

Predictive Modeling Process

Andrea Widjaja & Priscilla Hartono

11/04/2016

Abstract

This report is the second project of STAT 159, Reproducible and Collaborative Statistical Data Science, taught by Professor Gaston Sanchez in the fall of 2016 at UC Berkeley. STAT 159 focuses on the collaborative work and the ability to reproduce valuable data.

This paper consists of the six parts: abstract, introduction, data, methods, analysis, results, conclusions. We reproduced results from Chapter 6: Linear Model Selection and Regularization of **An Introduction to Statistical Learning* by Gareth James, Daniel Witten, Trevor Hastie, and Robert Tibshirani using *Credit.csv** dataset.

Introduction

So far, in this course, we have come across response Y and a set of variables X_1, X_2, \dots, X_p . We used the least squares model in a regression setting to observe their relationship. Continuing on from that, we want to approach more linear model framework to better yield predictive accuracy and model interpretability.

Prediction accuracy provides the true relationship between the response and the predictors. Are they approximately linear? Does the least squares estimates have a low bias? This depends on n as compared to p . If $n \gg p$, this will yield low variance, hence will produce good results. However, when this is not the case, there can be a lot of variability in the least squares fit, resulting in overfitting and poor predictions. It is even worse if $p > n$. There is no longer a unique least squares coefficient estimate since the variance is infinite.

Model interpretability removes irrelevant variables from a multiple regression model. This allows us to obtain a model that is easier to interpret as the coefficient estimate gets set to zero. However, in least squares, it is unlikely that it yields any coefficient estimate that is exactly zero. Hence, in this paper, we will discuss other models to fix this.

We will cover three methods:

- *Subset selection* This approach identifies predictors we believe are related to a response and fitting them into a model using least squares on the reduced set of variables.
- *Shrinkage*
This approach fits all predictors to a model. However, the estimated coefficients are shrunk towards zero relative to the least squares estimates. This will lead to a lower variance, which may even go until zero.
- *Dimension reduction*
This approach projects predictors into M -dimensional subspace where $M < p$. We do this by computing M linear combinations and use them as predictors to fit a linear regression model by least squares.

Data

For this project, the data set we will use is `__Credit.csv__`, which can be downloaded from the author of *An Introduction to Statistical Learning*'s website <http://www-bcf.usc.edu/~gareth/ISL/Credit.csv>.

The *Credit* data set records `_balance*` as the response variable with several quantitative and qualitative variables as the predictors.

The `*balance` response variable records the average credit card debt for `_400*` individuals according to the quantitative and qualitative predictors.

The qualitative variables are:

- *gender* (F/M)
- *student* (status: yes/no)
- *married* (status: yes/no)
- *ethnicity* (caucasian, african american, asian)

The quantitative variables are:

- *age*
- *cards* (number of credit cards)
- *education* (years of education)
- *income* (in thousands of dollars)
- *limit* (credit limit)
- *rating* (credit rating)

Methods

Goal: ways in which simple linear model can be improved by replacing plain least square fitting with alternative fitting procedures to get better prediction accuracy and model interpretability.

Ordinary Least Squares (OLS)

Ordinary least squares regression are typically known as the benchmark, in which it is a method that is used to estimate the unknown parameters in a linear regression model.

In the standard multiple linear model,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_p X_p + \epsilon$$

it is important to know that Y represents the response variable, X_i represents the i_{th} predictor variable, and β_i quantifies the association between the i_{th} predictor variable and the response. In other words, we can say that β_i is the average effect on Y for a unit increase in X_i , holding all other predictors fixed.

For our case, our linear model is in the form,

$$Balance = \beta_0 + \beta_1 x Income + \beta_2 x Limit + \beta_3 x Rating + \dots + \beta_p x Ethnicity + \epsilon$$

Estimating the Regression Coefficients/Parameters

From the multi-linear model above, we have to estimate the values of the regression coefficients or parameters $\beta_0, \beta_1, \beta_2, \dots, \beta_p$, because they are unknown.

Given the coefficient estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$

We can get the prediction for Y:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$$

and its residual (estimate of the error term) be:

$$\epsilon_i = y_i - \hat{y}_i$$

The values $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p$ that minimizes the RSS, where

$$RSS = \sum_i \epsilon_i^2$$

$$RSS = \sum_i (y_i - \hat{y}_i)^2$$

$$RSS = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \dots - \hat{\beta}_p x_{ip})^2$$

is used to estimate the multiple least square regression coefficients.

Subset Selection

Subset selection controls variance by using a subset of the original variables. From the previous homework, we have learnt the basics of subset selection.

There are 3 classical methods: * Forward stepwise selection - Starting with no variables in the model with only the intercept, we then proceed by adding a variable with the lowest RSS (greatest additional improvement). This is continued until some stopping rule is satisfied. * Backward stepwise selection - Starting with all variables in the model, we proceed to the next step by removing the variable with the largest p-value. We refit the model, and remove the variable with the largest p-value again. We stop when all p-values are below a certain threshold. * Mixed selection - Combining Forward selection and Backward selection, Mixed selection start with no variables in the model, and add variables that provide the best fit one-by-one. As new predictors are added, the p-value for one of the variables in the model can get larger. If it rises above a certain threshold, then we should remove that variable. Continue to do this until all variables in the model have a low p-value, and all variables outside the model have a large p-value if added to the model.

Cross-validation

In order to implement the methods, we can use validation and cross-validation, to determine which of the models are the best. Meaning that, it can be used to estimate the accuracy of the different models in order to choose the best one. Cross validation involves breaking data sets into training data and testing data to assess how the results of a statistical analysis can become an independent data set. It uses the training data to estimate the test MSE.

Shrinkage Methods

The Shrinkage Method approach involves fitting a model involving all p predictors instead of fitting a linear model that contains a subset of the predictors, which is discussed in the section above. The core concept behind the shrinkage method is constraining or regularizing the coefficient estimates of predictor variables towards zero. Having very minimal coefficient estimates results in a reduction of the variance.

There are two Shrinkage methods techniques:

1. Ridge Regression
2. The Lasso

Ridge regression (RR)

Ridge regression is very much alike to the least squares fitting procedure that minimizes

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

The difference between ridge regression and RSS is that the coefficients estimates $\hat{\beta}^R$ are the values that minimizes

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

where $\lambda \geq 0$ is called a tuning parameter.

Just like the least square fitting, by making the RSS small, we can get the estimates that fit the data well. However, the second part of the equation $\lambda \sum_{j=1}^p \beta_j^2$, which is called the shrinkage penalty. The shrinkage penalty get smaller as β_1, \dots, β_n gets to zero, giving an effect of shrinking the estimates of β_j towards zero. Unlike least square fitting, ridge regression will generate a different set of coefficient estimates for $\hat{\beta}_\lambda^R$ for each value of λ .

When $\lambda = 0$, then the ridge regression will be equivalent to the ordinary least square fitting. When λ grows larger, however, the impact of the shrinkage penalty will grow, and the ridge regression coefficient estimates will approach zero. Additionally, there will also be an decrease in the flexibility of ridge regression fit, which leads to decreased variance and increased bias. The theory behind this is rooted from the bias-variance trade-off.

Lasso regression (LR)

Lasso coefficient overcomes the one limitation of ridge regression: includes all the predictors in the final model. The lasso coefficients minimizes

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

The lasso also shrinks the coefficient estimates towards zero. However, the ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly zero when λ is large. Hence, the lasso performs variable selection, which generates a much easier to interpret model. Lasso also yields sparse models, meaning it involves only a subset of the variables.

When we perform the lasso, we are trying to find the set of coefficient estimates that lead to the smallest RSS, subject to how large $\sum_{j=1}^p |\beta_j|$ can be. It is clear that the lasso produces simpler and more interpretable model. However, it assumes that a number of the coefficients are truly equal to zero, resulting in prediction error. This shows that the ridge regression will perform better when the response is a function of many predictors with coefficients of roughly equal size while lasso regression yields a reduction in variance and can generate more accurate predictions.

Dimension Reduction Methods

From the subset selection method and shrinkage methods above, we have seen ways in which variance can be controlled by either using a subset of the original variables, or by shrinking the coefficients towards zero. Now, we are going to explore another approach that involves transforming the predictors, and fitting them to a least square model, called dimension reduction.

There are two Dimension Reduction methods:

1. Principal Components Regression
2. Partial Least Squares Regression

Principal Components regression (PCR)

Principal components regression is one of popular methods of dimension reduction. For a linear transformation, it involves transforming data from a high-dimensional space to a space of fewer dimensions. In other words, it derives a low-dimensional set from a large set of variables. The first step of principal components regression is to construct the first M principal components, Z_1, \dots, Z_M , and use the components as predictors for the linear regression model that is fit with the ordinary least squares. A smaller number of principal components, where $M < P$ is enough to explain the variability in the data and the relationship with the response. Using $M < P$ predictors is not a feature of subset selection because each of the M principal components used is a linear combination of all the p predictors.

Assuming that the directions of X_1, \dots, X_p shows that the most variations are the directions associated with Y , often times, although the assumption is not guaranteed, it turns out to be a reasonable approximation which gives good results. If the assumption is true, then we can say that fitting a least squares model to Z_1, \dots, Z_M produces better results than fitting a least square model to Z_1, \dots, Z_P . Additionally, estimating a smaller number of coefficients can mitigate overfitting.

Before performing PCR, it is important to standardize each predictors. This ensures that all the variables are on the same scale, and avoids the effect the roles high-variance variables play in PCR.

Partial Least Squares regression (PLSR)

Partial least squares is a supervised alternative to PCR. PLS is a dimension reduction method, which identifies a new set of features that are linear combinations of the original features, and then fits a linear model via least squares using these new features. But it also identifies these new features in a supervised way meaning it makes use of the response Y to identify not only the approximate of the old features but also are related to the response. The PLS approach finds directions that explain both the response and predictors.

PLS places the highest weight on the variables that are most strongly related to the response.

Analysis

Exploratory Data Analysis

When building any model, we want to first understand it. For this dataset, we obtain descriptive statistics and summaries of all variables.

Quantitative variables

We want to compute the following:

- minimum, maximum, range
- median, first and third quartiles, interquartile range
- mean, standard deviation
- histogram, boxplot

Qualitative variables

We want to compute the following:

- table of frequencies with frequency
- table of frequencies with relative frequency

- create barcharts
- matrix of correlation
- scatterplot matrix
- anova's between Balance and all qualitative variables (gender, ethnicity, student, married)
- conditional boxplots between Balance and all qualitative variables

Pre-modelling Data Process

After understanding the data, we now want to start fitting the model. But before we do that, we have to:

- convert factors into dummy variable
- mean centering and standardization

Dummy Out Categorical Variables

This is the process of transforming each categorical variable into dummy variable. We do this because the function we will use later, `glmnet()`, cannot process data that contains factors. So, we use `model.matrix()` to expand factors to a set of dummy variable: `temp_credit <- model.matrix(Balance ~ ., data = credit)`

If we do `model.matrix()` only, there will be a vector of ones in the first column that we do not want. So, we assemble the full data using `cbind()`: `new_credit <- cbind(temp_credit[, -1], Balance = Balance)`

Mean Centering and Standardizing

This is the process of making each variable have mean zero and standard deviation one. We do this because we want variables to have comparable scales: `scaled_credit <- scale(new_credit, center = TRUE, scale = TRUE)`.

We then save it as a .csv file and use this data for further analysis: `write.csv(scaled_credit, file = "path/of/processed/data/scaled-credit.csv")`

Training and Testing Sets

We now want to build and evaluate a model on the data. Two steps are required:

1. training: taking a random sample for model building
2. test: assess the performance and understand the model performance

The code to do this is `set.seed()`.

Results

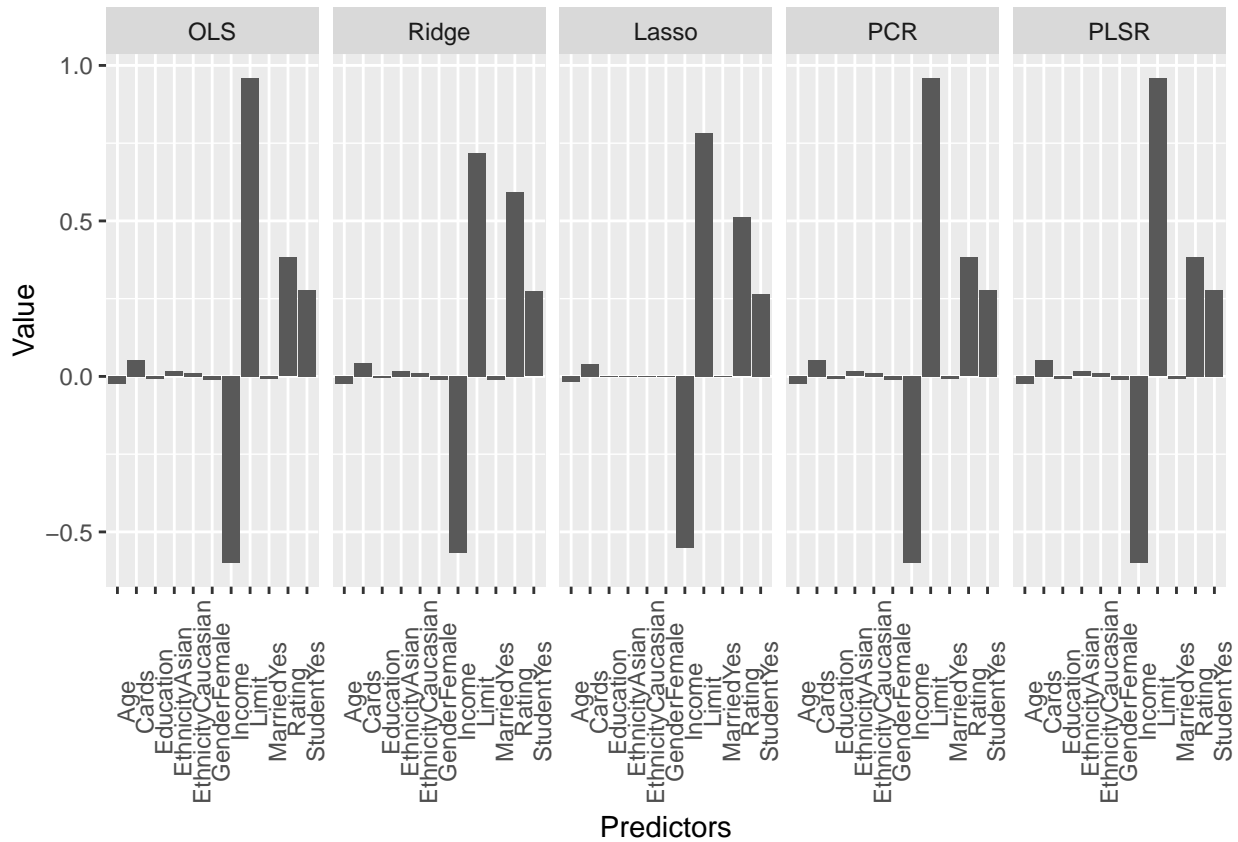
Coefficient Estimates

When conducting analysis, it is important to know the coefficient estimates for each model. Below are two representations of the regression coefficients needed to conduct analysis in table form and bargraph format.

From the table, we can see that OLS and PCR has similar coefficients, whereas Ridge, Lasso, and PLSR are similar.

Table 1: Coefficient Estimates for all models

	variables	OLS	Ridge	Lasso	PCR	PLSR
1	Income	-0.5982	-0.5687	-0.5514	-0.5982	-0.5981
2	Limit	0.9584	0.7187	0.7816	0.9584	0.9578
3	Rating	0.3825	0.5931	0.5111	0.3825	0.3831
4	Cards	0.0529	0.0443	0.0388	0.0529	0.0523
5	Age	-0.0230	-0.0254	-0.0168	-0.0230	-0.0234
6	Education	-0.0075	-0.0059	0.0000	-0.0075	-0.0076
7	GenderFemale	-0.0116	-0.0107	-0.0000	-0.0116	-0.0119
8	StudentYes	0.2782	0.2732	0.2661	0.2782	0.2782
9	MarriedYes	-0.0091	-0.0110	0.0000	-0.0091	-0.0086
10	EthnicityAsian	0.0160	0.0164	0.0000	0.0160	0.0159
11	EthnicityCaucasian	0.0110	0.0110	0.0000	0.0110	0.0111



From the bargraph, we can tell that there are 4 major variables, that significantly impact the regression for each models. These variables are: StudentYes, Rating, Income, and Limit.

Mean Squared Error

When looking at the training set mean squared errors for each model from the table produced below,

we can see that the best methods were PCR and PLS, yielding the lowest mean squared errors, at about 0.0397. Additionally, we can also tell that since the mean squared error of the OLS is largest at about 0.0448, it means that the OLS method is the worst among all the other methods. The second worst method is the Lasso method, in which it yields 0.0417 MSE, which is larger compared to Ridge method's MSE of 0.0412.

Table 2: MSE for all models test

	Model	MSE
1	OLS	0.0448
2	Ridge	0.0412
3	Lasso	0.0417
4	PCR	0.0397
5	PLS	0.0394

Conclusion

From the dataset Credit.csv, we were able to develop five different models to predict an individual's balance given variables `income`, `limit`, `rating`, `cards`, `age`, `education`, `gender`, `student`, `married`, and `ethnicity`.

The five models were ordinary least squares (OLS), ridge regression (RR), lasso regression (LR), principal components regression (PCR), and partial least squares regression (PLSR). Each of them had their pros and cons, and we were able to observe and evaluate each of these.