Technical University of Denmark

02441 Applied Statistics and Statistical Software

Case 1: Effect of hardness and detergent on enzymatic catalysis

Group 5:
Begoña Bolós Sierra, s193036
Laura Sans Comerma, s192437
Jorge Montalvo Arvizu, s192184

January 2020

Summary

Given the importance of high catalytic activity of enzymes in the industry, it is desired to test and discover the best combination of enzyme given certain different parameters to obtain the best performance of stain removal capacity, i.e. enzyme concentration, enzyme type, detergent presence and hardness. An experiment was used to measure the performance of five different enzymes exposed to different conditions. Our general linear model explains almost 96% of variance of the response and the variables enzyme concentration, enzyme type, and detergent presence are of significant importance in the washing process; and two of their interactions. It is concluded that hardness does not affect the enzyme performance and the unique combination of using enzyme type A, with presence of detergent and with a concentration of 15 nM shows the best catalytic activity, thus removing stains better in the stain removal process. Finally, the experiment is affected by the systematic error of the day when the experiment was carried on.

Contents

1 Introduction **3**

2 Data **4**

3 Statistical Analysis **8**
3.A Model Selection 8
3.B Predictions 9
3.C Systematic Error 9

4 Results **10**
4.A Final model visualization 12
4.B Systematic Error 12

5 Conclusion **14**

6 Appendix **15**
6.A R Code 15

List of Figures

1 Pair plot of variables 5
2 Box plots of Response vs. independent variables 6
3 Box Plot of response against enzyme type and concentration 6
4 Scatter plots of response by concentration 7
5 Enzyme concentration behaviour tests 7
6 Box Cox of dependent variable 10
7 Model diagnostics of model lm3l 11
8 Visualization of the final model lm3l with prediction intervals. 12
9 Data Transformation Plot 13
10 Residuals plot and Q-Q plot of lm4b model 14

List of Tables

1 Variable Types 4
2 Variables summary statistics 4
3 ANCOVA results table of model lm3l 11
4 Model results summary of model lm3l 12
5 ANCOVA results of model lm4b 13

1 Introduction

Isolated enzymes are widely used in the industry as part of detergents to remove stains at different temperatures. The first patent that described the use of enzymes for that purpose was in 1913. Since then the enzyme business has grown around 10% per year, in which the market leader is the Danish company Novozymes, followed by DuPont. The manufacturers of detergents are the main customers needing the 34% of the produced enzymes [1]. Therefore, the sector has developed to incorporate new enzyme complexes and formulations, new methods to produce them in large-scale as fermentation and through biotechnological methods like protein engineering in order to optimize them. As a result of these optimizations, the cost of enzyme production has decreased significantly. The top used enzymes in the sector are proteases, amilases, celullases and lipases [1]-[2].

In the washing process, the customers seek for a high performance of the enzyme, which means that the enzyme is able to remove high amounts of protein from the surfaces. This performance in washing processes depends on some factors as the concentration of the enzyme used, the active components (detergents) and the concentration of some ions like Ca^{++}, which determines the hardness. In this experiment, the Surface Plasmon Resonance technology (SPR) was used to measure the performance of five different enzymes, exposed to different conditions of hardness and detergent. The results of the experiments are collected in the data set detailed in **Section 2**.

The aim of this project is to analyze the performance of the different enzymes given the different conditions in the experiment. We aim to answer the following main questions in this report:

1. How does hardness & detergent presence influence the catalytic activity?
2. Is the catalytic activity dependent on the amount of enzyme present?
3. Are there any differences in performance among the enzymes in this study regarding the factors mentioned above?
4. Are there indications of systematic errors due to one enzyme per experiment, how would this affect the model?

In order to answer these questions the data set is going to be analyzed using R version 3.6.2 (Dark and Stormy Night) through the RStudio IDE. In **Section 2**, the data set is inspected and visualized in order to identify the behaviour of the variables to obtain an initial hypothesis of the enzymatic performance; also to assess any necessary transformation on the variables. Then, in **Section 3** the process of model construction and selection is explained; different statistical tests will be carried out in order to ensure that the variables and their interactions included in the model accurately represent the dependant variable without using non-statistically significant variables. The most used test to obtain conclusions from the data set is the ANCOVA test, due to the presence of categorical and continuous variables. However, linear regression, ANOVA, normality, and correlation tests are also performed throughout this study. Finally, in **Section 4** the results of the study are presented and in **Section 5** we conclude and propose future analysis.

2 Data

The data consists of a cross-sectional structured data set of 160 samples with 80 different experimental combinations in seven variables; one dependent and six independent. In column order, the variables are the following: *RunDate*, *Cycle*, *Response*, *Enzyme*, *EnzymeConc*, *DetStock* and *CaStock*. Table 1 summarizes the variable types and their status in this study.

Variable	Status	Type
RunDate	Independent	Discrete nominal (day number)
Cycle	Independent	Numerical discrete (1 to 34 in steps of 1)
Response	Dependent	Continuous ratio
Enzyme	Independent	Discrete nominal (A, B, C, D, E)
Enzyme Concentration	Independent	Continuous ratio
DetStock	Independent	Discrete binary (0 or 1)
CaStock	Independent	Discrete binary (0 or 1)

Table 1: Variable names, status and types.

After a thorough review of the dataset, it is found no missing or atypical values. However, the variable *RunDate* is not in YYMMDD format as specified in the data description of the assignment, a missing zero at the beginning of each observation is assumed, e.g. 81203 refers to December 3, 2008. The *Response* variable refers to the amount of protein that is removed by the enzyme, which is the same as the capacity of removing stains, i.e. our dependent variable. To take a closer view of the data, the basic summary statistics were performed and are shown in Table 2.

RunDate	Cycle	Response	Enzyme	EnzymeConc	DetStock	CaStock
Days :5	Min. : 1.00	Min. : 0.10	A:32	Min. : 0.00	Det+:80	Ca+:80
	1st Qu.: 9.00	1st Qu.: 94.60	B:32	1st Qu.: 1.88	Det0:80	Ca0:80
	Median :17.50	Median : 322.40	C:32	Median : 5.00		
	Mean :17.38	Mean : 431.60	D:32	Mean : 6.25		
	3rd Qu.:25.25	3rd Qu.: 662.70	E:32	3rd Qu.: 9.38		
	Max. :34.00	Max. :1588.00		Max. :15.00		
	Std. :9.67	Std. :393.02		Std. :5.75		

Table 2: Summary statistics of the raw data.

It is seen from the table that the standard deviations of *Response* and *EnzymeConc* are quite large compared to their mean value. Also, the data set is balanced, by having the observations ($N = 160$) divided into equal groups by grouping the experiments by *RunDate*, *Cycle*, *Enzyme*, *DetStock*, and/or *CaStock*.

Figure 1 shows a so-called pairs plot where the possible relationships of each variable against each other can be seen. It is seen that the protein removal activity is modified depending on the enzyme type, enzyme concentration, and presence of detergent. On the contrary, it seems that the presence of calcium-ions does not affect the response. Also, there seems to be some an heteroscedasticity relation between *Cycle* and *Response*. Any any correlation is not present between independent variables since there's only one continuous variable and the rest are categorical.

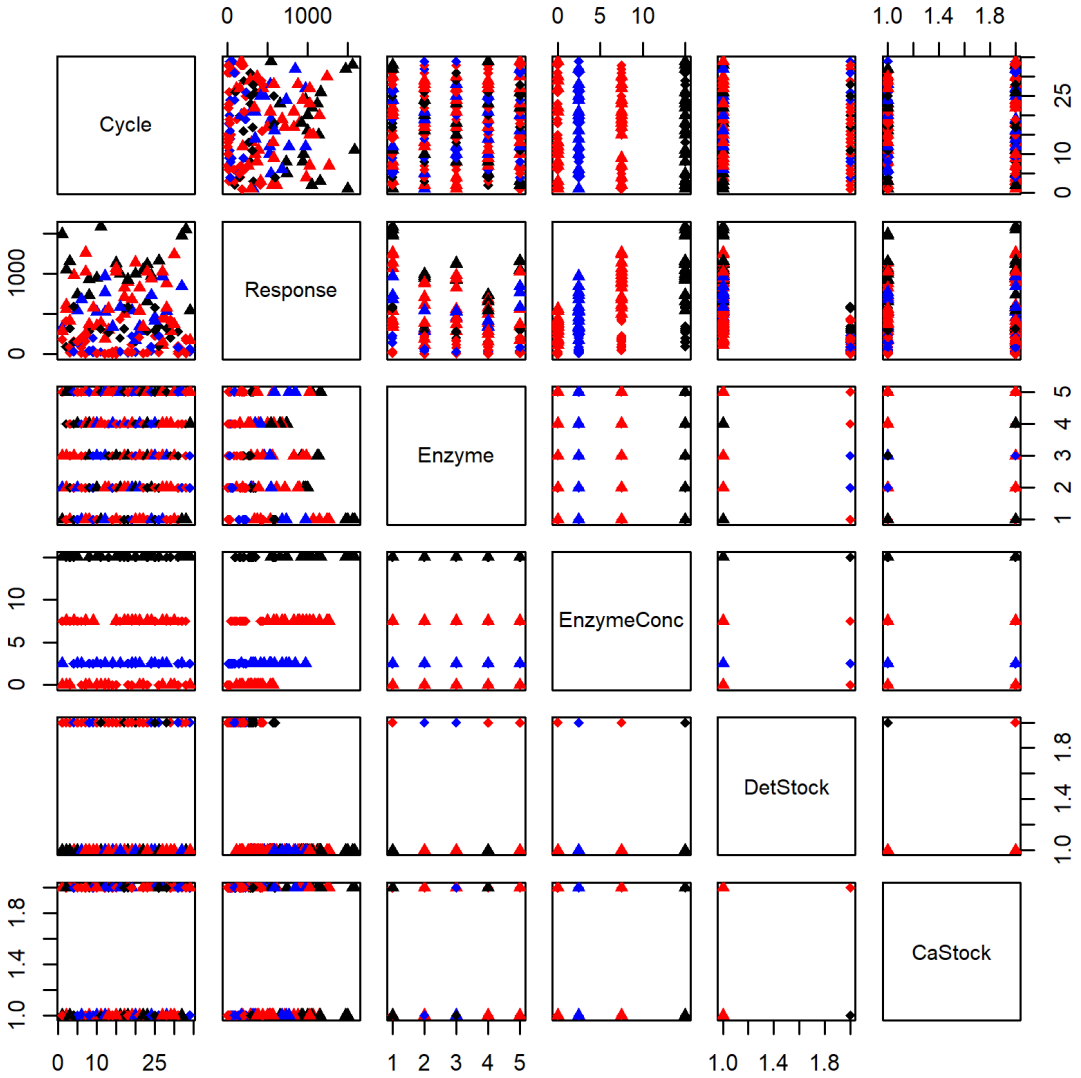

Figure 1: Pairs plot of the variables used in the statistical analysis.

In Figures 2 and 3, it is further observed the influence on the response dependant on the enzyme types, enzyme concentration, and presence or absence of detergent and calcium ions. In Figure 2, in the panel B the *EnzymeConc* seems to affect the response, which increases with the concentration, meaning that with more enzyme concentration there's more stain removal. Furthermore, in the panel C of the same figure, the detergent (in the left) seems to have an impact on the response and on the contrary, calcium ions presence or absence (hardness condition) does not seem to matter on the protein removal activity. In Figure 3, the Enzyme A appears to have the biggest influence and enzyme D appears to have the smallest differences between the response regarding the four different concentrations of the enzyme. Regarding enzyme concentration, it seems that the response of the enzyme increases when increasing the enzyme concentration too. Also, regarding detergent and calcium ions presence, it seems that the response is dependent on detergent presence but not on calcium ions presence. Finally, the response data seems to be normally distributed when looking at box plots A and C, but not on box plot B, indicating a possible transformation on the variable *EnzymeConc*.

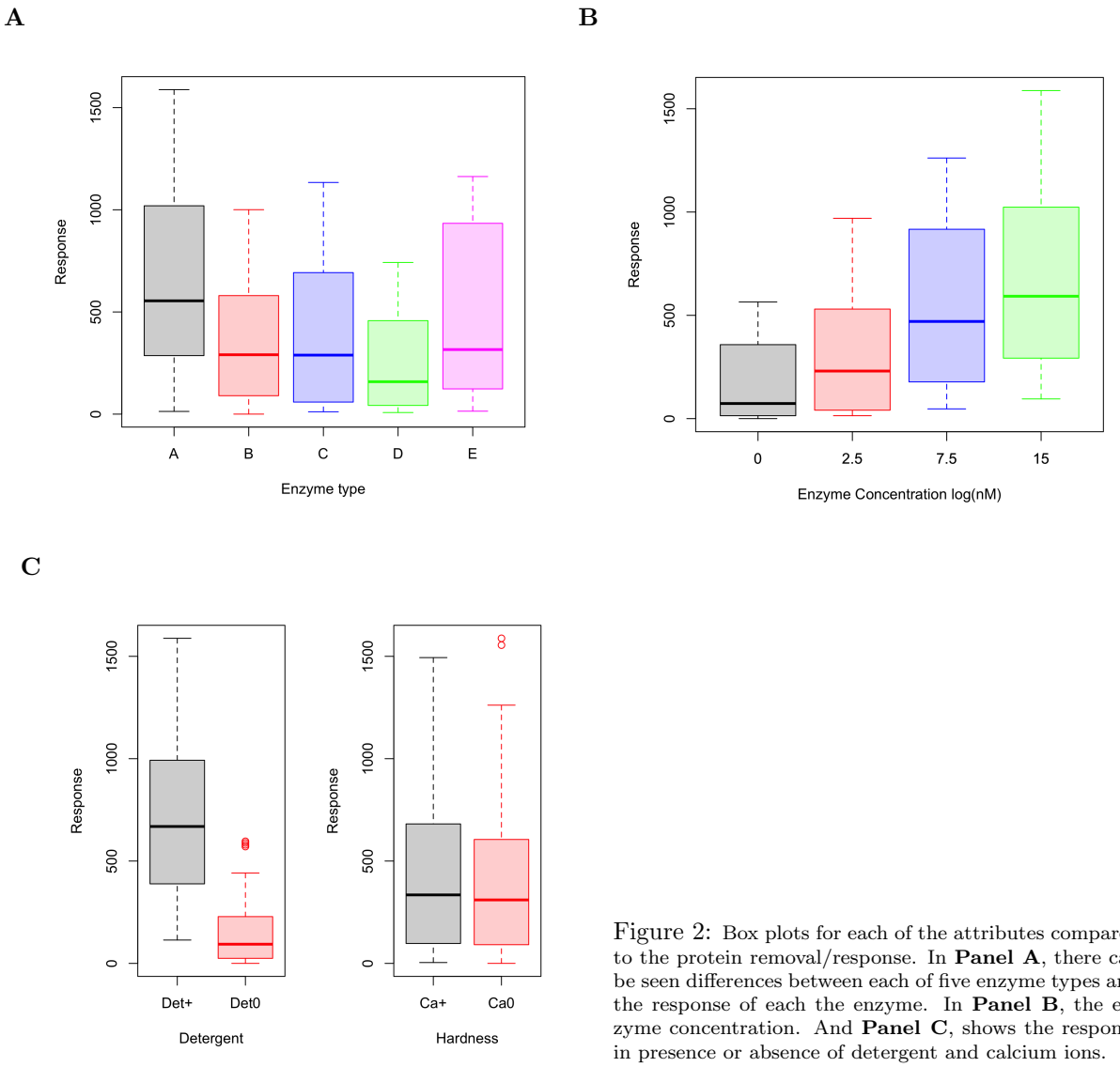

Figure 2: Box plots for each of the attributes compared to the protein removal/response. In **Panel A**, there can be seen differences between each of five enzyme types and the response of each the enzyme. In **Panel B**, the enzyme concentration. And **Panel C**, shows the response in presence or absence of detergent and calcium ions.

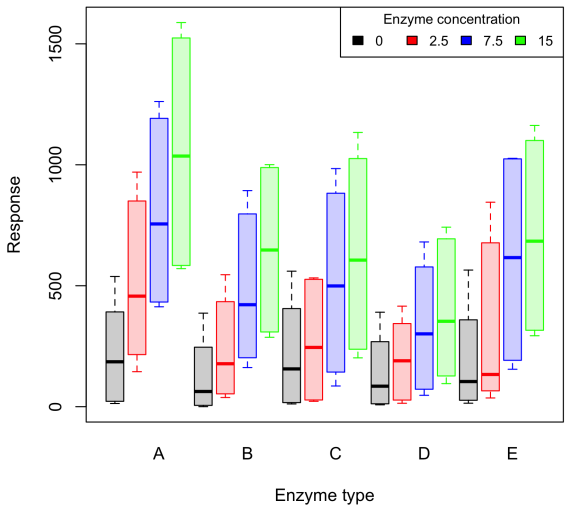

Figure 3: Grouped box plot of each enzyme type and enzyme concentration to see the distribution of the response.

In the left panel of Figure 4, it can be seen that there is a clear difference in the response level for each of the enzyme types when adding detergent or not. On the other hand, in the right panel (Calcium ions), there is no difference on adding or not calcium on the response level, indicated by the mix of diamonds (Ca0) and triangles (Ca+) throughout the plots.

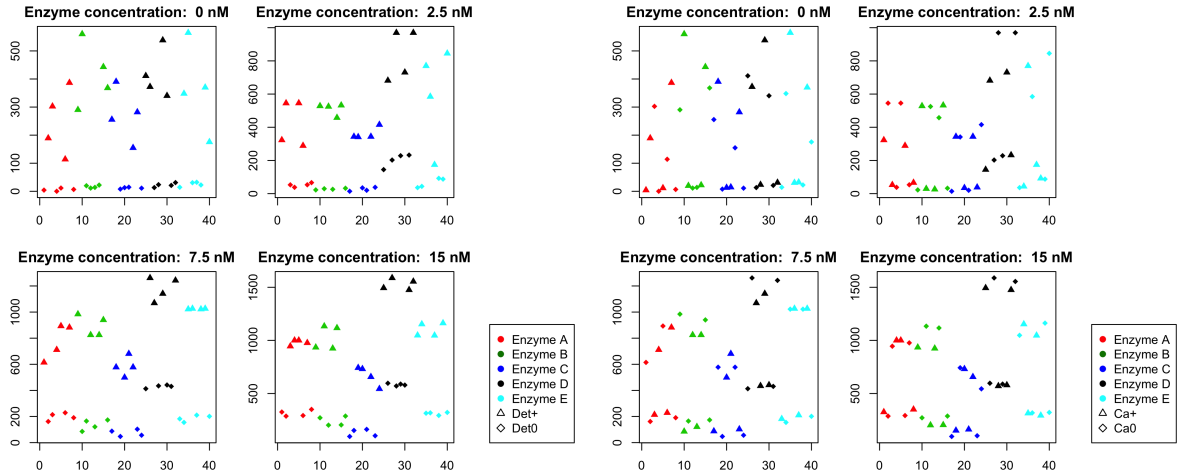

Figure 4: In the left panel, there is a representation of the enzyme concentration in each of the subplots, divided by color (type of enzyme) and the shape of the point distinguishes between Det0 (diamond) or Det+ (triangle). Moreover, in the right panel, there is again a representation of the enzyme concentration in each of the subplots, divided by type of enzyme and the shape of the point distinguishes between Ca0 (diamond) or Ca+ (triangle).

Given the hint that boxplot B of Figure 3 gave us about the non-linear behaviour of the variable *EnzymeConc*, it is investigated on Figure 5. The plot shows the raw data of the enzyme concentration indicated by four points (0, 2.5, 7.5 and 15 nM) and the different transformations that might match the raw data. This figure makes clear that the best fit for our data is either a logarithmic transformation (given the exponential behaviour) or the square root transformation (given the x^2 behaviour).

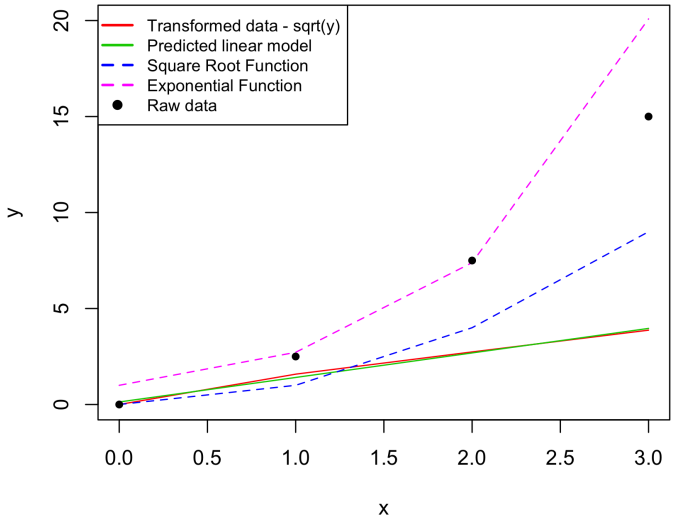

Figure 5: Plot of the concentration points (0, 2.5, 7.5 and 15 nM) and different transformations to the data.

3 Statistical Analysis

3.A Model Selection

To begin selecting the statistical model, we started by removing the variable *RunDate* from the data frame. This was done due to the fact that there was a 1-to-1 relationship between variables *RunDate* and *Enzyme*, thus it was decided that enzyme type was the important variable to test, instead of the day when the experiment was carried on.

To perform the statistical inference of the study it was used the general linear model, with which it could be carried out regression (when fitting the data to a continuous variable), one-way ANOVA (when comparing the mean of different groups) and two-way ANOVA (when comparing the mean of different groups at the second level or more), and ANCOVA (when mixing continuous and categorical variables). All of these statistical models are performed on the base of four main assumptions, which are tested with each model diagnostics:

1. Normality of residuals
2. Variance homogeneity
3. Variance should be independent of location
4. Linear relationship between independent variable and dependent variable

After cleaning the model and taking into account the assumptions of our statistical inference model, it was undertaken the following procedure to investigate a possible necessary transformation on the dependent variable *Response*:

1. Construct the **maximal model**[1]
2. Run diagnostics and Box-Cox function (check assumptions)
3. If lambda is different than 1 (and not inside the confidence interval), transform the dependent variable *Response*
4. If lambda is 1 (or inside the confidence interval), do not transform the dependent variable
5. Reduce the model by running the step function (backward selection) using the transformed dataset
6. Do sanity-check by running Box Cox function again on the reduced model, lambda should be 1 (or inside the confidence interval)

Box Cox transformation was used to find the optimal lambda value to find the optimal transformation to our data set. This transformation parts on the idea of optimizing the likelihood of the observations of the model and the formula of the transformation, after finding the optimal lambda, is the following:

$$f(y, \lambda) = (y^{\lambda} - 1)/\lambda$$
$$f(y, \lambda) = y^{\lambda}$$

The second formula was used as recommended by the professor since there was problem with negative values doing the full formula transformation.

After this procedure, it was proceeded with a check for a transformation in any of the independent variables. The maximal model was reduced again with backward selection and ran the diagnostics to see any relation between the residuals of the model against each independent variable. Then, the diagnostic plots were assessed to check for the compliance of the assumptions and decided on the transformation. From **Section 2**, there is a hint that it could be a log or sqrt transformation of the variable *EnzymeConc*.

Once both transformations were implemented, then the choice of the final model could be done by performing forward-selection instead of just using the step function (backward-selection). That way, the

[1]The maximal model is a linear model with the dependent variable subject to all possible interactions between independent variables, without taking into account the statistically significance of the variables and/or interactions. For this case, there were interactions between continuous and categorical variables.

interaction of each independent variable with the response can be seen.

ANCOVA was used in the general linear model, since there was a mix of categorical (*DetStock*) and continuous variables (*EnzymeConc*). The model can be explained with the following formula:

$$Y_{gi} = \alpha_g + \beta_g x_{gi} + \varepsilon_{gi}$$

where Y_{gi} is the response variable for each level for each group, α_g is the intercept of the model of each group, β_g is the slope in the model for each group, x_{gi} are the observations in each level for each group, and ε_{gi} is the noise.

However, one-way and two-way ANOVA were also performed when testing for detergent and hardness on response, the equation is very similar than ANCOVA but without the slopes:

$$Y_{gi} = \alpha_g + \varepsilon_{gi}$$

After the selection of the final model through forward-selection, a sanity-checked was done with a backward-selection of the 'full interactions' model and it confirmed the obtaining of the same final model. The diagnostics were ran again to check compliance with our assumptions of the general linear model, checked for normality on the residuals, and performed a final Box Cox transformation. At the end outliers were tested and observations that made the residuals' distribution non-normal were removed.

3.B Predictions

Once the final model was selected, the visualization of the final model was performed against the actual response. This visualization is useful to infer predictions from the model. In order to simplify the visualization, the data was divided in two depending on the presence or absence of detergent (*DetStock*). Subsequently, the function expand.grid was used in order to create combinations between the variables *DetStock*, *EnzymeConc* and *Enzyme*. In this case, 100 combinations were done with the variables above.

Moreover, the data was back-transformed to complete the visualization, due to the dependent (*Response*) and independent (*EnzymeConc*) variables were transformed with the Box Cox's lambda and the square root, respectively. In order to do so, the dependent variable (*Response*) was raised to 1/lambda and the independent variable (*EnzymeConc*) was squared. Finally, a function was created manually to perform the prediction intervals for each *Enzyme*. That has been archived using the functions predict and matlines.

3.C Systematic Error

Systematic errors are those caused due to inconsistencies in the experiments, such as environmental factors or machinery incorrectly calibrated. To ensure that the data set is free of this kind of errors the *RunDate* could be used to analyse them. This an important thing to take into account because if the errors are not checked they could void the rest of the statistical analysis performed. This is the reason why all experiments must have negative and positive controls.

There is an easy solution to this issue. It can be done by comparing the response of each day at concentration of enzyme equal to zero. The null enzyme concentration is used as a negative control and leaves out the influence of the enzyme type.

The same protocol as for the main statistical model was followed to check whether the *Response* is dependent on *RunDate* and its interactions with the other attributes. A maximal model linear was built adding the attributes *RunDate*, *DetStock* and *Cycle* against the *Response*. After this, statistical interference was performed using ANCOVA. Followed by the Box Cox function transformation and a diagnostics test to check assumptions, some of the attributes were left out of the model due to p-values > 0.05. Lastly, the final model was used to test for significance of the *RunDate* on the *Response*.

In addition, a Shapiro-Wilk normality test was performed to have a final proof of normal residuals of our final model.

4 Results

Regarding transformations, an optimal transformation parameter λ of 0.4646 was found, as shown in Figure 6. With this value the dependent variable *Response* was transformed with the Box Cox transformation formula from **Section 3**. As for the independent variables, both logarithmic and square root transformations were tested given the residuals of the response as a function of *EnzymeConc*; lower residuals were obtained when using the square root transformation and thus selected as the transformation of *EnzymeConc*.

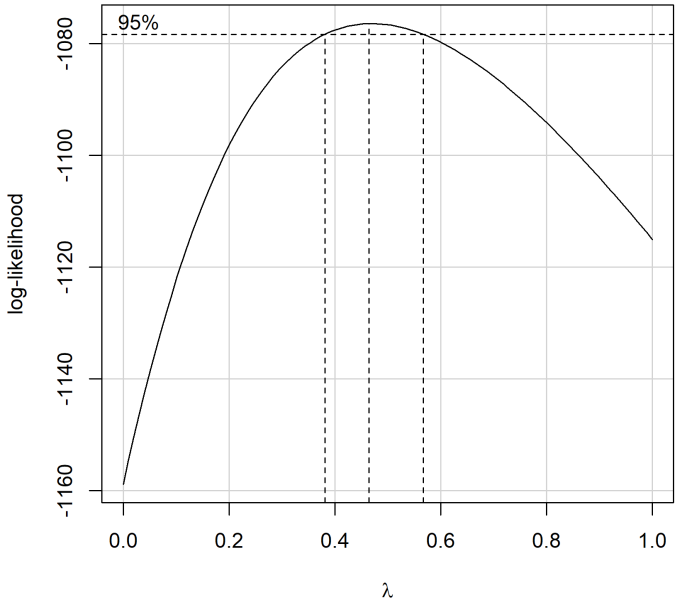

Figure 6: Optimal lambda of Box Cox transformation.

During our first one-way and two-way ANOVA on the general linear model with the transformed data, it is found that the variable *CaStock*, which is explained as the hardness, didn't show significant effect on response ($p>0.05$) and matched with the graphical visualization it is shown in Figure 4 of **Section 2**. However, *DetStock* did show statistical significance on response ($p<0.05$) and thus it is kept in the model.

Then, through the ANCOVA, it is found that *Enzyme* and *EnzymeConc* also were statistically significant on response ($p<0.05$). This time also one of the interactions was significant, being the interaction between *DetStock* and *Enzyme* the one it is kept in the forward-selected model.

Subsequently, tested all the interactions between our three selected variables *Enzyme*, *EnzymeConc*, and *DetStock*. Therefore obtaining the final model **lm3l**. The variables of this model are all of them significant (p-value 0.05) and the model presents a R-squared of 0.9572, thus explaining almost 96% of variance of the response. The model excludes the hardness or the presence of calcium. Table 3 shows the results of the ANCOVA analysis, there we can see the three variables and the two selected interactions in the model. We can see that the most important variable, thus affecting the response of the enzyme the most, is detergent presence.

	Sum Sq	Df	F value	Pr(>F)
data$DetStock	5368.09	1	2710.60	$< 2.2 \times 10^{-16}$ ***
data$Enzyme	831.20	4	104.93	$< 2.2 \times 10^{-16}$ ***
data$EnzymeConc	2508.13	1	1266.47	$< 2.2 \times 10^{-16}$ ***
data$DetStock:data$Enzyme	49.34	4	6.23	0.0001196 ***
data$Enzyme:data$EnzymeConc	149.73	4	18.90	1.703×10^{-12} ***
Residuals	283.20	143		

Table 3: ANCOVA results of the final model lm3l. Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'

Figure 7 shows the model diagnostics of the final model. We can see the first subplot as 'stars in the sky' where the variance of the residuals is homogeneous throughout the plot and no patterns are seen, the second subplot shows normality of the residuals in the qq-plot (we also tested with Shapiro-Wilk test for normality obtaining a value of p0.05, thus normality) since the points lay along the straight line with no skewness or kurtosis, the third subplot shows similar behaviour as subplot one as 'stars in the sky' with a straight horizontal line and no clear patterns, finally, subplot four shows no clear patterns and no observations with high leverage or within high Cook's distance.

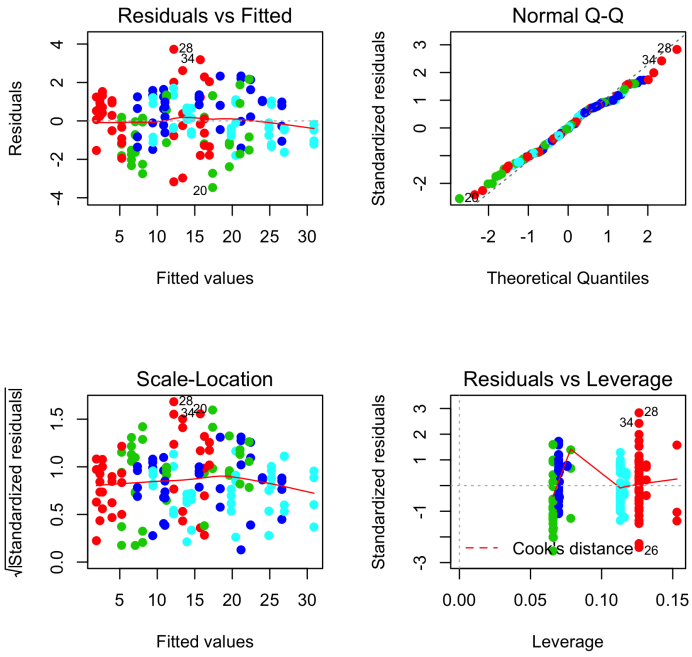

Figure 7: Model diagnostics.

Table 4 shows the summary of the final linear model, we can see the intercept α for categorical variables as the sum of the intercept and the groups; while the slopes β are the sum of the continuous variables and its interaction with the categorical variables, as shown before in **Section 3**. We can see that some interactions aren't statistically significant, being the interaction *DetStock* with *Enzyme* the less significant of the model's parameters with enzyme type B and D not statistically significant on the response when there's an absence of detergent.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	16.2881	0.5001	32.57	$< 2 \times 10^{-16}$ ***
data$DetStockDet0	-11.0271	0.4975	-22.16	$< 2 \times 10^{-16}$ ***
data$EnzymeB	-4.0849	0.7073	-5.78	4.60×10^{-08} ***
data$EnzymeC	-0.5490	0.7073	-0.78	0.43894
data$EnzymeD	-2.9034	0.7073	-4.11	6.76×10^{-05} ***
data$EnzymeE	0.6653	0.7436	0.89	0.37245
data$EnzymeConc	3.7698	0.1735	21.72	$< 2 \times 10^{-16}$ ***
data$DetStockDet0:data$EnzymeB	0.7056	0.7036	1.00	0.31766
data$DetStockDet0:data$EnzymeC	-2.0253	0.7036	-2.88	0.00461 **
data$DetStockDet0:data$EnzymeD	-0.0088	0.7036	-0.01	0.99004
data$DetStockDet0:data$EnzymeE	-1.9639	0.7168	-2.74	0.00693 **
data$EnzymeB:data$EnzymeConc	-0.4968	0.2454	-2.02	0.04480 *
data$EnzymeC:data$EnzymeConc	-1.3212	0.2454	-5.38	2.92×10^{-07} ***
data$EnzymeD:data$EnzymeConc	-1.9421	0.2454	-7.91	6.27×10^{-13} ***
data$EnzymeE:data$EnzymeConc	-1.1911	0.2503	-4.76	4.71×10^{-06} ***

Table 4: Summary of the linear model lm3l. Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'

4.A Final model visualization

After obtaining the final model **lm3l**, the visualization of the data was performed and the prediction intervals generated for each *Enzyme*. The visualization of this data is divided in two different plots depending on the absence (Fig. 8, Det0) or presence (Fig. 8, Det+) of detergent.

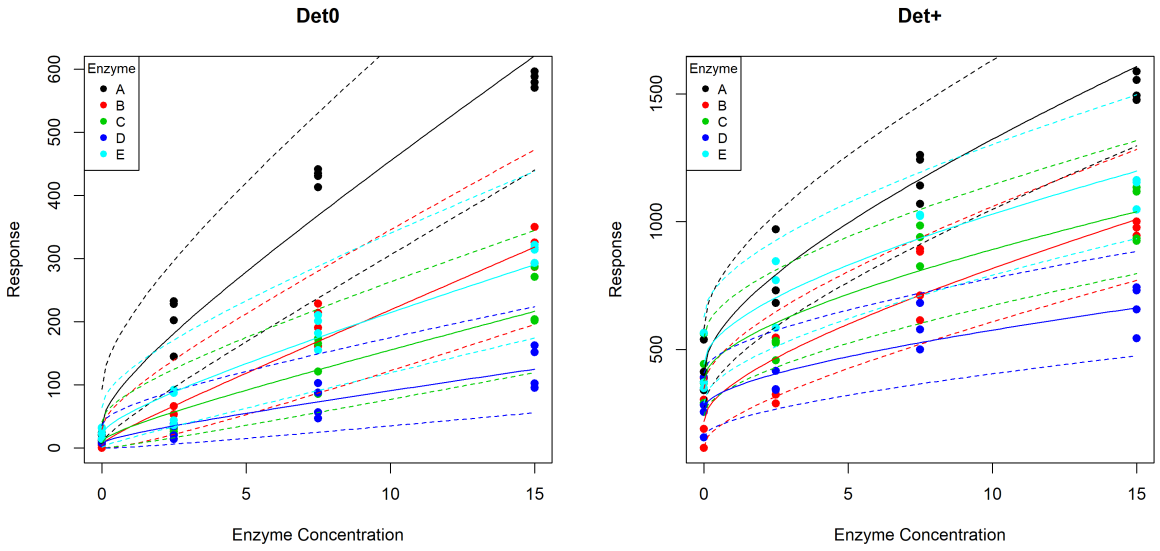

Figure 8: Visualization of the final model (lm3l) with the prediction intervals for each enzyme in absence (Det0) and presence (Det+) of detergent.

For both plots in Fig. 8, the dependent variable *Response* is represented against the independent variable *EnzymeConc* for each *Enzyme*. The plots show 95% prediction intervals for all enzymes in the two conditions (presence and absence of detergent). As expected, the model predicts a lower responses when the detergent is not present in the experiment, due to the presence of detergent had generated a higher protein removal in the experiment results. The predict intervals show a big range, thus referring to the big uncertainty of the model given the number of observations (2) per unique combination.

4.B Systematic Error

After building a the maximal model, it was found out that only the attribute *DetStock* was significant (p-value ≈ 0). After the Box Cox function was performed, Fig. 9 shows the transformation of *Response* with the optimal transformation parameter λ of 0.26263, to after reduce the model by running the step

function using the previously transformed data and to build the final model, **lm4b**.

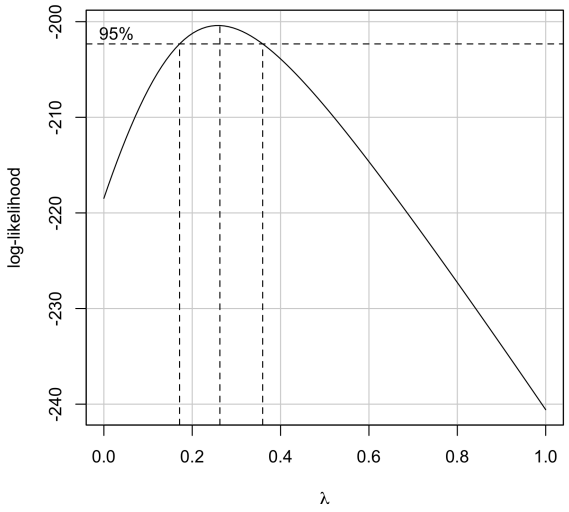

Figure 9: Optimal lambda of Box Cox transformation for lm4a model.

The final model included only the *DetStock* and *RunDate*, all the variables of the model were significant, p-value < 0.05 (Table 5, and the model presented a R-squared of 0.9347, thus explaining 93% of the variance of the response.

	Sum Sq	Df	F value	Pr(>F)
data$DetStock	66.70	1	463.55	$< 2.2 \times 10^{-16}$ ***
data$RunDate	3.38	4	5.88	0.0010 **
Residuals	4.89	34		

Table 5: ANCOVA results of the final model lm4b for the systematic error check. Signif. codes: 0 '***', 0.001 '**', 0.01 '*', 0.05 '.'

Figure 10 shows the diagnostics plot and the Q-Q distribution of the lm4b model. It can be seen that that the data was normally distributed as the plots indicate. However a Shapiro-Wilk normality test was also performed to ensure this.

```
Shapiro-Wilk normality test

data:  lm4b$residuals
W = 0.95511, p-value = 0.1138
```

The Shapiro-Wilk test gave a p-value > 0.05 indicating that the data was normally distributed.

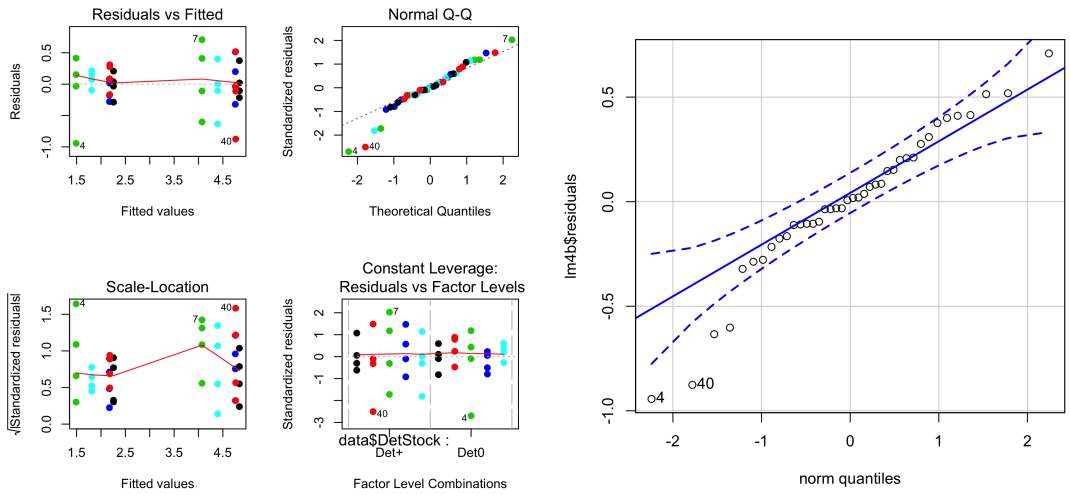

Figure 10: On the left, lm4b model diagnostics and on the right, Q-Q plot lm4b model distribution.

When piecing all this information together, the check of the dependence of the *RunDate* fails. This means that it must be confirmed that the date of the experiment does affect statistically the *Response*. Hence, the experimental data is affected from the start, because the experiment design and the data gathering depends also on the day, forcing that any analysis done after it will show different results.

To ensure that one can minimize the systematic errors, it is necessary to make multiple replicates, and all of them run each of the dates of the experiment, and in addition also take in to account the machinery calibration and environmental parameters such as temperature or humidity that might affect the results. When designing an experiment it is also very important to add positive and negative controls, that will help, like in this case, to verify the veracity of the analysis results.

5 Conclusion

Given the statistical analysis, we found that hardness does not have an effect on enzyme performance and the optimal model includes the variables and interactions from the final model **lm3l**, i.e. *DetStock*, *Enzyme*, *EnzymeConc*, *DetStock:Enzyme*, and *Enzyme:EnzymeConc*. Detergent presence is the variable that most affects the enzyme performance and enzyme performance also depends on enzyme type. Finally, we conclude that enzyme type A shows the highest catalytic properties and enzyme D shows the lowest performance; further enzyme A with concentration of 15 nM and with a presence of detergent is the highest performing combination of the assays. However, given the results of the systematic part in **Section 4** we encountered systematic bias given the day variation of the experiments, thus affecting the results.

For future work, we propose to analyze the enzyme concentration as a continuous variable to see if the results would have been different. Also, regarding the systematic error, a tighter control of the experiment should be taken to prevent different conditions between tests and more experiments per unique combination should be used, thus avoiding the high uncertainty given by only having two observations per unique combination.

6 Appendix

6.A R Code

```
##############################################################################
## Case 1:                                                                  ##
## Effect of hardness and detergent on enzymatic catalysis                  ##
##############################################################################

# Authors: Begona Bolos Sierra, Laura Sans Comerma, Jorge Montalvo Arvizu


# Load Data ------------------------------------------------------------------

require("car")
require("xtable")

# Load data and clean
data <- read.table("~/Github/02441_Applied_Statistics/Case1/2_Data/SPR.txt", header =
    TRUE, sep="\t")
df <- data
df$Stock <- as.factor(paste(as.character(df$DetStock),as.character(df$CaStock)))


# Transformation -------------------------------------------------------------

# Testing
# Set a new wd to save the images
# setwd("~/Github/02441_Applied_Statistics/Case1/4_Images")

#png(filename="test1_1.png", width=1750, height=1550, res=300)
par(mfrow=c(1,1))
y <- sort(unique(data$EnzymeConc))
x <- 0:3
par(mfrow=c(1,1))
y2 <- sqrt(y)
plot(x,y2, col=2, type="l", lwd=1, ylab='y', xlab='x',ylim=c(0,20),cex.axis=1)
lm <- lm(y2~x)
lines(x, predict(lm), col=3,lwd=1)
lines(x,x^2, col=4, lty= 2, lwd=1)
lines(x,exp(x), col=6, lty= 2, lwd=1)
points(y~x, pch=19,cex=0.7)
legend("topleft", legend = c("Transformed data - sqrt(y)", "Predicted linear model", "
    Square Root Function", "Logarithmic Function","Raw data"),
       col = c(2,3,4,6,1), lty=c(1,1,2,2,NA),lwd=1,pch=c(NA,NA,NA,NA,19), cex=0.8)
#dev.off()

# Could be square root or log, log gives a lower residual in this graph but in the model
    the residual is better with the sqrt

# Summary Statistics ---------------------------------------------------------

# Structure and summary of both data frames
str(data)
sum1 <- summary(data)
print(xtable(sum1, type = "latex"), file = "summary1.tex")
data <- data[,-1]

# Data Visualization ---------------------------------------------------------
#Set up colors
cols <- c("black","red", "blue", "green")
col_bg <- adjustcolor(cols, alpha = 0.2)
cols2 <- c("black","red", "blue", "green",6)
col_bg2 <- adjustcolor(cols2, alpha = 0.2)
cols3 <- c("black","red")
col_bg3 <- adjustcolor(cols3, alpha = 0.2)

# Pairs plot
#png(filename="pairs_1.png", width=1750, height=1750, res=300)
pairs(data, col=round(as.numeric(data$EnzymeConc))+2, pch=as.numeric(data$DetStock)+16)
#dev.off

# Plot Response - Stock
```

```
#png(filename="bp_response_stock_1.png", width=1750, height=1750, res=300)
par(mfrow=c(1,2))
plot(data$Response~data$DetStock, ylab="Response", xlab="Detergent",
     col=col_bg3, medcol=cols3, whiskcol=cols3, staplecol=cols3, boxcol=cols3, outcol=
         cols3,outbg=cols3)
plot(data$Response~data$CaStock, ylab="Response", xlab="Hardness",
     col=col_bg3, medcol=cols3, whiskcol=cols3, staplecol=cols3, boxcol=cols3, outcol=
         cols3,outbg=cols3)
#dev.off()

# Response - Enzyme
#png(filename="bp_response_enzyme.png", width=1750, height=1750, res=300)
boxplot(Response~Enzyme, data=df, xlab="Enzyme type", ylab="Protein removal (RU)",
        col=col_bg2, medcol=cols2, whiskcol=cols2, staplecol=cols2, boxcol=cols2, outcol=
            cols2, outbg=cols2)
#dev.off()

# Response - EnzymeConcentration
#png(filename="bp_response_conc.png",width=1750, height=1750, res=300)
a <- boxplot(Response~EnzymeConc, data=data, xlab="Enzyme Concentration log(nM)", ylab="
    Protein removal (RU)",
             col=col_bg, medcol=cols, whiskcol=cols, staplecol=cols, boxcol=cols, outcol=
                 cols,outbg=cols,
             names=c(0,2.5, 7.5,15))
axis(side= 1, at=seq_along(a$names), tick = FALSE, labels = a$names)
#dev.off()

# Response - Enzyme - EnzymeConcentration
#png(filename="bp_response_enzyme_conc.png", width=1750, height=1750, res=300)
par(mfrow = c(1,1))
b <- boxplot(Response ~  EnzymeConc + Enzyme, data = data, xaxt = "n", xlab="Enzyme type"
    ,
             col= col_bg, medcol=cols, whiskcol=cols, staplecol=cols, boxcol=cols, outcol
                 =cols,outbg=cols,
             names =c("","","A","","","","B","","","","C","","","","D","","","","E",""))
axis(side= 1, at=seq_along(b$names), tick = FALSE, labels = b$names)
legend("topright",title="Enzyme concentration", legend = c(0, 2.5, 7.5, 15), fill =cols,
    horiz =TRUE, cex=0.8)
#dev.off()

y <- sort(unique(data$EnzymeConc))

# Scatter Plot Detergent
y <- sort(unique(data$EnzymeConc))
#png(filename="response_det.png", width=2000, height=1750, res=300)
par(mfrow=c(2,2), oma =c(1,1,1,7), mar = c(3,2,2,2))
for (i in y){
  plot(data$Response[data$EnzymeConc==i], pch=as.numeric(data$DetStock[data$EnzymeConc==i
      ])+16, col=as.numeric(data$Enzyme[data$EnzymeConc==i]), ylab="Response", xlab="
      Observations", main=paste("Enzyme concentration: ",i,"nM"))
}
par(xpd=NA)
legend(x=50, y=1150,legend = c("Enzyme A","Enzyme B","Enzyme C", "Enzyme D", "Enzyme E",
    "Det+", "Det0"),
       col=c("red","#008000","blue","black",5,1,1), pch=c(19,19,19,19,19,2,5))
#dev.off()

# Scatter Plot Calcium
#png(filename="response_calcium.png", width=2000, height=1750, res=300)
par(mfrow=c(2,2), oma =c(1,1,1,7), mar = c(3,2,2,2))
for (i in y){
  plot(data$Response[data$EnzymeConc==i], pch=as.numeric(data$CaStock[data$EnzymeConc==i
      ])+16, col=as.numeric(data$Enzyme[data$EnzymeConc==i]), ylab="Response", xlab="
      Observations", main=paste("Enzyme concentration: ",i,"nM"))
}
par(xpd=NA)
legend(x=50, y=1150,legend = c("Enzyme A","Enzyme B","Enzyme C", "Enzyme D", "Enzyme E",
    "Ca+", "Ca0"),
       col=c("red","#008000","blue","black",5,1,1), pch=c(19,19,19,19,19,2,5))
#dev.off()


# BoxCox of maximal model ------------------------------------------------
```

```
126
127 # Maximal model with full interactions
128 lm1 <- lm(data$Response~data$Cycle*data$Enzyme*data$EnzymeConc*data$DetStock*data$CaStock
        )
129
130 # Diagnostics
131 Anova(lm1)
132 summary(lm1)
133 #png(filename="lm1.png", width=1750, height=1750, res=300)
134 par(mfrow=c(2,2))
135 plot(lm1, col=as.numeric(data$EnzymeConc)+1, pch=19)
136 #dev.off()
137
138 # Residuals
139 #png(filename="lm1_residuals.png", width=1750, height=1750, res=300)
140 par(mfrow=c(1,1))
141 plot(lm1$residuals~data$EnzymeConc, col=as.numeric(data$DetStock)+1, pch=19)
142 #dev.off()
143
144 # BoxCos Transformation
145 #png(filename="lm1_boxcox.png", width=1750, height=1750, res=300)
146 par(mfrow=c(1,1))
147 bc <- boxCox(lm1, lambda = seq(0, 1, by = 0.05))
148 lam1 <- bc$x[which.max(bc$y)]
149 #dev.off()
150
151 data$Response <- data$Response^lam1
152
153
154 # Model selection (2) ------------------------------------------------------------
155
156 # Model with full interactions
157 lm2a <- lm(data$Response~data$Cycle*data$Enzyme*data$EnzymeConc*data$DetStock*data$
        CaStock)
158 step(lm2a, k=3.8)
159
160 # Reduced model
161 lm2b <- lm(formula = data$Response ~ data$Enzyme + data$EnzymeConc + data$DetStock + data
        $Enzyme:data$EnzymeConc)
162
163 # Testing the model -------------------------------------------------------------
164
165 # Model diagnostics
166 Anova(lm2b)
167 summary(lm2b)
168
169 #png(filename="lm2.png", width=1750, height=1750, res=300)
170 par(mfrow=c(2,2))
171 plot(lm2b, col=as.numeric(data$Enzyme)+1, pch=19)
172 #dev.off()
173
174 # Residuals
175 #png(filename="lm2_residuals.png", width=1750, height=1750, res=300)
176 par(mfrow=c(1,1))
177 plot(lm2b$residuals~data$EnzymeConc, col=as.numeric(data$Enzyme), pch=19)
178 #dev.off()
179
180 # Checking BoxCox
181 #png(filename="lm2_boxcox.png", width=1750, height=1750, res=300)
182 par(mfrow=c(1,1))
183 bc2 <- boxCox(lm2b, lambda = seq(0, 2, by = 0.05))
184 lam <- bc2$x[which.max(bc2$y)]
185 #dev.off()
186
187 # Transforming the enzyme concentration
188 data$EnzymeConc <- sqrt(data$EnzymeConc)
189
190
191 # Model selection (3) ------------------------------------------------------------
192
193 # Testing response given detergent and hardness for data
194 lm3a <- lm(data$Response~data$DetStock)
195 Anova(lm3a)
```

```r

lm3b <- lm(data$Response~data$CaStock)
Anova(lm3b)

lm3c <- lm(data$Response~data$DetStock+data$CaStock)
Anova(lm3c)

lm3d <- lm(data$Response~data$DetStock*data$CaStock)
Anova(lm3d)
step(lm3d)

# Hardness doesn't seem to be significant, thus we keep increasing the complexity of the model without it by adding enzyme concentration

lm3e <- lm(data$Response~data$EnzymeConc)
Anova(lm3e)

lm3f <- lm(data$Response~data$DetStock+data$EnzymeConc)
Anova(lm3f)

lm3g <- lm(data$Response~data$DetStock*data$EnzymeConc)
Anova(lm3g)
step(lm3g)

# Interaction between detergent and enzyme concentration doesn't seem to be significant, add enzyme type

lm3h <- lm(data$Response~data$DetStock+data$Enzyme+data$EnzymeConc)
Anova(lm3h)

lm3i <- lm(data$Response~(data$DetStock+data$Enzyme)*data$EnzymeConc)
Anova(lm3i)

lm3j <- lm(data$Response~data$DetStock*(data$Enzyme+data$EnzymeConc))
Anova(lm3j)

drop1(lm3i, test="F")
drop1(lm3j, test="F")

lm3i <- update(lm3i, ~.-data$DetStock:data$EnzymeConc)
Anova(lm3i)

lm3j <- update(lm3j, ~.-data$DetStock:data$EnzymeConc)
Anova(lm3j)

BIC(lm3i, lm3j)
AIC(lm3i, lm3j)

# lm3i is better

# Full interactions without hardness
lm3k <- lm(data$Response~data$DetStock*data$Enzyme*data$EnzymeConc)
step(lm3k, test="F")

lm3l <- lm(formula = data$Response ~ data$DetStock + data$Enzyme + data$EnzymeConc + data$DetStock:data$Enzyme + data$Enzyme:data$EnzymeConc)
Anova(lm3l)

lm3m <- lm(data$Response~data$Cycle)
Anova(lm3m)

anova(lm3l, lm3i)
#p value is less than 0.05 so model lm3l is selected

# Testing the model ------------------------------------------------------------

# Diagnostics
Anova(lm3l)
summary(lm3l)
#png(filename="lm3.png", width=1750, height=1750, res=300)
par(mfrow=c(2,2))
plot(lm3l, col=as.numeric(data$Enzyme)+1, pch=19)
#dev.off()
```

```r

# Residuals
#png(filename="lm3_residuals.png", width=1750, height=1750, res=300)
par(mfrow=c(1,1))
plot(lm3l$residuals~data$EnzymeConc, col=as.numeric(data$DetStock), pch=19)
#dev.off()

# Testing BoxCox
#png(filename="lm3_boxcox.png", width=1750, height=1750, res=300)
par(mfrow=c(1,1))
bc3 <-boxCox(lm3l, lambda = seq(0, 2, by = 0.05))
lam <- bc3$x[which.max(bc3$y)]
#dev.off()

#png(filename="lm3_qqplot.png", width=1750, height=1750, res=300)
par(mfrow=c(1,1))
qqPlot(lm3l)
#dev.off()

# Should we remove the outliers?
data <- data[-c(147,160),]
lm3l <- lm(formula = data$Response ~ data$DetStock + data$Enzyme + data$EnzymeConc + data
    $DetStock:data$Enzyme + data$Enzyme:data$EnzymeConc)
Anova(lm3l)
summary(lm3l)

#png(filename="lm3_qqplot_wo_outliers.png", width=1750, height=1750, res=300)
par(mfrow=c(1,1))
qqPlot(lm3l$residuals)
#dev.off()
shapiro.test(lm3l$residuals)

# Residuals are normally distributed since p-value > 0.05

#png(filename="lml3.png",width=1750, height=1750, res=300)
par(mfrow=c(2,2))
plot(lm3l, col=as.numeric(as.factor(data$EnzymeConc))+1, pch=19)
#dev.off()


# Confidence Interval ------------------------------------------------------

lm3l <- lm(formula = Response ~ DetStock + Enzyme + EnzymeConc + DetStock:Enzyme + Enzyme
    :EnzymeConc, data=data)

# New x-data
new_data <- sqrt(seq(0, 15, length.out = 100))
new_data_grid <- expand.grid(EnzymeConc = new_data, Enzyme = levels(data$Enzyme),
    DetStock = levels(data$DetStock))

# Predictor plots
# Det0
#png(filename="pred_lm3l_det0.png", width=1750, height=1750, res=300)
par(mfrow=c(1,1))
det0 <- data[data$DetStock == 'Det0',]
new <- (det0$EnzymeConc)^2
plot((det0$Response^(1/lam1))~new, col= det0$Enzyme, pch=19, main="Det0", xlab="sqrt
    Enzyme Concentration", ylab="Response (BoxCox transformed)")

for (Enzyme in c(1,2,3,4,5)) {
  x0 <- new_data_grid[new_data_grid$Enzyme==levels(new_data_grid$Enzyme)[Enzyme]&new_data
      _grid$DetStock=="Det0",]

  pred0 <- predict(lm3l,
                   newdata = x0,
                   interval = "prediction")
  matlines (x0$EnzymeConc^2, pred0^(1/lam1), lty = c(1,2,2), lw = 1, col = Enzyme)
}
legend("topleft", legend=levels(data$Enzyme), col=1:nlevels(data$Enzyme),
       title="Enzyme", pch = 19, cex = 0.8)
#dev.off()

# Predictor plots
```

```
# Det+
#png(filename="pred_lm3l_detplus.png", width=1750, height=1750, res=300)
par(mfrow=c(1,1))
detplus <- data[data$DetStock == 'Det+',]
new <- (detplus$EnzymeConc)^2
plot((detplus$Response^(1/lam1))~new, col= detplus$Enzyme, pch=19, main="Det+", xlab="
    sqrt Enzyme Concentration", ylab="Response (BoxCox transformed)")

for (Enzyme in c(1,2,3,4,5)) {
  x0 <- new_data_grid[new_data_grid$Enzyme==levels(new_data_grid$Enzyme)[Enzyme]&new_data
      _grid$DetStock=="Det+",]

  pred0 <- predict(lm3l,
                   newdata = x0,
                   interval = "prediction")
  matlines (x0$EnzymeConc^2, pred0^(1/lam1), lty = c(1,2,2), lw = 1, col = Enzyme)
}
legend("topleft", legend=levels(data$Enzyme), col=1:nlevels(data$Enzyme),
       title="Enzyme", pch = 19, cex = 0.8)
#dev.off()


# Adding time ------------------------------------------------------------------

data <- read.table("~/Github/02441_Applied_Statistics/Case1/2_Data/SPR.txt", header =
    TRUE, sep="\t")
data <- data[data$EnzymeConc==0,]
data <- data[,-c(4,5)]
data$RunDate <- as.factor(data$RunDate)

# We can't just add run time because it's 1-to-1 with enzyme type
pairs(data, col=as.numeric(data$RunDate), pch=19)

# Maximal model
lm4a <- lm(data$Response~data$RunDate*data$Cycle*data$DetStock)
Anova(lm4a)

# Diagnostics
#png(filename="lm4a.png", width=1750, height=1750, res=300)
par(mfrow=c(2,2))
plot(lm4a, pch=19)
#dev.off()

# BoxCox
#png(filename="lm4a_boxcox.png", width=1750, height=1750, res=300)
par(mfrow=c(1,1))
bc4 <- boxCox(lm4a, lambda = seq(0, 1, by = 0.05))
lam <- bc4$x[which.max(bc4$y)]
#dev.off()

# Transforming the response
data$Response <- data$Response^lam

# Adding RunDate
lm4b <- lm(data$Response~data$DetStock*data$RunDate)
step(lm4b, k=3.8)
drop1(lm4b, test="F")

#lmtest <- lm(data$Response~data$DetStock+data$RunDate+data$CaStock)
#Anova(lmtest)

lm4b <- update(lm4b, ~.-data$DetStock:data$RunDate)
Anova(lm4b)

#png(filename="lm4b.png", width=1750, height=1750, res=300)
par(mfrow=c(2,2))
plot(lm4b, pch=19, col=as.numeric(data$RunDate))
#dev.off()

#png(filename="lm4b_residuals.png", width=1750, height=1750, res=300)
par(mfrow=c(1,1))
plot(lm4b$residuals, pch=19, col=as.numeric(data$RunDate))
#dev.off()
```

```
404
405 shapiro.test(lm4b$residuals)
406
407 #png(filename="lm4a_qqplotresiduals.png", width=1750, height=1750, res=300)
408 qqPlot(lm4b$residuals)
409 #dev.off()
410
411 kruskal.test(data$Response ~ data$RunDate)
412
413 ##############################################################################
```

References

[1] W. Rähse, "Production of tailor-made enzymes for detergents.," *Chembioeng Reviews*, vol. 1, no. 1, pp. 27–39, 2014.

[2] F. Rigoldi, S. Donini, A. Redaelli, E. Parisini, and G. A, "Review: Engineering of thermostable enzymes for industrial applications.," *APL Bioengineering*, vol. 1, no. 2, 2018.