Case 1
Effect of hardness and detergent on enzymatic catalysis

Frederik Boe Hüttel, s151689
Helle Rus Povlsen, s134885
Christian Glissov, s146996

January 10, 2018

DTU Compute - Institute for Mathematics and Computer science
Applied statistics and statistical software 02441 Jan 2018
Lasse Engbo Christiansen

Technical University of Denmark

Summary

This paper will present an analysis to describe and make inference about the enzymatic performance in the presence and absence of water hardness and detergent. Data from an existing study will be used. A brief summary of the data will be made and relevant descriptive statistics will be explained. Statistical tools and models will be used to find the enzyme which is most effective at removing stains. Enzyme A is the best enzyme to use and enzyme D is the worst based on the overall findings. Incorporating detergents and increasing the amount of enzyme concentration will improve the removal of stains. For future studies it would be relevant to include influence of performing several cycles in the same day. Furthermore the influence of performing the experiments on different days should also be taken into account to avoid systematic errors. More levels of enzyme concentration should also be included.

Contents

Introduction

Enzymes have long been known to improve detergents. Employing enzymes in combination with detergents improves the efficiency of the detergent. The enzyme contributes by degrading the substance of a strain and thus increases the surface availability for the interaction with detergent. The performance of enzymes may be affected by conditions, such as concentration of Calcium-ions. Certain ions function as co-factors that enhance enzyme activity and therefore the hardness of water may influence the performance of a particular enzyme. Finally, it is interesting to investigate how the concentration of an enzyme influence the performance.
The objective of testing enzyme performance in different conditions is to determine the optimal enzyme for every relevant condition. These conditions may vary according to geographical region, such as hardness of water, or according to market price since higher concentrations of enzymes are more costly.
Assessment of the enzyme performance is evaluated by the amount of protein removed from a surface, using the Surface Plasmon Resonance technology (SPR)[1]. A control experiment is conducted in which a reference enzyme with no degradative function for typical strain substances is included. This ensures that the amount of protein in form of enzyme is constant in the SPR assay. In other words, the reference enzyme holds the assumption of "one enzyme per experiment" true. Hence the null-performance can be used as a reference to the performance of the active enzymes. The data for the reference enzyme is not included in our study.

Description of the data set

The data set contains data collected during a 10 day period. The data consist of experiments of different combinations of Enzymes, the concentration of the enzymes, Detergent and Hardness of the water. The aim of the experiments is to estimate the response of the enzymes under certain combinations.
The data set has 5 different Enzymes that has been tested. Each enzyme was tested with 16 different combinations and the combinations was repeated twice. Resulting in 32 observations for each enzyme resulting in 160 observations in total. The data set contains 7 attributes. 2 continuous variable and 5 categorical variables. The attribute *Cycle* will be regarded as a continuous variable, because *Cycle* has 34 discrete levels, which in essence means it can be regarded as a continuous variable. Although the cycle has 34 levels only 32 experiments where conducted. The two remaining cycles has been used to test a reference enzyme. The data set does not contain the conditions or results for these two cycles. This results in a maximum observed value of 34 cycles but the count of the cycles are 32 within each enzyme (see table 2).
The attribute Response is the amount of removed protein given the combination of the other factors and is, therefore, our response variable as the attribute name indicates. The signs "+" and "0" in the attributes *DetStock* and *CaStock* indicate the presence of detergent and whether the water was hard or not. There is also a relationship bewtween the RunDate and the Enzyme attributes. Only 1 enzyme

[1] Case_det.pdf

	Factor	Continuous	Continuous	Factor	Factor	Factor	Factor
Id	RunDate	Cycle	Response	Enzyme	EnzymeConc	DetStock	CaStock
1	03/12/08	1	323	B	2.5	Det+	Ca+
2	03/12/08	2	614	B	7.5	Det+	Ca0
3	03/12/08	3	326	B	15	Det0	Ca+
4	03/12/08	4	162	B	7.5	Det0	Ca0
5	03/12/08	5	545	B	2.5	Det+	Ca0
6	03/12/08	6	214	B	7.5	Det0	Ca+

Table 1: Head of dataset, with the first 6 observations of the dataset.

was run each day, resulting that the date attribute and the enzyme attribute can be interchanged, as all observations with enzyme B will have the same date and so on.

RunDate	Cycle	Response	Enzyme	EnzymeConc	DetStock	CaStock
25/11/08: 32	Min: 1.00	Min: 0.1	A: 32	0:40	Det+:80	Ca+: 80
27/11/08: 32	1st Qu. 9.00	1st Qu:94.6	B: 32	2.5:40	Det0:80	Ca0: 80
03/12/08: 32	Median: 17.50	Median: 322.4	C:32	7.5:40		
05/12/08: 32	Mean: 17.38	Mean: 431.6	D:32	15:40		
08/12/08: 32	3rd Qu.:25.25	3rd Qu.: 662.7	E:32			
	Max: 34	Max: 1588				

Table 2: Summary of the 7 attributes in the dataset.

From Table (2) it can be seen that that there is an equal distribution of all the enzyme concentrations, with or without detergent and with or without Hard water. It can be seen that there is 80 observations with detergent and 80 without. The same can be seen for the hard water attribute. There is also an equal distribution between the enzymes, so all the enzymes has been run the same amount of times. In figure (1) the pairs plot of the 4 attributes Enzyme, EnzymeConc, CaStock and DetStock can be seen. Since all 6 plots contains points in each corner and as well as point at all the intersections, because it's plotting discrete values. This suggest that all the combinations of the the 4 variables has been tested.
As the response variable will be reponse variable in our statistical model a box plot of the enzymes and their response can be seen in Figure (2, left). This plot it suggest that their could be a difference in the mean values of the enzymes, however this plot does not take into account the different levels of concentration, which could have impact on the response attribute. Especially since the concentration of 0 is measured for all enzymes, this is assumed to be the same. So the plot (Figure (2, right)) looks at the enzymes and their response at different concentration levels. This plot also suggest a difference at the different concentration levels.However at the Enzyme concentration 0 there appears to be some similarities between the means, which was assumed. As well as (Figure (2, left)) it appears there is a difference in the response at higher concentrations. It also appears that some enzymes doesn't benefit from a higher concentration for example Enzyme E, which doesn't appear to change much from a concentration of 7.5 to 15.
For the other categorical variables DetStock and CaStock, the difference in the response mean is also looked at. from these blot it appears there is a significant

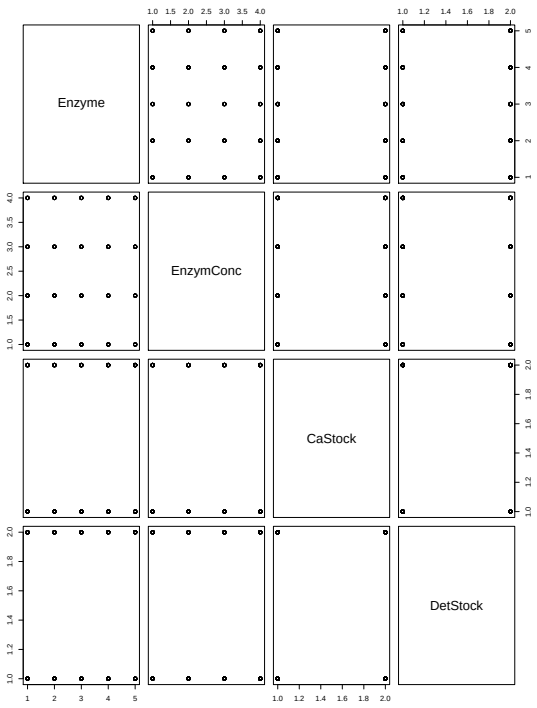

Figure 1: Figure showing the pairs plot of the 4 attributes, Enzyme, EnzymeConc, CaStock and DetStock

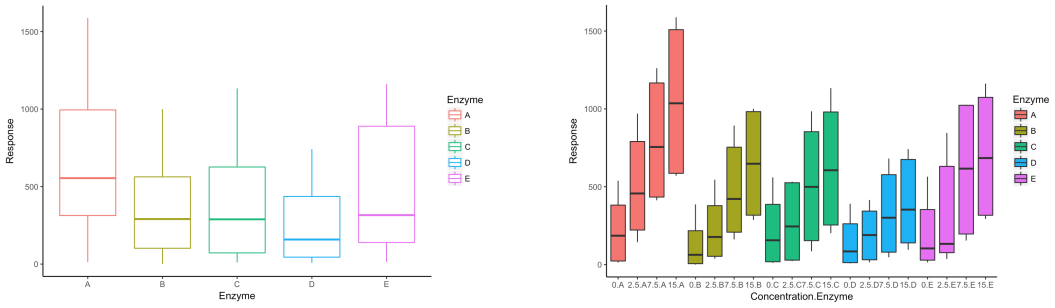

Figure 2: Boxplot of response based on enzyme and concentration

difference in the response based on whether or not detergent is present an observation. It also appears that there is no significance based on the hardness of water.

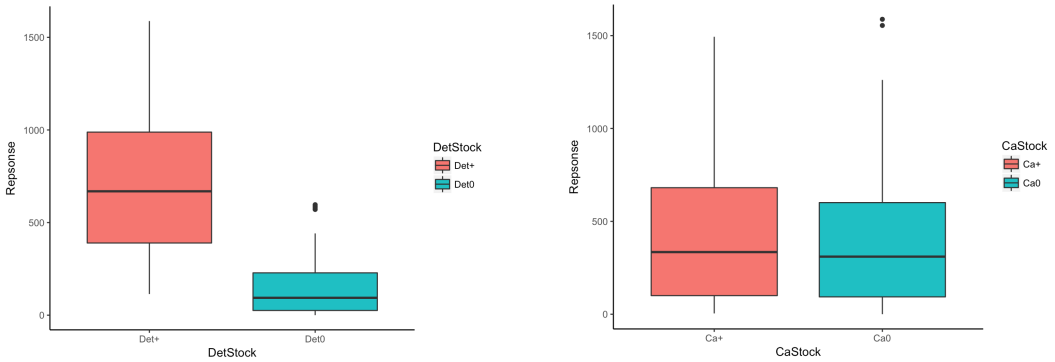

Figure 3: Boxplot of Response based on DetStock and CaStock

To see if there are any interactions and to further facilitate the assumption that interactions should be included into the statistical analysis an interesting thing to look at is an interaction plot. To keep things fairly short and simple only the interaction between enzyme concentration and the enzyme looked at:

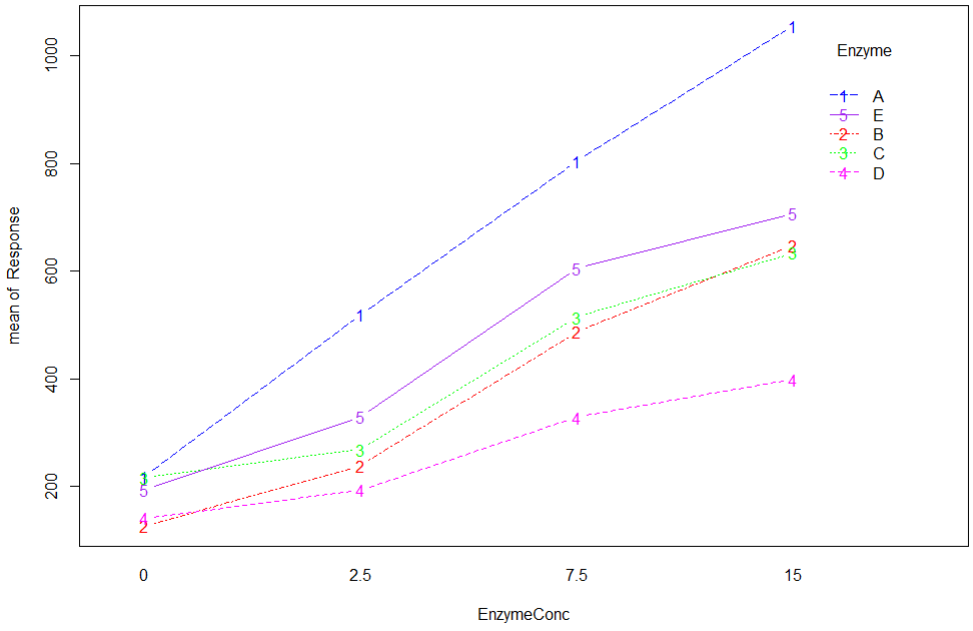

Figure 4: Interaction between enzyme and enzyme concentration

It can be seen that since the lines are not parallel and crosses each other it would be relevant to incorporate interactions to the model, which also makes intuitively sense.

Statistical analysis

As explained previously, the data consists of 2 continuous variables and 5 categorical variables of which the response variable is categorical. The aim is to model the

performance of enzyme with the explanatory variables being enzyme concentration and presence of detergent and calcium ions. The remaining variables RunDate and Cycle are not expected to influence the performance of the enzymes, but are merely considered to describe the protocol of the experiment. However, to account for systemic errors these two variables must be accounted for, eg. by including them in the initial model.
Since the explanatory variables are categorical the experiment is of a factorial design. Using analysis of variance the interactions between variables are evaluated by testing "whether the response to one factor level depends on the level of another factor".Depending on the number of factors in the final model an n-factor-way ANOVA is carried out. In an analysis of variance the individual treatment means are fitted on varying conditions and the relevant departures from the fitted mean form the error sum of squares. When the individual treatment means are significantly different, the respective residuals will decrease compared to the overall mean. Therefore the error sum of squares decrease as more variance is explained by the explanatory variables.
In order to determine which factors that may contribute to reduce the unexplained variance, a linear model can be formulated. In a one-factor design with two levels the resulting model would have the following equation:

$$y = a + bx_1 + cx_2 + \mathrm{e}, \qquad \mathrm{e} \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

where x_1 and x_2 are the levels of the factor called x.
With two or more explanatory variables the levels of the factors may interact. To investigate possible interactions two-way factor plots will be applied. However, with several variables, multiple-way interactions are possible. Often the modeller will limit the initial model to consist of two-way interactions, since more interactions make the evaluation of the model very complex.
Exceptions to this rule of thumb arise when physical interaction between multiple factors is reasonable. In this case a sound hypothesis could be that calcium ions affect the enzymes which in combination with detergent will remove stains more efficiently. This yields a three-way interaction. In addition, the concentration of enzyme may influence the performance of the tested enzymes differently, resulting in a four-way interaction. The risk of including high degrees of interactions is over-fitting the model. Therefore it is essential to remain critical of types and numbers of interactions when fitting the initial model. The approach is then to start with the most complex, however, reasonable model and reduce it by removing non-significant interactions one at a time until arriving at the least adequate model. Single levels cannot be eliminated from the model as long as the level exists in an interaction significant for the response factor. This process has been automated in R by the given function `stepP`.

A model must satisfy the assumptions of normality of residuals, variance homogeneity and residuals being independent and identically distributed. The assumptions are checked by visually inspecting graphs of residuals vs fitted values, the qq-plot, the scale-location plot and the residuals vs. leverage plot. If a model does not satisfy the assumptions the response variable can be transformed by a power transformation. The optimal power is deduced from the λ-value that maximizes

the log-likelihood function of a Box-Cox transformation to normality. Once the model has been fully reduced, the validity of the power transformation must be confirmed. If the confidence interval of the optimal λ-value covers the value one, the initial transformation is still adequate for the final model.

An approach to further reduce a model can be based on whether effect sizes of two parameters are not significantly different. To check whether there is a difference between two levels of a factor, an estimated t-test can be done on the effect size and the standard error of the parameter from the statistical model. If the difference between the effect size is within 2 times the standard error there is not a significant difference between the two levels. In such a case the number of levels in the corresponding factor may be reduced by collapsing the insignificant levels.

The final step of model fitting is the assessment of outliers. Again visual inspection of the assumption plots may reveal accentuated observations that appear extreme. In order to establish if an observation is an outlier the response for that specific set of condition combinations must be compared to responses of equal or highly similar conditions.

The equation of a fitted model differs from eq. (1). According to the treatment contrast convention, the model Intercept covers a reference mean based on the first level of each factor. The remaining parameters estimated in the model are denoted as the difference between the reference mean and the relevant factor level mean coined effect size.

Results

As described in section the data was initially fitted by a multi-way factor design due to biochemical meaningful combinations of factors. In section , two-way interactions were substantiated with interaction plots as in figure 4. This means that the model may contain two-way interactions between the different factor levels. However, the relevant interactions can only be evaluated from the significance of the effect sizes in the final model.
Before the test the model must satisfy the assumptions of normality, variance homogeneity and residuals being IID. From the residuals vs fitted values plot we notice that the residuals (figure 10) in the model seems to be clumped together at the low values of the fitted values. There also seems to be an upwards trend at the start of the scale-location plot. This indicates that a transformation is needed. By optimising the power coefficient λ, it was estimated that a cubic transformation $\lambda = 0.3$ was a good choice. This power transformation will scale the response variable and reduce the variance of large values and increase the variance for small values. After transforming the response variable a new model is created. A quick check of the residuals confirms that the residuals have improved (figure 11). The initial complex model is reduced by removing all non-significant explanatory factor interactions and variables. Reducing the function made the model significantly less complex, due to removing non-significant coefficients.

In assessing for outliers it is also observed from figure 11 that observation 147 seems to be an extreme observation. In order to compare the observation with similar data points a subset of the data was extracted based on the conditions of the observation 147. Values for enzyme E at 2.5nM concentration, tested with detergent in combination with hard and soft water, since the calcium ion did not appear to influence performance (fig. 3). Looking into the data from table 3 it is clear that observation 147 is an outlier, it is a lot lower than the other 4 responses for experiments with detergent for enzyme E:

Table 3

RunDate	Cycle	Response	Enzyme	EnzymeConc	DetStock	CaStock
081127	12	770	E	2.5	Det+	Ca+
081127	16	585	E	2.5	Det+	Ca0
081127	20	174	E	2.5	Det+	Ca+
081127	32	845	E	2.5	Det+	Ca0

A number of reasons might explain why observation 147 is so low compared to the other values, however lack of information restricts us to look further into this case, but something seems to have gone wrong. Based on the above table, observation 147 is removed from the data. Since we have removed an observation we need to fit a new model. All of the previous steps are repeated and once again from the residual vs fitted values plots we see that we need to transform the data once again. Since observation 147 didn't have an extreme leverage and didn't change the residual vs fitted values plot much, we use the same cubic transformation (figure 5).

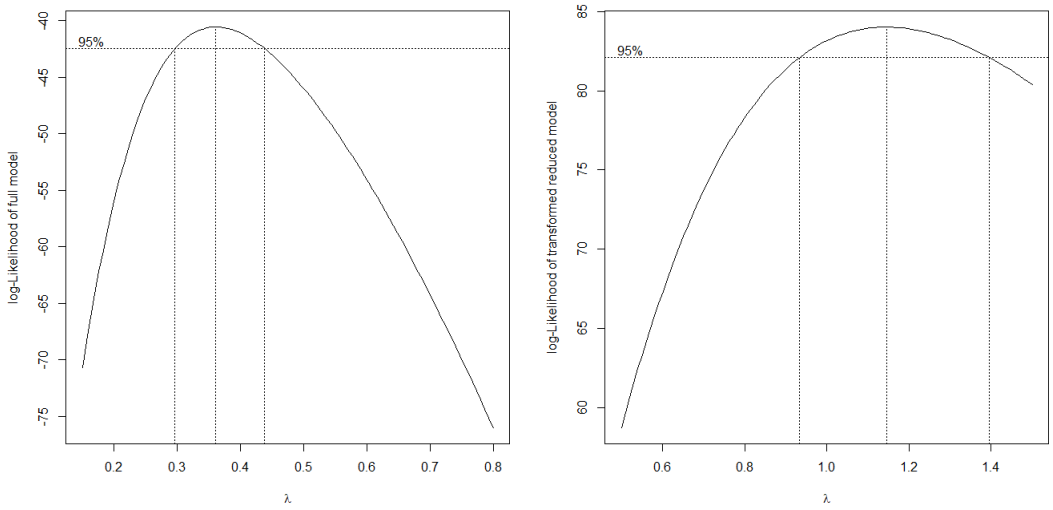

Figure 5: Result of Boxcox, notice the plot to the right. 1 is in the interval of λ, this means no further transformation of the data is needed.

Finally the model is reduced by removing non-significant coefficients to minimize the model complexity. To accept the final model the assumptions needs to be checked. Once again a residual vs fitted plot is made:

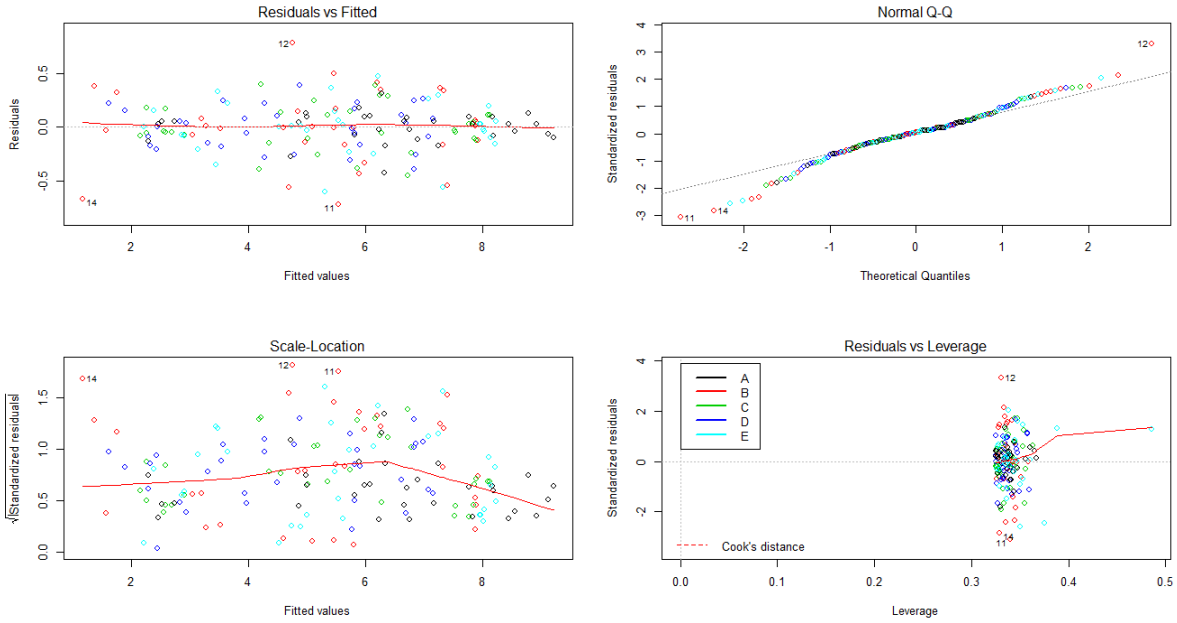

Figure 6: Top left: Residuals seems to be IID. Bottom left: Variance homogeneity seems to be satisfied. Top right: Normality seems to be satisfied. Bottom right: No significant outliers or values with extreme leverages. To see each enzyme class, notice the legend in the bottom left plot.

The last thing to check is the residual plot of each independent variable

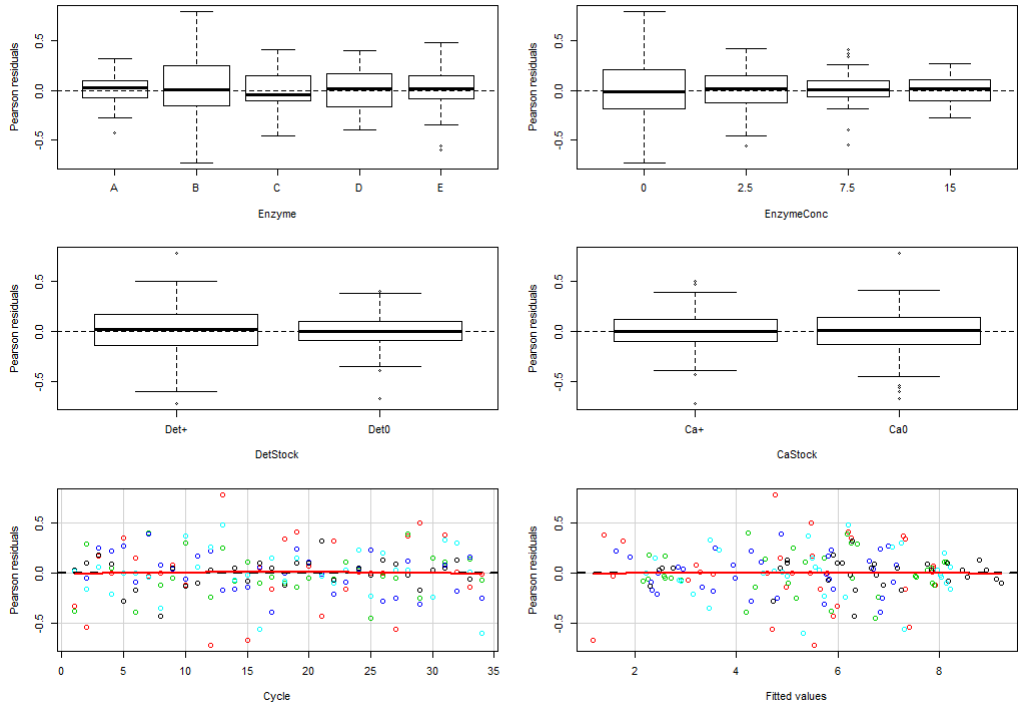

Figure 7: Residuals for each variable. Colors refer to enzymes, using the same color scheme as previously. Enzyme B = red.

Notice that the residuals for enzyme B (red) seems to have larger residuals and if looking at figure (6) one can see that enzyme B seems to have observations lying far away from the other residuals, this might be a systematic error of the experiment, which will be further discussed later.
From the plots the assumptions hold and the final model is valid. Because of the size of the model it can be seen in the appendix (13) along with all the effects of the chosen factors and interactions. From the reduced model it can be seen that the interaction between the continuous variable, cycle, and no detergent is significant. This seems to indicate yet another systematic error. The estimate of the effect size is positive if detergent is used, therefore the response would increase the more cycles being run. Looking at the coefficients in appendix from figure (13) it can be seen that when detergent is not added and hard water is not being used (both highly significant, $p_{det0} \approx 0 < 0.05$ and $p_{Ca0} \approx 0.0052 < 0.05$) it will affect the response negatively due to the negative estimates, with detergent being significantly lower than calcium (resp. -4.018 vs -0.453) and also because of all the interactions. Thus adding detergent will greatly increase the response and water containing calcium will slightly increase the response. Looking at the data from the ANOVA of the model in figure (4) it can be seen that enzymes are highly significant ($p_{enz} \approx 0 < 0.05$) and we accept the alternative hypothesis that not all the means are equal, this means that there is a relationship between which enzyme is being used and the catalytic activity.

Table 4: ANOVA of the final model

Coefficient	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
Enzyme	4	51.59	12.9	154.7008	<2.2e-16
EnzymeConc	3	191.10	63.70	764.1053	<2.2e-16
DetStock	1	364.12	364.12	4367.7246	<2.2e-16
CaStock	1	0.48	0.48	5.7420	0.018333
Cycle	1	0.29	0.29	3.5121	0.063703
Enzyme:EnzymeConc	12	11.23	0.94	11.2251	4.001e-14
Enzyme:DetStock	4	2.96	0.74	8.8889	3.200e-06
Enzyme:CaStock	4	1.27	0.32	3.8078	0.006240
EnzymeConc:DetStock	3	5.02	1.67	20.0619	2.356e-10
EnzymeConc:CaStock	3	2.64	0.88	10.5566	3.987e-06
DetStock:CaStock	1	0.20	0.20	2.3579	0.127657
DetStock:Cycle	1	0.90	0.90	10.8204	0.001367
Enzyme:EnzymeConc:DetStock	12	2.53	0.21	2.5243	0.005835
EnzymeConc:DetStock:CaStock	3	0.79	0.26	3.1732	0.027296
Residuals	105	8.75	0.08		

The same goes for enzyme concentration, which is also highly significant. In addition the interactions between which enzyme is being chosen and other factors such as detergent used and the concentration of the enzyme chosen etc. are significant. To check whether there is a difference within each enzyme, an estimated t-test can be done on the effect size and the standard error from the statistical model (13). If the difference between the estimate is within 2 times the std. error there

is not a significant difference between the two coefficients. it can also be seen that the standard error is roughly 0.22, times two is 0.44, and all the differences in estimate is bigger than this, resulting in a significant difference between the different concentration of the enzymes. This is also indicated from 2, where the response is growing with the concentration. To see this effect more clearly it would be interesting to look at how the prediction is being affected given a new data set of specified values for each of the factors. Making a plot of the back transformed predictions we get the following:

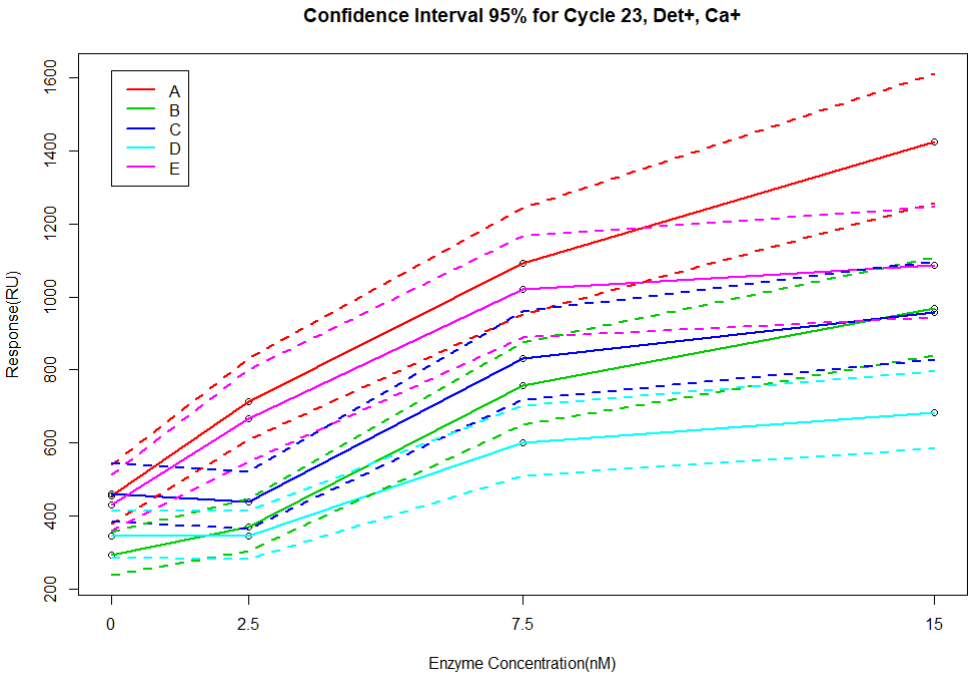

Figure 8: Prediction plot, the legend shows the color of the different enzymes used, the x-axis is the enzyme concentration used and the y-axis is the catalytic activity for water with detergent and calcium. The stippled lines cover the confidence intervals of the respective enzymes.

It can be seen that enzyme A, the reference enzyme, is quite a bit better. For enzyme concentration at 15 nM enzyme A even seems to be significantly better than the other enzymes due to no overlapping confidence intervals. The prediction is based on a cycle 23 with detergent and calcium. It can clearly be seen that the more concentration of the enzyme is added the greater is the catalytic activity, notice also that even though no enzyme was used we still have quite a high catalytic response, this is mainly due to the detergent being used.

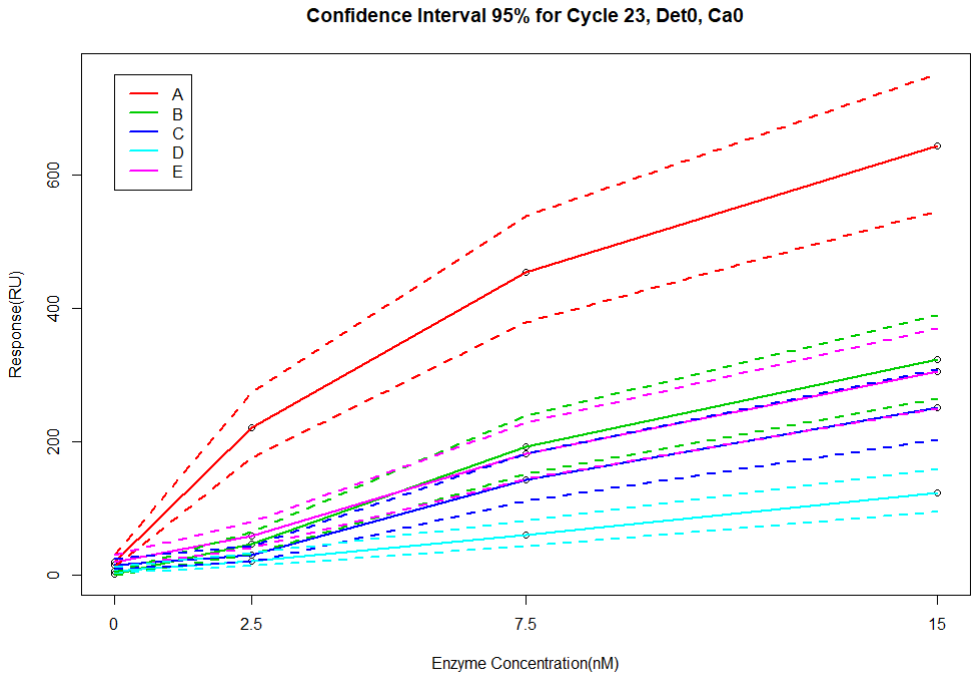

Figure 9: Prediction plot, the legend shows the color of the different enzymes used, the x-axis is the enzyme concentration used and the y-axis is the catalytic activity for water without detergent and calcium. The stippled lines cover the confidence intervals of the respective enzymes.

Now, looking at the prediction without calcium and detergent, but still for cycle 23, we see that the response is a lot lower. This is also what would be expected from the negative estimates when no detergent and calcium is used. Enzyme A seems to be significantly better in all cases which also seemed to be the case based on the box plots in figure (2). The reason figure (8) has a higher catalytic activity than figure (9) is because if adding detergent and calcium most of the negative estimates of the interactions and non-significant cofficients, seen in figure (13), will disappear.
Another interesting thing is looking at the 0 enzyme concentration it can be noticed that there is a difference between each response value. This was not expected, due to the assumption that the same amount of detergent was used with the hard water and therefore an enzyme concentration of 0 should not change the response values. This might indicate yet another systematic error and might indicate a difference between certain factors used for each day.
The catalytic activity seems to level off at some point. However due to not enough levels of enzyme concentration a conclusion based on when the extra amount of added enzyme will be negligible beneficial to the catalytic activity can't be given.

Discussion

From the experiments on the effect of hardness and detergent on enzymatic catalysis, the final linear model allows interpretation on the effect size of included factors and interactions. Moreover, the final model allows for a discussion of which variables were included in contrast to what was expected. We investigated the response to five explanatory variables: type of enzyme, enzyme concentration, use of detergent, presence of calcium ions, and cycle. As mentioned in section cycle was only included to account for systematic errors. Thus, the factor `runDate` was excluded since it correlated with the type of enzyme since each test day was dedicated to one enzyme.

Between the variables up to three-way interactions were accepted. Three-way interactions are seen between the type of enzyme, enzyme concentration and detergent as well as between enzyme concentration, use of detergent and presence of calcium ions.

The model reference has a rather large positive effect size and constitutes the following combination of factor levels: enzyme A, however at zero concentration, with the use of detergent in hard water. With increasing enzyme concentration, the A enzyme results in increased performance due to the positive additive effects. Also in combination with no detergent and no calcium the performance increases significantly, disregarding the three-way interaction of concentration, lack of detergent and lack of calcium.
The remaining enzymes primarily have negative effect sizes and thus yields a lower performance no matter what combination of conditions. However, when including the confidence intervals as in figure 8 it becomes clear that enzyme A is not significantly different in performance under every set of conditions.

One of the overall objectives of this analysis was to detect the effect of hardness and detergent on the enzyme performance. Based on the statistical model (fig. 13) it appears that detergent has a significant impact on the response. From the model it can be seen that when detergent is absent (DetStock == Det0) there is a significant negative impact on the response estimate resulting in a lower cleansing performance. In other words, there is a significant improvement in using detergent along with the enzymes.
In the interaction between varying concentrations and no detergent there is a significantly positive impact on the response estimate. However, for other enzymes than A the interactions have negative effect sizes resulting in a worse response. From the model it can be seen that the absence of Ca^{++} has a significant negative impact on the response estimate. The absence of Ca^{++} also has a negative impact through interactions with enzymes other than A. This supports the understanding of calcium as a co-factor that enhances the activity of enzymes. Without calcium the activity is not as great and therefore the performance of the respective enzyme is diminished.

Another of the overall objectives of this analysis was to detect whether the amount

of enzyme affects the performance. Indeed the overall trend of the prediction plots, no matter the combination of detergent and calcium, the performance increases with increasing enzyme concentration (fig. **??**. However, as mentioned in section the performance at enzyme concentraion 0nM varies according to enzyme, see fig 8. Since a concentration of 0 nM indicates no enzyme present this variation in performance is very odd, and may be an indication of a systemic error. Since the each enzyme has been run in different experiments of different days, this may be the cause of the deviation in performance. However, further tests are needed to confirm this.

Another systematical error may be found in the experiment of enzyme B. As already mentioned earlier the residuals for the enzyme B seems to be larger than the other groups based on the colors from figure 6 and from the residual plot in figure 7. Additionally, from the statistical model (13) it can be seen that there is a significant difference between the enzyme B and enzyme A with concentration 0 nM, which would be assumed to be non significant as just described in previous paragraph. Enzyme D also appears to be significantly different at concentration 0 nM, however the residuals do not vary the the same extend as enzyme B, see figure 7. But to return to the enzyme B it based on the residuals and the significant difference of B at the concentration 0 there appears that there could have been some systematically errors on that given day for the enzyme B.
Yet another systematical error could be the machine itself. From the statistical model (13) it appears that cycle does not have a significant impact on the response. However, in the interaction between cycle and lack of detergent there is a significant impact on the performance. Thus, when there is no detergent the cycle has a significant impact, this could be an indication that some sort of residue accumulates during the course of the experiment, which positively affects the performance. Such a residue might be calcium, which has a positive impact on the performance as discussed earlier.

The minimal adequate model that we arrived at could not be further reduced by collapsing levels of factors, since all levels were significantly different. This was assessed by comparing differences in effect sizes to the double standard error of the parameter.

One last attempt to derive the best model was to exclude the outlier from observation 147. We cannot evaluate the two models on the same scale. An ANOVA test or the AIC would have been appropriate if the models where nested. So to evaluate which model is superior we will look at figure (6) for the model without the outlier and Figure (12) for the model with the outlier. Looking at these two figures it appears that the residual is more normally distributed. The scale-locations plot are quite similar, however the scale of the scale-locations plots on the model without the outlier figure (6) is smaller meaning an overall lower residual. This results in the model without the outlier being superior. However, the users of the model must consider whether the experiment needs to be repeated or if they truely believe, that the observation is an error. Otherwise, we would recommend the full model.

Conclusion

Based on the analysis it can be concluded that enzyme A had an overall better catalytic activity compared to the other enzymes. Where D had the worst overall catalytic activity. Including detergent seems to significantly improve the removal capabilities of each enzyme. There was also a slightly better removal capability when washing in hard water rather than soft water. Future studies should include the influence of running several cycles after each other. It might also be relevant to include more than one experiment per day or randomise the enzymes used each day to avoid the differences in the enzyme concentration of 0. Finally more levels of concentration, detergent and calcium should be included. This would further facilitate the results.

Appendix

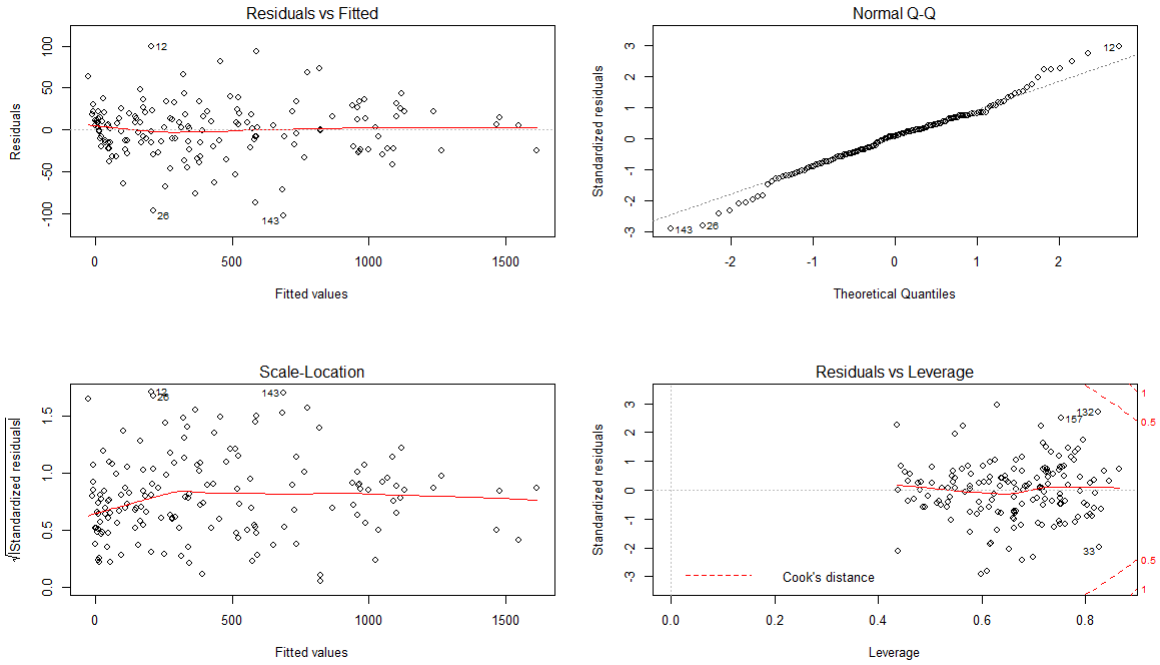

Figure 10: Plots to see if assumptions is fulfilled, full model

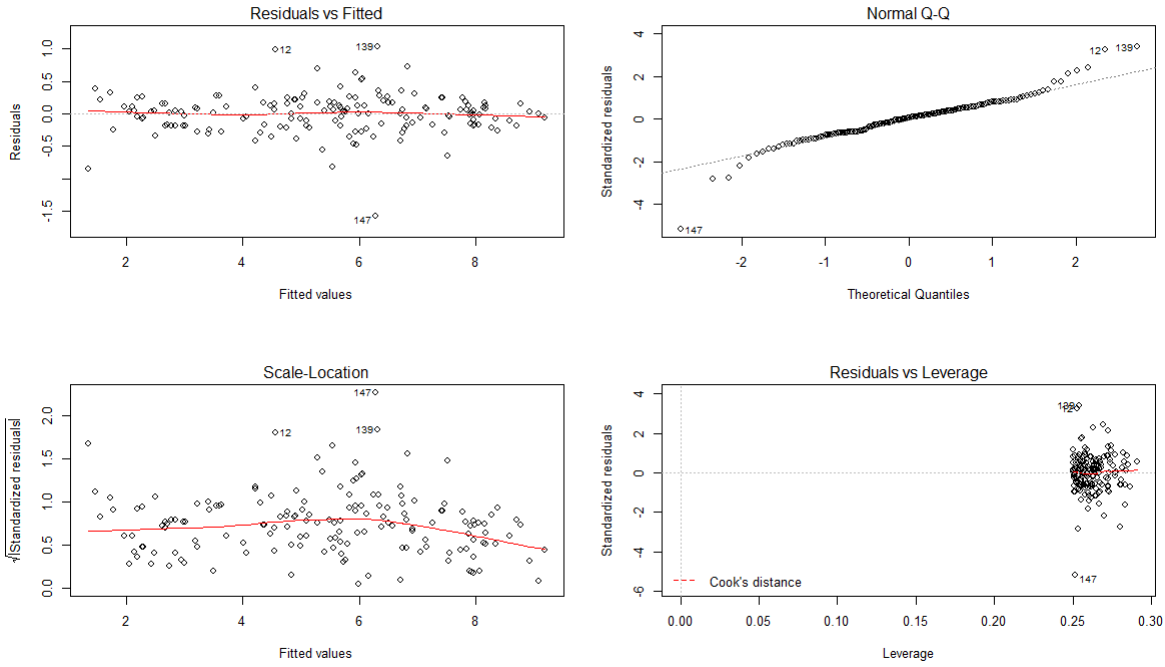

Figure 11: Plots to see if assumptions is fulfilled for the cubic transformed and reduced model

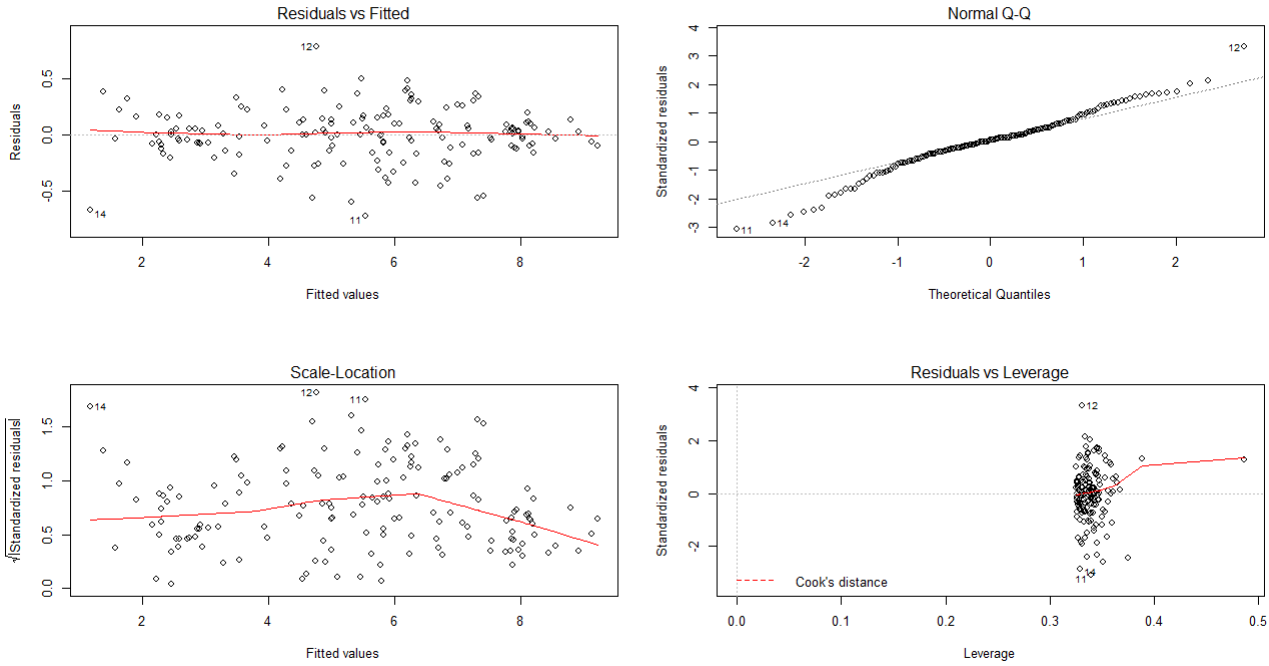

Figure 12: Plots to see if assumptions is fulfilled for the cubic transformed and reduced model, outlier removed

```
Call:
lm(formula = Response^(0.3) ~ Enzyme + EnzymeConc + DetStock +
    CaStock + Cycle + Enzyme:EnzymeConc + Enzyme:DetStock + Enzyme:CaStock +
    EnzymeConc:DetStock + EnzymeConc:CaStock + DetStock:CaStock +
    DetStock:Cycle + Enzyme:EnzymeConc:DetStock + EnzymeConc:DetStock:CaStock,
    data = SPRwithout147)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72235 -0.10978  0.01134  0.13004  0.78311

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               6.368758   0.171854  37.059  < 2e-16 ***
EnzymeB                                  -0.773409   0.218722  -3.536 0.000606 ***
EnzymeC                                   0.025100   0.216841   0.116 0.908070
EnzymeD                                  -0.500795   0.217477  -2.303 0.023264 *
EnzymeE                                  -0.097728   0.219149  -0.446 0.656556
EnzymeConc2.5                             0.911829   0.224406   4.063 9.36e-05 ***
EnzymeConc7.5                             1.885547   0.224826   8.387 2.47e-13 ***
EnzymeConc15                              2.567170   0.224067  11.457  < 2e-16 ***
DetStockDet0                             -4.018187   0.238211 -16.868  < 2e-16 ***
CaStockCa0                               -0.452999   0.158677  -2.855 0.005189 **
Cycle                                    -0.004326   0.003646  -1.187 0.238050
EnzymeB:EnzymeConc2.5                    -0.510208   0.292304  -1.745 0.083828 .
EnzymeC:EnzymeConc2.5                    -1.004439   0.288730  -3.479 0.000734 ***
EnzymeD:EnzymeConc2.5                    -0.912311   0.289249  -3.154 0.002100 **
EnzymeE:EnzymeConc2.5                    -0.048414   0.302644  -0.160 0.873212
EnzymeB:EnzymeConc7.5                    -0.074284   0.290388  -0.256 0.798598
EnzymeC:EnzymeConc7.5                    -0.661980   0.288937  -2.291 0.023956 *
EnzymeD:EnzymeConc7.5                    -0.836780   0.289249  -2.893 0.004641 **
EnzymeE:EnzymeConc7.5                    -0.064246   0.290292  -0.221 0.825277
EnzymeB:EnzymeConc15                     -0.197720   0.290388  -0.681 0.497445
EnzymeC:EnzymeConc15                     -1.022306   0.289249  -3.534 0.000609 ***
EnzymeD:EnzymeConc15                     -1.246765   0.289777  -4.303 3.80e-05 ***
EnzymeE:EnzymeConc15                     -0.592780   0.293366  -2.021 0.045865 *
EnzymeB:DetStockDet0                     -0.096517   0.289787  -0.333 0.739752
EnzymeC:DetStockDet0                     -0.143580   0.288995  -0.497 0.620351
EnzymeD:DetStockDet0                      0.309762   0.289311   1.071 0.286764
EnzymeE:DetStockDet0                      0.431023   0.290837   1.482 0.141332
EnzymeB:CaStockCa0                       -0.319637   0.145630  -2.195 0.030376 *
EnzymeC:CaStockCa0                       -0.072635   0.145444  -0.499 0.618539
EnzymeD:CaStockCa0                       -0.503248   0.144925  -3.472 0.000750 ***
EnzymeE:CaStockCa0                       -0.356090   0.146793  -2.426 0.016980 *
EnzymeConc2.5:DetStockDet0                1.398603   0.317106   4.411 2.50e-05 ***
EnzymeConc7.5:DetStockDet0                1.642756   0.317181   5.179 1.08e-06 ***
EnzymeConc15:DetStockDet0                 1.526870   0.316692   4.821 4.84e-06 ***
EnzymeConc2.5:CaStockCa0                  1.066894   0.186674   5.715 1.03e-07 ***
EnzymeConc7.5:CaStockCa0                  0.782264   0.183080   4.273 4.26e-05 ***
EnzymeConc15:CaStockCa0                   0.787382   0.183190   4.298 3.86e-05 ***
DetStockDet0:CaStockCa0                   0.272238   0.182704   1.490 0.139208
DetStockDet0:Cycle                        0.017193   0.005016   3.428 0.000870 ***
EnzymeB:EnzymeConc2.5:DetStockDet0       -0.184719   0.411811  -0.449 0.654678
EnzymeC:EnzymeConc2.5:DetStockDet0       -1.088673   0.409028  -2.662 0.008999 **
EnzymeD:EnzymeConc2.5:DetStockDet0       -0.920696   0.408700  -2.253 0.026355 *
EnzymeE:EnzymeConc2.5:DetStockDet0       -1.594220   0.419036  -3.804 0.000239 ***
EnzymeB:EnzymeConc7.5:DetStockDet0       -0.155872   0.409516  -0.381 0.704250
EnzymeC:EnzymeConc7.5:DetStockDet0       -0.982086   0.408480  -2.404 0.017958 *
EnzymeD:EnzymeConc7.5:DetStockDet0       -1.312081   0.408706  -3.210 0.001760 **
EnzymeE:EnzymeConc7.5:DetStockDet0       -1.410974   0.409462  -3.446 0.000819 ***
EnzymeB:EnzymeConc15:DetStockDet0         0.088285   0.409501   0.216 0.829725
EnzymeC:EnzymeConc15:DetStockDet0        -0.492124   0.409976  -1.200 0.232696
EnzymeD:EnzymeConc15:DetStockDet0        -0.768214   0.409247  -1.877 0.063274 .
EnzymeE:EnzymeConc15:DetStockDet0        -0.777030   0.412875  -1.882 0.062605 .
EnzymeConc2.5:DetStockDet0:CaStockCa0    -0.790969   0.261198  -3.028 0.003096 **
EnzymeConc7.5:DetStockDet0:CaStockCa0    -0.509701   0.258689  -1.970 0.051434 .
EnzymeConc15:DetStockDet0:CaStockCa0     -0.390763   0.258757  -1.510 0.134007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2887 on 105 degrees of freedom
Multiple R-squared:  0.9864,    Adjusted R-squared:  0.9795
F-statistic: 143.7 on 53 and 105 DF,  p-value: < 2.2e-16
```

Figure 13: The model and all of its coefficients

```
library(readr)
library(car)
library(ggplot2)
library(reshape2)
library(MASS)
SPR <- read_delim("Path_To_Data",
                  "\t", escape_double = FALSE, trim_ws = TRUE)

## Convert factos to factors
SPR$Enzyme <- as.factor(SPR$Enzyme)
SPR$DetStock <- as.factor(SPR$DetStock)
SPR$CaStock <- as.factor(SPR$CaStock)
SPR$EnzymeConc <- as.factor(SPR$EnzymeConc)
SPR$RunDate <- as.factor(SPR$RunDate)

###### DESCRIPTIVE STATISTICS ######
## Summary of the data
summary(SPR)
str(SPR)
par(mfrow=c(1,1))
plot(SPR)

## Boxplot
dat.m <-melt(SPR,id.vars='Enzyme', measure.vars=c('Response'))
ggplot(dat.m) +
    geom_boxplot(aes(x=Enzyme, y=value, color=variable))

##Pretty boxplots.
## setup dataframes for plotting
df <- data.frame(f1=SPR$Enzyme,f2=SPR$EnzymeConc)
df$f1f2 <- interaction(df$f2,df$f1)

## Plotting
ggplot(aes(y = SPR$Response, x = f1f2,fill=f1), data = df) +
 geom_boxplot()+
 theme(axis.line=element_line(colour="black"),
 panel.background = element_blank())+
 labs(x = "Concentration.Enzyme",y="Response",fill="Enzyme")

df <- data.frame(f1=SPR$Enzyme,f2=SPR$DetStock)
df$f1f2 <- interaction(df$f2,df$f1)
ggplot(aes(y = SPR$Response, x = f1f2,fill=f1), data = df) +
  geom_boxplot()+
  theme(axis.line = element_line(colour = "black"),
  panel.background = element_blank())+
  labs(x = "Detergent",y="Response",fill="Enzyme")

df <- data.frame(f1=SPR$EnzymeConc,f2=SPR$DetStock)
df$f1f2 <- interaction(df$f2,df$f1)
ggplot(aes(y = SPR$Response, x = f1f2,fill=f1), data = df) +
```

```
  geom_boxplot()+
  theme(axis.line = element_line(colour = "black"),
  panel.background = element_blank())+
  labs(x = "Detergent",y="Response",fill="Concentration")

df <- data.frame(f1=SPR$Enzyme,f2=SPR$CaStock)
df$f1f2 <- interaction(df$f2,df$f1)
ggplot(aes(y = SPR$Response, x = f1f2,fill=f1), data = df) +
  geom_boxplot()+
  theme(axis.line = element_line(colour = "black"),
  panel.background = element_blank())+
  labs(x = "Hardness",y="Response",fill="Enzyme")

df <- data.frame(f1=SPR$EnzymeConc,f2=SPR$CaStock)
df$f1f2 <- interaction(df$f2,df$f1)
ggplot(aes(y = SPR$Response, x = f1f2,fill=f1), data = df) +
  geom_boxplot()+
  theme(axis.line = element_line(colour = "black"),
  panel.background = element_blank())+
  labs(x = "Hardness",y="Response",fill="Enzyme")

# This is something.
df <- data.frame(f1=SPR$DetStock,f2=SPR$CaStock)
df$f1f2 <- interaction(df$f2,df$f1)
ggplot(aes(y = SPR$Response, x = f1f2,fill=f1), data = df) +
  geom_boxplot()+
  theme(axis.line = element_line(colour = "black"),
  panel.background = element_blank())+
  labs(x = "Hardness",y="Response",fill="Enzyme")

pairs(SPR)
cols <- c("blue","red","green","magenta","purple")
#x.factor, trace.factor, response
with(SPR, interaction.plot(DetStock, Enzyme, Response,
 type="b",col=cols))
with(SPR, interaction.plot(DetStock, EnzymeConc, Response,
 type="b",col=cols))
with(SPR, interaction.plot(CaStock, Enzyme,Response,
 type="b",col=cols))
with(SPR, interaction.plot(CaStock, EnzymeConc, Response,
 type="b",col=cols))
with(SPR, interaction.plot(EnzymeConc, Enzyme, Response,
 type="b",col=cols))

##### REGRESSION ANALYSIS #####

## Full model with all attributes and 3 way interactions
fullLinearModel <- lm(Response ~
  (Enzyme + EnzymeConc + DetStock + CaStock + Cycle)^3, SPR)
par(mfrow=c(2,2))
```

```
##Looking at assumptions
plot(fullLinearModel)
##Looking at transformation
par(mfrow=c(1,1))
boxcox(fullLinearModel,
 lambda = seq(0.2, 0.5, length.out = 20))

fullLinearModel <- lm(Response^(0.30) ~
 (Enzyme + EnzymeConc +DetStock + CaStock + Cycle)^3, SPR)


##Better
par(mfrow=c(2,2))
plot(fullLinearModel)
reducedFullLinearModel<-stepP(fullLinearModel)$object
## 1 included in confidence interval, transformation is fine
boxcox(reducedFullLinearModel,
 lambda=seq(0.5,1.50,length.out = 20))

par(mfrow=c(2,2))
plot(reducedFullLinearModel)
par(mfrow=c(1,1))
##Normality is fine
qqPlot(reducedFullLinearModel$residuals)
summary(reducedFullLinearModel)
anova(reducedFullLinearModel)
#residuals look fine
residualPlots(reducedFullLinearModel)

## Observations 147 appers to be an outlier.
## Let's look at it.
t147 <- SPR[147,] # it its enzyme E
t147
E <- SPR[SPR$Enzyme == 'E',]
E[E$EnzymeConc=='2.5',]
# observation 147 has an concentration of 2.5
# so we look at that to compare
t147

# It appears it is an outlier
# so we remove and create a new Fullmodel
SPRwithout147 <- SPR[-147,]

# Create fullLinearModel without outlier.
wo147fullLinearModel <- lm(Response ~
 (Enzyme + EnzymeConc + DetStock + CaStock + Cycle)^3,
 SPRwithout147)

par(mfrow=c(2,2))
plot(wo147fullLinearModel)
```

```
par(mfrow=c(1,1))
boxcox(wo147fullLinearModel)
boxcox(wo147fullLinearModel,
 lambda = seq(0.25, 0.5, length.out = 20))
wo147fullLinearModel <- lm(Response^(0.3) ~
 (Enzyme + EnzymeConc + DetStock + CaStock + Cycle)^3,
 SPRwithout147)

par(mfrow=c(2,2))
reducedWo147fullLinearModel<-stepP(wo147fullLinearModel)$object

par(mfrow=c(1,1))
boxcox(reducedWo147fullLinearModel,
 lambda = seq(0.5, 1.5, length.out = 20))

par(mfrow=c(2,2))
plot(reducedWo147fullLinearModel, col=as.numeric(SPR$Enzyme))
par(mfrow=c(1,1))
qqPlot(reducedWo147fullLinearModel$residuals)
residualPlots(reducedWo147fullLinearModel)
summary(reducedWo147fullLinearModel)
anova(reducedWo147fullLinearModel)
## Everything looks fine, however
## slight downward trend in scale location,
## but data fairly evenly spread

##### PREDICTION #####
## Transform back function
transform.back<-function(y){
  y^(1000/300)
}
##Predicting response based on enzymeconc
pred.data <- expand.grid(Enzyme=levels(SPRwithout147$Enzyme),
 EnzymeConc=levels(SPRwithout147$EnzymeConc),
 CaStock="Ca+", DetStock="Det+", Cycle=23)
pred.mod<-transform.back(predict(reducedWo147fullLinearModel,
 newdata=pred.data, interval="confidence"))
pred.mod2<-transform.back(predict(reducedWo147fullLinearModel,
 newdata=pred.data, interval="prediction"))

par(mfrow=c(1,1))
plot(c(0,2.5,7.5,15)[as.numeric(pred.data$EnzymeConc)],
 pred.mod[,1],
 ylim = range(pred.mod), ylab = "Response(RU)",
 xlab = "Enzyme␣Concentration(nM)",
 main="Confidence␣Interval␣95%␣for␣Cycle␣23,␣Det+,␣Ca+")

matlines(c(0,2.5,7.5,15),
 matrix(pred.mod[,1], nrow = 4, byrow = TRUE),
 lty=1, lwd=2, col=c(2:6))
```

```
matlines(c(0,2.5,7.5,15),
 matrix(pred.mod[,2], nrow = 4, byrow = TRUE),
 lty=2, lwd=2, col=c(2:6))
matlines(c(0,2.5,7.5,15),
 matrix(pred.mod[,3], nrow = 4, byrow = TRUE),
 lty=2, lwd=2, col=c(2:6))
legend(0,1620,c("A", "B", "C", "D", "E"),
lty=c(1,1,1,1,1),lwd=c(2.5,2.5,2.5,2.5,2.5),col=c(2:6))

##Predicting response based on enzymeconc
pred.data <- expand.grid(Enzyme=levels(SPRwithout147$Enzyme),
 EnzymeConc=levels(SPRwithout147$EnzymeConc),
 CaStock="Ca0", DetStock="Det0", Cycle=23)
pred.mod <- transform.back(predict(reducedWo147fullLinearModel,
 newdata=pred.data, interval="confidence"))
pred.mod2 <- transform.back(predict(reducedWo147fullLinearModel,
 newdata=pred.data, interval="prediction"))

par(mfrow=c(1,1))
plot(c(0,2.5,7.5,15)[as.numeric(pred.data$EnzymeConc)],
 pred.mod[,1],
 ylim = range(pred.mod), ylab = "Response(RU)",
 xlab = "Enzyme␣Concentration(nM)",
 main="Confidence␣Interval␣95%␣for␣Cycle␣23,␣Det0,␣Ca0")
matlines(c(0,2.5,7.5,15),
 matrix(pred.mod[,1], nrow = 4, byrow = TRUE),
 lty=1, lwd=2, col=c(2:6))
matlines(c(0,2.5,7.5,15),
 matrix(pred.mod[,2], nrow = 4, byrow = TRUE),
 lty=2, lwd=2, col=c(2:6))
matlines(c(0,2.5,7.5,15),
 matrix(pred.mod[,3], nrow = 4, byrow = TRUE),
 lty=2, lwd=2, col=c(2:6))
legend(0,750,c("A", "B", "C", "D", "E"),
lty=c(1,1,1,1,1),lwd=c(2.5,2.5,2.5,2.5,2.5),col=c(2:6))
```