UPPSALA
UNIVERSITET

A comparative analysis of ridge and lasso regularization in predicting student performance

Max Hansson and Martin Holmqvist

Project work statistical programming

2023

Abstract

This study investigates the predictive performance of ridge and lasso regularization in high school student mathematics performance. Using a dataset from Portuguese high schools, the research compares the effectiveness of ridge and lasso in predicting final grades and classifying pass/fail outcomes. Lasso regularization outperforms ridge in predicting final grades, indicating its ability for variable selection and creating a simpler, more accurate model. Diagnostic plots reveal challenges with non-normal residuals, with lasso exhibiting better behavior. However, statistical tests do not show a significant difference between the models, suggesting that observed differences could be due to randomness. This study emphasizes the importance of considering regularization techniques in predicting student outcomes, offering insights into potential improvements for educational support systems.

Contents

1 Introduction

The school system lays the foundation of our society. It is shown that education is positively related to economic growth (Hanushek and Kimko, 2000). Hence, it is of interest to increase the level of education to yield growth. It has also been shown that education is related to life expectancy of individuals (Psacharopoulos, 1994). Therefore, it is important to give each individual a fair chance to succeed in school. This is not always the case, school systems around the world often fail to provide struggling students the education they need to pass (OECD, 2016). Some students need extra support to do well in school. Creating good prediction models could help schools react in time and work preemptively.

Hence the ability to predict student performance, if solved, would contribute to great opportunities for parents, teachers and students. It could be used to find key indicators associated with good grades, help students who are at risk of struggling, and teach them the most important parts of planning for success in their studies. Being able to provide early help and guidance is crucial for the educational system to function properly and to reduce the number of failing students. With the recent success and popularity of machine learning techniques, it is interesting to investigate their potential to predict the performance of students and compare different techniques to each other.

Previous research has been made using the same data set, where the general goal of the research has been predicting the final grades accurately. The models have been fairly accurate, but still not better than a naive model that just predicts the last grade for the final one as well (Cortez, 2014). We will study the predictive and classifying performances of 4 models, using two different regularization techniques. We will use ridge and lasso regularization to predict final grades and penalized logistic regression to predict if a student will pass or fail the course also using ridge and lasso regularization. The purpose of this study is to highlight the differences between these methods and perhaps contribute to future research by discussing the pros and cons of each technique.

Our research question is:

- Is there a significant difference in performance between ridge and lasso regularization when predicting student grades?

:

The research question will be answered by comparing the predictive performances between ridge regression and lasso regression predicting Portuguese students' final grades in math. This will be evaluated by a paired t-test. We will also compare the classifying performance of a ridge logistic regression and lasso logistic regression classifying if students will pass or fail in math. This will be evaluated using a Mcnemar´s test.

2 Data

To examine and answer the research questions, data with information regarding high school student's performance in the subject of mathematics will be used. In addition to the students' grades which are measured on a scale between $0 - 20$, the data set contains 33 variables regarding demographic, social, and school-related variables. The data was collected from two different schools in Portugal between 2005 and 2006 by Paulo Cortez and Alice Silva. A survey was handed out to 788 students with questions regarding the variables in table 1. This resulted in a dataset with 395 complete data points. The students included in the study were all last-year students in high school. Cortez had problems mapping the students' answers with their grades which is why there is quite a big loss in data. To pass the course, students need 10 or higher as their grade. Cortez and Silva (2008).

Table 1: Description of the variables in the dataset.

Variable	Type	Description
School	binary	Gabriel Pereira or Mousinho da Silveira
sex	binary (Female or Male)	student's sex
excluding female in models		
age	numeric (15 to 22)	student's age
address	binary (Urban or Rural)	student's home address type
famsize	binary (≤ 3 or > 3)	family size
Pstatus	binary (Together or Alone)	parent's cohabitation status
Medu	categorical (0 - 4)[1]	mother's education
Fedu	categorical (0 - 4)[1]	father's education
Mjob	nominal[2]	mother's job
Fjob	nominal[2]	father's job
reason	nominal[3]	reason to choose this school
guardian	nominal[4]	student's guardian
traveltime	numeric (1 to 4)[5]	home to school travel time
studytime	numeric (1 to 4)[6]	weekly study time
failures	numeric (1 to 4)[7]	number of past class failures
schoolsup	binary (Yes or no)	extra educational support
famsup	binary (Yes or no)	family educational support
paid	binary (Yes or no)	extra paid classes
activities	binary (Yes or no)	extra-curricular activities
nursery	binary (Yes or no)	attended nursery school
higher	binary (Yes or no)	wants to take higher education
internet	binary (Yes or no)	Internet access at home
romantic	binary (Yes or no)	with a romantic relationship
famrel	numeric (1 to 5)[8]	quality of family relationships
freetime	numeric (1 to 5)[9]	free time after school
goout	numeric (1 to 5)[9]	going out with friends
Dalc	numeric (1 to 5)[9]	workday alcohol consumption
Walc	numeric (1 to 5)[9]	weekend alcohol consumption
health	numeric (1 to 5) [10]	current health status
absences	numeric (0 to 93)	number of school absences
G1	numeric (0 to 20)	first period grade
G2	numeric (0 to 20)	second period grade
G3	numeric (0 to 20)	final grade

[1] 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education.
[2] Teacher, health care related, civil services (e.g. administrative or police), at home or other.
[3] close to home, school reputation, course preference or other
[4] mother, father or other
[5] 1 – $<$ 15min , 2 – 15 to 30 min, 3 – 30 min to 1 hour or 4 – $>$ 1 hour
[6] 1– $<$ 2hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – $>$ 10hours
[7] n if $1 < n < 3$ else 4 if $n > 4$
[8] from 1 – very bad to 5 – excellent
[9] from 1 – very low to 5 – very high
[10] from 1 – very bad to 5 – very good

2.1 Datacleaning

The data set is downloaded from the UCI Machine Learning Repository database and is open for adaptation (Cortez, 2014). Since the data set obtained from the UCI is the cleaned data we have to trust Cortez and Silva with the cleaning and that they had no intention to create any bias. Since many of the variables are measured on a 1-5 Likert scale and are categorical data, these have been recoded to dummy variables. This results in 42 independent variables to predict the grades. No variables have been removed from the original data since the regularization techniques we will investigate will be able to handle and do the selection.

2.2 Exploratory analysis

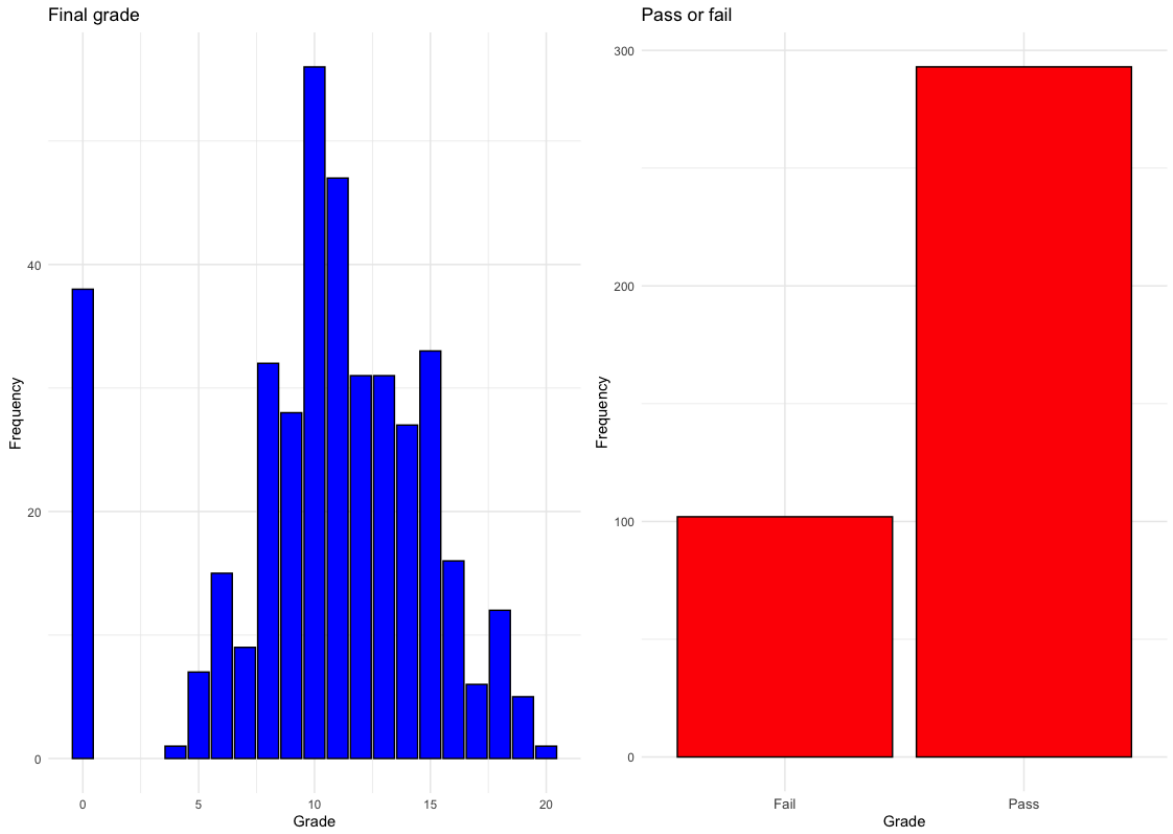

Figure 1: Barplot of final grades in math and barplot of how many students failing and passing the course

Figure 1 shows the distributions of the final grades plotted to the left in the figure. Except for the large number of students who got a grade of 0, the data looks approximately normal. The most common grade is 10 which is the lowest possible grade while still passing the course. Figure 1 also shows the amount of students that passed or failed in math to the right. Approximately 25% of the students failed which makes our data a bit imbalanced but it should not be an issue when training the models since we still have more than 100 observations in the "Fail" category.

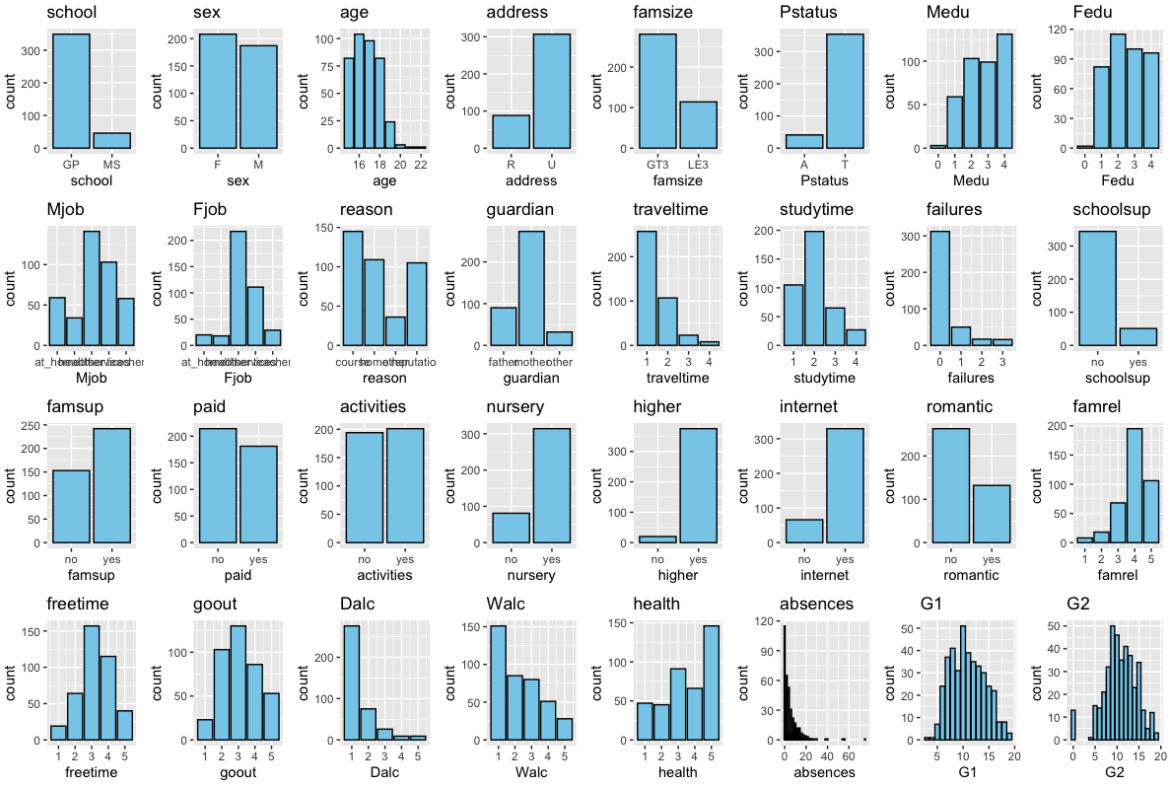

Figure 2: Distribution plots of independent variables

In figure 2 we have plotted the distributions for all the independent variables in the dataset. Some features look a bit imbalanced. However, since the models are equipped to handle many variables and high amounts of multicollinearity, all variables will be included.

In Figure 3 we can see a plot that shows the variance inflation factor values on a regular OLS model. The dotted line shows a variance inflation factor (VIF) of 5. This is generally regarded as a good threshold for when the VIF value should be considered high. Some other thresholds have been suggested, such as a VIF value of 10, but there is no real consensus here (O'brien, 2007). Here we can see that some high levels of multicollinearity could cause the linear regression model to suffer from inconsistent and high variance estimators. Thus, regularization like ridge and lasso would seem appropriate.

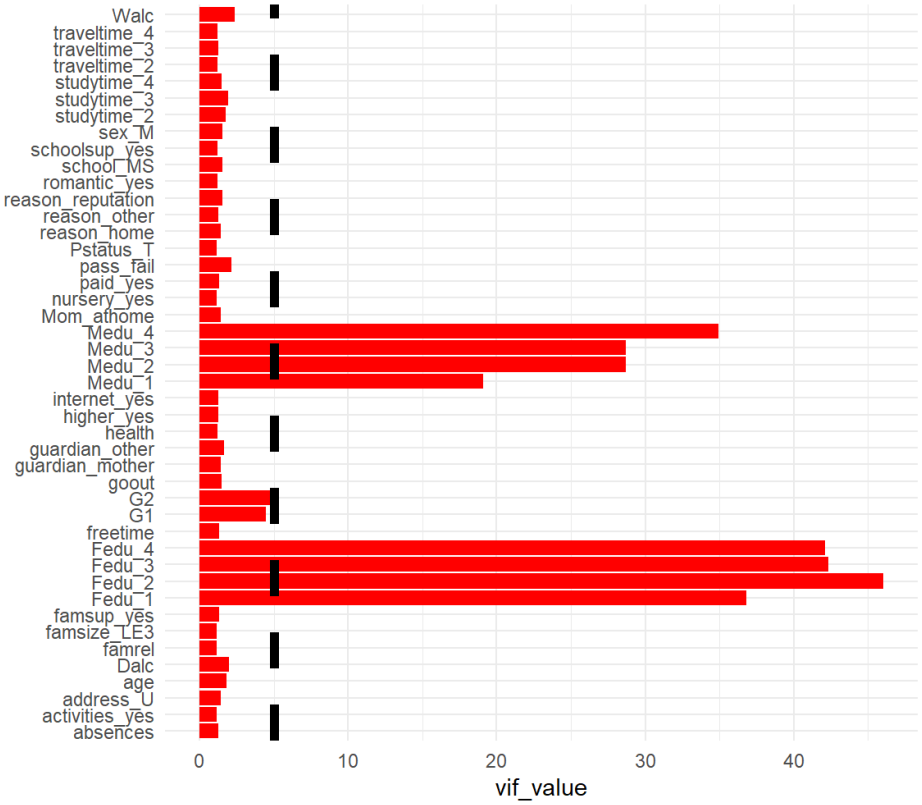

Figure 3: VIF values of data

3 Methods

Here we will present the statistical methods being used for the study as well as the method for the statistical analysis

3.1 Penalized regression

One of the statistical methods to be used in this study are ridge regression and lasso regression, which are different forms of penalized regression. They are both based on ordinary least squares regression. This means that certain assumptions need to be fulfilled such as linearity and homoscedasticity in the data.

Equation 1 shows the loss function OLS minimizes. This can lead to overfitting and a bad out-of-sample error when making predictions. When fitting data with high multicollinearity, the variance of the coefficients is inflated which then leads to a less accurate model. Ridge and lasso regularization are useful tools to avoid overfitting a model and will reduce the variance of the coefficients. This is achieved by adding a penalty term to the OLS regression. OLS

regression minimizes the in-sample error.

$$\text{in-sample error} = \sum_{i=1} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{1}$$

Ridge Regression, instead minimizes (2)

$$\text{in-sample error} + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

Finally, lasso regression minimizes (3)

$$\text{in-sample error} + \lambda \sum_{j=1}^{p} |\beta_j| \tag{3}$$

In ridge regression, the variable coefficients will tend to 0 as the penalty term, λ, increases. In lasso regression, the variable coefficients will not just tend to zero, they will become zero once λ is large enough. This means that lasso regression will produce a simpler model that reduces the number of parameters and can remove parameters that are not useful. Ridge on the other hand can never remove a parameter altogether since the penalty term takes the square of the coefficients instead of the absolute value (Tibshirani, 1996). In general, lasso tends to perform better when a lot of the variables are bad predictors of the dependent variable. Ridge is proved to outperform lasso when data suffers from severe multicollinearity due to the properties of the penalty term (Herawati et al., 2018).

3.2 Penalized Logistic Regression

When the aim is to classify a binary variable, penalized logistic regression is appropriate to use. It is based on regular logistic regression where the logistic regression equation is written in (4)

$$\pi_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}} \tag{4}$$

To estimate the β vector, logistic regression optimizes the log-likelihood function, given as (5)

$$l(\beta) = \sum_{i=i}^{n} x_i log(\pi_i) + (1 - x_i) log(1 - \pi_i) \tag{5}$$

This model could however face challenges such as overfitting and multicollinearity since it will optimize the coefficients on the training data. Penalized logistic regression, similar to

its linear regression counterparts, introduces regularization terms to address these issues. A regularization term to the probability function is introduced.

The lasso β vector is obtained by minimizing (6)

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = -\sum_{i=1}^{n}[(y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i))] + \lambda \sum_{j=1}^{p} |\beta_j| \tag{6}$$

The ridge β vector is obtained by minimizing (7)

$$\hat{\boldsymbol{\beta}}_{\text{Ridge}} = -\sum_{i=1}^{n}[(y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i))] + \lambda \sum_{j=1}^{p} {\beta_j}^2 \tag{7}$$

Similar to ridge and lasso regression, penalized logistic regression allows for the tuning of the regularization parameter to deal with the multicollinearity and overfitting issues (Pereira et al., 2016).

3.3 ROC and AUC

To decide the cutoff points of our logistic models created as in 3.2, an ROC curve will be plotted. From this ROC curve, optimal cut-off points can be retrieved which maximizes the accuracy of the model on the test data. In addition to the ROC curve, some performance-evaluating metrics such as sensitivity, specificity, and area under the curve (AUC) will be analyzed. Sensitivity and specificity are evaluation metrics showing the true positive and true negative rates of the predictions. A true positive corresponds to the model predicting true for a value that is indeed true, and a true negative is when the model correctly predicts a negative value for the dependent variable. (Hajian-Tilaki, 2013) Sensitivity is calculated as in (8)

$$\text{Sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False positives}} \tag{8}$$

Specificity is calculated as in (9)

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}} \tag{9}$$

3.4 K-fold - Cross validation

When optimizing a model you generally optimize some loss function such as Mean Squared Error (MSE) or misclassification rate to evaluate the model's performance. However, when minimizing the MSE for example in our training data, our model may suffer from overfitting, and as a result struggle to predict unseen data. K-fold cross-validation helps estimate the out-of-sample error to prevent overfitting a model. The data is split into k different groups.

The model is then trained on the first $k-1$ groups and then tested on the group that was left out which practically works as "test data". The loss function for this training split is then measured. This process is repeated until all groups have been tested as "test data". The final k-fold cross-validation error metric is calculated in equation 10.

$$Errormetric = \frac{1}{k}\sum_{i=1}^{k} errormetric_i. \tag{10}$$

This approach gives us a good idea of how our model would perform on unseen data (Hastie et al., 2009). Determining the optimal number of groups involves a balance between bias and variance. While there is no real consensus of a recommended value, common choices for k are often 5 or 10, influenced by factors such as the nature of the dataset and available computational resources (Wong and Yeh, 2019).

3.5 McNemar´s test

To compare two binary classifying models predicting the same thing it is common to use Mcnemar´s test (Dietterich, 1998). It is a nonparametric test that compares the two models' misclassification rates for each model, say A and B are the different misclassification rates. The null hypothesis for the test is that the misclassification rate is equal for the models and the alternative hypothesis is that they are different.

$$\chi^2 = \frac{(|B-C|-1)^2}{B+C} \tag{11}$$

Equation 11 is used to test the null hypothesis and it follows χ^2 distribution. Hence the null can be rejected under a given significance level (Raschka, 2018).

3.6 Analysis method

To investigate if ridge or lasso regression is better at predicting student performance, four different models will be created. Each model will have the same variables, see table 1. The regression models will be evaluated in terms of mean squared error and residual standard error. The models will also be evaluated using different diagnostic plots to check the assumptions for a linear model. A paired t-test will be conducted to test if the predictions of the two models are equal or significantly different with a 5% significance level. Here we choose the paired t-test over a regular t-test with the test of means since we have paired observations and is interested of the predictions are significantly different or not. Similarly, the logistic regression models will be evaluated in terms of accuracy, sensitivity as well as specificity. The threshold value for the logistic models will be found using a ROC curve, and in addition the

AUC value will be presented and analyzed. A Mcnemar´s test will be performed to test if the misclassification rate is equal for the to models or significantly different with a 5% significance level. Hence the regularization techniques will be tested for two different prediction tasks. This should provide more information on what regularization technique is best suited for this type of data. No in depth interpretations of the coefficients of each model will be done but only a comparison between them.

Table 2: The 4 different models to be made and evaluated

Model	**Subject**	**Evaluation metric**	**Objective**
Ridge Regression	Math	MSE	Predict final grade
Lasso Regression	Math	MSE	Predict final grade
Ridge Logistic Regression	Math	Accuracy	Predict Pass/fail
Lasso Logistic Regression	Math	Accuracy	Predict Pass/fail

3.6.1 Training of the models

The data set is randomly divided into a train and test. Train set consisting of 70% of the data and a test set of the remaining 30%. The train and test data are the same for all models meaning that they are trained and tested on the same data. The hyperparameter lambda for each model will be optimized using 10-fold cross-validation testing a large sequence of values for lambda. This is important since we want the best model for each regularization technique to make a fair comparison between them. Each model is then to be evaluated on the test data using the evaluation metric in table 2. Since the objective of the study is to compare ridge and lasso we look at what model performs better given the same data and the same objective i.e. which one performs better at predicting or classifying something given a dataset.

4 Implementation

To create the ridge and lasso models the "glmnet" package will be used. This is a good fit for our analysis since the package has functions that specifically tune and fit penalized regression and logistic regression models. The function "cv.glmnet" is used to tune all of our models, here we have specified what kind of model we want (regression och logistic regression in our case), if the model should use ridge or lasso as regularization, and what error metric it should minimize. Here we choose to use MSE for the regression models and misclassification error for the logistic regression models. The function has a standard option to test a default sequence of lambdas of which the models were optimized first. After realizing the regression models had a low lambda we specified the function to cross-validate 10 000 values on lambda between 1 and 0. For the logistic models, we kept the default sequence when cross validating.

Optimizing the hyperparameter lambda is an important task to make a fair comparison between the ridge and lasso models.

4.1 Simulation design

The package "glmnet" will be tested through a simulation to make sure it works as intended. Since both penalized regression and penalized logistic regression are used we need to make a simulation that proves both these models are working properly. Theory suggests that ridge and lasso regression will outperform OLS when predicting a variable when there are high levels of multicollinearity in the data. It is also known that ridge regression should outperform lasso as well when the multicollinearity is severe. Knowing this, it seems reasonable to simulate data with high levels of multicollinearity and compare the ridge and lasso models‘ predictive performances to a benchmark OLS models predictive performance. The expected results will then be, because of the inflated variance, OLS will struggle while lasso will perform better and ridge the best. This method also applies to regular logistic regression and penalized logistic regression using ridge and lasso, where the outcome variable is binary.

The design of the simulation will be to simulate a dataset of 2 000 observations with 20 independent variables and one dependent variable to predict. The variables for the penalized regression models and OLS will be drawn from a normal multivariate distribution. The dependent variable for the logistic distributions will be drawn from a Bernoulli distribution. The hyperparameters were tuned using cross-validation with a training set consisting of 50% of the dataset and the remaining 50% being the test set. This is to be done 5 000 times and then we check the mean of the error metric over these 5 000 iterations to check that ridge model is performing best and the OLS or logistic model is performing worse. This procedure will be done 3 times with the multicollinearity in the datasets first being set to be around 0.4, then 0.6 and lastly 0.8. As the multicollinearity increases, the difference between the models´ performances should increase.

An additional way of testing the package is by doing the same simulation but setting the penalty term in each model to zero. When doing this, all 3 models should fit the same model since to penalization term is gone which makes penalized regression into a regular OLS model and penalized logistic regression into a regular logistic model.

4.1.1 Regression simulation results

When running the simulation for the regression models the mean MSE is presented in figure 4. The mean MSE is the highest for OLS and lowest for ridge for all three levels of multicollinearity which is what we expected. We can also see that the difference in mean MSE between the OLS and the penalized models is increasing when increasing the multicollinearity in the data, as expected.

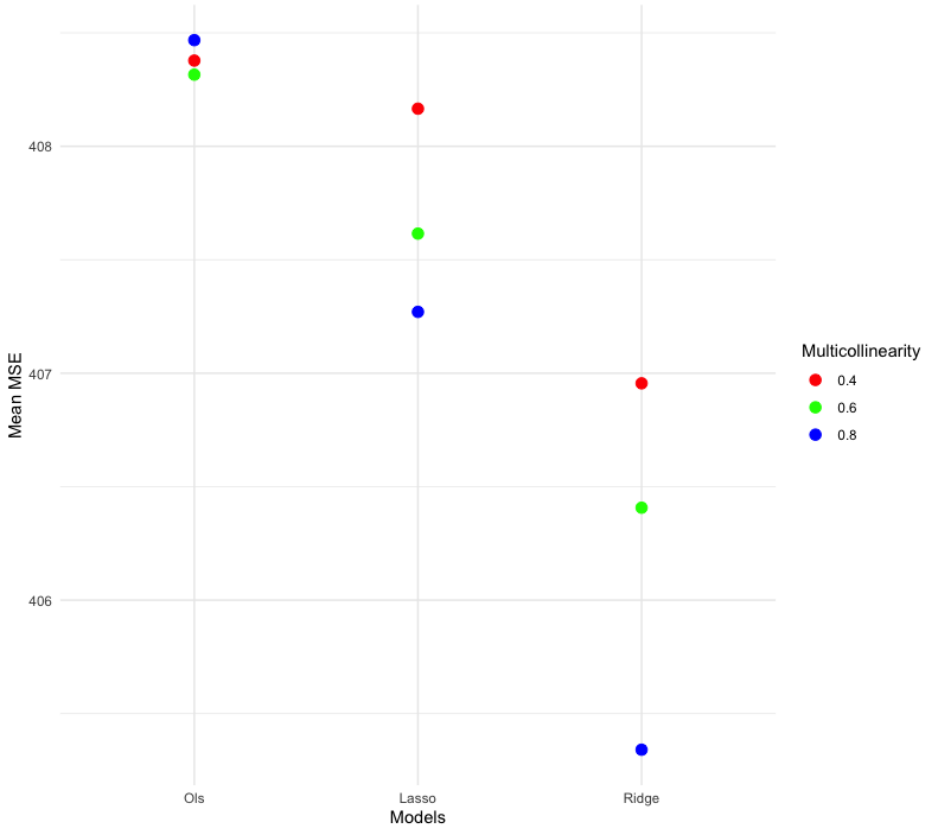

Figure 4: Mean MSE from simulation with different amounts of multicollinearity

When running the simulation where λ is set to zero we can see in table 3 that the mean MSE is exactly the same for the OLS and the penalized models as it should be for the penalized models to work as they should.

Table 3: Mean MSE from simulation when testing $\lambda = 0$

	OLS	Ridge	Lasso
mean MSE	405.0829	405.0829	405.0829

Hence we can state that the "glmnet" package works properly when optimizing and fitting penalized regression models.

4.1.2 Penalized logistic regression results

When simulating the penalized logistic regression models the mean accuracy is presented in figure 5. The mean accuracy is highest for the ridge model and lowest for OLS for all three levels of multicollinearity. We can also see that the difference in mean accuracy between the OLS and the penalized models is increasing when increasing multicollinearity.

When running the simulation where λ is set to zero we can see in table 4 that the mean

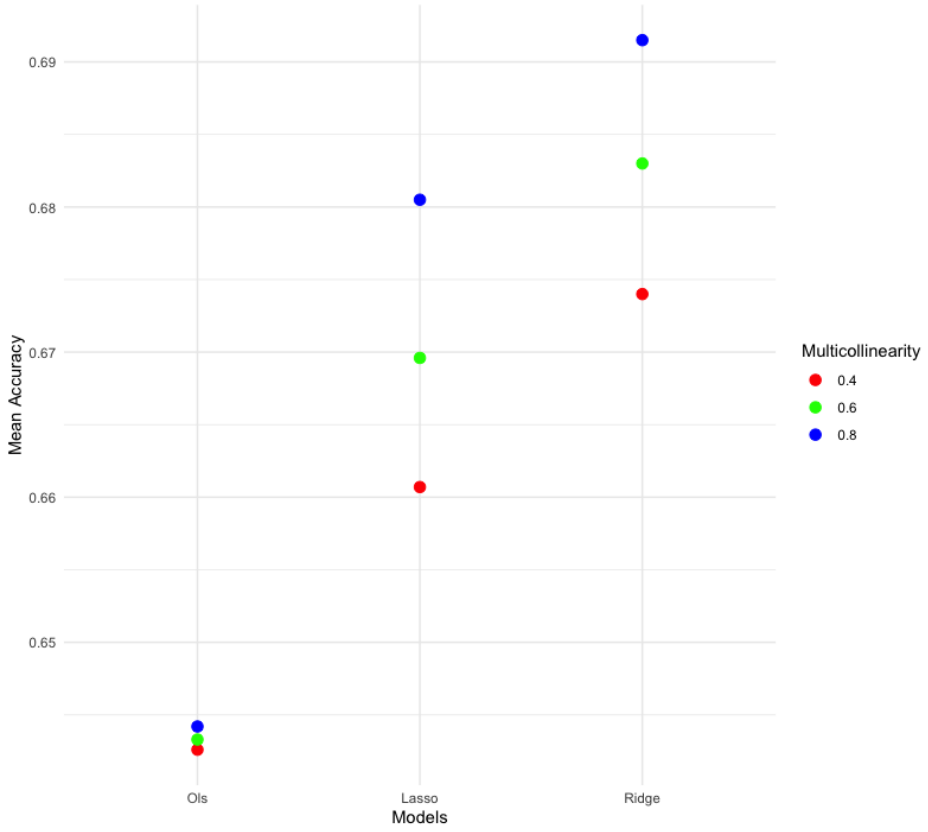

Figure 5: Mean accuracy from simulation with different amounts of multicollinearity

accuracy is exactly the same for all models as it should be for the penalized models to works as they should. In conclusion, we state that the "glm.net" package works properly when optimizing and fitting penalized logistic regression models.

Table 4: Mean accuracy from simulation when testing $\lambda = 0$

	OLS	Ridge	Lasso
mean accuracy	0.663935	0.663935	0.663935

5 Results

Table 5 presents the MSE values and residual standard errors for the penalized regression models. Here we see that lasso outperformed ridge when predicting the final grades for math. This may indicate that there are a lot of bad predictors that lasso removed, creating a more simple model based only on the best predictors. This suspicion is confirmed when looking at table 6, where it is clear that lasso has taken out a lot of variables from the models. Before creating the models, we expected G1 and G2, the previous year's grades, to be the best predictors for the final grade. When looking at table 6, you can indeed see that these variables have large coefficients while many others have been set to zero. Further, note that

lasso not only performed better in terms of MSE but also in terms of residual standard error, meaning it also had a lower spread among the residuals compared to ridge.

Since no feature selection was made when making the models it would have been interesting to investigate how the ridge model would have performed using the variables the lasso model kept. We hypothesize that the performance would increase since there would be less noise in the model. This was however not the purpose of this study and hence it has not been investigated further.

	Ridge	Lasso
(Intercept)	2.17	-0.50
age	-0.34	-0.15
famrel	0.43	0.33
freetime	-0.06	-
goout	0.07	-
Dalc	-0.09	-
Walc	0.04	-
health	0.03	-
absences	0.08	0.06
G1	0.34	0.16
G2	0.82	0.97
Mom_athome	-0.10	-
dad_athome	-0.12	-
school_MS	0.44	-
sex_M	0.16	-
address_U	0.22	-
famsize_LE3	0.42	0.16
Pstatus_T	-0.20	-
nursery_yes	-0.50	-0.07
internet_yes	-0.67	-0.22
schoolsup_yes	0.62	0.15
famsup_yes	0.05	-
paid_yes	0.22	-
activities_yes	-0.52	-0.18
higher_yes	0.25	-
romantic_yes	-0.32	-0.08
Medu_1	-0.38	-
Medu_2	-0.56	-0.24
Medu_3	-0.10	-
Medu_4	-0.20	-
Fedu_1	0.19	-
Fedu_2	-0.35	-0.17
Fedu_3	0.30	0.08
Fedu_4	-0.25	-
reason_home	-0.22	-
reason_other	0.38	-
reason_reputation	0.10	-
guardian_mother	0.25	0.04
guardian_other	0.28	-
traveltime_2	-0.13	-
traveltime_3	-0.75	-
traveltime_4	1.99	0.91
studytime_2	0.00	-
studytime_3	0.24	-
studytime_4	-0.87	-0.45

(a) Regression

	Ridge	Lasso
(Intercept)	-2.30	-10.80
age	-0.24	-0.21
famrel	0.29	0.34
freetime	-0.11	-
goout	-0.14	-0.12
Dalc	-0.11	-
Walc	0.07	-
health	-0.12	-0.09
absences	-0.01	-
G1	0.34	0.13
G2	0.38	1.35
Mom_athome	0.28	-
dad_athome	0.02	-
school_MS	0.23	-
sex_M	0.04	-
address_U	0.02	-
famsize_LE3	0.09	-
Pstatus_T	-0.44	-0.67
nursery_yes	-0.21	-
internet_yes	0.00	-
schoolsup_yes	-0.07	-
famsup_yes	-0.27	-
paid_yes	0.08	-
activities_yes	-0.09	-0.06
higher_yes	1.03	0.78
romantic_yes	-0.37	-0.22
Medu_1	0.53	1.75
Medu_2	-0.23	-
Medu_3	-0.11	-
Medu_4	0.00	-
Fedu_1	-0.10	-
Fedu_2	-0.04	-
Fedu_3	-0.07	-
Fedu_4	0.10	-
reason_home	-0.02	-
reason_other	0.61	0.76
reason_reputation	-0.06	-
guardian_mother	0.08	-
guardian_other	-0.55	-0.65
traveltime_2	-0.18	-0.26
traveltime_3	-0.06	-
traveltime_4	1.62	1.28
studytime_2	-0.12	-
studytime_3	-0.38	-0.30
studytime_4	-0.49	-0.22

(b) Logistic

Figure 6: Ridge and lasso model comparison

The ROC curves presented in figure 7 show the different levels of sensitivity and specificity at different thresholds. When inspecting the plots we can see that the lasso model has a larger area under the curve and peaks further to the top left part of the plot, where the accuracy is the highest. In table 6 the values for sensitivity and specificity have been gathered using the best threshold, where the best threshold was chosen as the one closest to the top left of the

ROC curve. This gives the best combination of sensitivity and specificity available for the models. Here we can see that they have a very similar threshold of about 0.5. Lasso has a higher sensitivity. Further, lasso beats ridge in terms of specificity. This means that the lasso model is better than ridge when it comes to correctly classifying students that passed, but also outperforms ridge when it comes to classifying failing students. Given that accurately identifying failing students is a crucial task for this model, lasso again outshines ridge in terms of performance. Looking at figure 6 b, we can here see that lasso removed several coefficients completely as in the regression model. Hence the conclusion that the model has a lot of bad variables that lasso removes and that ridge cannot, thus outperforming ridge.

Table 5: MSE and RSE Table for the regression models

	Math ridge	Math lasso
MSE	3.23	2.31
RSE	1.80	1.52

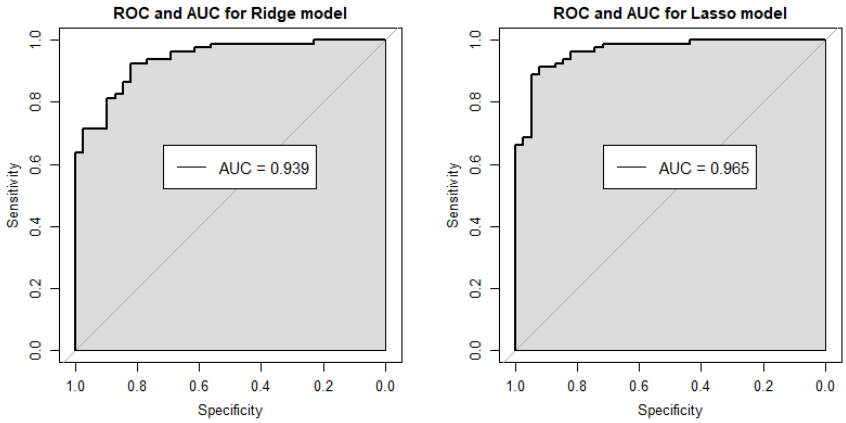

Figure 7: ROC and AUC the penalized logistic regression models

Table 6: ROC and AUC Table

	Threshold	Sensitivity	Specificity	Accuracy
1	0.509	0.925	0.821	0.891
2	0.597	0.912	0.923	0.916

We have also created diagnostic plots for all penalized regression models looking at figure 8 and 9. These are used to investigate the assumptions made when creating linear models to see why it is the lasso model performs better than the ridge model. The assumptions that we need to check are:
- Linear relationship between the independent and dependent variables.
- Independence of the residuals
- Homoscedasticity of the variance of the residuals
- Normality of the residuals

- The residuals vs fitted plot can indicate if there is a non-linear relationship among the residuals in the data. We observe no unexpected patterns. Diagonal patterns in the

plot are observed, but this is caused by the response variable being an integer between 0 and 20.

- The normal Q-Q plot investigates the normality of the residuals. It appears that the lower quantile residuals cause some problems, especially for the ridge model. This result is sensible considering that lasso outperformed ridge, and both models assume normality among the residuals. The ridge models do not seem to fit the data as well as the lasso models.
- The scale-location plot allows us to investigate the assumption of homoscedasticity of the variance of the residuals. This plot looks okay for all models, though some individually large square root residuals are observed.
- The last plot shows a histogram of the distribution of the residuals. This confirms the normal Q-Q plot, where the ridge model seems to violate the normality assumption, and the lasso model looks better.

In summary, these plots align with the results tables presented, where lasso outperformed ridge. It is evident that the lasso model fits the data better and can fulfill the assumptions put on a linear regression model better than ridge.

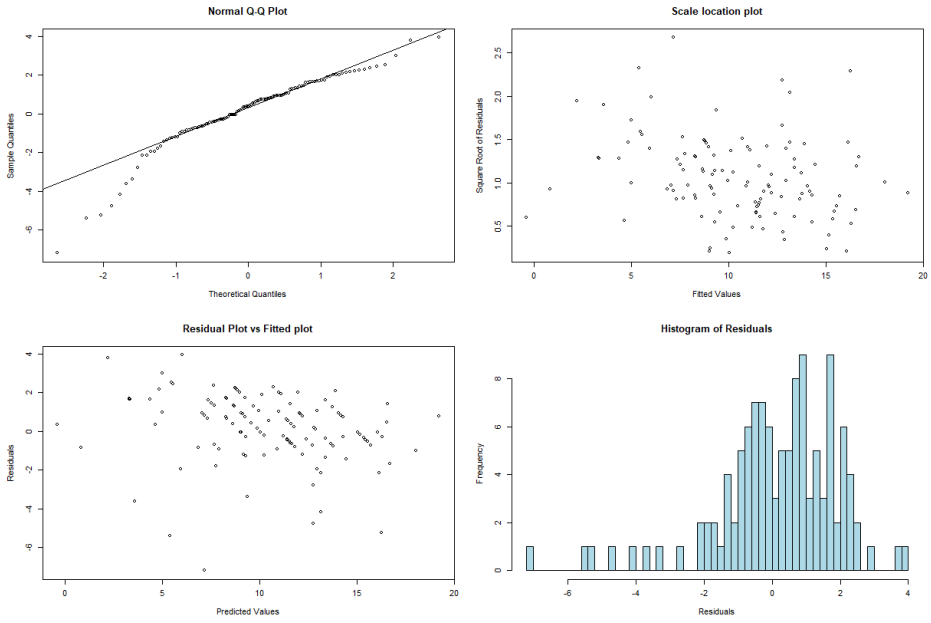

Figure 8: Diagnostic plots for ridge regression

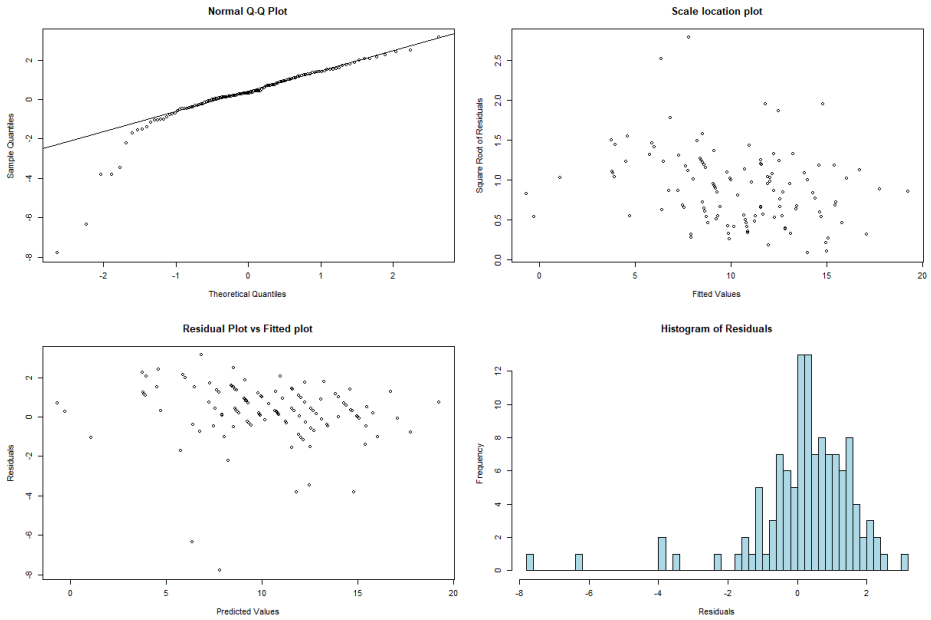

Figure 9: Diagnostic plots for lasso regression

Finally, to see if the differences between the model's predictions are significant, a paired t-test was performed for the penalized regression models, and a McNemars test was performed for the logistic classification models. The paired t-test had the null hypothesis that there was no difference between the model residuals. The test was performed at the 5 % significance level. Seen in table 7 the test could not reject the null hypothesis. Seen in table 8 the McNemars test was also performed at the 5 % significance level, but with the null hypothesis that the misclassification rate was equal for both models. Again, we could not reject the null hypothesis. Thus, since both tests were insignificant, the difference in results could be due to randomness rather than actual performance differences between the models.

Table 7: Paired t-test for penalized regression models

estimate	statistic	p.value	parameter	conf.low	conf.high	method	alternative
-0.1	-1.3	0.1819	118.0	-0.2	0.0	Paired t-test	two.sided

Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05

Table 8: Mcnemar's test on the misclassification rate for the penalized logistic models

statistic	p.value	parameter	method
0.4	0.5465	1.0	McNemar's Chi-squared test with continuity correction

Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05

6 Conclusion

When making this report, the question sought to answer was if there is a difference in predicting performance between ridge and lasso regularization, and if so, which one is preferable. Of course, this report alone will not be able to answer these questions in the general setting, but rather for this specific data set and perhaps similar data sets as well. In terms of results, our findings indicated a notable advantage for the lasso models. We concluded that lasso was better since the data seemed to be made up of a lot of unnecessary independent variables that made the model too complex. This problem is solvable for lasso as it can practice variable selection while ridge is stuck with all the worst predictors. When looking at diagnostic plots both models seemed to struggle with some non-normal residuals, where, however, lasso again looked better than ridge.

Despite all this, the paired t-test, as well as the McNemar‘s test were both insignificant, meaning we could not reject that there is no difference between the model's performances. Thus, even though lasso outperformed ridge across the board in terms of results, we must interpret these results with caution. We cannot be sure that the difference was not caused by random variation, instead of actual difference in model quality. Therefore the answer to our research question becomes that there is no difference between ridge and lasso when predicting student grades.

Some clear limitations of the study must be highlighted. First of all, the results of this report should not be generalized to other data sets or even regularization models without careful consideration. What we have tested is to specifically predict student grades, and cannot extrapolate the results towards other prediction tasks.

To enhance the robustness of future research, it is recommended to explore other data sets. It could also be interesting to work with more variable selection to improve the performance of the models, and particularly boost the performance of the ridge model. Further, a more in depth analysis of coefficients could broaden the understanding.

References

P. Cortez. Student performance. http://archive.ics.uci.edu/dataset/320/student+performance?fbclid=IwAR2pAaavyYUAKKvzjLKVyxTSV5ylZwDE1Fp7wqktSlRGxzhOqXyFe9emQA0, 2014.

P. Cortez and A. M. G. Silva. Using data mining to predict secondary school student performance. 2008.

T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

K. Hajian-Tilaki. Receiver operating characteristic (roc) curve analysis for medical diagnostic test evaluation. *Caspian Journal of Internal Medicine*, 4(2):627–635, Spring 2013. PMID: 24009950; PMCID: PMC3755824.

E. A. Hanushek and D. D. Kimko. Schooling, labor-force quality, and the growth of nations. *American economic review*, 90(5):1184–1208, 2000.

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

N. Herawati, K. Nisa, E. Setiawan, N. Nusyirwan, and T. Tiryono. Regularized multiple regression methods to deal with severe multicollinearity. *International Journal of Statistics and Applications*, 8(4):167–172, 2018.

OECD. *Low-Performing Students*. 2016. doi: https://doi.org/https://doi.org/10.1787/9789264250246-en. URL https://www.oecd-ilibrary.org/content/publication/9789264250246-en.

R. M. O'brien. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41:673–690, 2007.

J. M. Pereira, M. Basto, and A. F. Da Silva. The logistic lasso and ridge regression in predicting corporate failure. *Procedia Economics and Finance*, 39:634–641, 2016.

G. Psacharopoulos. Returns to investment in education: A global update. *World development*, 22(9):1325–1343, 1994.

S. Raschka. Model evaluation, model selection, and algorithm selection in machine learning. *arXiv preprint arXiv:1811.12808*, 2018.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

T.-T. Wong and P.-Y. Yeh. Reliable accuracy estimates from k-fold cross validation. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1586–1594, 2019.