Results

Ridge

We perform ridge regresssion on the centered Credit training data, and obtain the following result: The best lambda equals 0.01, and the test mse is 0.0492226. The coefficient of fitting the full data is following:

	1
(Intercept)	-0.00
Income	-0.57
Limit	0.71
Rating	0.62
Cards	0.04
Age	-0.03
Education	-0.00
GenderFemale	-0.01
StudentYes	0.27
MarriedYes	-0.02
EthnicityAsian	0.02
EthnicityCaucasian	0.01

Lasso

Then we fit lasso regression on the centered Credit training data. The best lambda is 0.01, and the lasso test MSE is 0.0486758. The refitting coefficients is the following:

	x
(Intercept)	0.00
Income	-0.55
Limit	0.78
Rating	0.51
Cards	0.04
Age	-0.02
Education	0.00
GenderFemale	-0.00
StudentYes	0.27
MarriedYes	0.00
EthnicityAsian	0.00
EthnicityCaucasian	0.00

We observe some coefficients could reduce to zero because of the special regularizing term the lasso regression has.

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 10, and the resulting test MSE is 0.0441615. The coefficients of PCR model refitting on full data set is:

	x
1	-0.00
2	0.25
3	0.27
4	0.27
5	0.01
6	0.06
7	-0.01
8	0.00
9	-0.00
10	0.01
11	-0.01
12	0.00

PLSR

We slightly change our method to PLSR. The optimal number of principle components is 11, and the resulting test MSE is 0.0491513 The coefficient of refitting PLSR model on full dataset is:

	x
1	-0.60
2	0.96
3	0.38
4	0.05
5	-0.02
6	-0.01
7	-0.01
8	0.28
9	-0.01
10	0.02
11	0.01

OLS

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

The R-square is 0.9551016. The Residual Standard Error is 0.2148752

Below is the table for OLS regression summary. As we can see, certain coefficients comes with a relatively high p-value, like education and ethnicity which suggests that they may not be significant. Also, if we look at the absolute value of the estimated coefficients, we can see that income, limit and ratins dominate the change in reponse variable, which suggests that we should try principal components regression that will ignore trivial variables.

{r results= ‘asis’, echo = FALSE} library(xtable) library(Matrix) options(xtable.comment = FALSE, xtable.table.placement = “H”) load(“../data/OLS/OLS-regression.RData”) print(xtable(ols_reg_sum$coefficients, caption = ‘OLS Coefficients’,digits = c(0,5,5,5,5)), comment = FALSE) “‘

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.00	0.01	0.00	1.00
Income	-0.60	0.02	-33.31	0.00
Limit	0.96	0.16	5.82	0.00
Rating	0.38	0.17	2.32	0.02
Cards	0.05	0.01	4.08	0.00
Age	-0.02	0.01	-2.09	0.04
Education	-0.01	0.01	-0.69	0.49
GenderFemale	-0.01	0.01	-1.07	0.28
StudentYes	0.28	0.01	25.46	0.00
MarriedYes	-0.01	0.01	-0.82	0.41
EthnicityAsian	0.02	0.01	1.19	0.23
EthnicityCaucasian	0.01	0.01	0.83	0.41

Table 1: OLS coefficients