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

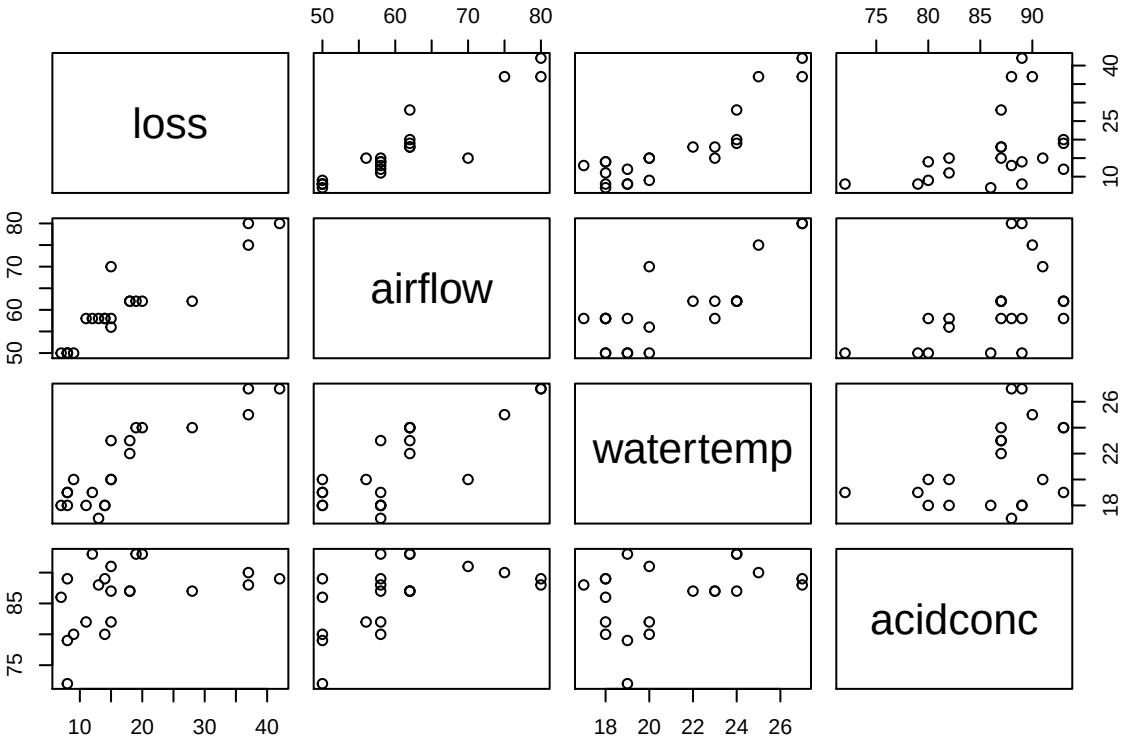
Start by loading the data

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

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

Let's plot the data prior to analysis.

```
pairs(pro)
```



Making simple regression for each variable

airflow

```
lm1 <- lm(loss~airflow, data = pro)
summary(lm1)
```

```
##
## Call:
## lm(formula = loss ~ airflow, data = pro)
##
## 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 ***
## 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
```

Airflow is significant.

watertemp

```
lm2 <- lm(loss~watertemp, data = pro)
summary(lm2)
```

```
##
## Call:
## lm(formula = loss ~ watertemp, data = pro)
##
## 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 ***
## 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
```

Watertemp is significant.

acidconc

```
lm3 <- lm(loss~acidconc, data = pro)
summary(lm3)
```

```
##
## Call:
## lm(formula = loss ~ acidconc, data = pro)
##
## 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
## 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
```

Acidconc is not significant.

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

```
lm4 <- lm(loss~., data = pro)
summary(lm4)
```

```
##
## Call:
## lm(formula = loss ~ ., data = pro)
##
## 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 **
## airflow       0.7156     0.1349   5.307  5.8e-05 ***
## watertemp     1.2953     0.3680   3.520  0.00263 **
## 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
```

4. Is there evidence of multicollinearity?

```
library(kableExtra)
kable(cor(pro))
```

	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

The pairwise correlations as seen in the table above indicate multicolinearity. Several independent variables appear to be highly correlated to each other.

Another way to investigate multicolinearity is to look at the correlation of the estimated regression parameters.

```
summary(lm4, cor = TRUE)
```

```
##
## Call:
## lm(formula = loss ~ ., data = pro)
##
## 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 **
## airflow       0.7156     0.1349   5.307  5.8e-05 ***
## watertemp     1.2953     0.3680   3.520  0.00263 **
## 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
##
## Correlation of Coefficients:
##           (Intercept) airflow watertemp
## airflow    0.18
## watertemp -0.15       -0.74
## acidconc  -0.90       -0.34    0.00
```

The regression parameters/coefficients indicate high correlation, especially for Watertemp and Airflow. It is interesting that the intercept is correlated to acidconc. This correlation is due to the large distance of the acidconc measurements to zero. The correlation dissappears when acidconc is centered around 0.

```r
lm5 <- lm(loss~ I(airflow-mean(airflow))+
                I(watertemp-mean(watertemp))+
                I(acidconc-mean(acidconc)), data = pro)

summary(lm5, cor = TRUE)
```

```
##
## Call:
## lm(formula = loss ~ I(airflow - mean(airflow)) + I(watertemp -
##     mean(watertemp)) + I(acidconc - mean(acidconc)), data = pro)
##
## 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)                     17.5238     0.7078  24.760 8.91e-15 ***
## I(airflow - mean(airflow))       0.7156     0.1349   5.307 5.80e-05 ***
## I(watertemp - mean(watertemp))   1.2953     0.3680   3.520  0.00263 **
## I(acidconc - mean(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
##
## Correlation of Coefficients:
##                                (Intercept) I(airflow - mean(airflow))
## I(airflow - mean(airflow))      0.00
## I(watertemp - mean(watertemp))  0.00       -0.74
## I(acidconc - mean(acidconc))    0.00       -0.34
##                                I(watertemp - mean(watertemp))
## I(airflow - mean(airflow))
## I(watertemp - mean(watertemp))
## I(acidconc - mean(acidconc))    0.00
```

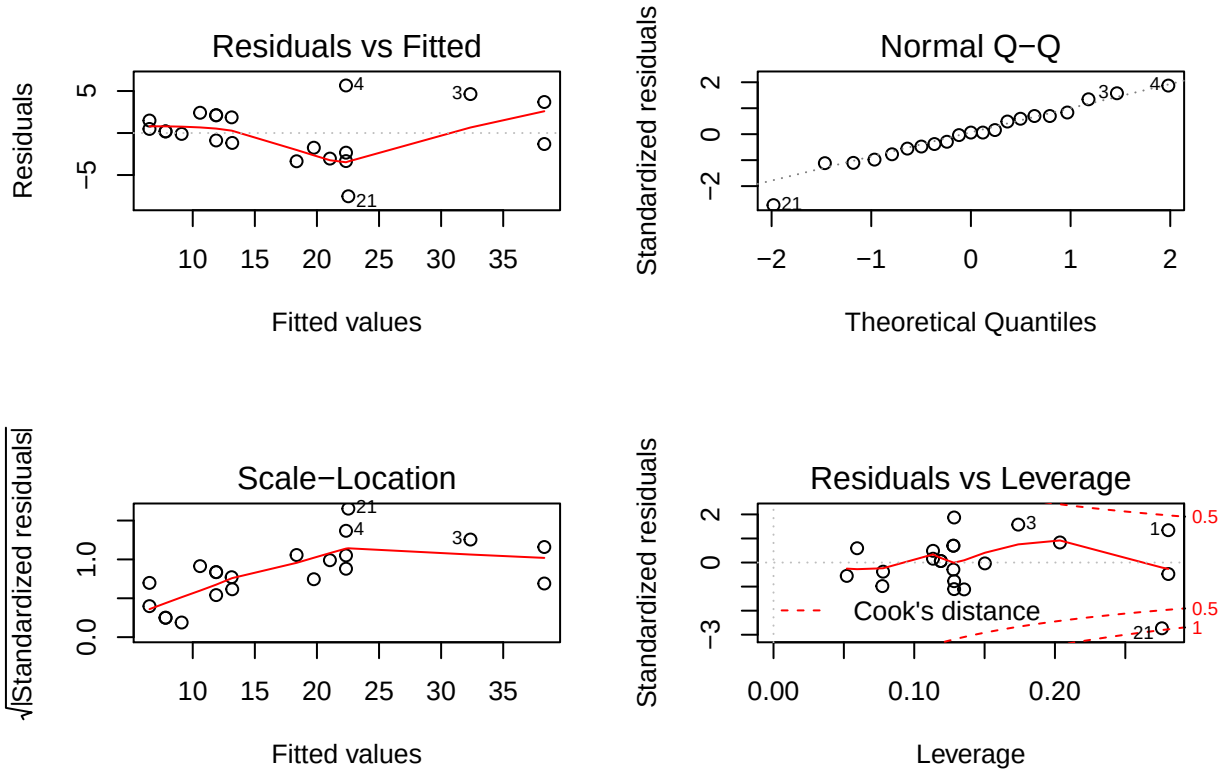
5. Plot the residuals and analyse the results. Which x-variable should be removed if we want to reduce the model?

```r
lm4b <- update(lm4, .~.-acidconc)
summary(lm4b)
```

```
##
## Call:
## lm(formula = loss ~ airflow + watertemp, data = pro)
##
## 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 ***
## airflow       0.6712     0.1267   5.298 4.90e-05 ***
## 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
```

```
par(mfrow=c(2,2))
plot(lm4b)
```



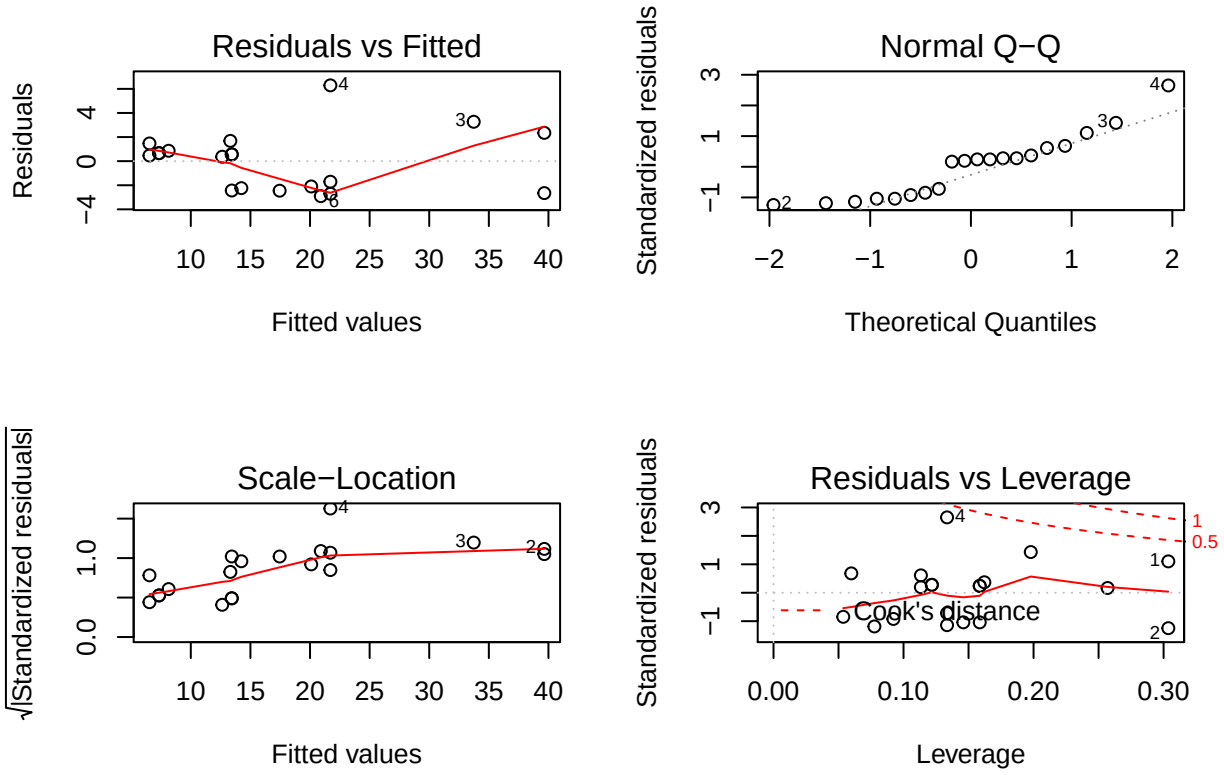
Sample 21 has quite high Cook's distance. Remove and re-model.

```
lm4c <- lm(loss~airflow+watertemp, data = pro[-c(21),])
summary(lm4c)
```

```
##
## Call:
## lm(formula = loss ~ airflow + watertemp, data = pro[-c(21), ])
##
## 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 ***
## airflow       0.8630     0.1140   7.568 7.70e-07 ***
## 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
```

```
par(mfrow=c(2,2))
plot(lm4c)
```



One could also consider to remove data points 4, 1 and 2. However, due to low number of measurements this is done here.