# STAT 8330 FALL 2015 ASSIGNMENT 3

*Peng Shao*

*September 22, 2015*

▶ **Exercises 5.8. Solution.**

(a). The code for spliting data is list at the end of this assignment.

(b). The test MSE for linear regression model is $1.1085313 \times 10^6$.

(c). The test MSE for ridge regression model is $1.0545268 \times 10^6$.

(d). The test MSE for lasso regression model is $1.0395033 \times 10^6$, and the number of non-zero coefficient estimates is 13.

(e). The test MSE for PCR model is $1.3256164 \times 10^6$, and $M = 16$.

(f). The test MSE for PLS model is $1.2799224 \times 10^6$, and $M = 16$.

(g). Compared to the standard deviation of variables Apps, which is $1.497846 \times 10^7$, the standard errors of five model is not so big, which are at most less than 10% of $sd(Apps)$. So the model the prediction accuracy of the models is good enough. This is not surprising. If we compute the correlation between the variable Apps, and the rest variables (as shown below), there are three variables – Accept, Enroll, F.Undergrad – which have really large cofficients of correlation, so linear model is reasonable choice. Comparing within these five models, ridge regression and lasso regression are about 5%-7% better than the traditional linear regression, while the PCR and PLS are 10% worse than the traditional linear regression. Furthermore, the lasso regression is a little better than the ridge regression, but not so significantly.

```
##      Accept      Enroll F.Undergrad
##   0.9434506   0.8468221   0.8144906
```

▶ **Exercises 2. Solution.**

(a). The code for spliting data is list at the end of this assignment

(b). The test MSE for linear regression model is 59.3651103.

(c). Results for best subset selection are listed below.

```
## $`Number of Variables`
## [1] 3
##
## $`Name of Variables`
## [1] "(Intercept)" "rad"         "lstat"
##
## $`Coefficients of Variables`
## (Intercept)         rad       lstat
##  -3.7604819   0.4750033   0.2041807
##
## $`test MSE`
## [1] 55.89099
```

(d). Results for ridge regression are listed below.

```
## $Lambda
## [1] 0.5412185
##
## $`test MSE`
## [1] 58.24609
```

(e). Results for lasso regression are listed below.

```
## $Lambda
## [1] 0.2512114
##
## $`test MSE`
## [1] 56.26698
##
## $`Non-zero Coefficient Estimates`
##  (Intercept)           zn           rm          dis          rad
## -4.102999495  0.008212353  0.736684216 -0.177832783  0.430151095
##        lstat         medv
##  0.129211862 -0.104816617
##
## $`Name of Variables with Zero Coefficient Estimates`
## [1] "indus"   "chas"    "nox"     "age"     "tax"     "ptratio" "black"
```

(f). The test MSE for PCR model is 59.3651103, and $M = 13$.

(g). The test MSE for PLS model is 59.2842902, and $M = 9$.

(h). Compared to the standard deviation of variables crim, which is 73.9865782, the standart errors of these six models is fairly large, roughly about 80% of $sd(crim)$. They are not much better than predicting the response without any model just by using the distribution of response. The differences between the performance of these models is not apparent, and best model is the linear regression just by using the exhaustive best subset selection. To be noticed, the predictors in the best model is not the variables with top two highest coefficient of correlation. We should think that it probably indicates that there exists some multicollinearity problems in this dataset. Hence, the multicollinearity makes the model selection and shrinkage for linear model difficult, and the small coefficients of correlation worsen this situation.

```
##     indus       nox       rad       tax     lstat
## 0.4065834 0.4209717 0.6255051 0.5827643 0.4556215
```

▶ **Exercises 3. Solution.**

From the results below, we can see that the standard deviantion of $\log(area + 1)$ is 1.398436, while the least test MSE is 1.9282168 of the ridge regression, which has more variation than the data itself. Even though we can say that the best model among these model is the ridge regression based on the CV MSE and the test MSE, but it is very difficult to say it is a useful model, since it has lower accuracy for prediction than just guessing based the distribution itself. Actually, we may think that the linear regression model is not suitable for this data set. To verify it, we can compute the correlations between the response and predictor, and none of the coefficients of correlation is larger than 0.01. Thus, all the predictors we use seem more likely to be noise features. This can be shown in the result of PCR and PLS. The NAs of test MSE of PCR and PLS are not some computing problem, it is just because the best model based on this two shrinkage method is the model with only intercept, which means that none of these variables can explain the variation of response.

On the other hand, it is obviously we need a scaling process before doing PCR and PLS because the ranges of different variables varies so much at a first glance of the summary of the data set. However, this data also has two categorical variable, and it is meaningless to scale the categorical variable of indicator variable. I cannot figure a reasonable to deal with this issue.

There is another issue in this data set which needs to be noticed. The observation is so unbalanced with respect to the time variable, i.e., some months have few observation. This is why we cannot perform a best subset selection for linear regression based on CV. For example, there is only one observation in November in the data set. If the cross-validation put this observation into test set, then there is no observation in November in training set, and model fitting will not estimate the coefficient of the indicator variable for

November. Then we will have some problem in predicting the test set using the model because we do not know how to deal with the November indicator variable.

```
## $`standard deviation of 'log_area'`
## [1] 1.398436
##
## $ridge.cv
## [1] 1.979572
##
## $lasso.cv
## [1] 1.981853
##
## $pcr.cv
## [1] 1.987031
##
## $plsr.cv
## [1] 1.987031
##
## $lm.mse
## [1] 2.226952
##
## $ridge.mse
## [1] 1.928217
##
## $lasso.mse
## [1] 1.930522
##
## $pcr.mse
## [1] NA
##
## $plsr.mse
## [1] NA
```

```
Correlations
```

```
##           X           Y       month         day        FFMC         DMC
##  0.06199491  0.03883821  0.03997446  0.02881217  0.04679856  0.06715274
##          DC         ISI        temp          RH        wind        rain
##  0.06635976 -0.01034688  0.05348655 -0.05366216  0.06697349  0.02331131
```

# Appendix

```
# 6.9(a)
set.seed(1)
train.ind <- sample(1:nrow(College), nrow(College)/2)
test.ind <- -train.ind
College.train <- College[train.ind, ]
College.test <- College[test.ind, ]
# 2. (a)
set.seed(1)
train=sample(c(TRUE,FALSE), nrow(Boston),rep=TRUE)
train.Boston <- Boston[train, ]
test.Boston <- Boston[!train, ]
```