HW08

Zixin Ouyang

11/6/2017

Exericse 1

```
library(readr)
leukemia = read_csv("/Users/Constance/Downloads/leukemia.csv", progress = FALSE)
```

```
y = as.factor(leukemia$class)
X = as.matrix(leukemia[, -1])
```

```
library(glmnet)
```

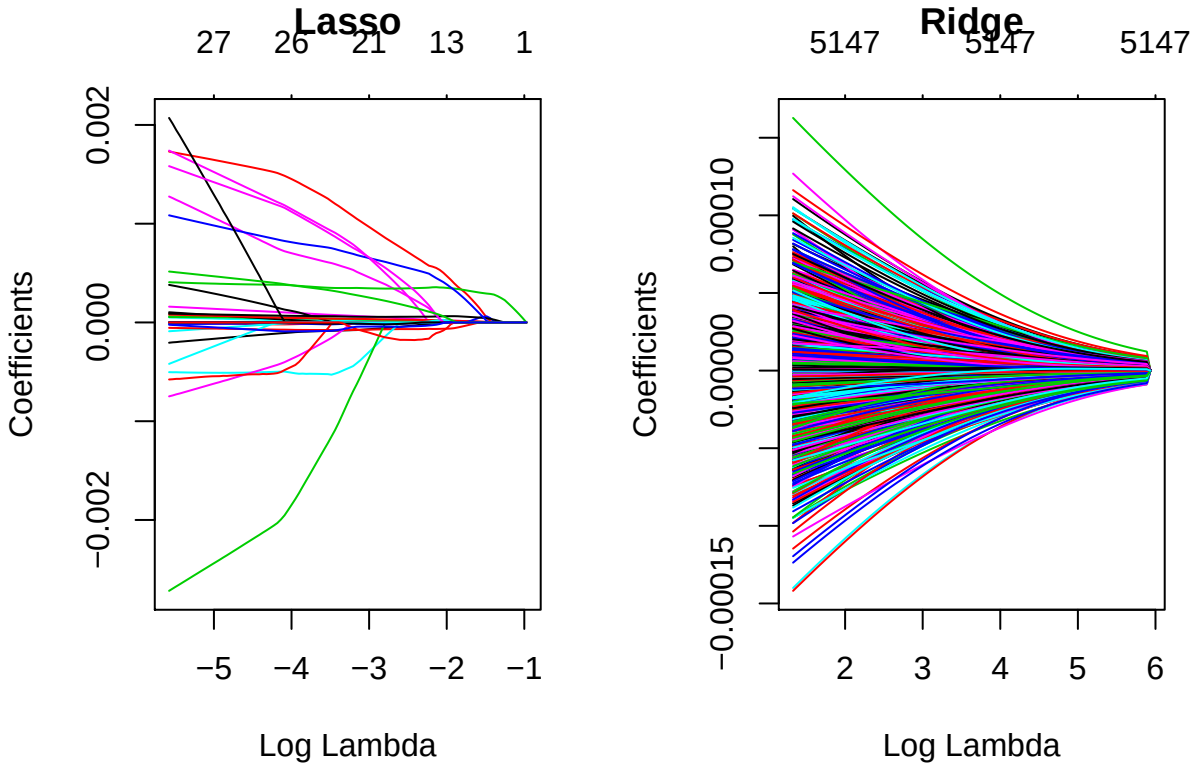
```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-13
```

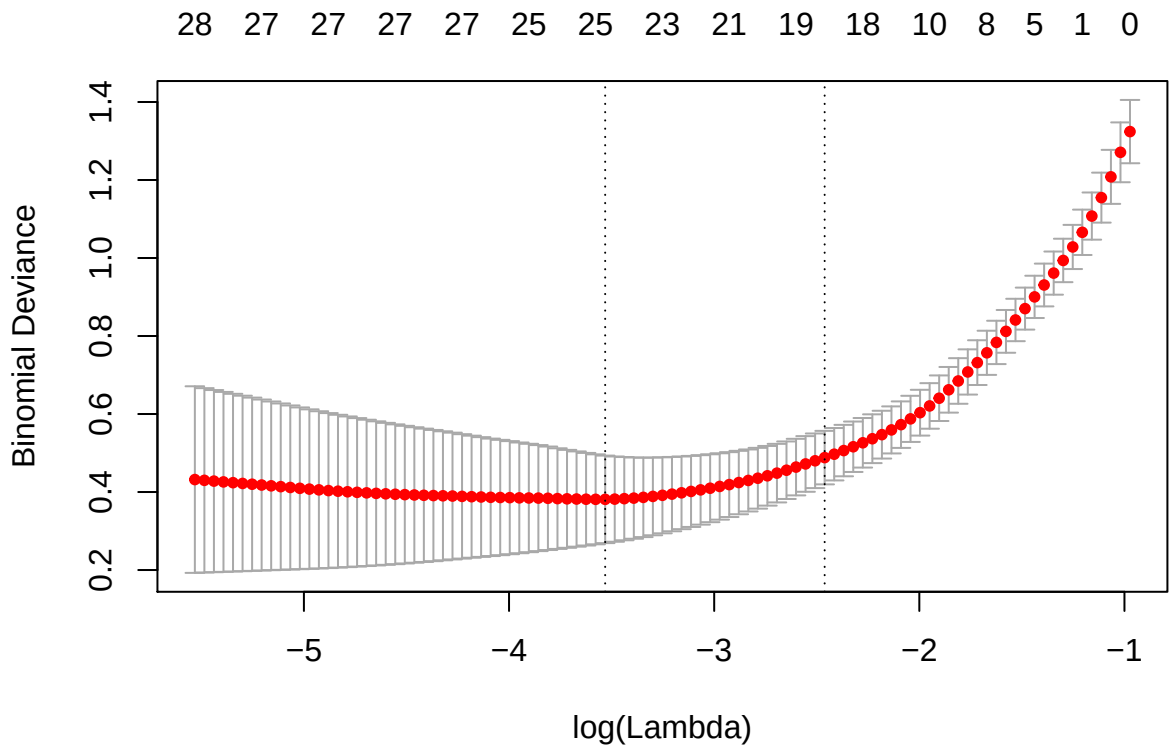
```
fit_lasso = glmnet(X, y, alpha = 1, family="binomial")
fit_ridge = glmnet(X, y, alpha = 0, family="binomial")
```

```
par(mfrow = c(1, 2))
plot(fit_lasso, xvar = "lambda", main = "Lasso")
plot(fit_ridge, xvar = "lambda", main = "Ridge")
```



```
fit_cv = cv.glmnet(X, y, family = "binomial", alpha = 1)
```

```
plot(fit_cv)
```




```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

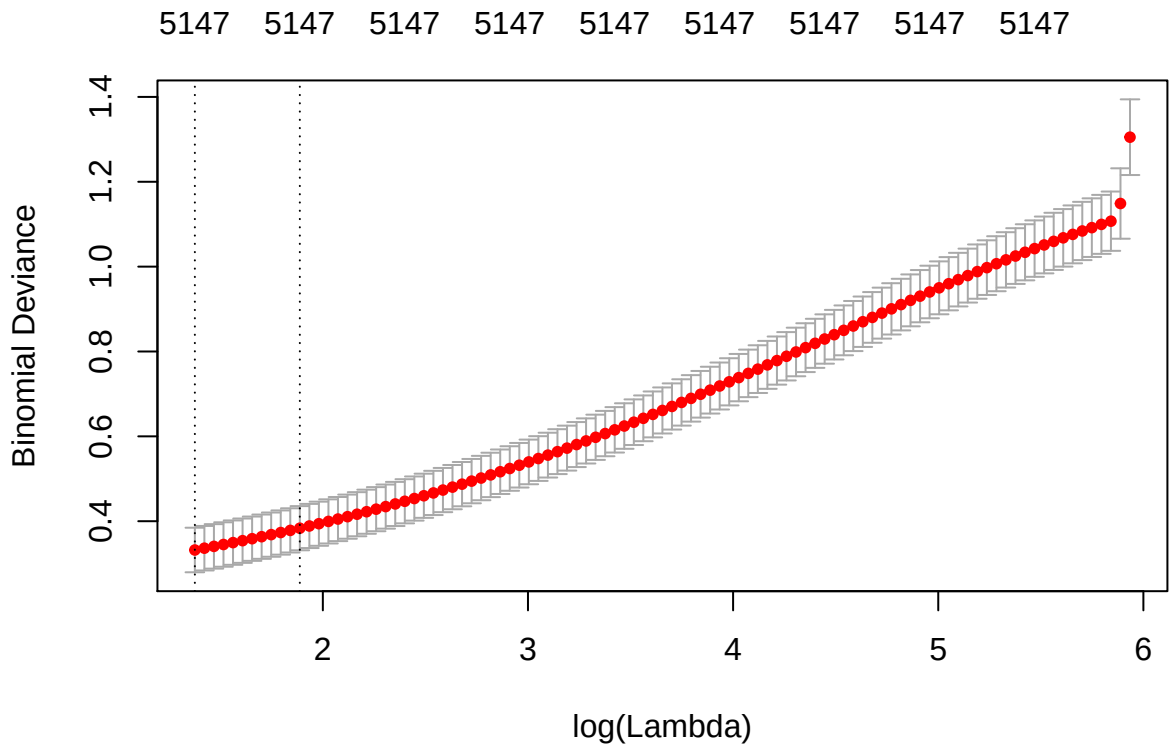
```
cv_5 = trainControl(method = "cv", number = 5)
lasso_grid = expand.grid(alpha = 1,
                         lambda = c(fit_cv$lambda.min,
                                    fit_cv$lambda.1se))
```

```
set.seed(659017838)
fit_lasso = train(X, y,
  method = "glmnet",
  trControl = cv_5,
  tuneGrid = lasso_grid
)
```

```
lasso_accuracies = fit_lasso$results$Accuracy
```

```
fit_cv2 = cv.glmnet(X, y, family = "binomial", alpha = 0)
```

```
plot(fit_cv2)
```




```
ridge_grid = expand.grid(alpha = 0,
                         lambda = c(fit_cv2$lambda.min, fit_cv2$lambda.1se))
```

```
set.seed(659017838)
fit_ridge = train(X, y,
  method = "glmnet",
  trControl = cv_5,
  tuneGrid = ridge_grid
)
```

```
ridge_accuracies = fit_ridge$results$Accuracy
```

```
set.seed(659017838)
fit_knn = train(X, y,
  method = "knn",
  trControl = cv_5,
  preProcess = c("center", "scale")
)
```

```
knn_accuracies = fit_knn$results$Accuracy
```

Model	Cross-Validated Accuracy	Standard Deviation
fit_lasso	0.9152381	0.0600831
fit_lasso	0.9009524	0.0818854
fit_ridge	0.9428571	0.0931315
fit_ridge	0.9428571	0.0931315
fit_knn	0.8885714	0.0637633
fit_knn	0.8742857	0.0935929
fit_knn	0.832381	0.0956776

Exercise 2

```
set.seed(42)
library(caret)
library(ISLR)
index = createDataPartition(College$Outstate, p = 0.75, list = FALSE)
college_trn = College[index, ]
college_tst = College[-index, ]
```

```
set.seed(659017838)
linear_mod = train(
  Outstate ~ .,
  data = college_trn,
  method = "lm",
  trControl = trainControl(method = "cv", number = 5)
)
```

```
set.seed(659017838)
elnet1 = train(
  Outstate ~ . ,
  data = college_trn,
  method = "glmnet",
  trControl = cv_5,
  tuneLength = 10
)
```

```
set.seed(659017838)
elnet2 = train(
  Outstate ~ .^2 ,
  data = college_trn,
  method = "glmnet",
  trControl = cv_5,
  tuneLength = 10
)
```

```
set.seed(659017838)
knn_mod = train(
  Outstate ~ .,
  data = college_trn,
  method = "knn",
  trControl = trainControl(method = "cv", number = 5),
  preProcess = c("center", "scale")
)
```

```
set.seed(659017838)
knn_mod2 = train(
  Outstate ~ .^2,
  data = college_trn,
  method = "knn",
  trControl = trainControl(method = "cv", number = 5),
  preProcess = c("center", "scale")
)
```

```
library(randomForest)
set.seed(659017838)
```

```
rf_mod = train(
  Outstate ~ .,
  data = college_trn,
  method = "rf",
  trControl = trainControl(method = "cv", number = 5)
)
```

```
get_best_result = function(caret_fit) {
  best = which(rownames(caret_fit$results) == rownames(caret_fit$bestTune))
  best_result = caret_fit$results[best, ]
  rownames(best_result) = NULL
  best_result
}
```

```
models = list(linear_mod, elnet1, elnet2, knn_mod, knn_mod2, rf_mod)
```

```
rmse = function(actual, predicted) {
  sqrt(mean((actual - predicted) ^ 2))
}
```

```
get_rmse = function(model, data, response) {
  rmse(actual = data[, response],
       predicted = predict(model, data))
}
```

```
test_rmse = sapply(models, get_rmse, data = college_tst, response = "Outstate")
```

Model	Cross-Validated RMSE	Test RMSE
linear_mod	1974.7423144	2069.0871599
elnet1	1970.6806276	2090.9905103
elnet2	1844.394725	1964.7392123
knn_mod	1905.530466	2007.5781377
knn_mod2	1971.5648606	2060.8032232
rf_mod	1749.1556016	1713.9861725

Exercise 3

```
uiuc_outstate = College[rownames(College)=="University of Illinois - Urbana",'Outstate']
```

Question	Answer
(a)	72 observations are in the dataset, and 5147 predictors are in the dataset.
(b)	Yes. We see a nice U-shaped CV error curve
(c)	No. This plot suggests that if we were to try smaller lambda, we could achieve a lower devience (error).
(d)	KNN performs worse. This is expected in a high-dimensional setting due to the curse of dimensionality.
(e)	The model with a ridge penalty should be choicen because it has the largest accuracy.
(f)	I prefer the random forest model because it has the smallest cross-validated RMSE and test RMSE.
(g)	The first elastic net model: alpha: 0.1, lambda: 82.9579489 The second elastic net model: alpha: 0.1, lambda: 220.8521303 Both have an alpha value of 0.1, so are in-between, but are closer to ridge.

Question	Answer
(h)	Yes. Yes. A lower error is found using scaled predictors.
(i)	Without interactions seems to work better. Adding all the interactions creates a high dimensional dataset.
(j)	The dataset is from year 1995, and the out-of-state tuition at UIUC at that time was 7560.