In [ ]:
```
STAT 420 HW7 Donghan Liu Netid: donghan2
```

In [ ]:
```
Question 1
```

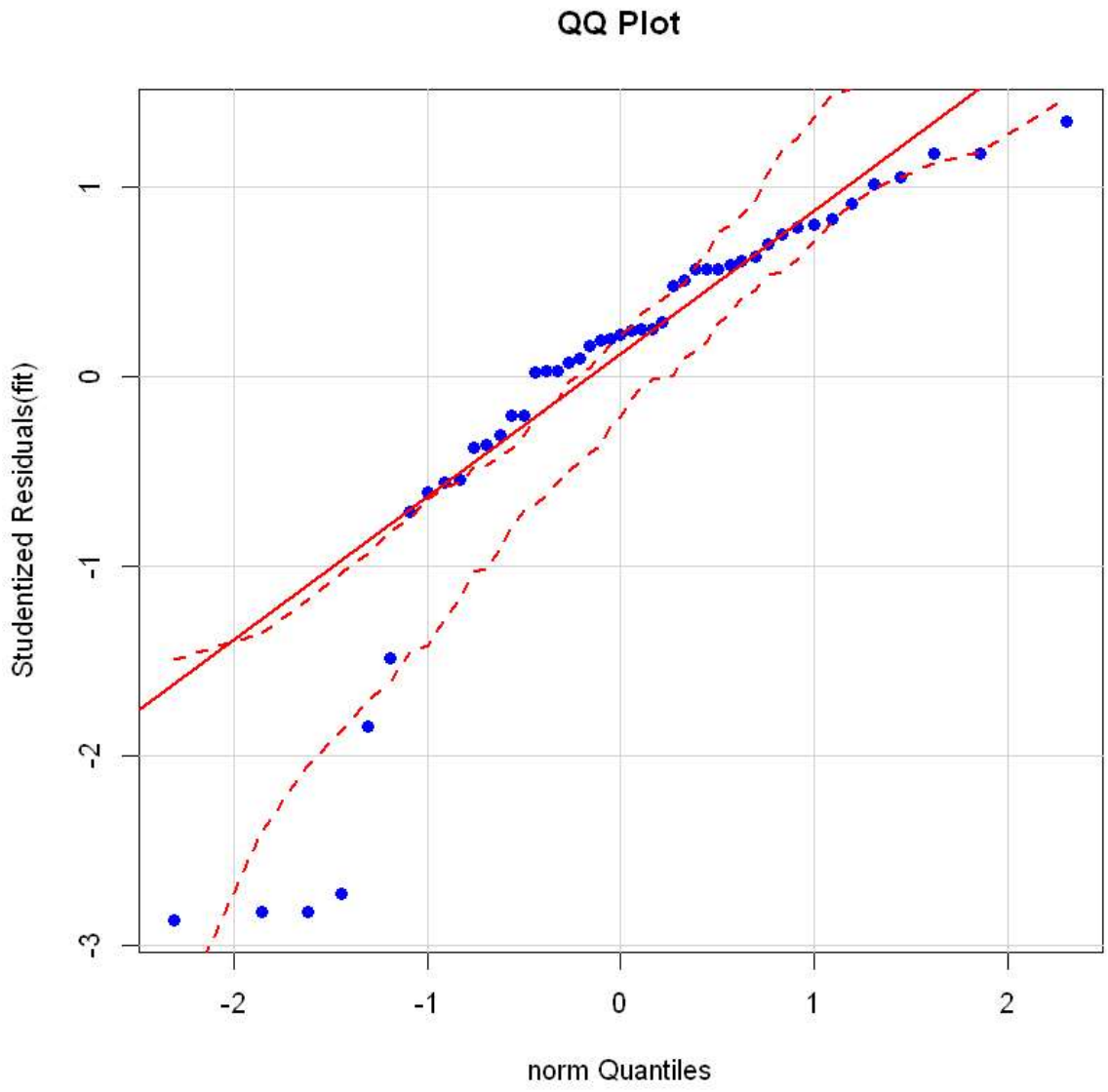
In [61]:
```
library(faraway)
data(star)
```

In [2]:
```
install.packages("car", repos = "http://cran.us.r-project.org")
```

```
package 'car' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Hans\AppData\Local\Temp\RtmpSoeMF3\downloaded_packages
```

In [3]:
```
library(car)
fit = lm(temp~light, data = star)
qqPlot(fit,distribution = "norm",main="QQ Plot", col = "blue", pch = 19, cex =
 1)
# As we could see the qq-plot below, these points are approximately line up
# in the predictive line, but it is not actually line up in the range of
# two dotted line, and there are two samll gaps in the continous points
# There are four points might be the main cause of inaccurate result for this
# kind of situation, which are these points that located in the bottom
# left corner, so we might consider remove them in order to fit the model
# normalitily.
```

```
Attaching package: 'car'

The following objects are masked from 'package:faraway':

    logit, vif
```

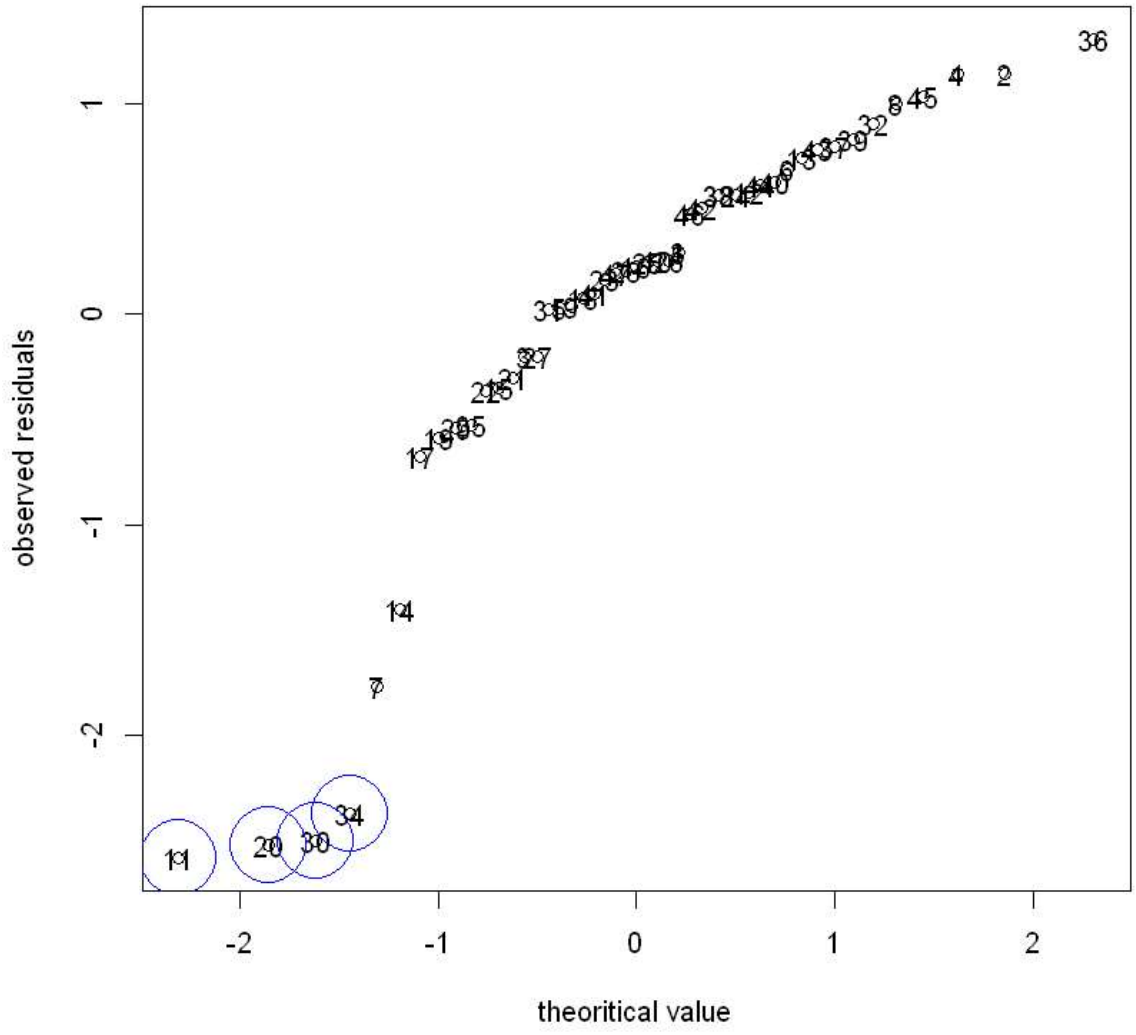


In [55]:
```
fit = lm(temp~light, data = star)
X = cbind(1,star$light)
which.max(cooks.distance(fit))
summary(influence.measures(fit))
# Then, we have to take a look at which points that are much higher than
# the cook'd cut-off line, which are four points in total, as we stated
# above. In the following table, the four points' index that seem far
# away from cut off line.
# By applying summary(influence.measures(fit)), we have the results
# below and conclude that the items with index of 11,20,30,34 have
# relatively high cook's d value and deviate most from the normal
# distribution, thus, removing these four points would be considerable.
```

34: 34

```
Potentially influential observations of
         lm(formula = temp ~ light, data = star) :

   dfb.1_ dfb.lght dffit   cov.r   cook.d hat
11  0.49  -0.55    -0.70_*  0.79_*  0.21   0.06
20  0.61  -0.66    -0.79_*  0.81_*  0.27   0.07
30  0.74  -0.79    -0.90_*  0.83_*  0.35   0.09
34  0.91  -0.96    -1.05_*  0.88    0.49   0.13_*
```

In [59]:
```
res = fit$residuals
perct = (rank(res)-0.5)/length(res)
z = qnorm(perct)
sigma = sqrt(sum(fit$residuals^2)/fit$df.residual)
plot(z, res/sigma, xlab = "theoritical value", ylab = "observed residuals")
text(z, res/sigma, c(1:length(res),pos=3))
y=res/sigma
points(z[c(11,20,30,34)],y[c(11,20,30,34)],cex=6,col="blue")
```

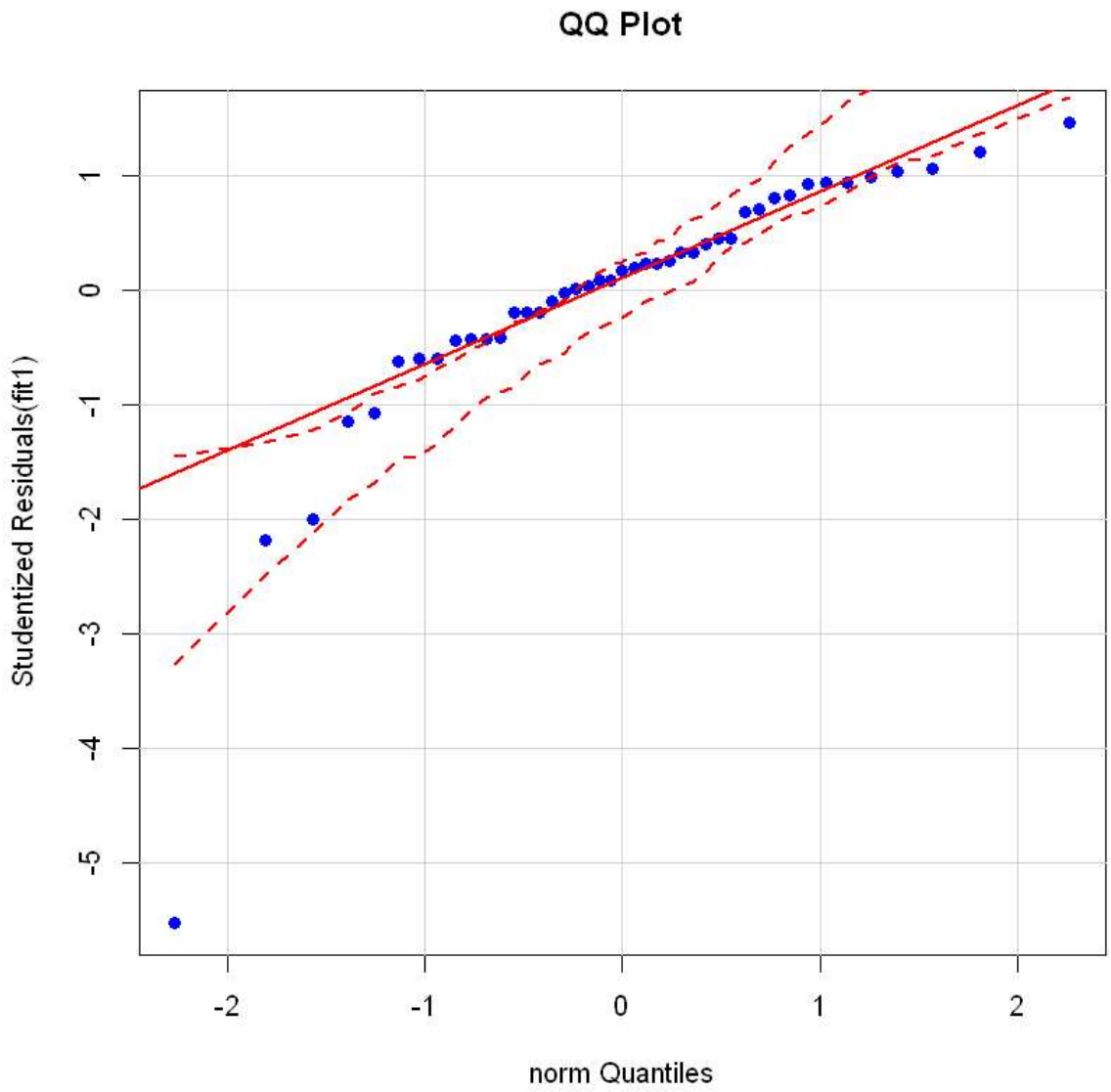


In [6]:
```
#b
res = fit$residuals
shapiro.test(res)
# We are determing the normaility of this dataset by using shapiro-wilk test
# As the normality test shows, W = 0.8474, which indicate that p-value
# for it is 2.208e-05, in other words, it is far less than 0.05, which
# is the cut-off line for significance. Thus, we would like to conclude
# that this dataset is more unlikely to be normal distribution.
```

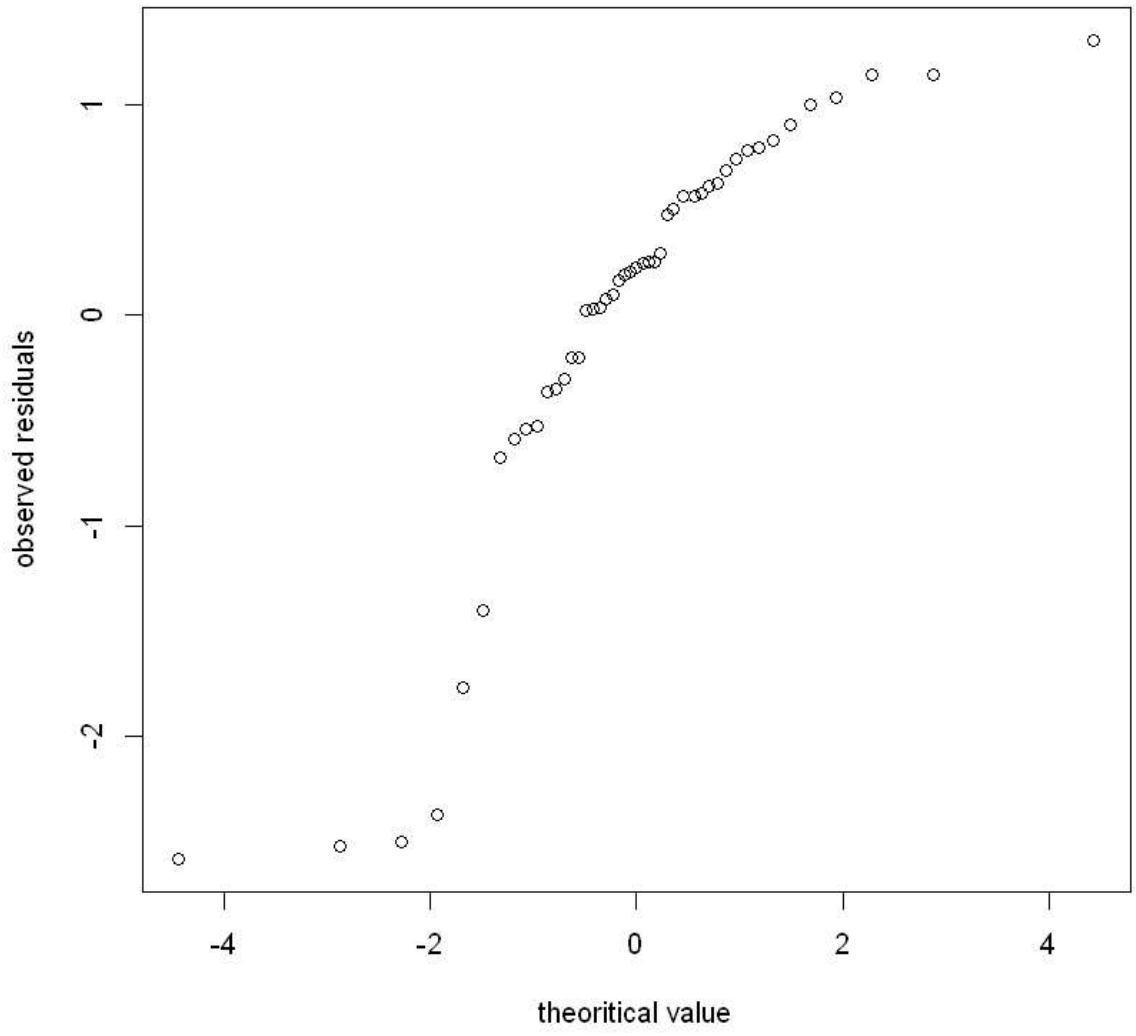
```
        Shapiro-Wilk normality test

data:  res
W = 0.8474, p-value = 2.208e-05
```

In [7]:
```
star1 = star[-c(11,20,30,34),]
fit1 = lm(temp~light, data = star1)
res1 = fit1$residuals
shapiro.test(res1)
qqPlot(fit1,distribution = "norm",main="QQ Plot", col = "blue", pch = 19, cex
= 1)
# In the result that generated below, the p-value is 3.676e-05, which has
# slight increasing than the one that we have not remove those outliers
# in other words, it indicates that removing those points make this model
# closer to normality, but it still lower than 0.05.Moreover, the two gaps
# in the graph becomes samller.
# Thus, we could say that this model does not follow normal distribution, even
 though
# the p-value incresed.
```

```
        Shapiro-Wilk normality test

data:  res1
W = 0.84388, p-value = 3.676e-05
```

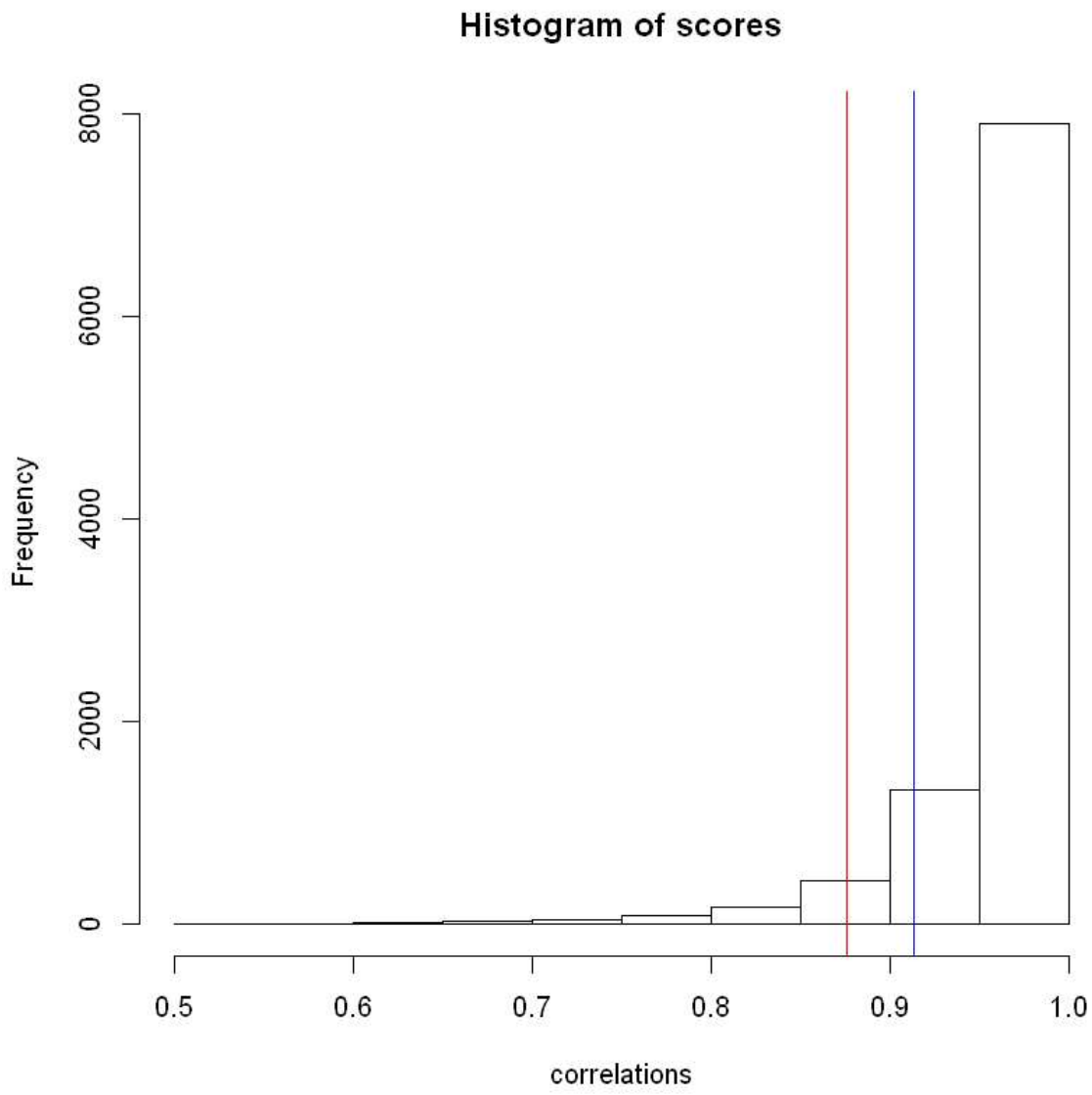


In [8]:
```
#c
res = fit$residuals
perct = (rank(res)-0.5)/length(res)
t = qt(perct,df = 3)
sigma = sqrt(sum(fit$residuals^2)/fit$df.residual)
plot(t, res/sigma, xlab = "theoritical value", ylab = "observed residuals")
# By generating a qq plot that follow t distribution, the graph seem like
# a line but not exactly lined up, in general, it is okay to admit that
# this is a t-distribution model and we will test whether it follows the
# t distribution in #d.
```



In [23]:
```
#d
res = fit$residuals
lg.test = function(res){
    perct = (rank(res)-0.375)/(length(res)+0.25)
    z = qt(perct,df = 3)
    c(cor(res,z),length(res))
}
lg.test(res)
scores = rep(NA, 10000)

for (i in 1:10000)
{
  x = rt(length(res),df = 3 )
  scores[i] = lg.test(x)[1]
}

lgcrit = quantile(scores, prob = 0.05)
hist(scores, xlab = "correlations")
abline( v = lgcrit, col = "red")
abline( v = lg.test(res)[1], col = "blue")
# For the graph that shown below, the cut off line is labled as red, and
# the generated line from data is marked as blue, obsivously, blue line is
# on the right side of the red line, and the left side of
# red line means the reject region, which states that it does not follow
# t distribution. Whereas, in this case, we have the blue is not in the
# reject ragion, thus, we could conclude that this dataset are likely follow
# the t distribution.
```

0.913246778798205 47



In [11]:
```
#2a
n = 1000
x1 = runif(n)
x2 = runif(n, 0, 2)
x3 = runif(n)
y = 2 + x1 + 2.5 * x2^2 + x1 * x3 + rnorm(n)
fit2 = lm(y~x1+x2+x3)
install.packages("lmtest", repos = "http://cran.us.r-project.org")
library(lmtest)
```

```
package 'lmtest' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Hans\AppData\Local\Temp\RtmpSoeMF3\downloaded_packages

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

    as.Date, as.Date.numeric
```

In [12]:
```
library(lmtest)
bptest(fit2)
# As we can see, the p-value = 0.4788, which is much greater than 0.05,
# to put it another way, it is highly insignificant and we would like
# to express that accepting the homoscedasticity and reject heteroscedasticity
# with only consider the bp test's p-value.
```

```
        studentized Breusch-Pagan test

data:  fit2
BP = 2.4809, df = 3, p-value = 0.4788
```

In [13]:
```
#b
white.fit = lm(fit2$residuals^2 ~ x1+x2+x3 + I(x1^2)+I(x2^2)+I(x3^2) + x1*x2+x
1*x3+x2*x3)
white = summary(white.fit)$r.squared*n
1-pchisq(white, 9)
# In the result below, the p-value is 0.00788, it is clearly to see that
# it is less than 0.05, so we could say that heteroscedasticity could be
# accepted. However, due to that the natural disadvantage, we got the differen
t
# results between a and b. The reason of that might because of model misspecif
ication.
```

```
0.0078768273824491
```

In [50]:
```
#c
res = fit2$residuals
X = cbind(1,x1,x2,x3)
sw = solve(t(X)%*%X)%*%t(X)%*%diag(res^2)%*% X %*%solve(t(X)%*%X)
round(sw,5)

#reduce the bias
round((sw*n)/(n-4),5)
```

		x1	x2	x3
	0.01681	-0.00947	-0.00528	-0.00973
x1	-0.00947	0.01859	-0.00026	-0.00015
x2	-0.00528	-0.00026	0.00538	0.00010
x3	-0.00973	-0.00015	0.00010	0.01894

		x1	x2	x3
	0.01688	-0.00951	-0.00530	-0.00976
x1	-0.00951	0.01866	-0.00027	-0.00015
x2	-0.00530	-0.00027	0.00540	0.00010
x3	-0.00976	-0.00015	0.00010	0.01902

In [53]:
```
# t test
summary(fit2)
t = (1.47389-0)/(sqrt(0.01866))
t
pvalue = pt(t, df=998)
1-pvalue
# As the above table shows that, the reduced bias variance for x1 is 0.02032,
 so we could use this value to calculate the t test
# for determining if beta1 is equal to zero. By applying summary(), the estima
ted paramater for x1 is 1.47389, and t formula
# is (beta1-0)/(sqrt(var)), then use pt(), we get the result of 0, which indic
ates that beta1 is more unlikely to be zero.
```

```
Call:
lm(formula = y ~ x1 + x2 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4618 -0.8606 -0.0146  0.8148  3.8882 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.15082    0.12486   1.208    0.227    
x1           1.47389    0.13710  10.751  < 2e-16 ***
x2           4.90603    0.06833  71.801  < 2e-16 ***
x3           0.58673    0.13882   4.227 2.59e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.256 on 996 degrees of freedom
Multiple R-squared:  0.8429,    Adjusted R-squared:  0.8425 
F-statistic:  1782 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
10.7896970925699
```

```
0
```