Using LASSO to select features

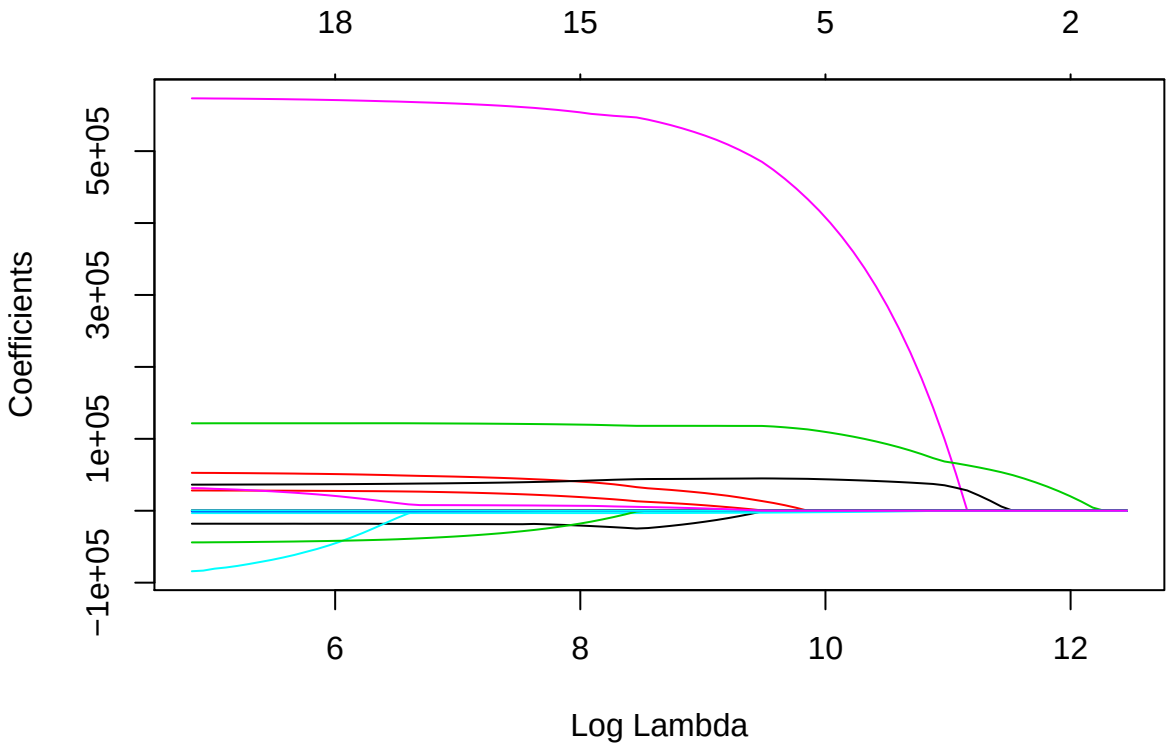
Guan-Yuan Wang

2020/9/9

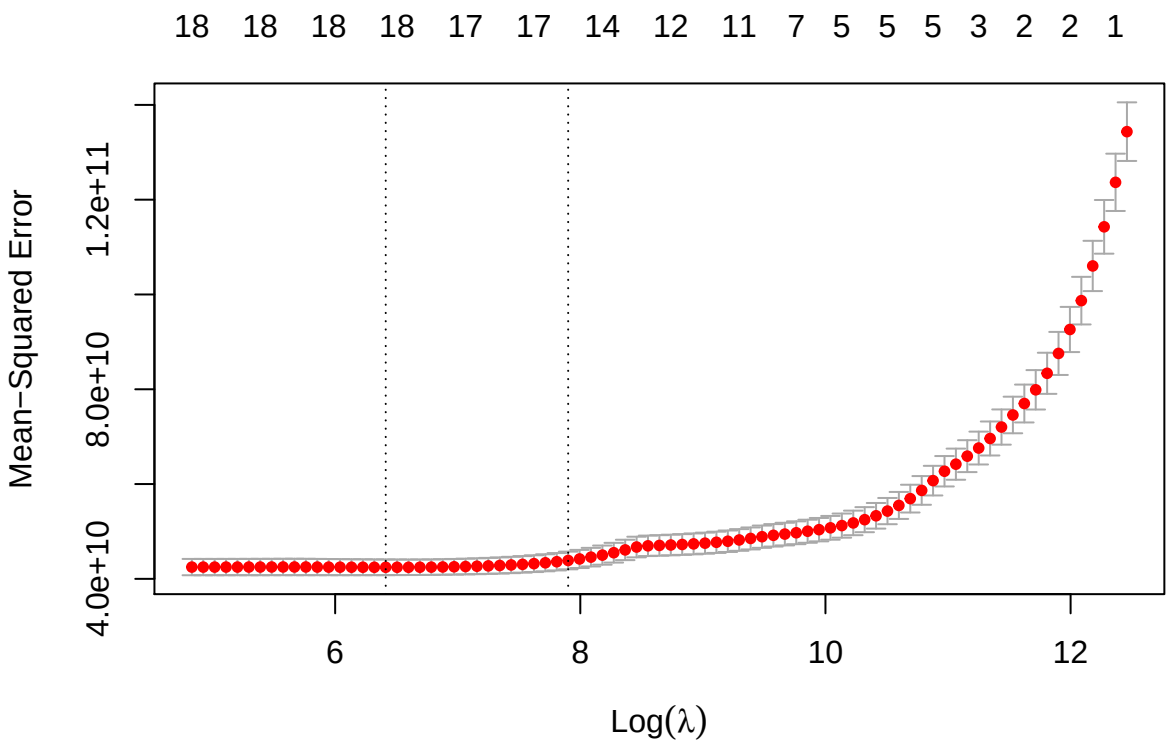
```
data <- read.csv("kc_house_data.csv")

### Create new features
data <- data %>%
  mutate(sqft_living_sqrt = sqrt(sqft_living),
         sqft_lot_sqrt = sqrt(sqft_lot),
         bedrooms_square = bedrooms * bedrooms,
         floors_square = floors * floors)

data <- data %>% select(!c(lat, long, zipcode, id, date))
```

lasso

```
x <- as.matrix(select(data, -price))
y <- data$price

lasso <- glmnet(x, y, alpha = 1)
plot(lasso, xvar = "lambda")
```




```
lassoTuning <- cv.glmnet(x, y, alpha = 1)
plot(lassoTuning)
```




```
lambdas <- 10^seq(2, 6, by = .5)
lasso_reg <- cv.glmnet(x, y, alpha = 1, lambda = lambdas,
                       standardize = TRUE, nfolds = 5)
lambda_best <- lasso_reg$lambda.min
lambda_best
```

```
## [1] 316.2278
```

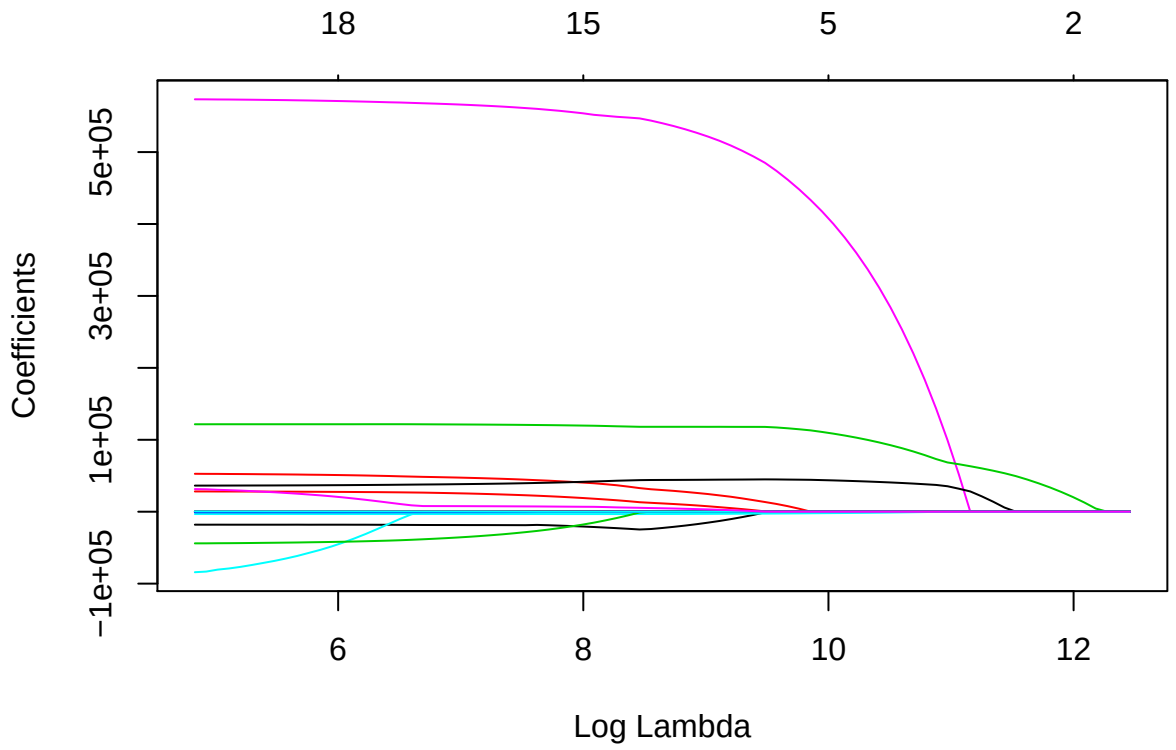
```
lassoModel <- glmnet(x, y, alpha = 1, lambda = lambda_best, standardize = TRUE)
lassoModel$beta
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                              s0
## bedrooms         -1.726392e+04
## bathrooms         5.143191e+04
## sqft_living       5.784472e+02
## sqft_lot          7.899482e-01
## floors           -6.639610e+04
## waterfront        5.715154e+05
## view              3.655980e+04
## condition         2.792739e+04
## grade             1.215865e+05
## sqft_above        .
## sqft_basement     1.163534e+01
## yr_built         -3.229322e+03
```

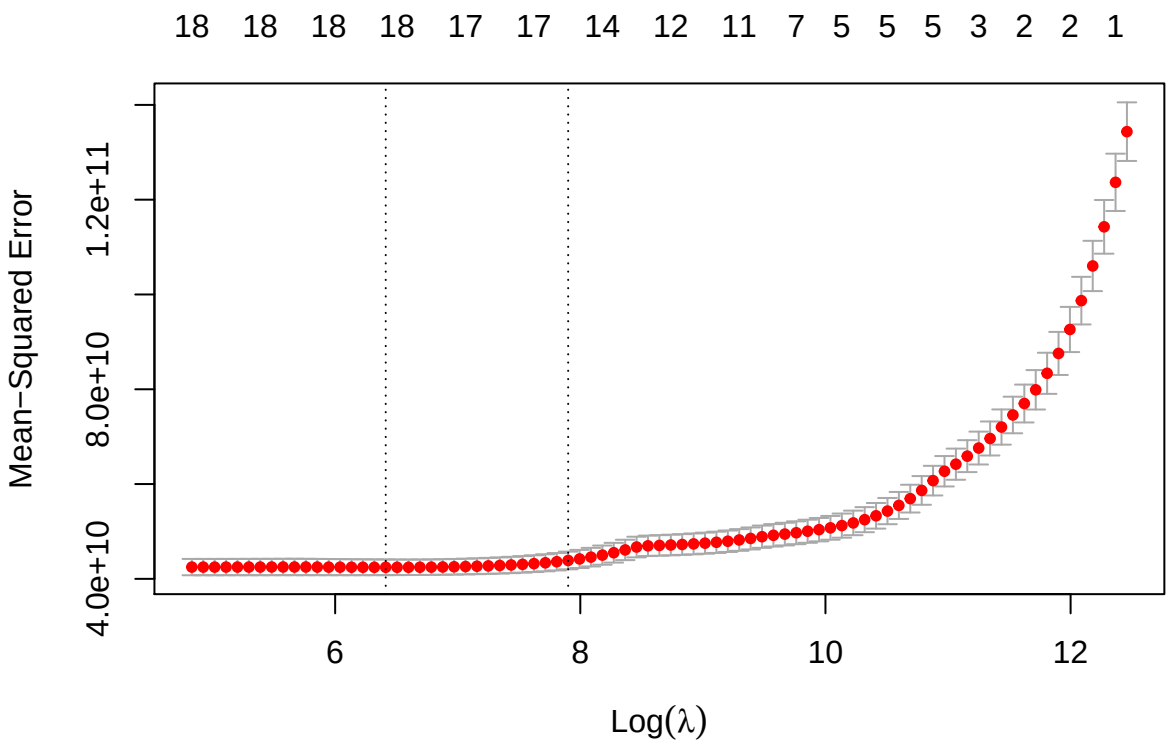
```
## yr_renovated      1.987511e+01
## sqft_living15     5.057571e+01
## sqft_lot15       -1.078246e-01
## sqft_living_sqrt -4.343205e+04
## sqft_lot_sqrt    -7.093889e+02
## bedrooms_square   3.617751e+02
## floors_square     2.651277e+04
```

ridge

```
x <- as.matrix(select(data, -price))
y <- data$price

ridge <- glmnet(x, y, alpha = 0)
plot(lasso, xvar = "lambda")
```




```
ridgeTuning <- cv.glmnet(x, y, alpha = 0)
plot(lassoTuning)
```




```
lambdas <- 10^seq(2, 6, by = .5)
ridge_reg <- cv.glmnet(x, y, alpha = 0, lambda = lambdas,
                       standardize = TRUE, nfolds = 5)
lambda_best <- ridge_reg$lambda.min
lambda_best
```

```
## [1] 1000
```

```
ridgeModel <- glmnet(x, y, alpha = 0, lambda = lambda_best, standardize = TRUE)
ridgeModel$beta
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                               s0
## bedrooms         -2.365978e+04
## bathrooms         5.241055e+04
## sqft_living       2.870146e+02
## sqft_lot          8.691286e-01
## floors           -1.000278e+05
## waterfront        5.759901e+05
## view              3.717815e+04
## condition         2.734347e+04
## grade             1.207724e+05
## sqft_above        2.364450e+02
## sqft_basement     2.432515e+02
## yr_built         -3.286474e+03
```

```
## yr_renovated      1.966144e+01
## sqft_living15     5.089858e+01
## sqft_lot15       -1.203348e-01
## sqft_living_sqrt -3.757624e+04
## sqft_lot_sqrt    -7.546226e+02
## bedrooms_square   7.111344e+02
## floors_square     3.549269e+04
```