

# Ridge Regression

*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= 442.280363620333"
```

$$RSS = \sum_{i=1}^n (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \|\mathbf{w}\|^2$$

## Ridge Regression

Try ridge regression 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
```

```
## Loaded glmnet 3.0-1
```

```
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
```

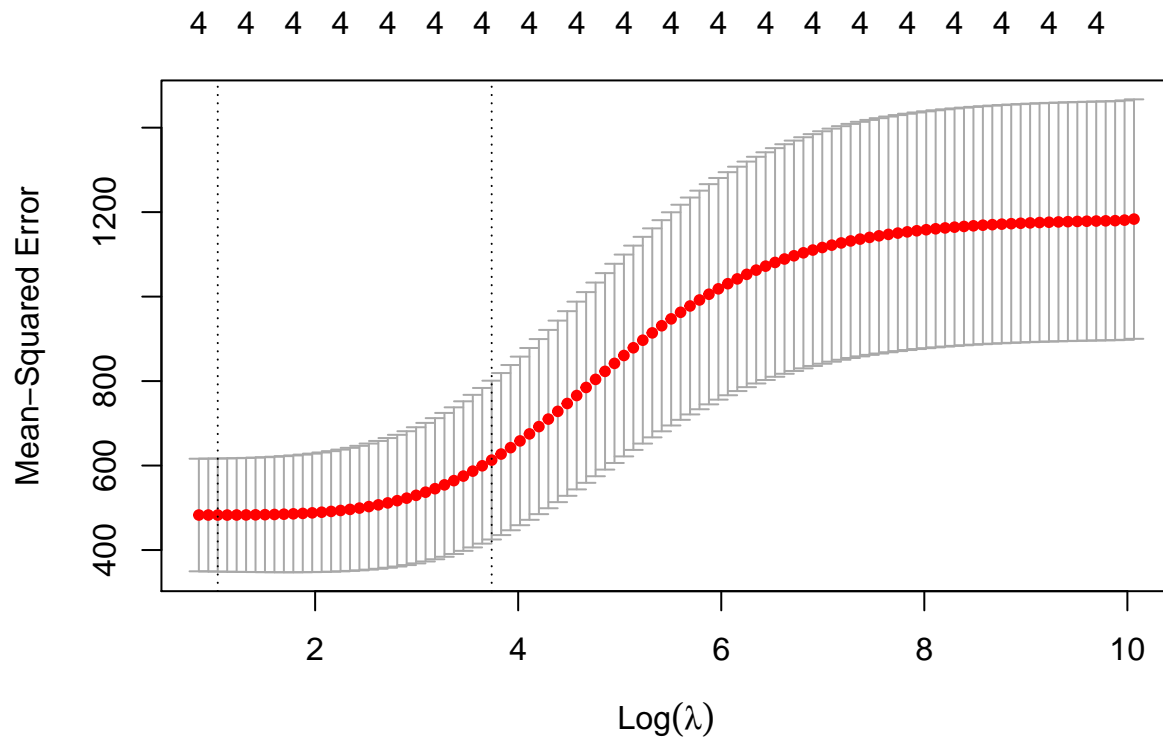
```
rm <- glmnet(train_x, train_y, alpha=0)
```

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

```
set.seed(1)
```

```
cv_results <- cv.glmnet(train_x, train_y, alpha=0)
```

```
plot(cv_results)
```



```
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.

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

```
## [1] "mse for linear regression = 442.280363620333"
```

```
coef(lm1)
```

```
## (Intercept)      Solar.R      Wind      Temp      Month
## -49.30679241  0.05146527 -3.79359396  1.82388553 -3.20124134
```

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

```
## [1] "mse for ridge regression = 433.811098270252"
```

```
coef2
```

```
## (Intercept)      Solar.R      Wind      Temp      Month
## -43.6687215  0.0526576 -3.6690376  1.6724750 -2.5421064
```