# Predictive Statistical Modeling Process

*Bret Hart and Yuyu Zhang*

*November 4, 2016*

### Abstract

The aim of this report is to perform a predictive modeling process applied on the data set named *Credit*. Specifically, the project performs exploratory data analysis, pre-modeling data processing and builds multiple linear regression models to predict the variable **Balance** in terms of ten predictors such as **Income**, **Age**, **Education**, **Gender**, **Ethinicity** etc. The project is largely based on chapter 6: *Linear Model Selection and Regularization* from the book **An Introduction to Statistical Learning** by James et al.

### Introduction

The overall goal of this collaborative project is to extend the linear model framework used in previous homework. We used the data set *Credit* as an example to discuss some ways in which the simple linear model can be improved, by replacing plain least squares fitting with some alternative fitting procedures. With alternative fitting procedures we expect to yield better prediction accuracy and model interpretability.

Two common alternatives are shrinkage and dimension reduction methods. Shrinkage includes Ridge and Lasso regression and dimension reduction includes Principle Components and Partial Least Square regressions, which will both be discussed more in details in the **Method** section.

In this report, we first performed exploratory data analysis, then used all four regression methods predict the variable **Balance** in terms of ten predictors such as **Income**, **Age**, **Education**, **Gender**, **Ethnicity** etc. Advantages and disadvantages of each model are also discussed. ####**Data** We will be working with the Credit dataset. This data set holds credit information of 400 different people for 10 different variables/predictors. 10 variables can be divided into two categories: *quantitative* and *qualitative*. Quantitative variables include Income, limit, Rating, Cards, Age and Education. Qualitative/categorical (only can take certain values) variables include Gender, Students(yes/no), Married(yes/no) and Ethnicity.

### Methods

Shrinkage methods means fitting a model involving all p predictors with the estimated coefficients shrunken towards zero relative to the least squares estimates, which has the effect of reducing variance.

For Ridge regression, we use coefficient estimates values that minimizes RSS with the following equation:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2, \qquad (6.5)$$

As for Lasso regression, we use the following equation that also shrinks the coefficient estimates towards zero:

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

Dimension reduction methods means projecting the p predictors into a M-dimensional subspace, where M<p. Then these M projections are used as predictors to fit a linear regression model by least squares.

For Principle Components regression is a technique reducing the dimension of a n × p data matrix X which involves identifying linear combinations, or directions, that best represent the predictors X1,. . . ,Xp. As for Partial Least square, it's a supervised alternative to PCR in that it attempts to find directions that help explain both the response and the predictors.

### Analysis

Thus, we fit the four different regression models (Ridge, Lasso, Principal Component, and Partial Least Squares) to the Credit data set, in attempt to find a model which best predicts future Balances based upon more readily observable predictors, which will in turn inform future decisions. We also fit a standard multiple linear regression model over the data set in order to create a "control" of sorts to compare our models to. If they cannot outperform multiple regression, they aren't very good representations of the data.

We begin by running the ridge and lasso regressions on the data. To do this, we use the package glmnet.

We carry out a basic 10-fold cross validation to make a set of ridge and lasso regression objects. Then, we find the 'best' lambda which minimizes the equations above. We find that for the ridge regression, our minimizing lambda is 0.0132194, and for our lasso regression, our minimizing lambda is 0.01. We then use these lambda values to create finalized ridge and lasso regressions across the whole data set.

For the Principal Component and Partial Least Squares Regression models, we follow a similar procedure, but use the package pls.

We again carry out a 10-fold cross validation to make a set of regression objects. This time, our "tuning parameter" is the number of components we desire to include in our regression models. This is determined by examining the minimum outputs of error over cross validation for each of the possible number of components (up to the number of predictors). For the analyses, we find that the best number of components is 11 for pcr and 11 for pls. We will touch more on why these numbers are interesting later in the paper.

Now, we have our 4 fitted models. Additionally, we have our standard multiple linear regression model to compare our results to. Let's see how our models do and what this tells us!

### Results

We have our 5 models. First, let's look at the coefficients our various models came up with and see how they compare. This should be interesting, as we're examining both shrinkage and dimension reduction methods - the lasso regression may have some coefficients that are completely 0. It should be noted that the pls package by default does not include intercepts in the pcr and pls regressions, but with a few slight tweaks of code the intercepts can be included as well.

Table 1: Coefficient table of the 5 models

| | OLS | Ridge | Lasso | PCR | PLS |
|---:|---:|---:|---:|---:|---:|
| (Intercept) | 0.00000 | 0.00000 | 0.00000 | 0.00000 | 0.00000 |
| Income | -0.59817 | -0.55945 | -0.55166 | -0.59817 | -0.59817 |
| Limit | 0.95844 | 0.70064 | 0.92505 | 0.95844 | 0.95844 |
| Rating | 0.38248 | 0.60175 | 0.36787 | 0.38248 | 0.38248 |
| Cards | 0.05286 | 0.04402 | 0.04500 | 0.05286 | 0.05286 |
| Age | -0.02303 | -0.02605 | -0.01666 | -0.02303 | -0.02303 |
| Education | -0.00747 | -0.00568 | 0.00000 | -0.00747 | -0.00747 |
| Gender.Male | 0.01159 | 0.01039 | 0.00000 | 0.01159 | 0.01159 |
| StudentYes | 0.27815 | 0.27198 | 0.26681 | 0.27815 | 0.27815 |
| MarriedYes | -0.00905 | -0.01122 | 0.00000 | -0.00905 | -0.00905 |
| EthnicityAsian | 0.01595 | 0.01622 | 0.00000 | 0.01595 | 0.01595 |
| EthnicityCaucasian | 0.01101 | 0.01098 | 0.00000 | 0.01101 | 0.01101 |

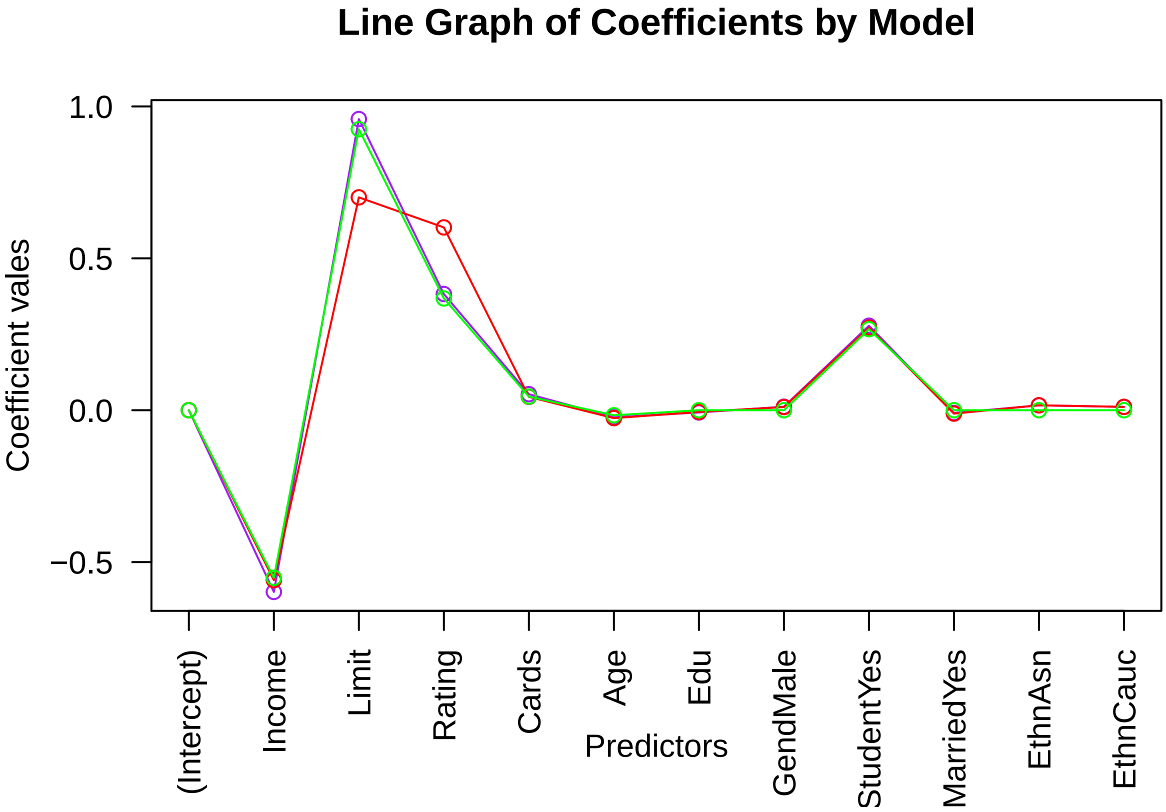
We see that the fuss over intercepts isn't too crucial, though, as they're all 0 (Not to imply that an NA value == 0, but it makes a lot of sense given that our data is standardized!) Overall, most of the coefficients are extremely similar across the models, while the lasso regression seems to have dropped many of the smaller ones.

A careful observer will note that PCR, PLS, and OLS are suspiciously similar - why might this be? Perhaps due to a faulty parameter minimization criterion, we selected the 'best' PCR and PLS models as those with 11 components. But, we have 11 predictors! Thus, the "components" for these 2 models simply were the predictors themselves, with no proxy representation by components at all! Maybe this data was particularly linear, or standardized data is less relevant for PCR/PLS analysis. I don't know enough to draw much from this result, but it is definitely interesting that for both PLS and PCR, the components that "best" represented our predictors were just the predictors themselves! In all likelihood, though, this was due to an erroneous metric of what a "best model" was in each of these shrinkage methods. We merely used the min of $validation$PRESS, and I quite frankly have not read enough of the pls documentation to conclude with confidence that this was a reductionist approach.

So, although we used the pls package and carried out an entirely different procedure than the simple lm() function for OLS, it seems that we ended up with 3 standard multiple linear regression models. Oh well! At least the lasso and ridge regressions are different!

To visualize the coefficients in a different way, we can plot their respective coefficient values on a line graph. We don't plot PCR and PLS as they are the same as OLS, so they would only serve to clutter up the visualization.

The Ridge Regression is Red, the OLS Purple, and the Lasso, Green.



We see that at most of the coefficients, all our models are incredibly similar. The ridge regression offers the only major deviations, at Limit and Rating. Additionally, some of the coefficients very close to 0 in the other two models are actually made 0 in the lasso model. But truthfully, having models so similar is not the most interesting result!

To actually compare our models, though, we can check the different MSEs that each model exhibited when checked against our test data set, as each model was fit on a smaller training subset to be able to test prediction strength in this very way! Even though our MSE for OLS, PCR, and PLS may turn out very similar, it's still worth examining the MSEs across the models. It is again worth noting that we only carried out a test/training set procedure over our non-OLS methods, so we have no OLS MSE as we simply fit the OLS to the entire data set.

Table 2: MSE table of the 4 advanced models

| | MSEs |
|---|---|
| Ridge | 0.04011 |
| Lasso | 0.04050 |
| PCR | 0.04123 |
| PLS | 0.04123 |

Logically, our PLS and PCR regressions give the same output. However, in an interesting turn, our ridge and lasso regressions, which actually were different from the OLS, actually perform better! Perhaps all this model fitting wasn't for naught! The MSE values are quite small all around due to the small values of the data that we're working with, but it the lasso and ridge still offer a noticable improvement in predictive power, as deemed by our test data!

### Conclusions

While the final results may not have been riveting, it's still worth noting that the ridge and lasso regressions definitely outperformed the basic OLS function in predictive power. We find the ridge to have the slightest edge over the other models, but not to a particularly extreme degree. The standardized credit data set may just be pretty linear - none of the models deviated very strongly from the basic OLS method. It may not be a juicy result, but it's still a worthwhile conclusion to draw from the data - with some extensive analysis, we can see something that we weren't able to see at the outset, i.e., that the 11 predictors have a somewhat linear role in predicting Balance. This, on its own, could be seen as a pretty profound conclusion!

A couple final thoughts - is there a better way to choose the number of components in plsr and pcr? I find it hard to believe, but possible, that under both regression methods there was no increase in strength at all over just the basic OLS method. Is choosing more components sort of like how R^2 increases proportionately with the number of predictors you have? It's hard to say, and I'm not confident enough myself to state anything too assuredly here. Additionally - why did the ridge regression have those 2 deviations in coefficient values from OLS and Lasso? I don't know for certain, but that slight alteration actually served to increase the quality of the model, so perhaps the lambda tuning parameter piece of the ridge regression function performed the best, mathematically, over this data set. At least the ridge can predict a little better than OLS! We performed something of actual value! Maybe, even in the real world, some things are simply pretty darn linear…