# Influential Factors on Student Performance

Katherine Jin - 505650329

## I. Introduction

As a student or people related, people would care about student performance and ways to improve. Some people consider hours of studying will be the most important factor, while some consider a good rest is the key. This project is conducted to answer the question: what can help improve student performance and how big of an impact it can have.

This short project will examine five possible factors, hours studied, previous scores, extracurricular activities, sleep hours, and sample question papers practiced, and look closely on the relationship of them with the performance index. The dataset contains 10000 observations of student records.

## II. Data Description

The dataset contains the following predictors:

1. Hours Studied: The total number of hours spent studying by each student.
2. Previous Scores: The scores obtained by students in previous tests.
3. Extracurricular Activities: Whether the student participates in extracurricular activities (Yes or No).
4. Sleep Hours: The average number of hours of sleep the student had per day.
5. Sample Question Papers Practiced: The number of sample question papers the student practiced.

The response of the dataset will be the following:

- Performance Index: A measure of the overall performance of each student. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

Notice that the predictor “extracurricular activities” has the value as character type, to make the computation easier, 0 and 1 is used to replace “No” and “Yes”.

Table 1: Summary of Variables

| Statistics\ Variable | Hours Studied | Previous Scores | Extracurricular Activities | Sleep Hours | Sample Question Papers Practiced | Performance Index |
|---|---|---|---|---|---|---|
| Mean | 4.9929 | 69.4457 | 0.4948 | 6.5306 | 4.5833 | 55.2248 |
| Median | 5 | 69 | 0 | 7 | 5 | 55 |
| Standard deviation | 2.589309 | 17.34315 | 0.499998 | 1.695863 | 2.867348 | 19.21256 |
| Range | (1, 9) | (40, 99) | (0, 1) | (4, 9) | (0, 9) | (10,100) |

From the above summery table, we notice that, for every variable, the mean is roughly the midpoint of the range, which implicates the distributions are roughly symmetric. However,

one thing needs to be noticed is that the units of variables are not the same, thus it is not applicable to evaluate their influence base on the same standard.

## III. Fit Linear Regression Model

First step is to check is any two of the predictor variables are highly correlated.

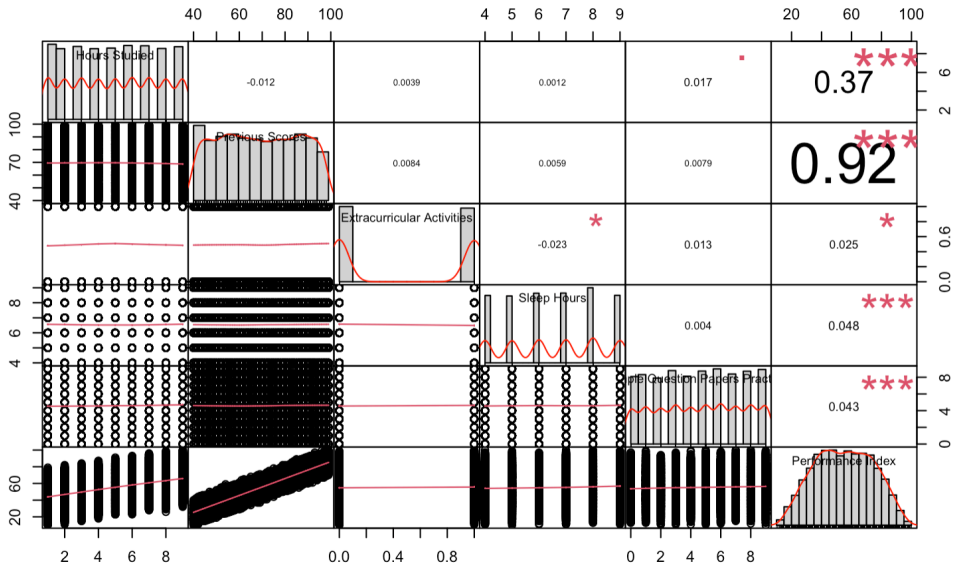


Figure 1: Visualize Correlation

According to figure 1, none of the two predictors have high correlation. Additionally, "previous score" has high correlation with the response. Therefore, all predictors can be included in the full model.

*Step 1: ANOVA for the Full Model*

The first step is to ensure if any predictor is significant to the regression model. Therefore, the hypothesis test is as the following:

$$H_0: Performance\ Index = \beta_0 \quad (reduced\ model)$$

$$H_1: Performance\ Index = \beta_0 + \beta_1 Hours\ Studied + \beta_2 Previous\ Scores + \beta_3 Extracurricular\ Activities + \beta_4 Sleep\ Hours + \beta_5 Sample\ Question\ Papers\ Practiced \quad (full\ model)$$

After applying F-test, the p-value is 2.2e-16, which is significantly smaller than any reasonable $\alpha$. Therefore, at least one predictor is significant.

*Step 2: Partial F-test*

The second step is to exclude potential non-significant predictor and apply partial f-test to figure out if the reduced model is better or not. The predictor "extracurricular activities" has the lowest correlation with response based on the result from Figure 1. Thus, the hypothesis will be as the following:

$$H_0: Performance\ Index = \beta_0 + \beta_1 Hours\ Studied + \beta_2 Previous\ Scores + \beta_3 Sleep\ Hours + \beta_4 Sample\ Question\ Papers\ Practiced \quad (reduced\ model)$$

$$H_1: Performance\ Index = \beta_0 + \beta_1 Hours\ Studied + \beta_2 Previous\ Scores + \beta_3 Extracurricular\ Activities + \beta_4 Sleep\ Hours + \beta_4 Sample\ Question\ Papers\ Practiced \quad (full\ model)$$

The resulting p-value is 2.2e-16, which is significantly smaller than any reasonable $\alpha$. Therefore, the full model is better than the reduced model.

*Step 3: T-Test on the Full Model*

According to Figure 1, it is likely that predictors does not have collinearity. Thus, the result from t-test would be valid. After applying t-test on the full model, the p-value for all predictors are significantly small, which indicates that every predictor should be included in the regression model.

The coefficient estimates for each predictor is as the following:

Table 2: Coefficient Estimates for Predictors

| Predictor | Coefficient Estimate |
|---|---|
| Hours Studied | 2.852982 |
| Previous Scores | 1.018434 |
| Extracurricular Activities | 0.612898 |
| Sleep Hours | 0.480560 |
| Sample Question Papers Practiced | 0.193802 |

Although predictors might have different unit of measurement, some conclusion can still be conducted. Since “Hours Studied” and “Sleep Hours” have the same unit, the coefficient can be directly compared. “Previous Scores” has unit as points, which is considered easiest to obtain. Thus, it is reasonable to conclude “Previous Scores” affect “Performance Index” the most.

*Step 4: Model Selection*

The best fitted model should have the least BIC value and the greatest $R^2_{adj}$ value. Therefore, after calculating all possible combination of regression models, the best model is the full model, same as the conclusion of the above steps.

As conclusion, the best fitted model is:

$$Performance\ Index = \beta_0 + \beta_1 Hours\ Studied + \beta_2 Previous\ Scores + \beta_3 Extracurricular\ Activities + \beta_4 Sleep\ Hours + \beta_5 Sample\ Question\ Papers\ Practiced$$

## IV. Diagnostic of Assumption

There are assumptions for a linear regression model. Independence of observations, normality of the error term, constance variance, and enough sample size.

1. *Independence of observation*

Since the independence of student records are difficult to check, it is assumed that every student independently study and sleep to obtain the data.

2. *Enough sample size*

The sample size is 10000, way larger than the number of predictors. Therefore the assumption is satisfied.

3. *Normality and Homocsedacity*

According to Figure 2, three scatter plots all have points randomly scatter with no obvious pattern. In the plot “Residuals vs Fitted” and “Residuals vs Leverage”, points are scattered around 0. These inplicates that residuals approximately follow normal distribution centered at 0 with constant variance. The Normal Q-Q is roughly a straight line, which also implicates the residuals follow normal distribution.

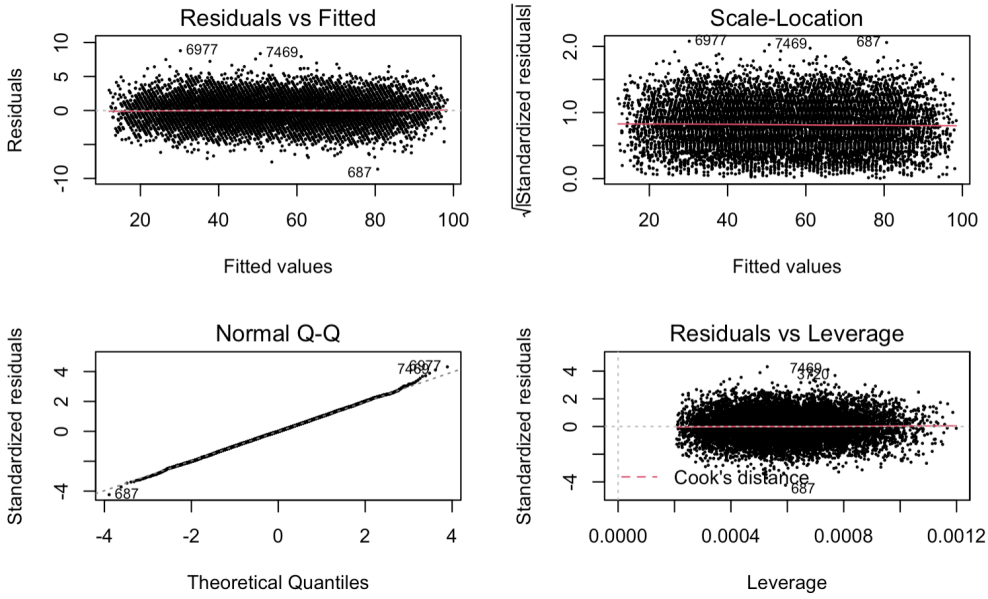


Figure 2: Standard Plot for Diagnostics

## *Diagnostic on Influencial Points*

Three tools are often used in identifying influential points and potential outliers: standardized residuals, leverage points, and Cook’s distance. These metrics will be applied one by one in the following:

1. *Standardized Residuals:*

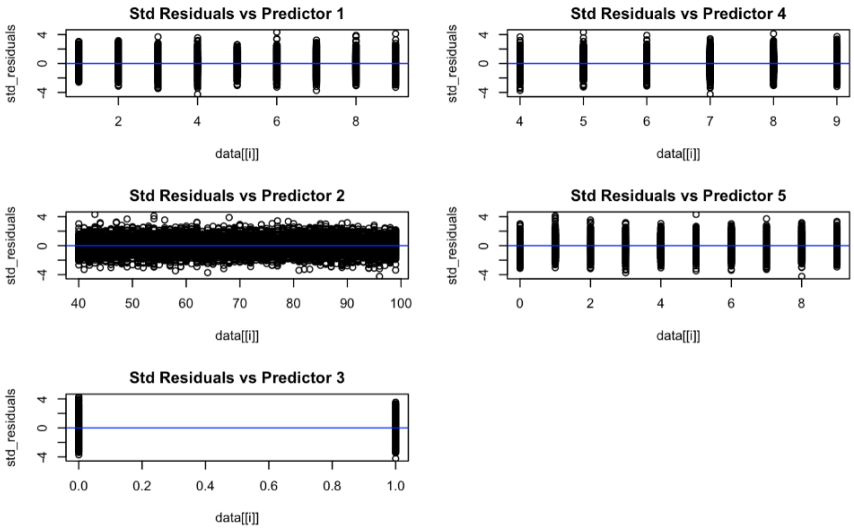


Figure 3: Plot Residuals Against Predictors

According to Figure 3, every predictor seems to have constant standard residuals.

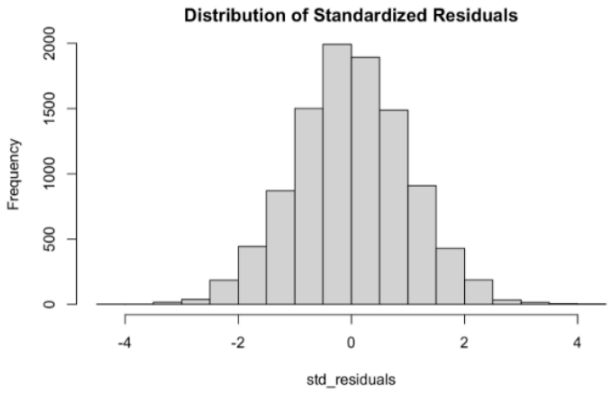


Figure 4: Distribution of Standardized Residuals

According to Figure 4, the distribution of standardized residuals seems to be approximately normal. Points lie outside the range of [-2, 2] are potential outliers. After the calculation, there are 477 outliers out of 10000 data points.

2. *Leverage*

Any points that have a leverage that $\frac{2 \times (number\ of\ predictors + 1)}{number\ of\ observations}$ will be identified as a high leverage point. After the calculation, there is no leverage points in the dataset.

3. *Cook's Distance*

Any points that have a Cook's distance that larger than $\frac{4}{number\ of\ observations}$ will be identified as potential influential points. After the calculation, there are 493 points with high Cook's distance out of 10000 data points.

## V. Conclusion

Based on the inference made above, the impact of studying hours, previous scores, extracurricular activities, sleep hours, and sample question papers all have statistically significant impact on student performance. Additionally, if one has high previous test score, it is more likely to also get a high score when evaluating the performance. Therefore, in order to improve the student performance, its essential to keep studying and preparing for a length of time. Other than study, having good rest and do extracurricular activities will also, somehow, improve the performance.

In conclusion, student performance is a process of accumulating in a long-term. It is more important to develop a good study habits and methods, and then the performance will be naturally improved.