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)	0.00000	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 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 exmaple, the corresponding values of categorical variables `eduaction`, `EthnicityAisan` , `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 regresssion on the centered Credit training data, and obtain the following result:

The coefficient of fitting the full data is following:

	Estimate
(Intercept)	-0.00397
Income	-0.56688
Limit	0.70600
Rating	0.61764
Cards	0.03790
Age	-0.02998
Education	-0.00314
GenderFemale	-0.00722
StudentYes	0.27387
MarriedYes	-0.02427
EthnicityAsian	0.01594
EthnicityCaucasian	0.01480

Table 2: Ridge Coefficients

Using the outputs of the 10-fold cross-validation with minimum validation error, the λ we get is $\lambda = 0.01$ and the test mse is 0.0492226.Among the 11 variables, the corresponding absoluted 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)	0.00000
Income	-0.55137
Limit	0.78139
Rating	0.51119
Cards	0.03883
Age	-0.01676
Education	0.00000
GenderFemale	-0.00000
StudentYes	0.26607
MarriedYes	0.00000
EthnicityAsian	0.00000
EthnicityCaucasian	0.00000

Table 3: Lasso Coefficients

Using the outputs of the 10-fold cross-validation with minimum validation error, the λ we get is $\lambda = 0.01$, and the lasso test MSE is 0.0486758.

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, 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	0.24927
Limit	0.26975
Rating	0.26996
Cards	0.00928
Age	0.05568
Education	-0.01226
GenderFemale	0.00292
StudentYes	-0.00029
MarriedYes	0.01183
EthnicityAsian	-0.01355
EthnicityCaucasian	0.00002

Table 4: PCR Coefficients

PLSR

We slightly change our method to PLSR. The optimal number of principle components is 8, and the resulting test MSE is 0.0497093 The coefficient of refitting PLSR model on full dataset is:

	Estimate
Income	-0.60
Limit	0.85
Rating	0.49
Cards	0.05
Age	-0.02
Education	-0.01
GenderFemale	-0.01
StudentYes	0.28
MarriedYes	-0.01
EthnicityAsian	0.01
EthnicityCaucasian	0.01