Question 1 (8 points) There is provided the dataset ”SwedenEcon.csv” which has been worked with in class. You will work with that dataset for this question:

1. Apply LOESS to the columns of ”NominalExport” and ”ExportGrowth” separately. Each plot should have the original data and the smoothing that LOESS provides (or LOWESS is also an approach to use).
2. Redraw those plots with a logarithmic scale for the y-axis values
3. Build a simple linear regression model and a polynomial regression model with powers up to the value of 3 to fit the prediction where the ’dependent is GDPbyActivityBasicPrices’ and ’independent is ExportGrowth’. Plot the fits and the original data (3 lines).
4. Compare the 2 models using the F-test (anova(model1,model2,test=”F”)) and report on the decision for the choice of model.
5. Use the AIC function for the 2 models and report which model you choose on the outputs it provides.

In [108]:
```
install.packages("ggpubr")
install.packages("datarium")
install.packages("glmnet")
install.packages("readxl")
install.packages("caret")
install.packages("mlbench")
install.packages("psych")
install.packages("ggplot2")
install.packages("DAAG")
install.packages("MASS")
install.packages("relaimpo")
install.packages("TTR")
install.packages("tseries")
install.packages("data.table")
install.packages("MTS")
install.packages("ggfortify")
```

```
Warning message:
"package 'ggpubr' is in use and will not be installed"

package 'datarium' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Andrew\AppData\Local\Temp\RtmpI5MZBx\downloaded_packages

Warning message:
"package 'glmnet' is in use and will not be installed"Warning message:
"package 'readxl' is in use and will not be installed"Warning message:
"package 'caret' is in use and will not be installed"Warning message:
"package 'mlbench' is in use and will not be installed"Warning message:
"package 'psych' is in use and will not be installed"Warning message:
"package 'ggplot2' is in use and will not be installed"Warning message:
"package 'DAAG' is in use and will not be installed"Warning message:
"package 'MASS' is in use and will not be installed"Warning message:
"package 'relaimpo' is in use and will not be installed"Warning message:
"package 'TTR' is in use and will not be installed"Warning message:
"package 'tseries' is in use and will not be installed"Warning message:
"package 'data.table' is in use and will not be installed"Warning message:
"package 'MTS' is in use and will not be installed"

package 'ggfortify' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Andrew\AppData\Local\Temp\RtmpI5MZBx\downloaded_packages
```

In [109]:
```
library(ggpubr)
#Library(tidyverse)
library(readxl)
library(glmnet)
library(caret)
library(mlbench)
library(psych)
library(ggplot2)
library(DAAG)
library(MASS)
library(relaimpo)
library(TTR)
library(tseries)
library(data.table)
library(MTS)
library(ggfortify)
```

```
Warning message:
"package 'ggfortify' was built under R version 3.6.3"
```

```
Error: package or namespace load failed for 'ggfortify' in loadNamespace(j <- i[[1L]], c(lib.loc, .libPaths()), ver
sionCheck = vI[[j]]):
 namespace 'dplyr' 0.8.0.1 is already loaded, but >= 0.8.2 is required
Traceback:

1. library(ggfortify)
2. tryCatch({
 .     attr(package, "LibPath") <- which.lib.loc
 .     ns <- loadNamespace(package, lib.loc)
 .     env <- attachNamespace(ns, pos = pos, deps, exclude, include.only)
 . }, error = function(e) {
 .     P <- if (!is.null(cc <- conditionCall(e)))
 .         paste(" in", deparse(cc)[1L])
 .     else ""
 .     msg <- gettextf("package or namespace load failed for %s%s:\n %s",
 .         sQuote(package), P, conditionMessage(e))
 .     if (logical.return)
 .         message(paste("Error:", msg), domain = NA)
 .     else stop(msg, call. = FALSE, domain = NA)
 . })
3. tryCatchList(expr, classes, parentenv, handlers)
4. tryCatchOne(expr, names, parentenv, handlers[[1L]])
5. value[[3L]](cond)
6. stop(msg, call. = FALSE, domain = NA)
```

# Since you mention on question one, you want the original data, i did not take off NA

1. Apply LOESS to the columns of ”NominalExport” and ”ExportGrowth” separately. Each plot should have the original data and the smoothing that LOESS provides (or LOWESS is also an approach to use).

In [110]:
```
SwedenEcon<- read.csv('C:/Users/Andrew/Desktop/SwedenEcon-1.csv')
str(SwedenEcon)
```

```
'data.frame':   216 obs. of  15 variables:
 $ Year                                : int  1800 1801 1802 1803 1804 1805 1806 1807 1808 1809 ...
 $ PrivateGrowthConsumptionVolume      : num  NA -2.66 3.98 1.09 11.92 ...
 $ PrivateGrowthConsumptionVolumeNominal: num  142 138 136 132 149 ...
 $ GovernmentConsumptionGowthVolume    : num  NA -1.39 4.49 8.09 -6.15 ...
 $ GovernmentConsumptionGowthNominal   : num  13.4 12.7 12.7 13.1 12.6 ...
 $ NominalExport                       : num  15.8 20.9 22.3 20.8 19.2 ...
 $ ExportGrowth                        : num  NA 32.54 8.5 -9.73 -9.71 ...
 $ NominalImport                       : num  14.8 19 14.5 13.8 23.5 ...
 $ ImportGrowth                        : num  NA 24.49 -15.43 -5.26 66.81 ...
 $ GDPbyActivityBasicPrices            : num  110 113 116 110 109 ...
 $ GDPbyActivityCorrectedForChanges    : num  -4.449 3.812 0.545 -0.736 4.485 ...
 $ DividendIndex                       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ GovShortYield                       : num  NA NA NA NA NA NA NA NA NA NA ...
 $ AgricultureEmployeeNumber           : num  NA NA NA NA NA NA NA NA NA NA ...
 $ GovernmentServiceEmployeeNumber     : num  NA NA NA NA NA NA NA NA NA NA ...
```

In [111]:
```
head(SwedenEcon, 10)
summary(SwedenEcon)
```

| Year | PrivateGrowthConsumptionVolume | PrivateGrowthConsumptionVolumeNominal | GovernmentConsumptionGowthVolume | GovernmentConsumpti |
|---|---|---|---|---|
| 1800 | NA | 141.7600 | NA | |
| 1801 | -2.6552163 | 138.3895 | -1.3924178 | |
| 1802 | 3.9828373 | 135.7893 | 4.4855156 | |
| 1803 | 1.0860552 | 132.0632 | 8.0936523 | |
| 1804 | 11.9233158 | 148.9401 | -6.1537242 | |
| 1805 | -4.2713947 | 144.6281 | 3.0147266 | |
| 1806 | -1.2301890 | 165.0820 | -7.7087453 | |
| 1807 | -5.2216500 | 160.7495 | -0.5106037 | |
| 1808 | 4.0913070 | 203.2462 | -0.9651354 | |
| 1809 | -0.5753731 | 206.9595 | -15.3344311 | |

```
     Year      PrivateGrowthConsumptionVolume
Min.   :1800   Min.   :-10.2622
1st Qu.:1854   1st Qu.: -0.7811
Median :1908   Median :  2.2500
Mean   :1908   Mean   :  2.1255
3rd Qu.:1961   3rd Qu.:  4.4713
Max.   :2015   Max.   : 26.1188
               NA's   :2
PrivateGrowthConsumptionVolumeNominal GovernmentConsumptionGowthVolume
Min.   :    132.1                     Min.   :-33.64338
1st Qu.:    585.1                     1st Qu.:  0.03575
Median :   2595.0                     Median :  2.11766
Mean   : 177589.1                     Mean   :  2.35570
3rd Qu.:  44404.1                     3rd Qu.:  4.83320
Max.   :1811947.0                     Max.   : 29.33966
NA's   :1                             NA's   :2
GovernmentConsumptionGowthNominal NominalExport        ExportGrowth
Min.   :     12.6                 Min.   :     12.7   Min.   :-44.718
1st Qu.:     40.2                 1st Qu.:     81.6   1st Qu.: -1.267
Median :    246.3                 Median :    646.8   Median :  4.688
Mean   :  94434.9                 Mean   : 150340.1   Mean   :  4.470
3rd Qu.:  12538.6                 3rd Qu.:  18100.5   3rd Qu.:  9.385
Max.   :1030997.0                 Max.   :1743745.0   Max.   : 76.547
NA's   :1                         NA's   :1           NA's   :2
NominalImport        ImportGrowth     GDPbyActivityBasicPrices
Min.   :     13.8   Min.   :-45.090   Min.   :    108.7
1st Qu.:     69.6   1st Qu.: -2.325   1st Qu.:    438.8
Median :    712.6   Median :  5.200   Median :   2077.1
Mean   : 132527.5   Mean   :  5.311   Mean   : 140083.9
3rd Qu.:  17538.4   3rd Qu.: 10.659   3rd Qu.:  29060.9
Max.   :1600463.0   Max.   :225.572   Max.   :1813900.4
NA's   :1           NA's   :2         NA's   :15
GDPbyActivityCorrectedForChanges DividendIndex     GovShortYield
Min.   :-10.5558                 Min.   : 0.0180   Min.   : 0.200
1st Qu.:  0.5156                 1st Qu.: 0.0460   1st Qu.: 3.750
Median :  2.9639                 Median : 0.0615   Median : 5.000
Mean   :  2.3776                 Mean   : 0.9897   Mean   : 5.173
3rd Qu.:  4.4715                 3rd Qu.: 0.2152   3rd Qu.: 6.000
Max.   : 13.0124                 Max.   :15.9690   Max.   :14.150
NA's   :16                       NA's   :74        NA's   :59
AgricultureEmployeeNumber GovernmentServiceEmployeeNumber
Min.   : 129211           Min.   :  49276
1st Qu.: 519458           1st Qu.:  89943
Median :1115254           Median : 141695
Mean   : 899337           Mean   : 397322
3rd Qu.:1233428           3rd Qu.: 510837
Max.   :1334343           Max.   :1436600
NA's   :65                NA's   :65
```

In [112]:
```
Nexport=SwedenEcon[,6]
NexportMat=as.matrix(Nexport)
dim(NexportMat)
is.matrix(NexportMat)
head(NexportMat)
```

216 1

TRUE

15.82142

20.92968

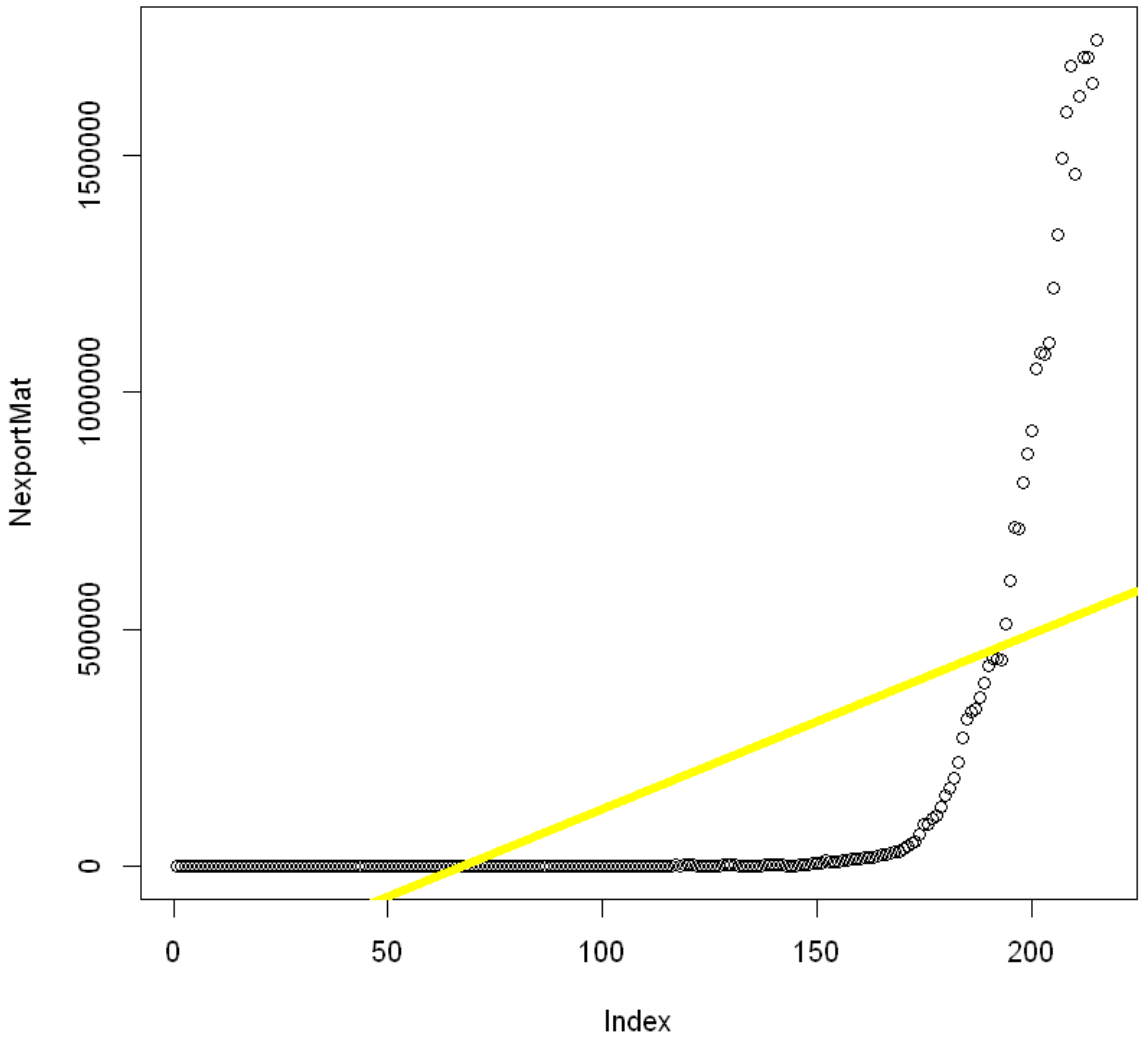
22.33747

20.82065

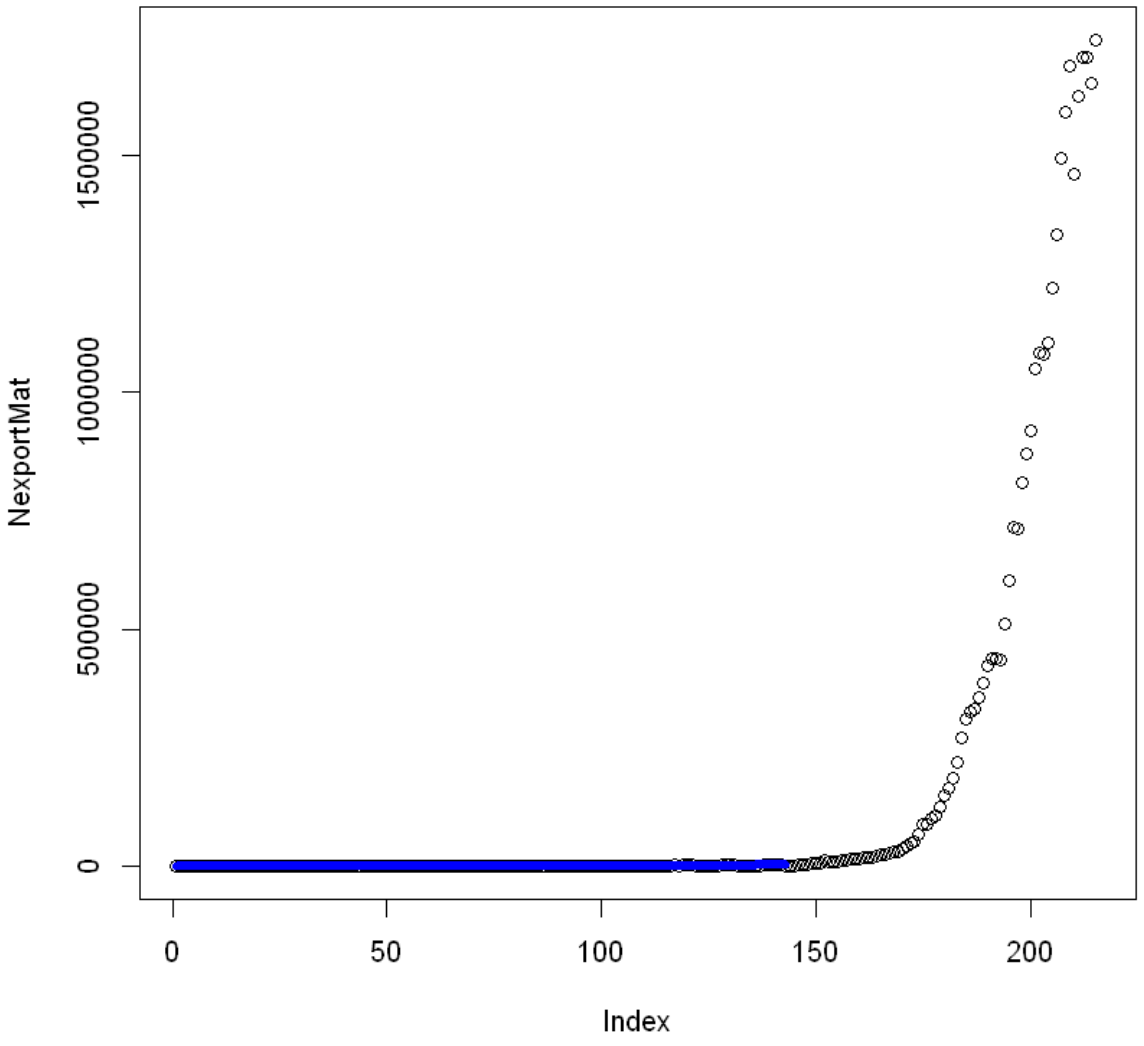
19.16201

18.33463

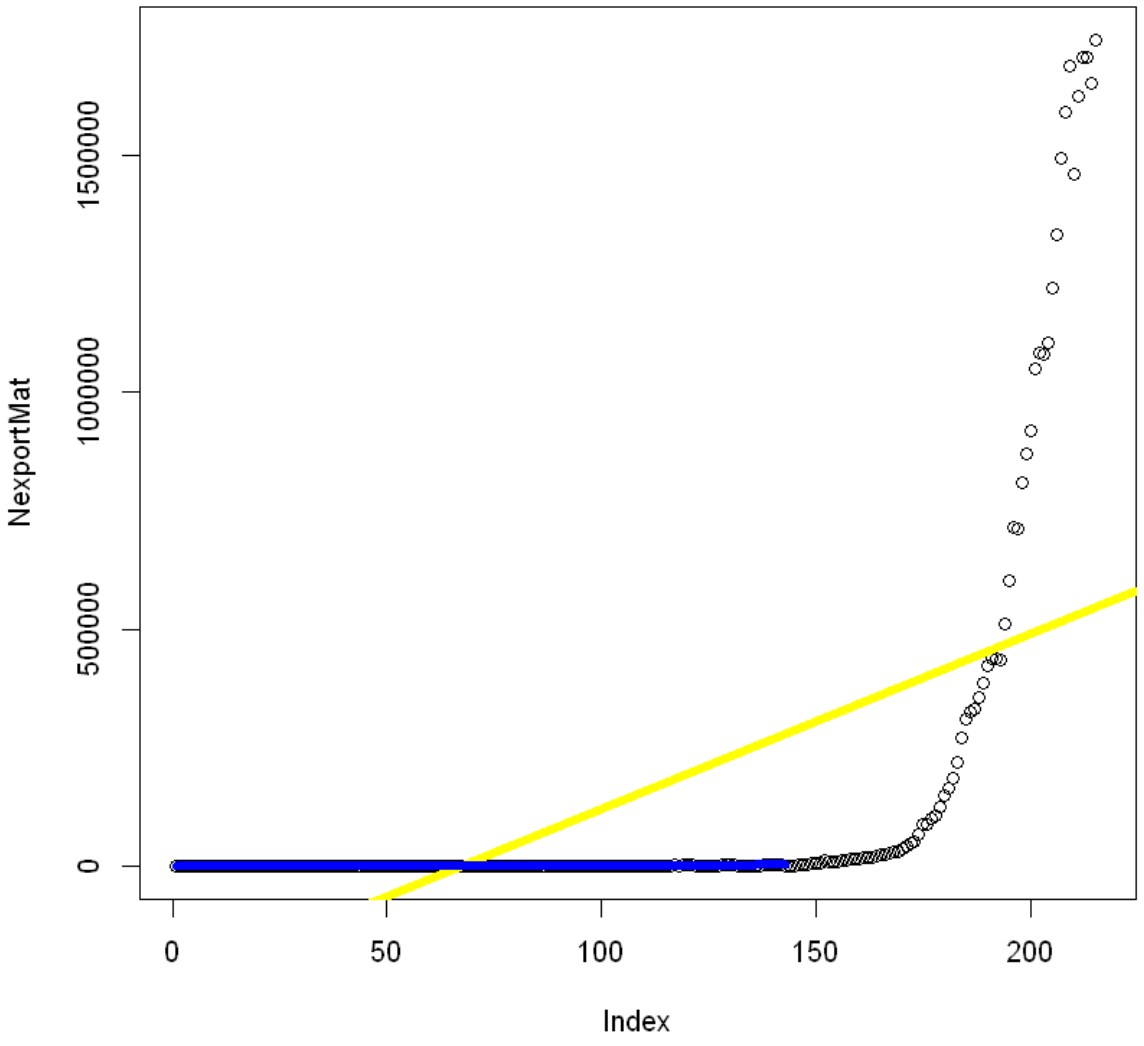
In [113]:
```
plot(NexportMat)
x=seq(1,nrow(NexportMat))
lmM1=abline(lm(NexportMat ~x), col="yellow", lwd=5)
```



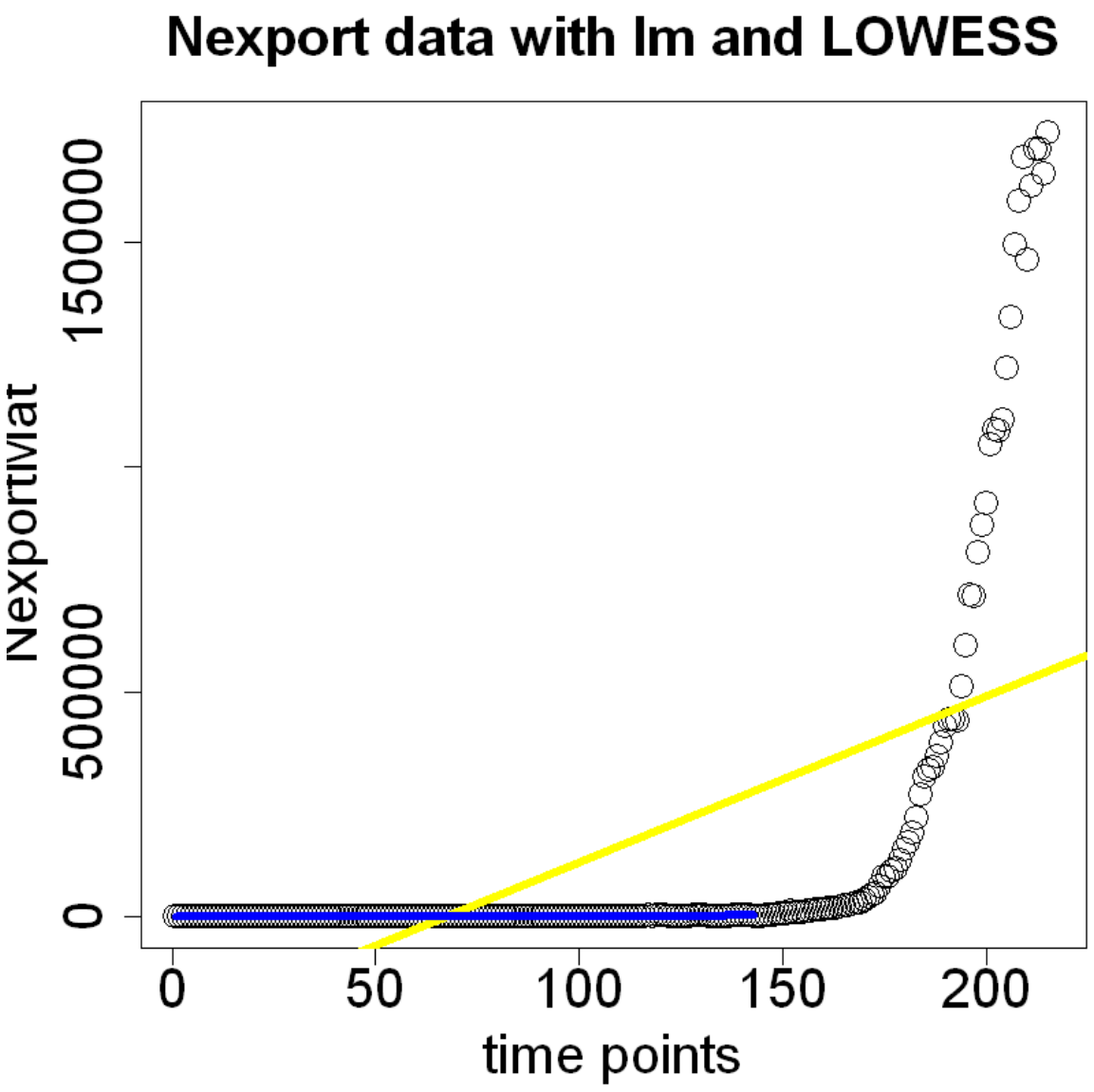
In [114]:
```
#view the LOWESS
plot(NexportMat)
x=seq(1,nrow(NexportMat))
loessM1=lines(lowess(NexportMat), col="blue", lwd=5)
```



In [115]:
```
#view both of them together on the same plot
plot(NexportMat)
lmM1=abline(lm(NexportMat ~x), col="yellow", lwd=5)
loessM1=lines(lowess(NexportMat), col="blue", lwd=5)
```



In [116]:
```
#change the plot parameters
plot(NexportMat, main="Nexport data with lm and LOWESS",
    xlab="time points", cex=2,cex.lab=2, cex.axis=2, cex.main=2)
lmM1=abline(lm(NexportMat ~x), col="yellow", lwd=5)
loessM1=lines(lowess(NexportMat), col="blue", lwd=5)
```



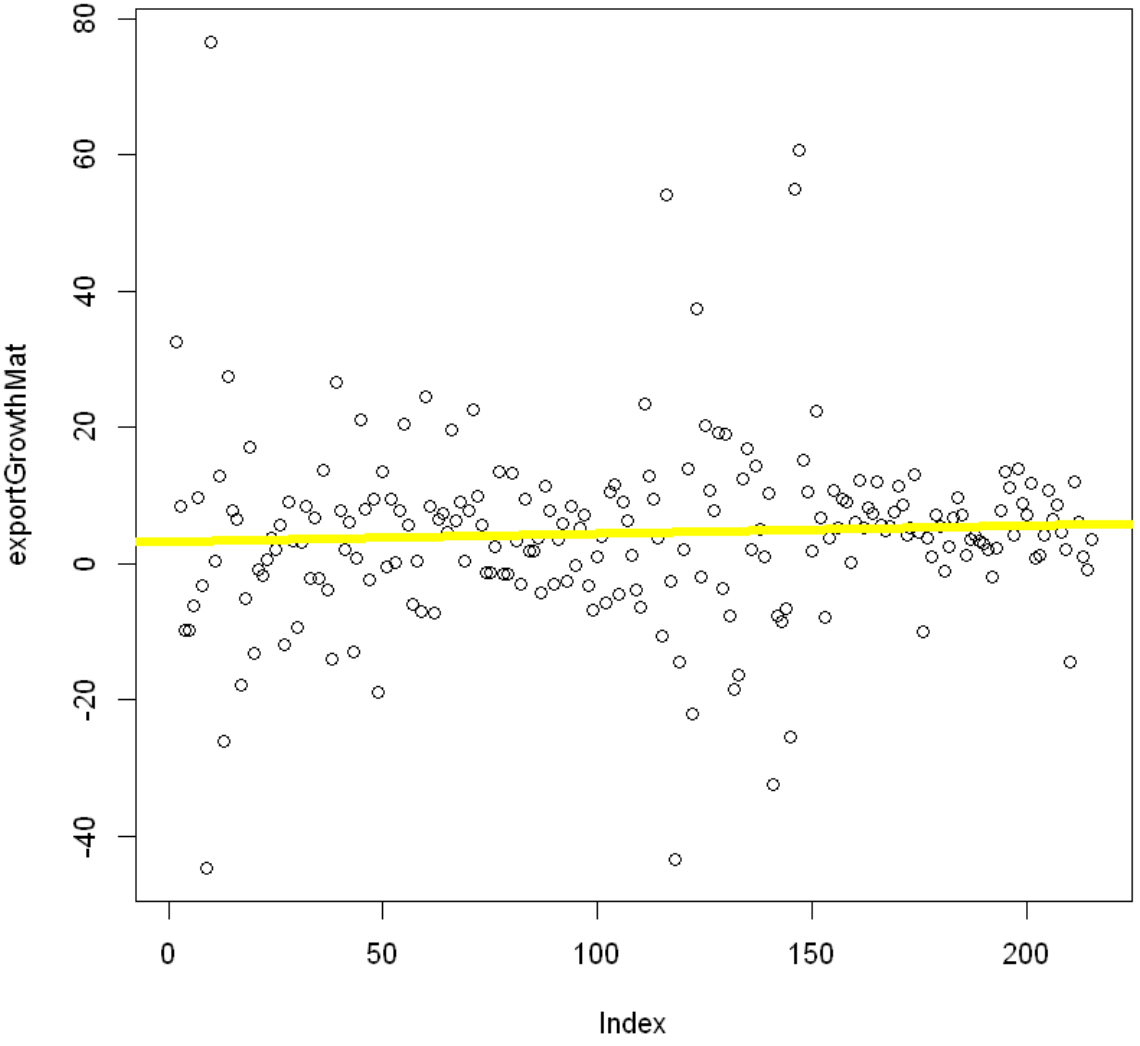

In [117]:
```
#now move to ExportGrowth variable
exportGrowth=SwedenEcon[,7]
exportGrowthMat=as.matrix(exportGrowth)
dim(exportGrowthMat)
is.matrix(exportGrowthMat)
head(exportGrowthMat)
```

216 1

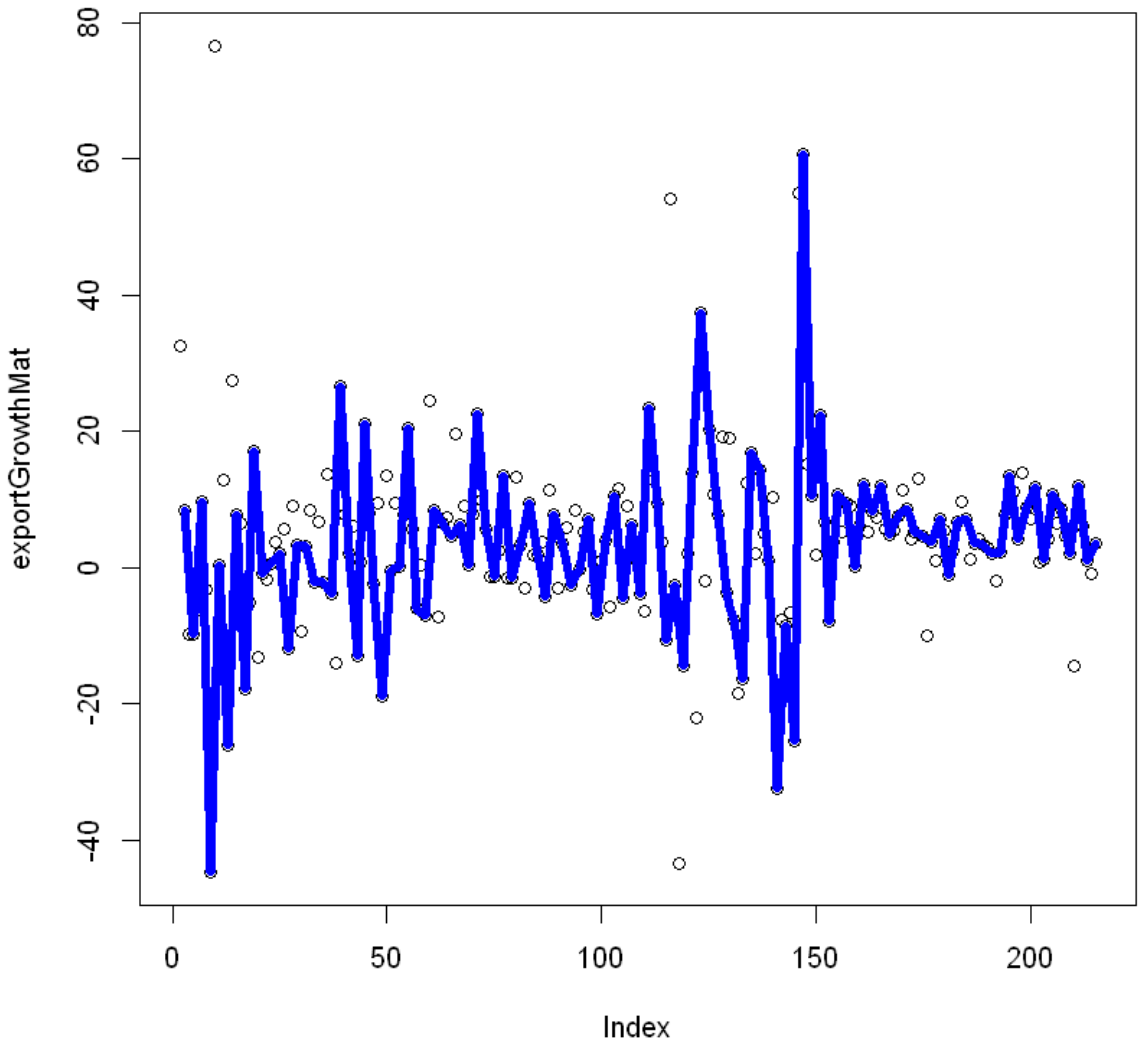
TRUE

| |
|---|
| NA |
| 32.542291 |
| 8.500821 |
| -9.728642 |
| -9.706756 |
| -6.110455 |

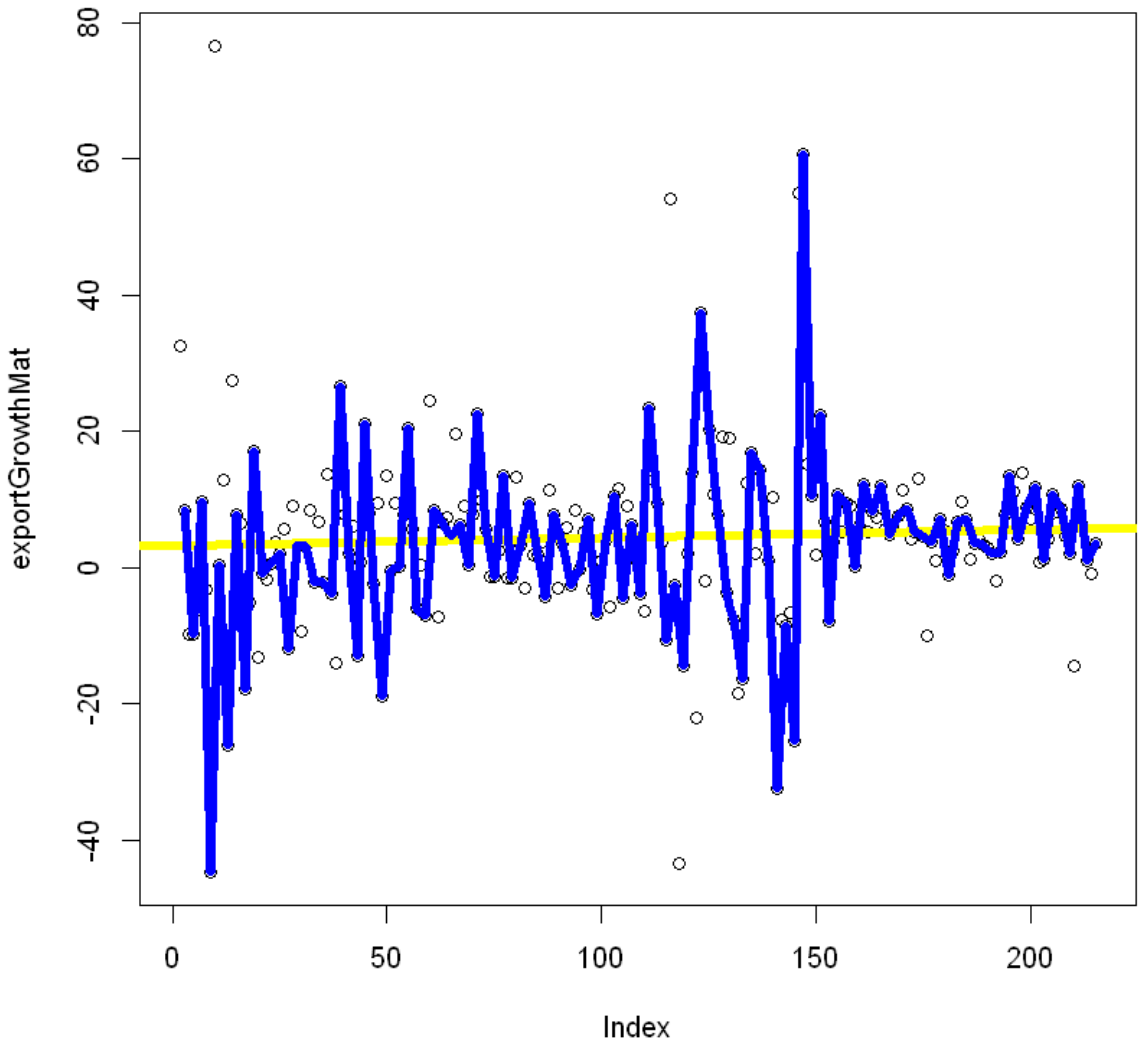
In [118]:
```
plot(exportGrowthMat)
x=seq(1,nrow(exportGrowthMat))
lmM1=abline(lm(exportGrowthMat ~x), col="yellow", lwd=5)
```



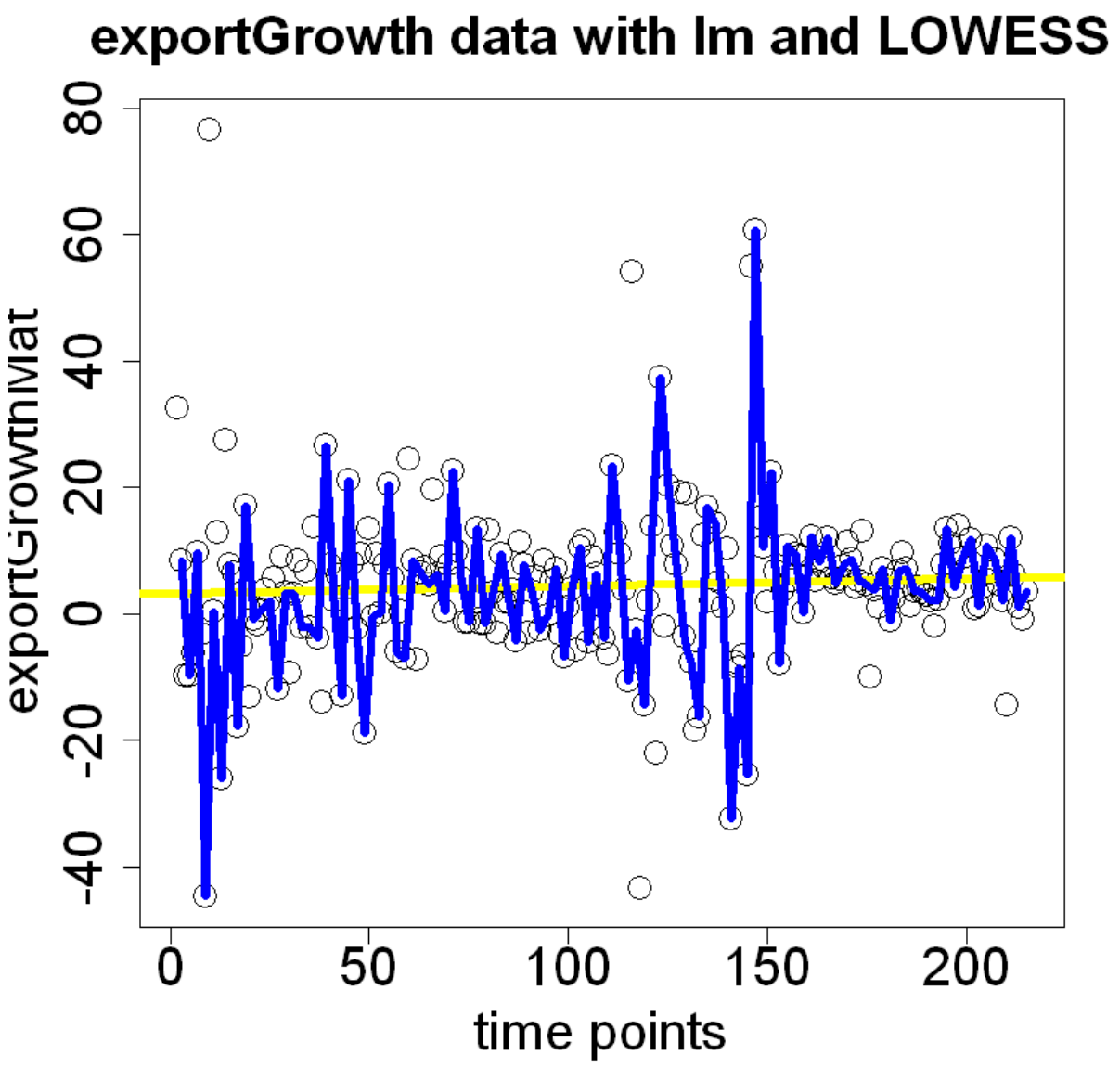
In [119]:
```
#view the LOWESS
plot(exportGrowthMat)
x=seq(1,nrow(exportGrowthMat))
loessM1=lines(lowess(exportGrowthMat), col="blue", lwd=5)
```



In [120]:
```
#view both of them together on the same plot
plot(exportGrowthMat)
lmM1=abline(lm(exportGrowthMat ~x), col="yellow", lwd=5)
loessM1=lines(lowess(exportGrowthMat), col="blue", lwd=5)
```



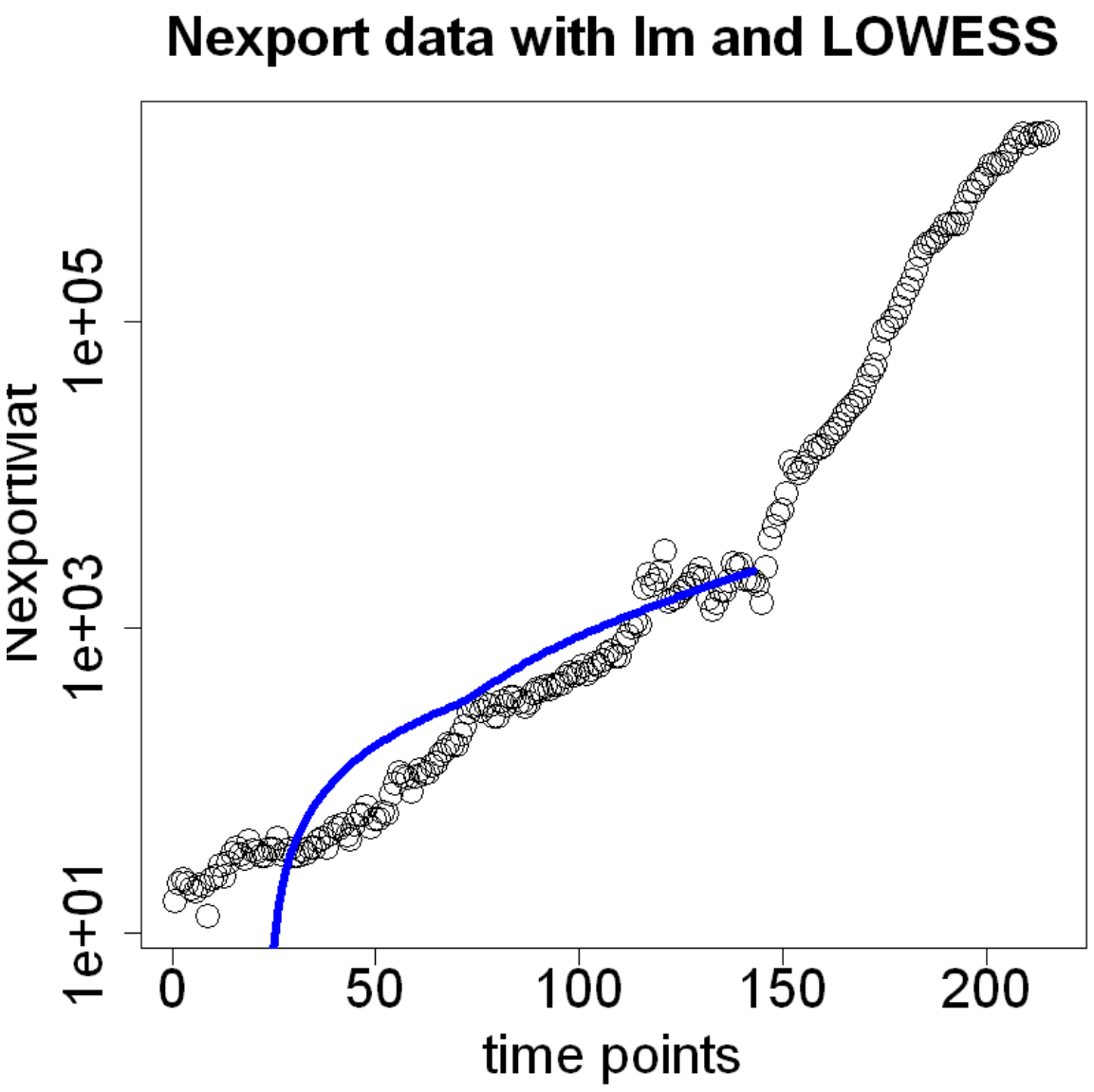
In [121]:
```
#change the plot parameters
plot(exportGrowthMat, main="exportGrowth data with lm and LOWESS",
    xlab="time points", cex=2,cex.lab=2, cex.axis=2, cex.main=2)
lmM1=abline(lm(exportGrowthMat ~x), col="yellow", lwd=5)
loessM1=lines(lowess(exportGrowthMat), col="blue", lwd=5)
```




In [ ]:

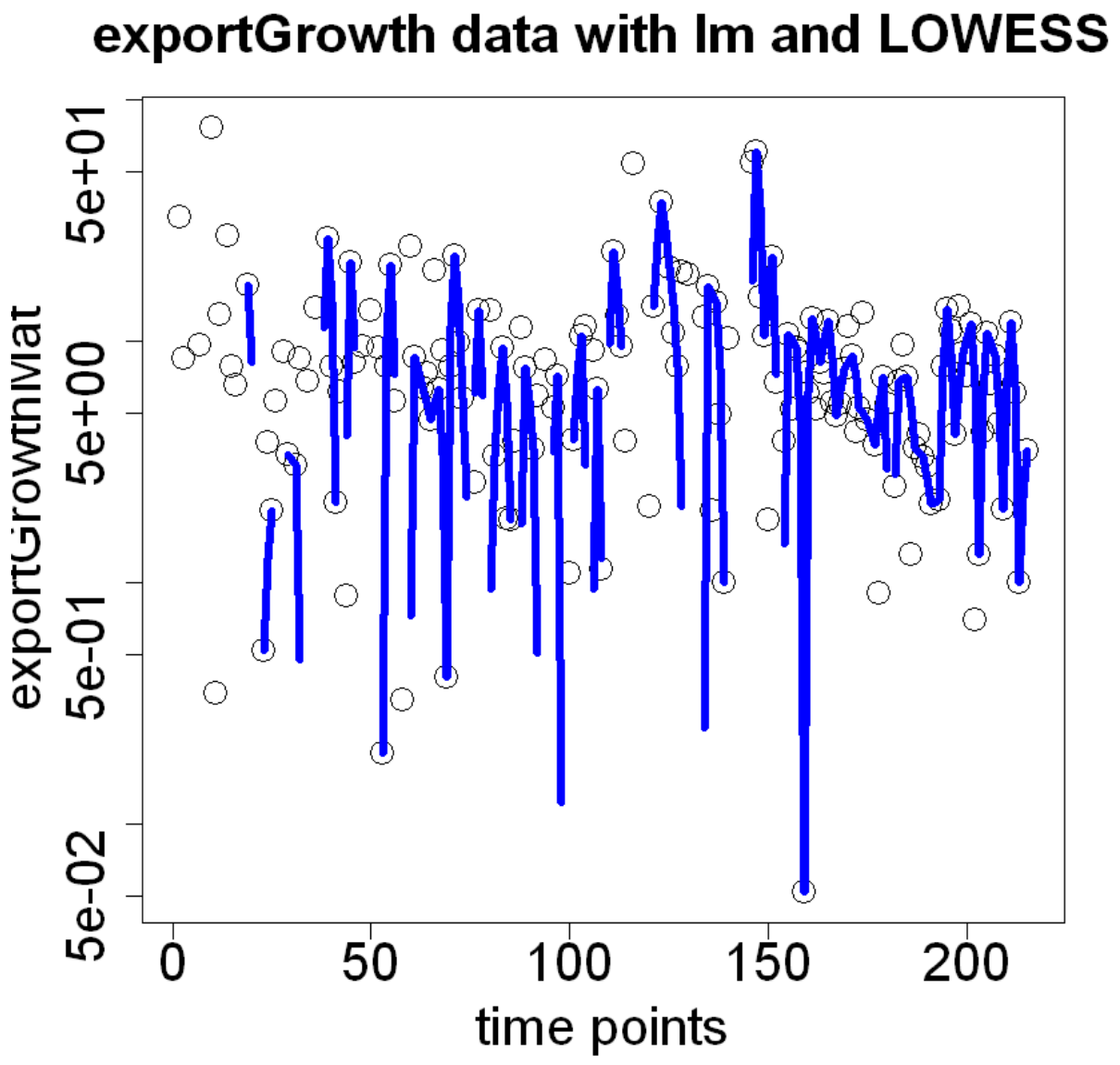
1. Redraw those plots with a logarithmic scale for the y-axis values

In [122]:
```
#apply log scale for the y axis
plot(NexportMat, main="Nexport data with lm and LOWESS",
    xlab="time points", cex=2,cex.lab=2, cex.axis=2, cex.main=2, log="y")
lmM1=abline(lm(NexportMat ~x), col="yellow", lwd=5)
loessM1=lines(lowess(NexportMat), col="blue", lwd=5)
```



In [123]:
```
#apply log scale for the y axis
plot(exportGrowthMat, main="exportGrowth data with lm and LOWESS",
    xlab="time points", cex=2,cex.lab=2, cex.axis=2, cex.main=2, log="y")
lmM1=abline(lm(exportGrowthMat ~x), col="yellow", lwd=5)
loessM1=lines(lowess(exportGrowthMat), col="blue", lwd=5)
```

```
Warning message in xy.coords(x, y, xlabel, ylabel, log):
"60 y values <= 0 omitted from logarithmic plot"
```



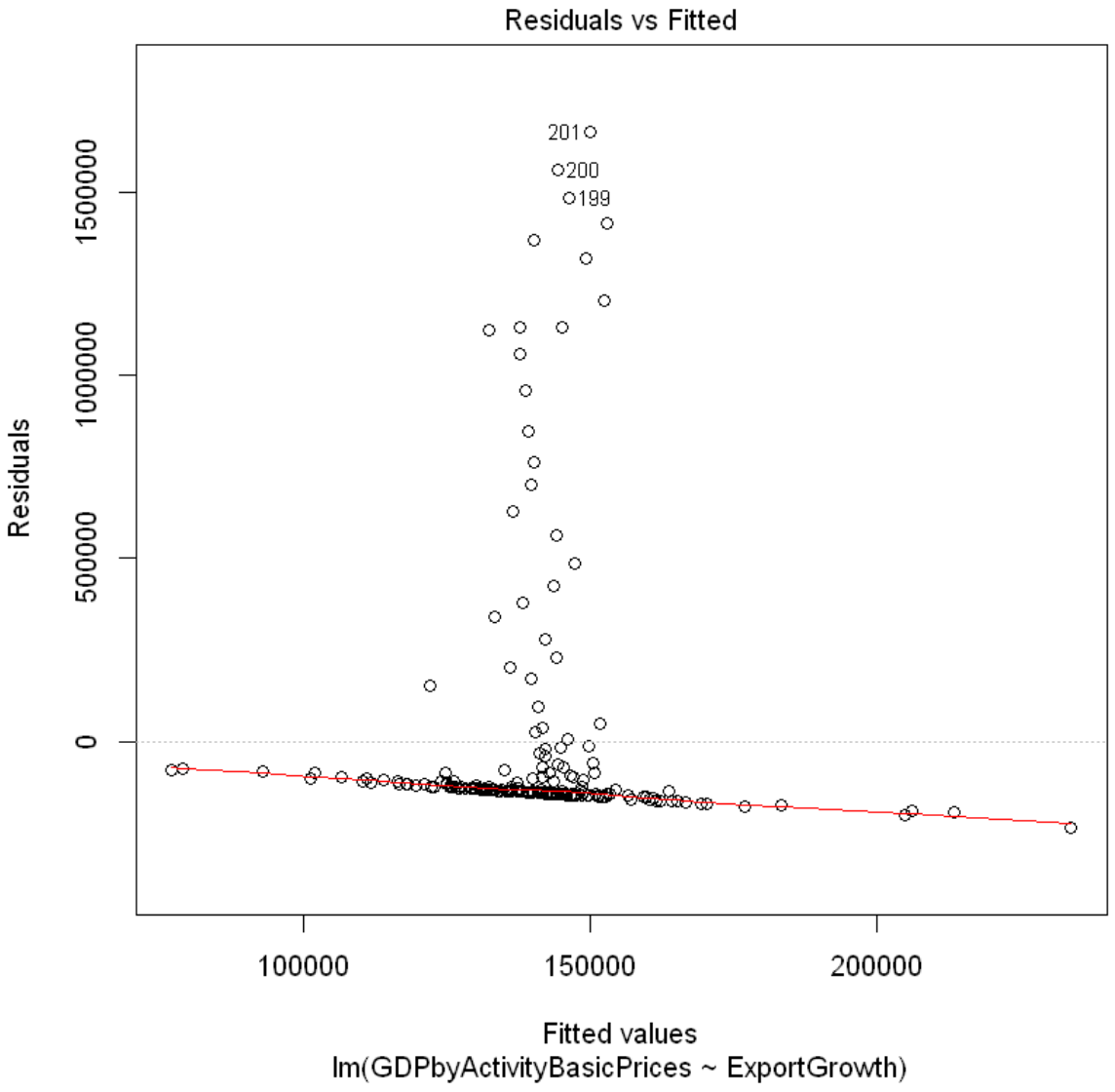

In [ ]:
```
```

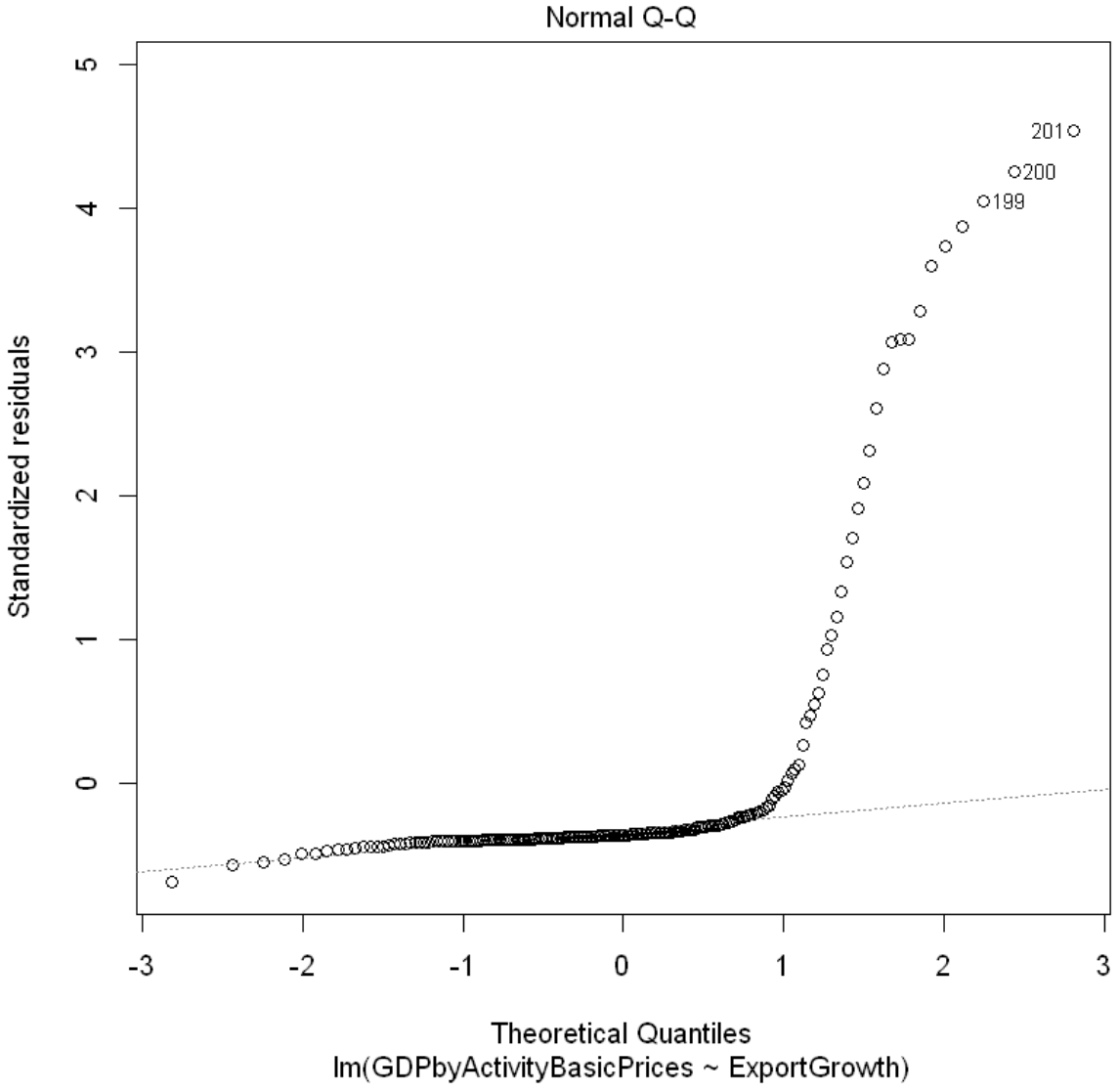
1. Build a simple linear regression model and a polynomial regression model with powers up to the value of 3 to fit the prediction where the 'dependent is GDPbyActivityBasicPrices' and 'independent is ExportGrowth'. Plot the fits and the original data (3 lines).

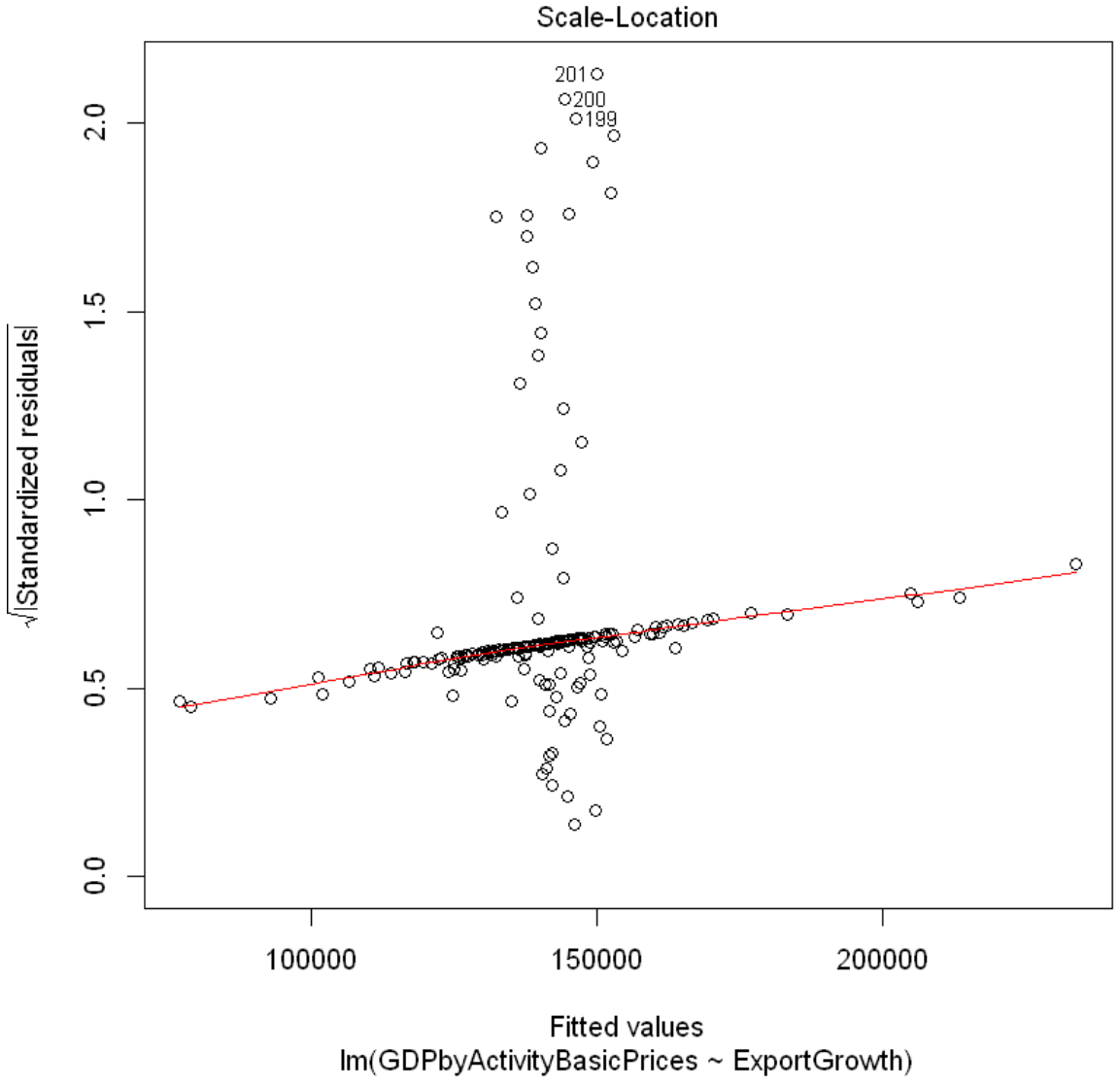
In [124]:
```
#linear regression model
model<-lm(GDPbyActivityBasicPrices ~ ExportGrowth, data=SwedenEcon)
model
```

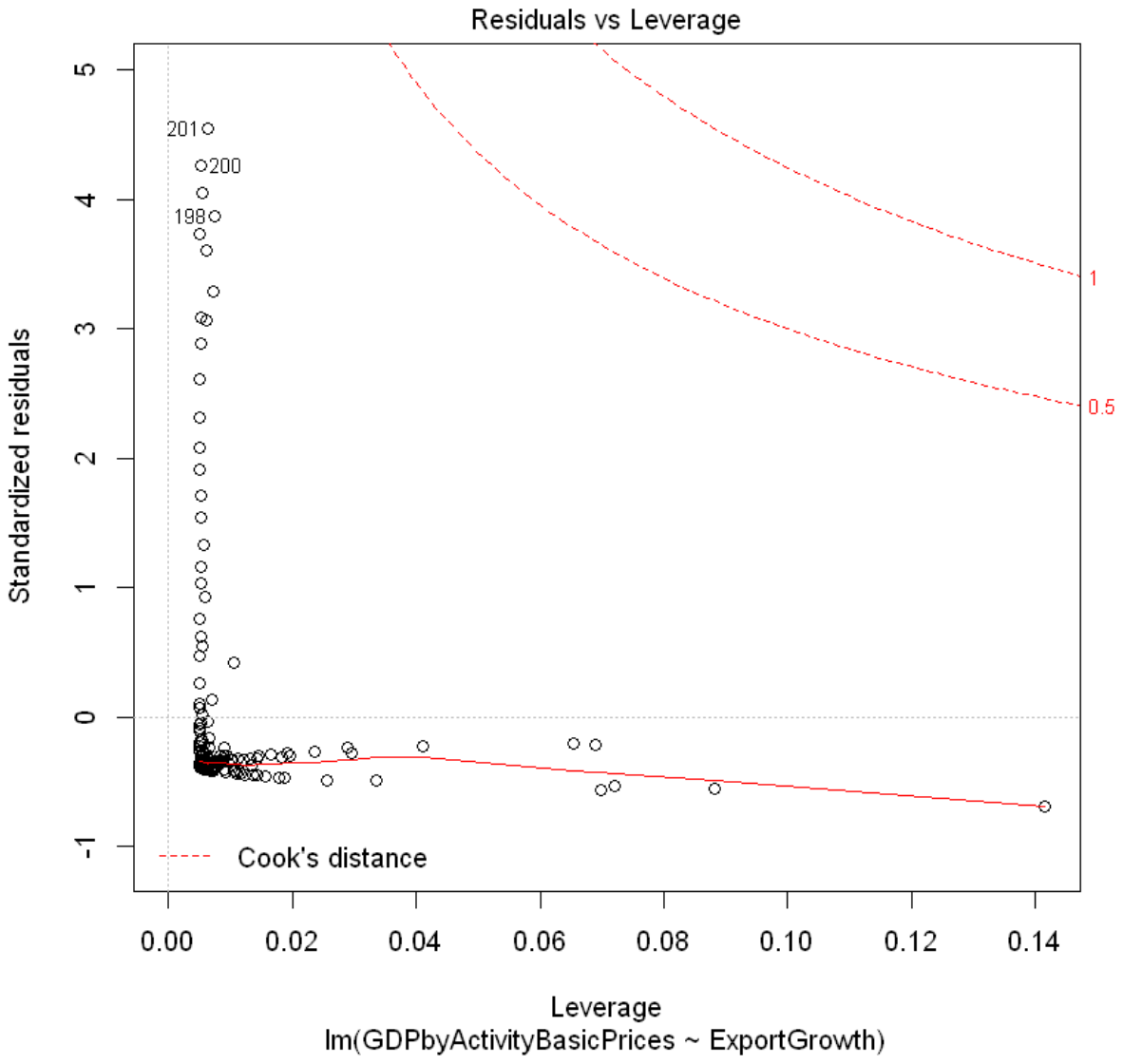
```
Call:
lm(formula = GDPbyActivityBasicPrices ~ ExportGrowth, data = SwedenEcon)

Coefficients:
 (Intercept)  ExportGrowth
      134888          1295
```

In [125]:
```
plot(model)
```


Residuals vs Fitted
201
200
199
Residuals
0
500000
1000000
1500000
100000
150000
200000
Fitted values
lm(GDPbyActivityBasicPrices ~ ExportGrowth)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(GDPbyActivityBasicPrices ~ ExportGrowth)
201
200
199
-3
-2
-1
0
1
2
3
0
1
2
3
4
5


Scale-Location
201
200
199
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100000
150000
200000
Fitted values
lm(GDPbyActivityBasicPrices ~ ExportGrowth)




In [126]:
```
summary(model)
```

```
Call:
lm(formula = GDPbyActivityBasicPrices ~ ExportGrowth, data = SwedenEcon)

Residuals:
    Min      1Q  Median      3Q     Max 
-233838 -143299 -131413  -96448 1663864 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    134888      27366   4.929 1.74e-06 ***
ExportGrowth     1295       1886   0.686    0.493    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 367400 on 198 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.002374,  Adjusted R-squared:  -0.002665 
F-statistic: 0.4711 on 1 and 198 DF,  p-value: 0.4933
```

**my poly function doesn't work, i have to use alternative way**

In [127]:
```
#my poly function doesn't work, i have to use alternative way
#polynomial regression model with powers up to the value of 3 to fit the prediction
#Plot the fits and the original data (3 lines).
m1<-lm(SwedenEcon$GDPbyActivityBasicPrices ~ SwedenEcon$ExportGrowth)
m2<-lm(SwedenEcon$GDPbyActivityBasicPrices ~ SwedenEcon$ExportGrowth + I(SwedenEcon$ExportGrowth^2))
m3<-lm(SwedenEcon$GDPbyActivityBasicPrices ~ SwedenEcon$ExportGrowth + I(SwedenEcon$ExportGrowth^2)+
       I(SwedenEcon$ExportGrowth^3))
```

```
Your code contains a unicode char which cannot be displayed in your
current locale and R will silently convert it to an escaped form when the
R kernel executes this code. This can lead to subtle errors if you use
such chars to do comparisons. For more information, please see
https://github.com/IRkernel/repr/wiki/Problems-with-unicode-on-windows
```

In [128]:
```
#mean squred errors (MSE)
print(mean(summary(m1$residuals^2)))
print(mean(summary(m2$residuals^2)))
print(mean(summary(m3$residuals^2)))
```

```
[1] 493035291072
[1] 482967339938
[1] 481784155686
```

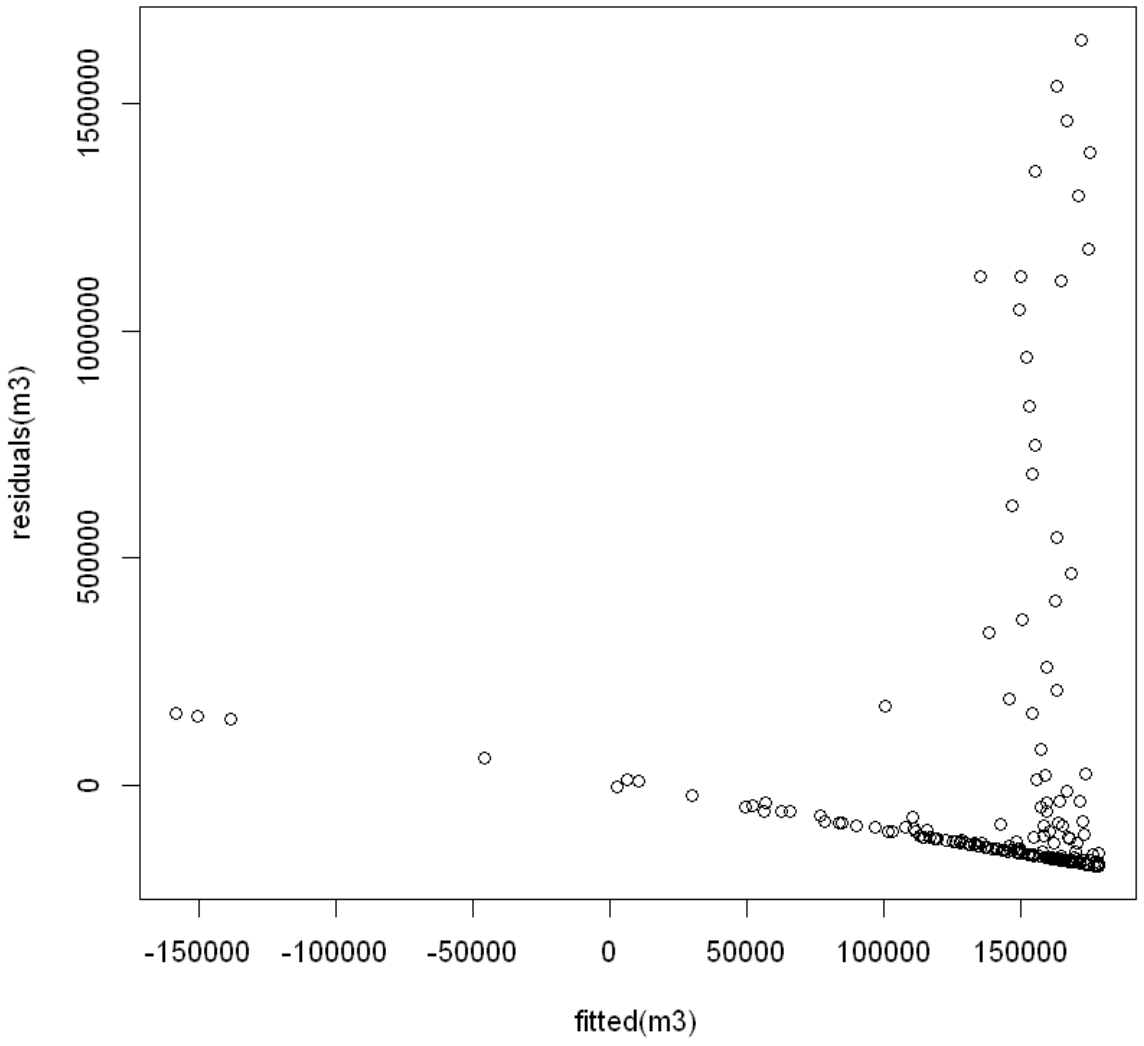
In [129]:
```
summary(m3)
```

```
Call:
lm(formula = SwedenEcon$GDPbyActivityBasicPrices ~ SwedenEcon$ExportGrowth +
    I(SwedenEcon$ExportGrowth^2) + I(SwedenEcon$ExportGrowth^3))

Residuals:
    Min      1Q  Median      3Q     Max
-178506 -161250 -134540  -65896 1641571

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  1.427e+05  3.062e+04   4.662 5.77e-06 ***
SwedenEcon$ExportGrowth      3.487e+03  2.757e+03   1.265    0.207
I(SwedenEcon$ExportGrowth^2) -7.922e+01  7.576e+01  -1.046    0.297
I(SwedenEcon$ExportGrowth^3) -2.314e-01  1.435e+00  -0.161    0.872
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 366200 on 196 degrees of freedom
  (16 observations deleted due to missingness)
Multiple R-squared:  0.01912,   Adjusted R-squared:  0.004108
F-statistic: 1.274 on 3 and 196 DF,  p-value: 0.2846
```

In [130]:
```
confint(m3, level=0.95)
```

| | 2.5 % | 97.5 % |
|---|---|---|
| **(Intercept)** | 82365.950513 | 2.031314e+05 |
| **SwedenEcon$ExportGrowth** | -1949.393345 | 8.923091e+03 |
| **I(SwedenEcon$ExportGrowth^2)** | -228.633926 | 7.018710e+01 |
| **I(SwedenEcon$ExportGrowth^3)** | -3.061038 | 2.598144e+00 |

In [131]:
```
plot(fitted(m3), residuals(m3))
```



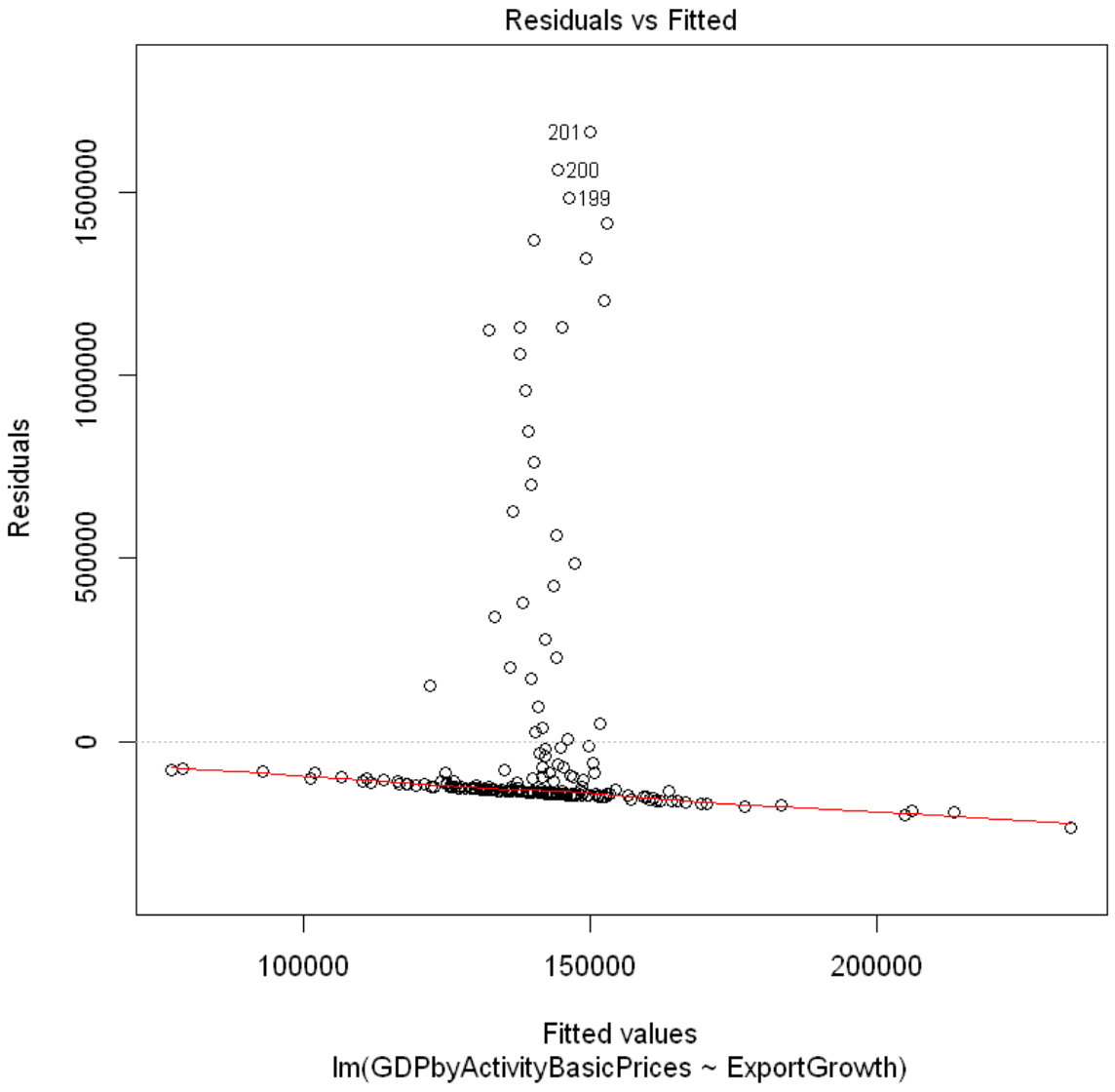

In [132]:
```
#use a few statement to fit the linear model and use the model to predict some values at points we have seen and no
t seen
m1<-lm(GDPbyActivityBasicPrices~ExportGrowth, data=SwedenEcon)
new.df<-data.frame(ExportGrowth=c(10.5))
predict(m1, new.df)
```

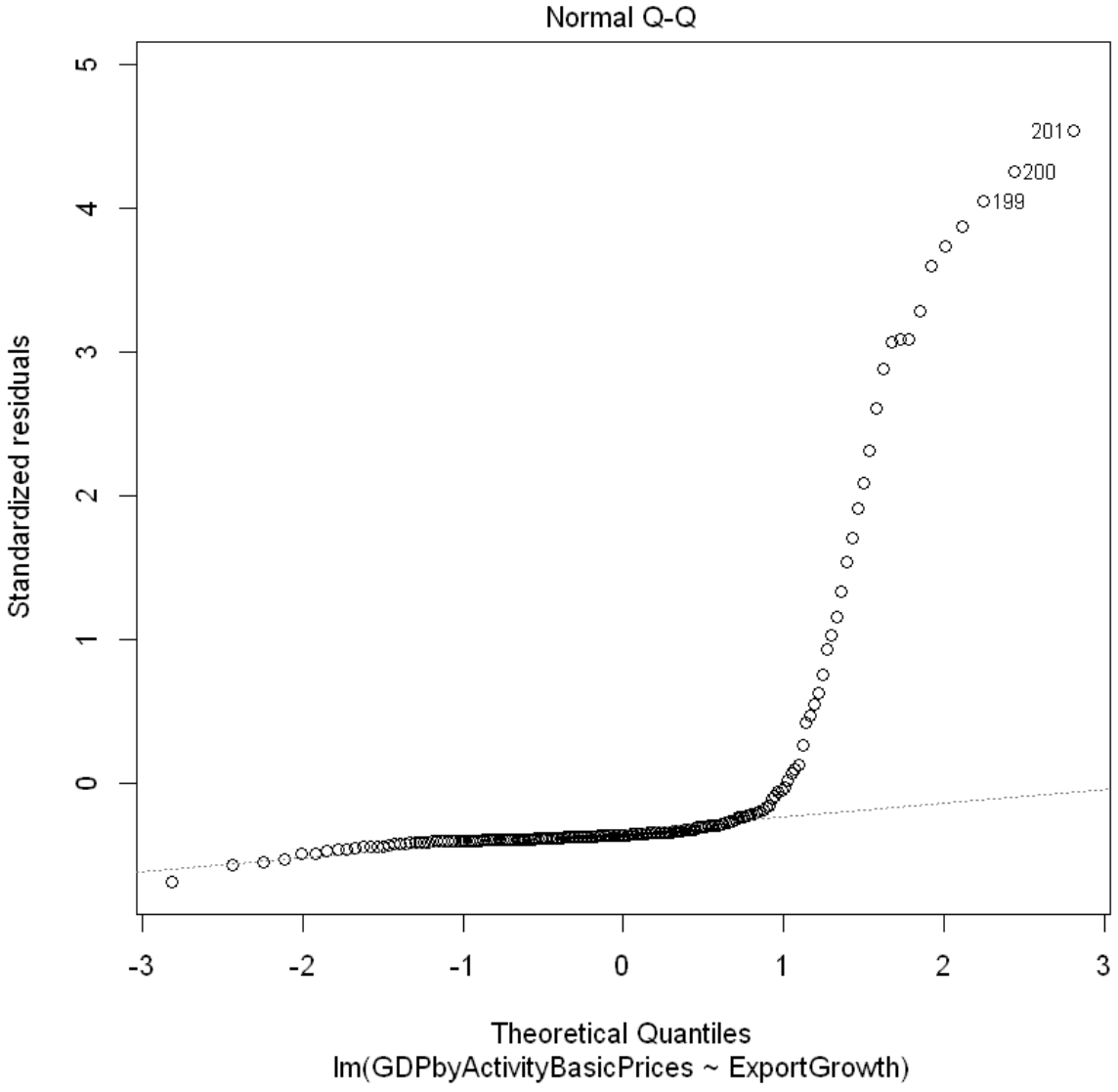
**1:** 148482.67500375

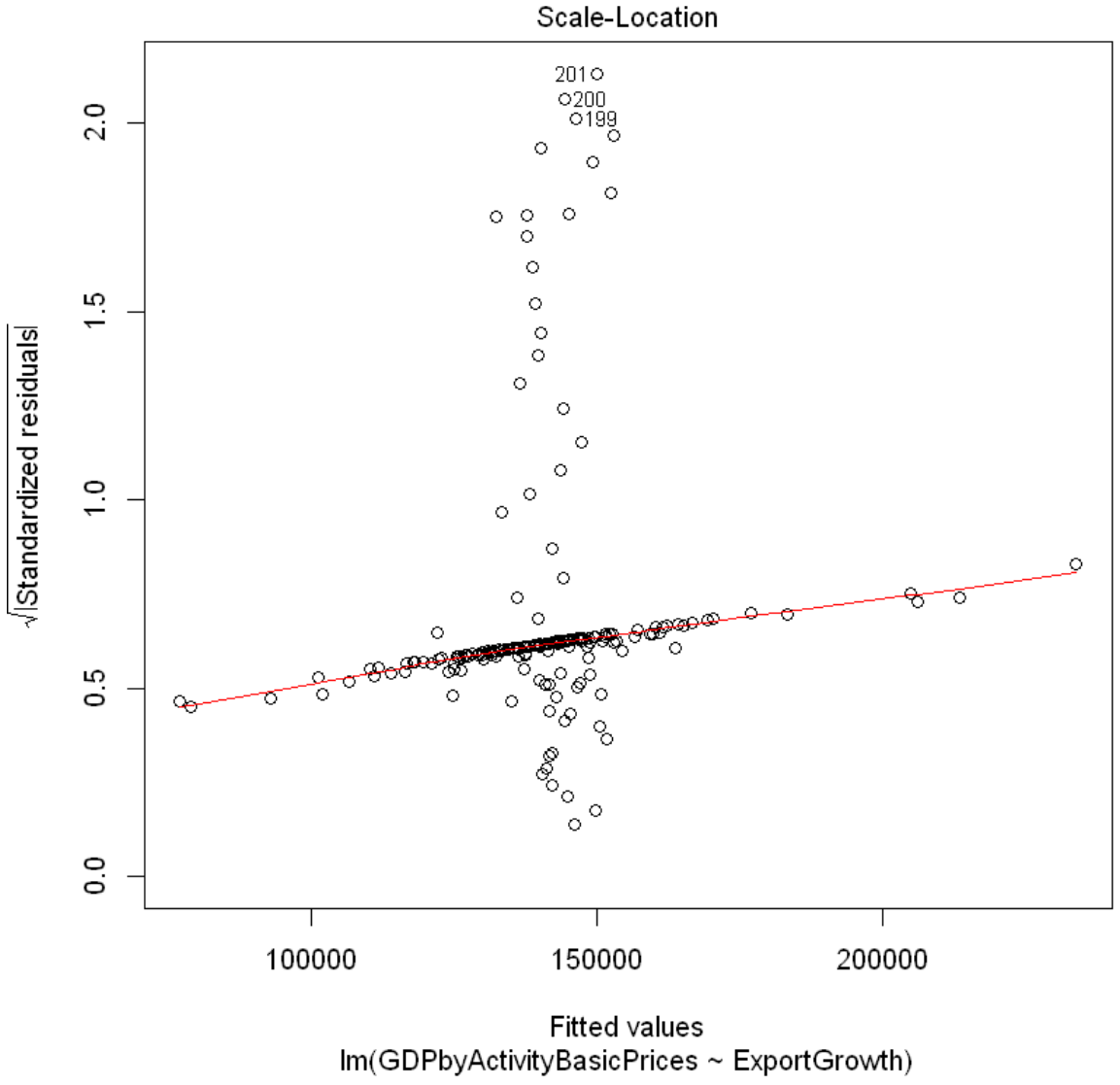
In [133]:
```
#let's look at the prediction interval around a point that we specify
m1<-lm(GDPbyActivityBasicPrices~ExportGrowth, data=SwedenEcon)
new.df<-data.frame(ExportGrowth=c(10.5))
predict(m1, new.df, interval="prediction")
```

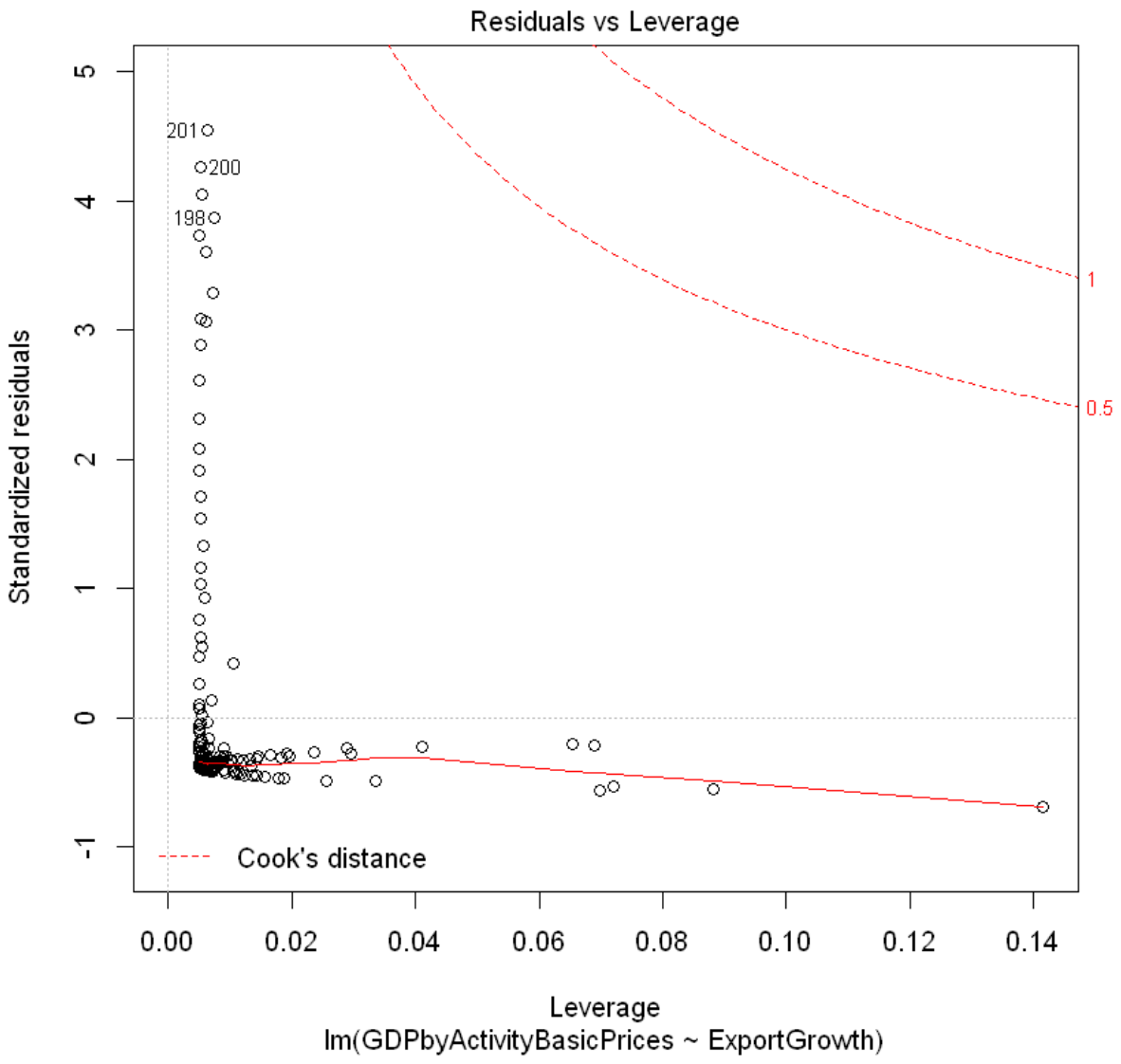
| fit | lwr | upr |
|---|---|---|
| 148482.7 | -578273.7 | 875239 |

In [134]:
```
plot(m1)
```


Residuals vs Fitted
201
200
199
Residuals
0
500000
1000000
1500000
100000
150000
200000
Fitted values
lm(GDPbyActivityBasicPrices ~ ExportGrowth)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(GDPbyActivityBasicPrices ~ ExportGrowth)
201
200
199
-3
-2
-1
0
1
2
3
0
1
2
3
4
5


Scale-Location
201
200
199
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
100000
150000
200000
Fitted values
lm(GDPbyActivityBasicPrices ~ ExportGrowth)



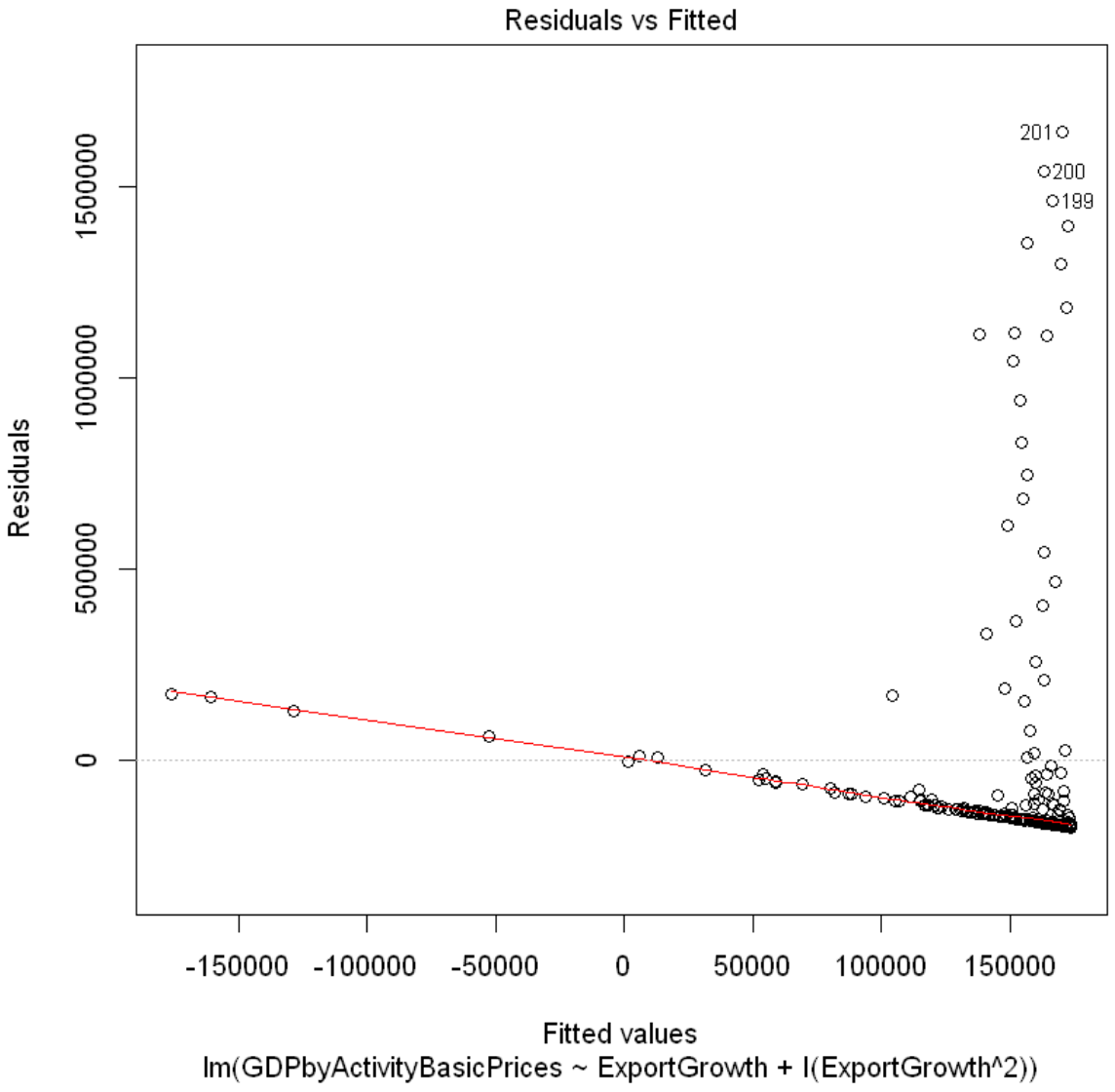

these code do not work for me plot(x=1:20, y=GDPbyActivityBasicPrices, col="blue", cex=2, cex.axis=1.5, cex.lab=2) new.df<-data.frame(ExportGrowth=c(1:20)) ynew<-predict(m1, new.df) points(ynew, col="red", cex=2)

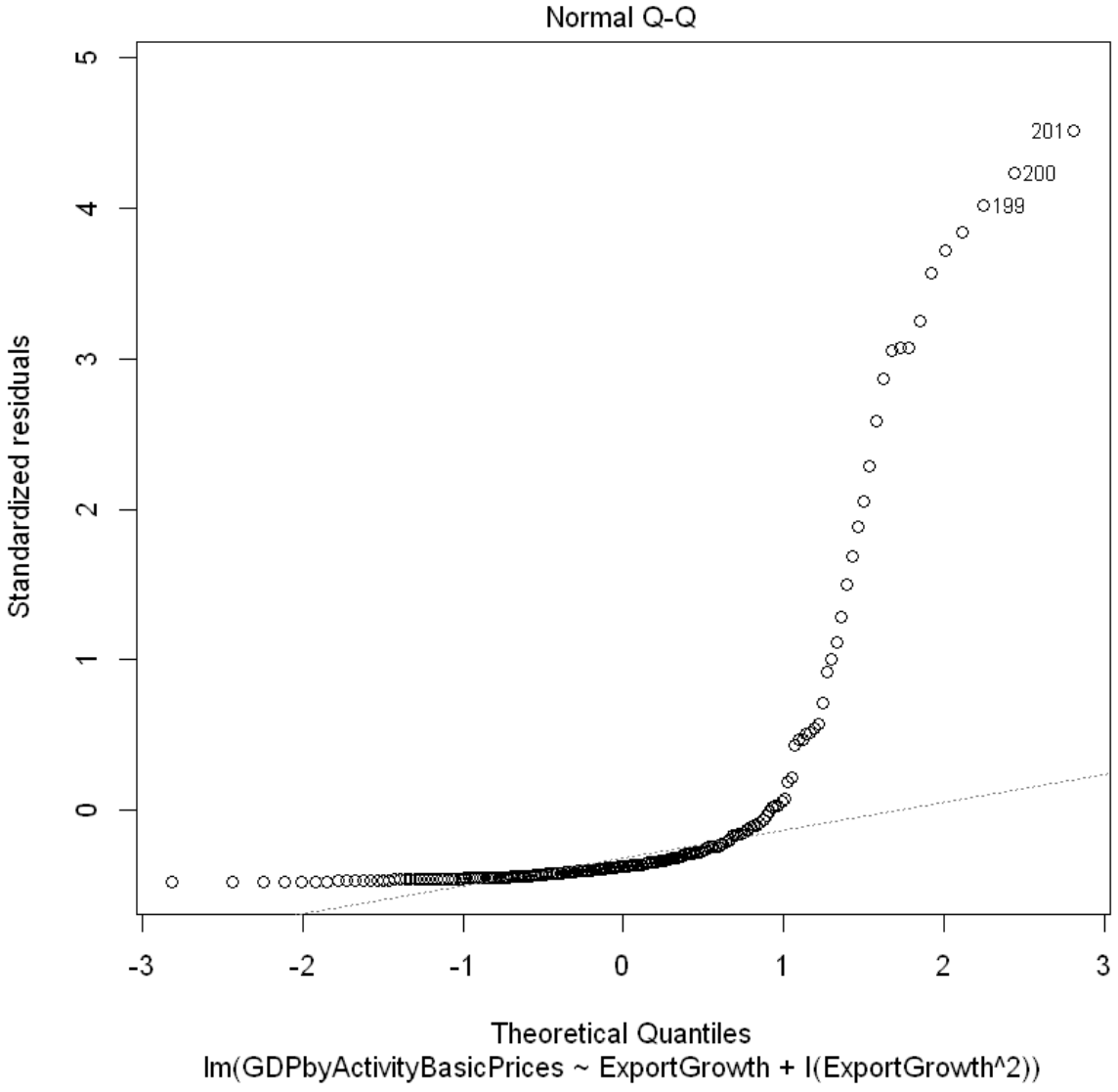
In [135]:
```
#Let's look at the prediction interval around a point that we specify
m2<-lm(GDPbyActivityBasicPrices~ExportGrowth+I(ExportGrowth^2), data=SwedenEcon)
new.df<-data.frame(ExportGrowth=c(10.5))
predict(m2, new.df, interval="prediction")
```

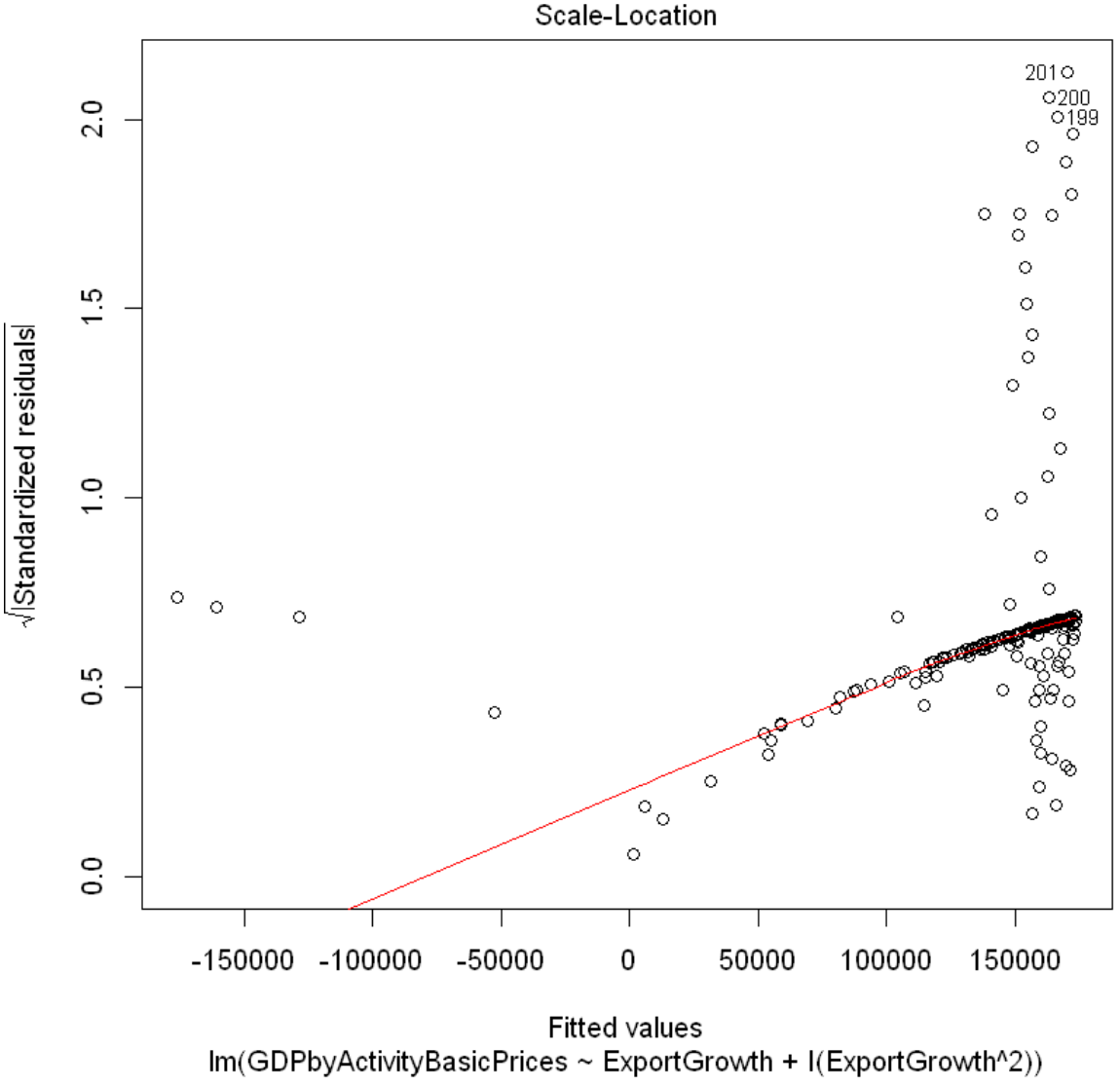
| fit | lwr | upr |
|---|---|---|
| 168744.4 | -554114.3 | 891603 |

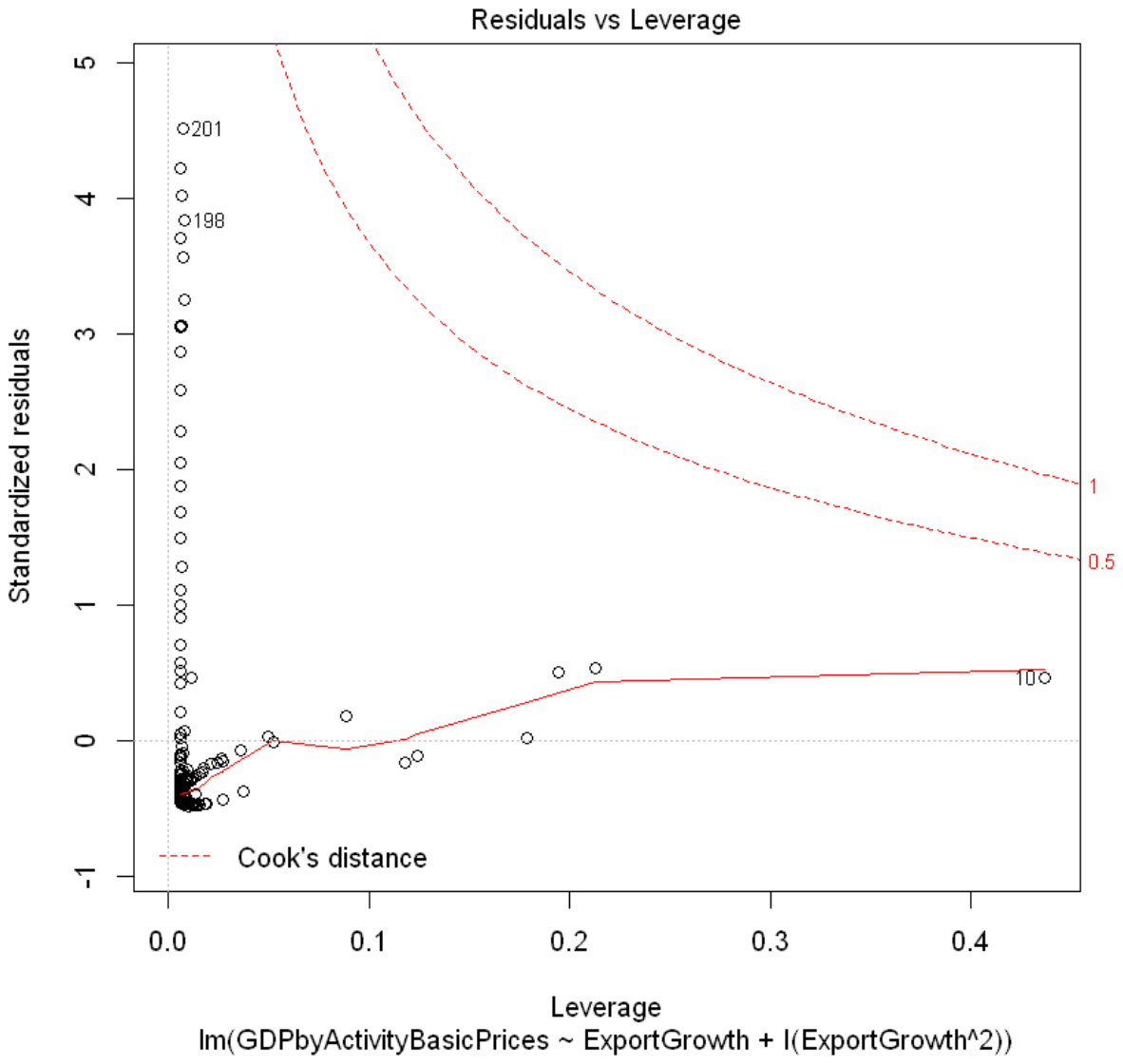
these code do not work for me: plot(x=1:20, y=GDPbyActivityBasicPrices, col="blue", cex=2, cex.axis=1.5, cex.lab=2) new.df<-data.frame(ExportGrowth=c(1:20)) ynew<-predict(m2, new.df) points(ynew, col="red", cex=4)

In [136]:
```
plot(m2)
```


Residuals vs Fitted
Residuals
1500000
1000000
500000
0
201
200
199
-150000
-100000
-50000
0
50000
100000
150000
Fitted values
lm(GDPbyActivityBasicPrices ~ ExportGrowth + I(ExportGrowth^2))


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(GDPbyActivityBasicPrices ~ ExportGrowth + I(ExportGrowth^2))
201
200
199
-3
-2
-1
0
1
2
3
0
1
2
3
4
5


Scale-Location
201
200
199
2.0
1.5
1.0
0.5
0.0
√|Standardized residuals|
-150000
-100000
-50000
0
50000
100000
150000
Fitted values
lm(GDPbyActivityBasicPrices ~ ExportGrowth + I(ExportGrowth^2))

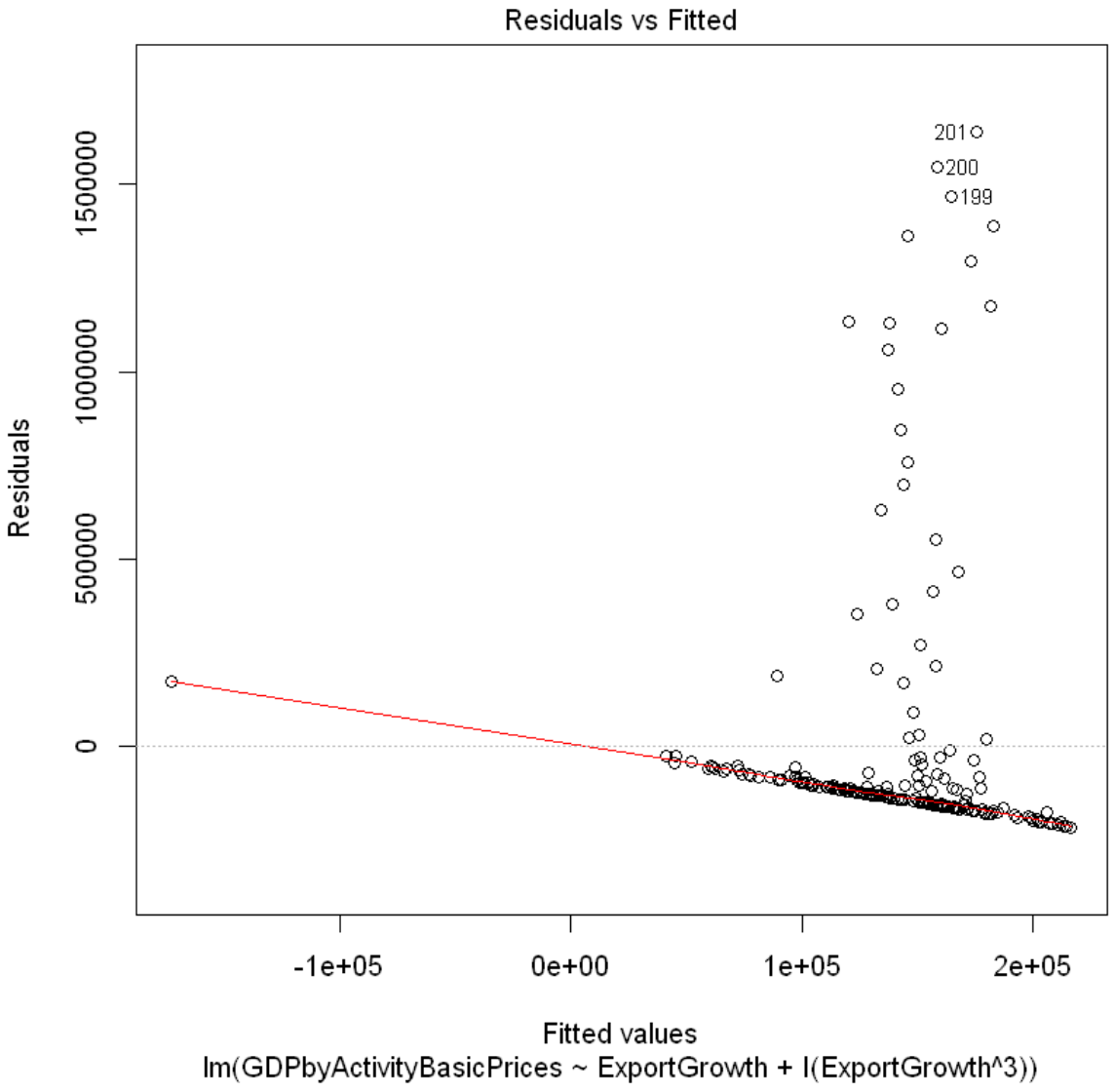


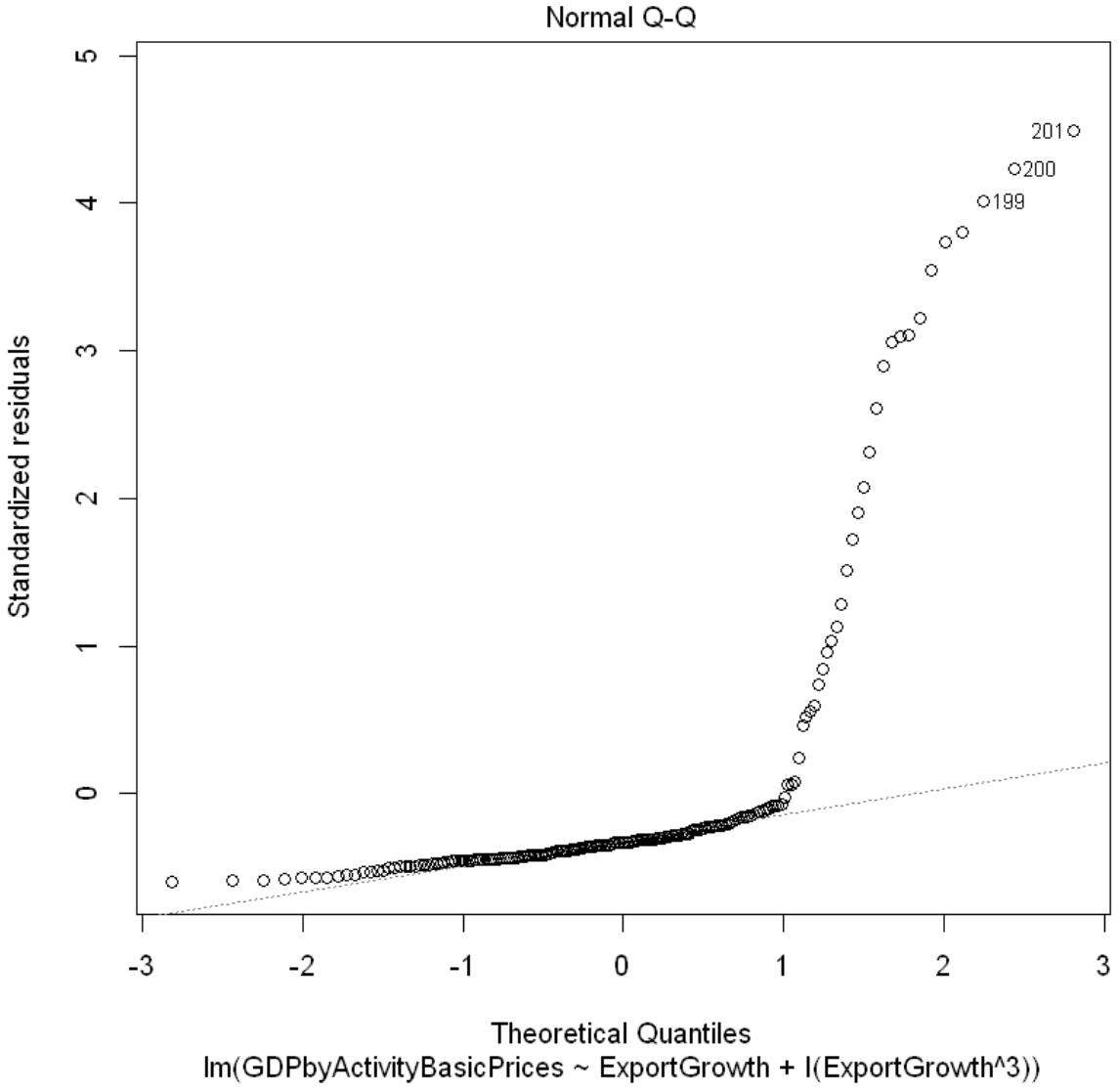

In [137]:
```
#Let's look at the prediction interval around a point that we specify
m3<-lm(GDPbyActivityBasicPrices~ExportGrowth+I(ExportGrowth^3), data=SwedenEcon)
new.df<-data.frame(ExportGrowth=c(10, 10.5))
predict(m3, new.df, interval="prediction")
```

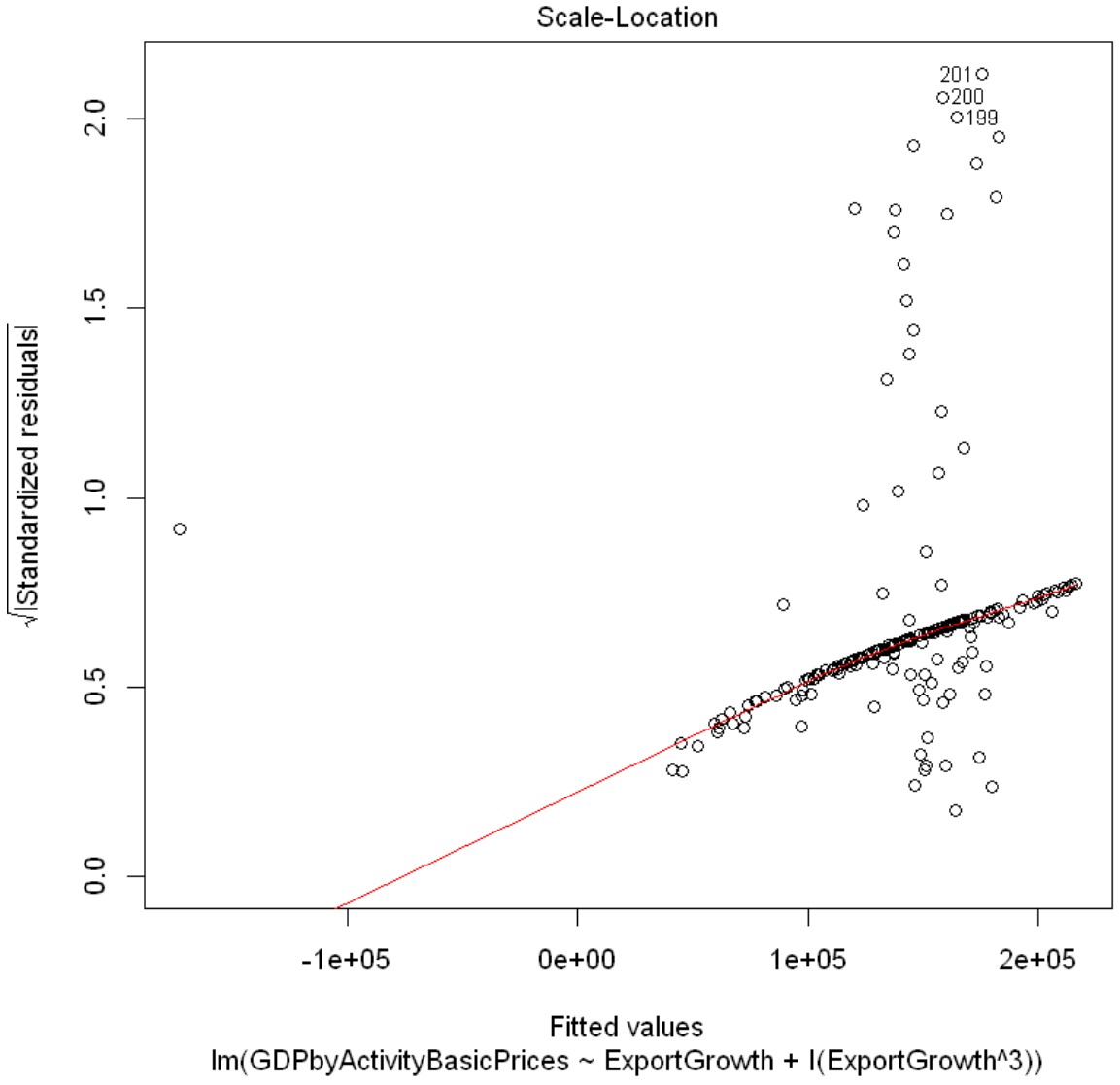
| fit | lwr | upr |
|---|---|---|
| 169057.0 | -555917.8 | 894031.7 |
| 170919.9 | -554172.1 | 896011.9 |

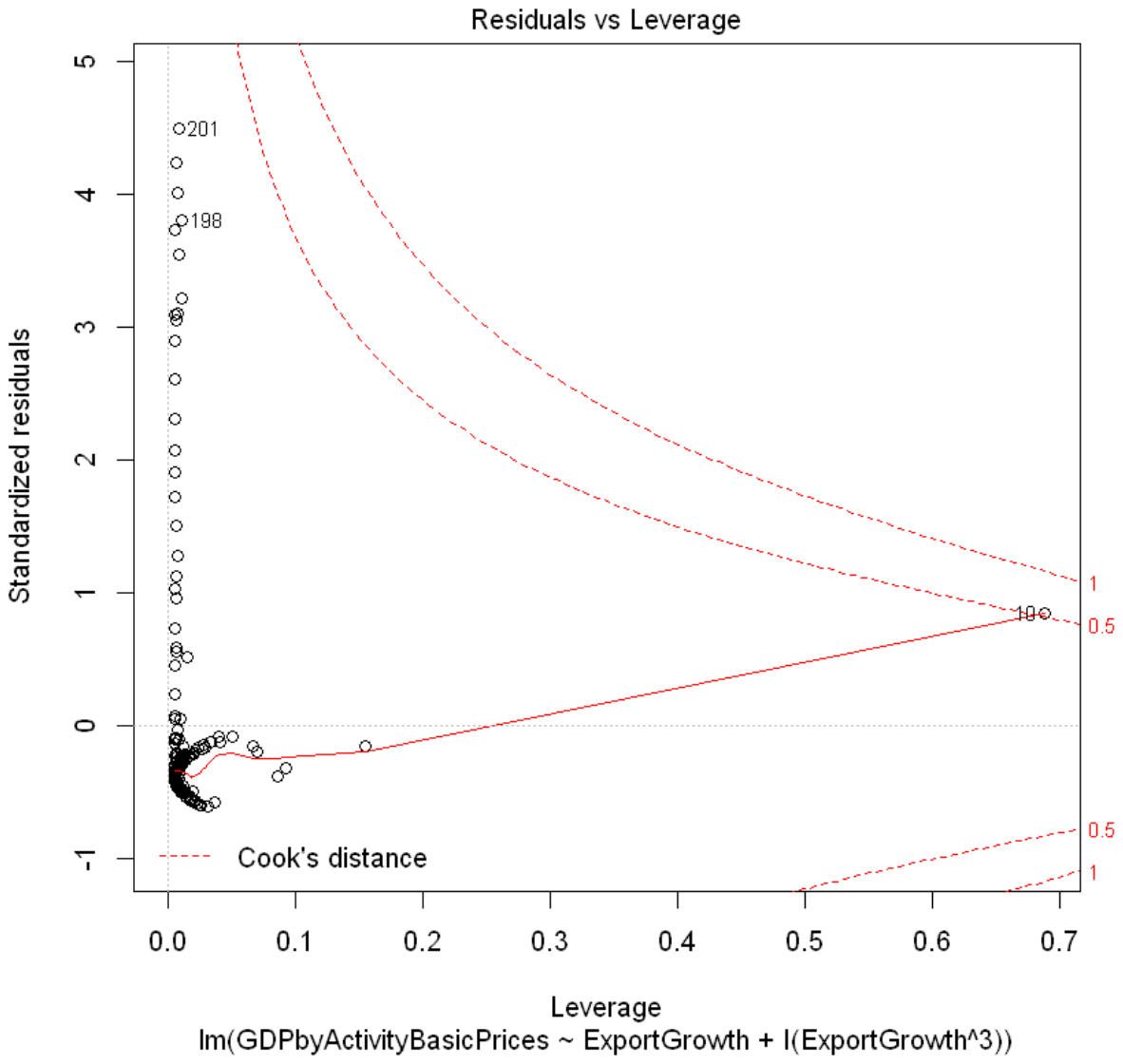
these code do not work for me: plot(x=1:20, y=GDPbyActivityBasicPrices, col="blue", cex=2, cex.axis=1.5, cex.lab=2) new.df<-data.frame(ExportGrowth=c(1:20)) ynew<-predict(m3, new.df) points(ynew, col="red", cex=4)

In [138]:
```
plot(m3)
```


Residuals vs Fitted
Residuals
0
500000
1000000
1500000
201
200
199
-1e+05
0e+00
1e+05
2e+05
Fitted values
lm(GDPbyActivityBasicPrices ~ ExportGrowth + I(ExportGrowth^3))


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(GDPbyActivityBasicPrices ~ ExportGrowth + I(ExportGrowth^3))
201
200
199
-3
-2
-1
0
1
2
3
0
1
2
3
4
5


Scale-Location
201
200
199
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
-1e+05
0e+00
1e+05
2e+05
Fitted values
lm(GDPbyActivityBasicPrices ~ ExportGrowth + I(ExportGrowth^3))




In [ ]:

1. Compare the 2 models using the F-test (anova(model1,model2,test="F")) and report on the decision for the choice of model.

In [139]:
```
anova(m1, m2, test="F")
```

| Res.Df | RSS | Df | Sum of Sq | F | Pr(>F) |
|---|---|---|---|---|---|
| 198 | 2.673331e+13 | NA | NA | NA | NA |
| 197 | 2.628802e+13 | 1 | 445292116497 | 3.336978 | 0.06925228 |

we can see that the 'Df' is that there is a single additional parameter and the p value is bigger than 0.05 since there was no change in a perfect fit

In [ ]:

1. Use the AIC function for the 2 models and report which model you choose on the outputs it provides.

In [140]:
```
AIC(m1, m2)
```

| | df | AIC |
|---|---|---|
| **m1** | 3 | 5697.298 |
| **m2** | 4 | 5695.939 |

I will choose m2 because the model has smaller AIC

In [141]:
```
#########################################################################################################################
#####
```

Question 2 (10 points) You will use the dataset ”UN98.csv” for this question:

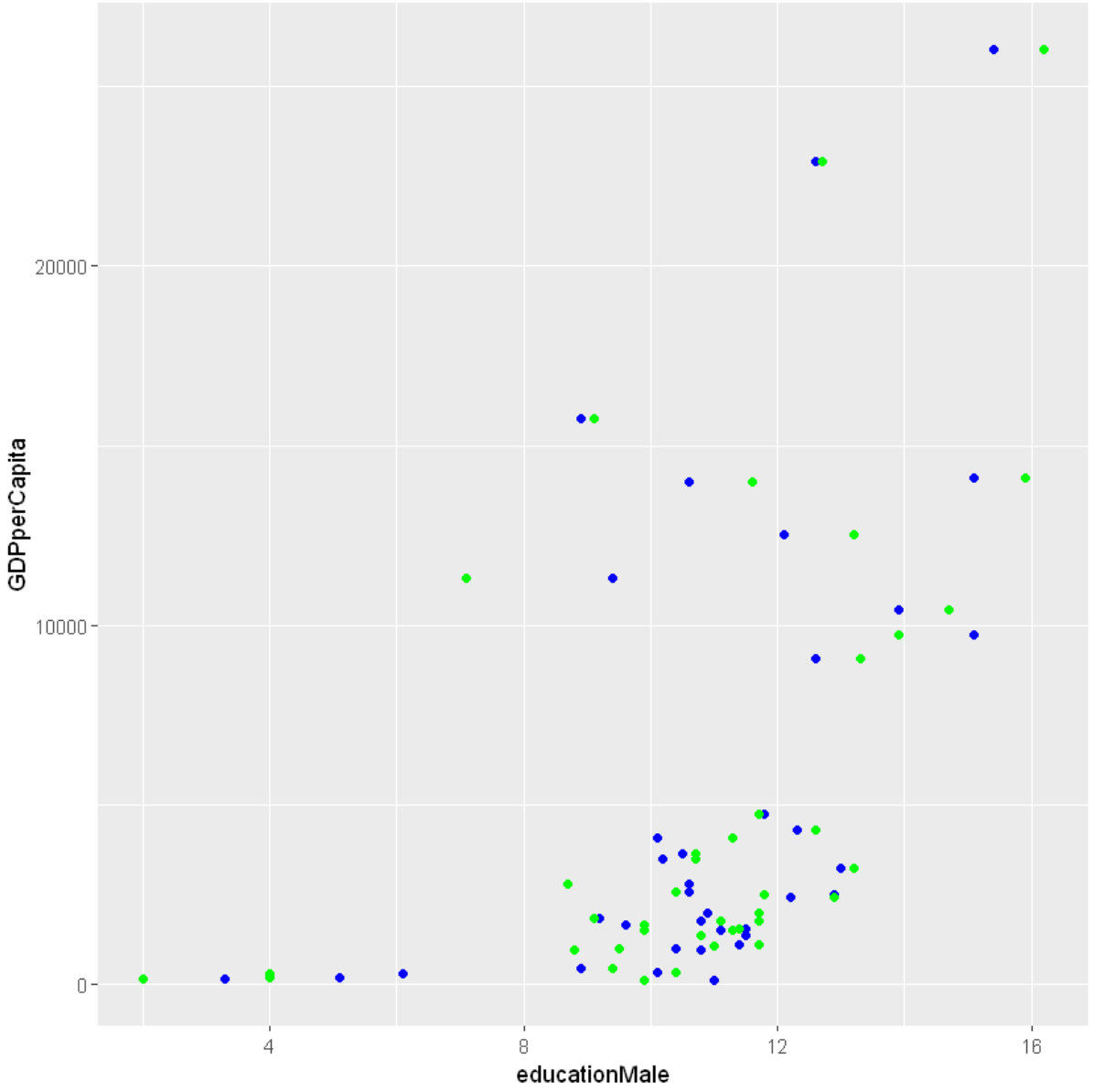
1. In a 2D scatterplot display the points for (educationMale,GDPperCapita) in the color blue, and add to that same plot the points for (educationFemale,GDPperCapita) in green.
2. In the dataset check (validate) that the column ’region’ is a Factor and if it is not a factor make it so.
3. Use the library, ’nnet’, ( library(nnet), install.packages(”nnet”) ) to use the function ’multinom’ to do multinomial logistic regression where the dependent is the ’region’, and the independents are ’infantMortality’ and ’GDPperCapita’. What can you conclude from the model production?
4. Can you use the ’predict’ function with your model on a new data point where ’infantMortality = 10’ and ’GDPperCapita = 20000’ and report on the output? This same function was not shown for multinom’s use but it was shown for binary logitic regression and for decision trees etc. If it cannot be used say why or else apply it as well to report on the results.
5. Build a decision tree where the dependent is the ’region’, and the independents are ’infantMortality’ and ’GDPperCapita’. Display the tree.
6. Can you use the ’predict’ function with your decision tree model on a new data point where ’infantMortality = 10’ and ’GDPperCapita = 20000’ and report on the output?

1. In a 2D scatterplot display the points for (educationMale,GDPperCapita) in the color blue, and add to that same plot the points for (educationFemale,GDPperCapita) in green.

In [142]:
```
UN98<- read.csv('C:/Users/Andrew/Desktop/UN98-1.csv')
str(UN98)
```

```
'data.frame':   207 obs. of  14 variables:
 $ X                    : Factor w/ 207 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ region               : Factor w/ 5 levels "Africa","America",..: 3 4 1 3 4 1 2 2 4 5 ...
 $ tfr                  : num  6.9 2.6 3.81 NA NA 6.69 NA 2.62 1.7 1.89 ...
 $ contraception        : int  NA NA 52 NA NA NA 53 NA 22 76 ...
 $ educationMale        : num  NA NA 11.1 NA NA NA NA NA NA 16.3 ...
 $ educationFemale      : num  NA NA 9.9 NA NA NA NA NA NA 16.1 ...
 $ lifeMale             : num  45 68 67.5 68 NA 44.9 NA 69.6 67.2 75.4 ...
 $ lifeFemale           : num  46 74 70.3 73 NA 48.1 NA 76.8 74 81.2 ...
 $ infantMortality      : int  154 32 44 11 NA 124 24 22 25 6 ...
 $ GDPperCapita         : int  2848 863 1531 NA NA 355 6966 8055 354 20046 ...
 $ economicActivityMale  : num  87.5 NA 76.4 58.8 NA NA 74.4 76.2 65 74 ...
 $ economicActivityFemale: num  7.2 NA 7.8 42.4 NA NA 56.2 41.3 52 53.8 ...
 $ illiteracyMale       : num  52.8 NA 26.1 0.264 NA NA NA 3.8 0.3 NA ...
 $ illiteracyFemale     : num  85 NA 51 0.36 NA NA NA 3.8 0.5 NA ...
```

In [143]:
```
#take off NA
UN98NoNA=na.omit(UN98)
p<-ggplot()+
    geom_point(data=UN98NoNA, aes(x=educationMale, y=GDPperCapita), col="blue")+
    geom_point(data=UN98NoNA, aes(x=educationFemale, y=GDPperCapita), col="green")
print(p)
```




In [ ]:

1. In the dataset check (validate) that the column 'region' is a Factor and if it is not a factor make it so.

In [144]:
```
is.factor(UN98NoNA$region)
```

TRUE

column 'region' is a Factor

In [ ]:

1. Use the library, 'nnet', ( library(nnet), install.packages("nnet") ) to use the function 'multinom' to do multinomial logistic regression where the dependent is the 'region', and the independents are 'infantMortality' and 'GDPperCapita'. What can you conclude from the model production?

In [145]:
```
install.packages("nnet")
library(nnet)
```

```
Warning message:
"package 'nnet' is in use and will not be installed"
```

In [146]:
```
levels(UN98NoNA$region)
contrasts(UN98NoNA$region)
```

'Africa' 'America' 'Asia' 'Europe' 'Oceania'

| | America | Asia | Europe | Oceania |
|---|---|---|---|---|
| **Africa** | 0 | 0 | 0 | 0 |
| **America** | 1 | 0 | 0 | 0 |
| **Asia** | 0 | 1 | 0 | 0 |
| **Europe** | 0 | 0 | 1 | 0 |
| **Oceania** | 0 | 0 | 0 | 1 |

In [147]:
```
#built multi nomial regression
model<-multinom(region ~ infantMortality + GDPperCapita, family=binomial(link='logit'), data=UN98NoNA)
summary(model)
coef(model)
```

```
# weights:  20 (12 variable)
initial  value 62.768079
iter  10 value 43.455051
iter  20 value 35.327942
iter  30 value 35.102365
iter  40 value 35.081895
final  value 35.081616
converged

Call:
multinom(formula = region ~ infantMortality + GDPperCapita, data = UN98NoNA,
    family = binomial(link = "logit"))

Coefficients:
        (Intercept) infantMortality  GDPperCapita
America   8.4005450      -0.2114168  0.0007464656
Asia      5.7103749      -0.1589191  0.0008987694
Europe   10.5592752      -0.3494420  0.0006553723
Oceania   0.5501728      -0.0331512 -0.0004236457

Std. Errors:
         (Intercept) infantMortality GDPperCapita
America 0.0006918875      0.02409275 0.0003169448
Asia    0.0006992323      0.02523785 0.0003164302
Europe  0.0019344713      0.04442970 0.0003227298
Oceania 0.0002508197      0.02256019 0.0010608704

Residual Deviance: 70.16323
AIC: 94.16323
```

| | (Intercept) | infantMortality | GDPperCapita |
|---|---|---|---|
| **America** | 8.4005450 | -0.2114168 | 0.0007464656 |
| **Asia** | 5.7103749 | -0.1589191 | 0.0008987694 |
| **Europe** | 10.5592752 | -0.3494420 | 0.0006553723 |
| **Oceania** | 0.5501728 | -0.0331512 | -0.0004236457 |

From the summary of the model, From the coeficient result, we can see infantMortality and region is negative relationship, increase on region and infantMortality decrease. Africa is 0 and other region is 1, it means other regions has lower infant Mortality.

In [ ]:
```
```

1. Can you use the 'predict' function with your model on a new data point where 'infantMortality = 10' and 'GDPperCapita = 20000' and report on the output? This same function was not shown for multinom's use but it was shown for binary logitic regression and for decision trees etc. If it cannot be used say why or else apply it as well to report on the results.

In [148]:
```
new<-data.frame(infantMortality = c(10), GDPperCapita = c(20000))
print(new)
print(predict(model, new))
```

```
  infantMortality GDPperCapita
1              10        20000
[1] Asia
Levels: Africa America Asia Europe Oceania
```

Yes I can predict region under the condition of 'infantMortality = 10' and 'GDPperCapita = 20000'. It predict in Asia.

In [ ]:
```
```

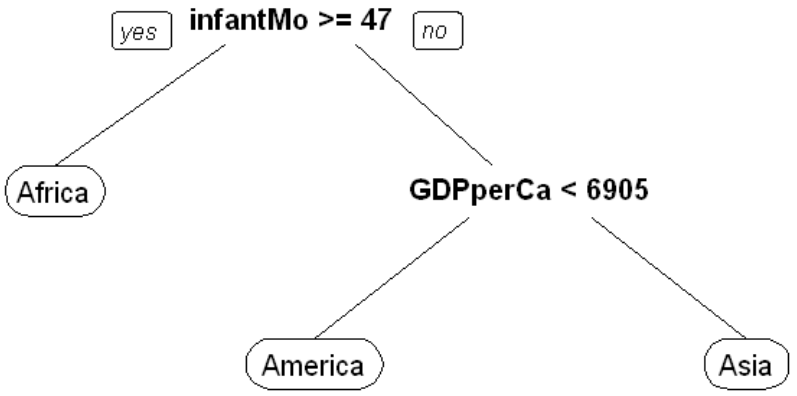
1. Build a decision tree where the dependent is the 'region', and the independents are 'infantMortality' and 'GDPperCapita'. Display the tree.

In [149]:
```
#build decision tree
install.packages("rpart")
install.packages("rpart.plot")
library(rpart)
library(rpart.plot)
```

```
Warning message:
"package 'rpart' is in use and will not be installed"Warning message:
"package 'rpart.plot' is in use and will not be installed"
```

In [150]:
```
regionTree=rpart(region ~ infantMortality + GDPperCapita, data=UN98NoNA)
#draw the tree model
prp(regionTree)
regionTree
```

```
n= 39

node), split, n, loss, yval, (yprob)
      * denotes terminal node

1) root 39 26 America (0.23 0.33 0.26 0.15 0.026)
  2) infantMortality>=46.5 11  3 Africa (0.73 0 0.18 0 0.091) *
  3) infantMortality< 46.5 28 15 America (0.036 0.46 0.29 0.21 0)
    6) GDPperCapita< 6904.5 19  8 America (0.053 0.58 0.16 0.21 0) *
    7) GDPperCapita>=6904.5 9  4 Asia (0 0.22 0.56 0.22 0) *
```



In [151]:
```
summary(regionTree)
```

```
Call:
rpart(formula = region ~ infantMortality + GDPperCapita, data = UN98NoNA)
  n= 39

         CP nsplit rel error    xerror      xstd
1 0.3076923      0 1.0000000 1.0000000 0.1132277
2 0.1153846      1 0.6923077 0.7307692 0.1200566
3 0.0100000      2 0.5769231 0.7307692 0.1200566

Variable importance
infantMortality    GDPperCapita
             60              40

Node number 1: 39 observations,    complexity param=0.3076923
  predicted class=America  expected loss=0.6666667  P(node) =1
    class counts:     9    13    10     6     1
   probabilities: 0.231 0.333 0.256 0.154 0.026
  left son=2 (11 obs) right son=3 (28 obs)
  Primary splits:
      infantMortality < 46.5   to the right, improve=5.992507, (0 missing)
      GDPperCapita    < 1550.5 to the left,  improve=3.000000, (0 missing)
  Surrogate splits:
      GDPperCapita < 1448.5 to the left,  agree=0.846, adj=0.455, (0 split)

Node number 2: 11 observations
  predicted class=Africa   expected loss=0.2727273  P(node) =0.2820513
    class counts:     8     0     2     0     1
   probabilities: 0.727 0.000 0.182 0.000 0.091

Node number 3: 28 observations,    complexity param=0.1153846
  predicted class=America  expected loss=0.5357143  P(node) =0.7179487
    class counts:     1    13     8     6     0
   probabilities: 0.036 0.464 0.286 0.214 0.000
  left son=6 (19 obs) right son=7 (9 obs)
  Primary splits:
      GDPperCapita    < 6904.5 to the left,  improve=1.760652, (0 missing)
      infantMortality < 27     to the right, improve=1.281119, (0 missing)
  Surrogate splits:
      infantMortality < 8.5    to the right, agree=0.821, adj=0.444, (0 split)

Node number 6: 19 observations
  predicted class=America  expected loss=0.4210526  P(node) =0.4871795
    class counts:     1    11     3     4     0
   probabilities: 0.053 0.579 0.158 0.211 0.000

Node number 7: 9 observations
  predicted class=Asia     expected loss=0.4444444  P(node) =0.2307692
    class counts:     0     2     5     2     0
   probabilities: 0.000 0.222 0.556 0.222 0.000
```

In [ ]:
```
```

1. Can you use the 'predict' function with your decision tree model on a new data point where 'infantMortality = 10' and 'GDPperCapita = 20000' and report on the output?

In [152]:
```
new<-data.frame(infantMortality = c(10), GDPperCapita = c(20000))
print(new)
print(predict(regionTree, new))
```

```
  infantMortality GDPperCapita
1              10        20000
  Africa   America      Asia    Europe Oceania
1      0 0.2222222 0.5555556 0.2222222       0
```

Yes, I can predict region under the condition of 'infantMortality = 10' and 'GDPperCapita = 20000', the region 56% chance is in Asia and equal 22% chance are in either America or Europe.

In [153]:
```
##########################################################################################################################
#####
```

Question 3 (8 points) There is another dataset 'books.csv' from Kaggle on the data collected from GoodReads.com which registers activity on a website where people read and share opinions, rating etc.

1. Produce a box plot for the 'average rating', 'num pages', 'ratings count', and 'text reviews count'.
2. Produce the same type of plot as before but make it a 'violin box plot' as shown in class.
3. Is the average rating correlated with the number of pages?
4. Make a model to predict the average rating from the number of pages and the ratings count (dependent is the 'avarage rating' and the independents 'num pages' and 'ratings count'). Is the model significant?

1. Produce a box plot for the 'average rating', 'num pages', 'ratings count', and 'text reviews count'.

In [154]:
```
b<- read.csv('C:/Users/Andrew/Desktop/books.csv')
```

In [155]:
```
install.packages("microbenchmark")
install.packages("magrittr")
install.packages("tidyr")
install.packages("dplyr")
library(microbenchmark)
library(tidyr)
library(dplyr)
library(magrittr)
```

```
Warning message:
"package 'microbenchmark' is in use and will not be installed"Warning message:
"package 'magrittr' is in use and will not be installed"also installing the dependency 'dplyr'

Warning message:
"package 'dplyr' is in use and will not be installed"

package 'tidyr' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\Andrew\AppData\Local\Temp\RtmpI5MZBx\downloaded_packages

Warning message:
"package 'dplyr' is in use and will not be installed"Warning message:
"package 'tidyr' was built under R version 3.6.3"

Error: package or namespace load failed for 'tidyr' in loadNamespace(j <- i[[1L]], c(lib.loc, .libPaths()), version
Check = vI[[j]]):
 namespace 'dplyr' 0.8.0.1 is already loaded, but >= 0.8.2 is required
Traceback:

1. library(tidyr)
2. tryCatch({
 .     attr(package, "LibPath") <- which.lib.loc
 .     ns <- loadNamespace(package, lib.loc)
 .     env <- attachNamespace(ns, pos = pos, deps, exclude, include.only)
 . }, error = function(e) {
 .     P <- if (!is.null(cc <- conditionCall(e)))
 .         paste(" in", deparse(cc)[1L])
 .     else ""
 .     msg <- gettextf("package or namespace load failed for %s%s:\n %s",
 .         sQuote(package), P, conditionMessage(e))
 .     if (logical.return)
 .         message(paste("Error:", msg), domain = NA)
 .     else stop(msg, call. = FALSE, domain = NA)
 . })
3. tryCatchList(expr, classes, parentenv, handlers)
4. tryCatchOne(expr, names, parentenv, handlers[[1L]])
5. value[[3L]](cond)
6. stop(msg, call. = FALSE, domain = NA)
```

In [156]:

```
#take off NA
bNoNA=na.omit(b)
#change factor variabes to numeric
average_rating_int=as.integer(bNoNA$average_rating)
X..num_pages_int = as.integer(bNoNA$X..num_pages)

is.factor(average_rating_int)
class(average_rating_int)

is.factor(X..num_pages_int)
class(X..num_pages_int)
```

FALSE

'integer'

FALSE

'integer'

In [157]:

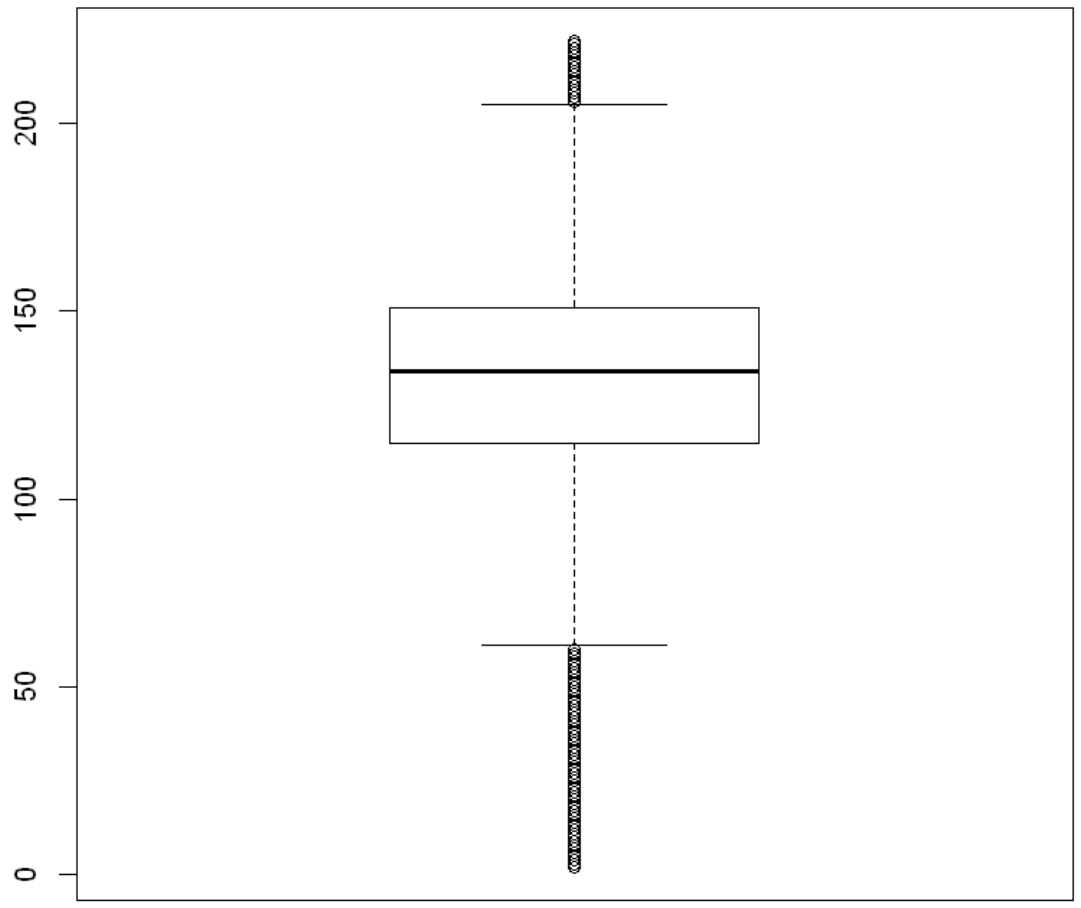
```
summary(bNoNA)
str(bNoNA)
head(bNoNA)
```

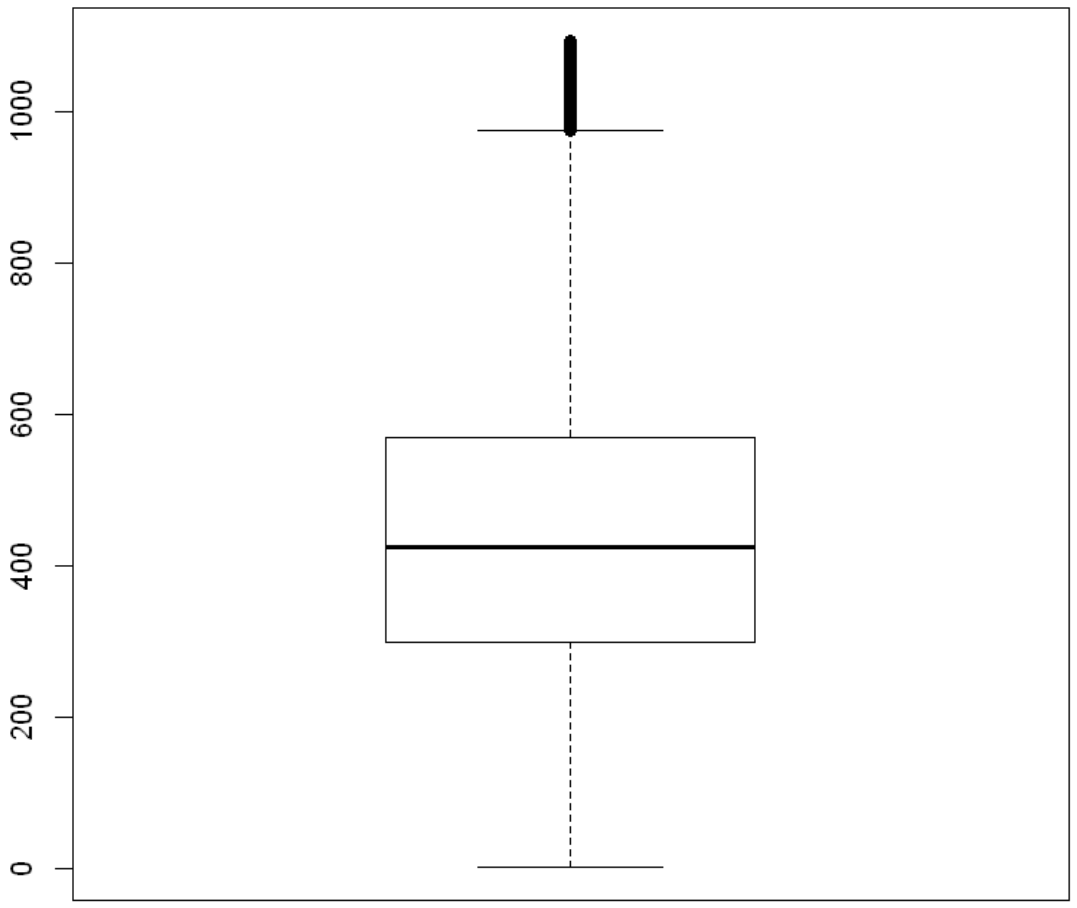
```
    bookID                                 title                  authors     
 Min.   :    1   'Salem's Lot                :   11   Agatha Christie :   69  
 1st Qu.:10621   One Hundred Years of Solitude:   11   Stephen King    :   66  
 Median :21322   The Brothers Karamazov      :   10   Orson Scott Card:   48  
 Mean   :22161   The Iliad                   :   10   Rumiko Takahashi:   46  
 3rd Qu.:33322   A Midsummer Night's Dream   :    9   P.G. Wodehouse  :   42  
 Max.   :47709   A Tale of Two Cities        :    9   Terry Brooks    :   40  
                 (Other)                     :13659   (Other)         :13408  
 average_rating         isbn               isbn13      language_code  
 4.00   :  261   0.00      :    1   0008987059752:    1   eng    :10594  
 3.96   :  231   000100039X:    1   0020049130001:    1   en-US  : 1699  
 3.95   :  222   0001713191:    1   0023755004321:    1   spa    :  419  
 3.89   :  219   0002005883:    1   0034406054602:    1   en-GB  :  341  
 3.98   :  219   0002259834:    1   0049086007763:    1   ger    :  238  
 3.99   :  217   0002261987:    1   0073999140774:    1   fre    :  209  
 (Other):12350   (Other)   :13713   (Other)      :13713   (Other):  219  
  X..num_pages   ratings_count     text_reviews_count
 288    :  252   Min.   :      0   Min.   :    0.0  
 320    :  248   1st Qu.:     83   1st Qu.:    7.0  
 192    :  244   Median :    630   Median :   40.0  
 256    :  236   Mean   :  17759   Mean   :  533.6  
 352    :  230   3rd Qu.:   4737   3rd Qu.:  222.0  
 224    :  227   Max.   :5629932   Max.   :93619.0  
 (Other):12282                                      

'data.frame':   13719 obs. of  10 variables:
 $ bookID            : int  1 2 3 4 5 8 9 10 12 13 ...
 $ title             : Factor w/ 12428 levels "","'Salem's Lot",..: 3808 3809 3812 3806 3811 3813 11750 3815 11082
11081 ...
 $ authors           : Factor w/ 7606 levels "","A.B. Yehoshua-Hillel Halkin",..: 3044 3044 3044 3036 3044 3044 724
8 3036 1696 1696 ...
 $ average_rating    : Factor w/ 222 levels "","  Jr.-C.S. Lewis-P.G. Wodehouse-Michael Moorcock-L. Sprague de Camp-
Fletcher Pratt-Eric Knight-Mervyn Peake-Pier"| __truncated__,..: 194 187 185 179 193 215 107 211 176 176 ...
 $ isbn              : Factor w/ 13720 levels "","0.00","000100039X",..: 4973 4890 4932 4931 4953 4958 10416 4980 6
001 3356 ...
 $ isbn13            : Factor w/ 13720 levels "","0008987059752",..: 4996 4913 4955 4954 4976 4981 10428 5003 6024
3379 ...
 $ language_code     : Factor w/ 36 levels "","9780674842113",..: 14 14 14 14 14 14 13 14 14 14 ...
 $ X..num_pages      : Factor w/ 1093 levels "","0","1","10",..: 794 1007 426 463 555 361 204 442 956 956 ...
 $ ratings_count     : int  1944099 1996446 5629932 6267 2149872 38872 18 27410 3602 240189 ...
 $ text_reviews_count: int  26249 27613 70390 272 33964 154 1 820 258 3954 ...
 - attr(*, "na.action")= 'omit' Named int  4012 5689 7058 10604 10672
  ..- attr(*, "names")= chr  "4012" "5689" "7058" "10604" ...
```

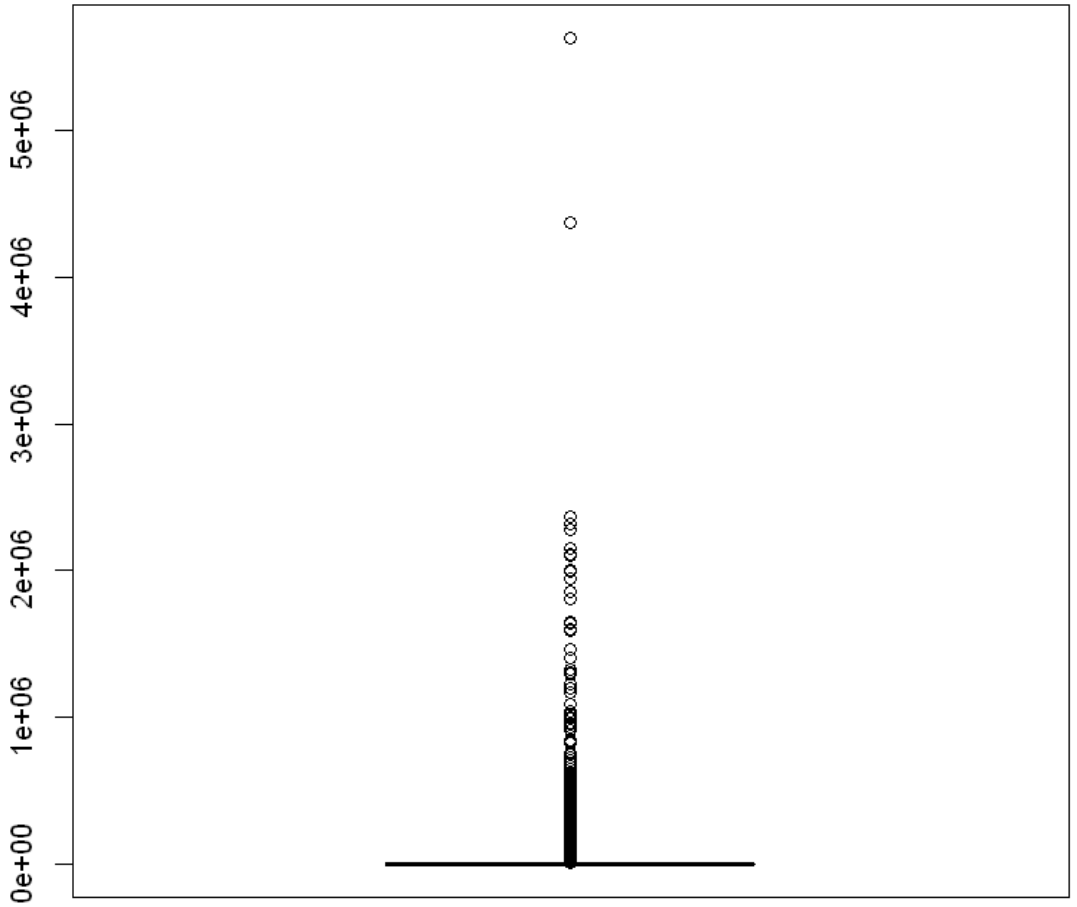
| bookID | title | authors | average_rating | isbn | isbn13 | language_code | X..num_pages | ratings_count | text_reviews_count |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Harry Potter and the Half-Blood Prince (Harry Potter #6) | J.K. Rowling-Mary GrandPrÃ© | 4.56 | 0439785960 | 9780439785969 | eng | 652 | 1944099 | 26249 |
| 2 | Harry Potter and the Order of the Phoenix (Harry Potter #5) | J.K. Rowling-Mary GrandPrÃ© | 4.49 | 0439358078 | 9780439358071 | eng | 870 | 1996446 | 27613 |
| 3 | Harry Potter and the Sorcerer's Stone (Harry Potter #1) | J.K. Rowling-Mary GrandPrÃ© | 4.47 | 0439554934 | 9780439554930 | eng | 320 | 5629932 | 70390 |
| 4 | Harry Potter and the Chamber of Secrets (Harry Potter #2) | J.K. Rowling | 4.41 | 0439554896 | 9780439554893 | eng | 352 | 6267 | 272 |
| 5 | Harry Potter and the Prisoner of Azkaban (Harry Potter #3) | J.K. Rowling-Mary GrandPrÃ© | 4.55 | 043965548X | 9780439655484 | eng | 435 | 2149872 | 33964 |
| 8 | Harry Potter Boxed Set Books 1-5 (Harry Potter #1-5) | J.K. Rowling-Mary GrandPrÃ© | 4.78 | 0439682584 | 9780439682589 | eng | 2690 | 38872 | 154 |

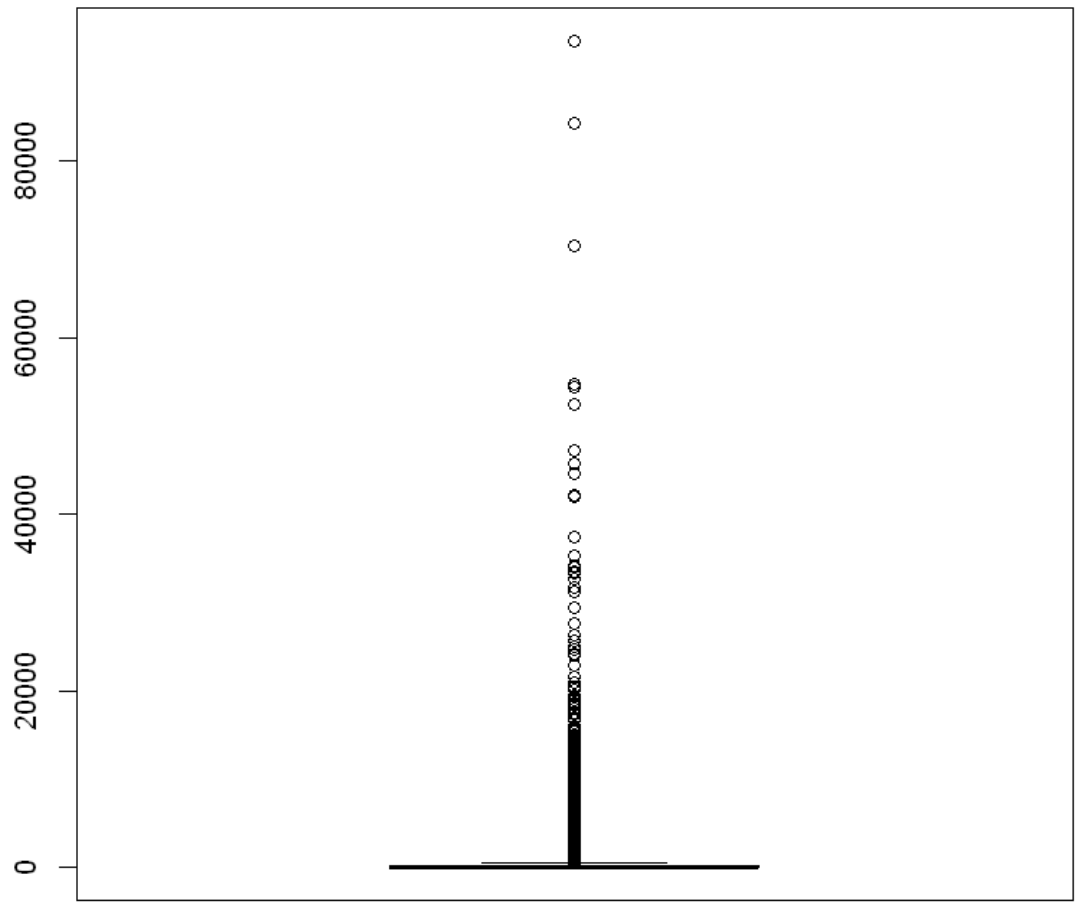
In [158]:

```
boxplot(average_rating_int, main="average rating")
boxplot(X..num_pages_int, main="num pages")
boxplot(b$ratings_count, main=" ratings count")
boxplot(b$text_reviews_count, main="text reviews count")
```

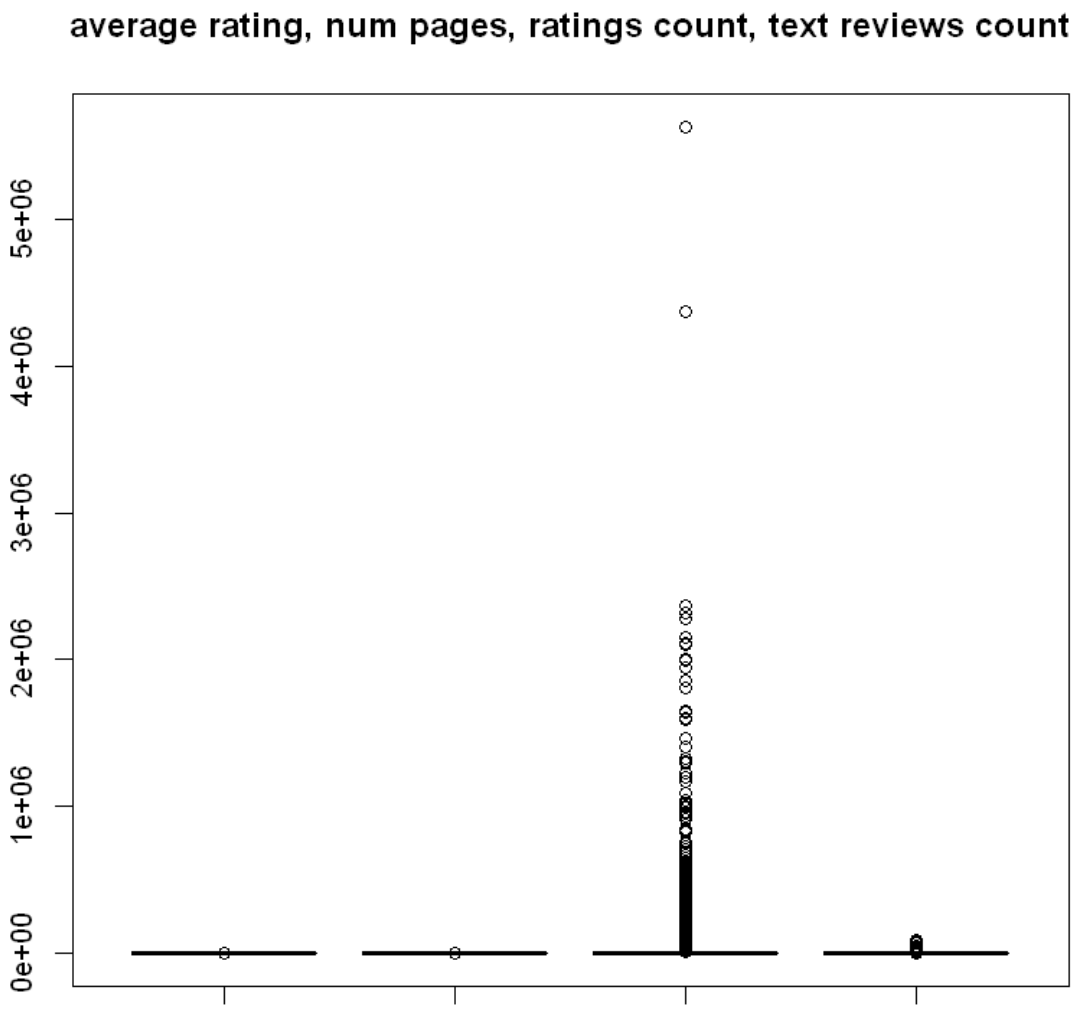

average rating
0
50
100
150
200


num pages
0
200
400
600
800
1000


ratings count
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06


text reviews count
0
20000
40000
60000
80000

In [159]:
```
#not sure you asking for one plot for these 4 variables or 4 individual plots, made this one just in case
boxplot(b$average_rating, b$X..num_pages,b$ratings_count,b$text_reviews_count,
        main="average rating, num pages, ratings count, text reviews count"  )
```




In [ ]:

1. Produce the same type of plot as before but make it a 'violin box plot' as shown in class.

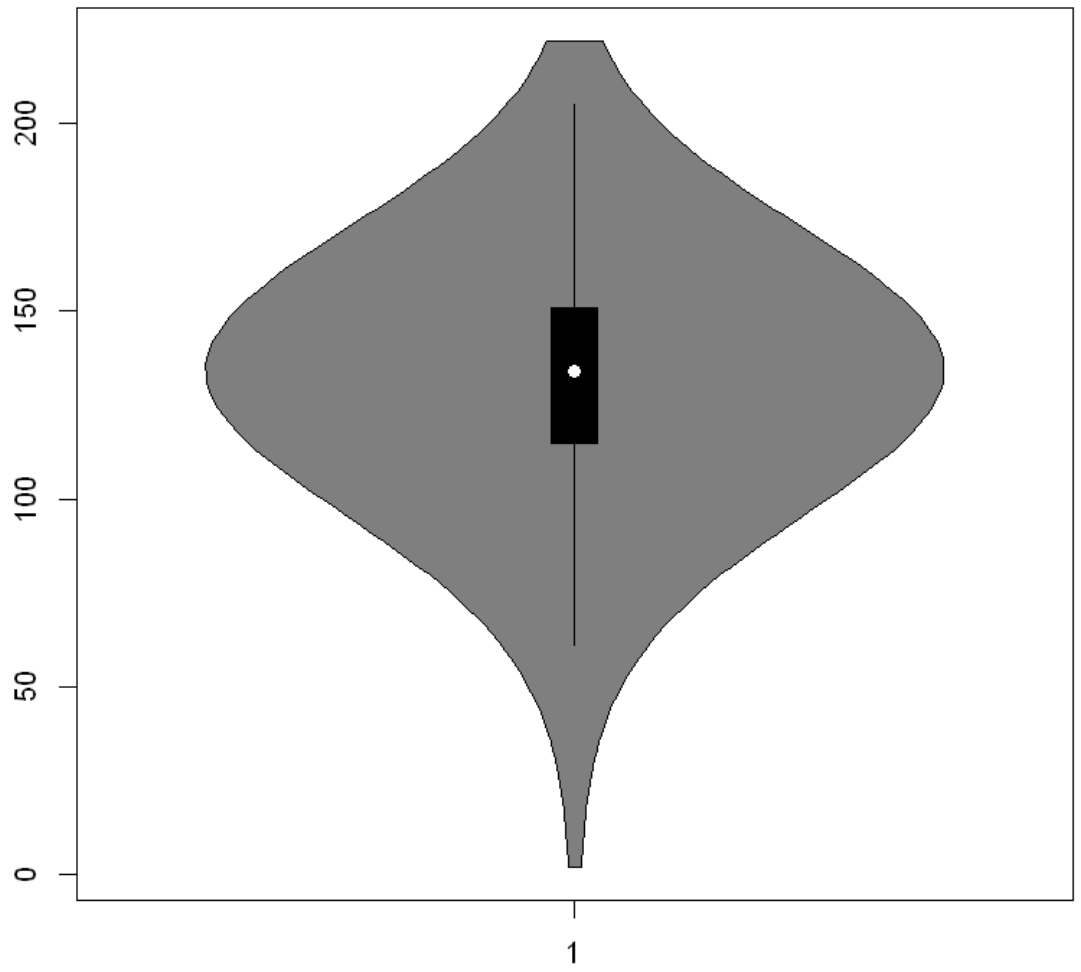
In [160]:

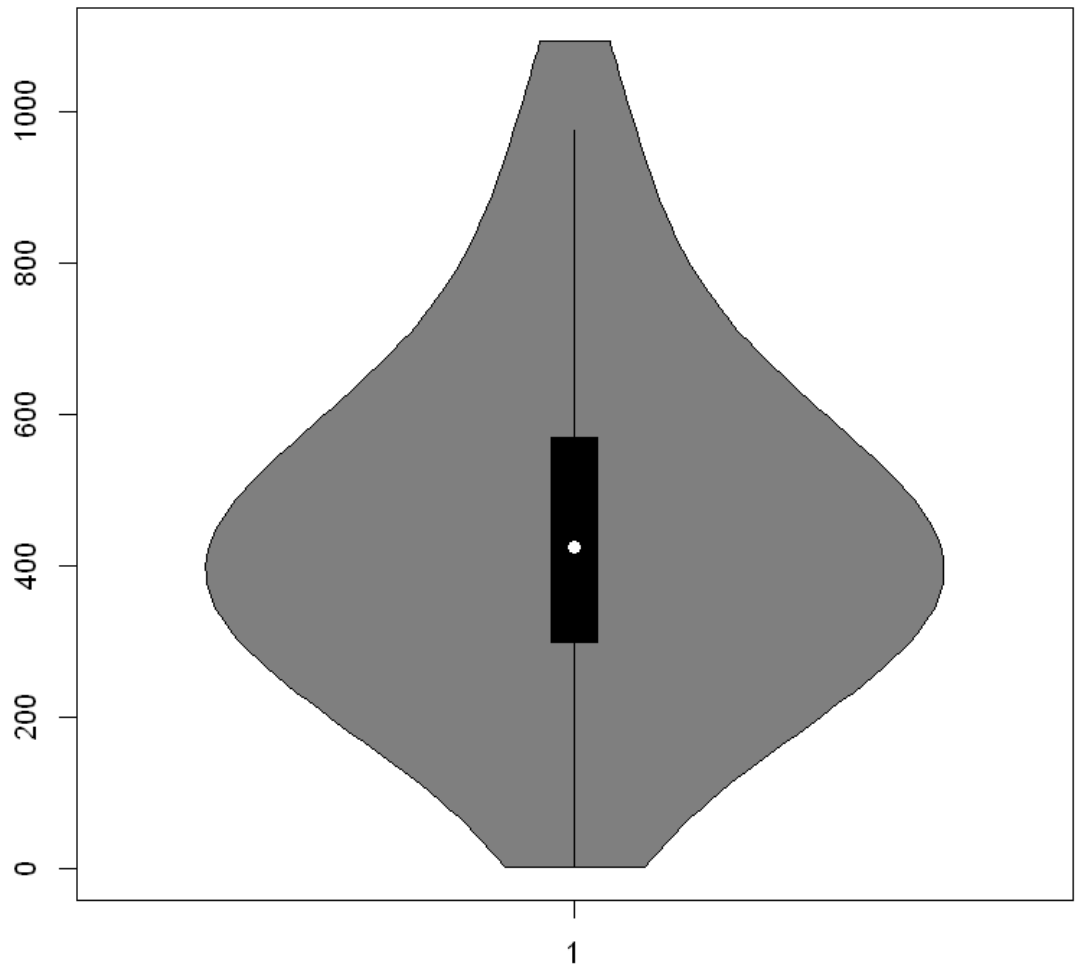
```
install.packages("sm")
install.packages("zoo")
install.packages("vioplot")
install.packages("GGally")
install.packages("htmlwidgets")
install.packages("rpivotTable")
install.packages("gplots")
install.packages("graphics")
install.packages("corrplot")
library(sm)
library(vioplot)
library(zoo)
library(GGally)
library(htmlwidgets)
library(rpivotTable)
library(gplots)
library(graphics)
library(corrplot)
```

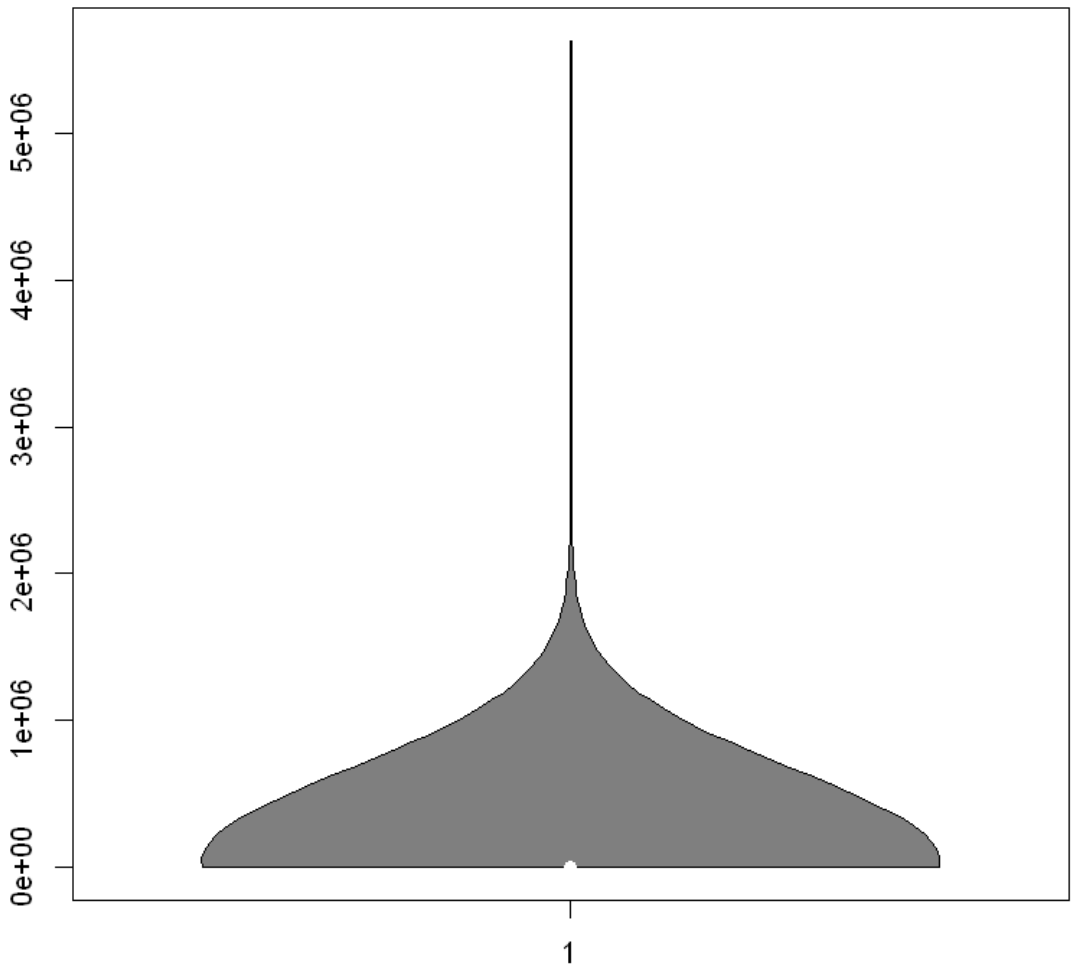
```
Warning message:
"package 'sm' is in use and will not be installed"Warning message:
"package 'zoo' is in use and will not be installed"Warning message:
"package 'vioplot' is in use and will not be installed"Warning message:
"package 'GGally' is in use and will not be installed"Warning message:
"package 'htmlwidgets' is in use and will not be installed"Warning message:
"package 'rpivotTable' is in use and will not be installed"Warning message:
"package 'gplots' is in use and will not be installed"Warning message:
"package 'graphics' is not available (for R version 3.6.1)"Warning message:
"package 'graphics' is a base package, and should not be updated"Warning message:
"package 'corrplot' is in use and will not be installed"
```

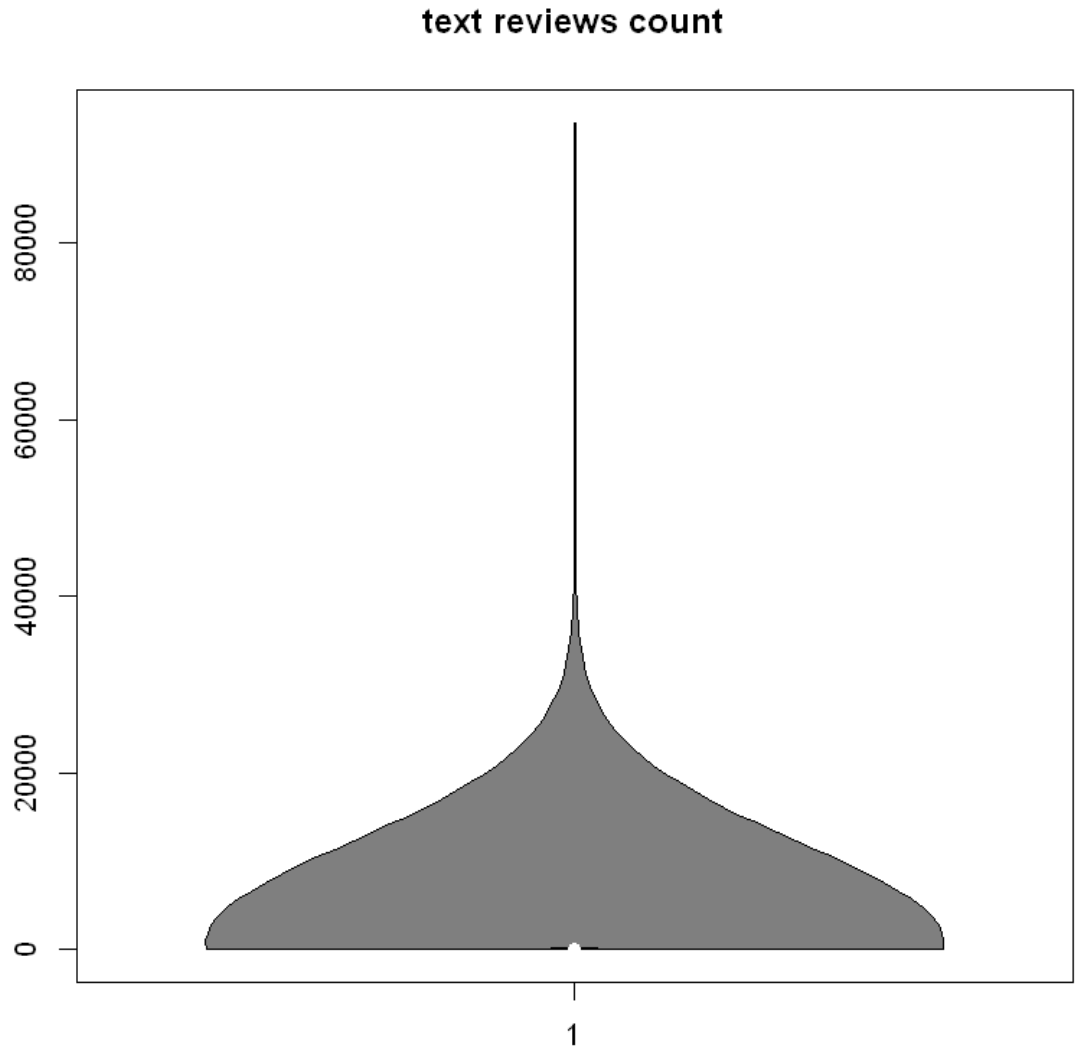
In [161]:

```
vioplot(average_rating_int, main="average rating")
vioplot(X..num_pages_int, main="num pages")
vioplot(b$ratings_count, main=" ratings count")
vioplot(b$text_reviews_count, main="text reviews count")
```


average rating
200
150
100
50
0
1


num pages
0
200
400
600
800
1000
1


ratings count
0e+00
1e+06
2e+06
3e+06
4e+06
5e+06
1




In [ ]:

1. Is the average rating correlated with the number of pages?

In [162]:
```
#because both of these two variables are factor in original data, use chi square to test the correlation
tb=table(b$average_rating, b$X..num_pages)
chisqResult=chisq.test(tb)
chisqResult
```

```
Warning message in chisq.test(tb):
"Chi-squared approximation may be incorrect"

        Pearson's Chi-squared test

data:  tb
X-squared = 312209, df = 241332, p-value < 2.2e-16
```

From the chi square test, the p value is very small, states variable average rating correlated with the number of pages.

In [163]:
```
#use pearson's test to test if the converted two numeric varialbes are correlated
cor.test(average_rating_int, X..num_pages_int, method=c("pearson", "kendall", "spearman"))
```

```
        Pearson's product-moment correlation

data:  average_rating_int and X..num_pages_int
t = 9.8044, df = 13717, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.06678064 0.10001534
sample estimates:
       cor
0.08342118
```

From the Pearson's test, the p value is very small, states the numerica variable average rating correlated with the numerica variable number of pages.

In [ ]:
```

```

1. Make a model to predict the average rating from the number of pages and the ratings count (dependent is the 'avarage rating' and the independents 'num pages' and 'ratings count'). Is the model significant?

In [164]:
```
install.packages("dplyr")
library(dplyr)
```

```
Warning message:
"package 'dplyr' is in use and will not be installed"
```

In [165]:
```
#build multi regression model using converted numberic variables
modelglm<-glm(average_rating_int ~ X..num_pages_int + ratings_count, data=bNoNA)
summary(modelglm)
```

```
Call:
glm(formula = average_rating_int ~ X..num_pages_int + ratings_count,
    data = bNoNA)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-136.584   -16.671     1.684    19.465    94.340

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.266e+02  5.714e-01 221.570  < 2e-16 ***
X..num_pages_int 1.096e-02  1.126e-03   9.732  < 2e-16 ***
ratings_count    1.172e-05  2.247e-06   5.214 1.87e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 883.5907)

    Null deviance: 12228453  on 13718  degrees of freedom
Residual deviance: 12119330  on 13716  degrees of freedom
AIC: 132007

Number of Fisher Scoring iterations: 2
```

From the summary of the model, we can see the p value are very small, the model is statiscal significant.

In [166]:
```
#make categories from average rating
bNoNA$average_ratingGroup=ntile(bNoNA$average_rating, 4)
head(bNoNA$average_ratingGroup, 10)
#make categories from number of page
bNoNA$X..num_pagesGroup=ntile(bNoNA$X..num_pages, 4)
head(bNoNA$X..num_pagesGroup, 10)
```

```
4 4 4 4 4 4 1 4 4 4

4 4 3 3 3 2 1 3 4 4
```

In [167]:
```
install.packages("nnet")
library(nnet)
```

```
Warning message:
"package 'nnet' is in use and will not be installed"
```

In [168]:
```
modelmultiA<-multinom(average_ratingGroup ~ X..num_pagesGroup + ratings_count, data=bNoNA)
summary(modelmultiA)
```

```
# weights:  16 (9 variable)
initial  value 19018.572340
iter  10 value 18940.197564
final  value 18940.025065
converged

Call:
multinom(formula = average_ratingGroup ~ X..num_pagesGroup +
    ratings_count, data = bNoNA)

Coefficients:
  (Intercept) X..num_pagesGroup ratings_count
2  -0.4261228         0.1688427  1.303554e-06
3  -0.5240801         0.2044145  1.689709e-06
4  -0.5270193         0.2048218  1.762504e-06

Std. Errors:
   (Intercept) X..num_pagesGroup ratings_count
2 1.606716e-12      4.019918e-12  4.061006e-07
3 1.547282e-12      3.892790e-12  3.912453e-07
4 1.551771e-12      3.897852e-12  3.899449e-07

Residual Deviance: 37880.05
AIC: 37898.05
```

In [169]:
```
modelmultiB<-multinom(average_ratingGroup ~ ratings_count, data=bNoNA)
summary(modelmultiB)
```

```
# weights:  12 (6 variable)
initial  value 19018.572340
iter  10 value 19000.738745
iter  10 value 19000.738744
iter  10 value 19000.738744
final  value 19000.738744
converged

Call:
multinom(formula = average_ratingGroup ~ ratings_count, data = bNoNA)

Coefficients:
  (Intercept) ratings_count
2 -0.01825108  1.416013e-06
3 -0.02552116  1.808376e-06
4 -0.02740774  1.881022e-06

Std. Errors:
   (Intercept) ratings_count
2 1.729308e-12  4.152665e-07
3 1.678784e-12  4.008923e-07
4 1.684662e-12  3.996655e-07

Residual Deviance: 38001.48
AIC: 38013.48
```

In [170]:
```
modelmultiC<-multinom(average_ratingGroup ~ X..num_pagesGroup, data=bNoNA)
summary(modelmultiC)
```

```
# weights:  12 (6 variable)
initial  value 19018.572340
iter  10 value 18956.154320
iter  10 value 18956.154297
iter  10 value 18956.154297
final  value 18956.154297
converged

Call:
multinom(formula = average_ratingGroup ~ X..num_pagesGroup, data = bNoNA)

Coefficients:
  (Intercept) X..num_pagesGroup
2  -0.4138021         0.1707184
3  -0.5065104         0.2070364
4  -0.5083656         0.2076432

Std. Errors:
  (Intercept) X..num_pagesGroup
2  0.05804869        0.02176307
3  0.05858073        0.02179188
4  0.05859458        0.02179406

Residual Deviance: 37912.31
AIC: 37924.31
```

In [171]:
```
#tried A B C 3 models with different vaiables, AIC of model A is the smallest
# i am going to use model A to predict
#try a prediction use random data frame
new<- data.frame(X..num_pagesGroup=c(4), ratings_count=c(2500))
print(new)
str(new)
predict(modelmultiA, new)
```

```
  X..num_pagesGroup ratings_count
1                 4          2500
'data.frame':   1 obs. of  2 variables:
 $ X..num_pagesGroup: num 4
 $ ratings_count    : num 2500

3
```

▶ **Levels**:

I created two groups factor variables to predict average rating group (1-4). For an example, scenario num page group 4 and ratings count 2500, the model prediction of average rating group is 3.

In [ ]:
```

```

In [172]:
```
##########################################################################################################################
####
```

Question 4 (6 points) There are 2 datasets; 'hr.7257.csv' and 'hr.11839.csv'. These datasets are recordings of heart rates (heart beat data) from 2 subjects at the same period on the same conditions.

1. Produce a single 2 column dataset from the 2 datasets.
2. Produce 2 histograms from the values in each column.
3. Produce a single plot with the 2 histograms overlayed with a vertical line for each mean value of the distributions.
4. Use the T-test to see if there is significant difference between the distributions.

1. Produce a single 2 column dataset from the 2 datasets.

In [173]:
```
h7257<- read.csv('C:/Users/Andrew/Desktop/hr.7257.csv')
str(h7257)
h11839<- read.csv('C:/Users/Andrew/Desktop/hr.11839.csv')
str(h11839)
```

```
'data.frame':   1799 obs. of  1 variable:
 $ X91.4634: num  91.5 91.2 91.9 91.2 89.8 ...
'data.frame':   1799 obs. of  1 variable:
 $ X84.2697: num  84.3 84.1 85.7 87.2 87.1 ...
```

In [174]:
```
summary(h7257)
summary(h11839)
```

```
    X91.4634
 Min.   : 80.21
 1st Qu.: 92.42
 Median : 98.25
 Mean   : 96.64
 3rd Qu.:101.36
 Max.   :104.89

    X84.2697
 Min.   : 73.44
 1st Qu.: 88.86
 Median : 92.21
 Mean   : 92.60
 3rd Qu.: 96.40
 Max.   :106.76
```

In [175]:
```
#colomun combine to one dataset called c
c<-cbind(h7257,h11839)
head(c)
str(c)
```

| X91.4634 | X84.2697 |
|---|---|
| 91.4634 | 84.2697 |
| 91.1834 | 84.0619 |
| 91.8788 | 85.6542 |
| 91.1772 | 87.2093 |
| 89.7992 | 87.1246 |
| 90.3571 | 86.8726 |

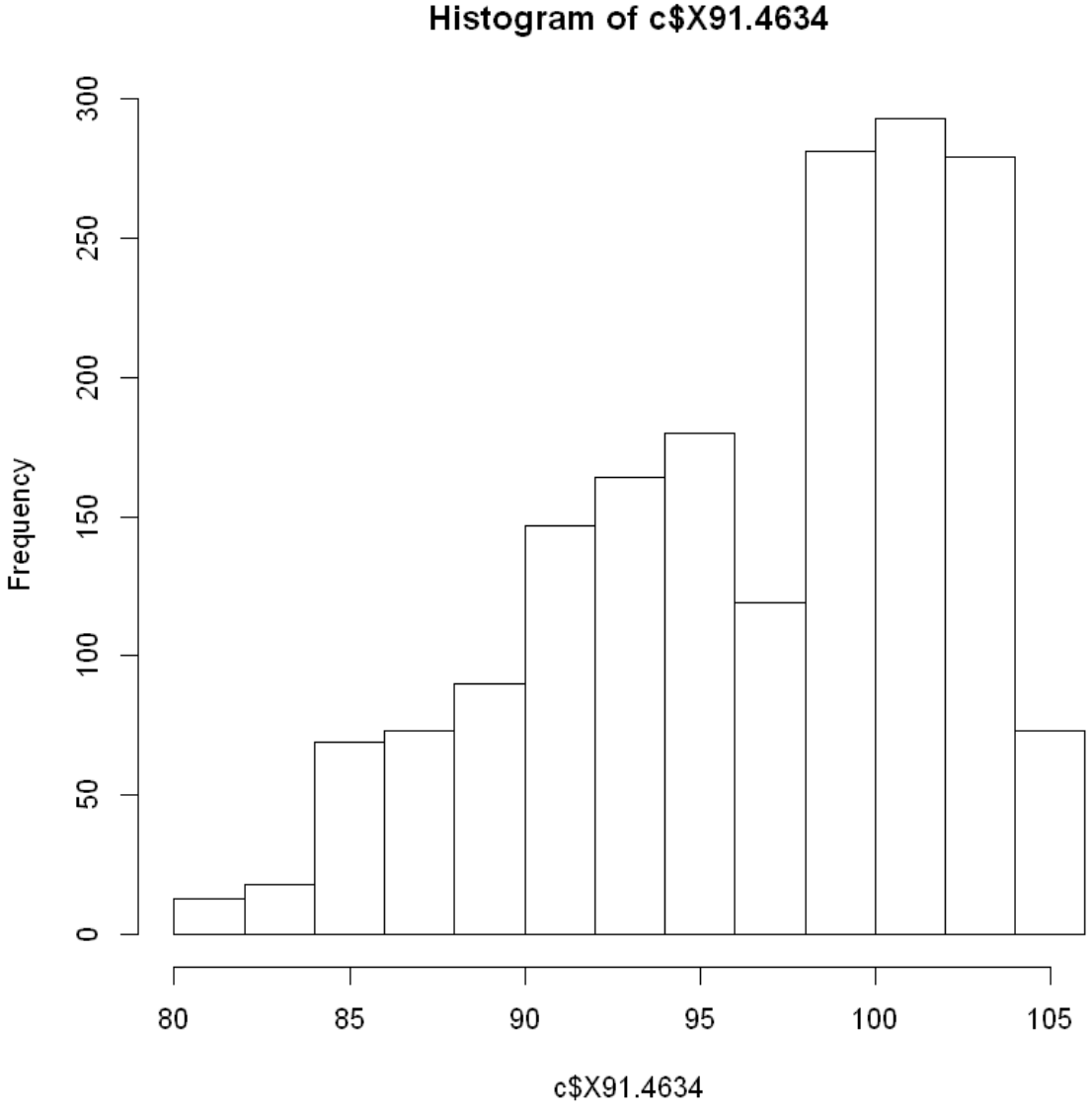
```
'data.frame':   1799 obs. of  2 variables:
 $ X91.4634: num  91.5 91.2 91.9 91.2 89.8 ...
 $ X84.2697: num  84.3 84.1 85.7 87.2 87.1 ...
```

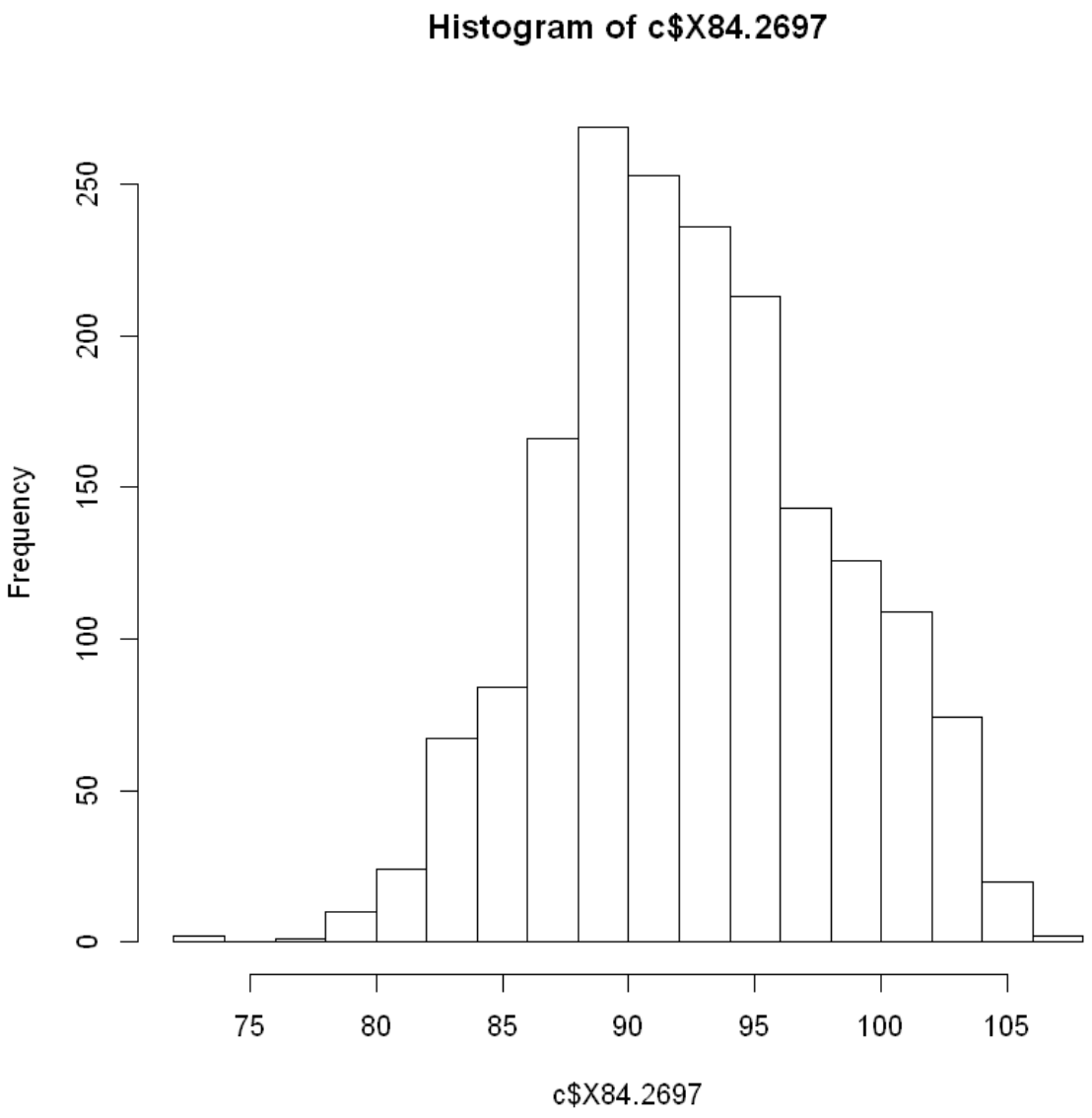
In [ ]:
```

```

1. Produce 2 histograms from the values in each column.

In [176]:

```
h1<-hist(c$X91.4634)
h2<-hist(c$X84.2697)
```


Histogram of c$X91.4634
Frequency
0
50
100
150
200
250
300
80
85
90
95
100
105
c$X91.4634



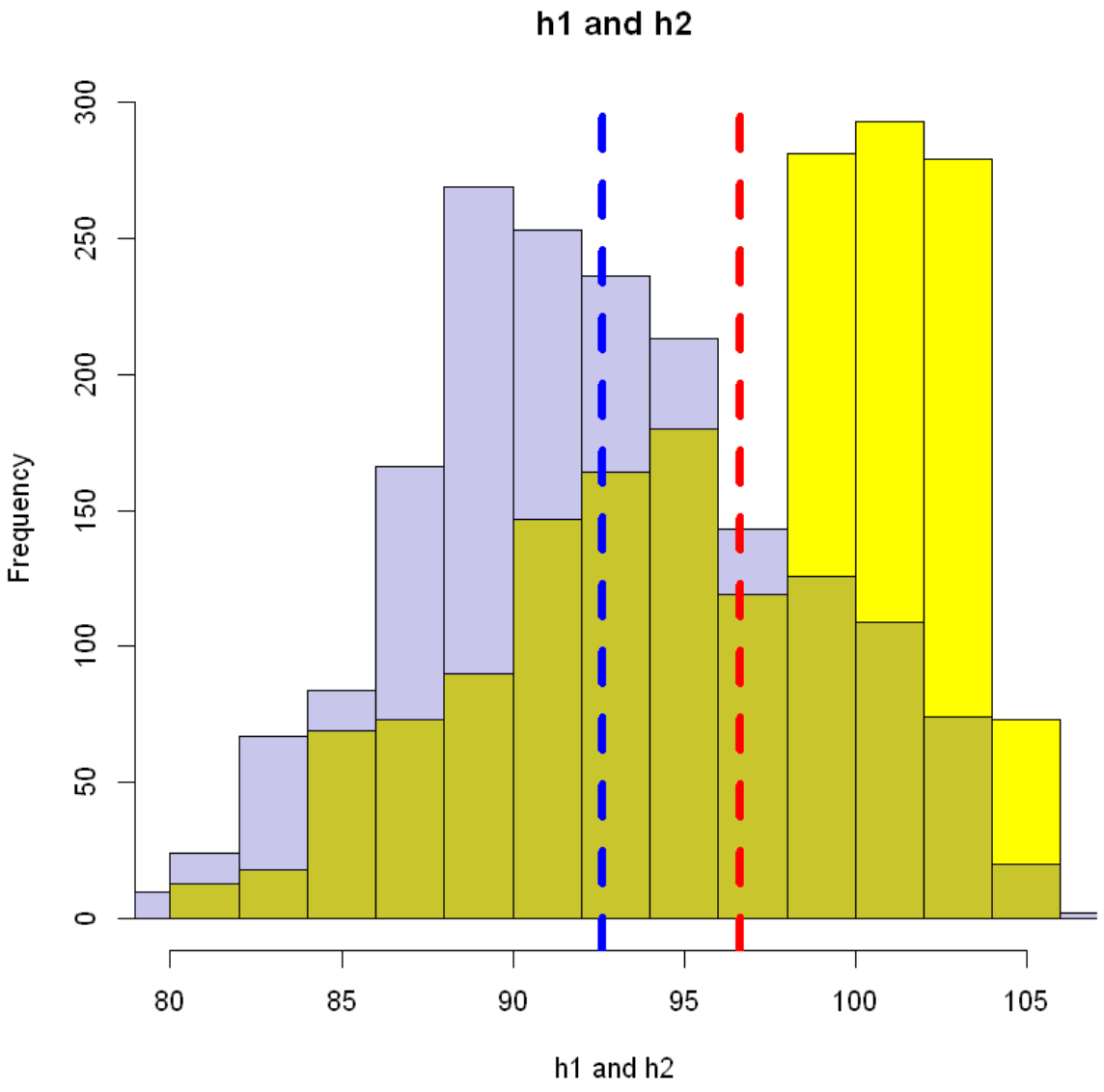

In [ ]:

1. Produce a single plot with the 2 histograms overlayed with a vertical line for each mean value of the distributions.

In [177]:
```
install.packages("plyr")
library(plyr)
```

```
Warning message:
"package 'plyr' is in use and will not be installed"
```

In [178]:
```
plot(h1, col="yellow", main="h1 and h2", xlab="h1 and h2")
plot(h2, col=rgb(0.15, .1, .7, 1/4), xlim=c(-5, 15), add=TRUE )
abline(v=mean(c$X91.4634), col="red", lwd=5, lty=2)
abline(v=mean(c$X84.2697), col="blue", lwd=5, lty=2)
```




In [ ]:
```
```

1. Use the T-test to see if there is significant difference between the distributions.

In [179]:
```
t.test(c$X91.4634, c$X84.2697)
```

```
        Welch Two Sample t-test

data:  c$X91.4634 and c$X84.2697
t = 21.68, df = 3591.4, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 3.674287 4.404927
sample estimates:
mean of x mean of y
 96.64035  92.60074
```

From the result of T-test, we can see the p value is very small, h1 mean is 96.64 and h2 mean is 92.60. They are significant difference in means.

In [ ]:

In [ ]: