STAT425_Homework7

Giang Le

11/1/2021

Problem 1

```
# Load my data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
questionnaire <- read.csv("questionnaire.csv")
summary(questionnaire)
```

```
##       rate          color              lot.size
##  Min.   :25.00   Length:19          Min.   :100.0
##  1st Qu.:27.50   Class :character   1st Qu.:194.5
##  Median :29.00   Mode  :character   Median :264.0
##  Mean   :29.47                      Mean   :270.7
##  3rd Qu.:31.50                      3rd Qu.:329.5
##  Max.   :35.00                      Max.   :473.0
```

```
# Fit a SLR model separately for each group
questionnaire.blue=lm(rate ~ lot.size, data=questionnaire[questionnaire$color=='blue',]);
questionnaire.green=lm(rate ~ lot.size, data=questionnaire[questionnaire$color=='green',]);
summary(questionnaire.blue)
```

```
##
## Call:
## lm(formula = rate ~ lot.size, data = questionnaire[questionnaire$color ==
##     "blue", ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6036 -0.0847  0.3028  1.4163  3.3214
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.66936    2.53461  13.284 3.21e-06 ***
## lot.size    -0.01991    0.01010  -1.972   0.0893 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.711 on 7 degrees of freedom
## Multiple R-squared:  0.3571, Adjusted R-squared:  0.2652
## F-statistic: 3.888 on 1 and 7 DF,  p-value: 0.08926
```

```
summary(questionnaire.green)
```

```
##
## Call:
## lm(formula = rate ~ lot.size, data = questionnaire[questionnaire$color ==
##     "green", ])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.353 -1.524  0.023  1.144  4.031
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.330340   2.287265  15.447 3.07e-07 ***
## lot.size    -0.017904   0.007173  -2.496   0.0372 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.232 on 8 degrees of freedom
## Multiple R-squared:  0.4378, Adjusted R-squared:  0.3675
## F-statistic:  6.23 on 1 and 8 DF,  p-value: 0.03717
```

The regression line corresponds to the blue questionnaires is $\hat{rate} = 33.66936 - 0.01991 * lot.size$. The regression line corresponds to the green questionnaires is $\hat{rate} = 35.330340 - -0.017904 * lot.size$

Test whether the interaction term is statistically significant. State the hypotheses, decision rule and conclusion.

```
# Run a full model with an interaction term.
question.full = lm(rate ~ lot.size + color + lot.size:color, data = questionnaire)
summary(question.full)
```

```
##
## Call:
## lm(formula = rate ~ lot.size + color + lot.size:color, data = questionnaire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6036 -1.1694  0.3028  1.2939  4.0309
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         33.669361   2.306804  14.596 2.85e-10 ***
## lot.size            -0.019907   0.009189  -2.167   0.0468 *
## colorgreen           1.660979   3.422258   0.485   0.6344
## lot.size:colorgreen  0.002003   0.012136   0.165   0.8711
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.467 on 15 degrees of freedom
```

```
## Multiple R-squared:   0.41,  Adjusted R-squared:  0.292
## F-statistic: 3.475 on 3 and 15 DF,  p-value: 0.04282
```

The hypothesis H_0 is that the partial slope of the interaction term is 0. The alternative hypothesis H_a is that the partial slope of the interaction term is not 0 and there is an interaction between color and lot.size.

According to the table above, the p-value of the interaction term is $0.8711 > 0.05$. Conclusion: So at the significance level 0.05, we fail to reject H_0. It is likely that there is no interaction between color and lot.size.

Does the response rate vary according to the questionnaire color? We ran an additive model and found that the color predictor is significant at 0.1 level. So the questionnaire color only has an additive effect on the rate at 0.1 level, (only changing the intercept) and both blue and green groups have the same slope.

In other words, the response rate does not vary according to the questionnaire color as the slope of lot.size in this SLR is the same across the two groups.

```
question.additive = lm(rate ~ lot.size + color, data=questionnaire)
summary(question.additive)
```

```
##
## Call:
## lm(formula = rate ~ lot.size + color, data = questionnaire)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.5111 -1.1048  0.2277  1.1838  4.1708
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.400026   1.580154  21.137 4.07e-13 ***
## lot.size    -0.018759   0.005817  -3.225   0.0053 **
## colorgreen   2.189577   1.169044   1.873   0.0795 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.391 on 16 degrees of freedom
## Multiple R-squared:  0.4089, Adjusted R-squared:  0.3351
## F-statistic: 5.535 on 2 and 16 DF,  p-value: 0.01489
```

Problem 2

```
# Load the dataset
install.packages("ISLR",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/3_mgdyf12z7dvb8rt4d60nt80000gn/T//RtmpBH5KJp/downloaded_packages
```

```
library("ISLR")
head(Hitters)
```

```
##                   AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun
## -Andy Allanson      293   66     1   30  29    14     1    293    66      1
## -Alan Ashby         315   81     7   24  38    39    14   3449   835     69
## -Alvin Davis        479  130    18   66  72    76     3   1624   457     63
## -Andre Dawson       496  141    20   65  78    37    11   5628  1575    225
## -Andres Galarraga   321   87    10   39  42    30     2    396   101     12
```

```
## -Alfredo Griffin    594  169     4   74  51    35    11   4408  1133     19
##                   CRuns CRBI CWalks League Division PutOuts Assists Errors
## -Andy Allanson       30   29     14      A        E     446      33     20
## -Alan Ashby         321  414    375      N        W     632      43     10
## -Alvin Davis        224  266    263      A        W     880      82     14
## -Andre Dawson       828  838    354      N        E     200      11      3
## -Andres Galarraga    48   46     33      N        E     805      40      4
## -Alfredo Griffin    501  336    194      A        W     282     421     25
##                   Salary NewLeague
## -Andy Allanson        NA         A
## -Alan Ashby        475.0         N
## -Alvin Davis       480.0         A
## -Andre Dawson      500.0         N
## -Andres Galarraga   91.5         N
## -Alfredo Griffin   750.0         A
```

```
# Run a full model with the listed predictors
model1 <- lm(Salary ~ AtBat + HmRun + RBI + Years + CHits + CRuns + CWalks + Assists, data=Hitters)
summary(model1)
```

```
##
## Call:
## lm(formula = Salary ~ AtBat + HmRun + RBI + Years + CHits + CRuns +
##     CWalks + Assists, data = Hitters)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1005.86  -188.99   -53.37   118.74  2088.81
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  90.41578   86.08554   1.050   0.2946
## AtBat         0.08554    0.30322   0.282   0.7781
## HmRun        -6.34027    5.50323  -1.152   0.2504
## RBI           6.12587    2.44463   2.506   0.0128 *
## Years       -21.92174   11.79091  -1.859   0.0642 .
## CHits         0.04035    0.24725   0.163   0.8705
## CRuns         0.91604    0.53527   1.711   0.0882 .
## CWalks       -0.08545    0.25199  -0.339   0.7348
## Assists      -0.06946    0.17930  -0.387   0.6988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 349.3 on 254 degrees of freedom
##   (59 observations deleted due to missingness)
## Multiple R-squared:  0.4186, Adjusted R-squared:  0.4003
## F-statistic: 22.86 on 8 and 254 DF,  p-value: < 2.2e-16
```

After fitting a model, I conduct variable selection with the leaps package.

```
# Calling leaps
library(leaps)
b = regsubsets(Salary ~ AtBat + HmRun + RBI + Years + CHits + CRuns + CWalks + Assists, data=Hitters)
rs = summary(b)
```

```
# Calling which to find the models and which variables are selected or not.
rs$which
```

```
##   (Intercept) AtBat HmRun   RBI Years CHits CRuns CWalks Assists
## 1        TRUE FALSE FALSE FALSE FALSE FALSE  TRUE  FALSE   FALSE
## 2        TRUE FALSE FALSE  TRUE FALSE FALSE  TRUE  FALSE   FALSE
## 3        TRUE FALSE FALSE  TRUE  TRUE FALSE  TRUE  FALSE   FALSE
## 4        TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE  FALSE   FALSE
## 5        TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE   TRUE   FALSE
## 6        TRUE FALSE  TRUE  TRUE  TRUE FALSE  TRUE   TRUE    TRUE
## 7        TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE   TRUE    TRUE
## 8        TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE   TRUE    TRUE
```

```
# Then I examine the R^2 and other criteria such as Cp, AIC, and BIC.
rs$adjr2
```

```
## [1] 0.3139878 0.3964733 0.4061688 0.4086044 0.4068839 0.4047643 0.4025994
## [8] 0.4003103
```

```
which.max(rs$adjr2)
```

```
## [1] 4
```

```
rs$cp # wants lowest
```

```
## [1] 39.569720  4.663592  1.469781  1.431709  3.182850  5.098657  7.026633
## [8]  9.000000
```

```
which.min(rs$cp)
```

```
## [1] 4
```

```
# I calculated BIC and AIC by hand.
n=dim(Hitters)[1]
msize = 2:9;
BIC = n*log(rs$rss/n) + msize*log(n);
which.min(BIC)
```

```
## [1] 3
```

```
AIC = n*log(rs$rss/n) + 2*msize;
which.min(AIC)
```

```
## [1] 4
```

According to these results, the largest adjusted R^2 is 0.4086044, corresponding to the model with 4 variables and one intercept. The retained variables are HmRun, RBI, Years, CRuns.

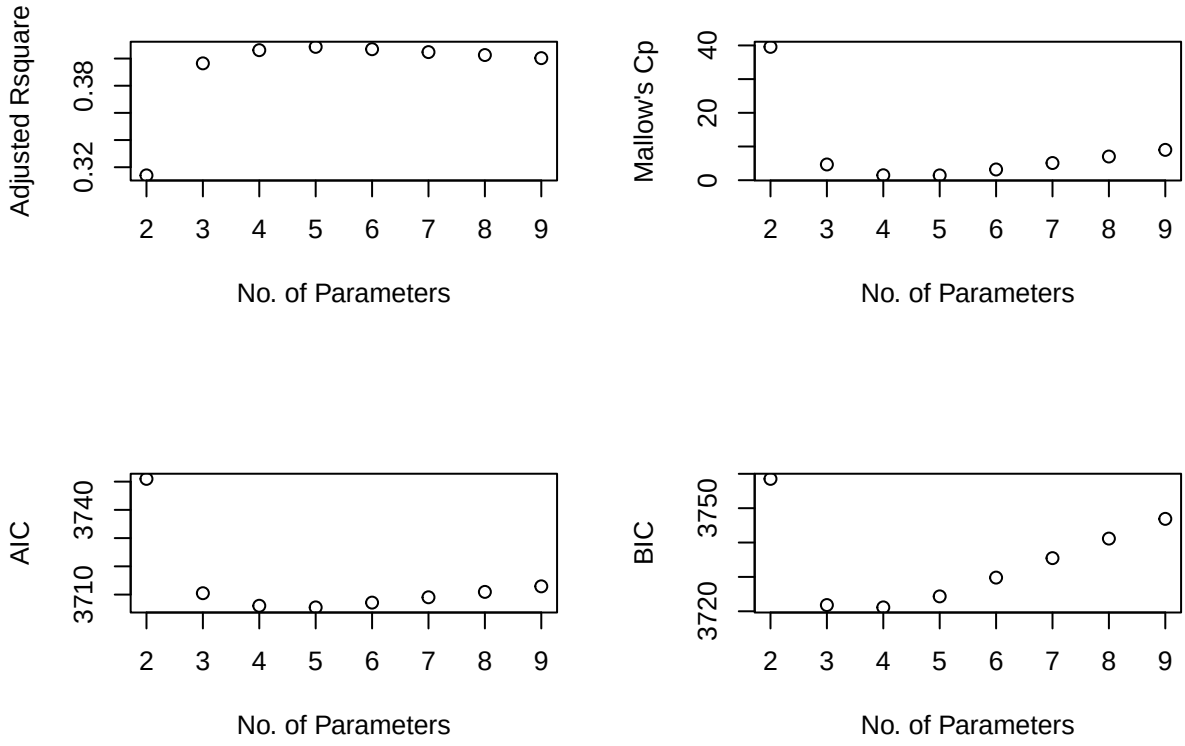
The smallest Cp is 1.431709, , corresponding to the model with 4 variables and one intercept. The retained variables are HmRun, RBI, Years, CRuns.

The model corresponding to the lowest BIC is model 3 (3 variables and one intercept). The retained variables are RBI, Years, CRuns.

The model corresponding to the lowest AIC is model 4. The retained variables are HmRun, RBI, Years, CRuns.

```
# Verification with plots
par(mfrow=c(2,2))
plot(msize, rs$adjr2, xlab="No. of Parameters", ylab = "Adjusted Rsquare");
plot(msize, rs$cp, xlab="No. of Parameters", ylab = "Mallow's Cp");
```

```
plot(msize, AIC, xlab="No. of Parameters", ylab = "AIC");
plot(msize, BIC, xlab="No. of Parameters", ylab = "BIC");
```



I use step to conduct variable selection. Explain which variables are removed and in which order.

```
step(model1, direction="both")
```

```
## Start:  AIC=3089.13
## Salary ~ AtBat + HmRun + RBI + Years + CHits + CRuns + CWalks +
##     Assists
##
##           Df Sum of Sq      RSS    AIC
## - CHits    1      3250 31001838 3087.2
## - AtBat    1      9712 31008299 3087.2
## - CWalks   1     14035 31012622 3087.2
## - Assists  1     18315 31016902 3087.3
## - HmRun    1    161990 31160577 3088.5
## <none>                 30998587 3089.1
## - CRuns    1    357430 31356017 3090.1
## - Years    1    421856 31420443 3090.7
## - RBI      1    766331 31764918 3093.6
##
## Step:  AIC=3087.16
## Salary ~ AtBat + HmRun + RBI + Years + CRuns + CWalks + Assists
##
##           Df Sum of Sq      RSS    AIC
## - AtBat    1      8790 31010627 3085.2
## - Assists  1     16847 31018685 3085.3
## - CWalks   1     24954 31026792 3085.4
## - HmRun    1    188374 31190212 3086.8
## <none>                 31001838 3087.2
```

```
## + CHits    1      3250 30998587 3089.1
## - Years    1    499749 31501587 3089.4
## - RBI      1    835204 31837041 3092.2
## - CRuns    1   2349973 33351810 3104.4
##
## Step:  AIC=3085.23
## Salary ~ HmRun + RBI + Years + CRuns + CWalks + Assists
##
##           Df Sum of Sq      RSS    AIC
## - Assists  1     10275 31020903 3083.3
## - CWalks   1     30289 31040916 3083.5
## - HmRun    1    217236 31227863 3085.1
## <none>                 31010627 3085.2
## + AtBat    1      8790 31001838 3087.2
## + CHits    1      2329 31008299 3087.2
## - Years    1    593227 31603854 3088.2
## - RBI      1   1683638 32694266 3097.1
## - CRuns    1   2678699 33689326 3105.0
##
## Step:  AIC=3083.32
## Salary ~ HmRun + RBI + Years + CRuns + CWalks
##
##           Df Sum of Sq      RSS    AIC
## - CWalks   1     30371 31051274 3081.6
## - HmRun    1    216046 31236948 3083.1
## <none>                 31020903 3083.3
## + Assists  1     10275 31010627 3085.2
## + AtBat    1      2218 31018685 3085.3
## + CHits    1      1579 31019323 3085.3
## - Years    1    584302 31605205 3086.2
## - RBI      1   1834198 32855101 3096.4
## - CRuns    1   2671588 33692491 3103.1
##
## Step:  AIC=3081.58
## Salary ~ HmRun + RBI + Years + CRuns
##
##           Df Sum of Sq      RSS    AIC
## <none>                 31051274 3081.6
## - HmRun    1    248730 31300004 3081.7
## + CWalks   1     30371 31020903 3083.3
## + CHits    1     11877 31039397 3083.5
## + Assists  1     10357 31040916 3083.5
## + AtBat    1      4995 31046279 3083.5
## - Years    1    614897 31666170 3084.7
## - RBI      1   1999718 33050992 3096.0
## - CRuns    1   4837160 35888434 3117.7
##
## Call:
## lm(formula = Salary ~ HmRun + RBI + Years + CRuns, data = Hitters)
##
## Coefficients:
## (Intercept)        HmRun          RBI        Years        CRuns
##     97.2979      -6.6629       6.6091     -22.0623       0.9332
```

The best model contains 4 predictors and an intercept. The predictors are HmRun, RBI, Years, and CRuns. The variables removed are AtBat, CHits, CWalks, and Assists.

The order in which variables are removed is: CHits -> AtBat -> Assists -> CWalks.