Module 10 Assignment RMD

1

### Problem 1  
df <- MPV::table.b2  
n <- 20  
set.seed(1); chosen\_rows <- sort(sample(seq(1, nrow(df)), n))  
df1 <- df[chosen\_rows,]

1.1

# Reference: https://stats.stackexchange.com/questions/347652/default-stepaic-in-r  
# (a) forward selection  
model\_0 <- lm(y~1, data = df1)  
model\_1 <- lm(y~., data = df1)  
forward1 <- MASS::stepAIC(model\_0,  
 scope = list(upper=model\_1, lower=model\_0),  
 direction = c('forward'))

## Start: AIC=131.07  
## y ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + x4 1 9125.4 3574.7 107.72  
## + x1 1 5965.9 6734.1 120.38  
## + x5 1 1707.9 10992.2 130.18  
## <none> 12700.1 131.07  
## + x3 1 253.6 12446.4 132.67  
## + x2 1 0.2 12699.9 133.07  
##   
## Step: AIC=107.72  
## y ~ x4  
##   
## Df Sum of Sq RSS AIC  
## + x3 1 1796.39 1778.3 95.753  
## + x5 1 1308.09 2266.6 100.606  
## + x1 1 397.82 3176.8 107.358  
## <none> 3574.7 107.718  
## + x2 1 1.41 3573.3 109.710  
##   
## Step: AIC=95.75  
## y ~ x4 + x3  
##   
## Df Sum of Sq RSS AIC  
## + x2 1 201.056 1577.2 95.354  
## + x1 1 173.115 1605.2 95.705  
## <none> 1778.3 95.753  
## + x5 1 90.108 1688.2 96.713  
##   
## Step: AIC=95.35  
## y ~ x4 + x3 + x2  
##   
## Df Sum of Sq RSS AIC  
## + x1 1 398.90 1178.3 91.522  
## <none> 1577.2 95.354  
## + x5 1 13.93 1563.3 97.176  
##   
## Step: AIC=91.52  
## y ~ x4 + x3 + x2 + x1  
##   
## Df Sum of Sq RSS AIC  
## + x5 1 131.21 1047.1 91.161  
## <none> 1178.3 91.522  
##   
## Step: AIC=91.16  
## y ~ x4 + x3 + x2 + x1 + x5

1.2

# (b) backward elimination  
backward1 <- MASS::stepAIC(model\_1,  
 # scope = list(upper=model\_1, lower=model\_0),  
 direction = c('backward'))

## Start: AIC=91.16  
## y ~ x1 + x2 + x3 + x4 + x5  
##   
## Df Sum of Sq RSS AIC  
## <none> 1047.1 91.161  
## - x5 1 131.2 1178.3 91.522  
## - x2 1 255.5 1302.6 93.528  
## - x3 1 332.5 1379.7 94.677  
## - x1 1 516.2 1563.3 97.176  
## - x4 1 3550.1 4597.2 118.750

1.3

# (c) stepwise regression  
step1 <- MASS::stepAIC(model\_0,  
 scope = list(upper=model\_1, lower=model\_0),  
 direction = c('both'))

## Start: AIC=131.07  
## y ~ 1  
##   
## Df Sum of Sq RSS AIC  
## + x4 1 9125.4 3574.7 107.72  
## + x1 1 5965.9 6734.1 120.38  
## + x5 1 1707.9 10992.2 130.18  
## <none> 12700.1 131.07  
## + x3 1 253.6 12446.4 132.67  
## + x2 1 0.2 12699.9 133.07  
##   
## Step: AIC=107.72  
## y ~ x4  
##   
## Df Sum of Sq RSS AIC  
## + x3 1 1796.4 1778.3 95.753  
## + x5 1 1308.1 2266.6 100.606  
## + x1 1 397.8 3176.8 107.358  
## <none> 3574.7 107.718  
## + x2 1 1.4 3573.3 109.710  
## - x4 1 9125.4 12700.1 131.073  
##   
## Step: AIC=95.75  
## y ~ x4 + x3  
##   
## Df Sum of Sq RSS AIC  
## + x2 1 201.1 1577.2 95.354  
## + x1 1 173.1 1605.2 95.705  
## <none> 1778.3 95.753  
## + x5 1 90.1 1688.2 96.713  
## - x3 1 1796.4 3574.7 107.718  
## - x4 1 10668.1 12446.4 132.669  
##   
## Step: AIC=95.35  
## y ~ x4 + x3 + x2  
##   
## Df Sum of Sq RSS AIC  
## + x1 1 398.9 1178.3 91.522  
## <none> 1577.2 95.354  
## - x2 1 201.1 1778.3 95.753  
## + x5 1 13.9 1563.3 97.176  
## - x3 1 1996.0 3573.3 109.710  
## - x4 1 10843.4 12420.6 134.628  
##   
## Step: AIC=91.52  
## y ~ x4 + x3 + x2 + x1  
##   
## Df Sum of Sq RSS AIC  
## + x5 1 131.2 1047.1 91.161  
## <none> 1178.3 91.522  
## - x1 1 398.9 1577.2 95.354  
## - x2 1 426.8 1605.2 95.705  
## - x3 1 1896.7 3075.0 108.707  
## - x4 1 4035.6 5213.9 119.267  
##   
## Step: AIC=91.16  
## y ~ x4 + x3 + x2 + x1 + x5  
##   
## Df Sum of Sq RSS AIC  
## <none> 1047.1 91.161  
## - x5 1 131.2 1178.3 91.522  
## - x2 1 255.5 1302.6 93.528  
## - x3 1 332.5 1379.7 94.677  
## - x1 1 516.2 1563.3 97.176  
## - x4 1 3550.1 4597.2 118.750

1.4

# (d) all possible regressions  
best1 <- leaps::regsubsets(x = y~., data = df1, nvmax = 5)  
best1\_sum <- summary(best1)  
p.m <- 2:6  
aic <- n \* log(best1\_sum$rss / n) + 2 \* p.m  
data.frame(  
 Adj.R2 = which.max(best1\_sum$adjr2),  
 CP = which.min(best1\_sum$cp),  
 BIC = which.min(best1\_sum$bic),  
 AIC = which.min(aic)  
)

## Adj.R2 CP BIC AIC  
## 1 5 4 4 5

### Problem 2  
df <- MPV::table.b1  
n <- 20  
set.seed(1); chosen\_rows <- sort(sample(seq(1, nrow(df)), n))  
df2 <- df[chosen\_rows,]

2.1

# (a)  
# PRESS residuals  
beta\_hat\_calc <- function(X, y) {  
 X <- as.matrix(X)  
 beta\_hat <- solve(t(X) %\*% X) %\*% t(X) %\*% y  
 return(beta\_hat)  
}  
H\_calc <- function(X) {  
 X <- as.matrix(X)  
 H <- X %\*% solve(t(X) %\*% X) %\*% t(X)  
 return(H)  
}  
y\_hat\_calc <- function(H, y) {  
 y\_hat <- H %\*% y  
 return(y\_hat)  
}  
e\_calc <- function(y, y\_hat) {  
 e <- y - y\_hat  
 return(e)  
}  
X <- df2[,2:ncol(df2)]  
ones <- rep(1, n)  
X <- cbind(ones, X)  
y <- df2$y  
beta\_hat <- beta\_hat\_calc(X=X, y=y)  
H <- H\_calc(X)  
y\_hat <- y\_hat\_calc(H=H, y=y)  
e <- e\_calc(y=y, y\_hat=y\_hat)  
H\_diag <- diag(H)  
  
PRESS\_res <- sapply(1:n, function(x) e[x] / (1 - H\_diag[x]))  
PRESS <- sum(PRESS\_res^2)  
sum(e^2)

## [1] 38.56744

2.1.1

# 2.1.1  
data.frame(e=e,PRESS=PRESS\_res)

## e PRESS  
## 1 2.603555099 3.326864455  
## 2 0.689629058 1.748870047  
## 4 2.530430030 8.877899013  
## 5 0.084051420 0.204607506  
## 6 -1.636850024 -3.257775145  
## 7 -1.802280589 -2.611010953  
## 9 0.832731273 2.736284742  
## 10 -2.353290395 -4.971565865  
## 11 1.197279727 3.116432957  
## 12 1.169352737 1.656639103  
## 14 -0.849504725 -1.593749179  
## 17 -0.657157608 -1.301927252  
## 18 -0.725923065 -1.792462885  
## 19 0.789040347 1.454843630  
## 21 -1.682754382 -4.522626657  
## 22 1.160479603 2.323335875  
## 23 0.695241265 1.231994527  
## 25 0.001838846 0.004286539  
## 26 -0.357409797 -0.600428788  
## 28 -1.688458818 -2.716326614

2.2

# (b)  
set.seed(1)  
chosen\_row\_subset <- sort(sample(chosen\_rows, size = length(chosen\_rows)/2))  
df2b <- df[chosen\_row\_subset,]  
n <- 10  
X <- df2b[,2:ncol(df2)]  
ones <- rep(1, n)  
X <- cbind(ones, X)  
y <- df2b$y  
beta\_hat <- beta\_hat\_calc(X=X, y=y)  
H <- H\_calc(X)  
y\_hat <- y\_hat\_calc(H=H, y=y)  
e <- e\_calc(y=y, y\_hat=y\_hat)  
H\_diag <- diag(H)  
  
PRESS\_res\_b <- sapply(1:n, function(x) e[x] / (1 - H\_diag[x]))  
PRESS\_b <- sum(PRESS\_res\_b^2)  
sum(e^2)

## [1] 2.260452e-19

2.2.1

# 2.2.1  
data.frame(e=e,PRESS=PRESS\_res\_b)

## e PRESS  
## 1 6.255441e-11 -6.081513  
## 2 1.152571e-10 75.408150  
## 4 1.309424e-10 -1820.098765  
## 5 1.915428e-10 84.497208  
## 9 1.492748e-10 76.014699  
## 14 1.444853e-10 -571.796134  
## 18 1.536167e-10 -54.872145  
## 19 1.609957e-10 72.296341  
## 23 1.667431e-10 80.964313  
## 26 1.858318e-10 -49.305526

2.2.2

# predictive powers  
deleted\_rows <- chosen\_rows[!(chosen\_rows %in% chosen\_row\_subset)]  
df2c <- df[deleted\_rows,]  
X2 <- df2c[,2:ncol(df2c)]  
X2 <- cbind(ones, X2)  
y\_hat2 <- as.matrix(X2) %\*% as.matrix(beta\_hat)  
y2 <- df2c$y  
e2 <- e\_calc(y=y2, y\_hat=y\_hat2)  
sum(e2^2)

## [1] 5389.1