SMM637 Answers - Shrinkage Methods

In this exercise, we will predict the number of applications received using the other variables in the `College` data set.

1. Split the data set into a training set and a test set.

```
library(ISLR)
set.seed(11)

train.size <- dim(College)[1] / 2
train <- sample(1:dim(College)[1], train.size)
test <- -train
College.train <- College[train, ]
College.test <- College[test, ]
```

2. Fit a linear model using least squares on the training set, and report the test error obtained.

```
lm.fit <- lm(Apps ~., data = College.train)
lm.pred <- predict(lm.fit, College.test)
mean((College.test[, "Apps"] - lm.pred)^2)
```

   Test RSS is 1,026,096.

3. Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.

   Pick λ using `College.train` and report error on `College.test`

```
library(glmnet)
train.mat <- model.matrix(Apps ~., data = College.train)
test.mat <- model.matrix(Apps ~., data = College.test)
grid = 10^seq(4, -2, length = 100)
mod.ridge = cv.glmnet(train.mat, College.train[, "Apps"], alpha = 0,
                      lambda = grid, thresh = 1e-12)
lambda.best = mod.ridge$lambda.min
lambda.best

ridge.pred <- predict(mod.ridge, newx = test.mat, s = lambda.best)
mean((College.test[, "Apps"] - ridge.pred)^2)
```

   Test RSS is slightly lower than that of OLS, 1,026,069.

4. Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

   Pick λ using `College.train` and report error on `College.test`

```
mod.lasso <- cv.glmnet(train.mat, College.train[, "Apps"], alpha = 1, lambda = gr
lambda.best <- mod.lasso$lambda.min
lambda.best
lasso.pred = predict(mod.lasso, newx = test.mat, s = lambda.best)
mean((College.test[, "Apps"] - lasso.pred)^2)
```

Again, Test RSS is slightly lower than that of OLS, 1,026,036.
The coefficients look like

```
mod.lasso <- glmnet(model.matrix(Apps ~., data = College), College[, "Apps"], alp
predict(mod.lasso, s = lambda.best, type = "coefficients")
```

5. Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these 3 approaches?