Danmarks Tekniske Universitet

02441
Applied Statistics and statistical software (R)

Effect of hardness and detergent on enzymatic catalysis

Agnieszka Golinska (s151222)
Manuel Montoya Catala (s162706)
Paulina Jaworska (s151330)

January 2017

Contents

1 Summary

In this report a statistical analysis of enzyme performance in removing stains from surfaces is performed. Thourghout the report, the relationship between catalytic activity and other variables is investigated and a linear model is formulated, described and gradually reduced, eliminating the parameters that appeared to be insignificant. First, water hardness and detergent influence on protein washout was analysed. Results showed that detergent is significant, while calcium ions are not significant regarding the response. Next, concentration of enzyme was added to the model, as well as the type of enzyme. The concentration was found to have an influence on the enzyme performance, while the type of enzyme used was significant in the model with detergent. After the analysis, systematic errors were discussed and the final model is presented.

2 Introduction

Stains on textiles can be removed using enzymes and their combination with other factors like detergents and water hardness. In our case, the effect of hardness and detergent on enzymatic catalysis is analyzed. In order to measure the performance of enzyme, Surface Plasmon Resonance technology method was used. Thanks to that, it was possible to measure the amount of protein that is removed from the surface. Data for the project was collected for 10 days. Each experiment lasted 2 days and it included 1 enzyme under four conditions (Det0Ca0, Det0Ca1, Det1Ca0, Det1Ca1) for each concentration of enzyme (0nM, 2.5nM, 7.5nM, 15nM). The test was replicated and each of the replicates was examined in random order.

In this project, we perform statistical tests and train linear regression models in which the response variable y is the Response and the explanatory variables $X = \{x_1, x_2, ...\}$ are the rest of the variables (Enzyme, CatStock...). Some of the explanatory variables are discrete and other are continuous so we have a final mixed linear models with factors and continuous regressions.

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

where, β_i are the coefficients, x_p are the explanatory variables, and ϵ is the error term to capture other sources of variation. We investigate the relationship of the variable y, which in our case is the amount of protein being removed from the surface, and other variables such as: water hardness, detergent presence, enzyme concentration, and enzyme type. The report consists of answers to the following questions:

- How does hardness and detergent influence the catalytic activity?
- Is the catalytic activity dependent on the amount of enzyme present?
- Are there any differences in performance among the enzymes in this study regarding the factors mentioned above?
- Are there indications of systematic errors due to one enzyme per experiment, how would this affect the model?
- Find a linear model that captures the relationship between the response and the explanatory variables.

3 Description of the data

The data file (spr.txt) contains columns with:

- RunDate: Run date in the YYMMDD format.
- Cycle: Cycle number within run.
- Response: Amount of protein removed by enzyme. In RU where 1 RU equals $10^{-6} g/m^2$.
- Enzyme: 5-level factor.
- EnzymeConc: Enzyme concentration in nM, either 4-level categorical factor or numeric.
- DetStock: With or without detergent (2-level factor).
- CaStock: Hardness (with or without Ca^{++}).

Looking at the structure of the data we see that the run date and cycle number are integers, response and enzyme concentrations are numeric, while enzyme types, detergent presence and water hardness are factors (with 5, 2, and 2 levels, respectively).

There are 160 observations of 7 variables. Table 1 presents the first 5 observations in a tabular form.

	RunDate	Cycle	Response	Enzyme	EnzymeConc	DetStock	CaStock
1	81203	1	323.0	B	2.5	Det+	Ca+
2	81203	2	614.4	B	7.5	Det+	Ca0
3	81203	3	325.6	B	15.0	Det0	Ca+
4	81203	4	161.7	B	7.5	Det0	Ca0
5	81203	5	545.3	B	2.5	Det+	Ca0

Table 1: Example of some of the observations.

Table 2 shows the summary of the data that we use in our analysis. From the table we can obtain different information of the model, like the means, medians and quantiles. For example, the mean of the response in the model, in other words the average amount of protein removed by enzyme, is 431.6 RU, while the mean on enzyme concentration is 6.25 nM. There are 32 observations for each of 5 enzymes (A, B, C, D, E), 80 observations.

RunDate	Cycle	Response	Enzyme	EnzymeConc	DetStock	CaStock
Min.: 81125	Min.: 1.00	Min.: 0.1	A:32	Min.: 0.000	Det+:80	Ca+:80
1st Qu.: 81127	1st Qu.: 9.00	1st Qu.: 94.6	B:32	1st Qu.: 1.875	Det0:80	Ca0:80
Median:81203	Median :17.50	Median : 322.4	C:32	Median : 5.000		
Mean:81174	Mean:17.38	Mean: 431.6	D:32	Mean: 6.250		
3rd Qu.:81205	3rd Qu.:25.25	3rd Qu.: 662.7	E:32	3rd Qu.: 9.375		
Max.:81208	Max.:34.00	Max.:1588.0		Max.:15.000		

Table 2: Summary of the data.

We will initially consider a transformation of the Response variable to reduce possible non-normality of the errors of the linear models that we will further train with the data. We will use the Box-Cox transformation in this regard. The Box-Cox plot λ value obtained lies around 0.5 which indicates that a squared root transformation of the variable should be performed. The next figure shows the Box-Cox plot for the original and transformed Response variable. We can appreciate how after the transformation, the λ value lies around 1.

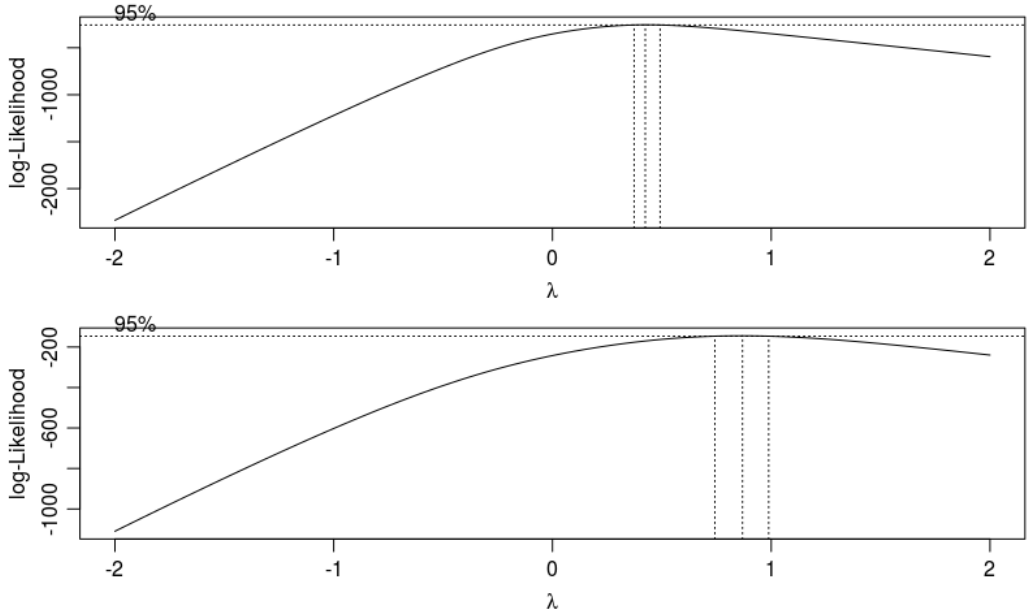

Figure 1: BoxCox plot of the original and transformed data

We can also look at the histogram of the response variable before and after the transformation. The next figure shows this distribution. We can appreciate how the transformation makes the distribution of the Response variable more Gaussian-like.

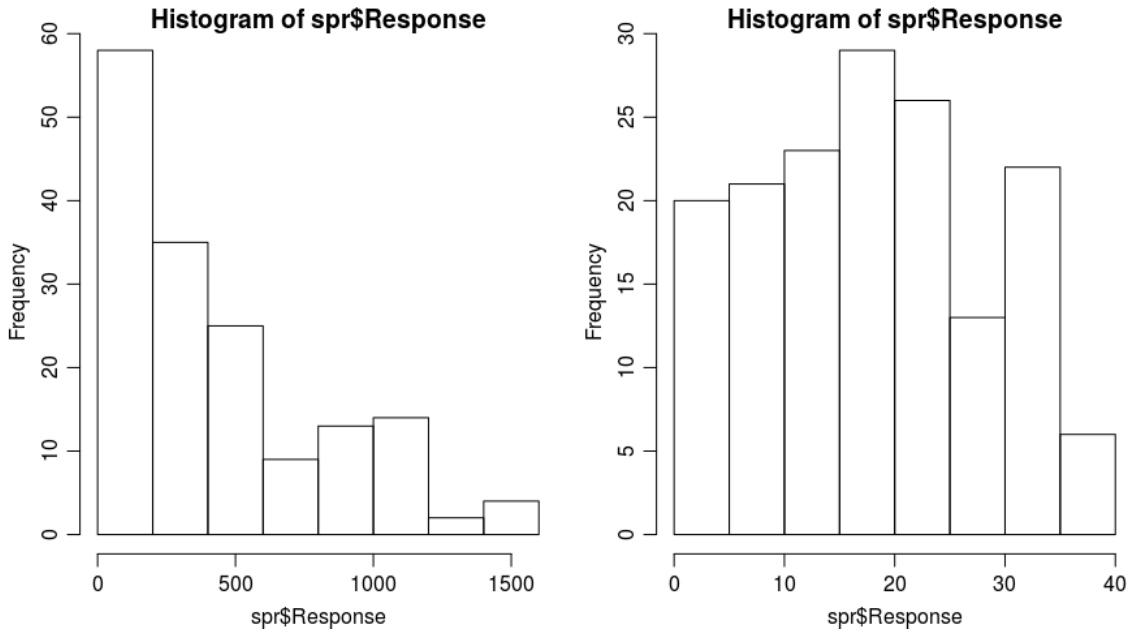

Figure 2: Histogram of the original and transformed data.

During the following steps we will study the statistical relationships between the response variable and the explanatory variables. We will mainly fit linear models in which we will assume that the different samples are independent and the residual is Gaussian noise.

4 Statistical analysis

4.1 How does hardness and detergent influence the catalytic activity?

Figure 3 shows the distribution of the data taking into account four situations: presence of calcium ions and detergent in the water, presence of detergent without calcium ions, presence of calcium ions but without detergent, and water without calcium ions and detergent. Top figure is a boxplot, while the bottom figure is a scatter plot of the model.

Looking at Figure 3 it is easily seen that calcium ion presence doesn't influence the mean μ nor the variance σ of the distribution (boxplots). Regarding the spread of each distribution we can observe that the presence of detergent in the water influences the Response that we get, changing its mean and variance. When we add detergent to the water we obtain a bigger spread of the distribution and a higher mean (better result - higher amount of protein that is removed from a surface). We will run statistical tests now to computationally check these effects.

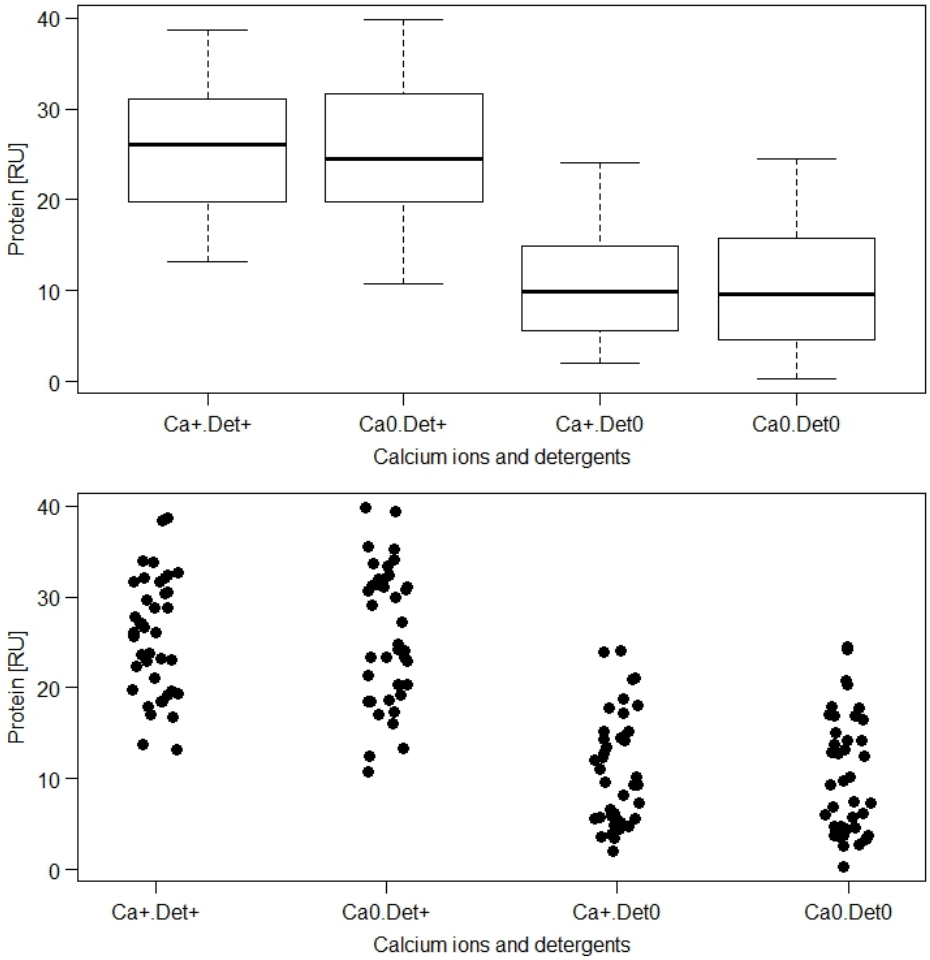

Figure 3: Distribution of the data divided in the factors: calcium ions (water hardness) and detergent.

In order to make a statistical analysis of how the hardness of the water and presence of detergent influence the catalytic activity we fit a linear model where the response variable depends on both of them and their interaction.

$$Response = Ca^{++} + Detergent + Ca^{++} : Detergent + \epsilon$$

The coefficients obtained are as shown in Table 3:

(Intercept)	CaStockCa0	DetStockDet0	CaStockCa0:DetStockDet0
5.022383	0.004488	-1.845154	-0.135660

Table 3: Coefficients obtained by the model.

The properties of the coefficients obtained are shown in the next table (Table 4). From the p-value we can check the significance of the contribution of the factors. We assume that a parameter is significant when its p-value $< \alpha$, where α is a significance level equal to 0.05. All of the parameters with p-value $>$ 0.05 are considered insignificant.

	Estimate Std.	Error	t-value	Pr($>\|t\|$)
(Intercept)	25.6373	1.0549	24.304	$<$ 2e-16
CaStockCa0	0.2044	1.4918	0.137	0.891
DetStockDet0	-14.6962	1.4918	-9.851	$<$ **2e-16**
CaStockCa0:DetStockDet0	-0.7431	2.1097	-0.352	0.725

Table 4: Results of the T-statistic for the parameters of the model.

From Table 4 we see that the factor Cat has no significant effect on the Response (its p-value is a lot bigger than 0.05) and the same applies for the interaction between water hardness and detergent presence, it is not significant. This means that we can eliminate the factor Cat from our model.

	Estimate Std.	Error	t-value	Pr($>\|t\|$)
(Intercept)	25.8231	0.9110	28.347	$<$2e-16
CaStockCa0	-0.1672	1.0519	-0.159	0.874
DetStockDet0	-15.0678	1.0519	-14.324	$<$ **2e-16**

Table 5: Results of the T-statistic for the reduced model.

Table 5 shows the T-statistic for the reduced model. Results from the test confirm conclusions drawn from distribution plots shown in Figure 3. The p-value of the detergent is below 0.05, which means that the detergent is a significant factor that has an influence on the result. Calcium ion presence in the water is not significant and the model can be reduced again.

4.2 Is the catalytic activity dependent on the amount of enzyme present?

To investigate if the catalytic activity is dependend on the amount of enzyme present we used the same models as mentioned in section above. Ij this case, our explanatory variable is not a factor but a numeric variable.

Looking at the boxplots of the distribution of enzyme concentration in Figure 4, we see that the higher the concentration the bigger the spread of the distribution. For example, having enzyme concentration of 15, the mean μ of the response is highest, as well as the variance of the distribution σ (comparing to lower concentrations).

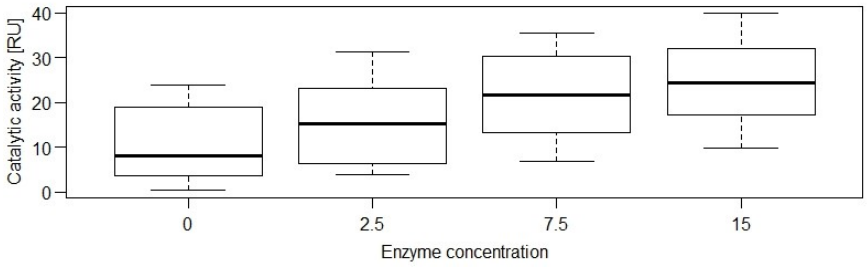

Figure 4: Distribution of each of the four enzyme concentrations.

Figure 5 shows the scatter plots of the distribution of each enzyme concentration in different combinations with water hardness and detergent presence. Based on this plot it can be concluded that calcium ions don't have significant influence on the result. The response changes with detergent presence and with increase of enzyme concentration. Best result is obtained when the concentration of enzyme is highest (15) and detergent is present.

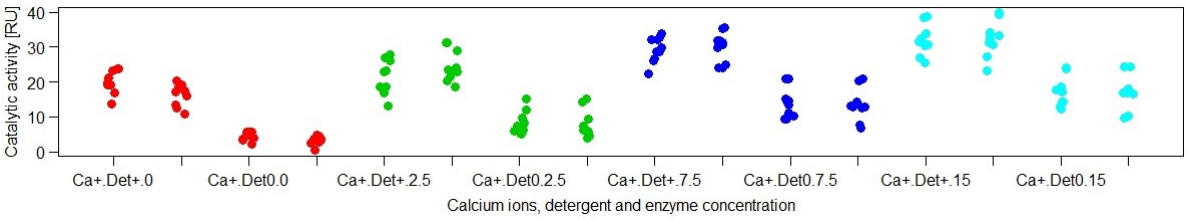

Figure 5: Distribution of the data divided in three factors: calcium ions, detergent and enzyme concentration.

Investigating enzyme concentration as a numeric effect, we could add it to the model and perform a T-test, where the null hypotesis is that the coefficients are 0.

$$Response = Ca^{++} * Detergent * Concentration + \epsilon$$

In this model, we account for the interactions between the factors as well. Table 6 shows the results of the test.

	Estimate Std.	Error	t-value	Pr($>\|t\|$)
(Intercept)	20.296774	0.993506	20.429	< 2e-16
CaStockCa0	-0.551204	1.405030	-0.392	0.695
DetStockDet0	-14.644752	1.405030	-10.423	< **2e-16**
EnzymeConc	0.854487	0.117188	7.292	**1.57e-11**
CaStockCa0:DetStockDet0	-0.306376	1.987012	-0.154	0.878
CaStockCa0:EnzymeConc	0.120894	0.165728	0.729	0.467
DetStockDet0:EnzymeConc	-0.008237	0.165728	-0.050	0.960
CaStockCa0:DetStockDet0:EnzymeConc	-0.069879	0.234375	-0.298	0.766

Table 6: Results of the T-statistic for the full model with enzyme concentration added.

Results in the table above suggest that only detergent and enzyme concentration are significant (p-values below 0.05). Neither calcium ions nor second-order interactions are significant which means they don't have influence on the response of our model. The interactions are not significant meaning that the they not add more information (reduce uncertainty) when they are jointly taken into account .It can be concluded that the catalytic activity is dependent on the amount of enzyme present.

4.3 Are there any differences in performance among the enzymes in this study regarding the factors mentioned above?

In order to check if there are any differences in performance of enzymes, a boxplot showing the influence of 5 enzymes on catalytic activity was made (Figure 6). Thanks to it, we can observe that enzyme A has the best effect, while D is the worst of all the enzymes.

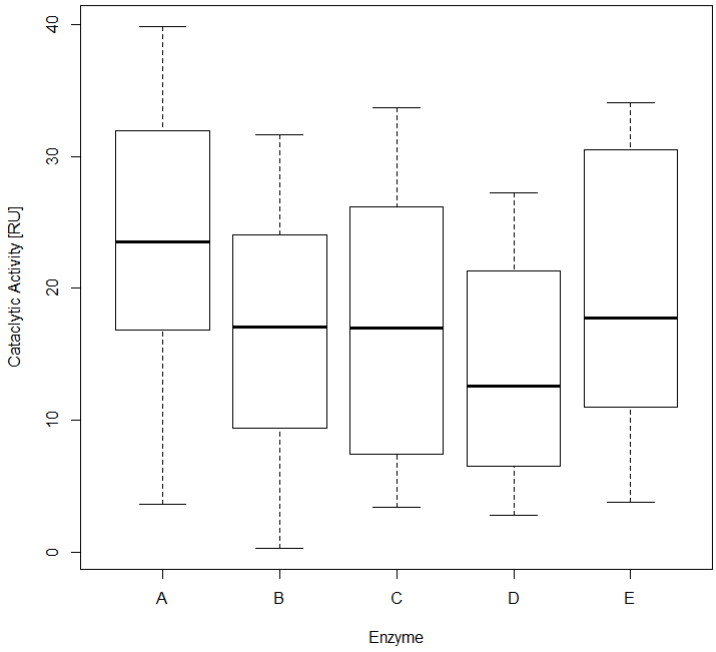

Figure 6: Influence of different enzymes on Catalytic Activity.

Figure 7 shows an interaction plot of the enzymes and concentration. We can notice that all of the enzymes follow the same pattern - the highest concentration the better it influences catalytic activity. On both Figure 7 and Figure 8 it is clearly visible that enzyme A, in comparison to the rest, has the best influence in all four concentrations.

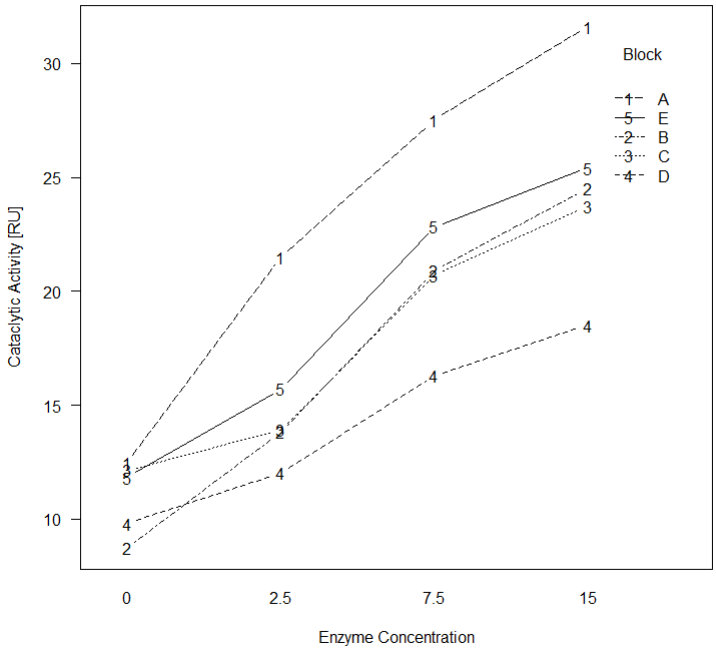

Figure 7: Interaction between enzymes and concentration

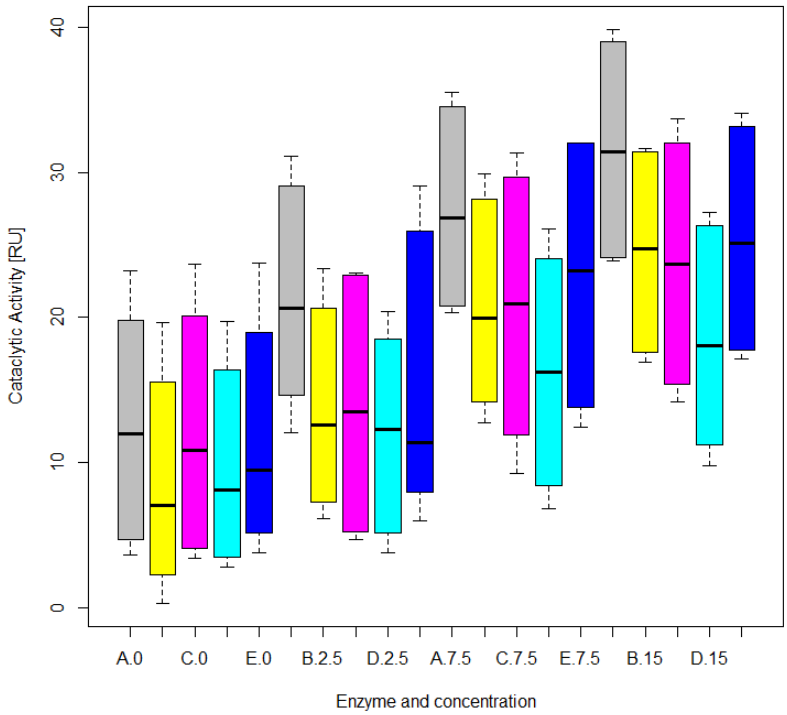

Figure 8: Interaction of different enzymes in different concentrations

Figure 9 shows the influence of enzymes together with calcium ions and detergent. Both of them indicate that enzyme A has highest influence, while enzyme D has lowest influence on the result. From detergent-enzyme interaction plot we can see that detergent is an important factor. However, the lines for different enzyme concentrations don't have exactly the same slope, which means that there are some higher order interactions. Calcium ion-enzyme interaction plot shows that water hardness is not an important factor regarding the result.

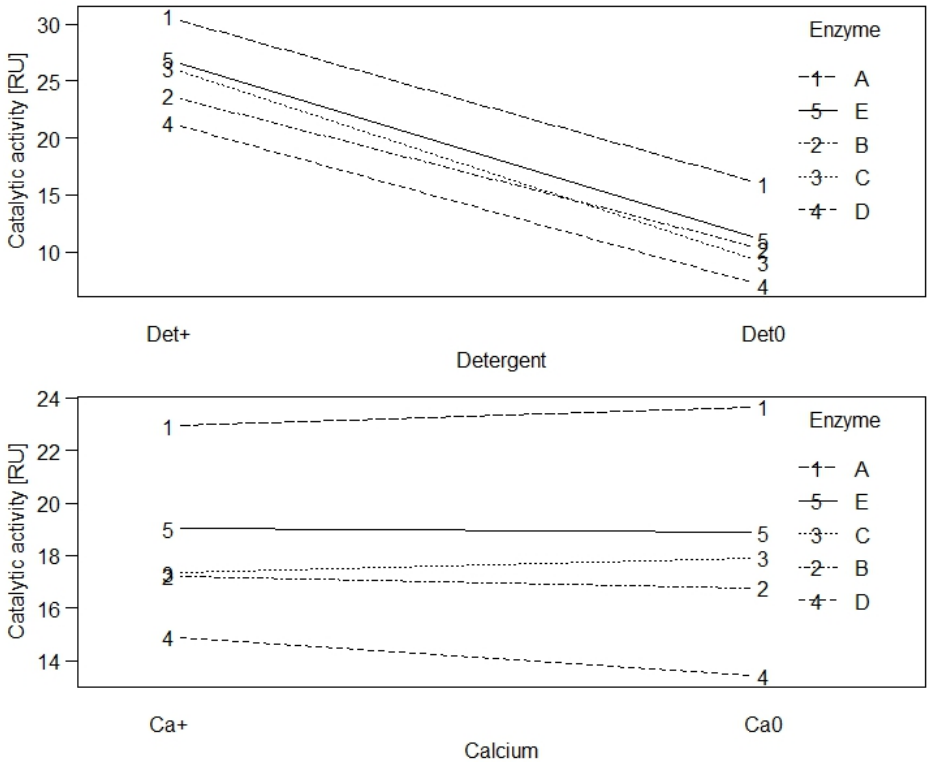

Figure 9: Catalytic activity resulting from interaction of different enzymes with water hardness and detergent.

In order to investigate behaviour of the enzymes while taking into account the factors presented before, linear models have been built. First, the model included only the enzyme variable and the second variable (concentration, detergent or a presence of calcium). They looked as below:

$$Response = Enzyme + Concentration + \epsilon$$

$$Response = Enzyme + Detergent + \epsilon$$

$$Response = Enzyme + Ca^{++} + \epsilon$$

Performing the t-test for all the three models we observed that enzymes, concentration, as well as detergent have significance contribution to the response. While calcium has no influence. In the next step we extended our linear models, adding a possible interaction between the factors. The first model obtained like that was as follow:

$$Response = Enzyme + Concentration + Enzyme : Concentration + \epsilon$$

Next it was needed to carry out the t-test to check the significance of the enzymes, concentration and their relation in such model. The results are displayed in Table 7.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	16.0244	2.1590	7.422	7.98e-12
EnzymeB	-5.4215	3.0532	-1.776	0.0778
EnzymeC	-3.4197	3.0532	-1.120	0.2645
EnzymeD	-5.4749	3.0532	-1.793	0.0750
EnzymeE	-2.6951	3.0532	-0.883	0.3788
EnzymeConc	1.1614	0.2547	4.561	**1.05e-05**
EnzymeB:EnzymeConc	-0.1390	0.3601	-0.386	0.7000
EnzymeC:EnzymeConc	-0.3571	0.3601	-0.992	0.3230
EnzymeD:EnzymeConc	-0.5826	0.3601	-1.618	0.1078
EnzymeE:EnzymeConc	-0.2617	0.3601	-0.727	0.4685

Table 7: Results of the T-statistic for the model with Enzyme and Concentration.

Here we can notice that the only enzyme that is significant is A with a p-value equal to 7.98e-12. The rest of the enzymes, even though they have low p-values (enzyme B - 0.0778, enzyme D - 0.0750), they don't have significant influence on the response. The same situation is with the ratio between enzymes and concentration. However, t-test shows that concentration itself has a significant influence on the catalytic activity. The interaction EnzymeD:EnzymeConc has a value of 0.1078, which we consider a little significant, meaning that when the Enzyme is D, the concentration influences the response in a slight different way than the other Enzymes.

The next linear model takes into account another factor - detergent and presents as below.

$$Response = Enzyme + Detergent + Enzyme : Detergent + \epsilon$$

Results from the t- test for this model can be seen in Table 8.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	30.6404	1.5087	20.309	<2e-16
EnzymeB	-6.9226	2.1336	-3.245	**0.00145**
EnzymeC	-4.4627	2.1336	-2.092	**0.03816**
EnzymeD	-9.3490	2.1336	-4.382	**2.20e-05**
EnzymeE	-3.7700	2.1336	-1.767	0.07927
DetStockDet0	-14.7138	2.1336	-6.896	**1.39e-10**
EnzymeB:DetStockDet0	1.2640	3.0174	0.419	0.67588
EnzymeC:DetStockDet0	-2.3775	3.0174	-0.788	0.43198
EnzymeD:DetStockDet0	0.4654	3.0174	0.154	0.87764
EnzymeE:DetStockDet0	-1.1219	3.0174	-0.372	0.71056

Table 8: Results of the T-statistic for the model with Enzyme and detergent.

Thanks to this, table we can observe that when we include detergent to the linear model, almost all of the enzymes have influence on the response (only enzyme E has a p - value higher than 0.5). Detergent itself is also significant. However, the relation between enzymes and detergents doesn't cause significant changes in response.
The last linear model that we need to check the performance among enzymes consists of the presence of calcium.

$$Response = Enzyme + Ca^{++} + Enzyme : Ca^{++} + \epsilon$$

Below Table (9) presents results from the t - test of given model.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	22.9189	2.4649	9.298	<2e-16
EnzymeB	-5.7059	3.4859	-1.637	0.1038
EnzymeC	-5.5679	3.4859	-1.597	0.1123
EnzymeD	-7.9945	3.4859	-2.293	**0.0232**
EnzymeE	-3.8803	3.4859	-1.113	0.2674
CaStockCa0	0.7291	3.4859	0.209	0.8346
EnzymeB:CaStockCa0	-1.1693	4.9299	-0.237	0.8128
EnzymeC:CaStockCa0	-0.1672	4.9299	-0.034	0.9730
EnzymeD:CaStockCa0	-2.2436	4.9299	-0.455	0.6497
EnzymeE:CaStockCa0	-0.9013	4.9299	-0.183	0.8552

Table 9: Results of the T-statistic for the model with Enzyme and Ca^{++}.

From this table we can learn that including Ca^{++} to the linear model causes that only Enzymes A and D are significant. Neither calcium itself, neither a ratio between Enzymes and Calcium have an influence on the response.

4.4 Are there indications of systematic errors due to one enzyme per experiment, how would this affect the model?

Systematic errors usually are the errors that come from the measuring instruments. If these instruments are not calibrated correctly or have a deformation they can output a transformed version of the original

signal, usually by means of a bias and a scale transformation.

In this experiment, the test of the different enzimes is performed in different days and we want to check if there is a systematic error due to time (the washing machine sensors might be degradating over time). For every day, there has been a control tests where no concentration of the Enzymes have been used, so if the statistical properties of these tests change significantly from day to day we may have a systematic error.

The next Figure shows the boxplot of the Response for the different days, for those tests with 0 concentration of enzyme. It can be appreciated that maybe the variance of the distributions is high enough to invalidate the difference in the mean values, being this difference due to the stochastic sampling. We perform an One-Way Anova test to check this, the F value is 0.3344 and the probability of getting an F value more extreme than that (p-value) is 0.8529, so the difference between the days is not significant. There are no systematic errors in this experiment.

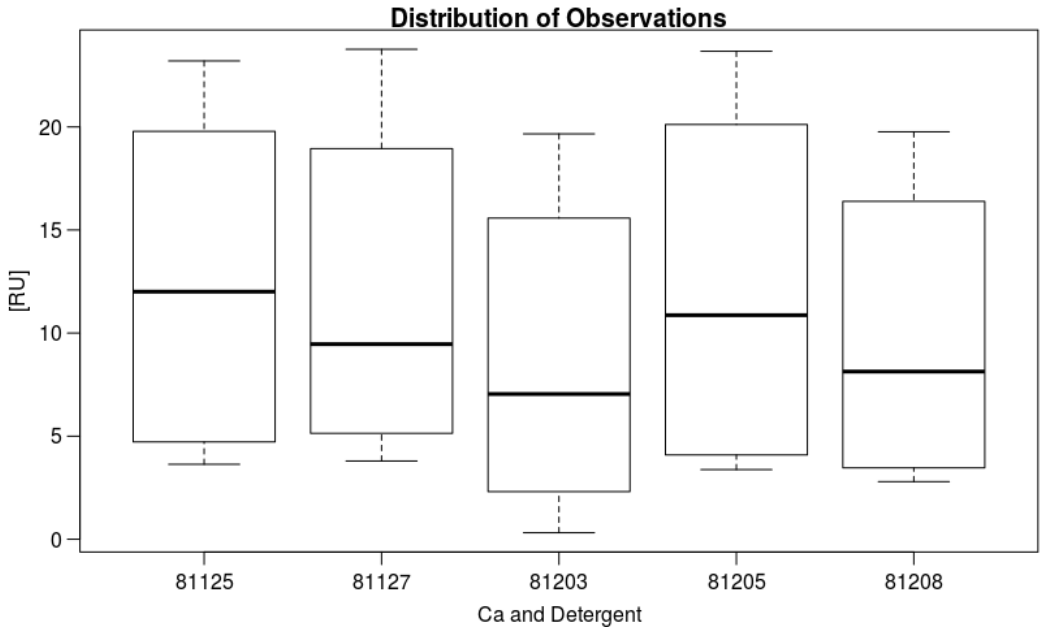

Figure 10: Boxplot of the Response as a function of the days for the 0 concentration tests

4.5 Final model presented

Given the statistical analysis performed so far we can establish a starting point to generate a model for estimating the Response, given the rest of the variables. So far we have observed that:

- The hardness of the water does not significantly affect the Response. At least on its own and in combination of the other variables that we have. It could be the case that this variable would become informative in the presence of another complementary variable, but we do not have it.
- The amount of enzyme present significantly affects the Response variable. The more enzyme concentration we have, the more Response, almost in a linear manner.
- The performance of the enzymes also depends on the existence or not of detergent. If there is detergent, there is a higher response.
- The enzymes have different performances, being A the Enzyme with the best catalytic activity.
- There seems to be no systematic error due to time so we can rule this variable out.

So the initial model presented is a linear model that depends on the factors Enzyme and Detergent and in the numeric variable Enzyme concentration. We have discarded the time and hardness of water. We add a residual term ϵ to account for the unexplained variances and we initially assume it is Gaussian with mean 0 and with the same variance for all the samples.

$$Response = Enzyme * Det * EnzymeConc + \epsilon$$

We fit this general model and then reduce it using the function step() in R. This model has a residual standard error: 2.572 on 144 degrees of freedom with Multiple R-squared: 0.9405, Adjusted R-squared: 0.9347 and an AIC: 767.9355. We will now see the properties of the residuals to check the initial assumtions and detect possible benefitial transformations of the explanatory variables.

The next figure shows an analysis of the residuals for this model. We can observe that:

- The residuals are mostly independent of the estimated area place. Their means is very close to 0 and the variance could be uniform. at the extremes the variance seems to decreases that just might be because we have less samples, or maybe me make less errors in the exptremes.
- The QQ plots shows a distribution that is very close to Gaussian around 0 but in the extremes there appears to be some outliers, mainly sample 160.
- The standardized error is lesser for small values of the fitted model and then it increases to saturation. This could be because we have more samples for the 0 Concentration zone (each enzime has 0 Concentration tests, so it is independent of the enzyme), which yields the lowest values of Response. So since we have more samples in this zone, we might have a better estimate of the real parameters and the overall model is more biased towards this zone.
- We can see how the samples 147 and 160 have a very high Cook's Distance and seems like an outlier in the remaining graphs. So we should remove it along with the samples 26 and 90.

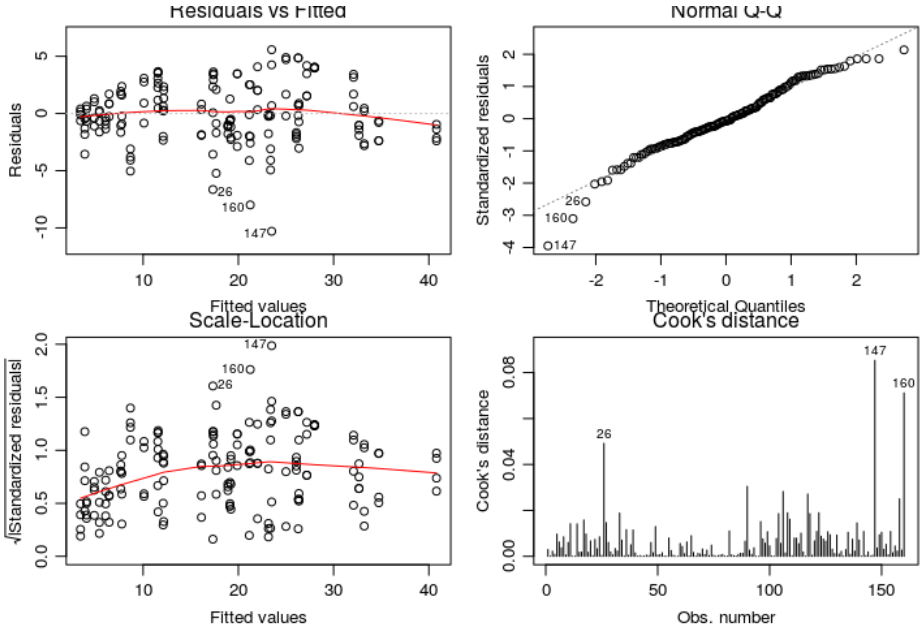

Figure 11: Residuals for the initial reduced model

These plots show the properties of all the residuals independently of the explanatory variables (they are marginalized) but we can also see if the residuals behave differently for the different variables to see if a transformation of these should be made. The next Figure shows the residuals as a function of the Enzyme concentration.

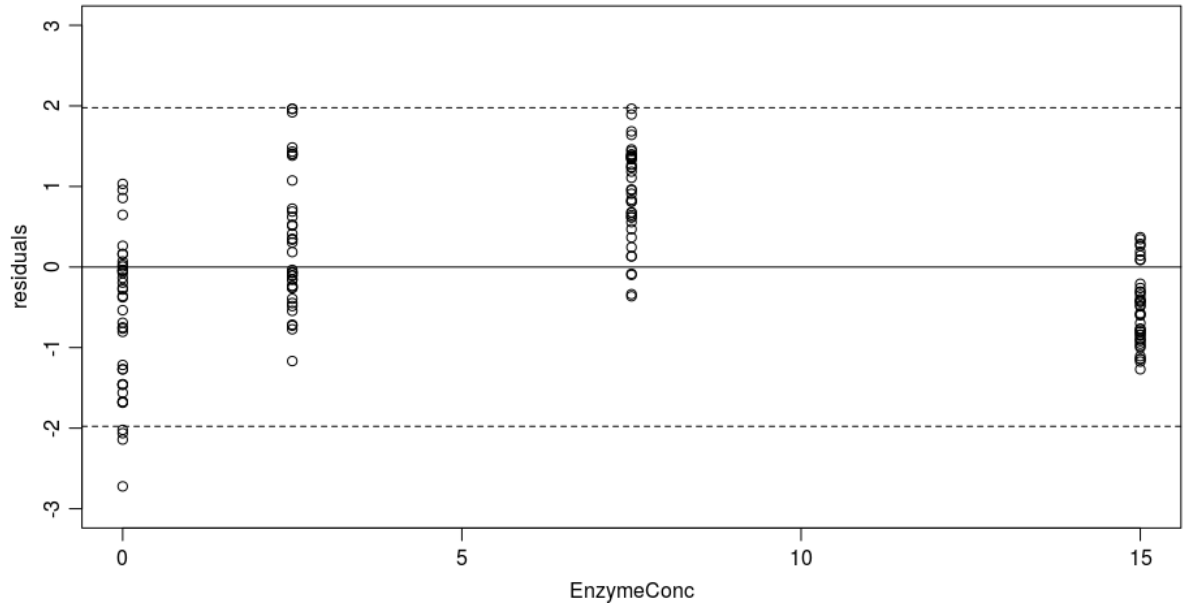

Figure 12: Residuals in terms of the Concentration

As we can see, the residuals have different mean depending on the Concentration value, we could perform a transformation of the variable to solve this. Taking the square root of the Enzyme Concentration we have the model:

$$Response = Enzyme * Det * sqrt(EnzymeConc) + \epsilon$$

Which we will also reduce in the same way as before. This model has a residual standard error: 1.939 on 144 degrees of freedom with Multiple R-squared: 0.9662, Adjusted R-squared: 0.9629 and an AIC: 678.0961.

As we can appreciate the error of the model has decreased. We will now see again the properties of the residuals in terms of the Enzyme concentration to check if they are independent. As we can appreciate in the next figure, now the residuals are less dependent on the Concentration level.

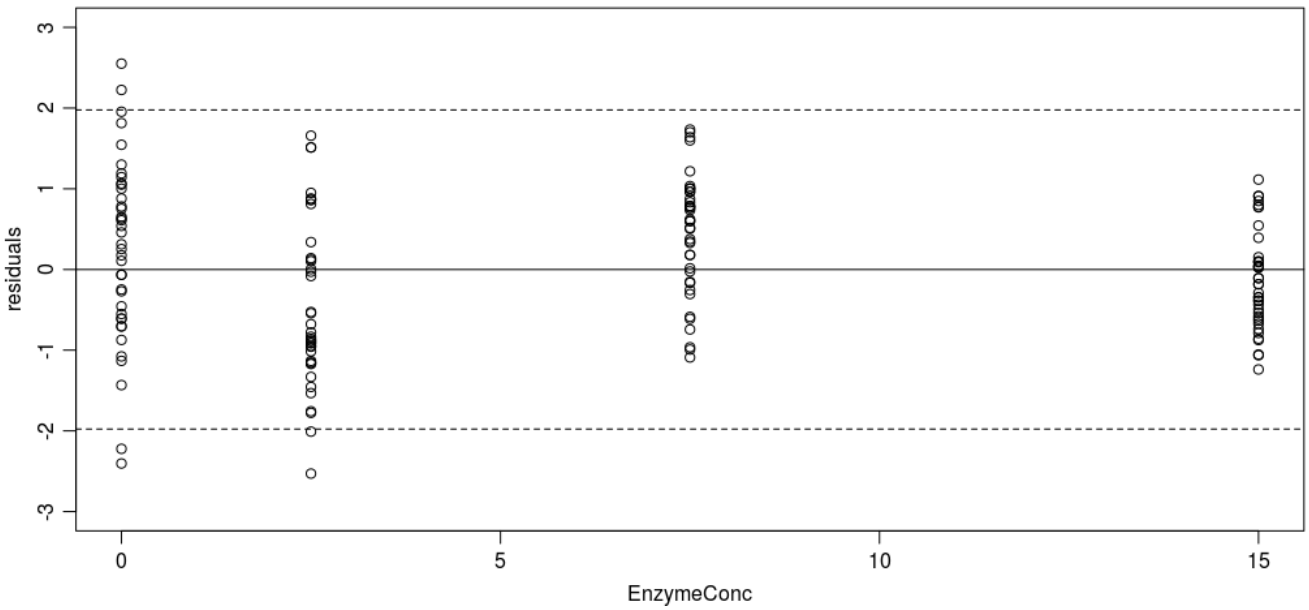

Figure 13: Residuals in terms of the Concentration

In the residuals of this model, the sample 160 is also an outlier so we decide to remove it from our dataset. Once removed, the sample 147 also is spotted as an outlier so it is also removed. The next graph shows the properties of this model once we have removed the outliers.

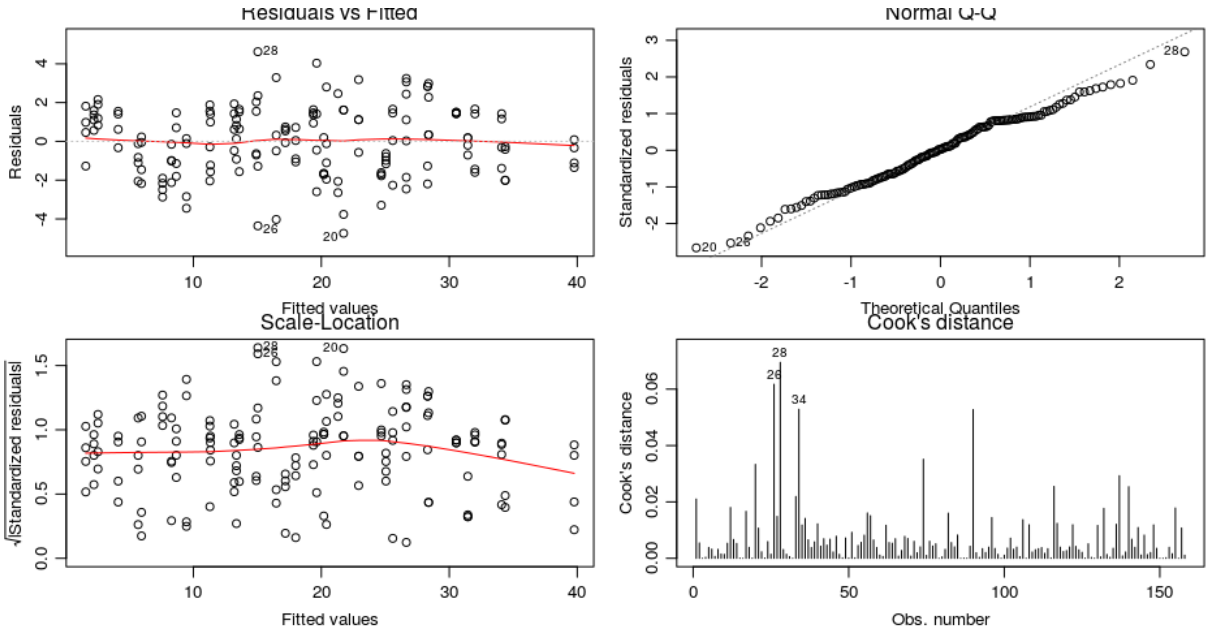

Figure 14: Residuals for the final model

As we can appreciate, the Gaussianity of the residual has improved so we prefer this model. We can also see how the residual error behaves depending on the Detergent and the Enzymes. This can be seen in the following figure. As we can observe, the mean of the residuals are not significantly different but the variances are. The variance of the residual increases in the presence of detergent.

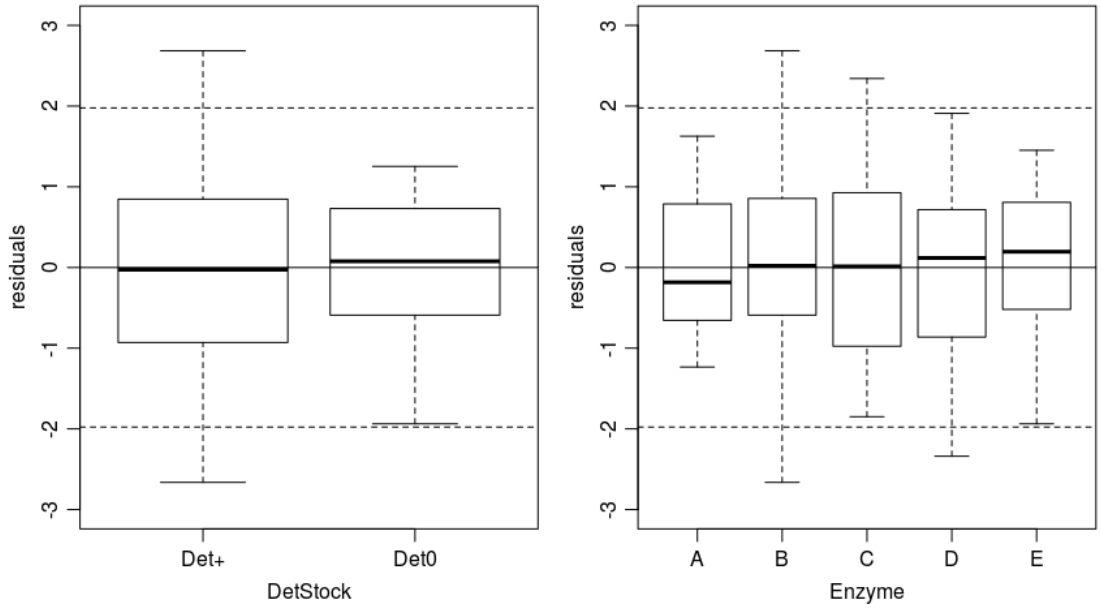

Figure 15: Residuals in terms of Detergent and Enzimes

This final reduced model has the equation:

$$\begin{aligned} Response = & DetStock + sqrt(EnzymeConc) + Enzyme + \\ & DetStock : Enzyme + sqrt(EnzymeConc) : Enzyme + \epsilon \end{aligned} \tag{1}$$

So in this final model there is some influence between the Enzyme and the Detergent and the Enzyme and the Concentration. The summary of this model is expressed in the following table.
We can appreciate how most of these parameters are significant at a 0.05 level of their t-statistic. Which is sufficient to incorporate their effects in the model.

	Estimate	Std..Error	Lower	Upper	$Pr(>\|t\|)$
(Intercept)	20.40	0.65	19.11	21.69	0.00
DetStockDet0	-14.71	0.65	-16.00	-13.43	0.00
I(sqrt(EnzymeConc))	5.00	0.23	4.55	5.45	0.00
EnzymeB	-5.36	0.93	-7.19	-3.53	0.00
EnzymeC	-0.76	0.93	-2.59	1.07	0.41
EnzymeD	-3.92	0.93	-5.76	-2.09	0.00
EnzymeE	0.90	0.97	-1.02	2.83	0.36
DetStockDet0:EnzymeB	1.26	0.92	-0.56	3.08	0.17
DetStockDet0:EnzymeC	-2.38	0.92	-4.20	-0.56	0.01
DetStockDet0:EnzymeD	0.47	0.92	-1.36	2.29	0.61
DetStockDet0:EnzymeE	-2.46	0.94	-4.32	-0.61	0.01
I(sqrt(EnzymeConc)):EnzymeB	-0.76	0.32	-1.40	-0.13	0.02
I(sqrt(EnzymeConc)):EnzymeC	-1.81	0.32	-2.44	-1.17	0.00
I(sqrt(EnzymeConc)):EnzymeD	-2.65	0.32	-3.28	-2.01	0.00
I(sqrt(EnzymeConc)):EnzymeE	-1.63	0.33	-2.27	-0.98	0.00

Table 10: Statistics of the parameters of the final model

5 Results and Conclusion

Throughout this report we studied the individual relationships between the Response variable and the explanatory variables and then we kept adding combinations of variables, increasing the complexity and expressiveness of our model, trying to capture all the significant linear patterns. In the "Final model presented" subsection some of these results and conclusions are summarised. In the following, we will elaborate them more.

First, a statistical analysis of water hardness and detergent influence on the catalytic activity was performed considering four situations: Ca+Det+, Ca0Det+, Ca+Det0, and Ca0Det0. Plotting the distribution of the data we could conclude that calcium ions don't influence the response, while detergent does. To confirm that, the t-test of the model was run. From the test we could easily seen that neither water hardness or interaction between water hardness and detergent are significant. The only significant parameter was the detergent itself.

In order to investigate how enzyme concentration influences the response that we get, we had to add it as a variable to our model. Plotting the enzyme concentration distribution we found out that higher concentration gives better results (more proteins are removed from the surface). Scatter plots of 16 combinations of calcium ions, detergent and enzyme concentration displayed that we obtain best response with high concentration of the enzyme and with presence of the detergent. Calcium ions didn't affect the results. T-test performed on the extended model confirmed those conclusions, showing that detergent and enzyme concentration are the significant variables in the model. None of the second-order interactions were found to be significant at this point.

Afterwards, we wanted to check the performance of different enzymes in the study regarding the factors mentioned above. In order to do that, we first made plots showing the influence of enzymes on catalytic activity, as well as interactions between types of enzymes and their concentration, detergent and calcium. We observed that enzyme A has the highest influence (removed the biggest number of proteins), while enzyme D is the worst of all of them. From plots we also learnt that the highest concentration and amount of detergents the better it influences catalytic activity. In the next step we created the linear models that consisted of enzymes and other factors, with and without interaction between them. From the t-test performed afterwards for every linear model we concluded that enzymes, concentration and detergent are significant for the response, while the calcium itself is not. However, after adding the relation between enzyme and other factors to the linear model, the situation has changed. Dependently on the factor used in the models, the significance of different enzymes, detergent and concentration varied. It is important to add that this doesn't apply to calcium, which doesn't have influence on the response in any of the cases.

Finally, we fitted a general linear model of the significant variables, and then reduced it with the step() function in R. We detected and removed some outliers and performed transformation of the data in order to make the residual independent and Gaussian like. After all the process we could find a linear model with a residual that met our assumptions and we could analyse the contribution of the different explanatory variables.

6 Evaluative discussion

During this report we analysed the relationships between the response variable Response and the explanatory variables, including how these explanatory variables interact with each other. The statistical significance of a given variable depends also on the other variables used in the model, since its influence could be already explained with another correlated variable so it is always hard to draw conclusions on the significance of variables and interactions when we have several of them. Nevertheless the individual studies performed showed solid significant conclusions about some of there relationships.

We have not divided the data into train a validation sets, this is a desirable thing to do to check that our system is not learning noise but since our models are simple, there is little risk of overfitting. More transformations of the variables and non-linear descriptive models could have been used but we kept this report simple.

A R code

```
setwd("C:/Users/Paulina/Desktop")
spr <- read.delim("C:/Users/Paulina/Desktop/SPR.txt")

#Summary statistics
str(spr)
head(spr)
summary(spr)

library(MASS)
#attach package
boxcox(Response~.,data=spr, plotit=TRUE)
#Explore a transformation on the response
spr$Response = sqrt(spr$Response)

#Plot the distribution of calcium ions and detergents
par(mfrow=c(2,1),mgp=c(2,0.7,0),mar=c(3,3,1,1))
boxplot(data=spr, Response~CaStock*DetStock, xlab = 'Calcium ions and detergents',ylab = 'Protein
stripchart(Response~CaStock*DetStock,data=spr, vertical=TRUE, method="jitter", xlab='Calcium ions

#Make a model
Model1a<-lm(Response~CaStock*DetStock,data=spr)
summary(Model1a)
#Eliminate the least significant parameter
drop1(Model1a,test="F")

#Reduced model (interaction not significant)
Model1b<-lm(Response~CaStock+DetStock,data=spr)
summary(Model1b)
drop1(Model1b,test="F")

#Reduced model (calcium ions not significant)
Model1c<-lm(Response~DetStock,data=spr)
summary(Model1c)
#Model is as simple as it can be
#test for assumptions
par(mfrow=c(2,2))
plot(Model1b,which=1:4)

#############################################################################################
#Enzyme concentration distribution
par(mfrow=c(2,1))
boxplot(Response~EnzymeConc,data=spr, xlab = 'Enzyme concentration',ylab = 'Catalytic activity [RU

#Enzyme concentration+calcium ions+detergent distribution
stripchart(Response~CaStock*DetStock*EnzymeConc,data=spr, vertical=TRUE, method="jitter",
           xlab='Calcium ions, detergent and enzyme concentration', ylab="Catalytic activity [RU]'
           las=1,col=c(2,2,2,2,3,3,3,3,4,4,4,4,5,5,5,5))

#Add enzyme concentration to the previous model
null <- lm(Response~1, data = spr)
Model2a<-lm(Response~CaStock*DetStock*EnzymeConc,data=spr)
Model2b<-step(Model2a, scope=list(lower=null, upper=Model2a), direction="both")
summary(Model2a)
summary(Model2b)

#############################################################################################
#Are there any differences in performance among the enzymes in this study regarding the factors me
#atach package
library(MASS)
```

```r
#Explore a transformation on the response
boxcox(Response~.,data=spr, plotit=TRUE)
spr$Response = sqrt(spr$Response)
par(mfrow=c(2,1),mgp=c(2,0.7,0),mar=c(3,3,1,1))

#plot to check how enzymes itself influeces the response
boxplot(Response~Enzyme,data=spr,ylab="Cataclytic Activity [RU]", xlab = "Enzyme")

#Interactio plot to check interaction between enzymes and concentration
interaction.plot( spr$EnzymeConc,spr$Enzyme,spr$Response, type=c("b"), ylab = "Cataclytic Activity

#Plot of the enzymes and concentrations
boxplot(Response~Enzyme + EnzymeConc, data=spr,ylab="Cataclytic Activity [RU]", xlab = "Enzyme and

#Linear models without interactions and t-tests
#to check significance of enzymes itself
ModelEE<-lm(Response~Enzyme, data = spr)
summary(ModelEE)
ModelEC<-lm(Response~Enzyme+EnzymeConc, data=spr)
summary(ModelEC)
ModelED<-lm(Response~Enzyme+DetStock, data=spr)
summary(ModelED)
ModelECal<-lm(Response~Enzyme+CaStock, data=spr)
summary(ModelECal)

#Linear models with interactions and t-tests
ModelE1<-lm(Response~Enzyme*EnzymeConc, data=spr)
summary(ModelE1)
ModelE2<-lm(Response~Enzyme*DetStock, data=spr)
summary(ModelE2)
ModelE3<-lm(Response~Enzyme*CaStock, data=spr)
summary(ModelE3)

## Checking a possible boxcox transformaiton of the data Response.
par(mfrow=c(2,1),mgp=c(2,0.7,0),mar=c(3,3,1,1))
library(MASS)  #atach package
boxcox(Response~.,data=spr, plotit=TRUE)  #Explore a transformation on the response
spr$Response = sqrt(spr$Response)
boxcox(Response~.,data=spr, plotit=TRUE)  #Explore a transformation on the response
## Checking for systematic errors
spr$RunDate = as.factor(spr$RunDate)
spr_0 = spr[spr[, "EnzymeConc"] == 0.0,]
boxplot(data=spr_0, Response~RunDate, xlab = 'Ca and Detergent',ylab = 'Protein
        [RU]', las = 1,main="Distribution of Observations")
fit <- aov(Response ~ RunDate, data=spr_0)
fit2 <- lm(Response ~ RunDate, data=spr_0)
anova(fit2)

## Building a final model
spr <- read.delim("./SPR.txt")
hist(spr$Response)
spr$Response = sqrt(spr$Response)
hist(spr$Response)
spr <- read.delim("./SPR.txt")
spr$Response = sqrt(spr$Response)
#Initial model
Model_F1<-lm(Response~DetStock*EnzymeConc*Enzyme,data=spr)
summary(Model_F1)
Model_F1R<-step(Model_F1, scope=list(lower=null, upper=Model_F1), direction="both")
summary(Model_F1R)
AIC(Model_F1R)
```

```
par(mfrow=c(2,2))
plot(Model_F1R,which=1:4)

#Testing model assumptions
par(mfrow=c(1, 1))
plot(rstandard(Model_F1R) ~ EnzymeConc, data = spr, ylim = c(-3,3), ylab="Standardized ...
     residuals")
abline(h = 0)
abline(h=c(-1, 1) * qt(.975, df = 140), lty = 2)

# Remove outlier
spr <- spr[-c(160),]
spr <- spr[-c(147),]

#Sqrt transformation
Model_F1<-lm(Response~DetStock*sqrt(EnzymeConc)*Enzyme,data=spr)
summary(Model_F1)
Model_F1R<-step(Model_F1, scope=list(lower=null, upper=Model_F1), direction="both", trace=1)
summary(Model_F1R)
AIC(Model_F1R)
par(mfrow=c(2,2))
plot(Model_F1R,which=1:4)

#Testing model assumptions
par(mfrow=c(1, 2))
plot(rstandard(Model_F1R) ~ DetStock, data = spr, ylim = c(-3,3), ylab="Standardized ...
     residuals")
abline(h = 0)
abline(h=c(-1, 1) * qt(.975, df = 140), lty = 2)

plot(rstandard(Model_F1R) ~ Enzyme, data = spr, ylim = c(-3,3), ylab="Standardized ...
     residuals")
abline(h = 0)
abline(h=c(-1, 1) * qt(.975, df = 140), lty = 2)

par(mfrow=c(2,2))
plot(Model_F1R,which=1:4)
which(duplicated(spr))
```