Advanced Dataanalysis and Statistical Modelling - Assignment 1: Dioxin Emission

Anders Launer Bæk (s160159)

10 Marts 2018

There are 5 incomplete rows within the given data. It has been chosen to exclude these rows in order to get a tidy data set. There are several common approaches to comprehend incomplete observations; one would be to replace the value by the expected mean value; another would be to sort the concentrations of `dioxin` and then perform linear interpolation.

The tidy data set includes 52 observations for further modelling.

This assignment is based upon in-sample modelling and therefore any kind of train, validate and test approaches has not been considered.

Q1

Figure 1 visualizes the `dioxin` concentration for the 52 observations.

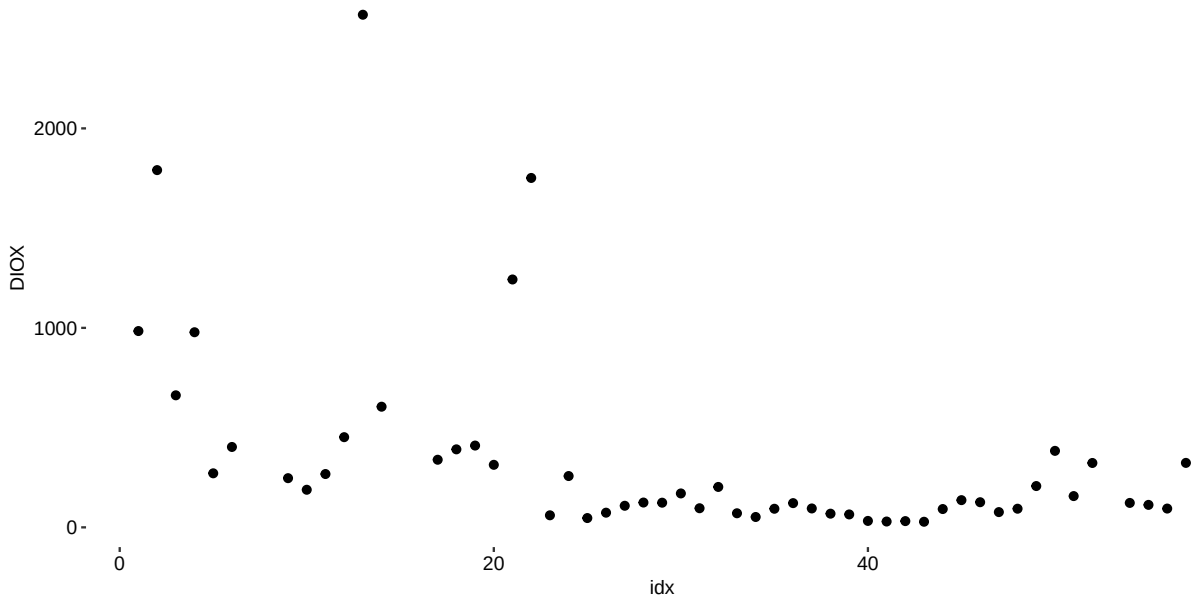

Figure 1: Concentration as function of observation.

As illustrated in the plot above there are few a odd observations. The `dioxin` concentrations above `1000` does not fit into the concentration pattern.

The output below illustrate selected statistics of the `dioxin` concentrations.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   28.13   92.96  146.62  347.36  350.09 2569.67
```

It is clear to see the great difference between the median and the mean value. There is a percentage gap of: 57.789%. The gap is an identification of a right skewed distribution.

This intuition is supported by the histogram plot in figure 2. The number of bins have been selected according to Freedman–Diaconis rule (1).

$$no\_bin\,(x,n) = 2\frac{IQR\,(x)}{\sqrt[3]{n}} \tag{1}$$

where x is vector of elements and n is the number of elements.

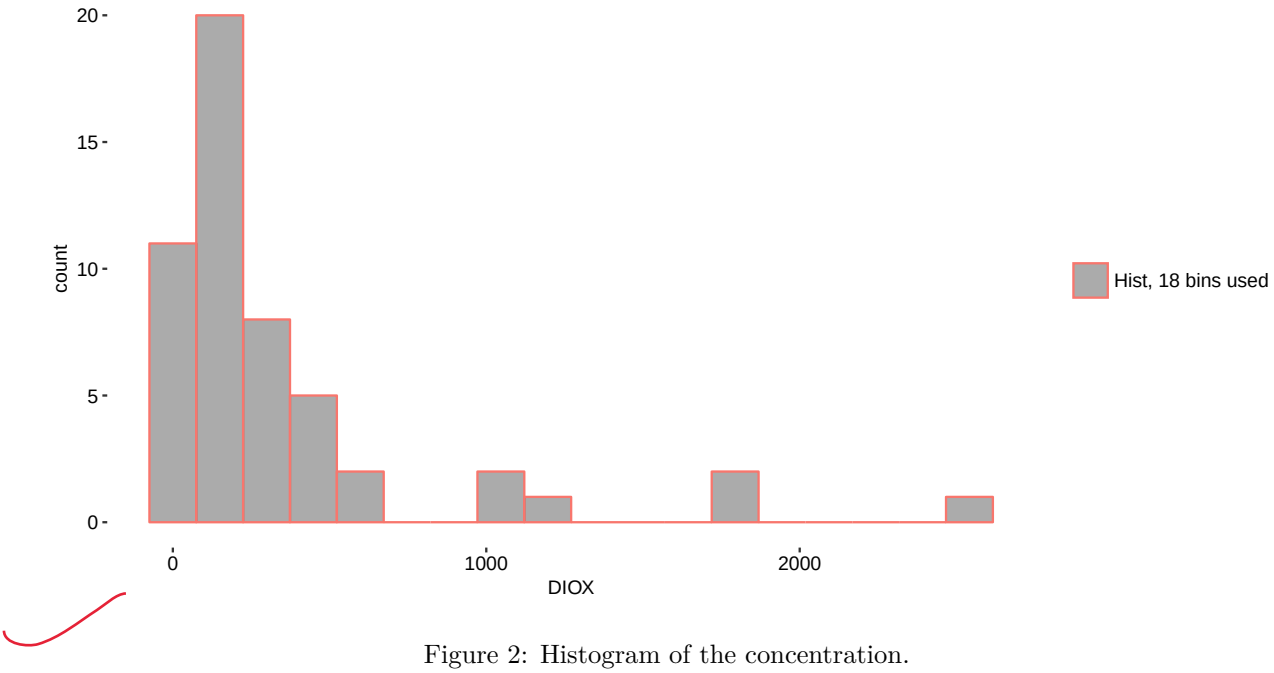

Figure 2: Histogram of the concentration.

The distribution of the `dioxin` concentrations are clearly right skewed in figure 2. Recalling lecture 1, this is normal for variables which only can be positive such as concentrations. The solution to improve this is to take the log transformation.

Figure 3 and figure 4 shows the log distribution which fits the normal distribution better.

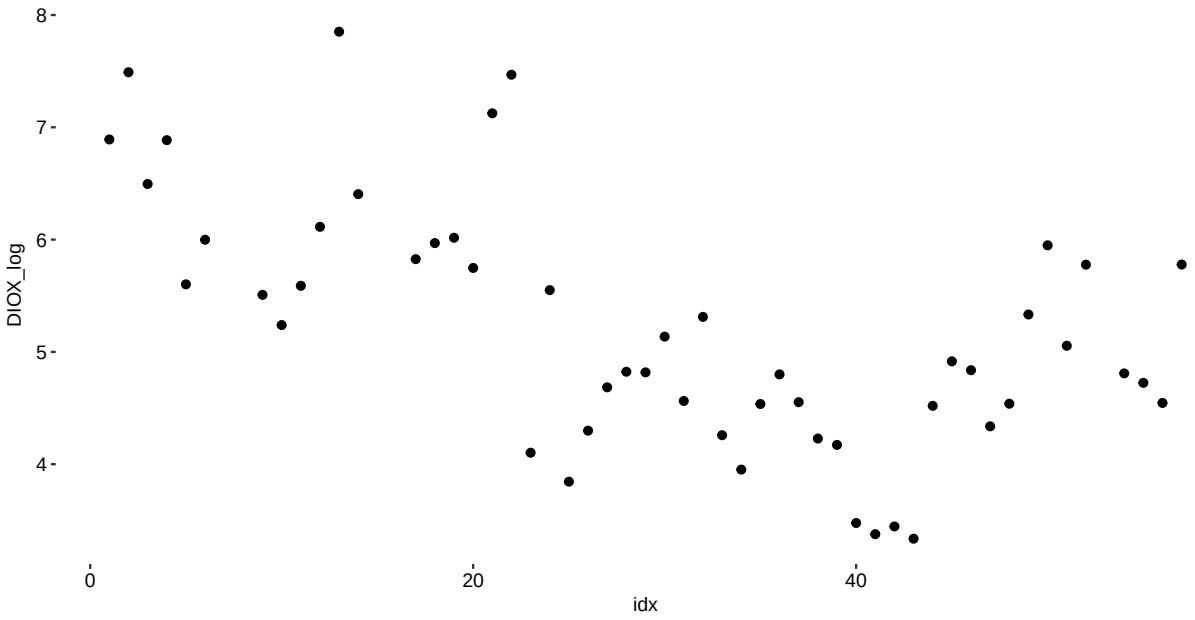

Figure 3: Log transformed concentration as a function of observation.

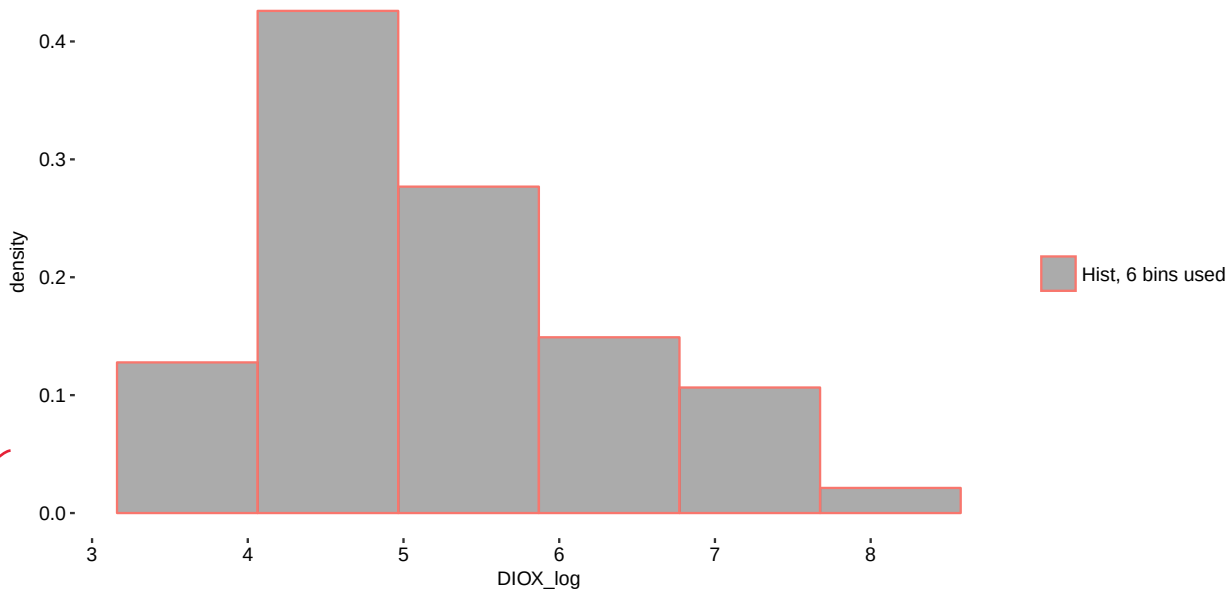

Figure 4: Histogram of the log transformed concentration.

The log distribution of the concentrations is not perfect but good enough for now. The output below shows a much lower percentage difference between the value of the median and of the value of the mean.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.337   4.532   4.985   5.204   5.857   7.852
```

The percentage difference is now: 4.2% after the log transformation.

Analysis of variables

Recalling the description of the variables in the assignment:

- Dependent variable
  - `DIOX`
- Block variables
  - `PLANT`
  - `TIME`
  - `LAB`
- Active variables
  - `OXYGEN` Oxygen surplus in gas
  - `LOAD` Plant load
  - `PRSEK` Air distribution (primary/secondary)
  - `O2`, `O2COR`
  - `NEFFEKT`
  - `QRAT`
- Passive variables
  - `QROEG` Gas flow (m3/h)
  - `TOVN` Combustion chamber temperature (oC)
  - `TROEG` Gas temperature (oC)
  - `POVN` Pressure in the chamber
  - `CO2` (ppm)
  - `CO` (ppm)
  - `SO2` (mg/m3)
  - `HCl` (mg/m3)
  - `H2O` (%)

Block variables

The following three plots in figure 5 can be used to see whether the given variable can be used as a good explanatory variable. If the distribution for a given level of a block variable is different compared to the distributions of the other levels, the block variable may be a suitable explanatory variable for the model. If the distribution of the levels in the block variable are similar, the block variable does only provide a bias which will be corrected in the intercept. It causes a simpler model when excluding such a variable.

Figure 5 shows multiple box whisker plots of the `PLAT`, `TIME` and `LAB` variables.

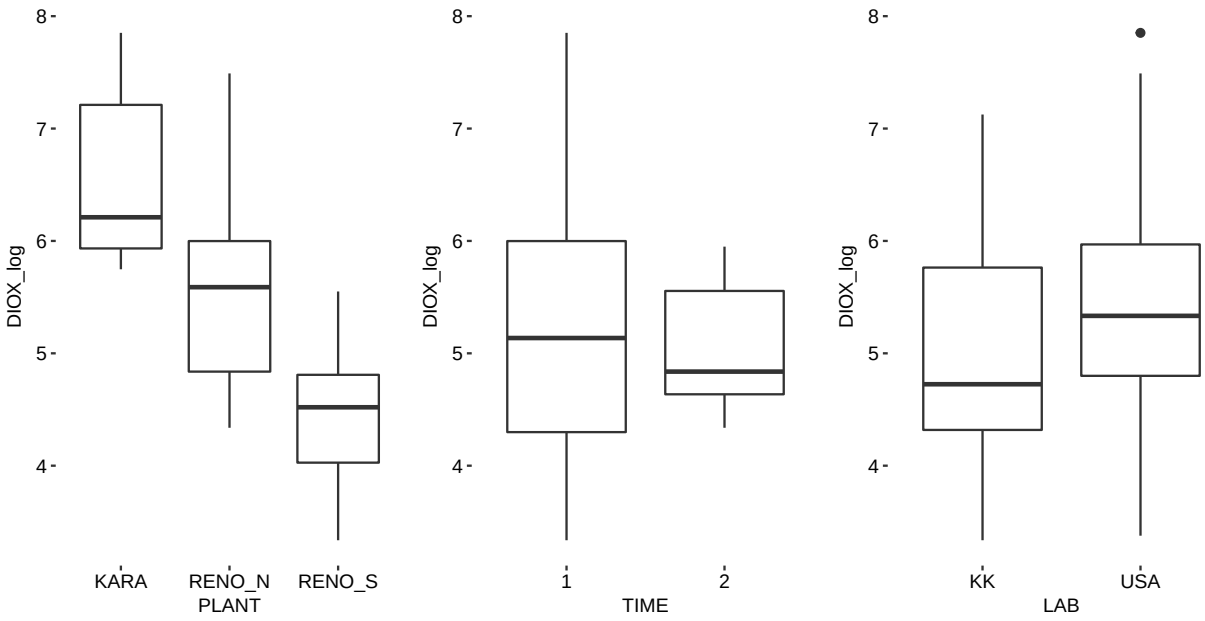

Figure 5: Box plot of the log transformed dioxin concentrations as a function of PLANT, TIME and LAB.

The measured log transformed concentrations does clearly depend on the given `PLAT`. This can identicate that the `PLANT` variable can be a good explanatory variable.

The difference between the two box whisker plots in figure 5(TIME) shows a much higher similarity despite the larger range when `TIME = 1`. The 25, 50 and 75 quantiles are closer compared to the previous plot in figure 5(PLANT).

The same previously mentioned pattern is again illustrated in the `LAB` variable. The variance of the `LAB=USA` is also larger compared to `LAB=KK`. See figure 5(LAB).

All block variables will be included in the initial "simple additive model". The archived knowledge from figure 5(PLANT) to figure 5(LAB) can be used in the backwards selection when reducing the model.

Passive variables

Figure 6 shows the correlation between the passive variables.

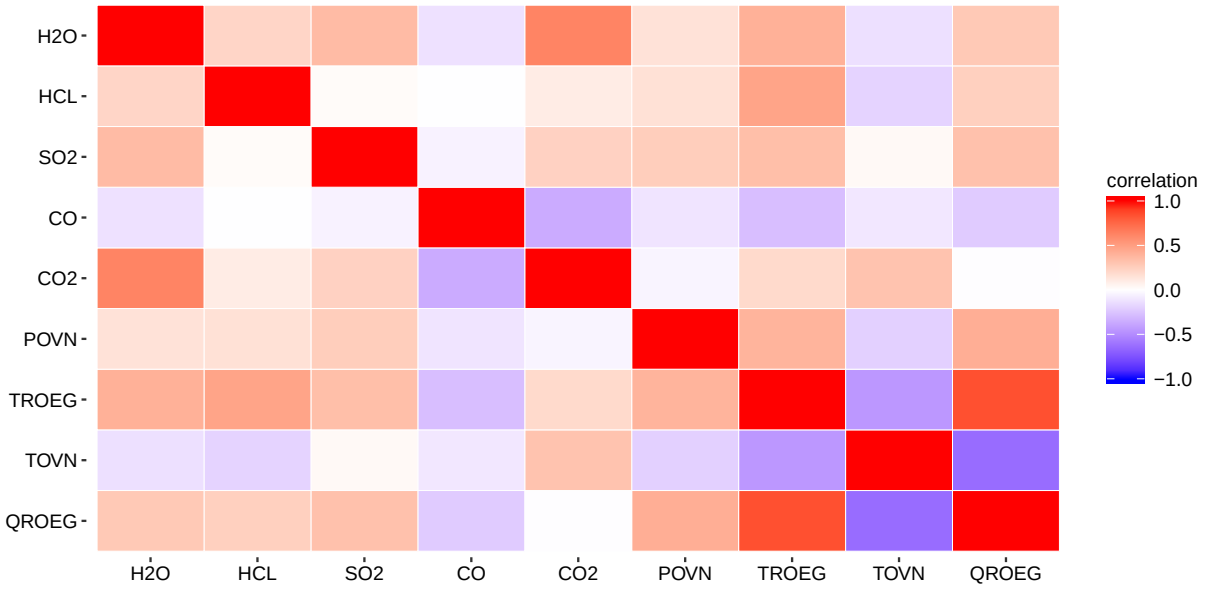

Figure 6: Correlation between passive variables.

As mentioned in previous section it is not wanted to have multiple variables which provide the same level of information. The output below reports a table with "head(5)" of the passive variables with the highest absolute correlation.

```
##    Var1  Var2 abs_value
## 1 TROEG QROEG 0.8416684
## 2  TOVN QROEG 0.6353504
## 3   H2O   CO2 0.6253262
## 4   HCL TROEG 0.4727127
## 5 TROEG  TOVN 0.4489704
```

`TROEG` and `QROEG`, `TOVN` and `QROEG`, `H2O` and `CO2` does have a high correlation close to |1|. The high correlations can be supported by the physical relations among the variables.

Figure 7 visualizes the distributions of all the passive variables. All the passive variables have a concentration unit apart from the following variables: TOVN, TROEG, POVN.

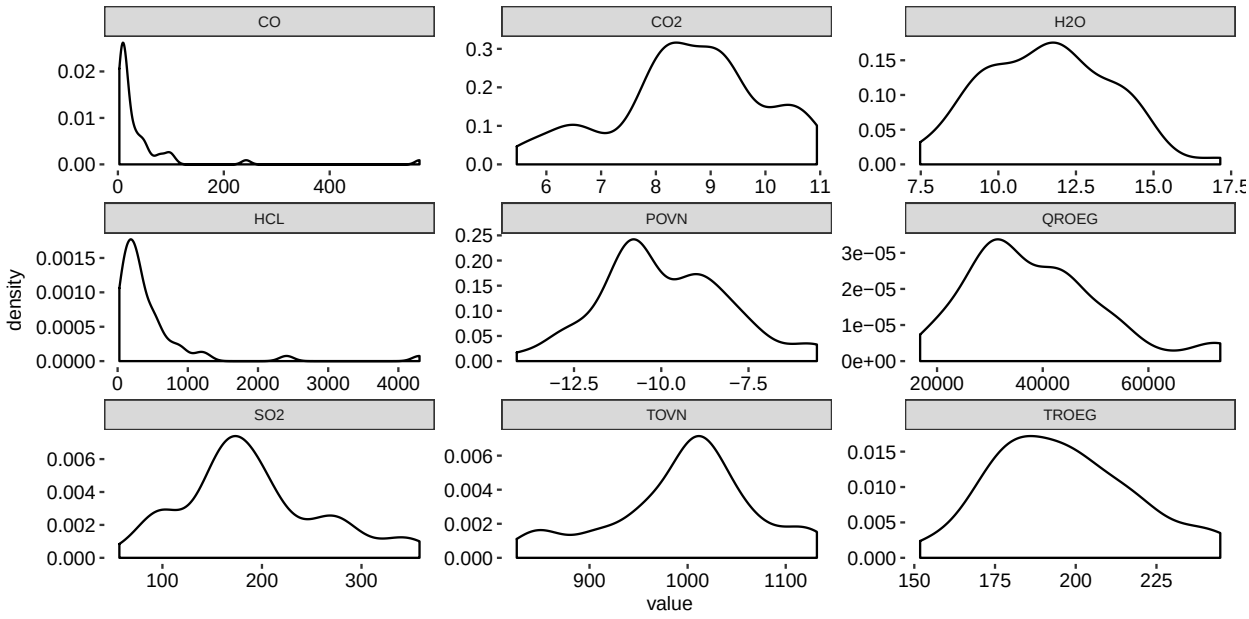

Figure 7: Distribution plot of the passive variables.

Recalling from lecture one, a log transformation of these passive variables will remove the right skeweness of their distributions. The passive variables has been divided into two groups; TOVN, TROEG, POVN and QROEG, CO2, CO, SO2, HCL, H2O.

Figure 8 visualizes a new distribution plot of the latter of the before mentioned groups.

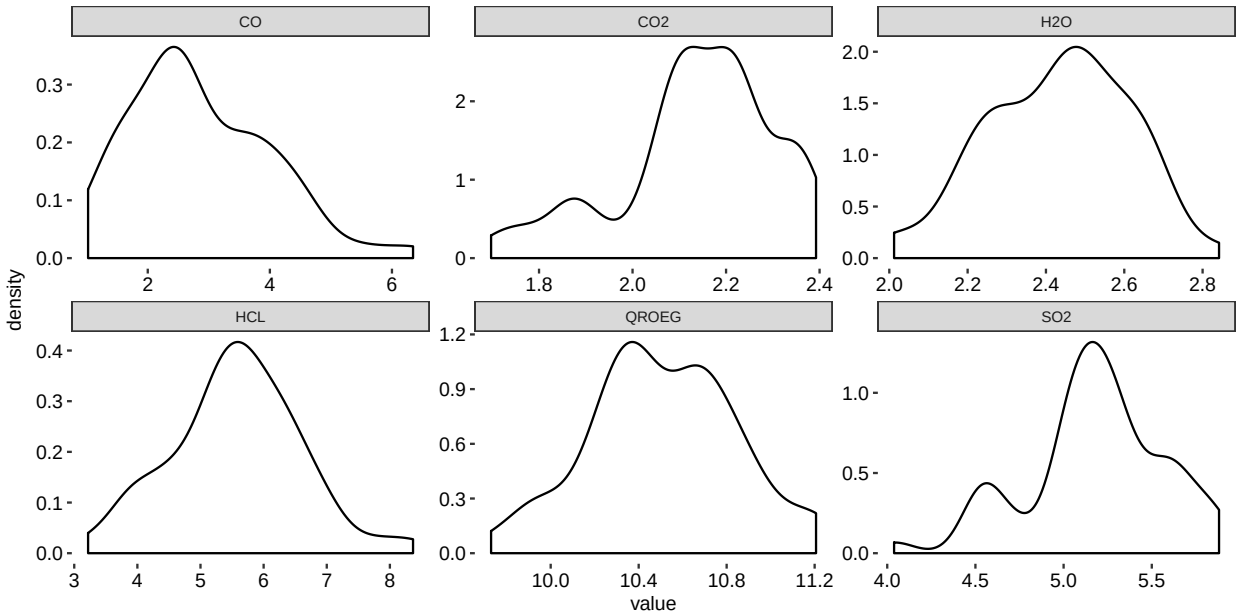

Figure 8: Distribution plot of the log transformed passive variables.

Active variables

Figure 9 shows the correlation between the numeric active variables.

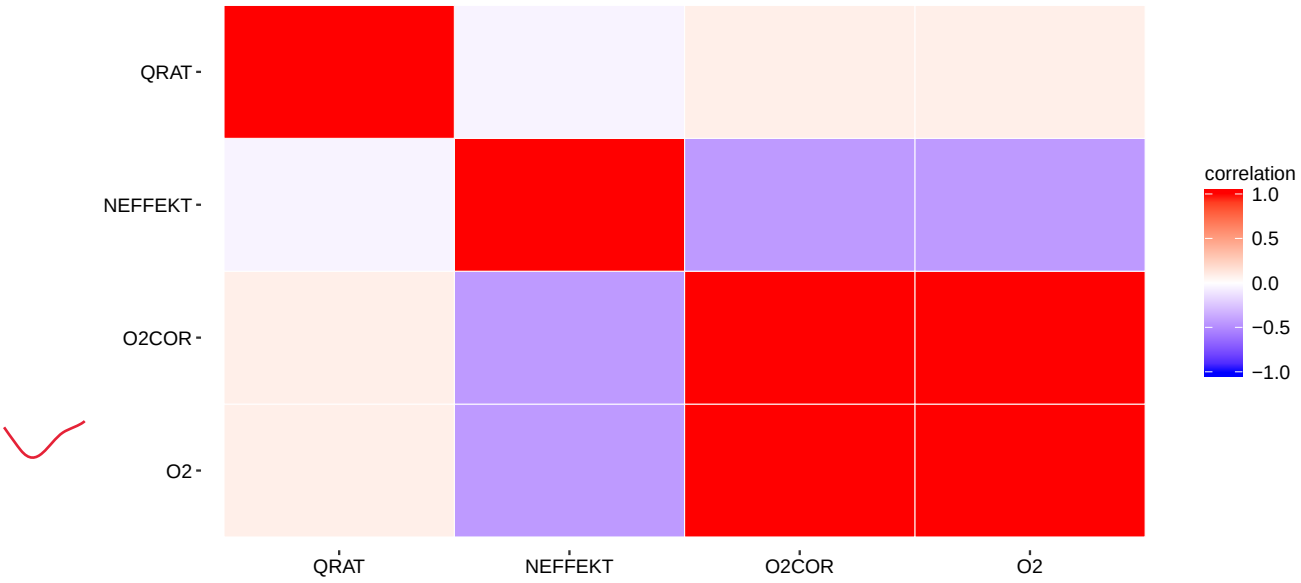

Figure 9: Correlation between passive variables.

Figure 10 visualizes a 3x3 box plots of `log_DIOX` as a function of the levels in `OXYGEN`, `LOAD` and `PRSEK`.

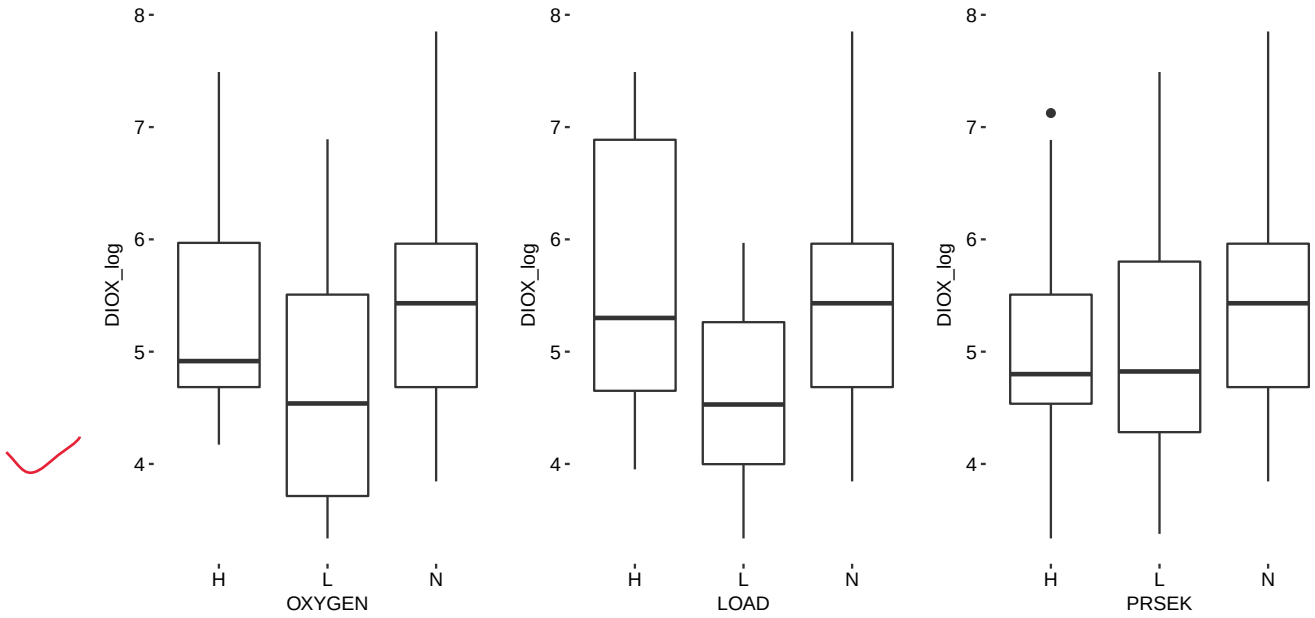

Figure 10: Box plot of the log transformed dioxin concentrations as a function of OXYGEN, LOAD and PRSEK.

Figure 11 shows the distribution plots of the numeric active variables.

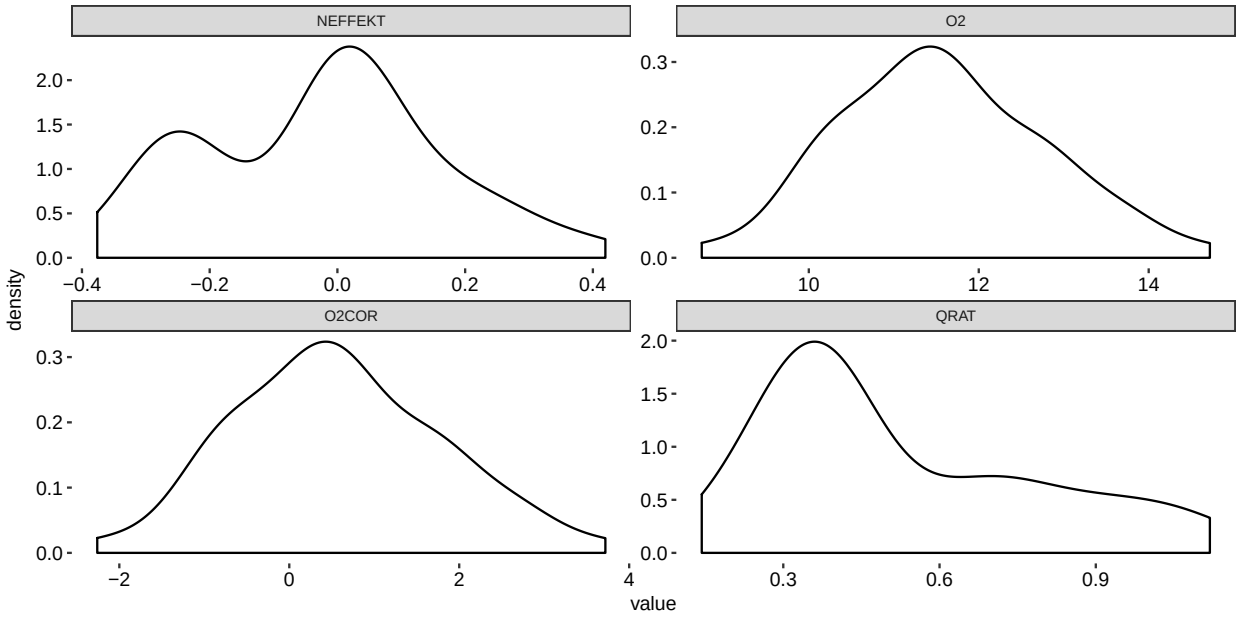

Figure 11: Distribution plot of the numeric active variables.

Q2

The following variables have been considered in the first simple additive model:

- OXYGEN
- LOAD
- PRSEK
- O2
- O2COR
- NEFFEKT
- QRAT
- PLANT
- TIME
- LAB

The notation of the additive model is given in (2).

$$Y_{DIOX_{log}} = X\beta + \epsilon \tag{2}$$

where β are the parameters which we want to estimate and X is the design matrix given below:

$$X = \begin{bmatrix} 1 & OXYGEN & LOAD & PRSEK & \cdots & PLANT & TIME & LAB \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

The underlying asspumtions for the linear model are as follows:

- Ther must be a linear and additive relation between the dependent variable and the explanatory varialbes. The effects of multiple explanatory varialbes are additive.
- The residuals must be independent and must not have any relation to the consecutive residauls.
- The residuals must be have a constant variance over time, as a function of the independent variable and as a function of the any of the explanatory variables.

- The residuals must follow: $\epsilon \sim \mathcal{N}(0, 1)$.

The systematic variance explained by the model may be in a higher degree biased and the prediction, confidence intervals may be misleading if any of these underlying asspumtions is violated.

β can be estimated by using several apporaches. It has been chosen to use the native `lm()` function in order to estimate the wanted β despite it is possible to estimate β in the following manner: $\widehat{\beta} = (X'X)^{-1} X'y$. The model object from `lm()` does provide more informative statistics out of the box.

```
## Single term deletions
##
## Model:
## DIOX_log ~ 1 + OXYGEN + LOAD + PRSEK + O2 + O2COR + NEFFEKT +
##     QRAT + PLANT + TIME + LAB
##         Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                8.468 -70.379
## OXYGEN   1    0.0674  8.535 -71.967  0.3182 0.5758707
## LOAD     1    0.0218  8.490 -72.245  0.1031 0.7498126
## PRSEK    1    0.0150  8.483 -72.286  0.0711 0.7911343
## O2       0    0.0000  8.468 -70.379
## O2COR    0    0.0000  8.468 -70.379
## NEFFEKT  1    0.7279  9.196 -68.091  3.4382 0.0710873 .
## QRAT     1    0.2185  8.686 -71.054  1.0321 0.3157634
## PLANT    2   26.9440 35.412   0.022 63.6389 3.736e-13 ***
## TIME     1    3.0174 11.485 -56.530 14.2534 0.0005196 ***
## LAB      1    1.7948 10.263 -62.383  8.4781 0.0058519 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output above shows the `drop1()` table of the fitted initial model.

A brief overview shows that the majority of the variables does not have a significant parameter estimate. `O2` and `O2COR` does not have an estimate at all. This the due to their multicollinearity. The correlation is illustrated in figure 9. A suitable counter approach would be to use Ridge regression or other regularization techniques which are shrinking the variance of the parameter or simply just remove one of the highly correlated variables.

Model reduction

It is needed to do a model reduction in order to optimize and simplify the model. This is an iterative process where one parameter will be removed at a time and the model will be re-estimated.

The iterative process

It has been chosen to follow the principle of backward elimination and using the highest value of `Pr(>F)` as an exclution criteria until all parameters have a significant estimate (`Pr(>F) < 0.05`). The `drop1(fit, test = "F")` function has been applied in order to comprehend the backward elimination.

The following outlines the backward elimination approach:

- Remove parameter: `O2` due to high collinearity between `O2` and `O2COR` (see figure 9).

```
## Single term deletions
##
## Model:
## DIOX_log ~ 1 + OXYGEN + LOAD + PRSEK + O2 + O2COR + NEFFEKT +
##     QRAT + PLANT + TIME + LAB
```

```
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                 8.468 -70.379
## OXYGEN    1    0.0674  8.535 -71.967  0.3182 0.5758707
## LOAD      1    0.0218  8.490 -72.245  0.1031 0.7498126
## PRSEK     1    0.0150  8.483 -72.286  0.0711 0.7911343
## O2        0    0.0000  8.468 -70.379
## O2COR     0    0.0000  8.468 -70.379
## NEFFEKT   1    0.7279  9.196 -68.091  3.4382 0.0710873 .
## QRAT      1    0.2185  8.686 -71.054  1.0321 0.3157634
## PLANT     2   26.9440 35.412   0.022 63.6389 3.736e-13 ***
## TIME      1    3.0174 11.485 -56.530 14.2534 0.0005196 ***
## LAB       1    1.7948 10.263 -62.383  8.4781 0.0058519 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Remove parameter: `PRSEK`
- Remove parameter: `LOAD`
- Remove parameter: `OXYGEN`
- Remove parameter: `QRAT`

```
## Single term deletions
##
## Model:
## DIOX_log ~ O2COR + NEFFEKT + PLANT + TIME + LAB
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                 8.802 -78.367
## O2COR     1     3.226 12.028 -64.130  16.492 0.0001928 ***
## NEFFEKT   1    10.198 19.000 -40.353  52.140 4.799e-09 ***
## PLANT     2    32.549 41.351  -1.916  83.205 7.621e-16 ***
## TIME      1     2.998 11.800 -65.125  15.326 0.0003037 ***
## LAB       1     2.103 10.905 -69.224  10.754 0.0020122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final model

The model in (3) can be conducted from the output above.

$$Y_{DIOX_{log}} = \begin{bmatrix} 1 & O2COR & NEFFEKT & PLANTRENO_N & PLANTRENO_S & TIME2 & LABUSA \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} 6.2682 \\ 0.2346 \\ 2.7591 \\ -0.5653 \\ -2.1133 \\ -0.7777 \\ 0.4031 \end{bmatrix} + \epsilon \tag{3}$$

Table 1 reports the estimated parameters and their belonging standard errors.

Table 1:

	Estimate	Std. Error
(Intercept)	6.2682361	0.1686851

	Estimate	Std. Error
O2COR	0.2345724	0.0577616
NEFFEKT	2.7590847	0.3821015
PLANTRENO_N	-0.5652863	0.2122601
PLANTRENO_S	-2.1133349	0.1846056
TIME2	-0.7776870	0.1986512
LABUSA	0.4031498	0.1229361

A residual analysis has been conducted in order to check the assumptions from the modelling part, see figure 12.

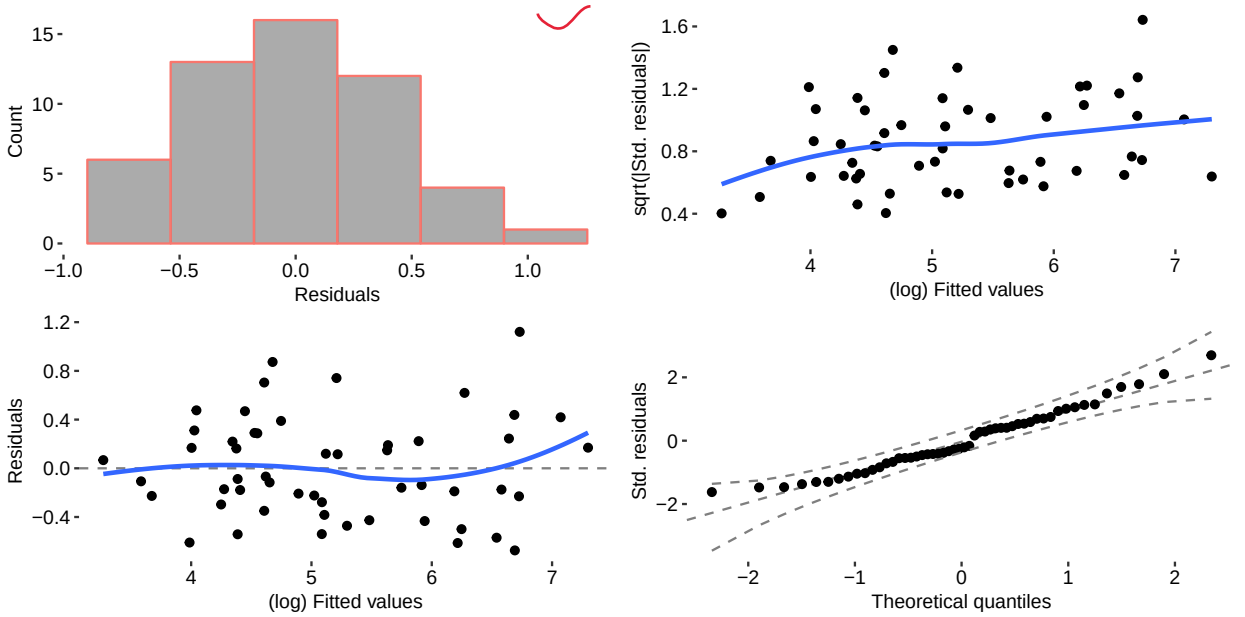

Figure 12: Subplot of four informative plots; A histogram of the residuals, a scatter plot with residuals as a function of the fitted values, a scatter plot of scale location and a normal QQ-plot.

It is possible to state that the estimated model does a reasonable job for modelling the dependent variable `DIOX_log` by considering the plots in figure 12. All though, it should be mentioned that:

- The histogram shows an acceptable distribution of the residuals.
- The scatter plot with the residuals as a function of the fitted values shows a similar acceptable distribution of the residuals. The residuals are not perfectly white noise, which is illustrated by the curvature of the blue line.
- The scale-location plot visualizes the standardized residuals as a function of the fitted values. It shows that the relative error grows as a function of the `DOIX` concentration.
- The QQ-plot states that the standardized residuals of the model have acceptable distribution. All residuals are within the two 95% confidence bands.

Q3

The parameters for a similar model with only the measured active variables are given below:

- O2
- O2COR
- NEFFEKT
- QRAT
- PLANT
- TIME
- LAB

The model is still on the same structure as (2) and the procedure for estimating the model is the same as ealier in question two. In order to reduce the amount of pages the model reduction will not be as comprehensive as ealier. The output below reports the initial fit of the model with measured active variables and block variables.

```
## Single term deletions
##
## Model:
## DIOX_log ~ 1 + O2 + O2COR + NEFFEKT + QRAT + PLANT + TIME + LAB
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                 8.611 -77.504
## O2        0    0.0000  8.611 -77.504
## O2COR     0    0.0000  8.611 -77.504
## NEFFEKT   1    9.5067 18.118 -40.825 48.5745 1.267e-08 ***
## QRAT      1    0.1904  8.802 -78.367  0.9731 0.3293097
## PLANT     2   29.1344 37.746  -4.659 74.4310 7.595e-15 ***
## TIME      1    3.1839 11.795 -63.144 16.2680 0.0002155 ***
## LAB       1    2.0453 10.657 -68.423 10.4505 0.0023262 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model reduction

The numbered list contains the excluded parameters in given order:

1. `O2` - removed w.r.t. multicollinearity.
2. `QRAT` - removed w.r.t. `Pr(>F)` in the `drop1()` table.

```
## Single term deletions
##
## Model:
## DIOX_log ~ O2COR + NEFFEKT + PLANT + TIME + LAB
##          Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                 8.802 -78.367
## O2COR     1     3.226 12.028 -64.130  16.492 0.0001928 ***
## NEFFEKT   1    10.198 19.000 -40.353  52.140 4.799e-09 ***
## PLANT     2    32.549 41.351  -1.916  83.205 7.621e-16 ***
## TIME      1     2.998 11.800 -65.125  15.326 0.0003037 ***
## LAB       1     2.103 10.905 -69.224  10.754 0.0020122 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The model is similar to the model (3) and table 1. All the estimated parameters and its derived statistics are identical after removing those two mentioned parameters. The residual plots are hereby similar to figure 12

and are therefore not included.

Q4

The following assumptions have been made in order to create a prediction of the `DIOX` concentration:

- `TIME = 1`

Table 2 reports the input values to the model object.

Table 2: Input values for the prediction.

O2COR	NEFFEKT	LAB	PLANT	TIME
0.5	-0.01	KK	RENO_N	1

The native `predict()` function takes the derived model object and the new input (table 2) as arguments.

It has been chosen to use the t-distribution rather than the normal distribution in order to calculate the 95% prediction intervals. The t-distribution does consider the degrees of freedom within the model and takes care of the estimated parameters. The chunk below illustrates the approach:

```
# predict the DIOX_log concentration
y_hat_log <- predict(fit_3_final, newdata = pred_df)
# calculate prediction intervals and do inverse log transform
y_hat_CI <- exp(y_hat_log + qt(1 - ALPHA, df = fit_3_final$df.residual) * sd(fit_3_final$residuals)/sqr
    c(-1, 0, 1))  # get lower CI, y_hat and higher CI
```

Table 3 reports the prediction based upon the inputs from the table above.

Table 3: Predicted DIOX concentration and its 95% prediction interval

Low.CI	y_hat	High.CI
297.6424	327.8792	361.1876

Q5

The model depends on the following operation conditions:

- `O2COR` with a coefficient value of: 0.2345724
- `NEFFEKT` with a coefficient value of: 2.7590847

It is requried to decrease the value of `O2COR` (oygen surplus in the gas `OXYGEN`) and the level of `NEFFEKT`in order to decrease the `dioxin` (plant load design variable `LOAD`) concentration based upon the latter reported model coefficients.

Several t-tests have been conducted in order to test wheter the suggestions does have a significant effect.

Minimizing the level plant load (`LOAD`)

- `LOAD` from `H` to `L`
  - H_0 : `H` = `L`
  - H_1 : `H` $\neq$ `L`

There is a significant difference in the mean values of `H` ($\mu = 5.6639074$, $\sigma = 1.2182038$) and `L` ($\mu = 4.5663185$, $\sigma = 0.870659$).

The following test statistics have been conducted; $t(27.1531516) = -2.9320771$, $p = 0.0067597$.

- `LOAD` from `N` to `L`
  - H_0 : `N` = `L`
  - H_1 : `N` $\neq$ `L`

There is a significant difference in the mean values of `N` ($\mu = 5.3463023$, $\sigma = 0.9474089$) and `L` ($\mu = 4.5663185$, $\sigma = 0.870659$).

The following test statistics have been conducted; $t(33.2928174) = -2.5679445$, $p = 0.0149011$.

- `LOAD` from `H` to `N`
  - H_0 : `H` = `N`
  - H_1 : `H` $\neq$ `N`

There is not a significant difference in the mean values of `H` ($\mu = 5.6639074$, $\sigma = 1.2182038$) and `N` ($\mu = 5.3463023$, $\sigma = 0.9474089$).

The following test statistics have been conducted; $t(27.8754662) = -0.8561114$, $p = 0.3992358$.

Minimizing the value of oxygen surplus (`OXYGEN`)

- `OXYGEN` from `H` to `L`
  - H_0 : `H` = `L`
  - H_1 : `H` $\neq$ `L`

There is not a significant difference in the mean values of `H` ($\mu = 5.4759389$, $\sigma = 1.1276074$) and `L` ($\mu = 4.7061769$, $\sigma = 1.1402662$).

The following test statistics have been conducted; $t(29.4185158) = -1.9155987$, $p = 0.0651814$.

- `OXYGEN` from `N` to `L`
  - H_0 : `N` = `L`
  - H_1 : `N` $\neq$ `L`

There is not a significant difference in the mean values of `N` ($\mu = 5.3463023$, $\sigma = 0.9474089$) and `L` ($\mu = 4.7061769$, $\sigma = 1.1402662$).

The following test statistics have been conducted; $t(26.9306231) = -1.7648334$, $p = 0.0889317$.

- OXYGEN from H to N
  - H_0 : H = N
  - H_1 : H $\neq$ N

There is not a significant difference in the mean values of H ($\mu = 5.4759389$, $\sigma = 1.1276074$) and N ($\mu = 5.3463023$, $\sigma = 0.9474089$).

The following test statistics have been conducted; $t(31.431892) = 0.3747394$, $p = 0.7103709$.

Evaluation of the suggestions

According to the test setups it is possible to:

- LOAD
  - Reject the H_0, which says that there is a significant effect of changing LOAD from H to L.
  - Reject the H_0, which says that there is a significant effect of changing LOAD from N to L
  - Accept the H_0, which says that there is not a significant effect of changing LOAD from H to N
- OXYGEN
  - Accept the H_0, which says that there is not a significant effect of changing OXYGEN from H to L.
  - Accept the H_0, which says that there is not a significant effect of changing OXYGEN from N to L.
  - Accept the H_0, which says that there is not a significant effect of changing OXYGEN from H to N.

From the concudted test the suggestion is to decreace the LOAD variable to L in the setup.

Q6

It has been chosen to derive hypothesis in order to test the block effects between the MSW plants and between the two laboratories.

Differences between the MSW plants

From the coefficients of model (3) is it possible to see the different values of the `PLANT` variable. The parametre estimates are tested:

- H_0 : `KARA` = `RENO_N` = `RENO_S`
- H_1 : `KARA` $\neq$ `RENO_N` $\neq$ `RENO_S`

Sub testing

- `KARA` and `RENO_N`

There is a significant difference in the mean values of `KARA` ($\mu = 6.5510716$, $\sigma = 0.817572$) and `RENO_N` ($\mu = 5.5951356$, $\sigma = 0.8548986$).

The following test statistics have been conducted; $t(13.2418282) = 2.7786554$, $p = 0.0154326$.

- `KARA` and `RENO_S`

There is a significant difference in the mean values of `KARA` ($\mu = 6.5510716$, $\sigma = 0.817572$) and `RENO_S` ($\mu = 4.3784017$, $\sigma = 0.6134072$).

The following test statistics have been conducted; $t(9.8889024) = 6.873594$, $p = 4.5795835 \times 10^{-5}$.

- `RENO_S` and `RENO_N`

There is a significant difference in the mean values of `RENO_S` ($\mu = 4.3784017$, $\sigma = 0.6134072$) and `RENO_N` ($\mu = 5.5951356$, $\sigma = 0.8548986$).

The following test statistics have been conducted; $t(35.9920337) = -5.3792542$, $p = 4.698601 \times 10^{-6}$.

Tt is possible to reject the H_0 by summaring the three above conducted tests. When H_0 is rejected it is meaningfull to include the `PLANT` variable in the model.

Differences between the two laboratories?

From the coefficients of model (3) is it possible to see, that `USA` (laboratory in USA) tends to measure a higher concentration of `DIOX_log`=0.4031498 compared to the `KK` (laboratory in Denmark). This is an expected parameter estimate according to the findings in figure 5.

- H_0 : `USA` = `KK`
- H_1 : `USA` $\neq$ `KK`

There is not a significant difference in the mean values of `USA` ($\mu = 5.4174061$, $\sigma = 1.1297621$) and `KK` ($\mu = 5.0064632$, $\sigma = 1.0406309$).

The following test statistics have been conducted; $t(48.7488072) = 1.3610486$, $p = 0.1797572$.

H_0 is therefore accepted.

Q7

The initial final model does include a second term for the numeric explanatory variables which can provide a better description of the total variation. The initial model does also include a interaction terms between `LAB:OXYGEN`, `LAB:LOAD`, `LAB:PRSEK` and `PLANT:OXYGEN`, `PLANT:LOAD`,`PLANT:PRSEK`.

The notation of the initial final model is given in (4).

$$Y_{DIOX_{log}} = X\beta + \epsilon \tag{4}$$

where β are the parameters up for estimation and X is the design matrix given below:

$$X = \begin{bmatrix} 1 & OXYGEN & LOAD & PRSEK & O2 & O2^2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

The following passive variables have been log-transformed: QROEG, CO2, CO, SO2, HCL, H2O.

```
# initial final model ready for model reduction
fit_7_1 <- lm(DIOX_log ~ 1 + OXYGEN + LOAD + PRSEK + O2 + O2COR + NEFFEKT +
    QRAT + PLANT + TIME + LAB + QROEG_log + CO2_log + CO_log + SO2_log + HCL_log +
    H2O_log + TOVN + TROEG + POVN + PLANT:OXYGEN + PLANT:LOAD + PLANT:PRSEK +
    LAB:OXYGEN + LAB:LOAD + LAB:PRSEK + I(O2^2) + I(O2COR^2) + I(NEFFEKT^2) +
    I(QRAT^2) + I(QROEG_log^2) + I(CO2_log^2) + I(CO_log^2) + I(SO2_log^2) +
    I(HCL_log^2) + I(H2O_log^2) + I(TOVN^2) + I(TROEG^2) + I(POVN^2), data = dioxin)
```

Model reduction

The numbered list represents the exclusion order of the variables. The exclusion order is determined by `Pr(>F)` in the drop1() table and the stopping criteria is `Pr(>F) > 0.05`.

1. `O2` and `I(O2^2)` due to high collinearity between `O2` and `O2COR` (see figure 9).
2. `PRSEK` unbalanced number of observations for each level.
3. `LAB:PRSEK` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
4. `PLANT:PRSEK` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
5. `I(CO2_log^2)` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
6. `OXYGEN:PLANT` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
7. `H2O_log` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
8. `CO2_log` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
9. `I(CO_log^2)` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
10. `QRAT` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
11. `I(QRAT^2)` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
12. `CO_log` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
13. `I(POVN^2)` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
14. `POVN` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
15. `OXYGEN:LAB` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
16. `I(NEFFEKT^2)` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
17. `TOVN` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
18. `I(TOVN^2)` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
19. `O2COR` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
20. `I(O2COR^2)` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
21. `OXYGEN` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
22. `TIME` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
23. `SO2_log` - removed w.r.t. `Pr(>F)` in the `drop1()` table.

24. `I(SO2_log^2)` - removed w.r.t. `Pr(>F)` in the `drop1()` table.
25. `LOAD` - removed w.r.t. the value of the std. error compared to the estimate in the `summary(fit)` table.
26. `PLANT:LOAD` - removed w.r.t. the value of the std. error compared to the estimate in the `summary(fit)` table.
27. `LAB:LOAD` - removed w.r.t. the value of the std. error compared to the estimate in the `summary(fit)` table.

Table 4 reports the estimated parameters and their belonging standard errors for the final reduced model.

Table 4:

	Estimate	Std. Error
(Intercept)	-232.5982086	72.4193561
NEFFEKT	2.2208624	0.8908746
PLANTRENO_N	-1.2708248	0.3315603
PLANTRENO_S	-2.5485366	0.2319093
LABUSA	0.4170668	0.1251247
QROEG_log	50.4165348	14.6338336
HCL_log	1.5088330	0.5826427
TROEG	-0.2578500	0.0670080
I(QROEG_log^2)	-2.4318330	0.7019233
I(HCL_log^2)	-0.1201058	0.0513099
I(H2O_log^2)	-0.4387311	0.1288584
I(TROEG^2)	0.0006917	0.0001761

Figure 13 illustrates the residual plots.

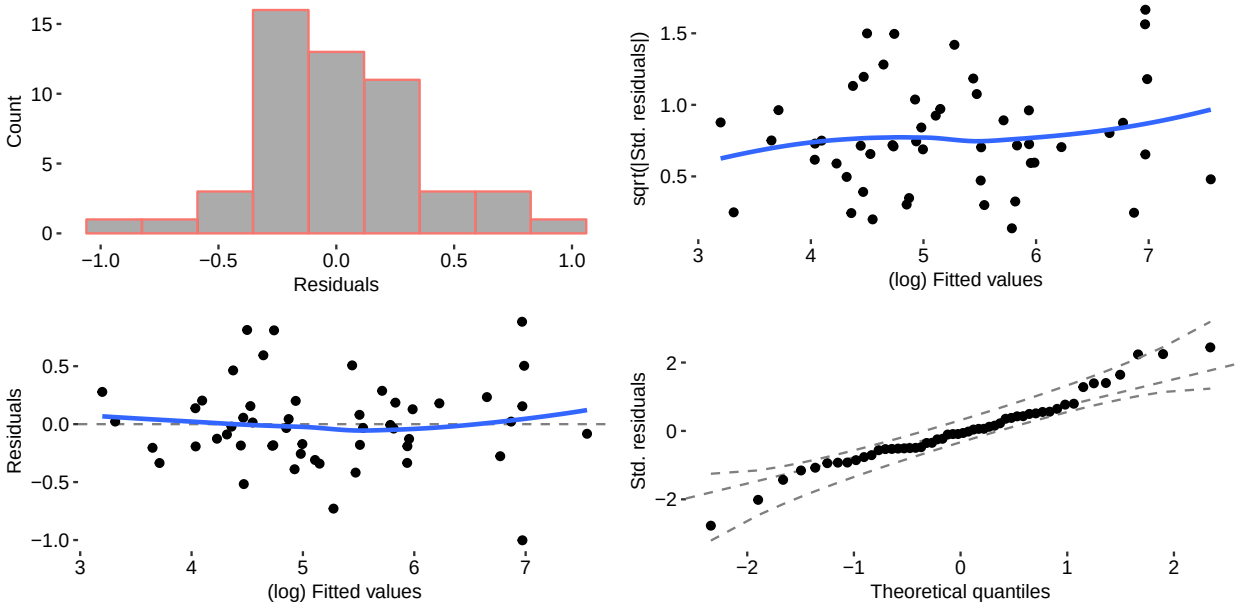

Figure 13: Subplot of four informative plots; A histogram of the residuals, a scatter plot with residuals as a function of the fitted values, a scatter plot of scale location and a normal QQ-plot.

It is possible to conclude that the reduced model does a reasonable job of modelling the dependent variable `DIOX_log` from the residual plots. All though, it should be mentioned:

- The histogram does not show a perfect bell shape.

- The scatter plot with the residuals as a function of the fitted values shows a acceptable normal distribution of the residuals. The residuals are closer to white noise compared to figure 12, which is illustrated by the placement and the curvature of the blue line.
- The scale-location plot shows a growth in the relative error as the function of the `DOIX` concentration increases.
- All standardized residuals but one are within the two 95% confidence bands in the QQ-plot.

Q8

The process of delevoping a model which describes the variation of the measured dioxin at three municipal solid waste plants in Denmark is successfully obtained. The final model is reported in table 4. The final model is depending on seven explanatory variables (three of these variables does have a second order term).

The parameters of the model are: `NEFFEKT`, `PLANT`, `LAB`, `QROEG_log`, `HCL_log`, `TROEG`, `QROEG_log^2`, `HCL_log^2`, `TROEG^2` and `H2O_log^2`.

The final model contains log-transformed passive variables, one active variable which is connected to the design setup (`LOAD`) and two block variables.

The inclusion of the block variables is caused by physical difference setups in three plants. The measured dioxin log-concentrations in the `RENO_N` plant tends to be -1.271 compared to the `KARA` plant. The measured dioxin log-concentrations in the `RENO_S` plant tends to be -2.549 compared to the `KARA` plant. Unfortunately the block variabel `LAB` has an influence on the measured dioxin concentration and therefore must be included in the model. The measured dioxin log-concentrations in the `USA` laboratory tends to be -0.4171 compared to `KK` laboratory. So grandmother, please use the `RENO_S` plant in order to minimize the release of dioxin from your waste.

The final model repects the underlying assumptions of the linear model which means the model is suitable for producing a prediction of the dioxin concentrations. This saves an expensive and protracted process of sending samples to the laboratory.

Q9

The precision (the lower variance) in the KK laboratory is visualized in the box plot in figure 5. Eq. (5) summarizes the assumptions for the final model in table 4.

$$y \sim \mathcal{N}(x\hat{\beta}, \sigma^2\Sigma) \tag{5}$$

where x is the design matrix, Σ is assumed to be the identity matrix ($\Sigma = I$) and σ^2 is estimated under the assumptions of a normal distributed linear model. But the KK laboratory has a smaller variance compared to the USA laboratory which is conflicting with the assumptions of $\Sigma = I$.

It is possible to include the individuel precision of the two laboratories by parameterize Σ. Then is it possible to estimate $\hat{\beta}$, find σ^2 and calculate the maximum likelihood. See eq. (6).

$$\begin{aligned}
\Sigma &= \begin{bmatrix} w_{KK} & & & \\ & w_{USA} & & \\ & & w_{USA} & \\ & & & \ddots \end{bmatrix} \\
\hat{\beta} &= \left(x^T\Sigma^{-1}x\right)^{-1}x^T\Sigma^{-1}y \\
\sigma^2 &= \frac{\left(y - x\hat{\beta}\right)^T\Sigma\left(y - x\hat{\beta}\right)}{n-k}
\end{aligned} \tag{6}$$

where $n - k$ is the degree of freedom in the model. It has been chosen to set $w_{KK} = 1$ and only use the optimizer to find the relative value of w_{USA} which is maximizing the likelihood. See the code chunk below:

```
# define reponds, design matrix and weight vector corresponding to the LAB
# variable.
y <- dioxin$DIOX_log %>% matrix()
x <- matrix(model.matrix(fit_7_final), nrow = dim(fit_7_final$model)[1])
weigths <- dioxin %>% mutate(LAB = ifelse(LAB == "USA", 0, 1)) %>% select(LAB)
# maximum likelihood function
MLE <- function(par, y, x, weigths) {
    SIG <- diag(c(ifelse(weigths == 1, 1, par)))
    beta_hat <- solve(t(x) %*% solve(SIG) %*% x) %*% t(x) %*% solve(SIG) %*%
        y
    sig2 <- t(y - x %*% beta_hat) %*% (y - x %*% beta_hat)/fit_7_final$df.residual
    # negative log-likelihood (using a minimztion algorithm)
    return(-sum(dnorm(y, mean = x %*% beta_hat, sd = sqrt(sig2 %*% diag(SIG)),
        log = TRUE)))
}
# optimize weights with a fixed value for KK=1 and a initial value of
# USA=1.03
w_USA_opt <- nlminb(start = 1, objective = MLE, y = y, x = x, weigths = weigths)$par
# re-estimate with optimal weight for LAB = USA
SIG <- diag(c(ifelse(weigths == 1, 1, w_USA_opt)))
beta_hat <- solve(t(x) %*% solve(SIG) %*% x) %*% t(x) %*% solve(SIG) %*% y
sig2 <- as.numeric(t(y - x %*% beta_hat) %*% (y - x %*% beta_hat)/fit_7_final$df.residual)
# uncertainty of the weight estimate
w_USA_opt <- w_USA_opt + qt(1 - ALPHA, df = fit_3_final$df.residual) * sqrt(sig2)/sqrt(length(diag(SIG)
    c(-1, 0, 1)  # get lower CI, y_hat and higher CI
```

The estimated weight is: $w_{USA} = 1.1322$ with its confidence interval of $[1.0371\,;\,1.2273]$ which describes the uncertainty in the estimate.

Table 5 reports the measured variance of the two laboratories and eq. (7) reports the relative variance between the KK laboratory and the USA laboratory.

Table 5: Measured variance.

LAB	var
KK	1.082913
USA	1.276362

$$1 + \frac{1.2764 - 1.0829}{1.0829} = 1.1786 \tag{7}$$

It can be seen that the measured relative variance of w_{USA} is within the estimated MLE confidence interval; see (7) and figure 14.

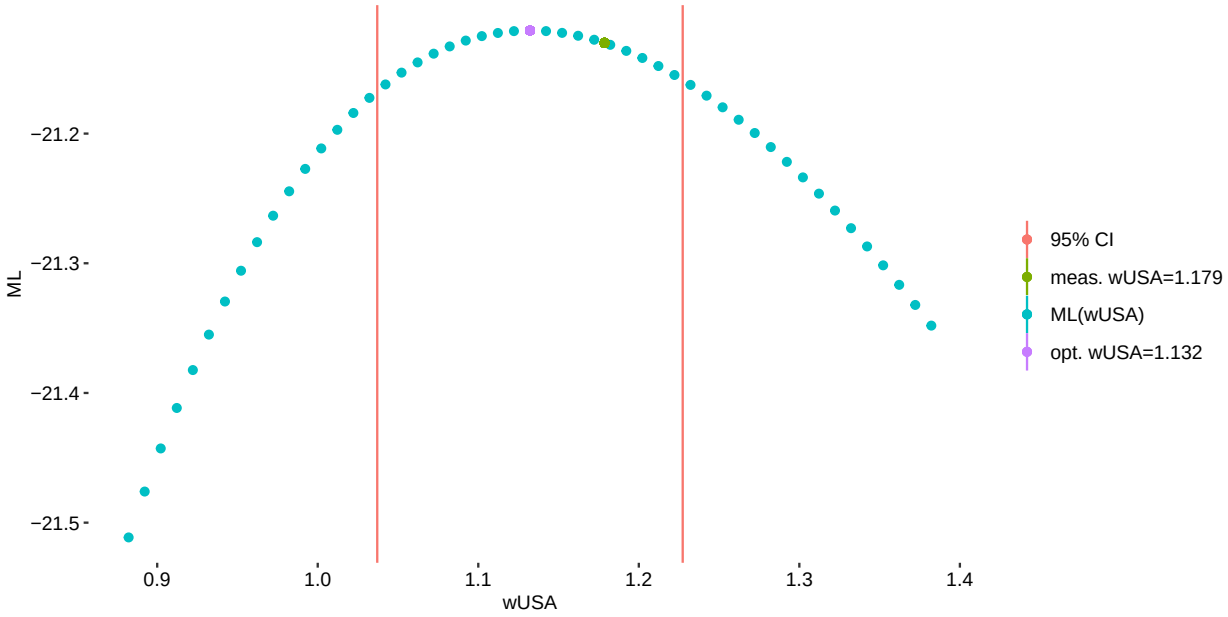

Figure 14: ML as a function of wUSA. MLE and measured relative variance of wUSA compared to wKK.