Report: Analysis of Student Performance

Objective

The objective of this analysis is to investigate the factors of student performance in an academic setting. We aim to build a predictive model using multiple linear regression to understand how factors such as study hours, previous scores, extracurricular activities, sleep hours, and sample question papers practiced influences student's performance.

Data summary

This table provides insights into the academic performance and related factors of students.

	Hours studied	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Paper Practiced	Performance Index
Min	1.000	40.00	0.000	4.000	0.000	10.00
1st quartile	3.000	54.00	0.000	5.000	2.000	40.00
Median	5.000	69.00	0.000	7.000	5.000	55.00
Mean	4.993	69.45	0.4948	6.531	4.583	55.22
3rd quartile	7.000	85.00	1.0000	8.000	7.000	71.00
Max	9.000	99.00	1.0000	9.000	9.000	100.00

Table 1.

Hours studied: the number of hours students spent studying.
Previous Scores: the scores students achieved in their previous assessments.
Extracurricular Activities: This is a categorical column, which indicates whether students are involved in extracurricular activities. A value of 0 mean no involvement, while a value of 1 means involvement.
Sleep Hours: The average number of hours of sleep students get per night.
Sample Question Paper Practiced: The number of sample question papers students have practiced.
Performance Index: an index or score representing the overall academic performance of students.

Model summary

We fitted a linear regression model with the response variable Performance Index and all five predictors, including hours studied, previous scores, extracurricular activities, sleep hours, and sample question papers practiced.

Residuals

Min	1st quartile	Median	3rd quartile	Max
-8.6333	-1.368	-0.0311	1.3556	8.7932

Table 2.

Coefficients

	Estimate	Std. Error	t value	Pr(> | t |)
Intercept	-34.07558	0.127143	-268.01	<2e-16
Hours studied	2.852982	0.007873	362.35	<2e-16
Previous scores	1.018434	0.001175	866.45	<2e-16
Extracurricular Activities	0.612898	0.040781	15.03	<2e-16
Sleep hours	0.480560	0.012022	39.97	<2e-16
Sample Question Papers Practiced	0.193802	0.007110	27.26	<2e-16

Table 3.

Model Fit Measures

R^2	Adjusted r^2	F-statistic	df1	df2	p-value
0.9888	0.9887	1.757e+05	5	9994	< 2.2e-16

Table 4.

In table 4, the model has an adjusted R-squared value of 0.9887, indicating that approximately 98.87% of the variance in the Performance Index is explained by the independent variables. Since the p values for all the predictors are less than 0.05, we conclude that the predictors are significant.
In table 3,
Hours Studied: For each additional hour studied, there is an increase of 2.853 points in the performance index, holding all other variables constant.

Previous Scores: For an increase of one unit in previous scores, there is an increase of 1.0184 in the performance index, holding all other variables constant.
Extracurricular Activities: Students who participate in extracurricular activities have, on average, a 0.613-point higher Performance Index compared to those who do not, holding all other variables constant.
Sleep Hours: For each additional hour of sleep, there is an increase of 0.481 points in the performance index, holding all other variables constant.
Sample Question Papers Practiced: Practicing 1 sample question paper more will lead to an increase of 0.194 points in the Performance Index, holding all other variables constant.

Assumption check

First, we will examine the residuals through a histogram. If it doesn't follow a normal distribution, it can lead to biased estimates, incorrect inferences, and inaccurate confidence intervals in regression analysis. This violates the key assumption of linear regression and can result in reduced model performance. Second, we will check for multicollinearity through a a correlation chart. If multicollinearity happens, it will be challenging to access the true effects of the predictors on the response.

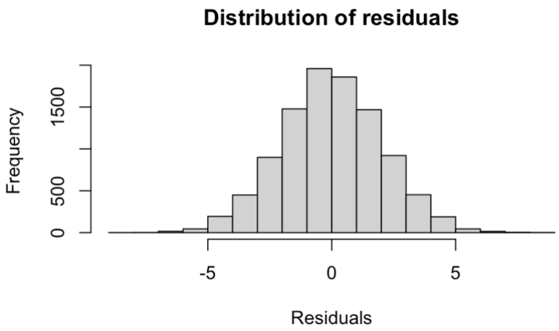


Figure 1.

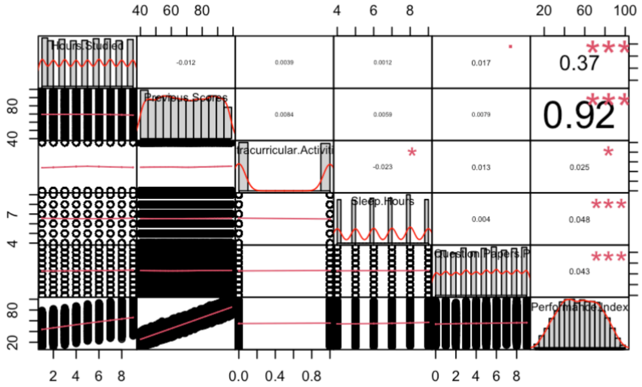


Figure 2.

In figure 1, it shows that the residuals follow a normal distribution satisfying the normality assumption.
In figure 2, it shows that correlation between all predictors is smaller than 0.8, so there is no correlation between the predictors. Therefore, we keep all five predictors in our model.

Assessment of Model Fit

We construct an F-test between two models to determine whether there is a significant linear relationship between the predictor variables and response variable. The null hypothesis states that there is no significant linear relationship between the predictor variables and the response variable. The alternative hypothesis suggests that there is a significant linear relationship between the response and at least one of the predictors. In table 4, Model 1 includes only response regressed on the intercept, while Model 2 is the full model.

$$H_0 : performance.index = \beta_0$$

$$H_a : performance.index = \beta_0 + hours.studied + previous.scores + extracurricular.activities + sleep.hours + sample.question.papers.practiced$$

ANOVA

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	9999	3690855				
2	9994	41514	5	3649341	175709	< 2.2e-16

Table 5.

In table 5, Since the p-value is smaller than 0.05, we reject the null hypothesis. So, there is a linear relationship between the response variable and at least one of the predictors.

Partial F-Test :
We construct a partial F-Test to test for linear association with a subset of predictors. Since the correlation between the response performance index and extracurricular activities is the smallest among all response and predictors, we remove extracurricular activities in the reduced model. The null hypothesis states that the performance index is not significantly affected by extracurricular activities in the regression model. The alternative hypothesis suggests that the performance index is significantly influenced by extracurricular activities. In table 6, model 1 includes the response regressed on all the predictors excluding extracurricular activities, while model 2 is the full model.

$$H_0 : performance.index = \beta_0 + hours.studied + previous.scores + sleep.hours + sample.question.papers.practiced$$

$$H_a : performance.index = \beta_0 + hours.studied + previous.scores + +extracurricular.activities + sleep.hours + sample.question.papers.practiced$$

ANOVA

Model	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)

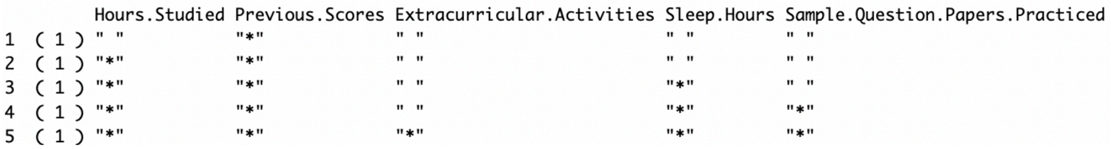
1	9995	42452				
2	9994	41514	1	938.25	225.88	< 2.2e-16

Table 6.

In table 6, since the p value is smaller than 0.05, we reject the null hypothesis. So, we keep the full model.

Model Selection

To choose the best model, we explore all possible subset of predictors. Table 7 presents the results of exploring different subsets of predictors to determine the best model based on adjusted R-squared, AIC, and BIC values. Each row represents a different model, with the number of predictors increasing from Model 1 to Model 5.



```
         Hours.Studied Previous.Scores Extracurricular.Activities Sleep.Hours Sample.Question.Papers.Practiced
1  ( 1 ) " "           "*"             " "                        " "         " "
2  ( 1 ) "*"           "*"             " "                        " "         " "
3  ( 1 ) "*"           "*"             " "                        "*"         " "
4  ( 1 ) "*"           "*"             " "                        "*"         "*"
5  ( 1 ) "*"           "*"             "*"                        "*"         "*"
```

Model	Predictors	Adjusted r^2	aic	bic
1	1	0.8375549	134328.4804	-18156.73
2	2	0.9858696	2558.9129	-42568.64
3	3	0.9876461	981.6685	-43903.97
4	4	0.9884935	229.8763	-44606.40
5	5	0.9887467	6.0000	-44820.68

Table 7.

We select our best model by exploring all possible subsets. Since the adjusted R squared is the largest and the aic and bic are the smallest for model 5, we should include all five predictors.

Run diagnostic

After choosing the best model, we perform diagnostic checks to ensure all assumptions are met through Residual vs Fitted plot, scale location plot, qq plot, residuals vs leverage plot and standardized residuals vs predictor plot.

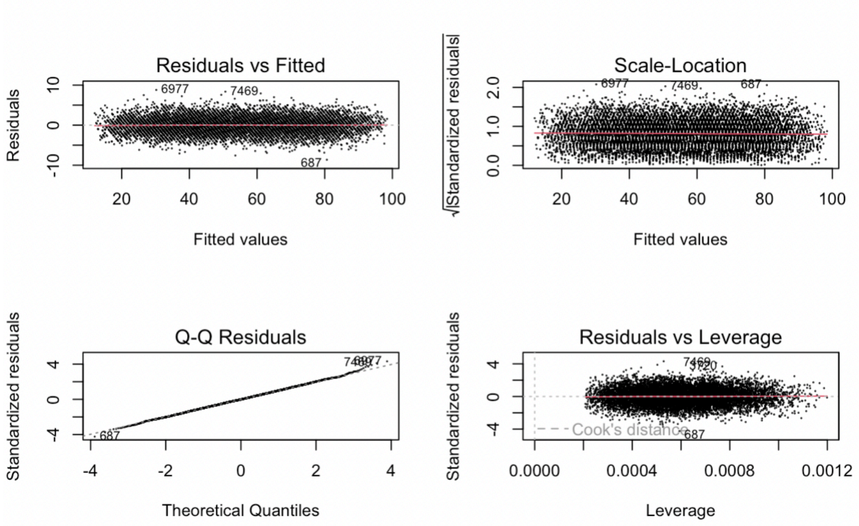


Figure 3.

Residuals vs Fitted: The points are randomly scattered around the horizontal line at y = 0, which indicates that the assumption of linearity is satisfied. This means that the relationship between the predictor variables and the response variable is linear.
Normal Q-Q: The points follow the diagonal line, which suggests that the residuals are normally distributed. This indicates that the assumption of normality of residuals is met.
Scale-Location: The points are randomly scattered around the horizontal line, which implies that the variance of the residuals is approximately constant across different levels of the predictor variables. This indicates that the assumption of constant variance of residuals is satisfied.
Residual vs leverage : The points are close to each other, with only a few points slightly further away. This suggests that there are no extreme leverage points in the model.

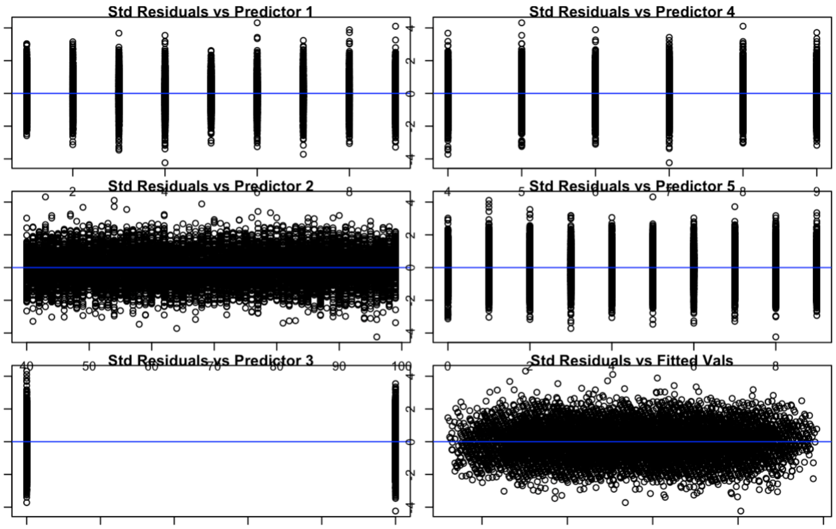
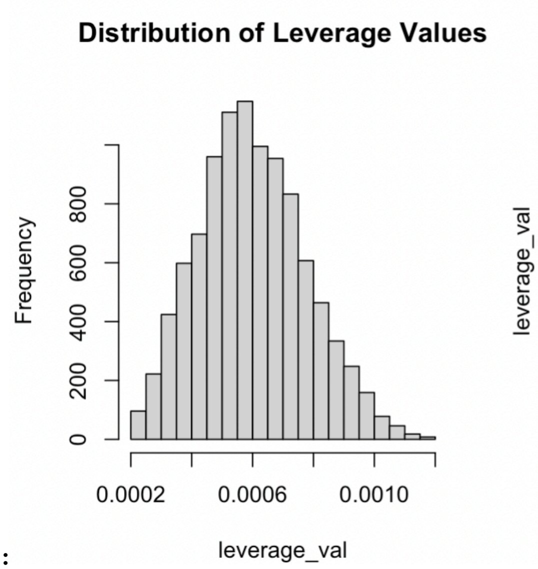


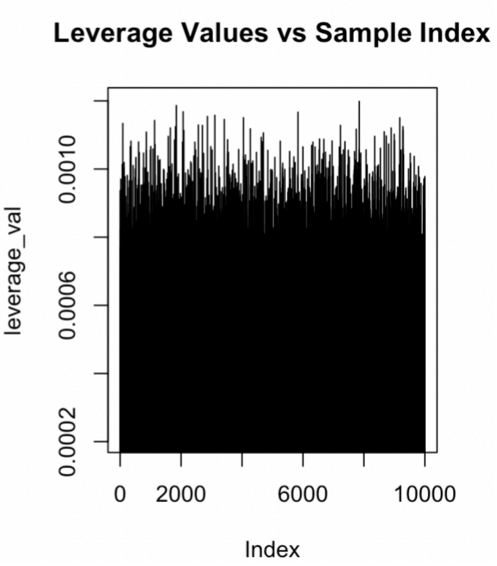
Figure 4.

In figure 4, the points are randomly scattered around the horizontal line at y = 0 with consistent variance, and there are no clear patterns or outliers, which suggests that the model is a good fit for the data, and the assumptions of linearity and constant variance are likely met. This suggests the model is valid. The distribution of leverage points approximately follows a normal distribution and there are no obvious spikes in the leverage value vs sample index graph which indicates that there are no leverage points.

Investigate high leverage, influential points, and outliers

We will use leverage plots to find high leverage points, cook's distance plots to identify influential points, and residuals vs fitted values plots to find outliers. By doing this, we ensure our model is reliable and our conclusions are trustworthy.





:
Figure 5 .

Figure 6.

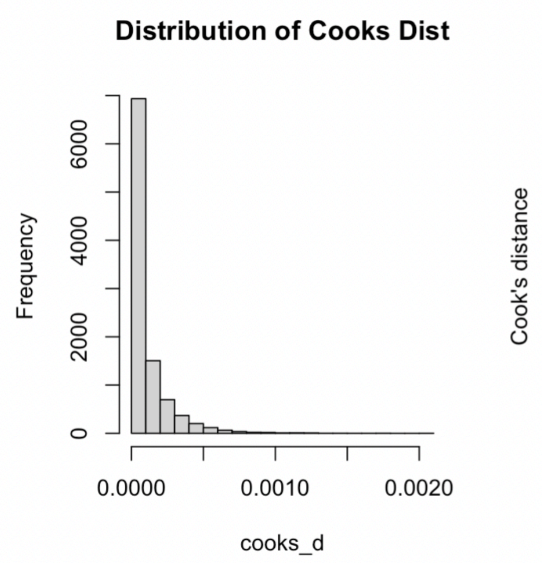


Figure 7.

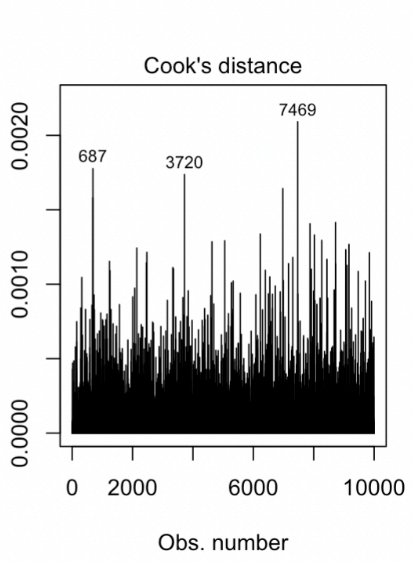


Figure 8.

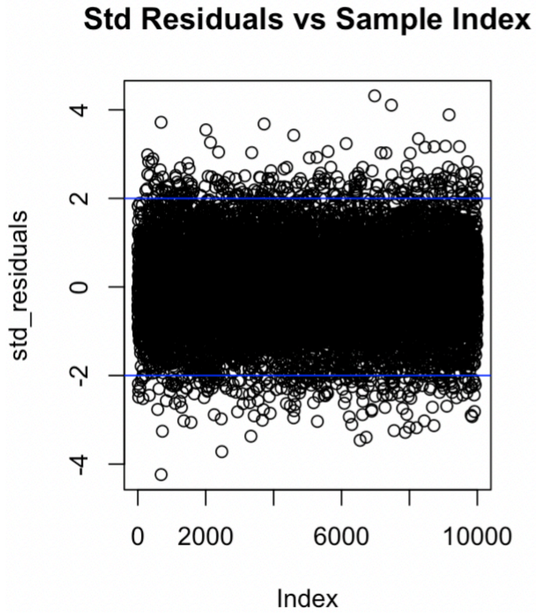


Figure 9.

In figure 5, the distribution of leverage points approximately follows a normal distribution and in figure 6, there are no obvious spikes in the graph which indicates that there are no leverage points.
In figure 7, Most points have a small cook's distance on the left graph. But figure 8 shows that there is still a few points that have a large cook's distance. Using r, we get that there is 493 points that have a large cook's distance.
In Figure 9, There are quite a few points that lie outside (-2,-2), which suggest that there are some outliers. Using r, we calculated that there is 477 outliers.

Conclusion

After analyzing the data with assumption checks, ANOVA, partial F-test, model selection, diagnostic assessments, and investigation of high leverage, influential points, and outliers, we conclude that the full model is the best fit.

Works Cited

https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression