AMS 325 Final Project Report

**Introduction**

In this project, we are going to analyze the data “[Student\_Performance.csv](https://www.kaggle.com/datasets/nikhil7280/student-performance-multiple-linear-regression/data)” using python. In this data, there are six variables: This is a 10,000x6 data set, which means there are 10,000 samples from students, with a total of six variables: Hours Studied (numeric, 1 to 9), Previous Scores (numeric, 40 to 99), Extracurricular Activities (categorical, Yes or No), Sleep Hours (numeric, 4 to 9), Sample Question Papers Practiced (numeric, 0 to 9), and Performance Index (numeric, 10 to 100). The goal is to fit a regression model using first five as possible independent variables and performance index as dependent variable. With the model, we will be able to predict future value of student performance.

**Find Model**

We want to find the relationship between the target variable performance index and independent numeric variables. Since Extracurriculars is a categorical variable, we dropped this column. The first thing is to know which model is the best for this dataset. We tried ordinary least squares method and fit a multiple linear regression model.

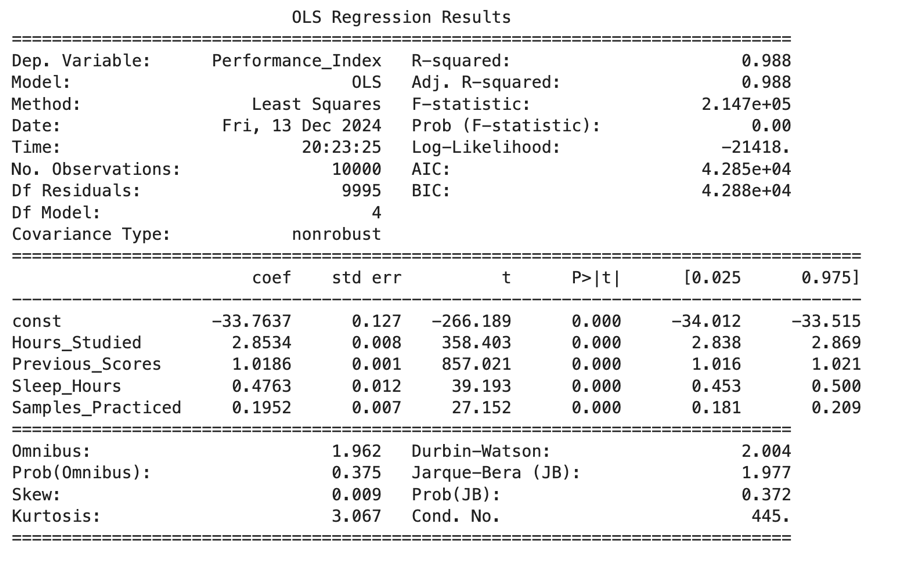


Fig.1 Output for *Performance index = ß0 + ß1 Hours Studied+ ß2 Previous Scores + ß4 Sleep Hours*

*+ ß5 Sample Question Papers Practiced + error*

The output shows adjusted R squared value equal to 0.988, meaning this model fits the dataset very well.

To check the model is linear, we want to check the residuals from the original data and the fitted values. In this case, we use QQ-plot to analyze the fitted model.

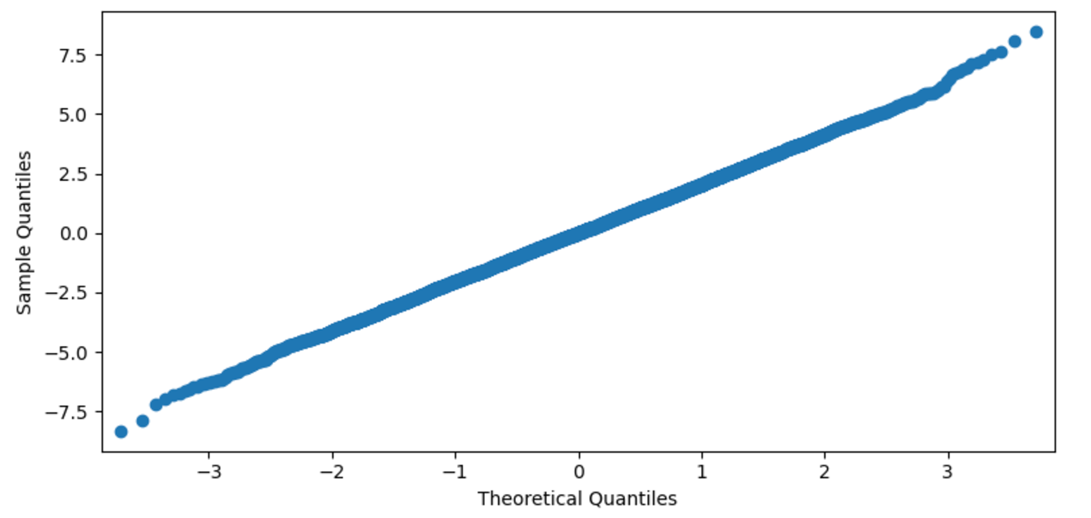


Fig. 2 QQ-plot of fitted model

The plot shows a linear pattern, indicating the error term e is normally distributed. Hence, we can conclude that this linear model is adequate for this dataset.

**Hypothesis**

We want to test that performance index is affected by Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, or Sample Question Papers Practiced. Our null hypothesis is: Performance index is not affected by other four independent variables, i.e. the coefficients are all equal to zero. The alternative hypothesis is: There is effect from independent variables. From figure 1, the P>|t| or the p-values of all five variables are closed to 0, and the absolute t values are relatively high. Therefore, we reject the null hypothesis and conclude that the coefficients are not zero.

**Multiple Linear Regression**

We fit a linear model using Hours Studied, Previous Scores, Extracurricular Activities, Sleep Hours, or Sample Question Papers Practiced as independent variables and Performance Index as dependent variables. In order to fit a good model, we should split the dataset into training dataset and testing dataset. We use the training dataset to fit our model and use testing to test the performance of this model. The model equation is:

*Performance Index = -33.76 + 2.85 \* Hours Studied + 1.02 \* Previous Scores + 0.48 \* Sleep Hours*

*+ 0.20 \* Samples Practiced*

Next, we apply the independent variables from the testing dataset to get the fitted value. The graph compares the true value of performance index and the predicted value.

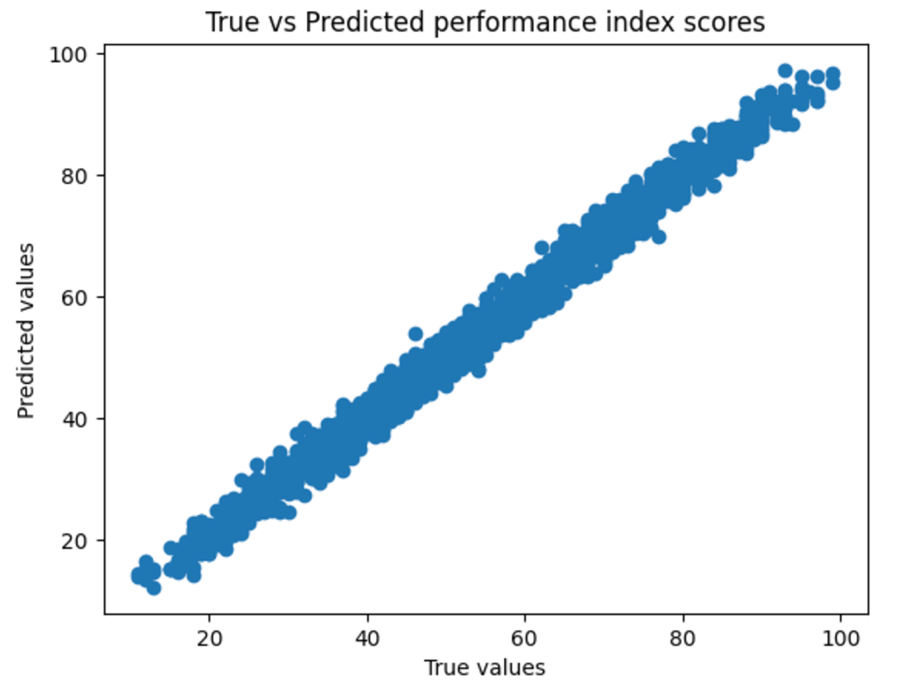


Fig. 3 Predicted Values vs True Values

The linear pattern shows a well-fitting of this linear model. The R squared of this model is 0.9884981216772581.

**Conclusion**

The dataset is well suited for multiple linear regression. From the output, we can see that the student Performance Index can be affected by Hours Studied, Previous Scores, Sleep Hours, Sample Question Papers Practiced. The model has a R squared value closed to 1, which means this model is well fitted and has a robust ability to make prediction. Recalling our model, the four variables all have positive effects the performance index. Therefore, if students want to improve their performance score, they should consider improving their behaviors on these four fields.