Analysis

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 Ethinicity 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.

	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 Regression

Lasso improves on Ridge by adding the incentive to render statistically insignificant estimates to 0. In this case, the λ we found that results in the smallest validation error is $\lambda = 0.01$, and we can see that a significant improve is that we have a number of regressors with 0 coefficient which makes the intepretation 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 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

Comparing the regression coefficients of five regression methods, we can see that the coefficients of Income, Cards, Age, Education, GenderFemale, StudentYes, MarriedYes, EthinicityAsian and EthnicityCaucasian are relatively close for each regression method while the coefficients of Limit and Rating are relatively

	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: PCR Coefficients

	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

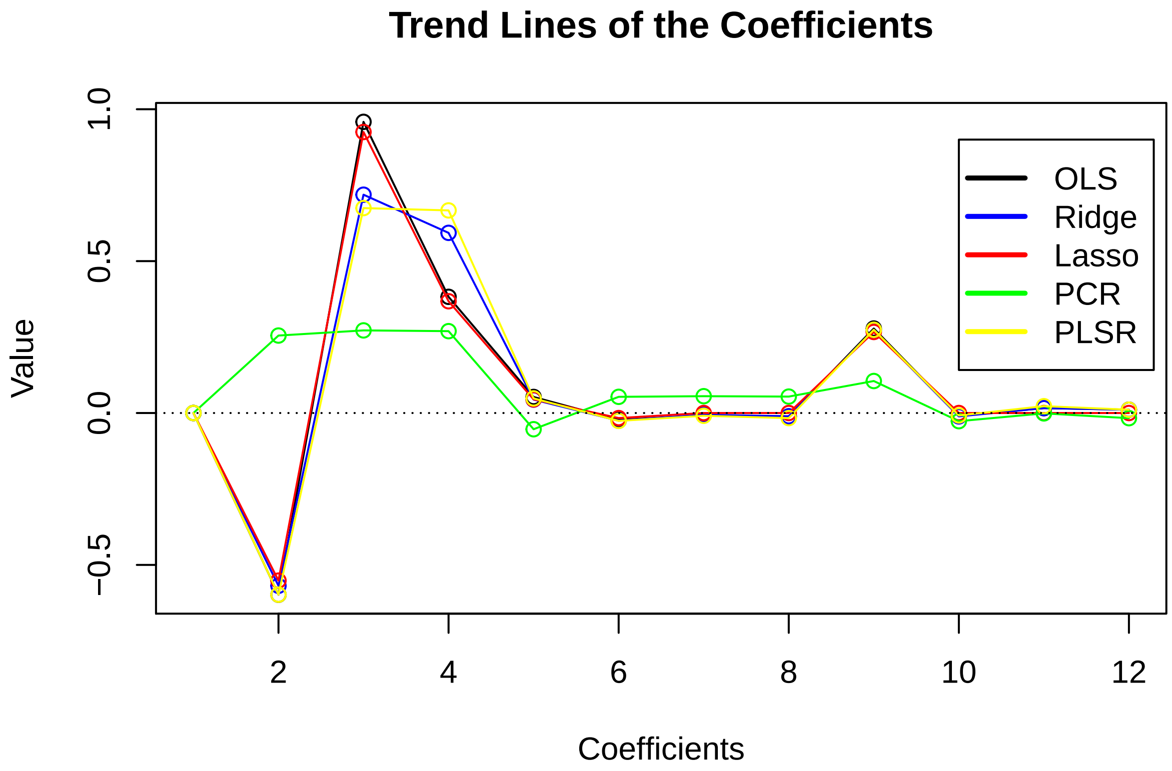
differed from each other. We can also see that the coefficients of Education, GenderFemale, MarriedYes, EthnicityAsian, and EthnicityCaucasion are very small and close to zero.

First, comparing the ols regression coefficients with the ridge regression coefficients, we can see that the ridge regression coefficients are generally smaller than the ols regression coefficients. This is because ridge regression method shrinks the coefficients of predictor variables and makes the coefficients closer to the true ones.

Second, comparing the lasso regression coefficients with the ridge regression coefficients, we can see that the lasso regression has several coefficients as zero. This is because lasso performs variable selection and yields model that involves only a subset of the variables. Lasso zeros out the coefficients of collinear variables. The advantage of lasso regression model over ridge regression model is that lasso regression method does both the parameter shrinkage and the variable selection and ridge regression will include all predictors in the final model and create a challenge in model interpretation.

	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 the Coefficients
Value
1.0
0.5
0.0
−0.5
2
4
6
8
10
12
Coefficients
OLS
Ridge
Lasso
PCR
PLSR