Assignment 5: Supervised Learning

Group 11:

Robert Locher (5747465**), Jose Enrique Leal Castillo (**9066381**), Niek Lieon (**6520448**)**

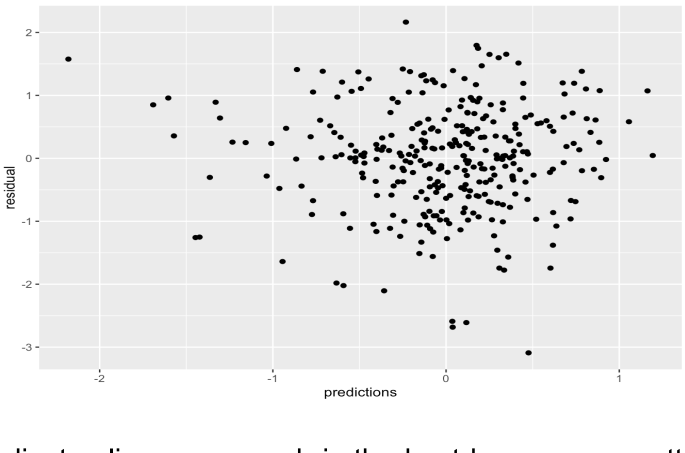
Data preparation:

First, we checked the data for missing values, but there were none. After examining the dataset, we found out some columns contained numerical data and others contained categorical data. Because a lot of machine learning models only work with numerical data, we decided to encode the nominal values of categorical columns to numerical values. Some of this categorical data was binary, while others contained more than two different values. For the binary columns, the names of the values were converted to 1 and 0. For the other categorical data, we chose to use one-hot encoding to convert the value to numeric. This means for each n different values in a column, this column will be transformed to n columns, where each column represents a distinct value.

To learn what features are important to predict the score of a student, we chose to implement the Lasso algorithm. To make sure each variable was assessed on an equal base, the data frame was normalized beforehand. The plot resulting from this algorithm shows the number of features to include in prediction algorithms should be somewhere between 14 and 20, as for this value the MSE is lowest.

Supervised learning method chosen:

We will first approach this prediction of scores with a linear model to see if the model needs more complexity A residual plot is done to validate if the linear model is correct



The residual plot indicates linear approach is the best because no patterns are observed in the data + the mean and median of residuals is close to 0

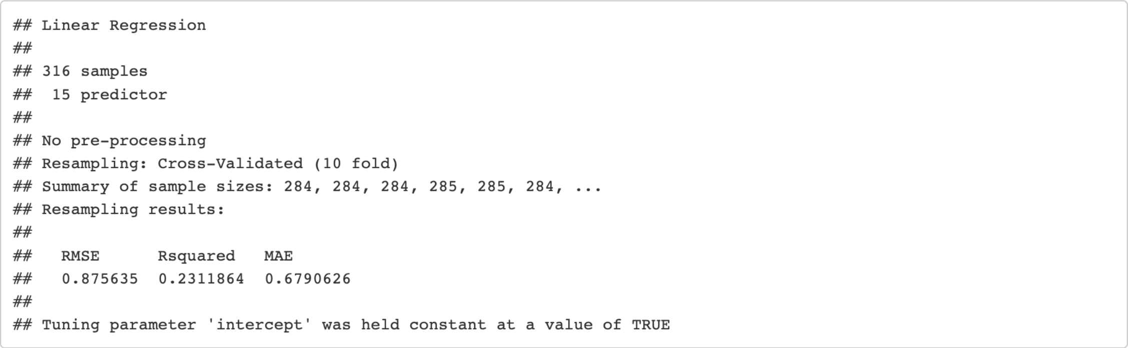
Assignment 5: Supervised Learning

Group 11:

Robert Locher (5747465**), Jose Enrique Leal Castillo (**9066381**), Niek Lieon (**6520448**)**

Validation strategy:

We evaluate cross validation performance of the model with train control RMSE



```
## Linear Regression
##
## 316 samples
##  15 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 284, 284, 284, 285, 285, 284, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   0.875635  0.2311864  0.6790626
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

additionally, we validate the model with the class constructed cross validation function obtaining a sweet and nice MSE of 0.70