# R Notebook

*Karen Mazidi*

Compare linear regression and ridge regression on the airquality data set.

### Data cleaning

First, remove rows with NAs using complete.cases(). Then remove the Day column.

```
df <- airquality[complete.cases(airquality[, 1:5]),]
df <- df[,-6]
```

### Train and test sets for linear regression

Divide into train and test sets, then create a model predicting Ozone from the other columns.

```
set.seed(1234)
i <- sample(1:nrow(df), .75*nrow(df), replace=FALSE)
train <- df[i,]
test <- df[-i,]
lm1 <- lm(Ozone~., data=train)
pred <- predict(lm1, newdata=test)
mse1 <- mean((pred-test$Ozone)^2)
print(paste("mse=", mse1))
```

```
## [1] "mse= 409.379992545846"
```

### Ridge Regression

Try ridge regession using glmnet.

First use the model.matrix() function to create a matrix of the predictors. Then split into test and train.

```
library(glmnet)
```

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

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

```
## Warning: package 'foreach' was built under R version 3.4.3
```

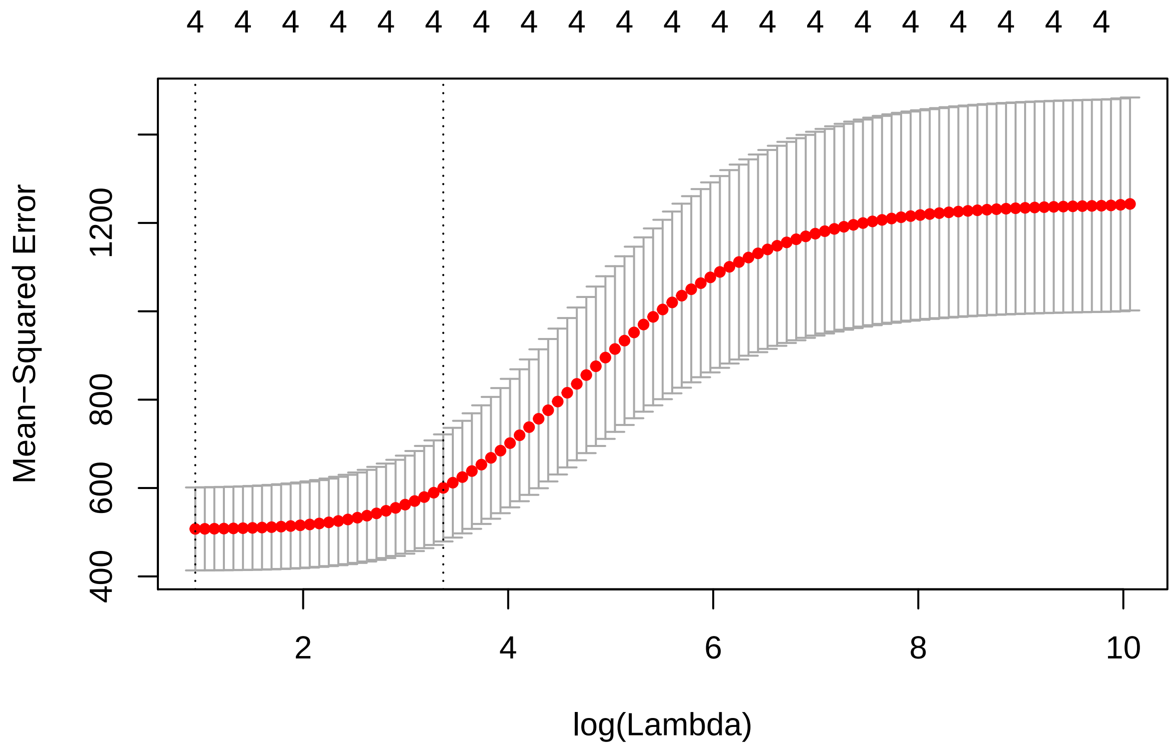
```
## Loaded glmnet 2.0-13
```

```
x <- model.matrix(Ozone~., df)[,-1]
y <- df$Ozone
train_x <- x[i,]
train_y <- y[i]
test_x <- x[-i,]
test_y <- y[-i]

# build a ridge regression model
rm <- glmnet(train_x, train_y, alpha=0)

# use cv to see which lambda is best
```

```r
set.seed(1)
cv_results <- cv.glmnet(train_x, train_y, alpha=0)
plot(cv_results)
```




```r
l <- cv_results$lambda.min

# get data for best lambda, which is the 99th
# as determined by looking at rm$lambda
pred2 <- predict(rm, s=l, newx=test_x)
mse2 <- mean((pred2-test_y)^2)
coef2 <- coef(rm)[,99]
```

### Compare mse and coefficients

The ridge regression got about 10% lower mse. Notice that its coefficients are smaller in absolute value.

```r
print(paste("mse for linear regression = ", mse1))
```

```
## [1] "mse for linear regression =  409.379992545846"
```

```r
coef(lm1)
```

```
##  (Intercept)      Solar.R         Wind         Temp        Month
## -66.85709002   0.08314323  -3.75229006   1.98524049  -3.27749222
```

```r
print(paste("mse for ridge regression = ", mse2))
```

```
## [1] "mse for ridge regression =  371.013801540814"
```

```r
coef2
```

```
##  (Intercept)      Solar.R         Wind         Temp        Month
## -60.80449134   0.08165752  -3.61256523   1.83183505  -2.60738344
```