Stats 159 Project 2 Report

Jie Sun, Stephen (Mingtao) Fang

Nov 4th, 2016

Abstract

In this project, we are trying to collaboratively perform a predictive modeling process with cross-validation on the Credit data set using five regression models, in a reproducible workflow. The motivation of the analysis is to find out how the variables (age, education, income etc.) affect the balance of a credit card user. With different regression model's analysis, we are hoping to compare our regression models with a benchmark regression model to find out the best regression model for this specific data set.

Introduction

This project is largely based on the chapter 5 and chapter 6 of the book Introduction to Statistical Learning by James El. As a case study, the book has introduced a credit data set for us to apply five different regression models. In the later sections, we are going to cover in details regarding the data set, the methods and the specific analysis.i

Data

The Credit data set records balance (average credit card debt for a number of individuals) as well as several quantitative predictors: age, cards (number of credit cards), education (years of education), income (in thousands of dollars), limit (credit limit), and rating (credit rating). This data can be downloaded through this link.

Methods

In this project, we explore two techniques - shrinkage and dimension reduction. The multiple linear regression model with ordinary least squares is used as a benchmark. We compare its results with other linear models that replace the least squares fitting with shrinkage or dimension reduction.

Shrinkage Methods

Shrinkage methods involve fitting a model using all predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance. Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform variable selection.

Ridge Regression

The ridge regression coefficient estimates are the values that minimize:

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

Ridge regression's advantage over least squares is rooted in the bias-variance trade-off. As λ (a tuning parameter) increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

But the ridge regression does have one obvious disadvantage. It will include all p predictors in the final model. Although l_2 penalty will shrink all of the coefficients towards zero, but it will not set any of them exactly to zero (unless $\lambda = \infty$). So it will not perform variable selection. This may not be a problem for prediction accuracy, but it can create a challenge in model interpretation in settings in which the number of variables p is quite large.

Lasso

The lasso coefficients minimize the quality:

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

Lasso overcomes the disadvantage of ridge regression by using l_1 penalty, which has the effect of forcing some of the coefficient estimates to be exactly equal to zero when λ is sufficiently large.

Dimension Reduction

Dimension reduction involves projecting the p predictors into a M-dimensional subspace, where $M < p$. This is achieved by computing M different linear combinations, or projections, of the variables. Then these M projections are used as predictors to fit a linear regression model by least squares. In this way, a low-dimensional set of features can be derived from a large set of variables.

Principal Components Regression

Using the principal components analysis (PCA), this regression approach involves constructing the first M principal components, $Z_1, ..., Z_M$, and then using these components as the predictors in a linear regression model that is fit using least squares. The key idea is that often a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.

The PCR approach focuses on identifying linear combinations, or directions, that best represent the predictors $X_1, ..., X_p$. That is, the response does not supervise the identification of the principal components. Consequently, PCR suffers from a drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.

Partial Least Squares Regression

The partial least squares (PLS) is a supervised alternative to PCR. Like PCR, PLS first identifies a new set of features $Z_1, ..., Z_M$ that are linear combinations of the original features, and then fits a linear model via least squares using these M new features. But unlike PCR, PLS identifies these new features in a supervised way — that is, it makes use of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response. Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

Analysis

Exploratory Phase

In order to perform data analysis, it is important for us to understand the data by obtaining descriptive statistics and summaries of the variables of our current data set. Therefore, for the exploratory phase, we are looking at the some quantitative results:

- Minimum, Maximum, Range
- Median, First and Third quartiles, and interquartile range (IQR)
- Mean and Standard Deviation
- Histograms and boxplots
- matrix of correlations among all quantitative variables.
- scatterplot matrix.
- anova's between `Balance` and all the qualitative variables
- conditional boxplots between `Balance` and the qualitative variables, that is, boxplots of `Balance` conditioned to each of `Gender`, `Ethnicity`, `Student`, and `Married`.

These results will help undestanding the data set. The images are all saved in the `images` folder of this project.

Pre-modelling Data Processing

In order to fit the data set with our intended regression models, it is important for us to perform data processing beforehand.

The first step in this data processing part is to dummy out the categorical variables. Since the `ridge` regression and `lasso` regression (implemented by `glmnet()` function) we are going to use in the next section do not support categorical variables, we need to transform these categorical variables into dummy variables. This transformation can be performed by setting binary indicators to expand a factor. For example, with `gender`, which has value of `male` and `female`, we can transform it into binary indicator of 1 or 0.

The second step in this data process is to perform mean centering and standardization because we want to make each variable have comparable scales. While the value of coefficients could be affected because of the measurement scale, we want to avoid favoring a specific coefficient by standardizing the variables, which means that we will transform each variable to have a mean of zero and a standard deviation of one.

Splitting Test Set and Train Set

To get a fair evaulation of how to regression model performs, we need to split the data set into train set and test set so we can avoid overfitting. According to the instruction of the project, we are splitting it into a test set of 100 rows of data and a train set of 300 rows of data. Moreover, to ensure the reproducibility of this project, we are also utilizing the `set.seed()` function to make the results reproducible.

Cross-validation

Through out the later sections of the project, we will be using 10-fold cross-validation to pick our best model. Since cross-validation employs a random sampling through the process, it is also essential to mention that we will utilize `set.seed()` to ensure that our project results are reproducible.

Results

To understand the data, we computed descriptive statistics and summaries of all variables and stored them in `data/eda-output.txt` in EDA phase. Because we are interested in studying the association between `Balance` and the rest of predictors, we also obtained matrix of correlations, scatterplot matrix, anova and conditional boxplots between Balance and the qualitative variables, which can be found in `images` folder.

After fitting all models, we summarize all regression coefficients including OLS in Table 1.

	OLS	Ridge	Lasso	PCR	PLSR
Income	-0.5982	-0.5687	-0.5517	-0.5982	-0.5981
Limit	0.9584	0.7187	0.9250	0.9584	0.9578
Rating	0.3825	0.5931	0.3679	0.3825	0.3831
Cards	0.0529	0.0443	0.0450	0.0529	0.0523
Age	-0.0230	-0.0254	-0.0167	-0.0230	-0.0234
Education	-0.0075	-0.0059	0.0000	-0.0075	-0.0076
GenderFemale	-0.0116	-0.0107	0.0000	-0.0116	-0.0119
StudentYes	0.2782	0.2732	0.2668	0.2782	0.2782
MarriedYes	-0.0091	-0.0110	0.0000	-0.0091	-0.0086
EthnicityAsian	0.0160	0.0164	0.0000	0.0160	0.0159
EthnicityCaucasian	0.0110	0.0110	0.0000	0.0110	0.0111

Table 1: Regression Coefficients for All Models

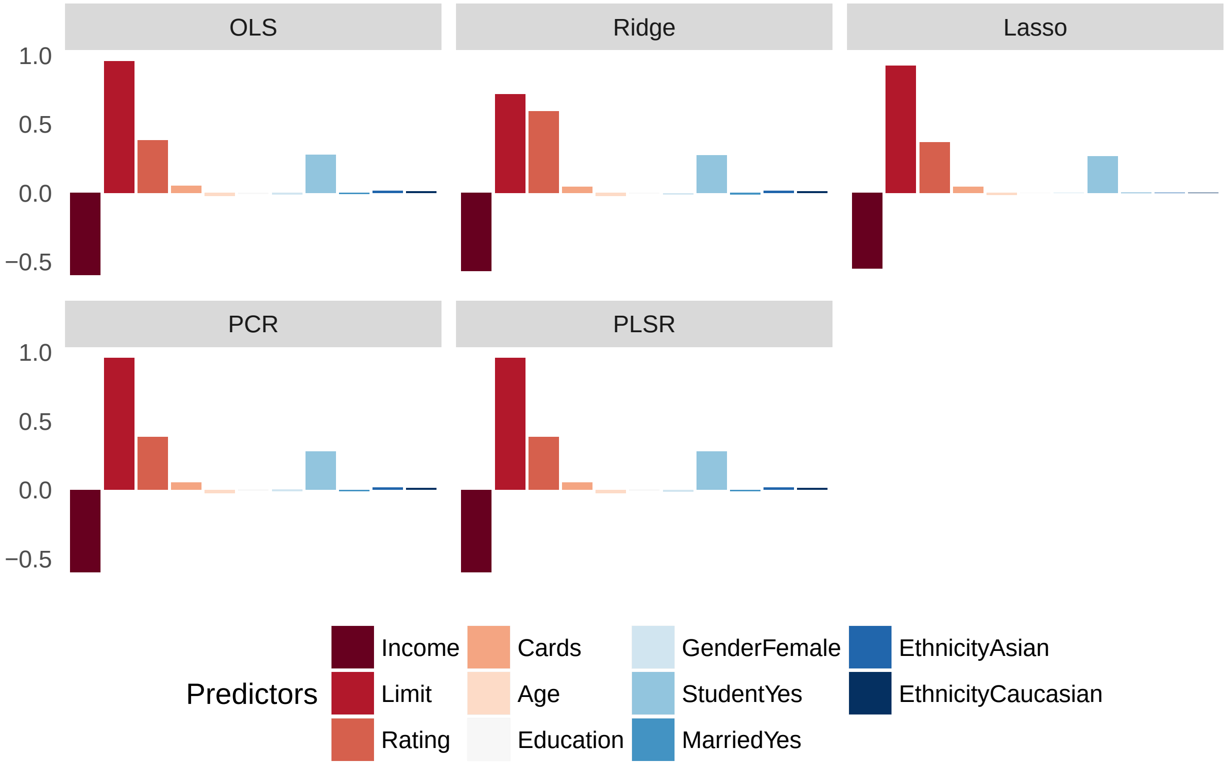
The test MSE value is also a good way to evaluate results of different regression techniques (Table 2).

	Ridge	Lasso	PCR	PLSR	OLS
Test MSE	0.05	0.05	1.59	1.73	1.80

Table 2: Test MSE Values for All Models

We can spot both difference and similarty in Chart 1 where all official coefficients are compared.

Chart 1. Barchart of Official Coefficients in Different Models



Conclusions

From Table 1, we can tell that coefficients for `Age`, `Education`, `Gender`, `Married` and `Ethnicity` are generally smaller than others, which means they're less influential on the balance of a credit card user. `Income`, `Limit`, `Rating` and `Student` are better predictors in this case. And `Income` negatively associated with `Balance`. The comparison is especially obvious in Lasso regression since Lasso eliminates some variables by setting coefficients to 0.

In terms of test MSE, Ridge = Lasso < PCR < PLSR. Shrinkage methods perform better than dimentaion reduction methods on this dataset. The takeaway here is that if you don't know which technique works better in the linear regression analysis, it's a good idea to try both on the dataset and choose the better one.

Also, we only used linear models. The next phase is probably to use some non-linear models and tune the corresponding parameters.