Exercize 10 Section 6.8 ISL

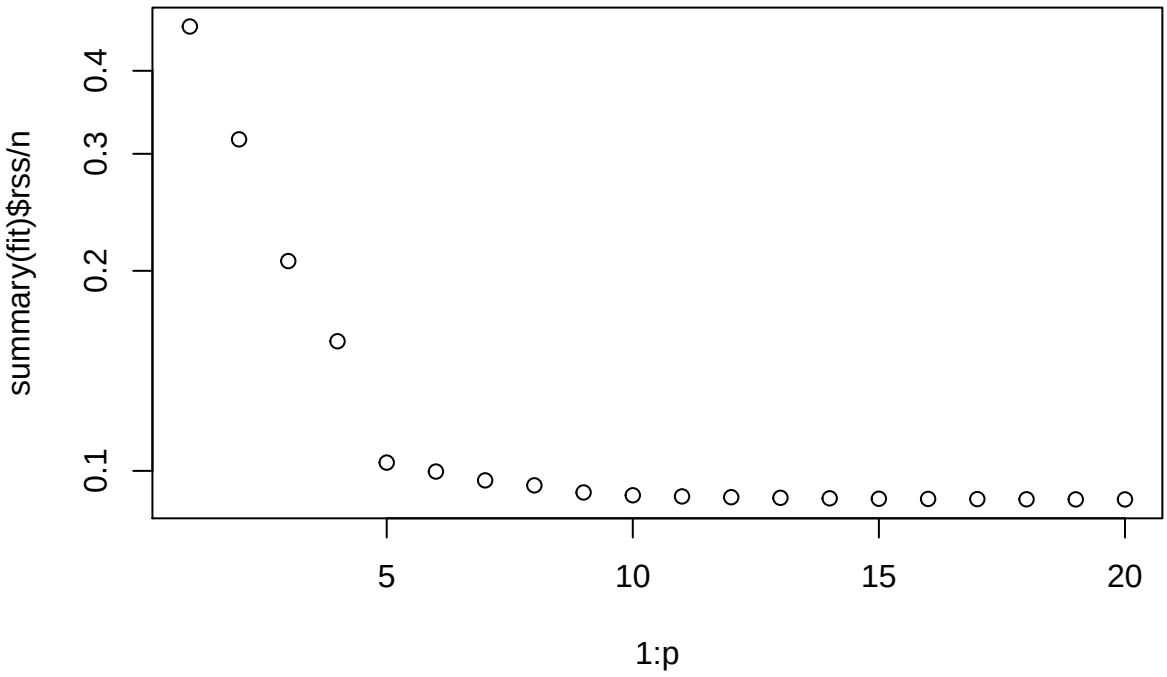
We have seen that as the number of predictors used in a model increases, the training error will necessarily decrease, but the test error may not. We will now explore this in a simulated data set.

(a) Generate a simulated data set as follows:

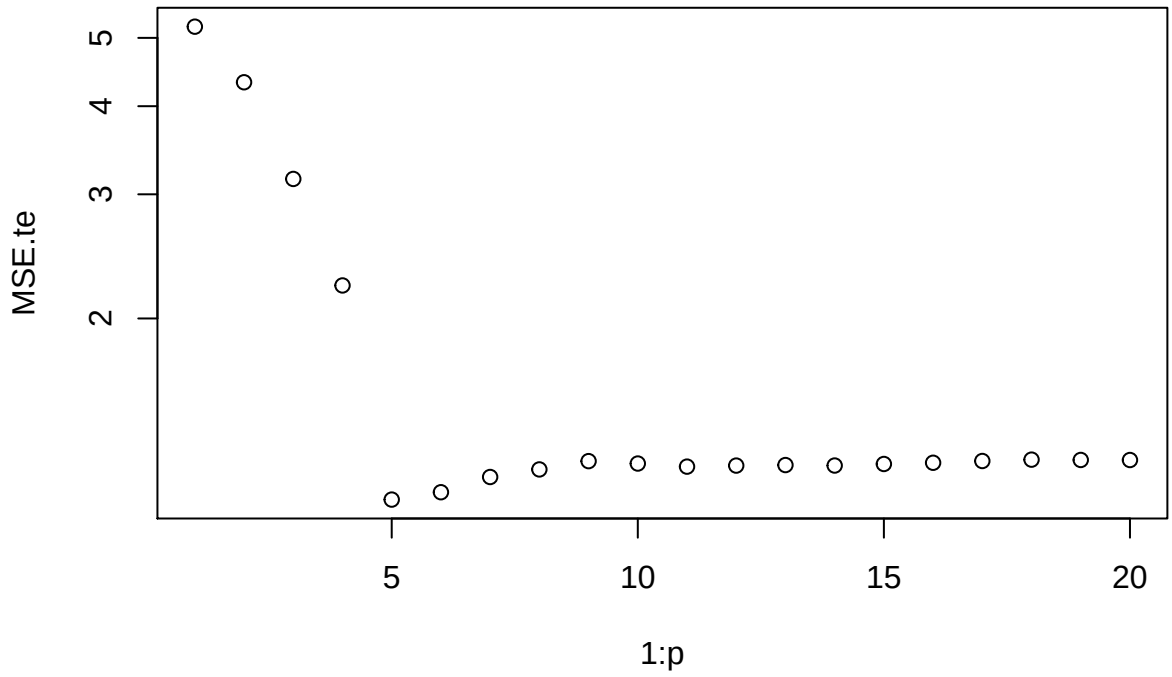
```
set.seed(123)
p = 20
n = 1000
X = matrix(rnorm(p*n), ncol=p)
beta = c(rep(1,p/4),rep(0,3*p/4))
y = X%*%beta + rnorm(n)
```

(b) Split your data set into a training set containing the first 100 observations and a test set containing the last 900 observations.

(c) Perform best subset selection on the training set, and plot the training set MSE associated with the best model of each size.



(d) Plot the test set MSE associated with the best model of each size.



(e) For which model size does the test set MSE take on its minimum value? Comment on your results.

```
## [1] 5
```

(f) Create a plot displaying $\sqrt{\sum_{j=0}^{p}(\beta_j - \hat{\beta}_j^k)^2}$ for a range of values of k from 1 to 20, where $\hat{\beta}_j^k$ is the jth coefficient estimate for the best model containing k coefficients. Comment on what you observe. How does this compare to the test MSE plot from (d)?