02441 Applied Statistics and Statistical Software

Exercise 2C - Process

The dataset process contains measurements of air flow, water temperature, and acid concentration of a process loss.

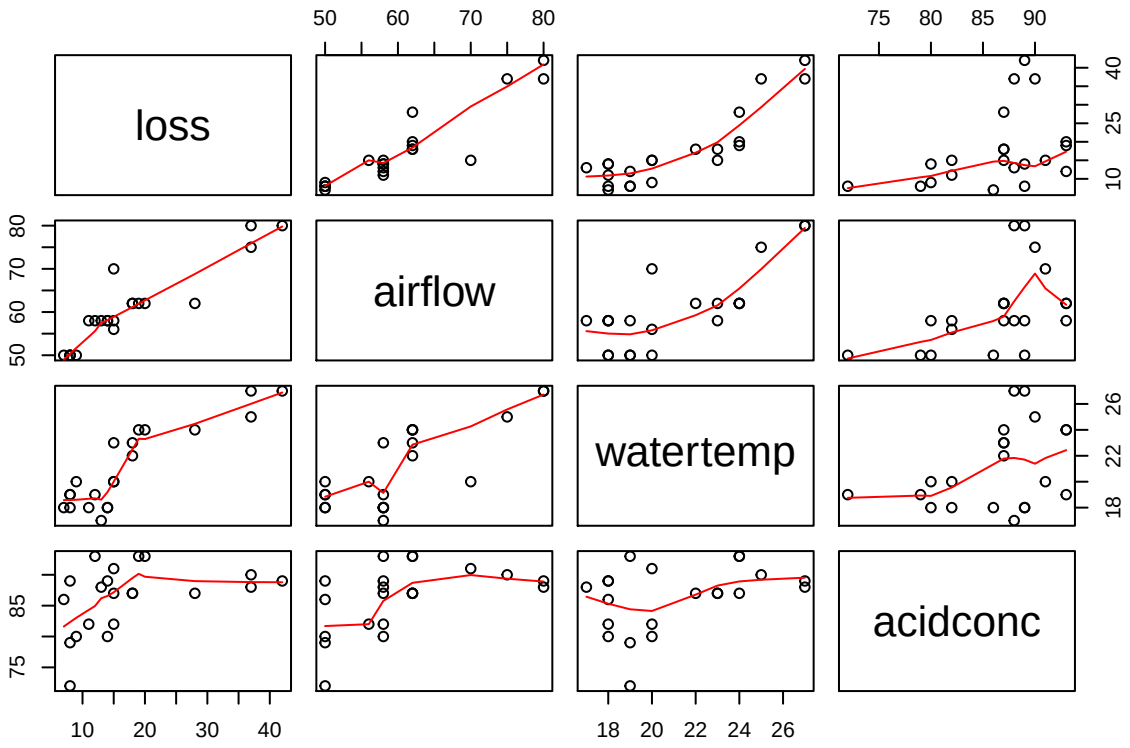
Variable name	Description
loss	loss from process
airflow	air flow
watertemp	temperature of water
acidconc	concentration of acid

1. Determine whether air flow, temperature of water, or concentration of acid influence on the process loss by a graphical comparison

Start by loading the data

```
process <- read.table("process.txt", header=TRUE)

plot(process, panel=panel.smooth)
```



2. Determine whether air flow, temperature of water or concentration of acid influence on the process loss by analysing each variable using simple linear regression

Making simple regression for each variable

```
lmair <- lm(process$loss~process$airflow)
lmwater <- lm(process$loss~process$watertemp)
lmacid <- lm(process$loss~process$acidconc)

summary(lmair)
```

```
##
## Call:
## lm(formula = process$loss ~ process$airflow)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.2896  -1.1272  -0.0459   1.1166   8.8728
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -44.13202    6.10586  -7.228 7.31e-07 ***
## process$airflow   1.02031    0.09995  10.208 3.77e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.098 on 19 degrees of freedom
## Multiple R-squared:  0.8458, Adjusted R-squared:  0.8377
## F-statistic: 104.2 on 1 and 19 DF,  p-value: 3.774e-09
```

```
summary(lmwater)
```

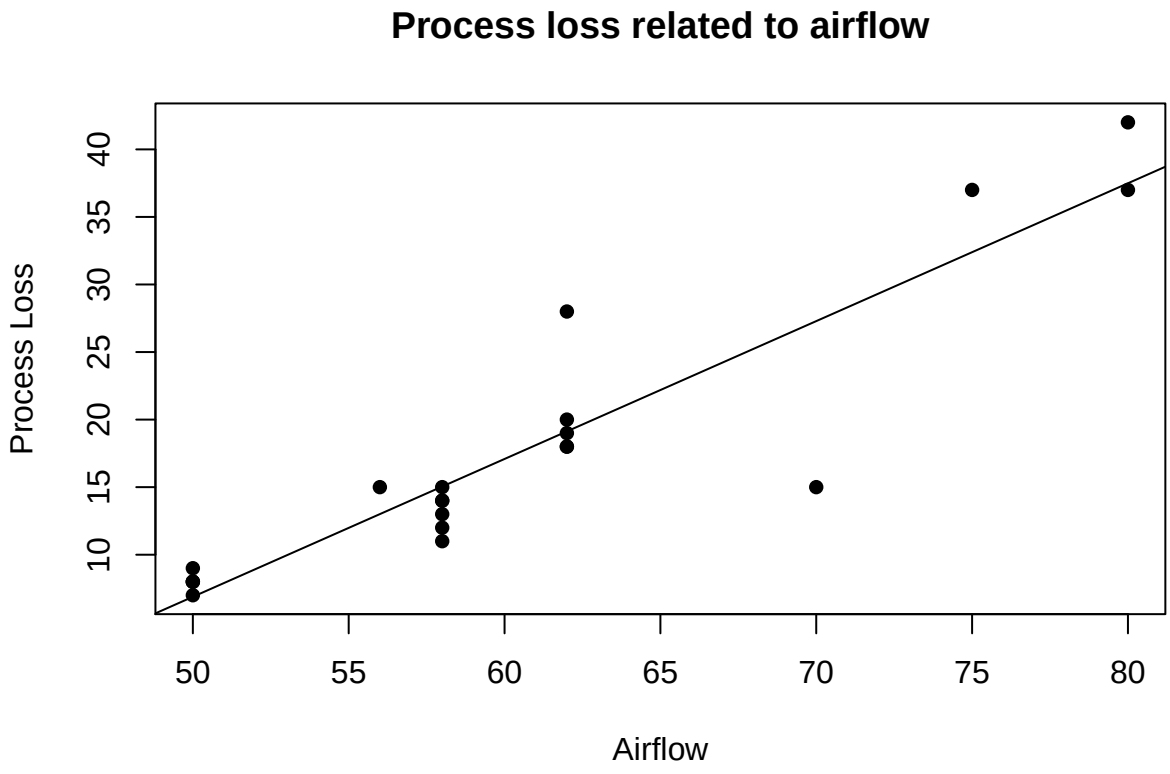
```
##
## Call:
## lm(formula = process$loss ~ process$watertemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.8904 -3.6206  0.3794  2.8398  8.4747
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -41.9109     7.6056  -5.511 2.58e-05 ***
## process$watertemp   2.8174     0.3567   7.898 2.03e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.043 on 19 degrees of freedom
## Multiple R-squared:  0.7665, Adjusted R-squared:  0.7542
## F-statistic: 62.37 on 1 and 19 DF,  p-value: 2.028e-07
```

```
summary(lmacid)
```

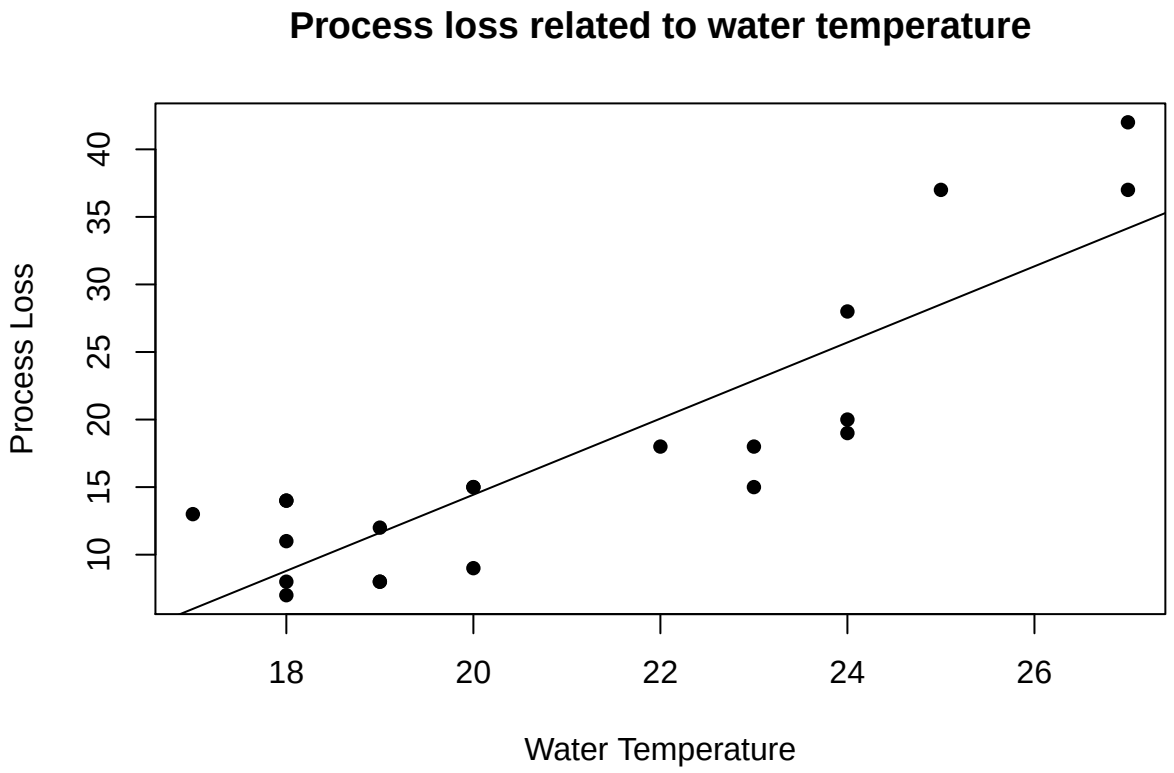
```
##
## Call:
## lm(formula = process$loss ~ process$acidconc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.584  -5.584  -3.066   1.247  22.416
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -47.9632    34.5044  -1.390   0.1806
## process$acidconc   0.7590     0.3992   1.901   0.0725 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.565 on 19 degrees of freedom
## Multiple R-squared:  0.1599, Adjusted R-squared:  0.1156
## F-statistic: 3.615 on 1 and 19 DF,  p-value: 0.07252
```

Plot the scatter plots

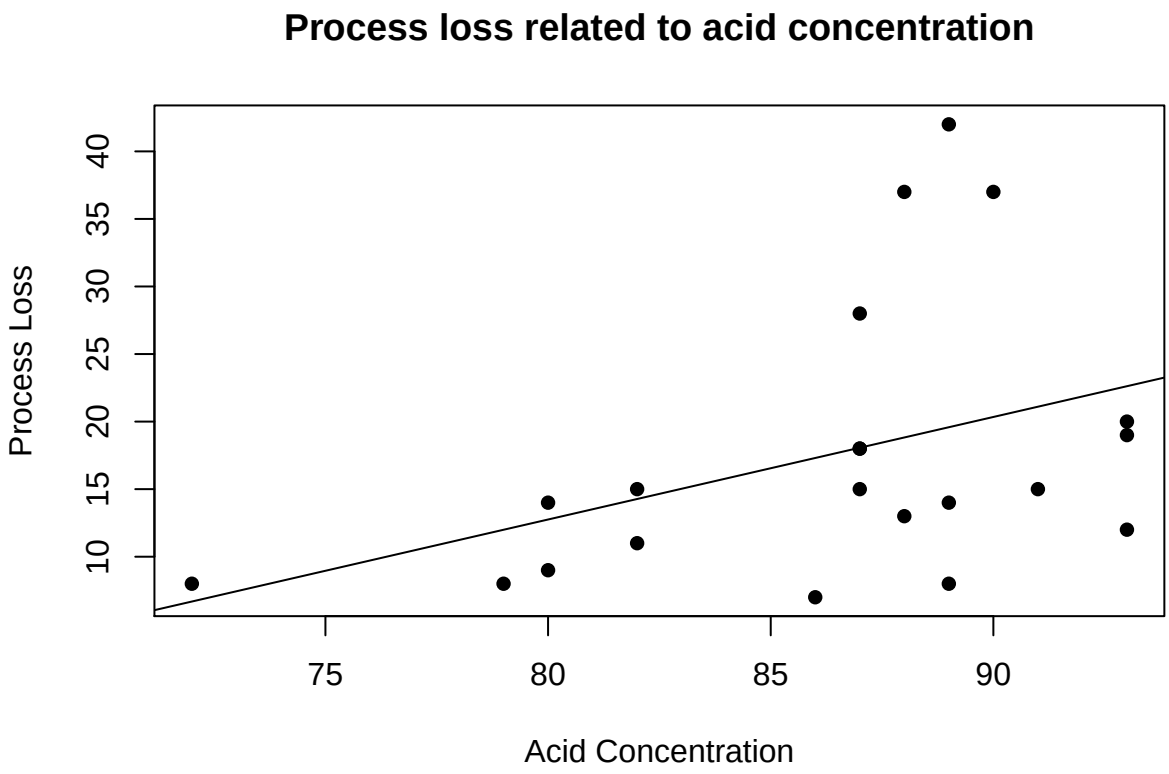
```
plot(process$loss~process$airflow, xlab="Airflow", ylab="Process Loss", main="Process loss related to a
abline(lmair)
```



```
plot(process$loss~process$watertemp, xlab="Water Temperature", ylab="Process Loss", main="Process loss r
abline(lmwater)
```




```
plot(process$loss~process$acidconc, xlab="Acid Concentration", ylab="Process Loss", main="Process loss r
abline(lmacid)
```



It appears like acid concentration doesn't have an influence in process loss because of the p-value in the linear regression and the visual inspection.

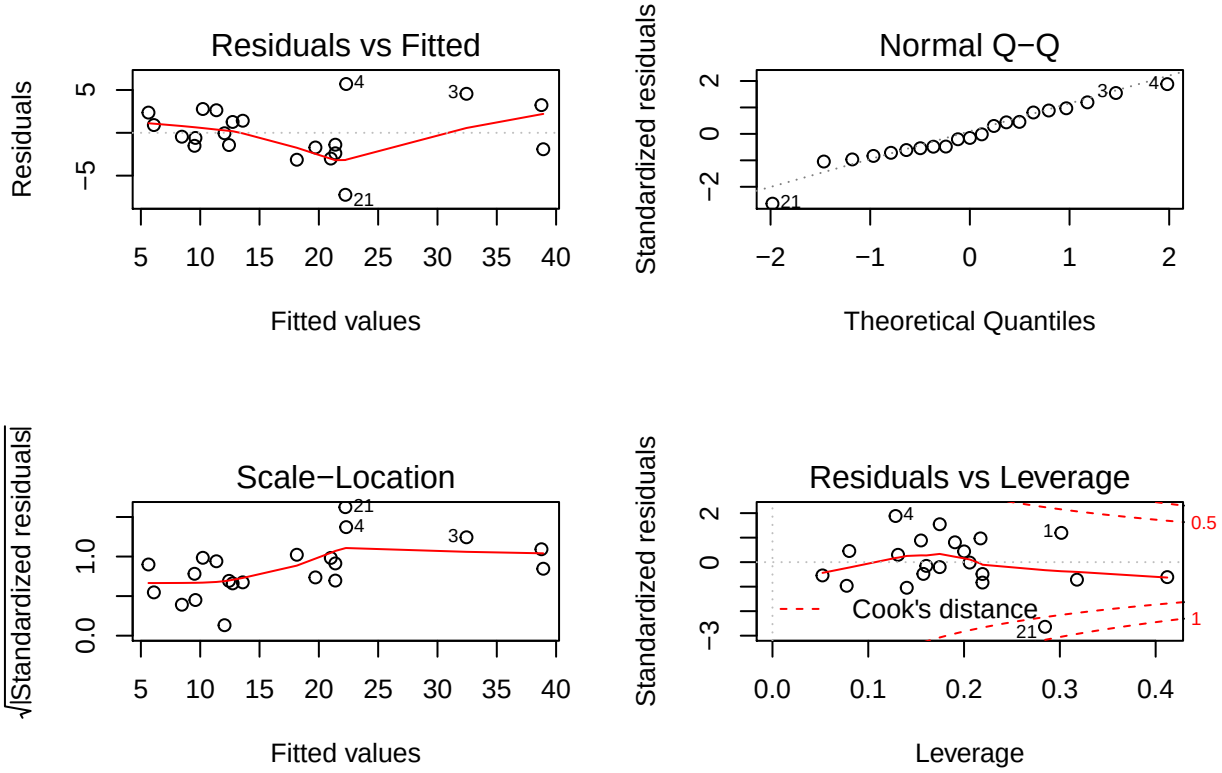
3. Determine whether air flow, temperature of water or concentration of acid influence on the process loss using multiple linear regression

```
lmall <- lm(process$loss~process$airflow+process$watertemp+process$acidconc)
summary(lmall)
```

```
##
## Call:
## lm(formula = process$loss ~ process$airflow + process$watertemp +
##     process$acidconc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.2377 -1.7117 -0.4551  2.3614  5.6978
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -39.9197    11.8960  -3.356  0.00375 **
## process$airflow     0.7156     0.1349   5.307  5.8e-05 ***
## process$watertemp   1.2953     0.3680   3.520  0.00263 **
## process$acidconc   -0.1521     0.1563  -0.973  0.34405
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.243 on 17 degrees of freedom
## Multiple R-squared:  0.9136, Adjusted R-squared:  0.8983
## F-statistic:  59.9 on 3 and 17 DF,  p-value: 3.016e-09
```

```
par(mfrow=c(2,2))
plot(lmall)
```

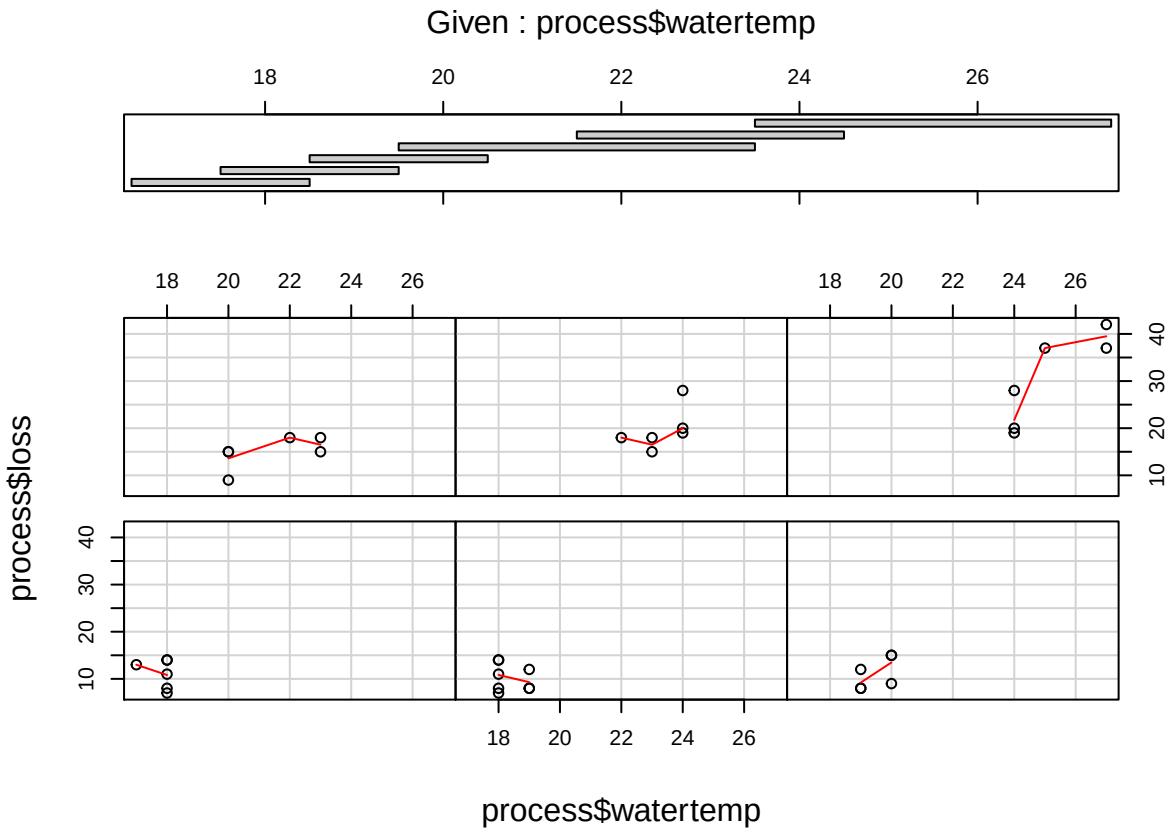


4. Is there evidence of multicollinearity?

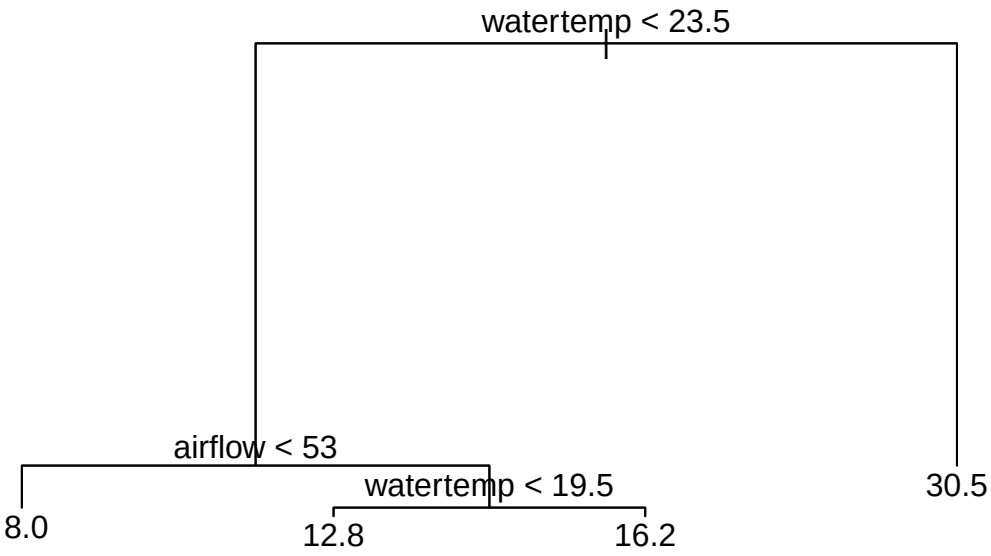
```
cor(process)
```

```
##                loss   airflow watertemp  acidconc
## loss       1.0000000 0.9196635 0.8755044 0.3998296
## airflow    0.9196635 1.0000000 0.7818523 0.5001429
## watertemp  0.8755044 0.7818523 1.0000000 0.3909395
## acidconc   0.3998296 0.5001429 0.3909395 1.0000000
```

```
coplot(process$loss ~ process$watertemp | process$watertemp, process, panel = panel.smooth)
```



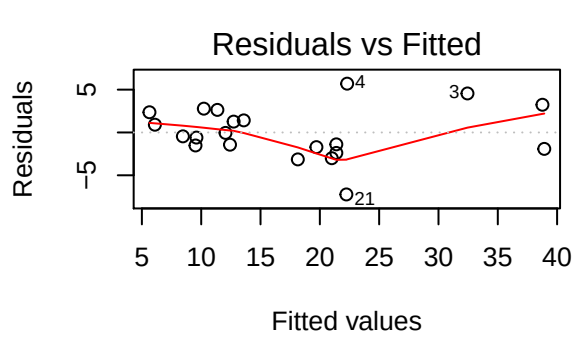

```
model <- tree(process$loss ~ ., process)
plot(model)
text(model)
```

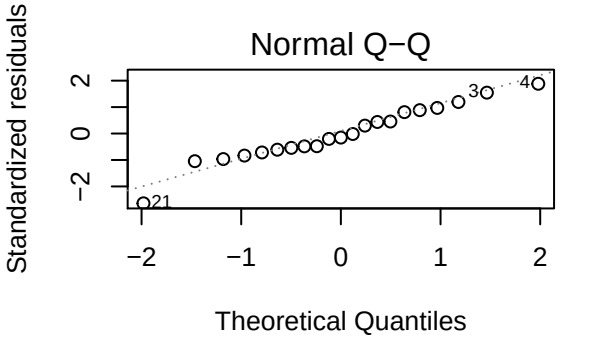


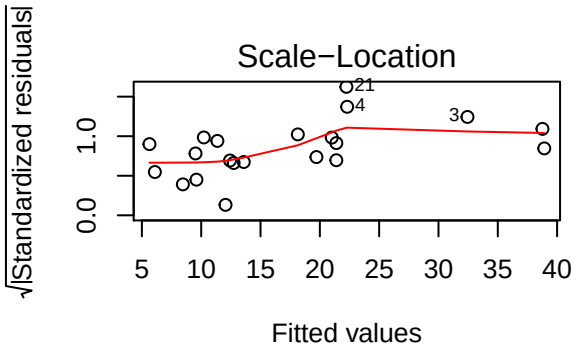
It looks like airflow and water temperature are collinear/positiviely correlated.

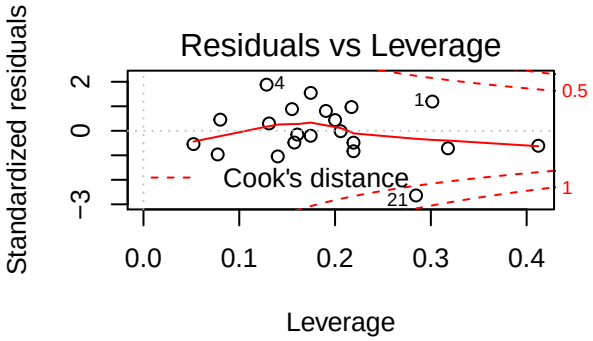
5. Plot the residuals and analyse the results. Which x-variable should be removed if we want to reduce the model?

```
par(mfrow=c(2,2))
plot(lmall)
```







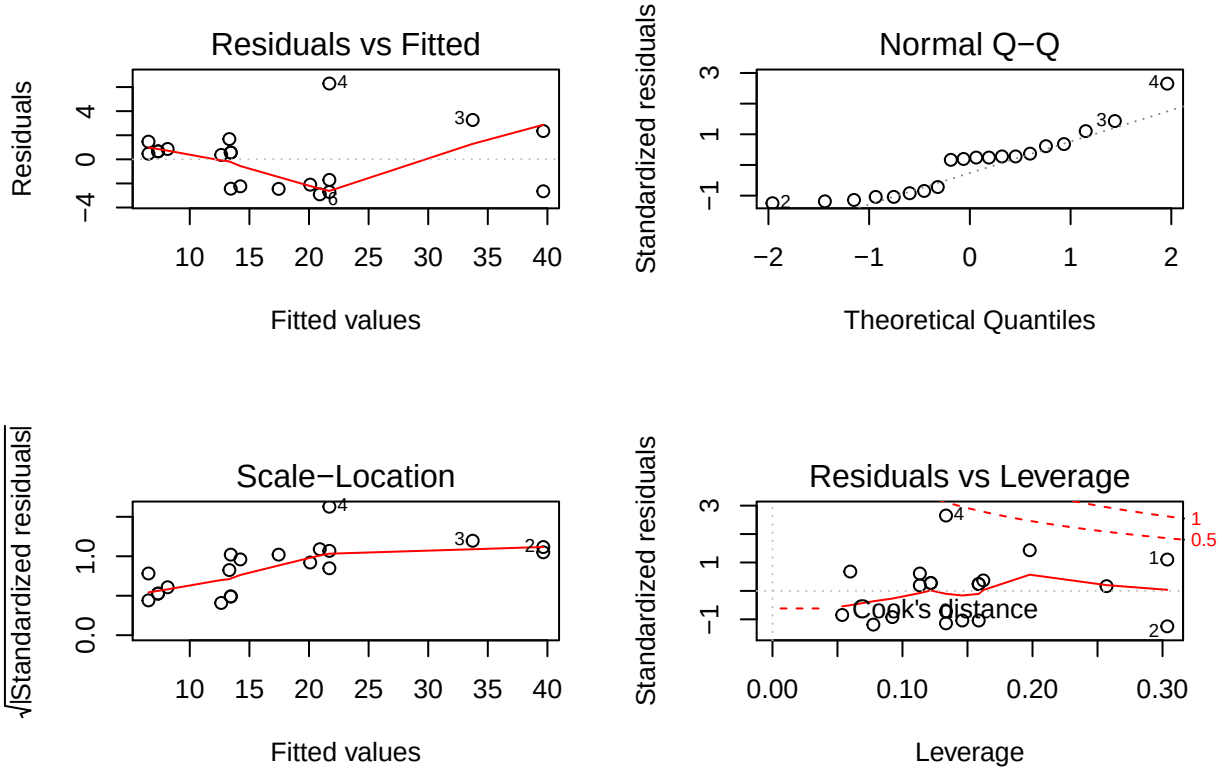



```
# New model without acid concentration
lmall2 <- lm(process$loss~process$airflow+process$watertemp)
summary(lmall2, correlation=TRUE)
```

```
##
## Call:
## lm(formula = process$loss ~ process$airflow + process$watertemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5290 -1.7505  0.1894  2.1156  5.6588
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -50.3588     5.1383  -9.801 1.22e-08 ***
## process$airflow     0.6712     0.1267   5.298 4.90e-05 ***
## process$watertemp   1.2954     0.3675   3.525  0.00242 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.239 on 18 degrees of freedom
## Multiple R-squared:  0.9088, Adjusted R-squared:  0.8986
## F-statistic: 89.64 on 2 and 18 DF,  p-value: 4.382e-10
##
## Correlation of Coefficients:
##                   (Intercept) process$airflow
```

```
## process$airflow   -0.31
## process$watertemp -0.34       -0.78
```

```
# New model without acid concentration and without the outlier
process2 <- process[-21,]
lmall3 <- lm(process2$loss~process2$airflow+process2$watertemp)
par(mfrow=c(2,2))
plot(lmall3)
```

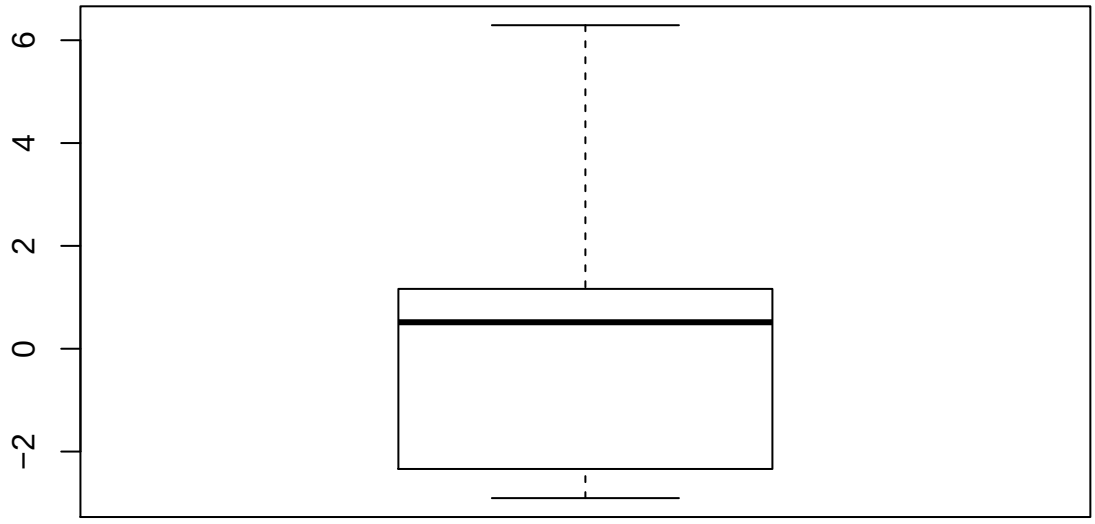



```
summary(lmall3, correlation=TRUE)
```

```
##
## Call:
## lm(formula = process2$loss ~ process2$airflow + process2$watertemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9052 -2.2893  0.5151  1.0123  6.2916
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -51.0760     4.0502 -12.611 4.69e-10 ***
## process2$airflow     0.8630     0.1140   7.568 7.70e-07 ***
## process2$watertemp   0.8033     0.3222   2.493   0.0233 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.549 on 17 degrees of freedom
## Multiple R-squared:  0.9464, Adjusted R-squared:  0.9401
## F-statistic: 150.2 on 2 and 17 DF,  p-value: 1.571e-11
##
## Correlation of Coefficients:
##                     (Intercept) process2$airflow
## process2$airflow    -0.30
## process2$watertemp -0.29        -0.83
```

```
par(mfrow=c(1,1))
boxplot(lmall3$residuals)
```



We should remove acid concentration and the outlier.