Introduction

Practical Problems Solved

Analysis of key factors of academic performance among different students

Objective of Research: Make use of regression analysis to identify most important factors that affect students' academic performance and assist educators and learners to better support themselves.

Methodes: Multi linear regression analysis

Data descripition

Generally, This dataset has a variety of variables about students' learning behaviors and performance.

The data set has 10,000 rows and 6 columns which means there are 10,000 samples and 6 variables.

The variables were listed as follow:

Hours.Studied: Study time, which represents the time students spend studying.

Previous.Scores: Previous scores, which represent the scores of students in previous exams.

Extracurricular.Activities: Extracurricular activities, which represent the extracurricular activities participated by students.

Sleep.Hours: Sleep time, which represents the number of sleep hours per day of students.

Sample.Question.Papers.Practiced: The number of sample questions practiced, which represents the number of sample questions practiced by students.

The response variables was listed as follow:

Performance Index: Response variable, the higher the performance index, the better the learning.

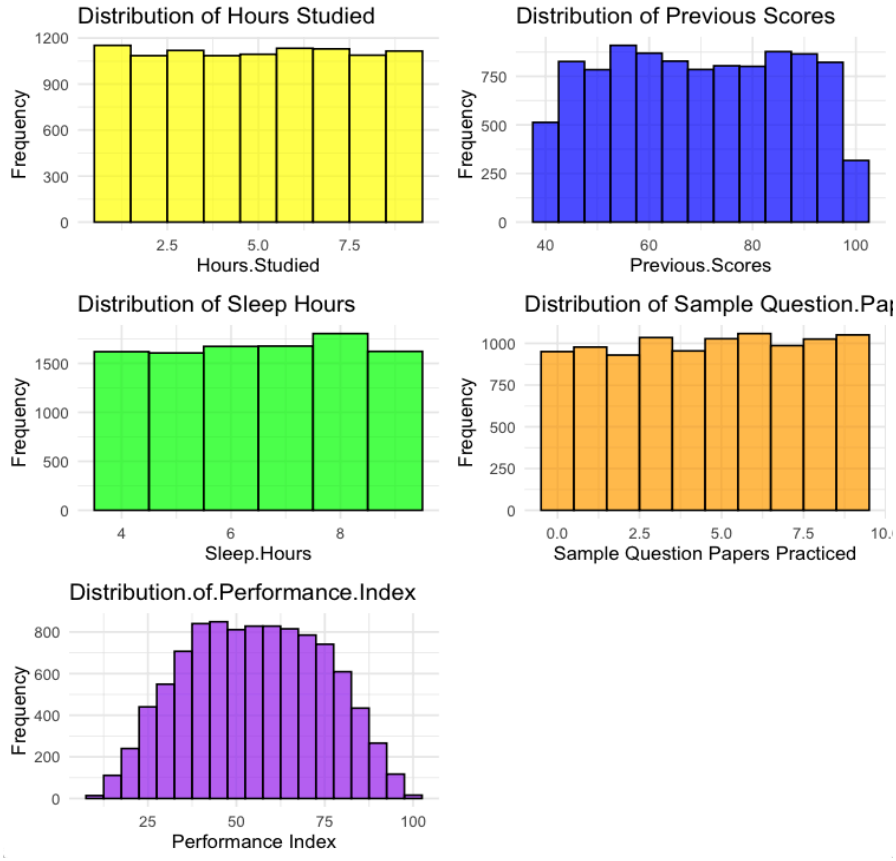
All numeric columns have 10,000 non-null values.

Among the variables, "Extracurricular Activities" is categorical. The other are numeric variables.

```
> head(df)
  Hours.Studied Previous.Scores Extracurricular.Activities Sleep.Hours
1             7              99                        Yes           9
2             4              82                         No           4
3             8              51                        Yes           7
4             5              52                        Yes           5
5             7              75                         No           8
6             3              78                         No           9
  Sample.Question.Papers.Practiced Performance.Index
1                                1                91
2                                2                65
3                                2                45
4                                2                36
5                                5                66
6                                6                61
```

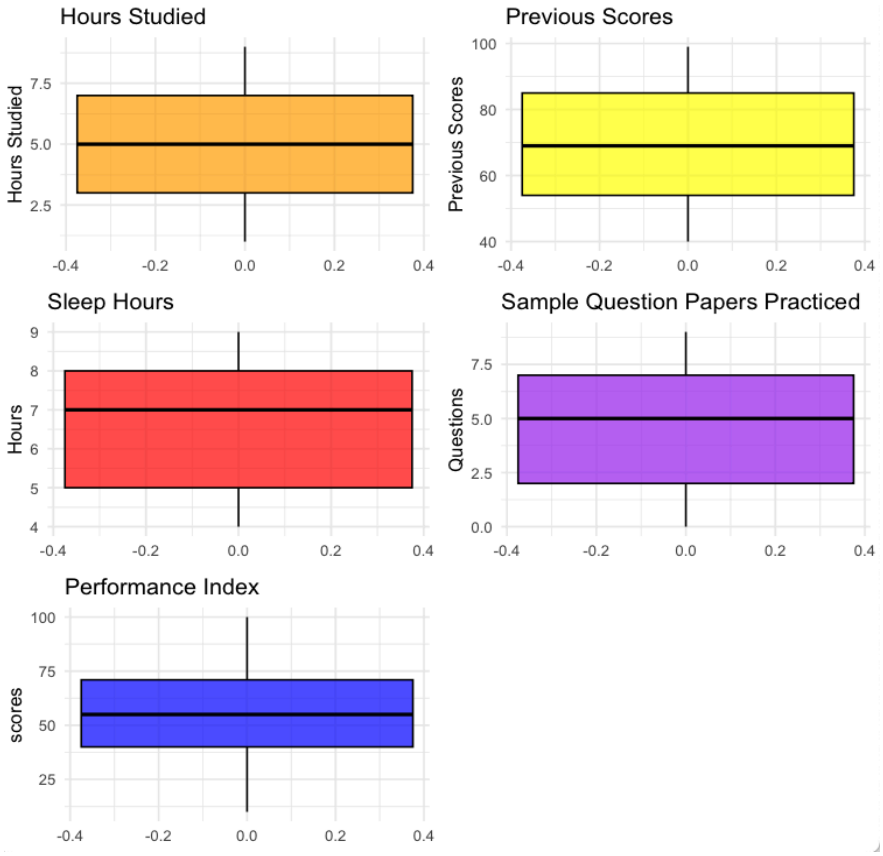
Exploratory data analysis

Distribution of Numerical Features



Based on the above histograms, “Hours Studied” and “Sample Question Papers Practiced” is similar with distribution. “Performance Index” seens have a normal distribution with many students' scores concentrated in the middle. "Previous Scores” is a little bit right skewed.

Boxplots of Numerical Features



Based on the above boxplots, there are some extreme values between “Previous Scores” and “Performance Index” .
The median and interquartile range of “Hours Studied” and “Sleep Hours" indicate that most students study 5 hours and sleep 7 hours on a daily basis.

Summary Statistics

```
> summary(df)
 Hours.Studied   Previous.Scores Extracurricular.Activities  Sleep.Hours
 Min.   :1.000   Min.   :40.00   Min.   :0.0000             Min.   :4.000
 1st Qu.:3.000   1st Qu.:54.00   1st Qu.:0.0000             1st Qu.:5.000
 Median :5.000   Median :69.00   Median :0.0000             Median :7.000
 Mean   :4.993   Mean   :69.45   Mean   :0.4948             Mean   :6.531
 3rd Qu.:7.000   3rd Qu.:85.00   3rd Qu.:1.0000             3rd Qu.:8.000
 Max.   :9.000   Max.   :99.00   Max.   :1.0000             Max.   :9.000
 Sample.Question.Papers.Practiced Performance.Index
 Min.   :0.000                    Min.   : 10.00
 1st Qu.:2.000                    1st Qu.: 40.00
 Median :5.000                    Median : 55.00
 Mean   :4.583                    Mean   : 55.22
 3rd Qu.:7.000                    3rd Qu.: 71.00
 Max.   :9.000                    Max.   :100.00
```

The mean of “Hours Studied” is 5 hours, and the standard deviation is about 2.6 hours. The maximum number of hours of study is 9 hours.

The mean of “Previous Scores” is 69.45 points, and the standard deviation of that is about 17.34 points.The maximum number of previous scores is 99 points.

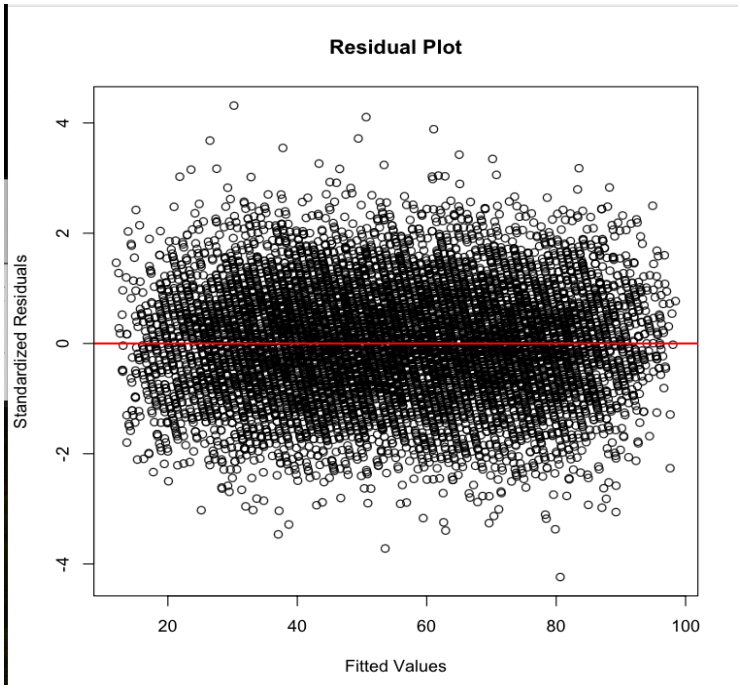
The mean of “Sleep Hours” is 6.53 hours, and the standard deviation of “Sleep hours” is 1.7 hours. The maximum number of sleeping hours is 9 hours.

The mean of “Sample Question Papers Practiced” is about 4.58 questions, and the standard deviation is about 2.87 questions. The maximum number of sample question papers practiced is 9 questions.

The mean of the “Performance Index” is 55.22 points, and the standard deviation is 19.21 points. The maximum number of performance indexes is 100 scores.

Outline of analysis

To decide which algorithm and models should be applied or implemented, I want to fit the data by making use of residual plot to choose either multi linear regression or multi regression. Specifically, if the residual plots indicate systematic tendency, the linear regression model will not be adequate and polynomial regression should be considered.



Based on the above residual plot, I think the residuals are randomly distributed over the range of fitted values, and they do not have a discernible pattern or trend. Therefore the assumption of linear regression is valid.

In terms of homoscedasticity, no funnel-shaped distribution can be found which means the variance consistency assumption of the error term is valid.

In general, from the observation of the residual plot, it is reasonable to use a linear regression model.

Data analysis

Model Parameter Selection

I used subset selection to choose the variables in the regression model. Specifically, an exhaustive selection algorithm was applied, which selects a subset containing from 1 to 5 variables.

This output below demonstrates the best variable selection. Each row represents a subset containing a different number of variables. The variables in each subset were marked with "*"indicating they are selected in the model.

```
> summary(best_subset)
Subset selection object
Call: regsubsets.formula(Performance.Index ~ ., df)
5 Variables  (and intercept)
                                Forced in Forced out
Hours.Studied                       FALSE      FALSE
Previous.Scores                     FALSE      FALSE
Extracurricular.Activities          FALSE      FALSE
Sleep.Hours                         FALSE      FALSE
Sample.Question.Papers.Practiced    FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         Hours.Studied Previous.Scores Extracurricular.Activities Sleep.Hours
1  ( 1 ) " "           "*"             " "                        " "
2  ( 1 ) "*"           "*"             " "                        " "
3  ( 1 ) "*"           "*"             " "                        "*"
4  ( 1 ) "*"           "*"             " "                        "*"
5  ( 1 ) "*"           "*"             "*"                        "*"
         Sample.Question.Papers.Practiced
1  ( 1 ) " "
2  ( 1 ) " "
3  ( 1 ) " "
4  ( 1 ) "*"
5  ( 1 ) "*"
```

Finally, In row 5, there are five variables which were marked with "*" which means there are five variables that can be selected as best variables in my linear regression model.

I used the name() function to generate the output of the best subsets regression model summary information.

```
> best.summary <- summary(best_subset)
> names(best.summary)
[1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
> best.summary$rsq
[1] 0.8375712 0.9858724 0.9876498 0.9884981 0.9887523
```

"which": the variables (features) included in each subset model.

"rsq": the R^2 value of each subset model, indicating the percentage of total variation explained by the model.

"rss": the residual sum of squares of each subset model, indicating the variation not explained by the model.

"adjr2": the adjusted R^2 value of each subset model, adjusting for the effect of the number of features on R^2.

"cp": Mallows' C_P value, one of the criteria used for model selection.

"bic": Bayesian Information Criterion, one of the criteria used for model selection.

"outmat": the matrix containing the features of each subset model.

"obj": internal object, usually containing intermediate data or results used for calculations.'

best. su,,ary$rsq is a vector containing R^2 values for each subset model, which indicates the goodness of fit of each model. The R^2 values for each subset model can be seen in the above screenshot.

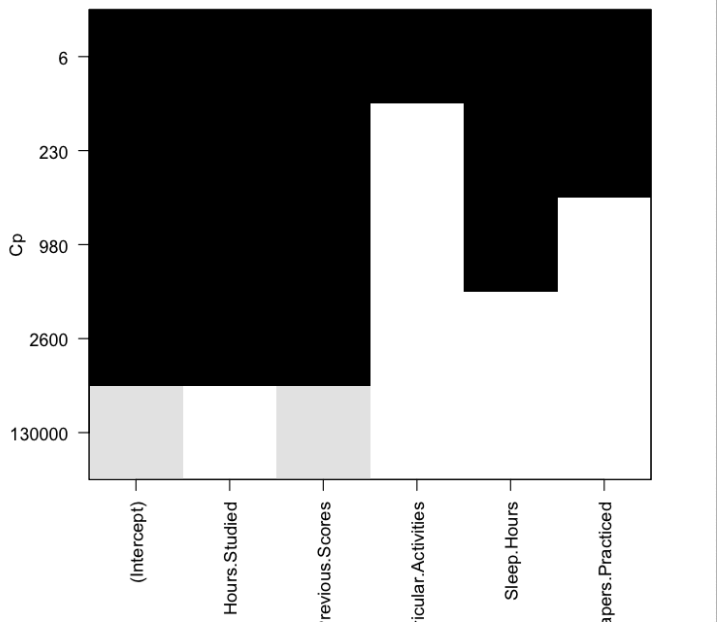
Among these values, the closer the value is to 1, the stronger the explanatory power of the model. The fifth subset model which contains five variables has the highest R^2 value which is 0.9887523. Therefore, I used this subset model to do the data analysis.

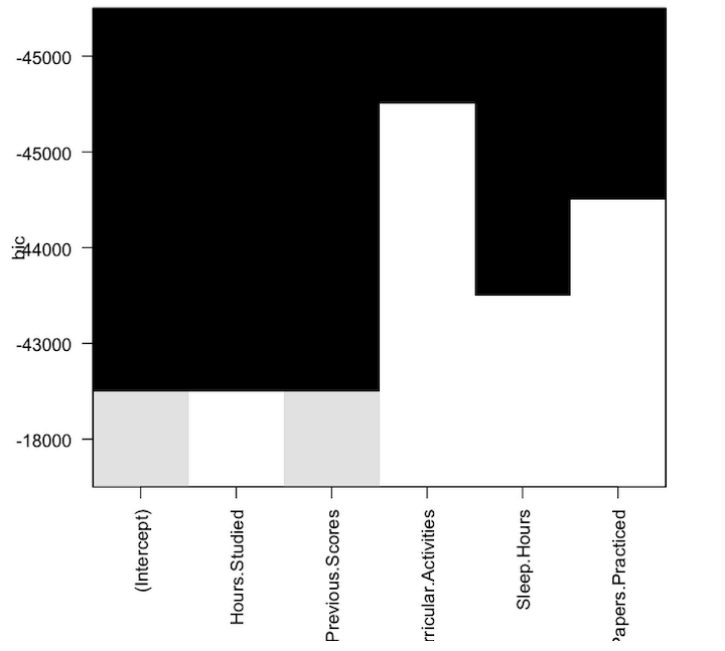
Similarly, to reconfirm which variables can be selected in the model, I used which.min() function to find the variables with minimium adjusted R^2.

```
> best_features <- best.summary$which[which.max(best.summary$adjr2), ]
> best_features
                     (Intercept)                     Hours.Studied
                            TRUE                              TRUE
                Previous.Scores       Extracurricular.Activities
                            TRUE                              TRUE
                    Sleep.Hours Sample.Question.Papers.Practiced
                            TRUE                              TRUE
```

Based on the above output, every feature has a Boolean value of TRUE, which means that the model can more accurately explain and do predictions with these five variables.

Although higher R^2 values indicate that the model has stronger explanatory power, R^2 value is not enough to choose the best model. Bayesian Information Criterion (BIC), and Mallow's C_P are also needed. I used the plot() function to visualize the BIC and C_P among 5 variables.





Based on the above plots, Black column means the BIC or C_P value when the feature is included. White area indicates the BIC or C_P value when the feature is not included.

In conclusion, after observation, I think the features "Extracurricular.Activities" and "Sleep.Hours" have less contribution to the model and could cause the model to overfit.

The other features, however, "Hours.Studied", "Previous.Scores", and "Sample.Question.Papers.Practiced" have lower BIC and Mallows' C_P values, which means that they have good contributions to the model.

Hypothesis test by using anova()

To better verify the conclusion I made previously and decide whether I should eliminate two irrelevant variables("Extracurricular.Activities" and "Sleep.Hours") , I did a hypothesis test.

Null Hypothesis:
There is no significant difference in the fitting effect between the simplified model and the complete model.
$\beta_{Extracurricular.Activities} = 0$, $\beta_{Sleep.Hours} = 0$

Alternative Hypothesis:
There is significant difference in the fitting effect between the simplified model and the complete model.
$\beta_{Extracurricular.Activities} \neq 0$, $\beta_{Sleep.Hours} \neq 0$

```
> # anova test
> full_model <- lm(Performance.Index ~ Hours.Studied + Previous.Scores +
Extracurricular.Activities + Sleep.Hours + Sample.Question.Papers.Practiced, data = df)
> reduced_model <- lm(Performance.Index ~ Hours.Studied + Previous.Scores +
Sample.Question.Papers.Practiced, data = df)
> print(anova_result)
Analysis of Variance Table

Model 1: Performance.Index ~ Hours.Studied + Previous.Scores +
Sample.Question.Papers.Practiced
Model 2: Performance.Index ~ Hours.Studied + Previous.Scores + Extracurricular.Activities +
    Sleep.Hours + Sample.Question.Papers.Practiced
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1   9996 48976
2   9994 41514  2    7462.6 898.28 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There are two different models that I made comparsion.

Simplified model (Model 1): Contains the three variables (“Hours.Studied”, “Previous.Scores”, and “Sample.Question.Papers.Practiced”).

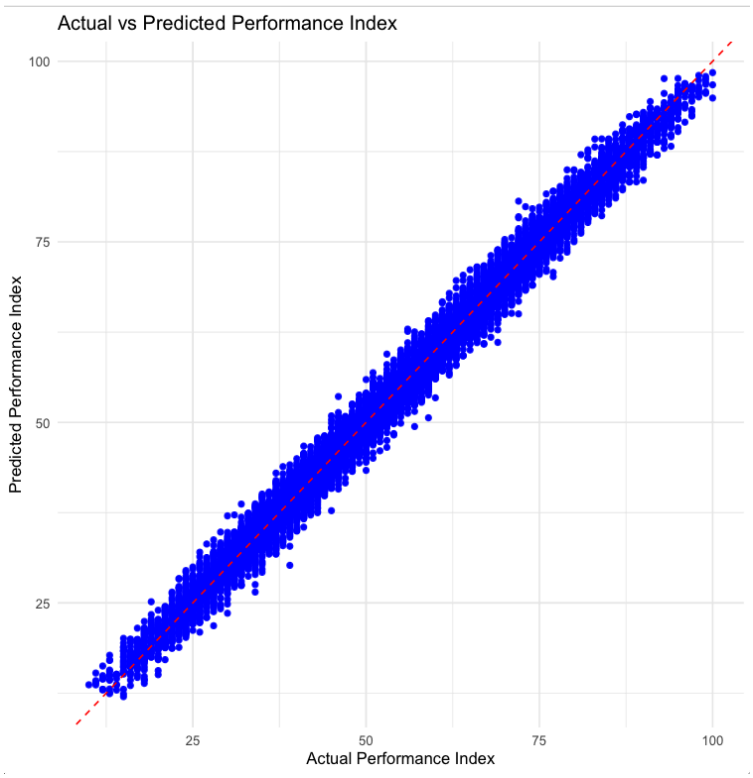
Complete model (Model 2): Contains the five variables(“Hours.Studied”, “Previous.Scores”, “Extracurricular.Activities”, “Sleep.Hours”, and “Sample.Question.Papers.Practiced”).

Based on the above output, RSS (Residual Sum of Squares) reduced from 48,976 to 41,514. The reduction means that the full model fits better than the reduced model. In terms of p-value, the p-value of the complete model is less than 2.2e-16, which is much smaller than 0.05, indicating that the adding variables has significantly improved the model.

In conclusion, adding the “Extracurricular.Activities” and “Sleep.Hours” significantly improved the fit of the model. Although these variables seemed to contribute less to the model in the previous BIC and Mallows' Cp analysis, they significantly made contributions to the explanatory power of the model based on analysis of variance (ANOVA). Therefore, I think the complete model is statistically significantly better than the simplified model.

Goodness of fit

In order to more intuitively show the degree of model fit, I used the ggplot() function to plot Relationship between” Actual Performance Index” and “Predicted Performance Index”.



In the above plot, X-axis is “Actual Performance Index” and Y-axis is “Predicted Performance Index”.
Scatter plot stands for observation, and the position of the point is the relationship between its actual value and predicted value.

Red dotted line represents the prediction result under ideal conditions, that is, the diagonal line (45 degree line) where the actual value is equal to the predicted value.

Based on the analysis, I found that most of the points are centered near the red dotted line, which means the predicted values given by the model are very close to the actual values.The data points are also closely distributed, meaning the consistency between the predicted values and the actual values is high.

In conclusion, I think his graph shows a regression model which has high accuracy so the model is able to predict the student performance index accurately.

Parameter Estimation

To better demonstrate the effect from each independent variable on the response variable (Performance.Index), I used summary() function to output the coefficient information of the linear regression model model_best.

```
> summary(model_best)

Call:
lm(formula = Performance.Index ~ Hours.Studied + Previous.Scores +
    Extracurricular.Activities + Sleep.Hours + Sample.Question.Papers.Practiced,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-8.6333 -1.3684 -0.0311  1.3556  8.7932

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -34.075588   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 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.038 on 9994 degrees of freedom
Multiple R-squared:  0.9888,  Adjusted R-squared:  0.9887
F-statistic: 1.757e+05 on 5 and 9994 DF,  p-value: < 2.2e-16
```

Based on the above output, I interpreted the coefficients and the p-values of each variable.

(Intercept): The intercept term, with an estimated value of -34.075588, which means that the value of Performance.Index is -34.07558 when all variables are zero.

Hours.Studied: The coefficient is 2.852982, which means that for every unit increase in study time, Performance.Index increases by 2.852982 units, and the coefficient is significant ($p < 2e-16$).

Previous.Scores: The coefficient is 1.018434, which means that for every unit increase in previous scores, Performance.Index increases by 1.018434 units, and the coefficient is significant ($p < 2e-16$).

Extracurricular.Activities: The coefficient is 0.612898, which means that for every unit increase in extracurricular activities, Performance.Index increases by 0.612898 units, and the coefficient is significant ($p < 2e-16$).

Sleep.Hours: The coefficient is 0.480560, which means that for every unit increase in sleep time, Performance.Index increases by 0.480560 units, and the coefficient is significant ($p < 2e-16$).

Sample.Question.Papers.Practiced: The coefficient is 0.193802, which means that for every unit increase in the number of practiced sample questions, Performance.Index increases by 0.193802 units, and the coefficient is significant ($p < 2e-16$).

In conclusion, all the independent variables significantly and statistically have positive effects on response variables (Performance.Index).

Interpretation and conclusion

Problem Background

The main purpose of doing this analysis is to figure out the factors that affect students' academic performance (Performance Index) and to make predictions by constructing a multiple linear regression model. The model I constructed includes five variables (Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, Sample Question Papers Practiced).

Data Analysis and Model Construction

Every variable’s distribution can be explored through data exploration and visualization. Afterward, I constructed a multiple linear regression model with the above variables as independent variables and academic performance (Performance Index) as the response variable.

Model Results

The effects of each variable on students' academic performance can be found in result of multiple linear regression, which are listed as follows:

Study time (Hours Studied): The longer the study time, the better the academic performance.

Previous Scores: There is a strong positive relationship between previous scores and current academic performance. Students who have high previous scores also have better current academic performance.

Extracurricular Activities: Students who participate in extracurricular activities have a slight improvement in academic performance.

Sleep Hours: Appropriate sleep hours have a positive impact on academic performance, and too long or too short sleep hours may be detrimental to academic performance.

Number of sample question papers practiced: The more sample questions exercises, the better the student's academic performance.

Conclusion

Based on previous analysis, I found that “study time”, “previous scores” and “the number of sample questions practiced” are the key factors that affect “students' academic performance”. Extracurricular activities and sleep hours however have a relatively small impact on academic performance.

Suggestions

Based on the above analysis, I sincerely come up with the following suggestions:

Increase study time: Students should be encourgaed to be more dedicated to study to improve academic performance.（Don't be lazy!)

Basic grades: Having solid basic knowledge improves students' basic grades very sufficiently. (A house built on sand will fall)

More exercises: Encourage students to do more exercises, consolidate knowledge through practice, and improve academic performance.

Reasonable amount of extracurricular activities: participating in extracurricular activities without affecting learning assists academic development.

Maintain proper sleep: Ensure that students have sufficient sleep time around 7 hours to maintain a metal state.

Finally I think these above conclusions and suggestions are not only applicable to the data set of this study I conducted, but also have its values to be referred to by most of the students in real life.Students can effectively improve their academic performance (GPA) or their standardized tests (SAT, A-level, AP test) by reasonably arranging time of study and leisure.

Appendix