

## Results

### OLS

At last we look at the ordinary least square regression: The coefficients of the model that includes all predictors is:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.000	0.011	0.000	1.000
Income	-0.598	0.018	-33.314	0.000
Limit	0.958	0.165	5.824	0.000
Rating	0.382	0.165	2.315	0.021
Cards	0.053	0.013	4.083	0.000
Age	-0.023	0.011	-2.088	0.037
Education	-0.007	0.011	-0.688	0.492
GenderFemale	-0.012	0.011	-1.075	0.283
StudentYes	0.278	0.011	25.459	0.000
MarriedYes	-0.009	0.011	-0.824	0.411
EthnicityAsian	0.016	0.013	1.190	0.235
EthnicityCaucasian	0.011	0.013	0.828	0.408

Table 1: OLS coefficients

The R-square is 0.9551016. The Residual Standard Error is 0.2148752. We find out that among these 11 coefficients, some of them have relatively high p-value. For example, the corresponding values of categorical variables `education`, `EthnicityAsian`, `EthnicityCaucasian`, `gender`, `Married status` are 0.4920746, 0.2347047, 0.4083088, 0.2832368, 0.4107256 which are far bigger than 0.05.

In addition, we also find that the absolute value of the estimated coefficient `Income`, `Limit`, and `Rating` are 0.5981715, 0.9584387, 0.3824789 which are statistically significant and thus make huge influence in response `balance`.

### Ridge

We perform ridge regression on the centered Credit training data, and obtain the following result:

The coefficient of fitting the full data is following:

	Estimate
(Intercept)	-0.004
Income	-0.567
Limit	0.706
Rating	0.618
Cards	0.038
Age	-0.030
Education	-0.003
GenderFemale	-0.007
StudentYes	0.274
MarriedYes	-0.024
EthnicityAsian	0.016
EthnicityCaucasian	0.015

Table 2: Ridge Coefficients

Using the outputs of the 10-fold cross-validation with minimum validation error, the  $\lambda$  we get is  $\lambda = 0.01$  and the test mse is 0.0492226. Among the 11 variables, the corresponding absolute estimated value of **Income**, **Limit**, and **Rating** are 0.5668824, 0.7059976, and 0.6176414, which are statistically significant and thus have huge influence on the response **Balance**.

## Lasso

Then we fit lasso regression on the centered Credit training data. The refitting coefficients is the following:

	Estimate
(Intercept)	0.000
Income	-0.537
Limit	0.743
Rating	0.534
Cards	0.035
Age	-0.015
Education	0.000
GenderFemale	0.000
StudentYes	0.262
MarriedYes	0.000
EthnicityAsian	0.000
EthnicityCaucasian	0.000

Table 3: Lasso Coefficients

Using the outputs of the 10-fold cross-validation with minimum validation error, the  $\lambda$  we get is  $\lambda = 0.0132194$ , and the lasso test MSE is 0.048962.

We observe some coefficients could reduce to zero because of the special regularizing term the lasso regression has. For example, **Education**, **Gender**, **Married status**, and **Ethnicity** all reduce to zero. Such reduction makes the interpretation much easier.

## PCR

Now we use a different method which focus on dimension reduction by unsupervised learning – the principle component methods to fit on the training data. In this case, we think the optimal number of principle components used is 11, and the resulting test MSE is 0.0438743. The coefficients of PCR model refitting on full data set is:

	Estimate
Income	0.249
Limit	0.270
Rating	0.270
Cards	0.009
Age	0.056
Education	-0.012
GenderFemale	0.003
StudentYes	-0.000
MarriedYes	0.012
EthnicityAsian	-0.014
EthnicityCaucasian	0.000

Table 4: PCR Coefficients

## PLSR

We slightly change our method to PLSR which also focus on dimension reduction, but in a supervised way. The optimal number of principle components is 11 by comparing validating errors for different Ms, and the thus resulting test MSE is 0.0491513 The coefficient of refitting PLSR model on full dataset is:

	Estimate
Income	-0.598
Limit	0.958
Rating	0.382
Cards	0.053
Age	-0.023
Education	-0.007
GenderFemale	-0.012
StudentYes	0.278
MarriedYes	-0.009
EthnicityAsian	0.016
EthnicityCaucasian	0.011

Table 5: PLSR Coefficients

## Comparing the Coefficient Estimates for 5 regression models

	ols	ridge	lasso	pcr	plsr
Income	-0.598	-0.567	-0.537	0.249	-0.598
Limit	0.958	0.706	0.743	0.270	0.958
Rating	0.382	0.618	0.534	0.270	0.382
Cards	0.053	0.038	0.035	0.009	0.053
Age	-0.023	-0.030	-0.015	0.056	-0.023
Education	-0.007	-0.003	0.000	-0.012	-0.007
GenderFemale	-0.012	-0.007	0.000	0.003	-0.012
StudentYes	0.278	0.274	0.262	-0.000	0.278
MarriedYes	-0.009	-0.024	0.000	0.012	-0.009
EthnicityAsian	0.016	0.016	0.000	-0.014	0.016
EthnicityCaucasian	0.011	0.015	0.000	0.000	0.011

Table 6: coefficients estimates for 5 regression models

## Comparing the MSE of 5 regression Models