Predictive Modeling on Credit Data

Kevin Liao and Thomas Sun

Abstract

This report applies predictive modeling techiques to find the relationship between credit card balance and various factors. Specifically, we use ridge regression, lasso regression, principal components regression, and partial least squares regression on credit card data to find the best fitting model to make accurate predictons on credit card balance. We find that all four methods produce models with similarly low mean squared errors, but ridge regression is the best performing model out of the four. From the coefficient estimates using ridge regression, we find that income, credit limit, credit rating, and student status are particularly useful for predicting credit card balance.

Introduction

The goal of the project is to use preditive modeling techniques in order to best predict how given inputs will affect the credit card balance of an individual. Specifically, there are ten predictors used to predict balance, of which can be classified as either quantitative or qualitative data. Quantitative variables include age, number of credit cards, years of education, income, credit limit, and credit rating. The qualitative variables are gender, student status, marital status, and ethnicity. All or some of these variables may be useful in predicting credit card balance.

In order to discover if there exists a relationship between the predictors and balance, one can use a variety of models applied to the dataset. A common model used is ordinary least squares regression (OLS). OLS estimators have the advantage of being unbiased given that the relationship between response and predictors is truly linear. In the *Credit* dataset, a multiple linear regression model can be used to fit the data.

However, OLS may have high variance and include irrelevant variables. Alternative methods may be used to improve the prediction accuracy and model interpretability. We use ridge regression, lasso regression, principal components regression (PCR), and partial least squares regression (PLSR) on the *Credit* in an attempt to find the best fitting model for predictive modeling.

Ridge regression and lasso regression, known as shrinkage methods, contrain the coefficient estimates and effectively shrinks them towards zero. Both tend to result in lower variance at the expense of more bias. Unlike ridge regression, lasso regression can have coefficient estimates equal to exactly zero, letting the resulting model only contain a subset of the predictors. Thus, lasso results are easier to interpret than results from ridge regression.

PCR and PLSR are methods that transform the predictors and reduces the number of coefficients needed to be estimated by least squares, known as dimension reduction. PCR uses principal components to explain the variance in the predictors, resulting in a single variable to predict the response on behalf of multiple variables. PLSR is similar to PCR, except PLSR also attempts to find a linear combination that best predicts the response in addition to explaining the original predictors well. Using either PCR and PLSR can prevent overfitting, a potential problem that arises in OLS.

In data with multiple dimensions, like in the *Credit* dataset, an OLS regression is prone to overfitting, especially when the regressors are highly collinear. We use ridge, lasso, principal component, and partial least squares regressions on *Credit* and try to pick the best model to fit the data. That is, find the model that is easy to interpret while having the best predictive capability.

Data

Data was obtained by downloading the *Credit* dataset made available by Gareth James on his website based on the related textbook, "An Introduction to Statistical Learning" by James et al. It contains the data `Balance`, credit card balances for a few hundred individuals, as well as data on several predictors associated with the individual. These variables include `Age`, `Cards` (number of credit cards), `Education` (number of years of education), `Income`, `Limit` (credit limit), `Rating` (credit rating), `Gender`, `Student` (student status), `Status` (marital status), and `Ethnicity`. `Gender`, `Student`, `Status`, and `Ethnicity` are qualitative variables, while the others are quantitative variables.

Methods

Exploratory Data Analysis

First, to better understand the *Credit* dataset, we perform exploratory data analysis. We obtain summary statistics of all the variables, as well as relevant plots to understand the distribution of each variable. We also find the matrix of correlations, scatterplot matrix, ANOVA between `Balance` and the qualitative variables, and a conditional boxplots of `Balance` conditioned to each qualitative variable.

Data Processing

Before fitting any of the models, we next conduct some data processing. The qualitative variables, `Gender`, `Student`, `Balance`, and `Ethnicity`, are categorical and thus need to be transformed into dummy variables, or binary indicators, to be used in the regression functions. The variables with two levels, `Gender`, `Student`, and `Married`, have one binary indicator for each, in which each observation takes a value of zero or one. `Ethnicity` has three levels so we create two binary indicators for this variable. We also mean center and standardize the data to remove different measurement scalings and be more comparable. So, each variable has a mean of zero and standard deviation of one.

Model Building

There are five regression models to be fitted to the *Credit* dataset, ordinary least squares, ridge, lasso, principal components, and partial least squares. To build the models, we use a training dataset, containing a random sample of 300 out of 400 observations in *Credit*. The remaining 100 observations will be the test set, used to test the performance of the model.

For the non-OLS models, we use ten-fold cross-validation to see how the predictive models would perform in practice. The training set is partitioned into ten equally sized subsamples. One fold is used as the testing set, while the other nine are used to fit the relevant model. This is then repeated for each fold so that all observations are tested exactly once. The cross validation is then used to tune a certain parameter, depending on the model, and then find the value of the parameter that results in the smallest cross validation error. This parameter is then selected as the "best" model, which is then fitted to the test set and finally the full dataset.

Regression Models

In order to find the relationship between *Credit* and the ten predictors to be used for predictive modelling, we assume the relationship between the independent and dependent variables is linear. The relationship is assumed to be the following:

$$Credit_i = \beta_0 + \beta_1 X_i + \beta_2 X_i + \ldots + \beta_{10} X_i + \epsilon_i$$

Where β_0 is the intercept and $\beta_1, \ldots, \beta_{10}$ are the regression coefficients for their associated predictor, and ϵ is the error term.

We first use an ordinary least squares method to be used as a benchmark for comparison between the other models. OLS estimates the coefficients by minimizing the residual sum of squares (RSS), defined as

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

For ridge regression, we minimize the RSS in addition to a shrinkage penalty, defined as

$$RSS + \lambda \sum_{j=1}^{p}\beta_j^2$$

Where λ is the tuning parameter. As $\lambda \to \infty$, the shrinkage penalty grows, effectively shrinking the coefficients $\beta_1, \ldots, \beta_{10}$ towards zero.

Lasso regression is another shrinkage method like ridge regression. The quantity we want to minimize is defined as

$$RSS + \lambda \sum_{j=1}^{p}|\beta_j|$$

This is similar to the quantity for ridge regression, except the penalty now contains $|\beta_j|$ instead of β_j^2. The tuning parameter λ is the same, except now sufficiently large λ may shrink coefficients to exactly zero.

For principal components regression, the method tries to reduce the dimensions $X_1, \ldots, X_p$ of the data matrix into principal components $Z_1, \ldots, Z_M$ and then using the components as the predictors for least squares regression. The M principal compenents will be the tuning parameter, and we use cross validation find which M produces the smallest mean squared error.

Lastly, partial least squares, also a dimension reduction method, tries to find $Z_1, \ldots, Z_M$ that approximate the original dimensions like PCR, but also tries to find new features related to the response *Balance*. The tuning parameter, the number of M directions, is the same as well. The M that is associated with the smalled mean squared error will be selected for the model.

Once all of the best models are identified, the test set will be used to compute the MSEs of each, and find which model performs best. The best model will finally be used on the full `Credit` data set to find official coefficients.

Analysis

OLS

First, let's look at our benchmark - OLS regression. We use full set of data that is mean centering and standardizing to fit the OSL model. OSL model will be served as our benchmark for comparison with later four different methods. So we calculate OSL model's MSE, which is equal to 0.0447862

Here is more information about OSL regression:

Table 1: Summary Table of OSL Regression

	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 1 exhibits estimates of all coefficients and their corresponding t-test and p-value.

From above information, we can conclude that predictors such as *Income*, *Limit*, *Rating*, *Cards*, and *Student* are extremely significant.

Ridge Regression

Next, let's look at ridge regression. We load mean centerized and standardized data before the analysis. Here are the steps: 1. Check mission value in data for both train and test set 2. Initialize lambda and use random seeds (set.seed()) for cross-validation 3. Use train set to conduct 10-fold cross-validation to find out the best tuning parameter

Let's look at plot of the cross-validation errors in terms of the tunning parameter

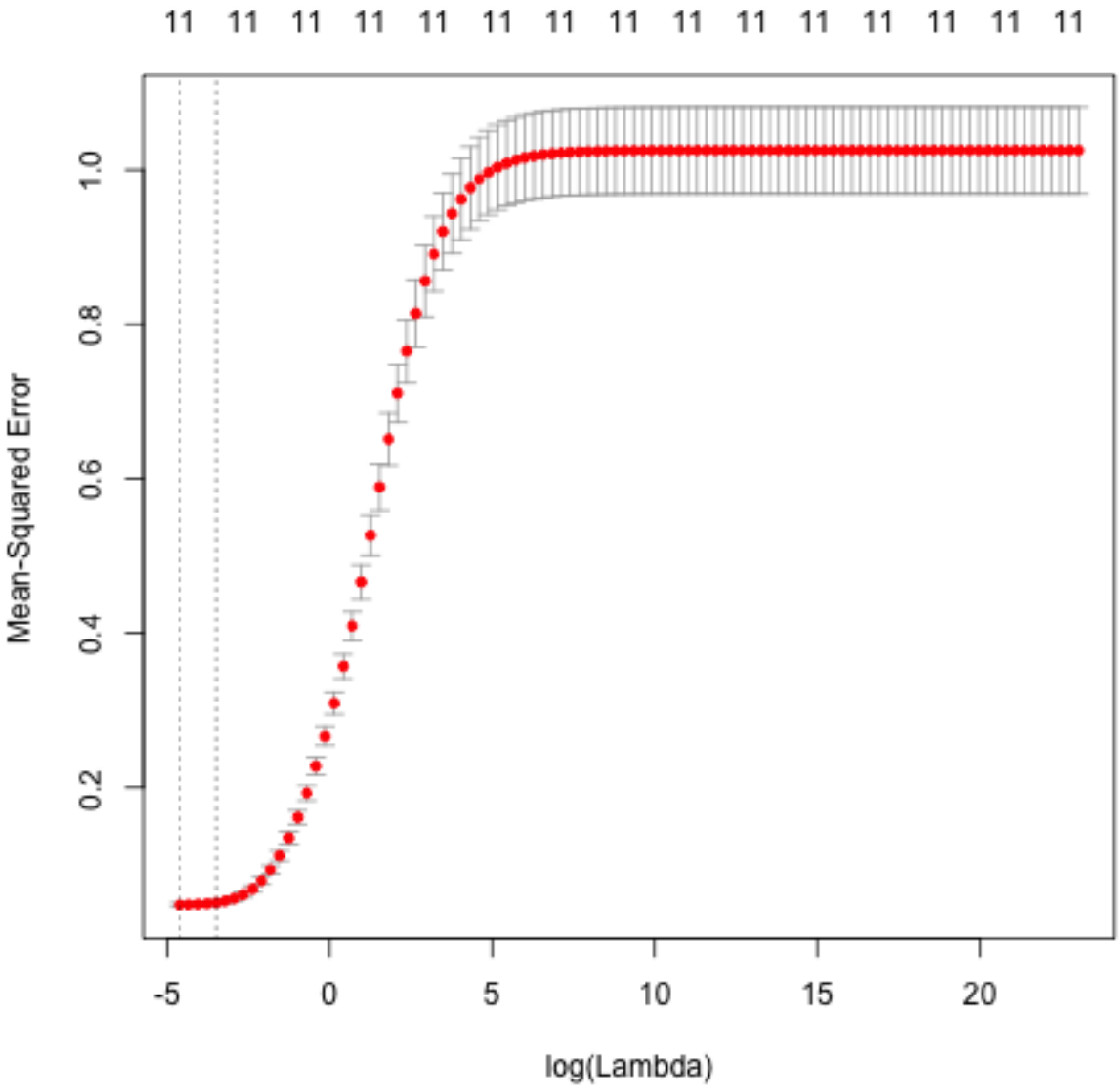


Figure 1: The cross-validation errors in terms of the tunning parameter

4. Find the λ for the best model: best $\lambda = 0.01$
5. Use the test set to compute the test Mean Square Error: test MSE $= 0.0454419$
6. Refit the model on the full data set using the best lambda and get official coefficients.

Here is the official coefficient table:

Table 2: Coefficient Table of Ridge Regression

	s0
(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

Lasso Regression

Next, let's look at lasso regression. We load mean centerized and standardized data before the analysis. Here are the steps: 1. Check mission value in data for both train and test set 2. Initialize lambda and use random seeds (set.seed()) for cross-validation 3. Use train set to conduct 10-fold cross-validation to find out the best tuning parameter

Let's look at plot of the cross-validation errors in terms of the tunning parameter

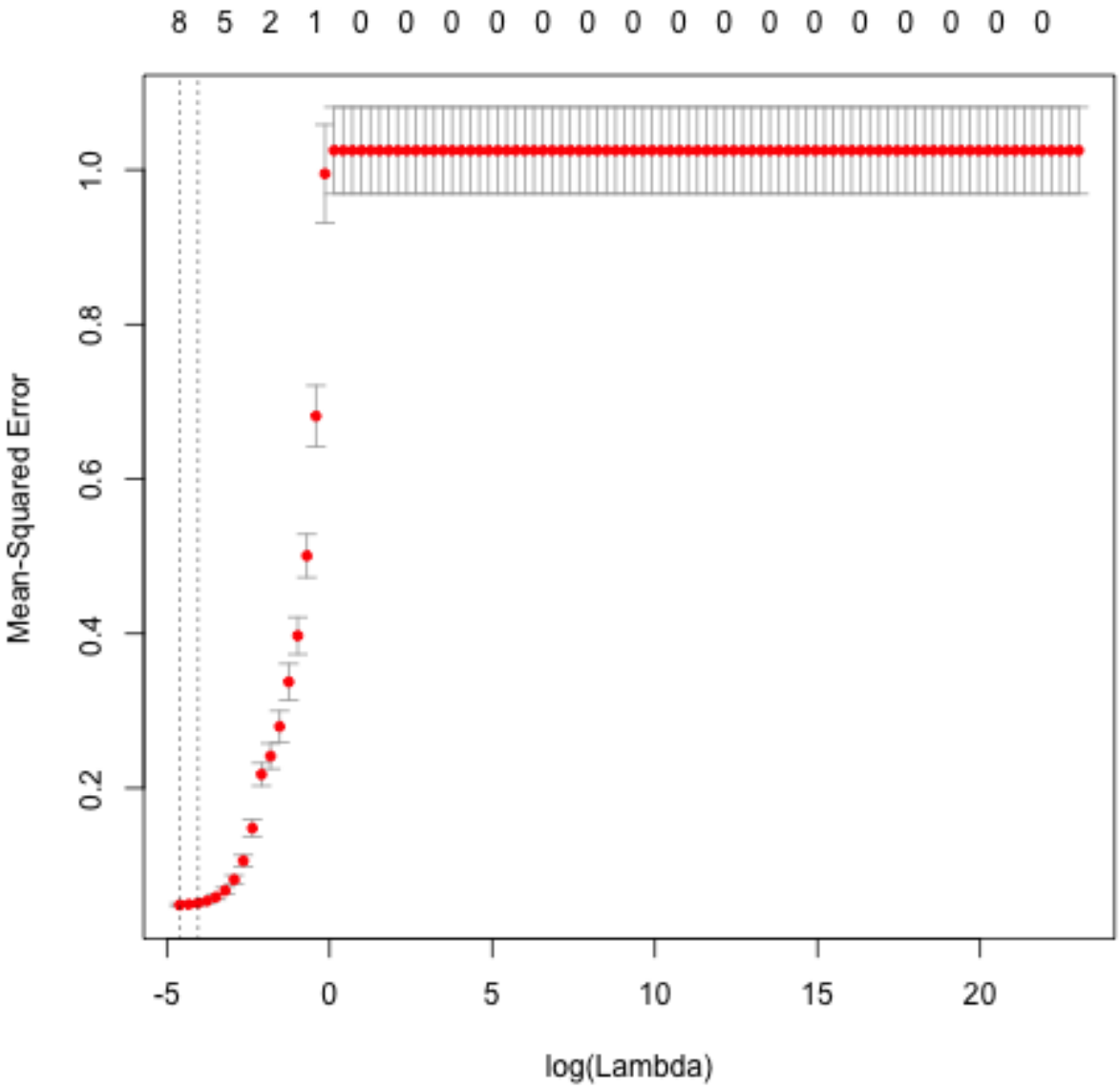


Figure 2: The cross-validation errors in terms of the tunning parameter

4. Find the λ for the best model: best $\lambda = 0.01$
5. Use the test set to compute the test Mean Square Error: test MSE $= 0.0467553$
6. Refit the model on the full data set using the best lambda and get official coefficients.

Here is the official coefficient table:

Table 3: Coefficient Table of Lasso Regression

	s0
(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

Principal Component Regression (PCR)

Next, let's look at PCR. We load mean centerized and standardized data before the analysis. Here are the steps: 1. Check mission value in data for both train and test set 2. Use random seeds (set.seed()) for cross-validation 3. Use train set to conduct 10-fold cross-validation to find out the best number of principal components used

Let's look at plot of the cross-validation errors in terms of the number of principal components used

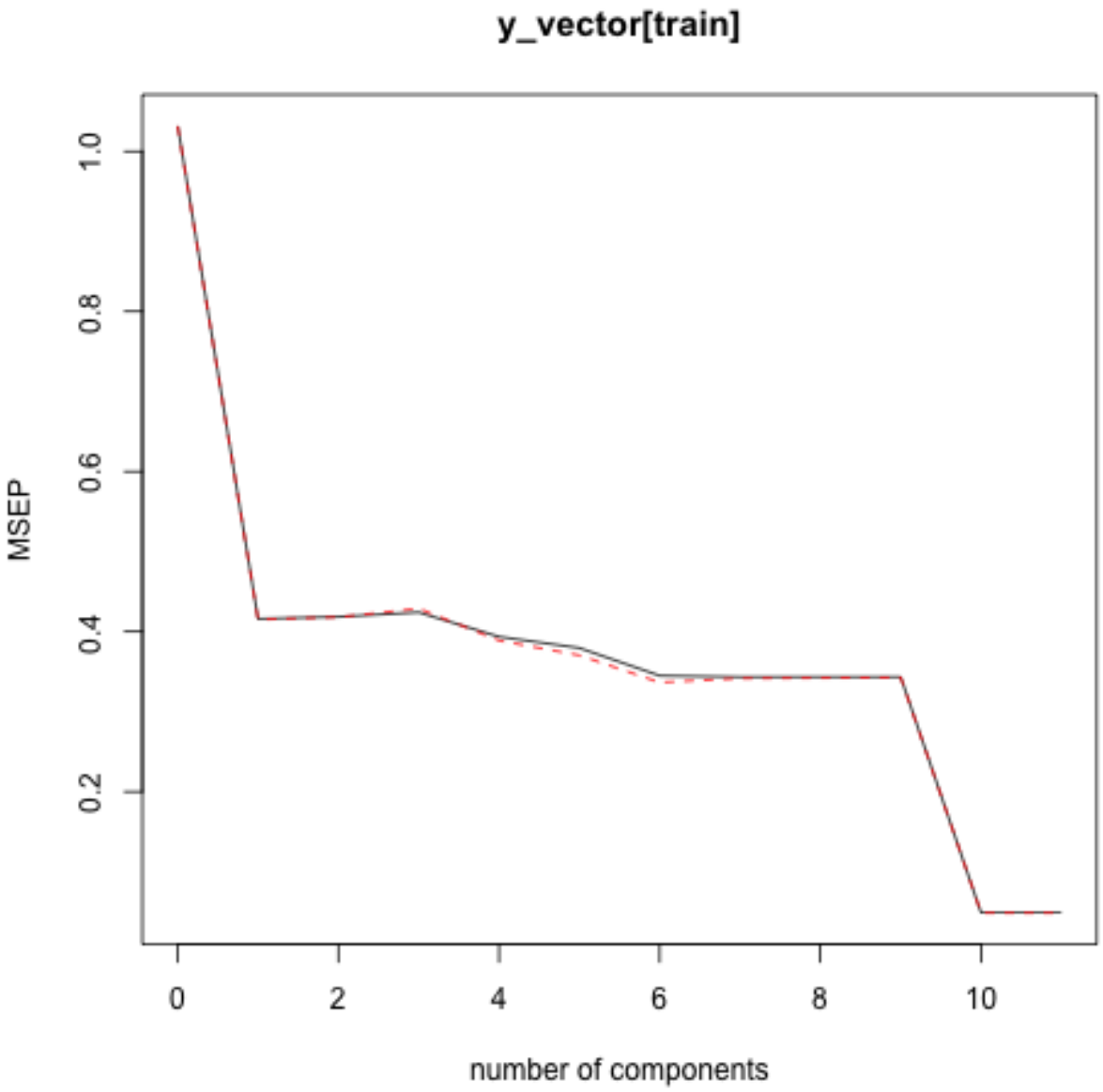


Figure 3: The cross-validation errors in terms of the number of principal components used

4. Find the number of principal components considered for the best model: M = 10
5. Use the test set to compute the test Mean Square Error: test MSE = 0.0460076
6. Refit the model on the full data set using M = 10and get official coefficients.

Here is the official coefficient table:

Table 4: Coefficient Table of PCR

	y_vector.10 comps
Income	-0.60
Limit	0.67
Rating	0.67
Cards	0.04
Age	-0.02
Education	-0.01
GenderFemale	-0.01
StudentYes	0.28
MarriedYes	-0.01
EthnicityAsian	0.02
EthnicityCaucasian	0.01

Partial Least Squares Regression (PLSR)

Next, let's look at PLSR. We load mean centerized and standardized data before the analysis. Here are the steps: 1. Check mission value in data for both train and test set 2. Use random seeds (set.seed()) for cross-validation 3. Use train set to conduct 10-fold cross-validation to find out the best number of partial least squares directions

Let's look at plot of the cross-validation errors in terms of the number of partial least squares directions

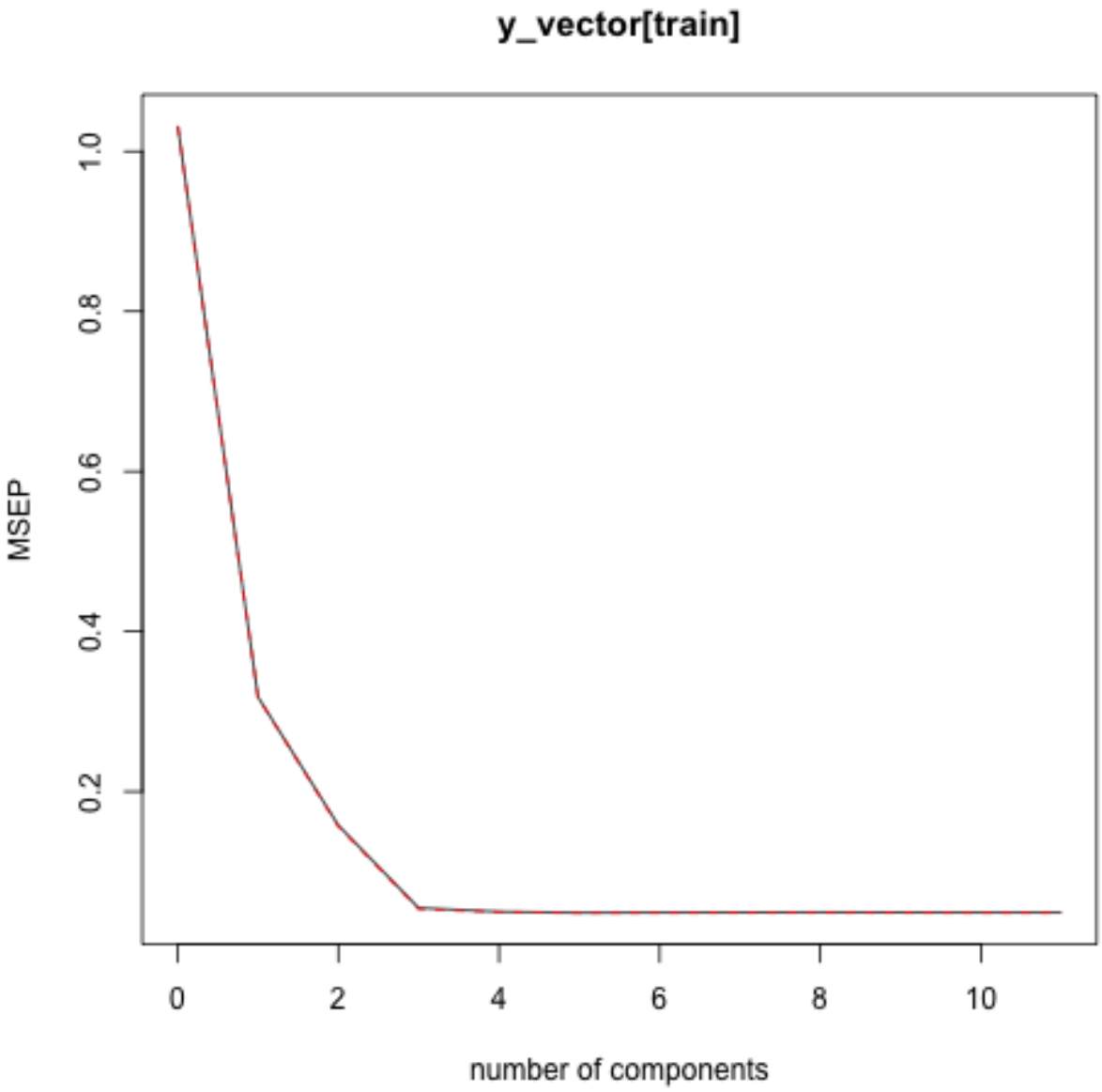


Figure 4: The cross-validation errors in terms of the number of partial least squares directions

4. Find the number of partial least squares directions for the best model: M = 5
5. Use the test set to compute the test Mean Square Error: test MSE = 0.0465753
6. Refit the model on the full data set using M = 5and get official coefficients.

Here is the official coefficient table:

Table 5: Coefficient Table of PLSR

	y_vector.5 comps
Income	-0.60
Limit	0.68
Rating	0.67
Cards	0.04
Age	-0.02
Education	-0.01
GenderFemale	-0.01
StudentYes	0.28
MarriedYes	-0.01
EthnicityAsian	0.02
EthnicityCaucasian	0.01

Results

Let's start with looking at regression coefficients for all methods including: ols, ridge, lasso, pcr, and plsr.

Table 6: Table of Regression Coefficients for All Methods

	OSL	ridege	lasso	PCR	PLSR
(Intercept)	0.000	0.000	0.000	0.000	0.000
Income	-0.598	-0.569	-0.552	-0.599	-0.599
Limit	0.958	0.719	0.925	0.671	0.676
Rating	0.382	0.593	0.368	0.671	0.666
Cards	0.053	0.044	0.045	0.040	0.041
Age	-0.023	-0.025	-0.017	-0.023	-0.023
Education	-0.007	-0.006	0.000	-0.006	-0.006
GenderFemale	-0.012	-0.011	0.000	-0.012	-0.012
StudentYes	0.278	0.273	0.267	0.276	0.277
MarriedYes	-0.009	-0.011	0.000	-0.011	-0.011
EthnicityAsian	0.016	0.016	0.000	0.017	0.019
EthnicityCaucasian	0.011	0.011	0.000	0.011	0.013

Table 1 has twelve rows (one intercept term and eleven predictors terms) and five columns (one column per regression methods: ols, ridge, lasso, pcr, and plsr).

From Table 1, the result shows that regression coefficients for ridge, lasso, pcr, and plsr are approximately closed to each other's value but a slightly different comparing to ols - our benchmark.

Not surprisingly, we have seen that some coefficients in lasso regression are zero because lasso regression allows coefficients to be zero to minimize the regression penalty.

Now let's look at another table with the test MSE values for the regression techniques: ridge, lasso, pcr, and plsr

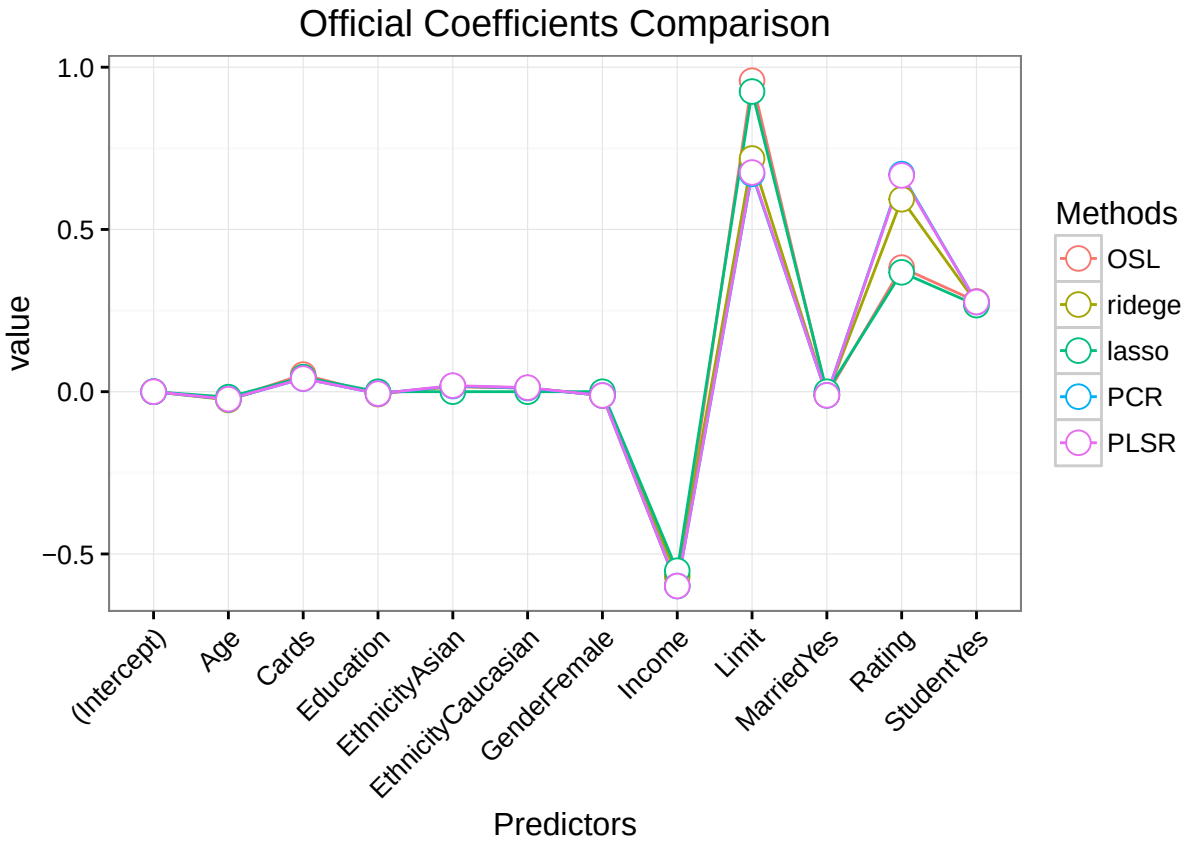
Table 7: Table of Test MSE for All Methods

	ridge	lasso	PCR	PLSR
MSE Value	0.045	0.047	0.046	0.047

Table 2 has only one row (Test MSE value) and four columns (one column per regression methods: ridge, lasso, pcr, and plsr).

From Table 2, the result shows that the model with lowest test Mean Square Error is Ridge Regression, which means that ridge regression acttually has the best performance when we test the prediction against the true value in testing set. So ridge regression is the best model in terms of measuring the fitness by MSE.

Now let's look at a plot in which the official coefficients are compared. We plot trend lines (i.e. the profiles) of the coefficients (one line connecting the coefficients of each method).



The graph displays comparison of coefficients of each predictor between different methods that we discussed earlier. In this visualization, we again confirm that there is some level of similarity between all methods.

Conclusions

After we explore the credit data, we are interested in learning relationship of Balance and the other 10 predictors. First, we have fitted a OLS model upon all of 10 predictors in credit data so that we could get more insight about the information hidden behind the data. And the summary statistics served as our benchmark. Next step is model selection. We will pick the best models according to test Mean Square Error. Two shrinkage methods (ridge regression and lasso regression) and two dimension reduction methods (principal component regression and partial least square regression). When we conducted our analysis, we would like to see test MSE improve each time. However, the test MSE from all 5 different methods are competitive with each other. Ridge regression method acchieved the lowest test MSE and is considered the best model among the other four methods.

References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning: With Applications in R. New York: Springer, 2013. Print.