Applied Linear Statistical Models

Homework 6

Qiuyang Yin, 2015011468

May 14th, 2017

Contents

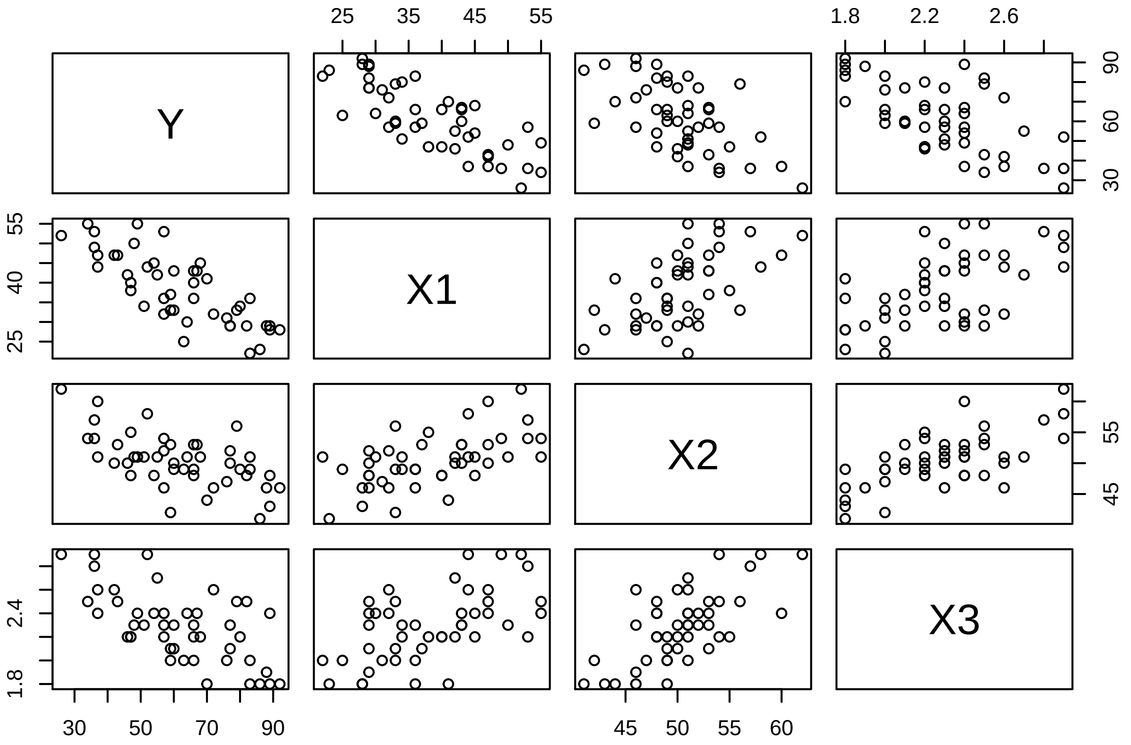
Problem 1

KNNL 6.15

b

- scatterplot
- correlation matirx
- interpret and state principal findings

```r
options(warn = -1)
dat=read.table("CH06PR15_300201404.txt")
colnames(dat)=c("Y","X1","X2","X3")
## scatter plot
pairs(dat)
```



```r
## correlation matrix
cor(dat)
```

```
##              Y         X1         X2         X3
## Y    1.0000000 -0.7867555 -0.6029417 -0.6445910
## X1 -0.7867555  1.0000000  0.5679505  0.5696775
## X2 -0.6029417  0.5679505  1.0000000  0.6705287
## X3 -0.6445910  0.5696775  0.6705287  1.0000000
```

The scatter plot and correlation matrix both demostrate that Y have a relatively strong negative relationship with both X1, X2 and X3, and X1 has the largest correlation with Y, indicating the smallest p-value in the t test and the largest extra sum of squares among variables.

In addition, X1 , X2, and X3 have a strong correlation between any two of them, indicating chances are that they may have multicollinearity problem in the latter discussion.

c

- Fit the model(6.5)
- State the estimated regression function
- Interpret b2

```
reg=lm(Y~X1+X2+X3,data = dat,x=TRUE,y=TRUE)
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = dat, x = TRUE, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## X1           -1.1416     0.2148  -5.315 3.81e-06 ***
## X2           -0.4420     0.4920  -0.898   0.3741
## X3          -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

So the estimated regression model is:

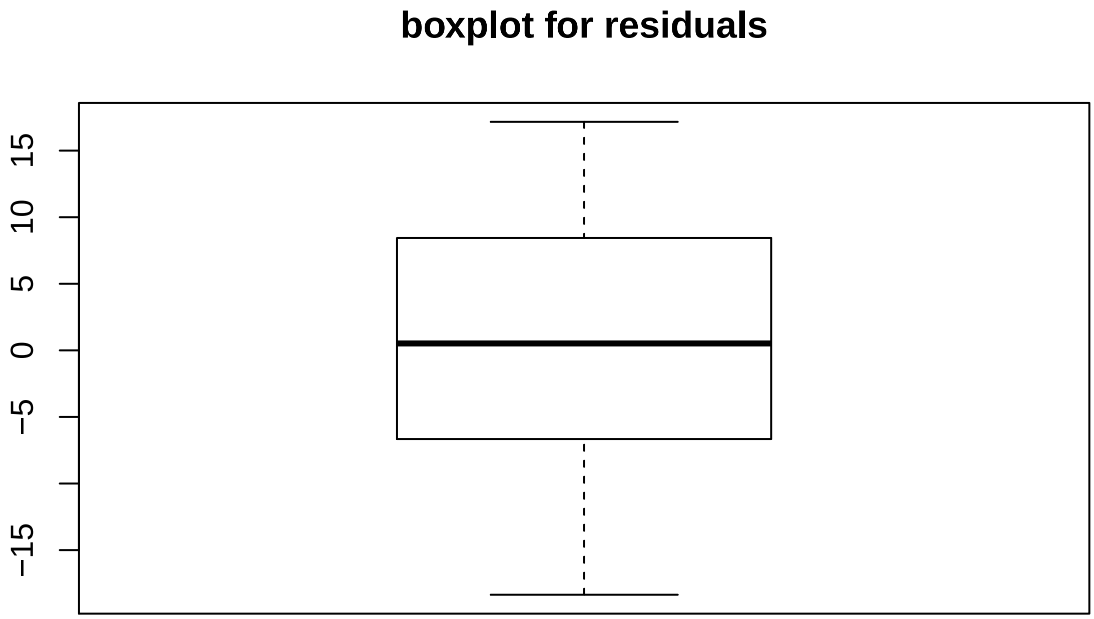
$$\hat{Y}_i = 158.4913 - 1.1416X_{i1} - 0.4420X_{i2} - 13.4702X_{i3}$$

And $b2 = -0.4420$ represents that, when X_2 has an unit increase, the mean response of Y is -0.4420. i.e, Y will decrease 0.4420 in average when X_2 increases in one.

d

- boxplot of residuals
- any outliers?

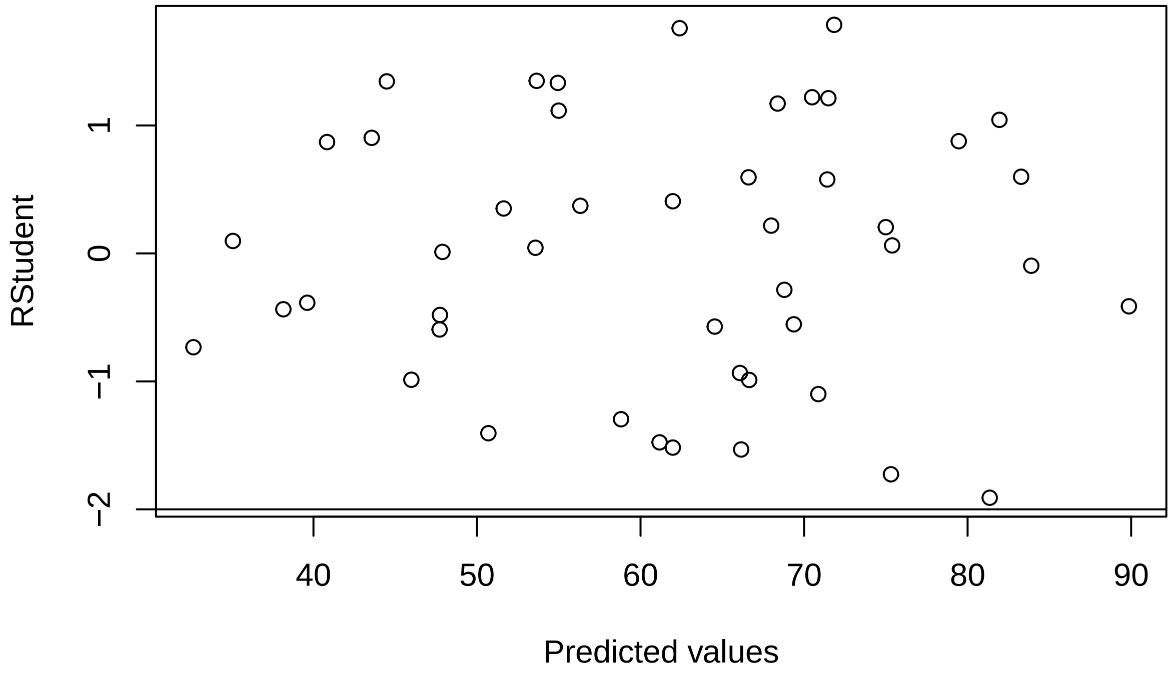
```
boxplot(reg$residuals,main="boxplot for residuals")
```



There doesn’t seem to have any outlier according to boxplot(no dots lie outside the border)

Let us judge outlier from studentized residuals:

```
plot(rstandard(reg)~reg$fitted.values, xlab="Predicted values", ylab="RStudent")
abline(h=2)
abline(h=-2)
```



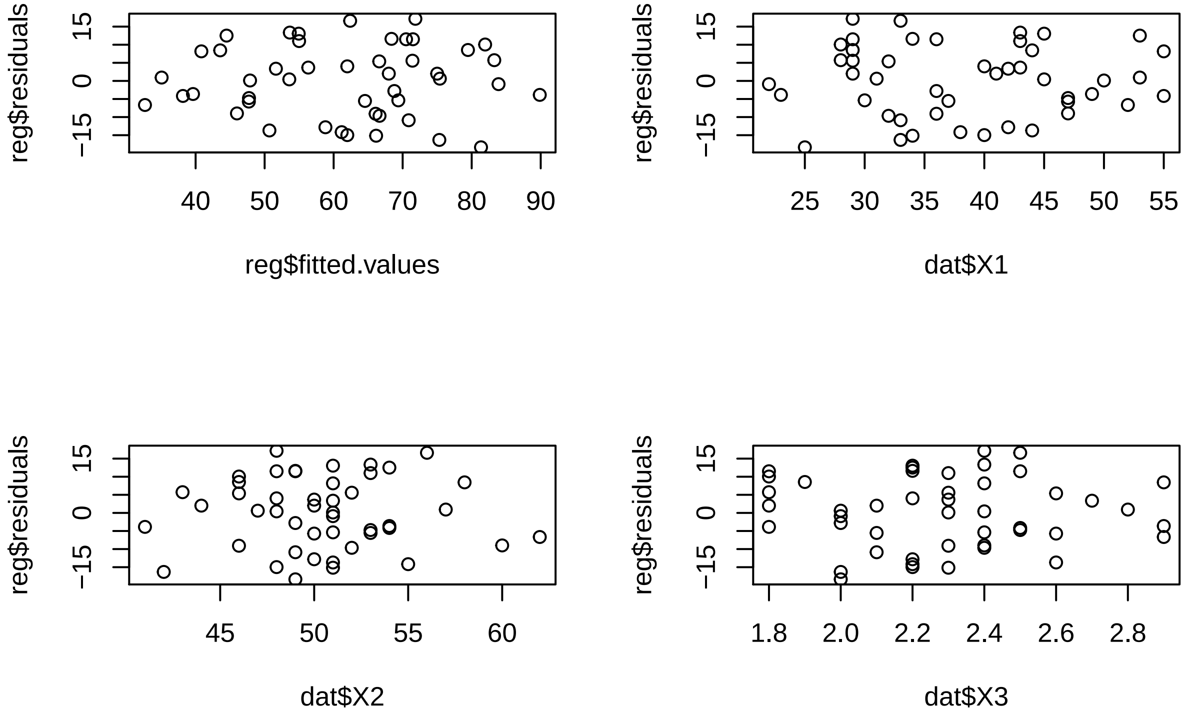
All the dots are between -2 and 2, so we can conclude that there are no outliers in this model.

e

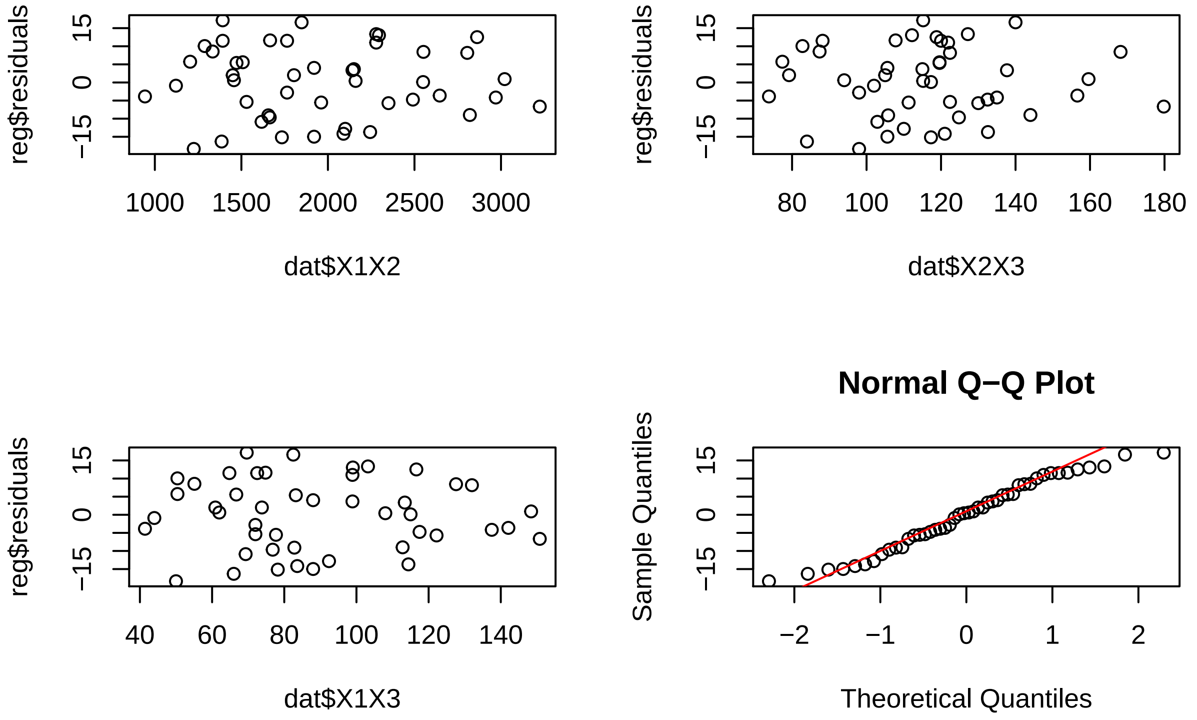
- residuals against $\hat{Y}$, each of variable and two-factor interaction term.
- a normal probability plot.
- interpret

To prevent to be too lengthy(if we plot 8 plots separately). I put four graphs in one plot. I don't think it will bother as long as we can observe the graph clearly.

```
dat$X1X2=dat$X1*dat$X2
dat$X1X3=dat$X1*dat$X3
dat$X2X3=dat$X2*dat$X3
plot.new()
par(mfrow=c(2,2))
plot(reg$residuals~reg$fitted.values)
plot(reg$residuals~dat$X1)
plot(reg$residuals~dat$X2)
plot(reg$residuals~dat$X3)
```



```r
plot(reg$residuals~dat$X1X2)
plot(reg$residuals~dat$X2X3)
plot(reg$residuals~dat$X1X3)
qqnorm(reg$residuals)
qqline(reg$residuals, col = "red")
```



```r
par(mfrow=c(1,1))
```

In terms of the seven plots related to residuals. Firstly, I haven't seen any obvious trend with any of the variable, so I can roughly conclude that the residuals are **independent**.

With regard to the heteroscedasticity(variance). I have spotted a slight increase in variance with the increase in $\hat{Y}$, and a slight decrease in variance with the increase in $X1$,$X1X2$ and $X1X3$. Maybe we need *Homogeneity of variance test* to reach a more precise conclusion.

With regard to the normality, the dots fit well in the Q-Q plots, except for several outliers. We can roughly conclude that the residuals are **normal**.

KNNL 6.16

a

- Test the regression relation use $\alpha = 0.10$
- state the alternatives, decision rule and the conclusion.
- p value
- what does the result imply about β_1 ,β_2 and β_3?

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1 : \beta_1^2 + \beta_2^2 + \beta_3^2 > 0$$

H_1 also represents that not all the $\beta_k(k = 1, 2, 3)$ equal zero.

We use the F test statistic:

$$F^* = MSR/MSE$$

And the decision rule is :

If $F^* \leq F(1-\alpha; p-1, n-p)$, conclude H_0

If $F^* < F(1-\alpha; p-1, n-p)$, conclude H_1

Here $p = 4$,$n = 46$,and $\alpha = 0.1$.

And the conclusion:

```r
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = dat, x = TRUE, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.3524  -6.4230   0.5196   8.3715  17.1601
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 158.4913    18.1259   8.744 5.26e-11 ***
## X1           -1.1416     0.2148  -5.315 3.81e-06 ***
## X2           -0.4420     0.4920  -0.898   0.3741
## X3          -13.4702     7.0997  -1.897   0.0647 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.06 on 42 degrees of freedom
## Multiple R-squared:  0.6822, Adjusted R-squared:  0.6595
## F-statistic: 30.05 on 3 and 42 DF,  p-value: 1.542e-10
```

The result lies in the bottom of the summary. $F^* = 30.05$, $p - value = 1.542e - 10$. Since p-value is smaller than 0.10, we reject H_0.

The test imply that at least one in the β_1 ,β_2 and β_3 is not zero.

b

- joint interval of β_1 ,β_2 and β_3 using $\alpha = 0.10$
- Interpret

Here I use the bofferroni method:

```r
confint(reg,level = 1-0.1/3)
```

```
##                   1.67 %     98.33 %
## (Intercept) 118.607631 198.3748720
## X1           -1.614248  -0.6689755
## X2           -1.524510   0.6405013
## X3          -29.092028   2.1517012
```

For Bonferroni joint confidence. $1-\alpha/3$ is appiled for $1-\alpha$, so 1-0.1/3 is used. We conclude that β_1 is between -1.614 and -0.6690 , β_2 is between -1.524 and 0.6405, and β_3 is between -29.092 and 2.1517. The familt confidence is at least 0.99.

c

- coefficient of the multiple determination
- indication?

```
summary(reg)$r.square
```

```
## [1] 0.6821943
```

The R^2 here is 0.6821943, indicating that 68.21% of variance have been explained. It is good to explain, however a little bit low for prediction.

KNNL 6.17

a

- interval estimate of mean satisfaction of 35,45,2.2, $\alpha = 0.1$.
- interpret

```
ans=predict(reg,se.fit = T,data.frame(X1=35,X2=45,X3=2.2),interval = "confidence",level = 0.90)
ans$fit
```

```
##        fit      lwr      upr
## 1 69.01029 64.52854 73.49204
```

We conclude with confidence coefficient 0.9 that the mean satisfaction when patient age is 35, severity of illness is 45, and anxiety level is 2.2 is somewhere between 64.52854 and 73.49204[1].

b

- interval estimate of prediction of 35,45,2.2, $\alpha = 0.1$.
- interpret

```
ans=predict(reg,se.fit = T,data.frame(X1=35,X2=45,X3=2.2),interval = "predict",level = 0.90)
ans$fit
```

```
##        fit      lwr      upr
## 1 69.01029 51.50965 86.51092
```

With confidence coefficient 0.90, we predict that the satisfaction for patient whose age is 35, severity of illness is 45, and anxiety level is 2.2 will be somewhere between 51.50965 and 86.51092[2].

[1] KNNL P54

[2] KNNL P59

Problem 2

KNNL 7.5

a

- obtain ANOVA table
- Extra sum of square into X2, X1|X2, and X3|X1,X2

```
reg2=lm(Y~X2+X1+X3,data = dat,x=TRUE,y=TRUE)
anova(reg2)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X2         1 4860.3  4860.3 48.0439 1.822e-08 ***
## X1         1 3896.0  3896.0 38.5126 2.008e-07 ***
## X3         1  364.2   364.2  3.5997   0.06468 .
## Residuals 42 4248.8   101.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So we can conclude that SSM(X2)=4860.3,SSM(X1|X2)=3896.0,SSM(X3|X1,X2)=364.2

b

- Test whether X_3 can be dropped from the regression given that X1 and X2 is retained.
- State the alternatives, decision rule and conclusion.
- P value

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

And the F statistic is :

$$F^* = \frac{MSR(X3|X1, X2)}{MSE(X1, X2, X3)}$$

If $F^* \leq F(1-\alpha; 1, n-4)$, conclude H_0

If $F^* < F(1-\alpha; 1, n-4)$, conclude H_1

There two methods to do the test. Firstly, we use the type 3 extra sum of squares.

```
require(car)
```

```
## Loading required package: car
```

```
Anova(reg,type = "3")# method 1
```

```
## Anova Table (Type III tests)
##
## Response: Y
##             Sum Sq Df F value    Pr(>F)
## (Intercept) 7734.5  1 76.4561 5.261e-11 ***
## X1          2857.6  1 28.2471 3.810e-06 ***
## X2            81.7  1  0.8072   0.37407
## X3           364.2  1  3.5997   0.06468 .
## Residuals   4248.8 42
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

And another way to do this is to use the general linear test:

```
full=lm(Y~X1+X2+X3,data = dat,x=TRUE,y=TRUE)
reduced=lm(Y~X1+X2,data = dat,x=TRUE,y=TRUE)
anova(reduced,full)# method 2
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     43 4613.0
## 2     42 4248.8  1    364.16 3.5997 0.06468 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both of the p-values are 0.06468. Since it is greater than 0.025, we can not reject H_0, so we conclude that X3 can be dropped from the regression line.

KNNL 7.6

- Test whether both X_2 and X_3 can be dropped from the regression funtion.
- state the alternatives, decision rule and conclusion.
- p-value.

$$H_0 : \beta_2 = \beta_3 = 0$$
$$H_1 : \beta_2^2 + \beta_3^2 \neq 0$$

And the F statistic is :

$$F^* = \frac{MSR(X2, X3|X1)}{MSE(X1, X2, X3)}$$

If $F^* \leq F(1-\alpha; 2, n-4)$, conclude H_0

If $F^* < F(1-\alpha, 2, n-4)$, conclude H_1

The first method is to calculate F based on the definition:

```
SSR=480.9+364.2# type 3 extra sum of squares.
sigma=summary(reg)$sigma
(F=SSR/2/(sigma^2))# method 1
```

```
## [1] 4.176928
```

And the second method is to use the general linear test:

```
full=lm(Y~X1+X2+X3,data = dat,x=TRUE,y=TRUE)
reduced=lm(Y~X1,data = dat,x=TRUE,y=TRUE)
anova(reduced,full)# method 2
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     44 5093.9
## 2     42 4248.8  2    845.07 4.1768 0.02216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is 0.02216, which is smaller than 0.025, we reject H_0 and conclude that neither X2 or X3 can be dropped.

KNNL 7.9

- Test whether $\beta_1 = -1.0$ and $\beta_2 = 0$,using $\alpha = 0.025$
- State the alternatives, decision rule and conclusion.

$$H_0 : \beta_1 = -1.0$$
$$H_1 : \beta_1 \neq -1.0$$

And the t statistic is :

$$t^* = \frac{b_1 + 1}{s(b_1)}$$

If $|t^*| \leq t(1 - \alpha/2, n - p)$, conclude H_0

Otherwise conclude H_1

```
result=summary(reg)
mdf=result$df[2]
sb1=result$coefficients[6]
b1=result$coefficients[2]
T1=(b1+1)/sb1
(p.value1=2*(1-pt(abs(T1),df = mdf)))
```

```
## [1] 0.5133169
```

The p-value is 0.51, so we can fail to reject H_0 and conclude that $\beta_1 = -1$

And another test is just identical to the previous one:

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

And the t statistic is :

$$t^* = \frac{b_2}{s(b_2)}$$

If $|t^*| \leq t(1-\alpha/2, n-p)$, conclude H_0

Otherwise conclude H_1

```
sb2=result$coefficients[7]
b2=result$coefficients[3]
T2=(b2)/sb2
(p.value2=2*(1-pt(abs(T2),df = mdf)))
```

```
## [1] 0.3740702
```

Since the p-value is 0.37, we fail to reject H_0 and conclude that $\beta_2 = 0$

Problem 3

KNNL 7.26

a

- Fit the first-order linear regression model for X1 and X2
- State the fitted regression model

```
reg3=lm(Y~X1+X2,data = dat,x=TRUE,y=TRUE)
summary(reg3)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = dat, x = TRUE, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1662  -8.5462  -0.4595   7.1342  17.2364
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 156.6719    18.6396   8.405 1.27e-10 ***
## X1           -1.2677     0.2104  -6.026 3.35e-07 ***
## X2           -0.9208     0.4349  -2.117   0.0401 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 10.36 on 43 degrees of freedom
## Multiple R-squared:  0.655,  Adjusted R-squared:  0.6389
## F-statistic: 40.81 on 2 and 43 DF,  p-value: 1.16e-10
```

$$\hat{Y}_i = 156.6719 - 1.2677X_{i1} - 0.9208X_{i2}$$

b

- Compare the results with Part 6.15c

```
reg=lm(Y~X1+X2+X3,data = dat,x=TRUE,y=TRUE)
result1=summary(reg)
result1$coefficients
```

```
##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept) 158.4912517 18.1258887  8.7439162 5.260955e-11
## X1           -1.1416118  0.2147988 -5.3147960 3.810252e-06
## X2           -0.4420043  0.4919657 -0.8984452 3.740702e-01
## X3          -13.4701632  7.0996608 -1.8972967 6.467813e-02
```

```
result2=summary(reg3)
result2$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 156.6718598 18.6396443  8.405303 1.273843e-10
## X1           -1.2676542  0.2103519 -6.026351 3.347580e-07
## X2           -0.9207881  0.4348935 -2.117273 4.005967e-02
```

We can see from the result above that the estimated coefficients of X1 and X2 change when we include another variable X3, which is the result of the multicolinearity.

c

- SSR(X1) and SSR(X1|X3)
- SSR(X2) and SSR(X2|X3)

```
(SSR.X1=anova(reg)$`Sum Sq`[1])
```

```
## [1] 8275.389
```

```
temp=lm(Y~X3+X1,data = dat,x=TRUE,y=TRUE)
(SSR.X1_X3=anova(temp)$`Sum Sq`[2])
```

```
## [1] 3483.891
```

```r
temp=lm(Y~X2,data = dat,x=TRUE,y=TRUE)
(SSR.X2=anova(temp)$`Sum Sq`[1])
```

```
## [1] 4860.26
```

```r
temp=lm(Y~X3+X2,data = dat,x=TRUE,y=TRUE)
(SSR.X2_X3=anova(temp)$`Sum Sq`[2])
```

```
## [1] 707.9971
```

So we can conclude that $SSR(X1) = 8273$ and $SSR(X1|X3) = 3483$, they are not equal. $SSR(X2) = 4860$ and $SSR(X2|X3) = 708$, they are not equal,too.

d

```r
cor(dat)[1:4,1:4]
```

```
##             Y         X1         X2         X3
## Y   1.0000000 -0.7867555 -0.6029417 -0.6445910
## X1 -0.7867555  1.0000000  0.5679505  0.5696775
## X2 -0.6029417  0.5679505  1.0000000  0.6705287
## X3 -0.6445910  0.5696775  0.6705287  1.0000000
```

Because of the multicolinearity between X1,X2 and X3. Consequences are that:

- extra sum of square may not equal sum of square

Just like what we have seen in part c, $SSR(X1)$ and $SSR(X1|X3)$, they are not equal. $SSR(X2)$ and $SSR(X2|X3)$, they are not equal,too.

- coeffenct standard error may become larger, leading to the failure in t test.
- coefficient may not make sense

Just like what we have seen in part b, the coefficient changes when another variable is included.

KNNL 7.29

a

$$\begin{aligned}
&SSR(X_1) + SSR(X_2, X_3|X_1) + SSR(X_4|X_1, X_2, X_3) \\
&= SSR(X_1) + SSR(X_1, X_2, X_3) - SSR(X_1) + SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3) \\
&= SSR(X_1, X_2, X_3, X_4)
\end{aligned} \tag{1}$$

b

$$\begin{aligned}
&SSR(X_2, X_3) + SSR(X_1|X_2, X_3) + SSR(X_4|X_1, X_2, X_3) \\
&= SSR(X_2, X_3) + SSR(X_1, X_2, X_3) - SSR(X_2, X_3) + SSR(X_1, X_2, X_3, X_4) - SSR(X_1, X_2, X_3) \\
&= SSR(X_1, X_2, X_3, X_4)
\end{aligned} \tag{2}$$

Problem 4

KNNL 8.16

a

- Explain how each regression coefficient in model (8.33) is interpreted here.

```
rm(list = ls())
dat2=read.table("CH01PR19_818607472.txt")
temp=read.table("CH08PR16_703809665.txt")
dat2=data.frame(dat2,temp)
colnames(dat2)=c("Y","X1","X2")
dat2$X1X2=dat2$X1*dat2$X2
```

The model 8.33 is:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 Xi2 + \epsilon_i$$

where X_{i1} stands for the ACT test score, and X_{i2} stands for whether studenthad chosen a major field of concentration at the time the application was submitted.

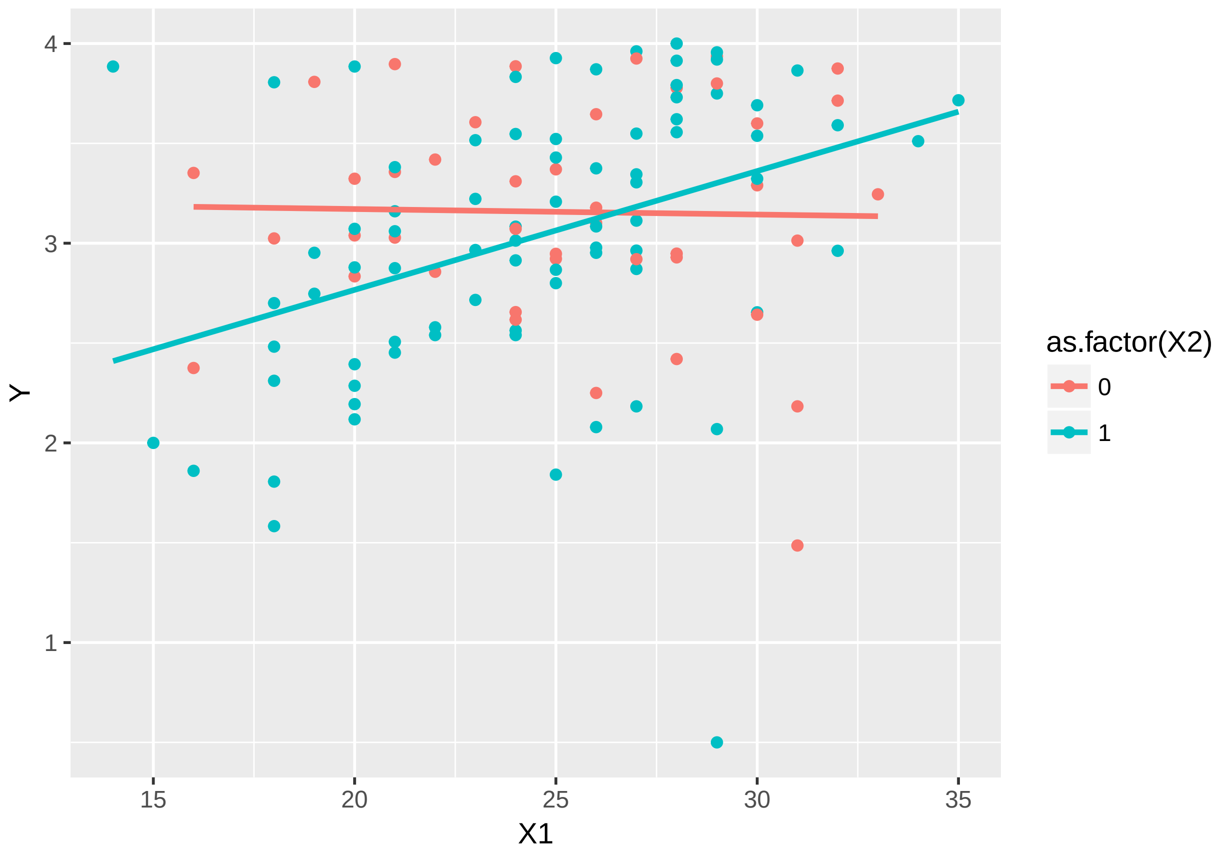
If model 8.33 is applied, β_0 stands for the intercept for student who have chosen the major field, while $\beta_0 + \beta_2$ stands for the intercept for student who have **not** chosen the major field. They share the same slope of β_1, which is the average response to the unit increase of ACT test score.

We may take a look at the model from the figure below:

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
ggplot(dat2, aes(y=Y,x=X1,col=as.factor(X2))) +
  geom_point() +
  geom_smooth(method = "lm",se=FALSE)
```



b

- Fit the regression model
- state the estimated regression function

```
reg=lm(Y~X1+X2,data = dat2,x=TRUE,y=TRUE)
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = dat2, x = TRUE, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70304 -0.35574  0.02541  0.45747  1.25037
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19842    0.33886   6.488 2.18e-09 ***
## X1           0.03789    0.01285   2.949  0.00385 **
## X2          -0.09430    0.11997  -0.786  0.43341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.6241 on 117 degrees of freedom
## Multiple R-squared:  0.07749,    Adjusted R-squared:  0.06172
## F-statistic: 4.914 on 2 and 117 DF,  p-value: 0.008928
```

So the estimated regression function is :

$$\hat{Y}_i = 2.20 + 0.04X_{i1} - 0.09X_{i2}$$

c

- Test whether X_2 can be dropped from the estimated regression model
- State the alternatives, decision rule and conclusion.

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

And the t statistic is :

$$t^* = \frac{b_2}{s(b_2)}$$

If $|t^*| \leq t(1-\alpha/2, n-p)$, conclude H_0

Otherwise conclude H_1

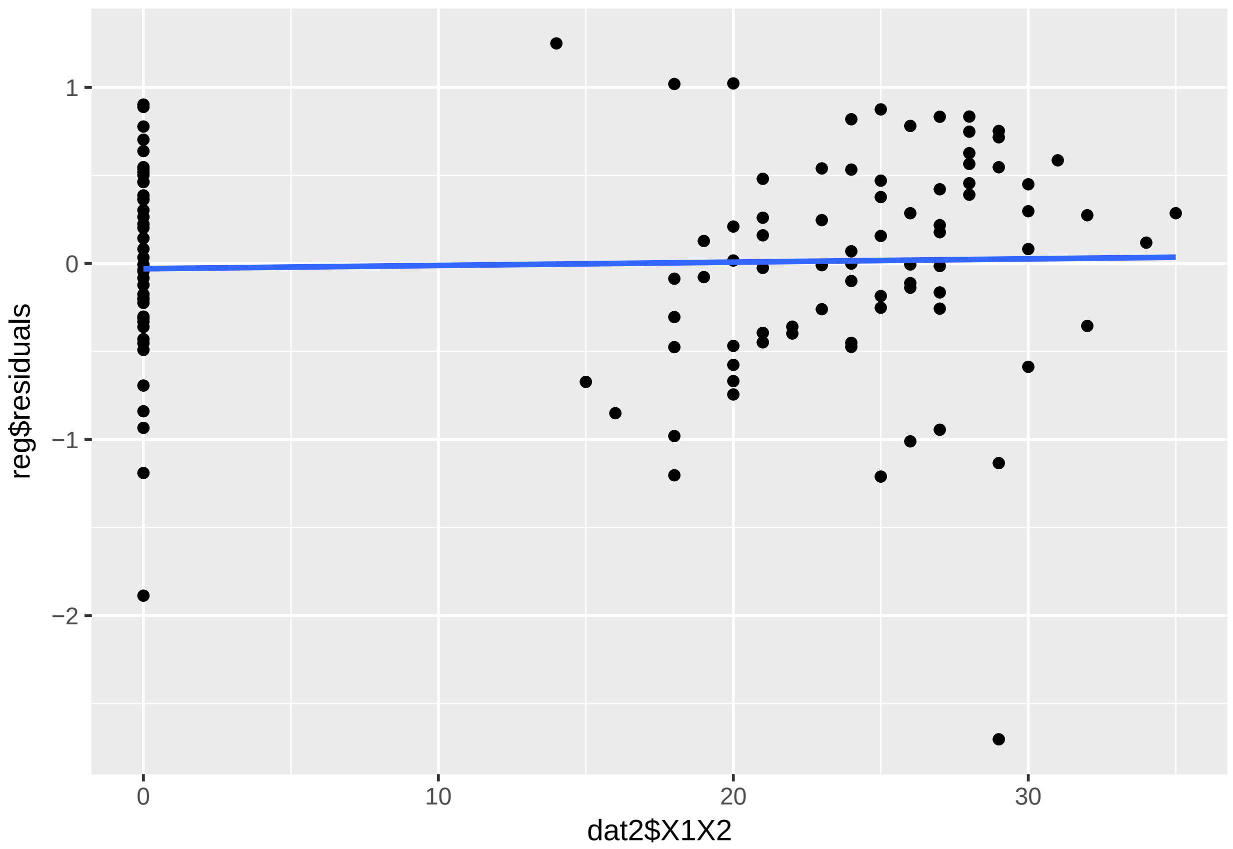
```r
result=summary(reg)
result$coefficients[3,4]
```

```
## [1] 0.4334062
```

Since the p-value of the t test is 0.433>0.01, we fail to reject H_0 and conclude that X_2 can be dropped from the regression model.

d

```r
par(mfrow=c(1,1))
ggplot(data.frame(reg$residuals,dat2$X1X2),aes(y=reg$residuals,x=dat2$X1X2))+
  geom_point()+
  geom_smooth(method="lm",se=FALSE)
```



According to smooth line, I don't think X1X2 is a helpful term, although there is a slight positive relationship.

KNNL 8.20

a

```
reg=lm(Y~X1+X2+X1X2,data = dat2,x=TRUE,y=TRUE)
summary(reg)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1X2, data = dat2, x = TRUE, y = TRUE)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.80187 -0.31392  0.04451  0.44337  1.47544
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.226318   0.549428   5.872 4.18e-08 ***
## X1          -0.002757   0.021405  -0.129   0.8977
## X2          -1.649577   0.672197  -2.454   0.0156 *
## X1X2         0.062245   0.026487   2.350   0.0205 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6124 on 116 degrees of freedom
## Multiple R-squared:  0.1194, Adjusted R-squared:  0.09664
## F-statistic: 5.244 on 3 and 116 DF,  p-value: 0.001982
```

So the estimated regression function is :

$$\hat{Y}_i = 3.226 - 0.003X_{i1} - 1.650X_{i2} + 0.062X_{i1}X_{i2}$$

b

- Test whether X_1X_2 can be dropped from the estimated regression model
- State the alternatives, decision rule and conclusion.

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

And the t statistic is :

$$t^* = \frac{b_3}{s(b_3)}$$

If $|t^*| \leq t(1 - \alpha/2, n - p)$, conclude H_0

Otherwise conclude H_1

```
result=summary(reg)
result$coefficients[4,4]
```

```
## [1] 0.02046149
```

Since p-value is 0.0205<0.05, we reject H_0 and conclude that the interaction term can not be dropped from the model.

Accually unlike model 8.33, here we have different slope for different groups of students: $\beta_1 + \beta_3$ for student having chosen major and β_1 for students haven't chosen major. The coefficient of β_3 stands for the difference in slope between two groups.

Problem 5

a

- Indicate which subset of predictors you would recommend as best for predicting patient satisfaction.

```
rm(list=ls())
dat=read.table("CH06PR15_300201404.txt")
colnames(dat)=c("Y","X1","X2","X3")
fit=lm(Y~X1+X2+X3,data = dat)
sigsqhat.big <- summary(fit)$sigma^2
library(leaps)
```

```
predictors = dat[,c("X1", "X2","X3")]
response = dat$Y

nmodels=7
leapSet = leaps(x=predictors, y=response, nbest = nmodels)

n = nrow(dat)
cp <- double(nmodels)
aic <- double(nmodels)
bic <- double(nmodels)
cvss <- double(nmodels)
adj.rsquared <- double(nmodels)
for( i in 1:nmodels){
  selectVarsIndex = leapSet$which[i,]
  newData <- cbind(response, predictors[, selectVarsIndex])
  newData <- as.data.frame(newData)
  selectedMod <- lm(response ~ ., data=newData)  # build model
  summary(selectedMod)
  adj.rsquared[i] <- summary(selectedMod)$adj.r.squared
  aic[i] <- AIC(selectedMod)
  bic[i] <- AIC(selectedMod, k = log(n))
  cvss[i] <- sum((selectedMod$residuals/(1 - hatvalues(selectedMod)))^2)
  cp[i] <- sum(selectedMod$residuals^2)/sigsqhat.big + 2 * selectedMod$rank - n
}

models = leapSet$which
colnames(models) = c("X1","X2","X3")
bestModels = cbind(adj.rsquared, cp, aic, bic, cvss, models)
bestModels
```

```
##   adj.rsquared        cp      aic      bic     cvss X1 X2 X3
## 1    0.6103248  8.353606 353.0717 358.5577 5569.562  1  0  0
## 1    0.4022134 35.245643 372.7561 378.2420 8451.432  0  0  1
## 1    0.3490737 42.112324 376.6735 382.1595 9254.489  0  1  0
## 2    0.6610206  2.807204 347.6030 354.9176 4902.751  1  0  1
## 2    0.6389073  5.599735 350.5100 357.8246 5235.192  1  1  0
## 2    0.4437314 30.247056 370.3874 377.7019 8115.912  0  1  1
## 3    0.6594939  4.000000 348.7273 357.8705 5057.886  1  1  1
```

We can conclude from all the criteria that model(X1,X3) is the best subset of predictor.(Largest in adj.rsquared, min in cp, aic,bic and PRESS)

b

- Do the four criteria identify the same best subset
- Does this always happen?

Yes!

No, this does not always happen, there are many examples when different criteria can lead to totally different 'best model'.TODO

c

No, since only there are only $2^3 = 8$ models.

Problem 6

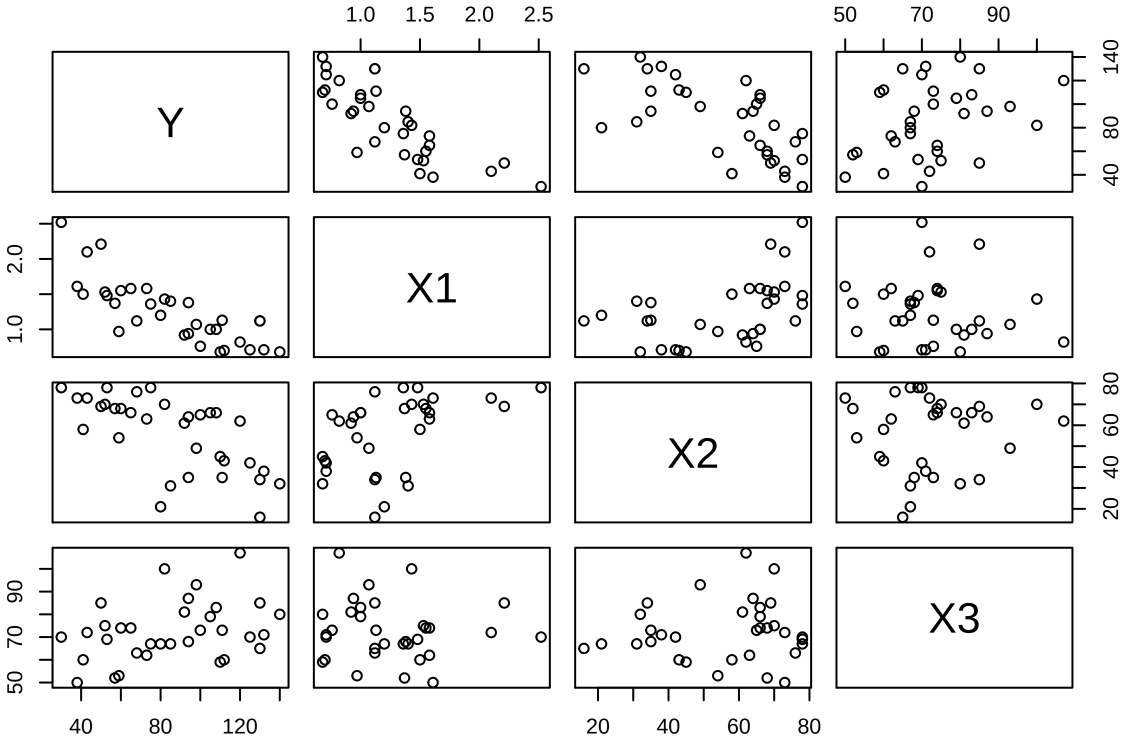
KNNL 9.15

```
rm(list = ls())
dat=read.table("CH09PR15_11905844.txt")
colnames(dat)=c("Y","X1","X2","X3")
```

b

- scatter plot and correlation matrix
- discuss

```
pairs(dat)
```



```
cor(dat)
```

```
##               Y          X1          X2          X3
## Y   1.0000000 -0.80181086 -0.66787239  0.34591487
## X1 -0.8018109  1.00000000  0.46773179 -0.08898262
## X2 -0.6678724  0.46773179  1.00000000  0.06848147
## X3  0.3459149 -0.08898262  0.06848147  1.00000000
```

We can conclude from the scatter plot and correaltion matrix that response variable (Y) has a strong negative relationship with X1 and X2, and a positive relationship with X3.

There may have some multicollinearity problems because the correaltion between X1 and X2 is 0.46, which is relatively high.

c

```r
fit=lm(Y~X1+X2+X3,data = dat)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = dat)
##
## 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
```

All of the p-values are small, so all of the predictors should be retained.

KNNL 9.16

a

```r
dat$X1=dat$X1-mean(dat$X1)
dat$X2=dat$X2-mean(dat$X2)
dat$X3=dat$X3-mean(dat$X3)
dat$X1X2=dat$X1*dat$X2
dat$X1X3=dat$X1*dat$X3
dat$X2X3=dat$X3*dat$X2
```

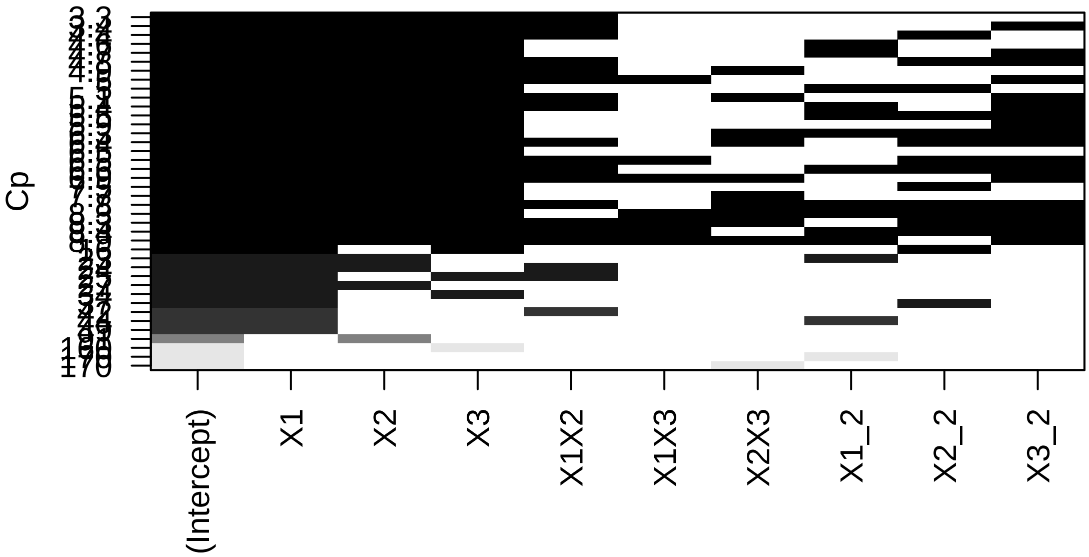
```r
dat$X1_2=dat$X1*dat$X1
dat$X2_2=dat$X2*dat$X2
dat$X3_2=dat$X3*dat$X3
predictors = dat[,c(-1)]
response = dat$Y
```

There are two criteria using Cp method:

- Cp is small
- Cp is near p

Here is an approach to visualizing, however I will not use this method:

```r
test=regsubsets(x=predictors,y=response, nbest = 5)
plot(test,scale = "Cp")#visualizing
```



Now let's begin:

Firstly

let's pick three smallest Cps and check their models:

```
require(leaps)


ans=leaps(x=predictors,y=response,nbest = 30,method = "Cp")
model=ans$which
colnames(model)=colnames(predictors)

(s=sort(ans$Cp)[1:3])
```

```
## [1] 3.302215 3.384990 4.447976
```

```
o=order(ans$Cp)[1:3]
model[o,]
```

```
##     X1   X2   X3 X1X2  X1X3  X2X3  X1_2  X2_2  X3_2
## 4 TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE
## 5 TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE  TRUE
## 5 TRUE TRUE TRUE TRUE FALSE FALSE FALSE  TRUE FALSE
```

So here three best models according to ‘smallest cp’ critera have been picked out. However, considering the second criteria, there Ps are:

```
ans$size[o]
```

```
## [1] 5 6 6
```

There is a large gap between cp and p.

Secondly

If we pick the models whose cp is nearest to cp:

```
(s=sort(abs(ans$Cp-ans$size))[1:5])
```

```
## [1] 3.552714e-15 2.274234e-02 9.576492e-02 1.388060e-01 1.443674e-01
```

```
o=order(abs(ans$Cp-ans$size))[1:5]
model[o,]
```

```
##     X1   X2   X3  X1X2 X1X3  X2X3  X1_2  X2_2  X3_2
## 9 TRUE TRUE TRUE  TRUE TRUE  TRUE  TRUE  TRUE  TRUE
## 6 TRUE TRUE TRUE FALSE TRUE FALSE  TRUE  TRUE FALSE
## 6 TRUE TRUE TRUE  TRUE TRUE  TRUE FALSE FALSE FALSE
## 7 TRUE TRUE TRUE  TRUE TRUE  TRUE FALSE  TRUE FALSE
## 8 TRUE TRUE TRUE  TRUE TRUE  TRUE  TRUE FALSE  TRUE
```

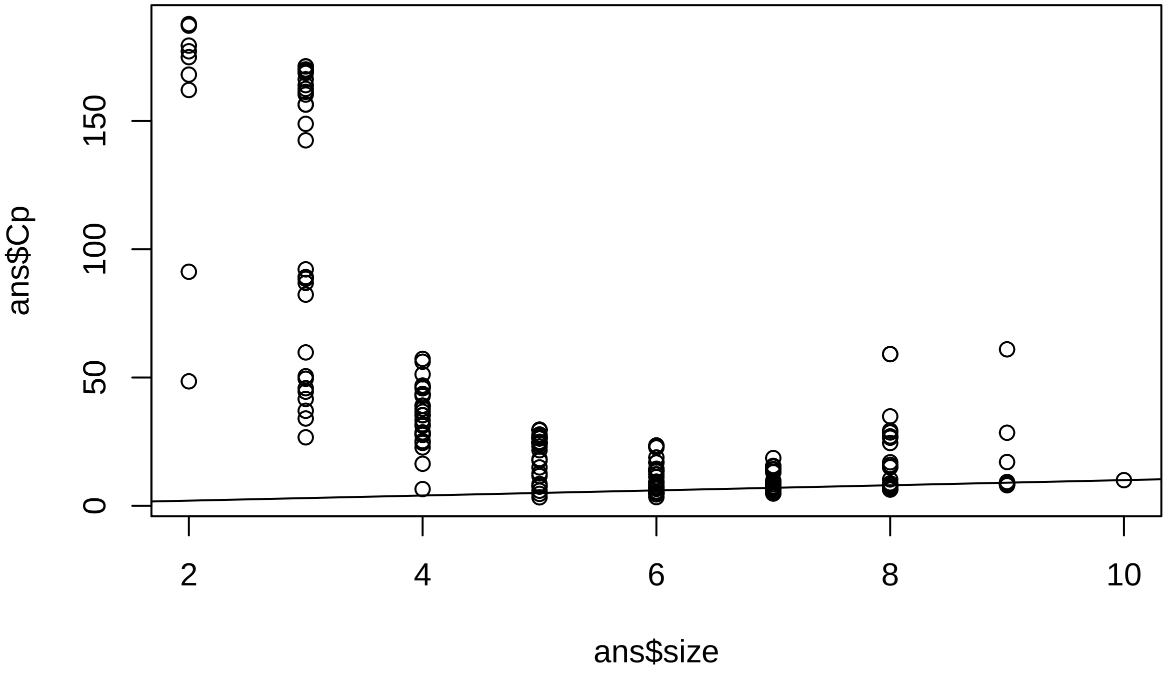
But their cps are too large:

```
ans$Cp[o]
```

```
## [1] 10.000000  7.022742  6.904235  7.861194  8.855633
```

Now let's check the cp~p plot:

```
plot(ans$size,ans$Cp)
abline(0,1)
```



Finally

So I choose the criteria below:

- |Cp-p| should smaller than 1
- Among all the models satisfy rule 1, pick smallest cp

```
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
## The following object is masked from 'package:car':
##
##     recode
```

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

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

```
tempdata=abs(ans$Cp-ans$size)
mydat=data.frame(ans$Cp,tempdata,ans$size,model)
mydat = mydat %>% filter(tempdata<1 & ans.size<=6)
mydat
```

```
##     ans.Cp  tempdata ans.size   X1   X2   X3  X1X2  X1X3  X2X3  X1_2  X2_2
## 1 4.639546 0.3604541        5 TRUE TRUE TRUE FALSE FALSE FALSE  TRUE FALSE
## 2 5.866135 0.8661352        5 TRUE TRUE TRUE FALSE FALSE FALSE FALSE FALSE
## 3 5.030767 0.9692331        6 TRUE TRUE TRUE FALSE FALSE FALSE  TRUE  TRUE
## 4 5.259446 0.7405539        6 TRUE TRUE TRUE  TRUE  TRUE FALSE FALSE FALSE
## 5 5.261941 0.7380588        6 TRUE TRUE TRUE  TRUE FALSE FALSE  TRUE FALSE
## 6 5.580645 0.4193550        6 TRUE TRUE TRUE FALSE FALSE  TRUE  TRUE FALSE
## 7 6.591733 0.5917333        6 TRUE TRUE TRUE FALSE  TRUE FALSE  TRUE FALSE
##    X3_2
## 1 FALSE
## 2  TRUE
## 3 FALSE
## 4 FALSE
## 5 FALSE
## 6 FALSE
## 7 FALSE
```

So the best subset according to my criteria above is :

- X1,X2,X3,X_1^2
- X1,X2,X3,X_1^2, X_2^2
- X1,X2,X3,X1X2, X1X3

However the term ‘near’ is quite subjective. If we set the threshold value |cp-p|=1.8:

```
tempdata=abs(ans$Cp-ans$size)
mydat=data.frame(ans$Cp,tempdata,ans$size,model)
mydat = mydat %>% filter(tempdata<1.8 & ans.size<=6)
mydat
```

```
##      ans.Cp  tempdata ans.size   X1   X2   X3  X1X2  X1X3  X2X3  X1_2
## 1  3.302215 1.6977851        5 TRUE TRUE TRUE  TRUE FALSE FALSE FALSE
## 2  4.639546 0.3604541        5 TRUE TRUE TRUE FALSE FALSE FALSE  TRUE
## 3  5.866135 0.8661352        5 TRUE TRUE TRUE FALSE FALSE FALSE FALSE
## 4  4.447976 1.5520241        6 TRUE TRUE TRUE  TRUE FALSE FALSE FALSE
## 5  4.683032 1.3169678        6 TRUE TRUE TRUE FALSE FALSE FALSE  TRUE
## 6  4.918929 1.0810705        6 TRUE TRUE TRUE  TRUE FALSE  TRUE FALSE
## 7  5.030767 0.9692331        6 TRUE TRUE TRUE FALSE FALSE FALSE  TRUE
## 8  5.259446 0.7405539        6 TRUE TRUE TRUE  TRUE  TRUE FALSE FALSE
## 9  5.261941 0.7380588        6 TRUE TRUE TRUE  TRUE FALSE FALSE  TRUE
## 10 5.580645 0.4193550        6 TRUE TRUE TRUE FALSE FALSE  TRUE  TRUE
## 11 6.591733 0.5917333        6 TRUE TRUE TRUE FALSE  TRUE FALSE  TRUE
## 12 7.131904 1.1319042        6 TRUE TRUE TRUE FALSE FALSE FALSE FALSE
## 13 7.212053 1.2120534        6 TRUE TRUE TRUE FALSE FALSE  TRUE FALSE
## 14 7.741465 1.7414651        6 TRUE TRUE TRUE FALSE  TRUE FALSE FALSE
##     X2_2  X3_2
## 1  FALSE FALSE
## 2  FALSE FALSE
## 3  FALSE  TRUE
## 4   TRUE FALSE
## 5  FALSE  TRUE
## 6  FALSE FALSE
## 7   TRUE FALSE
## 8  FALSE FALSE
## 9  FALSE FALSE
## 10 FALSE FALSE
## 11 FALSE FALSE
## 12  TRUE  TRUE
## 13 FALSE  TRUE
## 14 FALSE  TRUE
```

The best models selected is similar to the 'smallest cp' approach

- X1, X2, X3, X1X2
- X1, X2, X3, X1X2, X_2^2
- X1, X2, X3, X_1^2, X_2^2

TODO

b

We can see from the result above that accually there isn't much difference in Cp among 'best models'. When the threshold is 1, the three cps are 4.6,5.0 and 5.26. When the threshold is 1.8, the three cps are 3.3,4.44 and 4.64.

KNNL 9.19

a

- using foward stepwise regression with $\alpha = 0.10$ and $\alpha = 0.15$ [3]

```
min.model=lm(Y~1,data = dat)
biggest=(lm(Y~.,data = dat))
```

[3] http://stackoverflow.com/questions/22913774/forward-stepwise-regression

When $\alpha = 0.10$[4]

```
#a=0.10
m=qchisq(0.10,1,lower.tail=FALSE)
fwd.model = step(min.model, scope=list(lower=min.model, upper=biggest),
                 direction='both', k=m)
```

```
## Start:  AIC=228.59
## Y ~ 1
## 
##        Df Sum of Sq   RSS    AIC
## + X1    1   19927.0 11068 197.32
## + X2    1   13825.7 17170 211.81
## + X3    1    3708.8 27287 227.09
## + X1_2  1    2851.7 28144 228.11
## <none>              30996 228.59
## + X2X3  1    1878.3 29117 229.24
## + X1X2  1    1552.3 29443 229.60
## + X2_2  1    1236.6 29759 229.96
## + X1X3  1     124.9 30871 231.17
## + X3_2  1      38.1 30957 231.26
## 
## Step:  AIC=197.32
## Y ~ X1
## 
##        Df Sum of Sq     RSS    AIC
## + X2    1    3402.4  7666.1 187.90
## + X3    1    2355.3  8713.2 192.13
## + X2_2  1    1924.0  9144.5 193.72
## + X1X2  1    1267.6  9800.9 196.01
## <none>              11068.5 197.32
## + X1_2  1     860.5 10208.0 197.35
## + X2X3  1     669.1 10399.4 197.97
## + X3_2  1     142.8 10925.7 199.59
## + X1X3  1       1.9 11066.6 200.02
## - X1    1   19927.0 30995.5 228.59
## 
## Step:  AIC=187.9
## Y ~ X1 + X2
## 
##        Df Sum of Sq     RSS    AIC
## + X3    1    3166.1  4500.0 173.03
## + X1_2  1     848.8  6817.3 186.74
## <none>               7666.1 187.90
## + X1X2  1     597.8  7068.3 187.93
## + X2X3  1     151.8  7514.3 189.95
## + X1X3  1     120.7  7545.4 190.08
## + X2_2  1      30.3  7635.8 190.48
## + X3_2  1      10.0  7656.1 190.56
## - X2    1    3402.4 11068.5 197.32
## - X1    1    9503.8 17169.9 211.81
## 
```

[4] https://stats.stackexchange.com/questions/97257/stepwise-regression-in-r-critical-p-value

```
## Step:  AIC=173.03
## Y ~ X1 + X2 + X3
##
##         Df Sum of Sq     RSS    AIC
## + X1X2  1     744.0  3756.0 169.77
## + X1_2  1     553.0  3947.0 171.41
## + X3_2  1     377.9  4122.1 172.84
## <none>                4500.0 173.03
## + X2_2  1     150.2  4349.8 174.61
## + X2X3  1     112.0  4388.0 174.90
## + X1X3  1       2.2  4497.8 175.72
## - X3    1    3166.1  7666.1 187.90
## - X2    1    4213.2  8713.2 192.13
## - X1    1    7893.0 12393.0 203.75
##
## Step:  AIC=169.77
## Y ~ X1 + X2 + X3 + X1X2
##
##         Df Sum of Sq     RSS    AIC
## <none>                3756.0 169.77
## + X3_2  1     273.8  3482.2 169.98
## + X2_2  1     122.0  3634.0 171.38
## + X2X3  1      54.7  3701.3 171.99
## + X1X3  1       6.1  3749.9 172.42
## + X1_2  1       5.8  3750.2 172.42
## - X1X2  1     744.0  4500.0 173.03
## - X3    1    3312.3  7068.3 187.93
## - X2    1    3415.7  7171.7 188.41
## - X1    1    8135.2 11891.2 205.09
```

So the best model chosen is $Y \sim X_1 + X_2 + X_3 + X_1X_2$

If $\alpha = 0.15$

```
m=qchisq(0.15,1,lower.tail=FALSE)
fwd.model = step(min.model,scope=list(lower=min.model, upper=biggest),
                 direction='both', k=m)
```

```
## Start:  AIC=227.96
## Y ~ 1
##
##         Df Sum of Sq   RSS    AIC
## + X1    1   19927.0 11068 196.05
## + X2    1   13825.7 17170 210.54
## + X3    1    3708.8 27287 225.83
## + X1_2  1    2851.7 28144 226.85
## <none>              30996 227.96
## + X2X3  1    1878.3 29117 227.97
## + X1X2  1    1552.3 29443 228.34
## + X2_2  1    1236.6 29759 228.69
## + X1X3  1     124.9 30871 229.90
## + X3_2  1      38.1 30957 229.99
##
## Step:  AIC=196.05
## Y ~ X1
##
##         Df Sum of Sq     RSS    AIC
## + X2    1    3402.4  7666.1 186.00
## + X3    1    2355.3  8713.2 190.23
## + X2_2  1    1924.0  9144.5 191.82
## + X1X2  1    1267.6  9800.9 194.11
## + X1_2  1     860.5 10208.0 195.45
## <none>              11068.5 196.05
## + X2X3  1     669.1 10399.4 196.06
## + X3_2  1     142.8 10925.7 197.69
## + X1X3  1       1.9 11066.6 198.12
## - X1    1   19927.0 30995.5 227.96
##
## Step:  AIC=186
## Y ~ X1 + X2
##
##         Df Sum of Sq     RSS    AIC
## + X3    1    3166.1  4500.0 170.50
## + X1_2  1     848.8  6817.3 184.20
## + X1X2  1     597.8  7068.3 185.40
## <none>               7666.1 186.00
## + X2X3  1     151.8  7514.3 187.41
## + X1X3  1     120.7  7545.4 187.55
## + X2_2  1      30.3  7635.8 187.94
## + X3_2  1      10.0  7656.1 188.03
## - X2    1    3402.4 11068.5 196.05
## - X1    1    9503.8 17169.9 210.54
##
## Step:  AIC=170.49
## Y ~ X1 + X2 + X3
##
##         Df Sum of Sq     RSS    AIC
## + X1X2  1     744.0  3756.0 166.60
## + X1_2  1     553.0  3947.0 168.24
## + X3_2  1     377.9  4122.1 169.67
## <none>               4500.0 170.50
## + X2_2  1     150.2  4349.8 171.45
## + X2X3  1     112.0  4388.0 171.74
## + X1X3  1       2.2  4497.8 172.55
## - X3    1    3166.1  7666.1 186.00
## - X2    1    4213.2  8713.2 190.23
## - X1    1    7893.0 12393.0 201.85
##
## Step:  AIC=166.6
## Y ~ X1 + X2 + X3 + X1X2
##
##         Df Sum of Sq     RSS    AIC
## + X3_2  1     273.8  3482.2 166.18
## <none>               3756.0 166.60
## + X2_2  1     122.0  3634.0 167.59
## + X2X3  1      54.7  3701.3 168.19
## + X1X3  1       6.1  3749.9 168.62
## + X1_2  1       5.8  3750.2 168.62
## - X1X2  1     744.0  4500.0 170.50
## - X3    1    3312.3  7068.3 185.40
## - X2    1    3415.7  7171.7 185.88
## - X1    1    8135.2 11891.2 202.56
##
## Step:  AIC=166.18
## Y ~ X1 + X2 + X3 + X1X2 + X3_2
##
##         Df Sum of Sq     RSS    AIC
## <none>               3482.2 166.18
## - X3_2  1     273.8  3756.0 166.60
## + X2_2  1      88.3  3393.9 167.40
## + X1X3  1      52.8  3429.4 167.75
## + X2X3  1      46.5  3435.7 167.81
## + X1_2  1       2.2  3480.0 168.23
## - X1X2  1     639.9  4122.1 169.67
## - X2    1    2893.6  6375.8 184.06
## - X3    1    3548.2  7030.4 187.29
## - X1    1    8375.1 11857.3 204.54
```

So the best model chosen is $Y \sim X_1 + X_2 + X_3 + X_1X_2 + X_3^2$

b

```
ans=leaps(x=predictors,y=response,nbest = 30,method = "adjr2")
model=ans$which
colnames(model)=colnames(predictors)

(s=sort(ans$adjr2,decreasing = T)[1:3])
```

```
## [1] 0.8668497 0.8652362 0.8638250
```

```
o=order(ans$adjr2,decreasing = T)[1:3]
model[o,]
```

```
##     X1   X2   X3 X1X2  X1X3  X2X3  X1_2  X2_2 X3_2
## 5 TRUE TRUE TRUE TRUE FALSE FALSE FALSE FALSE TRUE
## 6 TRUE TRUE TRUE TRUE FALSE FALSE FALSE  TRUE TRUE
## 6 TRUE TRUE TRUE TRUE  TRUE FALSE FALSE FALSE TRUE
```

We can know from the result that the best model according to the adjusted R square is

$$Y \sim X_1 + X_2 + X_3 + X_1X_2 + X_3^2$$

which is identical to the foward stepwise regression when $\alpha = 0.15$