STATS 101A - Final Project: Predicting Academic Performance

Jade Liang, Riya Hariyaplar, Bhavya Sankarappan, Zijia Zhang,
Yubin Hong, Priscilla Lim, Rosalinda Chen

Introduction

The goal of this research is to analyze the extent to which these factors contribute to students' current academic performance, measured by the Performance Index. By including continuous variables, the study aims to provide a holistic view of the determinants affecting student outcomes.

The dataset "Student Performance (Multiple Linear Regression)" from Kaggle contains information about various factors influencing student academic performance. The dataset contains 10,000 observations and 6 variables. It includes data on students' previous academic scores, hours studied, sleep hours, participation in extracurricular activities, and whether they practiced sample question papers. This data is used to analyze the impact of these factors on the students' Performance Index, a measure of their current academic performance.

- Performance Index: A measure of the overall performance of each student from 10 to 100
- Hours Studied: The total number of hours spent studying by each student
- Previous Scores: The scores obtained by students in previous tests
- Extracurricular Activities: Whether the student participates in extracurriculars (Yes or No)
- Sleep Hours: The average number of hours of sleep the student had per day
- Sample Question Papers Practiced: The number of sample question papers the student practiced

Research Question: How do hours studied, previous academic scores, sleep hours, and the practice of sample question papers influence students' current academic performance, as measured by the Performance Index?

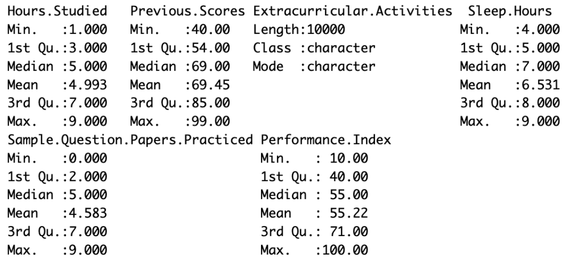
The model focuses on numerical variables because they offer detailed and quantifiable insights directly related to student performance. This choice simplifies the model by providing straightforward relationships and easy interpretation, enhancing its clarity and usability. Additionally, these numerical variables demonstrate significant correlations with the target variable and exhibit low multicollinearity, contributing to the model's statistical robustness and effectiveness.

Our paper looks at the summary statistics and distribution of the variables of the generalized full model, excluding Extracurricular Activities. We see that transformation, but not variable selection, is needed to better fit the variables. We tested three models and after investigating, a final model was chosen.

Data Description

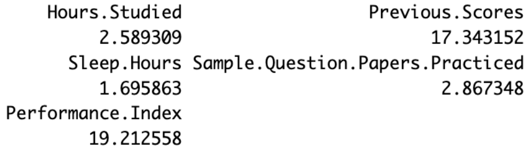
Summary Statistics

Figure 1 shows the general summary statistics for all the numerical variables (exclude Extracurricular.Activities) and the target variable Performance.Index. From this, we are able to see the mean, median, upper and lower quartile, and min and max of each numerical variable.



```
Hours.Studied   Previous.Scores Extracurricular.Activities  Sleep.Hours
Min.   :1.000   Min.   :40.00   Length:10000               Min.   :4.000
1st Qu.:3.000   1st Qu.:54.00   Class :character           1st Qu.:5.000
Median :5.000   Median :69.00   Mode  :character           Median :7.000
Mean   :4.993   Mean   :69.45                              Mean   :6.531
3rd Qu.:7.000   3rd Qu.:85.00                              3rd Qu.:8.000
Max.   :9.000   Max.   :99.00                              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
```

Figure 1. R output for data summary

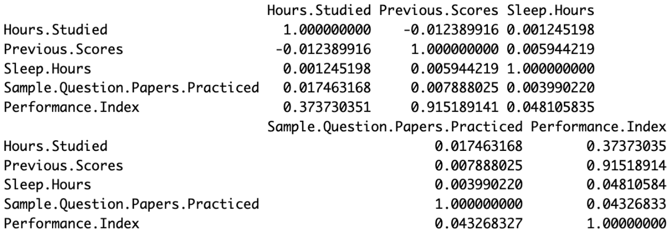


```
    Hours.Studied                  Previous.Scores
         2.589309                        17.343152
      Sleep.Hours Sample.Question.Papers.Practiced
         1.695863                         2.867348
Performance.Index
        19.212558
```

Figure 2. R output for standard deviation of numerical variables

Figure 2 shows the standard deviation of each numerical variable, which is basically the variation from the mean for each variable.

Lastly, we try to find the correlation between variables. Figure 3 shows the correlation matrix.



```
                                 Hours.Studied Previous.Scores Sleep.Hours
Hours.Studied                      1.000000000    -0.012389916 0.001245198
Previous.Scores                   -0.012389916     1.000000000 0.005944219
Sleep.Hours                        0.001245198     0.005944219 1.000000000
Sample.Question.Papers.Practiced   0.017463168     0.007888025 0.003990220
Performance.Index                  0.373730351     0.915189141 0.048105835
                                 Sample.Question.Papers.Practiced Performance.Index
Hours.Studied                                         0.017463168        0.37373035
Previous.Scores                                       0.007888025        0.91518914
Sleep.Hours                                           0.003990220        0.04810584
Sample.Question.Papers.Practiced                      1.000000000        0.04326833
Performance.Index                                     0.043268327        1.00000000
```

Figure 3. R output for correlation matrix

As seen from Figure 3, Previous.Scores shows the strongest positive correlation with the target variable Performance.Index, suggesting that a student's previous scores is likely to be strongly correlated to their performance index. In contrast, Sample.Question.Papers.Practiced shows the lowest positive correlation with Performance.Index, suggesting that increasing the number of sample question papers practiced is the least effective in improving a student's performance. Also, generally the correlation between the non-target variables is very small and close to 0. Hence, the weak relationships between non-target variables is unlikely to affect the relationship between the non-target variables and the target variable.

Distribution of Variables and Relationships among Variables

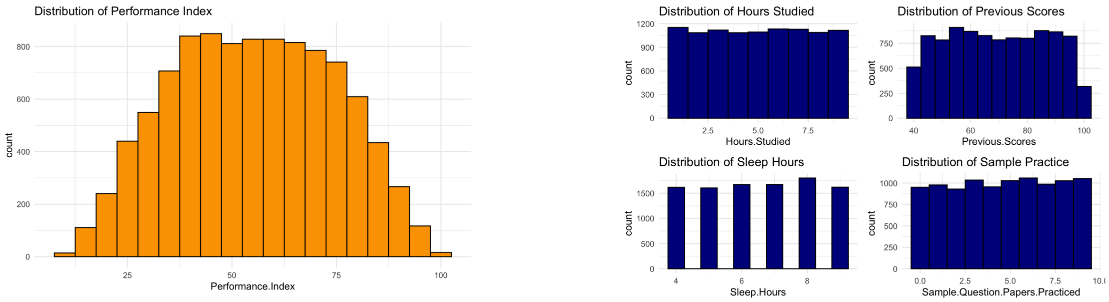


Figure 4. Distribution of variables

We see that our response variable (Performance.Index) is normally distributed from Figure 1. The four predictor variables are uniformly distributed.

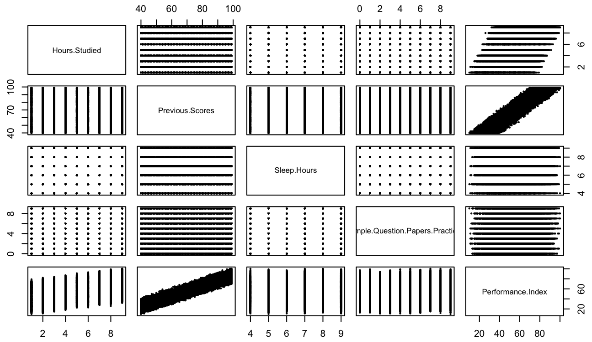
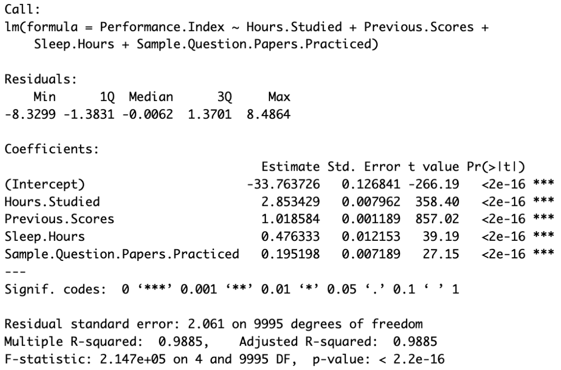


Figure 5. Scatterplot matrix of variables

Based on the scatter plot in Figure 2, Previous.Scores seem to have a strong positive linear relationship with Performance.Index. The variables Hours.Studied and Sample.Question.Papers.Practice appear to have a positive linear relationship with Performance.Index as well, but Sample.Question.Papers.Practice does not look as strong as the other two predictors. There does not seem to be an apparent relationship between Sleep.Hours and Performance.Index, which we will further investigate through our model. Additionally, there does not appear to be any clear relationship between the four predictor variables, which is a positive sign of a lack of multicollinearity. We will begin our analysis by fitting the data with a multiple linear regression model, given the linear relationship between some of our predictors variables and the response variable.

Results and Interpretation

Model 1:



```
Call:
lm(formula = Performance.Index ~ Hours.Studied + Previous.Scores +
    Sleep.Hours + Sample.Question.Papers.Practiced)

Residuals:
    Min      1Q  Median      3Q     Max
-8.3299 -1.3831 -0.0062  1.3701  8.4864

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -33.763726   0.126841 -266.19   <2e-16 ***
Hours.Studied                     2.853429   0.007962  358.40   <2e-16 ***
Previous.Scores                   1.018584   0.001189  857.02   <2e-16 ***
Sleep.Hours                       0.476333   0.012153   39.19   <2e-16 ***
Sample.Question.Papers.Practiced  0.195198   0.007189   27.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.061 on 9995 degrees of freedom
Multiple R-squared:  0.9885,	Adjusted R-squared:  0.9885
F-statistic: 2.147e+05 on 4 and 9995 DF,  p-value: < 2.2e-16
```

Figure 6. R output for model m1

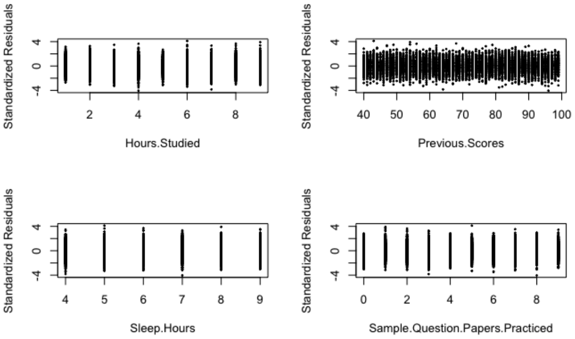
As seen from Figure 3, the multiple R^2 is 0.9885 which suggests that 98.85% of the variability in the performance index of the student is explained by the model m1. However, given that there are many terms in the model and the addition of each predictor results in a decrease in RSS and a resultant increase in R^2, this may be a misleadingly high R. Hence, we check the adjusted R, but as seen from Fig. 3, this is also 0.9885. Hence, we can conclude that a large portion of the variability of the performance index of the student is explained by the model m1.

We derive the linear regression equation from above:

$$\widehat{Performance.Index} = -33.76 + 2.85\ \textit{Hours.Studied} + 1.02\ \textit{Previous.Scores}$$
$$+\ 0.48\ \textit{Sleep.Hours} + 0.20\ \textit{Sample.Question.Papers.Practiced}$$

Diagnostic Plots for Model 1

The plots below show a random scatter of points around y = 0. It also appears to have constant variability. Thus it doesn't seem to violate any model assumptions.



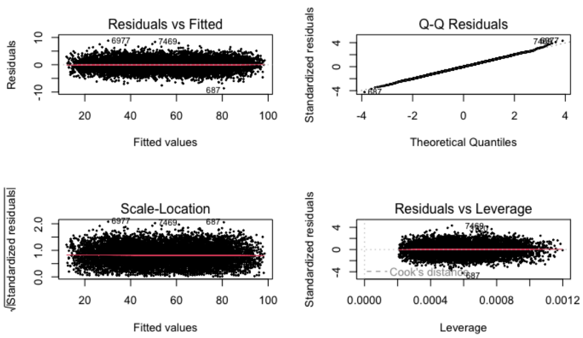


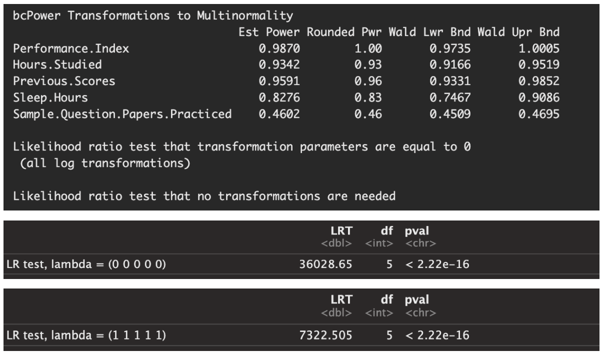
Figure 7. Standardized Residuals vs. Each Predictor

Figure 8. Diagnostic Plots for Model 1

In Figure 8, the residuals plot shows that the mean of residuals is around 0 and the plot doesn't show any pattern, which doesn't violate the model assumption of linearity. The Normal Q-Q plot implies there's normality of the errors as the plot appears to be a nearly straight line. The Scale-Location plot does not seem to have an apparent pattern, which indicates constant variance. In the Residuals vs Leverage plot, we feel confident to expand the range for outliers from [-2,2] to [-4,4] because we have a large dataset with 10,000 observations.Then, most of the points are within the range of [-4,4]. Therefore, none of the model assumptions seem to be violated.

Model 2: Model After Power Transformation

Using the Box-cox method, R rejects the null hypothesis that log transformation is needed for all variables. It also rejects the null hypothesis that no transformation is needed for all variables. According to the R output in figure 8, this means we need to perform power transformation for the variables Hours.Studied, Previous.Scores, Sleep.Hours, and Sample.Question.Papers.Practiced. No transformation is needed for Performance.Index as the suggested power for its power transformation is λ = 1.



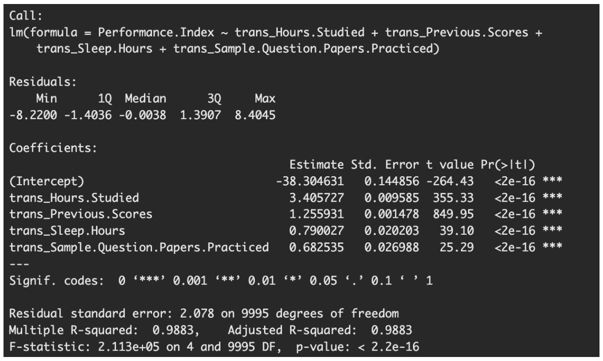
```
bcPower Transformations to Multinormality
                                 Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Performance.Index                   0.9870        1.00       0.9735       1.0005
Hours.Studied                       0.9342        0.93       0.9166       0.9519
Previous.Scores                     0.9591        0.96       0.9331       0.9852
Sleep.Hours                         0.8276        0.83       0.7467       0.9086
Sample.Question.Papers.Practiced    0.4602        0.46       0.4509       0.4695

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

```
                                        LRT     df   pval
                                      <dbl>  <int>  <chr>
LR test, lambda = (0 0 0 0 0)      36028.65      5  < 2.22e-16
```

```
                                        LRT     df   pval
                                      <dbl>  <int>  <chr>
LR test, lambda = (1 1 1 1 1)      7322.505      5  < 2.22e-16
```



```
Call:
lm(formula = Performance.Index ~ trans_Hours.Studied + trans_Previous.Scores +
    trans_Sleep.Hours + trans_Sample.Question.Papers.Practiced)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2200 -1.4036 -0.0038  1.3907  8.4045

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           -38.304631   0.144856 -264.43   <2e-16 ***
trans_Hours.Studied                     3.405727   0.009585  355.33   <2e-16 ***
trans_Previous.Scores                   1.255931   0.001478  849.95   <2e-16 ***
trans_Sleep.Hours                       0.790027   0.020203   39.10   <2e-16 ***
trans_Sample.Question.Papers.Practiced  0.682535   0.026988   25.29   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.078 on 9995 degrees of freedom
Multiple R-squared:  0.9883,    Adjusted R-squared:  0.9883
F-statistic: 2.113e+05 on 4 and 9995 DF,  p-value: < 2.2e-16
```

Figure 9. R Outputs for Box-cox Method

Figure 10. R Output for Model 2

After transforming the variables, Figure 9 shows that the adjusted R-squared for Model 2 shows a 0.0002 decrease compared to Model 2. This can indicate a slight improvement in overfitting.

Similar to figure 7, the scatter plots in figure 11 do not appear to violate any assumptions about the error term as there seems to be a random scatter around y = 0 and that the variance appears to be constant. Likewise, the plots in figure 12 also do not appear to violate any assumptions about the error term. Moreover, the added variable plots in figure 13 shows that all variables are significant.

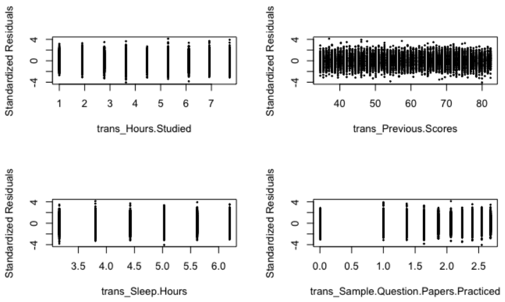


Figure 11. Standardized Residuals vs. Predictors

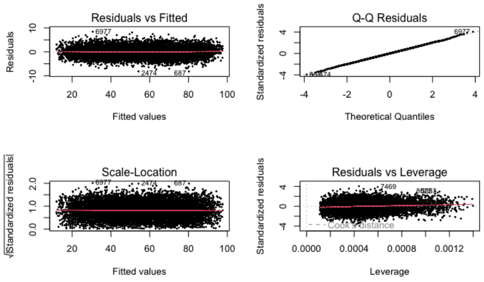


Figure 12. Diagnostic Plots for Model 2

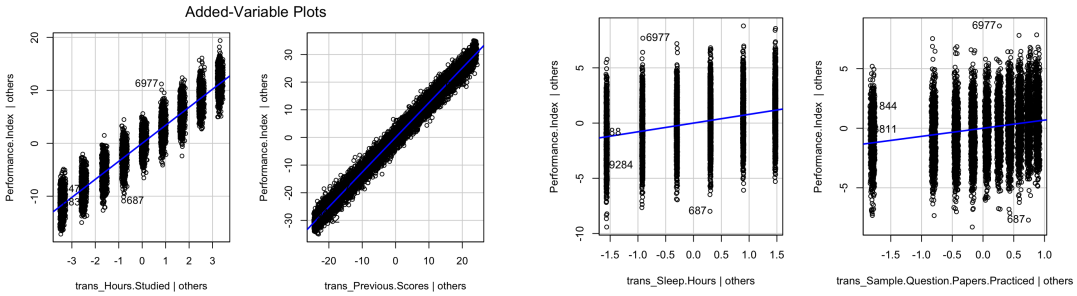
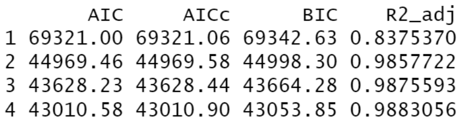


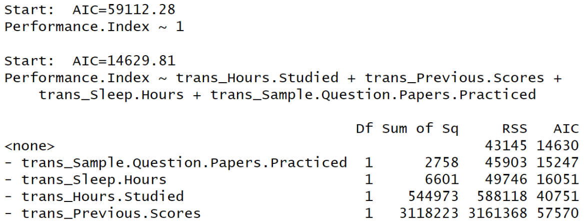
Figure 13. Added Variable Plots

Model 3: Model After Variable Selection

We tried all three approaches to see if variable selection would simplify our transformed model. First, we ran an all subsets regression and calculated AIC, AICc, BIC, and Radj^2 for all possible models. The results showed that the model with all the transformed predictors included is the best, having the smallest AIC, AICc, and BIC scores, while also having the highest Radj^2.



```
       AIC     AICc      BIC    R2_adj
1 69321.00 69321.06 69342.63 0.8375370
2 44969.46 44969.58 44998.30 0.9857722
3 43628.23 43628.44 43664.28 0.9875593
4 43010.58 43010.90 43053.85 0.9883056
```



```
Start:  AIC=59112.28
Performance.Index ~ 1

Start:  AIC=14629.81
Performance.Index ~ trans_Hours.Studied + trans_Previous.Scores + 
    trans_Sleep.Hours + trans_Sample.Question.Papers.Practiced

                                         Df Sum of Sq     RSS   AIC
<none>                                                  43145 14630
- trans_Sample.Question.Papers.Practiced  1      2758   45903 15247
- trans_Sleep.Hours                       1      6601   49746 16051
- trans_Hours.Studied                     1    544973  588118 40751
- trans_Previous.Scores                   1   3118223 3161368 57570
```

Figure 13. All Subsets for Model 3

Figure 14. Stepwise Regression for Model 3

Next, we conducted stepwise regression, both backward and forward, based on AIC. The backward regression also recommended the model with all the transformed predictors, while the forward regression recommended the model with no predictors. Comparing the AIC scores, the full model with all the transformed predictors has a much smaller AIC.

As a result, we can conclude that no variable selection is needed. The original model with all the transformed predictors is the best one.

Based on our result analysis, the best model is produced by transforming all our predictors (Hours Studied, Previous Scores, Sleep Hours, and Sample Question Papers Practiced) based on our response variable Performance Index.

The final model after transformation is:

$$Performance.Index = \beta 0 + \beta 1\ Hours.Studied^{0.93} + \beta 2\ Previous.Scores^{0.96}$$
$$+ \beta 3\ Sleep.Hours^{0.83} + \beta 4\ Sample.Question.Papers.Practiced$$

Interpretation of Model

Our predictors seem positively correlated with our response variable, although not necessarily a strictly linear relationship. This model emphasizes how factors like how study time, previous test scores, hours of sleep, number of sample papers practiced affect students' performance. Our analysis reinforces the well-known notion that factors like such improve academic performance, useful knowledge for students to keep in mind.

Discussion

The project developed a predictive model using numerical data to understand the factors influencing student performance, measured by the Performance.Index. Key variables included Previous.Scores, Hours.Studied, Sample.Question.Papers.Practiced, and Sleep.Hours. The model was based on multiple linear regression, chosen for its simplicity and interpretability.

The final model aligns well with educational research that highlights the importance of study habits, prior academic achievement, and adequate rest on student performance. The model's emphasis on Previous.Scores and Hours.Studied correlates with findings that consistent study time and strong foundational knowledge are critical to academic success. While the impact of practicing sample question papers was found to be weaker, it still reflects the real-world scenario where quality and diversity of study methods matter.

The model has several limitations: it relies solely on numerical variables, potentially missing qualitative factors like student motivation or teaching quality; it doesn't account for extracurricular activities or other personal influences; and it assumes linear relationships, which might not capture more complex interactions. To address these issues, future improvements could include incorporating categorical data such as study environments and teaching methods, expanding the sample size for better generalizability, and exploring advanced modeling techniques like non-linear models or machine learning to detect more intricate patterns. These enhancements would provide a more comprehensive understanding of the factors influencing student performance.