Analysis

In analysis part, we will see the results of the each methods we applied to a given data set.

1) Ordinary Least Square Method(OLS)

There are several steps in applying the Ordinary Least Square method.

STEP1. Using the result from F-statistics, ruling out the qualitative variables that is not significant in identifying the response variable. As you can see in the anova Tables 1-4, only student variable is signifant in evaluating the response variable balance. Thus, we are ruling out ethnicity, gender, and marrage variables.

STEP2. Apply the OLS on our training dataset on selected predictor variables: income, limit, rating, cards, ages, eduction, student as shown in Table 5. This also enables us to identify the optimal predictors to predict the balance: income, limit, cards, students in Table 6.

STEP3. Using the result from step2, apply linear regression model on opitmal predictor variables to estimate the balance: income, limit, cards, student. Figure 1 shows the residual plot of OLS on test dataset.

STEP4. To identify the best OLS model. we are going to compare the Adjusted R square of both linear models. Model1, using selected predictor yields the 0.95371. Model2 using optimal predictor yields 0.95366. Therefore, the second model using optimal predictors is the best OLS model because it uses few variables but produces almost same Adjusted R square value as the first model as shown in Table 6. Lastly, using the best model we found, apply it to full data set to the "official coefficients", which is displayed at the end of analysis section.

2) Ridge Regression(RR)

First, using ten-fold cross-validation, Ridge Regression selects the best model when $\lambda = 0.01$ and cross-validation plot is shown in Figure 2.

After finding out the best model, applying the best model to our test set to compute the Test Mean Square Error, which is 0.0479811, later use this figure to compare different method in the result section.

Lastly, using the best model we found, apply it to full data set to the "official coefficients", which is displayed at the end of analysis section.

3) Lasso Regression(LR)

Lasso Regression does exactly same procedure as Ridge Regression. It selecst the best model when $\lambda = 0.01$ and cross-validation plot is shown in Figure 3.

After finding out the best model, applying the best model to our test set to compute the Test Mean Square Error, which is 0.0515181, later use this figure to compare different method in the result section.

Lastly, using the best model we found, apply it to full data set to the "official coefficients", which is displayed at the end of analysis section.

4) Principle Components Regression(PCR)

Principle Component Regression selecst the best model when $M = 11$ where M represents the compression, reduced dimension and cross-validation plot is shown in Figure 4.

After finding out the best model, applying the best model to our test set to compute the Test Mean Square Error, which is 0.0474722, later use this figure to compare different methods in the result section.

Lastly, using the best model we found, apply it to full data set to the "official coefficients", which is displayed at the end of analysis section.

5) Partial Least Squares Regression(PLSR)

Principle Component Regression selecst the best model when $M = 8$ where M represents the compression, reduced dimension and cross-validation plot is shown in Figure 5.

After finding out the best model, applying the best model to our test set to compute the Test Mean Square Error, which is 0.0464641, later use this figure to compare different methods in the result section.

Lastly, using the best model we found, apply it to full data set to the "official coefficients", which is displayed at the end of analysis section.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
ethnicity	2.00	18454.20	9227.10	0.04	0.96
Residuals	397.00	84321457.71	212396.62		

Table 1: Anova analysis of ethnicity and Balance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
gender	1.00	38891.91	38891.91	0.18	0.67
Residuals	398.00	84301020.00	211811.61		

Table 2: Anova analysis of gender and Balance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
marriage	1.00	2714.77	2714.77	0.01	0.91
Residuals	398.00	84337197.14	211902.51		

Table 3: Anova analysis of marriage and Balance

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
student	1.00	5658372.27	5658372.27	28.62	0.00
Residuals	398.00	78681539.64	197692.31		

Table 4: Anova analysis of student and Balance

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0021	0.0124	-0.1680	0.8667
income	-0.5940	0.0211	-28.1295	0.0000
limit	1.0702	0.1919	5.5779	0.0000
rating	0.2676	0.1923	1.3917	0.1651
cards	0.0538	0.0150	3.5954	0.0004
age	-0.0078	0.0127	-0.6174	0.5374
education	-0.0115	0.0127	-0.9051	0.3662
student	0.2764	0.0129	21.4850	0.0000

Table 5: MODEL1 using selected predictor variables: income, limit, rating, cards, ages, eduction, student

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0021	0.0124	-0.1676	0.8670
income	-0.5946	0.0210	-28.3296	0.0000
limit	1.3361	0.0203	65.9319	0.0000
cards	0.0666	0.0122	5.4648	0.0000
student	0.2764	0.0128	21.5288	0.0000

Table 6: MODEL2 using Optimal predictor variables: income, limitm cards, student

Adjusted.R.Square	Value
Model1	0.9537
Model2	0.9537

Figure 1: Residuals plot of OLS on the test dataset

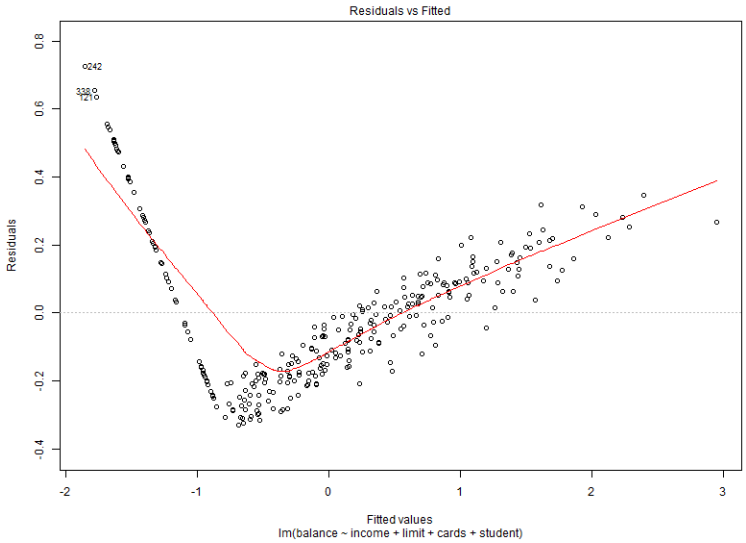

Figure 2: Plot of the cross-validtion errors of Ridge Regression on Lambda

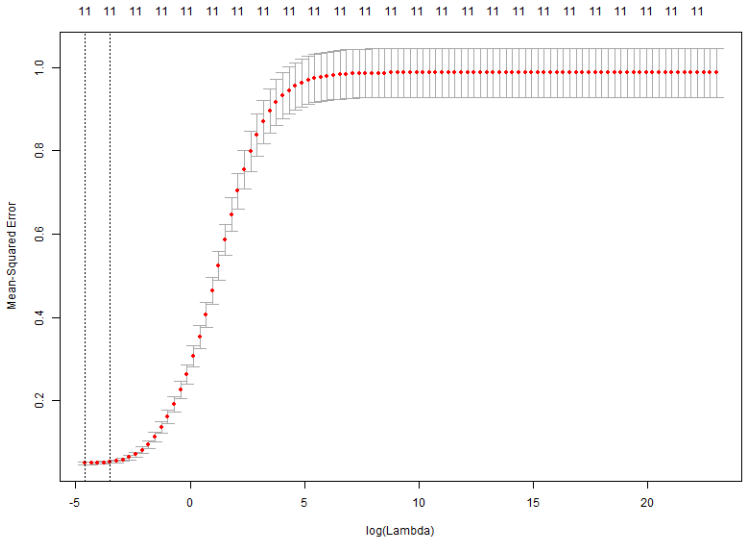

Figure 3: Plot of the cross-validtion errors of Lasso Regression on Lambda

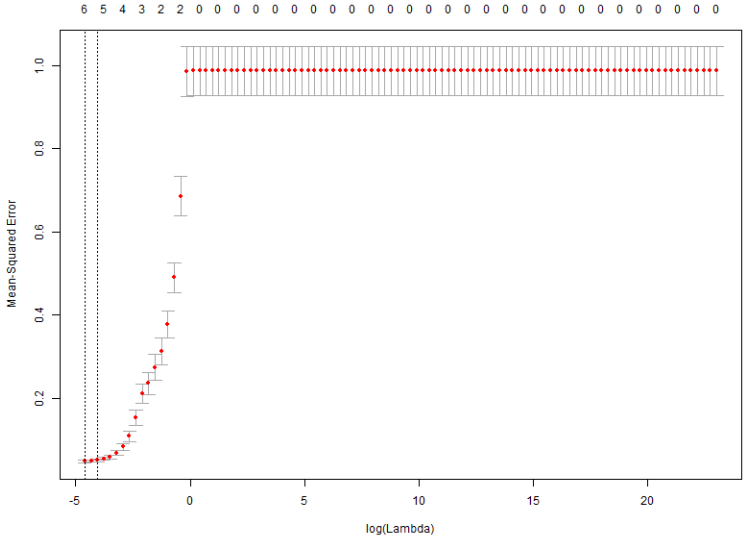

Figure 4: Plot of the cross-validtion errors of PCR Regression on M

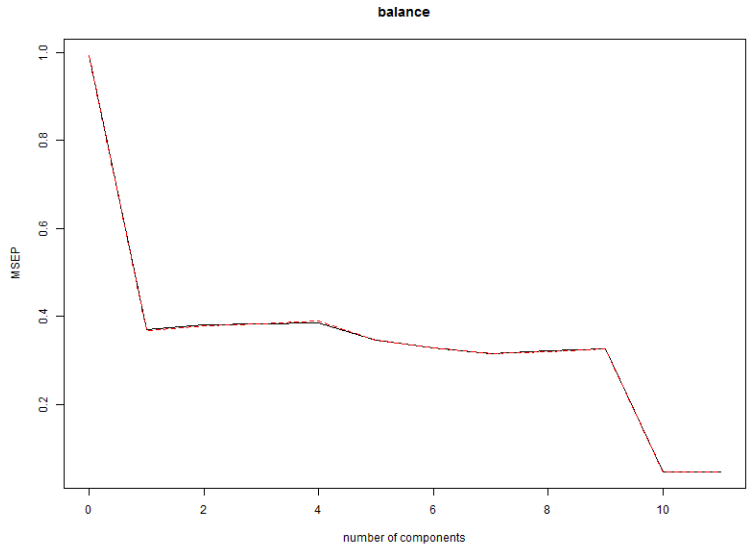

Figure 5: Plot of the cross-validtion errors of PLSR Regression on M