STAT 5630 Project Milestone 2

Lei Zhang, Yunsheng Lu, Yuxuan Wang

April 2023

1. EDA

In the original dataset, three response variables, "G1", "G2", and "G3", record the students' grades in Mathematics and Portuguese for each term. The correlation matrices show that the response variables are highly correlated. Since some students kept receiving 0 in the second and third exams, we assume they have completely given up, and their scores are hard to predict by the statistical model, so we only consider "G1" as the response variable for our project.

We checked the independence between predictors. Specifically, we performed Chi-square test when comparing two categorical variables, two-sample t-test if one of the variables is continuous, and F-test of linear regression if both are continuous variables. The red blocks in figure 1 imply the p-value for independence test is smaller than 0.05, thus indicating correlation. Figure 1 suggests we perform variable selection or shrinkage in the future due to collinearity.

2. Classification

To make the original problem a classification task, we divide the continuous response variable "grade" into 5 categories, A (16-20), B (14-15), C (12-13), D (10-11), and F (0-9).

The baseline estimation is created by randomly generating the class label according to the proportion of different classes in the training set.

We implemented several linear methods on the data, which include support vector machine (SVM), linear discriminant analysis (LDA), logistic regression (LR), and logistic regression with Lasso or Ridge penalty. The corresponding training and test error are shown in table 1. For LR with Lasso and Ridge penalty, we choose the tuning parameter by 10-fold cross-validation. The corresponding plots are shown in figure 2. For SVM, we select the tuning parameter "cost" by 10-fold cross-validation and the plot of validation error versus cost is shown in figure 3.

The result suggests LR+Lasso and LDA yield the best performance for math data and SVM for Português data. However, overall the misclassification rate is roughly high.

3. Regression

The baseline method is estimating the value by the mean grade of training set.

We again implemented several linear methods, including supported vector machine(SVM), ordinary least squares (OLS), linear regression with Lasso and Ridge penalty. For Ridge and Lasso regression, the parameter is selected by 10-fold cross-validation, shown in figure 4. For SVM, the parameter "cost" is selected by 10-fold cross-validation as well, as shown in figure 5. The corresponding mean squared error of training and test set is shown in table 2.

The table indicates that Lasso yields the best performance for math, and SVM for Português.

4. Discussion

We expect non-linear models like tree-based ones may yield more accurate estimation.

By comparing the results of the classification task and regression task, we suppose the data may not be suitable to be divided into 5 categories as the sample size is not large enough.

Due to space limitations, we are not able to include inference on parameters.

Appendix

A. Tables

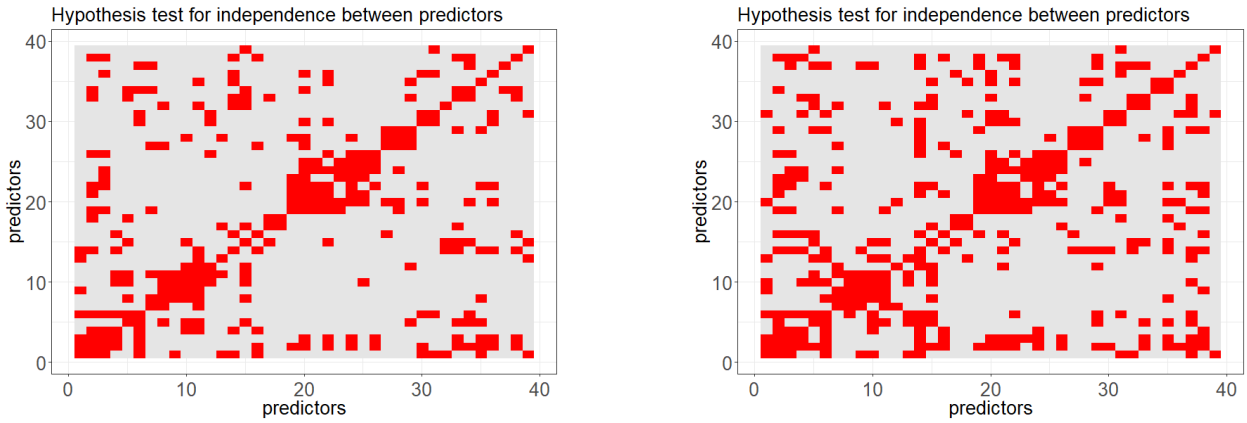
	math		Português	
	training error	test error	training error	test error
baseline	0.768	0.835	0.802	0.738
SVM	0.642	0.633	0.489	0.661
Logistic regression	0.446	0.633	0.462	0.715
LR + Lasso	0.579	0.620	0.493	0.700
LR + Ridge	0.582	0.696	0.489	0.692
LDA	0.487	0.620	0.484	0.692

Table 1: Performance of different methods for classification task

	math		Português	
	training MSE	test MSE	training MSE	test MSE
baseline	0.997	0.966	0.998	0.988
SVM	0.676	0.844	0.629	0.763
OLS	0.886	0.969	0.601	0.800
Lasso	0.678	0.782	0.641	0.774
Ridge	0.671	0.803	0.617	0.772

Table 2: Performance of different methods for regression task

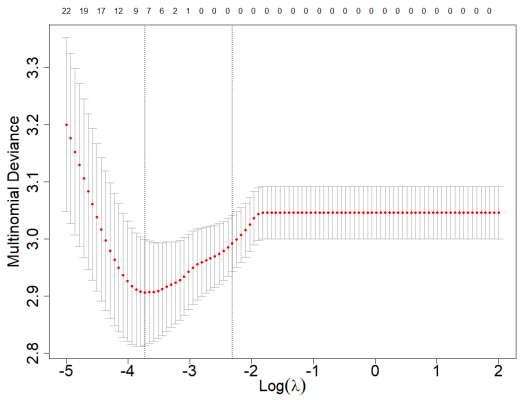
B. Figures



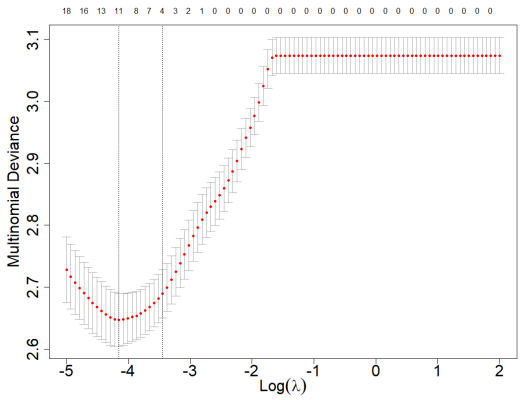
(a) Independence test for math data

(b) Independence test for Português data

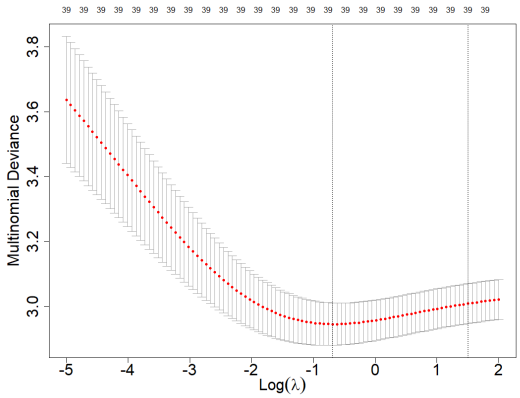
Figure 1: Result of Independence Test.
The red dots indicate the corresponding two variables are not independent. The grey dots indicate they are independent.



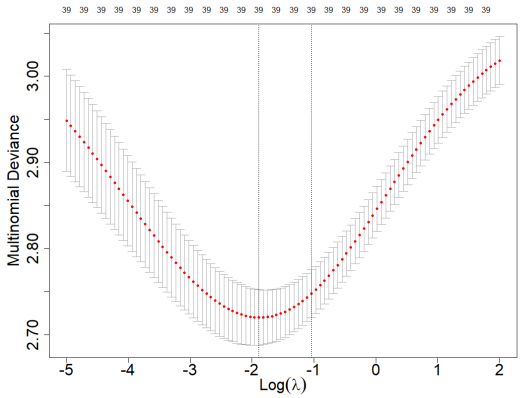
(a) LR + Lasso for math data



(b) LR + Lasso for Português data

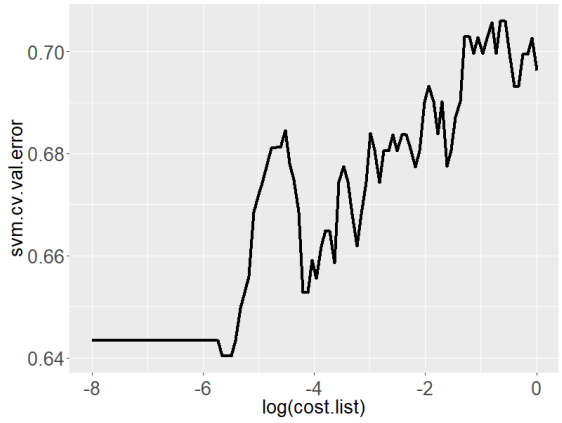


(c) LR + Ridge for math data

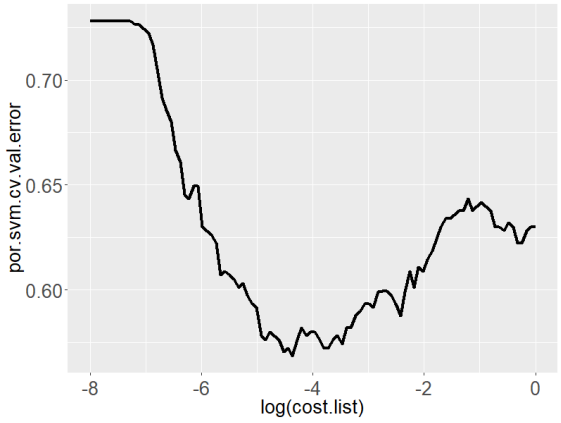


(d) LR + Ridge for Português data

Figure 2: The result of cross validation on tuning parameter λ for different penalty and data

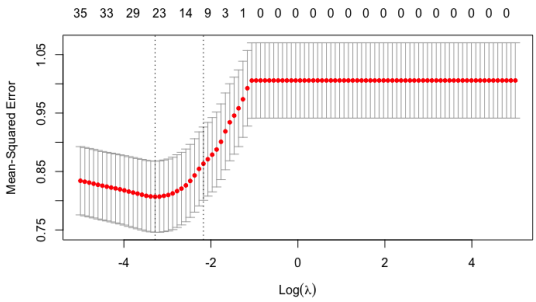


(a) Classification SVM for math data

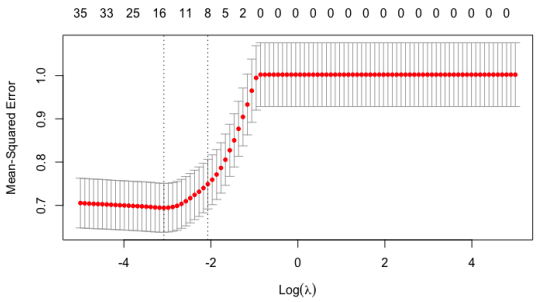


(b) Classification SVM for Português data

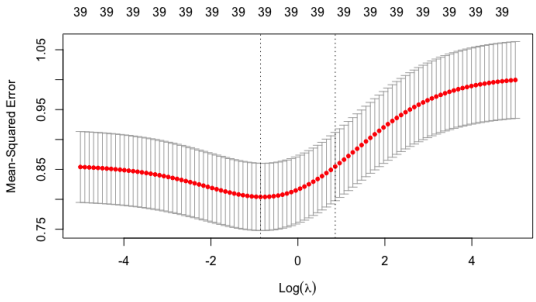
Figure 3: Cross Validation Result of Classification SVM.
The x-axis are log of tuning parameter "cost" and the y-axis is the mean of cross validation classification error.



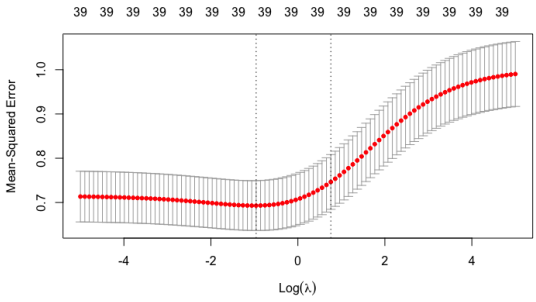
(a) Lasso regression for math data



(b) Lasso regression for Português data

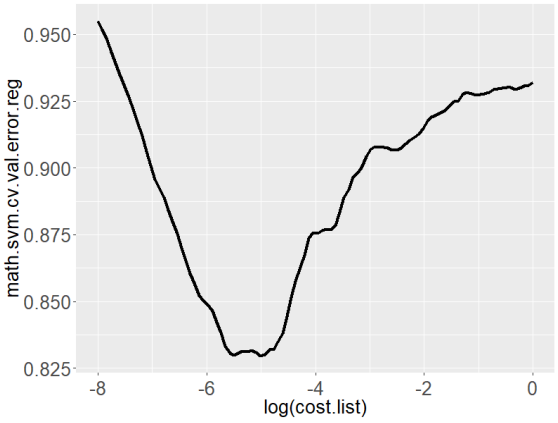


(c) Ridge regression for math data

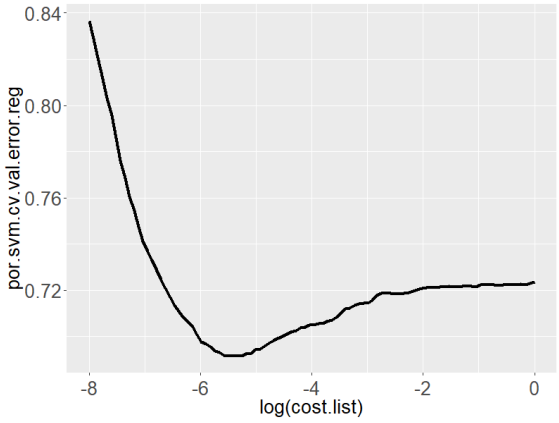


(d) Ridge regression for Português data

Figure 4: The result of cross validation on tuning parameter λ for different penalty and data



(a) Regression SVM for math data



(b) Regression SVM for Português data

Figure 5: Cross Validation Result of Regression SVM.
The x-axis are log of tuning parameter "cost" and the y-axis is the mean of cross validation mean squared error.