Homework #5

Eric Pettengill

We will predict the number of applications received using the other variables in the `College` data set

Tables for the test MSE, model parameters used(λ, M), and coefficients of the final model are printed in part (g).

```
library(ISLR)
library(tidyverse)
library(glmnet)
library(caret)
library(pls)

college <- College %>%
  filter(PhD <= 100) %>%
  filter(Grad.Rate <= 100)
```

(a) Split the data into a training and test set

```
set.seed(1110)
trn <- createDataPartition(college$Apps,
                           p = 0.5,
                           list = FALSE)

x <- model.matrix(Apps ~ ., data = college)[ ,-1]
y <- college$Apps

college.train <- college[trn, ]
college.test <- college[-trn, ]

x.train <- x[trn, ]
x.test <- x[-trn, ]
y.train <- y[trn]
y.test <- y[-trn]
```

(b) Fit a linear model using least squares on the training set, and report the test error.

```
# Linear model
lm.fit <- lm(Apps ~ ., data = college.train)

# predictions
lm.pred <- predict(lm.fit, college.test)

# MSE
lm.mse <- mean((lm.pred - college.test$Apps)^2)

# Final model coefficients on all data
lm.coef <- round(coef(lm(Apps ~ ., data = college)), 2)
```

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

```
set.seed(1110)

grid <- 10^seq(10, -2, length = 100)

# Ridge reg. on training
ridge.fit <- glmnet(x.train, y.train, alpha = 0, lambda = grid)

# CV for lambda
ridge.fit.cv <- cv.glmnet(x.train, y.train, alpha = 0, nfolds = 10)

# lambda value
ridge.lambda <- ridge.fit.cv$lambda.min

# predictions
ridge.pred <- predict(ridge.fit, s = ridge.lambda, newx = x.test)

# MSE
ridge.mse <- mean((ridge.pred - y.test)^2)

# Final model coefficients on all data
ridge.coef <- round(coef(glmnet(x, y, alpha = 0, lambda = ridge.lambda)),2)
```

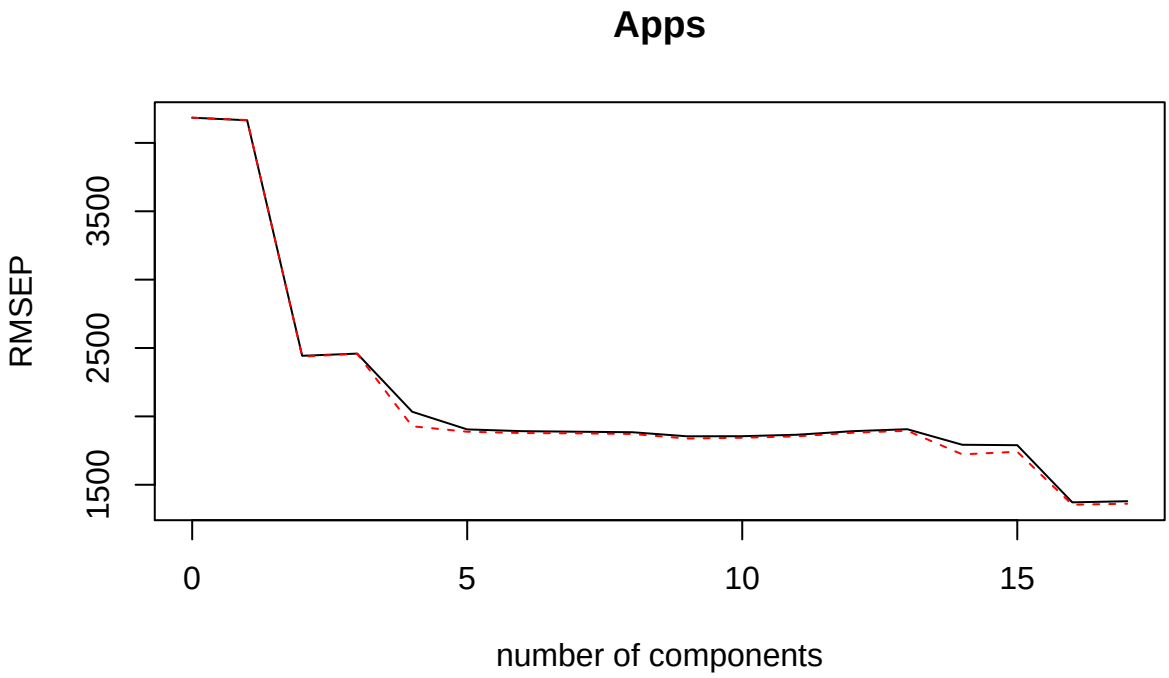
(d) Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

```
set.seed(1110)

# Lasso reg. on training
lasso.fit <- glmnet(x.train, y.train, alpha = 1, lambda = grid)

# CV for lambda
lasso.fit.cv <- cv.glmnet(x.train, y.train, alpha = 1, nfolds = 10)

# lambda value
lasso.lambda <- lasso.fit.cv$lambda.min

# predictions
lasso.pred <- predict(lasso.fit, s = lasso.lambda, newx = x.test)

# MSE
lasso.mse <- mean((lasso.pred - y.test)^2)

# Final model coefficients on all data
lasso.coef <- round(coef(glmnet(x, y, alpha = 1, lambda = lasso.lambda)),2)
```

(e) Fit a PCR model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

Below is the validation plot for principal components regression, as we can see the MSE for the cross validated value for M, the number of PC's, starts to level off around 10. The `summary(pcr.fit)` function chooses this number of PC's for us.

```
set.seed(1110)

# PCR on training
pcr.fit <- pcr(Apps ~ ., data = college.train, scale = TRUE, validation = "CV")
validationplot(pcr.fit)
```




```
# predictions
pcr.pred <- predict(pcr.fit, x.test, ncomp = 10)

# MSE
pcr.mse <- mean((pcr.pred - y.test)^2)

# Final model on all data
pcr.final <- pcr(Apps ~ ., data = college, scale = TRUE, ncomp = 10)
pcr.coef <- pcr.final$coefficients
```

(f) Fit an elastic net 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.

```
set.seed(1110)

## Elastic net on training
elnet.fit <- glmnet(x.train, y.train, alpha = 0.5, lambda = grid)

# CV for lambda
elnet.fit.cv <- cv.glmnet(x.train, y.train, alpha = 0.5, nfolds = 10)

# lambda value
elnet.lambda <- elnet.fit.cv$lambda.min

# predictions
elnet.pred <- predict(elnet.fit, s = elnet.lambda, newx = x.test)

# MSE
elnet.mse <- mean((elnet.pred - y.test)^2)

# Final model coefficients on all data
elnet.coef <- round(coef(glmnet(x, y, alpha = 0.5, lambda = elnet.lambda)),2)
```

(g) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much diference among the test errors resulting from these five approaches?

Table 1 contains the test MSE for each model used above, as well as the parameter used in each model, chosen by cross validation (λ for Ridge, lasso, and elastic net and M for PCR). Table 2 contains the coefficients for all models considered.

As we can see Ridge regression gives the lowest test MSE followed by Lasso, Elastic net, Linear model, and PCR is surprisingly much larger than the others. Plots of the MSE as a function of λ for the Ridge, Lasso, and Elastic net models is shown.

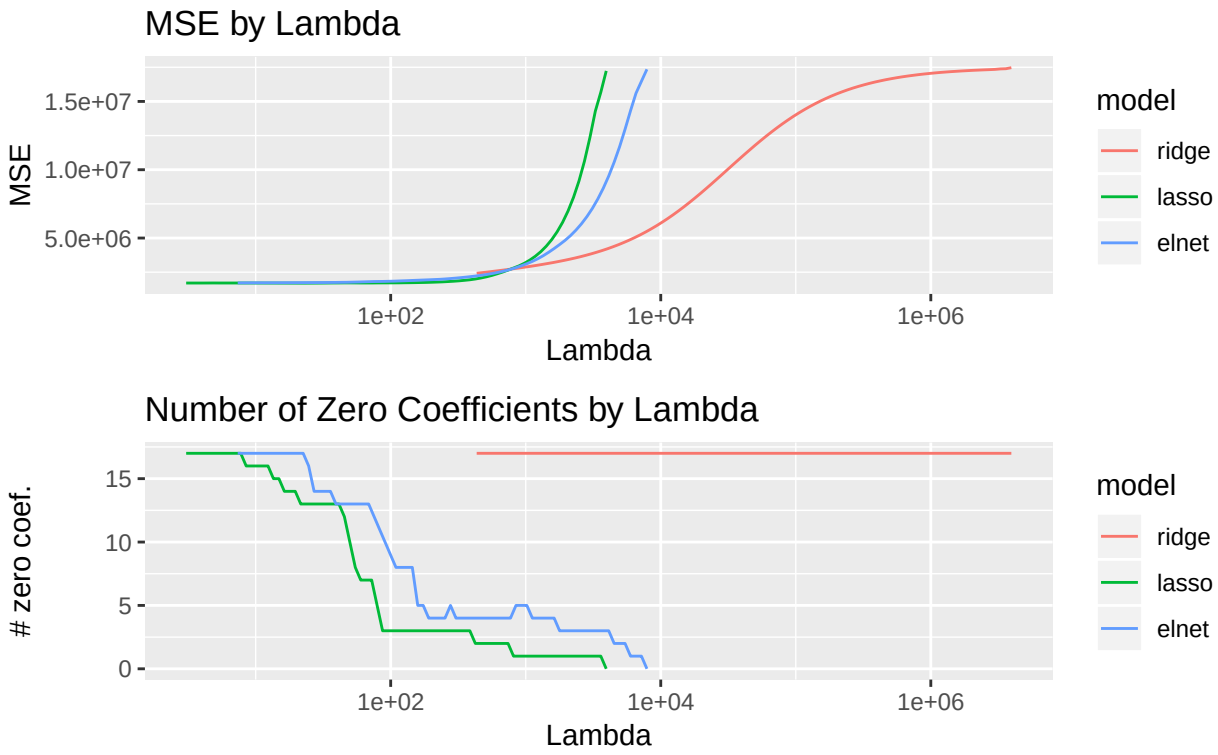
The number of zero coefficients by λ for the cross validated Ridge, Lasso, and Elastic net models are also shown below. This is not the same as the coefficients in table 2. Notice that as the value of λ increases the number of coefficients that goes to zero decreases for the Lasso and Elastic net, with the Elastic net consistently having more zero coefficients. The Ridge model however shrunk almost all coefficients in the cross validation step. From the final model coefficients in table 2, we can see that the Lasso model is the only one to shrink a coefficient to zero, having 3, while the Ridge and Elastic Net have some coefficients very close to zero.

	Model	Parameter	MSE
1	LM		1011849.30
2	Ridge	433.32	968894.16
3	Lasso	21.57	997565.86
4	Elastic Net	7.36	1001448.97
5	PCR	10.00	1588471.66

Table 1: Model Parameters/MSE

	LM	Ridge	Lasso	Elastic Net	PCR
(Intercept)	-435.35	-1552.06	-595.30	-493.22	-219.62
PrivateYes	-499.24	-532.33	-427.28	-496.02	1313.67
Accept	1.59	0.96	1.46	1.55	1182.51
Enroll	-0.88	0.50	-0.21	-0.65	191.86
Top10perc	49.84	24.24	33.85	46.40	137.72
Top25perc	-14.30	1.46	-2.66	-11.72	1054.14
F.Undergrad	0.06	0.08	0.00	0.03	-192.49
P.Undergrad	0.05	0.02	0.02	0.04	198.35
Outstate	-0.09	-0.02	-0.06	-0.08	342.43
Room.Board	0.15	0.20	0.13	0.15	9.20
Books	0.03	0.14	0.00	0.02	-71.74
Personal	0.03	-0.01	0.00	0.02	-112.90
PhD	-9.37	-3.67	-6.15	-8.78	-128.41
Terminal	-3.20	-4.63	-3.15	-3.12	-23.22
S.F.Ratio	15.42	12.63	4.48	13.84	-232.06
perc.alumni	0.09	-9.09	-1.02	-0.44	341.03
Expend	0.08	0.07	0.07	0.08	220.62
Grad.Rate	9.39	11.63	5.74	8.77	-219.62

Table 2: Model Coefficients


MSE by Lambda
MSE
1.5e+07
1.0e+07
5.0e+06
1e+02
1e+04
1e+06
Lambda
model
ridge
lasso
elnet

Number of Zero Coefficients by Lambda
zero coef.
15
10
5
0
1e+02
1e+04
1e+06
Lambda
model
ridge
lasso
elnet