```
library(ElemStatLearn)
data(prostate)
data1 <- prostate[which(prostate$train==TRUE),1:9]

# 1. Full Model:

lm.full <- lm(lpsa~.,data=data1)
summary(lm.full)

# Interpretation: The p-values in the summary show that the Null Hypothesis for the
# variables lcavol, lweight, lbph and svi are rejected (with a confidence level of 95%),
# which means that their coefficients in the linear regression are significantly
# different than 0. The four other variables as well as the intercept the Null Hypothesis
cannot
# be rejected.
# Multiple R-squared and Adjusted R-square are between 65%-70%, which means that the
model only
# accounts for this percentage of the variance.
# As the p-value of the F-statistics is less than 5%, the hypothesis that all linear
coefficients
# are simultaneously 0 is rejected. The significant variables show a positive dependency
on lpsa
# as these coefficients are estimated as >0.


# 2. Stepwise regression

lm.forward <- step(lm.full,direction="forward")
lm.backward <- step(lm.full,direction="backward")
lm.both <- step(lm.full,direction="both")            #same result as bakward

anova(lm.backward,lm.forward, lm.full)        #lm.forward=lm.full as no variables are
taken away

# Interpretation: The results through anova show that there is no important loss of
information from reducing
# the model (variable gleason is taken away). This can be seen as the p-value of nearly
88% shows
# we cannot reject the Hypothesis that the means of two models are the same.


# 3. Best subset regression

#a
library(leaps)
lm.regsub <- regsubsets(lpsa~.,data=data1,nbest=3,nvmax=8)
summary(lm.regsub)

#b
plot(lm.regsub)

# In the plot it can be seen that the minimum BIC value reached is -51. Thia is found in
# three different models. However, the prefered one would be the one with less variables:
# lpsa = b0 + b1*lcavol + b2*lweight


#c
results <- summary(lm.regsub)
str(results)
```

```
plot("Size of model","BIC",xlim=c(1,8),ylim=c(min(results$bic),max(results$bic)))
lines(results$bic[seq(1,22,3)],type="b",col=1) # First best model
lines(results$bic[seq(2,22,3)],type="b",col=2) # Second best model
lines(results$bic[seq(3,22,3)],type="b",col=3) # third best model

# Choosing the best model:
which(results$bic[seq(1,22,3)]==min(results$bic[seq(1,22,3)]))

# Again it can be seen that the best model (lowest BIC) is the one containing lcavol and
lweight
# as variables

lm.bestreg <- lm(lpsa~lcavol+lweight,data=data1)
summary(lm.bestreg)

# As it is seen from the p-values, the Null Hypothesis (coefficient = 0) is rejected in
the case
# lcavol and lweight. However the intercept may be accounted as zero. The F-statistic
shows that
# not all coefficients can be taken as zero.

# As the intercept is not significant in the above model, one could consider the
following model:
lm.bestreg1 <- lm(lpsa~lcavol+lweight-1,data=data1)
summary(lm.bestreg1)

# The multiple r-squared and adjusted r-squared show better results than above.

# 4. Comparison by MSE for test data

data2 <- prostate[which(prostate$train==FALSE),1:8]
y <- prostate[which(prostate$train==FALSE),9]

mse <- function(lm,data,y){
  yhat <- predict(lm,data)
  n <- dim(data)[1]
  mse <- 1/n*norm(as.matrix(yhat-y),"F")^2
}


rbind(mse(lm.full,data2,y),mse(lm.forward,data2,y),mse(lm.both,data2,y),
      mse(lm.bestreg,data2,y),mse(lm.bestreg1,data2,y))

# The minimal MSE is for the last model (lm.bestreg1), which can again be seen as the
best model
```