LINEAR REGRESSION: Homework #8

Professor Jingchen Liu

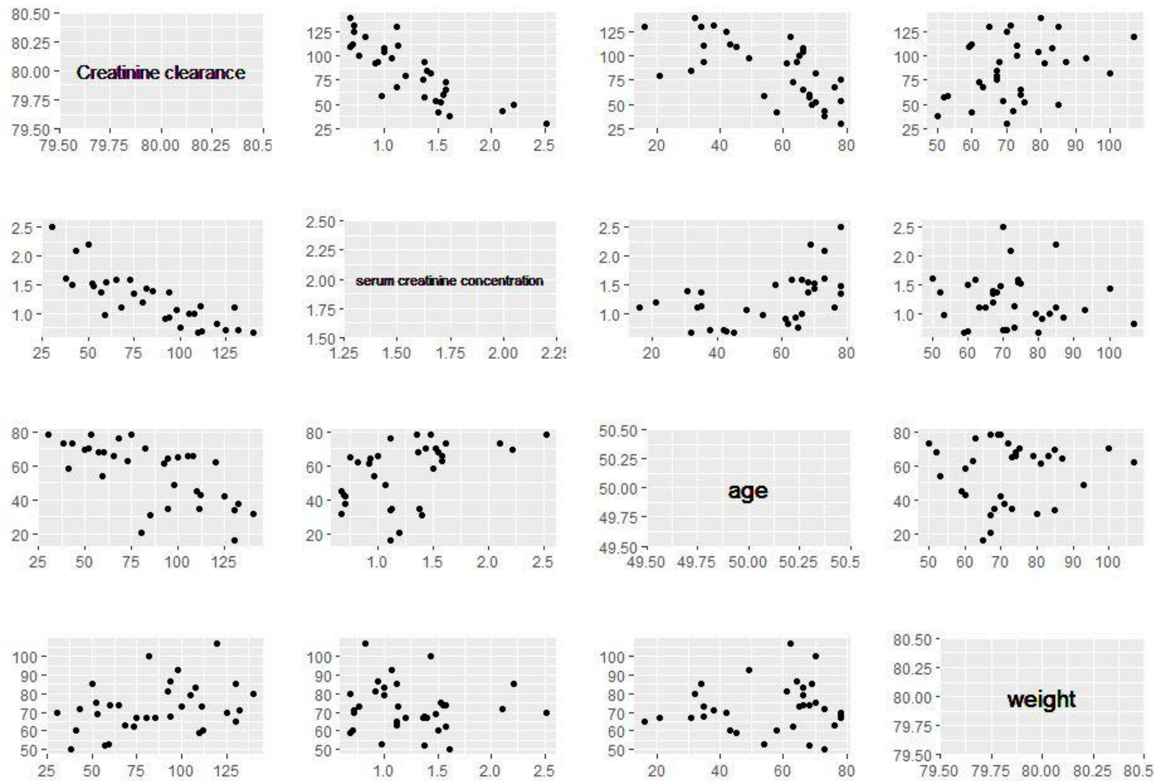
Fan Yang
UNI: fy2232

Problem 1 (9.9)

```
> ABCp <- function(data,col,MSE){
+   Y <- data[,1];n <- length(Y);p <- length(col) + 1
+   model <- lm(Y~., data = data[,col])
+   SSE <- sum((Y-model$fitted.values)^2)
+   AIC = n*log(SSE) - n*log(n) + 2*p
+   BIC = n*log(SSE) - n*log(n) + floor(log(n))*p
+   Cp = SSE/MSE - (n-2*p)
+   return (data.frame(AIC,Cp,BIC))
+ }
>
> MSE <- anova(lm(V1~V2+V3+V4,data=d9))$`Mean Sq`[4]
> col <- list(2,3,4,c(2,3),c(2,4),c(3,4),c(2,3,4))
> ABIC <- sapply(col,ABCp,data=d9,MSE=MSE)
> colnames(ABIC) <- col
>
> ABIC
    2        3        4        c(2, 3)  c(2, 4)  c(3, 4)  c(2, 3, 4)
AIC 220.5294 244.1312 240.2137 217.9676 215.0607 237.845  216.185
Cp  8.353606 42.11232 35.24564 5.599735 2.807204 30.24706 4
BIC 222.5294 246.1312 242.2137 220.9676 218.0607 240.845  220.185
```

Problem 2 (9.15)

(b)




```
> cor(d15[,2:4])
            V2         V3          V4
V2  1.00000000 0.46773179 -0.08898262
V3  0.46773179 1.00000000  0.06848147
V4 -0.08898262 0.06848147  1.00000000
```

The response variable and the variable serum creatinine concentration (X_1) show significant linear relationship. And with the variable age (X_2) also forms a linear line but it is not as linear as with X_1. As for the last variable weight (X_3), points are uniformly lied on the plot with little evidence of linearity.

In the correction matrix, X_1 and X_2 have a correction of 0.468, which may imply multicollinearity problem.

(c)

```
> model15c <- lm(V1~V2+V3+V4,data=d15)
> summary(model15c)
Call:
lm(formula = V1 ~ V2 + V3 + V4, data = d15)
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 ***
V2          -39.9393     5.6000  -7.132 7.55e-08 ***
V3           -0.7368     0.1414  -5.211 1.41e-05 ***
V4            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
```

According to the result above, F-statistic is large with p-value less than 0.01. And each variable coefficient shows significant contribution to the model. So, all the variables should be attained.

Problem 3 (9.16)

(a)

First compute a function to calculate the Cp criteria:

```
> Cp2 <- function(data,col,MSE){
+   Y <- data[,1];n <- length(Y);col <- unlist(col);p <- length(col) + 1
+   X <- data.frame(data[,col])
+   model <- lm(Y~.,data=X)
+   SSE <- sum((Y-model$fitted.values)^2)
+   Cp = SSE/MSE - (n-2*p)
+   return (Cp)
+ }
```

Then we list all the 511 kinds of combination of the 9 variables and store the combinations in a list called *col*.

Finally calculate all 511 Cp of the 511 combinations and list the three lowest value.

```
> model16 <- lm(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9,data=d16)
> MSE <- anova(model16)$'Mean Sq'[10]
> Cp <- sapply(col,Cp2,data=d16,MSE=MSE)
> head(sort(Cp),3)
[1] 3.302215 3.384990 3.674777
> head(order(Cp),3)
[1] 131 263 153
> col[131];col[263];col[153]
[1] 2 3 4 6
[1] 2 3 4 6 9
[1] 2 4 5 8
```

$X_1, X2, X3, X4, X5, X6, X7, X8, X_9$ represent variable combinations $\{X_1\}\{X2\}\{X3\}\{X1^2\}\{X1X2\}\{X2^2\}\{X2X3\}\{X3^2\}\{X1X3\}$ respectively. and correspond to column 2,3,4,5,6,7,8,9 respectively.

As we can see the three lowest value are 3.302215 3.384990 3.674777 and the three corresponding variables combination is

$$\{X_1, X_2, X_3, X_1X_2\}; \quad \{X_1, X_2, X_3, X_1X_2, X_3^2\}; \quad \{X_1, X_3, X_1^2, X_2X_3\};$$

(b)

The three lowest value are 3.302215 3.384990 3.674777, so there is little difference between the three subset models.

Problem 4 (9.19)

(a)

```
> my_stepwise(Y,X)
[1] "current predictor: "
[1] 1
[1] "current predictor: "
[1] 1 2
[1] "current predictor: "
[1] 1 2 3
[1] "current predictor: "
[1] 1 2 3 4
[1] "current predictor: "
[1] 1 2 3 4
[1] "final predictor: "
[1] 1 2 3 4
```

Using the stepwise function and get the best subset of variables. The result subset is 1,2,3,4 which corresponds to $\{X_1, X_2, X_3, X_1^2\}$
so the model becomes $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1^2$