

Statistics 159 Project 2 Report

Aoyi Shan, Yukun He

10/28/2016

Abstract

In this project, we explore the difference between various multiple regression methods and conclude on what approach should be used in order to fit our data better.

Introduction

In this project, our objective is to apply model selection methods introduced in Chapter 6, Linear Model Selection and Regularization, from the book “An Introduction to Statistical Learning” by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. In our analysis, we compare the fit of models generated by 5 different regression methods, regression model obtained by Ordinary Least Squares, Ridge regression, Lasso regression, Principal Components regression and Partial Least Squares regression. To evaluate how well each model fits the data, we perform 10-fold cross validation in the model construction process and select the model with minimum cross-validation errors in terms of the running parameter. The requirement for this project can be found at <https://github.com/ucb-stat159/stat159-fall-2016/blob/master/projects/proj02/proj02-predictive-modeling.pdf>.

Data

The data we used in this analysis can be downloaded from <http://www-bcf.usc.edu/~gareth/ISL/Credit.csv>, which is provided by the textbook we refer to throughout the project, An Introduction to Statistical Learning.

After downloading the original data, in order to have comparable scales, we first standardize the data by centering around means and dividing by their respective standard deviation. Then, since we will use cross validation to improve the accuracy of our fit, we divide the data into training set and test set. We have 400 observations in total, so we randomly select 300 entries to be in the train set and the remaining 100 in the test set.

Methods

Least Squares Method

OLS estimators are considered as our base case in this project. OLS estimators are obtained by minimizing the sum of squares $RSS = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^p (\beta_j x_{ij}))^2$ where β_0, \dots, β_p are our coefficient estimates. This is the most common regression method, and it works the best when we have homoskedastic and uncorrelated errors within the data.

Shrinkage Method

Similar to the OLS method, shrinkage method also fits a model containing all independent variables but with a focus on constraining the coefficient estimates toward 0. By shrinking the coefficient estimates, the variance in estimators is significantly reduced compared to that of OLS.

Ridge Regression

Instead of minimizing RSS, Ridge regression minimizes $RSS + \lambda \sum b_j^2$. This regression abandons the requirement of an unbiased estimator in order to obtain a more precise prediction intervals. The addition term is a shrinkage penalty and since it's a sum of coefficient estimators, it will be small when β s are all close to 0. Here, λ is called a tuning parameter and it controls how much RSS and shrinkage penalty affects the regression model. When $\lambda = 0$, ridge regression is the same as the least squares models. Because each λ corresponds to a different set of coefficient estimates, we usually try a wide range of λ s and pick the best tuning parameter with the smallest MSE.

One of the main features of ridge regression is the bias-variance trade-off. When $\lambda = 0$, which is the least square model, the variance is high the estimators are unbiased. As λ increases, variance decreases significantly but bias only slightly increases. Since $MSE = variance + squared\ bias$ along with the relative effect of the change, when λ is smaller than 10, the variance drops rapidly, with very little increase in bias, so MSE decreases.

Lasso Regression

The major difference between Lasso and Ridge is that Lasso method aims to minimize $RSS + \lambda \sum |\beta_j|$. This is to compensate for the fact that ridge regression will always generate a model with all predictors since it doesn't involve a process of variable reduction. Therefore, lasso improves on this by allowing some of the coefficient estimates to be exactly 0 if the tuning parameter is sufficiently large. Therefore, this variable selection process makes the model much easier to interpret with a reduced amount of predictors.

Comparison between Ridge and Lasso

In general, lasso is expected to outperform ridge when we have a small amount of predictors having significant coefficients, and the rest close to 0. Ridge regression will perform better when the response is affected by many regressors with equal-sized coefficients.

Dimension Reduction Methods

Dimension reduction method involves a transformation of variables before we fit in the least squares model. Instead of estimating $p + 1$ coefficients β_0, \dots, β_p when we have p regressors, it transforms the data to $Z_1, \dots, Z_M, M < p$ where Z_m represent linear combinations of the original predictors, so that we only need to estimate $M + 1$ coefficients $\theta_0, \theta_1, \dots, \theta_M$.

Principal Components Regression

In principal components regression, we first perform principal components analysis on the original data which constructs M principal components, Z_1, \dots, Z_M . Then, we use these components as the predictors and fit the least squares model to obtain coefficient estimates. The key assumption we hold in this regression model is that the directions in which X_1, \dots, X_p show the most variation are the directions that are associated with Y . If the assumption holds, then fitting a least squares model to Z_1, \dots, Z_M will lead to better results than fitting a least squares model to X_1, \dots, X_p , since most or all of the information in the data that relates to the

response is contained in Z_1, \dots, Z_m . With fewer predictors, we can also reduce the risk of overfitting. PCR performs the best when the first few principal components are sufficient to capture most of the variation in the predictors and their relationship with the response.

Partial Least Squares

In comparison to PCR regression, which uses unsupervised method to identify the principal components, partial least squares method takes the advantage of a supervised learning process. It uses the response variable Y to identify new vectors that are not only similar to existing regressors, but also select the ones that are related to the response. Therefore, as indicated in the textbook, the PLS approach attempts to find directions that help explain both the response and the predictors.

Analysis

Results

OLS Regression

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.00	0.01	-0.10	0.92
Income	-0.60	0.02	-29.58	0.00
Limit	0.95	0.19	5.12	0.00
Rating	0.38	0.19	2.05	0.04
Cards	0.05	0.01	3.55	0.00
Age	-0.03	0.01	-2.03	0.04
Education	-0.01	0.01	-1.18	0.24
GenderFemale	-0.02	0.01	-1.78	0.08
StudentYes	0.28	0.01	21.68	0.00
MarriedYes	-0.02	0.01	-1.35	0.18
EthnicityAsian	0.01	0.02	0.36	0.72
EthnicityCaucasian	-0.00	0.01	-0.32	0.75

Table 1: OLS Coefficients

Since OLS would be the base case in this project, we also use 10-fold cross validation in building the OLS model by first using the train set to fit the model, calculate MSE in the test set and then fit the model we select to the entire data set. From the OLS regression output, we noticed some coefficients come with a big p-value, which means that they are not statistically significant. Therefore, we can conclude that the constant term, Education, Gender, Marital Status and Ethnicity don't belong to this regression. Also we noticed that among the statistically significant regressors, some has very small coefficients, so the main factors influencing Balance are Income, Limit and Rating.

Ridge Regression

	Estimate
(Intercept)	0.00
Income	-0.57
Limit	0.72
Rating	0.59
Cards	0.04
Age	-0.03
Education	-0.01
GenderFemale	-0.01
StudentYes	0.27
MarriedYes	-0.01
EthnicityAsian	0.02
EthnicityCaucasian	0.01

Table 2: Ridge Coefficients

With Ridge regression, the λ we found that results in the smallest validation error is $\lambda = 0.01$. As we have a relatively small λ , we expect to see that the estimation with ridge is very similar to that of OLS but a little bit smaller due to the shrinkage effect.

Lasso Regression

	Estimate
(Intercept)	0.00
Income	-0.55
Limit	0.93
Rating	0.37
Cards	0.04
Age	-0.02
Education	0.00
GenderFemale	0.00
StudentYes	0.27
MarriedYes	0.00
EthnicityAsian	0.00
EthnicityCaucasian	0.00

Table 3: Lasso Coefficients

Lasso improves on Ridge because it adds the incentive to render statistically insignificant estimates to 0 by performing both variable selection and yields model that involves only a subset of the variables. In this case, the λ we found that results in the smallest validation error is $\lambda = 0.01$. We can see that a significant improvement is that we have a number of regressors with a coefficient of 0 which makes the interpretation much easier.

PCR Regression

	Estimate
Income	0.26
Limit	0.27
Rating	0.27
Cards	-0.05
Age	0.05
Education	0.06
GenderFemale	0.05
StudentYes	0.11
MarriedYes	-0.03
EthnicityAsian	-0.00
EthnicityCaucasian	-0.02

Table 4: PCR Coefficients

With PCR, we mainly focus on dimension reduction by unsupervised learning. By comparing validation errors for different Ms, we decide that the best M to use here is 10. Since we only have 11 predictors in this multiple regression, it is not a huge change, so that explains why PCR is very close to OLS. Therefore, we can say that the dimension of predictors for this regression almost cannot be reduced, there doesn't exist major principal components dominating the change in the response variable, so PCR doesn't help to improve that much on the OLS estimation.

PLS Regression

	Estimate
Income	-0.60
Limit	0.67
Rating	0.67
Cards	0.05
Age	-0.03
Education	-0.01
GenderFemale	-0.02
StudentYes	0.27
MarriedYes	-0.01
EthnicityAsian	0.02
EthnicityCaucasian	0.01

Table 5: PLS Coefficients

Similarly, PLS regression is also trying to reduce the dimension of predictors, but in an supervised way. By comparing validation errors for different Ms, we decide that the best M to use here is 4. Since we originally have 11 predictors in this multiple regression, this is a huge improvement, so this method successfully find the major principal components dominating the change in Y, and therefore reduce the risk of overfitting and therefore obtain a better fit.

Combined All Coefficients Together

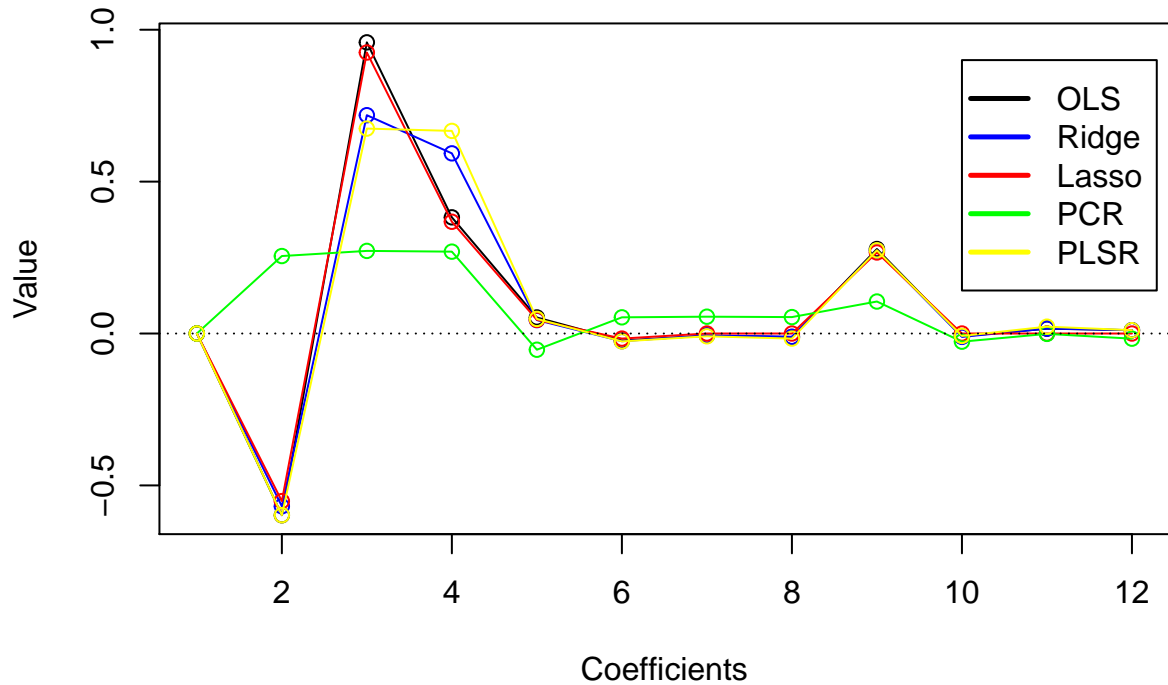
	ols	ridge	lasso	pcr	plsr
(Intercept)	0.00000	0.00000	0.00000	0.00000	0.00000
Income	-0.59817	-0.56871	-0.55166	0.25505	-0.59801
Limit	0.95844	0.71866	0.92505	0.27196	0.67458
Rating	0.38248	0.59306	0.36787	0.26946	0.66699
Cards	0.05286	0.04425	0.04500	-0.05334	0.04808
Age	-0.02303	-0.02538	-0.01666	0.05327	-0.02592
Education	-0.00747	-0.00588	0.00000	0.05542	-0.00908
GenderFemale	-0.01159	-0.01068	0.00000	0.05408	-0.01671
StudentYes	0.27815	0.27318	0.26681	0.10534	0.27394
MarriedYes	-0.00905	-0.01103	0.00000	-0.02681	-0.00800
EthnicityAsian	0.01595	0.01638	0.00000	-0.00120	0.02257
EthnicityCaucasian	0.01101	0.01101	0.00000	-0.01697	0.01101

Table 6: Regression Coefficients for 5 Regression Methods

	MSE
ols	0.05179
ridge	0.05259
lasso	0.05154
pcr	0.41615
plsr	0.05192

Table 7: MSE of 5 Regression Methods

Trend Lines of Coefficients for Different Regression Models



Conclusion

Combining the coefficients from all five regression models, we notice that only the coefficients for Limit, Rating and StudentYes varies and a number of regressors have coefficients close to 0. As we explained above, each model has its own feature and can lead to the optimal regression model under different circumstances. Here, for the data set we work on in this project, both our analysis and MSE shows that Partial Least Squares Regression model fits the best and leads to the most accurate prediction.