# CASE DETERGENT

Applied statistics and statistical software
02441

**Kasper Einarson s134604**
**Ran Wang s111503**
**Linards Kalnins s124612**

Technical University of Denmark

DTU compute

14-01-2015

# Summary

In this paper, 5 different enzymes denoted A,B,C,D,E and their ability to remove stains are being tested. Different factors varies the performance of the enzymes such as the concentration of Calcium-ions (hardness), surface active components (detergents) and the concentration of the enzyme. Statistical analyse of variance(ANOVA) and multi-linear regression is being carried out in a combination (ANCOVA) to test which of the factors are significant in the activity of the enzymes. It is shown how hardness appears to have no significant effect but the presence of detergent does matter significantly. It is concluded that enzyme A has the highest effect in response and enzyme E carry out the weakest effect. Furthermore it is shown that the amount of enzymes, the enzyme concentration, matters significantly to the removal of stains.

# Contents

# 1 Introduction

Surface Plasmon Resonance technology (SPR) is the technique used to determine how much protein from a surface is removed during a laundry wash. In this project we will test 5 different enzymes and show factors concerning the activity of the catalytic response. We will determine whether adding detergent, calcium or higher the concentration on enzymes have an impact of the efficiency of the stain removal given in amount of protein removed from a stain.

# 2 Description of Data

The first few lines of the dataset is shown in tabel 1 and gives an overview of the overall structure of data. RunDate is the date the experiment was carried out given in YYDDMM. Cycle is treatet as the numerical values from 1 to 34 levels for each run with different enzyme. The response is the amount of protein removed in $10^{-6} g/m^2$. 5 different enzymes are test denoted enzyme A,B,C,D and E which we treat as a 5 level factor. For each enzyme 4 different enzyme concentrations are tested (0, 2.5, 7.5, 15) treatet in this report as the numerical values. Finally each enzyme is tested either with or without Detergent (Det+ or Det0) and either with or without Calcium (Ca+ or Ca0) making both indicators a two level factors. From each experiment a reference enzyme is placed - data that is not included in this dataset.

| RunDate | Cycle | Response | Enzyme | EnzymeConc | DetStock | CaStock |
|---|---|---|---|---|---|---|
| 081203 | 1 | 323.0 | B | 2.5 | Det+ | Ca+ |
| 081203 | 2 | 614.4 | B | 7.5 | Det+ | Ca0 |
| 081203 | 3 | 325.6 | B | 15.0 | Det0 | Ca+ |
| 081203 | 4 | 161.7 | B | 7.5 | Det0 | Ca0 |
| 081203 | 5 | 545.3 | B | 2.5 | Det+ | Ca0 |

Table 1: Structure of data

following Figure 1 shows the catalytic activity at concentrations 0, 2.5, 7.5 and 15:

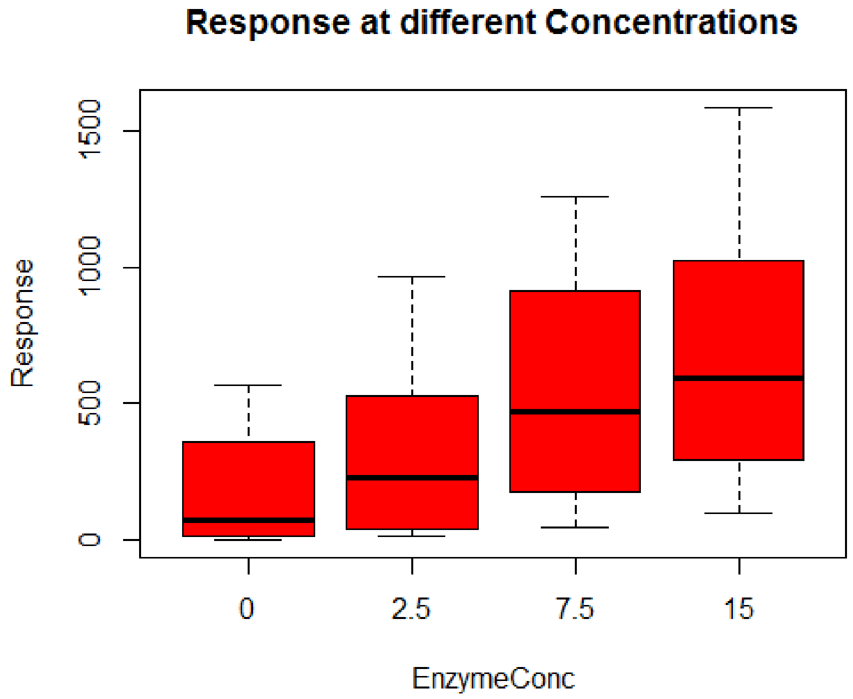


Figure 1: Response given different concentrations

It seems there's some tendency explaining higher concentration generally giving a higher response. To give an even closer look at the data,figure 2 shows response plotted as a function of the enzyme concentration where each enzyme is color represented:

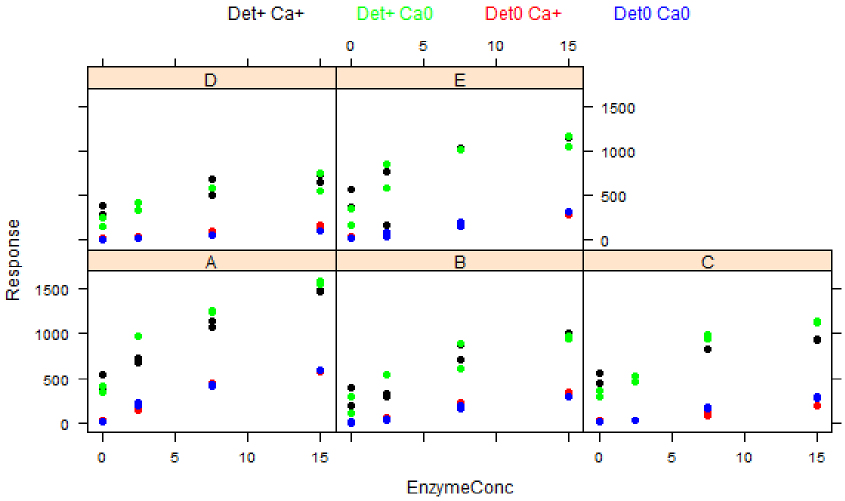


Figure 2: Response given concentrations

The explanatory variable here is the enzyme concentration and the dependent variables is the response. This indicates a tendency for all the enzymes where detergent is added generally gives a higher response. We also see that calcium doesn't seem to have a great effect since the red and the blue dots seems lower and they are somewhat the same. This is two observations that we will consider more in depth during the analysis.

## 3 Statistical Analysis

Because we want to compare several different variables including both factors and continuous variables, we want to combine the analysis of variance (ANOVA) and testing with linear models. The combination of these two is called ANCOVA and is able to compare variables while taking hidden influence from the covariate into consideration. We compare 5 enzymes, the performance of these enzymes and whether hardness and detergent has an effect. Assumptions for using ANCOVA includes constant variance, independent observations meaning no structure in the way the observations occur, linear regression between the covariate and the dependent variable and normally distributed residuals. We start by testing whether the residuals are normally distributed by making a multi-linear regression including all variables and their interaction. Figure 3 shows the plot of the model:

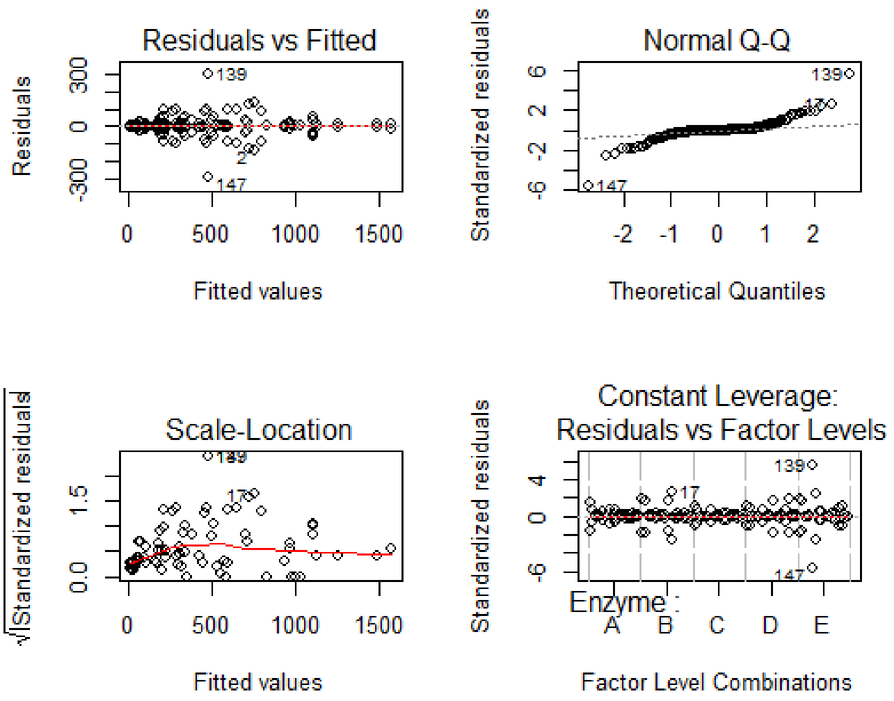


Figure 3: plot of Linear model including all interactions

Here we see the residuals behaving nice and it looks like the variance grows in the beginning but then remains somewhat constant. However from the QQ plot it seems like the values depart from normality far too early which indicates that a transformation of the dependent variable (the response) is needed.

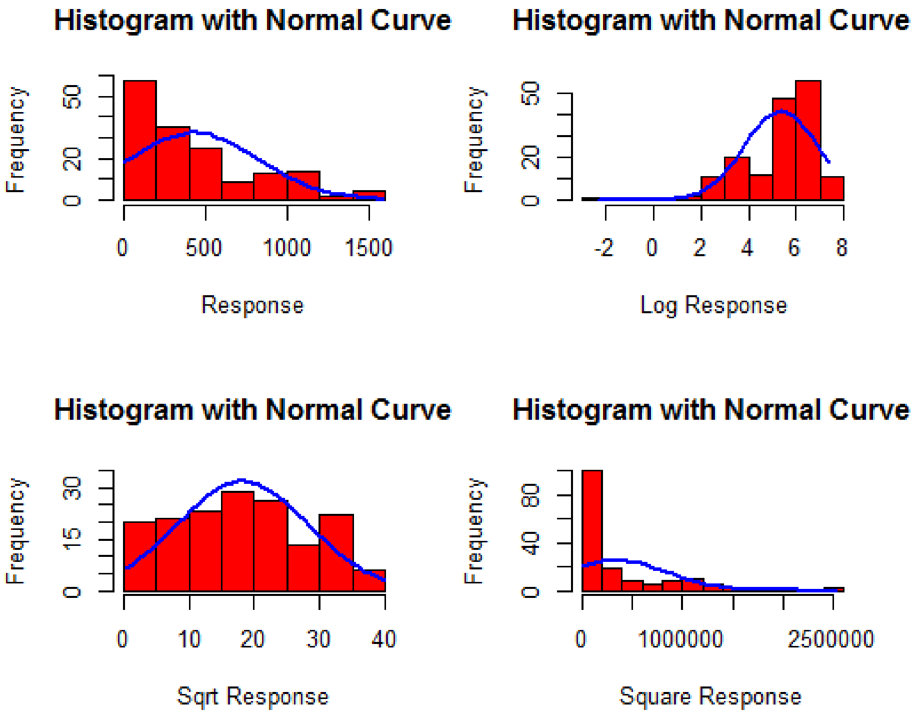


Figure 4: Histogram of 3 different transformations along with the individual normal curve

From figure 4 we see the distribution from our data with no transformation, log transformation, square root transformation and square transformation. It seems like the square root transformation is the most appropriate one of our 3 different transformation. The log transformation also seems reasonable to use. Given there's many transformation you could use, one could probably find a transformation that was even better, but within these three transformation we choose to work with the square root transformation. The linearity between the covariate and the dependent variable is achieved by taking the square root of the response, and the square root of the enzyme concentration.

Figure 6 shows the square root of the response with the square root of the enzyme concentration:

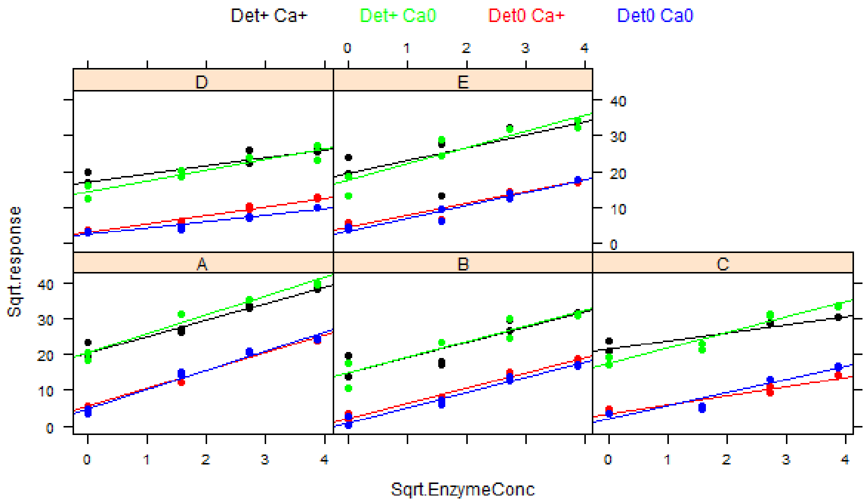
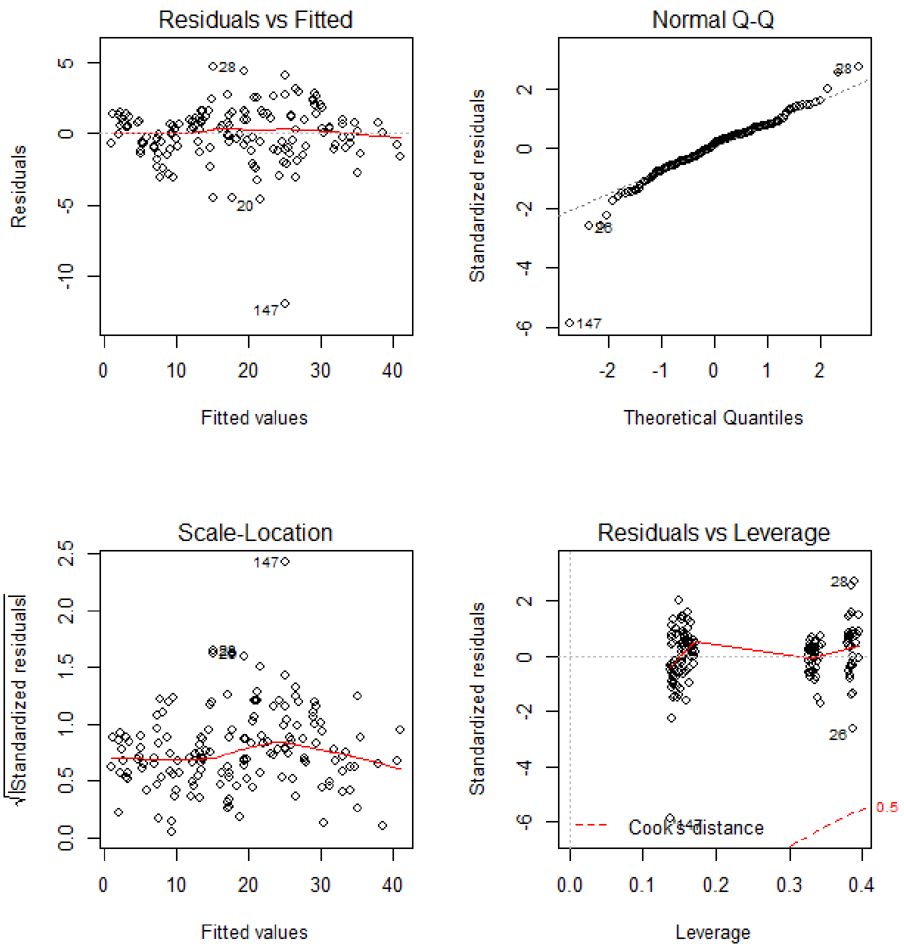


Figure 5: Square root of response as a function of square root of Enzyme concentration

From this we see a more clear linearity between our explanatory variable and the independent variable and thereby fulfilling one of the ANCOVA assumptions.

With the square root transformation our model is now given as in figure 6


Residuals vs Fitted

Normal Q-Q

Scale-Location

Residuals vs Leverage

Figure 6: Plot of first linear model with all interaction and square root transformned

The model in figure 6 shows an improvement towards normality given the QQ plot, no pattern in the residuals and the variance is also somewhat constant. We see that point 147 is a bit off the model. The measurement comes from enzyme E with concentration 2.5 and gives a response that seems significantly lower. figure 7 shows prediction interval for enzyme E and we see how the point is outside the prediction interval. This itself is not enough to draw any conclusions but for now we'll exclude it from the dataset, and come back to a better justification on our decision later.

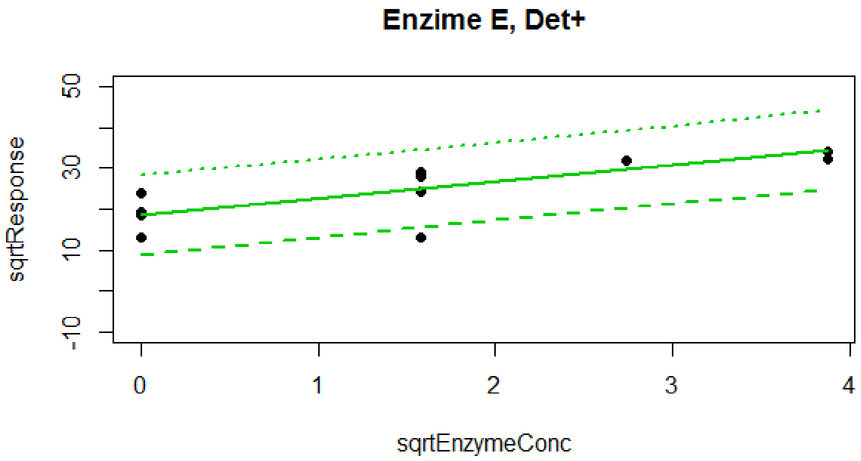


Figure 7: Plot enzyme E with detergent with prediction interval. observation 174 is out of the prediction interval and excluded from dataset

Now with our new, corrected, dataset. we'll proceed with the simplification of the linear model. This is done with R's function *step* and manually delete the terms that does not appear significant - using a 95 procent confidence interval. After the elimination our minimized model is giving in table 2

| Coefficients | Estimate Std | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| intercept | 23.34 | 0.56 | 43.90 | 2e-16 |
| EnzymeB | -6.29 | 0.602 | -10.20 | 2e-16 |
| EnzymeC | -5.615 | 0.6037 | -9.31 | 2e-16 |
| EnzymeD | -9.1163 | 0.6037 | -15.100 | 2e-16 |
| EnzymeE | -3.9533 | 0.672 | -6.75 | 1.09e-09 |
| DetStockDet0 | -15.203 | 0.381 | -39.728 | 2e-16 |
| sqrtEnzymConc | 3.6834 | 0.1322 | 27.648 | 2e-16 |

Table 2: Minimized multi- linear Model

In this tabel we see all the significant terms after elimination compared to EnzymeA (intercept) What we notice is, that no interaction involving Calcium is included and we therefore conclude that hardness has no significant influence on the response. Figure 2 showed this tendency and is now confirmed.

We saw that point 147 is a bit off the model. The measurement comes from enzyme E with concentration 2.5, Det+ and Ca+ giving a response that seemed significantly lower. Since we know calcium makes no significant different in response, we don't take this factor into consideration and thereby getting 4 reference observations:

| Sqrt.Response | Enzyme | EnzymeConc | DetStock | CaStock |
|---|---|---|---|---|
| 27.56 | E | 2.5 | Det+ | Ca+ |
| 24.18 | E | 2.5 | Det+ | Ca0 |
| 13.19 | E | 2.5 | Det+ | Ca+ |
| 29.07 | E | 2.5 | Det+ | Ca0 |

Table 3: Comparison of point 147 (line 3) and reference measurements

It's clear to see that point 147 given on line 3 of table 3 is significantly lower than the others and thereby we justify our previous decision on excluding it from the dataset.

From figure 2 it has been clear indicated that for all enzymes the response seems to get higher as the enzyme concentration gets higher. This is additionally approved by our minimized model given the "sqrtEnzymConc" is included as a significant term (very significant). Figure 8 shows a 3D plot of the interaction between the 5 different enzymes, the enzyme concentration and the response.

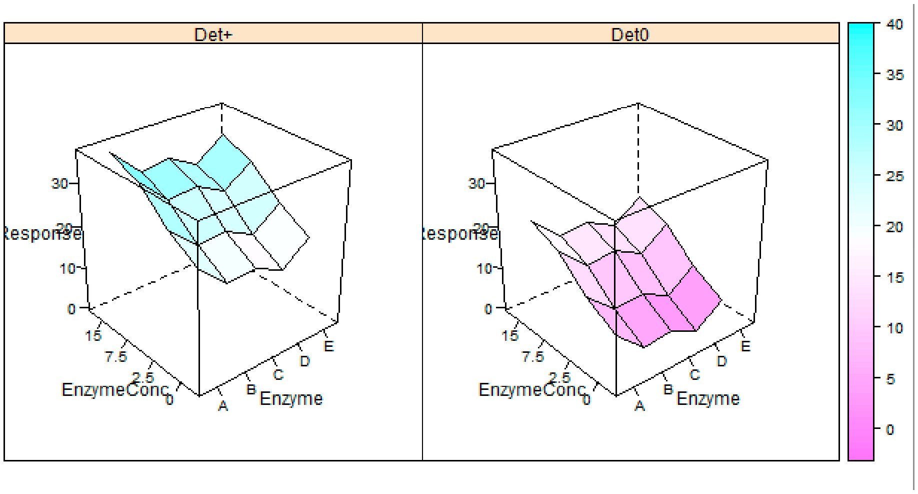


Figure 8: 3D plot of interaction between enzymes, concentration and response

Here we see two structures in a 3D space. First of all we see that the structure concerning Det+ is much higher in response similar to previous conclusions. We also see that in both cases A seems to have the highest response and B and D seems to give the lowest response. these aspects will be considered more into details later. We see that the two structures compared to each other is very similar which means, that the relationship between concentration and response doesn't change very much when adding detergent. When considering the performance among the enzymes we cancel out hardness since it has no significant effect and only look at each enzyme group with or without detergent (transformed back to original units)

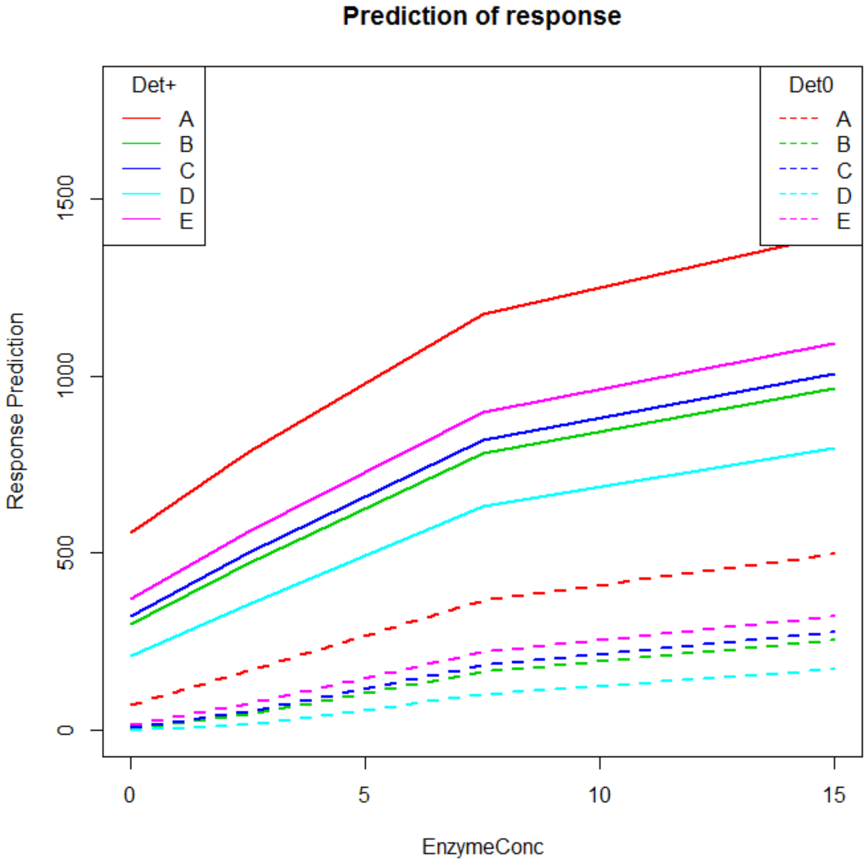


Figure 9: Enzyme response with or without detergent

We see from this plot, that enzyme A is the most efficient enzyme of them all giving the highest response both with and without detergent. We also see that enzyme E is the least efficient and that the enzymes generally gets more inefficient as we move from enzyme A to E. We see that the slope is steepest at concentrations 2.5 and 7.5 giving that interval of concentration the most response effect.

# 4 Systematic error and future work

To discover any systematic errors we first check whether the response of the experiments are independent of the cycle treated as numerical values. In order to take all hidden interaction into consideration we included cycle in the beginning linear model and saw from table 2 that the cycle wasn't significant. During the report we made the decision to leave out number 147 since there were arguments to why this would be a significantly error in measurement. If we chose to leave observation 147 in the dataset investigating normality within each enzyme it would result in a hasty conclusion concerning systematically errors

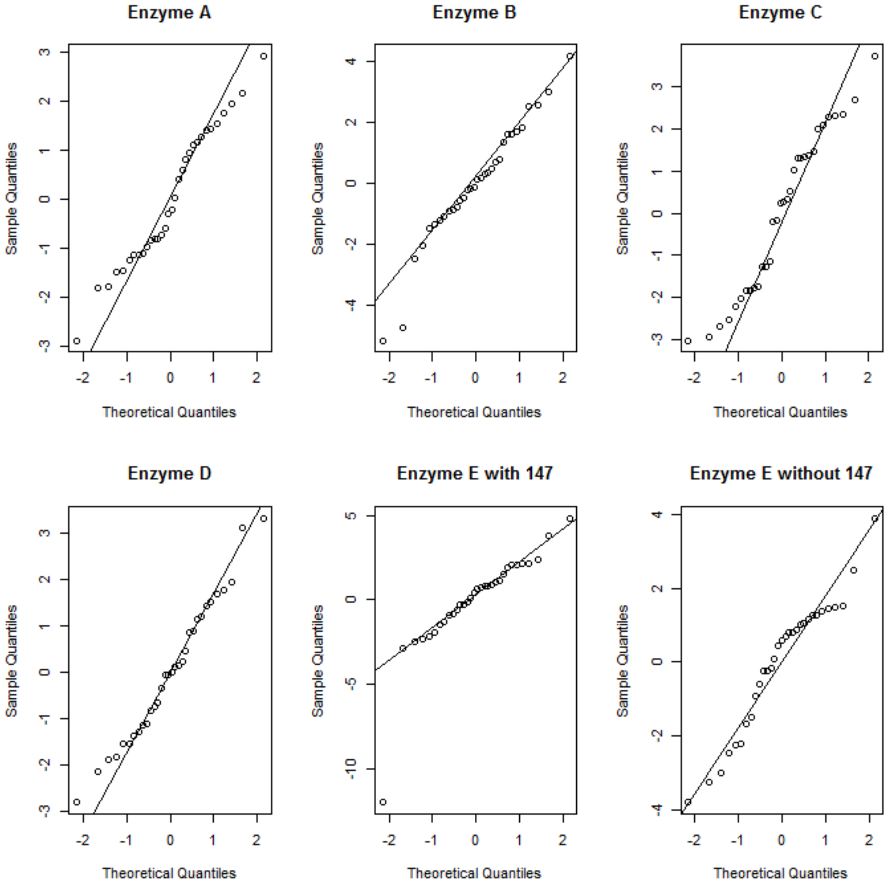


Figure 10: QQ plot of each enzyme group. Leaving obs. 147 in the dataset would have left us with wrong conclusion

The drastic change in slope in E with observation 147 could leave one to think of systematic errors but since observation 147 is only one measurement then we can NOT conclude this. All in all it is hard to determine outliers since we are limited to few replicates per combination of variables.

Since each combination of measurement with either concentration 0, 2.5, 7.5 or 15 and with or without detergent and calcium, is replicated 2 times it would be interesting to test whether there was a significance between the two measurements for all the observations. This could be a place to discover some systematic errors.
Future work could also include investigating whether the log transformation would have been significantly different from the square root transformation. When we did the histograms of the different transformations it looked as though the log transformation also could have been considered and it could be interesting to compare the two models.

In general the experiment would have been more reliable if we had changed the setup of the experiment to have more replicates per observation and each time have a control enzyme to compare performance with. This way we could easily have discovered any bias in the data and corrected it from potential systematic errors.

## 5 Conclusion

We have tested 5 different enzymes at different levels of either with or without detergent, with or without calcium and at different concentrations. Through our ANCOVA model we conclude that the adding of calcium does not make a significant different between the responses. The amount of enzymes does however matter significantly thus higher concentration gives higher response. Another important conclusion is, that adding detergent improves the performance of the enzymes significantly. The enzyme that performed best in the test was enzyme A and so the highest response can be found using enzyme A at 15 nM with detergent added. Using this combination improves the performance by approximately 40 procent compared to if one would use enzyme E at 15nM with detergent.

## 6 Appendix

Here are the code for all the plots and analysis done during this report.

```r
SPR<-read.table("SPR.txt",header=T)
summary(SPR)
str(SPR)
pairs(SPR)
SPR$lResponse<-log(SPR$Response)
SPR$sqrtResponse<-sqrt(SPR$Response)
```

```r
## first model####
lm3<-lm(sqrtResponse~ Enzyme * DetStock * CaStock * EnzymeConc,SPR)
summary(lm3)
lm4<-step(lm3)
drop1(lm4,test="F")
lm5<-update(lm4,~.-Enzyme:CaStock)
summary(lm5)
lm6<-step(lm5)
summary(lm6)
drop1(lm6,test="F")
lm7<-update(lm6,~.-DetStock:CaStock:EnzymeConc)
summary(lm7)
lm8<-step(lm7)
drop1(lm8,test="F")
lm9<-update(lm8,~.-Enzyme:DetStock )
summary(lm9)
lm10<-step(lm9)
summary(lm10)
drop1(lm10,test="F")
lm11<-update(lm10,~.-CaStock:EnzymeConc)
summary(lm11)
drop1(lm11,test="F")
lm12<-update(lm11,~.-CaStock )
summary(lm12)
drop1(lm12,test="F")
lm13<-update(lm12,~.-Enzyme:EnzymeConc)
summary(lm13)

par(mfrow=c(2,2))
plot(lm13,col=as.numeric(factor(SPR$EnzymeConc)))
anova(lm13,lm10)


## 2nd Model####
SPR$sqrtResponse<-sqrt(SPR$Response)
SPR$sqrtEnzymeConc<-sqrt(SPR$EnzymeConc)

##minus 147 model1 from every variables included reduced to the final####
SPR[-147,]
lm3<-lm(sqrtResponse~ Cycle + Enzyme * DetStock * CaStock * sqrtEnzymeConc,
        SPR[-147,])
summary(lm3)
plot(lm3)
lm4<-step(lm3)
drop1(lm4,test="F")
```

```r
lm5<-update(lm4,~.-Enzyme:CaStock:sqrtEnzymeConc)
summary(lm5)
lm6<-step(lm5)
summary(lm6)
lm9<-update(lm6,~.-DetStock:CaStock:sqrtEnzymeConc)
summary(lm9)
lm10<-step(lm9)
summary(lm10)
lm11<-update(lm10,~.- Enzyme:CaStock)
summary(lm11)
lm12<-step(lm11)
lm13<-update(lm12,~.- CaStock:sqrtEnzymeConc )
summary(lm13)
lm14<-step(lm13)
lm15<-update(lm14,~.-Enzyme:DetStock)
summary(lm15)
lm16<-step(lm15)
lm17<-update(lm16,~.-Enzyme:sqrtEnzymeConc)
summary(lm17)
anova(lm17)

##################### 3D graph###############
coef(lm13)
SPR$EnzymeConc<-as.factor(SPR$EnzymeConc)
pred.data<-expand.grid(Enzyme=levels(SPR$Enzyme),
                       EnzymeConc=levels(SPR$EnzymeConc),DetStock=levels(SPR$DetStock))
pred.data$Response<-predict(lm13,newdata=pred.data)
head(pred.data)
library(lattice) ## A flexible graphics package including many nice functions
wireframe(Response~Enzyme * EnzymeConc | DetStock,
          pred.data,scales = list(arrows = FALSE))
wireframe(Response~Enzyme * EnzymeConc | DetStock,
          pred.data,groups=as.numeric(pred.data$DetStock),
          scales = list(arrows = FALSE), drape=TRUE,colorkey=TRUE)

wireframe(Response~Enzyme * DetStock|EnzymeConc,
          pred.data,groups=as.numeric(pred.data$EnzymeConc),
          scales = list(arrows = FALSE), drape=TRUE,colorkey=TRUE)

wireframe(Response~ DetStock*EnzymeConc |Enzyme ,
          pred.data,groups=as.numeric(pred.data$Enzyme),
          scales = list(arrows = FALSE), drape=TRUE,colorkey=TRUE)

# SPR$EnzymeConc<-as.numeric(SPR$EnzymeConc)
SPR$sqrtEnzymeConc<-sqrt(SPR$EnzymeConc)
```

```r
##############2D graph##############################################
model<-lm(formula = sqrtResponse ~ (Enzyme + DetStock + sqrtEnzymeConc)^3,
          data = SPR)
summary(model)
## uncertianty: Response~ Enzyme+EnzymeConc|Det+
#### mat.height=seq(151,185,2)
pred.en<-expand.grid(Enzyme=levels(SPR$Enzyme),
                     EnzymeConc=levels(SPR$EnzymeConc),DetStock="Det0")
pred<-predict(lm17,newdata=pred.en,interval="c")^2
summary(pred)
par(mfrow=c(1,1)) #c(0,2.5,7.5,15)
matplot(c(0,2.5,7.5,15)[as.numeric(pred.en$EnzymeConc)],pred,type="n",
        ylim=c(0,1800),xlab="EnzymeConc", ylab="Response Prediction",
        main="Prediction of response")
matlines(c(0,2.5,7.5,15), matrix(pred[,1],nrow=4,byrow=TRUE),
         col=2:6,lty=rep(c(1,2,2),each=5),lwd=2)
legend("topleft",legend=levels(SPR$Enzyme),lty=1,col=2:6,title = "Det+")

# Barplot: response~Enzyme+EnzymeConc |Det+##
library(Hmisc)
barplot(matrix(pred[,1],nrow=5,dimnames=dn),beside=TRUE,col=2:6,
        legend.text=TRUE,args.legend=list(x="topleft"),
        main="95% confidence interval for Det+",
        ylab="Response",xlab="Enzyme concentration",ylim=c(0,max(pred)))
errbar((1:23)[-c(1:3)*6]+0.5,y=pred[,1],
       yminus=pred[,2],yplus=pred[,3],add=TRUE,pch=NA)

#############################################################
## uncertianty: Response~ Enzyme+EnzymeConc|Det0 ####
pred.en<-expand.grid(Enzyme=levels(SPR$Enzyme),
                     EnzymeConc=levels(SPR$EnzymeConc),DetStock="Det0")
pred<-predict(lm13,newdata=pred.en,interval="c")^2
par(mfrow=c(1,1))
matplot(c(0,2.5,7.5,15)[as.numeric(pred.en$EnzymeConc)],
        pred,type="n",ylim=range(pred))
# matlines(c(0,2.5,7.5,15), cbind(matrix(pred[,1],nrow=4,byrow=TRUE),
# matrix(pred[,2],nrow=4,byrow=TRUE),matrix(pred[,3],nrow=4,byrow=TRUE)),
# col=2:6,lty=rep(c(1,2,2),each=5),lwd=2)
# legend("topleft",legend=levels(SPR$Enzyme),lty=1,col=2:6)

matplot(c(0,2.5,7.5,15)[as.numeric(pred.en$EnzymeConc)],
        pred,type="n",ylim=range(pred),
        xlab="EnzymeConc", ylab="Response Prediction",
        main="Prediction of response Det0")
matlines(c(0,2.5,7.5,15), matrix(pred[,1],nrow=4,byrow=TRUE),
         col=2:6,lty=rep(c(2,2,2),each=5),lwd=2)
legend("topright",legend=levels(SPR$Enzyme),
       lty=2,col=2:6,title = "Det0")

# Barplot: response~Enzyme+EnzymeConc |Det0
dn<-list(Enzyme=levels(SPR$Enzyme),
         EnzymeConc=levels(SPR$EnzymeConc)) #Dimnames
barplot(matrix(pred[,3],nrow=5,dimnames=dn),
        beside=TRUE,col=2:6,legend.text=TRUE,
        args.legend=list(x="topleft"),main="95% confidence interval for Det0",
        ylab="Response",xlab="Enzyme concentration")
barplot(matrix(pred[,1],nrow=5,dimnames=dn),beside=TRUE,col=2:6,add=TRUE)
barplot(matrix(pred[,2],nrow=5,dimnames=dn),beside=TRUE,col=0,add=TRUE)
```

```
#Projekt 1:
#load data:
rm(list = ls(all = TRUE))
par(mfrow=c(1,1))
data <- read.table("SPR.txt",header=T)

par(mfrow=c(2,3))
plot(Response~En)
#look at linear model for data:
par(mfrow=c(2,2))
lm1 <- lm(Response~(Enzyme*EnzymeConc*DetStock*CaStock),data)
summary(lm1)
plot(lm1)
# not so good. Need transformation. but which one?
##################################################
#transformations:
data$Log.response <- log(data$Response)
data$Sqrt.response <- sqrt(data$Response)
data$Sq.response <- data$Response^2

#plot for reponse no transformation:
x1 <- data$Response
h1<-hist(x1, breaks=8, col="red", xlab="Response",
        main="Histogram with Normal Curve")
xfit1<-seq(min(x1),max(x1),length=40)
yfit1<-dnorm(xfit1,mean=mean(x1),sd=sd(x1))
yfit1 <- yfit1*diff(h1$mids[1:2])*length(x1)
lines(xfit1, yfit1, col="blue", lwd=2)
#plot for log response:
x2 <- data$Log.response
```

```
h2<-hist(x2, breaks=8, col="red", xlab=" Log Response",
         main="Histogram with Normal Curve")
xfit2<-seq(min(x2),max(x2),length=40)
yfit2<-dnorm(xfit2,mean=mean(x2),sd=sd(x2))
yfit2 <- yfit2*diff(h2$mids[1:2])*length(x2)
lines(xfit2, yfit2, col="blue", lwd=2)
#plot for sqrt response:
x3 <- data$Sqrt.response
h3<-hist(x3, breaks=8, col="red", xlab="Sqrt Response",
         main="Histogram with Normal Curve",ylim=c(0,35))
xfit3<-seq(min(x3),max(x3),length=40)
yfit3<-dnorm(xfit3,mean=mean(x3),sd=sd(x3))
yfit3 <- yfit3*diff(h3$mids[1:2])*length(x3)
lines(xfit3, yfit3, col="blue", lwd=2)

# plot for sq response:
x4 <- data$Sq.response
h4<-hist(x4, breaks=15, col="red", xlab="Square Response",
         main="Histogram with Normal Curve")
xfit4<-seq(min(x4),max(x4),length=40)
yfit4<-dnorm(xfit4,mean=mean(x4),sd=sd(x4))
yfit4 <- yfit4*diff(h4$mids[1:2])*length(x4)
lines(xfit4, yfit4, col="blue", lwd=2)

#using sqrt transformation but it seems
#as log transformation might also could be possible.
```

```
rm(list = ls(all = TRUE))
par(mfrow=c(1,1))
data<-read.table("SPR.txt",header=TRUE)


#include additional rows in data
data$Sqrt.response <- sqrt(data$Response)
data$Sqrt.EnzymeConc <- sqrt(data$EnzymeConc)
data$merged_Det_Ca<-paste(data$DetStock,data$CaStock)



xyplot(Sqrt.response~Sqrt.EnzymeConc|Enzyme,
       data=data, groups=merged_Det_Ca, type = c("p", "r"),
       auto.key=list(space="top", columns=4, cex.title=1,
      points=FALSE, col=c("black", "green", "red", "blue")),
      pch=19, col=c("black", "green", "red", "blue"))
```

```
#include lattice
library(lattice)

#read the data
SPR<-read.table("SPR.txt",header=TRUE)

SPR<-SPR[SPR$Enzyme=="E",]
SPR<-SPR[SPR$DetStock=="Det+",]
#SPR<-SPR[SPR$EnzymeConc == 2.5,]

dat<-SPR

##sqrt variables####
dat$sqrtResponse<-sqrt(dat$Response)
dat$sqrtEnzymeConc<-sqrt(dat$EnzymeConc)

#order because of something
dat = dat[order(dat$EnzymeConc),]

plot(sqrtResponse~sqrtEnzymeConc,dat,ylim=c(-10,50),
     pch=19,main="Enzime E, Det+",
     col=c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1))

lmA = lm(sqrtResponse~sqrtEnzymeConc,dat)
summary(lmA)

pred<-predict(lmA,interval="predict",newdata=data.frame(dat))
conf<-predict(lmA,interval="conf",newdata=data.frame(dat))

matlines(dat$sqrtEnzymeConc,pred,col=3 ,lwd=2)
```

```
# Lattice Examples
library(lattice)

data<-read.table("SPR.txt",header=TRUE)

data$merged_Det_Ca<-paste(data$DetStock,data$CaStock)
#View(data)


# xyplot(Response~EnzymeConc|Enzyme, data=data, groups=merged_Det_Ca,
#       auto.key=TRUE, pch=1, col=c("black", "green", "red", "blue"))


xyplot(Response~EnzymeConc|Enzyme, data=data, groups=merged_Det_Ca,
       auto.key=list(space="top", columns=4, cex.title=1,
      points=FALSE, col=c("black", "green", "red", "blue")),
      pch=19, col=c("black", "green", "red", "blue"))
```