A Regression Comparison

Bryana Gutierrez and Erica Wong

November 4, 2016

Abstract

This project was created for the purpose of comparing different regression models. In this report we attempt to describe the differences between ridge, lasso, principal components, partial least squares, and ordinary least squares regressions. We base our analysis on Chapter 6 *Linear Model Selection and Regularization* of the book **An Introduction to Statistical Learning** by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. You can find a link to their book and data set by clicking on this sentence.. This analysis looks and financial and demographic information to fit a model for credit balance.

Introduction

In this project, we are looking at different methods of regression, which are ridge, lasso, principal components, partial least squares, and ordinary least squares. Ridge and lasso regression are shrinkage methods and principal components and partial least squares are reduction methods. The purpose of comparing the different types of methods is so we can improve the linear model in our studies. Using other methods can allow for more prediction accuracy and model interpretation.

Ordinary least squares is most effective when the number of observations is much larger than the number of variables in the model. This is because when there are many samples, the variance will be closer to the true variance by the law of large numbers. When variance is low, we will have a more accurate prediction when using the model. However, when our sample size is not significantly larger than the number of variables in our model, variance becomes a problem because that means our predictions from the model may be all over the place and we may have over-fit our model. Another huge problem is that if the sample size is less than the number of variables we have in the model then we cannot use the ordinary least squares method because then our variance will be infinite. By using other methods, we are able to increase prediction accuracy in these scenarios.

We also want to use different methods because in some models, some variables are not correlated with the response variable. using ordinary least squares, we will leave those variables in the model, which can cause things to become more complicated. Using other models will set the coefficient of those uncorrelated variables to zero and when doing our analysis, the process will be easier because there will be less variables to interpret and the model will be easier to understand because we know that each the variables in the model have some significance.

This is why we want to look into different methods such as ridge, lasso, principal component, and partial least squares. These methods all bring some value to our model, but in different ways. So, we want to compare all of them with each other and see how the model produced using each method differs from one another.

Data

In this report, we are using data that originated from `Credit.csv`. In this data set, there are qualitative and quantitative variables. The qualitative variables are labeled as gender, student, status, and ethnicity. These variables are one's gender, if one is a student, if one is married, and one's ethnicity respectively. The quantitative variables in this are labeled as age, cards, education, income, limit, rating, and balance. These

variables are one's age, the number of credit cards one has, years of education, income measured in thousands of dollars, credit limit, credit rating, and one's average credit card debt respectively.

Specifically, when doing our analysis, we are using a data set that we made from `Credit.csv` called `scaled-credit.csv`. In `scaled-credit.csv`, all of the variables are the same as those in `Credit.csv`. However, we converted factors (qualitative variables) into dummy variables, centered the mean, and standardized all of the data. Since, we standardized the data, this makes our data more comparable because all of our variables now have comparable scales. This is really important because our β will be different depending on the scale that the variable is measured in. By centering and standardizing, we will not favor any coefficient. This is why we want to use the `scaled-credit.csv` instead of `Credit.csv` for our regression analysis.

Methods

Ordinary Least Squares Regression

We first perform multiple linear regression analysis on our `scaled_credit.csv` data. We use the linear model

$$y \approx \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \cdots + \beta_1 1 * x_1 1$$

to describe the relationship between `Balance` and financial and demographic information represented in the $x_i s$. Therefore, the linear model looks more like this:

$$Sales \approx \beta_0 + \beta_1 * Income + \beta_2 * Limit + \cdots + \beta_1 1 * EthnicityCaucasian$$

where β_0 is the intercept term and the $\beta_i s$ describe how each financial or demographic variable affects the sales.

$$\hat{\beta} = \begin{bmatrix} \hat{\beta_0} \\ \hat{\beta_1} \\ \hat{\beta_2} \\ \vdots \\ \hat{\beta_1 1} \end{bmatrix}$$

is the least squares estimate of β which contains the actual values of the $\beta_i s$. By the Gauss-Markov Theorem they are the best linear unbiased estimators. They are estimated by minimizing the sum of the residual squared errors (RSS):

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

where e_i is equal to $y_i - \hat{y_i}$. $\hat{y_i}$ is calculated by using the model and $\hat{\beta}$:

$$\hat{y_i} = X\hat{\beta}$$

where

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,11} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,11} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{400,1} & x_{400,2} & \dots & x_{400,11} \end{bmatrix}$$

$\hat{y_i}$ is the predicted y value. In terms of this analysis, $\hat{y_i}$ is the amount of predicted balance based off of the all the different predictor variables. Basically, minimizing the RSS would be minimizing the error of the prediction.

RSS can also be written as:

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta_0} - \hat{\beta_1} * x_{i,1} - \hat{\beta_2} * x_{i,2} - \cdots - \hat{\beta_1 1} * x_{i,11})$$

Minimizing this value over the $\hat{\beta}_i s$ results in

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

where Y is a vector with all the y values. Using the `Advertising.csv` data we replace the $y_i s$ with the `Balance` numbers, the $x_{i,1}$ with the `Income` numbers, the $x_{i,2}$ with the `Limit` numbers, and so on.

Ridge Regression

Ridge regression is a shrinkage method, which means that it will constrain/regularize the coefficient estimates. This effectively will cause the coefficient estimates to tend toward zero because it can help to reduce the variance. Ridge regression is very similar to ordinary least squares, but difference is in the way that the coefficients are estimated. In ridge regression, we are minimizing

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

whereas in ordinary least squares, we are minimizing

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

. From the two types of regression, you can see that in ridge regression, there is a term that ordinary least squares does not, $\lambda \sum_{j=1}^{p} \beta_j^2$, and this will shrink estimates of β_j towards zero because $\lambda \sum_{j=1}^{p} \beta_j^2$ is smallest when all of the betas are close to zero. λ when greater than equal to zero is known as the tuning parameter. The purpose of the turning parameter is to control the impact of RSS and $\lambda \sum_{j=1}^{p} \beta_j^2$ on the regression coefficients.When the tuning parameter is large, the ridge regression coefficients will tend towards zero and when the tuning parameter is equal to zero, ridge and least squares will give the same coefficient estimates.

Ridge regression is useful because of the bias-variance trade-off. As λ increases flexibility of ridge regression decreases, so the variance of a model will decrease as well. However, this will increase the bias. Ridge regression gets around this because when λ increases, there will be a large decrease in variance, but only a slight increase in bias. So, mean squared error (MSE) will only increase slightly instead of a lot because MSE takes into account bias. This is why, we want to look at our results comparing ridge regression coefficients and ordinary least squares coefficients.

Lasso Regression

Lasso regression is very similar to ridge regression when it comes to the process. However, in lasso regression we minimize

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

. This is called a shrinkage method because we are putting a limit to the "size" of the β coefficients. Unlike in ridge regression, this method of minimization could result in estimations of the β_j as exactly zero. This means Lasso regression performs a variable selection and only fits a model to a subset of the data. In order to calculate the best λ for the minimization, we used ten-fold cross validation. Cross validation involves partitioning the data into subsets, fitting a model with one subset, and validating it on the the other. We select a λ based off of the validation with the least amount of error. When we ran our cross validation on the `scaled-credit.csv` data set, we came up with a lambda of .

Principal Components Regression

Principal components regression (PCR) is a dimension reduction method. What this means is that we are transforming the predictor variables and then fitting a least squares regression model to it. The idea behind PCR is that there are a small number of "principal components" within all of the variables that can explain the variability in the data as well as the relationship with the response variable. This model is better than using ordinary least squares because if we are able to find the variables that most closely relate to the variability and response variable, we can avoid over-fitting the model and our model will not pick up any unnecessary variability.

Under PCR, if as the number of variables in the model increases, variance also increases, but bias decreases. PCR tends to fit the model better when the first couple of principal components capture most of the variability and explain most of the response. It is also crucial to standardize our variables because the scale of each variable may effect variance and thus affect the fit of the model produced in PCR.

Partial Least Squares Regression

Partial least squares regression, PLSR is an extension of principal components regression. It also works at finding the best linear combinations, but takes into account both the response and predictor variables. The idea behind this is to find a set of data Z_i that would create a better fit for the data than the original data would be. These Z_i are calculated using the scaled data from `scaled-data.csv`. The process of calculating these values is iterative, so we begin with the first valueZ_1. Mathematically, the formula is

$$Z_1 = \sum_{j=1}^{p} \phi_{j1} X_j$$

where X_j is the jth explanatory variable and ϕ_{j1} is the coefficient when you perform simple linear regression of Y onto X_j. This process gives more weight to values that have more of a relationship with the response variable. Then to calculate Z_2 you regress each of the variables on Z_1. The residuals from these predictions will be the new set variables, $\tilde{X}_i$ on which we will perform the same process as for the X_i. So we will say

$$Z_2 = \sum_{j=1}^{p} \phi_{j2} \tilde{X}_j$$

where ϕ_{j2} is the coefficient from simple linear regression of Y onto $\tilde{X}_j$. This process in continued M times. M is chosen in the cross validation step. Then once we have $Z_1, Z_2, \ldots, Z_M$ we now fit a linear model in the same way as in PCR.

Analysis

Ordinary Least Squares Regression

The coefficients for Ordinary Least Squares are found in the *Regression Coefficients* Table. Although the Gauss-Markov Theorem states that these are the "best" estimates since they have the least variance for unbiased linear estimators, sometimes you can achieve less variance if you are okay with biased estimates. The rest of the regressions used in this report are biased, but the trade off in variance might be worth the bias. OLS yields an MSE of 0.0517916

Ridge Regression

When doing ridge regression, we started by looking at a ten-fold cross-validation. From cross validation, we were able to find the best model, which included finding our λ or tuning variable. In the plot *MSE Plot of Ridge Regression*, it shows the relationship between MSE and $\log(\lambda)$.

	OLS	Ridge	Lasso	PC	PLS
Intercept	-0.0012	0.0000	0.0000	0.0000	0.0000
Income	-0.5964	-0.5687	-0.5517	-0.5989	-0.5989
Limit	0.9545	0.7187	0.9250	0.6714	0.6849
Rating	0.3831	0.5931	0.3679	0.6706	0.6570
Cards	0.0510	0.0443	0.0450	0.0404	0.0414
Age	-0.0259	-0.0254	-0.0167	-0.0233	-0.0228
Education	-0.0147	-0.0059	0.0000	-0.0060	-0.0051
GenderFemale	-0.0220	0.0107	0.0000	0.0116	0.0124
StudentYes	0.2761	0.2732	0.2668	0.2764	0.2770
MarriedYes	-0.0171	-0.0110	0.0000	-0.0112	-0.0100
EthnicityAsian	0.0055	0.0164	0.0000	0.0174	0.0146
EthnicityCaucasian	-0.0047	0.0110	0.0000	0.0112	0.0087

Table 1: Regression Coefficients

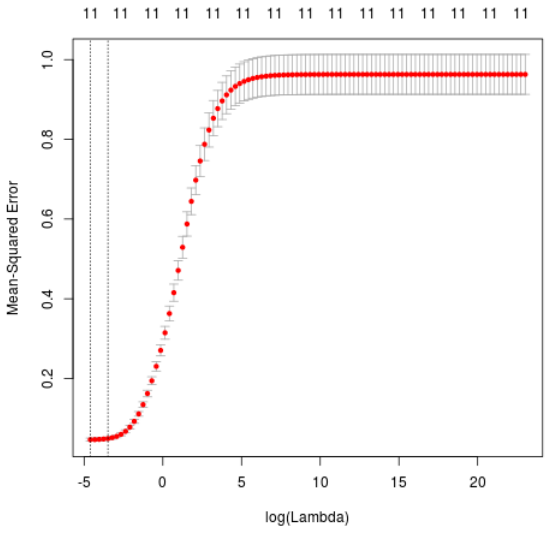

Figure 1: MSE Plot of Ridge Regression

From running our cross-validation on the train data set, we find that $\lambda = 0.01$. When comparing the coefficients of ridge regression and ordinary least squares, we find that all of the ridge regression coefficients are very similar to ordinary least squares except for the Rating coefficient. In ridge, the coefficient is 0.7186579 and in ordinary least squares, the coefficient is -0.0259231. Other than this difference being rather larger, the other coefficients tend to be smaller than that of ordinary least squares. Finally, when comparing the MSEs between the two methods, we found that the MSE of ridge regression is larger.

When doing PCR, we started by using a ten-fold cross-validation. From the cross validation, we were able to find the best model which was 10. When comparing coefficients between PCR and ordinary least squares, we find that they are very similar to one another. However, PCR's coefficients tends to be smaller than that of ordinary least squares. Additionally, the MSE of PC is larger than that of ordinary least squares.

Lasso Regression

In lasso regression we found a λ that capped the coefficients. The right lambda was chosen based off the MSE, the mean squared error. The *MSE Plot of Lasso Regression* plot shows the relationship between MSE

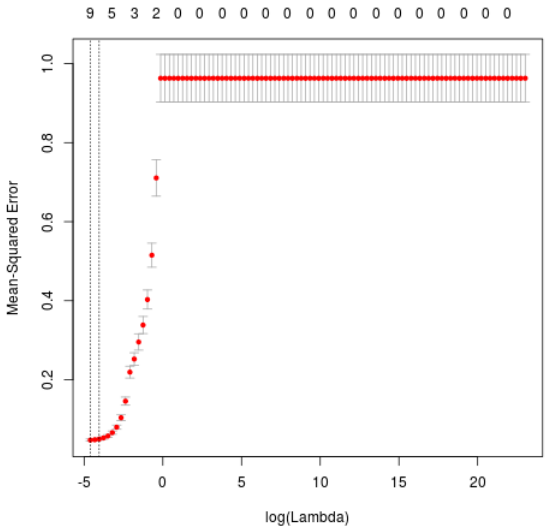

Figure 2: MSE Plot of Lasso Regression

and the log of lambda.

We use the λ that gives the smallest MSE. This is given by the left-most value on this graph. We got this analysis from the training data set, which helped us fit the best model.

In order to test the model, we used the testing data set. We fit the model using the λ from above and calculated the MSE. This will test how accurate of a fit the model is. When we did this, we got an MSE of 0.0515445.

Then using the full data set, we came up the following coefficients in the *Regression Coefficients* Table. Although we got the λ from the training data set, we got these coefficients from the entire data set, `scaled-credit.csv`. Some of the coefficients have been set to zero by the lasso regression analysis. As mentioned earlier, lasso regression has a dimension reduction component and will only fit the data to the variables that fit the MSE criteria. Our analysis shows six beta coefficients that have been set to zero. The rest of the lasso coefficients also tend to be smaller than those from OLS. This is due to the added restriction of minimizing $\lambda \sum_{j=1}^{p} |\beta_j|$. In addition, the MSE of lasso regression is smaller than that of OLS.

Principal Components Regression

When doing PCR, we started by using a ten-fold cross-validation. In the plot *MSEP Plot of Principal Components Regression*, we see the relationship between MSEP (mean squared error of predictions) and the number of components.

From the cross validation, we were able to find the best model which was 10. When comparing coefficients between PCR and ordinary least squares, we find that they are very similar to one another, except in Limit and Rating where there is a larger difference. When looking at PCR's coefficients, we notice that PCR's coefficients were most similar to that of ridge regression and PLSR. Overall PCR's coefficients tends to be smaller than that of ordinary least squares. Finally, the MSE of PC is 0.0519968 and is larger than that of ordinary least squares.

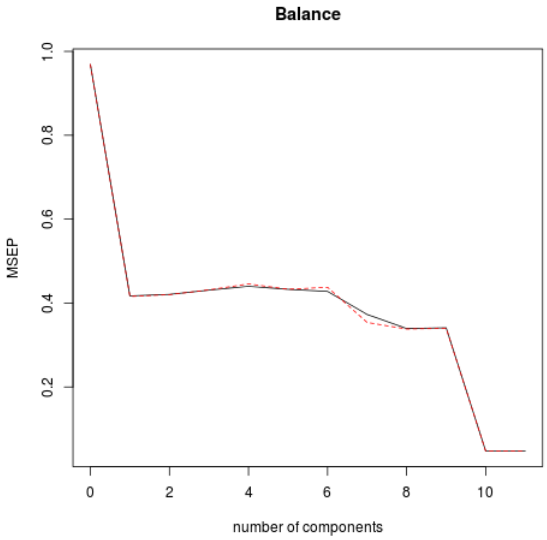

Figure 3: MSEP Plot of Principal Components Regression

Partial Least Squares Regression

In PLS regression, we have to find right M, which is the number of components to be used. This is chosen based off the MSE, the mean squared error. The *MSEP Plot of Partial Least Squares* plot shows the relationship between MSEP (mean squared error of predictions) and the number of components.

We use the M that gives the smallest MSEP/MSE. We received this minimization from cross validation. We got this analysis from the training data set, which helped us fit the best model.

In order to test the model, we used the testing data set. We fit the model using the M from above and calculated the MSE. This will test how accurate of a fit the model is. When we did this, we got an MSE of 0.0519598.

Then using the full data set, we came up the coefficients in the *Regression Coefficients* Table. The coefficients for partial least squares tend to vary the most from the OLS coefficients. For variables such as `Limit`, the PLS coefficient is smaller, but for variables such as `Rating`, the PLS coefficient is larger than that of OLS. The MSE for PLS is slightly larger that than of OLS as seen in the following section.

Results

The plot `Estimated Regression Coefficients by Variable and Regression` compares the coefficients of each variable. This representation allows us to compare how each type of regression fits a a model to the data. For instance we can clearly see that all but OLS regression has an intercept of exactly zero. Also, we can see that Lasso regression results in some of the coefficients being exactly zero. We can also see that some variables result in estimates that are fairly the same regardless of the type of regression like `Income` and `StudentYes`. While some variables like `Rating` and `Limit` have coefficient estimates that vary widely by different regression types.

When looking at the table called Test MSE Values for the Regression Techniques, we found that all of the mean squared errors (MSE) are very close to one another with the maximum difference being 0.0010482. We found of all the regressions the one with the smallest MSE is lasso regression, where MSE = 0.0515445. Since lasso regression has the smallest MSE, this means that it provides the best fit model for the credit data set. On the other hand, the one with the largest MSE is Ridge Regression, where MSE = 0.0525927. This means

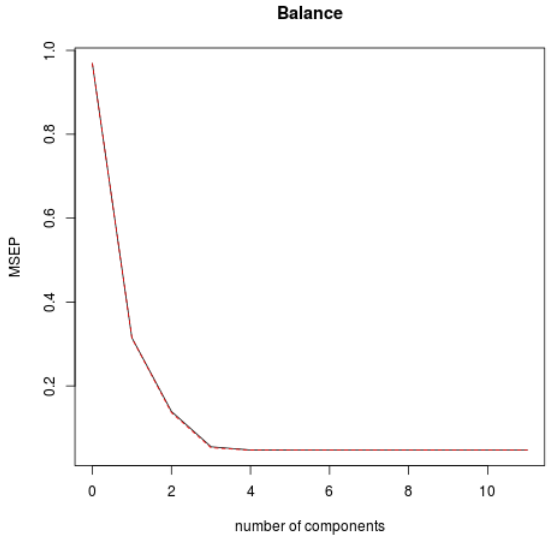

Figure 4: MSEP Plot of Partial Least Squares Regression

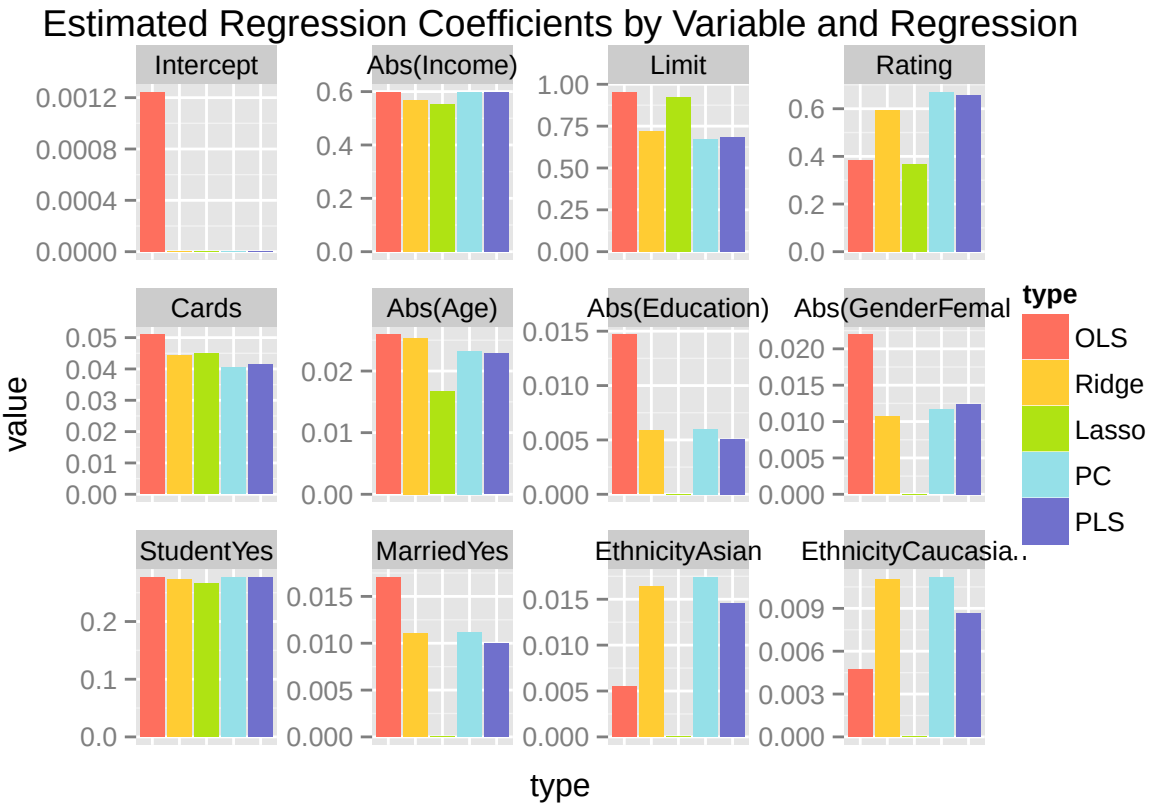

Figure 5: Estimated Regression Coefficients by Variable and Regression

Table 2: Test MSE Values for the Regression Techniques

Regression	MSE
OLS	0.0517916
Ridge	0.0525927
Lasso	0.0515445
PCR	0.0519968
PLSR	0.0519598

that ridge regression provides the worst fit model of the regressions that we looked at for the Credit data set. Finally, looking at the other MSEs, we noticed that PCR and PLSR had the most similar MSEs with one another.

Conclusion

In this project, we learn why it is important to try different models. From looking at the Credit data set, we found that while OLS gives the best unbiased model, we found an even better model using lasso regression. While lasso regression increases bias, it decreases variability. With a smaller variance, the predictions are closer to the true model meaning that this may be a better model.

Additionally, with ridge regression, lasso regression, PCR, and PLSR, we fit models that will eliminate/narrow down the variables so that only impactful and correlated variables are taken into account. This is important because eliminating some of the predictor variables can cause for more meaningful analysis. For example, if there was only a certain amount of time that was allotted for analysis, by eliminating variables that do not have a correlation with the response variable, we have more time to do meaningful analysis on the variables. Figuring out which variables have an impact on the response variable can be helpful for others who may use our data in the future and do analysis there.