Regression Analysis on Student Performance data obtained from Portuguese language classes and its effect on final grade performance

Multiple Linear Regresion

Final Project Report: MDA 9159-9160

Zerun Xiao and Jasleen kaur Saini

TABLE OF CONTENTS

1-INTRODUCTION

Background

Education plays an important role in society as it is crucial for the sustainable development of economic progress (Stephens, 2018). Recent years, Portugal has undergone a major economic crisis that is arose from aging populations and low productivity. The Portuguese Government see the importance of enhancing the educational level of the population in order to increase the competitiveness (Williams, 2017). With the abundance of data and the rise of quantitative techniques, the combination of statistical modelling and the domain knowledge could help organizations assess the process and create valuable insights. It is shown that there is an increasing use of large datasets by education economists (Rogge et al., 2017).

Motivation

Statistical analysis can be used in education and learning to improve the quality of learning process, evaluate the performance of the student, improve based on feedback, and enrich the learning experience. While it is hard to predict student performance because there are a variety of factors that will affects the student's final grade. Therefore, the purpose of this study is to predict students' academic performance by performing statistical modeling such as multiple linear regression and determine what are the most important predictors. The study is organized as follows: data cleaning, exploratory analysis, feature selection, model building, model diagnostic and model performance.

2-General Description of Dataset

2.1-Know the Data

2.1.1 General Description

The dataset access student performance in Portuguese language classes. It was built from school reports and questionnaires from two Portuguese institutions of secondary education. The dataset contains attributes of student's personal information, family information, leisure habits and study habits. The response variable is final year grade of the Portuguese language.

2.1.2 Variable Description

There are 649 students in the sample, 30 predicator variables, and three response variables. By looking at the structure of the data, we have 17 categorical variables and 16 integer features including the response variable G3. Below is the description details of features.

Attribute	Description (Domain)
sex	student's sex (binary: female or male)
age	student's age (numeric: from 15 to 22)
school	student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*)
address	student's home address type (binary: urban or rural)
Pstatus	parent's cohabitation status (binary: living together or apart)
Medu	mother's education (numeric: from 0 to 4[a])
Mjob	mother's job (nominal[b])
Fedu	father's education (numeric: from 0 to 4[a])
Fjob	father's job (nominal[b])
guardian	student's guardian (nominal: mother, father or other)
famsize	family size (binary: ≤ 3 or > 3)
famrel	quality of family relationships (numeric: from 1 – very bad to 5 – excellent)
reason	reason to choose this school (nominal: close to home, school reputation, course preference or other)
traveltime	home to school travel time (numeric: 1 – $<$ 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – $>$ 1 hour).
studytime	weekly study time (numeric: 1 – $<$ 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10 hours)
failures	number of past class failures (numeric: n if $1 \leq n < 3$, else 4)
schoolsup	extra educational school support (binary: yes or no)
famsup	family educational support (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
paidclass	extra paid classes (binary: yes or no)
internet	Internet access at home (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
freetime	free time after school (numeric: from 1 – very low to 5 – very high)
goout	going out with friends (numeric: from 1 – very low to 5 – very high)
Walc	weekend alcohol consumption (numeric: from 1 – very low to 5 – very high)
Dalc	workday alcohol consumption (numeric: from 1 – very low to 5 – very high)
health	current health status (numeric: from 1 – very bad to 5 – very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	final grade (numeric: from 0 to 20)

a 0 – none, 1 – primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
b teacher, health care related, civil services (e.g. administrative or police), at home or other.

2.2-Cleaning and Shaping Data

Since the interest of target is only final grade, so 'G1', 'G2' are dropped. Variables like 'fam_size' and 'address' does not seem to affect the final grade, so they are dropped as well. Next is to check missing values, it turns out that there was no missing value in our dataset. Then, I renamed the names of columns to more explicit names so that they can be easily recognized in the analysis phase. Now, the dataset contains 649 observations of 28 predictors and one response variables

2.3-Data Visualization

2.3.1 Response Variable

The response variable is the final grade of Portuguese language class, so the grade distribution can be shown in the histogram format to give us a general overview of the response variable. It can be shown from Figure 1 that there were some students clustered in the grade of 0, by looking at the statistics, there were 15 of 649 students got grade of 0. It is not clear that this phenomenon is an outlier or indeed there is such several students perform poorly. The distribution of the final grade is a little bit right skewed. By looking at the summary of the G3, it shows that minimum grade is 0, the maximum is 19, the median is 12, and the mean is 11.19.

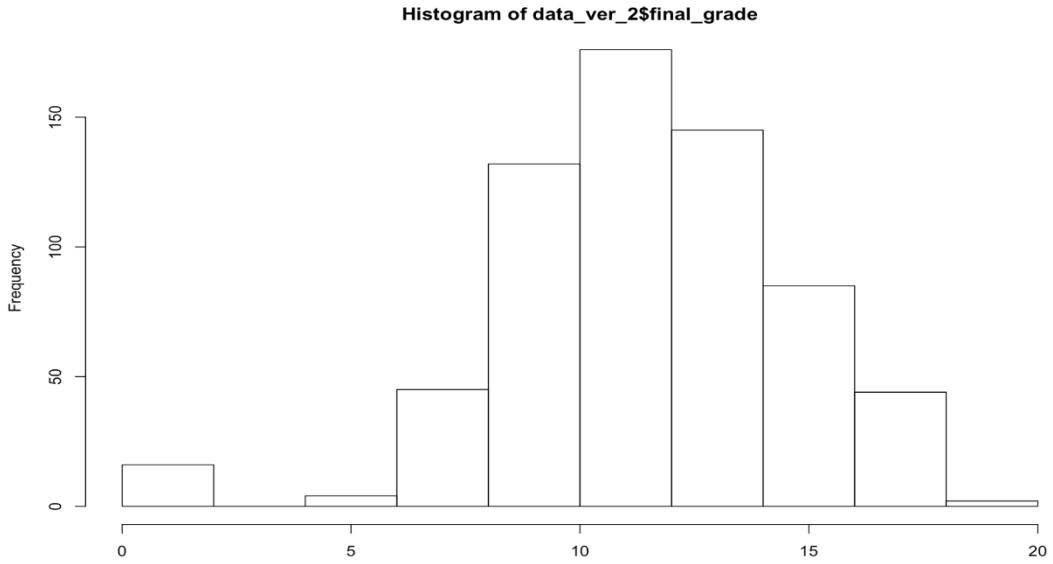


Fig1: Histogram of final grade distribution in the Portuguese Class

2.3.2 Correlation Matrix

The Pearson correlation test is performed between pairs for numerical data to remove the highly correlated variables, because they will affect the significance of the whole model. The threshold is set to 0.5, so we removed one of two variables that has correlation either greater or less than |0.5|. From the correlation matrix below, it can be seen that 'workday_alcoholic' and 'weekend_alcoholic' has a correlation of 0.56, 'mom_edu' and 'dad_edu' are also highly correlated with a correlation of 0.57.

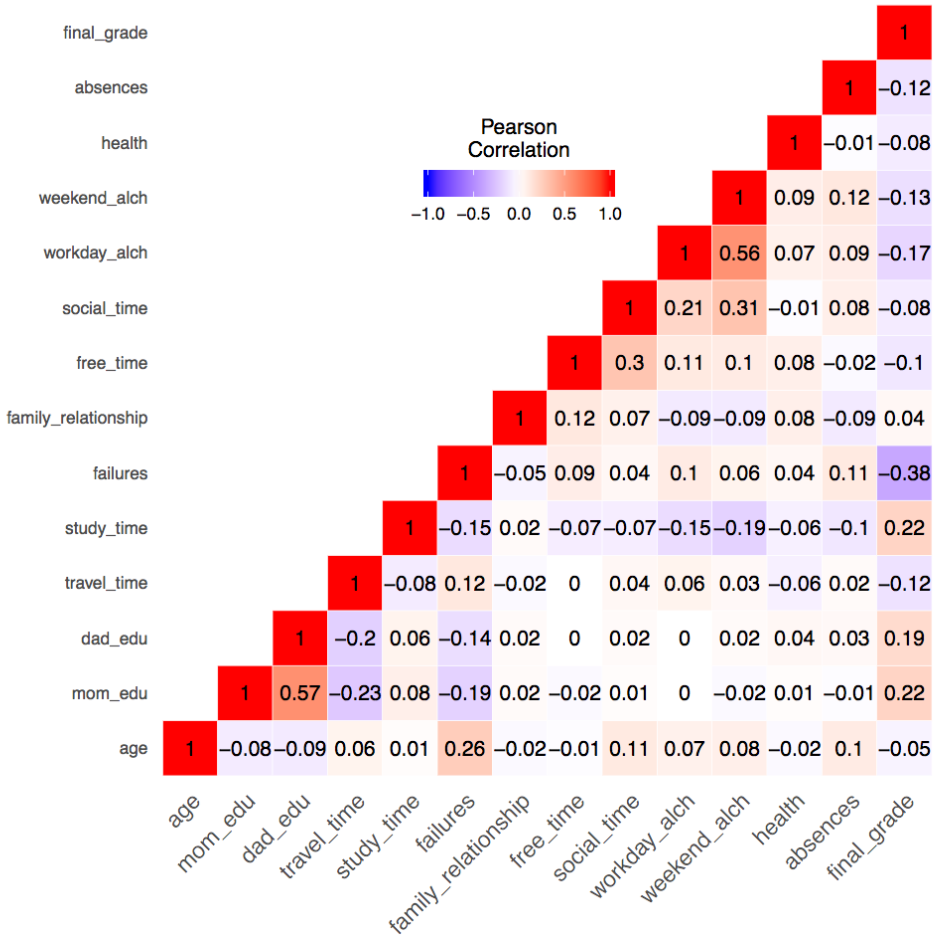


Figure 2: Correlation Matrix Between Numerical Variable

Some of the plots of predictors versus response and their interpretation are described.

First, we have gender distribution plot: There were 59% of students are female, and 41% of students are male in the class. For the grade distribution between different genders, it can be shown from figure 3 that female has higher median than male. It seems that female performs better and has more higher grades than male, but there is not huge difference.

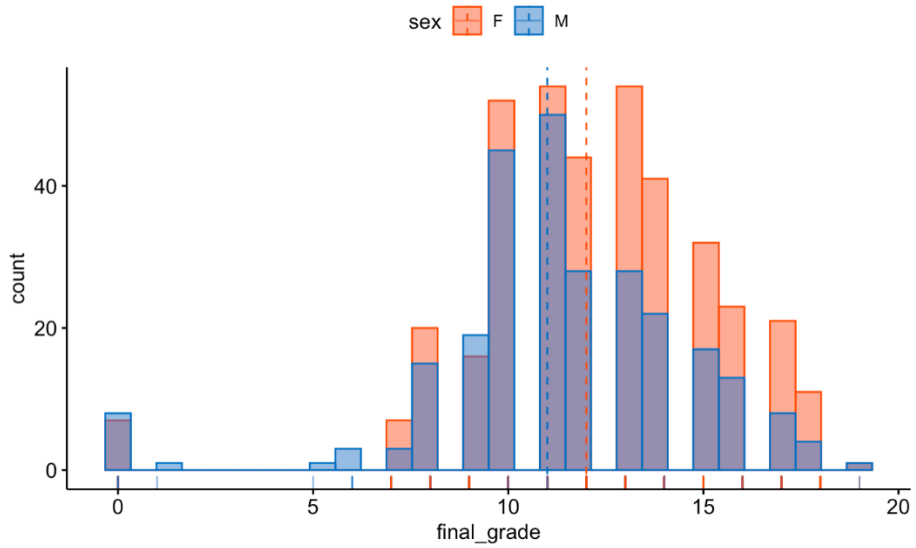


Figure 3: Distribution Plot of G3 vs. Sex

Second, for the variable 'higher' that refers to the intention to take higher education, it is obvious that students who has intention of higher education had a higher mean score than those who does not as shown below.

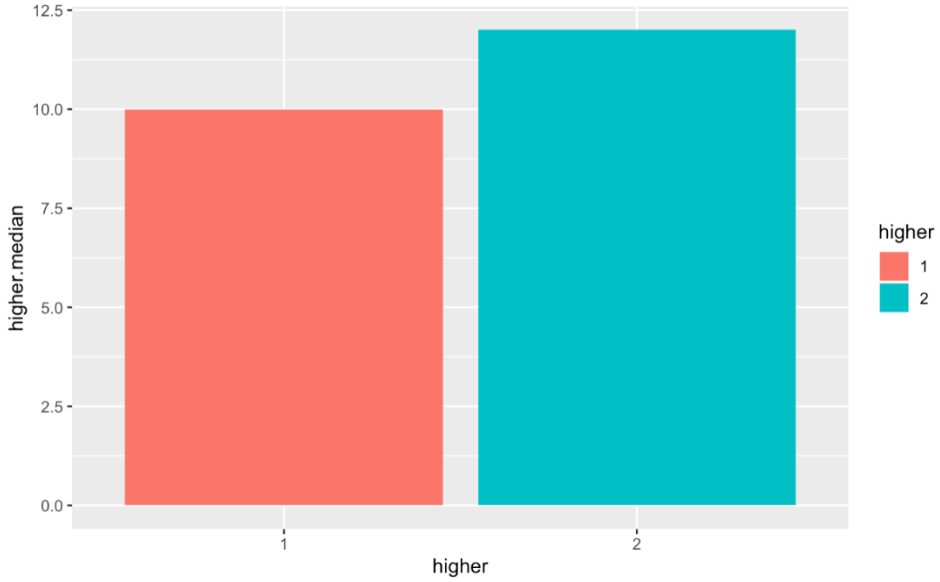


Figure 4: Mean Score Comparison Between Higher Education Intention

Third, for the variable, 'mom's edu' showed the most relevant trends. From the grade distribution plot among different levels of mom's education in figure 5, as mom's educational level increases, final grade G3 increases linearly as indicated in graph below.

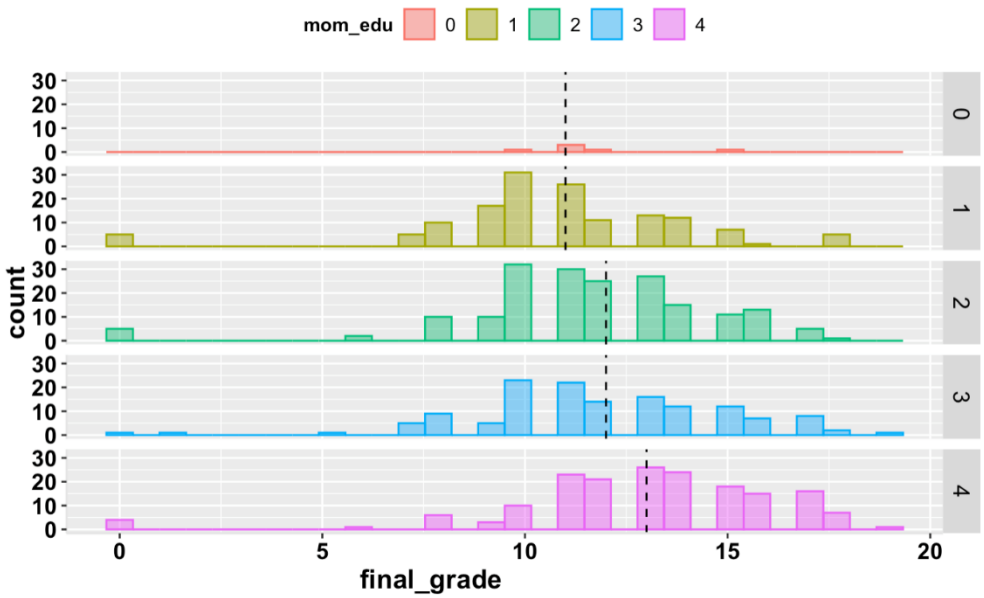


Figure 5: Distribution Plot of G3 vs. Mom_Edu

2.4-Outliers and Influential Points

We checked outliers of numerical predictors like ‘mom_edu’, ‘study_time’, ‘workday_alco’, and ‘health’ by using boxplot. There is no significant number of outliers found, so we skipped removing outliers from data.

3-MULTIPLE LINEAR REGRESSION

3.1-Feature selection

For the sake of our understanding we divided our data into following categories.

CATEGORY	DESCRIPTION	PREDICTORS
1	Student's information	school,age,sex,nursery,higher,romantic,reason,health
2	Student's family information	parents_cohab,mom_job,dad_job, guardian,family_sup,family_relationshp,mom_edu,mom_edu
3	Student's study habits	study time, travel time, failures, education support ,paid, absences
4	Student's leisure interests	activites,internet, free_time,social_time,social_time,workday_alch,weekend_alch

Feature selection involves observing correlation of predictors with each other. Since we found our correlation matrix before, this helps in finding which predictors are important and which can be eliminated to form a model such that we have least number of predictors that can explain the model. Although this seem more random, but according to correlation matrix we found that when response is fitted with all predictors majority of predictors have higher p-values indicating their correlation with each other.

We built few fit models initially to determine how certain predictors behave with and without predictors the have correlation with.

1- G3 is fitted with Mjob
2- G3 with Medu and Mjob
3- G3 with Dalch
4- G3 with Dalch,Walc

These predictors were chosen because it seemed they were highly correlated and either of the mom education or mom job can be used to predict the overall model. Below are the p-values for first two cases. Below two models tell that for fit_sample_2 mom_job is not significant for prediction when we have mom_edu model but lower value of mom_job in fit_model_1 shows due to lower p-value we found sufficient evidence to reject the null hypothesis, thus mom_job is significant when alone in model.

```
Call:
lm(formula = final_grade ~ mom_job, data = data_ver_2)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1389  -1.6705  -0.0444   1.9556   6.9556

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.0444     0.2727  40.494  < 2e-16 ***
mom_job2      2.0181     0.5325   3.789 0.000165 ***
mom_job3      0.6261     0.3366   1.860 0.063345 .
mom_job4      1.1026     0.3850   2.864 0.004321 **
mom_job5      2.0944     0.4625   4.529 7.06e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.169 on 644 degrees of freedom
Multiple R-squared:  0.04377,   Adjusted R-squared:  0.03783
F-statistic:  7.37 on 4 and 644 DF,  p-value: 8.305e-06
```

```
Call:
lm(formula = final_grade ~ mom_job + mom_edu, data = data_ver_2)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1761  -1.6740   0.0598   2.0553   7.3984

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.49738    1.29343   8.889   <2e-16 ***
mom_job2     0.91902    0.60256   1.525    0.128
mom_job3     0.33857    0.34366   0.985    0.325
mom_job4     0.51852    0.41735   1.242    0.215
mom_job5     0.74992    0.58551   1.281    0.201
mom_edu1    -0.89579    1.30871  -0.684    0.494
mom_edu2    -0.16196    1.30742  -0.124    0.901
mom_edu3     0.03525    1.31908   0.027    0.979
mom_edu4     0.92882    1.33965   0.693    0.488
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.14 on 640 degrees of freedom
Multiple R-squared:  0.0669,    Adjusted R-squared:  0.05524
F-statistic: 5.736 on 8 and 640 DF,  p-value: 4.534e-07
```

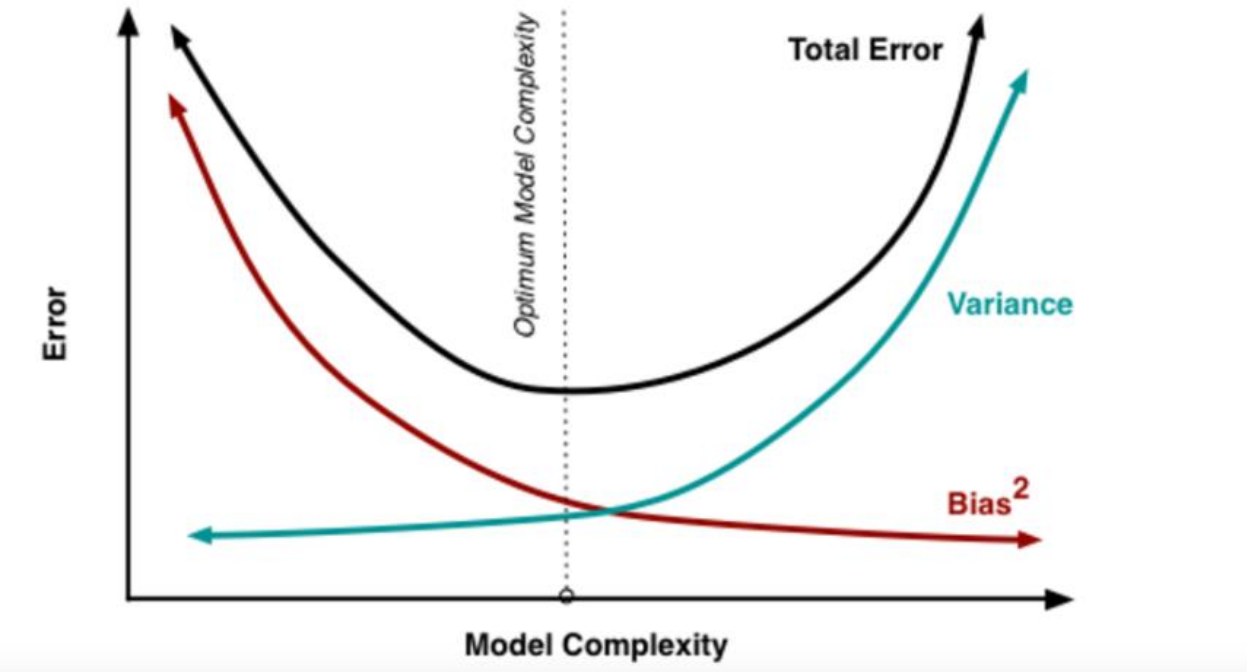
These results indicated most of predictors become less significant when regressed with all other predictors against response variable. This basis helped us eliminate predictors that had collinearity with other predictors. Also, we observe no quadratic relation between pairs of predictors thus eliminating the need for quadratic model. On the contrary we can consider some interaction models where Dedu and Djob seemed to have some general collinearity and Djob and Mjob seem to have some collinearity. Thus, we regress our model including these interaction terms as well. We can get certain models by eliminating and adding interaction terms in models but since there seemed enough correlation between some categories, we use LASSO shrinkage method to obtain fit model and feature selection.

Before proceeding with model constructing, we found that shrinking some unimportant predictors to zero would simplify our model and still perform better so we use shrinkage methods. This forms basis of bias-variance tradeoff. In the Ordinary Least Squares (OLS) approach, we estimate them as betas in such a way, that the sum of squares of residuals is as small as possible. In other words, we minimize the following loss function:

$$L_{OLS}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 = ||y - X\hat{\beta}||^2$$

The OLS estimator has the desired property of being unbiased. However, it can have a huge variance. Specifically, this happens for the dataset we are using:
1-The predictor variables are highly correlated with each other;
2-There are many predictors.

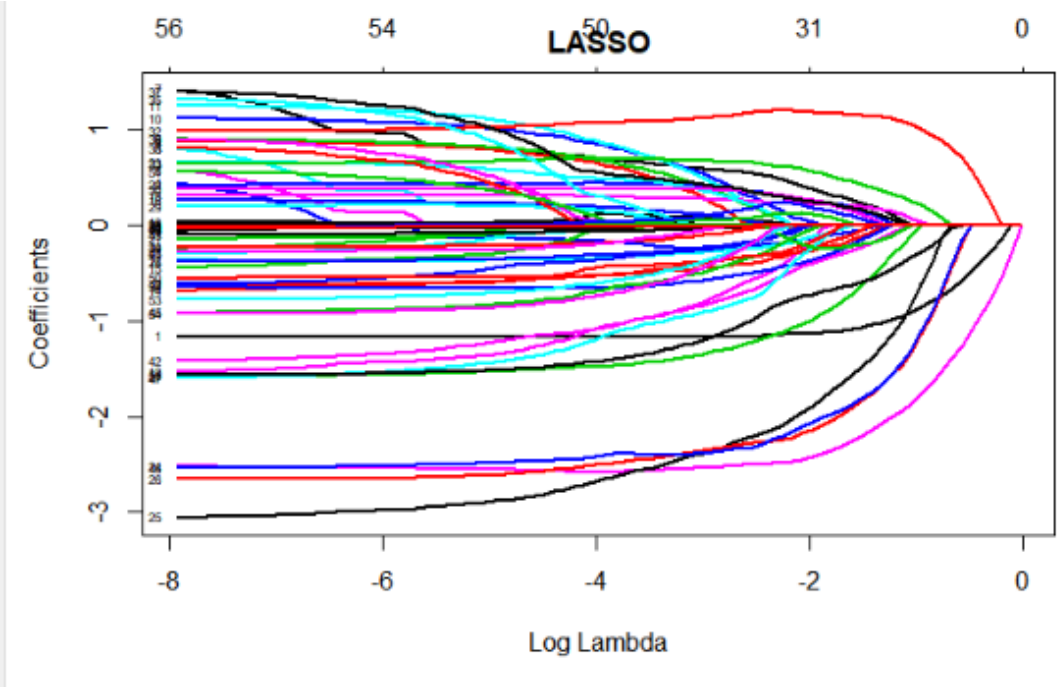
So general solution is reduced variance at the cost of introducing some bias based on complexity graph below.



We use LASSO and ridge to perform this and both shrink beta coefficients in a way that it reduces variance and increases bias by introducing some penalty for adding predictors to models. If we increase number of predictors both will penalize the model, but LASSO will eliminate certain predictors that are not useful thus complexity of model and achieving lower variance. We use LASSO and ridge over AIC, BIC or adjusted R squared because our aim was to predict how well the model will perform on future predictions and these metrics were more useful if only getting important predictors was our aim of study and we would miss effect of removed predictors on our response. So we choose to analyze and predict response G3 based on important predictors and reducing beta coefficients for all to non-zero(ridge) or for some to zero (LASSO). Below is the loss function for LASSO which we target to minimize.

$$L_{lasso}(\hat{\beta}) = \sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda \sum_{j-1}^{m}|\hat{\beta}_j|.$$

The lambda is regularization parameter which we choose using K-fold cross validation by splitting data into test and training set and thus getting LASSO fit model as below.



3.2-Model Construction

We had already chosen LASSO and ridge models. Based on our data exploration and predictors obtained from LASSO we attempted to build certain models by choosing some predictors from each category. Finally, we use MSE metric to compare models that have lowest MSE as it indicates how far are our actual response values from fitted values. We have constructed 4 new models apart from LASSO, ridge and linear models based on removing predictor from each category

MODELS	VARIABLES USED
CATEGORY 1	Removing reason and romantic predictors
CATEGORY 2	Removing dad_job,family_sup
CATEGORY 3	Removing travel_time
CATEGORY 4	Removing weekend_alch

4-Summary of Output and Results

From our models and exploratory analysis, we find some interesting outputs:

- Parents' education and job levels demonstrate a positive correlation with the academic performance of students.
- Female students achieved a higher score than male students across all three grades.
- Students who consume less alcohol over the weekend / schooldays perform better academically
- Number of past class failures, student age, higher, reason are significant predictors for final grade's multiple regression model.

The full linear model has large number of predictors who become significantly less when they are regressed with other predictors against response variable. More than 50 percent of predictors from our dataset are significant and can't be dropped because of which our choice of Ridge and LASSO over AIC, BIC seems justified. Also, we see lasso tends to do well if there are a small number of significant parameters and the others are close to zero (when only a few predictors influence the response). Ridge works well if there are many large parameters of about the same value (when most predictors impact the response) which seem like our case.
We also got lower MSE from models where some of the predictors were removed as compared to linear model but still higher than that from LASSO or Ridge, so we choose LASSO over other models because of its lower complexity, better performance and lower variance over other models as shown in below output. We also tend to find if this model holds all model assumptions by plotting fitted vs residual plots like below clearly showing these got violated.

```
Different MSE are

MSE from Linear model is: 9.172137

MSE from Ridge model is: 8.259019

MSE from LAsso model is: 8.376249

MSE from model without category 1 variables is: 9.152325

MSE from model without category 2 variables is: 8.991174

MSE from model without category 3 variables is:: 9.128974

MSE from model without category 4 variables is: 8.969887
```

Our final model can be expressed as:

final_grade=beta1*school+beta2*sex+beta3*health+beta4*higher+beta5*romantic+beta6*reason
+beta7*dad_edu+beta8*mom_job+beta9*dad_job+beta10*family_sup+beta11*family_relationship
+beta12*travel_time+beta13*study_time+beta14*failures+beta15*absences+beta16*edu_sup
+beta17*activities+beta18*internet+beta19*free_time+beta20*social_time+beta21*workday_alch+beta22*weekend_alch

5-Discussions of findings and references

Discussions

Our interest to find model that could predict well on future observations is returned from ridge model
Based on MSE score we get least scores from ridge and lasso. Both tend to vary in number of predictors used for modeling. Since the performance for both is not much different but LASSO provide simpler model that has low variance and more bias as compared to other models, we can say it would perform well for future predictions also. Since our final model obtained has following predictors that are significant in predicting final grade G3 of student we can discuss few questions:

1-Does one gender excel another?

Yes, the mean percent female score is higher compared to male and from plot we have statistical evidence to prove our assumption derived from descriptive statistics.

2-Does intention to study higher help to excel scores?

Yes, the mean percent of students who has want to do higher studies has higher than the mean percent of students who does not and from plot we have statistical evidence of median scores to prove our assumption derived from descriptive statistics.

3- Does family dynamics affect the final-grades of student?

Yes, majority of family dynamics affect the final grade of students.

4-What all variables can be ignore while predicting final grade of students?

Address, nursery, dad education can be dropped while predicting final grades of students.

References

Dataset- https://www.kaggle.com/uciml/student-alcohol-consumption#student-por.csv

Jonathan, W. (2017), "Addressing the challenges in Portuguese higher education". Retrieved from https://www.hks.harvard.edu/sites/default/files/centers/mrcbg/files/81_Technical%20Report%20-%20Final.pdf.

Stephens, M. D. (Ed.). (2018). *Universities, education and the national economy* (Vol. 30). Routledge.

Vanthienen, J., & De Witte, K. (Eds.). (2017). *Data analytics applications in education*. CRC Press.

6-Limitations and further questions raised by model

Our model doesn't seem to follow any model assumptions as indicate by bp and Shapiro test

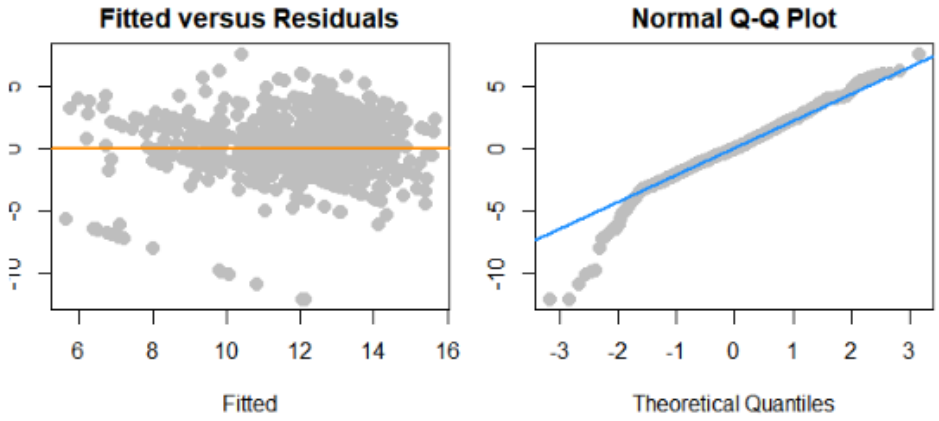
```
        studentized Breusch-Pagan test

data:  fit_final
BP = 94.138, df = 52, p-value = 0.0003135


        Shapiro-Wilk normality test

data:  resid(fit_final)
W = 0.95988, p-value = 2.572e-12
```

Another limitation is our data is almost categorical and non-linear and this doesn't seem to match most of the regression methods we used. There is no random error in data and so below plots represent the distribution of data which is almost non- linear non normalized and doesn't hold any equal variance. No quadratic or interaction models could be used as there was no relationship between most of the predictors and due to lack of this continuous data limited regression methods could be followed to obtain an optimal model as observed from graph below.



The dataset has a lot of factors whose levels are not justified like mom job, dad job, age, school which can vary from sample to sample. This forms very limited number of samples that can be drawn or test scenarios for predicting how response variable will behave with respect to predictors for future observations.