hw5 Q11

Q11a.

Try out some of the regression methods explored in this chapter, such as best subset ### selection, the lasso, ridge regression, and PCR. Present and discuss results

```
library(MASS)
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
library("glmnet")
```

```
## Warning: package 'glmnet' was built under R version 4.1.3
```

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

```
## Loaded glmnet 4.1-3
```

```
library("leaps")
```

```
## Warning: package 'leaps' was built under R version 4.1.3
```

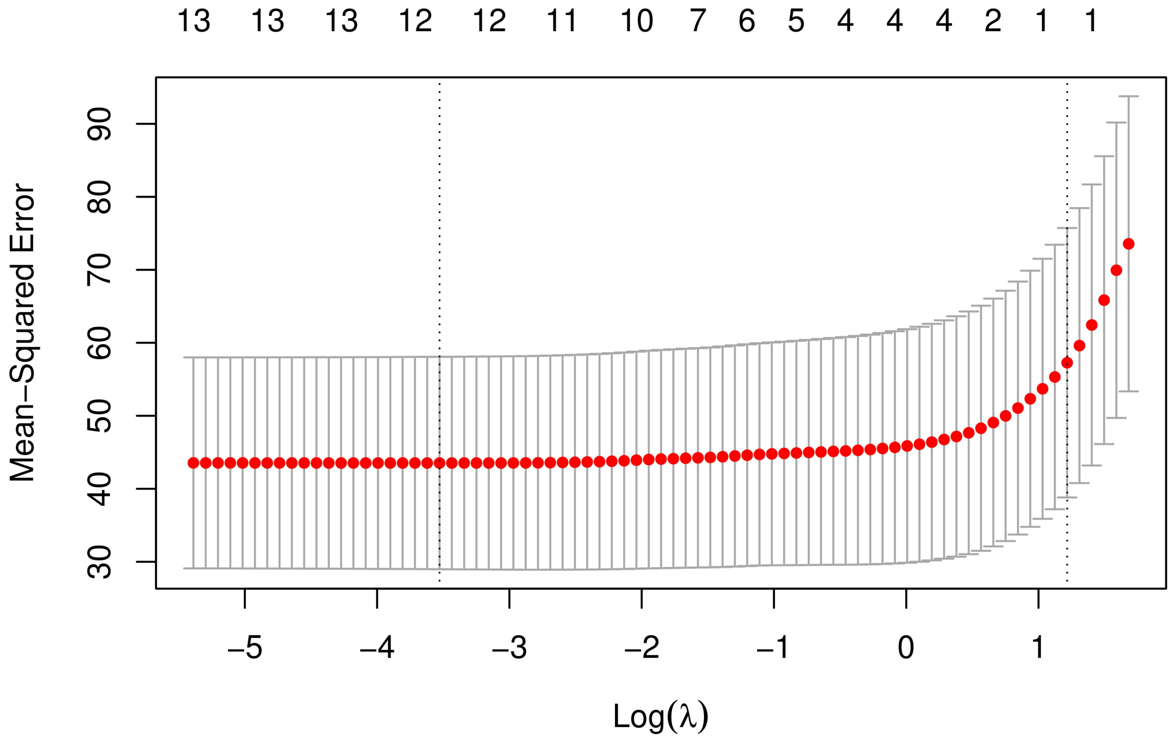
```
data(Boston)
# 506 obs 14 vars

x <- model.matrix(crim ~ . - 1, data = Boston)
y <- Boston$crim
lasso_method <- cv.glmnet(x, y, type.measure = "mse")
coef(lasso_method)
```

```
## 14 x 1 sparse Matrix of class "dgCMatrix"
##                    s1
## (Intercept) 1.4186414
## zn          .
## indus       .
## chas        .
## nox         .
## rm          .
## age         .
## dis         .
## rad         0.2298449
## tax          .
## ptratio      .
## black        .
## lstat        .
## medv         .
```

```
sqrt(lasso_method$cvm[lasso_method$lambda == lasso_method$lambda.1se])
```

```
## [1] 7.567599
```

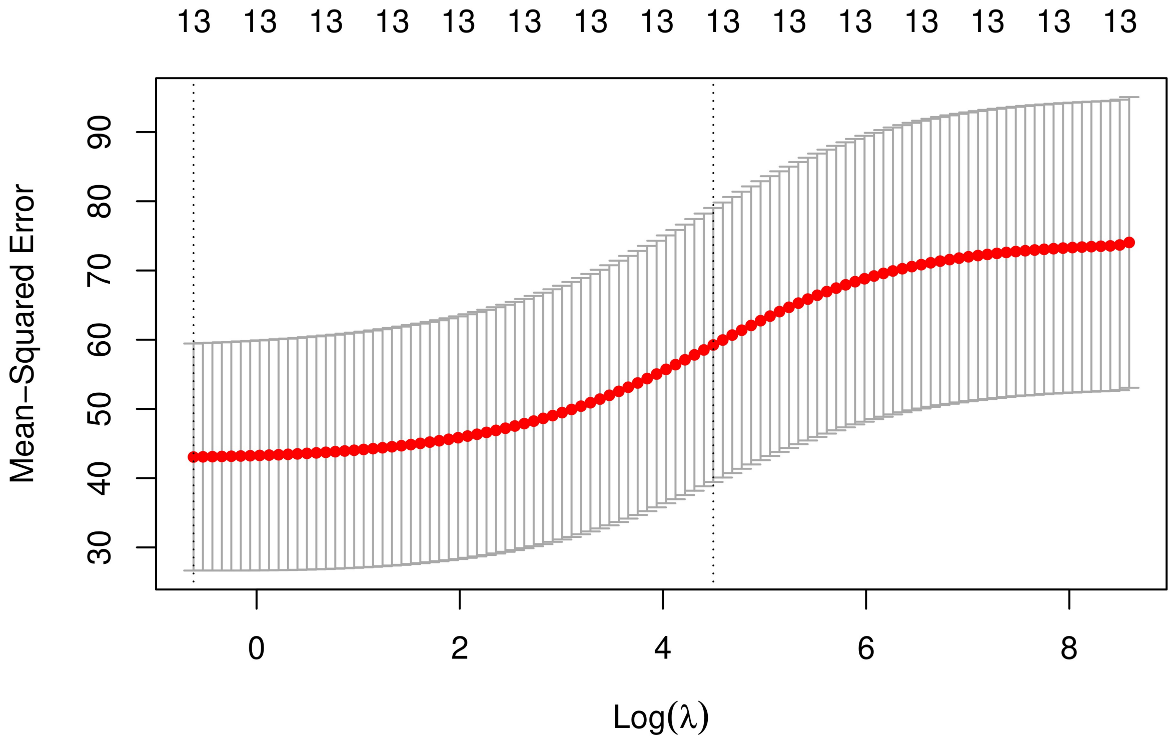
```
plot(lasso_method)
```



```
# Ridge Regression method
x <- model.matrix(crim ~ . - 1, data = Boston)
y <- Boston$crim
ridge_method <- cv.glmnet(x, y, type.measure = "mse", alpha = 0)
sqrt(ridge_method$cvm[ridge_method$lambda == ridge_method$lambda.1se])
```

```
## [1] 7.69679
```

```
plot(ridge_method)
```



The ridge error (7.78) is slightly higher than the lasso (7.39).

Q11b.

Propose a model (or set of models) that seem to perform well on

this data set, and justify your answer. Make sure that you are

evaluating model performance using validation set error, crossvalidation

From the models used above, the lasso seems to be the best method that performs well on this data set sincec the MSE is lower than the ridge model.

##11c. ### Does your chosen model involve all of the features in the data set? Why or why not? No, it does not. It contains all but the last column of variables because it had the lowest cross validated MSE.