Predictive Modeling Process

Atul Lanka, Rushil Sheth

November 4th, 2016

Abstract

This project is based on the data set and statistical concepts introduced in Chapter 6, **Linear Model Selection and Regularization**, from "An Introduction to Statistical Learning" (by James et al.). The primary objective is to perform mulitple predictive modeling processes applied to the data set *Credit*, as well as learn about the relationship between `Balance` and the rest of the quantitative and qualitative variables. The five models considered are *Ordinary Least Squares*, *Ridge Regression (RR)*, *Lasso Regression (LR)*, *Principal Components Regression (PCR)*, and *Partial Least Squares REgression (PLSR)*. Using cross-validation, the most ideal model is evaluated from their respective Minimum Square Errors on the test sets.

Introduction

The primary purpose of this report is to determine the best model for predicting `Balance` given the ten different predictors, both quantitative and qualitative from the *Credit* data set. The distributions of these variables will be examined through summaries and plots, and five aforementioned regression models will be applied to the data (*Ridge Regression (RR)*, *Lasso Regression (LR)*, *Principal Components Regression (PCR)*, and *Partial Least Squares REgression (PLSR)*). The means of comparing and analyzing five models will be by studying their respective coefficients and mean squared errors (MSE) calculated from the 10-fold cross-validation.

We will discuss our data, methods, analysis, and our main conclusions throughout this paper. These sections will also include diagrams such as tables and graphs to help the reader gain a better understanding of our data and to visualize the outcomes of the different methods.# Data

We are examining the data set **Credit.csv**, which can be downloaded (here), from the book *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

The data set contains both qualitiative and quantitative variables. The qualitative variables are gender, student, status, and ethnicity. Additonally, the quantitative variables are balance, age, cards, education, income, limit, and rating.

The report is based on **Credit.csv** data set, but all of our analysis is done using **scaled-credit.csv**, which is a standardized version of the original credit data set. To create this new data set we perform mean centering and standardizing(mean zero and sd is one) because we want to have comparable scales to prevent favoritism for certain coefficients. We also scale our data because *glmnet()*, the function we will use for ridge and lasso regressions, does not take factors as data.

Once we perform analysis on **scaled-credit.csv** using our various regression methods we create various Rdata files with summary statistics for each respective method. These Rdata files are:

- OLS_regression.Rdata
- ridge-regression.Rdata
- lasso.Rdata
- pc-regression.Rdata
- plsr_regression.Rdata

We also create various text files containing summary statistics for each respective method. These text files are:

- ols_regression_output.txt
- ridge-regression-output.txt
- Lasso_regression.txt
- pc-regression-output.txt
- plsr_regression.txt

Finally, **train_test.Rdata** contains the training and testing data sets for the regression models. The training data contains 300 elements and the testing data contains 100 elements. The elements of each set are randomized and meant to be reproducible since we set the seed in the script, which produced the 2 sets. The different use of these two sets will be explained in the next section.# Methods

We built various linear models from the **Credit.csv** data set. We predicted the variable 'Balance' in terms of ten predictors: 'gender', 'student', 'status', and 'ethnicity', 'age', 'cards', 'education', 'income', 'limit', and 'rating'.

We created 5 different scripts for running various regression methods: Ordinary Least Squares, Ridge regression, Lasso Regression, Principal Components Regression, and Partial Least Squares Regression.

Our first method, was a simple linear model, Oridnary Least Squares. In **ols_regression.R**, using the lm() where the y variable is Balance and the x

variable is the combination of the the rest of the variables, the results from the OLS regression methods are output. This script's outputs will be used as a basis for all other regression models to compare to. We read in the scaled data set, which in turn is used for the regression model.

Our next two methods, Ridge and Lasso regression, are shrinkage methods. The script, **ridge-script.R**, performs a ridge regression and **lasso_regression.R** performs a lasso regression. The steps for these two methods are nearly identical:

1. Load **scaled_credit.csv**, and 'library(glmnet)' and set the seed, for reproducibility sake.
2. Create a 'x' and 'y' varible from the training set, read in from my Rdata file containing training and testing set indicies.
3. Run 'cv.glmnet()' which performs 10-fold cross-validation and outputs an intercept term and standardizes the variables by default. For the function arguements use the 'x' and 'y' from above, 'lambda = 10^seq(10, -2, length = 100)', 'intercept = FALSE', and 'standardize = FALSE' because our **scaled_credit.csv** is already standardized. For ridge regression we use 'alpha =0' and for lasso we use 'alpha =1'.
4. 'cv.glmnet()' will output a list of models. We decide the best one based of the minimum lambda and then save this lambda value as well as the coefficients associated with it.
5. Next we plot the model and save it to a png.
6. Once we identified the best model we use the **test_set** to calculate the test MSE, which will eventually help us compare the performances of all the models.
7. Finally we refit the model to the **scaled_data.csv** which is our entire data set using the lambda from step 4. We save the coefficient estimates and use it in the *Results* section of the report.

Our next two methods, Principal Components(PCR) and Partial Least Squares regression(PLSR) are dimension reduction methods performed by **pcr-script.R** and **plsr_regression.R** respectively. The steps for these two regression are very similar to those above, so naturall this outline will not go into as much detail.

1. Load **scaled_credit.csv**, and 'library(pls)' and set the seed, for reproducibility sake.
2. Create a 'x' and 'y' varible from the training set, read in from my Rdata file containing training and testing set indicies.
3. Run 'pcr()' or 'plsr()' depending on which model you want, and use arguements 'Balance ~ .', 'data=train_set', 'validation = CV' and 'scale = TRUE'.

4. We decide the best model using 'which.min(PCR/PLSR_model*validation*PRESS)' where 'MODEL' is the name of the pcr or plsr model depending on the script.

5. Next we plot the model and save it to a png.

6. Once we identified the best model we use the **test_set** to calculate the test MSE, which will eventually help us compare the performances of all the models.

7. Finally we refit the model to the **scaled_data.csv** which is our entire data set using the lambda from step 4. We save the coefficient estimates and use it in the *Results* section of the report. analysis....need to read the book for this results... what happened??

what can we take from this??