

Solution Paper 2

2022-04-22

##Q1A

#(a)

k = 4

p = 0.3

set.seed(08)

x = rnbinom(100,k,p) + k

IQR(x)

[1] 8

#(b)

(qnbinom(0.75,k,p)+k) - (qnbinom(0.25,k,p)+k)

[1] 7

Both the answers differ only by 1 so the answers are quite near too each other.

#(c)

If the number of simulate values increase , the sample moments / IQR approximate to the population moments/ IQR

##Q1B

#(i)

M = matrix(0,2,2)

M[,1] = c(29,31)

M[,2] = c(200-29,180-31)

M

[,1] [,2]

[1,] 29 171

[2,] 31 149

fisher.test(M, conf.level = 0.99,alternative = "less")

##

Fisher's Exact Test for Count Data

##

data: M

p-value = 0.2788

alternative hypothesis: true odds ratio is less than 1

99 percent confidence interval:

0.000000 1.631599

sample estimates:

odds ratio

0.8155726

#NOTE TO CHECKER : Alternative can be "greater" if the order of the rows in matrix is different.

#(ii)

#Since the p - value is greater than 5%, we do not have sufficient evidence to reject H_0 , i.e. Marketing campaign was not effective.

##Q1C

#(i)

```
model0 = lm(circumference~1, data = Orange)
```

#(ii)

```
model1 = update(model0, .~.+age + I(age^2))
```

#(iii)

R^2 always increases when the number of explanatory variables increase in the model , even if they are not significant. So , we should use R^2 to compare models with different number of explanatory variables

#We will use Adj R^2 to compare both models

```
summary(model0)$adj.r
```

```
## [1] 0
```

```
summary(model1)$adj.r
```

```
## [1] 0.8269442
```

#Since Adj R^2 increases, model1 is better than model0

##Q2A

```
storm = c(9, 6, 6, 3, 2, 5, 4, 6, 2, 1, 4, 6, 3, 7, 8, 10, 9, 7)
```

#(i)

```
poisson.test(sum(storm),length(storm), conf.level = 0.99)$conf.int
```

```
## [1] 4.132187 7.027941
```

```
## attr(,"conf.level")
```

```
## [1] 0.99
```

#(ii)

```
mean(storm) +
```

```
  c(-1,1)*qnorm(0.005,lower = F)*sqrt(mean(storm)/length(storm))
```

```
## [1] 4.027811 6.861078
```

#Normal is a symettric distribution where as Poisson is skewed

#As a result, the interval , in part (i) is wider than the approximate distribution in (ii)

##Q2B

#(a)

```
model0 = glm(weight~Time + Diet,
              data = ChickWeight,
              family = Gamma("log"))
```

```

#(b)
modell1 = update(model0, .~.+Time:Diet)
#(c)
anova(model0,model1, test = "Chi")

## Analysis of Deviance Table
##
## Model 1: weight ~ Time + Diet
## Model 2: weight ~ Time + Diet + Time:Diet
##   Resid. Df Resid. Dev Df Deviance   Pr(>Chi)
## 1         573      28.338
## 2         570      27.024   3    1.3137 2.327e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

#As the p -value is less than 5%, we can conclude that interaction term is significant.

##Q2C
claims = 0:8
freqs = c(28,126,0,111,31,30,0,0,39)
#(i)
lambdahat = sum(claims*freqs)/sum(freqs)
lambdahat

## [1] 2.863014

#(ii)
exp.prob = c(dpois(0:7,lambdahat),ppois(7,lambdahat,lower = F))
chisq.test(freqs,p = exp.prob)$expected

## Warning in chisq.test(freqs, p = exp.prob): Chi-squared approximation may
## be
## incorrect

## [1] 20.840197 59.665769 85.411956 81.511867 58.342398 33.407017 15.940791
## [8]  6.519815  3.360190

#Combining last two values
exp.prob2 = c(dpois(0:6,lambdahat), ppois(6,lambdahat,lower = F))
freqs = c(28,126,0,111,31,30,0,39)
test = chisq.test(freqs,
                  p = exp.prob2)
#Since we estimated Lambda.
dof = test$parameter - 1
pchisq(test$statistic,dof,lower = FALSE)

##      X-squared
## 4.476701e-59

#Since p - value is almost 0 , we have enough evidence to reject H0.
#The claims are not from a Poi(Lambda)

```

```

#Q3A
policy_data = read.csv("PolicyData.csv")
attach(policy_data)
##(a)
table(Claim)

## Claim
##   0   1   2   3
## 458 149  36   7

nrow(policy_data) - table(Claim)[1]

##   0
## 192

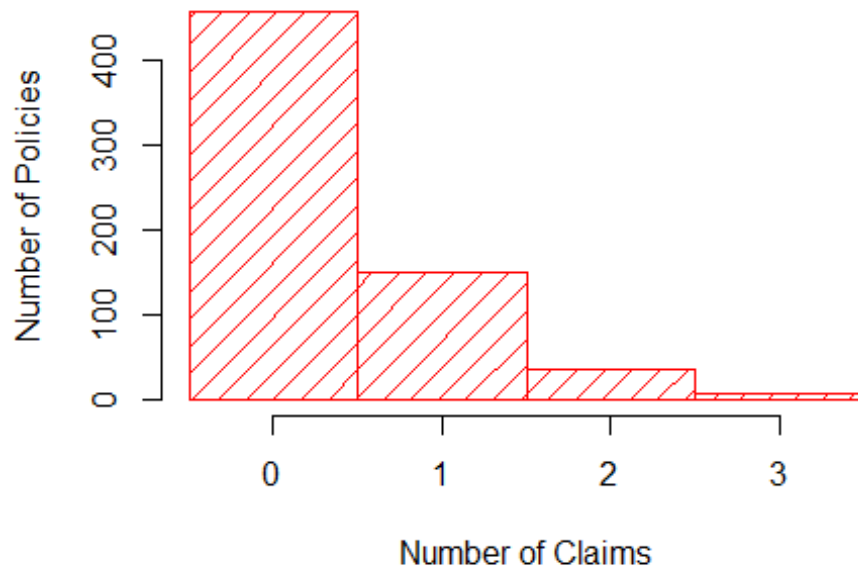
#OR
sum(table(Claim)[-1])

## [1] 192

##(b)
hist(Claim, breaks = -0.5:3.5,
     main = "Histogram of Claims",
     xlab = "Number of Claims",
     ylab = "Number of Policies",
     density = 10,
     col = "red")

```

Histogram of Claims



#Poisson or Negative Binomial could provide a good fit

##(c)

```
poisson.test(sum(Claim),length(Claim),  
             conf.level = 0.95)
```

##

Exact Poisson test

##

data: sum(Claim) time base: length(Claim)

number of events = 242, time base = 650, p-value < 2.2e-16

alternative hypothesis: true event rate is not equal to 1

95 percent confidence interval:

0.3268739 0.4222903

sample estimates:

event rate

0.3723077

#Since the 95% Confidence interval does not contain 0.30 , we can conclude that 0.35 is more appropriate value to model the number of claims.

##(d)

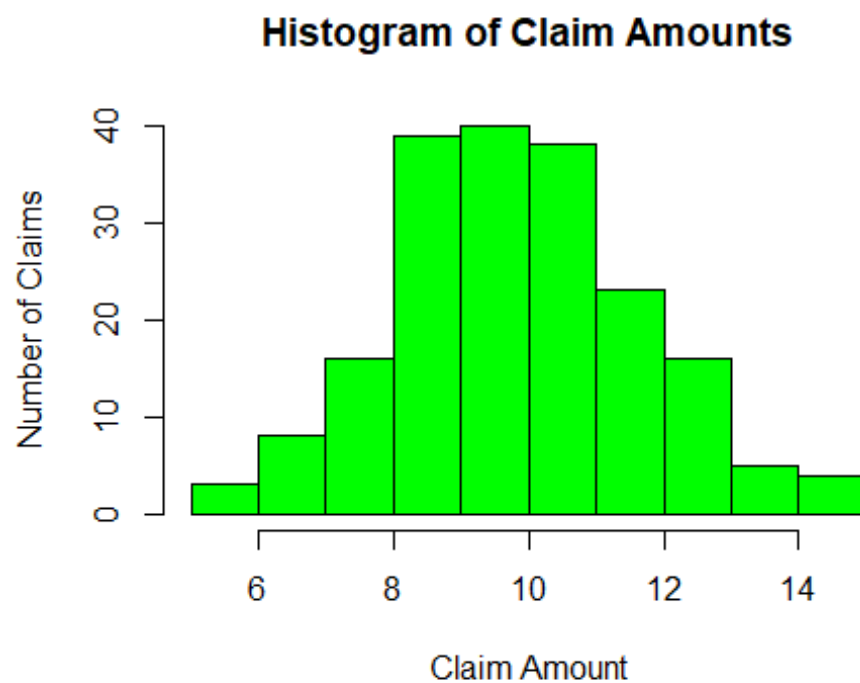
```
log_claim = log(Amount[Claim !=0])
```

```
mean(log_claim)
```

```
## [1] 9.835205
```

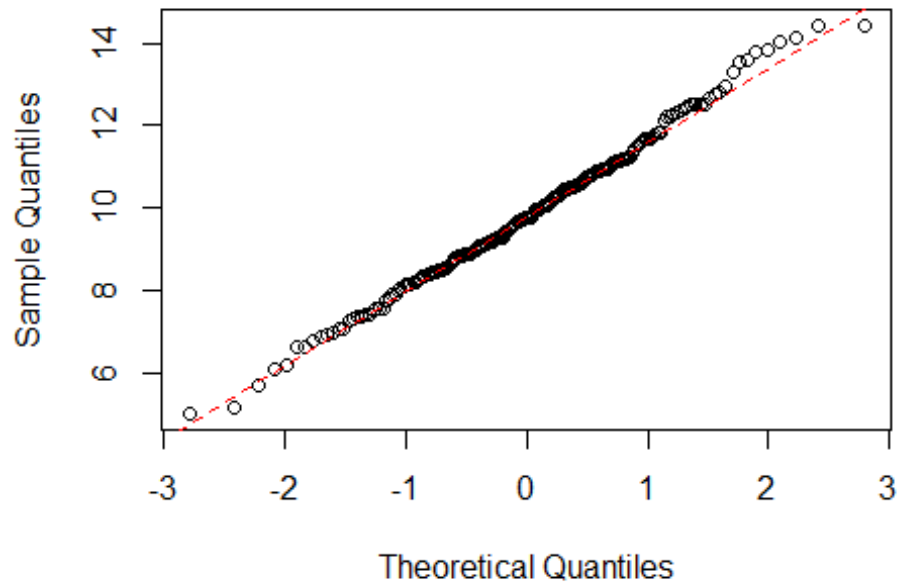
```
var(log_claim)
```

```
## [1] 3.425705  
median(log_claim)  
## [1] 9.774659  
##(e)  
hist(log_claim,  
      col = "green",  
      main = "Histogram of Claim Amounts",  
      xlab = "Claim Amount",  
      ylab = "Number of Claims")
```



```
qqnorm(log_claim)  
qqline(log_claim, col = "red", lty = 2)
```

Normal Q-Q Plot



##(f)

#The Log Claim Amount seems to follow a Normal Distribution as the shape of the Histogram is symmetric and most points on the QQ - Plot Lie on the QQ Line.

#Additionally, even the mean and median are approximately equal which is true for the Normal Distribution

##(g)

#H0 : $\mu = 10$

##vs

#H1 : $\mu > 10$

```
t.test(log_claim,
      mu = 10,
      alternative = "greater",
      conf.level = 0.9)
```

##

One Sample t-test

##

data: log_claim

t = -1.2337, df = 191, p-value = 0.8906

alternative hypothesis: true mean is greater than 10

90 percent confidence interval:

9.663428 Inf

sample estimates:

mean of x

9.835205

#Since , the p - value is greater than 10%, we do not have enough evidence to reject the H_0 , so we will conclude that $\mu = 10$

##(h)

```
M = table(Cust_Exp,Claim)
```

#H0 : Cust_Exp and Claim are independent and have no dependency on one another

##vs

#H1 : Cust_Exp and Claim are dependent on each other

```
chisq.test(M)
```

```
## Warning in chisq.test(M): Chi-squared approximation may be incorrect
```

```
##
```

```
## Pearson's Chi-squared test
```

```
##
```

```
## data: M
```

```
## X-squared = 47.749, df = 9, p-value = 2.846e-07
```

#Since the p - value is less than 5% , we can conclude that two variables are not independent and that policyholder experience changes with the number of claims

##(i)

```
binom.test(length(log_claim[log_claim>log(10000)]),  
            length(log_claim),  
            p = 0.25,  
            alternative = "greater")
```

```
##
```

```
## Exact binomial test
```

```
##
```

```
## data: length(log_claim[log_claim > log(10000)]) and length(log_claim)
```

```
## number of successes = 116, number of trials = 192, p-value < 2.2e-16
```

```
## alternative hypothesis: true probability of success is greater than 0.25
```

```
## 95 percent confidence interval:
```

```
## 0.5425955 1.0000000
```

```
## sample estimates:
```

```
## probability of success
```

```
## 0.6041667
```

#Since the 95% Confidence does not contain 0.25 , we can conclude that more than 25% of claims are large.

```
detach(policy_data)
```

#Q3B

```
indices_return = read.csv("Indices>Returns.csv")
```

##(a)

```
indices_return$Return_Direction = ifelse(indices_return$Sensex> 0,  
"Positive","Negative")
```

```
attach(indices_return)
```

##(b)

```
colnames(indices_return)
```

```
## [1] "Month"          "Sensex"          "BM"              "CD"
## [5] "EN"              "FM"              "FI"              "HC"
## [9] "IN"              "IT"              "TE"              "UT"
## [13] "Return_Direction"

indices_return$NumberReturnDirection = ifelse(indices_return$Return_Direction
== "Positive",1,0)
modelA = glm(NumberReturnDirection~BM+CD+EN+FM+FI+HC+IN+IT+TE+UT, family =
binomial(),
            data = indices_return)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(modelA)
```

```
##
## Call:
## glm(formula = NumberReturnDirection ~ BM + CD + EN + FM + FI +
##      HC + IN + IT + TE + UT, family = binomial(), data = indices_return)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27544  -0.00117   0.00000   0.01354   1.75651
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.0086     0.7315  -1.379   0.16796
## BM              7.7977    16.5255   0.472   0.63703
## CD            -87.5335    42.6785  -2.051   0.04027 *
## EN             93.9675    38.3193   2.452   0.01420 *
## FM             41.1745    20.1436   2.044   0.04095 *
## FI            172.8807    60.8192   2.843   0.00448 **
## HC             -6.4294    13.9394  -0.461   0.64463
## IN              4.1735    18.2152   0.229   0.81877
## IT             78.3494    30.9307   2.533   0.01131 *
## TE             29.9111    13.4184   2.229   0.02581 *
## UT            -14.4767    23.0602  -0.628   0.53015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 223.213  on 163  degrees of freedom
## Residual deviance:  32.905  on 153  degrees of freedom
## AIC: 54.905
##
## Number of Fisher Scoring iterations: 11
```

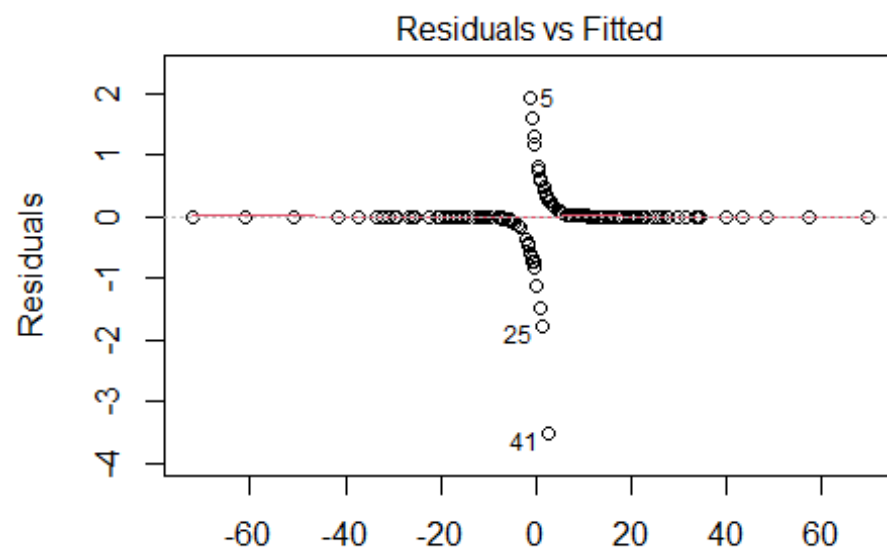
```
##(c)
```

```
#At 95% Level , CD , EN , FM , FI , IT and TE affect the Return Direction
#At 99% Level only FI sector affects the direction of Return.
```

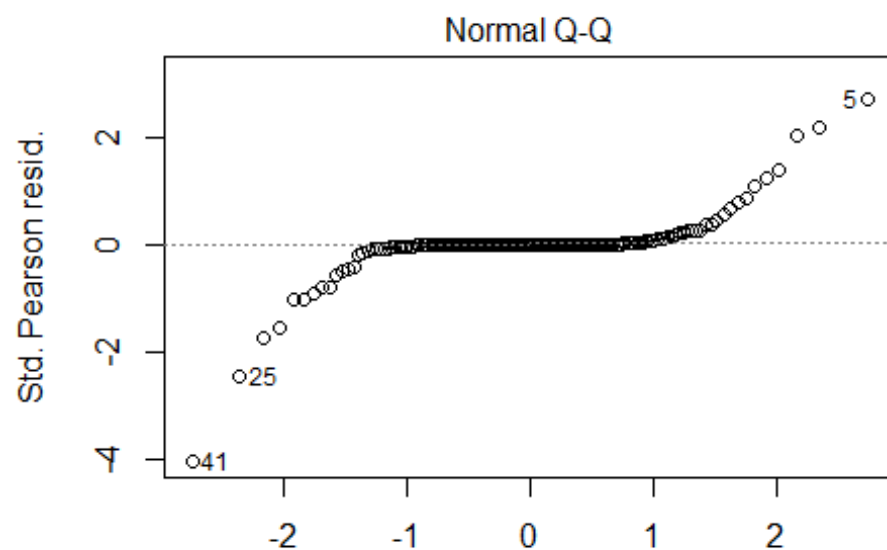
##(d)

##(e)

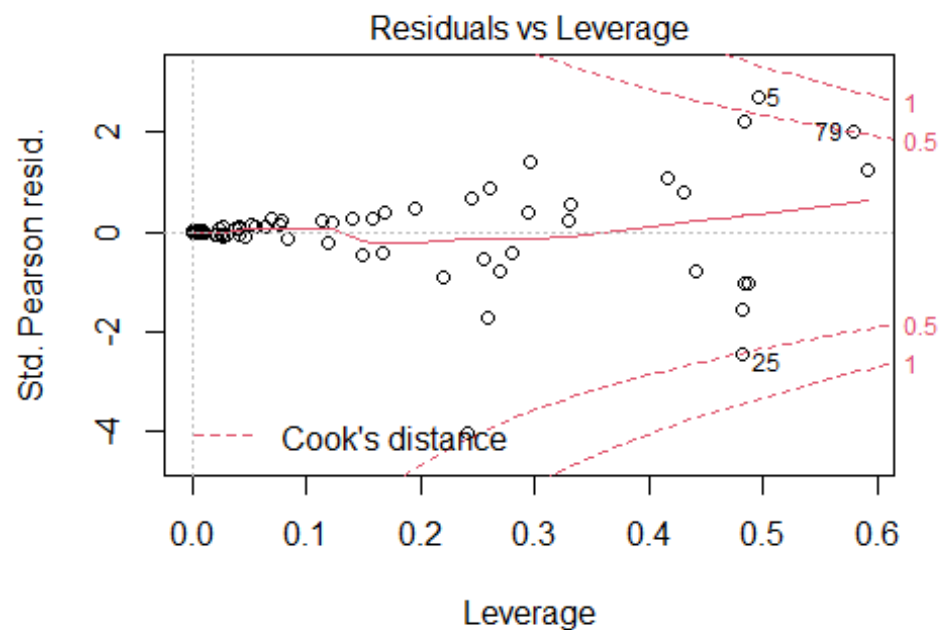
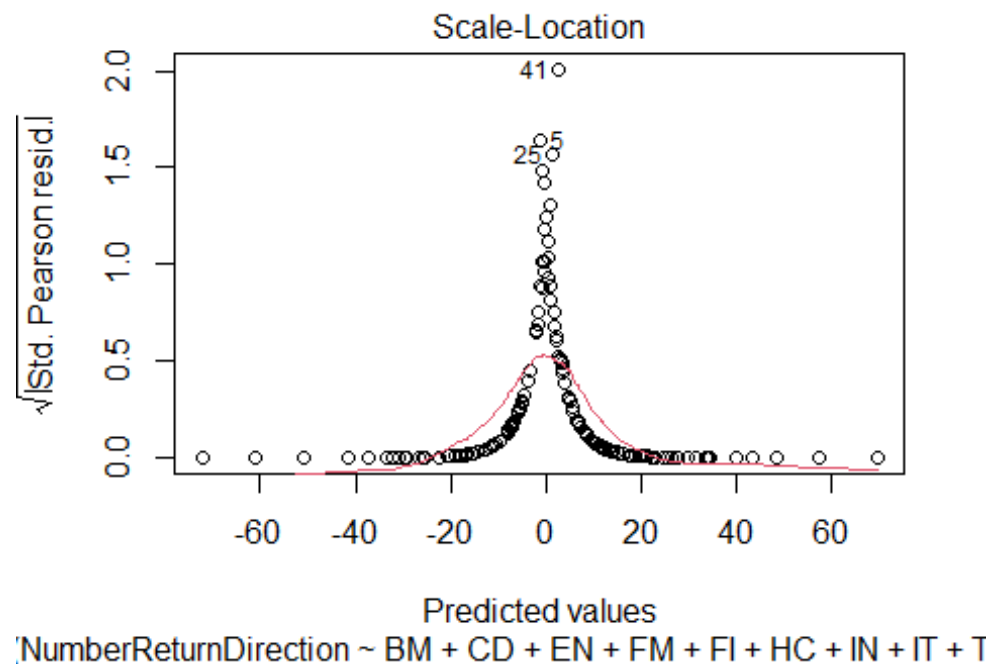
plot(modelA)



Predicted values
 $\text{NumberReturnDirection} \sim \text{BM} + \text{CD} + \text{EN} + \text{FM} + \text{FI} + \text{HC} + \text{IN} + \text{IT} + \text{T}$



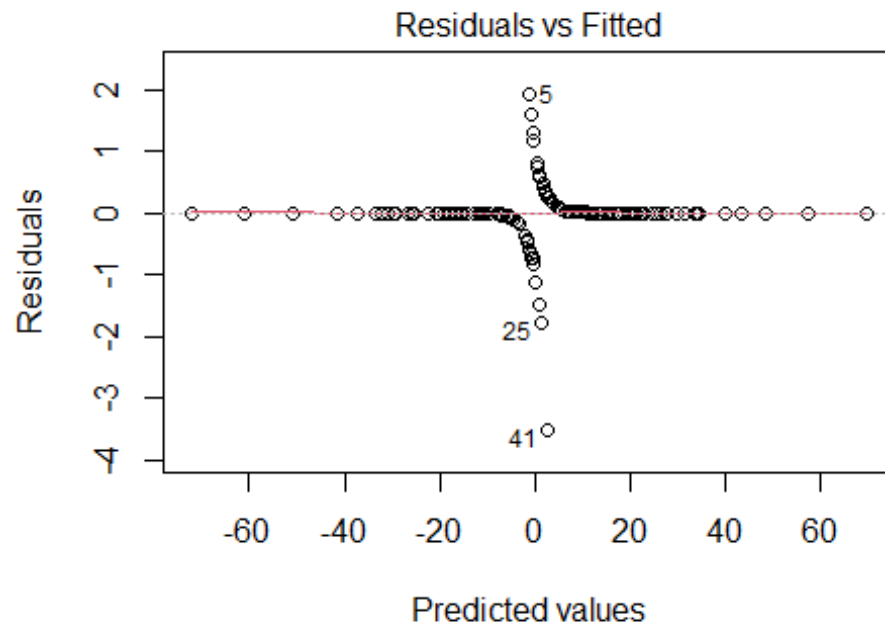
Theoretical Quantiles
 $\text{NumberReturnDirection} \sim \text{BM} + \text{CD} + \text{EN} + \text{FM} + \text{FI} + \text{HC} + \text{IN} + \text{IT} + \text{T}$



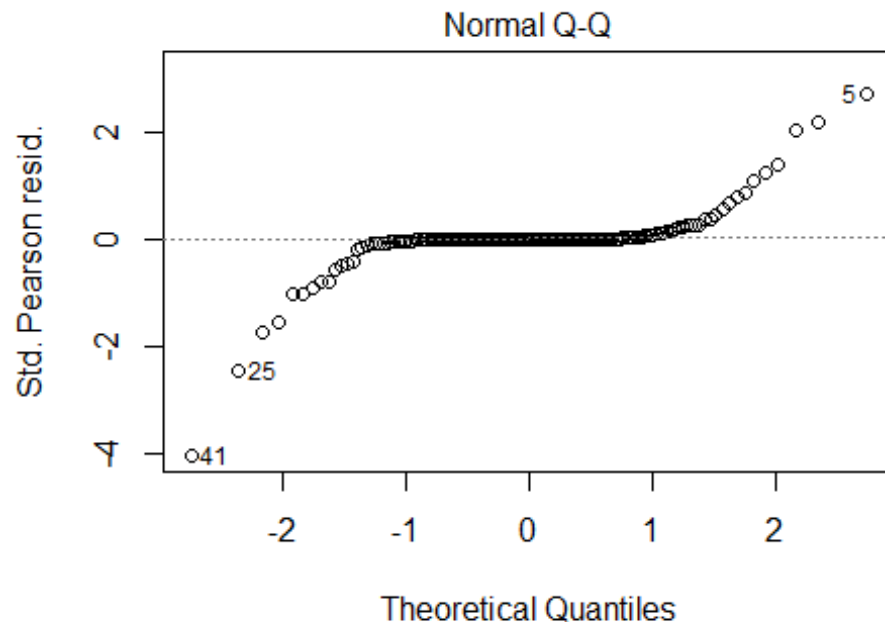
```
indices_return[5,1]
```

```
## [1] "Jun-06"
```

```
#The most significant outlier is from June 2006  
##(f)  
plot(modelA,1)
```



```
[NumberReturnDirection ~ BM + CD + EN + FM + FI + HC + IN + IT + T  
plot(modelA,2)
```



`NumberReturnDirection ~ BM + CD + EN + FM + FI + HC + IN + IT + T`

*#There is a clear pattern between the fitted and the residual values.
 #This means that the model is not appropriate
 #The Normal QQ Plot also indicates the Pearson residuals do not follow a Normal Distribution, which is expected as we used a Non - Normal family argument.
 #ALL this with the added point of having influential outliers in the data. It seems model can be improved and other data processes applied to fitting a GLM model.*

##(g)

`summary(modelA)`

##

Call:

`glm(formula = NumberReturnDirection ~ BM + CD + EN + FM + FI + HC + IN + IT + TE + UT, family = binomial(), data = indices_return)`

##

Deviance Residuals:

	Min	1Q	Median	3Q	Max
##	-2.27544	-0.00117	0.00000	0.01354	1.75651

##

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
## (Intercept)	-1.0086	0.7315	-1.379	0.16796
## BM	7.7977	16.5255	0.472	0.63703
## CD	-87.5335	42.6785	-2.051	0.04027 *
## EN	93.9675	38.3193	2.452	0.01420 *

```

## FM          41.1745    20.1436    2.044  0.04095 *
## FI          172.8807    60.8192    2.843  0.00448 **
## HC          -6.4294    13.9394   -0.461  0.64463
## IN           4.1735    18.2152    0.229  0.81877
## IT          78.3494    30.9307    2.533  0.01131 *
## TE          29.9111    13.4184    2.229  0.02581 *
## UT         -14.4767    23.0602   -0.628  0.53015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 223.213  on 163  degrees of freedom
## Residual deviance: 32.905  on 153  degrees of freedom
## AIC: 54.905
##
## Number of Fisher Scoring iterations: 11

modelB = update(modelA, .~.-1-BM-HC-IN-UT)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(modelB)

##
## Call:
## glm(formula = NumberReturnDirection ~ CD + EN + FM + FI + IT +
##      TE - 1, family = binomial(), data = indices_return)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28117  -0.00606   0.00002   0.02189   1.60681
##
## Coefficients:
##      Estimate Std. Error z value Pr(>|z|)
## CD    -81.26     28.73  -2.829  0.004676 **
## EN     70.39     23.28   3.024  0.002497 **
## FM     27.99     12.87   2.175  0.029663 *
## FI    145.56     38.40   3.790  0.000151 ***
## IT     61.85     18.69   3.309  0.000935 ***
## TE     32.03     11.36   2.821  0.004794 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 227.352  on 164  degrees of freedom
## Residual deviance: 35.948  on 158  degrees of freedom
## AIC: 47.948
##
## Number of Fisher Scoring iterations: 10

```

###Students could also do backward selection removing one at a time.

##(h)

```
anova(modelB,modelA,test = "Chi")
```

```
## Analysis of Deviance Table
```

```
##
```

```
## Model 1: NumberReturnDirection ~ CD + EN + FM + FI + IT + TE - 1
```

```
## Model 2: NumberReturnDirection ~ BM + CD + EN + FM + FI + HC + IN + IT +  
##      TE + UT
```

```
##      Resid. Df Resid. Dev Df Deviance Pr(>Chi)
```

```
## 1          158          35.948
```

```
## 2          153          32.905   5    3.0425    0.6934
```

#Since the pvalue is greater than 5%, we can conclude that modelB is better than modelA.