Predictive Modeling

Jacqueline Liu, Steven Chen

November 4, 2016

Abstract

This paper will go through several different types of linear regression models that are good alternatives to the standard simple linear model $y = Ax + b$. The different regression models utilizes different fitting methods, as oppose to using Ordinary Least Squares (OLS) in the simple and multiple linear regression. A similar explanation can be found in Chapter 6, *Linear Model Selection and Regularization*, of the book An Introduction to Statistical Learning by James et al. After reading, one can easily follow the same steps and achieve the same results. However, our analysis will be slightly different due to our methods of preprocessing (see *Data* section).

Introduction

Previous we have explored different types of linear regression models, but all were utilizing Ordinary Least Squares method to fit the model. Now we want to try and see whether other fitting methods and models will give us better predictions. We will consider the following regression methods:

- Ridge Regression (RR)
- Lasso Regression (LR)
- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLSR)

We will perform analysis on the data set Credit. In each analysis, we will tune the parameter(s) for each respective model with 10 fold cross-validation and then evaluate the best models by using Minimum Square Errors (MSE) as our loss function on the test set. A more detailed explaination on the models is proved in the *Methods* section.

Data

The primary data set we are using is Credit, which contains 400 bank customers and their personal information. *Balance*, which states the current balance of the customer, is used as an dependent variable and the rest are used as predictor variables (or features) for our models. They consist of:

- *Income*: a quantitative variable stating customer's income
- *Limit*: a quantitative variable stating customer's credit limit
- *Rating*: a quantitative variable stating customer's credit rating
- *Cards*: a quantitative variable stating the number of cards that the customer has
- *Age*: a quantitative variable stating customer's age
- *Education*: a quantitative variable stating the number of years of education of the customer
- *Gender*: a qualitative variable stating customer's gender
- *Student*: a qualitative variable stating whether the customer is currently a student
- *Married*: a qualitative variable stating the customer's marital status
- *Ethnicity*: a qualitative variable stating customer's ethnicity

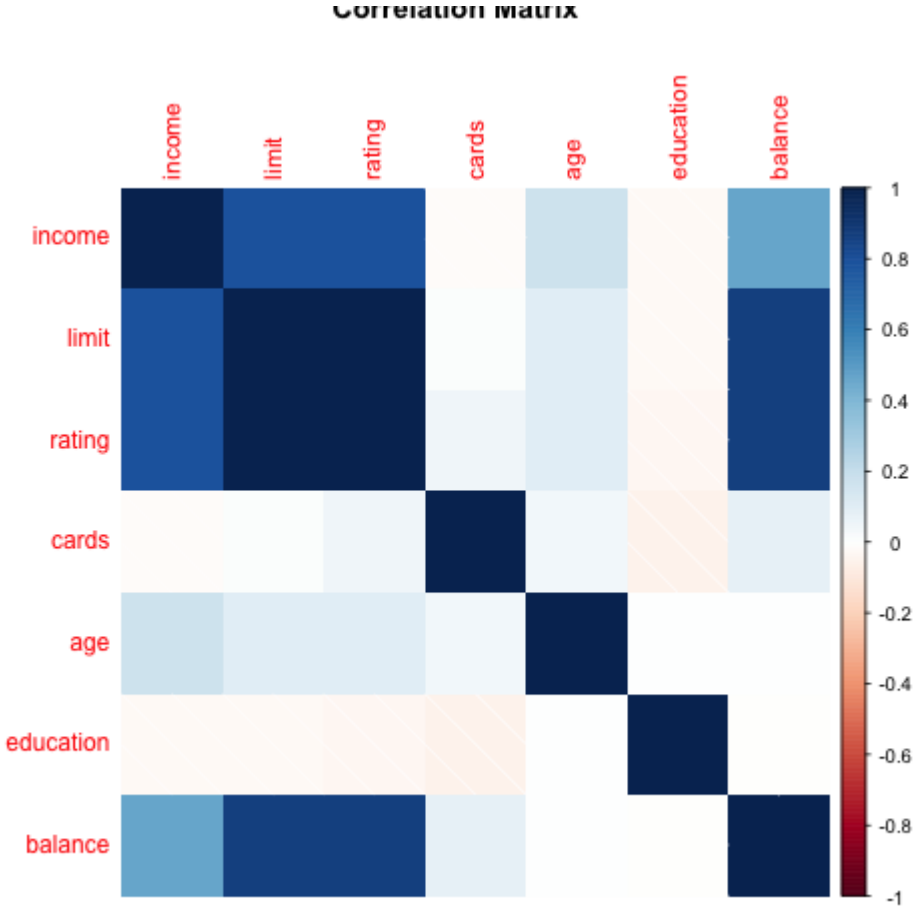

Figure 1: Correlation Matrix

We'll pick the best models with optimal parameters in predicting *Balance* using the above variables.

Before we get into the regression methods used, let's do a preliminary analysis on the correlations between the quantitative variables.

Based on the correlation visualization in Figure 1, we can see that *Income*, *Limit*, and *Rating* have strong correlations with each other. This might present a problem when we use these variables during linear regression, but they also have a strong correlation with *Balance*, suggesting these variables will be important for our predictions. On the other hand, we see there is little correlation between *Balance* and *Cards*, *Age*, and *Education*, so we expect these to have a smaller weight in our final models. We can demonstrate the same relationships with the following scatterplot matrix as well (see Figure 2).

Preprocessing

Our regression models require qualitative variables, so to include the quantitative columns, we binarized the values. For instance, the `Student` column takes on two values (*Yes* or *No*), so we created a binary column called `StudentYes` that takes on the value 1 if the customer is a student, and 0 otherwise. This was done via the `model.matrix` function in R. Because the features are all measured in different units, we also mean-center and standarize the columns so larger values aren't given undue weight in the regression coefficients.

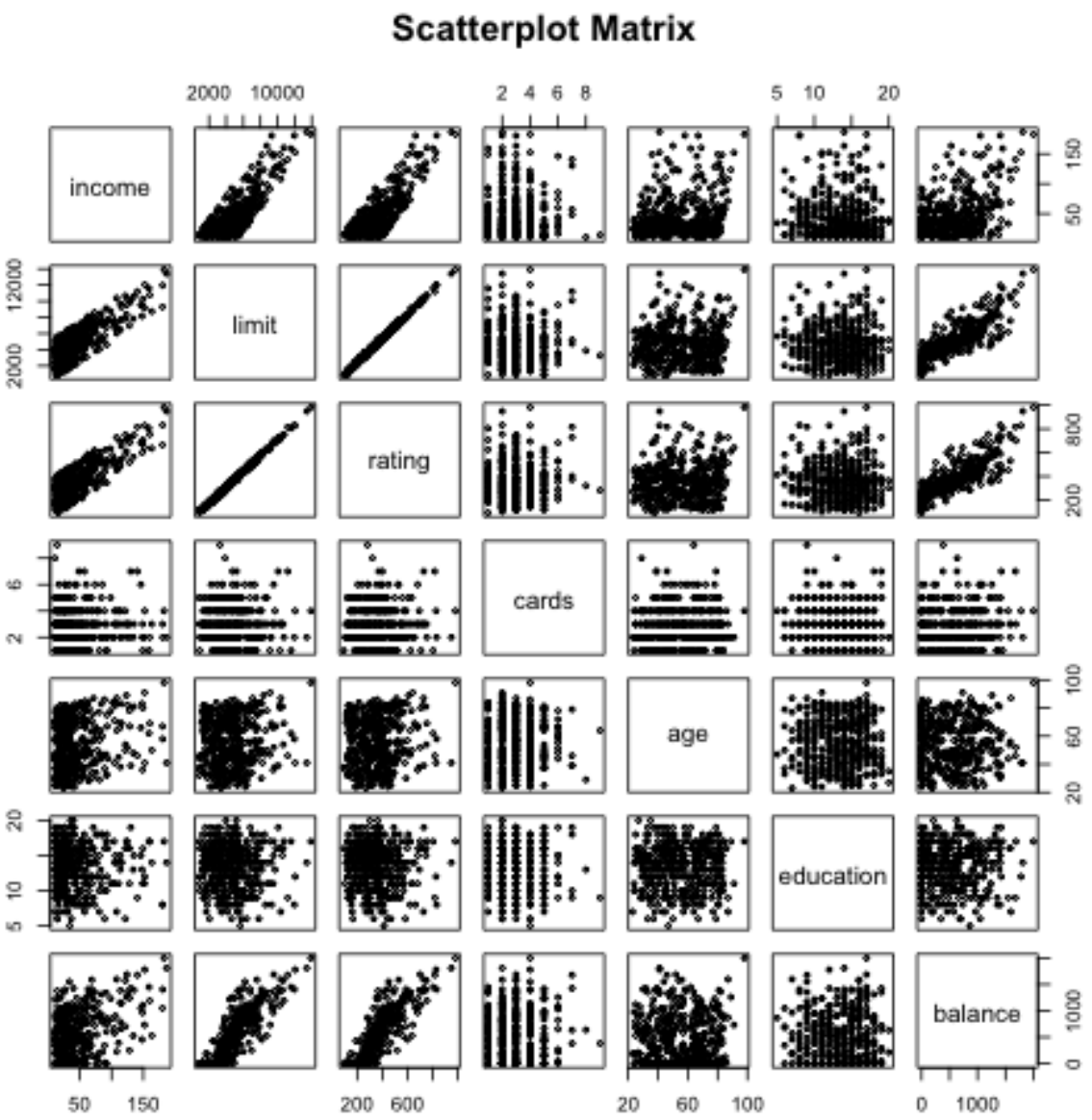

Figure 2: Scatterplot Matrix

Methods

Ordinary Least Square (OLS)

Previous, we have predominantly used OLS to estimate the coefficients for a linear model, and the goal is to find the set of coefficients that will minimize the sum of the squares of the predicted values and the actual values of the dependent variable. The formula we want to minimize can be express as follows:

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

Since this method is fundamental and unbiased, we will use this model as a benchmark to evaluate the other methods.

Shrinkage Methods

Shrinkage methods are used to mainly constrain the coefficients estimates of the model to prevent overfitting. By penalizing larger coefficients, the resulting model has a smaller coefficients and thuslower variance in predictions. As a consequence, however, the bias of the model increases so it is important to choose the shrinkage parameter carefully, usually through cross-validation. We will use *Ridge Regression* and *Lasso Regression* to see whether shrinkage will help improve our predictions.

Ridge Regression (RR)

Ridge regression is an example of shrinkage method applied to least squares regression. The formula we want to minimize can be express as follows:

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

where λ is a positive parameter controlling the effect of shrinkage. Higher value means there is a higher penalty for large coefficients and vice versa. For example, when $\lambda = 0$, there is no penalty on the coefficients, therefore it is regular OLS. As λ increases, the coefficients get closer to 0. Since it is a parameter, we will use cross-validation to tune an optimal value for λ.

Lasso Regression (LR)

Lasso Regression is another shrinkage method in least square regression. It can be expressed as:

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

λ has the same effect here as it does in ridge regression, penalizing larger coefficients to produce smaller values. However, unlike ridge, lasso encourages feature selection; the shape of the constraint regions tend to produce coefficients that are exactly zero. To learn more, read Chapter 6, Section 2.2 *The Lasso* in the *Introduction to Statistical Learning* book.

Dimension Reduction Methods

Dimension Reduction method uses OLS to fit the coefficients. However, instead of using the original predictors, it will first create a new vector of predictors from linear combinations of the original predictors. The purpose is to reduces the number of variables we are using as predictors in our regression, lowering the chance of overfitting and speeding up the computation of our models. There are a number of different ways to achieve dimensionality reduction, and here we will discuss *Principal Components Regression* and *Partial Least Squares Regression.*

Principal Components Regression (PCR)

Principal Components Regression uses Principal Components Analysis (PCA) as a *unsupervised* dimension reduction step prior to linear regression. Benefits of using PCA include being able to identify structures and relations, reducing the chance of overfit, and lowering dimensions. PCA transforms a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called *principal components.* The first principal component tries to account for as much variability in the data as possible, and all subsequent principal components have the highest variance possible given the constraint that they must be orthongonal to the preceding components. As a result, PCA produces a new set of predictors with size *less than or equal* to the original set, denoted as variable M, meaning that a smaller amount of predictors is enough to explain most of the variation in the data.

With lower dimension, we can lower the risk of overfitting our model. Since M will be the parameter of PCR, we will again use cross-validation to tune an optimal value for M.

Partial Least Square Regression (PLSR)

Like all dimensionality reduction methods, the benefits of PLSR include reducing the chance of overfitting and lowering the dimensionality of the dataset. Unlike PCR, however, it is a *supervised* learning method and uses that additional information to its advantage. Instead of choosing vectors that capture the most variablity in the features, it choose vectors that capture the most variability in both the features AND the response variable. The first PLS direction is chosen simply by taking the coefficients from simple linear regression, which are proportional to the correlations between the features and the response variable. The second PLS direction is then found by another linear re gression, this time on the residuals of the features regressed on the first direction.

PLSR is very useful when the features are highly collinear or when there are more features than observations. Its weight matrix reflects the covariance between the features and the response variables, while the weight matrix of PCA reflects the covariance between features themselves.

Analysis

Ridge Regression (RR)

We fit the training data using 10-fold cross-validation with many different values for the λ parameter of ridge regression, from 10^{10} to 10^{-10}. Figure 3 is the plot of Mean Square Error (MSE) and λ values during cross-validation. We choose the λ with the smallest associated error, which is 2.7825594×10^{-6}. We the calculate the MSE of the model with this λ using the test set, which is 0.0457619.

Lasso Regression (LR)

Like above, we use 10-fold cross-validation to fit lasso models with λ values from 10^{10} to 10^{-10} with the training data. Figure 4 is the plot of MSE against the tuning parameter values. The min error came from

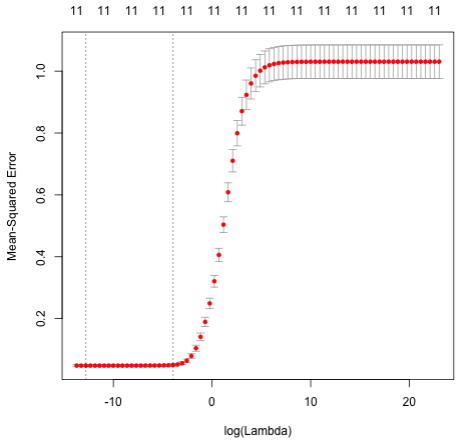

Figure 3: Ridge: MSE vs Shrinkage Param

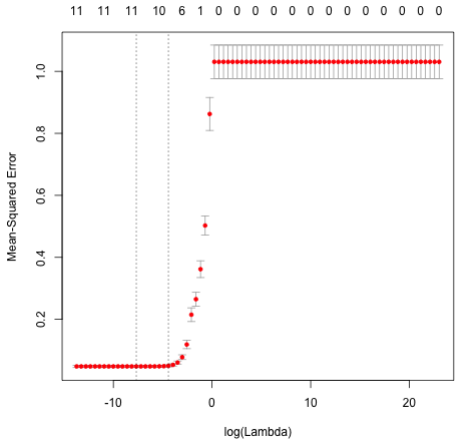

Figure 4: Lasso: MSE vs Shrinkage Param

the λ of 4.6415888×10^{-4}, producing a MSE on the test set of 0.0456889.

Principal Components Regression (PCR)

We fit the training data using cross-validation on the numbers of principal components. Figure 5 is the plot of Mean Square Error (MSE) of the predictions and number of components during cross-validation. Again, we choose the best number of components (resulting in minimum error), which is 11. We then calculated the MSE of the 11-component model using the test set, which turned out to be 0.0462546.

Partial Least square Regression (PLSR)

Again, we fit the training data using cross-validation on the numbers of components. Figure 6 is the plot of the MSE and number of components during cross-validation. The minimum error was achieved by 9 components. This resulted in a MSE 0.0463485 on the test set.

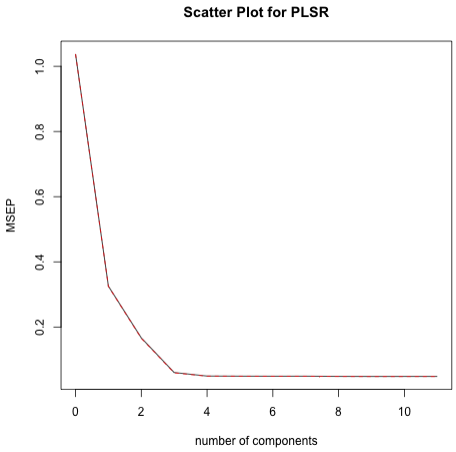

Figure 5: MSE vs Num Components

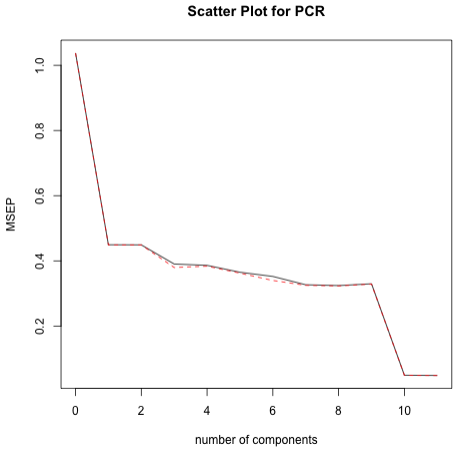

Figure 6: MSE vs Num Components

Results

After performing 10-fold cross-validation on all regression methods, the tuning parameters were chosen on the models with the smallest MSE. Ridge regression had a shrinkage paramater of 2.7825594×10^{-6}, while lasso used 4.6415888×10^{-4}. In principal components analysis, the best number of components was 11, while in partial least square regression, it was 9. After the models, with these optimal parameters, were fitted on the entire dataset, the coefficients in Table 1 were calculated.

	ols	ridge	lasso	pcr	plsr
(Intercept)	0.0000				
Age	-0.0230	-0.0230	-0.0231	-0.0230	-0.0234
Cards	0.0529	0.0560	0.0533	0.0529	0.0523
Education	-0.0075	-0.0078	-0.0075	-0.0075	-0.0076
EthnicityAsian	0.0160	0.0156	0.0159	0.0160	0.0159
EthnicityCaucasian	0.0110	0.0110	0.0110	0.0110	0.0111
GenderFemale	-0.0116	-0.0116	-0.0116	-0.0116	-0.0119
Income	-0.5982	-0.5983	-0.5970	-0.5982	-0.5981
Limit	0.9584	1.0312	0.9662	0.9584	0.9578
MarriedYes	-0.0091	-0.0085	-0.0090	-0.0091	-0.0086
Rating	0.3825	0.3099	0.3737	0.3825	0.3831
StudentYes	0.2782	0.2786	0.2780	0.2782	0.2782

Table 1: Coefficients of all regression models

Just looking at the raw numbers, the coefficients look very similar; a majority are the same or don't differ until the fourth decimal place. This is also seen in the Figure 7.

To compare the models, the mean-squared error for each are displayed in Table 2.

	MSE
ols	0.0448
ridge	0.0458
lasso	0.0457
pcr	0.0463
plsr	0.0463

Table 2: Mean-squared error of all regression models

Conclusions

Though the correlation matrix showed weak relationships between certain features and *Balance*, it seems the best model was still ordinary least squares, which conducts no shrinkage or dimensionality reduction. This is further emphasized in the results of the principal components regression, which choose to retain use 11 components, equal to the number of features. As mentioned earlier, the coefficients for all five methods were very similar, proving that shrinkage and dimensionality reduction were not significantly beneficial. This suggests that the original data was not highly collinear nor prone to overfitting. However, we must keep in mind that though the methods discussed and demonstrated in this report are very useful, they are limited by their constraint to linear relationships. In many datasets, higher order relationships are common; in that case, other models may be more appropriate. Linear regression models can still be used, but usually requires some sort of feature engineering or use of kernels with the dataset.

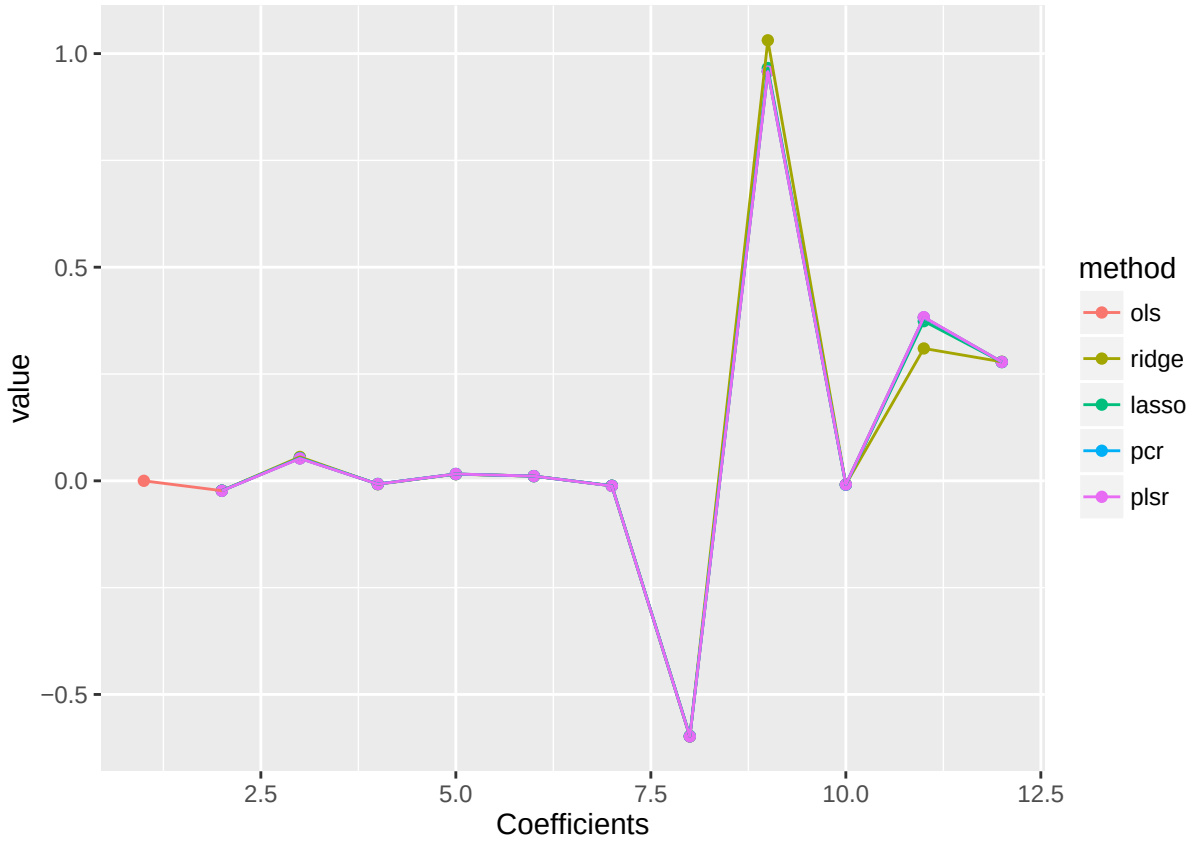

Figure 7: Coefficients by Regression Method