Advertising Data Regression Fits

Elly Wang, Lily Li

November 5, 2016

Abstract

A prevalent problem among banking institutions is predicting the behavior of their clients such as paying their bills promptly. In this report, we will explore the relationship between an individual's characteristics and their payment behavior to develop a prediction model.

The motivation of this report is to develop a prediction model based on the characteristics of a credit card owner. The dataset and more information can be found in the book, *An Introduction to Statistical Learning*.

Introduction

The prediction models are built upon the credit data set, which contains:

Quantitative Variables:

- **balance**: average credit card debt for a number of individuals
- **age**: age of the individual
- **cards**: number of credit cards owned by the individual
- **education**: years of education of the individual
- **income**: in thousands of dollars
- **limit**: credit limit
- **rating**: the individual's credit rating

Qualitative Variables:

- **gender**
- **student**
- **married**: marital status
- **ethnicity**: Caucasian, African American or Asian

Some questions we would like to explore include:

1. What are the distributions of the different categories of the data?
2. How well do each model fit the data?
3. What is the MSE of each model?

Data

The credit data used in this project includes both categorical and quantitative data.

To take a closer look at each variables included in the **Credit** dataset, we looked at the distributions of each quantitative variable (Figures 1 and 2) and the conditional distribution of each categorical variable against Balance (Figure 3).

As shown through the historgrams, the distribution of Incomes, Limits, Ratings, # of cards, and balances are slightly skewed to the right. A possible explanation for the similar distribution could be that all five factors

are highly correlated – Individuals with higher incomes are more likely to have higher credit limit and better ratings.

	Income	Limit	Rating	Cards	Age	Education	Balance
Income	1	0.7921	0.7914	-0.0183	0.1753	-0.0277	0.4637
Limit		1	0.9969	0.0102	0.1009	-0.0235	0.8617
Rating			1	0.0532	0.1032	-0.0301	0.8636
Cards				1	0.0429	-0.0511	0.0865
Age					1	0.0036	0.0018
Education						1	-0.0081
Balance							1

Table 1: Correlation Matrix for Credit Data

Examining the correlation matrix and scatterplot matrix (Figure 4), we see that there is an extremely high correlation between Limit and Rating. In addition, the correlations between Income and Limit and Income and Rating are also fairly high. The high correlations could be potentional problems in the OLS regression due to almost collinearity.

In addition to histograms for quantitative variables in credit, we also examined the conditional boxplots of categorical variables against Balance (Figure 3). The boxplots show that the balance is right-skewed for most groups, and with exception of students who have a fairly symmetric distribution for balances. In addition to the distribution, we also note that excluding the students, the averages for all groups in each categorical variable are about the same. For the student variable, we see that the students have much higher average for balance when compared to the non-students, and that makes sense because usually students have little to no income and still have to pay tuition on their own.

Methods

Standardization

The dataset includes categories measured in different scales. To prevent any biased weighting, we want to standardize the dataset before building any models upon it.

First, we create dummy variables for the qualitative variables discussed in the introduction. Next, we want to standardize the absolute quantitites to relative quantities (mean centering). This means that each variable will have mean zero, and standard deviation one. One reason to standardize variables is to have comparable scales. When you perform a regression analysis, the value of the computed coefficients will depend on the measurement scale of the associated predictors.

Regression Models

Ordinary Least Squares Regression (OLS)

Based on the Gauss-Markov theorem, OLS is the best linear unbiased estimator. However, if predictors (regressors) are correlated, the stability of the $\hat{beta}$ decreases, meaning, every estimate of β could be very different and not converge to the true population coefficient.

$$Balance = \beta_1 Income + \beta_2 Limit + \beta_3 Rating + \beta_4 Cards + \beta_5 Age... + \beta_1 1 EthnicityCaucasian$$

Ridge Regression (RR)

RR is a variation of the minimization in OLS Regression but with a constraint of $||\beta||_2^2 < c^2$.

In vector form: $min\beta\ ||y - A\beta||_2^2 + \lambda||\beta||_2^2$

A difference in behavior of RR is that as λ increases, more weight is given to the second term in the minimization. This means that with a large λ, the β will be small.

The main advantage of RR is that it takes multicollinearity into account and does automatics parameter selection.

Lasso Regression (LR)

LR is a variation of the minimization in OLS Regression but with a constraint of $||\beta||_1 < c$. With c, the constraint shape becomes a diamond and any pairs of β will likely contain zeros. Unlike RR, there is no explicit form of β.

In vector form: $min\beta\ ||y - A\beta||_2^2 + \lambda||\beta||_1$

The main advantage of LR is that it performs both parameter shrinkage through feature selection (sparsify regressors/predictors) and variable selection automatically.

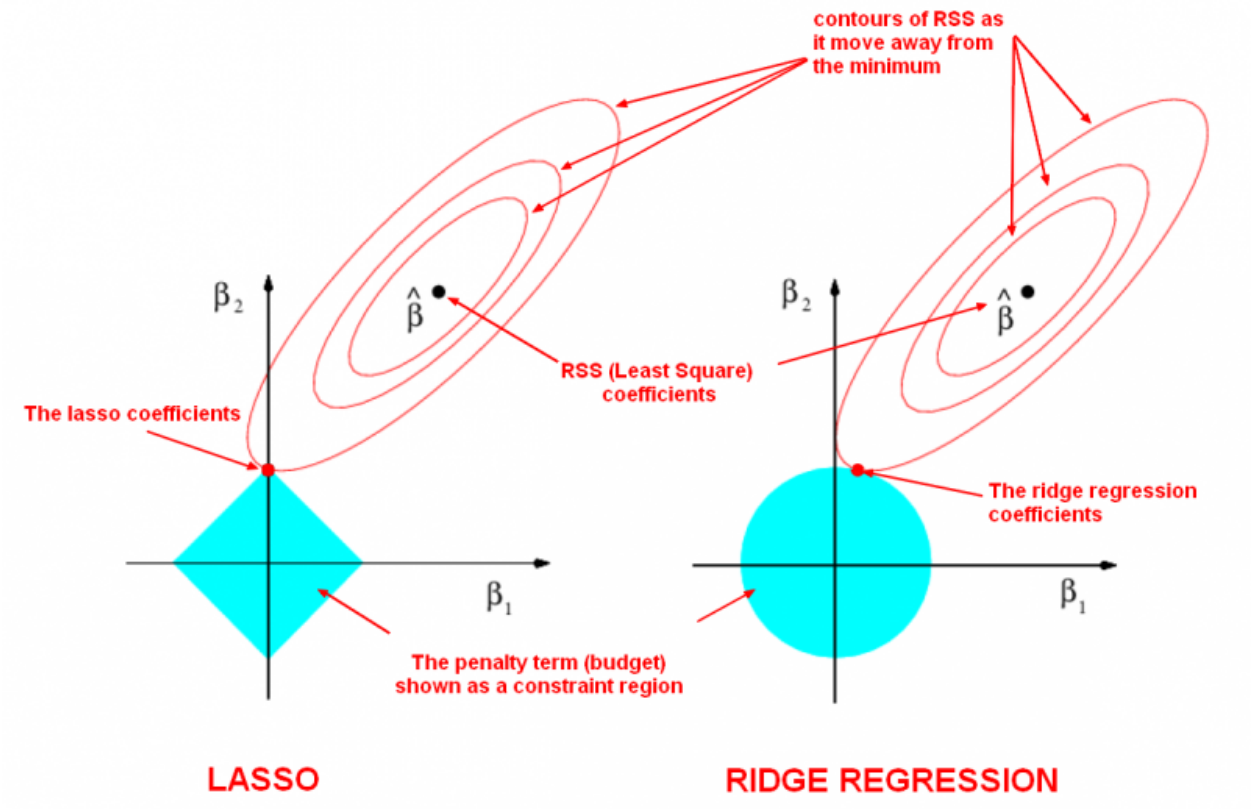


Figure 1: a visual representation of Ridge and Lasso Regressions, *Nicolas Gerard*

Principal Component Regression (PCR)

PCR is based on principal component analysis. The goal of this method is to reduce the dimensions (created by the set of data points in n-dimensional space).

To do so, we want define a direction, the first principal component, that maximizes the the variability in the data set and set the second principal component perpendicular to this first principal component. As a result, each data point's coordinates will change to this new coordinate system.

Having dimensions with the greatest variance will maximize preservation of distances between the data points. It's important because physical distances also represent similarity.

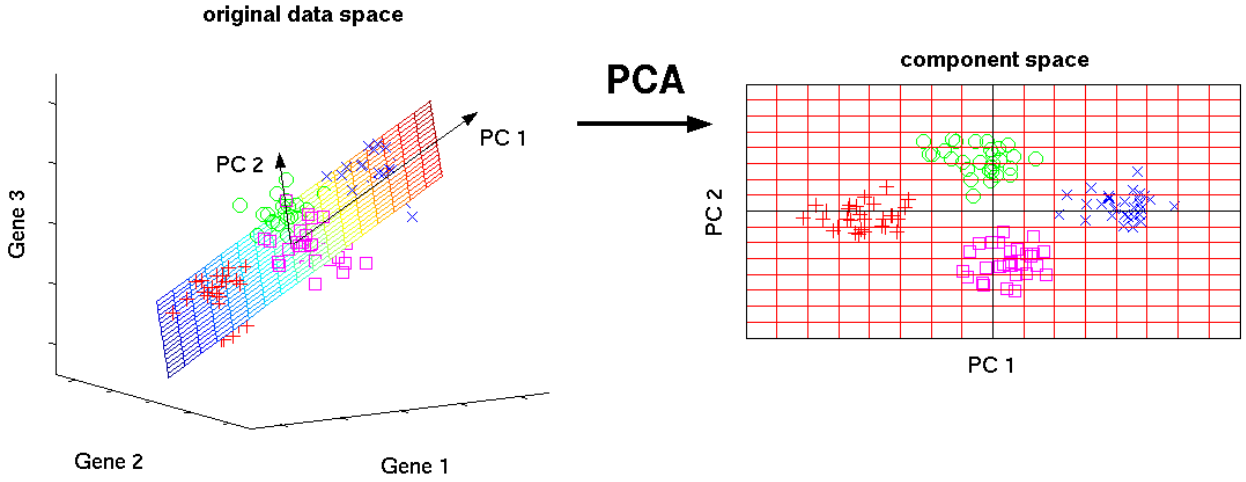


Figure 2: A visual representation of PCA with reduction from 3 dimensions to 2 dimensions, *NLPCA*

Partial Least Squares Regression (PLSR)

PLSR, similar to PCR, is also a dimensionality reduction method. While PCR finds hyperplanes of maximum variance between the predictors and responses, PLSR projects the predicted variables and the observable variables into a new space.

The main advantage is that PLSR uses the annotated label to maximize inter-class variance. It takes into account of the classes and tries to reduce the dimension while maximizing the separation of classes.

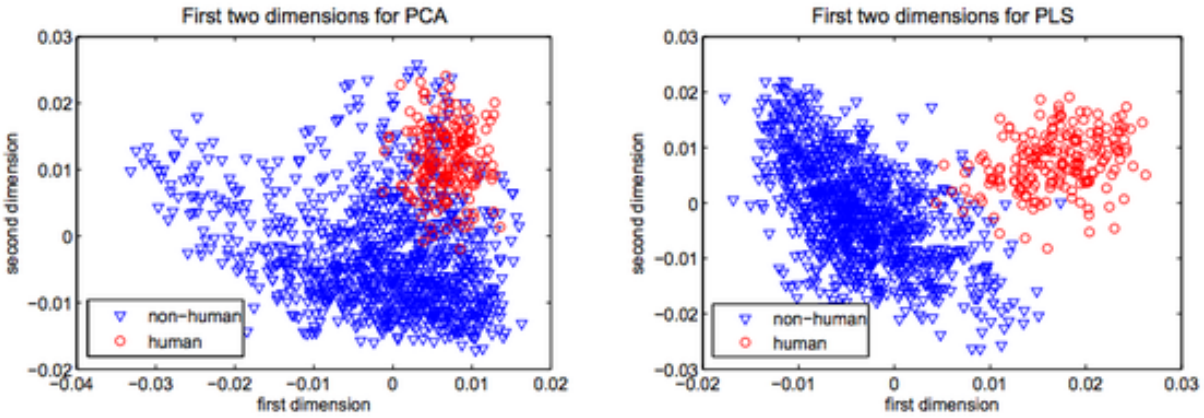


Figure 3: A visual representation of PCA and PLS differences, *Gustavo Fuhr*

Cross Validation and Train-Test Sets

Because we have limited amount of observations to build and test the model and we want to prevent bias, we will build and test the model using different subsets of the whole data set.

We built train sets of 300 out of the total 400 observations and test sets of the remaining 100 by random sampling (without replacement). We repeated this process 10 times for a 10 fold cross validation when we ran the regressions.

Analysis

Using the regression methods above, we performed analysis on the credit data set as described in this section.

OLS

To conduct the Ordinary Least Squares (OLS) regression, we first ran the regression using `lm()` on the training set. With the obtained coefficients, we predicted the balance for the test set and compared it with the actual balance data from the test set.

The table below shows the OLS coefficients we obtained from running OLS on the entire set of data (trained and test set).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0000	0.0107	0.00	1.0000
Income	-0.5982	0.0180	-33.31	0.0000
Limit	0.9584	0.1646	5.82	0.0000
Rating	0.3825	0.1652	2.32	0.0211
Cards	0.0529	0.0129	4.08	0.0001
Age	-0.0230	0.0110	-2.09	0.0374
Education	-0.0075	0.0109	-0.69	0.4921
GenderFemale	-0.0116	0.0108	-1.07	0.2832
StudentYes	0.2782	0.0109	25.46	0.0000
MarriedYes	-0.0091	0.0110	-0.82	0.4107
EthnicityAsian	0.0160	0.0134	1.19	0.2347
EthnicityCaucasian	0.0110	0.0133	0.83	0.4083

Table 2: Multiple Ordinary Linear Regression (OLS)

Ridge and Lasso Regressions

To run Ridge and Lasso Regressions in R, we used the package `glmnet`. In order to find the best lambda values for the regressions, we created an array of 100 lambda values that ranges from 0.01 to 10^{10} and used `cv.glmnet()` to run 10-fold cross validations on all those lambda values. With the function `lambda.min`, we were able to find that the best lambda values are 0.01 and 0.01 for ridge and lasso regression respectively.

Using the best lambda for each regression, we predicted the balance values with the predict function in R and compared them with the actual balance values in the test set to get the test MSE.

Lastly, we ran the regressions on the entire credit data using the best lambda values and got the final coefficient values show in the combined table in Table 3 under results.

Principal Components Regression and Partial Least Squares Regression

To perform Principal Components Regression (PCR) and the Partial Least Squares Regression (PLSR) in R, we used the package `pls`.

To begin, we ran PCR and PLSR on our train data set with cross validation. The functions to do so are `pcr()` and `pls()` for PCR and PLSR, respectively. Then, with the cross validation object, we used `which.min()` function to find the component that has the least predicted residual error sum of squares (PRESS) from the

cross-validation, and marked that as the best model. The best models we found for PCR and PLSR are 11 and 7, respectively.

Using the best model, we predicted the balance values for the test set. Comparing the predicted values with the actual values, we calculated the mean squared error to measure fitness.

Lastly, we ran the regressions on the entire credit data to get the final coefficients for each of the regression models (as shown below in Table 3).

Results

Using the procedures described above in Analysis, we obtained the following results for the final coefficient estimates.

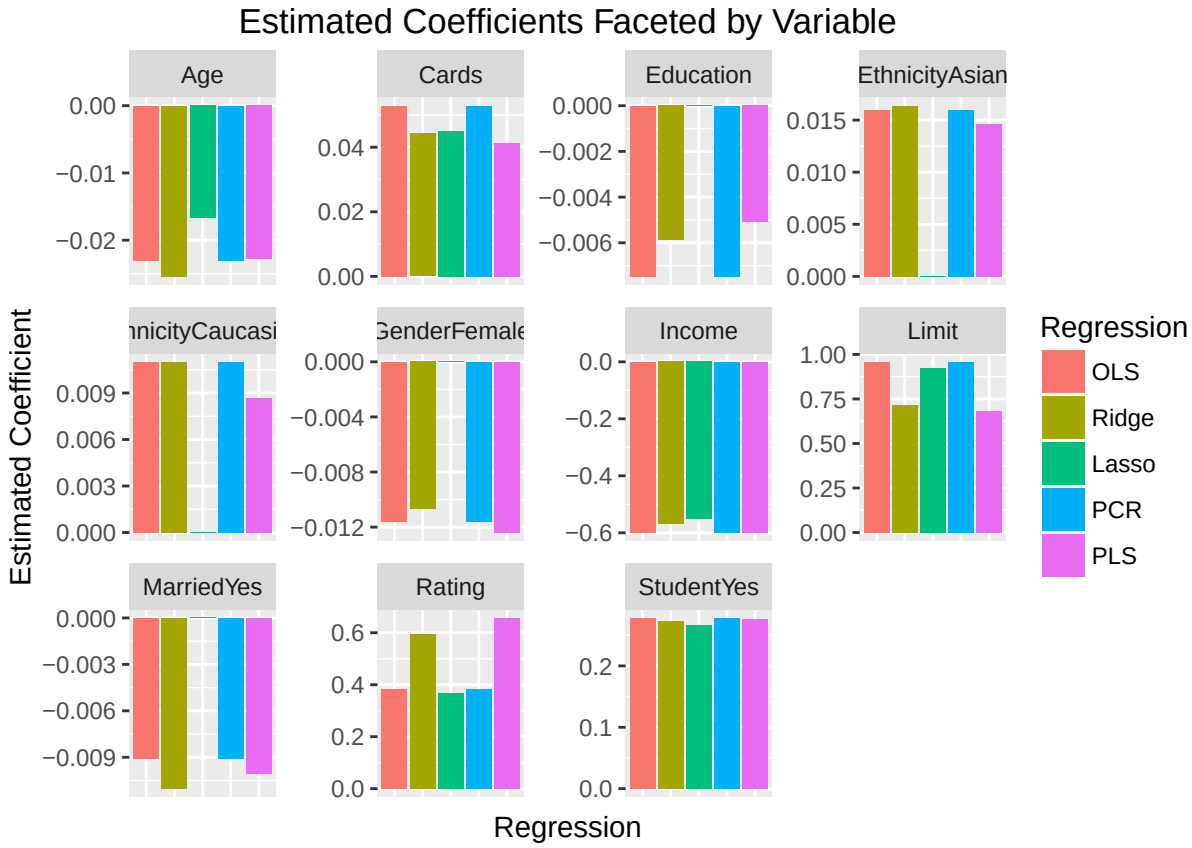
	OLS	Ridge	Lasso	PCR	PLS
(Intercept)	0.0000	0.0000	0.0000		
Income	-0.5982	-0.5687	-0.5517	-0.5982	-0.5989
Limit	0.9584	0.7187	0.9250	0.9584	0.6849
Rating	0.3825	0.5931	0.3679	0.3825	0.6570
Cards	0.0529	0.0443	0.0450	0.0529	0.0414
Age	-0.0230	-0.0254	-0.0167	-0.0230	-0.0228
Education	-0.0075	-0.0059	0.0000	-0.0075	-0.0051
GenderFemale	-0.0116	-0.0107	0.0000	-0.0116	-0.0124
StudentYes	0.2782	0.2732	0.2668	0.2782	0.2770
MarriedYes	-0.0091	-0.0110	0.0000	-0.0091	-0.0100
EthnicityAsian	0.0160	0.0164	0.0000	0.0160	0.0146
EthnicityCaucasian	0.0110	0.0110	0.0000	0.0110	0.0087

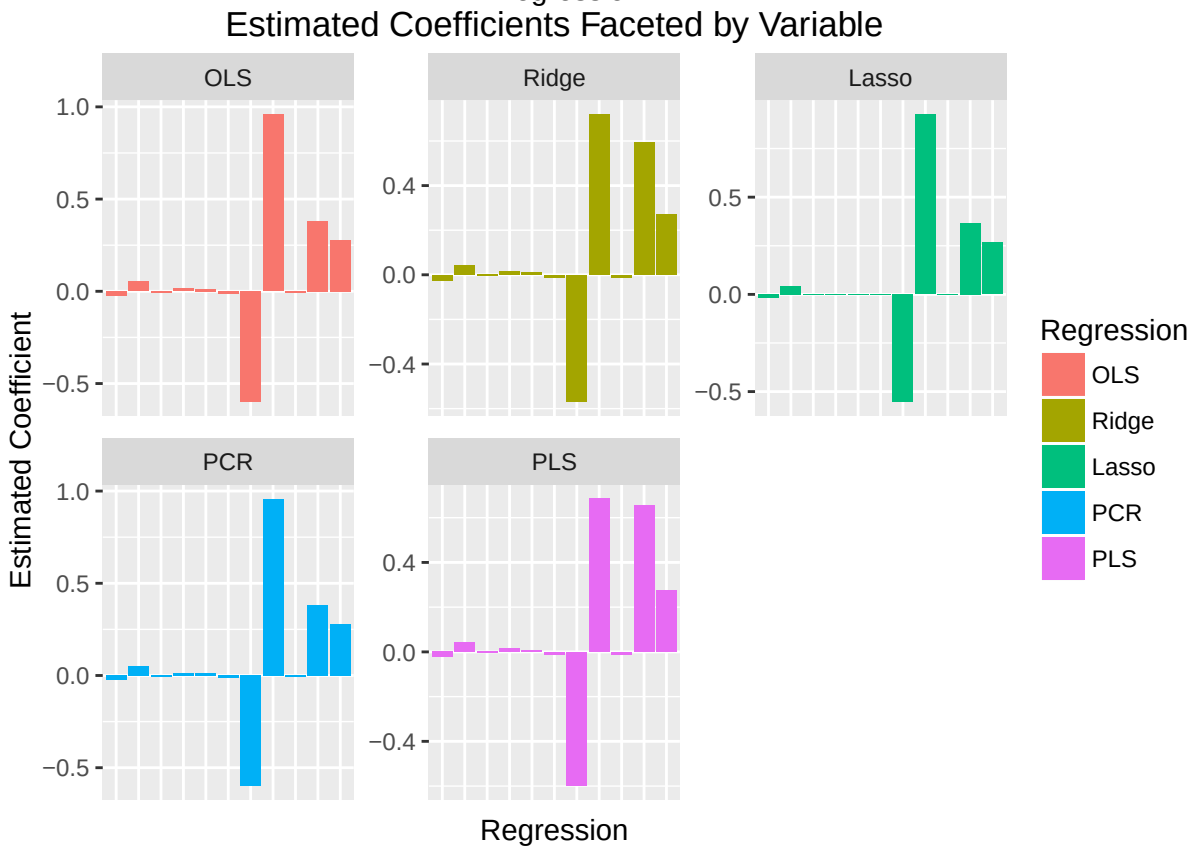
Table 3: Estimates of coefficients for all regressions

Comparing them side by side, we noticed that most coeffiicents are about the same. In particular, we noticed that the coefficient estimates for OLS and PCR came out to the be the same.

	OLS	Ridge	Lasso	PCR	PLS
MSE	0.055549994	0.055607096	0.055151748	0.055549994	0.056046873

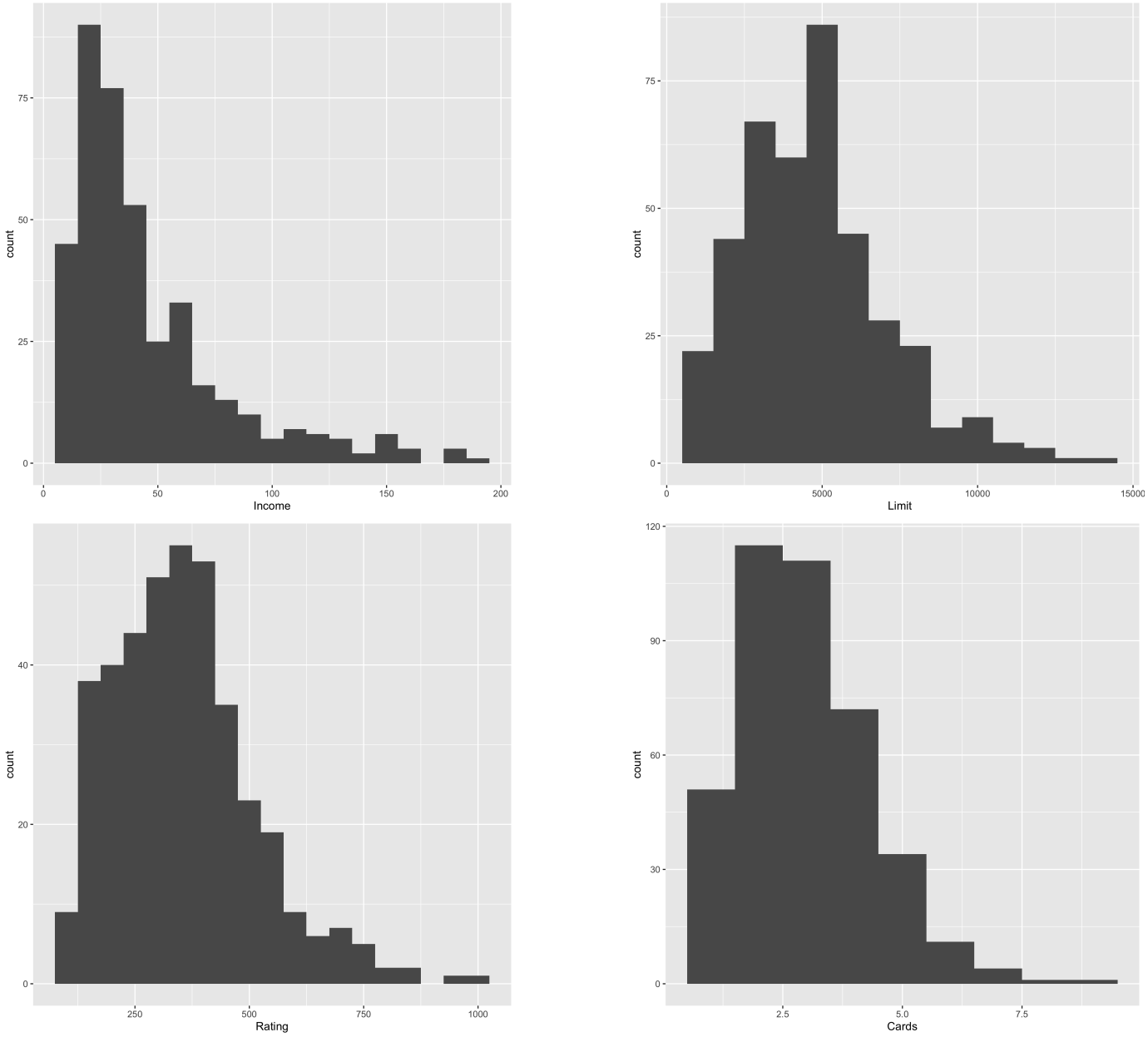
Table 4: Mean Squared Errors for all regression on test set

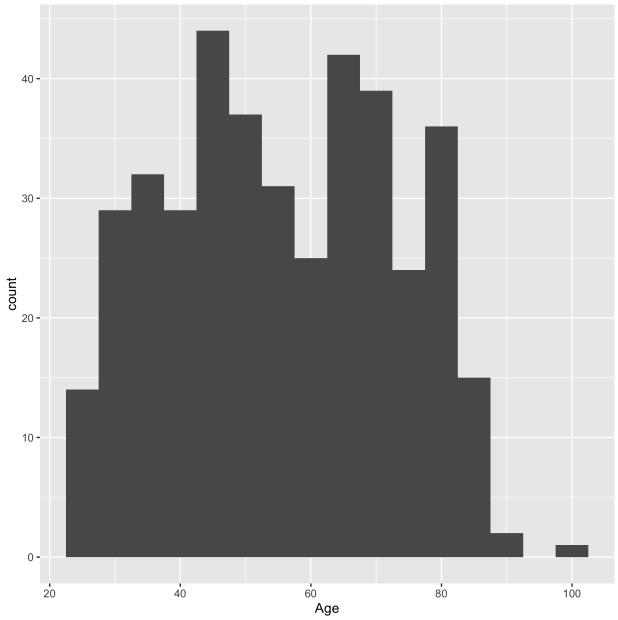

Estimated Coefficients Faceted by Variable
Age
Cards
Education
EthnicityAsian
nicityCaucasi
GenderFemale
Income
Limit
MarriedYes
Rating
StudentYes
Estimated Coefficient
Regression
OLS
Ridge
Lasso
PCR
PLS


Estimated Coefficients Faceted by Variable
OLS
Ridge
Lasso
PCR
PLS
Estimated Coefficient
Regression
OLS
Ridge
Lasso
PCR
PLS

Appendix

Figure 1: Histograms for quantitative variables in Credit




count
40
30
20
10
0
20
40
60
80
100
Age

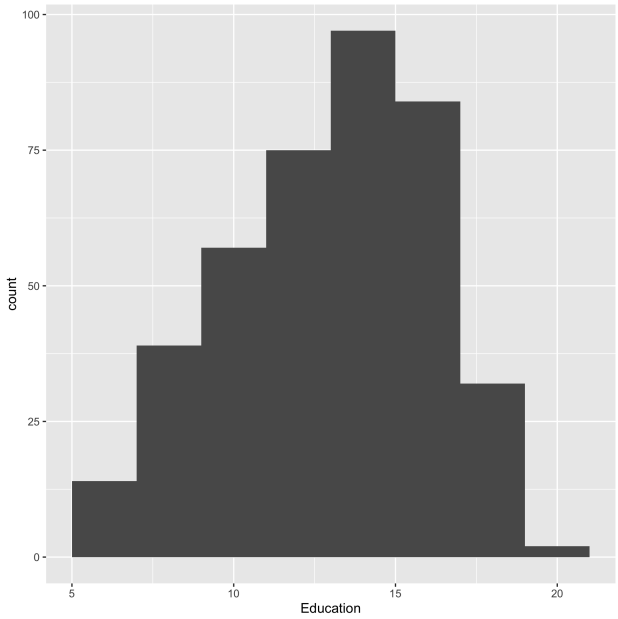

count
100
75
50
25
0
5
10
15
20
Education

Figure 2: Distribution of Balance

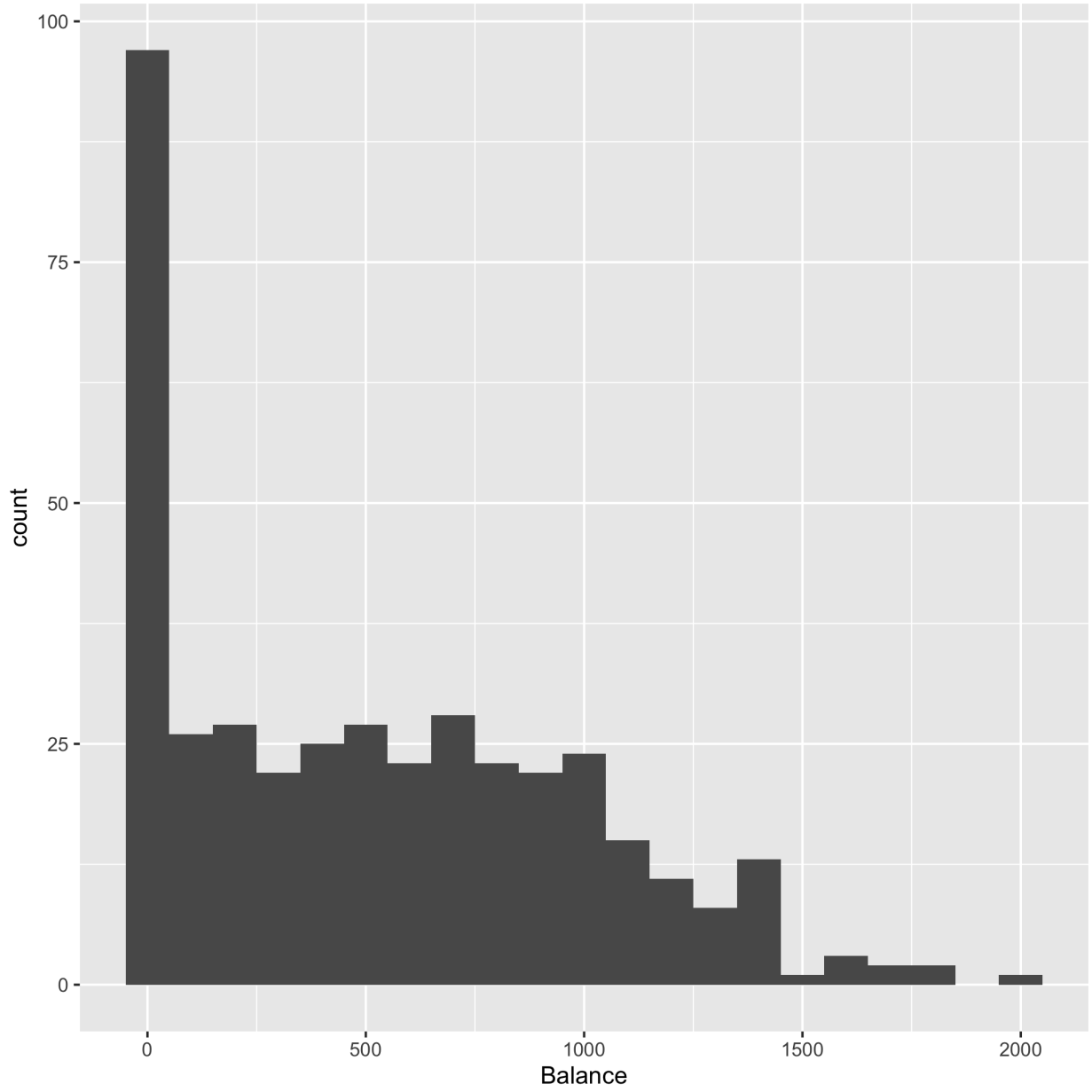


Figure 4: Fig 2: Balance

Figure 3: Conditional boxplots for categorical variables in Credit and Balance

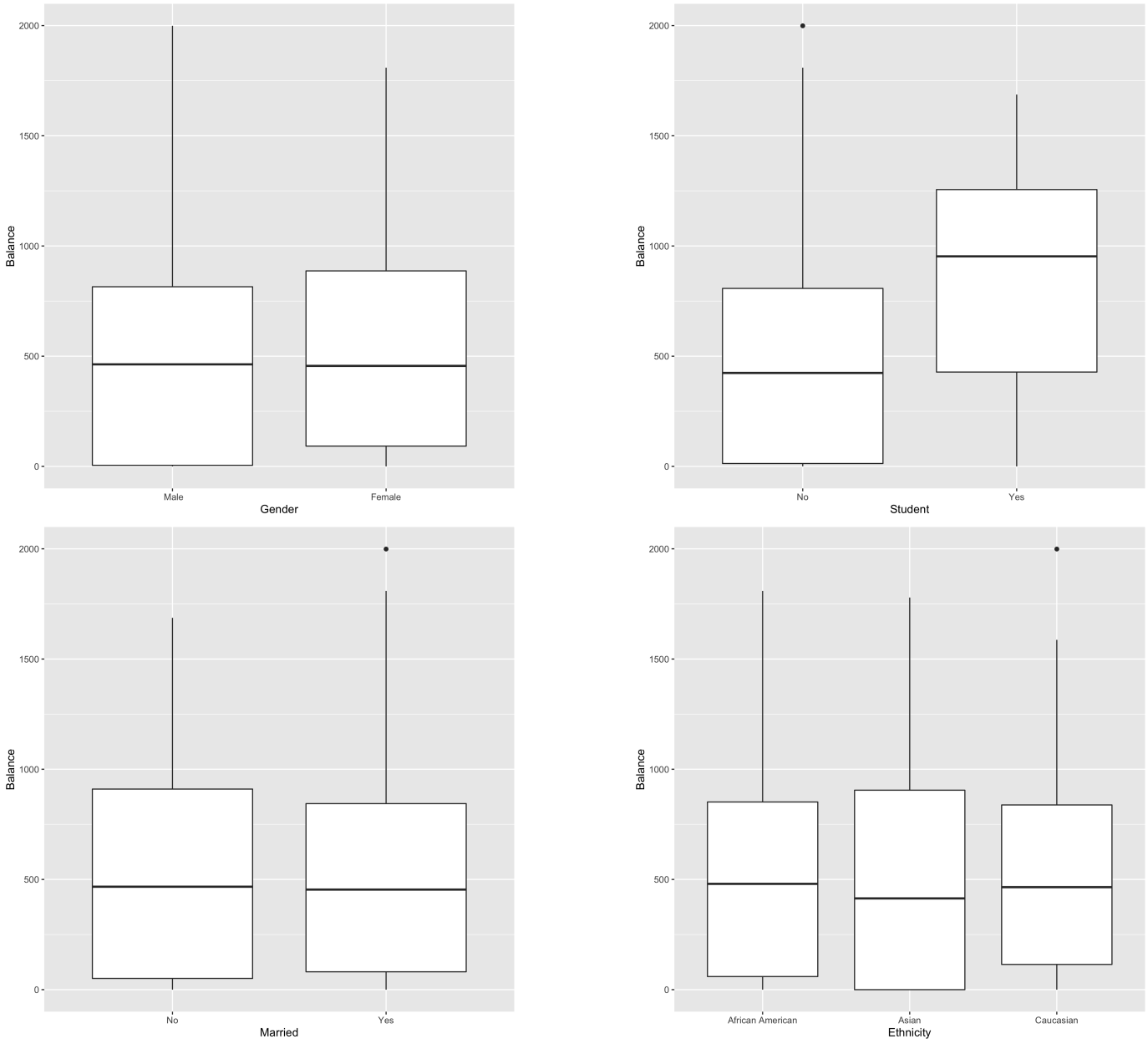


Figure 4: Scatterplot Matrix for all quantiative variables

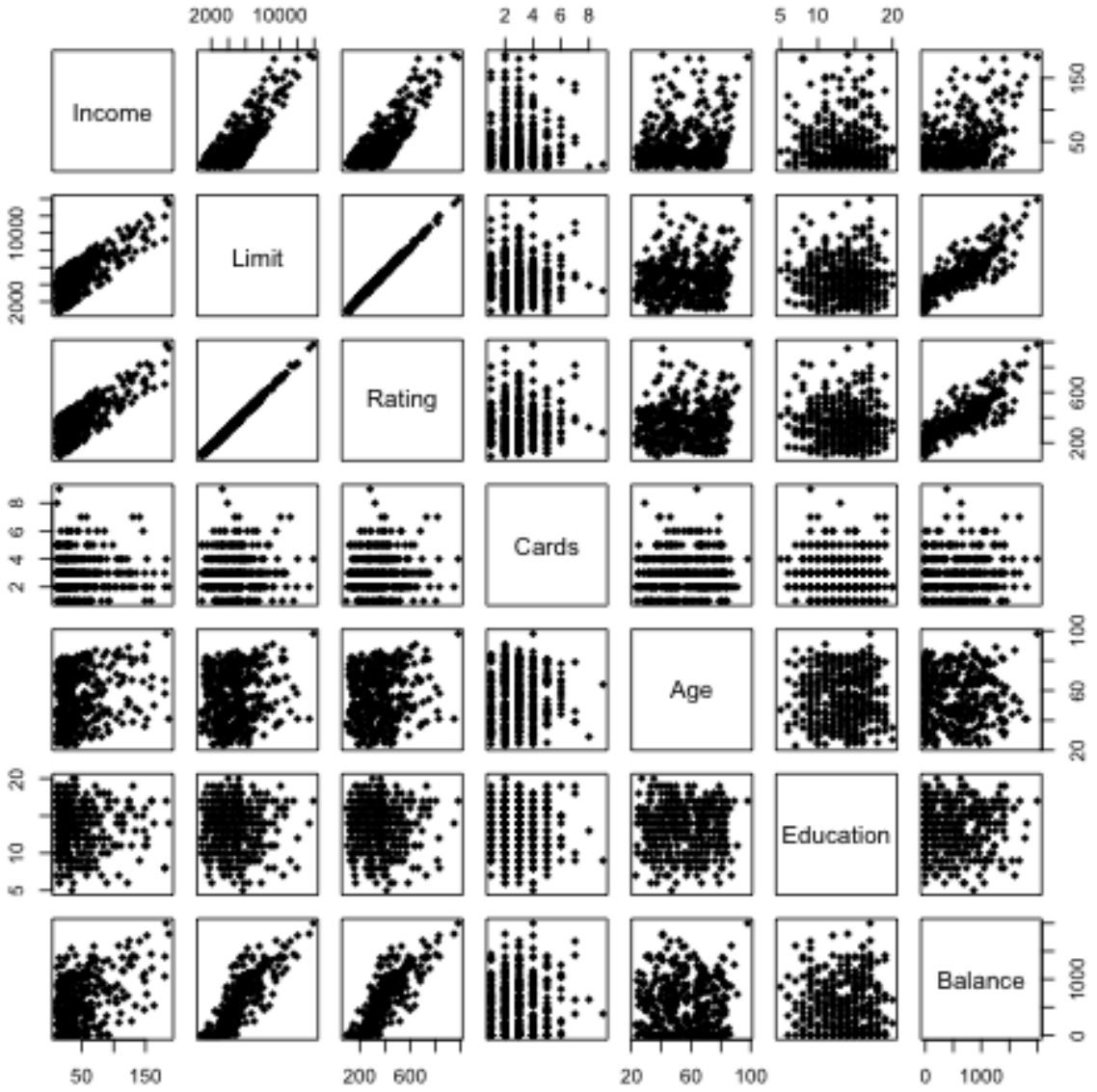


Figure 5: Fig 4: Scatterplot Matrix