00-abstract

Abstract

This report aims to find a predictive modelling process for the Credit balance. The data analysis focuses on ten different factors that could influence credit card balance. We use five different models of regression such as exploratory data analysis, pre-modeling data processing and multiple linear regression models in order to make a decision on which of the methods is the best predictor for the balance in terms of ten predictors such as Income, Age, Education, Gender and others.

Introduction

In this project we focus on the five regressions learned such as Least-Squares regression, Ridge regression, Lasso regression, Principal Component regression, and Partial Least-Squares regression. We approach the project by considering these types of regression, so that we can compare them for the simple purpose of improving the accuracy of the predicion of the linear model in this case. By using other models this will make the coefficient of the other variables equal to zero and during our analysis, the whole process will be more efficient because there will be less variables to interpret and easier visually.

Another reason to use different kind of methods is that not every variable is connected with the response variable. To avoid using least-square method, we use other models that will set the coeff. equal to zero and making our analysis easier to interpret. Other methods include ridge, lasso, partial least square and principal component. They would all go towards the same value but the way to reach that would be different depending on method.

Before running any kind of regression in the dataset, we first focus on the exploratory data analysis by observing and reproducing mediums of visualization like statistical diagrams, histograms, graphs. This could affect on our ability to evaluate each regression's performance.

The reproducibility and collaborativeness included in the project is vital not only to an eloquent flow of our progress but also on the results we get at the end. Just like we were able to analyse, produce and observe patterns in our project, other users may find this helpful and contribute to the advancement of project in the future. This comes as a result of collaborative efforts from both parts.

The variables used in this dataset are qualitative {Gender, Student, Married, and Ethnicity} and quantative {Income, Limit, Rating, Cards, Age, Education, and Balance}. We are using data that originated from Credit.csv. The quantitative variables in this are labeled as age, cards, education, income, limit, rating, and balance. These variables are one's age, the no. of credit cards one has, years of education, income in dollars, credit limit, credit rating, and one's average credit card debt. Balance is the response variable.

One important part of our job was to standardize all of the data for a more comprehensibile and easier approach. For our analysis, based on the Credit.csv, we created scaled-credit.csv, which has the same variables as the original file. However, not every content is the same since we can still make changes to scaled-credit.csv, especially depending to the scale variables are measured in this case.

Methods

Ordinary Least Squares Regression

Apply multiple linear regression by using the **lm()** function to find the relationship between Balnace and the 11 predictors of Income, Limit, Rating and more.

$$Result = \beta_0 + \beta_1 * Income + \beta_2 * Limit + \beta_3 * Rating + \cdots + \beta_{10} * EthnicityCaucasian$$

Shrinkage Methods

Ridge regression

Ridge regression is similar to OLS with the coefficients estimated by minimizing a slightly different quantity.By minimzing RSS, we can find the coefficient estimates that fit the data well.

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

However, the shrinkage penalty is that $\lambda\sum_{j=1}^{p}\beta_j^2$ is small when β are close to zero. Ridge regression will produce a different set of coefficient estimates for each value of λ .

Lasso regression

Lasso also minimize the quantity by the following:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$$

Lasso also shrink the coefficient estimates to zero. However, the penalty has the effect of focing some of the coefficient estimates to equal to zero when the tuning pararmeter is too large. Thus, lasso is better on feature selection.

Dimension Reduction Methods

Principal Components regression (PCR)

In principal compoenents regression, the method is construct $Z_1 + \cdots + Z_{M*}$ and then use these components as the predictors in a linear regression model. With the small amount of principal components, it can explain the variability of the data. PCR performs well when the first few principal components have enough information on the variation in the predictors and relationship with the response. The response does not supervise the principal components.

Partial Least Squares regression (PLSR)

Partial Least Square is a supervised way of PCR. PLS is a dimension reduction method and fits a linear model through least square using $Z_1 + \cdots + Z_{M*}$. PLS also make use of response Y to identify new features. PLS will try to find ways to explain the trend and pattern of reponse and predictor variables.

Credit to An Introduction to Statistical Learning

Analysis

```r
library(xtable)
```

```
## Warning: package 'xtable' was built under R version 3.2.3
```

```r
library(readr)
library(Matrix)
options(xtable.floating = FALSE)
load('../data/ols.RData')
load('../data/ridge.RData')
load('../data/lasso.RData')
load('../data/pcr.RData')
load('../data/plsr.RData')
```

OLS

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.00	0.01	0.00	1.00
scaledCreditAIncome	-0.60	0.02	-33.31	0.00
scaledCreditALimit	0.96	0.16	5.82	0.00
scaledCreditARating	0.38	0.17	2.32	0.02
scaledCreditACards	0.05	0.01	4.08	0.00
scaledCreditAAge	-0.02	0.01	-2.09	0.04
scaledCreditAEducation	-0.01	0.01	-0.69	0.49
scaledCreditAGenderFemale	-0.01	0.01	-1.07	0.28
scaledCreditAStudentYes	0.28	0.01	25.46	0.00
scaledCreditAMarriedYes	-0.01	0.01	-0.82	0.41
scaledCreditAEthnicityAsian	0.02	0.01	1.19	0.23
scaledCreditAEthnicityCaucasian	0.01	0.01	0.83	0.41

[1] 0.04478619

The ordinary least squares regression MSE is 0.04478619.

Ridge

	1
(Intercept)	0.00
Income	-0.57
Limit	0.72
Rating	0.59
Cards	0.04
Age	-0.03
Education	-0.01
GenderFemale	-0.01
StudentYes	0.27
MarriedYes	-0.01
EthnicityAsian	0.02
EthnicityCaucasian	0.01

[1] 0.0525927

The ridge regression MSE is 0.0525927. When comparing the ridge regression with OLS, the results are similar.

Lasso

	1
(Intercept)	0.00
Income	-0.55
Limit	0.93
Rating	0.37
Cards	0.04
Age	-0.02
Education	0.00
GenderFemale	0.00
StudentYes	0.27
MarriedYes	0.00
EthnicityAsian	0.00
EthnicityCaucasian	0.00

```
[1] 0.05154446
```

The lasso regression MSE is 0.05154446 The lasso regression and ridge regression have the same results.

Principal Components Regression

	x
1	-0.60
2	0.96
3	0.38
4	0.05
5	-0.02
6	-0.01
7	-0.01
8	0.28
9	-0.01
10	0.02
11	0.01

```
[1] 0.0517916
```

The PCR MSE is 0.05199678. The PCR coefficients and OLS coefficients are similar. They also shared the similar results with ridge regression.

Partial Least Squares Regression

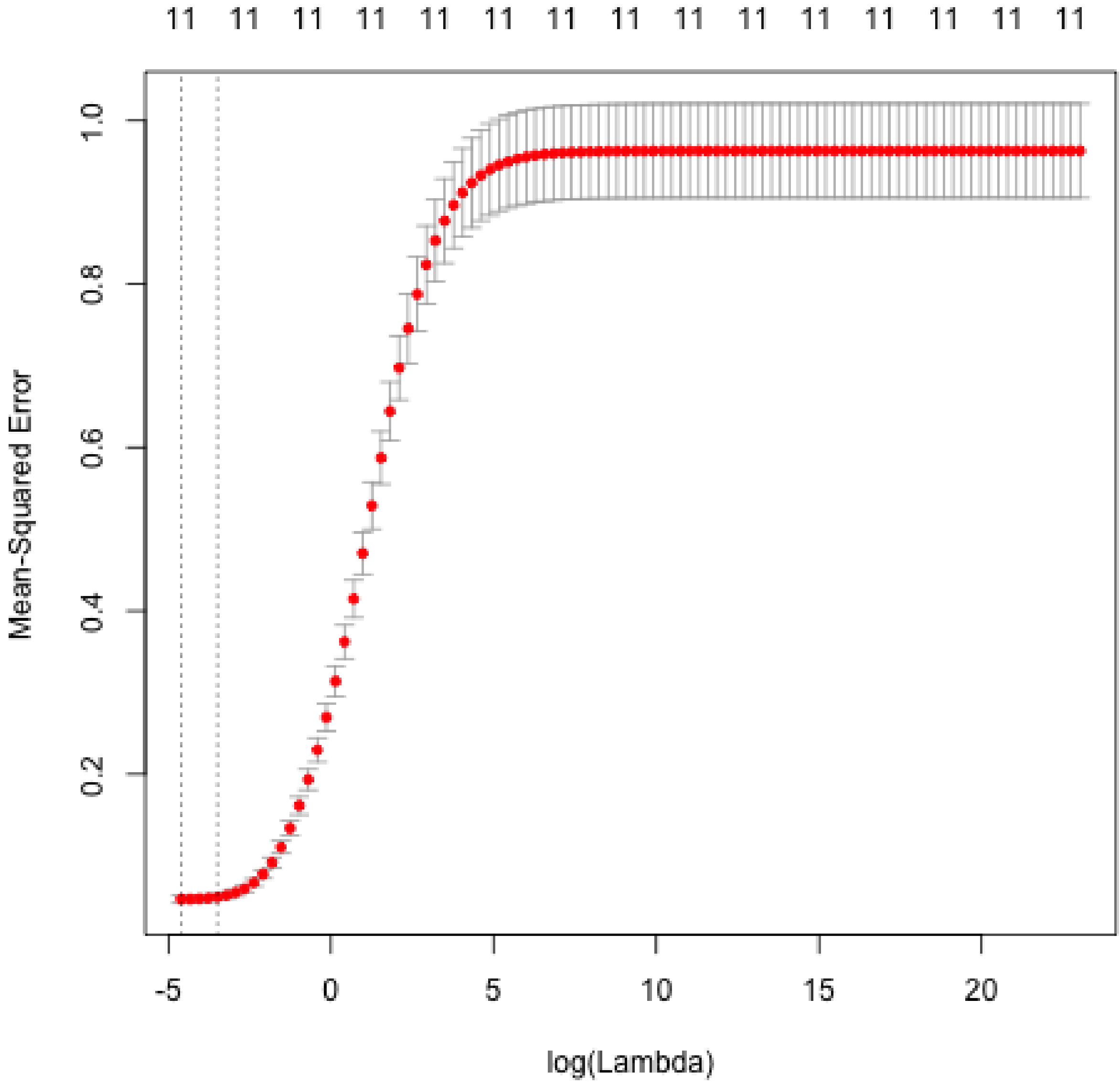
	x
1	-0.60
2	0.96
3	0.38
4	0.05
5	-0.02
6	-0.01
7	-0.01
8	0.28
9	-0.01
10	0.02
11	0.01

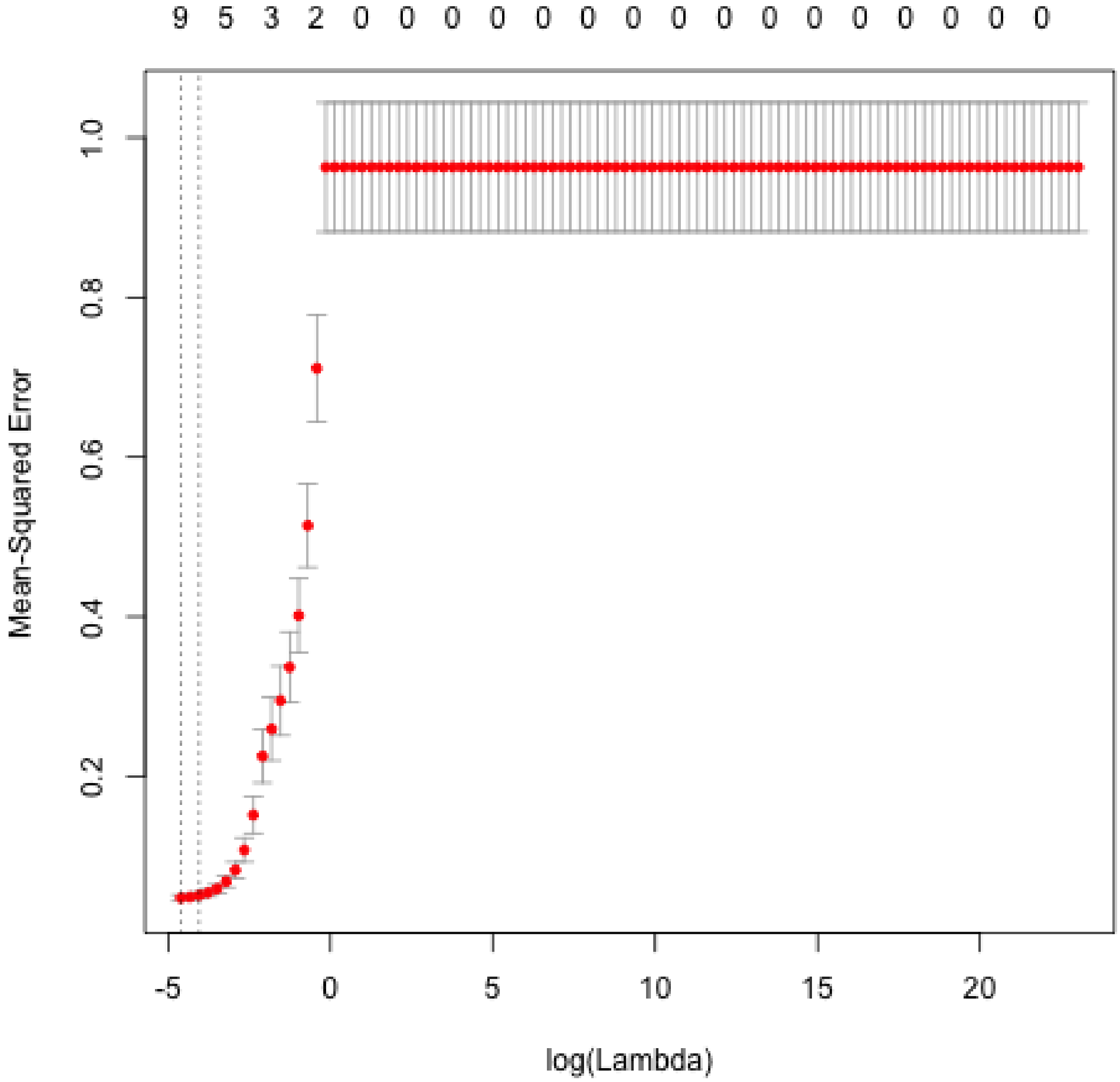
```
[1] 0.0517916
```

The PLSR MSE is 0.0517916. The PLSR coefficients have some difference from PCR. When comparing with second and third components, PCR and PLSR have different results. The majority others remains the same.

From the analysis, we can see that the MSE are similar to each other. The OLS has the smallest MSE, which is the best fit model for the credit data set. The ridge regression has the largest MSE among the five regression, which is the least efficient model. From the tables above, most of the coefficients are similar except for PLSR.

In the Lasso Regression, the Education, Gender female, Ethnicity Asian and Ethnicity Caucasian, Married Yes are 0 for the coefficient. The regression is trying to minimize the predictors.





When comparing the cross validation on both ridge and lasso, lasso cross validation is less consistent than ridge cross validation.

Conclusions

This project uses five regression model to determine the best fit model for the Credit dataset. Based on the result, the OLS perform the best among all the models. The second best is the lasso regression. However, the MSE among all models are very similar and they shared similar coefficients. That means none of the models is far off from the rest.

Credit

An Introduction to Statistical Learning: With Applications in R. New York: Springer, 2013. Print.