Lab 5: Regularization and Dimension Reduction

We will use the `Hitters` data set in the `ISLR` package to predict `Salary` for baseball players.

```
library(ISLR)
library(tidyverse)
library(knitr)

str(Hitters)
```

```
## 'data.frame':    322 obs. of  20 variables:
##  $ AtBat    : int  293 315 479 496 321 594 185 298 323 401 ...
##  $ Hits     : int  66 81 130 141 87 169 37 73 81 92 ...
##  $ HmRun    : int  1 7 18 20 10 4 1 0 6 17 ...
##  $ Runs     : int  30 24 66 65 39 74 23 24 26 49 ...
##  $ RBI      : int  29 38 72 78 42 51 8 24 32 66 ...
##  $ Walks    : int  14 39 76 37 30 35 21 7 8 65 ...
##  $ Years    : int  1 14 3 11 2 11 2 3 2 13 ...
##  $ CAtBat   : int  293 3449 1624 5628 396 4408 214 509 341 5206
...
##  $ CHits    : int  66 835 457 1575 101 1133 42 108 86 1332 ...
##  $ CHmRun   : int  1 69 63 225 12 19 1 0 6 253 ...
##  $ CRuns    : int  30 321 224 828 48 501 30 41 32 784 ...
##  $ CRBI     : int  29 414 266 838 46 336 9 37 34 890 ...
##  $ CWalks   : int  14 375 263 354 33 194 24 12 8 866 ...
##  $ League   : Factor w/ 2 levels "A","N": 1 2 1 2 2 1 2 1 2 1 ...
##  $ Division : Factor w/ 2 levels "E","W": 1 2 2 1 1 2 1 2 2 1 ...
##  $ PutOuts  : int  446 632 880 200 805 282 76 121 143 0 ...
##  $ Assists  : int  33 43 82 11 40 421 127 283 290 0 ...
##  $ Errors   : int  20 10 14 3 4 25 7 9 19 0 ...
##  $ Salary   : num  NA 475 480 500 91.5 750 70 100 75 1100 ...
##  $ NewLeague: Factor w/ 2 levels "A","N": 1 2 1 2 2 1 1 1 2 1 ...
```

0.1 Data Processing

1. Remove records with missing values from the data (Hint: `complete.cases()` is useful)

Use `model.matrix` to create an X matrix for all predictors that contains dummy variables for categorical predictors (for predicting `Salary`). You can specify this as a formula in the `model.matrix` call, e.g.

```
x <- model.matrix(y ~ ., data)[, -1] # remove the y column
```

3. Create a Y vector of `Salary` information.

0.2 Ridge Regression

The `glmnet()` function in the `glmnet` package can perform both ridge regression and the lasso. This is done with the specification of a parameter `alpha`. If `alpha = 0` then a ridge regression model is fit and if `alpha = 1` then the lasso is fit.

By default, `glmnet` performs ridge regression for an automatically selected range of values, but we can instead pass a vector of values.

1. Create a vector of λ values from $\lambda = .01$ to $\lambda = 10^10$ of length 100.
2. Fit a ridge regression model for each λ in your grid.

Note, by default `glmnet` will standardize the X variables.

3. Make a line plot of coefficient corresponding to each λ. You should have an individual line for each variable with coefficient value on the y-axis and λ on the x axis. What happens to your coefficients as λ increases?

4. Use `cv.glmnet` to perform 10-fold cross validation and get an estimate of the test MSE for each λ in your grid. Which λ would you choose and why?

0.3 Lasso

1. Fit the lasso model for each λ in your grid.

2. Make a line plot of coefficient corresponding to each λ. You should have an individual line for each variable with coefficient value on the y-axis and λ on the x axis. (Hint: `coef` may be a useful function). What happens to your coefficients as λ increases?

3. Use `cv.glmnet` to perform 10-fold cross validation and get an estimate of the test MSE for each λ in your grid. Which λ would you choose and why?

0.4 Principal Components Regression

The `pcr()` function in the `pls` package can perform principal components regression.

1. Fit the PCR model using the `pcr` command. A couple tips: a) setting `scale = TRUE` will standardize your data prior to fitting the model, and b) setting `validation = TRUE` will perform 10-fold cross validation for each value of M.

2. Create a plot of the CV MSE (note root MSE is reported) vs. M.

3. When does the smallest cross-validation error occur? Which M would you choose for your final model?

4. The `summary` function also provides the *percentage of variance explained* in the predictors and the response using M principal components. How many principal components would we need to explain at least 80% of the variability in the predictors?

5. How much variability in Y is explained for your chosen value of M?