase then score will be decrease. Scores and absentees are two important data points in this study. The hypothesis developed for the first case involving absenteeism and student scores. Let us consider here the hypothesis which ais given below:

Null Hypothesis (H_0): Absentees does not affect Scores.

Alternative Hypothesis (H_a): Absentees affect the Scores

We have developed two hypotheses, and now we will use a statistically independent t-test to see if the null or alternate hypothesis is correct. The

independent t-test is performed to see if there is any correlation between absentee and student score. The t-test are used to test the hypothesis that the regression coefficients produced in basic linear regression are accurate. The two-sided hypothesis that the true slope, β_1, equals some constant value, $\beta_{1,0}$, is tested using a statistic based on the t distribution. Let's look at the test of hypotheses using the generally used 5% level of significance. The hypothesis test statements are written as follows:

$$H_0: \beta_1 = \beta_{1,0} = 0$$

$$vs. H_a: \beta_1 < 0, \alpha = 0.05$$

From the "Critical Value of" with $(d.f. = 30 - 2 = 28)$ degree of freedom $t_{0.05}$ so the rejection region is$_{[-\infty, -1.701]}$ The value of the standardized test is:

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{S_\epsilon / \sqrt{SS_{dd}}} = \frac{-4.54227 - 0}{13.85263/\sqrt{153.3667}} = -4.06075$$

This is located in the rejection zone. In favor of H_0, we reject H_0 At the 5% level of significance, the statistics support the conclusion that β_1 is negative, implying that as the number of absentees grows, the average final exam score declines. As previously stated, the figure $(\beta_1 = -4.54227)$ is a point estimate of how much one additional absentee affects the average final exam result. The average reduces by around 4.54227 points for each subsequent absentee.

The frequency of absentees has been visualized using a histogram in figure 3.

Figure 3: Histogram absentees with their frequency

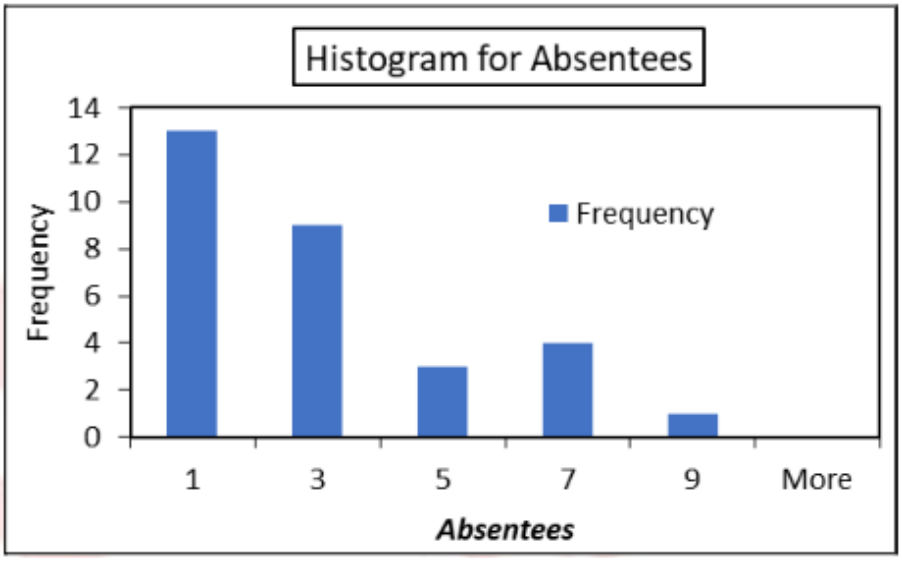

Figure 3:
Histogram absentees with their frequency

We can observe from this histogram that the majority of students have absentees in the range of zero percent to three. For β_1, we can expand this point estimate to a confidence interval. "Critical Values of" with *d.f*=30-2=28 degrees of freedom, $t_{a/2}=t_{0.025}=2.048$at the 95 percent confidence level. Based on our sample data, the 95 percent confidence interval for β_1 is:

$$\hat{\beta}_1 \pm t_{\alpha/2}\frac{S_\epsilon}{\sqrt{SS_{dd}}} = -4.54227 \pm 2.048 \times \frac{13.85263}{\sqrt{153.3667}} = -5.23 \pm 2.29 = (-7.52, -2.94)$$

We are 95 percent positive that, among all students who have ever taken this course, the average final test score drops by 2.94 to 7.52 points for each extra class missed. If we focus on the sub-population of all students who had exactly five absences. We may estimate the average final test score for those students using the least squares regression equation $\bar{s} = -4.54 \times 5 + 90.23 = 67.53$

This is also our best estimate of a student's final exam grade if he/she is absent five times. The average final test score for all students with five absences has a 95% confidence interval of:

$$\hat{s}_p \pm t_{\alpha/2}S_\epsilon\sqrt{\frac{1}{n}+\frac{(s_p-\bar{s})^2}{SS_{dd}}} = 67.53 \pm (2.048) \times (13.85263)\sqrt{\frac{1}{30}+\frac{(5-2.566667)^2}{153.3667}}$$

$$\hat{s}_p \pm t_{\alpha/2}S_\epsilon\sqrt{\frac{1}{n}+\frac{(s_p-\bar{s})^2}{SS_{dd}}} = 67.53 \pm 7.609393 = (75.14, 59.92)$$

According to this confidence interval, the true mean final test score for all students who miss class precisely five times over the semester is expected to be between 59.92 and 75.14. If a student misses exactly five classes during the semester, his final exam score is predicted to be in the interval with 95 percent certainty.

$$\hat{s}_p \pm t_{\alpha/2}S_\epsilon\sqrt{1+\frac{1}{n}+\frac{(s_p-\bar{s})^2}{SS_{dd}}} = 1 + 67.53 \pm (2.048) \times (13.85263)\sqrt{\frac{1}{30}+\frac{(5-2.566667)^2}{153.3667}}$$

$$\hat{s}_p \pm t_{\alpha/2}S_\epsilon\sqrt{\frac{1}{n}+\frac{(s_p-\bar{s})^2}{SS_{dd}}} = 67.53 \pm 29.37296 = (96.90, 38.16)$$

This prediction interval indicates that this student's final exam score will most likely fall somewhere between 38.16 and 96.90. Unlike the 95 percent confidence interval for the average score of all students with five absences, which provided useful information, this interval is so large that it reveals almost nothing about the final exam score of any particular student. The existence of the extra summand 1 under the square sign in the prediction interval can have a dramatic effect in this case. Finally, the coefficient of determination, r^2 , estimates the fraction of the variability in students' final exam scores that is explained by the linear relationship

between that score and the number of absences. Since we've already calculated ., we can readily deduce:

$$r^2 = (-0.608803016)^2 = 0.370641112 \approx 0.37$$

As a result, the regression model explains 37 percent of the variability in the yield data, demonstrating a good fit of the regression model. Despite the fact that there is a strong link between attendance and final test performance. Although we can estimate the average score of students who miss a specific number of classes with reasonable accuracy, the number of absentees accounts for less than half of the entire range in exam scores in the sample. This is hardly surprising, given that student exam performance is influenced by a variety of factors other than attendance.

A residual plot is a graph in which the residuals are displayed on the vertical axis and the independent variable is displayed on the horizontal axis. A linear regression model is appropriate for the data if the dots in a residual plot are randomly distributed across the horizontal axis; otherwise, a nonlinear model is more suited. Table 3 shows the output of linear regression model and residuals.

Figure 4 also depicts that the residual plot displays a haphazard pattern. Some residual points are positive, while others are negative. This random pattern implies that the data is well-fit by proposed a linear model.

V. Conclusion

The findings of this study revealed that absence has a major impact on academic achievement. The research was carried out in-depth, with a lot of data visualization and statistical modelling included in the publication. The findings revealed a moderately negative relationship between the number of absences and the final score (r=-0.6088, p=0.00036 which is les than 0.05) exam scores between students who missed less than and equal to 22% of their classes and students who missed more than 23% of their classes (*t-test*=-4.06075 and the p is less then 0.05) The key finding was that if a student misses one class, their final test grades are projected to drop by 4.54 percent on average. It is believed that the findings of this study would help the colleges and university plan for students who will graduate on time. Furthermore, this study has the potential to raise student knowledge about the impact of missing courses on their academic performance.

Table 3:
Output of Linear Regression Model and residuals

Students	Predicted Score (y)	Residuals
1	76.59834819	0.401651815
2	62.97152793	-34.97152793
3	58.42925451	5.570745490
4	81.1406216	15.85937840
5	76.59834819	4.401651815
6	53.88698109	18.11301891
7	85.68289502	4.317104977
8	90.22516844	0.774831558
9	67.51380135	-15.51380135
10	62.97152793	12.02847207
11	85.68289502	-0.682895023
12	81.1406216	-5.140621604
13	85.68289502	-21.68289502
14	76.59834819	-36.59834819
15	67.51380135	-1.513801348
16	72.05607477	16.94392523
17	85.68289502	11.31710498
18	85.68289502	13.31710498
19	90.22516844	-0.225168442
20	90.22516844	6.774831558
21	81.1406216	10.8593784
22	90.22516844	-1.225168442
23	81.1406216	-12.1406216
24	85.68289502	0.317104977
25	76.59834819	2.401651815
26	85.68289502	-6.682895023
27	90.22516844	-0.225168442
28	62.97152793	2.028472071
29	85.68289502	-5.682895023
30	81.1406216	16.8593784

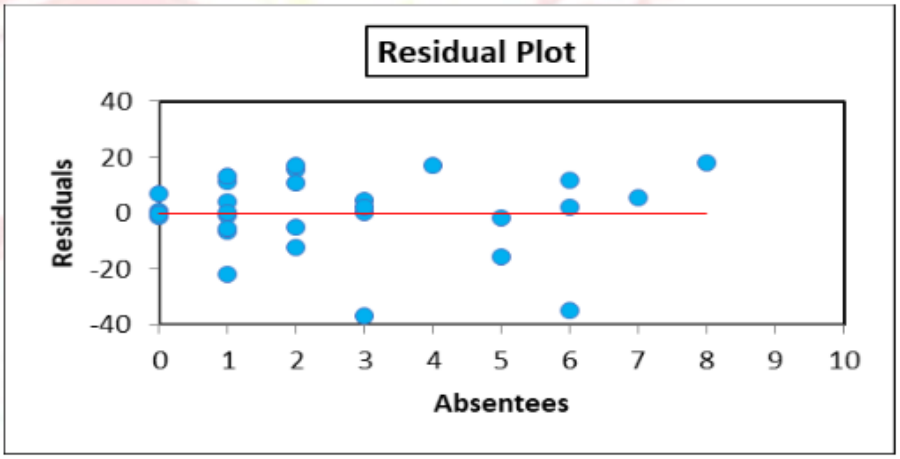

Figure 4:
Residual plot of proposed regression model

VI. References

Nurhafizah Ahmad, Ahmad Zia Ul-Saufie, Siti Asmah Mohamed, Hasfazilah Ahmat and Mohd Fahmi Zahari. The Impact of Class Absenteeism on Student's Academic Performance using Regression Models. Conference Proceeding of the 25th National Symposium on Mathematical Sciences, Published by the American Institute of Physics, (June 2018), pp. 1-5. https://aip.scitation.org/doi/abs/10.1063/1.5041712.

Jenq-Foung "Jf" Yao and Tsu-Ming Chiang. Correlation Between Class Attendance and Grade. Journal of Computing Sciences in Colleges, Volume 27, Issue 2, (December 2011), pp. 142–147. https://dl.acm.org/doi/abs/10.5555/2038836.2038857.

Chengcheng Zhang and Fei Wang. Research on Correlation Analysis between Test Score and Classroom Attendance Based on Linear Regression Model. Proceeding of 2010 2nd International Conference on Industrial Mechatronics and Automation (ICIMA2010), IEEE, (May 2010), pp. 545-548. https://ieeexplore.ieee.org/document/5538079.

Suleiman Obeidat, Adnan Bashir, Wisam Abu Jadayil. The Importance of Class Attendance and Cumulative GPA for Academic Success in Industrial Engineering Classes. World Academy of Science, Engineering and Technology International Journal of Humanities and Social Sciences Vol:6, No:1, (2012), pp. 120-123. https://publications.waset.org/2974/the-importance-of- class-attendance-and-cumulative-gpa-for-academic-success-in-industrial-engineering-classes.

Meenakshi Narula and Pankaj Nagar. Relationship Between Students' Performance and Class Attendance in a Programming Language Subject in a Computer Course. International Journal of Computer Science and Mobile Computing. Vol. 2, Issue. 8, August 2013, pg.206 – 210. https://www.ijcsmc.com/docs/papers/August2013/abstract s/V2I8201318.pdf.

Anna Lukkarinen, Paula Koivukangas, Tomi Seppala. Relationship between class attendance and student performance. 2nd International Conference on Higher Education Advances, HEAd´16, Procedia - Social and

Behavioral Sciences 228 (2016), pp. 341 – 347. https://www.sciencedirect.com/science/article/pii/S18770 42816309776.

Lucrecia Santibanez and Cassandra M. Guarino. The Effects of Absenteeism on Academic and Social- Emotional Outcomes: Lessons for COVID-19. Educational Researcher, Vol. 50 No. 6, (February 25, 2021), pp. 392–400. https://journals.sagepub.com/doi/10.3102/0013189X21994488

Yousaf Latif Khan, Sohail Khursheed Lodhi, Shahzad Bhatti and Waqas Ali. Does Absenteeism Affect Academic Performance Among Undergraduate Medical Students? Evidence From "Rashid Latif Medical College (RLMC)." Advances in Medical Education and Practice, Vol. 10, Issue 1, (2019), pp. 999–1008. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6897060/