02424 - Advanced Dataanalysis and Statistical Modelling
Assingment 1: Dioxin emission

Tymoteusz Barcinski - s221937
Soren Skjernaa - s223316

March 10, 2023

1 Introduction

The purpose of the assignment is to establish a model for the variation of measured dioxin emission at a number of solid waste plants based on the data collected during conducted experiments. The dependent variable is the concentration of dioxin in ng per m3. The independent variables are divided into 3 groups: block effects, active variables and passive variables.

2 Exploratorive Analysis (Question 1)

There are 57 observations and 21 columns in the dataset. One column OBSERV is the index of observation and was excluded from the analysis. Figure 1 presents the distribution of the dependent variable DIOX. It exhibits a highly positively skewed distribution and hence the upcoming exploratory analysis was conducted in the original and log domain to reveal interesting relationships. The log transformation was chosen to ease the interpretation. It will be argued in section 4 why the dependent variable was chosen to be included in the model in the transformed domain.

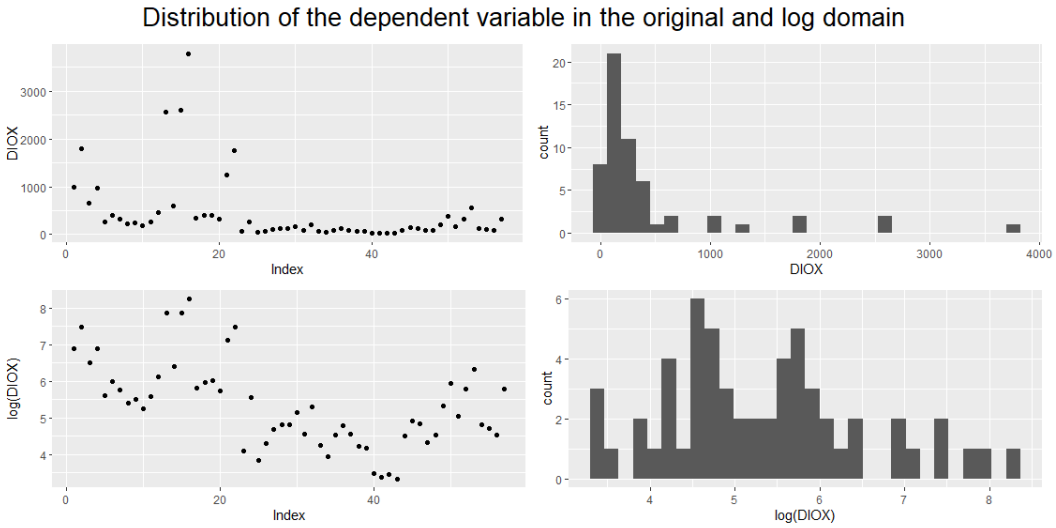

Figure 1

There are 6 missing values of explanatory variables: 2 for the variable PRSEK, 1 for the variable CO, and 3 for the variable SO2. We refer to the procedure of handling those missing values in subsequent sections where those features were used in the models. There are 3 variables described as active variables that varied based on the experimental conditions. There are continuous and discretized versions of those variables available and Figure 2 shows the relation between them. Figure 3 presents the relation between discretized versions of the active variable with relation to the dependent variable in the original and log domain.

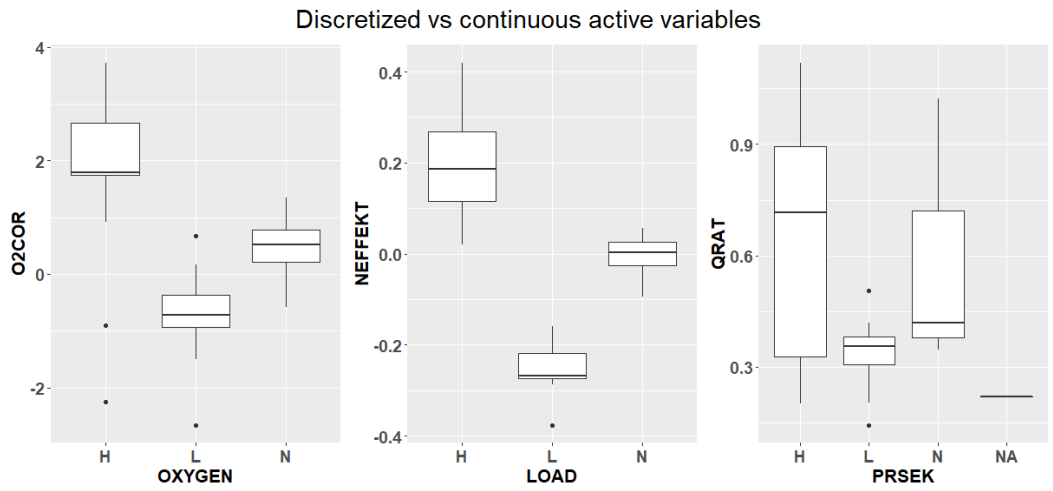

Figure 2

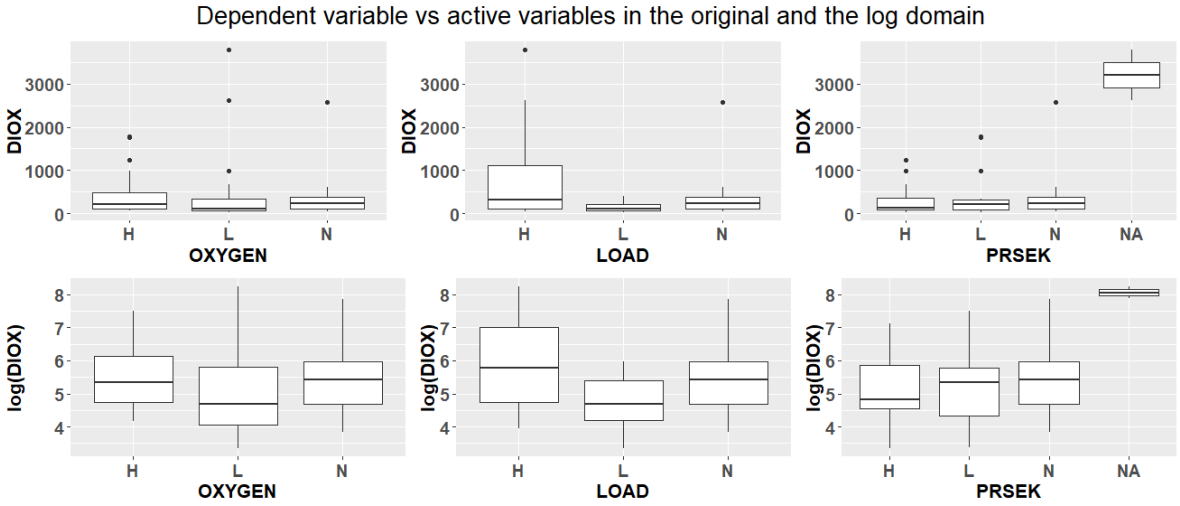

Figure 3

We see that the dependent variable changes with different experimental conditions. There are 3 variables classified as block variables which indicate different conditions under which the experiments were conducted. Figure 4 presents the relation between those variables and the dependent variable. We see that the dependent variable varies for different plants and laboratories. For the plant RENO_N, the experiments were conducted at two different times. We see that the feature TIME influences the dependent variable. As the discrete active variable corresponds to the planned levels of the experiment, while the continuous active variables correspond to the actual measured levels of the experiment, our analysis will mainly focus on the continuous ones. Figure 5 presents the pairwise relationships between continuous active variables and dependent variables for different laboratories. We see that NEFFEKT has the highest correlation with the dependent variable. The correlations between observations from different laboratories do not differ significantly for each active variable. Figure 6 presents the pairwise relationships between continuous active variables and dependent variable for different plants and the time of the experiment. We see that the observations from different plants form clusters compared to the transformed dependent variable (scatter plots on the left side). The correlations between observations from different plants differ for the plot of O2COR vs log(DIOX).

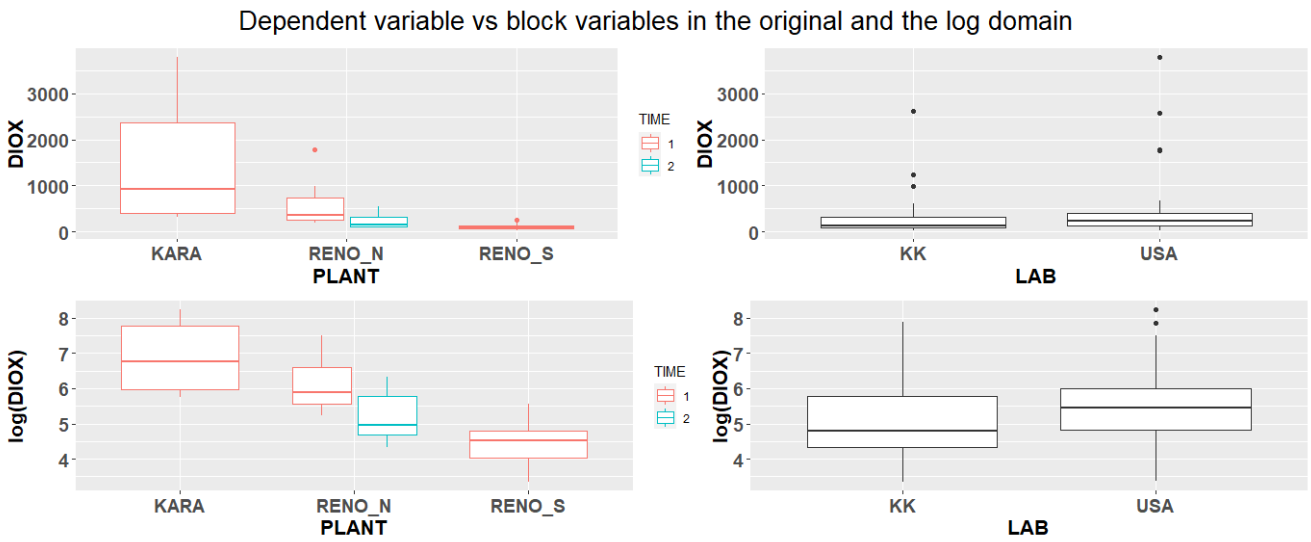

Figure 4

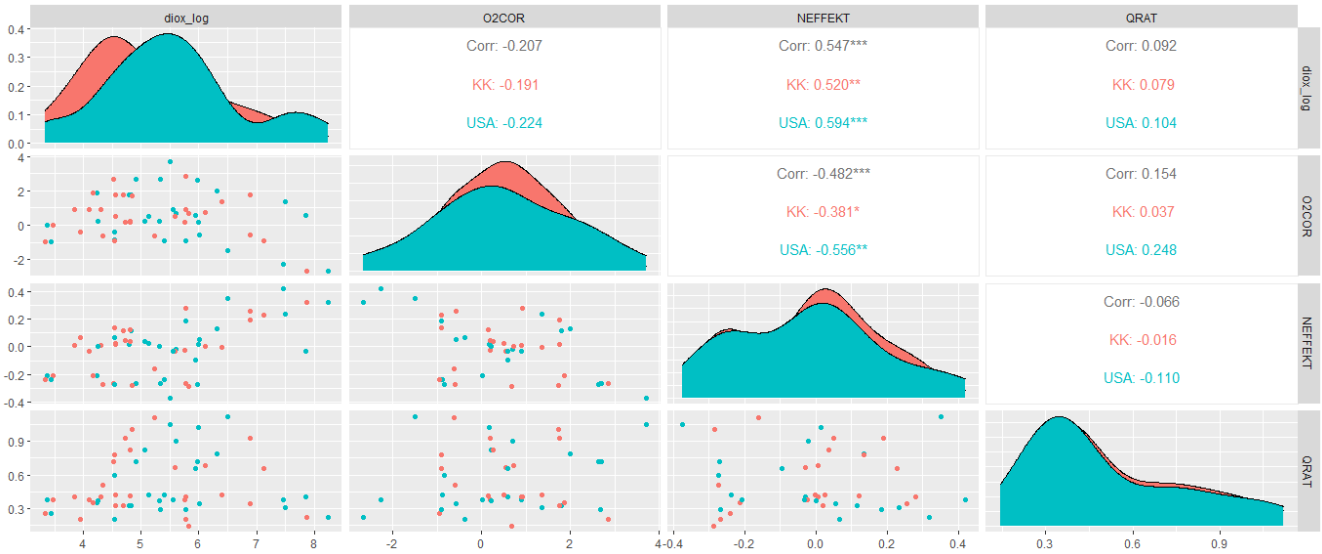

Figure 5: Pairwise scatter plots between continuous active variables and the dependent variable in the log domain for different plants. The pairwise correlations were also presented.

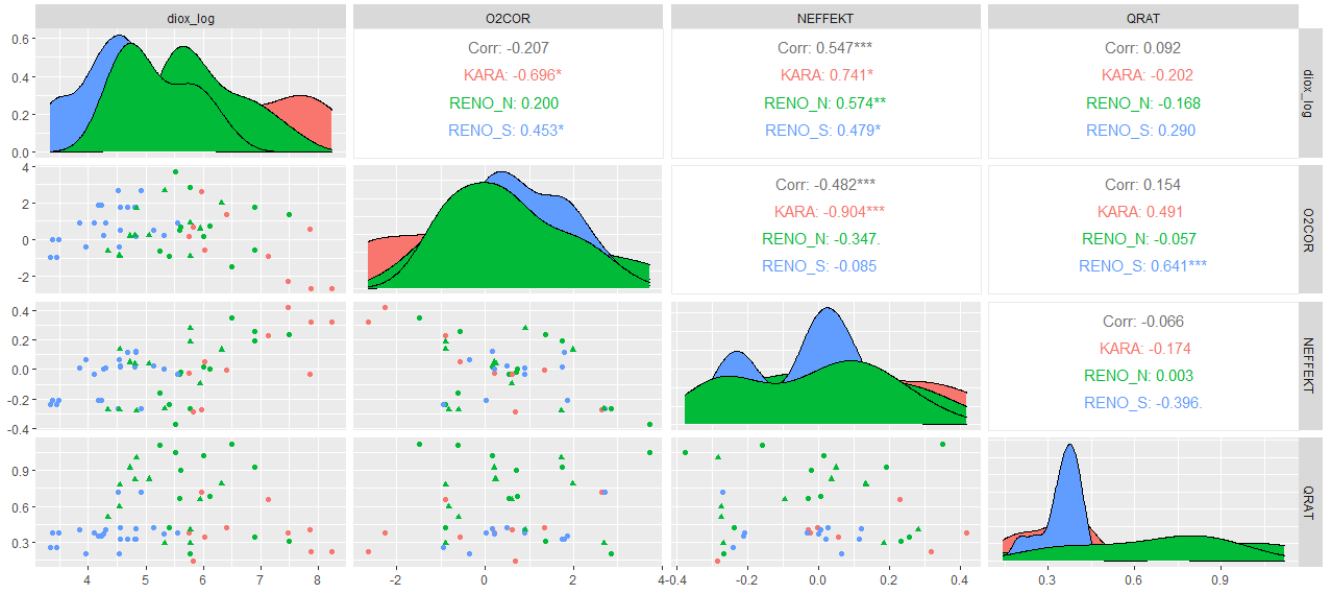

Figure 6: Pairwise scatter plots between continuous active variables and the dependent variable in the log domain for different plants. The pairwise correlations were also presented. For plant RENO_N second time of the experiment was indicated by the triangle shape of points.

3 General Linear Model

In this assignment, the general linear model will be used. The model assumptions can be summarized with the following formula:

$$\boldsymbol{Y} \sim N_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}\right) \tag{1}$$

which states that the dependent variable $\boldsymbol{Y}$ is a linear function of the parameters $\boldsymbol{\beta}$. If not stated otherwise, in this assignment we further assume that $\boldsymbol{\Sigma} = \boldsymbol{I}$. This assumption implies that the errors are normally distributed with a common constant variance, the errors are pairwise independent, and the errors are independent of the mean. Hence, when performing models' diagnostics in the further section, we will focus on the analysis of the residuals. As stated in the assignment's description, we use $\alpha = 5\%$ as the level of significance for statistical testing. Note that in our dataset the block effects are categorical variables (PLANT, TIME, LAB) and will enter the general linear model as factors. Measured active variables and passive variables are continuous variables. Note, that none of the factors are balanced.

4 Simple additive model using discretized active variables (Questions 2)

We fit a simple additive regression model to the data, using the active variables and the block effects, according to the following formula:

$$\begin{aligned}\mathrm{DIOX}_i = \beta_0 &+ \beta_1(\mathrm{OXYGEN}_i) + \beta_2(\mathrm{LOAD}_i) + \beta_3(\mathrm{PRSEK}_i) \\ &+ \beta_4\left(\mathrm{PLANT}_i\right) + \beta_5\left(\mathrm{TIME}_i\right) + \beta_6\left(\mathrm{LAB}_i\right) + \epsilon_i\end{aligned}$$

where $i = 1, 2, \ldots, 57$ and ϵ_i are i.i.d. error terms with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. This assumption is not necessarily correct to make, e.g. as observations may be correlated with plants or there may be different variances for the two labs. The last example is addressed in section 8. We return to the problem of missing values from the previous section. It was verified that 2 missing values of the variable PRSEK correspond to the same observation which was sent to different laboratories. Since the discretized versions of active variables are supposed to be used in the model in this section, we decide to input the missing value of PRSEK based on its observed measurement. Hence, the imputed value was "Low".

4.1 Model diagnostics

Figure 7 presents the model diagnostics plots. Based on the plots in the top panel, we conclude that the assumption of constant variance of residuals is violated. The variance increases with higher fitted values. To deal with this problem, we calculated the boxcox power transformation which indicated the transformation with $\lambda = 0$, which corresponds to a logarithmic transformation, was within 95% confidence interval. We decided to use this logarithmic transformation. Figure 8 presents the model diagnostics plots with a transformed dependent variable. We see that the problem with heteroscedastic variance was not present, and we concluded that the linear model assumptions were satisfied. Hence, we could perform the model reduction part.

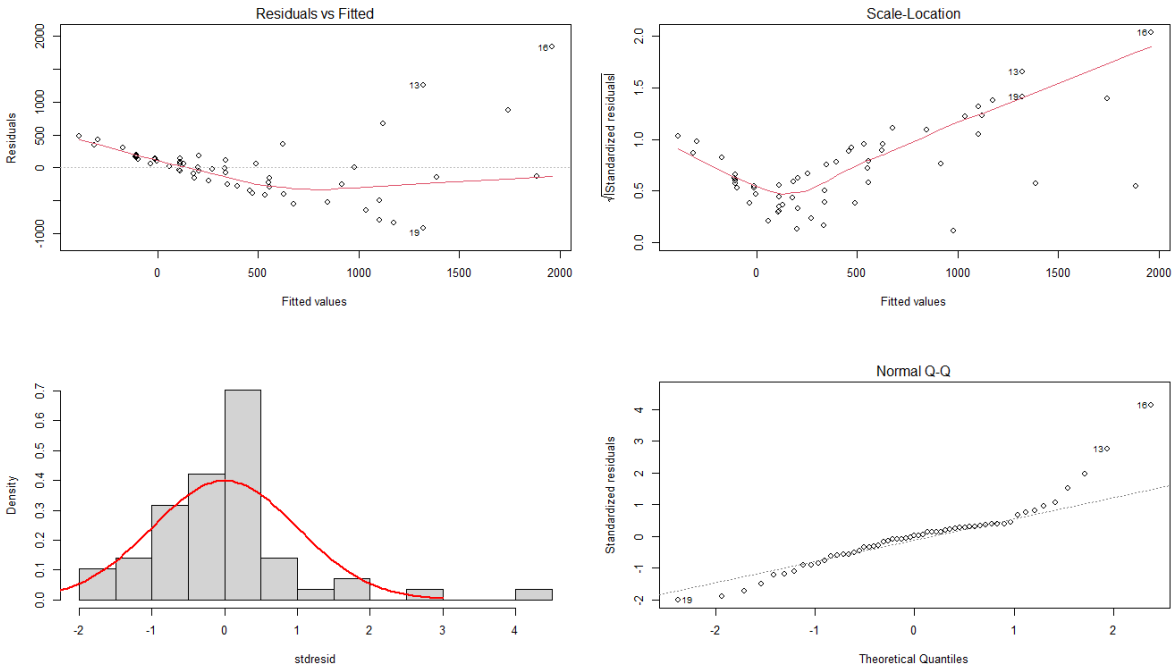

Figure 7: Left to right: (1) Plot of residuals against fitted values. (2) Plot of the square root of the absolut value of standardized residuals against fitted values. (3) Histogram of the standardized residuals overlaid with the $\mathcal{N}(0, 1)$ density function. (4) Normal QQ-plot of the standardized residuals.

4.2 Model reduction

Using a significance level of $\alpha = 0.05$ we now check for significant effects using a standard likelihood ratio test. As there are multiple possible hypothesis chains, but no higher-order terms or interactions, we reduce the model using type III partitioning. The resulting ANOVA table is

	Df	RSS	F value	p-value
OXYGEN	1	13.430	6.9905	0.01104
LOAD	1	21.961	41.9180	4.708e-08
PRSEK	1	12.534	3.3219	0.07460
PLANT	2	56.502	91.6725	¡ 2.2e-16
TIME	1	16.758	20.6150	3.781e-05
LAB	1	13.843	8.6808	0.00495

Table 1: ANOVA table resulting from type III partitioning of the model deviance for the simple additive model in the log domain

From this we see that the active variable PRSEK is insignificant. Removing PRSEK and fitting the resulting model yields the following ANOVA table

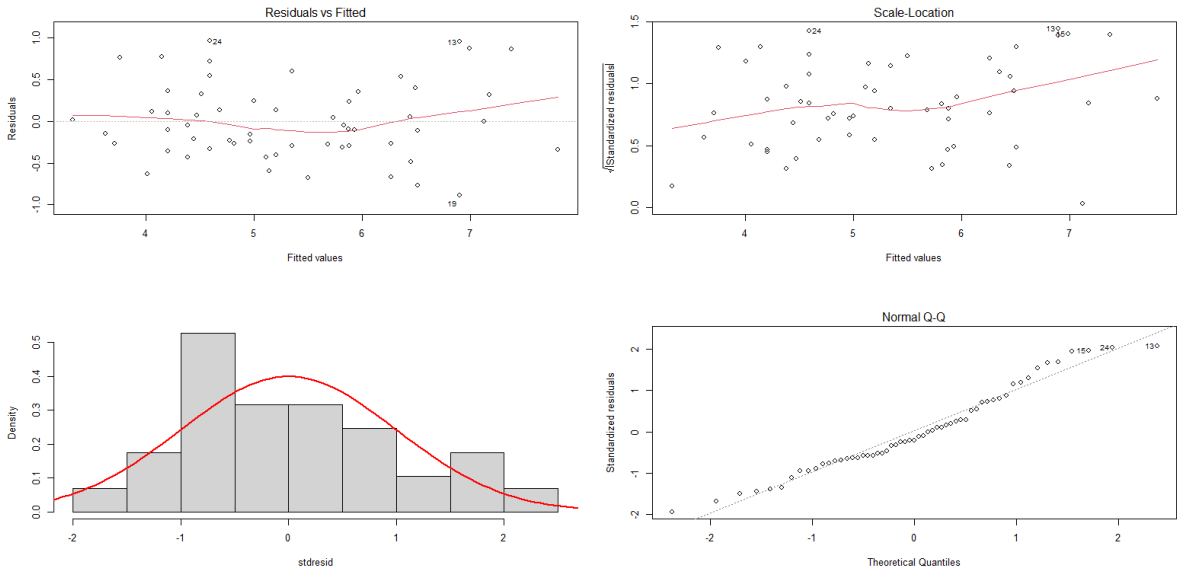

Figure 8: Left to right: (1) Plot of residuals against fitted values. (2) Plot of the square root of the absolut value of standardized residuals against fitted values. (3) Histogram of the standardized residuals overlaid with the $\mathcal{N}(0,1)$ density function. (4) Normal QQ-plot of the standardized residuals.

	Df	RSS	F value	p-value
OXYGEN	1	14.069	5.9971	0.017953
LOAD	1	22.872	40.4103	6.578e-08
PLANT	2	58.706	90.2474	¡ 2.2e-16
TIME	1	17.569	19.6824	5.175e-05
LAB	1	14.611	8.1179	0.006392

Table 2: ANOVA table resulting from type III partitioning of the model deviance for the simple additive model in the log domain

As all dependent variables are seen to be significant we arrive at the final model

$$\begin{aligned}\log(\text{DIOX}_i) = \beta_0 + \beta_1(\text{OXYGEN}_i) + \beta_2(\text{LOAD}_i) \\ + \beta_4\,(\text{PLANT}_i) + \beta_5\,(\text{TIME}_i) + \beta_6\,(\text{LAB}_i) + \epsilon_i\end{aligned}$$

where $i = 1, 2, \ldots, 57$ and ϵ_i are i.i.d. error terms with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

From the model fit, we obtain the following parameter estimates (presented with their standard deviations and 95% confidence intervals):

Parameter	Estimate	S.d.	Lower CI	Upper CI
β_0 (Intercept)	7.2983	0.2110	6.8743933	7.72226363
$\beta_{1,\,\text{OXYGENL}}$ (OXYGENL)	-0.4086	0.1669	-0.7439369	-0.07330552
$\beta_{1,\,\text{OXYGENN}}$ (OXYGENN)	-0.7567	0.1802	-1.1188749	-0.39459185
$\beta_{2,\,\text{LOADL}}$ (LOADL)	-1.0645	0.1675	-1.4010111	-0.72798342
$\beta_{2,\,\text{LOADN}}$ (LOADN)	NA	NA	NA	NA
$\beta_{4,\,\text{PLANTRENO\_N}}$ (PLANTRENO_N)	-0.6636	0.2174	-1.1005268	-0.22667515
$\beta_{4,\,\text{PLANTRENO\_S}}$ (PLANTRENO_S)	-2.3452	0.1929	-2.7328393	-1.95758105
β_5 (TIME2)	-0.9160	0.2065	-1.3309877	-0.50111020
β_6 (LABUSA)	0.3829	0.1344	0.1128303	0.65293541

Table 3: Parameter estimates for the model simple additive model with uncertainties.

Note that the NA for the estimate of $\beta_{2,\,\text{LOADN}}$ is a result of the fact, that the experimental conditions didn't change when the categorical variable LOAD was "Normal" compared to the categorical variable OXYGEN was "Normal".

5 Simple additive model using continuous active variables (Questions 3-6)

As a further exploratory analysis we now fit a regression model to the data, using the measured active variables and the block effects, before we turn our attention to a more detailed model with interactions and higher-order effects.

Motivated by the log-transformation from the previous analysis, we fit a regression model according to the following formula:

$$\begin{aligned}\log(\text{DIOX}_i) = \beta_0 + \beta_1 \cdot \text{O2COR}_i + \beta_2 \cdot \text{NEFFEKT}_i + \beta_3 \cdot \text{QRAT}_i \\ + \beta_4 \left(\text{PLANT}_i\right) + \beta_5 \left(\text{TIME}_i\right) + \beta_6 \left(\text{LAB}_i\right) + \epsilon_i\end{aligned}$$

where $i = 1, 2, \ldots, 57$ and ϵ_i are i.i.d. error terms with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. As mentioned previously this assumptions is not necessarily correct.

5.1 Model diagnostics

Before we reduce the model, by testing for insignificant effects, we first examine the residuals of the model. As this is not our final model we will not go fully in depth with the model diagnostics here.

Plot (1) and (2) from figure 9, of the residuals against the fitted values, gives no cause for concern as the residuals are evenly distributed around zero.

Looking at plot (3) and (4), of the histogram and QQ-plot of the standardized residuals, we see that while it is not problematic to assume that the errors follow a normal distribution, the tail of the residuals at the positive end is a bit heavier than expected. However, we conclude that the model assumptions are not violated.

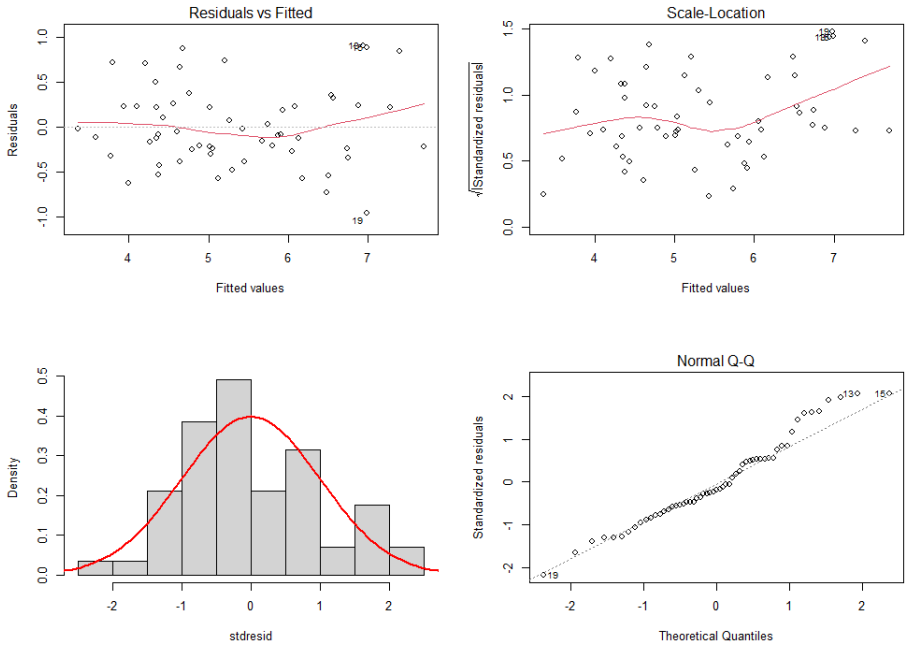

Figure 9: Left to right: (1) Plot of residuals against fitted values. (2) Plot of the square root of the absolut value of standardized residuals against fitted values. (3) Histogram of the standardized residuals overlaid with the $\mathcal{N}(0, 1)$ density function. (4) Normal QQ-plot of the standardized residuals.

5.2 Model reduction

Using a significance level of $\alpha = 0.05$ we now check for significant effects using a standard likelihood ratio test. As there are multiple possible hypothesis chains, but no higher order terms or interactions, we reduce the model using a type II partitioning. The resulting ANOVA table is

	Df	RSS	F value	p-value
O2COR	1	13.69	12.56	0.0009
NEFFEKT	1	24.15	59.60	<0.0001
QRAT	1	11.29	1.78	0.1885
PLANT	2	49.64	87.10	<0.0001
TIME	1	14.91	18.05	0.0001
LAB	1	13.27	10.69	0.0020

Table 4: ANOVA table resulting from type III partitioning of the model deviance.

From this we see that the active variable QRAT is insignificant. Removing QRAT and fitting the resulting model yields the following ANOVA table

	Df	RSS	F value	p-value
O2COR	1	13.85	11.30	0.0015
NEFFEKT	1	24.68	59.26	<0.0001
PLANT	2	51.20	88.35	<0.0001
TIME	1	15.09	16.82	0.0002
LAB	1	13.66	10.49	0.0021

Table 5: ANOVA table resulting from type III partitioning of the model deviance.

As all dependent variables are seen to be significant we arrive at the final model

$$\begin{aligned}\log(\text{DIOX}_i) = \beta_0 + \beta_1 \cdot \text{O2COR}_i + \beta_2 \cdot \text{NEFFEKT}_i \\ + \beta_4 (\text{PLANT}_i) + \beta_5 (\text{TIME}_i) + \beta_6 (\text{LAB}_i) + \epsilon_i \end{aligned} \quad (2)$$

where $i = 1, 2, \ldots, 57$ and ϵ_i are i.i.d. error terms with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

5.3 Interpretation of the model

From the model fit, we obtain the following parameter estimates (presented with their standard deviations and 95% confidence intervals):

Parameter	Estimate	S.d.	Lower CI	Upper CI
β_0 (Intercept)	6.50	0.17	6.17	6.83
β_1 (O2COR)	0.18	0.05	0.07	0.29
β_2 (NEFFEKT)	2.90	0.38	2.14	3.66
$\beta_{4\text{RENO\_N}}$ (PLANT RENO_N)	-0.74	0.21	-1.16	-0.32
$\beta_{4\text{RENO\_S}}$ (PLANT RENO_S)	-2.30	0.19	-2.68	-1.92
$\beta_{5\text{TIME2}}$ (TIME 2)	-0.80	0.19	-1.19	-0.41
$\beta_{6\text{USA}}$ (LAB USA)	0.41	0.13	0.16	0.66

Table 6: Parameter estimates for model (2) with uncertainties.

Note that as the dioxin concentrations were log-transformed the parameters cannot be interpreted directly.

5.3.1 Active variables (Question 5)

Examining the effects of the active variables, we found that QRAT (ratio between primary and secondary air) had no significant effect on the dioxin concentrations. Transforming the parameter estimates for O2COR and NEFFEKT, we get the following table

Parameter	Estimate	Lower CI	Upper CI
e^{β} (O2COR)	1.20	1.08	1.34
$e^{\beta_2 \cdot 0.01}$ (NEFFEKT)	1.0294	1.0216	1.0373

Table 7: Back transformed parameter estimates for the active variables in model (2).

Thus, a unit increase in the oxygen surplus (O2COR), yields a 20% (8%, 34%) increase in the dioxin concentration. As the variable NEFFEKT measures the normalized plant load we see that a load of 1% more than the designed load (which corresponds to an increase of 0.01 unit in the scale of this variable which is in percent originally) yields a 2.94% (2.16%, 3.73%) increase in the dioxin concentration.

Thus we conclude that to reduce the dioxin emission, the oxygen surplus and plant load should be reduced. Note, that we have not investigated if there is an interaction between the oxygen surplus and plant load, which could change the conclusion if both parameters are changed simultaneously.

5.3.2 Block effects (Question 6)

From the analysis we found that there is a significant difference between the three plants, as well as between the two laboratories. Transforming the parameter estimates for PLANT and LAB, we find the following estimates:

Parameter	Estimate	Lower CI	Upper CI
$e^{\beta_{4\text{RENO\_N}}}$ (PLANT RENO_N)	0.48	0.31	0.73
$e^{\beta_{4\text{RENO\_S}}}$ (PLANT RENO_S)	0.10	0.07	0.15
$e^{\beta_{6\text{USA}}}$ (LAB USA)	1.504	1.168	1.937

Table 8: Back transformed parameter estimates for the block effects in model (2).

Thus, we see that for two identical samples, (i.e. same plant, O2COR etc) the USA laboratory estimates the dioxin concentration 50.4% (16.8%, 93.7%) higher compared to the KK laboratory.

Looking at the plant estimates, we see that the KARA plant had the highest dioxin emissions with the emissions at the RENO_N plant being 52.3% (27.3%, 68.7%) lower than at the KARA plant and the emissions at the RENO_S plant being 90.0% (85.3%, 93.2%) lower than at the KARA plant.

An overview of the differences between the two laboratories and the three plants is found in figure 10, where the estimated marginal means for the log-transformed dioxin emissions, and the corresponding 95% confidence intervals are plotted.

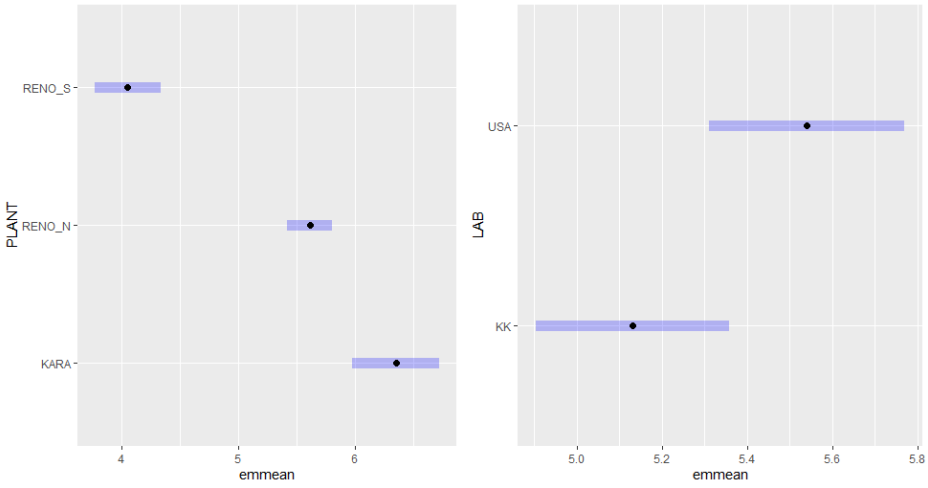

Figure 10: Estimated marginal means of log(DIOX) for the PLANT and LAB effects.

5.3.3 Prediction of a new observation (Question 4)

Using the found parameter estimates we may predict a new dioxin concentration.

Assume we run a new experiment under the conditions: O2COR = 0.5, NEFFEKT = -0.01 during time 1 at the RENO_N plant, and then analyze it at the KK lab. We would then expect the log dioxin concentration to be (with a 95% prediction interval):

$$\log(\text{DIOX}) = 5.82 \quad (4.82, 6.83)$$

This means that the we predict the dioxin concentration to be 338.55 (124.31, 922.01) ng. pr. m^3.

6 Final model with higher order terms and passive variables (Question 7)

On the basis of the previous sections, we found that a model of the log-transformed dioxin concentration was preferable, and that the active variables O2COR and NEFFEKT was significant, together with the block effects PLANT, TIME and LAB. In this section, we will try to expand upon these findings by analyzing the addition of the passive variables and possible interactions and higher order effects to our model.

6.1 Specification of the model

Using model (2) as our base model, we try to find a model with a more suitable fit to the given data. We use Stepwise-Regression for our model selection strategy with a type II partitioning of the model deviance. That is, given a search scope of possible effects, which may be added to the model, we alternate between a forward selection strategy and a backwards elimination strategy, thus alternately adding significant effects and removing insignificant effects. As we used a type II partitioning, no higher order terms or interactions, were added if the lower order terms were not already included in the model. The search scope for our model selection was the following:

- All the passive variables.
- log(CO) and log(HCL).
- Second order polynomials of all the active variables.
- Second order polynomials of all the active variables interacting with the block effects.
- Second order polynomials of log(HCL) and log(CO).

Examining the distributions of the passive variables, we observed that the CO and HCL concentrations were highly skewed. As a log-transform of these two passive variables yielded a more uniform distribution, we chose to include log(CO) and log(HCL) in the search scope to stabilize the model. Based on the correlations observed in figure 6 we chose to include possible interactions between the block effects and the active variables. Furthermore, initial data exploration led us to suspect a significant effect of log(HCL). Thus we decided to include polynomials of this passive variable.

As the passive variables SO2 and CO contained missing data (obs 7, 8 and 53) we chose to exclude these three observations from the model selection. A possible improvement on the model selection could be to impute the missing data by modelling it, but this strategy was not pursued further as deemed out of the scope of the course.

The Stepwise-Regression strategy resulted in no effects being dropped during the model selection. As such the procedure equaled a forward model selection. Table 9 shows the order the different effects were added to the final model along with their p-values at the time of addition.

The final model was described by the following equation

$$\begin{aligned}
\log(\text{DIOX}_i) = \beta_0 &+ \beta_1 \cdot \text{O2COR}_i + \beta_2(\text{TIME}_i) \cdot \text{NEFFEKT}_i && \textit{(active variables)} \\
&+ \beta_3\,(\text{PLANT}_i) + \beta_4\,(\text{TIME}_i) + \beta_5\,(\text{LAB}_i) && \textit{(block effects)} \\
&+ \beta_6 \cdot \text{CO2}_i + \beta_7 \cdot \text{TROEG}_i + \beta_8 \cdot \text{POVN}_i + \beta_9 \cdot \log(\text{HCL})_i + \beta_{10} \cdot \log(\text{HCL})_i^2 && \textit{(passive variables)} \\
&+ \epsilon_i && \textit{(error term)}
\end{aligned}$$

Effect	df	F-value	p-value
poly(log(HCL), 1)	1	10.71	0.002
CO2	1	7.83	0.008
TROEG	1	6.09	0.018
TIME:poly(NEFFEKT, 1)	1	5.65	0.022
POVN	1	7.55	0.009
poly(log(HCL), 2)	1	9.17	0.004

Table 9: Results of the Stepwise-regression strategy. The effects in the table are listed in the order they were added to the final model.

where $i = 1, 2, \ldots, 57$ and ϵ_i are i.i.d. error terms with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. As the passive variables CO and SO2 did not enter the final model, we reestimated the final model using all 57 observations.

6.2 Model diagnostics

A residual analysis was carried out on the final model reestimated with all 57 observations. Figure 11 shows the overall results of the model diagnostics. Plot (1) and (2) shows that the standardized residuals are independent of the fitted values, and thus that the assumption of variance homogeneity of the residuals is valid. Plot (3) and (4) shows that the standardized residuals are approximately $\mathcal{N}(0, 1)$ distributed, thus the assumption of the error terms being normal distributed is valid.

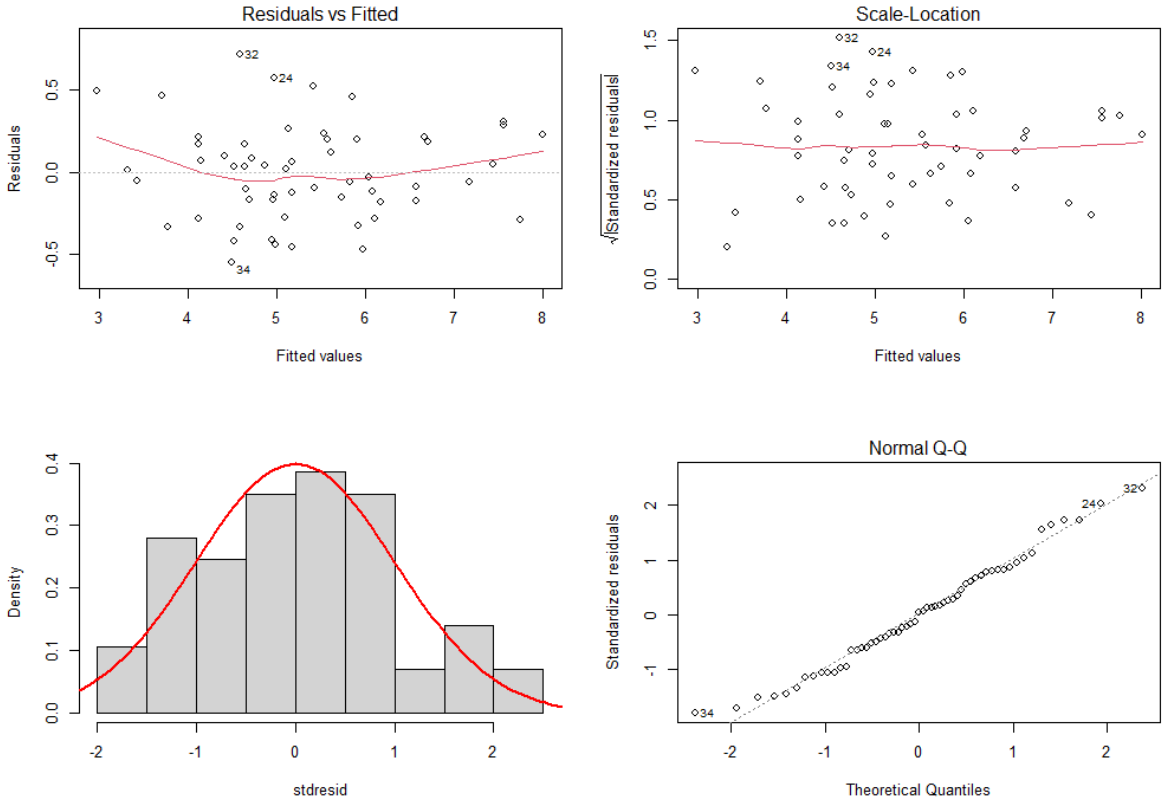

Figure 11: Left to right: (1) Plot of residuals against fitted values. (2) Plot of the square root of the absolut value of standardized residuals against fitted values. (3) Histogram of the standardized residuals overlaid with the $\mathcal{N}(0, 1)$ density function. (4) Normal QQ-plot of the standardized residuals.

Figures 12 and 13 shows the standardized residuals plotted against the different independent variables of the model. As the standardized residuals are evenly spread out, with no distinct pattern detected in the plots, we find that the model is a good fit for the data.

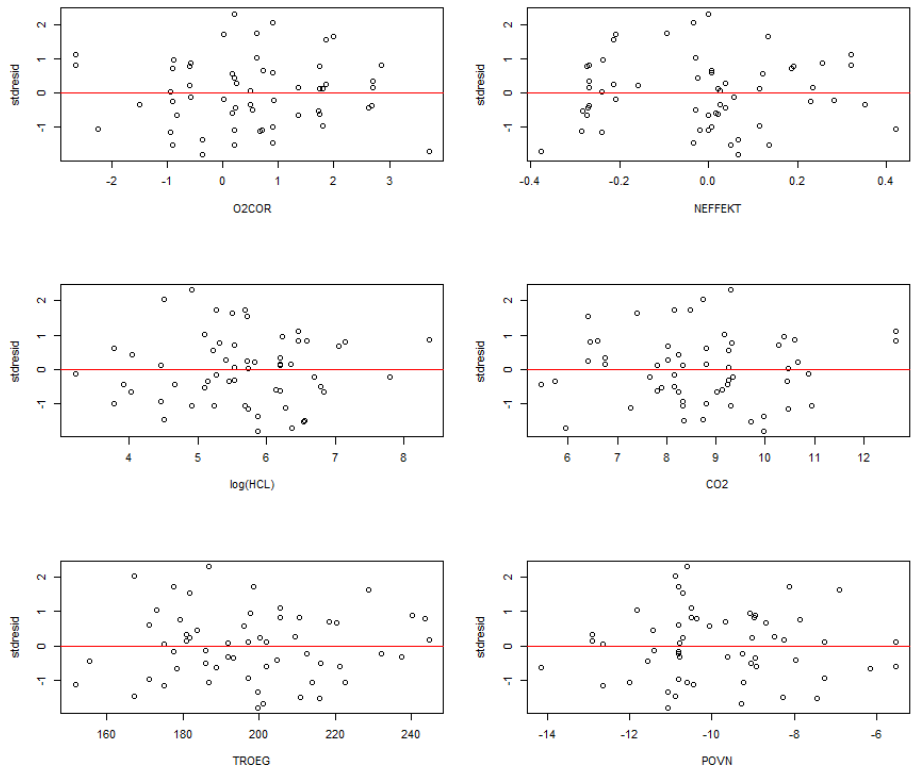

Figure 12: Plots of the standardized residuals against the different active and passive variables.

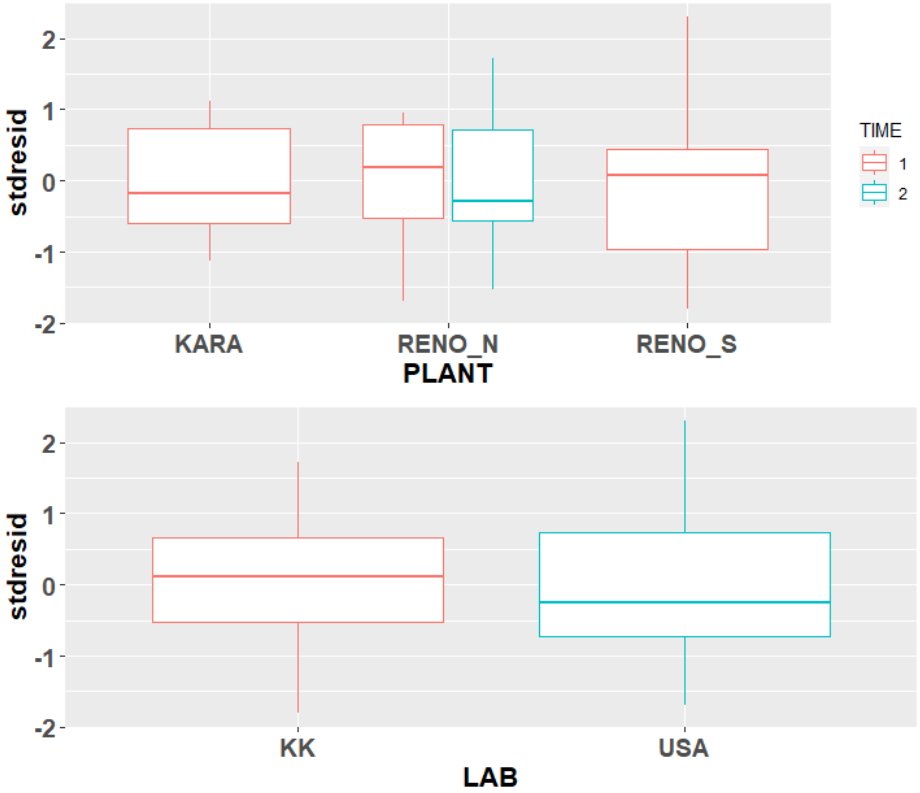

Figure 13: Boxplots of the standardized residuals agianst the different block effects.

Finally we examine the model for outliers and influential observations. Figure 14 shows a plot of the studentized residuals against the leverages of the observations. Cutoff lines for detection of outliers are drawn at the levels ± 2 and a cutoff for detection of influential observations of $2 \cdot$ "number of parameters" / "number of observations" is used. These cutoffs are chosen according to recommendations in (Conradsen et al., 2019). We see that while observations 24 and 32 are outliers, they are not influential. Thus we have decided to include them in the final model, but we recommend discussing these observations with the experimenters. Observations 1 and 19 are influential, but not considered outliers. Examining the data we see that these observations corresponds to the highest and lowest HCL concentrations respectively. As they are not considered outlier we choose to include them in the final model, but again we recommend discussing these observations with the experimenters and the effect of HCL on the dioxin emission with domain experts.

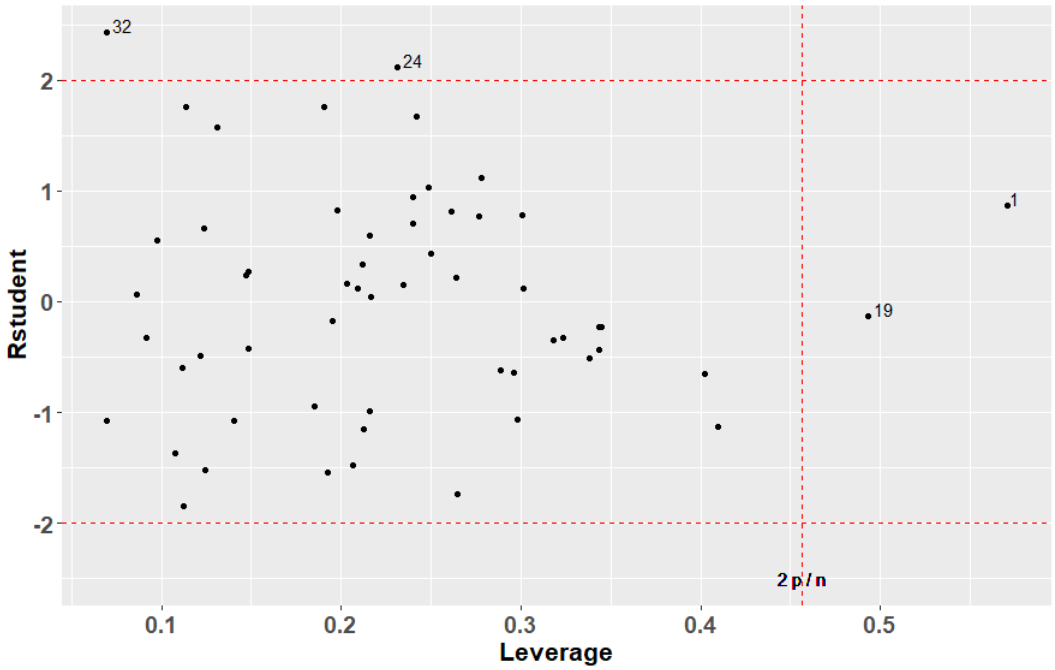

Figure 14: Plot of the studentized residuals, against the leverage of the observations. For the vertical line through $x = 2p/n$, p equals the number of parameters i.e. ($p = 13$), and n equals the number of observations (i.e. $n = 57$).

6.3 Parameter estimates for the final model

From the model fit, we obtain the following parameter estimates (presented with their standard deviations and 95% confidence intervals):

Parameter	Estimate	S.d.	Lower CI	Upper CI
β_0 (Intercept)	7.68	1.54	4.58	10.79
β_1 O2COR	0.64	0.09	0.46	0.82
$\beta_{2\text{TIME}_1}$ TIME1:poly(NEFFEKT, 1)	8.00	1.01	5.97	10.03
$\beta_{2\text{TIME}_2}$ TIME2:poly(NEFFEKT, 1)	5.30	1.14	3.00	7.61
$\beta_{3\text{RENO\_N}}$ PLANT RENO_N	-0.27	0.29	-0.85	0.31
$\beta_{3\text{RENO\_N}}$ PLANT RENO_S	-2.35	0.15	-2.66	-2.05
$\beta_{4\text{TIME}_2}$ TIME 2	-0.73	0.17	-1.08	-0.39
$\beta_{4\text{USA}}$ LAB USA	0.45	0.09	0.26	0.64
β_6 CO2	0.30	0.08	0.14	0.45
β_7 TROEG	-0.03	0.01	-0.04	-0.01
β_8 POVN	-0.09	0.03	-0.16	-0.02
β_9 poly(log(HCL), 2)1	3.23	0.49	2.23	4.22
β_{10} poly(log(HCL), 2)2	-1.18	0.39	-1.96	-0.40

Table 10: Parameter estimates for the final model with uncertainties.

Note that the parameter $\beta_{3\text{RENO\_N}}$ is not significant as 0 belongs to the estimated confidence interval (-0.85, 0.31) for the parameter. This implies that there is no significant difference between the KARA and RENO_N plants.

From the model we get the following estimate of the variance of the error terms (presented with 95% Wald confidence intervals):

$\hat{\sigma}^2$	Lower CI	Upper CI
0.10	0.07	0.17

Table 11: Estimate of model variance with uncertainties.

Comparing tables 6 and 10, we see that the final model is in agreement with our previous model in regards to which active variables and block effects are significant, and the sign of the estimated parameters.

7 Summary of our findings (Question 8)

From our final model we found that the dioxin emissions at the MSV plants, depended on the oxygen surplus (O2COR) and the plant load (NEFFEKT) during operation of the incinerators, however the ratio between primary and secondary air (QRAT) had no significant effect. To reduce the concentration of dioxin, the plants can reduce the oxygen surplus or the load during incineration. We found that an increase in the oxygen concentration of 1 resulted in a 89.65% increase in dioxin concentration. We found that the loads effect on the dioxin concentration was different between the two times the experiment was carried out. The first time, a 1% increase in plant load (which corresponds to an increase of 0.01 unit in the scale of this variable which is in percent originally) yielded a 8.33% increase in dioxin concentration. The second time, a 1% increase in plant load yielded a 5.44% increase in the dioxin concentration.

Besides this, our analysis found that there was a significant difference between dioxin emissions at the three plants, even when correcting for other effects. Here the KARA plant had the highest emissions, with the emissions at RENO_N being 23.66% lower, and the emissions at RENO_S being 90.46% lower than the KARA plant. Furthermore, we found a significant difference between the analysis results of the two laboratories with the USA laboratory giving a 56.83% higher estimate of the dioxin concentrations.

Finally our analysis found that the dioxin emissions also depended on the CO_2 and HCL concentrations, and the gas temperature (TROEG) and chamber pressure (POVN). Especially the HCL concentrations had a large effect on the dioxin emissions.

8 Model with weighted uncertainty w.r.t. LAB (Question 9)

We want to investigate whether the precision between laboratories is different. We define a precision parameter (called also weight) τ as the ratio of the variances between the USA laboratory and KK laboratory, which may be interpreted as how many times the KK laboratory is more precise than the USA laboratory. We consider the final model from the previous section with an additional structure on the covariance matrix $\mathbf{\Sigma}$ between observations than in previous sections where we assumed that $\mathbf{\Sigma} = \boldsymbol{I}$. Namely, our model has the form:

$$\log(\boldsymbol{Y}) \sim N_n\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbf{\Sigma}\right); \quad \mathbf{\Sigma} = \begin{bmatrix} \frac{1}{\tau} & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 \\ 0 & \frac{1}{\tau} & \cdots & 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\tau} & 0 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & \cdots & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \tag{2}$$

Note, that the observations were ordered in the model formulation based on the two laboratories. The model is parameterized by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \tau)$. We are interested in an inference only on the τ parameter and hence we consider

the profile likelihood of it and treat the rest of the parameters as nuisance parameters. The profile likelihood is defined as:

$$L_P(\tau; \log(\boldsymbol{y})) = \sup_{\boldsymbol{\beta}, \sigma^2} L((\tau, \boldsymbol{\beta}, \sigma^2); \log(\boldsymbol{y})) \tag{3}$$

where the maximization is performed at a fixed value of τ and the vector $\log(\boldsymbol{y})$ is the vector of the log of the dependent variable DIOX. Therefore we solve the following optimization problem to estimate τ:

$$\hat{\tau} = \sup_{\tau} L_P(\tau; \log(\boldsymbol{y})) \tag{4}$$

The uncertainty associated with the estimation is captured by the profile likelihood function. We consider two types of confidence intervals: Wald's and likelihood-based. The following table presents the results

$\hat{\tau}$	Wald Lower CI	Wald Upper CI	Profile Lower CI	Profile Upper CI
1.91	-0.03	3.85	0.65	4.95

Table 12: Estimate of the precision parameter with uncertainties.

Figure 15 presents the profile likelihood and the quadratic approximation at the optimum. We are interested in testing whether the precision parameter is different than 1. Formally, we test a null hypothesis $H_0 : \tau = 1$ against the alternative $H_1 : \tau \neq 1$. Since the asymptotic distribution of the Wald test statistic requires the quadratic approximation to be good at a point where the testing is performed (here $\tau = 1$), we can't trust this test. Hence, the Wald's confidence interval for τ can not be trusted as well. We see that the profile likelihood is asymmetric which can't be captured by quadratic approximation. Therefore, we use the likelihood ratio test, which is equivalent to checking whether the estimate of precision parameter $\hat{\tau}$ belongs to the profile likelihood confidence interval. We see that it's not the case, and hence we fail to reject the null hypothesis that $\tau = 1$. Therefore, there is no evidence to conclude that the precision in laboratories is different. We further calculated the p-value for the Likelihood Ratio Test and found that it was equal to 0.232 confirming our conclusion.

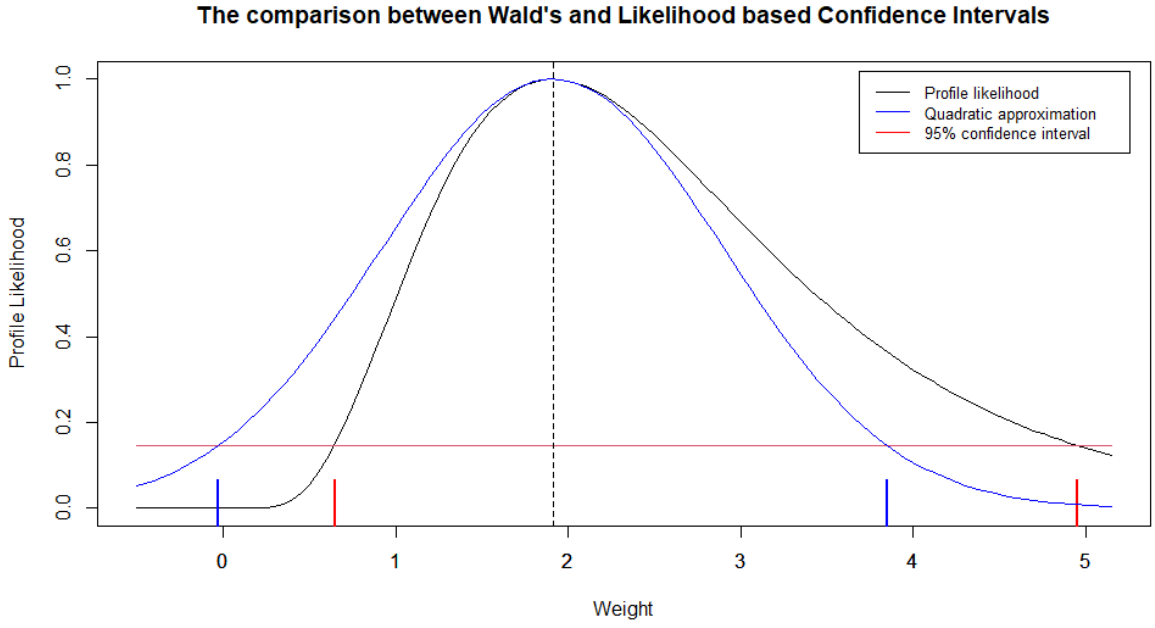

Figure 15: Plot of the profile likelihood for the parameter weight θ. The vertical line indicates the value that maximizes the likelihood. Blue rugs correspond to Wald's confidence interval, whereas red rugs correspond to the Likelihood-based confidence intervals.

9 References

1. Conradsen, K., Christensen, A.M., Nielsen, A.A., Ersbøll, B.K. (2019, v.0.94). *Multivariate Statistics: For the Technical Sciences*. DTU Compute Lyngby.

2. Madsen H., Thyregod P. (2011). *Introduction to General and Generalized Linear Models.* CRC Press.

10 Appendix

```
# File Description ------------------------------------------------------------
#
#   S ren Skjernaa - s223316
#   Tymotuesz Barcinski - s221937
#   10/03-2023
#
#   Advanced Dataanalysis and Statistical Modelling
#   Assignment 1
#
#   Note: The data file for the analysis is assumed to be found at the relative
#         path "data/assignment_1_dioxin.csv".
#
#------------------------------------------------------------------------------

# Library imports -------------------------------------------------------------

# Plotting
library(ggplot2)
library(gridExtra)

# Post-hoc analysis
library(emmeans)
library(multcomp)
library(xtable)

# Model diagnostics
library(MASS)
library(MESS)
library(nortest)
library(influence.ME)



# . ---------------------------------------------------------------------------
# Initial data exploration ----------------------------------------------------

## Data preprocessing ---------------------------------------------------------
DATA <- read.table("data/assignment_1_dioxin.csv", sep=",", header=TRUE)

summary(DATA)
DATA$OBSERV <- factor(DATA$OBSERV)
DATA$TIME <- factor(DATA$TIME)
DATA$PLANT <- factor(DATA$PLANT)
DATA$LAB <- factor(DATA$LAB)
summary(DATA)

attach(DATA)


### Tabulating the data -------------------------------------------------------

# We have N = 57 observations

# Planned active variables
table(OXYGEN)
table(LOAD)
table(PRSEK, useNA ="ifany")

table(OXYGEN, LOAD)
table(OXYGEN, PRSEK)
table(LOAD, PRSEK)

# Block effects
table(PLANT)
table(TIME)
table(LAB)

table(PLANT, TIME)    # Only RENO_N have repeated measurements.
table(PLANT, LAB)
table(LAB, TIME)

# Measured active variables
table(O2COR, OXYGEN)
table(NEFFEKT, LOAD)
table(QRAT, PRSEK)

### Summary statistics for the outcome ----------------------------------------
```

```
mean(DIOX)
sd(DIOX)

tapply(DIOX, OXYGEN, mean)
tapply(DIOX, OXYGEN, sd)

tapply(DIOX, LOAD, mean)
tapply(DIOX, LOAD, sd)

tapply(DIOX, PRSEK, mean)
tapply(DIOX, PRSEK, sd)

tapply(DIOX, PLANT, mean)
tapply(DIOX, PLANT, sd)

tapply(DIOX, LAB, mean)
tapply(DIOX, LAB, sd)

tapply(DIOX, PRSEK, mean)
tapply(DIOX, PRSEK, sd)



### Other interisting summary statistics ------------------------------------

tapply(QRAT, PLANT, mean)
tapply(QRAT, PLANT, sd)

tapply(NEFFEKT, PLANT, mean)
tapply(NEFFEKT, PLANT, sd)

### Plotting the data -------------------------------------------------------

# Scatter plots of all variables
plot(DATA)


# Boxplot of the measured active variables vs the planned active variables
g1 <- ggplot(DATA, aes(x=OXYGEN, y=O2COR, colour=OXYGEN)) +
  geom_boxplot() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="none") +
  xlab("OXYGEN") + ylab("O2COR")

g2 <- ggplot(DATA, aes(x=LOAD, y=NEFFEKT, colour=LOAD)) +
  geom_boxplot() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="none") +
  xlab("LOAD") + ylab("NEFFEKT")

g3 <- ggplot(DATA, aes(x=PRSEK, y=QRAT, colour=PRSEK)) + geom_boxplot() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="none") +
  xlab("PRSEK") + ylab("QRAT")

grid.arrange(g1, g2, g3, nrow=1, ncol=3)


# Boxplots
g1 <- ggplot(DATA, aes(x=OXYGEN, y=DIOX, colour=OXYGEN)) +
      geom_boxplot() +
      theme(axis.text=element_text(size=14,face="bold"),
            axis.title=element_text(size=16,face="bold"),
            legend.position="none") +
      xlab("OXYGEN") + ylab("DIOXIN")

g2 <- ggplot(DATA, aes(x=LOAD, y=DIOX, colour=LOAD)) +
      geom_boxplot() +
      theme(axis.text=element_text(size=14,face="bold"),
            axis.title=element_text(size=16,face="bold"),
            legend.position="none") +
      xlab("LOAD") + ylab("DIOXIN")

g3 <- ggplot(DATA, aes(x=PRSEK, y=DIOX, colour=PRSEK)) + geom_boxplot() +
      theme(axis.text=element_text(size=14,face="bold"),
            axis.title=element_text(size=16,face="bold"),
            legend.position="none") +
      xlab("PRSEK") + ylab("DIOXIN")


g4 <- ggplot(DATA, aes(x=PLANT, y=DIOX, colour=TIME)) +
```

```
      geom_boxplot() +
      theme(axis.text=element_text(size=14,face="bold"),
            axis.title=element_text(size=16,face="bold"),
            legend.position="right") +
      xlab("PLANT") + ylab("DIOXIN")

g5 <- ggplot(DATA, aes(x=LAB, y=DIOX, colour=LAB)) +
      geom_boxplot() +
      theme(axis.text=element_text(size=14,face="bold"),
            axis.title=element_text(size=16,face="bold"),
            legend.position="none") +
      xlab("LAB") + ylab("DIOXIN")

grid.arrange(g1, g2, g3, g4, g5, nrow=2, ncol=3)


# Interaction plot for the planned active variables
par(mfrow=c(2,2))
interaction.plot(OXYGEN, LOAD, DIOX,
                 legend=TRUE, bty="n", col=1:8, xtick = TRUE)
interaction.plot(OXYGEN, PRSEK, DIOX,
                 legend=TRUE, bty="n", col=1:8, xtick = TRUE)
interaction.plot(LOAD, PRSEK, DIOX,
                 bty="n", col=1:3, xtick = TRUE)
par(mfrow=c(1,1))


# Scatter plots of the measured active variables (investigate different slope for PLANT)
g1 <- ggplot(DATA, aes(x=O2COR, y=log(DIOX), colour=LAB)) +
      geom_point() +
      theme(axis.text=element_text(size=14,face="bold"),
            axis.title=element_text(size=16,face="bold"),
            legend.position="right") +
      xlab("02COR") + ylab("DIOXIN")

g2 <- ggplot(DATA, aes(x=NEFFEKT, y=log(DIOX), colour=PLANT)) +
  geom_point() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("NEFFEKT") + ylab("DIOXIN")

g3 <- ggplot(DATA, aes(x=QRAT, y=log(DIOX), colour=PLANT)) +
  geom_point() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("QRAT") + ylab("DIOXIN")

grid.arrange(g1, g2, g3, nrow=2, ncol=2)


# Scatter plots of the measured active variables (investigate different slope for LAB)
g1 <- ggplot(DATA, aes(x=O2COR, y=log(DIOX), colour=LAB)) +
  geom_point() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("02COR") + ylab("DIOXIN")

g2 <- ggplot(DATA, aes(x=NEFFEKT, y=log(DIOX), colour=LAB)) +
  geom_point() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("NEFFEKT") + ylab("DIOXIN")

g3 <- ggplot(DATA, aes(x=QRAT, y=log(DIOX), colour=LAB)) +
  geom_point() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("QRAT") + ylab("DIOXIN")

grid.arrange(g1, g2, g3, nrow=2, ncol=2)

# Scatter plots of the measured active variables (investigate different slope for TIME)
g1 <- ggplot(DATA, aes(x=O2COR, y=log(DIOX), colour=TIME)) +
  geom_point() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("02COR") + ylab("DIOXIN")
```

```
g2 <- ggplot(DATA, aes(x=NEFFEKT, y=log(DIOX), colour=TIME)) +
  geom_point() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("NEFFEKT") + ylab("DIOXIN")

g3 <- ggplot(DATA, aes(x=QRAT, y=log(DIOX), colour=TIME)) +
  geom_point() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("QRAT") + ylab("DIOXIN")

grid.arrange(g1, g2, g3, nrow=2, ncol=2)


# . ------------------------------------------------------------------------
# Simple additive model - planned active effects ------------------------

### Specification of the initial model ----------------------------------

# We insert the missing values in PRSEK
table(PRSEK, useNA ="ifany")  # Observation 15 and 16 are missing
tapply(QRAT, PRSEK, mean, useNA="ifany")  # Observations seems to have been "L"
DATA$PRSEK[15:16] <- "L"
attach(DATA)


APV1 <- lm(DIOX ~ OXYGEN  + LOAD + PRSEK + PLANT + TIME + LAB)
summary(APV1)


### Model diagnostics ---------------------------------------------------

stdresid <- rstandard(APV1)

# Model diagnostics plots
par(mfrow=c(2, 2))
plot(APV1, which=1:4)
par(mfrow=c(1,1))

# BOXCOX transformation
boxcox(APV1)


### Specification of the log transformed model --------------------------
APV2 <- lm(log(DIOX) ~ OXYGEN  + LOAD + PRSEK + PLANT + TIME + LAB)
summary(APV2)


### Updated model diagnostics -------------------------------------------

stdresid <- rstandard(APV2)

# Wally plot
# wallyplot(APV2)

# Model diagnostics plots
par(mfrow=c(2, 2))                                                    # Note obs 15 and 16 influential.
plot(APV2, which=1:4)
par(mfrow=c(1,1))

# Normality plots
par(mfrow=c(1, 2))
hist(stdresid, main="", probability=TRUE, breaks=10)
curve(dnorm, -3, 3, col="red", lwd=2, add=TRUE)
plot(APV2, which=2)
par(mfrow=c(1, 1))


# BOXCOX transformation
boxcox(APV2)



### Model reduction -----------------------------------------------------

drop1(APV2, test = 'F')
APV2 <- update(APV2, . ~ . - PRSEK)

drop1(APV2, test = 'F')
```

```
# . ------------------------------------------------------------------------
# Simple additive model - measured active effects ------------------------

### Specification of the initial model ------------------------------------
AMV1 <- lm(DIOX ~ O2COR  + NEFFEKT + QRAT + PLANT + TIME + LAB)
summary(AMV1)


### Model diagnostics -----------------------------------------------------


# Model diagnostics plots
par(mfrow=c(2, 2))
plot(AMV1, which=1:4)
par(mfrow=c(1,1))

# BOXCOX transformation
boxcox(AMV1)


### Specification of the log transformed model ----------------------------
AMV2 <- lm(log(DIOX) ~ O2COR  + NEFFEKT + QRAT + PLANT + TIME + LAB)
summary(AMV2)

AMV3 <- lm(log(DIOX) ~ O2COR*NEFFEKT*QRAT + PLANT + TIME + LAB)

### Updated model diagnostics ---------------------------------------------

stdresid <- rstandard(AMV2)

# Wally plot
# wallyplot(AMV2)

# Model diagnostics plots
par(mfrow=c(2, 2))
plot(AMV2, which=c(1,3))
# par(mfrow=c(1,1))

# Normality plots
# par(mfrow=c(1, 2))
hist(stdresid, main="", probability=TRUE, breaks=10)
curve(dnorm, -3, 3, col="red", lwd=2, add=TRUE)
plot(AMV2, which=2)
par(mfrow=c(1, 1))

# BOXCOX transformation
boxcox(AMV2)


### Model reduction -------------------------------------------------------

drop1(AMV2, test = 'F')
xtable(drop1(AMV2, test = 'F')[c(1,3,5,6)])
AMV2 <- update(AMV2, . ~ . - QRAT)                                  # QRAT is not significant.

drop1(AMV2, test = 'F')
xtable(drop1(AMV2, test = 'F')[c(1,3,5,6)])
summary(AMV2)


### Interpretation of results ---------------------------------------------

# Question 4:
new_obs <- data.frame(TIME = factor(1), PLANT = "RENO_N", LAB = "KK",
                      O2COR = 0.5, NEFFEKT = -0.01, QRAT = 0.5)

new_pred1 <- predict(AMV2, new_obs, interval = "predict")
new_pred2 <- exp(new_pred1)
# We get at new prediction of DIOX = 338.55 with the prediction interval
# (124.31; 922.01).


# Question 5:
results1 <- data.frame(Parameter_estimate = summary(AMV2)$coef[,1],
                       Standard_deviation = summary(AMV2)$coef[,2],
                       Lower_CI = confint(AMV2)[,1],
                       Upper_CI = confint(AMV2)[,2])
xtable(round(results1, digits=2))

results2 <- exp(results1)[2:3,c(1,3,4)]
xtable(round(results2, digits=2))

# Notice that the operating condition QRAT is insignificant, while the other
```

```
# estimates are:
# O2COR = 0.1829
# NEFEKT = 2.9013.

# Remeber however that the response is log-transformed, so we get that
# a unit increase in O2COR yields a relative increase in DIOX of
# exp(0.1829) = 1.201 relative increase in DIOX from one unit increase in O2COR
# exp(2.9013) = 18.198 relative increase in DIOX from one unit increase in NEFFEKT.

# Thus lowering O2COR yiels a small decrease in DIOX and lowering NEFEKT yields
# a drastic decrease in DIOX.


# Question 6:
results3 <- exp(results1)[c(4,5,7),c(1,3,4)]
xtable(round(results3, digits=2))
remove(results1, results2, results3)

# Both lab and plant factor are significant, so there is a difference between the
# plants and the labs respectively.
#
# Looking at the parameters we see that
# KARA = 6.502
# RENO_N = 5.762
# RENO_S = 4.203
# so we get:
# exp(6.502-5.762) a relative increase in DIOX of 2.096 going from RENO_N to KARA.
# exp(5.762-4.203) a relative increase in DIOX of 4.754 going from RENO_S to RENO_N.
#
#
#
# Looking at the parameters we see that
# USA = 0.408
# so we get:
# exp(0-0.408) a relative increase in DIOX of 0.665 going from USA to KK.


# Emmeans.
em1 <- emmeans(AMV2, pairwise ~ PLANT, adjust="none")
em2 <- emmeans(AMV2, pairwise ~ LAB, adjust="none")

g1 <- plot(em1)
g2 <- plot(em2)

grid.arrange(g1, g2, nrow=1, ncol=2)


# . ---------------------------------------------------------------------------
# More advanced model ---------------------------------------------------------


### Finding the missing data ---------------------------------------------------

is.na(DATA)     # SO2: Obs 7, 8, 53 and CO: obs 8
DATA2 <- DATA[-c(7,8,53),]
is.na(DATA2)


### Model reduction (Forward selection) ----------------------------------------

# Use the data set with missing values removed
attach(DATA2)

AM3 <- lm(log(DIOX) ~ O2COR  + NEFFEKT + PLANT + TIME + LAB)

scope <- ~ . + QROEG + TOVN + TROEG + POVN + CO2 + CO + SO2 + HCL + H2O +       # Passive variables
  poly(O2COR,2) + poly(NEFFEKT,2) +  poly(QRAT,1) + poly(QRAT,2) +              # Active variables higher order terms
  PLANT:poly(O2COR,1)    + PLANT:poly(O2COR,2) +                                # PLANT interactions
  PLANT:poly(NEFFEKT,1) + PLANT:poly(NEFFEKT,2) +
  PLANT:poly(QRAT,1)     + PLANT:poly(QRAT,2) +
  LAB:poly(O2COR,1)      + PLANT:poly(O2COR,2) +                                # LAB interactions
  LAB:poly(NEFFEKT,1)    + PLANT:poly(NEFFEKT,2) +
  LAB:poly(QRAT,1)       + PLANT:poly(QRAT,2) +
  TIME:poly(O2COR,1)     + TIME:poly(O2COR,2) +                                 # TIME interactions
  TIME:poly(NEFFEKT,1)   + TIME:poly(NEFFEKT,2) +
  TIME:poly(QRAT,1)      + TIME:poly(QRAT,2) +
  poly(log(HCL),1) +  poly(log(HCL),2) +                                        # Higher order effects of log(HCL) an
  PLANT:poly(log(HCL),1) +  PLANT:poly(log(HCL),2) +
  LAB:poly(log(HCL),1) +  LAB:poly(log(HCL),2) +
  TIME:poly(log(HCL),1) +  TIME:poly(log(HCL),2) +
  poly(log(CO),1) +  poly(log(CO),2) +
  PLANT:poly(log(CO),1) +  PLANT:poly(log(CO),2) +
  LAB:poly(log(CO),1) +  LAB:poly(log(CO),2) +
  TIME:poly(log(CO),1) +  TIME:poly(log(CO),2)
```

```
add1(AM3, scope, test = "F")

AM3 <- update(AM3, . ~ . + poly(log(HCL), 1))
drop1(AM3, test = "F")
add1(AM3, scope, test = "F")

AM3 <- update(AM3, . ~ . + CO2)
drop1(AM3, test = "F")
add1(AM3, scope, test = "F")

AM3 <- update(AM3, . ~ . + TROEG)
drop1(AM3, test = "F")
add1(AM3, scope, test = "F")

AM3 <- update(AM3, . ~ . + TIME:poly(NEFFEKT, 1))
drop1(AM3, test = "F")
add1(AM3, scope, test = "F")

AM3 <- update(AM3, . ~ . + POVN)
drop1(AM3, test = "F")
add1(AM3, scope, test = "F")

AM3 <- update(AM3, . ~ . + poly(log(HCL), 2))
drop1(AM3, test = "F")
add1(AM3, scope, test = "F")

summary(AM3)
AM3 <- update(AM3, . ~ . - poly(log(HCL), 1) + poly(log(HCL), 2) -
                           NEFFEKT + TIME:poly(NEFFEKT, 1))
summary(AM3)



# Formula for the final model:
formula(AM3)


#Reestimate the final model with all observations
attach(DATA)
AM3 <- lm(log(DIOX) ~ O2COR + PLANT + TIME + LAB + CO2 + TROEG + POVN +
                      poly(log(HCL), 2) + TIME:poly(NEFFEKT, 1))


### Model diagnostics ------------------------------------------------------

stdresid <- rstandard(AM3)

# Model diagnostics plots
par(mfrow=c(2, 2))
plot(AM3, which=c(1,3,4))
#par(mfrow=c(1,1))

# BOXCOX transformation
# boxcox(AM3)

# Normality plots
#par(mfrow=c(1, 2))
hist(stdresid, main="", probability=TRUE, breaks=10)
curve(dnorm, -3, 3, col="red", lwd=2, add=TRUE)
plot(AM3, which=2)
par(mfrow=c(1, 1))

# Check for outliers
temp <- sort(stdresid)                                                  # Note observations 11 and 17 both ha
tail(temp, 20)                  # 1 obs with std.resid > 2.             # std resid above numerically > 2, ar
temp[1:20]                      # 1 obs with std.resid < -2             # they are among the most influential
length(temp) * (2 * pnorm(-2)) # Expect around 2.5 std.resid numerically > 2.
remove(temp)


# Check influence measures
infl <- influence.measures(AM3)$infmat
dim(infl)

# Leverage by residual plot
plot(AM3, which=4)

studresid <- studres(AM3)
leverage <- infl[,ncol(infl)]
cutoff <- 2 * nrow(summary(AM3)$coef) / length(studresid)
df_leverage_plot <- data.frame(obs=1:length(studresid),
                               Leverage = leverage, Rstudent = studresid)
```

```
ggplot(df_leverage_plot, aes(x=Leverage, y=Rstudent, label=obs)) +
  geom_point() +
    geom_text(aes(label=ifelse(abs(Rstudent)>2 | Leverage > cutoff,
                               as.character(obs),'')),hjust=-0.3,vjust=0) +
  geom_hline(yintercept=-2, linetype="dashed", color = "red") +
  geom_hline(yintercept=2, linetype="dashed", color = "red") +
  geom_vline(xintercept=cutoff, linetype="dashed", color = "red") +
    geom_text(aes(x=cutoff, y=-2.5, label = "2 p / n",hjust = 0.5)) +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="none")


# Normality tests
shapiro.test(stdresid)
lillie.test(stdresid)
cvm.test(stdresid)
ad.test(stdresid)

# Wally plot
# wallyplot(AM3)

# Plot of standardized residuals against the different regressors
par(mfrow=c(3,2))
plot(O2COR, stdresid)
abline(h=0, col="red")
plot(NEFFEKT, stdresid)
abline(h=0, col="red")
plot(log(HCL), stdresid)
abline(h=0, col="red")
plot(CO2, stdresid)
abline(h=0, col="red")
plot(TROEG, stdresid)
abline(h=0, col="red")
plot(POVN, stdresid)
abline(h=0, col="red")
par(mfrow=c(1,1))

# Boxplots
g1 <- ggplot(DATA, aes(x=PLANT, y=stdresid, colour=TIME)) +
  geom_boxplot() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="right") +
  xlab("PLANT") + ylab("stdresid")

g2 <- ggplot(DATA, aes(x=LAB, y=stdresid, colour=LAB)) +
  geom_boxplot() +
  theme(axis.text=element_text(size=14,face="bold"),
        axis.title=element_text(size=16,face="bold"),
        legend.position="none") +
  xlab("LAB") + ylab("stdresid")


grid.arrange(g1, g2, nrow=2, ncol=1)


### Post-Hoc analysis ----------------------------------------------------------

# Summary of models
summary(AM3)


# Parameters with confidence intervals.
results1 <- data.frame(Parameter_estimate = summary(AM3)$coef[,1],
                       Standard_deviation = summary(AM3)$coef[,2],
                       Lower_CI = confint(AM3)[,1],
                       Upper_CI = confint(AM3)[,2])
xtable(round(results1, digits=2))

# Variance estimate with confidence intervals.
sigma2 <- summary(AM3)$sigma^2
n <- length(stdresid)
k <- nrow(results1)
lower <- (n-k) * sigma2 / qchisq(0.975, df = n-k)
upper <- (n-k) * sigma2 / qchisq(0.025, df = n-k)

results2 <- data.frame(Variance = sigma2, Lower_CI = lower, Upper_CI = upper)
xtable(round(results2, digits=2))


################ QUESTION 9 ########################
model_optimization <- AM3
```

```
design_matrix = model.matrix(model_optimization)
n = dim(design_matrix)[1]
p = dim(design_matrix)[2]
y = log(df$DIOX)

objective = function(theta){
  y_hat = design_matrix %*% theta[1:p]
  Sigma_weighted = diag(n)
  diag(Sigma_weighted)[design_matrix[, "LABUSA"] == 0] = 1/theta[p+2]
  # print(diag(Sigma_weighted))
  # diag(Sigma_weighted)[design_matrix[, "LABUSA"] == 1] = (1 - theta[p+2])
  Sigma_weighted = theta[p+1]*Sigma_weighted
  result = sum(dmvnorm(y, mean = y_hat, sigma = Sigma_weighted, log = TRUE))
  return(-result)
}
theat_initial = c(rep(1, length(model_optimization$coefficients)), "sigma_squared" = 1,
                  "weight" = 1)
opt <- nlminb(theat_initial, objective)
opt$par

hessian_matrix = hessian(objective,opt$par)
standard_error = sqrt(diag(solve(hessian_matrix)))

c(opt$par["weight"] - qnorm(0.975)*standard_error[length(standard_error)],
  opt$par["weight"] + qnorm(0.975)*standard_error[length(standard_error)])

### PROFILE LIKELIHOOD CI
profile_objective <- function(weight){
  fun.tmp <- function(theta_inner, weight_input){
    objective(c(theta_inner, weight_input))
  }
  theat_initial_inner_opt = c(rep(1, length(model_optimization$coefficients)),
                              "sigma_squared" = 1)
  nlminb(theat_initial_inner_opt, fun.tmp, weight_input = weight)$objective
}
profile_objective(10)

p1 <- seq(-0.5, opt$par["weight"]*2.7, by = 0.01)
logLp1 <- sapply(p1, profile_objective) ## note sapply!
logLp1 <- logLp1 - min(logLp1) ## normalization
L_CI_lower = min(p1[exp(-logLp1) > exp(-qchisq(0.95,df=1)/2)])
L_CI_upper =  max(p1[exp(-logLp1) > exp(-qchisq(0.95,df=1)/2)])

obs_hessian <- hessian(profile_objective, opt$par["weight"])[1, 1]
quadratic_approximation <-
  exp( -0.5* obs_hessian * (p1 - opt$par["weight"])^2)
quadratic_approximation <- quadratic_approximation / max(quadratic_approximation)

par(mfrow = c(1,1))
plot(p1, exp(-logLp1), type = "l",
     xlab="Weight", ylab="Profile Likelihood",
     main="The comparison between Wald's and Likelihood based Confidence Intervals")
axis(side=1, at=seq(0, 10, by=1))
lines(p1, rep(exp(-qchisq(0.95,df=1)/2), length(p1)), col = 2)
rug(L_CI_lower, ticksize = 0.1, lwd = 2, col = "red")
rug(L_CI_upper, ticksize = 0.1, lwd = 2, col = "red")
c(L_CI_lower, L_CI_upper)
lines(p1, quadratic_approximation, col = "blue")
abline(v = opt$par["weight"], lty = 2)
rug(opt$par["weight"] - qnorm(0.975)*standard_error[length(standard_error)],
    ticksize = 0.1, lwd = 2, col = "blue")
rug(opt$par["weight"] + qnorm(0.975)*standard_error[length(standard_error)],
    ticksize = 0.1, lwd = 2, col = "blue")
legend("topright", 95,
       legend=c("Profile likelihood", "Quadratic approximation",
                "95% confidence interval"),
       col=c("black", "blue", "red"), lty = 1:1, cex=0.8,
       inset = 0.02)

c(opt$par["weight"] - qnorm(0.975)*standard_error[length(standard_error)],
  opt$par["weight"] + qnorm(0.975)*standard_error[length(standard_error)])
c(L_CI_lower, L_CI_upper)

wald_statistic = (opt$par["weight"] - 1)/standard_error[length(standard_error)]

# LRT
ll_full <- -profile_objective(as.numeric(opt$par["weight"]))
ll_test <- -profile_objective(1)
LRT <- -2*(ll_test - (ll_full))
p <- 1 - pchisq(LRT, df = 1)
p
```