slides

Introduction

The aim of this statistical analysis project is to compare different kind of linear regression models for the Balance variable in credit data. Such linear regressions are:

1. Principal Component Regression
2. Lasso
3. Ridge
4. Ordinary Least Squared
5. Partial Least Squared Regression

Data

From the set of data that we analyzed we observe that we have these kind of variables:

- Qualitative Variables(4) -1. Gender -2. Student -3. Married Ethnicity
- Quantitative Variables(7) -1. Income -2. Limit -3. Rating -4. Cards -5. Age -6. Education -7. Balance

Types of Regression

Ridge Regression

Ridge regression is similar to OLS with the coefficients estimated by minimizing a slightly different quantity. By minimzing RSS, we can find the coefficient estimates that fit the data well. Ridge regression will produce a different set of coefficient estimates for each value of λ .

Lasso regression

Lasso also shrink the coefficient estimates to zero. However, the penalty has the effect of focing some of the coefficient estimates to equal to zero when the tuning pararmeter is too large. Thus, lasso is better on feature selection.

Ordinary Least Squares (OLS)

Apply multiple linear regression by using the lm() function to find the relationship between Balnace and the 11 predictors of Income, Limit, Rating and more.

Dimension Reduction Methods

Principal Components regression (PCR)

PCR performs well when the first few principal components have enough information on the variation in the predictors and relationship with the response. The response does not supervise the principal components.

Partial Least Squares regression (PLSR)

Partial Least Square is a supervised way of PCR. PLS is a dimension reduction method and fits a linear model through least square using $Z_1 + \cdots + Z_{M*}$. PLS will try to find ways to explain the trend and pattern of reponse and predictor variables.

Analysis

Comparing results

OLS

*The ordinary least squares regression MSE is 0.044683.

Ridge

*The ridge regression MSE is 0.0525927. When comparing the ridge regression with OLS, the results are similar.

Lasso

*The lasso regression MSE is 0.05154446 The lasso regression and ridge regression have the same results.

Principal Components Regression

*The PCR MSE is 0.05199678. The PCR coefficients and OLS coefficients are similar. They also shared the similar results with ridge regression.

Partial Least Squares Regression

*The PLSR MSE is 0.0517916. The PLSR coefficients have some difference from PCR. When comparing with second and third components, PCR and PLSR have different results. The majority others remains the same.

Results

From the analysis, we can see that the MSE are similar to each other. The OLS has the smallest MSE, which is the best fit model for the credit data set. The ridge regression has the largest MSE among the five regression, which is the least best fit model. When comparing the cross validation on both ridge and lasso, lasso cross validation is less consistent than ridge cross validation.

Conclusion

This project uses five regression model to determine the best fit model for the Credit dataset.It is highly important to compare each of the results each method gives for a more accurate and comprehensible approach. Based on the result, the OLS perform the best among all the models. The second best is the lasso regression. However, the MSE among all models are very similar and they shared similar coefficients. That means that most of the models are pretty optimal.