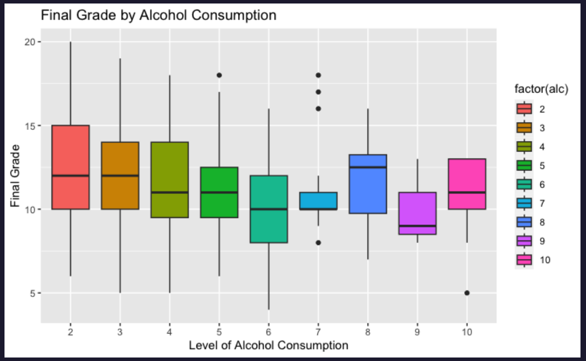
"DATA621 Final Project"

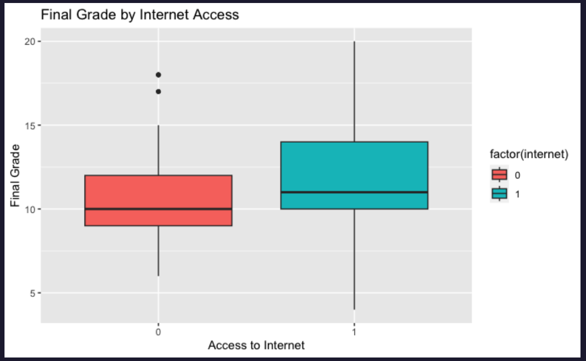
Predicting Student Achievement

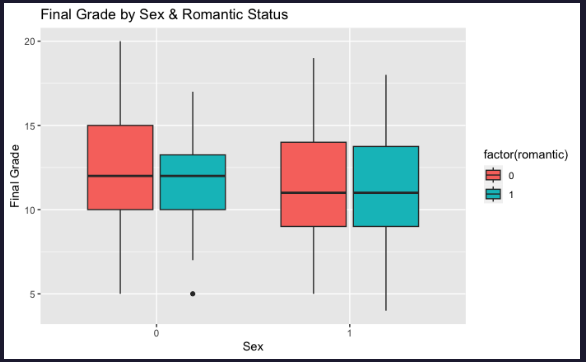
Ivan Tikhonov, Seung Min Song, Alice Friedman

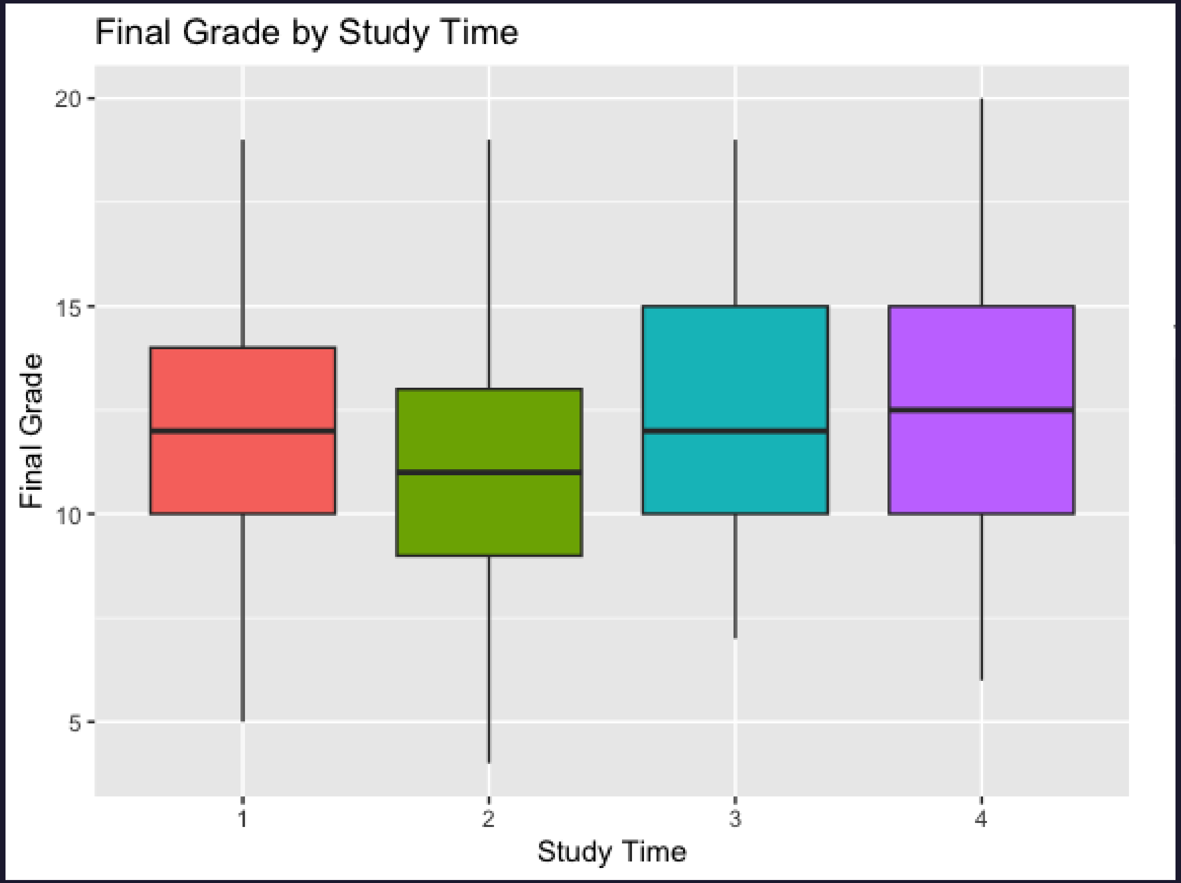
Abstract

- This study investigates the external factors affecting students' academic performance in secondary schools' math and Portuguese language courses.
- Using a quantitative dataset obtained through a survey, the study analyzes various external variables, such as students' health, social and demographic factors, study habits, and other relevant factors to understand their impact on academic performance.
- Results indicate that nonparametric methods may be useful to model this complex phenomenon.


Final Grade by Alcohol Consumption


Final Grade by Internet Access


Final Grade by Sex & Romantic Status



Introduction

- The data used in this study was collected from a survey of two secondary schools in Portugal and includes information on various external variables and their impact on students' grades, such as study time – shown at left.
- This study attempts to predict academic achievement in the form of a final grade averaged across two core subjects, Math and Portuguese.

Literature Review

SUMMARY

- Previous studies have used versions of linear modeling of ANOVA to understand the impact of social, demographic, physical, and cognitive factors on education.
- Complex models have typically performed the best in explaining variation, but resutlts have varied widely based on the exact study design as well as what subjects have studied.

CITATIONS

Application of Multiple Linear Regression Identifying Contributing Factors in Stu- dents'Academic Achievement Authors: Dg Siti Nurisya Sahirah Binti Ag Isha and Siti Rahayu Binti Mohd Hashim Proceedings of the International Conference on Mathematical Sciences and Statistics 2022 (ICMSS 2022)At: Selangor, Malaysia Mathematics With Economics Programme, Faculty of Science and Natural Resources, December 2022 https://www.researchgate.net/publication/366929441_Application_ of_Multiple_Linear_Regression_in_Identifying_Contributing_Factors_in_Students%27_Academic_ Achievement

A Study on Academic Achievement and Personality of Secondary School Students Authors: Dr. Suvarna V. D.and Dr H. S. Ganesha Bhata1 Research in Pedagogy, v6 n1 p99-108, 2016 https://files. eric.ed.gov/fulltext/EJ1149330.pdf

Predictors of Academic Performance in High School Students: The Longitudinal ASAP Study Authors: Marie-Maude Dubuc, Mylene Aubertin-Leheudr, and Antony D. Karelis Published online 2022 May 1, International Journal of Exercise https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9365103/ #:~:text=Finally%2C%20psychological%20factors%20such%20as,42%2C%2043%2C%2048)

Data Collection & Preparation

- The data used comes from survey data containing grades and student information initially divided into two data sets representing student grades in Portuguese and Spanish

- In order to proceed with the results, we first identified students in both classes based on identical characteristics across the other 30 features and merged the tables into a single table using an inner join

- The resulting table is one observation per student. Fields which are specific to one class or another are aggregated or merged in the final data set:

  - `final_grade` is the average of final grades across both classes

  - `absences_mean` is the mean of absences reported across both classes

  - `paid_tutor` is used to indicate wheteher a paid tutor was used in either subject

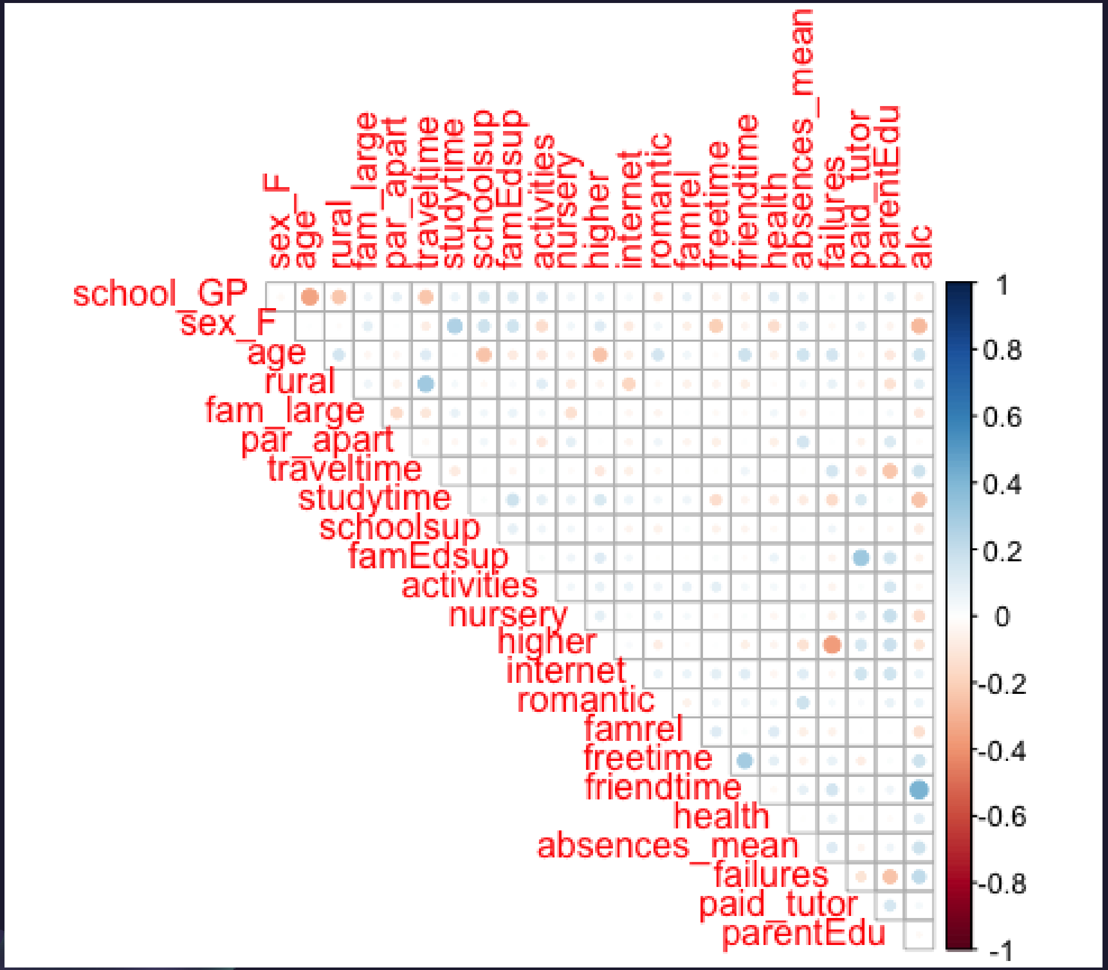
  - `missed_exam` is a field to indicate whether a student missed their final exam (score of 0) in either class.

- Finally, we removed students from consideration who had a 0 in their final grade, as that is most likely to missing the final exam. This is because we presumed to indicate absence rather than student potential for achievement

KAGGLE LINK:

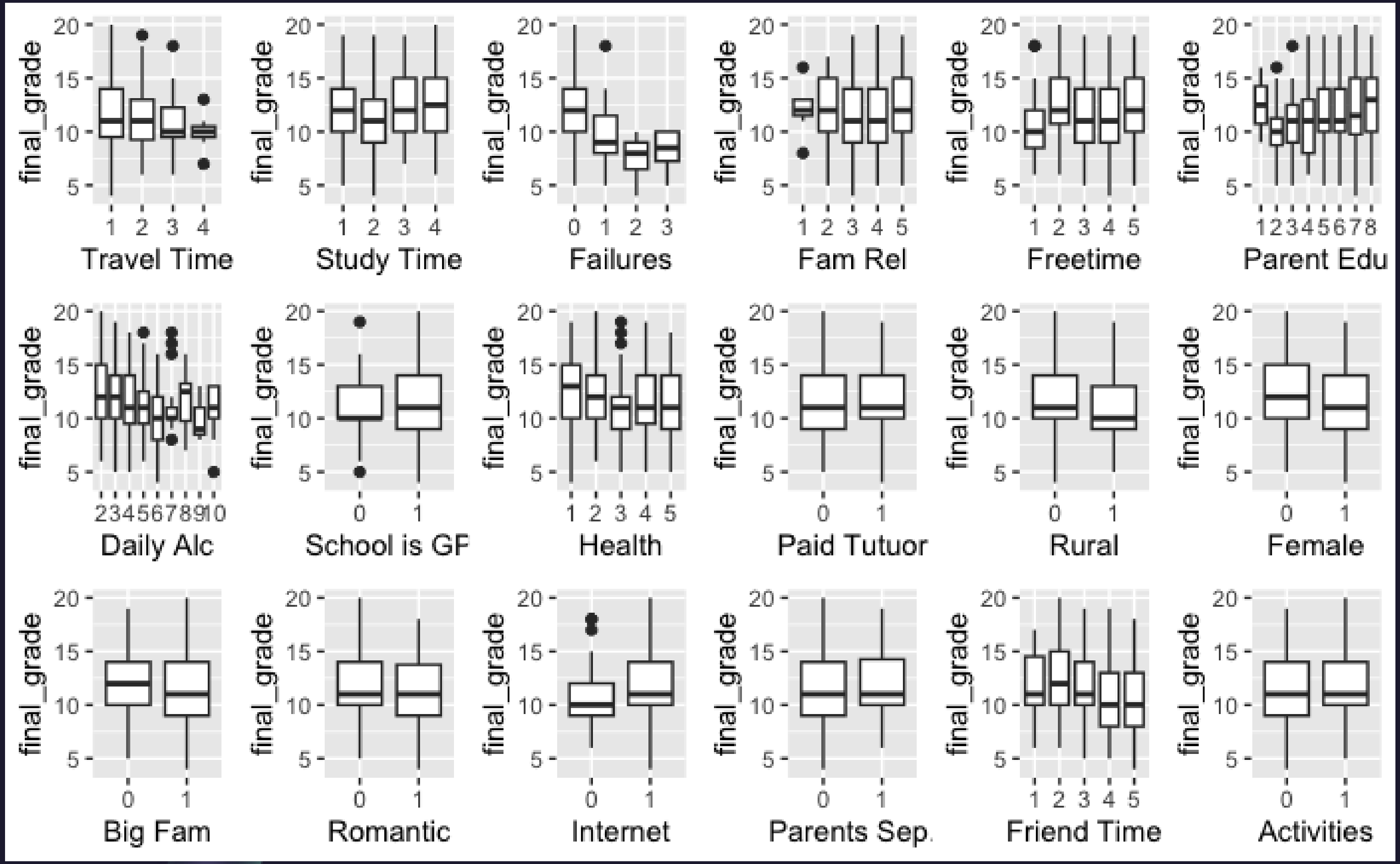
[HTTPS://WWW.KAGGLE.COM/DATASETS/UCIML/STUDENT-ALCOHOL-CONSUMPTION](https://www.kaggle.com/datasets/uciml/student-alcohol-consumption)

Data Exploration



- After data preparation, the correlation plot at right show that we have no remaining multicollinearities, which allows us to proceed with feature selection methods such as forward selection.

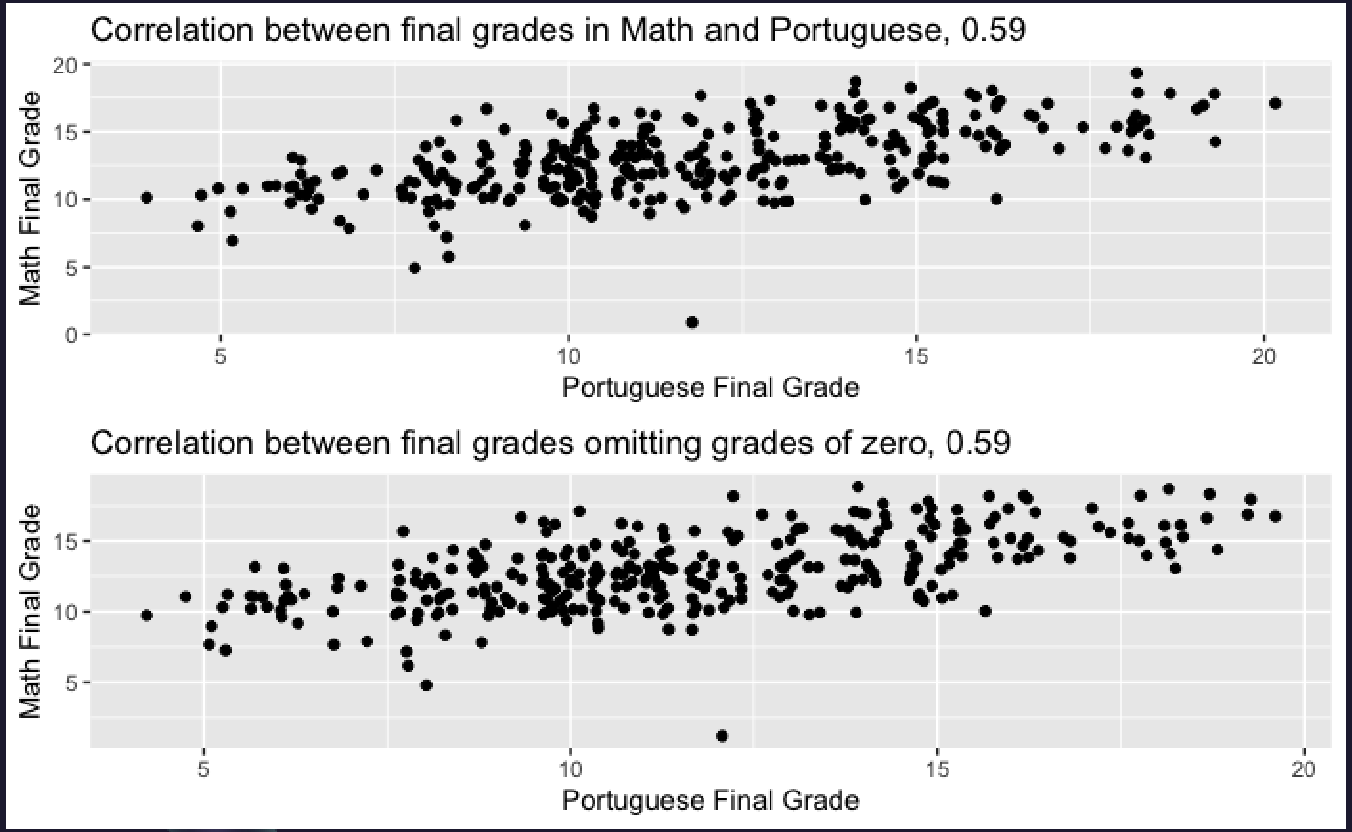
Data Exploration



- Plotting the mean of the target in relation to variance in the features allows us to get a sense of whether or not a feature is likely to be predictive of the target.

- Here, failures stands out as appearing to have the largest impact on the mean final grade. We expect to include this feature in our final model.

Data Exploration

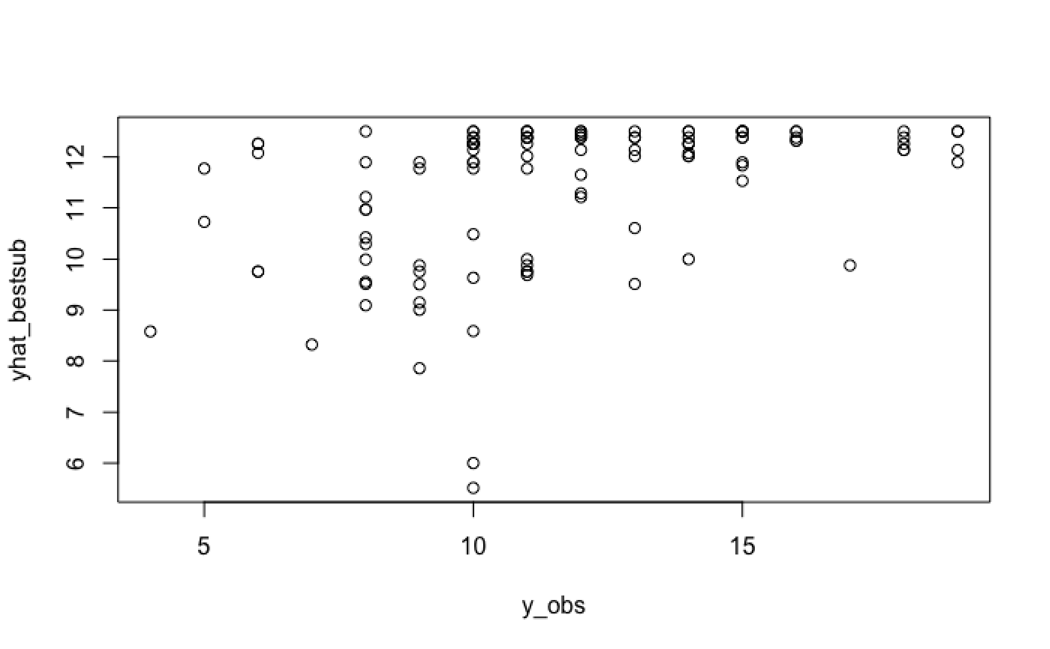


- Here we check our assumption about the impact of our data preparation: Because we can see that final grades are highly correlated for students who take both, it is reasonable to combine them as a single target to understand variation in student achievement. Excluding zeroes has not major impact on that results.

Linear Model: Best Subset

- A best subset approach was first attempted, resulting in a model with 3 features:
  - Failures
  - Mean Absences
  - School Support
- This model has an RMSE of 3.10 when tested using the test set.

Predicted vs Actual Results



Linear Model: Stepwise

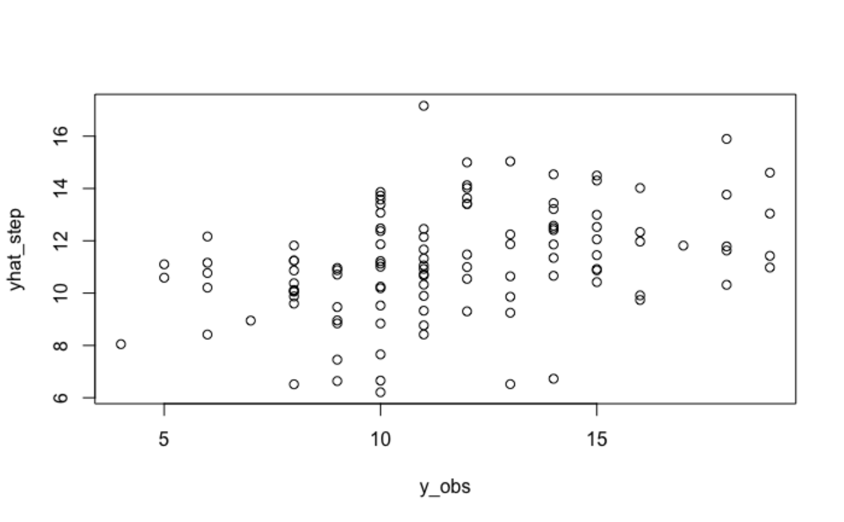
- We can also attempt to use stepwise regression to see if there is a different result. Interestingly, this results in a different model -- perhaps because it is judged on AIC rather than BIC, which penalizes the inclusion of additional features more.

- When tested on the test set, RMSE is actually higher (3.31) indicating perhaps it is overfit.

•The best stepwise model is close to the full model the following features:

- Failures
- Mean Absences
- School Support
- School is GP
- Sex_F
- Mother's job
- Father's job
- Guardian
- Study time
- Family relations
- Friend time
- Health
- Parent Education

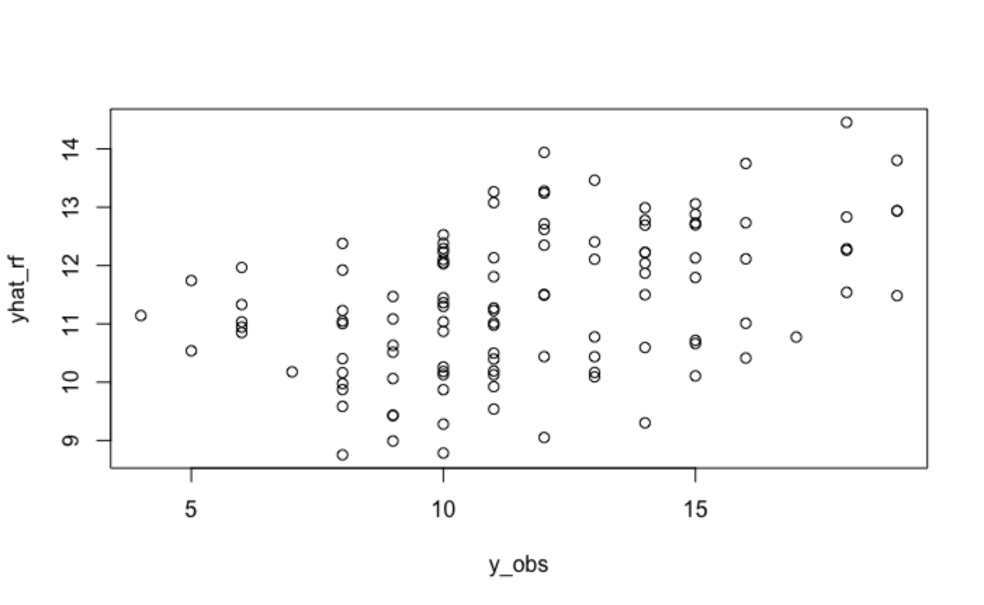
Predicted vs Actual Results



Random Forrest

- We next attempted Random Forrest, which is a nonparametric method of feature selection.
- When tested on the test set, RMSE is similar to the linear model (3.10) and correlation between predicted and actual values is 42%.

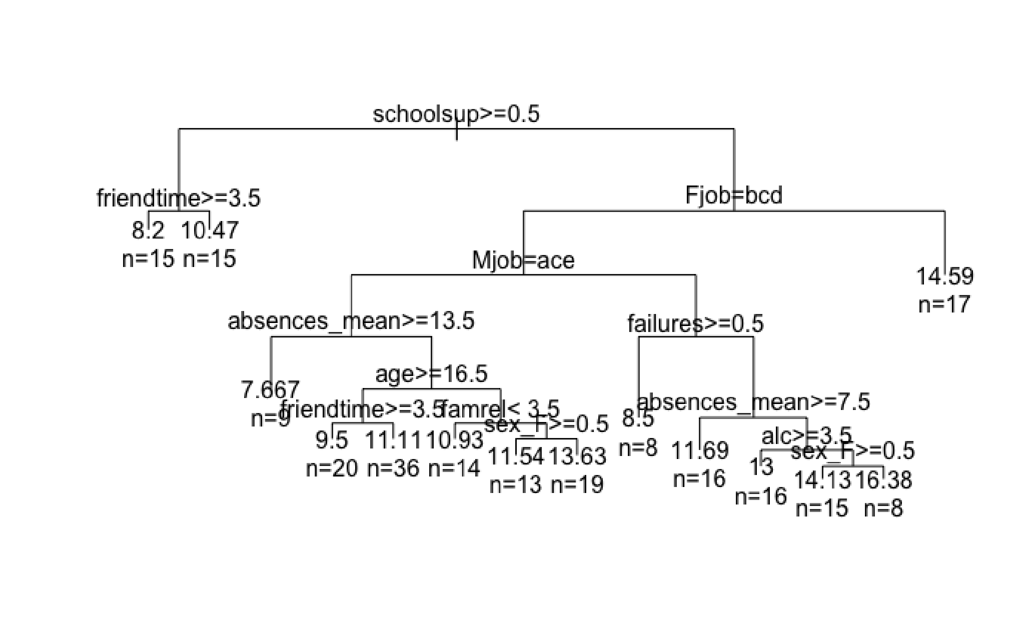
Predicted vs Actual Results



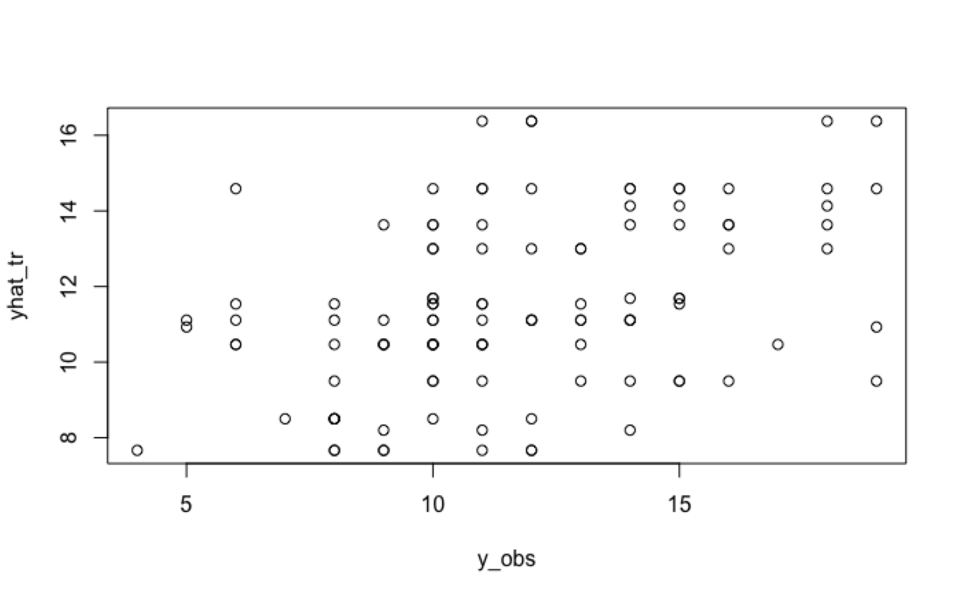
Decision Tree

- Finally we attempted Decision Tree

- When tested on the test set, RMSE is similar to the linear model (3.24) and correlation between predicted and actual values is 41%.



Predicted vs Actual Results



Conclusion

Models	Correlation between Predicted and Actual	RMSE
Best Subset	0.4238 372	3.107402
Stepwise Selection	0.3715614	3.312686
Random Forrest	0.4283920	3.108460
Decision Tree	0.4144219	3.244109

The best model is not one with the highest R-squared, but the one that performs the best on test data. In fact, all of these models performed very similarly.

Ultimately, the choice of the best model depends upon the goal of the study. For a policy maker, the simplest linear model may be preferred as it results in the clearest and most defensible policy prescriptions:

* Intervene with students who have previously failed

* Offer school support

* Intervene to prevent or mitigate absences

Thank You