Frame5

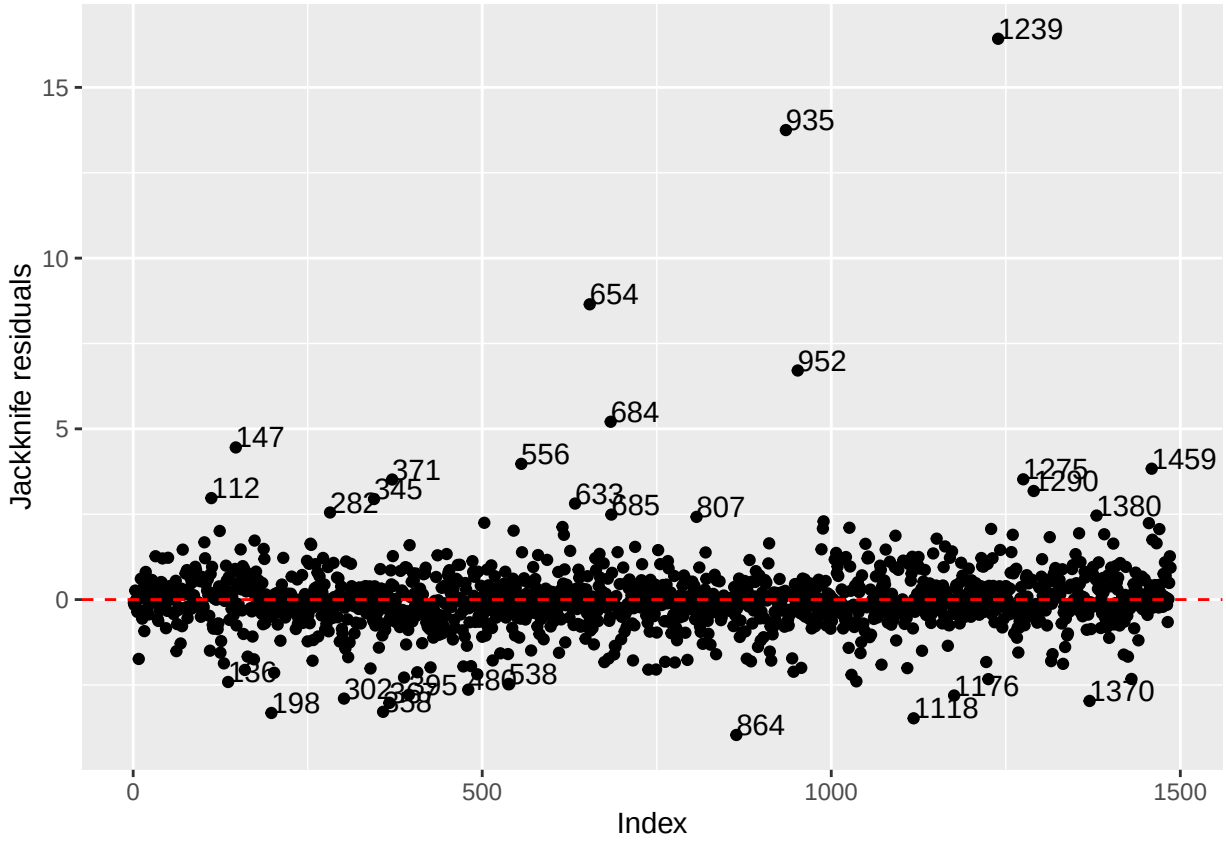
Ottó Hólm Reynisson

15.september 2016

Residuals

Begin by looking at the residuals from this model

```
indexPlotJackResiduals(lm.all)
```

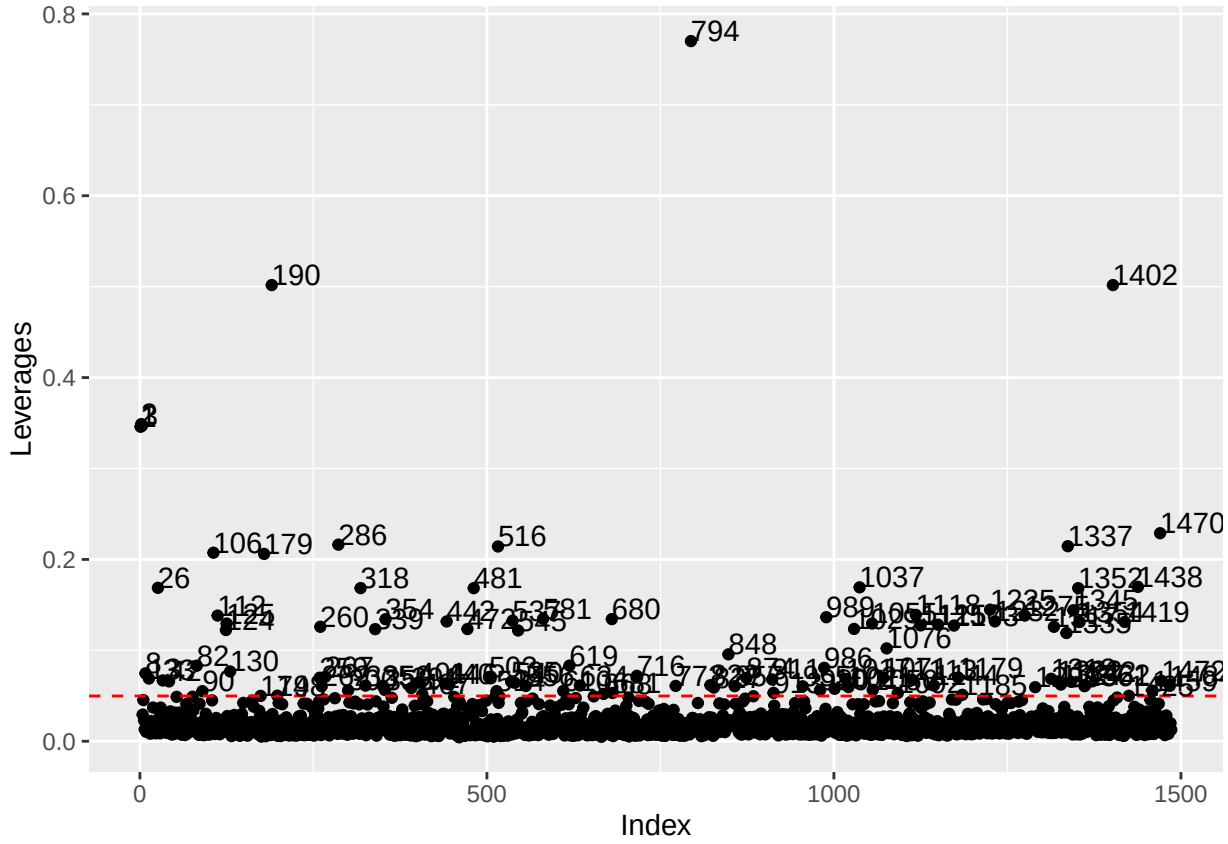


Here the blue line represent 2 standard deviation from 0 and green line 3 standard deviation away from 0. To help us identify the possible outliers we mark points 2 sd from 0 with their index.

Leverages

The next thing to do is looking at the leverages, that is the measure of how far independent variable values of an observation are from those of the other observation.

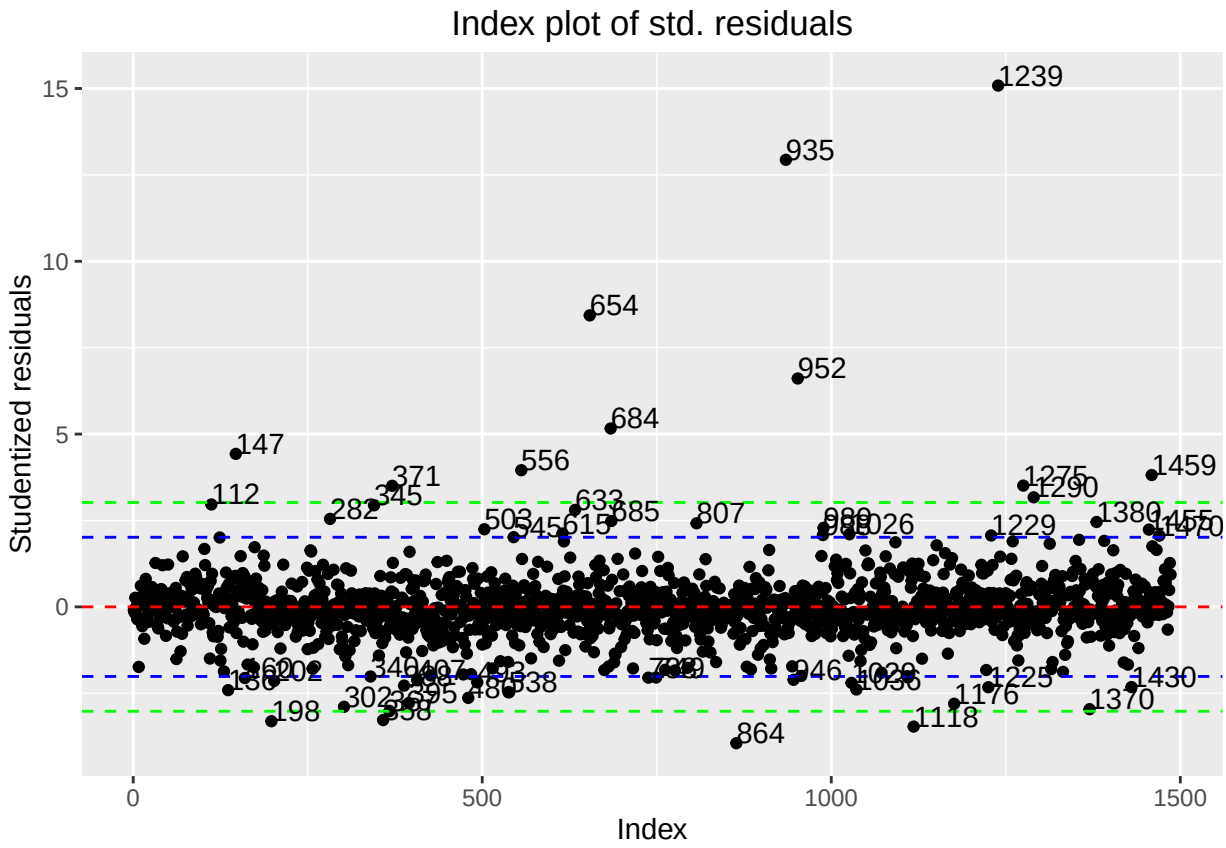
```
indexPlotLeverage(lm.all)
```



Studentized residuals

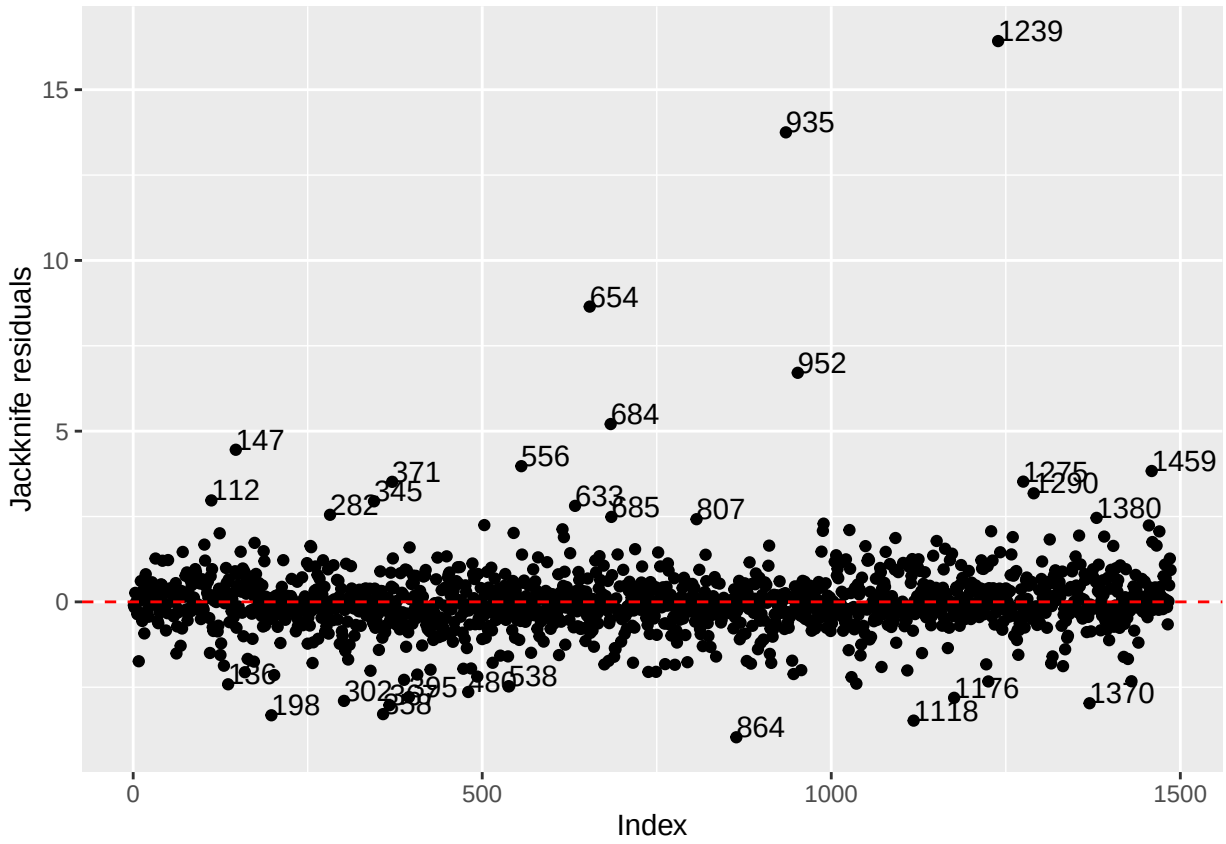
Studentized residuals are sometimes preferred in residual plots as they have been standardized to have equal variance. They are also a big part in the Jackkiife residulas that follows

```
indexPlotStResiduals(lm.all)
```

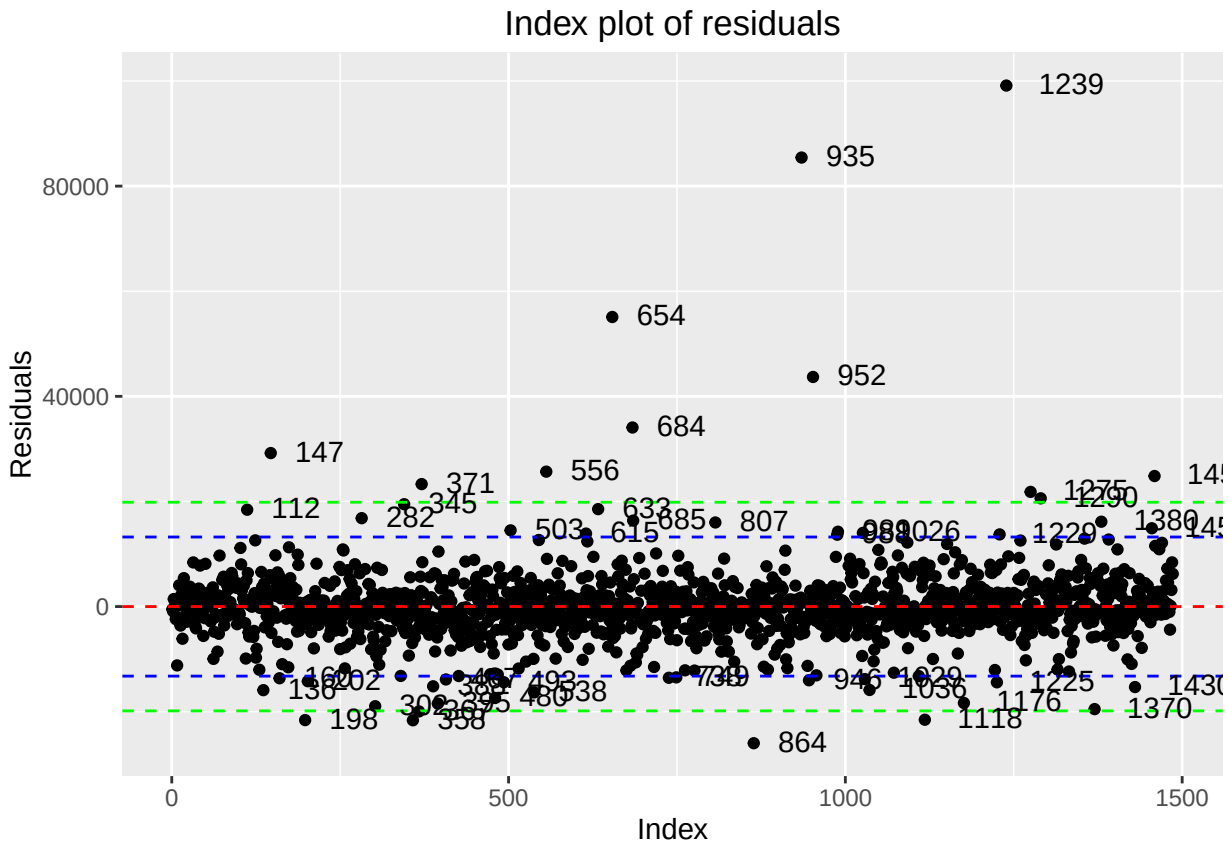


Blue and green line represent as before 2 and 3 sd from zero.

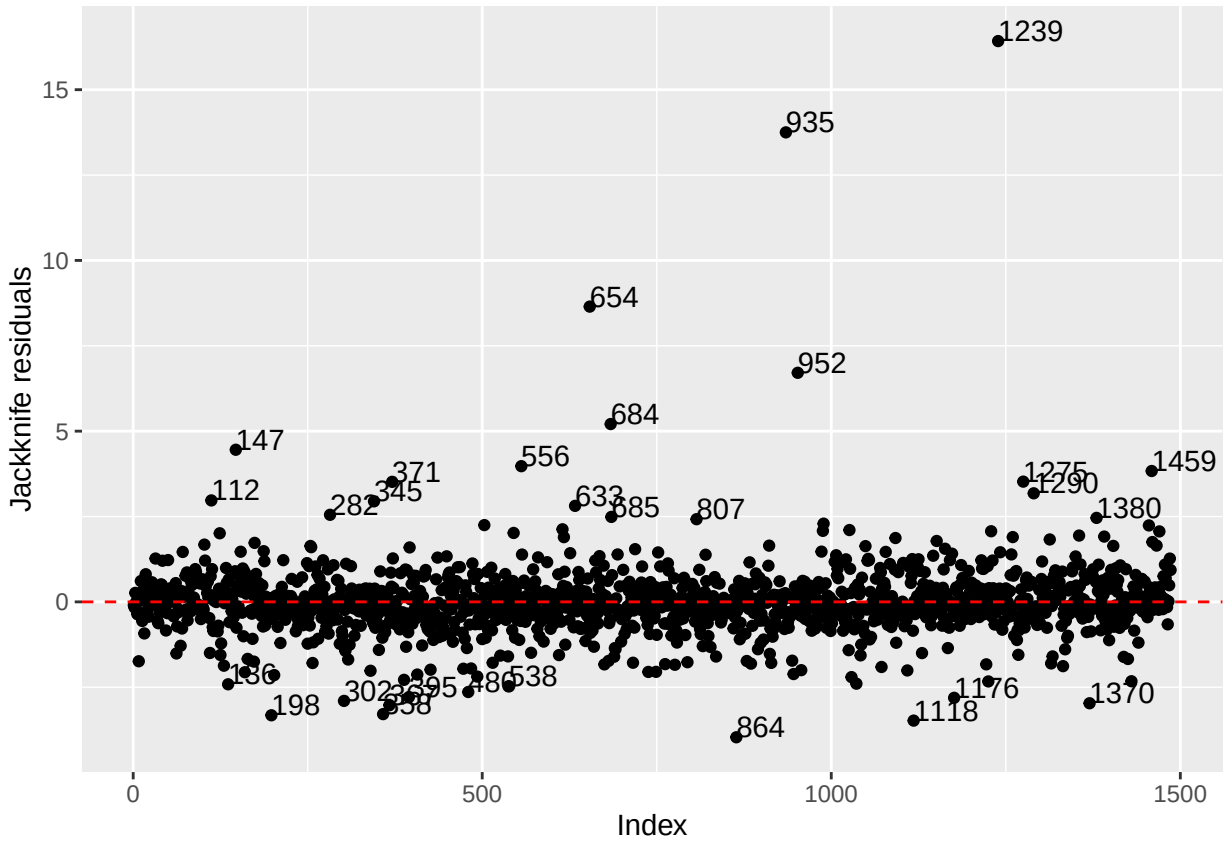
```
indexPlotJackResiduals(lm.all)
```



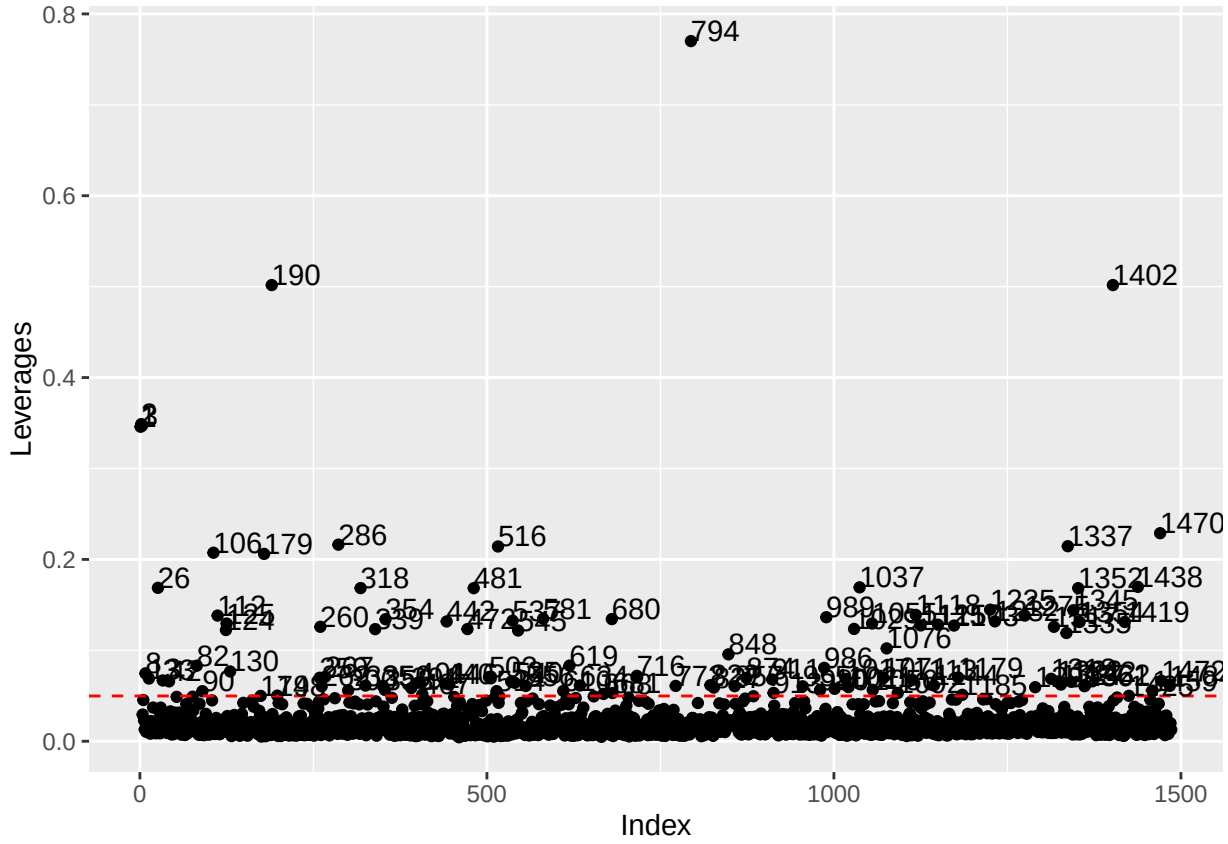

```
######################  Föll fyrir residuals og annað er í outliers.R skránni  ######

# funcions to plot residuals and leverage
indexPlotResiduals(lm.all)
```



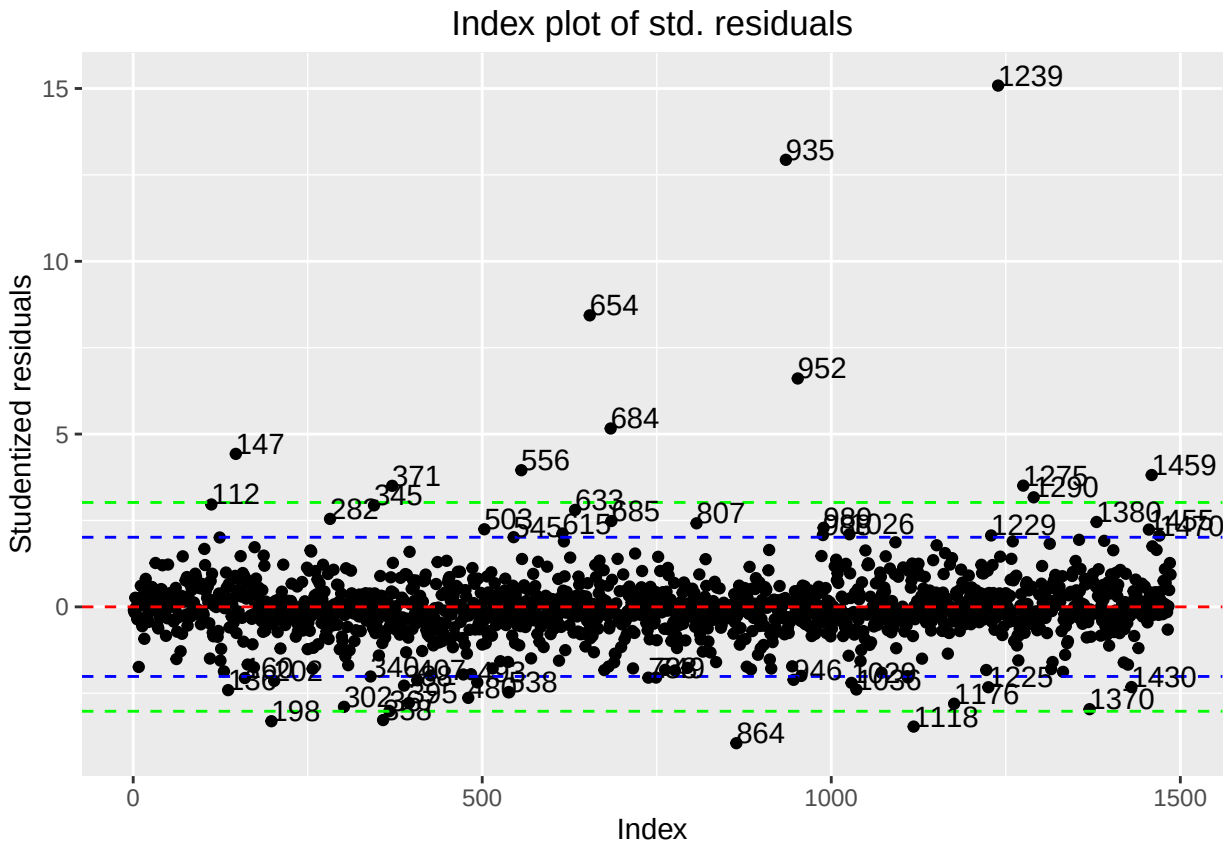

```
indexPlotJackResiduals(lm.all)
```



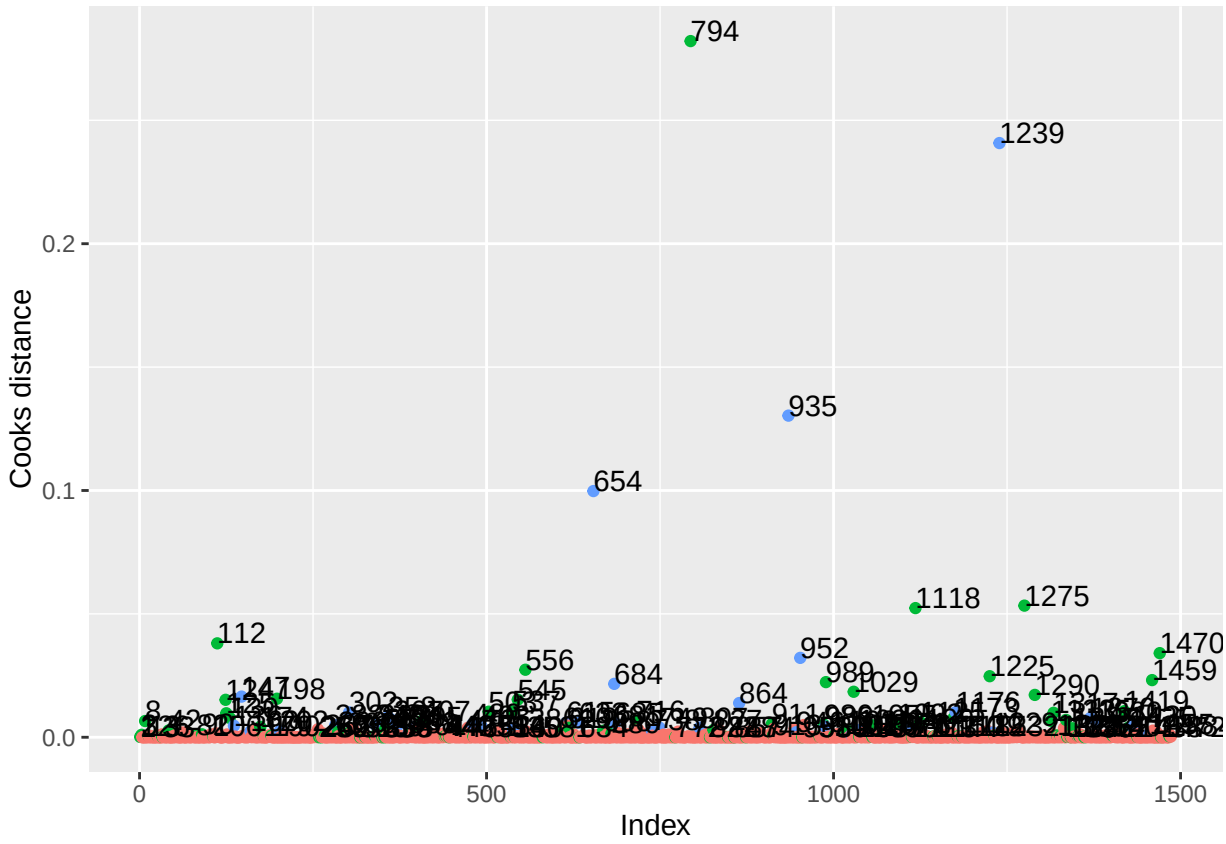

```
indexPlotLeverage(lm.all)
```




```
indexPlotStResiduals(lm.all)
```




```
# blue dots have high residuals and green have high leverage
indexPlotCookdistance(lm.all)
```




```
##################################### R squared adjusted fyrir lm.all
Radj.all <- CalculateRadjusted(lm.all, test) # 0.7990392
```

```
################################### lm.allNoOutliers Here I take out all outliers and get a new model
lm.allNoOutlier <- removeOutliersWStdResMoreThanThree(lm.all)
Radj.allNoOutliers <- CalculateRadjusted(lm.allNoOutlier, test) # 0.7328961
```

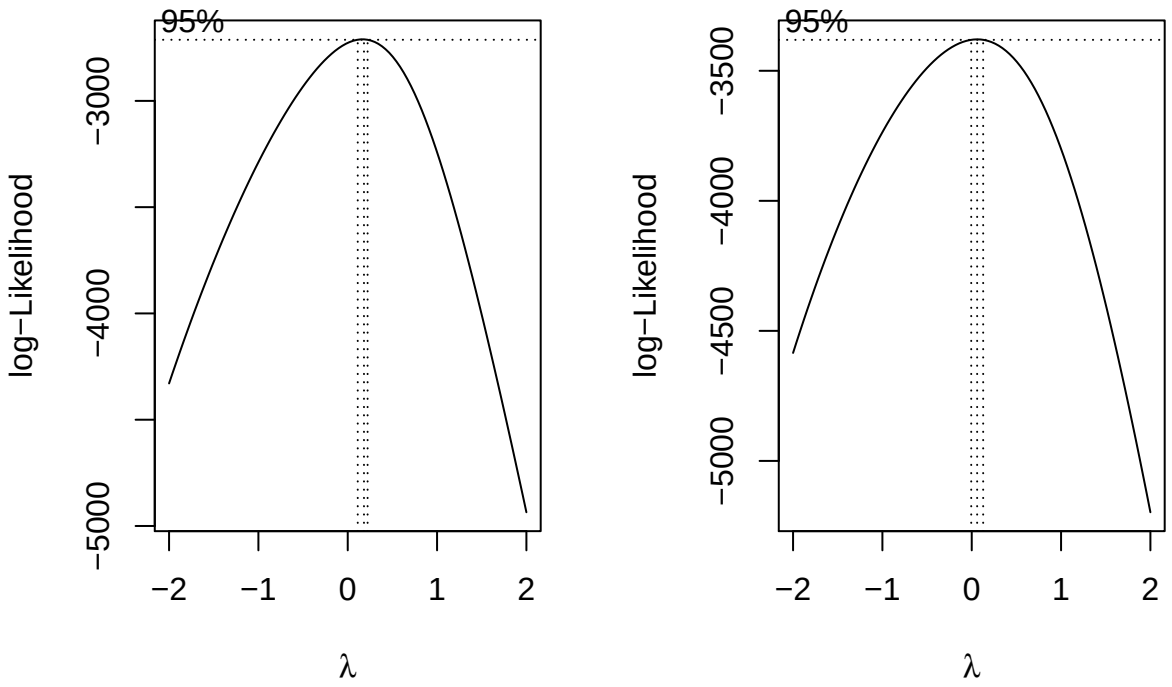
```
################################### lm.allNoInfluential Here I take out all influential points and get
lm.allNoInfluential <- removeInfluential(lm.all,maxCookDistance = 4/n)
Radj.allNoInfluential <- CalculateRadjusted(lm.allNoInfluential, test) # 0.7327364
```

Transformation

We know that nuvirdi has an unusually heavy tale so we'll start by transforming our response variable using boxcox.

```
par(mfrow=c(1,2))
lm.noOutlier <- lm(nuvirdi ~ ibm2, data=train)
boxcox(lm.all)
boxcox(lm.noOutlier)
```

```
par(mfrow=c(1,1))
```



Mynd 1: Boxcox plot for the model and the model without outliers. No outlier plot is on the right

```
trainNO <- train
bcAll <- boxcox(lm.all, plotit = FALSE)
bcNO <- boxcox(lm.noOutlier, plotit = FALSE)
# Get the highest value from the boxcox plots
lambdaAll <- with(bcAll, x[which.max(y)])
lambdaNO <- with(bcNO, x[which.max(y)])
# Transform the train and test datasets
trainBCAll <- bcTransF(train, lambdaAll)
testBCAll <- bcTransF(test, lambdaAll)
trainBCNO <- bcTransF(trainNO, lambdaNO)
testBCNO <- bcTransF(test, lambdaNO)

lm.allBc <- lm(nuvirdi ~ ., data = trainBCAll)
lm.noOutlierBc <- lm(nuvirdi ~ ibm2, data = trainBCNO)
Radj.allBc <- CalculateRadjLambda(lm.allBc, testBCAll, lambdaAll) # 0.8488961 rétt: 0.7643883
Radj.allNOBc <- CalculateRadjLambda(lm.noOutlierBc, testBCNO, lambdaNO)
```

Here we get $R_{adj.fullmodel} = 0.7673279$ and $R_{adj.NoOutliers} = 0.5459739$ by using the Boxcox method on the response variable. We can see from our ggpairs plot that the explanatory variable $ibm2$ has a heavy tale so lets try to transform that variable as well. Here we use a log transformation for the $ibm2$ variable.