HOMEWORK 6

Yiqiao Yin [YY2502]

Contents

PROBLEM 1 (Ch9, Q9)

Let us load the data:

```
data = read.table("C:/Users/eagle/OneDrive/Course/CU Stats/STATS GR6101 - Applied Statistics I/Data/CH0
colnames(data) = c("Y", "X1", "X2", "X3")
datacopy = data
performance = rbind()
```

(1) Let us compute the evaluation criteria:

```
for (i in c("0", "1", "2", "3", "1_2", "1_3", "2_3", "1_2_3")) {
  i = as.numeric(unlist(strsplit(i, "_")))
  if (as.numeric(i) == 0) {
    MODEL = lm(Y~1L, data = data)
  } else {
    data = data.frame(cbind(data[,1], datacopy[,-1][,i]))
    colnames(data) = c("Y", paste0("X", i))
    MODEL = lm(Y~., data = data)
  }
  eval_R2 = summary(MODEL)$r.square
  eval_AIC = nrow(data) * log(sum(MODEL$residuals^2)) - nrow(data) * log(nrow(data)) + 2*ncol(data)
  eval_Cp = olsrr::ols_mallows_cp(MODEL, lm(Y~., data = datacopy))
  performance = rbind(performance, cbind(
    Formula = paste0("Y=", paste0(paste0("X", i), collapse = "+")),
    R2 = eval_R2,
    AIC = eval_AIC,
    Cp = eval_Cp ))
}; performance
```

```
## Warning in if (as.numeric(i) == 0) {: the condition has length > 1 and only the
## first element will be used

## Warning in if (as.numeric(i) == 0) {: the condition has length > 1 and only the
## first element will be used

## Warning in if (as.numeric(i) == 0) {: the condition has length > 1 and only the
## first element will be used

## Warning in if (as.numeric(i) == 0) {: the condition has length > 1 and only the
## first element will be used

##      Formula      R2                   AIC                Cp
## [1,] "Y=X0"       "0"                  "268.915461105549" "88.1562338191455"
## [2,] "Y=X1"       "0.618984251996021" "220.529390822719" "8.35360628199045"
## [3,] "Y=X2"       "0.363538735911057" "244.13120196195"  "42.1123236337672"
## [4,] "Y=X3"       "0.415497545878045" "240.213723332691" "35.2456429948055"
## [5,] "Y=X1+X2"    "0.654955853888437" "217.967647227459" "5.59973485144706"
## [6,] "Y=X1+X3"    "0.67608638253165"  "215.060654177041" "2.80720376735253"
## [7,] "Y=X2+X3"    "0.46845446298584"  "237.845006316576" "30.2470562751665"
## [8,] "Y=X1+X2+X3" "0.68219433328074"  "216.184962183753" "4"
```

Similar results can be found from the following:

```
olsrr::ols_step_all_possible(lm(Y~., data=data))
```

```
##   Index N Predictors  R-Square Adj. R-Square Mallow's Cp
## 1     1 1         X1 0.6189843     0.6103248    8.353606
## 3     2 1         X3 0.4154975     0.4022134   35.245643
## 2     3 1         X2 0.3635387     0.3490737   42.112324
## 5     4 2      X1 X3 0.6760864     0.6610206    2.807204
## 4     5 2      X1 X2 0.6549559     0.6389073    5.599735
## 6     6 2      X2 X3 0.4684545     0.4437314   30.247056
## 7     7 3   X1 X2 X3 0.6821943     0.6594939    4.000000
```

(2) The *R-square* tends to go up as we increase the number of features. This is not necessarily true for AIC. As a matter of fact, AIC penalizes the model more as we increase the number of parameters. We can see this from the formula of AIC.

(3) If we start with X_1, we get AIC of 220. Then we start adding X_2 or X_3. Here $X_1 + X_3$ has lower AIC at 215. Then we see if we can keep adding in order to lower AIC. In this example, we can no longer do that. Hence, the optimal model is $Y \sim X_1 + X_3$.

The other direction is the following. We start with the full model $Y \sim X_1 + X_2 + X_3$ with AIC of 216. We subtract the variables one by one. Subtracting X_1 gives us AIC of 237. Subtracting X_2 gives us AIC of 215. Subtracting X_3 gives us AIC of 217. The lowest AIC is model $Y \sim X_1 + X_3$.

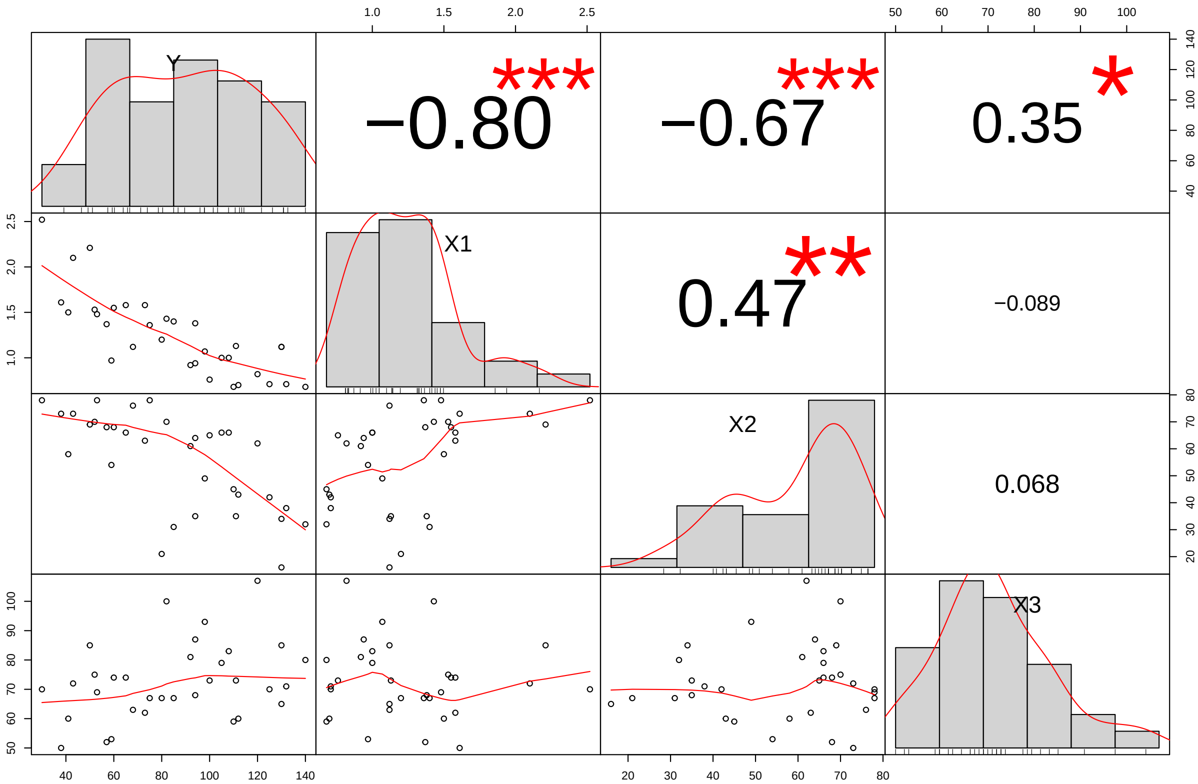
PROBLEM 2 (Ch9, Q15b,c)

Let us load the data first

```r
data2 = read.table("C:/Users/eagle/OneDrive/Course/CU Stats/STATS GR6101 - Applied Statistics I/Data/CH
colnames(data2) = c("Y", paste0("X", 1:3))
```

(1) Correlation and Scatter Plot

```r
PerformanceAnalytics::chart.Correlation(data2)
```



From results above, we have correlation matrix of the X varaibles to be the following

$$\begin{bmatrix} 1 & 0.47 & -0.089 \\ 0.47 & 1 & 0.068 \\ -0.089 & 0.068 & 1 \end{bmatrix}$$

and the scatter plot are above as well.

From the results, we can say that there is some correlation between X_1 and X_2, i.e. $\text{cor}(X_1, X_2) = 0.47$. We can also confirm this idea from $\text{cor}(Y, X_1) = -0.8$ and $\text{cor}(Y, X_2) = -0.67$. In other words, the response variable Y is both correlated with X_1 and X_2.

(2) Fit multiple linear regression model:

```r
MODEL2 = lm(Y~., data = data2)
summary(MODEL2)
```

```
##
## Call:
```

```
## lm(formula = Y ~ ., data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.668  -7.002   1.518   9.905  16.006
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 120.0473    14.7737   8.126 5.84e-09 ***
## X1          -39.9393     5.6000  -7.132 7.55e-08 ***
## X2           -0.7368     0.1414  -5.211 1.41e-05 ***
## X3            0.7764     0.1719   4.517 9.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.46 on 29 degrees of freedom
## Multiple R-squared:  0.8548, Adjusted R-squared:  0.8398
## F-statistic: 56.92 on 3 and 29 DF,  p-value: 2.885e-12
```

From above regression results, we have the following model

$$Y = 120.0473 - 39.9393X_1 - 0.7368X_2 + 0.7764X_3$$

We observe that both X_1 and X_2 have regression coefficients to be negative and they both have signifitantly high *t-value*. Since from part (1) we discussed the potential of multicollinearity, it is with doubt that we should keep both of these covariates.

To see if our doubts need to be put in action, we can confirm the idea by the following:

```
data2copy = data2
performance2 = rbind()
for (i in c("0", "1", "2", "3", "1_2", "1_3", "2_3", "1_2_3")) {
  i = as.numeric(unlist(strsplit(i, "_")))
  if (as.numeric(i) == 0) {
    MODEL = lm(Y~1L, data = data2)
  } else {
    data2 = data.frame(cbind(data2[,1], data2copy[,-1][,i]))
    colnames(data2) = c("Y", paste0("X", i))
    MODEL = lm(Y~., data = data2)
  }
  eval_R2 = summary(MODEL)$r.square
  eval_AIC = nrow(data2) * log(sum(MODEL$residuals^2)) - nrow(data2) * log(nrow(data2)) + 2*ncol(data2)
  eval_Cp = olsrr::ols_mallows_cp(MODEL, lm(Y~., data = data2copy))
  performance2 = rbind(performance2, cbind(
    Formula = paste0("Y=", paste0(paste0("X", i), collapse = "+")),
    R2 = eval_R2,
    AIC = eval_AIC,
    Cp = eval_Cp ))
}; performance2
```

```
## Warning in if (as.numeric(i) == 0) {: the condition has length > 1 and only the
## first element will be used

## Warning in if (as.numeric(i) == 0) {: the condition has length > 1 and only the
```

```
## first element will be used

## Warning in if (as.numeric(i) == 0) {: the condition has length > 1 and only the
## first element will be used

## Warning in if (as.numeric(i) == 0) {: the condition has length > 1 and only the
## first element will be used
```

```
##      Formula      R2                  AIC                 Cp
## [1,] "Y=X0"       "0"                 "233.887977887098" "168.750046029594"
## [2,] "Y=X1"       "0.642900659975658" "195.906515893374" "42.330609607"
## [3,] "Y=X2"       "0.446053532146372" "210.395299451315" "81.6508324516934"
## [4,] "Y=X3"       "0.119657096239573" "225.682333013619" "146.848535547972"
## [5,] "Y=X1+X2"    "0.752670606738983" "185.785847477995" "22.4040576883597"
## [6,] "Y=X1+X3"    "0.718889720038503" "190.010706255522" "29.1517913617011"
## [7,] "Y=X2+X3"    "0.600167376151661" "201.636572312908" "52.866585017839"
## [8,] "Y=X1+X2+X3" "0.854818556608976" "170.205535378933" "4"
```

We observe that the model $Y \sim X_1 + X_2 + X_3$ have the best *R-square*, *AIC*, and *Cp* scores.

PROBLEM 3 (Ch9, Q16)

(1) Fit model and select the best three accordign to *Mallow's Cp*:

```
Y = data2$Y
X1 = data2$X1; X2 = data2$X2; X3 = data2$X3
center1 = X1 - mean(X1); center1_sq = center1^2
center2 = X2 - mean(X2); center2_sq = center2^2
center3 = X3 - mean(X3); center3_sq = center3^2
MODEL3 = lm(
  Y~center1+center2+center3+center1*center2+center1*center3+center2*center3+
    center1*center3+center1_sq+center2_sq+center3_sq)
MODELLIST = olsrr::ols_step_all_possible(MODEL3)
MODELLIST[order(MODELLIST$cp, decreasing = FALSE), ][1:3, ]
```

```
##     Index N                                          Predictors  R-Square
## 133   130 4             center1 center2 center3 center1:center2 0.8788215
## 265   256 5 center1 center2 center3 center3_sq center1:center2 0.8876545
## 262   257 5 center1 center2 center3 center2_sq center1:center2 0.8827571
##     Adj. R-Square Mallow's Cp
## 133     0.8615103    3.302215
## 265     0.8668497    3.384990
## 262     0.8610455    4.447976
```

From above results, we conclude that the models are

$$Y \sim (X_1 - \bar{X}_1) + (X_2 - \bar{X}_2) + (X_3 - \bar{X}_3) + (X_1 - \bar{X}_1) * (X_2 - \bar{X}_2) \text{ with Cp=3.30}$$

$$Y \sim (X_1 - \bar{X}_1) + (X_2 - \bar{X}_2) + (X_3 - \bar{X}_3) + (X_3 - \bar{X}_3)^2 + (X_1 - \bar{X}_1) * (X_2 - \bar{X}_2) \text{ with Cp=3.38}$$

$$Y \sim (X_1 - \bar{X}_1) + (X_2 - \bar{X}_2) + (X_3 - \bar{X}_3) + (X_2 - \bar{X}_2)^2 + (X_1 - \bar{X}_1) * (X_2 - \bar{X}_2) \text{ with Cp=4.45}$$

(2) In terms of the measure of *Mallow's Cp*, the numerical difference is quite small.

PROBLEM 4 (Ch9, Q19a)

Let us conduct forward stepwise regression. First, let us reconstruct the data with the appropriate columns.

```
data4 = data.frame(cbind(
  Y,
  center1, center2, center3,
  center1*center2, center1*center3, center2*center3,
  center1^2, center2^2, center3^2 ))
colnames(data4) = c("Y", "C1", "C2", "C3",
                    "C12", "C13", "C23", "C1_sq", "C2_sq", "C3_sq")
```

Next, let us screen models using stepwise regression and set to *forward*. Note alpha level is 0.1.

```
MODEL4 <- lm(Y~1, data = data4)
summary(SignifReg::SignifReg(MODEL4, alpha = 0.1, scope = data4, criterion = "AIC", direction = "forwar
```

```
##
## Call:
## lm(formula = Y ~ C1 + C2 + C3 + C12, data = data4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.198  -4.867   0.761   5.074  17.657
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  82.0934     2.4122  34.033  < 2e-16 ***
## C1          -47.3114     6.0752  -7.788 1.75e-08 ***
## C2           -0.6760     0.1340  -5.046 2.44e-05 ***
## C3            0.7951     0.1600   4.969 3.02e-05 ***
## C12           0.8620     0.3660   2.355   0.0258 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.58 on 28 degrees of freedom
## Multiple R-squared:  0.8788, Adjusted R-squared:  0.8615
## F-statistic: 50.77 on 4 and 28 DF,  p-value: 1.959e-12
```

Last, let us repeat the above and set alpha level to 0.15.

```
MODEL4 <- lm(Y~1, data = data4)
summary(SignifReg::SignifReg(MODEL4, alpha = 0.15, scope = data4, criterion = "AIC", direction = "forwa
```

```
##
## Call:
## lm(formula = Y ~ C1 + C2 + C3 + C12, data = data4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.198  -4.867   0.761   5.074  17.657
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  82.0934     2.4122  34.033  < 2e-16 ***
## C1          -47.3114     6.0752  -7.788 1.75e-08 ***
## C2           -0.6760     0.1340  -5.046 2.44e-05 ***
## C3            0.7951     0.1600   4.969 3.02e-05 ***
## C12           0.8620     0.3660   2.355   0.0258 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.58 on 28 degrees of freedom
## Multiple R-squared:  0.8788, Adjusted R-squared:  0.8615
## F-statistic: 50.77 on 4 and 28 DF,  p-value: 1.959e-12
```

We observe that both alpha levels produce the same stepwise results. The optimal model is the following:

$$Y \sim (X_1 - \bar{X}_1) + (X_2 - \bar{X}_2) + (X_3 - \bar{X}_3) + (X_1 - \bar{X}_1) * (X_2 - \bar{X}_2)$$

with regression coefficients produced we have

$$Y = 82.09 - 47.31(X_1 - \bar{X}_1) - 0.67(X_2 - \bar{X}_2) + 0.795(X_3 - \bar{X}_3) + 0.86(X_1 - \bar{X}_1) * (X_2 - \bar{X}_2)$$

and we are done.