Goldsmiths
UNIVERSITY OF LONDON

Insights and Conclusions from Detailed Regression Analysis

A Report submitted for Statistics and Statistical Data Mining Module

December 12, 2023

Module leader: Dr Daniel Stamate

Carlos Manuel De Oliveira Alves
Student ID: 33617310

Dataset and Models Overview

The dataset used for this regression analysis primarily concerns student academic performance, with the dependent variable being the final grade (G3). The dataset includes various predictors such as age, travel time, family relationship quality, alcohol consumption, absences, and previous grade performances (G1, G2).

Several models were constructed to analyse the data

1. Model M1: This linear regression model is built on the training set. It includes multiple predictors like age, travel time (traveltime.x, traveltime.y), family relationship quality (famrel.x, famrel.y), daily alcohol consumption (Dalc.x, Dalc.y), absences (absences.x, absences.y), previous grades (G1.x, G2.x), and others. It aims to understand the linear relationship between these variables and the final grade (G3).

2. Model M2: A polynomial regression model focusing on the relationship between travel time to school (traveltime.x) and final grades (G3).

3. Model M3: Another polynomial regression model, but this one examines the relationship between age and final grades (G3).

4. Model M4: This model is similar to M2 and M3 but explores the relationship between student absences (absences.x) and final grades (G3).

5. Model M5: Constructed using the k-Nearest Neighbour (k-NN) algorithm with different values of k (1, 2, 3, 4) to find the best k value based on performance on the validation set.

6. Model Mb: The best model selected from M1, M2, M3, M4, and M5 based on their performance on the validation set.

Results of the Models

- Model M1 showed a solid ability to explain the variance in G3 with a notable R-squared value. However, both x and y versions of the same variables with opposite signs might indicate potential data issues or multicollinearity.

- Models M2, M3, and M4 demonstrated that while there are significant relationships between each pair of variables (travel time, age, absences, and final grades), these relationships are not strong enough to predict final grades accurately on their own, as indicated by the low R-squared values.

- Model M5: The k-NN algorithm revealed that k=3 provided the best performance based on the validation set, though the method's overall effectiveness in predicting final grades was not apparent.

- Model Mb: was identified as Model M1 after comparing all models' performance metrics (RMSE and R-squared) on the validation dataset.

Conclusion

The analysis revealed that while travel time, age, and absences are related to student grades, they are not solid or straightforward predictors of academic performance. Model M1 emerged as the best model in this context, demonstrating the importance of simultaneously considering a multitude of variables to understand and predict student academic outcomes effectively.

The selected Model Mb (M1) showed a better fit and predictive accuracy when evaluated on the test dataset, suggesting its reliability in predicting student academic performance. However, this study also underscores the complexity of the factors affecting academic performance and the need for comprehensive models incorporating a broader range of variables to understand better and predict student outcomes.