Final Project

Introduction

This data is designed to examine the factors influencing academic student performance, which consists of 10,000 student records with information about various predictors and a performance index.

In the following sections, we will test five factors (Hour Studies, Previous Scores, Extracurricular Activities, Sleep Hours, and Sample Question Papers Practiced) to analyze the effect on the performance Index.

Data Description

The Predictor Variables:

1) Hours Studies: The total number of hours spent studying by each student
2) Previous Scores: The score obtained by students in the previous test
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 Variable:

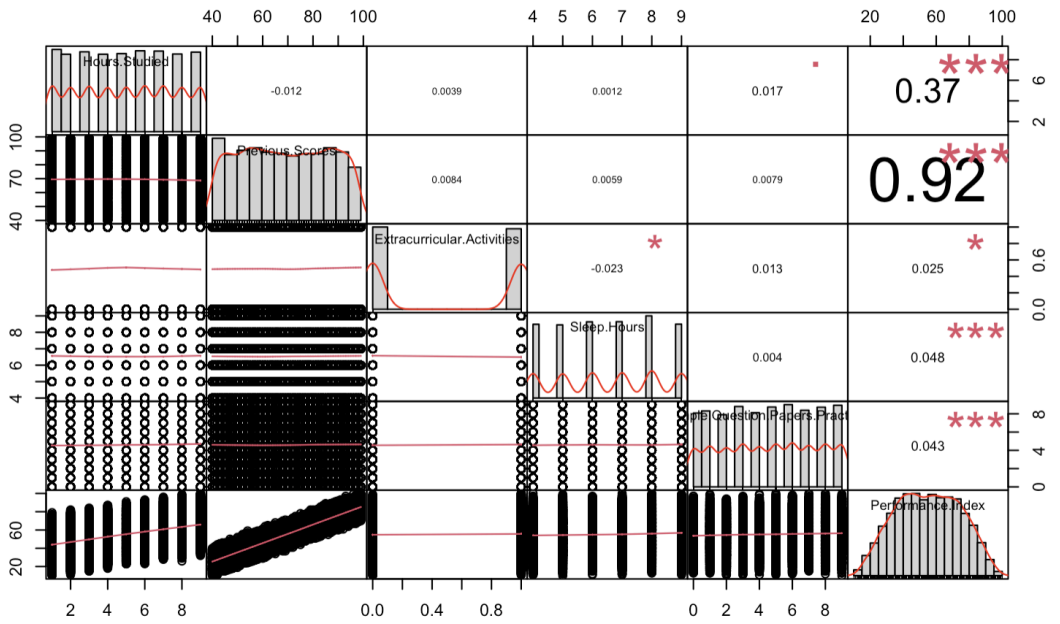
Performance Index: A measure of the overall performance of each student. The performance index represents the student's academic performance and has been rounded to the nearest integer. The index ranges from 10 to 100, with higher values indicating better performance.

	Hours Studies	Previous Scores	Extracurricular Activities	Sleep Hours	Sample Question Paper Practiced	Performance Index
Mean	4.9929	69.4457	0.4948	6.5306	4.5833	55.2248
Standard Deviation	2.58931	17.3432	0.4999	1.6959	2.8673	19.2126
Variance	6.70425	300.785	0.2499	2.8760	8.2217	369.1124
Correlation to response	0.3737	0.9152	0.0245	0.0481	0.0433	1.0000

Notice: For Extracurricular Activities, I set the "Yes" equal to 1 and "No" equal to 0 to summarize the variables better.

Fit Linear Regression Model

Visual Correlation



From this figure, we can see that the predictor variable Previous Scores has the highest correlation with the response variable Performance Index. Moreover, we also get to know there are no two predictor variables that have a high correlation, so we can include all predictor variables in the model.

ANOVA for the Entire Model

```
Analysis of Variance Table

Model 1: Performance.Index ~ 1
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   9999 3690855
2   9994   41514  5   3649341 175709 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the figure, we set the null hypothesis which means there is no significant difference between the means of the groups being compared, and the alternative hypothesis is there is at least one significant difference between the means of the groups. The P-value we get is 2.2e-16,

which means we have enough evidence to reject the null hypothesis and conclude there is at least one predictor is significant.

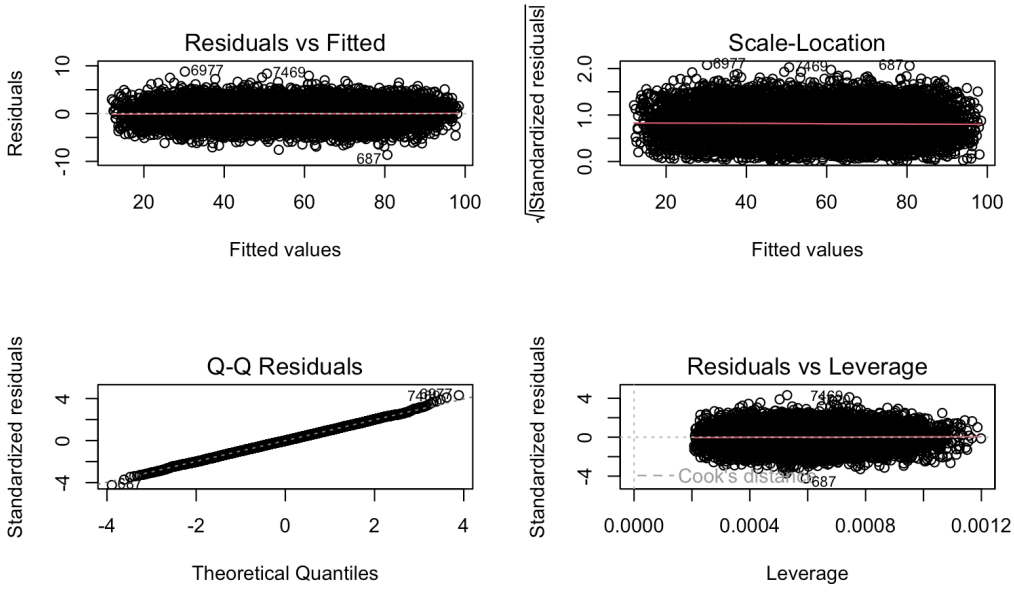
Partial F-test

```
Call:
lm(formula = Performance.Index ~ ., data = data)

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
```

From this figure, we can know the estimates for each predictor and see the p-values of all predictor variables are less than 0.05, which means they are all significant. Therefore, we can conclude that the full model is better than the reduced model.

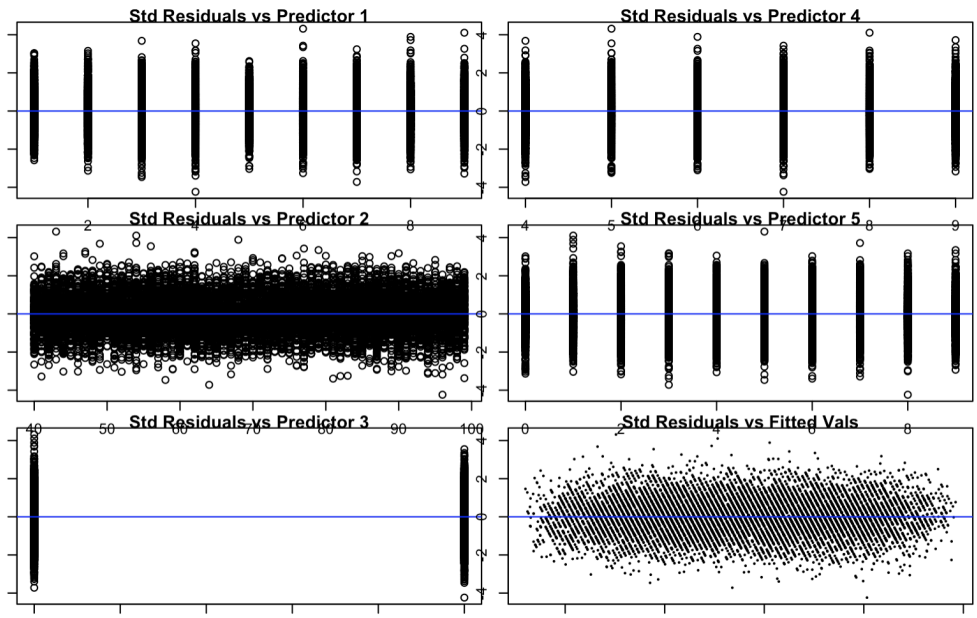
Diagnostics of Assumption

Four Model Diagnosis Plot



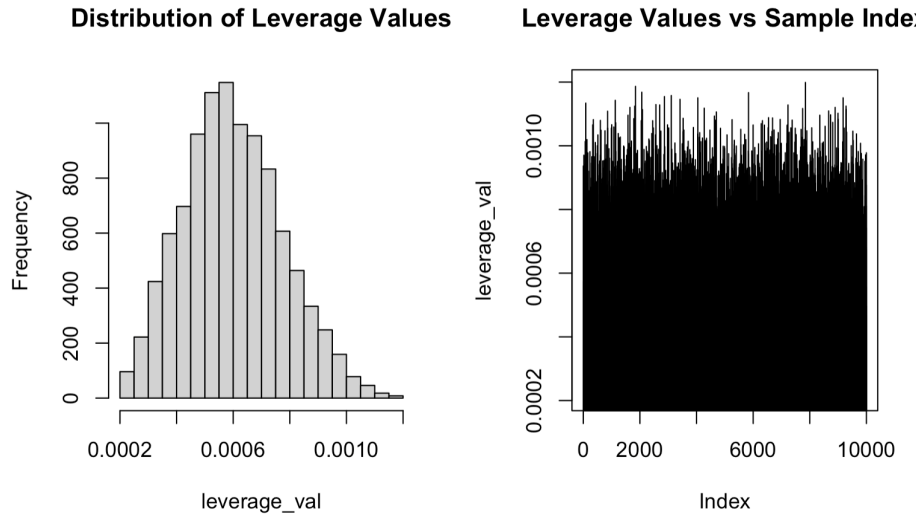
The plots of the standardized residuals vs. predictor of predictors show the mean of residuals is 0 and constant variability, so this model is valid.

Standardized Residuals



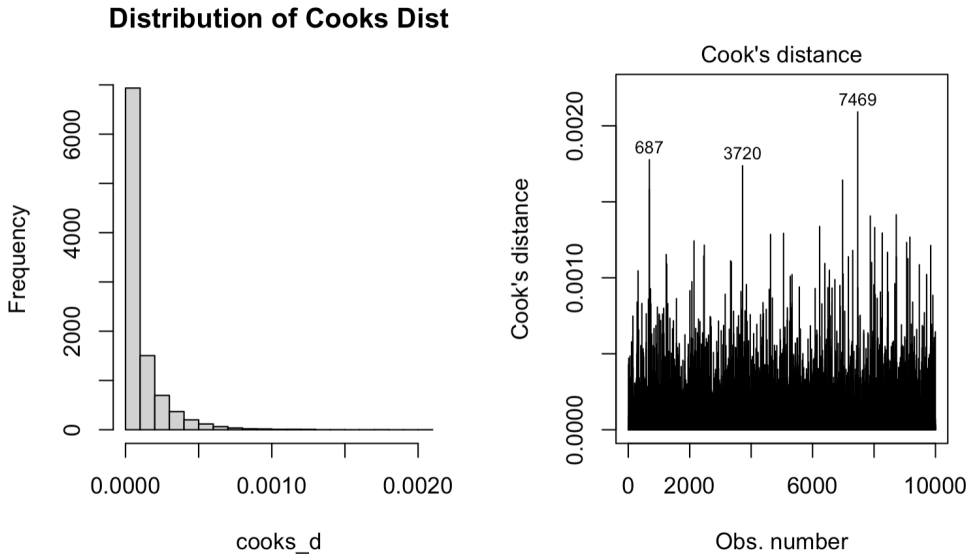
As we know, the plot of the standardized residuals against any predictor of the predictors will show 1) a random scatter of points around the horizontal axis, and 2) constant variability as we look along the horizontal axis. Therefore, the plots of the standardized residuals vs. predictors show that the model is valid.

Leverage



Leverage points for MLR are defined when $h_{ii} > 2 * average(h_{ii}) = 2 * \frac{p+1}{n}$. After calculation, there are no high leverage points.

Cook's Distance



Influential points for MLR are defined when Cook's distance is greater than $\frac{4}{n-2}$. After calculation, there are 493 points with a large Cook's distance.

.

Conclusion

Based on the information we get above, all predictor variables are significant for the response variable Performance Index, which shows Hour Studies, Previous Scores, Extracurricular Activities, Sleep Hours, and Sample Question Papers Practiced all impact the Performance. People who get a high score in previous exams have a big chance to perform better.

All in all, performing great is not major influenced by a single factor. If the student wants to get a high-performance index, they should spend time studying, get good grades on previous scores, join curricular activities, get enough sleep, and practice sample questions. Then, they can get a great performance.

Data Cited

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