PSTAT 126 HW 5

Tamjid Islam

5/28/2020

1. Using the divusa dataset in the faraway package with divorce as the response and the other variables as predictors, implement the following variable selection methods to determine the "best" model:

```r
# install.packages('faraway')
library(faraway)
```

(a) Stepwise regression with AIC

```r
library(faraway)
data(divusa)
model <- lm(divorce ~ year + unemployed + femlab + marriage + birth + military, data = divusa)
reduced <- lm(divorce ~ 1, data = divusa)
step(reduced, scope = list(lower = reduced, upper = model))
```

```
## Start:  AIC=268.19
## divorce ~ 1
##
##              Df Sum of Sq     RSS    AIC
## + femlab      1   2024.42  418.10 134.28
## + year        1   1888.22  554.31 155.99
## + birth       1   1272.98 1169.54 213.48
## + marriage    1    697.17 1745.36 244.31
## + unemployed  1    108.33 2334.19 266.69
## <none>                    2442.53 268.19
## + military    1      0.84 2441.68 270.16
##
## Step:  AIC=134.28
## divorce ~ femlab
##
##              Df Sum of Sq     RSS    AIC
## + birth       1    113.73  304.38 111.83
## + year        1     29.70  388.41 130.60
## + marriage    1     13.34  404.76 133.78
## <none>                     418.10 134.28
## + military    1      1.93  416.17 135.92
## + unemployed  1      1.48  416.62 136.00
## - femlab      1   2024.42 2442.53 268.19
##
## Step:  AIC=111.83
## divorce ~ femlab + birth
##
##              Df Sum of Sq     RSS     AIC
## + marriage    1     94.54  209.84  85.196
```

```
## + unemployed  1     44.43  259.94 101.683
## + year        1     15.54  288.84 109.798
## <none>                     304.38 111.834
## + military    1      0.87  303.50 113.613
## - birth       1    113.73  418.10 134.278
## - femlab      1    865.16 1169.54 213.483
##
## Step:  AIC=85.2
## divorce ~ femlab + birth + marriage
##
##              Df Sum of Sq     RSS     AIC
## + year        1     26.76  183.08  76.691
## + unemployed  1      6.85  202.99  84.639
## + military    1      5.66  204.18  85.089
## <none>                     209.84  85.196
## - marriage    1     94.54  304.38 111.834
## - birth       1    194.92  404.76 133.781
## - femlab      1    949.45 1159.29 214.805
##
## Step:  AIC=76.69
## divorce ~ femlab + birth + marriage + year
##
##              Df Sum of Sq    RSS     AIC
## + military    1    20.957 162.12  69.330
## <none>                    183.08  76.691
## + unemployed  1     0.651 182.43  78.417
## - year        1    26.761 209.84  85.196
## - marriage    1   105.757 288.84 109.798
## - femlab      1   137.509 320.59 117.829
## - birth       1   183.446 366.53 128.140
##
## Step:  AIC=69.33
## divorce ~ femlab + birth + marriage + year + military
##
##              Df Sum of Sq    RSS     AIC
## <none>                    162.12  69.330
## + unemployed  1     1.925 160.20  70.410
## - military    1    20.957 183.08  76.691
## - year        1    42.054 204.18  85.089
## - marriage    1   126.643 288.77 111.779
## - femlab      1   158.003 320.13 119.718
## - birth       1   172.826 334.95 123.203
```

```
##
## Call:
## lm(formula = divorce ~ femlab + birth + marriage + year + military,
##     data = divusa)
##
## Coefficients:
## (Intercept)       femlab        birth     marriage         year     military
##    405.6167       0.8548      -0.1101       0.1593      -0.2179      -0.0412
```

The stepwise regression model using the AIC method is: lm(formula = divorce ~ femlab + birth + marriage + year + military, data = divusa)

(b) Best subsets regression with adjusted R^2

```r
library(leaps)
mod <- regsubsets(cbind(divusa$year, divusa$unemployed, divusa$femlab, divusa$marriage,
                        divusa$birth, divusa$military), divusa$divorce)
summary.mod <- summary(mod)
summary.mod$which
```

```
##   (Intercept)     a     b    c     d     e     f
## 1        TRUE FALSE FALSE TRUE FALSE FALSE FALSE
## 2        TRUE FALSE FALSE TRUE FALSE  TRUE FALSE
## 3        TRUE FALSE FALSE TRUE  TRUE  TRUE FALSE
## 4        TRUE  TRUE FALSE TRUE  TRUE  TRUE FALSE
## 5        TRUE  TRUE FALSE TRUE  TRUE  TRUE  TRUE
## 6        TRUE  TRUE  TRUE TRUE  TRUE  TRUE  TRUE
```

```r
summary.mod$adjr2
```

```
## [1] 0.8265403 0.8720158 0.9105579 0.9208807 0.9289506 0.9287914
```

The best subsets regression with adjusted R^2 is where there is the largest R^2 value. In this case it involves our first, third, forth, fifth and sixth predictor. Thus, year, femlab, marriage, birth and military are the the "best" for this model which the same as part a.

(c) Best subsets regression with

```r
summary.mod$cp
```

```
## [1] 109.695444  62.001274  22.692257  12.998703   5.841314   7.000000
```

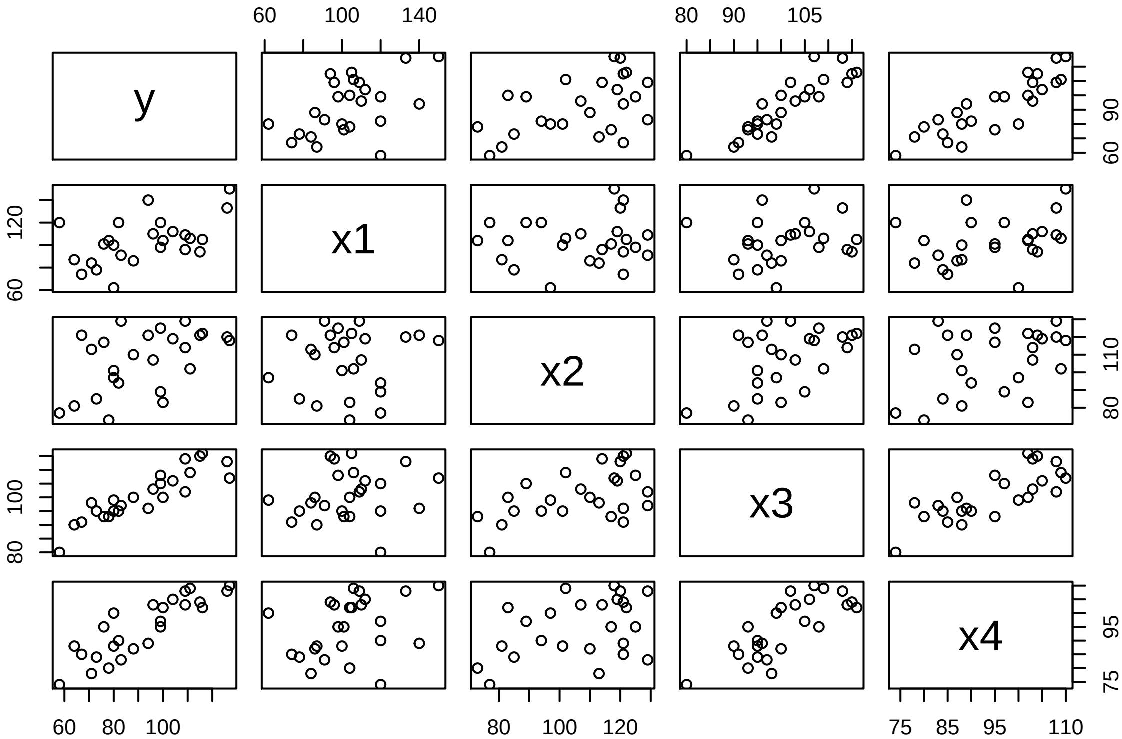
The "best" subsets regression with adjusted Mallow's C_p is 5.841314. It is the model that consists of all predictors except the second predictor. Thus, year, femlab, marriage, birth and military are the "best" for this model which is the same as a) and b).

2. Refer to the "Job proficiency" data posted on Gauchospace.

```r
setwd('~/Documents')
jobp <- read.csv('Job proficiency.csv', sep = ',')
```

(a) Obtain the overall scatterplot matrix and the correlation matrix of the X variables. Draw conclusions about the linear relationship between Y and the predictors.

```r
pairs(y ~ x1 + x2 + x3 + x4, data = jobp)
```




```r
cor(jobp)
```

```
##            y        x1        x2        x3        x4
## y  1.0000000 0.5144107 0.4970057 0.8970645 0.8693865
## x1 0.5144107 1.0000000 0.1022689 0.1807692 0.3266632
## x2 0.4970057 0.1022689 1.0000000 0.5190448 0.3967101
## x3 0.8970645 0.1807692 0.5190448 1.0000000 0.7820385
## x4 0.8693865 0.3266632 0.3967101 0.7820385 1.0000000
```

Based on these matrices, we can see that there is a positive linear trend betweem y and x3, y and x4, and a slight positive linear relationship between y and x1. There isn't a clear linear relationship between y and x2.

(b) Using only the first order terms as predictors, find the four best subset regression models according to the R^2 criterion.

```r
library(leaps)
mod <- regsubsets(cbind(jobp$x1, jobp$x2, jobp$x3, jobp$x4), jobp$y)
summary.mod <- summary(mod)
summary.mod$which
```

```
##   (Intercept)     a     b    c     d
## 1        TRUE FALSE FALSE TRUE FALSE
## 2        TRUE  TRUE FALSE TRUE FALSE
## 3        TRUE  TRUE FALSE TRUE  TRUE
## 4        TRUE  TRUE  TRUE TRUE  TRUE
```

```r
summary.mod$rsq
```

```
## [1] 0.8047247 0.9329956 0.9615422 0.9628918
```

(c) Since there is relatively little difference in R^2 for the four best subset models, what other criteria would you use to help in the selection of the best models? Discuss.

Since there is relatively little distance some better observations can be made by looking at the best subset model based on adjusted R^2 which will look at the largest adjusted R^2 value. Another option could be the MSE, where the smallest MSE value would be the best model. Other options could be looking at the AIC method, BIC method or using adjusted Mallow's C_p.

3. Refer again to the "Job proficiency" data from problem 2.

(a) Using stepwise regression, find the best subset of predictor variables to predict job proficiency. Use α limit of 0.05 to add or delete a variable.

```r
mod0 <- lm(jobp$y ~ 1)
add1(mod0, ~. + jobp$x1 + jobp$x2 + jobp$x3 + jobp$x4, test = 'F')
```

```
## Single term additions
##
## Model:
## jobp$y ~ 1
##         Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                9054.0 149.30
## jobp$x1  1    2395.9 6658.1 143.62  8.2763  0.008517 **
## jobp$x2  1    2236.5 6817.5 144.21  7.5451  0.011487 *
## jobp$x3  1    7286.0 1768.0 110.47 94.7824 1.264e-09 ***
## jobp$x4  1    6843.3 2210.7 116.06 71.1978 1.699e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Add x3 into model.

```r
mod1 <- update(mod0, ~. + jobp$x3)
summary(mod1)
```

```
##
## Call:
## lm(formula = jobp$y ~ jobp$x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.6908  -6.1073  -0.8528   2.6658  22.6010
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -106.1328    20.4472  -5.191 2.91e-05 ***
## jobp$x3        1.9676     0.2021   9.736 1.26e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.768 on 23 degrees of freedom
## Multiple R-squared:  0.8047, Adjusted R-squared:  0.7962
## F-statistic: 94.78 on 1 and 23 DF,  p-value: 1.264e-09
```

```r
add1(mod1, ~. + jobp$x1 + jobp$x2 + jobp$x4, test = 'F' )
```

```
## Single term additions
##
## Model:
## jobp$y ~ jobp$x3
##         Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>                1768.02 110.469
```

```
## jobp$x1  1   1161.37  606.66  85.727  42.116 1.578e-06 ***
## jobp$x2  1     12.21 1755.81 112.295   0.153   0.69946
## jobp$x4  1    656.71 1111.31 100.861  13.001   0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Add x1 into the model.

```r
mod2 <- update(mod1, ~. + jobp$x1)
summary(mod2)
```

```
##
## Call:
## lm(formula = jobp$y ~ jobp$x3 + jobp$x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3489 -2.8086 -0.4546  2.8981 12.6469
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -127.59569   12.68526  -10.06 1.09e-09 ***
## jobp$x3        1.82321    0.12307   14.81 6.31e-13 ***
## jobp$x1        0.34846    0.05369    6.49 1.58e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.251 on 22 degrees of freedom
## Multiple R-squared:  0.933,  Adjusted R-squared:  0.9269
## F-statistic: 153.2 on 2 and 22 DF,  p-value: 1.222e-13
```

```r
add1(mod2, ~. + jobp$x2 + jobp$x4, test = 'F')
```

```
## Single term additions
##
## Model:
## jobp$y ~ jobp$x3 + jobp$x1
##         Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                606.66 85.727
## jobp$x2  1     9.937 596.72 87.314  0.3497 0.5605965
## jobp$x4  1   258.460 348.20 73.847 15.5879 0.0007354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Add x4 to the model.

```r
mod3 <- update(mod2, ~. + jobp$x4)
summary(mod3)
```

```
##
## Call:
## lm(formula = jobp$y ~ jobp$x3 + jobp$x1 + jobp$x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4579 -3.1563 -0.2057  1.8070  6.6083
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.20002    9.87406 -12.578 3.04e-11 ***
## jobp$x3        1.35697    0.15183   8.937 1.33e-08 ***
## jobp$x1        0.29633    0.04368   6.784 1.04e-06 ***
## jobp$x4        0.51742    0.13105   3.948 0.000735 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.072 on 21 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.956
## F-statistic:   175 on 3 and 21 DF,  p-value: 5.16e-15
```

```r
add1(mod3, ~. + jobp$x2, test = 'F')
```

```
## Single term additions
##
## Model:
## jobp$y ~ jobp$x3 + jobp$x1 + jobp$x4
##         Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>                348.20 73.847
## jobp$x2  1     12.22 335.98 74.954  0.7274 0.4038
```

x4 has a p-value higher than $\alpha = 0.05$, so it will not be used for the model.

```r
finalmod <- lm(y ~ x3 + x1 + x4, data = jobp)
summary(finalmod)
```

```
##
## Call:
## lm(formula = y ~ x3 + x1 + x4, data = jobp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4579 -3.1563 -0.2057  1.8070  6.6083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.20002    9.87406 -12.578 3.04e-11 ***
## x3             1.35697    0.15183   8.937 1.33e-08 ***
## x1             0.29633    0.04368   6.784 1.04e-06 ***
## x4             0.51742    0.13105   3.948 0.000735 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.072 on 21 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.956
## F-statistic:   175 on 3 and 21 DF,  p-value: 5.16e-15
```

The best subset of predictor variables for stepwise regression is x3, x1, and x4. Therefore y ~ x3 + x1 + x4.

(b) How does the best subset obtained in part (a) compare with the best subset from part (b) of Q2?

Our part 3(a) best subset matches with one of the four best subset in part 2(b). However, based on the R^2 subset for part 2(b) it seems that the best model out for the four presented is the second one containing two predictors based on R^2 since it has the largest difference amongst the others. In part part 3(a), there are three predictors that represent the best model for this stepwise regression.

4. Refer to the "Brand preference" data posted on Gauchospace.

```
setwd('~/Documents')
brand <- read.csv('brand preference.csv', sep = ',')
```

a) Obtain the studentized deleted residuals and identify any outlying Y observations.

```
fit <- lm(y ~ x1 + x2, data = brand)
rs <- rstudent(fit)
rs
```

```
##           1           2           3           4           5           6
## -0.04085498  0.06128781 -1.36059879  1.38602483 -0.36694571 -0.66490618
##           7           8           9          10          11          12
## -0.76716157  0.50461264  0.46506694 -0.60436295  1.82302030  0.97784298
##          13          14          15          16
## -1.13966417 -2.10272640  1.48973208  0.24572878
```

```
which(abs(rs)>3)
```

```
## named integer(0)
```

There are no outliers since the absolute value of all of our externally studentized residuals are not greater than 3.

b) Obtain the diagonal elements of the Hat matrix, and provide an explanation for any pattern in these values.

```
h <- hatvalues(fit)
h
```

```
##      1      2      3      4      5      6      7      8      9     10     11
## 0.2375 0.2375 0.2375 0.2375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375 0.1375
##     12     13     14     15     16
## 0.1375 0.2375 0.2375 0.2375 0.2375
```

The hatvalues start at 0.2375 for the first 4 values, then goes to 0.1375 for the next 8 values then goes back to 0.2375 for the last 4 values. This basically calculates the the seperation of predictor variables from the mean. Therefore, it makes sense that the first 4 and last 4 values are larger than the middle 8 values based on the data as they are farther away from the mean. Thus, they are less likely to be accurate.

c) Are any of the observations high leverage point?

```
p <- sum(h)
n <- length(brand$y)
which(h > 3*p /n)
```

```
## named integer(0)
```

There are no observations with high leverage points.

5. The data below shows, for a consumer finance company operating in six cities, the number of competing loan companies operating in the city (X) and the number per thousand of the company's loans made in that city that are currently delinquent (Y):

i:	1	2	3	4	5	6
X_i:	4	1	2	3	3	4
Y_i:	16	5	10	15	13	22

Assume that a simple linear regression model is applicable. Using matrix methods, find

```
n <- 6
c <- c(1, 1, 1, 1, 1, 1)
Xi <- c(4, 1, 2, 3, 3, 4)
Yi <- c(16, 5, 10, 15, 13, 22)
```

(a) The appropriate X matrix.

```
X <- matrix(c(c, Xi) , ncol = 2)
X
```

```
##      [,1] [,2]
## [1,]    1    4
## [2,]    1    1
## [3,]    1    2
## [4,]    1    3
## [5,]    1    3
## [6,]    1    4
```

(b) Vector b of estimated coefficients.

```
tXX <- matrix(c(n, sum(Xi), sum(Xi), sum(Xi^2)), nrow = 2, ncol = 2)
tXY <- matrix(c(sum(Yi), sum(Xi*Yi)), ncol=1)
b <- solve(tXX) %*% tXY
b
```

```
##           [,1]
## [1,] 0.4390244
## [2,] 4.6097561
```

(c) The Hat matrix H.

```
H <- X %*% solve(tXX) %*% t(X)
H
```

```
##             [,1]       [,2]       [,3]      [,4]      [,5]        [,6]
## [1,]  0.36585366 -0.1463415 0.02439024 0.1951220 0.1951220  0.36585366
## [2,] -0.14634146  0.6585366 0.39024390 0.1219512 0.1219512 -0.14634146
## [3,]  0.02439024  0.3902439 0.26829268 0.1463415 0.1463415  0.02439024
## [4,]  0.19512195  0.1219512 0.14634146 0.1707317 0.1707317  0.19512195
## [5,]  0.19512195  0.1219512 0.14634146 0.1707317 0.1707317  0.19512195
## [6,]  0.36585366 -0.1463415 0.02439024 0.1951220 0.1951220  0.36585366
```

6. In stepwise regression, what advantage is there in using a relatively large α value to add variables? Comment briefly.

The advantage in having a relatively large α value is to allow more predictor variables to be involved to become the best model. It makes it easier on the restrictions by adding or removing predictors to create the best model.