Project 2 Report

Shuotong Wu, Siyu Chen

11/03/2016

Abstract

In this project, we try out different muitiple regression models and find the best approach to fit the data by comparision.

Introduction

In this project, we will utilize the model selection method used in Chapter 6, Linear Model Selection and Regularization, from the book "An Introduction to Statistical Learning" by Gareth James, Deniela Witten, Trevor Hastie and Robert Tibshirani.

The following five models will be applied to our dataset: Ordinary Least Squares, Ridge regression, Lasso regression, Principal Components regression and Partial Least Squares regression.

We will first split the data into training set and test set by 3:1 ratio. For each regression method, we then use a 10-fold cross validation in the training set trying to find "best" model by selecting the one with minimum cross validation errors. Once we get the model with optimal hyper-parameters for each regression method, we will apply them to test set and compare their MSE. Finally we will refit the best model on the full dataset and get our final parameter estimate.

Data

The dataset we use can be found here.

We clean the data by turning categorical variables into dummies, mean centering and standardizing.

We have 400 entries in total and use 3:1 ratio for train-test-split, which gets us a training set with 300 entries and test set with 100 entries.

Methods

In this section, we will introduce the five linear regression method we use.

Ordinary Least Squares Regression Method

OLS estimators is the most common regression method. OLS estimators minimizes residual sum of squares: $RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}(\beta_j x_{ij}))^2$ where β_i are coefficient estimates. The OLS estimator is optimal among linear unbiased estimators when the errors are homoscedastic and uncorrelated.

Ridge Regression

Ridge regression imposes a penalty on the size of coefficients on OLS. Instead of minimizing RSS, Ridge regression minimizes $RSS + \lambda \sum bj^2$. This model solves a regression model where the loss function is the linear least squares function and regularization is given by the l2-norm. λ is complexity parameter that controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage and thus the coefficients become more robust.

Lasso Regression

Lasso regression estimates sparse coefficients. It contains a l1 prior as regularizer. Lasso method minimize $RSS + \lambda \sum |\beta_j|$. Compared to ridge regression, lasso regression has tendency to prefer solutions with fewer parameters values, efficiently reducing the number of variables upon which the solution is dependent.

Principal Components Regression

Typically, Principal Components Regression considers regresses based on a standard linear regression model, but uses PCA for estimating the unknown regression coefficients in the model. We first perform PCA(privipal components analysis) and then use these components as the regressors to obtain coefficient estimates. PCR performs really well if major pricipal components are more determined in the relationship between x and y than trivial components.

Partial Least Squares

Instead of finding hyperplanes of maximum variance between y and x in PCR, Partial Least Squares finds a linear regression model by projecting y and x to a new space. It outperforms standard regression when there are more variables than observations. And PLS finds the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space so that it tries to explain both y and x.

Analysis

Exploratory Data Analysis

We first explore our data. We calcluated statistics and summaries for quantitative variables, like income, age, and ratings. We also made histograms and boxplots to get a better idea of how data is distributed. For the qualitative variables like gender, ethnicity, we have the table of frequency for each of them. We also generate barplot to futher illustrate the distribution. Correlation matrix for quantitative variables is calculated as well. To learn the relationship between balance and qualitative variables, we first conduct anova(Analysis of variance), and then make boxplots for each pair of them.

Pre-modeling Data Processing

Once we get a rough idea of our original dataset, we decide to process our data before modeling. We dummy out categorical variables, center data around mean and standardize it. We export the processed data for use of modeling.

Regression Models

As mentioned in the data section, we have 400 entries in total, and randomly select 300 entries as training set, the rest as test set. We conduct the following process to 5 different methods discussed in methods section. We first use a 10-fold cross validation to look for good hyper-parameters. Then we select the best hyper-parameter set by minimum cv error, and fit it to test set to calculate MSE. We also plot the cross-validation errors in terms of the tunning parameter to visualize which parameter gives the "best" model. We at last fit the best model we have to the whole dataset.

For ordinary least squares linear model, we use lm(). The results is set as a standard.

For ridge regression and lasso regression, we use library *glmnet*. For parameters of the regression function, we set alpha to 0 for ridge regression and alpha to 1 for lasso regression. Intercept and standardize are both set to false since data is already mean centered and standardized. Cross validation is done by using *cv.glmnet*() function, which performs 10-fold cross validation by default. We select models by minimum lambda value and calculate mean squared error on test set.

For principal components regression and partial least square regression, we use library *pls*, pcr() and plsr(). Cross validation is done by setting regression methods' parameter validation to "CV". We plot validation errors using method validationplot(val.type = "MSEP") and select the best model. Similarly as above, we calculate MSE on test set and refit model to the whole dataset.

Then we compare the models we get from model building process, utilizing the tables, summaries and plots we saved. The result analysis detail is in the next section.

Results

OLS Regression

Below is the table for OLS regression summary. As we can see, certain coefficients comes with a relatively high p-value, like education and ethnicity which suggests that they may not be significant. Also, if we look at the absolute value of the estimated coefficients, we can see that income, limit and ratins dominate the change in reponse variable, which suggests that we should try principal components regression that will ignore trivial variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.00362	0.01209	-0.29959	0.76470
Income	-0.59941	0.02062	-29.07175	0.00000
Limit	0.96718	0.18327	5.27747	0.00000
Rating	0.38908	0.18391	2.11557	0.03524
Cards	0.05246	0.01444	3.63167	0.00033
Age	-0.02982	0.01228	-2.42904	0.01575
Education	0.00164	0.01201	0.13637	0.89163
GenderFemale	-0.01341	0.01208	-1.11036	0.26777
StudentYes	0.27845	0.01217	22.88639	0.00000
MarriedYes	-0.02735	0.01229	-2.22622	0.02677
EthnicityAsian	-0.00292	0.01539	-0.18998	0.84946
EthnicityCaucasian	0.00492	0.01490	0.33033	0.74139

Table 1: OLS Coefficients

Ridge Regression

	Estimate
(Intercept)	0.00000
Income	-0.56803
Limit	0.70284
Rating	0.60819
Cards	0.04363
Age	-0.02545
Education	-0.00580
GenderFemale	-0.01066
StudentYes	0.27307
MarriedYes	-0.01114
EthnicityAsian	0.01645
EthnicityCaucasian	0.01103

Table 2: Ridge Coefficients

The hyper-parameter λ we get from cross validation with minimum validation error is $\lambda = 0.01$. The estimated coefficients are consistent with ones we get from OLS. Income, limit and ratings still dominate the change in reponse. Other trivial variabls, like cards and gender are given less influence because of introducing l2-norm in ridge regression.

Lasso Regression

	Estimate
(Intercept)	0.00000
Income	-0.55137
Limit	0.78139
Rating	0.51119
Cards	0.03883
Age	-0.01676
Education	0.00000
GenderFemale	-0.00000
StudentYes	0.26607
MarriedYes	0.00000
EthnicityAsian	0.00000
EthnicityCaucasian	0.00000

Table 3: Lasso Coefficients

The hyper-parameter λ we get from cross validation with minimum validation error is $\lambda = 0.01$. Comparing to the estimates in ridge regression, we can see that a lot of regressors are given 0 estimate, which greatly reduce the number of regressors. As discussed in the methods section, this is caused by the tendency lasso regression holds to prefer solutions with fewer paramters than ridge regression. Dominated variables are given much more power and trivial variables are given none power. Whether reducing the number of regressors is an improvement or not will be discussed when we compare all our 5 models.

Principal Components Regression

	Estimate
Income	-0.59887
Limit	0.67141
Rating	0.67064
Cards	0.04043
Age	-0.02327
Education	-0.00600
GenderFemale	-0.01165
StudentYes	0.27636
MarriedYes	-0.01116
EthnicityAsian	0.01741
EthnicityCaucasian	0.01119

Table 4: PCR Coefficients

The best hyper-parameter for number of components we get is 10 by cross validation. The range space for our hyper-parameter is 1 to 11 since we only have 11 variables. Reduce 1 dimension is trivial. Hence, the results from PCR are similar to those of OLS. As we will see later, PCR does not improve MSE neither.

Partial Least Squares Regression

	Estimate
Income	-0.59814
Limit	0.95782
Rating	0.38314
Cards	0.05227
Age	-0.02340
Education	-0.00759
GenderFemale	-0.01193
StudentYes	0.27818
MarriedYes	-0.00865
EthnicityAsian	0.01594
EthnicityCaucasian	0.01106

Table 5: PLS Coefficients

The best hyper-parameter for number of components we get is 9 by cross validation. We can see some improvement, but 9 is still very close to 11, our original number of regressor.

Comparision of 5 models

	MSE
ols	0.05784
ridge	0.05741
lasso	0.05537
pcr	0.05817
plsr	0.05736

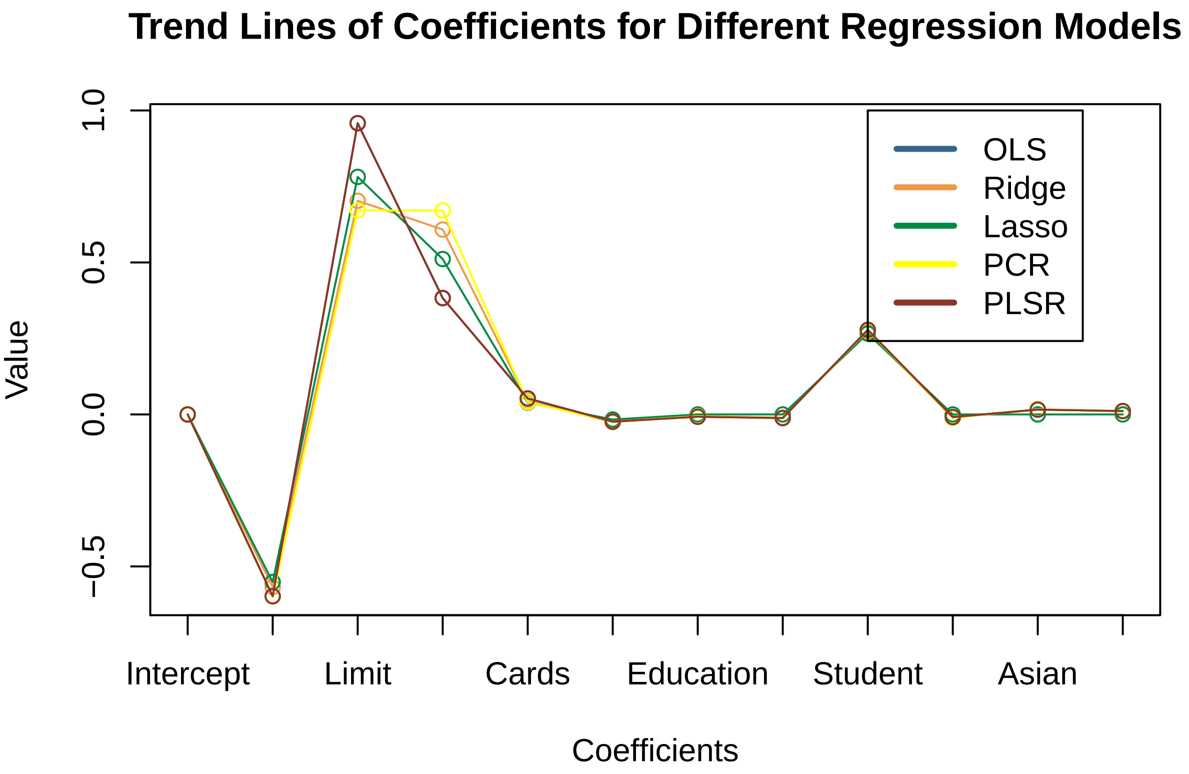
Table 6: MSE of 5 Regression Methods

As we can see in the table, our lasso model perform the best. PLSR improves a little bit than PCR but is still not good enough. Ridge regression's result is similar to OLS, which is consistent with what we dicussed above.

Below is the table for coefficients for all 5 models, and trend lines for each of coefficient.

	ols	ridge	lasso	pcr	plsr
(Intercept)	0.00000	0.00000	0.00000	0.00000	0.00000
Income	-0.59817	-0.56803	-0.55137	-0.59887	-0.59814
Limit	0.95844	0.70284	0.78139	0.67141	0.95782
Rating	0.38248	0.60819	0.51119	0.67064	0.38314
Cards	0.05286	0.04363	0.03883	0.04043	0.05227
Age	-0.02303	-0.02545	-0.01676	-0.02327	-0.02340
Education	-0.00747	-0.00580	0.00000	-0.00600	-0.00759
GenderFemale	-0.01159	-0.01066	-0.00000	-0.01165	-0.01193
StudentYes	0.27815	0.27307	0.26607	0.27636	0.27818
MarriedYes	-0.00905	-0.01114	0.00000	-0.01116	-0.00865
EthnicityAsian	0.01595	0.01645	0.00000	0.01741	0.01594
EthnicityCaucasian	0.01101	0.01103	0.00000	0.01119	0.01106

Table 7: Estimated Coefficients of 5 Regression Methods



Conclusion

When we look at the coefficients table for all our 5 models, it's clear that limit, rating, student are the dominating factors in all models. Lasso regression model give the least power to trival variables, which is 0 to reduce the number of regressors, among all models. And it performs the best. It suggests that those trival factors may not have a strong relationship with balance. Other than lasso, the rest 4 models performs similarly. Those regression techniques are not helpful for our dataset. With the smallest MSE, we at last choose our lasso regression model for this dataset.