PSTAT 126 Homework 5

Shaiyon Hariri

6/5/2020

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:

```
library(faraway)
library(leaps)
```

(a) Stepwise regression with AIC

```
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
```

This metric suggests that a 5 parameter model that only excludes unemployed is the best model.

(b) Best subsets regression with adjusted R squared

```r
mod <- regsubsets(subset(divusa, select=-c(divorce)), divusa$divorce)
summary.mod <- summary(mod)
summary.mod$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
```

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

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

This metric suggests that the same 5 parameter model is the best, as it has the highest adjusted R squared.

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

```r
summary.mod$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
```

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

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

This metric suggests that the 5 parameter model that only excludes unemployed is the best model.

```r
bestModel <- lm(divorce ~ year + femlab + marriage + birth + military, data=divusa)
summary(bestModel)
```

```
##
## Call:
## lm(formula = divorce ~ year + femlab + marriage + birth + military,
##     data = divusa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7586 -1.0494 -0.0424  0.7201  3.3075
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 405.61670   95.13189   4.264 6.09e-05 ***
## year         -0.21790    0.05078  -4.291 5.52e-05 ***
## femlab        0.85480    0.10276   8.318 4.29e-12 ***
## marriage      0.15934    0.02140   7.447 1.76e-10 ***
## birth        -0.11012    0.01266  -8.700 8.43e-13 ***
## military     -0.04120    0.01360  -3.030  0.00341 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.511 on 71 degrees of freedom
## Multiple R-squared:  0.9336, Adjusted R-squared:  0.929
## F-statistic: 199.7 on 5 and 71 DF,  p-value: < 2.2e-16
```

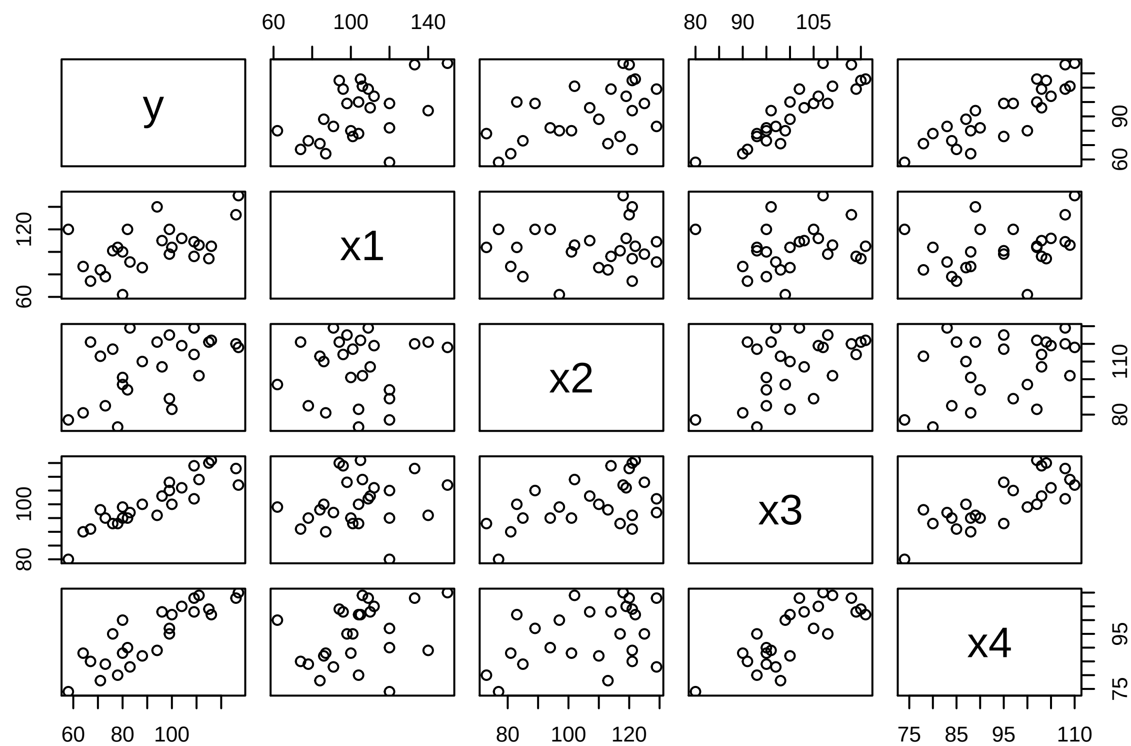
All of the tests suggest the 5 parameter model including all predictors except unemployed is the "best" model.

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

```r
job <- read.csv("Job proficiency.csv")
```

(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(job)
```



```r
cor(job)
```

```
##            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
```

x3 and x4 have an immediately noticeable and strong positive linear relationship with the response. x1 has a slight positive linear relationships with y, while x2 does not seem to have a significant linear relationship with y.

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

```r
mod <- regsubsets(subset(job, select=-c(y)), job$y)
summary.mod <- summary(mod)
summary.mod$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
```

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

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

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

The best subset model based on adjusted R squared, the stepwise AIC method, Mallow's Cp metric + more could all help select the most optimal model.

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

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

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

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

```
## job$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 to the model, as it has the highest F value.

```r
model <- update(baseline, ~. + job$x3)
add1(model, ~. + job$x1 + job$x2 + job$x4, test='F')
```

```
## Single term additions
##
## Model:
## job$y ~ job$x3
##        Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>              1768.02 110.469
## job$x1  1   1161.37  606.66  85.727  42.116 1.578e-06 ***
## job$x2  1     12.21 1755.81 112.295   0.153   0.69946
## job$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 to the model, as it has the highest F value.

```r
model <- update(model, ~. + job$x1)
add1(model, ~. + job$x2 + job$x4, test='F')
```

```
## Single term additions
##
## Model:
## job$y ~ job$x3 + job$x1
##        Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>              606.66 85.727
## job$x2  1     9.937 596.72 87.314  0.3497 0.5605965
## job$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, as the p value is low.

```r
model <- update(model, ~. + job$x4)
add1(model, ~. + job$x2, test='F')
```

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

Do not add x2 to the model, as the p value is high. Therefore, the model containing all the predictors except x2 is the final model.

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

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

```
##
## 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 ***
## x1             0.29633    0.04368   6.784 1.04e-06 ***
## x3             1.35697    0.15183   8.937 1.33e-08 ***
## 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
```

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

It is consistent with the findings in Q2, (b), as the change in R squared is very minimal when the last predictor (x2) is added to the model. This suggests that x2 does not add much predictive power to the model, and should be disregarded.

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

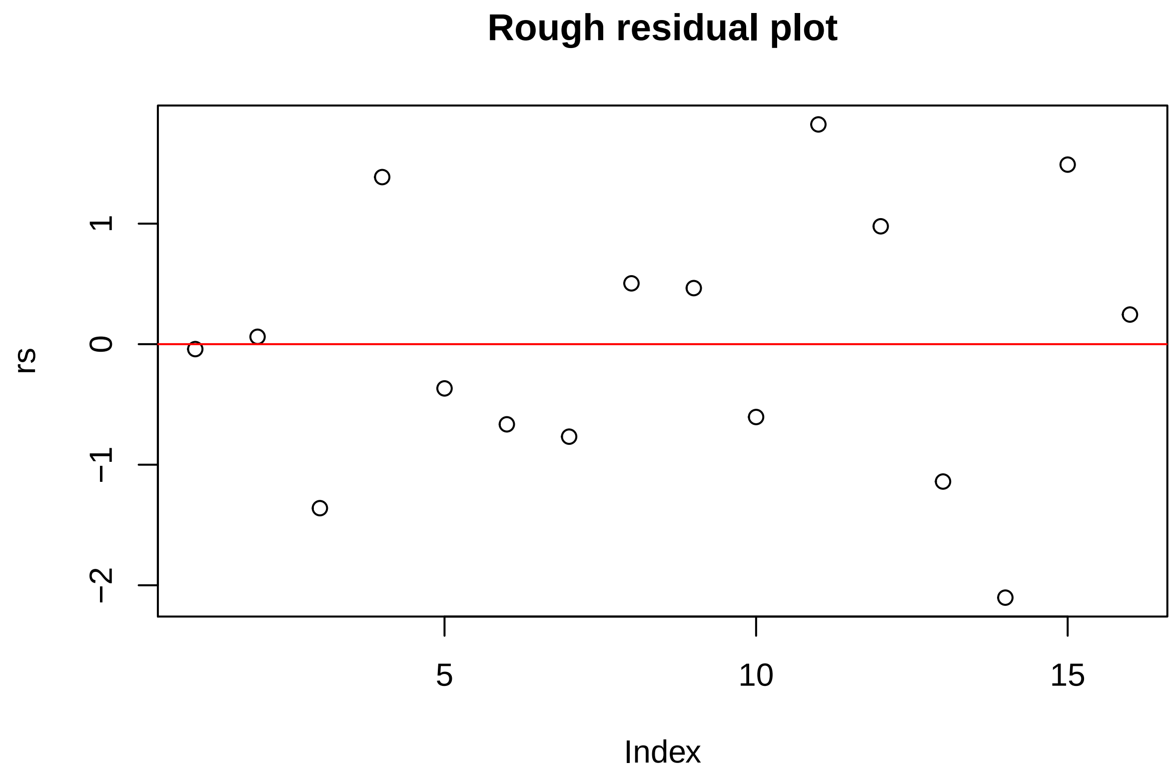
```
brand <- read.csv("brand preference.csv")
```

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

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

```
plot(rs, main="Rough residual plot")
abline(0,0, col="red")
```



There does not seem to be any significant outliers in the residuals.

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

```
hat <- hatvalues(model)
hat
```

```
##      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
```

There are 4 0.2375 values surrounding 8 0.1375 values. The diagonals measure the separation the values have to the mean, so it checks out that the first and last 4 are greater than the middle.

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

```
p <- sum(hat)
n <- length(brand$y)

which(hat > (3*p)/n)
```

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

No.

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):

```r
xi <- c(4, 1, 2, 3, 3, 4)
yi <- c(16, 5, 10, 15, 13, 22)
n <- length(xi)
```

(a) The appropriate X matrix.

```r
X <- matrix(c(rep(1, n), xi), nrow=n)
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
tXX <- matrix(c(n, sum(xi), sum(xi), sum(xi**2)), nrow=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.

```r
hat <- X %*% solve(tXX) %*% t(X)
hat
```

```
##             [,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 alpha value to add variables? Comment briefly.

A large value for alpha encourages more predictors to be added to the model than less, leading to the model having potentially increased predictive power. The statistician can manually add or remove borderline valuable predictors based on judgement rather than the stepwise regression removing them automatically.