Math 564 Assignment 4

2022-10-22

———————–Mohammed Wasim R D(A20497053)——————

Problem1

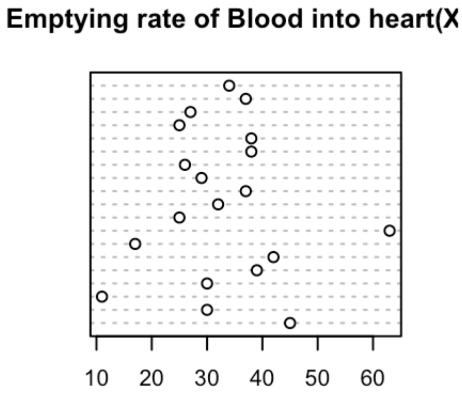
Reading data and renaming columns

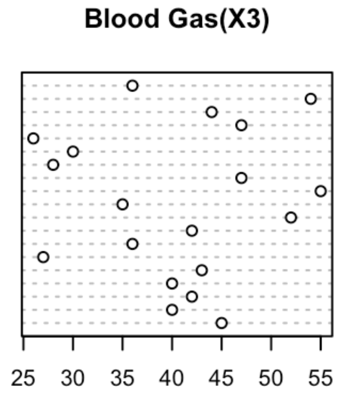
a.

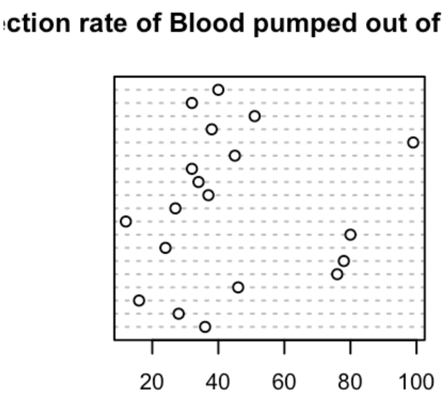
```
lung<-read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%20%209%20Da
ta%20Sets/CH09PR13.txt",header = F)
colnames(lung)[1] ="Y"
colnames(lung)[2]="X1"
colnames(lung)[3]="X2"
colnames(lung)[4]="X3"
head(lung)
```

```
##    Y X1 X2 X3
## 1 49 45 36 45
## 2 55 30 28 40
## 3 85 11 16 42
## 4 32 30 46 40
## 5 26 39 76 43
## 6 28 42 78 27
```

```
par(mfcol=c(2,3))
dotchart(lung$X1,main="Emptying rate of Blood into heart(X1)")
dotchart(lung$X2,main="Ejection rate of Blood pumped out of heart (X2)")
dotchart(lung$X3,main="Blood Gas(X3)")
```



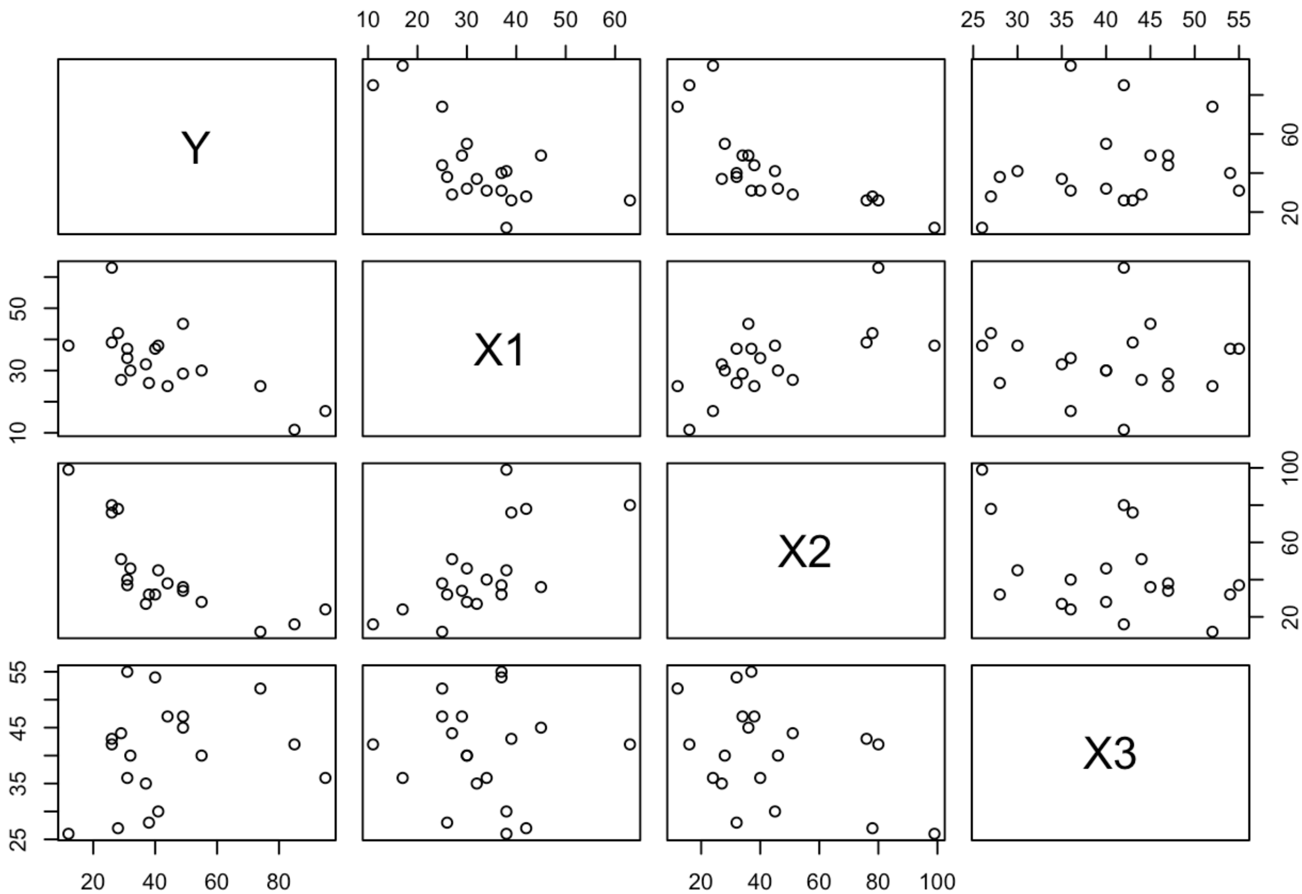




From the above graphs we can say that X1 is distributed from 10 to 60 and has one outlier , in X2 most of the data lies between 20 to 80 and has couple of outliers, also X3 is distributed between 25-55 ——————————————————————————————–

b.

```
pairs(lung)
```




```
cor(lung)
```

```
##              Y          X1         X2          X3
## Y   1.0000000 -0.66504734 -0.7475706  0.22386504
## X1 -0.6650473  1.00000000  0.6528513 -0.04613927
## X2 -0.7475706  0.65285127  1.0000000 -0.42348025
## X3  0.2238650 -0.04613927 -0.4234803  1.00000000
```

According to the scatter plot matrix, there is a negative linear association between Y and X1 and Y and X2, with Y and X3 showing little to no relationship. X1 and X2 appear to have a weak linear relationship, but even though they can

display multicollinearity, as can X2 and X3 and X1, which also appear to have some linear relationships.

---

c.

```
multiple_lm<-lm(Y~X1+X2+X3,data=lung)
summary(multiple_lm)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = lung)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.075 -12.064  -0.988   7.707  32.315
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 87.18750   21.55246   4.045  0.00106 **
## X1          -0.56448    0.42791  -1.319  0.20691
## X2          -0.51315    0.22449  -2.286  0.03723 *
## X3          -0.07196    0.45457  -0.158  0.87633
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.42 on 15 degrees of freedom
## Multiple R-squared:  0.6141, Adjusted R-squared:  0.5369
## F-statistic: 7.957 on 3 and 15 DF,  p-value: 0.002083
```

$\hat{Y} = 87.18750 - 0.56448X_1 - 0.51315X_2 - 0.07196X_3$ ## We can easily see that all of the predictor values should be kept because our linear model appears to benefit greatly from this.

Problem2

```
regsubsets <- lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2), lung)
summary(regsubsets)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2), data = lung)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.802  -6.452  -3.246   6.327  23.624
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 139.053349  16.647279   8.353 8.26e-07 ***
## X1           -2.996057   1.000293  -2.995  0.00964 **
## X2           -1.288050   0.598022  -2.154  0.04916 *
## I(X1^2)       0.034978   0.012516   2.795  0.01433 *
## I(X2^2)       0.007049   0.004987   1.414  0.17935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.58 on 14 degrees of freedom
## Multiple R-squared:  0.8061, Adjusted R-squared:  0.7507
## F-statistic: 14.55 on 4 and 14 DF,  p-value: 6.851e-05
```

```
regsubsets1 <- lm(Y ~ X1 + X2 + X1*X2, lung)
summary(regsubsets1)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1 * X2, data = lung)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3075  -6.6602  -0.5824   4.6284  24.0398
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 134.399866  15.981599   8.410 4.63e-07 ***
## X1           -2.133022   0.522157  -4.085 0.000975 ***
## X2           -1.699330   0.363669  -4.673 0.000300 ***
## X1:X2         0.033347   0.009283   3.592 0.002667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.58 on 15 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7507
## F-statistic: 19.06 on 3 and 15 DF,  p-value: 2.233e-05
```

```
regsubsets2<- lm(Y ~ X1 + X2 + X1*X2 + I(X2^2), lung)
summary(regsubsets2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1 * X2 + I(X2^2), data = lung)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.658  -4.802  -2.591   4.641  24.694
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 135.928530  16.422075   8.277 9.19e-07 ***
## X1           -1.867312   0.657434  -2.840  0.01310 *
## X2           -2.003727   0.577445  -3.470  0.00375 **
## I(X2^2)       0.003859   0.005618   0.687  0.50335
## X1:X2         0.029384   0.011073   2.654  0.01889 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.78 on 14 degrees of freedom
## Multiple R-squared:  0.799,  Adjusted R-squared:  0.7416
## F-statistic: 13.91 on 4 and 14 DF,  p-value: 8.741e-05
```

The three top models have $R2a, p$ values around 0.75 after testing several first-order and second-order term combinations.

b. The three best subset models' $R2a, p$ values don't differ significantly from one another. The best and worst of the three are separated by $0.01 dollars.

```
library(leaps)
lp_forder_subsets = regsubsets(formula(Y ~ scale(X1, center=TRUE, scale=FALSE) + scal
e(X2, center = TRUE, scale = FALSE) + scale(X3, center = TRUE, scale = FALSE) + I(sca
le(X1^2, center = TRUE, scale = FALSE)) + I(scale(X2^2, center = TRUE, scale = FALSE)
) + I(scale(X3^3, center = TRUE, scale = FALSE)) + scale(X1 * X2, center = TRUE, scal
e = FALSE) + scale(X2 * X3, center = TRUE, scale = FALSE) + scale(X1 * X3, center = T
RUE, scale = FALSE)), data = lung)

lp_forder_summary = summary(lp_forder_subsets, all.best = TRUE)
lp_forder_summary
```

```
## Subset selection object
## Call: regsubsets.formula(formula(Y ~ scale(X1, center = TRUE, scale = FALSE) +
```

```
##     scale(X2, center = TRUE, scale = FALSE) + scale(X3, center = TRUE,
##     scale = FALSE) + I(scale(X1^2, center = TRUE, scale = FALSE)) +
##     I(scale(X2^2, center = TRUE, scale = FALSE)) + I(scale(X3^3,
##     center = TRUE, scale = FALSE)) + scale(X1 * X2, center = TRUE,
##     scale = FALSE) + scale(X2 * X3, center = TRUE, scale = FALSE) +
##     scale(X1 * X3, center = TRUE, scale = FALSE)), data = lung)
## 9 Variables  (and intercept)
##                                              Forced in Forced out
## scale(X1, center = TRUE, scale = FALSE)          FALSE      FALSE
## scale(X2, center = TRUE, scale = FALSE)          FALSE      FALSE
## scale(X3, center = TRUE, scale = FALSE)          FALSE      FALSE
## I(scale(X1^2, center = TRUE, scale = FALSE))     FALSE      FALSE
## I(scale(X2^2, center = TRUE, scale = FALSE))     FALSE      FALSE
## I(scale(X3^3, center = TRUE, scale = FALSE))     FALSE      FALSE
## scale(X1 * X2, center = TRUE, scale = FALSE)     FALSE      FALSE
## scale(X2 * X3, center = TRUE, scale = FALSE)     FALSE      FALSE
## scale(X1 * X3, center = TRUE, scale = FALSE)     FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          scale(X1, center = TRUE, scale = FALSE)
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) "*"
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
##          scale(X2, center = TRUE, scale = FALSE)
## 1  ( 1 ) "*"
## 2  ( 1 ) "*"
## 3  ( 1 ) "*"
## 4  ( 1 ) "*"
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) "*"
##          scale(X3, center = TRUE, scale = FALSE)
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
##          I(scale(X1^2, center = TRUE, scale = FALSE))
```

```
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) "*"
## 5  ( 1 ) "*"
## 6  ( 1 ) " "
## 7  ( 1 ) " "
## 8  ( 1 ) " "
##          I(scale(X2^2, center = TRUE, scale = FALSE))
## 1  ( 1 ) " "
## 2  ( 1 ) "*"
## 3  ( 1 ) " "
## 4  ( 1 ) "*"
## 5  ( 1 ) " "
## 6  ( 1 ) " "
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
##          I(scale(X3^3, center = TRUE, scale = FALSE))
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
##          scale(X1 * X2, center = TRUE, scale = FALSE)
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) "*"
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
##          scale(X2 * X3, center = TRUE, scale = FALSE)
## 1  ( 1 ) " "
## 2  ( 1 ) " "
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) "*"
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
##          scale(X1 * X3, center = TRUE, scale = FALSE)
## 1  ( 1 ) " "
## 2  ( 1 ) " "
```

```
## 3  ( 1 ) " "
## 4  ( 1 ) " "
## 5  ( 1 ) " "
## 6  ( 1 ) "*"
## 7  ( 1 ) "*"
## 8  ( 1 ) "*"
```

```
summary_bestsubset = data.frame(lp_forder_summary$which, lp_forder_summary$adjr2, lp_
forder_summary$rsq)
colnames(summary_bestsubset) = c("Y","X1","X2","X3","X1sqr","X2sqr","X3sqr","X1X2","X
2X3", "X1X3", "Radjsqr", "Rsqr")

summary_bestsubset
```

```
##      Y    X1    X2    X3 X1sqr X2sqr X3sqr  X1X2  X2X3  X1X3   Radjsqr
## 1 TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE 0.5329124
## 2 TRUE FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE 0.6416257
## 3 TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE 0.7506631
## 4 TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE 0.7506701
## 5 TRUE  TRUE FALSE  TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE 0.7354895
## 6 TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE 0.7379080
## 7 TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE 0.7194864
## 8 TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE 0.6926137
##        Rsqr
## 1 0.5588617
## 2 0.6814450
## 3 0.7922193
## 4 0.8060768
## 5 0.8089646
## 6 0.8252720
## 7 0.8285750
## 8 0.8292298
```

```
top3_bestsubset = summary_bestsubset[order(summary_bestsubset$Radjsqr, decreasing=TRU
E),1:11]
top3_bestsubset[1:3,]
```

```
##      Y   X1    X2    X3 X1sqr X2sqr X3sqr  X1X2  X2X3  X1X3   Radjsqr
## 4 TRUE TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE 0.7506701
## 3 TRUE TRUE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE 0.7506631
## 6 TRUE TRUE FALSE  TRUE FALSE FALSE  TRUE  TRUE  TRUE  TRUE 0.7379080
```

The top 3 best subset having R-Sqr(a,p) value are: 0.7506701, 0.7506631 and 0.7379080

##Problem 3

```
cosmetic<-read.table("http://www.cnachtsheim-text.csom.umn.edu/Kutner/Chapter%2010%20
Data%20Sets/CH10PR13.txt")
colnames(cosmetic)[1] ="Y"
colnames(cosmetic)[2]="X1"
colnames(cosmetic)[3]="X2"
colnames(cosmetic)[4]="X3"
head(cosmetic)
```

```
##       Y  X1  X2  X3
## 1 12.85 5.6 5.6 3.8
## 2 11.55 4.1 4.8 4.8
## 3 12.78 3.7 3.5 3.6
## 4 11.19 4.8 4.5 5.2
## 5  9.00 3.4 3.7 2.9
## 6  9.34 6.1 5.8 3.4
```

a.

```
cosmetic_lm<-lm(Y~X1+X2+X3,data=cosmetic)
summary(cosmetic_lm)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = cosmetic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4217 -0.9115  0.0703  1.1420  3.5479
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0233     1.2029   0.851   0.4000
## X1            0.9657     0.7092   1.362   0.1809
## X2            0.6292     0.7783   0.808   0.4237
## X3            0.6760     0.3557   1.900   0.0646 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.825 on 40 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7223
## F-statistic: 38.28 on 3 and 40 DF,  p-value: 7.821e-12
```

Regression model is

$Y = 1.0233 + 0.9657X_1 + 0.6292X_2 + 0.6760X_3 + \epsilon$

```
cosmetic_lm$fitted.values
```

```
##         1         2         3         4         5         6         7         8
## 12.523331 11.247489  9.232069 12.005132  8.594978 12.861599 15.557943  9.971563
##         9        10        11        12        13        14        15        16
##  7.134733  7.893822 11.358116  7.858766  5.889261 11.179547  8.541178  2.974351
##        17        18        19        20        21        22        23        24
##  6.328273 10.691463  7.226361 10.817552  9.962447  8.217182  4.404481 12.971120
##        25        26        27        28        29        30        31        32
##  8.623391  9.705286 11.580732  9.295027 12.812187  8.381651  5.989922  5.777189
##        33        34        35        36        37        38        39        40
## 11.658598  3.133836  9.459454 10.213518 16.930693  7.134775  6.559452  6.221696
##        41        42        43        44
##  8.324271  9.802153  9.014906 13.198505
```

b.

```
anova(cosmetic_lm)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq  F value    Pr(>F)
## X1         1 365.56  365.56 109.7054 4.994e-13 ***
## X2         1   5.07    5.07   1.5215   0.22459
## X3         1  12.03   12.03   3.6113   0.06461 .
## Residuals 40 133.29    3.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(Y ~ X3, data= cosmetic),cosmetic_lm)
```

```
## Analysis of Variance Table
##
## Model 1: Y ~ X3
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     42 400.00
## 2     40 133.29  2    266.71 40.021 2.848e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
F_cosmetic = qf(1 - 0.05,2,40)
F_cosmetic
```

```
## [1] 3.231727
```

$H_0(null\ Hypothesis) : \beta_1 = \beta_2 = \beta_3 = 0$
$H_a(Alternate\ Hypothesis) :\ not\ all\ \beta_k = 0(where\ k = 1, 2, 3)$ Decision :
If $F* \leq F(1-\alpha; 2, 40)$, then conclude H_0 If $F* > F(1-\alpha; 2, 40)$, then conclude H_a

Conclusion : Here, $F* = 40.021 > F(0.95; 2, 40)$ = 3.231727, therefore, conclude H_a i.e.not all $\beta_k \neq 0(k = 1, 2, 3)$

---

c.

```
SSE_X1X2X3 = sum(cosmetic_lm$residual^2)
SSR_X1X2 <- sum(lm(Y ~X1 + X2, data=cosmetic)$residual^2) - SSE_X1X2X3
SSR_X1X2
```

```
## [1] 12.03326
```

```
SSR_X2X3 <- sum(lm(Y ~X2 + X3, data=cosmetic)$residual^2) - SSE_X1X2X3
SSR_X2X3
```

```
## [1] 6.177849
```

```
SSR_X3X1 <- sum(lm(Y ~X3 + X1, data=cosmetic)$residual^2) - SSE_X1X2X3
SSR_X3X1
```

```
## [1] 2.177503
```

```
F1 <- SSR_X2X3/(SSE_X1X2X3/cosmetic_lm$df.residual)
F2 <- SSR_X3X1/(SSE_X1X2X3/cosmetic_lm$df.residual)
F3 <- SSR_X1X2/(SSE_X1X2X3/cosmetic_lm$df.residual)
cat("F1 : ",F1)
```

```
## F1 :  1.854008
```

```
cat("\nF2 : ",F2)
```

```
##
## F2 :  0.6534814
```

```
cat("\nF3 : ",F3)
```

```
##
## F3 :  3.611251
```

$H_0(null\ Hypothesis) : \beta_k = 0\ H_a(Alternate\ Hypothesis) :\ \beta_k \neq 0(where\ k = 1, 2, 3)$ Test :

$$F^* = \frac{\frac{SSR(X_k|X_j, forj=1,2,3, j\neq k)}{1}}{\frac{SSE(X_1,X_2,X_3)}{n-p}} = \frac{MSR(X_k|X_j, forj=1,2,3, j\neq k)}{MSE(X_1,X_2,X_3)}$$

If $F* \leq F(1-\alpha;1,40)$, then conclude H_0; If $F* > F(1-\alpha;1,40)$, then conclude H_a;

Conclusion :

Here, $F^* = \begin{cases} 1.854008, \ k=1 \\ 0.6534814, \ k=2 \\ 3.611251,, \ k=3 \end{cases} > F(0.95;1,40) = 0.2513963$

conclude H_a i.e.not all $\beta_k \neq 0 (k = 1,2,3)$

```
cat("",df(0.95,1,cosmetic_lm$df.residual) )
```

```
##  0.2513963
```

---

d.

```
cor(cbind(cosmetic$X1,cosmetic$X2, cosmetic$X3) )
```

```
##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.9744313 0.3759509
## [2,] 0.9744313 1.0000000 0.4099208
## [3,] 0.3759509 0.4099208 1.0000000
```

---

e. By $b1$, we can estimate. Given that $X1$ and $X2$ might be linear, as shown in (d), $X1$ is practically fixed when $X2$ is fixed ($b10$), and the sales expectation when $X1$ is increased by 1000 while $X2$ and $X3$ are maintained constant ($beta1$). and the data can be in conflict. Therefore, the data might not be appropriate for the research goal.

##Problem 4

a.

```
lung_X1X2 <- lm(Y ~ X1 + X2 + X1 * X2, data=lung)
summary(lung_X1X2)
```

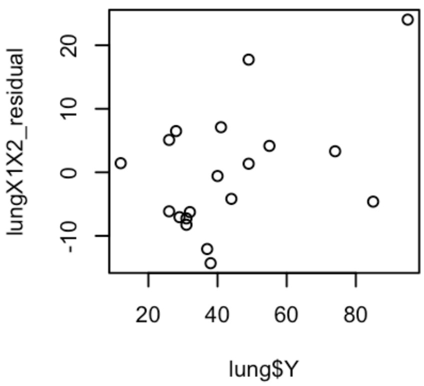
```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X1 * X2, data = lung)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3075  -6.6602  -0.5824   4.6284  24.0398
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 134.399866  15.981599   8.410 4.63e-07 ***
## X1           -2.133022   0.522157  -4.085 0.000975 ***
## X2           -1.699330   0.363669  -4.673 0.000300 ***
## X1:X2         0.033347   0.009283   3.592 0.002667 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.58 on 15 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7507
## F-statistic: 19.06 on 3 and 15 DF,  p-value: 2.233e-05
```

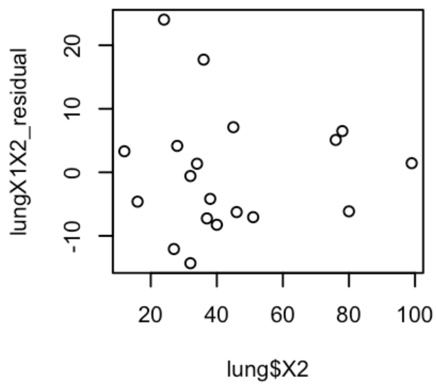
$\hat{Y} = 134.399866 - 2.133022X_1 - 1.699330X_2 + 0.033347X_1 6X_2$

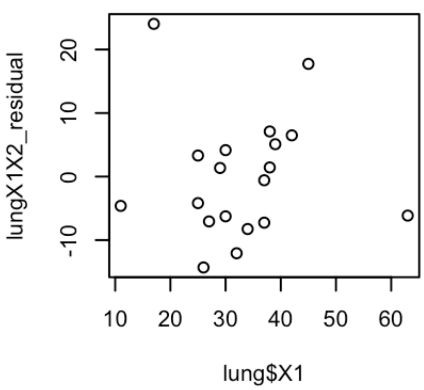
```
lungX1X2_residual <- lung_X1X2$residuals
lungX1X2_residual
```

```
##           1           2           3           4           5           6
##  17.7397360   4.1604873  -4.6164306  -6.2589963   5.0963276   6.4897650
##           7           8           9          10          11          12
##  24.0398190  -6.1423593   3.3135205 -12.0731285  -7.2549936   1.3547714
##          13          14          15          16          17          18
## -14.3075045   7.1013141   1.4369254  -4.1795022  -7.0613714  -0.5824338
##          19
##  -8.2559462
```

```
par(mfcol=c(2,3))
plot(lung$Y, lungX1X2_residual)
plot(lung$X1, lungX1X2_residual)
plot(lung$X2, lungX1X2_residual)
```

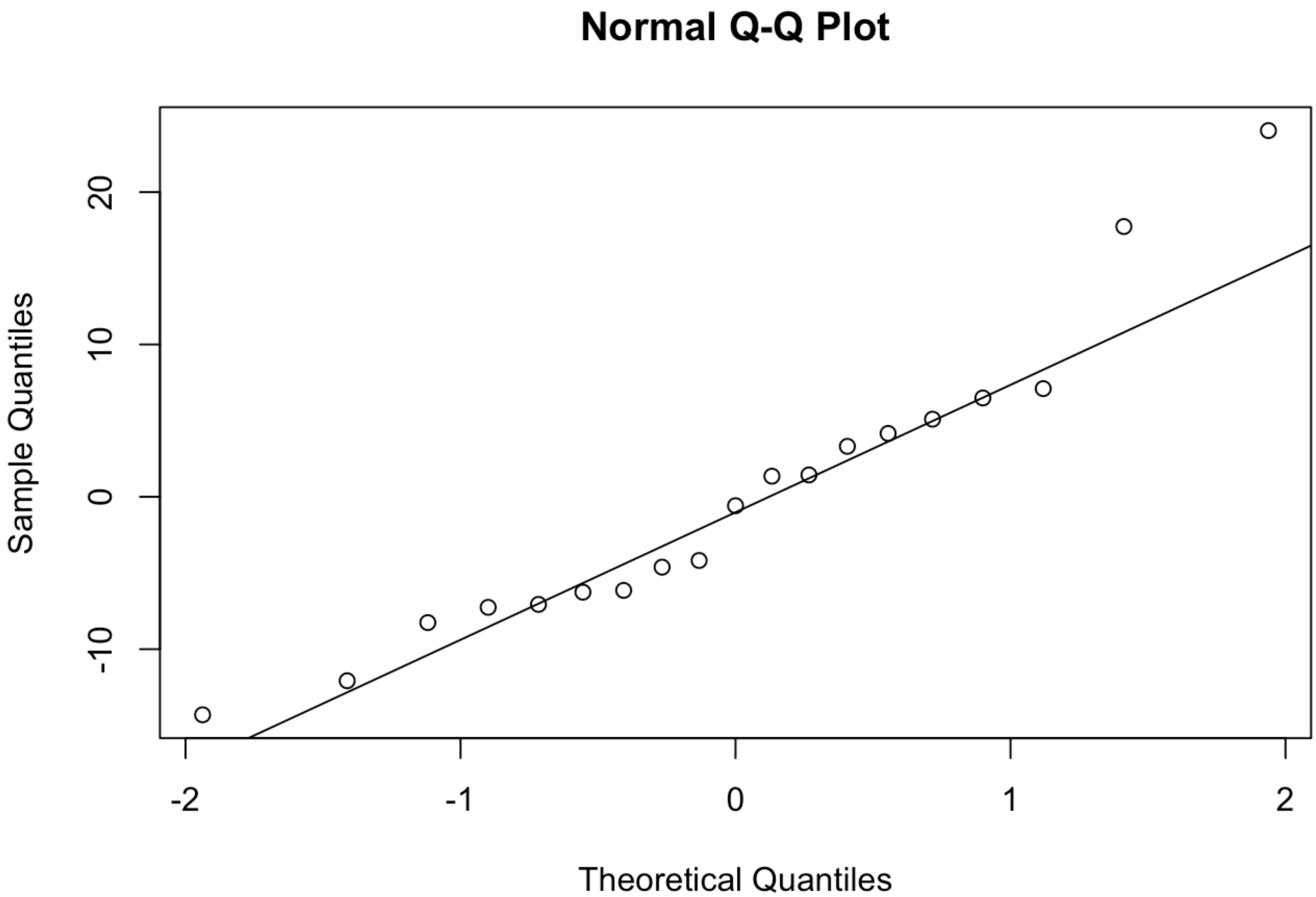






b.

```
qqnorm(lungX1X2_residual)
qqline(lungX1X2_residual)
```




```
Sd <- summary(lung_X1X2)$sigma
Sd
```

```
## [1] 10.58447
```

```
dim(lung)
```

```
## [1] 19  4
```

```
n <- 19
ExpVals <- sapply(1:n, function(k) Sd * qnorm((k-.375)/(n+.25)))
ExpVals
```

```
##  [1] -19.535807 -14.563904 -11.609083  -9.358099  -7.466997  -5.789137
##  [7]  -4.245931  -2.788423  -1.382169   0.000000   1.382169   2.788423
## [13]   4.245931   5.789137   7.466997   9.358099  11.609083  14.563904
## [19]  19.535807
```

```
cor(ExpVals, sort(lungX1X2_residual))
```

```
## [1] 0.9633751
```

It appears fair that there is a correlation of 0.9633751 between the ordered residuals and expected values under normality.

---

c.

```
library(car)
```

```
## Loading required package: carData
```

```
vif(lung_X1X2)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##        X1        X2     X1:X2
##  5.431477 11.639560 22.474469
```

VIF(X1) = 5.431477 , VIF(X2) = 11.639560, VIF(X1X2) = 22.474469

All of the VIF values for the predictors are clearly larger than 5, which is potentially worrying. and yes, it is proof that significant multicollinearity exists.

---

d.

```
del_res_lung <- round(rstudent(lung_X1X2),3)
del_res_lung
```

```
##      1      2      3      4      5      6      7      8      9     10     11
##  2.209  0.399 -0.629 -0.605  0.517  0.662  3.314 -1.779  0.338 -1.223 -0.715
##     12     13     14     15     16     17     18     19
##  0.128 -1.457  0.692  0.182 -0.402 -0.709 -0.057 -0.802
```

```
n = 19
p = 3
ifelse(del_res_lung > qt(0.9987,14), "outlier", "no outlier")
```

```
##            1            2            3            4            5            6
## "no outlier" "no outlier" "no outlier" "no outlier" "no outlier" "no outlier"
##            7            8            9           10           11           12
## "no outlier" "no outlier" "no outlier" "no outlier" "no outlier" "no outlier"
##           13           14           15           16           17           18
## "no outlier" "no outlier" "no outlier" "no outlier" "no outlier" "no outlier"
##           19
## "no outlier"
```

t(0.9987;14) = 3.65 If $|t_i| <= 3.65$, Conclude no outliers, otherwise outliers, Conclusion: It appears that all observe values cannot be defineted as outliers by Bonferroni outlier test.

---

e.

```
hatmatrix <- round(lm.influence(lung_X1X2)$hat,3)
hatmatrix
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## 0.276 0.083 0.539 0.085 0.176 0.174 0.218 0.878 0.193 0.102 0.112 0.068 0.075
##    14    15    16    17    18    19
## 0.093 0.480 0.090 0.144 0.139 0.077
```

```
ifelse(hatmatrix> 2*4/19, "outlier", "no outlier")
```

```
##            1            2            3            4            5            6
## "no outlier" "no outlier"    "outlier" "no outlier" "no outlier" "no outlier"
##            7            8            9           10           11           12
## "no outlier"    "outlier" "no outlier" "no outlier" "no outlier" "no outlier"
##           13           14           15           16           17           18
## "no outlier" "no outlier"    "outlier" "no outlier" "no outlier" "no outlier"
##           19
## "no outlier"
```

In cases 3, 8, and 15, the diagonal elements of the hat matrix are greater than double the mean leverage value. They are considered as outliers by rule of thumb

---

f.

```
Dfits_DBeta <- cbind(
  "DFFITS"  <- round(dffits(lung_X1X2), 4),
  "DFBETA0" <- round(dfbetas( lung_X1X2)[,1], 4),
  "DFBETA3" <- round(dfbetas( lung_X1X2)[,2], 4),
  "DFBETA1" <- round(dfbetas( lung_X1X2)[,3], 4),
  "DFBETA4" <- round(dfbetas( lung_X1X2)[,3], 4),
  "Cook's D" <- round(cooks.distance( lung_X1X2), 4))
Dfits_DBeta[c(3,7,8,15),]
```

```
##       [,1]    [,2]    [,3]    [,4]    [,5]   [,6]
## 3  -0.6802 -0.6519  0.5919  0.4334  0.4334 0.1205
## 7   1.7486  1.4541 -1.2776 -0.7415 -0.7415 0.4589
## 8  -4.7798 -1.5469  1.1866  3.1623  3.1623 4.9908
## 15  0.1749 -0.0155 -0.0353  0.0771  0.0771 0.0082
```

DFBETAS scores for cases 3, 8, and 15 are all significantly below 1, indicating non-influential.

DFFITS readings for cases 3, 8, and 15 have absolute values of -0.6802, -4.7798, and 0.1749, respectively, which all exceed the cut-off value of 0.8.

Therefore we determine that none of the outlier X observations are significant by examining the Cook's distance.