Technical University Of Denmark

Applied Statistics and Statistical Software (02441)

Detergent Case: Effects of hardness and detergent on enzymatic catalysis

February 25, 2023

Authors

Marie Murmann Kragh S203566

Lasse Schnell Danielsen S203512

Johanna Munch Haraldsdottir S204657

Summary

The laundry industry is very dependent on the use of enzymes. Their catalytic efficiency has a big impact on the user product. The performance of the enzymes depends on various factors. In this study, the goal has been to investigate the relationship between different enzymes, enzyme concentration, and the addition of both detergent and calcium to the catalytic efficiency. The catalytic efficiency has been measured using Surface Plasmon Resonance technology (SPR).

The data were modeled using a General Linear Model (GLM). It was derived from a maximal model with backward selection. Here the model selection was based on F-statistics using a significance level of 5%.
Based on this, a model was made that approximately fulfilled the model assumptions and therefore could be used for estimation and investigation of the variables influence and correlation.
From this model we found the different enzymes used in the study to have significantly different catalytic efficiency across enzyme concentrations. In addition to this, for the different enzyme concentrations, the use of Calcium was found to result in lower catalytic efficiency. Lastly, the use of detergent was found to result in a constant increase in catalytic efficiency not dependent on enzyme concentrations. This is possibly due to it itself having a catalytic effect.

Contents

1 Introduction

The catalytic ability of enzymes is of high importance in the detergent industry, with the aim of removing stains in textile surfaces. Various factors can affect the enzyme performance in the wash process. Especially interesting variables of investigation could be Calcium-ions (hardness), surface active components (detergents) and the enzyme concentration. In the experiment the enzyme performance has been measured as the ability of removing protein from surface. The method used to determine the performance is called Surface Plasmon Resonance technology (SPR) and is used to detect real-time information of binding events. As all methods, this could carry a variances as well to the data.

Firstly is the parameters compared, and the influence of the variables investigated. A transformation of the data is examined and a general linear model is created. Lastly will insignificant parameters be removed through model selection, and an analysis of the performance among the enzymes be made.

2 Description of Data

This Case investigates how different variables affect the efficiency of laundry powder. The data is collected over a period of ten days, where each experiment took two days and included one enzyme. The four conditions tested for each experiment were (Det0, Ca0), (Det0, Ca+), (Det+, Ca0) and (Det+, Ca+) tested at four enzyme concentrations (0 nM, 2.5 nM, 7.5 nM and 15 nM). All experiments were replicated and analyzed in random order. The data consists of 160 observations for each variable. There are 6 independent variables and a dependent variable called response which represents the catalytic response (Table 1).

Table of data		
Rundate	Actual date of experiment	YYMMDD format
Cycle	Cycle number within run	Numeric value
Enzyme	Type of enzyme	5-level factor
EnzymeConc	Enzyme concentration [nM]	Numeric value
DetStock	Detergent + / 0	2-level factor
CaStock	Hardness (Ca + / 0)	2-level factor
Response	Catalytic activity	Numeric value

Table 1: Table of data, 6 different independent variables that affect the laundry powder, and a dependent response variable.

3 Statistical Analysis

3.1 Descriptive Analysis

To investigate the relationship between the 6 variables and response is each variable plotted against one another(see figure 1). In the relation between response and respectively rundate and enzyme there is a clear separation into colour groups. This indicates a strong influence of the enzyme group on the catalytic response since each color group represents an enzyme. The same tendency is seen between the enzyme and the enzyme concentration.

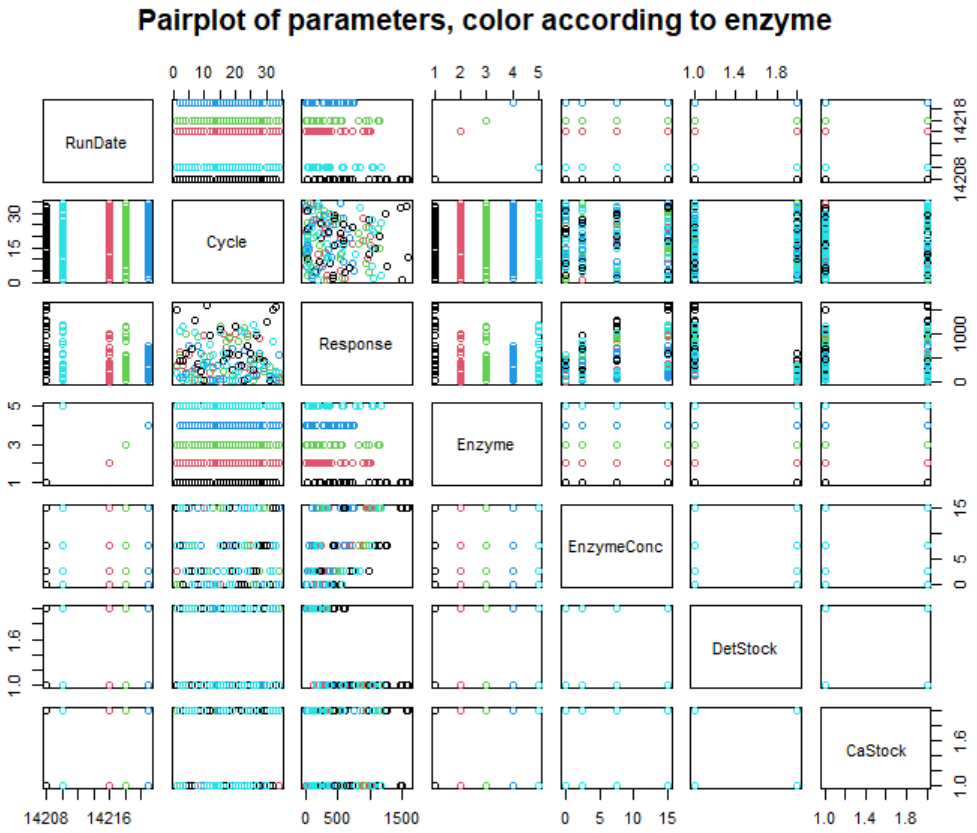


Figure 1: Pairsplot showing all variables plotted pairwise aginast each other. All plots has been colored by enzyme type.

To compare and illustrate the relationship between the catalytic response and the different variables, different box plots were created. The different enzymes seem to carry different catalytic responses(figure 2). The highest median response is found for enzyme A and the lowest is seen for Enzyme D. Overall this boxplot could indicate that the enzymes have a significant influence on the catalytic response.

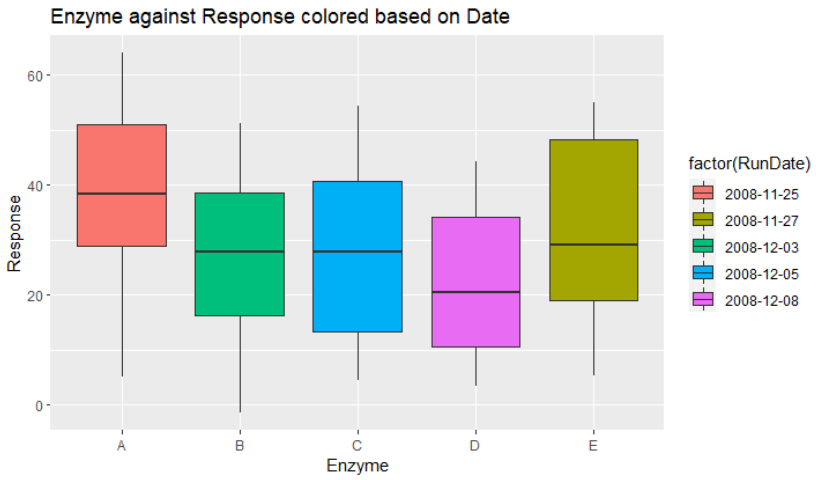


Figure 2: Boxplot of catalytic response against enzyme type. Each Box has been colored by the variable Rundate.

When comparing responses affected by detergent or not, a difference is seen between the medians as well as the first and third quantiles(figure 3). However does the adjective calcium not seem to have a significant difference when looking at the catalytic response. This is illustrated by the medians and the quantiles being very close in range of one another(Figure 3). When adding the factor of the different enzymes to the relation between detergent addition and catalytic response, a tendency of a higher response is seen when detergent has been added (Figure 3). The same tendency can however not be seen for the addition of calcium (hardness) from this illustration of the data.

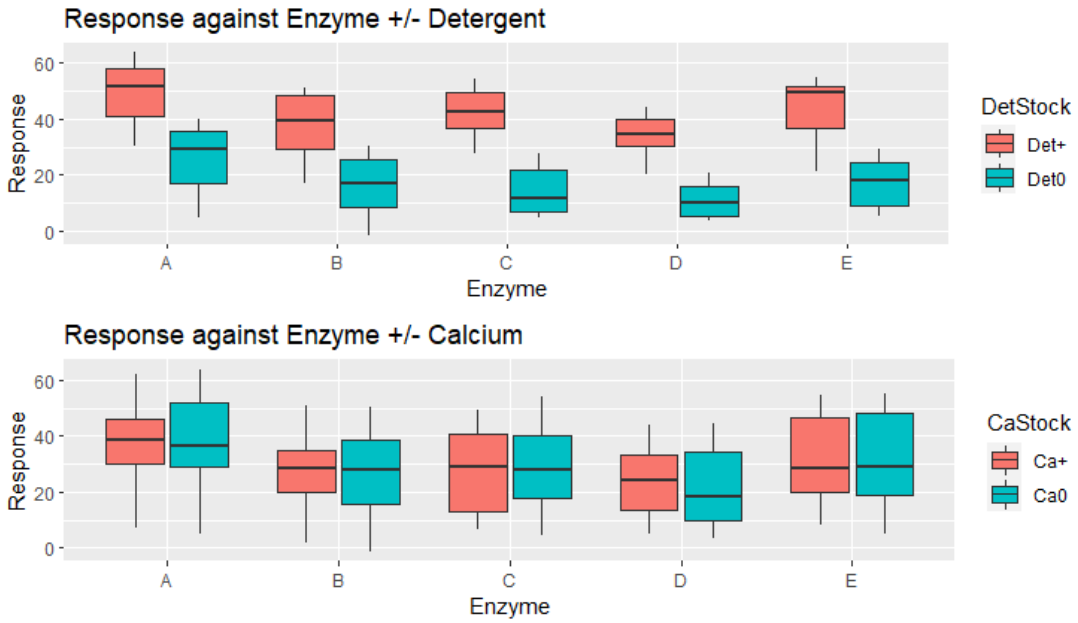


Figure 3: Boxplot of response against enzyme, Where each plot has been colored by respectively detergent stock and Calcium stock for each enzyme type.

3.2 Model assumptions and transformation of model

From the variables is a general linear model (GLM) created which acts as our maximum model. The model assumptions are checked to ensure that the model is valid for analysis and representation of the data. When checking the model assumptions, the residuals and their distribution is investigated (Figure 4).

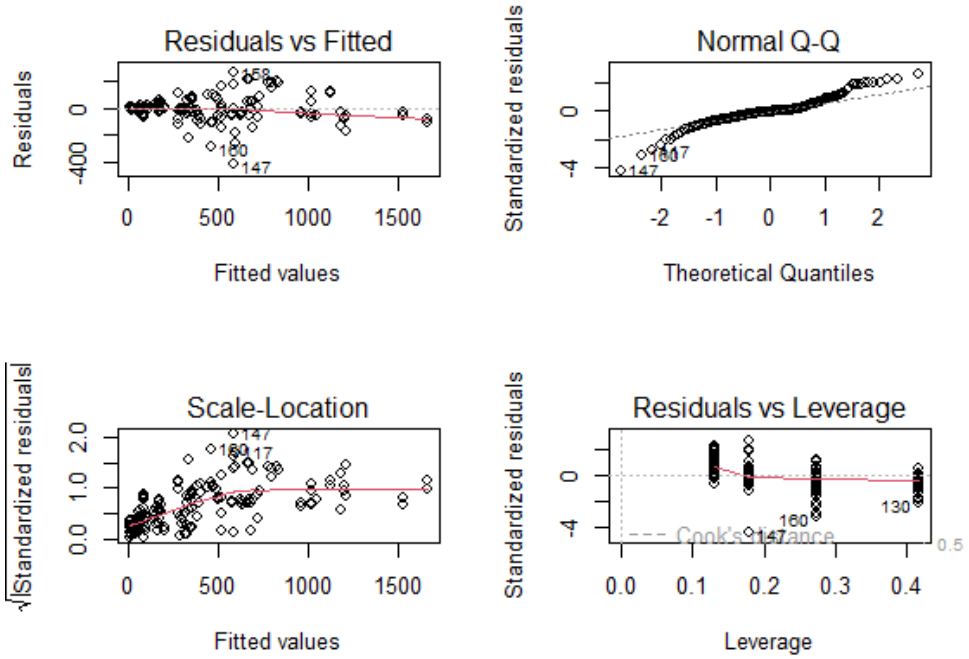


Figure 4: A plot of the regression model made to check model assumptions of the residuals. All axes title are present for each of the plots.

Generally does the residuals not seem to fit on the model assumptions of independent and identically distributed residuals (i.i.d.), since they seem to be clustered more to the left side. Additionally does the QQ-plot not show a linear tendency, which is seen at both the start and end of the data points - respectively below and above the tendency line. Based on this conclusion is a transformation of the model performed.

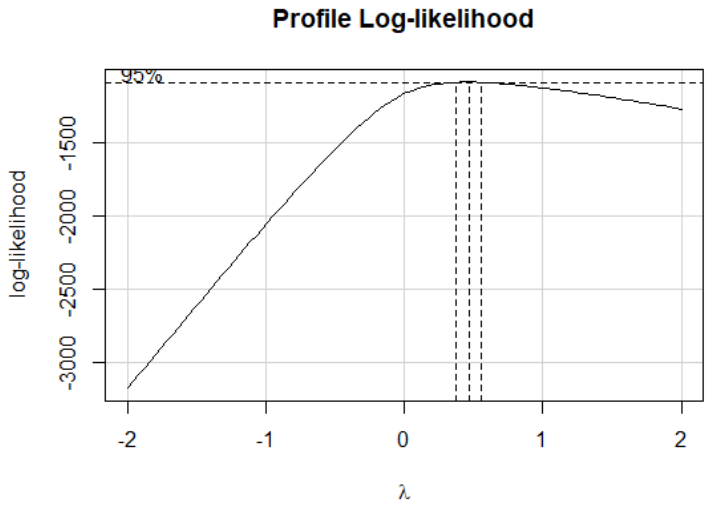


Figure 5: BoxCox plot showing the correlation coefficient with a maximum at $\lambda = 0.46$.

To investigate which transformation is required for the optimal model the BoxCox method is used[1] based on a resemblance to a normal distribution of the target variable - the catalytic response (Figure 5). The optimal Lambda as the maximum of the graph is then utilized in the transformation of the catalys response by the formula; $f(y, \lambda) = \frac{y^{\lambda}-1}{\lambda}$. The optimal λ is found as 0.46. From the transformation, a new model is made which by the BoxCox definition should be more suitable for representing the data. The model assumptions for the transformed model are then checked once more, to validate the distribution of the residuals (Figure 6).

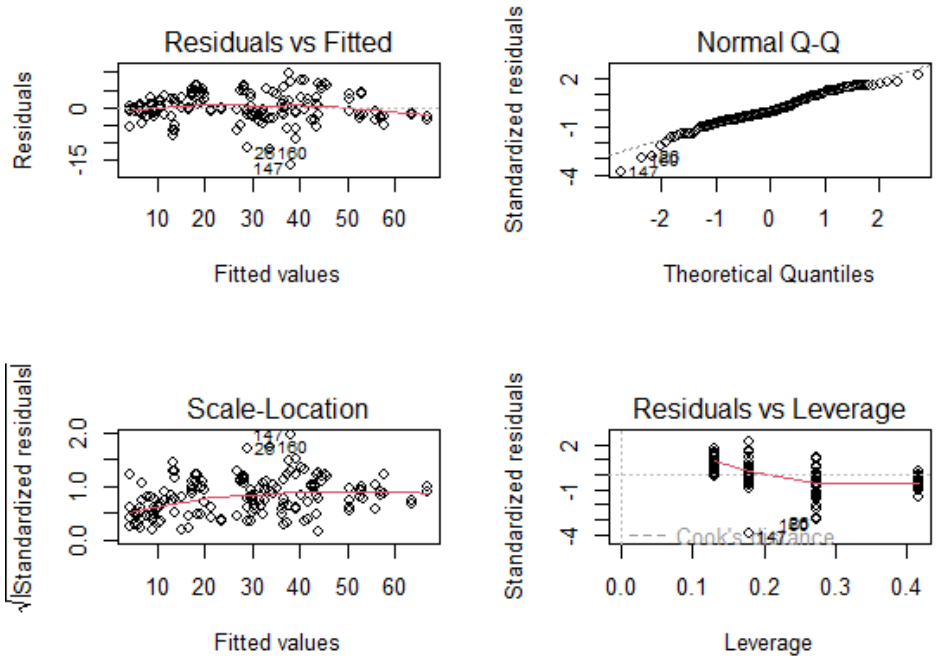


Figure 6: Residual plot after transformation

The residuals of the transformed data are much more scattered and evenly located compared to the non-transformed model, hence showing a more normal distribution (figure 6). This is clear in the Residuals Vs. fitted values as well as the Scale-location plot. The linear tendency of the normal distribution is also more distinct in the QQ plot. Since the transformed response is accepted as better, the transformed value will from now on be referred to as response. The correlation between the transformed response and the variables is then investigated for potential transformations. Since the Enzyme concentration is the only continuous variable, and therefore the only one relevant to investigation of correlation. Hence, a possible transformation of the enzyme concentration.

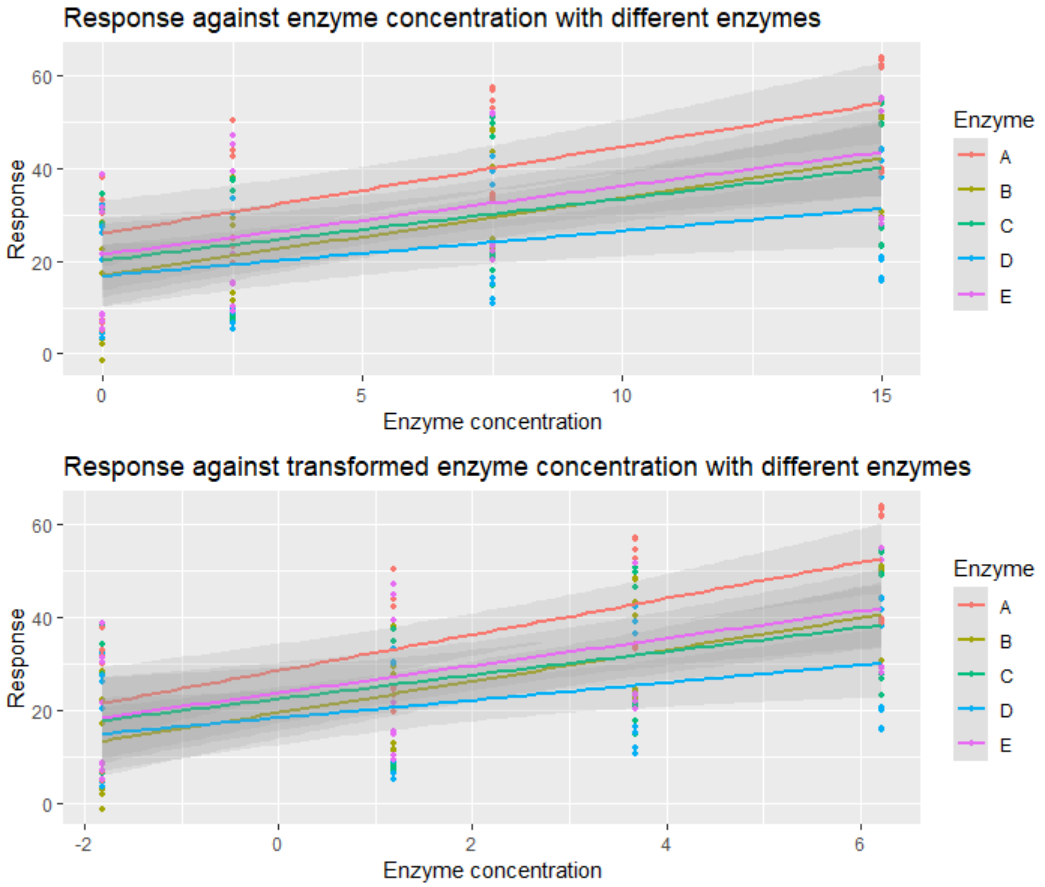


Figure 7: Scatterplot of the transformed Response as a function of the Enzyme concentration respectively non-transformed and transformed.

The correlation between the transformed response and the enzyme concentration is close to linear, but from the data points a tendency from the power function can be seen, hence improved. To transform the x-value the R function Boxtidwell() is used. The maximum λ value is found to be 0.55, and is therefore the optimal value to use for transformation of enzyme concentration. This has been calculated from the same formula used for y-value transformation. One could however conclude, that since both λ values are close to 0.5, the square-root function would also have been a reasonable estimation of transformation. When comparing the transformed and non-transformed enzyme concentrations is it clear, that the linearity is heavily increased at the transformed x-value (Figure 7). This is especially clear when looking at the positions of the data points.

3.3 Model selection

When performing an analysis of covariance(Ancova) on the transformed maximum model it is discovered that the model has aliased coefficients - variables that are perfectly correlated. When investigating these aliased coefficients is it further discovered that rundate and enzyme contribute with the same knowledge. Based on these discoveries rundate is removed from the model, this has been done to make the estimation of parameters more precise as well as avoid the failure of showing significance. The multicollinearity[2] is also indicated in figure 2, where each date is coupled with a specific enzyme indicating a clear correlation between date and enzyme.

The model is reduced by using a program for backward selection based on the F-statistics. This is done by the step function in R which leaves us with only significant variables and interactions when using a significance level at $\alpha = 5\%$.

$$Y_{gi} = \mu_g + \beta_g x_{gi} + \varepsilon_{gi} \tag{1}$$

$$Y_{gi} = \mu_0+\mu_B+\mu_C+\mu_D+\mu_E+\mu_{Det0}+\mu_{Ca0}+(\beta_0+\beta_B+\beta_C+\beta_D+\beta_E+\beta_{Ca0})\cdot x_{gi}+\varepsilon_{gi} \qquad \varepsilon \sim N(0,\sigma^2) \quad \varepsilon \text{ is i.i.d} \tag{2}$$

On the reduced model an analysis of covariance is performed. The variable "CaStock" does not appear to be significant (Table 2). However, since it is a part of a significant interaction, the "EnzymeConc:CaStock" interaction, it can not be reduced.

	Sum Sq	Df	F value	Pr(>F)
Enzyme	3760.98	4	71.31	1.85e-33
EnzymeConc	12225.07	1	927.15	2.31e-65
DetStock	24443.91	1	1853.83	3.11e-85
CaStock	4.58	1	0.35	0.56
Enzyme:EnzymeConc	661.13	4	12.54	8.44e-09
EnzymeConc:CaStock	65.16	1	4.94	0.03
Residuals	1938.28	147		

Table 2: ANCOVA made on the reduced transformed model.

It is expected that the reduced model should perform better since all relevant insignificant variables has been removed. To further investigate this The AIC from both the non-reduced and the reduced model is compared with k=2 - a lower AIC indicates a better fit on the data it was created from with a penalty for the number of parameters, to avoid overfitting.

	df	AIC
Not transformed	41.00	1983.57
Transformed Reduced	14.00	881.17

Table 3: AIC Comparison between non-transformed maximum model and the final reduced transformed model.

When comparing the AIC from the non-transformed and the transformed model a heavy decrease appears, indicating that the transformed model fits the data much better than the non-transformed (Table 3).

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
Not transformed	104.00	1183.49				
Transformed Reduced	147.00	1938.33	-43.00	-754.84	1.54	0.04

Table 4: Anova Comparison between non-transformed maximum model and the final reduced transformed model.

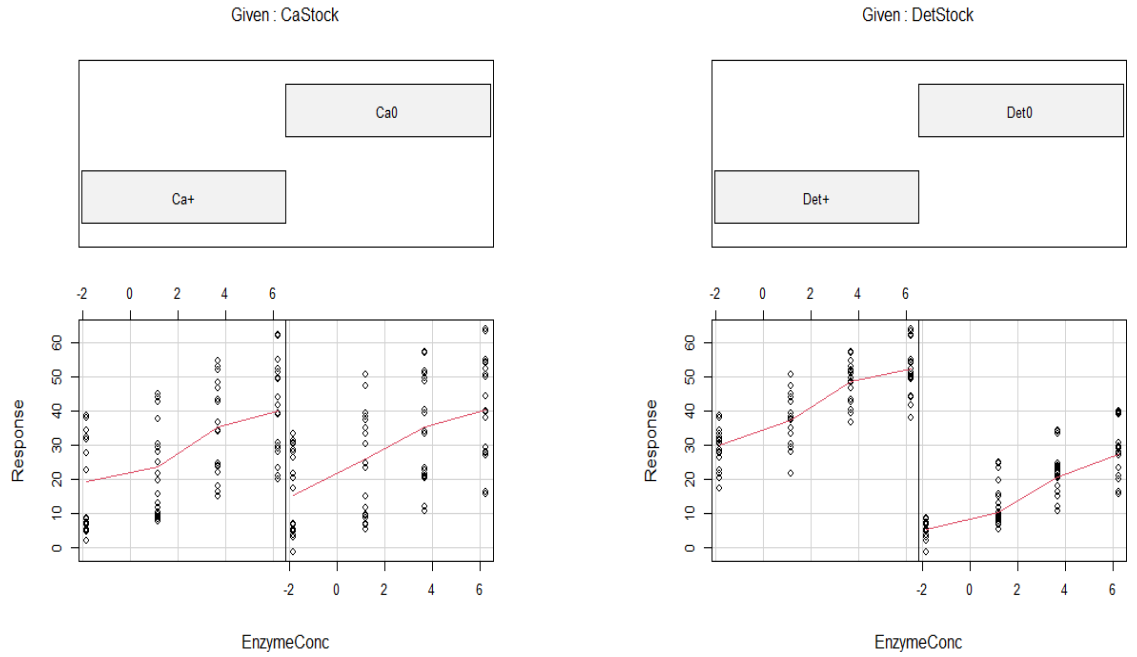
To investigate whether there is a significant difference between the non-transformed maximum model and

the final reduced transformed model an ANOVA (analysis of variance) is performed (Table 4). The models are significantly different. As a result of the models being significantly different, it is important to actually choose the better one of them. Since the transformed and reduced model appears to have a better fit of data based on the AIC values that model is chosen as the optimal one.

4 Results

4.1 Co-plots

Based on the model we have found a significant relationship between Detergent use and enzyme concentration and between Calcium use and enzyme concentration. To investigate these interactions different coplots have been created.



(a) Coplot of Response aginast Enzyme cncentration with hardness as a factor

(b) Coplot of Reponse against enzyme concentration with detergent as a factor

Figure 8

Based on the coplots of hardness and detergent against enzyme concentration and response, is a significant relationship found. The coplot for hardsness (figure 8a) shows a change in slope which indicates that there is an interaction between hardness and enzyme concentration. This interaction is also found significant in table 4 with a p-value=0.03. Based on figure 8b we can observe a change in intercept and slope when running the experiments with and without detergents. This relationship is significant based on our model.

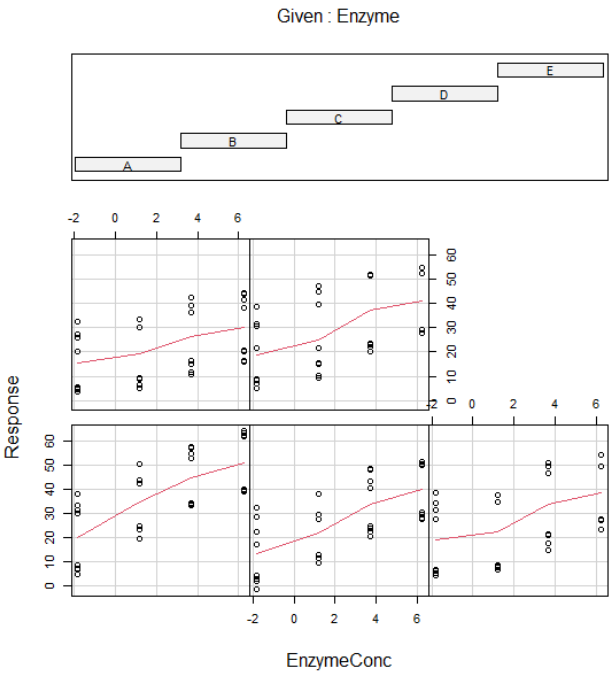


Figure 9: Coplot of Response agianst enzyme concentration with Enzyme as factor

The coplot of response agianst enzyme concentration shows an indication of an interaction between enzyme and enzyme concentration (Figure 9). This is illustrated as a change in slope depending of the enzyme. When comparing with the final ANOVA model (Table 2) the potential interaction is supported by the corresponding P-value being significant at a significance level at $\alpha = 5\%$.

4.2 Backtransformation of model

All variables are back-transformed to enable better and more correct interpretation of the data at the original scale (Table 6). The data is back-transformed using the power of two function. This is valid based on the BoxCox analysis with a maximum λ close to 0.5 in each case. For all back-transformed variables are the confidence interval (CI) calculated as well with a significance level at $\alpha = 5\%$, meaning 2.5% at each end of the data. CI is shown for both the transformed and the back-transformed data (Table 5 og 6).

	2.5 %	97.5 %
(Intercept)	39.82	43.52
EnzymeB	-11.35	-6.80
EnzymeC	-8.37	-3.82
EnzymeD	-12.49	-7.94
EnzymeE	-7.09	-2.54
EnzymeConc	3.20	4.13
DetStockDet0	-25.86	-23.59
CaStockCa0	-2.77	0.11
EnzymeB:EnzymeConc	-1.09	0.11
EnzymeC:EnzymeConc	-1.93	-0.73
EnzymeD:EnzymeConc	-2.59	-1.38
EnzymeE:EnzymeConc	-1.55	-0.34
EnzymeConc:CaStockCa0	0.05	0.81

Table 5: Transformed variables with corresponding confidence interval at a significance level at 5%.

Names	2.5%	97.5%
(Intercept)	1585.59	3587179.51
EnzymeB	128.84	2142.73
EnzymeC	70.08	213.93
EnzymeD	155.95	3976.58
EnzymeE	50.20	41.50
EnzymeConc	10.25	292.06
DetStockDet0	668.48	309455.87
CaStockCa0	7.67	0.00
EnzymeB:EnzymeConc	1.20	0.00
EnzymeC:EnzymeConc	3.73	0.28
EnzymeD:EnzymeConc	6.68	3.65
EnzymeE:EnzymeConc	2.39	0.01
EnzymeConc:CaStockCa0	0.002	0.43

Table 6: Back-transformed variables with corresponding confidence interval at a significance level at 5%.

The back-transformed model is illustrated with corresponding confidence and prediction intervals (Figure 10). When focusing on the y-axis it is clear, that the addition of detergent increases the y-value very much compared to the data with no added detergent - hence, an increase in catalytic response. The plot of data containing both detergent and Calcium (hardness) obtain a lower enzyme concentration resulting in a increase in catalytic response compared to the data containing detergent but no Calcium (hardness) (Figure 10). When comparing the plots of data both not containing detergent no big difference in catalytic activity is observed. The addition/removal of calcium seems to increase the enzyme concentration for some enzymes while decreasing it for others when no detergent is added. However, calcium does calcium seem to have a bigger influence when detergent is added, which maybe could be caused by the intercept being higher.

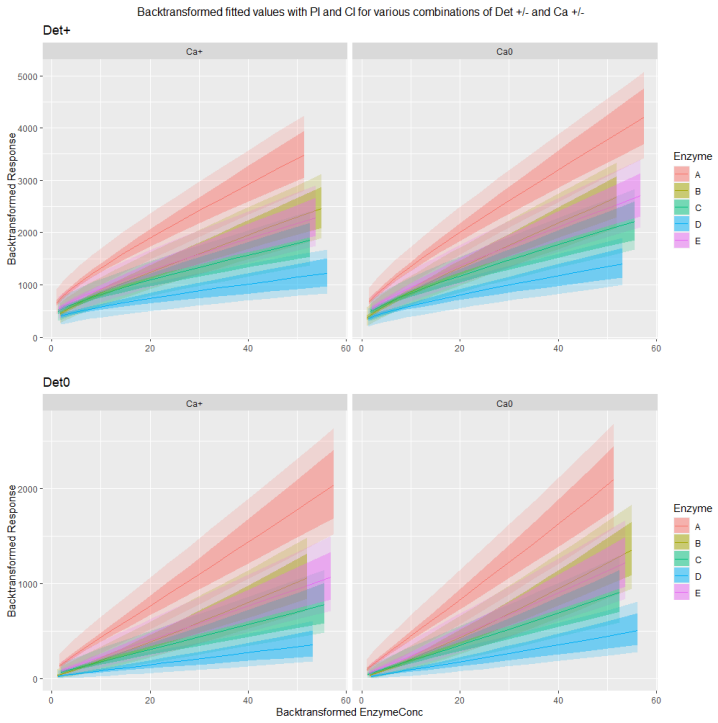


Figure 10: The backtransformed model with the confidence intervals

5 Conclusion

In conclusion, our study aimed to investigate the effects of six different factors on the enzymatic activity of five different enzymes. Through creating a general linear model, we found the effect of detergent on the intercept to be significant at a significant level of 5%. This could be due to its own catalytic properties. However, the level of detergent did not have an effect on the slope of the model, and therefore did not change the efficiency of the enzymes.

Additionally, we found that the level of calcium had a small effect on the slope, with addition of calcium lowering the efficiency of the enzymes. Even though the level of effect was found to be small, it was still significant based on the analysis with a $p - value = 0.03$. Along with this it was found that it had no significant effect on the intercept, which is in line with the fact that there is no enzymatic activity without the enzyme. Therefore, based on our findings, we can conclude that calcium, in contrast to detergent, does not possess catalytic properties of its own. Instead, it acts as an inhibitor of the enzyme, reducing its efficiency.

Enzyme concentration was found to have a positive effect on response, with higher concentrations resulting in higher response. The effect of enzyme concentration on response was found to be very significant. However, enzyme concentration did not have any effect on the intercept of the model, as expected since there is no enzymatic activity without the enzyme.

It was also found that the efficiency of the enzymes increased differently at different concentrations, with enzyme A having the highest efficiency and enzyme D having the lowest. The efficiencies of the enzymes were equally affected by calcium, but not by detergent.

Finally, it is noted that there may be systematic errors present in the study, such as differences in results due to testing on different days. The only potential systematic error which could be investigated would be caused by cycle. However, this variable did not show significant influence on the final model. The design of experiment should have been different for a final conclusion on systematic errors.

References

[1] L. E. Christiansen and A. Baum, "02441: Transformations and more," pp. 1–7, 12 2019.

[2] A. Baum, "02441: Multicolinearity," pp. 1–6, 12 2019.

Apendix

- R-script; Case1pdf.pdf