Predictive Modeling Processes on Credit Data

Alexander Lee and Youngshin Kim

November 4, 2016

Abstract

This project functions as an introduction to ridge regression, lasso regression, principal components regression, and partial least squares regression, particularly in R. We use R to run each of these regressions and compare the coefficients determined by these regressions to those generated by the original least squares regression.

Introduction

This project aims to predict data using differnet methods. Methods used are ols, ridge, lasso, principal component regression, and partial least sqaures regression. We want to be able to predict variable 'Balance' from ten predictors. Before we begin, we will first make some necessary changes to our dataset using data manipulation in R.

Data

The data we use in this project is the credit data and it includes variable Balance, the variable we are trying to predict, and ten predictors that we use to predict Balance. These predictors are Income, Limit, Rating, Cards, Age, Education, GenderFemale, StudentYes, MarriedYes, EthnicityAsian, EthnicityCaucasian. The last two predictors were originally one predictor, but it was expanded as a result of dummifying the values.

Methods

We use five regression methods in this project: OLS, ridge, lasso, pcr, plsr and check the method that gives us the lowest MSE. OLS is the most simple regression method that tries to fit a line that minimizes the residual sum of squares. The rest are all regularization method. Ridge and lasso regressions are shrinkage methods that put a penalty to our linear model and we try to find the best model with the smallest lambda. Lasso is different from ridge in that lasso does variable selection. It cuts out predictors that are not useful to predicting Balane. PCR and PLSR regressions perform dimension reduction, and are useful when the predictors are correlated to each other.

Analysis

For OLS, we simply fit the lm function in R. For the rest of the methods, we first use 10 fold cross-validation to fit method on train set and find the best model based on minimum lambda for ridge/lasso and minimum prediction error sum of squares for pcr and plsr. Next, we apply the best model to test set and calculate MSE. Finally, we choose the regression method that gives us the lowest MSE.

	OLS	Ridge	Lasso	PCR	PLSR
Income	-7.80	-0.57	-0.55	-0.60	-0.60
Limit	0.19	0.71	0.85	1.03	1.02
Rating	1.14	0.60	0.43	0.30	0.31
Cards	17.72	0.05	0.05	0.06	0.06
Age	-0.61	-0.02	-0.01	-0.02	-0.01
Education	-1.10	-0.01	0.00	-0.01	-0.01
GenderFemale	-10.65	-0.00	0.00	-0.00	-0.00
StudentYes	425.75	0.27	0.27	0.28	0.28
MarriedYes	-8.53	-0.01	0.00	-0.01	-0.01
EthnicityAsian	16.80	0.02	0.00	0.01	0.01
EthnicityCaucasian	10.11	0.03	0.01	0.03	0.03

Table 1: Regression Coefficients for All Methods

Results

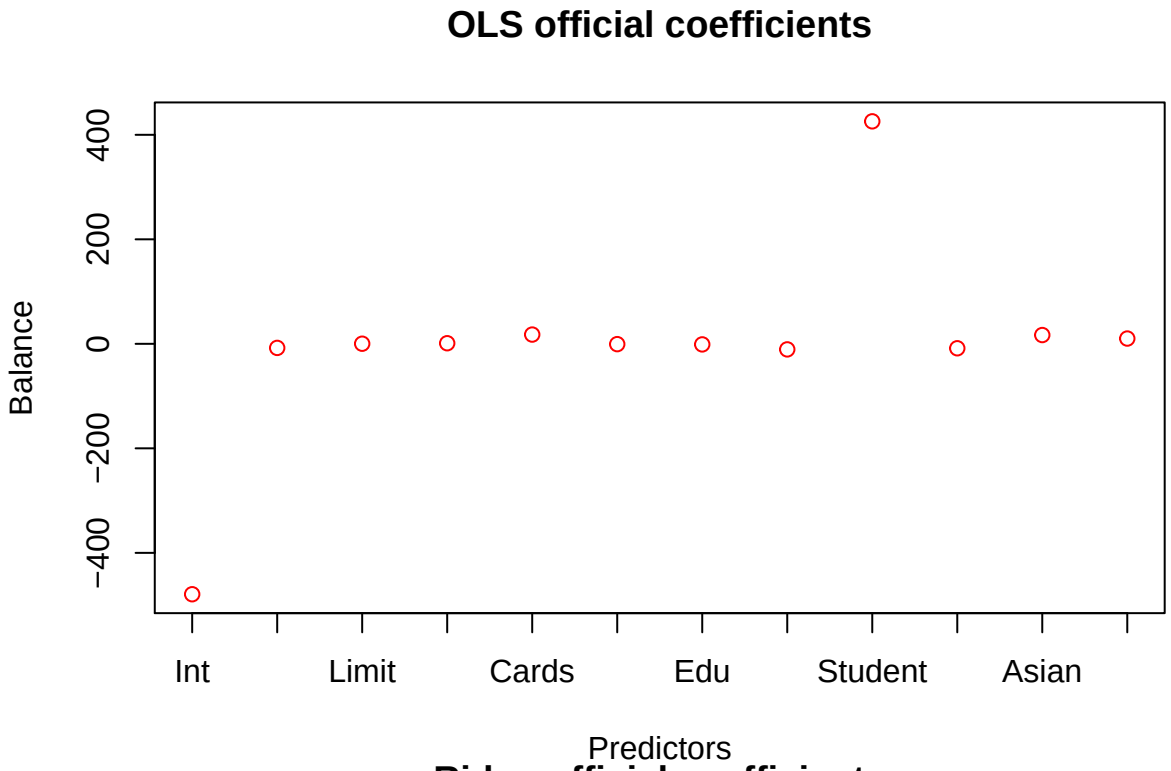
The table above displays the regression coefficients of the best model for each method. There are some things that stand out. First, the OLS coefficients seem very different from the others. This was expected, because OLS is the weakest method in this project. Second, ridge/lasso and pcr/plsr seem to share similar regression coefficients values. This is also expected, because each pair is similar to each other.

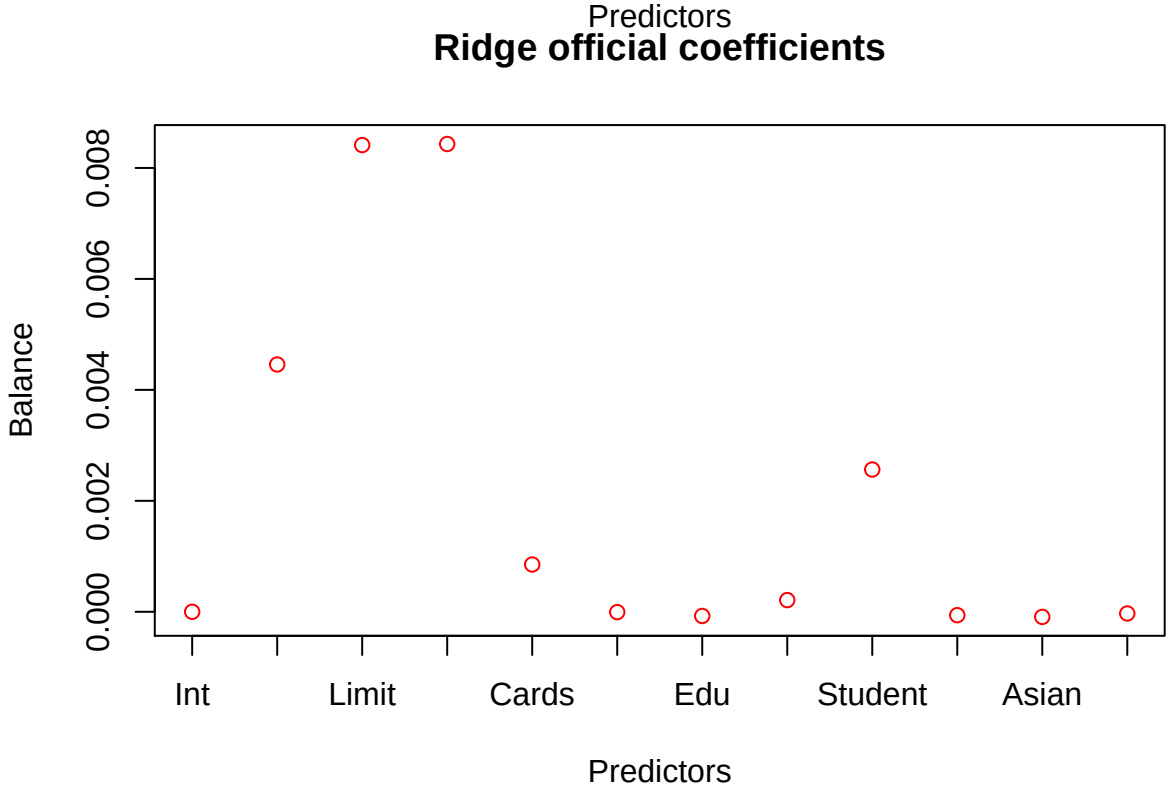
	Ridge	Lasso	PCR	PLSR
MSE	0.97	1.00	0.05	0.05

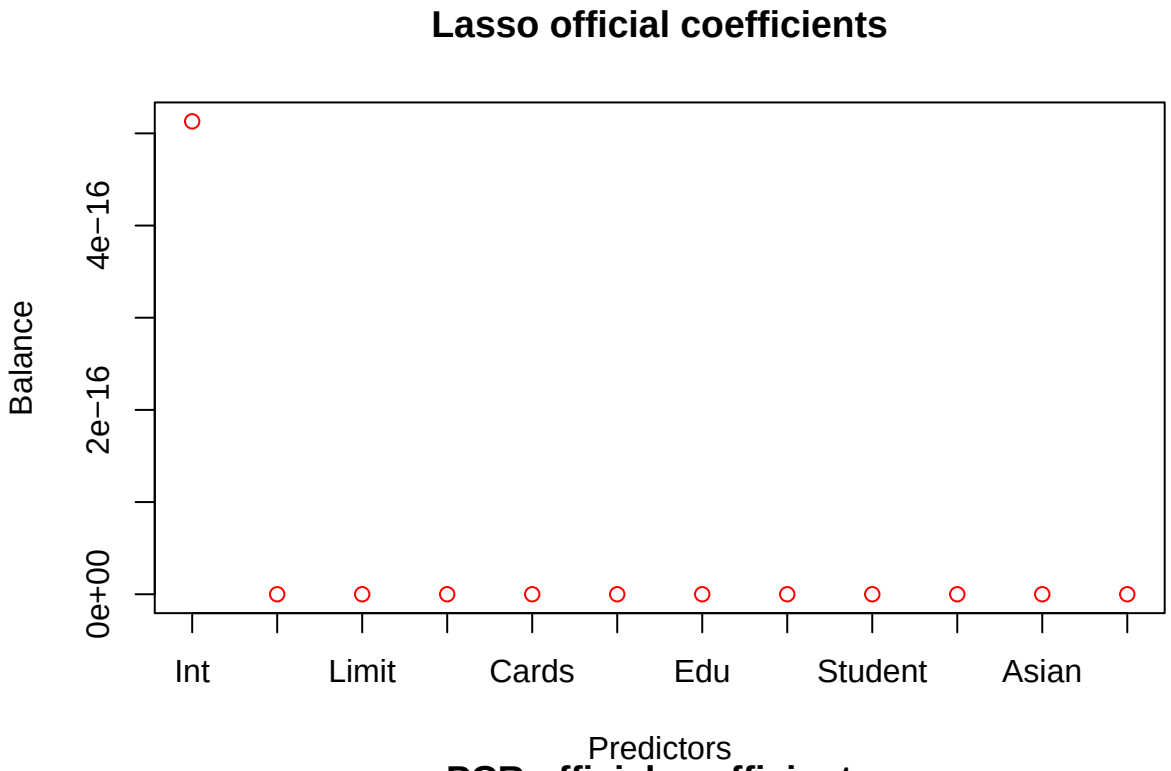
Table 2: MSEs for All Methods

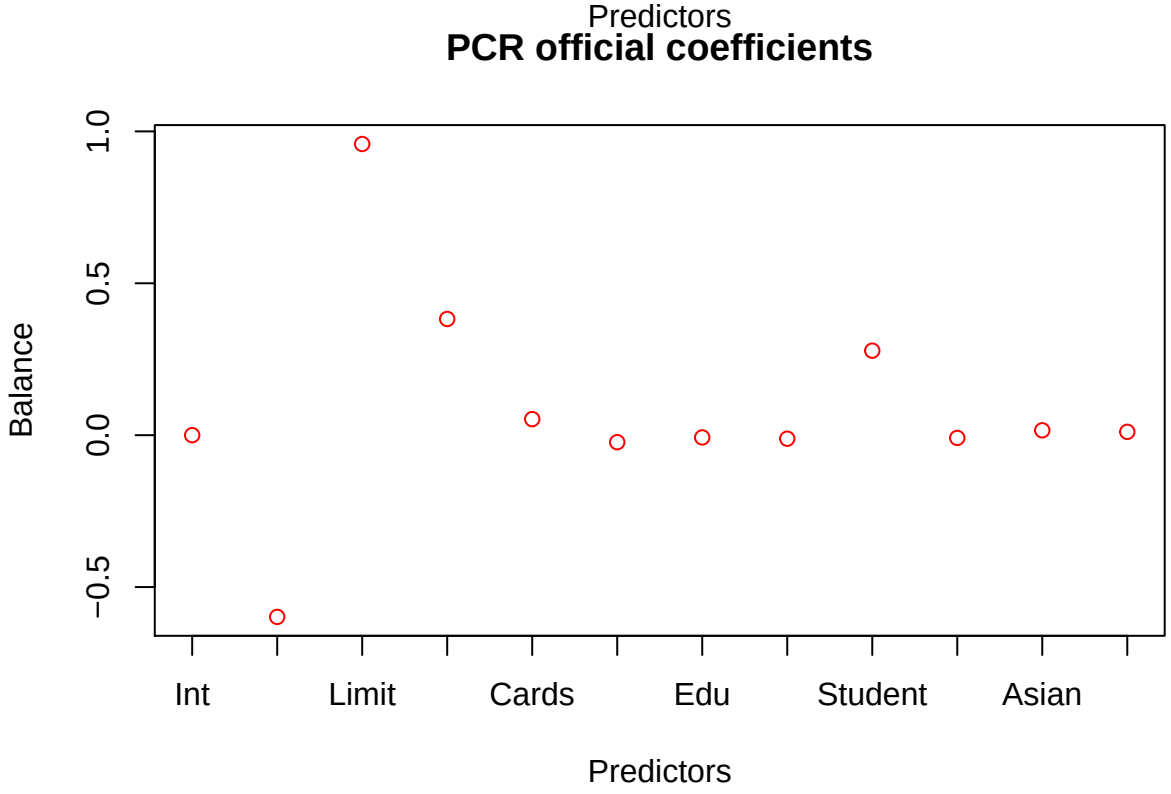
The table above shows the mean squared errors for each method. These values were calculated by applying the best model to the test set and calculating the mean of squared difference between estimated value and actual value (Balance). It is clear that PCR and PLSR have low MSEs compared to the other methods.

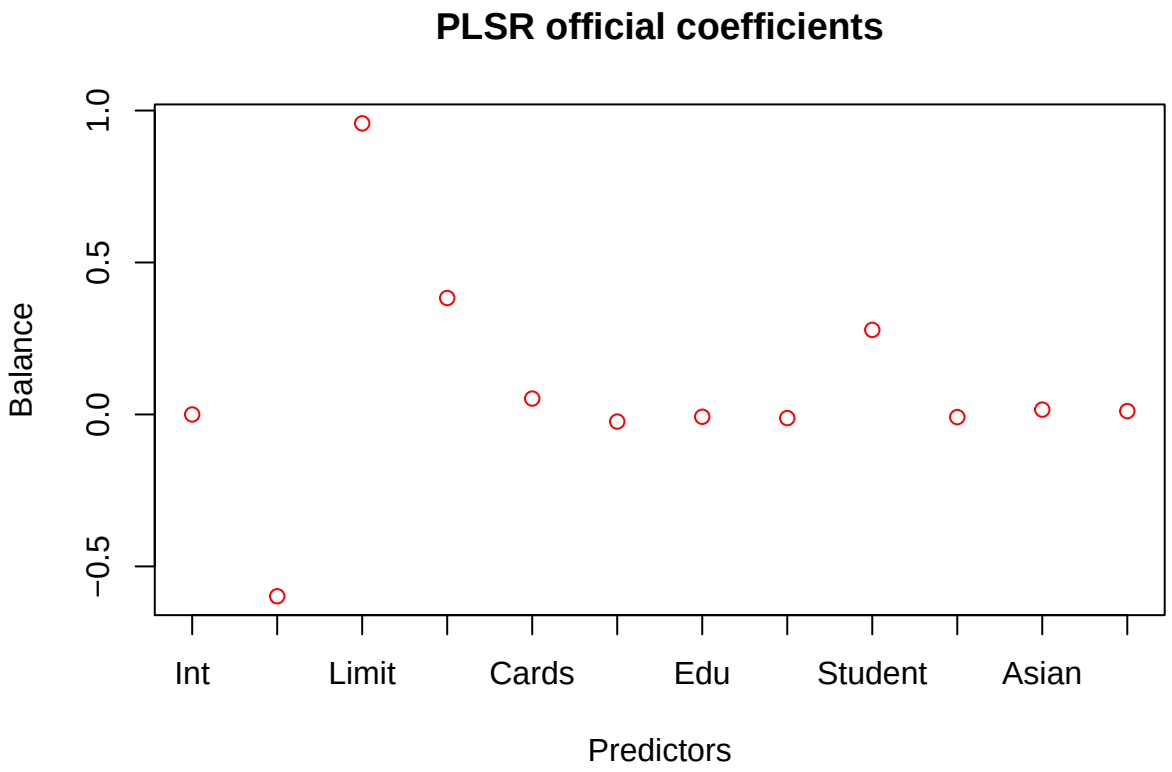
```
Loading required package: Matrix
```


OLS official coefficients
Balance
400
200
0
−200
−400
Int
Limit
Cards
Edu
Student
Asian
Predictors


Ridge official coefficients
Balance
0.008
0.006
0.004
0.002
0.000
Int
Limit
Cards
Edu
Student
Asian
Predictors


Lasso official coefficients
Balance
0e+00
2e−16
4e−16
Int
Limit
Cards
Edu
Student
Asian
Predictors


PCR official coefficients
Balance
−0.5
0.0
0.5
1.0
Int
Limit
Cards
Edu
Student
Asian
Predictors



These are the 'official' regression coefficients that were calculated by applying the best model to the full data set.

Conclusion

From the tables and figures above, we can conclude that PCR/PLSR give us the best estimate of Balance. We can make this conclusion, because PCR/PLSR gave us the lowest MSE (0.05), which means the predicted values and the actual values weren't that much different from each other. Also, from the figures above, we can assume that there were only a couple predictors that were actually important in estimating Balance. The most noticeable predictors are Limit and Student. It's clear in the figures that the coefficients are high for these predictors.