Homework 4

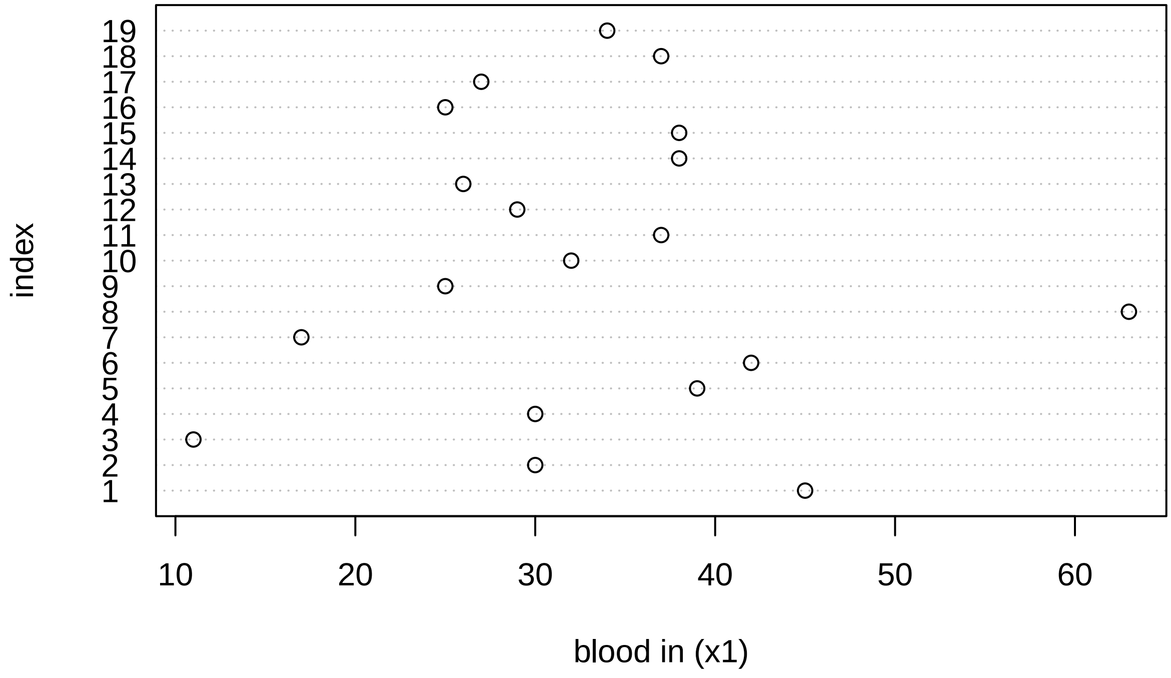
Sohaib Syed

2022-10-18

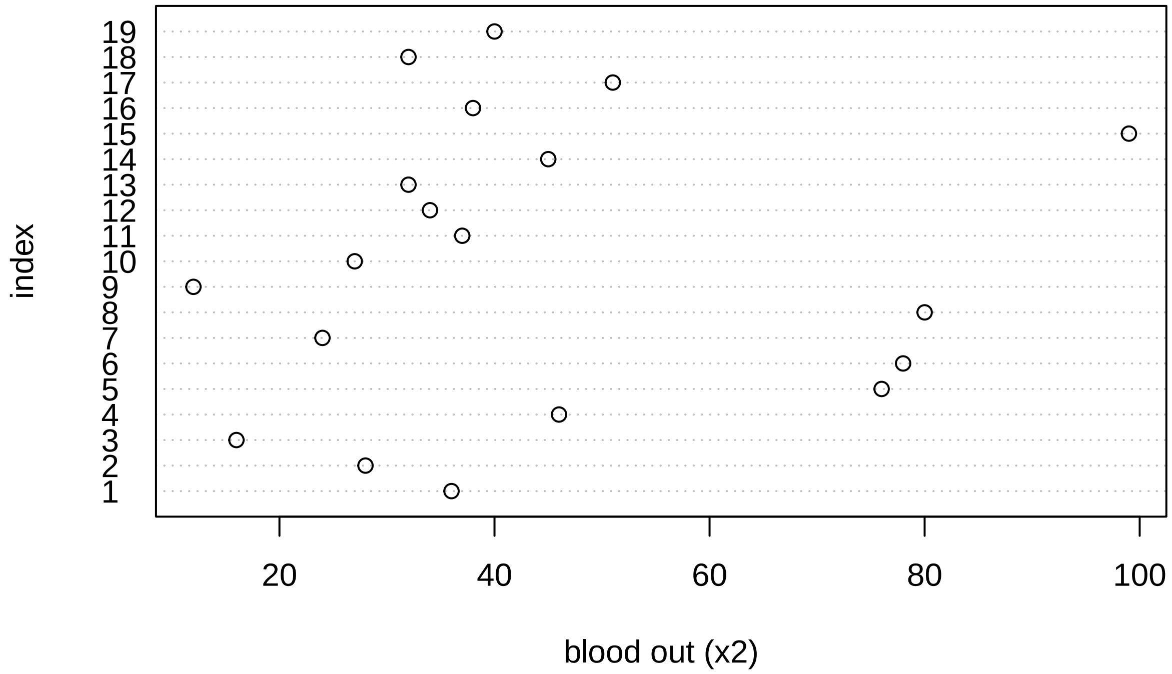
Problem 1

a

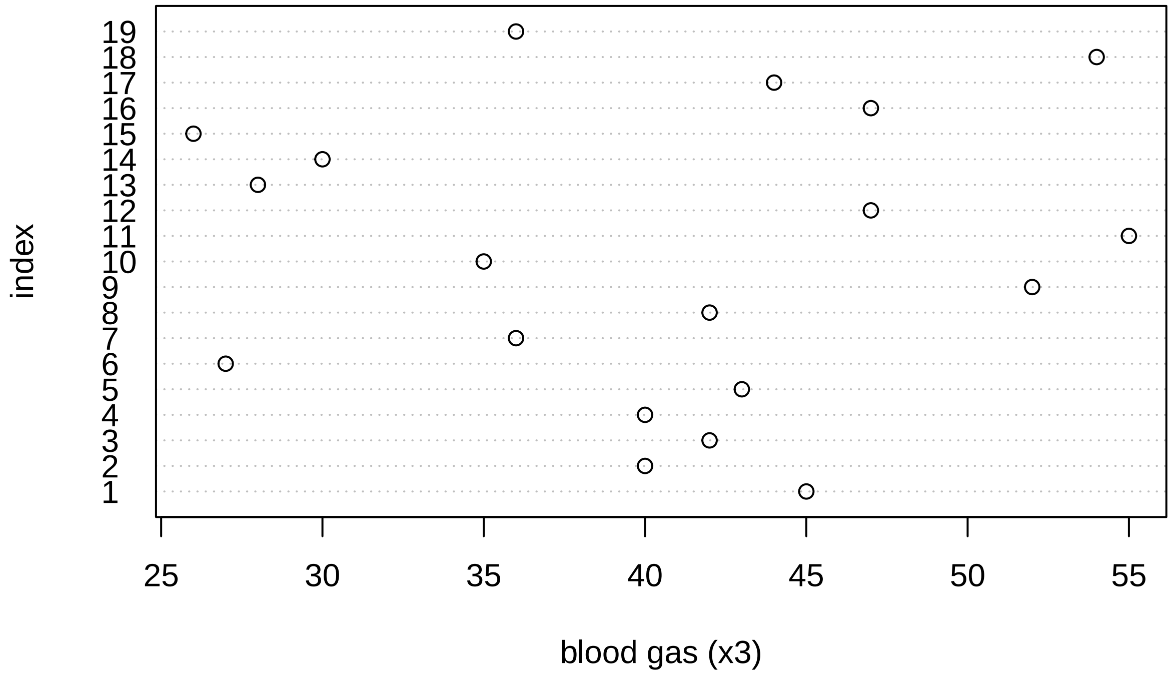
```
lung_data<-read.table('CH09PR13.txt',col.names = c('y','x1','x2','x3'))
dotchart(lung_data$x1, labels=1:nrow(lung_data),xlab = 'blood in (x1)',ylab = 'index')
```



```
dotchart(lung_data$x2,labels=1:nrow(lung_data),xlab = 'blood out (x2)',ylab = 'index')
```



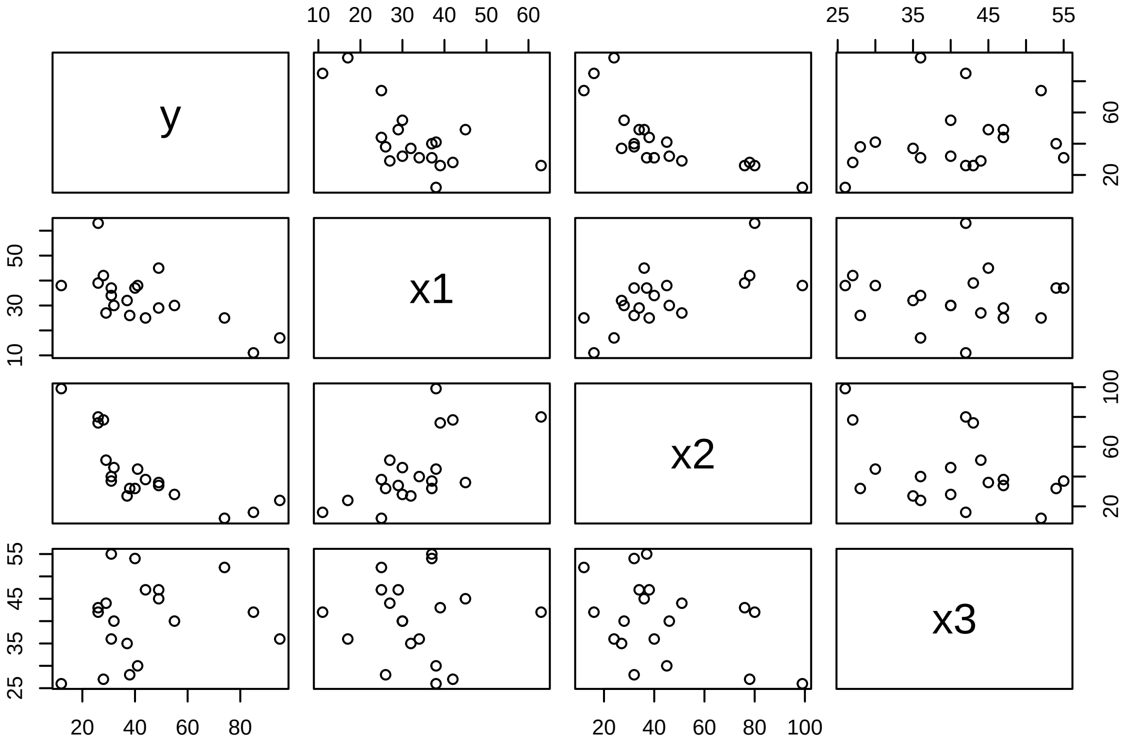
```
dotchart(lung_data$x3,labels=1:nrow(lung_data),xlab = 'blood gas (x3)',ylab = 'index')
```



its noticeable that for the dot charts for x1 and x2 show most data around the values of 25-40 on the x-axis. That being said, points 3,7 and 8 are outliers for X1, and points 5,6,8,15 are outliers for 15

b

```
pairs(~y+.,data=lung_data)
```



```
cor(lung_data)
```

```
##              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
```

the scatter plots show that x1 and x2 each have a negative correlation with y. x3 does not have correlation with y. x1 and x2 seem to have positive correlation, while x3 is not correlated to x1 or x2. there might be multi-collinnearity between x1 and x2.

c

```
lungfit1<-lm(y~.,data=lung_data)
summary(lungfit1)
```

```
##
## Call:
## lm(formula = y ~ ., data = lung_data)
##
## 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
```

Y=87.1875-56448x1-.51315x2-.07196x3

No, not all predictors should be retained.

Problem 2

a

```r
lung_data['x1x2']<-data.frame(c(lung_data$x1*lung_data$x2))
lung_data['x1x3']<-data.frame(c(lung_data$x1*lung_data$x3))
lung_data['x2x3']<-data.frame(c(lung_data$x2*lung_data$x3))
lung_data['x1^2']<-data.frame(c(lung_data$x1*lung_data$x1))
lung_data['x2^2']<-data.frame(c(lung_data$x2*lung_data$x2))
lung_data['x3^2']<-data.frame(c(lung_data$x3*lung_data$x3))

best9<-regsubsets(x=lung_data[,2:10], y=lung_data$y,nbest=9,
nvmax=9,method="exhaustive")
report9<-summary(best9)
order(report9$adjr2, decreasing=TRUE)[1:3] # best models at output matrix row 28 19 and 29
```

```
## [1] 28 19 29
```

```r
report9$adjr2[order(report9$adjr2, decreasing=TRUE)[1:3]]
```

```
## [1] 0.7506701 0.7506631 0.7485086
```

The best model is x1,x2,x1^{2,x2}2 The second best is x1,x2,x1x2 The third best is x1,x3,x2x3,x1ˆ2

b

There is not much difference between the three. However, compared to the first and second bect model, model 3 is lower than the others.

Problem 3

a

The regression model to be used should be Y_hat=B0+B1x1+B2x2+B3X3+e

```
cosmetic_data<-read.table('CH10PR13.txt',col.names = c('y','x1','x2','x3'))
cosmeticfit<-lm(y~.,data=cosmetic_data)
summary(cosmeticfit)
```

```
##
## Call:
## lm(formula = y ~ ., data = cosmetic_data)
##
## 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
```

Y=1.0233+.9657x1+.6292x2+.6760x3

b

Hypothesis

H0: b1=b2=b3=0 Ha: bk != 0 where k=(1,2,3)

Decision Rule

F-Ratio > F alpha,p-q,n-p then reject H0

```
cosmestic_anova<-anova(cosmeticfit)
cosmestic_anova
```

```
## 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
```

```
SSR_over_pminusq<-mean(sum(cosmestic_anova[1:3,'Sum Sq']))/3
MSE<-cosmestic_anova[4,'Mean Sq']
SSR_over_pminusq/MSE
```

```
## [1] 38.27938
```

```
qf(.95,3,40)
```

```
## [1] 2.838745
```

```
SSR_over_pminusq/MSE>qf(1-.05,3,40)
```

```
## [1] TRUE
```

Conclusion

Reject H0

c

Hypothesis

H0: bk=0 Ha: bk != 0 k=(1,2,3)

Decision Rule

T|obs| > t alpha/2,n-p then reject H0

```
# getting values for coefficient and std.error from summary of fitted model part a
b1_tobs<-.9657/.7092
b2_tobs<-0.6292/0.7783
b3_tobs<-0.6760/0.3557
qt(1-.025,40)
```

```
## [1] 2.021075
```

```
b1_tobs>qt(1-.05/2,40)
```

```
## [1] FALSE
```

```
b2_tobs>qt(1-.05/2,40)
```

```
## [1] FALSE
```

```
b3_tobs>qt(1-.05/2,40)
```

```
## [1] FALSE
```

Conclusion

Accept H0 for each b1=0 b2=0 and b3=0

No; The conclusions for this individual tests don't correspond to part b

d

```
cor(cosmetic_data)
```

```
##            y        x1        x2        x3
## y  1.0000000 0.8417342 0.8424849 0.4740581
## x1 0.8417342 1.0000000 0.9744313 0.3759509
## x2 0.8424849 0.9744313 1.0000000 0.4099208
## x3 0.4740581 0.3759509 0.4099208 1.0000000
```

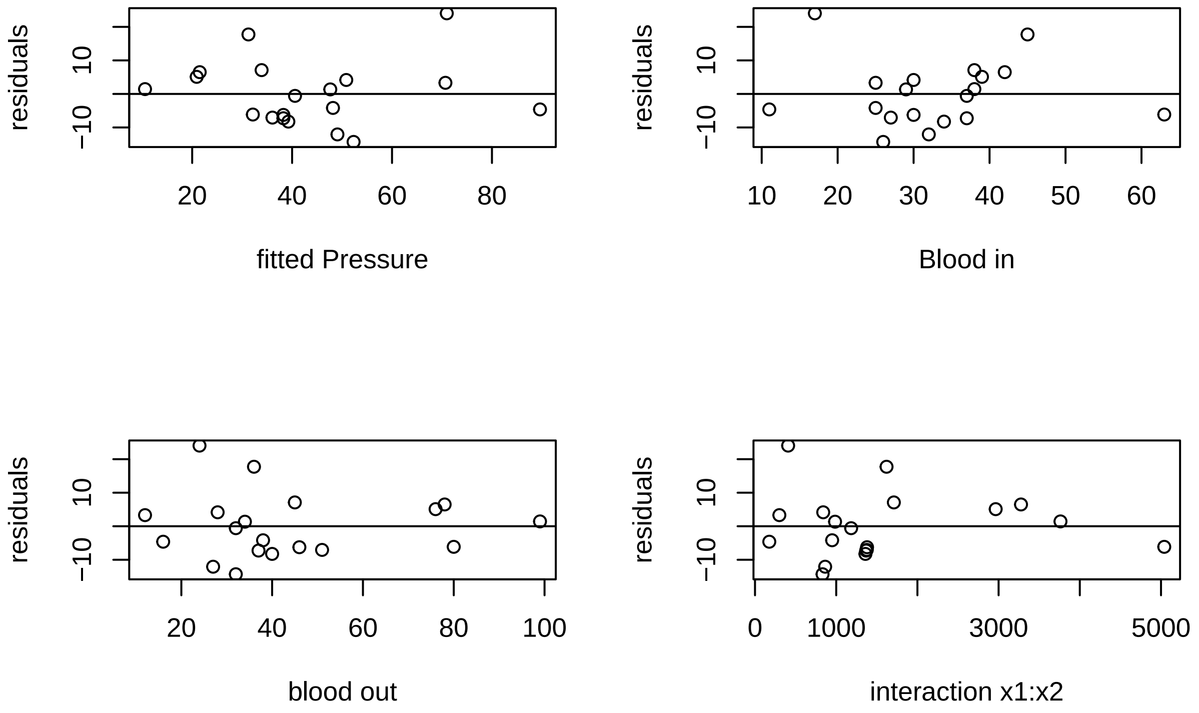
e

Parts b,c,d show that the data is not suitable because there is strong multi-collinearity in the data. The tests were contradicting and the correlation matrix shows strong correlation between predictors.

Problem 4

a

```
lung_subset_model<-lm(y~x1+x2+(x1*x2),data=lung_data)
par(mfrow=c(2,2))
plot(lung_subset_model$fitted.values,lung_subset_model$residuals,xlab='fitted Pressure',ylab='residuals
abline(0,0)
plot(lung_data$x1,lung_subset_model$residuals,xlab='Blood in',ylab='residuals')
abline(0,0)
plot(lung_data$x2,lung_subset_model$residuals,xlab='blood out',ylab='residuals')
abline(0,0)
plot(lung_data$x1*lung_data$x2,lung_subset_model$residuals,xlab='interaction x1:x2',ylab='residuals')
abline(0,0)
```



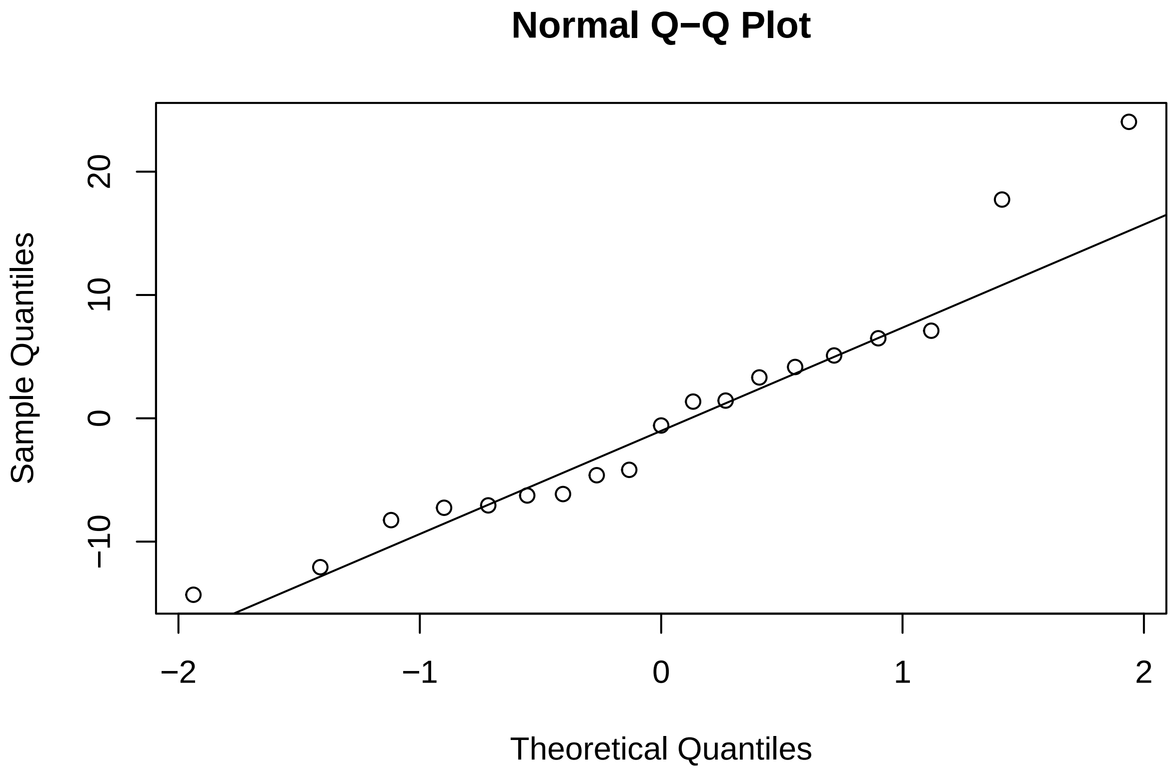
```
anova(lung_subset_model)
```

```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## x1         1 3577.1  3577.1  31.929 4.611e-05 ***
## x2         1 1384.4  1384.4  12.357  0.003124 **
## x1:x2      1 1445.8  1445.8  12.905  0.002667 **
## Residuals 15 1680.5   112.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The spread of the interaction term across the x-axis doesn't seem great. Also the variance of error for 'blood out' residuals does not seem constant as the vertical spread decreases as values increase on x-axis. There is a linear formation in the interaction term, so there may be need to get rid of one of the predictors since interaction may be present.

b

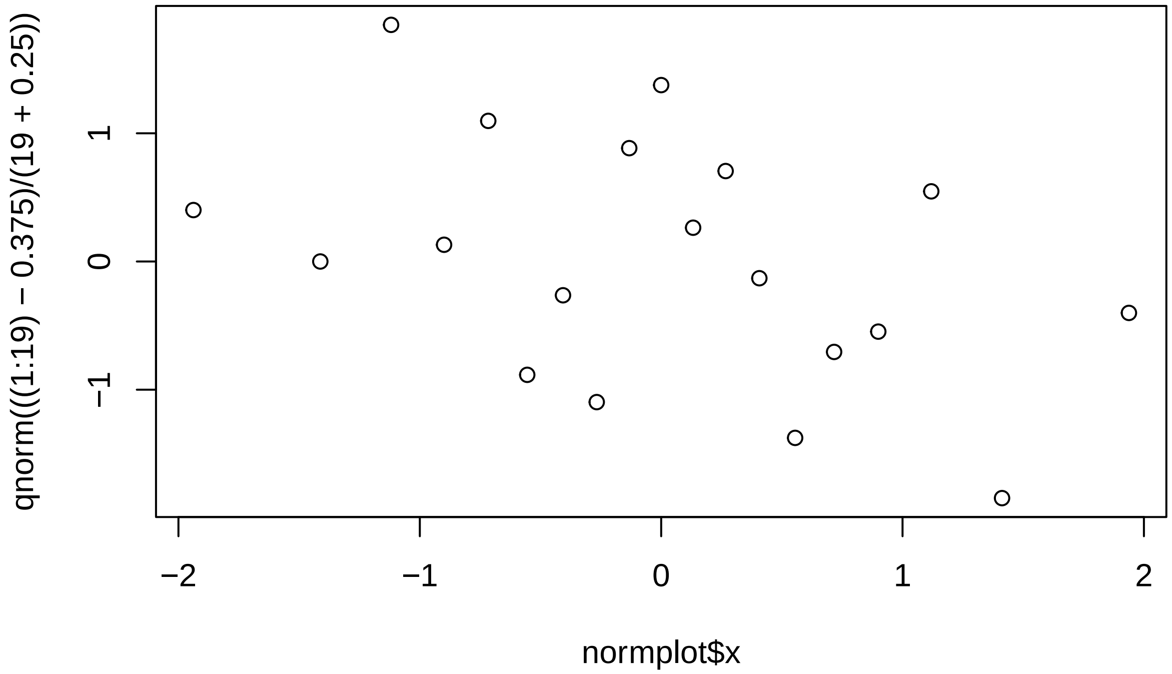
```
normplot<-qqnorm(resid(lung_subset_model))
qqline(resid(lung_subset_model))
```



```
cor(normplot$y,normplot$x)
```

```
## [1] 0.9642988
```

```
plot(normplot$x,qnorm(((1:19)-.375)/(19+.25)))
```



the normal assumption does not look reasonable as the ei vs zi plot is not a straight line.

c

```
library(car)
```

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

```
vif(lung_subset_model)
```

```
## 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
```

The maxVIF=22.474469 which is > 10 so there is strong evidence of multicollinearity here.

d

```
library(MASS)

studres(lung_subset_model)>qt(1-.05/(2*(length(studres(lung_subset_model)))),length(studres(lung_subset_
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##    14    15    16    17    18    19
## FALSE FALSE FALSE FALSE FALSE FALSE
```

decision rule:

If studentized residual is greater then t alpha/2*length(studentized residuals),n-p-1 then there are outliers.

conclusion

None of the studentized residuals are greater than t distribution so no outliers

e

```
influence(lung_subset_model)$hat > (2*(4/19))
```

```
##     1     2     3     4     5     6     7     8     9    10    11    12    13
## FALSE FALSE  TRUE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
##    14    15    16    17    18    19
## FALSE  TRUE FALSE FALSE FALSE FALSE
```

At i of 3, 8 and 15, the rule of thumb suggests that x1 x2 are outliers at those idices. Comparing to dot charts, I do see that for x1 3 and 8 are outliers, and then for x2 at 3,8 and 15 are all outliers. They should be the same because the diagonal of hat matrix measures the distance of that point to the centroid, and if the values are too far from centroid they should be also apparent in the dot chart.

f

```
M <- model.matrix(lung_subset_model)
H <- M%*%solve(t(M)%*%M)%*%t(M)
leverage<-diag(H)
t_value<-lung_subset_model$residuals*sqrt((19-4-1)/
(sum(lung_subset_model$residuals^2)*(1-leverage)-lung_subset_model$residuals^2))

DFFITS <- t_value*sqrt(leverage/(1-leverage))
DFFITS[c(3,7,8,15)]
```

```
##          3          7          8         15
## -0.6801824  1.7485509 -4.7797848  0.1748573
```

```
dfbetas(lung_subset_model)[c(3,7,8,15),]
```

```
##    (Intercept)          x1          x2       x1:x2
## 3   -0.6519371  0.59191342  0.43337176 -0.48191103
## 7    1.4541305 -1.27760852 -0.74151968  0.84752328
## 8   -1.5469080  1.18662253  3.16226530 -3.28579003
## 15  -0.0155059 -0.03525106  0.07714703 -0.01569977
```

```
CookD <-lung_subset_model$residuals^2/(4*sum(lung_subset_model$residuals^2)/lung_subset_model$df)*(leve
CookD[c(3,7,8,15)]
```

```
##           3           7           8          15
## 0.120515509 0.458917058 4.990814979 0.008170411
```

Using all three tests to conclude that points 7 and 8 are highly influential to the model. The DFFITS and DFBETAs are > 1 and the Cooks distance is also high percentages.