Predictive Modeling Process

Shirley Jin, Xiaoqian Zhu

Abstract

In this project, we are going to take a look at five different regression models, including, ordinary least squares, ridge, lasso, principal components, and partial least squares regressions. We are mainly following the analysis from *Linear Model Selection and Regularization*(chapter 6) of the *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Link to the book. Apply each model on the data set *Credit*, Link to the data set, we predict the variable `Balance` in terms of ten predictors such as `Income`, `Age`, `Education`, `Gender`, `Ethnicity`, etc and also describe the differences between these models. # Introduction

In this project we are going to follow the analysis from *Linear Model Selection and Regularization*(chapter 6) of the *An Introduction to Statistical Learning* by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Link to the book. Applying each model on the data set *Credit*, Link to the data set, we are going to predict the variable `Balance` in terms of ten predictors such as `Income`, `Age`, `Education`, `Gender`, `Ethnicity`, etc and also describe the differences between these models.Addtionally, this project involves working collaboratively in teams of two members.

In previous homework, we used simple and multiple linear regression for data analysis, while in this project, we consider some approaches for improving the simple linear model, by replacing plain least squares fitting with some alternative fitting procedures. *So, why might we want to use another fitting procedure instead of least squares?*

The linear model assumes that the true relationship between the response and the predictors is approximately linear and then, the least squares estimates will have low bias. If the number of observations is larger than the number of variables, the least squares estimates tend to have low variace and hence will perform well on test observations. However, if number of observations is not larger than the number of variables, then there can be a lot of variabiblity in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training. If the number of observations is smaller than the number of variables, then the method could not be used. This results in failures in *prediction accuracy*.

Moreover, It is often the case that some or many of the variables used in a multiple regression model are in fact not associated with the response. Including such irrelevant variables leads to unnecessary complexity in the resulting model. By removing these variables—that is, by setting the corresponding coefficient estimates to zero—we can obtain a model that is more easily interpreted. Now least squares is extremely unlikely to yield any coefficient estimates that are exactly zero.This results in failures in *Model Interpretability*.

Using alterantive fitting procedures can yield better prediction accuracy and model interpretability, by constraining or shrinking the estimated coefficients and performing feature selection or variable selection, and also reducing dimension.

In this report, we are going to focus shrinkage methods and dimension reduction method. Ridge and Lasso regressions are shrinkage methods. PCR and PLSR are dimension reduction methods. Functions to fit models with ridge and lasso regressions are available in the package `"glmnet"`. Functions to fit models with PCR and PLSR are available in the package `"pls"`. We use the multiple linear regression model via Ordinary Least Squares(OLS) based on the data set `credit` as the benchmark to compare the other methods.

Data

In this project, we are using the data set `Credit.csv`, which includes some qualitative and quantitative variables. The quantitative varaibels includes `age`, `cards`, `balance`, `income`, `limit`, `education`, `rating`,

which represent the sample's age, the number of credit cards, years of education, income measured in thousands of dollars, credit limit, credit rating and the sample's average credit card debt. The qualitative variables are `gender`, `student`, `status` and `ethnicity`, which represent whether the sample is a student, is married, and also the specific ethnicity.

However, before we can fit any model, we have to perform two major processing steps on the data set:

- convert factors into dummy variables
- mean centering and standardization

The first step involves transforming each categorical variable (`Gender`, `Student`, `Married`, and `Ethnicity`) into dummy variables. The main reason to do this is because the function `glmnet()` (used in ridge andlasso regressions) will not work if the input data contains factors.

The second processing step involves mean-centering and standardizing all the variables. This means that each variable will have mean zero, and standard deviation one. One reason to standardize variables is to have comparable scales. When you perform a regression analysis, the value of the computed coefficients will depend on the measurement scale of the associated predictors. A β coefficient will be different if the variable is measuredin grams or in kilos. To avoid favoring a certain coefficient, it is recommended to mean-center and standardize the variables.

The scaled data we get from the two steps is saved as a csv file, called `scaled-credit.csv`. This is the actual data we are going to use for the model building process.# Methods

Ordinary Least Squares Regression (OLS)

We first apply a multiple linear regression analysis on the data set `scaled_credit.csv`. In order to find the relationship between `Balance` and those financial and demogrphic variables, including `Age`, `Limit` and `Education`, we set up a linear model, which looks 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. In this case, those β coefficients are the least squares estimate of the actual values, estimated by minimizing the sum of the residual squared errors. Specifically, we are minimizing

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

, and 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. References from Wikipedia.

Bascially, minimizing the RSS would be minimizing the error of the prediction. Accoding to the Gauss-Markov Theorem, they are the best linear unbiased estimators in this model.

Shrinkage Methods

Ridge Regression

Ridge regression is a shrinkage method that constrain the coefficient estimates to help to reduce variance and thus get better estimation. In ridge regression, we are minimizing

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

, comparing to the OLS one, you can see the ridge regression has a term $\lambda \sum_{j=1}^{p} \beta_j^2$, which is called *shrinkage penalty*, which has the effect of shrinking the estimates of coefficients towards zero. λ in this term is called *tuning parameter*, which serves to control the relative impact on the regression coefficient estimates. When $\lambda = 0$, the penalty term will have no effect, and the ridge regression will produce the least squares estimates. However, when λ approaches infnity, the impact of the shrinkage penalty will grow, and hte ridge regression coefficient estimates will approach zero.

The Lasso

The *lasso* is also a shrinkage method and it is a relatively recent alternative to ridge regression that overcomes ridge's disadvantage, which we will explain specifically in the later section. Thus, the *lasso* is quite similar to ridge regression. The lasso coefficients minimize the quantity:

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

. As with ridge regression, the lasso shrinks the coefficient estimates towards zero. However, in the case of the lasso, the penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter is sufficiently large. Hence, much like best subset selection, the lasso performs variable selection. As a result, models generated from the lasso are generally much easier to interpret than those produced by ridge regression.

Additionally, to define a best λ for the minimization, we are going to use cross validation, which is model validation technique for assessing how the results of statistical analysis will generalize to an independent data set. In a prediction problem, a model is usually given a dataset of known data on which training data set is run , and a testing dataset against which the model is tested. The goal of cross validation is to define a dataset to "test" the model in the training phase, in order to limit problems like overfitting.

Dimension Reduction Methods

Principal Components Regression

Principal components regression (PCR) is a dimension reduction technique for regression, involves constructing the first M principal components, $Z_1 + \cdots + Z_{M*}$, and then using these components as the predictors in a linear regression model that is fit using least squares. The key idea is that often a small number of principal components suffice to explian most of the variability in the data, as well as the relationship with the response. This model is better than using OLS, because we can avoid overfitting in this model, and the model will not pick up any unnecessary variability. In PCR, the variance will increase as the number of variables in the model increases, while the bias will decrease.

Furthermore, the number of principal components, M, is also chosen by cross-validation. It is also crucial to standardize the variables when running PCR, because the scale of each variable may effect variance and thus affect the fit of the model produced.

Partial Least Squares

Partial least squares (PLS), is a supervised alternative to PCR. Like PCR, PLS is a dimension reduction method, while unlike PCR, PLS makes uses of the response Y in order to identify new features that not only approximate the old features well, but also that are related to the response. Roughly speaking, the PLS approach attempts to find directions that help explain both the response and the predictors.

We now describe how the first PLS direction is computed. After standardizing the p predictors, PLS computes the first direction Z_1 by setting each ϕ_{j1} in equal to the coefficient from the simple linear regression of Y

onto X_j. One can show that this coefficient is proportional to the correlation between Y and X_j. Hence, in computing

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

, PLS places the highest weight on the variables that are most strongly related to the response.

Analysis

This is the table of regression coefficients for all methods.

	Variables	OLS	Ridge	Lasso	PCR	PLSR
1	Intercept	0.000	0.000	0.000	0.000	0.000
2	Income	-0.598	-0.569	-0.552	-0.598	-0.598
3	Limit	0.958	0.719	0.925	0.958	0.958
4	Rating	0.382	0.593	0.368	0.382	0.382
5	Cards	0.053	0.044	0.045	0.053	0.053
6	Age	-0.023	-0.025	-0.017	-0.023	-0.023
7	Education	-0.007	-0.006	0.000	-0.007	-0.007
8	GenderFemale	-0.012	-0.011	0.000	-0.012	-0.012
9	StudentYes	0.278	0.273	0.267	0.278	0.278
10	MarriedYes	-0.009	-0.011	0.000	-0.009	-0.009
11	EthnicityAsian	0.016	0.016	0.000	0.016	0.016
12	EthnicityCaucasian	0.011	0.011	0.000	0.011	0.011

Table 1: Regression Coefficients for 5 Regression Methods

OLS Regression

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.00000	0.01074	0.00000	1.00000
Income	-0.59817	0.01796	-33.31357	0.00000
Limit	0.95844	0.16456	5.82412	0.00000
Rating	0.38248	0.16520	2.31522	0.02112
Cards	0.05286	0.01295	4.08301	0.00005
Age	-0.02303	0.01103	-2.08820	0.03743
Education	-0.00747	0.01086	-0.68767	0.49207
GenderFemale	-0.01159	0.01079	-1.07457	0.28324
StudentYes	0.27815	0.01093	25.45943	0.00000
MarriedYes	-0.00905	0.01099	-0.82351	0.41073
EthnicityAsian	0.01595	0.01340	1.19018	0.23470
EthnicityCaucasian	0.01101	0.01330	0.82777	0.40831

Table 2: OLS Coefficients

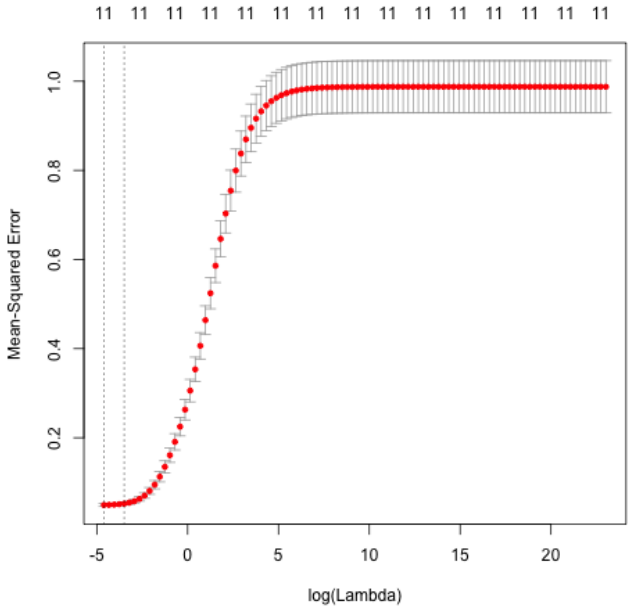
OLS is the benchmark for the comparison between different models. From the OLS regression results, we find that some coefficients have a very big p-value, thus they are not statistically significant. Therefore, Education, Gender, Marital Status and Ethnicity do not seem to have a relationship with Balance. Furthermore, we also find that some statistically significant regressors with small coefficients, which means that has very small economic effect on Balance. Thus, the main factors that may have relationship with Balance are Income, Limit and Rating.

Ridge Regression

	Estimates
(Intercept)	0.00000
Income	-0.56871
Limit	0.71866
Rating	0.59306
Cards	0.04425
Age	-0.02538
Education	-0.00588
GenderFemale	-0.01068
StudentYes	0.27318
MarriedYes	-0.01103
EthnicityAsian	0.01638
EthnicityCaucasian	0.01101

Table 3: Ridge Coefficients

In this Ridge regression, we found that results in the smallest validation error is $\lambda = 0.01$. This tuning parameter λ is relatively small. Comparing ridge regression coefficients with OLS coefficients, we find that the estimation with ridge is very similar to that of OLS and has relatively smaller coefficients, expect the Rating coefficient, is much larger in ridge.

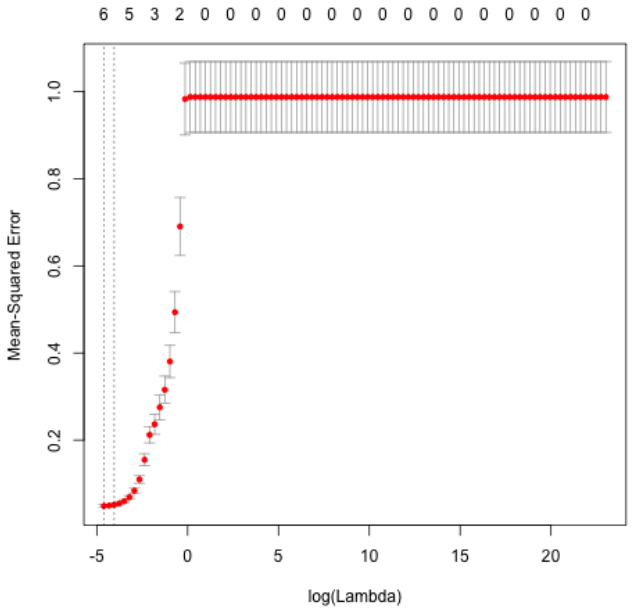


Lasso Regression

Lasso is an improved alternative to ridge, as it adds the incentive to render statistically insignificant estimate to 0 by performing both variable slection and only fitting the data to the variables that fit the MSE criteria. Here, the λ we find is $\lambda = 0.01$. Moreover, six of the coefficients 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 based on ridge regression.

	Estimates
(Intercept)	0.00000
Income	-0.55166
Limit	0.92505
Rating	0.36787
Cards	0.04500
Age	-0.01666
Education	0.00000
GenderFemale	0.00000
StudentYes	0.26681
MarriedYes	0.00000
EthnicityAsian	0.00000
EthnicityCaucasian	0.00000

Table 4: Lasso Coefficients

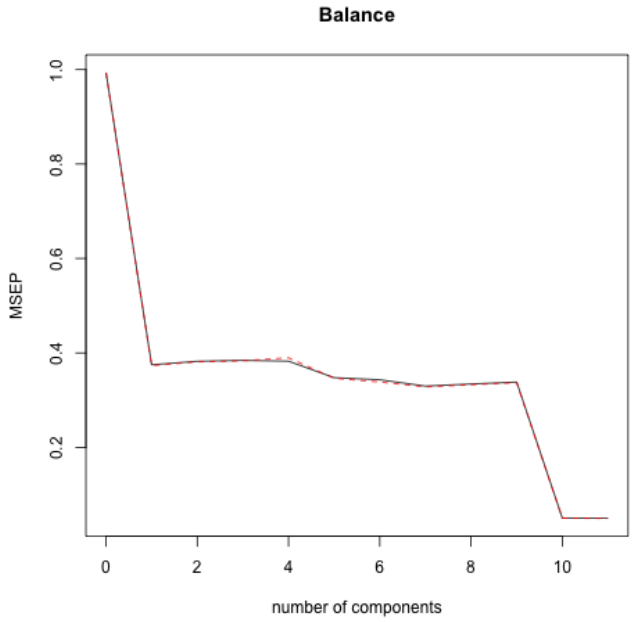


Principal Components Regression

In this case, we find that the best M is 11. Comparing coefficients between PCR and OLS, we find that they are almost same to each other. Additionally, the coefficients of PCR is very similar to ridge and PLSR regression.

	Estimates
Income	-0.59817
Limit	0.95844
Rating	0.38248
Cards	0.05286
Age	-0.02303
Education	-0.00747
GenderFemale	-0.01159
StudentYes	0.27815
MarriedYes	-0.00905
EthnicityAsian	0.01595
EthnicityCaucasian	0.01101

Table 5: PCR Coefficients

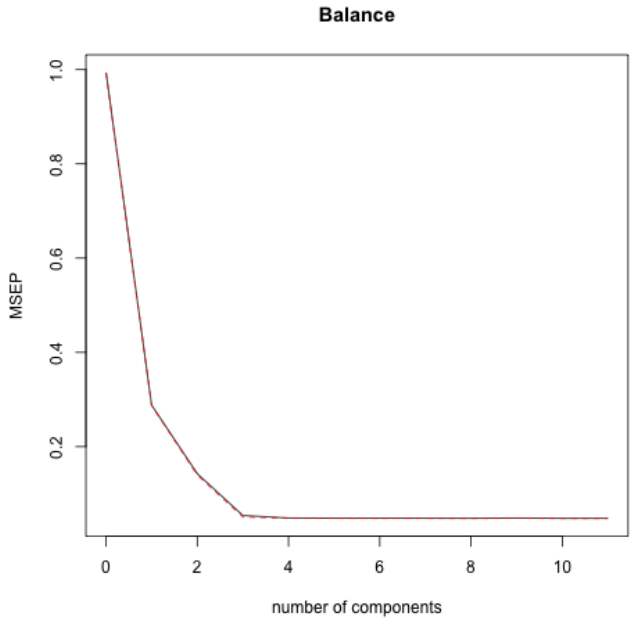


Partial Least Squares Regression

PLS is a supervised alternative to PCR. By comparing validation errors for different Ms, we find the best M is 10. Comparing coefficients of PLS to OLS coefficients, we can see they are almost same to each other.

	Estimates
Income	-0.59817
Limit	0.95844
Rating	0.38248
Cards	0.05285
Age	-0.02303
Education	-0.00748
GenderFemale	-0.01163
StudentYes	0.27816
MarriedYes	-0.00908
EthnicityAsian	0.01595
EthnicityCaucasian	0.01100

Table 6: PLS Coefficients



Results

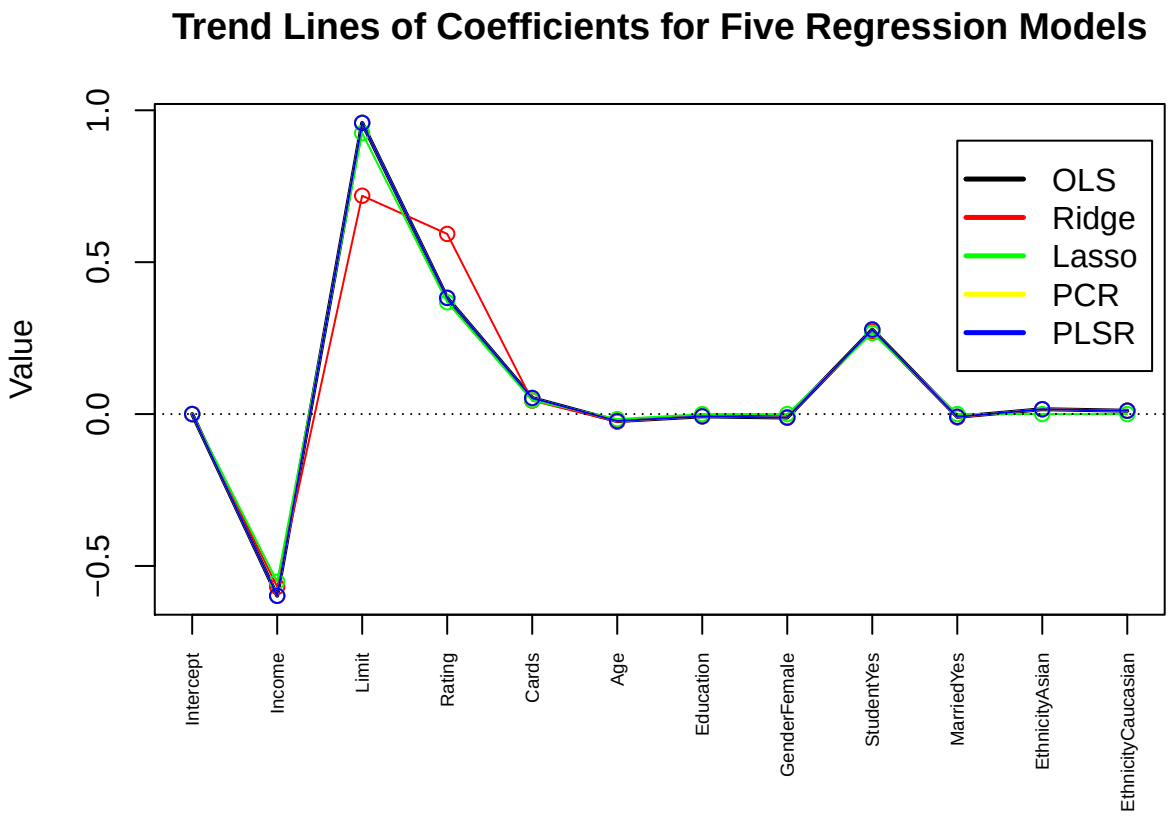
	MSE
ols	0.04479
ridge	0.04798
lasso	0.05152
pcr	0.04747
plsr	0.04754

Table 7: MSE of Five Regression Methods

From the MSE table above, we find that all the MSE are very close to each other and the maximum difference between the MSE of ols and that of lasso. The regression with the smallest MSE is ols regression, where MSE = 0.0447862. As ols has the smallest MSE, then we can conlcude that ols is the best fitted model

for the credit data set in this case. Aditionally, the one with the largest MSE is lasso regression, where MSE is euqual to 0.0515181. this means tha lasso regression provides the worst fitted model for the credit dataset.This is a very interesting finding that extended approaches get worse model than ols regression. This is may because we do not have a very large number of variables or low ratio of number of observations to number of variables. Moreover, it is also possible that preditors ae highly correlated, thus lasso does not yeild good results in presense of high collinearity, results in the worst fit model.

Furthermore, we also include a trendline plot in which the official coefficients are compared.



This trendline plot also allows us to compare how each type of regression models fit a model to the credit data set. We can see from the graph that the lines of PLSR, PCR and OLS are very close to each other, and even overlap. Lasso has the biggest variance with OLS generally. Ridge generally has smaller coefficients with OLS. Moreover, for variables like `Income`, it is almost same in each model, however, for varibles like `Rating` and `Limit`, these estimates vary widely in different regression methods.

Conclusions

In conclusion, in this project, we fitted 5 types of regression models on the dataset `credit.csv` and we found that even though the MSE from all the regression models are quite similar, the lasso regression model generated the lowest MSE, at 0.0515181. lasso regression also seems to have the least variance with OLS as shown from the trendline plot. This seems to indicate that lasso regression model is the best fitting model for predicting `balance` from the predictor variables in this dataset. In addition, the ridge regression model generated the highest MSE, indicating that it might be the worst fitting model.