Project 2 - Predictive Modeling for Stat 159

Dakota Lim & Kartikeya Gupta

November 2, 2016

Abstract

This report, which is an extension to what we have learnt in class, explores the relationship between "Balance" (average credit card debt) and a set of variables given in the Credit data set. It closely follows the analysis done in Chapter 6 - Linear Model Selection and Regularization from the book An Introduction to Statistical Machine Learning by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. The main idea of this report is to fit a predictive model that can help decribe the relationship better.

Introduction

This report explores the relation between the average credit card debt across people and a set of variables that include credit limit, age, income, ethnicity etc. While still sticking to a linear model, we do however, explore models other than the OLS. To create this predictive model, we will be testing two shrinkage reduction methods - Ridge and Lasso, and two dimension reduction methods - Partial Least Squares and Principle Component Analysis. The final model for prediction will be that which explains the most of variation in the data for the given least Mean Squared Error.

Data

For this model, we will be using the Credit.csv data set available on the Introduction to Statistical Learning website: http://www-bcf.usc.edu/~gareth/ISL. It is also the same data set used in Chapter 6 of the book.

This dataset contains "Balance" which is the average credit card debt across people and 10 explanatory variables, of which 6 are quantitative and 4 and qualitative.

The quantitative variables are: Income, Credit Limit, Credit Rating, No. of Cards, Age and Education. The qualitative variables are: Gender, Student Status, Married and Ethinicity.

Method

We start our analysis by first conducting basic Exploratory Data Analysis to understand the variables we are working with and then following on with Pre-modeling Data Processing to structure the data in such a way that allows us to fit the multiple models on this set.

For EDA, we consider the summary statistics and use some basic plots to understand the variables. For qualitative variables, we create fequency tables and chart these proportions. Given that we want to understand the relationship between variables, we plot a matrix of correlation for the quantitative variables, ANOVA between Balance and qualitative variables along with conditional boxplots for the same.

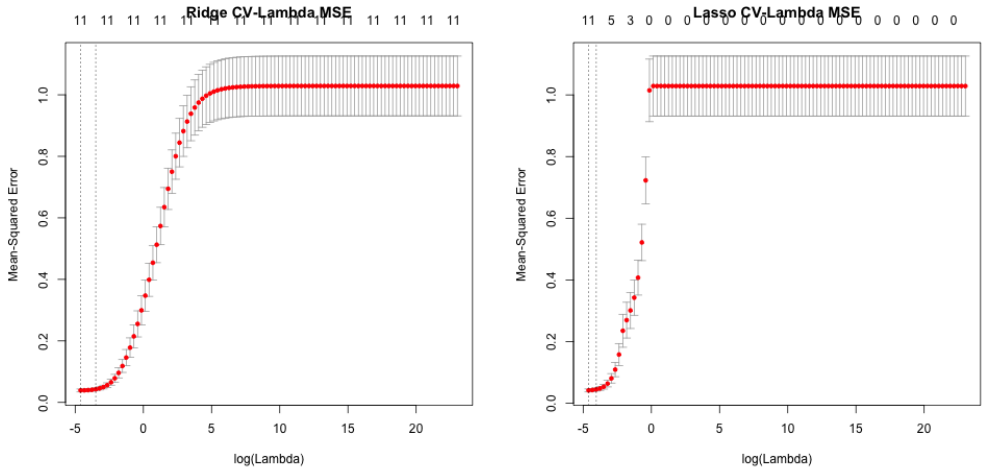
For Pre-modelling Data Processing, we follow the two major steps: 1. Coverting Factors into Dummy Variables so that they can be used in the regression. 2. Mean Centering and Standarization of Variables - This is done because the different models can only be compared when their coefficients represent the same scales. Hence, we mean center all the variables to 0 and the adjust the standard deviation to 1. Doing this given us comparable scales. Finally, we randomize this scaled data and take out a training set of size 3/4 of the data and the test set which is size 1/4 of the data.

Moving to the 5 regression models that we use:

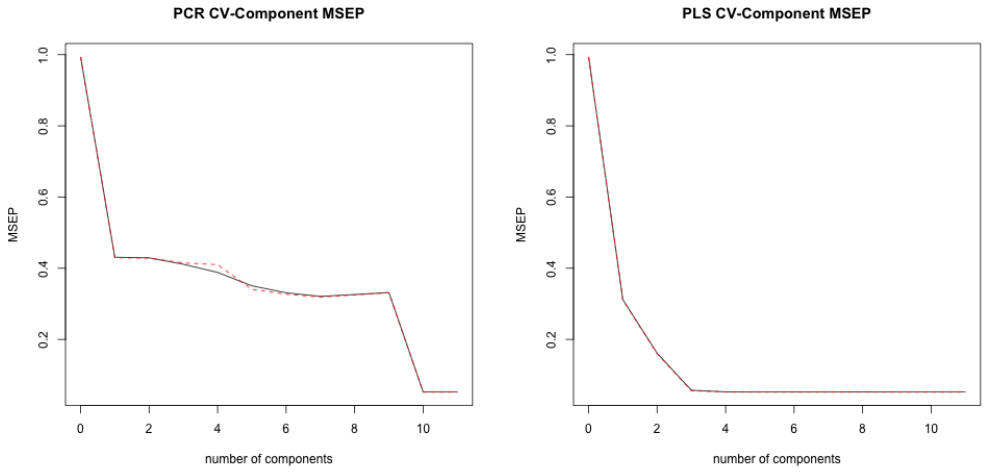
- An OLS model on the training data
- A Ridge model, using 10-fold cross validation to select lambda, on the training data
- A Lasso model, using 10-fold cross validation to select lambda, on the training data
- A PCR model, using 10-fold cross validation to select the number of components, on the training data
- A PLSR model, using 10-fold cross validation to select the number of components, on the training data

Each model was built using a scaled and centered version of the raw data. To ensure consistency we divide the data into test/training sets once and use the same sets to train and test each model.

Each model (excluding OLS) required a tuning parameter be definied, lambda values for Ridge and Lasso, and the number of components for PCR and PLS. To decide thse tuning parameters we used cross validation, below are the plots of prediction intervals for the MSE of the Ridge and Lasso methods as a function of different lambda values:

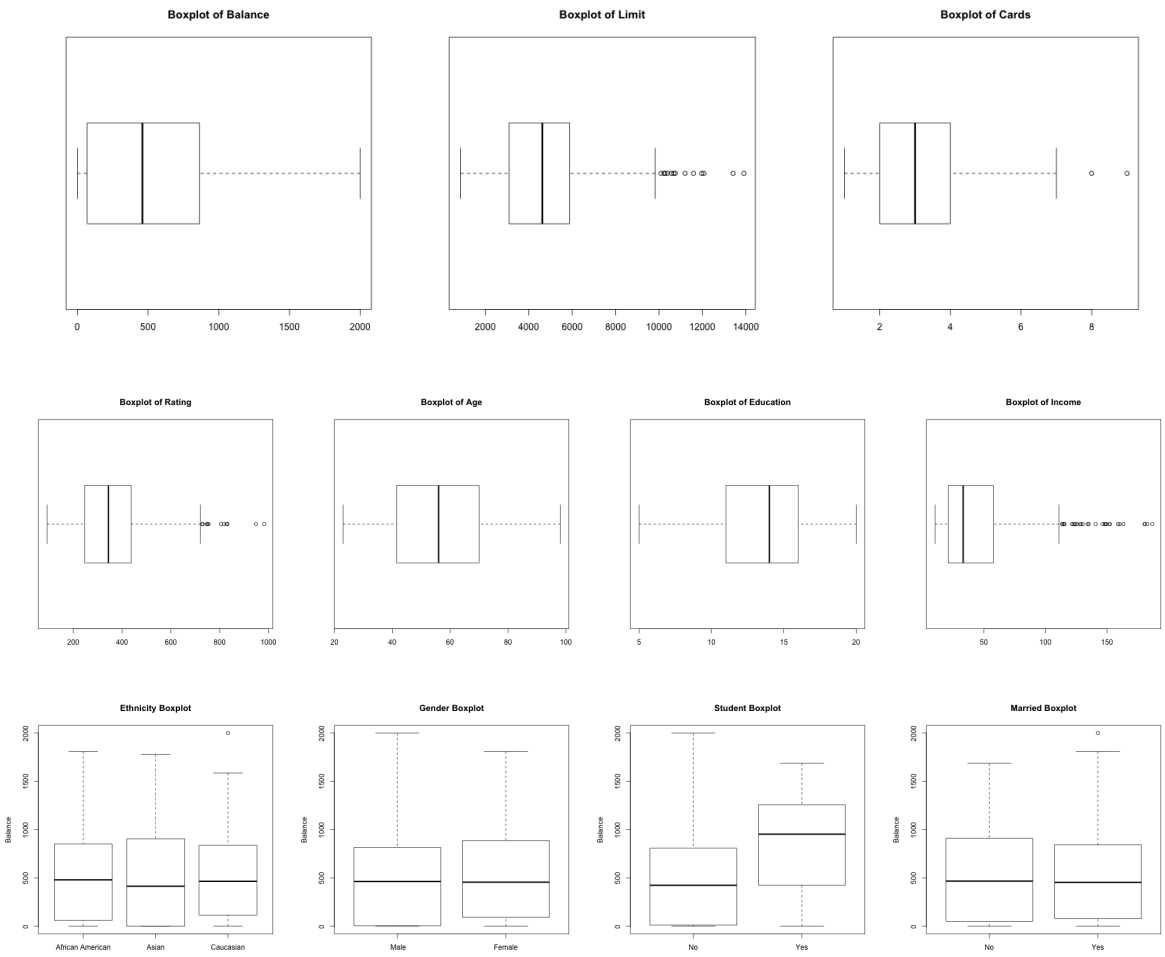


Next we consider the plots for MSEP as a function of the number of components taken in the PCR and PLS methods:

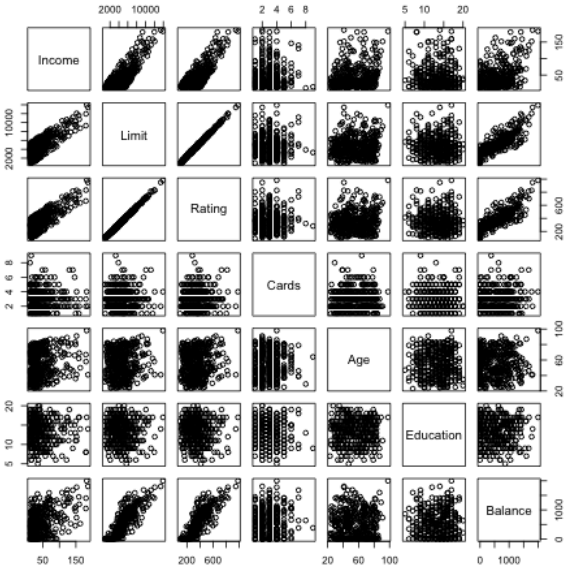


We simply take the minimal value (lambda or #-components depending on the model) of each plot and use that value as the tuning parameter for our full models.

Before we begin analyzing the coefficients from the regression models, we first take a look at the data.



Now to understand the relationship between each variable, we study the correlation matrix.

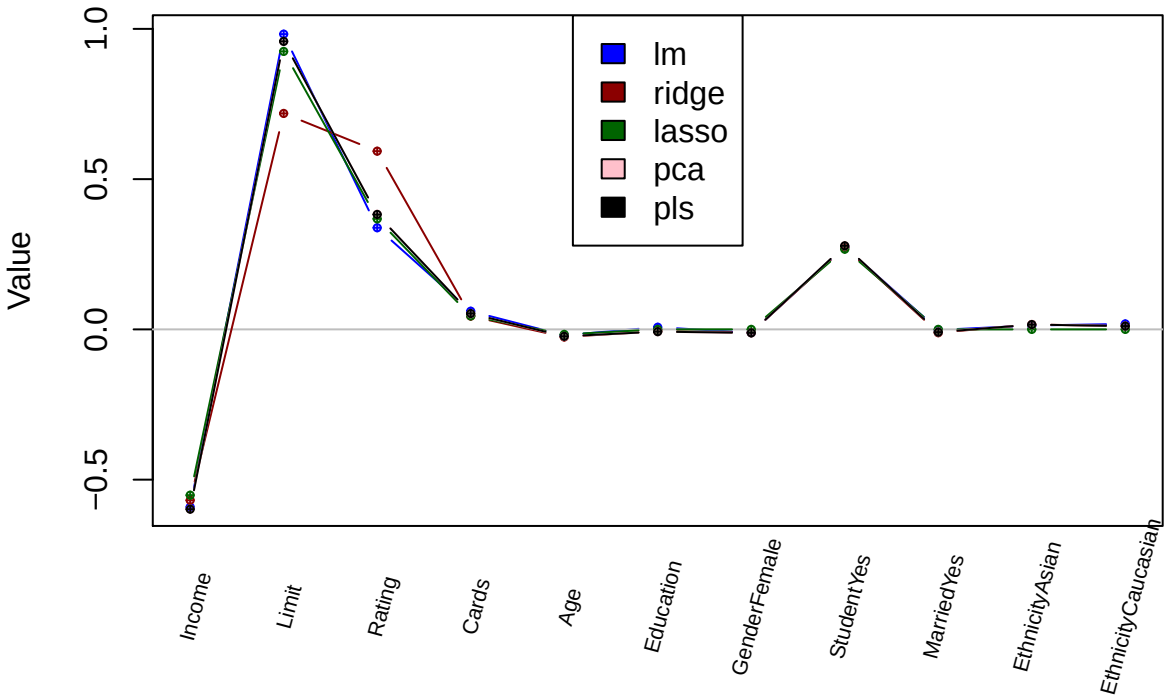


Analysis

Below is a table of the regression coefficients produced by each model on the full data along with a plot of all model coefficients:

Table 1: Model Coefficients

	lm	ridge	lasso	pls	pcr
Income	-0.59	-0.57	-0.55	-0.60	-0.60
Limit	0.98	0.72	0.93	0.96	0.96
Rating	0.34	0.59	0.37	0.38	0.38
Cards	0.06	0.04	0.04	0.05	0.05
Age	-0.02	-0.03	-0.02	-0.02	-0.02
Education	0.01	-0.01	0.00	-0.01	-0.01
GenderFemale	-0.01	-0.01	0.00	-0.01	-0.01
StudentYes	0.27	0.27	0.27	0.28	0.28
MarriedYes	-0.00	-0.01	0.00	-0.01	-0.01
EthnicityAsian	0.01	0.02	0.00	0.02	0.02
EthnicityCaucasian	0.02	0.01	0.00	0.01	0.01



The majority of the models produce similar values for each regressor, with the obvious exception of the ridge and lasso methods. The lasso coefficients for `Education`, `GenderFemale`, `MarriedYes`, `EthnicityAsian`, and `EthnicityCaucasian` are all 0, this is a direct result of the lasso penalty term being the L-1 norm of the Betas scaled by some lambda value. The ridge method appears to be giving much lower estimates of the `Limit` coefficient and higher values of the `Rating` coefficient.

From this we deduce that `Education`, `GenderFemale`, `MarriedYes`, `EthnicityAsian`, and `EthnicityCaucasian` are all *meaningless* in our lasso regression. Reducing the number of regressors in our model allows us to easily interpret the results and increases the overall effectiveness of our model. This is why we consider lasso-regression to be a shrinkage method, since it tends (as the number of regressors increases) to send the coefficients of regressors with little predictive power to 0. In short, the lasso-regression identified the `Education`, `GenderFemale`, `MarriedYes`, `EthnicityAsian`, and `EthnicityCaucasian` as unimportant variables and set their regression coefficients to 0.

The second shrinkage method is ridge-regression. In contrast to the lasso, the ridge pentaly term uses the L-2 norm of the Betas. While this may seem arbitrary, this reduces the probability of unimportant regressor coefficients being 0. Upon further inspection we note that the `Education`, `GenderFemale`, `MarriedYes`, `EthnicityAsian`, and `EthnicityCaucasian` regressors have comparatively low coefficient magnitueds, which further supports the thoery that these variables are less important to our regression. This leads us to believe both the Ridge and Lasso models will yield similar results.

For the quantitative predictors, there is a clear interpretation of the magnitude of their respective coefficients. The higher the magnitude, the more important the predictor. For the qualitative predictors, the distiction is not so clear. It would seems useless to assign a numerical weight to a qualitative value within a regression context, and for this reason we factor out all our qualitative vaiables. Now we consider weights to `0-1` indicator variables for each level of a particular regressor. For example `Ethnicity` has three levels: `Causcasian`, `Asian`, and `African American`; thus we create two (3 - 1 = 2) indicator variables `EthnicityCaucasian` and `EthnicityAsian` to represent the same information numerically (if both are 0 then the ethnicity was African American). Now we have the same interpretation for our qualitative regressors as we do our quantitative ones.

Next we consider the coefficients of the two dimmension reduction methods: Partical Component Regression (PCR) and Partial Least Squares (PLS). Both these methods produce nearly identical coefficeints, which we expect since they both use principle components effectively remove the correlation between any given regressors. Furthermore, PCR and PLS use similar numbers of components (10 and 9 respectively).

Results

Next we consider the training set mean squared errors of each model, to better address overall model effectiveness:

Table 2: Model Test MSE

	Model	TestMSE
1	OLS	0.0474
2	Ridge	0.0303
3	Lasso	0.0317
4	PCR	0.0397
5	PLS	0.0394

From this it is clear that OLS was the worst method, which we expected since our data clearly has correlated regressors (ex: Age and Education). The best methods were the ridge and lasso shrinkage methods, which yielded test MSEs of 0.0303 and 0.0317 respectively. The next best methods were the PCA and PLS dimmension reduction methods, which yielded test MSEs of 0.0397 and 0.0397.

Conclusion

Over the course of this project we have modeled the `Credit.csv` data in several ways. We began by cleaning and preparing the data, analyzing individual regressors (qualitative and quantitative) for any interesting facts, centering and standardizing the data, as well as factoring out levels of qualitative variables (i.e `Married` became `MarriedYes` with 0 for a `No` response and 1 for `Yes`). In preparation for building future models, we also partitioned the data into test and training sets. We chose not to cross-validate the linear model since it is simply a benchmark, we fully expect our models to outpreform this.

Next we created a baseline model to compare all other models to, for this we used a simple linear regression. This yielded a surprisingly good MSE of 0.04742. Our next approach was to try cross-validated ridge and lasso regressions, which are shrinkage methods. Through crossvalidation we found the optimal lamb Both ridge and lasso of these yielded much better cross-validated MSEs of 0.0302638 and 0.0317322 respectively.

Finally, we considered Partial Component Regression (PCR) and Partial Least Squares (PLS) which both yielded cross-validated MSE's of 0.0397271 and 0.0394221.

From this data is it clear that the ridge reggression model was the most effective at modeling the `Credit.csv` data.