Intermediate Report

Raven Brown,Josh Kim, Dorothy Stroh

Introduction

In statistical analysis, the combination of theoretical knowledge learned in the classroom with practical application is crucial for drawing reasonable and accurate conclusions from large, overwhelming datasets. Our goal is to formulate insightful questions specific to each dataset and use the appropriate statistical methods to analyze and interpret any underlying patterns.

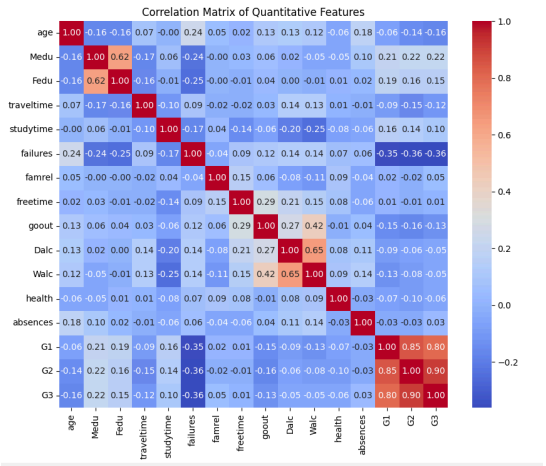
The dataset we will be evaluating in our report is student performance, which was taken from the UC Irvine Machine Learning Repository. Student performance contains both quantitative and qualitative data, so we will be using a number of classification methods (logistic regression, linear discriminant analysis, and naive bayes) as well as linear regression and multiple linear regression to evaluate the data.

Questions

What is the relationship between demographic/social/school-related factors and student performance? (Linear regression)
Do demographic or social factors such as free time and travel time, alcohol consumption, or parental education level influence the classification of students into different performance categories? (QDA)
What certain characteristics would a student need to achieve higher grades? (Naive Bayes)
Are we able to predict whether or not a student will pass or fail based on their internet access and/or whether or not they attended a nursery? (Logistic Regression)
Which variables are pertinent in determining a student's passing or failure? (Lasso/Ridge Regression).
What factors have the greatest influence on wine quality? (Decision Trees) (Wine Quality Dataset)

Linear Regression / Multiple Linear Regression

The relationships between different factors and student performance can be found using multiple linear regression and linear regression. We could just do a simple linear regression for each predictor, but the advantage of using multiple linear regression is that we can see all the predictors being considered simultaneously, allowing us to see how each variable affects student performance while accounting for the influence of other variables, so that will be the main focus of this section. The first step to take is preparing the data for linear regression analysis to ensure the data is in a format suitable for the model. In our case, we will be focusing on only the quantitative variables in the dataset to simplify the preprocessing stage.



When looking at the heatmap, there is a lack of evidence of correlation between grades and each quantitative variable (the correlation between predictors could be affecting the interpretation of individual coefficients), so we will be using all of them in our analysis. Below are the results from the

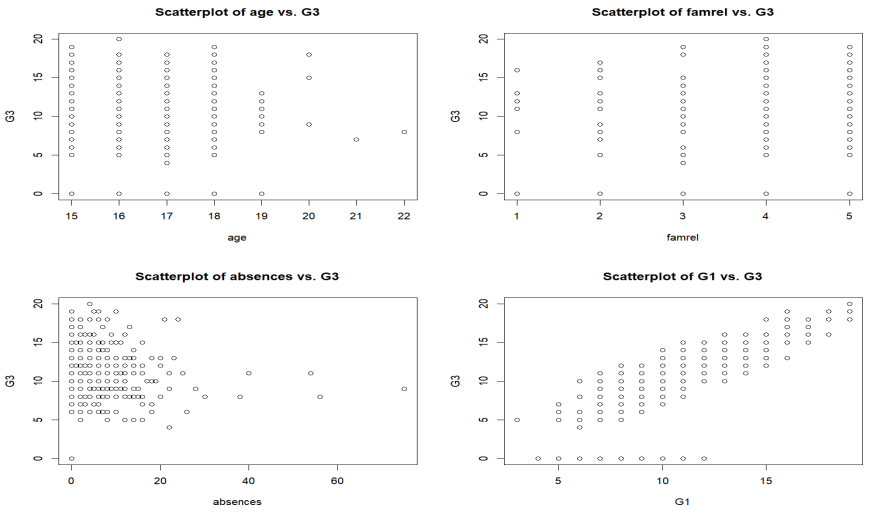
multiple linear regression model with only quantitative variables predicting student performance (G3) for math and Portuguese classes respectively.

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.6597 -0.4033  0.2559  0.9736  4.0970

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6689158  1.5411303  -0.434 0.664505
age         -0.1815062  0.0808776  -2.244 0.025396 *
Medu         0.1069982  0.1152672   0.928 0.353862
Fedu        -0.1404233  0.1143700  -1.228 0.220285
traveltime   0.1270661  0.1420398   0.895 0.371579
studytime   -0.1362063  0.1208403  -1.127 0.260388
failures    -0.2321985  0.1457021  -1.594 0.111847
famrel       0.3516407  0.1096758   3.206 0.001459 **
freetime     0.0537032  0.1037476   0.518 0.605016
goout        0.0004457  0.1003632   0.004 0.996459
Dalc        -0.1204910  0.1431980  -0.841 0.400638
Walc         0.1545526  0.1067973   1.447 0.148679
health       0.0555571  0.0700899   0.793 0.428475
absences     0.0414289  0.0123309   3.360 0.000859 ***
G1           0.1590346  0.0566607   2.807 0.005262 **
G2           0.9743079  0.0502614  19.385  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.881 on 379 degrees of freedom
Multiple R-squared:  0.8378,    Adjusted R-squared:  0.8314
F-statistic: 130.5 on 15 and 379 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-8.9458 -0.4653 -0.0674  0.6240  6.1431

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13153    0.84235  -0.156 0.875967
age          0.03180    0.04411   0.721 0.471269
Medu        -0.03160    0.05874  -0.538 0.590788
Fedu         0.03591    0.05956   0.603 0.546743
traveltime   0.09089    0.06910   1.315 0.188894
studytime    0.07662    0.06306   1.215 0.224782
failures    -0.23483    0.09564  -2.455 0.014345 *
famrel      -0.03844    0.05351  -0.718 0.472772
freetime    -0.03714    0.05111  -0.727 0.467652
goout       -0.01028    0.04904  -0.210 0.834100
Dalc        -0.09102    0.06903  -1.319 0.187800
Walc        -0.01915    0.05284  -0.362 0.717192
health      -0.04124    0.03506  -1.176 0.240014
absences     0.02448    0.01107   2.212 0.027343 *
G1           0.14292    0.03677   3.887 0.000112 ***
G2           0.87923    0.03467  25.357  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.253 on 633 degrees of freedom
Multiple R-squared:  0.853,     Adjusted R-squared:  0.8495
F-statistic: 244.8 on 15 and 633 DF,  p-value: < 2.2e-16
```

Low p-values suggest statistically significant relationships, and the coefficients tell the strength. Filtering out the variables with the most significant p-values ($p < 0.05$) gives us famrel (quality of family relations), absences, and age for math and failures and absences for Portuguese. This suggests that there is a significant relationship between these factors and their subjects and that these factors are not random.

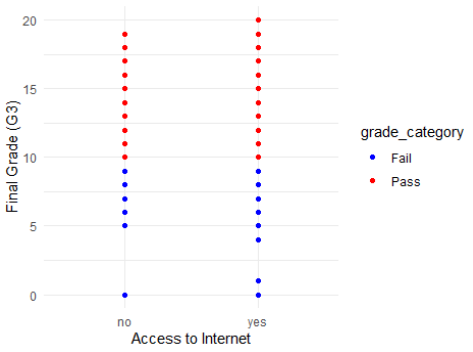
Analyzing the coefficients associated with each variable, for math, we can see that famrel and absences have a positive coefficient (very interesting result), while age has a negative coefficient. The positive coefficient suggests a positive linear relationship, meaning that students with better family relationships and skip school…? tend to have higher grades, and the negative coefficient suggests a negative relationship, meaning that younger students generally perform worse on average. We can apply the same logic to the Portuguese dataset. Intuitively, more absences should mean skipping out material and learning, which leads to worse instead of better grades. The coefficient (0.0414289) is relatively small in this case, suggesting a weak positive relationship, or possibly no real linear relationship between absences and grades.



To draw further conclusions, we utilize scatterplots of each significant predictor vs. student to help visualize the distribution of data points and identify any potential non-linear relationships. Looking at the plots, the only strong linearities we can see are from G1 and G2, and we can not definitively say there is a

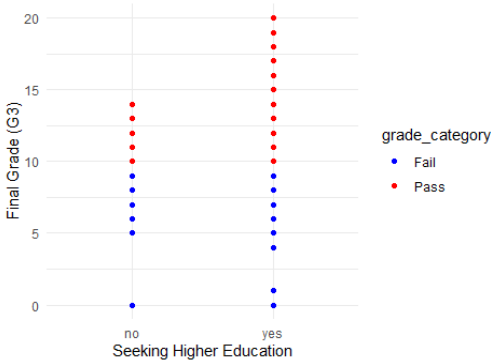
positive/negative linear relationship for the other predictors given the current data despite their coefficient values. To answer the question at hand, the model shows that factors such as quality of family relationships and absences are weakly associated with better grades, and factors such as age and failures are weakly associated with worse grades. However, we can strongly conclude that higher first and second period grades lead to better grades. Multiple linear regression is not a sufficient strategy to formulate a convincing answer to this question. The problem with MLR is that it assumes linear relationships. Since our data suggests otherwise, we may need to consider alternative techniques.

Logistic Regression



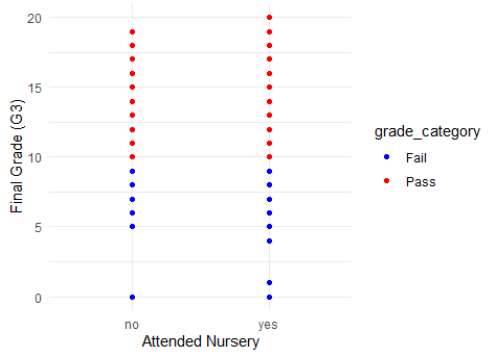

```
Call:
glm(formula = G3_binary ~ internet, family = "binomial", data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0085     0.1534   6.574 4.89e-11 ***
internetyes   0.3287     0.1757   1.871   0.0614 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1101.0  on 1043  degrees of freedom
Residual deviance: 1097.6  on 1042  degrees of freedom
AIC: 1101.6

Number of Fisher Scoring iterations: 4
```

Logistic Regression is another model used to analyze the relationships between independent variables and binary outcome variables. As seen above, logistic regression was applied to the "Student Performance" data set and was used to analyze the relationship between whether or not a student had access to the internet and their final grade as well as whether or not a student attended a nursery and their final grade. After converting "Final Grades" to binary (A grade 10 and above would result in a pass and 9 and below resulted in a fail), we were able to create a summary of each model. Based on the summary for the internet access model, we are given the coefficient estimates for the intercept and 'internetyes'. Since the coefficient for 'internetyes' > 0, this implies that an increase in this predictor variable is associated with an increase in the probability of the outcome variable(passing or failing) being equal to 1 (passing). To further support the idea that internet access has an impact on whether or not a student passes or fails, we can also examine the Pr(>|z|) value, which is 0.0614. This value for 'internetyes' is fairly close to 0.05 which also implies that this variable might be significant in predicting the probability in achieving a final grade greater than or equal to 10 (passing). However, this is not conventionally conclusive since we prefer the p-value to be below (0.05).




```
Call:
glm(formula = G3_binary ~ higher, family = "binomial", data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.06744    0.21212  -0.318    0.751
higheryes    1.50019    0.22744   6.596 4.22e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1101.0  on 1043  degrees of freedom
Residual deviance: 1059.3  on 1042  degrees of freedom
AIC: 1063.3

Number of Fisher Scoring iterations: 4
```

We see a similar trend when by following the same procedure with a different predictor variable, whether or not a student is seeking higher education. Since the coefficient for 'higheryes' > 0, this suggests that an increase in the outcome variable is associated with an increase in the predictor variable, which in this case is seeking higher education. The p-value, which is (4.22e-11). Since this value is really small, whether or not a student is seeking higher education is statistically significant when it comes to predicting the final grade value for that student.




```
Call:
glm(formula = G3_binary ~ nursery, family = "binomial", data = data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.29320    0.16828   7.685 1.53e-14 ***
nurseryyes  -0.03659    0.18779  -0.195    0.846
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1101  on 1043  degrees of freedom
Residual deviance: 1101  on 1042  degrees of freedom
AIC: 1105

Number of Fisher Scoring iterations: 4
```

On the other hand, when applying Logistic Regression to the same dataset, but analyzing the relationship between attending a nursery and final grades, we see an opposite reaction. Since the coefficient for 'nurseyyes' < 0, this implies that an increase in the predictor variable (nursery) will not have a substantial impact on the outcome variable (passing or failing). The Pr(>|z|) value further supports this idea because it is 0.846 which is not close to 0.05. This indicates that the predictor variable 'nursery' is not statistically relevant when predicting the probability of achieving a final grade greater than or equal to 10 (passing).

To conclude determining the significance of certain variables in comparison to final grade values, Logistic Regression is very useful because it provides coefficients, p-values, and standard deviations which allow us to examine the relationship between such predictor variables and outcome variables.

Multiple Logistic Regression

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -2.657e+01  2.157e+05   0.000    1.000
schoolMS          -1.227e-09  3.067e+04   0.000    1.000
sexM              -6.038e-12  2.611e+04   0.000    1.000
age                5.766e-10  1.070e+04   0.000    1.000
addressU           4.869e-11  2.849e+04   0.000    1.000
famsizeLE3        -7.697e-10  2.574e+04   0.000    1.000
PstatusT          -3.958e-09  3.704e+04   0.000    1.000
Medu              -1.415e-10  1.625e+04   0.000    1.000
Fedu               5.176e-10  1.447e+04   0.000    1.000
Mjobhealth        -4.901e-10  5.712e+04   0.000    1.000
Mjobother         -1.604e-09  3.375e+04   0.000    1.000
Mjobservices      -9.027e-10  4.000e+04   0.000    1.000
Mjobteacher       -2.293e-09  5.289e+04   0.000    1.000
Fjobhealth         1.654e-09  7.737e+04   0.000    1.000
Fjobother          3.396e-10  4.975e+04   0.000    1.000
Fjobservices       4.616e-11  5.215e+04   0.000    1.000
Fjobteacher        9.124e-09  6.973e+04   0.000    1.000
reasonhome        -8.871e-10  2.955e+04   0.000    1.000
reasonother       -2.183e-09  4.013e+04   0.000    1.000
reasonreputation  -7.538e-10  3.088e+04   0.000    1.000
guardianmother     6.044e-10  2.826e+04   0.000    1.000
guardianother      1.041e-09  5.409e+04   0.000    1.000
traveltime         3.374e-10  1.714e+04   0.000    1.000
studytime          2.152e-10  1.497e+04   0.000    1.000
failures          -7.072e-10  2.020e+04   0.000    1.000
schoolsupyes       1.701e-09  3.771e+04   0.000    1.000
famsupyes         -2.230e-09  2.432e+04   0.000    1.000
paidyes           -6.063e-10  2.883e+04   0.000    1.000
activitiesyes     -3.192e-10  2.336e+04   0.000    1.000
nurseryyes         1.344e-09  2.868e+04   0.000    1.000
higheryes          2.226e-09  4.441e+04   0.000    1.000
internetyes       -6.777e-10  3.012e+04   0.000    1.000
romanticyes        5.193e-10  2.445e+04   0.000    1.000
famrel             3.498e-10  1.250e+04   0.000    1.000
freetime           1.174e-11  1.197e+04   0.000    1.000
goout              2.315e-10  1.150e+04   0.000    1.000
Dalc              -3.169e-11  1.650e+04   0.000    1.000
Walc              -1.412e-10  1.267e+04   0.000    1.000
health             5.072e-11  8.264e+03   0.000    1.000
absences          -1.301e-11  1.960e+03   0.000    1.000
G1                -2.363e-09  7.649e+03   0.000    1.000
G2                 5.992e-10  9.621e+03   0.000    1.000
G3                -7.125e-11  7.774e+03   0.000    1.000
grade_categoryPass 5.313e+01  4.054e+04   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.1010e+03  on 1043  degrees of freedom
Residual deviance: 6.0569e-09  on 1000  degrees of freedom
AIC: 88

Number of Fisher Scoring iterations: 25
```

On the other hand, after applying multiple logistic regression, we are provided more information about the effect different variables have on whether or not a student passed their Math and Portuguese courses. From the above coefficient values, we can see that for a majority of the variables, the estimates are very close to zero with large standard errors. When the

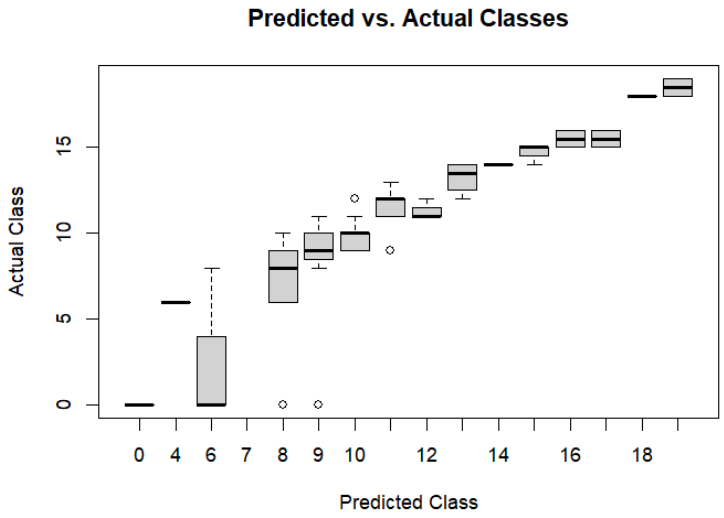
estimates are close to zero, this means that there was a very small change in the log odds of the response variable for a one-unit change in the predictor variable while holding all other variables constant. When the standard error is relatively large, there is a large amount of uncertainty in the estimated coefficient. With these factors acting together, this indicates that a majority of the variables listed are not statistically relevant predictors of the response variable. The last coefficient, grade_CategoryPass, however, has the opposite effect. This coefficient has a large estimate with a low p-value, which suggests that this coefficient in particular is a significant predictor of the response variable.

In conclusion, multiple logistic regression provides a better analysis of the given information than simple logistic regression. Not only is it a more flexible approach than simple logistic regression, but since we are including multiple predictor variables, this will allow us to obtain more reliable estimates of the relationships between our predictor and response variables. Simple logistic regression is more computationally expensive because of the number of variables in this case and this in turn could lead to poor accuracy and interpretability.

Quadratic Discriminant Analysis

	0	6	7	8	9	10	11	12	13	14	15	16	18	19
0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	3	0	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	4	1	1	5	3	3	0	0	0	0	0	0	0	0
9	1	0	0	1	4	0	2	0	0	0	0	0	0	0
10	0	0	0	0	4	8	1	1	0	0	0	0	0	0
11	0	0	0	0	1	0	1	2	1	0	0	0	0	0
12	0	0	0	0	0	0	3	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	1	1	2	0	0	0	0
14	0	0	0	0	0	0	0	0	0	2	0	0	0	0
15	0	0	0	0	0	0	0	0	0	2	6	0	0	0
16	0	0	0	0	0	0	0	0	0	0	2	2	0	0
17	0	0	0	0	0	0	0	0	0	0	1	1	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	1	0
19	0	0	0	0	0	0	0	0	0	0	0	0	1	1

The confusion matrix provided summarizes the performance of a classification model applied to a dataset. Each row of the matrix corresponds to the actual classes, while each column represents the predicted classes. In this case, the matrix reveals that for the class "0," three observations were correctly classified as class "0." Similarly, for class "4," one observation was correctly classified. However, there were misclassifications observed, such as one observation from class "6" being incorrectly classified as class "9." Moreover, multiple misclassifications were evident in class "8," with observations incorrectly classified as classes "4," "6," "9," "10," and "11." Overall, the confusion matrix provides valuable insights into the model's performance, highlighting areas of accurate classification and areas requiring improvement.



Plotting the predicted classes against the actual classes provides a visual comparison to evaluate the performance of the classification model. In the graph, each point represents an observation, where the x-coordinate indicates the predicted class assigned by the model, and the y-coordinate represents the actual class label from the test data. Ideally, the points would align along the diagonal, indicating perfect predictions. However, deviations from the diagonal signify misclassifications. This visualization helps to assess the model's accuracy and identify any patterns or trends in the misclassifications. Additionally, it provides insights into which classes the model struggles to predict accurately, guiding further investigation or model refinement.

K-Nearest Neighbors

Initially, employing the KNN algorithm with k = 1, the confusion matrix reveals that the predicted classes predominantly align with classes 8 or 10, regardless of the actual class. This suggests a bias in the predictions towards these particular grades. The overall accuracy rate for this model is approximately 49.37%.

Upon increasing k to 3, there is a noticeable enhancement in the model's performance. The confusion matrix indicates a broader distribution of correct predictions across different actual classes compared to k = 1. Similar to the previous model, the predicted classes remain largely focused on 8 or 10. The overall accuracy rate experiences a modest improvement, reaching approximately 53.16%.

Subsequently, exploring the KNN model with k = 5 results in a slightly different confusion matrix pattern, yet the predominant prediction of classes 8 or 10 persists. Despite some improvements observed in correctly predicting other classes, the overall accuracy rate decreases marginally to around 46.84%.

In conclusion, while the KNN model exhibits moderate performance in predicting the final grade (G3) based on the predictors G1 and G2, there are notable areas for improvement. The choice of k appears to influence the model's accuracy, with k = 3 yielding the highest accuracy among the tested values. However, the model demonstrates a bias towards predicting

classes 8 or 10, indicating a need for further refinement or feature exploration to enhance predictive capabilities.

Naive Bayes

The factors that influence the final grade (G3) performance in the dataset can be investigated using Naive Bayes. Using the Naive Bayes algorithm should help us identify the student characteristics with the strongest association with achieving a high G3 score because it works well with a large number of predictors and provides results based on class probabilities. However, it assumes independence between predictors, which may not be ideal since some predictors have a potential influence on others. Also, we should keep in mind that since we're using the entire dataset for only training and prediction, there's a big risk for overfitting.

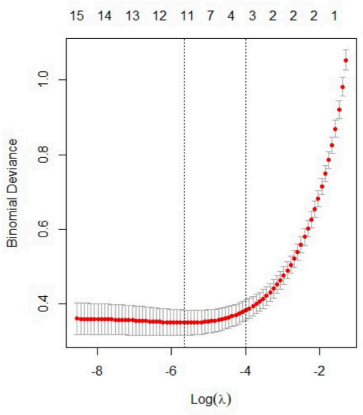
The approach was to fit a Naive Bayes model using all predictors except the target variable (G3 - the actual values) and we generated a confusion matrix to assess the performance of the model. The diagonal entries indicate the number of times the predicted class matched the actual number. The off-diagonal entries are misclassifications. It looks like the model has a higher misclassification rate in the middle ranges (9, 10, 11, 12) compared to extreme grades (16-20), suggesting that the distinguishing between grades in the middle is a little more challenging for the model.

In conclusion, while our analysis does not directly answer the question about the characteristics associated with achieving higher grades, it does tell us Naive Bayes' capability to accurately predict different student grades. The model achieved an overall accuracy rate of 40.15, which is moderately accurate compared to the actual grades in the dataset. I think Naive Bayes would've been sufficient in answering what grades can be more accurately predicted using the dataset instead of the question I was investigating.

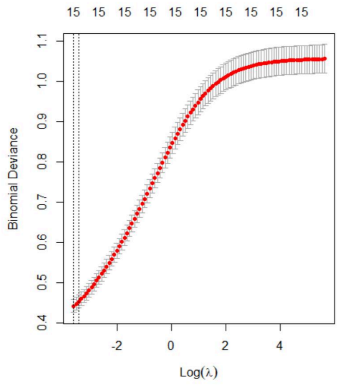
nb.class	0	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0	37	0	0	0	1	5	6	10	6	4	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	1	0	1	15	1	4	1	7	1	0	0	0	0	0	0	0	0	0
7	0	0	1	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	16	2	2	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	4	11	3	1	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	2	4	30	4	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	1	4	3	27	5	3	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	1	4	16	2	1	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	2	4	18	2	1	0	0	0	0	0
14	0	0	0	0	0	0	0	0	2	2	6	19	2	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	2	5	26	16	0	1	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	1	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	5	1
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Lasso vs. Ridge Regression

(Intercept)	-15.07053689
school	.
sex	.
age	.
address	.
famsize	.
Pstatus	.
Medu	.
Fedu	-0.12486133
Mjob	.
Fjob	.
reason	.
guardian	.
traveltime	0.27606036
studytime	-0.10577205
failures	-0.03234894
schoolsup	.
famsup	.
paid	.
activities	.
nursery	.
higher	.
internet	.
romantic	.
famrel	.
freetime	.
goout	-0.12715935
Dalc	-0.02353598
Walc	0.03130002
health	-0.03250040
absences	-0.02655831
G1	0.39124665
G2	1.38503652
grade_category	.



(Intercept)	-5.851869908
school	.
sex	.
age	-0.026838602
address	.
famsize	.
Pstatus	.
Medu	-0.006572741
Fedu	-0.019330787
Mjob	.
Fjob	.
reason	.
guardian	.
traveltime	0.181120023
studytime	-0.046999593
failures	-0.291653031
schoolsup	.
famsup	.
paid	.
activities	.
nursery	.
higher	.
internet	.
romantic	.
famrel	0.056062899
freetime	-0.038122469
goout	-0.108676923
Dalc	-0.024436938
Walc	0.039260253
health	-0.020394712
absences	-0.027065182
G1	0.356120532
G2	0.444239060
grade_category	.



Lasso Ridge

Both lasso and ridge regression are regularization methods that are used in linear regression to help with overfitting. Lasso incorporates a penalty term which in turn forces some coefficients to zero. This yields fewer models resulting in easier interpretation. Ridge regression introduces loss and penalty that shrink coefficients towards zero. Unlike lasso, ridge rarely shrinks coefficients exactly to zero, which means that all variables are kept in the final model.

For this dataset in particular, the choice between lasso and ridge regression depends on what we want our modeling outcome to be. Both methods help to identify useful predictors, but each has their own trade-offs. Lasso is a better option if: we suspect that there are only a few variables that are truly important for predicting the outcome; there are select features that lead to a more interpretable models; if we want a simple model that results in setting coefficients to zero; if we prefer a model with fewer predictors. Ridge is a better alternative if: there are highly correlated variables; we want more prediction accuracy; we want to keep all predictors but with reduced coefficients. In order to decide which is best for this dataset, we perform both methods and use cross-validation to see which is most appropriate.

Since lasso gets rid of some variables, the trade-off for this method is that it may introduce bias. Since ridge regression doesn't exclude non-valuable variables, this could also introduce bias.

Selections

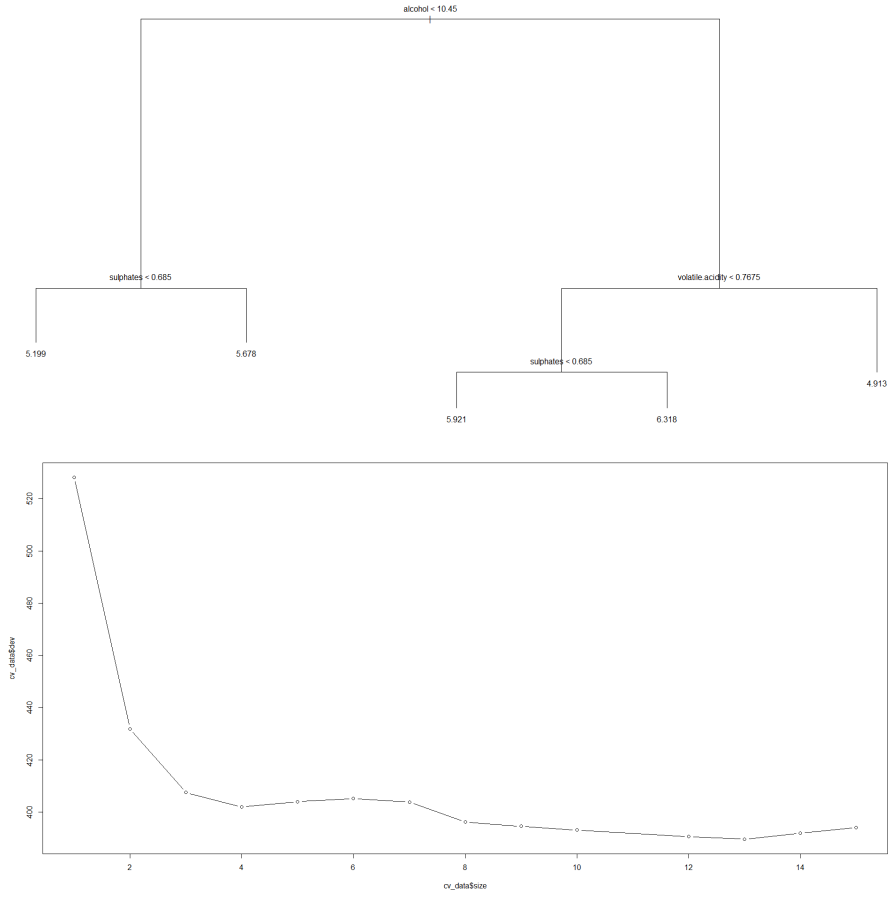
Best subset selection can be computationally expensive for large datasets, prompting us to rely on forward step selection and backward step selection. Backward stepwise selection, for instance, begins with a model incorporating all 41 variables from the "student_mat" dataset. At each step, the algorithm evaluates each variable's contribution to the model's fit, systematically removing the one that adds the least value until the stopping criterion is met. The resulting model, as seen in the summary output, iteratively sheds variables, ultimately retaining significant predictors like "romantic-yes" and "go-out," marked with an asterisk for their importance in predicting absences. This method effectively streamlines the model while preserving the most pertinent predictors. Conversely, forward stepwise selection commences with an empty model, gradually incorporating variables based on their contribution to enhancing the model's fit. The

summary output illustrates the sequential addition of variables like "age" and "sex-M," followed by others such as "traveltime" and "studytime," signifying their predictive value for absences.

Comparing these approaches, backward selection starts with a full model and eliminates variables, while forward selection builds up from an empty model by adding variables. In the "student_mat" dataset, backward selection identifies "romantic-yes" and "go-out" as significant predictors, whereas forward selection includes "age" and "traveltime" among others. Both methods aim to uncover the most relevant predictors while preventing overfitting, but their final models may differ based on the dataset and stopping criterion. Backward selection is advantageous for simplifying models with numerous predictors, while forward selection is efficient for datasets with fewer predictors.

Decision Trees

A decision tree can help identify the most important predictors that contribute to wine quality by visualizing the splits based on the “best” predictor variables, which are chosen by the greatest reduction in variance of “quality” within each partition (Note: The wine quality dataset is being used for this problem so that we have more information to work with in our regression decision tree as opposed to having a combination of regression/classification data).



These are the results of the pruned tree. Initially, our tree had multiple splits on the same variables, resulting in 25 terminal nodes, which told us that the model may be capturing too

much noise rather than actual patterns (overfitting), so pruning was necessary to simplify our tree by removing unnecessary branches. From our cross validation results, the rapid drop in deviance shows that the early splits in the tree are providing improvements in predictive performance and that the tree pruning is helping. What's strange is that even though there is a significant drop in deviance, there is a marginal increase in the MSE of the pruned tree… The difference may be negligible (.4998 to .5255), but this suggests that the pruning is not improving performance. The decision tree itself is pretty self explanatory. The terminal nodes are the predicted quality values given each decision split, and the internal nodes are the "important" predictors that are determined by our software, ranking higher in importance from bottom to top. For example, alcohol percentage is the most important predictor.

The decision tree makes splits based on the variable that best separates the data at each node. As a result, the choice of splitting variables may not always best represent the entirety of the dataset's information, especially if the tree is being pruned for simplicity. To address this, we planned on using the Boosting and Random Forest methods.

Boosting would sequentially refine the model by focusing on instances that were misclassified by previous models. It works by iteratively examining and combining small trees, which would enable a thorough exploration of the dataset's feature space and could potentially reveal associations that were overlooked by individual decision trees. We are using Random Forest because they typically produce accurate predictions by combining predictions from multiple trees, which are each trained on a random subset of features (also introduces possibly overlooked features as well!). Its results are also very useful and easy to interpret so it will help us identify which variables have the most significant impact on the predictions.

Bagging would not be applicable for our case because we are not interested in reducing the variance in our tree since it has a simple structure, and our dataset isn't overly complex. BART has a risk of being more complex and less interpretable than simpler methods like Random Forest or Boosting., which is what we need to understand the underlying relationships between variables and wine quality. Therefore, because of the relatively small size of our dataset, simpler methods like Random Forest and Boosting may potentially offer a more interpretable solution.

	%IncMSE	IncNodePurity
fixed.acidity	30.05076	53.39885
volatile.acidity	49.80421	133.27473
citric.acid	28.23075	56.91916
residual.sugar	24.79517	53.51817
chlorides	29.83235	63.58531
free.sulfur.dioxide	28.37954	47.80041
total.sulfur.dioxide	46.12629	80.11920
density	32.93686	67.14896
pH	26.57389	57.07587
sulphates	65.25864	145.24915
alcohol	86.19950	245.72310

A larger %IncMSE suggests that the feature is more important since its shuffling(?) results in a larger increase in prediction error. A higher IncNodePurity value means greater importance because it assesses the impact of each feature on the purity of nodes in the decision trees. When evaluating both metrics from the results of our Random Forest, we can conclude that alcohol, sulfates, volatile acidity, and total sulfur dioxide are the most important features that impact quality.

```
> summary(boost_model)
                                     var   rel.inf
alcohol                          alcohol 15.448703
volatile.acidity        volatile.acidity 12.509456
sulphates                      sulphates 11.419189
total.sulfur.dioxide total.sulfur.dioxide  9.881168
density                          density  9.244185
citric.acid                  citric.acid  9.108070
chlorides                      chlorides  7.815756
pH                                    pH  7.236859
residual.sugar            residual.sugar  6.359266
fixed.acidity              fixed.acidity  5.940932
free.sulfur.dioxide   free.sulfur.dioxide  5.036417
```

The rel.inf column represents the relative importance of each variable (higher values mean more important and lower values mean less important). We can conclude that alcohol percentage has the highest importance, indicating that it is the most influential in predicting wine quality, and volatile acidity, sulfates, and total sulfur dioxide are also significant. So from the results from our tree analysis and the results from the Boosting and Random Forest methods, we can confidently say that the quality depends on its alcohol percentage, volatile acidity, sulfates, and total sulfur dioxide.