PSTAT126_HW5

Celeste Herrera

6/06/2020

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:

(a) Stepwise regression with AIC

```r
library(faraway)
data("divusa")
year = divusa$year
divorce = divusa$divorce
unemployed = divusa$unemployed
femlab = divusa$femlab
marriage = divusa$marriage
birth = divusa$birth
military = divusa$military
mod0 <- lm(divorce ~ 1, data = divusa)
mod1 <- lm(divorce ~ year + unemployed + femlab + marriage + birth + military, data = divusa)
summary(mod1)
```

```
##
## Call:
## lm(formula = divorce ~ year + unemployed + femlab + marriage +
##     birth + military, data = divusa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9087 -0.9212 -0.0935  0.7447  3.4689
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 380.14761   99.20371   3.832 0.000274 ***
## year         -0.20312    0.05333  -3.809 0.000297 ***
## unemployed   -0.04933    0.05378  -0.917 0.362171
## femlab        0.80793    0.11487   7.033 1.09e-09 ***
## marriage      0.14977    0.02382   6.287 2.42e-08 ***
## birth        -0.11695    0.01470  -7.957 2.19e-11 ***
## military     -0.04276    0.01372  -3.117 0.002652 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.513 on 70 degrees of freedom
## Multiple R-squared:  0.9344, Adjusted R-squared:  0.9288
## F-statistic: 166.2 on 6 and 70 DF,  p-value: < 2.2e-16
```

```r
step(mod0, scope = list(lower=mod0, upper = mod1))
```

```
## 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 smallest AIC = 69.33 which gave us the best model of lm(formula = divorce ~ femlab + birth + marriage + year + military, data = divusa).

(b) Best subsets regression with adjusted R2

```
library(leaps)
models = regsubsets(cbind(year,unemployed,femlab,marriage, birth,military),divorce)

summary_model =summary(models)
summary_model$adjr2
```

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

```
summary_model$which
```

```
##   (Intercept)  year unemployed femlab marriage birth military
## 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
```

The best model with adj R^2 as the scale is model with the year, femlab, marriage, birth and military as the predictors.

(c) Best subsets regression with adjusted Mallow's Cp

```r
summary_model$cp
```

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

```r
summary_model$which
```

```
##   (Intercept)  year unemployed femlab marriage birth military
## 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
```

The best subsets regression model with adjusted Mallow's Cp is 5.841314. It The model consists of all the predictors such as year, femlab, marriage, birth and military except unemployed. Those predictors previously stated are the best for the model.

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

```r
getwd()
```

```
## [1] "/Users/celesteherrera/Documents/PSTAT 126"
```

```r
setwd("~/Documents/PSTAT 126")
job_proficiency = read.csv("Job proficiency.csv", header = TRUE)
```

(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. Also, is there a multicollinearity problem which is evident?

```r
getwd()
```

```
## [1] "/Users/celesteherrera/Documents/PSTAT 126"
```

```r
setwd("~/Documents/PSTAT 126")
job_proficiency = read.csv("Job proficiency.csv", header = TRUE)
library(corrplot)
```

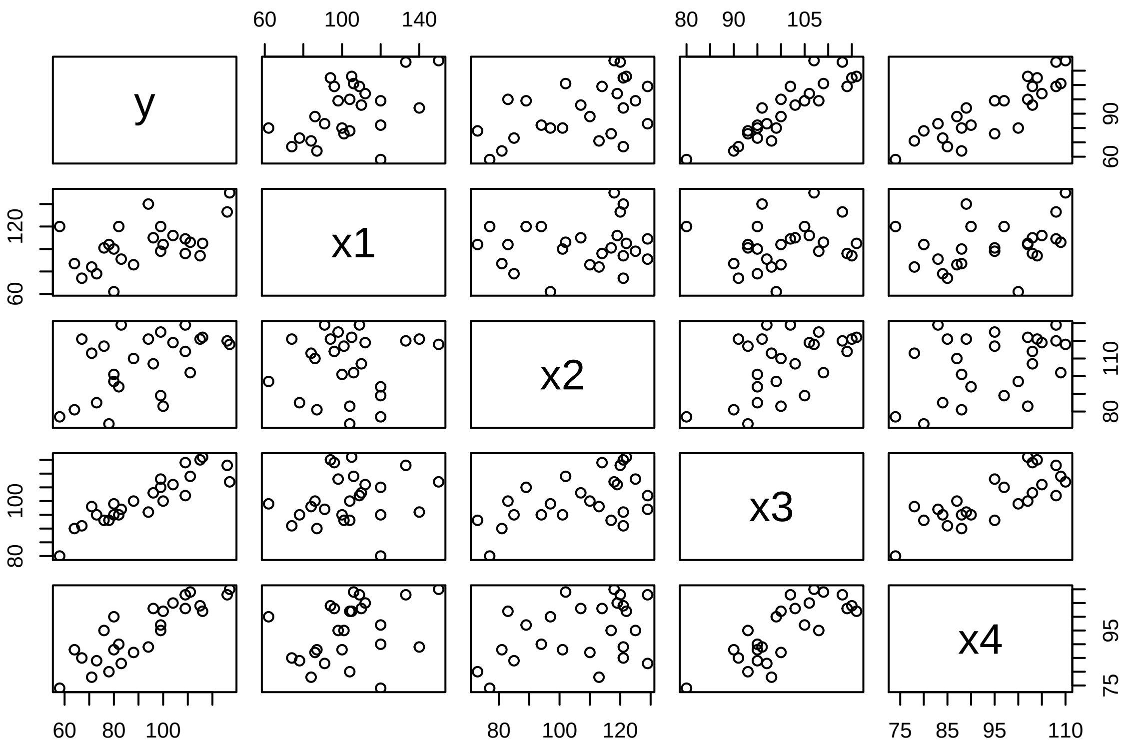
```
## corrplot 0.84 loaded
```

```r
library(Hmisc)
```

```
## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:faraway':
##
##     melanoma

## Loading required package: survival

## Warning: package 'survival' was built under R version 3.6.2

##
## Attaching package: 'survival'

## The following objects are masked from 'package:faraway':
##
```

```
##     rats, solder

## Loading required package: Formula

## Loading required package: ggplot2

##
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
y = job_proficiency$y
x1 = job_proficiency$x1
x2 = job_proficiency$x2
x3 = job_proficiency$x3
x4 = job_proficiency$x4
#scatterplot matrix
pairs(job_proficiency)
```




```r
#corrilation of the matrix
cor(job_proficiency)
```

```
##             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
```

It seems that x3, x4 and Y have strong linear relationship. While x1 has moderately strong relationship with Y. And x2 having the weakest relationship of all 4 with Y.

(b) Using only the first order terms as predictors, find the four best subset regression models according to the R2 criterion.

```
library(leaps)
mod = regsubsets(cbind(x1,x2,x3,x4),y)
summary.mod =summary(mod)
summary.mod$which
```

```
##   (Intercept)    x1    x2   x3    x4
## 1        TRUE FALSE FALSE TRUE FALSE
## 2        TRUE  TRUE FALSE TRUE FALSE
## 3        TRUE  TRUE FALSE TRUE  TRUE
## 4        TRUE  TRUE  TRUE TRUE  TRUE
```

```
summary.mod$rsq
```

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

The four best subset regression models according to the R^2 criterion is 0.8047247 0.9329956 because it has the biggest jump in value.

(c) Since there is relatively little difference in R2 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 such a small distance there is some better observations can be made significantly by looking at the best subset model based on the adjusted R^2 which will be reffering to look at the largest adjusted R^2 value. Another option that could have been is 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 Mallows Cp Criterion such as AICp and SBCp that I can use to help select the best model. They all place penalties for adding predictors.

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.

```
model1 <- lm(y ~ 1, data = job_proficiency)

add1(model1, ~.+x1 + x2 + x3 + x4, data = job_proficiency, test = 'F')
```

```
## Single term additions
##
## Model:
## y ~ 1
##        Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>              9054.0 149.30
## x1      1    2395.9 6658.1 143.62  8.2763  0.008517 **
## x2      1    2236.5 6817.5 144.21  7.5451  0.011487 *
## x3      1    7286.0 1768.0 110.47 94.7824 1.264e-09 ***
## 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
```

```
# since x3 has the smallest p-value and the largest F value
model2<- lm(y~x3, data =  job_proficiency)

add1(model2, ~.+x1 +x2 +x4, data = job_proficiency, test = 'F')
```

```
## Single term additions
```

```
##
## Model:
## y ~ x3
##        Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>              1768.02 110.469
## x1      1   1161.37  606.66  85.727  42.116 1.578e-06 ***
## x2      1     12.21 1755.81 112.295   0.153   0.69946
## 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
model3 = lm(y ~ x3+x1, data = job_proficiency)
summary(model3)
```

```
##
## Call:
## lm(formula = y ~ x3 + x1, data = job_proficiency)
##
## 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 ***
## x3             1.82321    0.12307   14.81 6.31e-13 ***
## 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
```

```
add1(model3, ~.+x2+x4, data = job_proficiency, test = 'F')
```

```
## Single term additions
##
## Model:
## y ~ x3 + x1
##        Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>              606.66 85.727
## x2      1     9.937 596.72 87.314  0.3497 0.5605965
## 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
model4 = lm(y ~ x3 +x1 +x4, data =  job_proficiency)
add1(model4, ~.+x2, data = job_proficiency, test = 'F')
```

```
## Single term additions
##
## Model:
## y ~ x3 + x1 + x4
```

```
##        Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>              348.20 73.847
## x2      1     12.22 335.98 74.954  0.7274 0.4038
```

```r
summary(model4)
```

```
##
## Call:
## lm(formula = y ~ x3 + x1 + x4, data = job_proficiency)
##
## 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
```

Finally, regressing y on all four predictors and x2 isn't significant to be included because the p-value is much larger than our alpha value 0.05 (0.4038 > 0.05). Thus it is deleted from the model. The best subset of predictor variables to predict job proficiency is (x1,x3,x4)

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

In 3a th best subset matches with one of the four best subset for 2b. Although, for the R^2 for 2b it seems that the model out of the four presented is the second one containing two predictors based on the R^2 since it is shown to have the biggest difference compared to the others. In 3a there are three predictors(x1, x3 and x4) the model for the stepwise regression

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

```r
getwd()
```

```
## [1] "/Users/celesteherrera/Documents/PSTAT 126"
```

```r
setwd("~/Documents/PSTAT 126")
brand_prefrence = read.csv("Brand preference.csv", header = TRUE)
```

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

```r
y= brand_prefrence$y
x1 = brand_prefrence$x1
x2 = brand_prefrence$x2
fit.all= lm(y~ x1 + x2, data = brand_prefrence)
(rsd.lm=round(rstudent(fit.all), 3))
```

```
##      1      2      3      4      5      6      7      8      9     10     11
## -0.041  0.061 -1.361  1.386 -0.367 -0.665 -0.767  0.505  0.465 -0.604  1.823
```

```
##     12     13     14     15     16 
##  0.978 -1.140 -2.103  1.490  0.246
```

```r
n=16
p=4
ifelse(rsd.lm > qt(1-0.95/2/n,n-p-1), "outlier", "Non-outlier")
```

```
##             1             2             3             4             5 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##             6             7             8             9            10 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            11            12            13            14            15 
## "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" "Non-outlier" 
##            16 
## "Non-outlier"
```

There are no outliers considering the absolute value of 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.

```r
h<-(h.lm=round(hatvalues(fit.all), 3))
h
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13 
## 0.238 0.238 0.238 0.238 0.138 0.138 0.138 0.138 0.138 0.138 0.138 0.138 0.238 
##    14    15    16 
## 0.238 0.238 0.238
```

The first 4 values start at 0.238 then follows with the next 8 values being 0.138 then lastly the last 4 values being 0.238. This calulation is the seperation of prediction variables from the mean. Therefore the data shows to be further awau from the mean states that it is less likely to be accurate.

(c) Are any of the observations high leverage point?

```r
p<-sum((h.lm=round(hatvalues(fit.all), 3)))
n<-length(brand_prefrence$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):

$$\begin{pmatrix} i:1 & 2 & 3 & 4 & 5 & 6 \\ X_i:4 & 1 & 2 & 3 & 3 & 4 \\ Y_i:16 & 5 & 10 & 15 & 13 & 22 \end{pmatrix}$$

Assume that a simple linear regression model is applicable. Using matrix methods, find (a) The appropriate X matrix.

```r
X = matrix(c(rep(1,times =6), 4,1,2,3,3,4), nrow = 6, ncol =2, byrow = FALSE)
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.

```r
Y= matrix(c(16,5,10,15,13,22),nrow = 6, ncol =1)
solve(t(X)%*%X)%*%t(X)%*%Y
```

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

(c) The Hat matrix H

```r
X%*% solve(t(X)%*%X) %*% t(X)
```

```
##             [,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.

In the Stepwise Regression the advantage in using a large alpha in the variables is that it will increase the overall R square value. Also easier to remove a relativley large or small value.