Predictive Modeling Process - Project Report

Author: Minsu Kim and Lingjie Qiao

Course: Statistics 159

Course Title: Reproducible and Collaborative Statistical Data Science

Instructor: Gaston Sanchez

Abstract

This paper summarizes the results of Stats 159 Reproducible and Collaborative Statistical Data Science Project 2: Predictive Modeling Process. After learning and applying multiple linear regression model via Ordinary Least Squares (OLS), we recognized some potential insufficiency of the ordinary least squares model and develop the idea of performing a predictive modeling process applied on the data set `Credit`. This project is largely based on the instructions provided in *Chapter 6: Linear Model Selection and Regularization* from book **"An Introduction to Statistical Learning"**.

The goal of this project is to present the use of predictive modeling process and utilize software tools that effectively communicate the results. We are therefore dedicated to maintain project reproducibility and provide both objective and personal reflections upon regression analysis.

---

1. Introduction

This project thoroughly explores the predictive modeling process. From previous study, in order to understand the relationship of one dependent variable with several independent variables, we fit a multiple linear regession with Ordinary Least Squares. However, since OLS may have high variance and include irrelevant variables, Predictive Modeling Process can improve the results in terms of Prediction Accuracy and Model Interpretability.

According to the book "An Introduction to Statistical Learning", alternative fitting procedures can yield better results in the following perspectives:

Prediction Accuracy: Provided that the true relationship between the response and the predictors is approximately linear, the least squares estimates will have low bias. If $n > p$—that is, if n, the number of observations, is much larger than p, the number of variables—then the least squares estimates tend to also have low variance, and hence will perform well on test observations. However, if n is not much larger than p, then there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training. And if $p > n$, then there is no longer a unique least squares coefficient estimate: the variance is infinite so the method cannot be used at all. By constraining or shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias. This can lead to substantial improvements in the accuracy with which we can predict the response for observations not used in model training. Model Interpretability: It is often the case that some or many of the variables used in a multiple regression model are in fact not associated with the response. Including such irrelevant variables leads to unnecessary complexity in the resulting model. By removing these variables—that is, by setting the corresponding coefficient estimates to zero—we can obtain a model that is more easily interpreted. Now least squares is extremely unlikely to yield any coefficient estimates that are exactly zero. In this chapter, we see some approaches for automatically performing feature selection or variable selection—that is, for excluding irrelevant variables from a multiple regression model.

The following analysis therefore utilizes four different kinds of regression models to find the best fitting model for predictive modeling process.

2. Data

We download the data set `Credit` from online link `http://www-bcf.usc.edu/~gareth/ISL/Credit.csv`, which is provided by the author of the book, "An Introduction to Statistical Learning". This data set records `Balance`, which is the average credit card debt for a number of individuals, as well as several predictors. The dataset has eleven variables - seven **quantitative** variables, `Income`, `Limit`, `Rating`, `Cards`, `Age`, `Education`, and `Balance`, and four **qualitative** variables, `Gender`, `Student`, `Married`, and `Ethnicity`. Our goal is to understand the relationship between `Balance` and these potential predictors with statistical fitting procedures.

2.1 Pre-modeling Data Processing

In order to fit the regression models, we first preprocess the dataset `Credit` with two steps: * convert factors into dummy variables - which avoids the problem of input data as factors * mean centering and standardization - which provides comparable scales for data analysis

---

3. Methods

In this section, we present the methods and procedure to build predive models and select the best model that predicts the target variable, balance. We fit and tune five different regression models namely OLS, Lasso, Ridge, PLSR and PCR.

3.1 Data preparation

We first consider to split the original dataset into train, validation and test. However, since modeling libraries internally support cross validation by splitting train into train and validation, we decide to split the original dataset into train and test data. It is important to hold out test data and fit models on train data because the held-out test data will later be used to measure the effectiveness of final models. If it is included in a training phase, the models overfit the data and bring about very optimistic MSE but do not generalize to future observations.

3.2 Evaluation

Since the goal of this analysis is to figure out which model works the best, we need to have an evaluation criterion to compare the effectiveness of each model. Thus, we choose a common evaluation criterion for regressiom model, Mean Square Error. To effectively calculate MSE for each model, we calculate MSE for each fold ten times using 10-fold cross-vaition and average them to obtain the averaged MSE for tuning parameters for each model. However, this process is automatically supported by libraries. Finally, we fit each model using test data and compare MSE from test data.

3.3 Model description and hypter-parameter tuning

While training each model, we need to find optimal paramters for each model. In order to effectively select hyper-parameters, we use 10-fold cross-validation. For lasso and ridge, lambda is the tuning parameter. It determines how much we will penalize models for high weights on predictors. If lambda is high, it penalizes models more and ends up generating sparse models. This kind of models is called shrinkage method because

they shrinken weights or even remove predictors by penalizing models. These models are especially good options when there are many predictors. By penalziing or removing unnecessary predictors, they provide more interpretable results. Thus, they are often utilized in genomic and pharametical analysis. For PCR and PLSR, the number of principal components is the tuning parameter. They both internally use the precipal component analysis to obtain pricipal components, which are linear combination of original predictors in dataset. Based on spectral theorem, the eigenvector corresponding to highest eigenvalue is the direction that explains the most about the variability in data, which is the first principal component. We need to find the optimal number of subsets of PCs that summarize the entire data set without harming accuracy. Again, we use 10-kold cross validation to obtain the optimal number of principal components.

3.4 Model comparison

As mentioned on 3.2 evaluation, we use MSE to compare models and select the best model. Although MSE does not provide an absolute means of model accuracy, it provides a relative measure to compare models. Thus, we finalize our model with the lowest MSE.

In summary, the following is the the procedure for each model.

1. Split the data into train and test, 80% and 20% repectively.
2. Train a model using train data with 10-fold cross-valdation.
3. Pick the optimal hyper-parameters.
4. Predict balance using the model with the optimal parameters.
5. Calculate Mean Square Error.
6. Record both Mean Square Error and coefficients.

4. Analysis

In this section, we present the result for each model by investigating hyper-parameters, coefficients and Mean Square Error. Lastly, we choose the best model based on Mean Square Error.

4.1 Baseline model

In order to set the baseline model, we first fit the ordinary least square regression using eleven predictors. Unlike other models, OLS does not have a tuning paramter. The p-values in coefficients indicate that Income, Limit, cards and StudentYes are statistically significant. Also, adjusted R-squared shows that this model explains well about the data.

4.2 Tuning parameter selection

As mentioned in methods section, we use 10-fold cross validation to tune hyper-paramters for each model.

Ridge regression penalizes predictors' weights by L2 norm. And lambda determines the magnitude of the penalty. Figure 1 shows that MSE increases as the lambda increments. Using cross validation, we finally obtain the minimum lambda that maximizes MSE, which is 0.01.

Lasso regression penalizes predictors' weights by L1 norm. Similar to Ridge regression, its lambda determines the magnitude of the penalty. Figure 2 shows that MSE increases significanly as lambda grows. Using cross validation, we finally obtain the minimum lambda that maximizes MSE, which is 0.01.

This is an interesting result in that both Ridge and Lasso have very small lambda. It means that both models end up penalizing a little bit. This makes sense because there are not many predictors and predictors are quite independant.

Principal Component Regression fits a linear regression on newly generated basis, principal components. The number of principal components is important. To obtain the best number, we use 10-fold cross validation and select one that minimizes RMSEP. Figure 3 shows that both ten and eleven PCs are very simliar and bring about the lowest RMSEP. So, we end up choosing ten pricipal components.

Partial least squares regression bears some relation to principal components regression. Instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. [1] Again, The

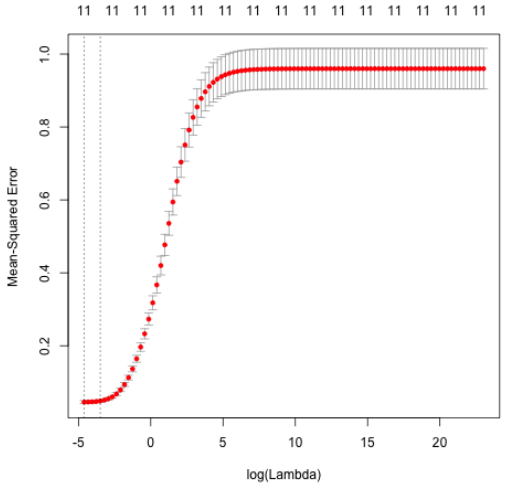


Figure 1: Lambda for Ridge

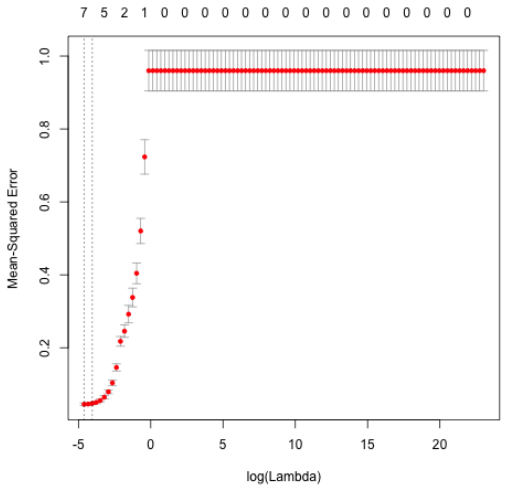


Figure 2: Lambda for Lasso

number of principal components is important. Similarly, to obtain the best number, we use 10-fold cross validation and select one that minimizes RMSEP. Figure 4 shows that four to eleven PCs bring out almost the same RMSEP. If our goal is to reduce dimensionality, we can select either four or five principal components. We end up choosing four.

4.3 Coefficients

Table 1 and Figure 5 show the coefficients for five different models. They show that cofficietns for predictors in five models are very simliar and there are slight differences in **Limit** and **Rating** predictors. It is noteworthy that there are a lot of zero weights in Lasso. So, Lasso effectively removes unnecessary predictors

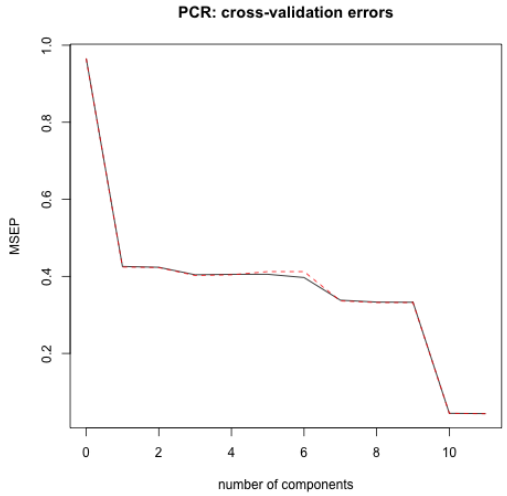


Figure 3: PCR cross validation

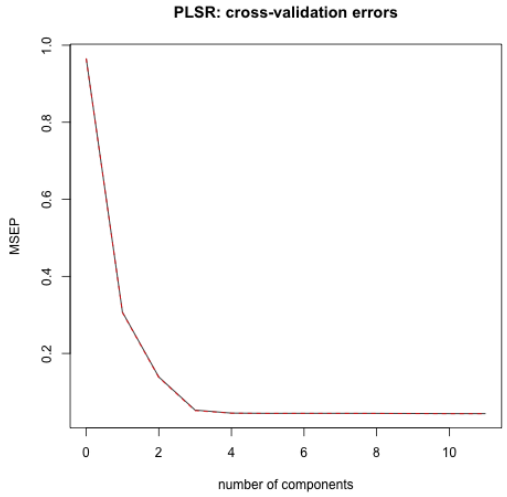
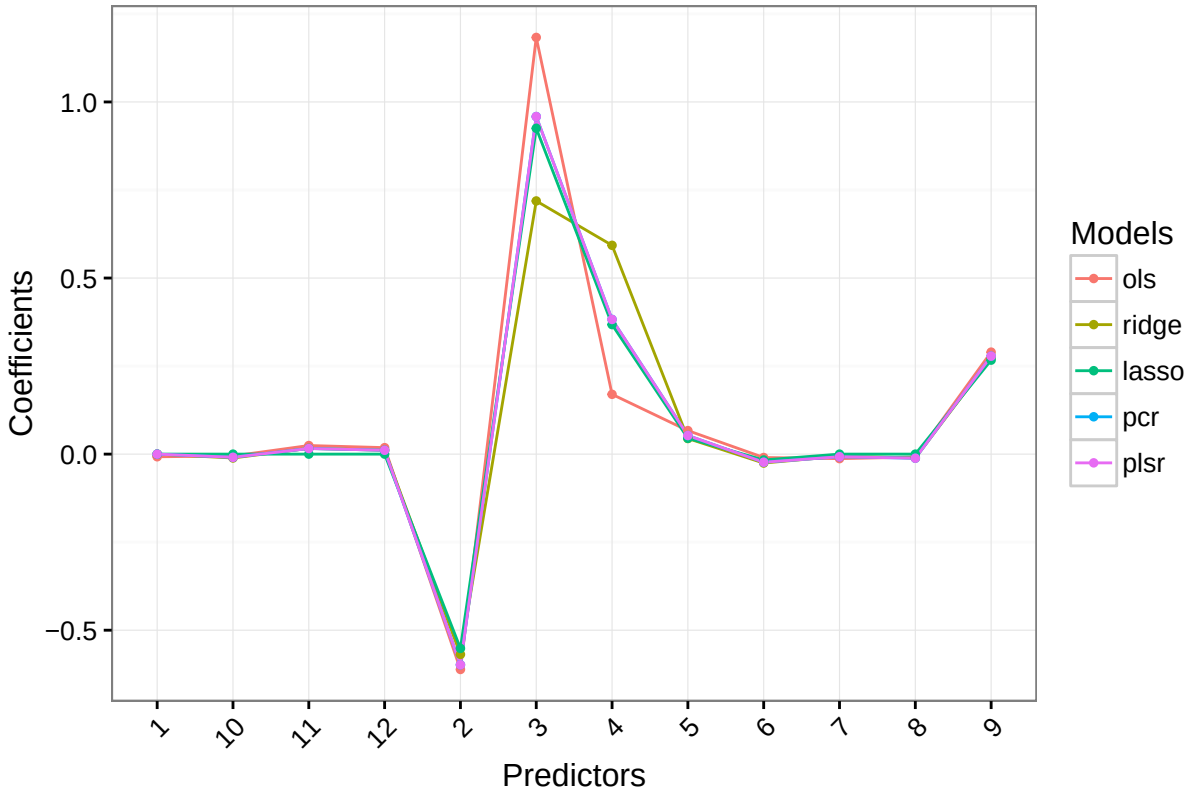


Figure 4: PLSR cross validation

Table 1: Coefficient Table

	ols	ridge	lasso	pcr	plsr
1	-0.01	0.00	0.00	0.00	0.00
2	-0.61	-0.57	-0.55	-0.60	-0.60
3	1.18	0.72	0.93	0.96	0.96
4	0.17	0.59	0.37	0.38	0.38
5	0.07	0.04	0.05	0.05	0.05
6	-0.01	-0.03	-0.02	-0.02	-0.02
7	-0.01	-0.01	0.00	-0.01	-0.01
8	-0.01	-0.01	0.00	-0.01	-0.01
9	0.29	0.27	0.27	0.28	0.28
10	-0.01	-0.01	0.00	-0.01	-0.01
11	0.02	0.02	0.00	0.02	0.02
12	0.02	0.01	0.00	0.01	0.01

Figure 5: Official Coefficients Comparison



4.4 Model Comparison and selection

Table 2 shows MSE for each model. This indicates that Ridge brings out the lowest MSE and Lasso and PCR give almost the same MSE and OLS gives the highest MSE. So, we decide to select Ridge regression as our best model.

Table 2: MSE Table

	mse
ols	0.065650
ridge	0.061220
lasso	0.062670
pcr	0.062090
plsr	0.063440

5. Results

5.1 Explanatory Data Analysis

In order to fully understand the data, we first obtain descriptive statistics and summaries of all variables in the `Credit.csv` data set. Some of the plots we obtained are displayed as following:

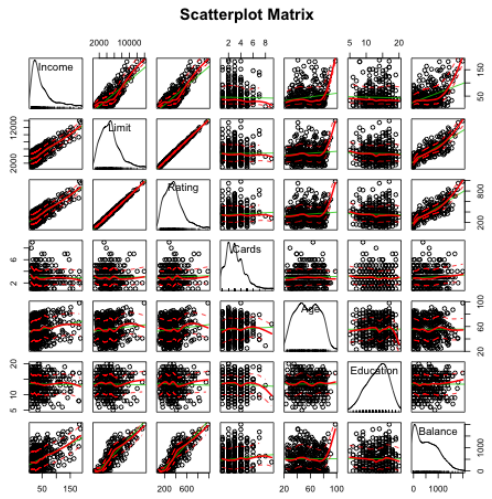


Figure 5: Scatterplot Matrix of all variables

For qualitative variables, we can also observe some conditional boxplot as displayed below:

6. Conclusion

In conclusions, we explore and compare the usage of different regression models on dataset `Credit` to understand the relationship between dependent variable `Balance` and ten potential predictors. Setting ordinary least squares as the benchmark, we look at two shrinkage regression methods (ridge and lasso) and two dimension reduction regression methods (PCR and PLSR) to find the best fitting model.

Reference

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

1. https://en.wikipedia.org/wiki/Partial_least_squares_regression

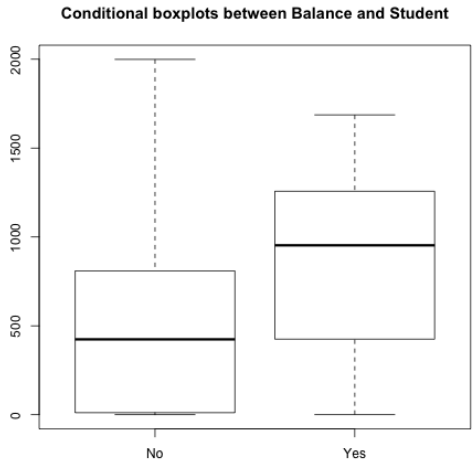


Figure 6: Conditional Doxplot on Student