DiaryProduction

```r
setwd("~/Desktop/PiE2")
library(car)
```

```
## Loading required package: carData
```

```r
library(tables)
```

```
## Loading required package: Hmisc
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```r
library(RcmdrMisc)
```

```
## Loading required package: sandwich
```

```
##
## Attaching package: 'RcmdrMisc'
```

```
## The following object is masked from 'package:Hmisc':
##
##     Dotplot
```

```r
library(car)
#library(Hmisc)
#library(lattice)
#library(survival)
#library(Formula)
#library(colorspace)
#library(ggplot2)
```

We first read the dataset

```r
diaryprod <- read.csv2("./Dades/diaryp.csv")
head(diaryprod)
```

```
##   Days PROD
## 1    6 28.5
## 2   13 34.8
## 3   20 36.8
## 4   27 39.4
## 5   34 35.9
## 6   41 40.8
```

```r
summary(diaryprod)
```

```
##       Days            PROD
##  Min.   :  6.0   Min.   :17.80
##  1st Qu.: 69.0   1st Qu.:25.20
##  Median :139.0   Median :29.50
##  Mean   :138.8   Mean   :30.15
##  3rd Qu.:202.0   3rd Qu.:35.70
##  Max.   :272.0   Max.   :40.80
```

```r
dim(diaryprod)
```

```
## [1] 37  2
```

The dataset has 37 rows and two columns. The first colimn corresponds to the number of days since the cow gave birth and, the second one to the milk production.

We define the variable log of the production

Next we define the logarithm of the days and we append this column to the dataset.

```r
diaryprod$lDays<-log(diaryprod$Days)
dim(diaryprod)
```

```
## [1] 37  3
```

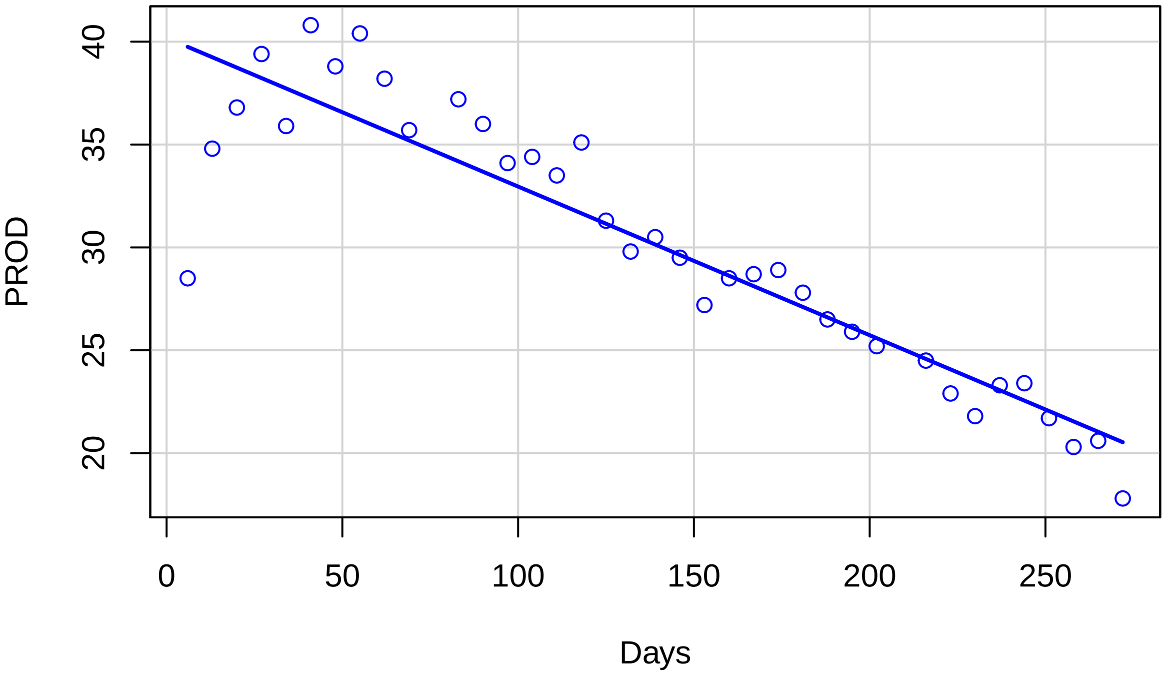
Descriptive Statistics

We do a summary of the data and we plot the poduction as a function fo the Days

```r
summary(diaryprod)
```

```
##       Days            PROD           lDays
##  Min.   :  6.0   Min.   :17.80   Min.   :1.792
##  1st Qu.: 69.0   1st Qu.:25.20   1st Qu.:4.234
##  Median :139.0   Median :29.50   Median :4.934
##  Mean   :138.8   Mean   :30.15   Mean   :4.655
##  3rd Qu.:202.0   3rd Qu.:35.70   3rd Qu.:5.308
##  Max.   :272.0   Max.   :40.80   Max.   :5.606
```

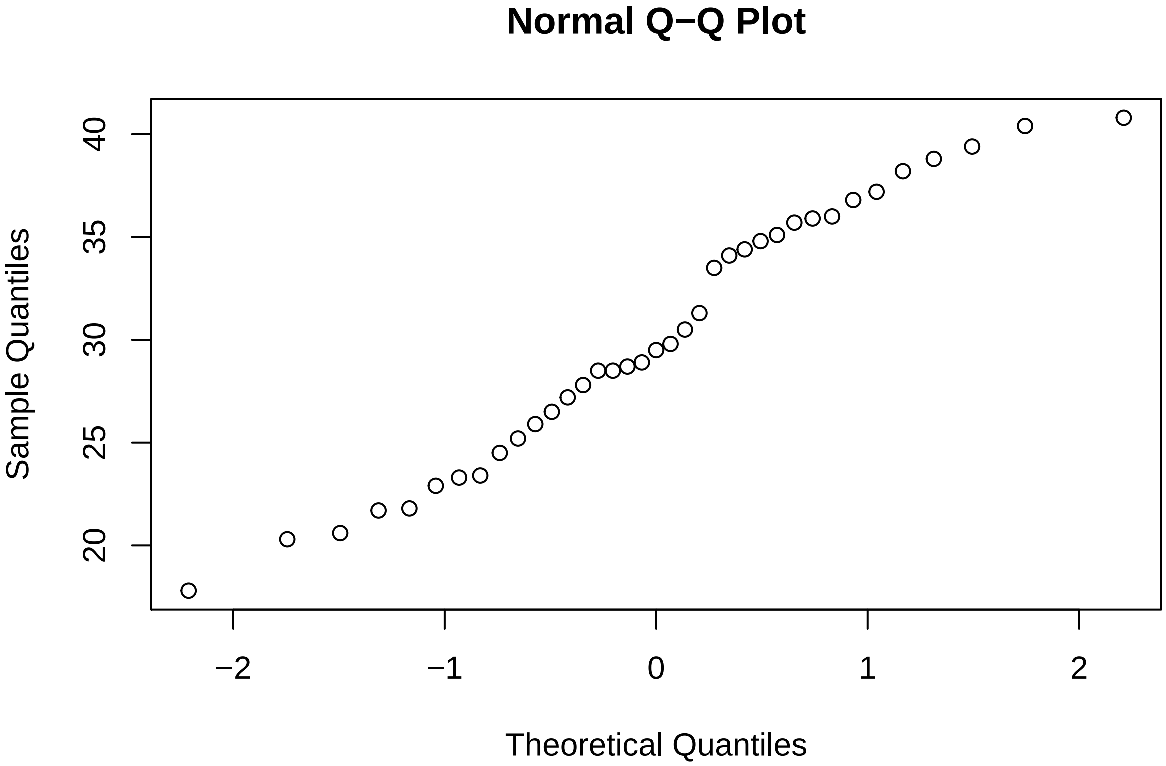
```r
sp(PROD~Days, smooth=F, boxplots=F, data=diaryprod)
```



We clearly see that in the long-term the production decreases with the days as it has sense to be. Nevertheless, an straight line is not able to adapt the concavity observed at the begining. It is reasonable that the think milk production will be small the first days and that once the cow is activated to producte the milk, the production will increase for a certain time to laterly decrease again.

In what follows we wil perform the *qqplot* of the variable PROD to conclude that it is not clear to follow a normal distribution, specially in the tails

```
qqnorm(diaryprod$PROD)
```



First model: Log-normal for the response and days and log(days) as explanatory

We first will fit the model proposed in the statement assuming a lineat model for the log of the production. This is equivalent to assume a a log-Normal distribution for the variable PROD. The fitted values will be compared to the observed ones graphically.

```
model1<-lm(log(PROD)~Days+I(log(Days)), diaryprod)
summary(model1)
```

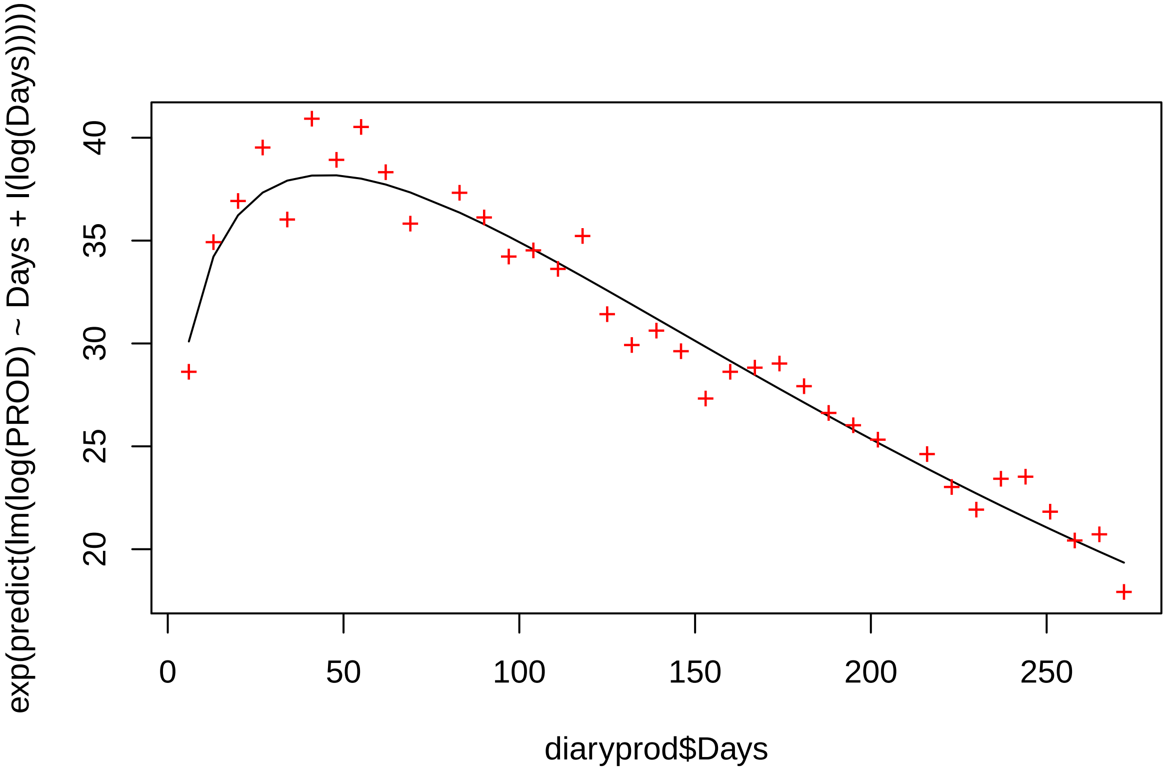
```
##
## Call:
## lm(formula = log(PROD) ~ Days + I(log(Days)), data = diaryprod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.09240 -0.03160  0.00324  0.02451  0.08269
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.0593493  0.0638354   47.93  < 2e-16 ***
## Days          -0.0046452  0.0002177  -21.34  < 2e-16 ***
## I(log(Days))   0.2081302  0.0192567   10.81 1.54e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04382 on 34 degrees of freedom
## Multiple R-squared:  0.9628, Adjusted R-squared:  0.9606
## F-statistic: 439.8 on 2 and 34 DF,  p-value: < 2.2e-16
```

We see that both variables: *days* and *log*(*days*) are significant. Both variables jointly explain 96% of the variability in the production.

Taking into account that the production mean is 30 we can say that the standard error estimation is very small.

Next plot contains the exponential of the predicted values compared to the observed values.

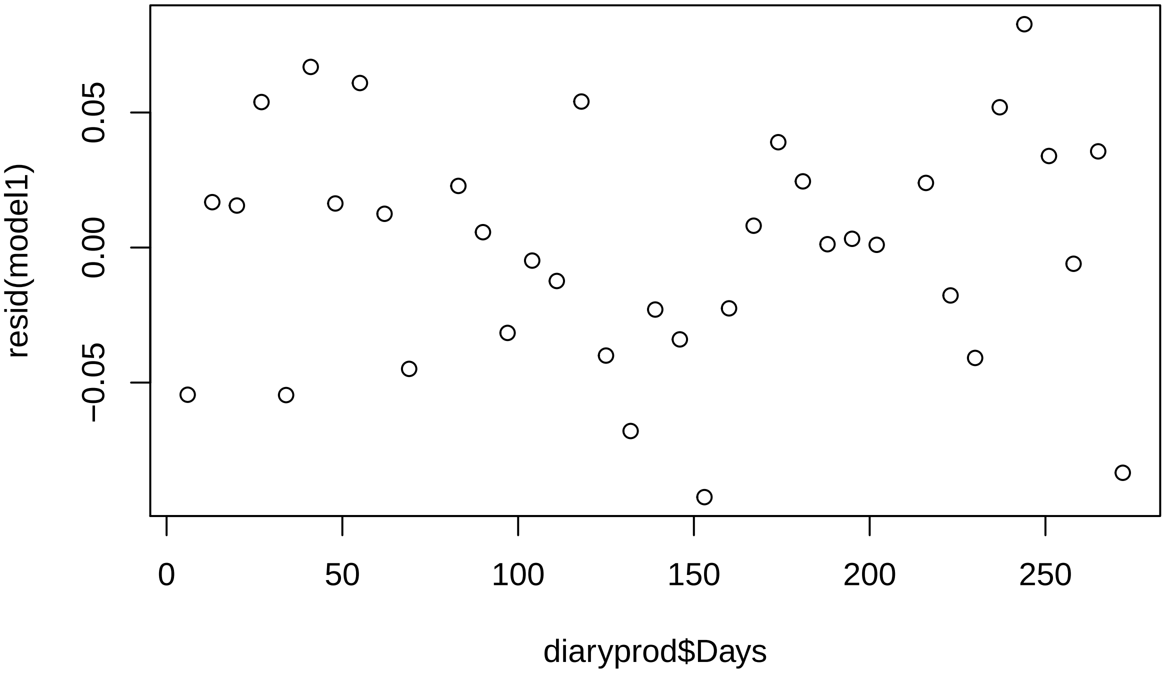
```
with(diaryprod,plot(diaryprod$Days,exp(predict(lm(log(PROD)~Days+I(log(Days))))),ty="l",ylim=c(min(PROD
with(diaryprod,points(Days,PROD,col="red",pch="+"))
```



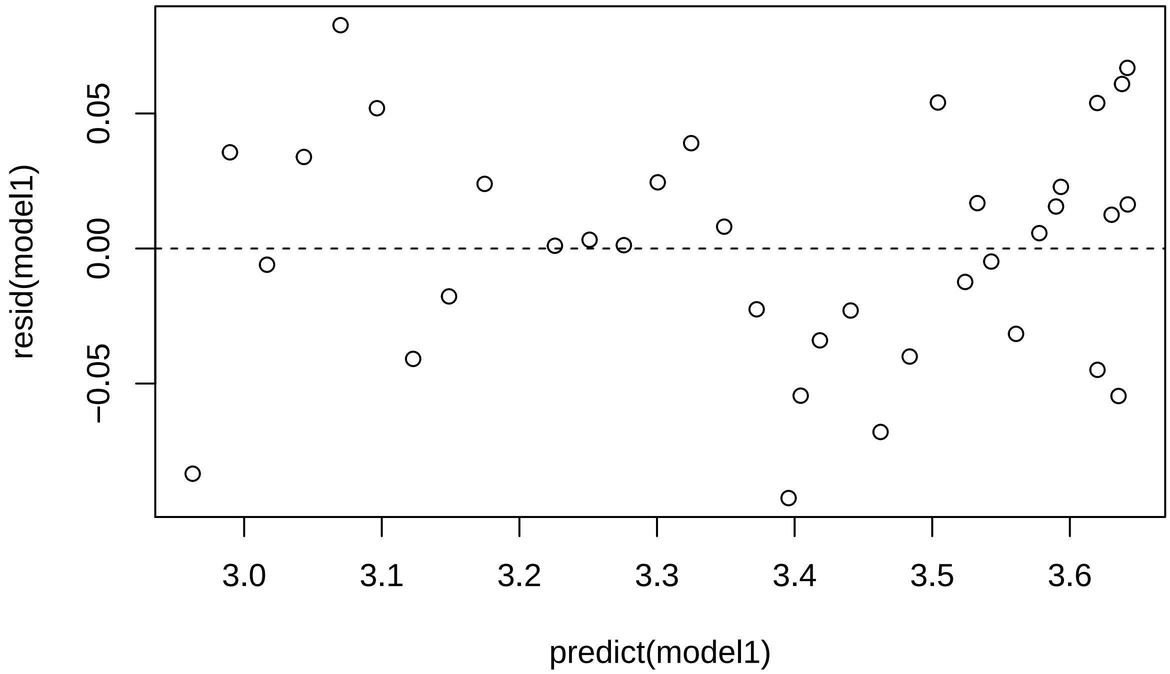
We clearly see that the function that relates the expectation of the production and the days variable is the appropiate one, since it is able to adapt the concavity at the begining maintaining the linearity in the tail.

Let us analyze the residuals of this linear model.

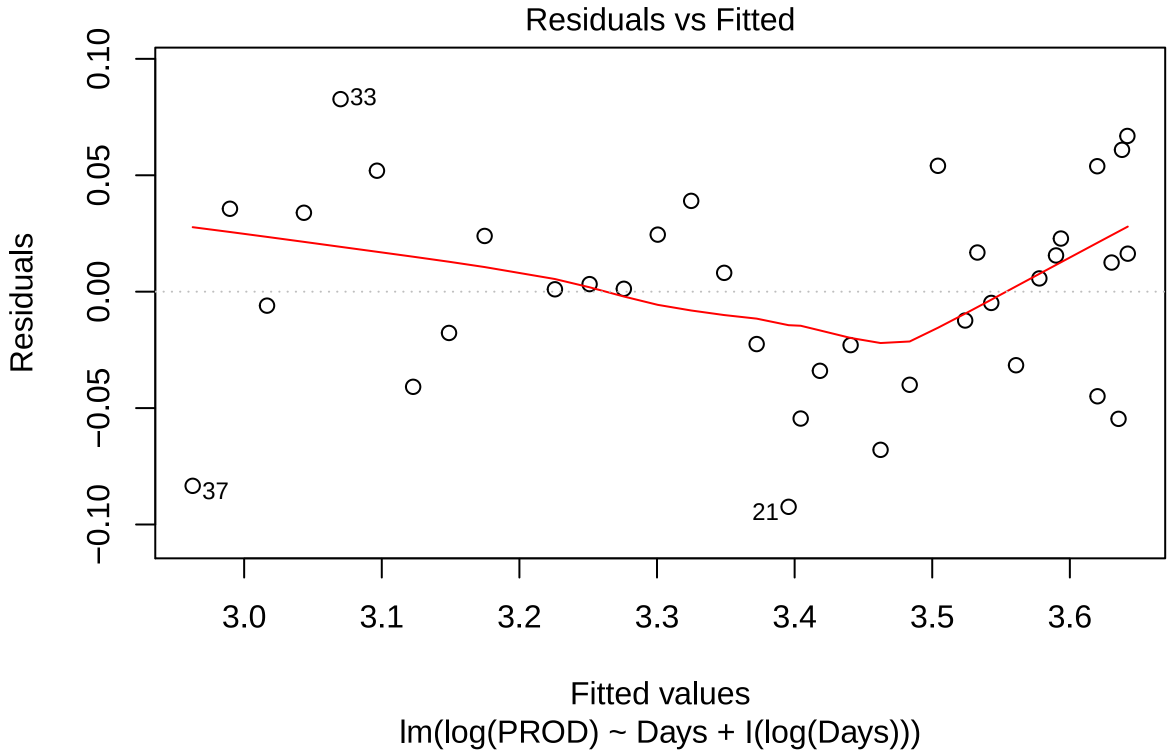
```
plot(diaryprod$Days,resid(model1))
```

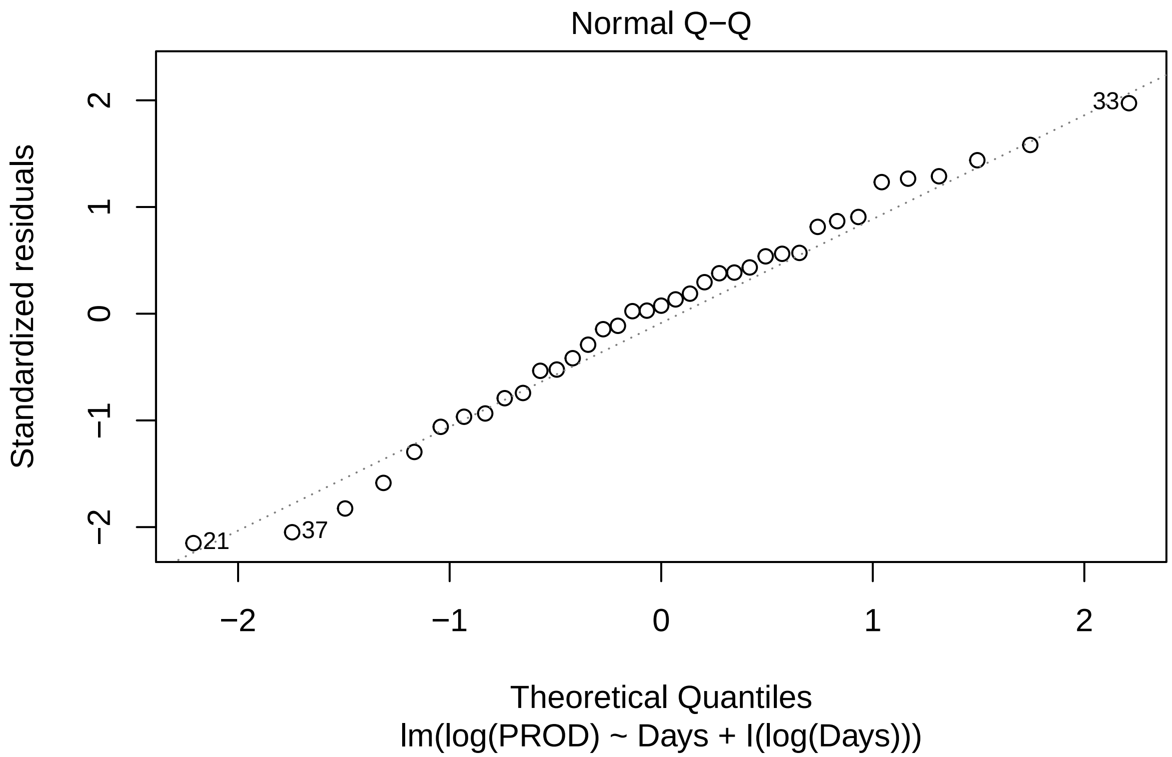


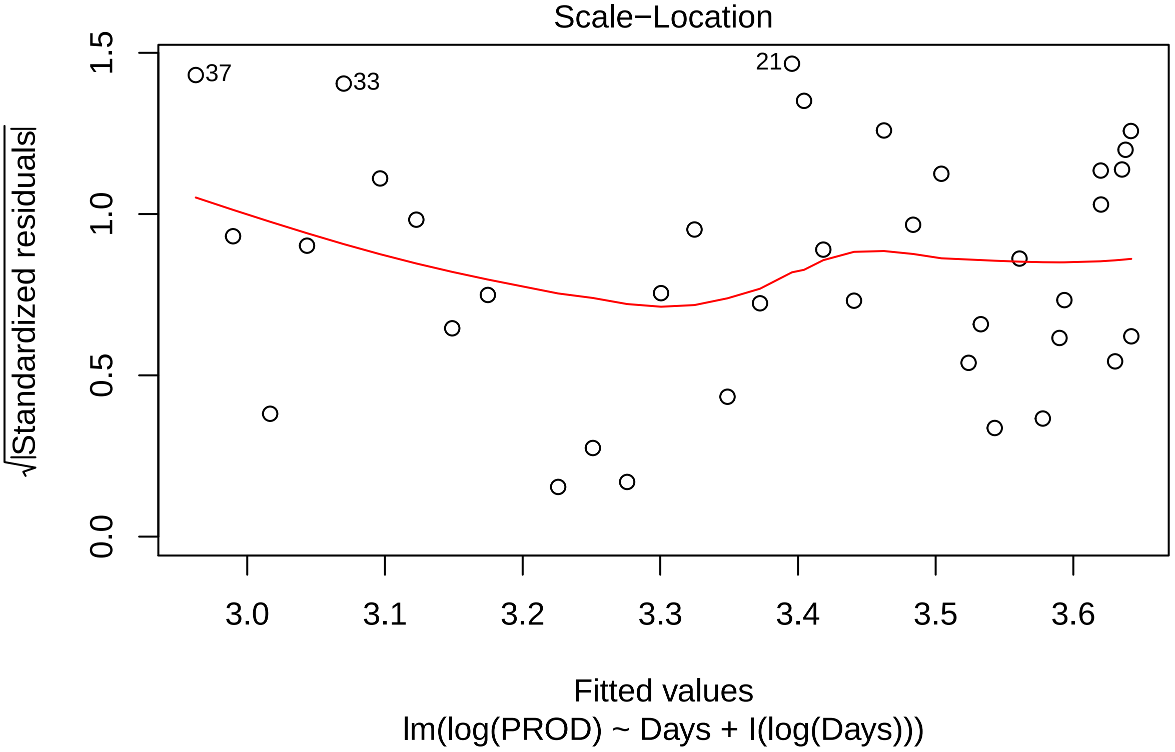
```
plot(predict(model1),resid(model1))
abline(h=0,lty=2)
```

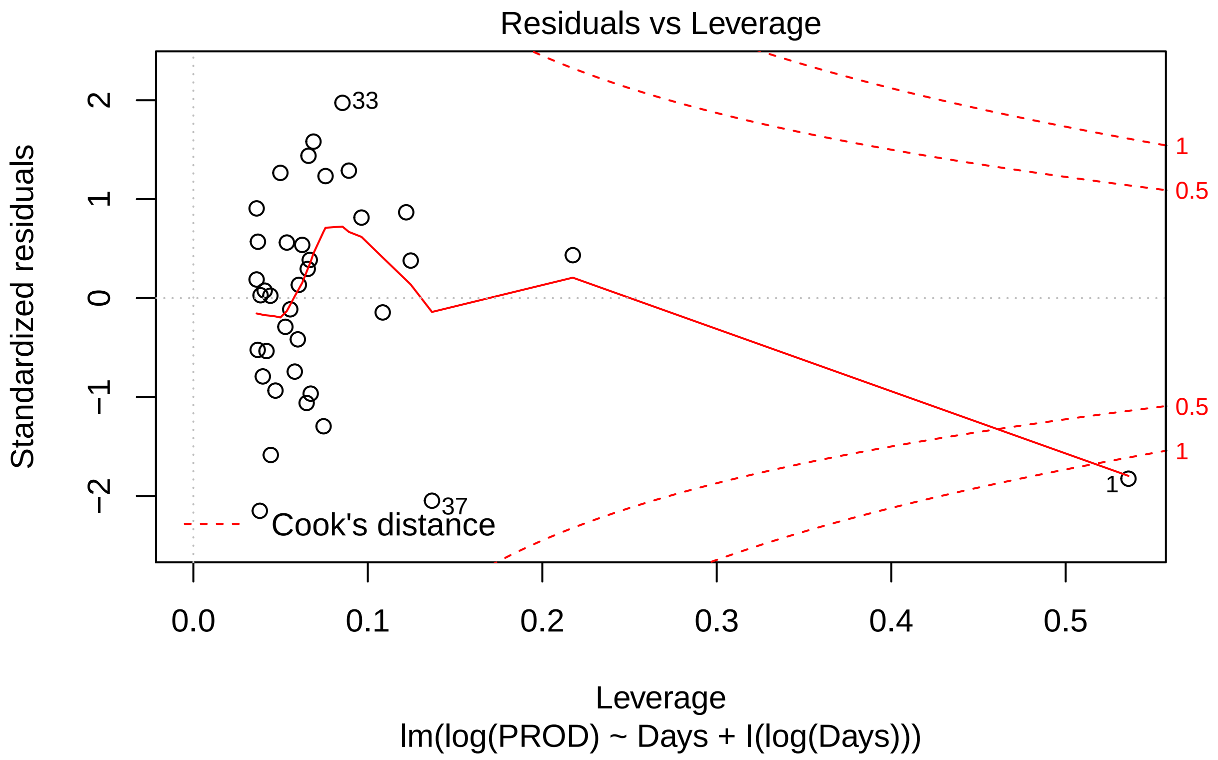

resid(model1)
0.05
0.00
−0.05
3.0
3.1
3.2
3.3
3.4
3.5
3.6
predict(model1)

```
plot(model1)
```


Residuals vs Fitted
Residuals
0.10
0.05
0.00
−0.05
−0.10
33
37
21
3.0
3.1
3.2
3.3
3.4
3.5
3.6
Fitted values
lm(log(PROD) ~ Days + I(log(Days)))


Normal Q−Q
Standardized residuals
−2
−1
0
1
2
21
37
33
−2
−1
0
1
2
Theoretical Quantiles
lm(log(PROD) ~ Days + I(log(Days)))


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
37
33
21
3.0
3.1
3.2
3.3
3.4
3.5
3.6
Fitted values
lm(log(PROD) ~ Days + I(log(Days)))



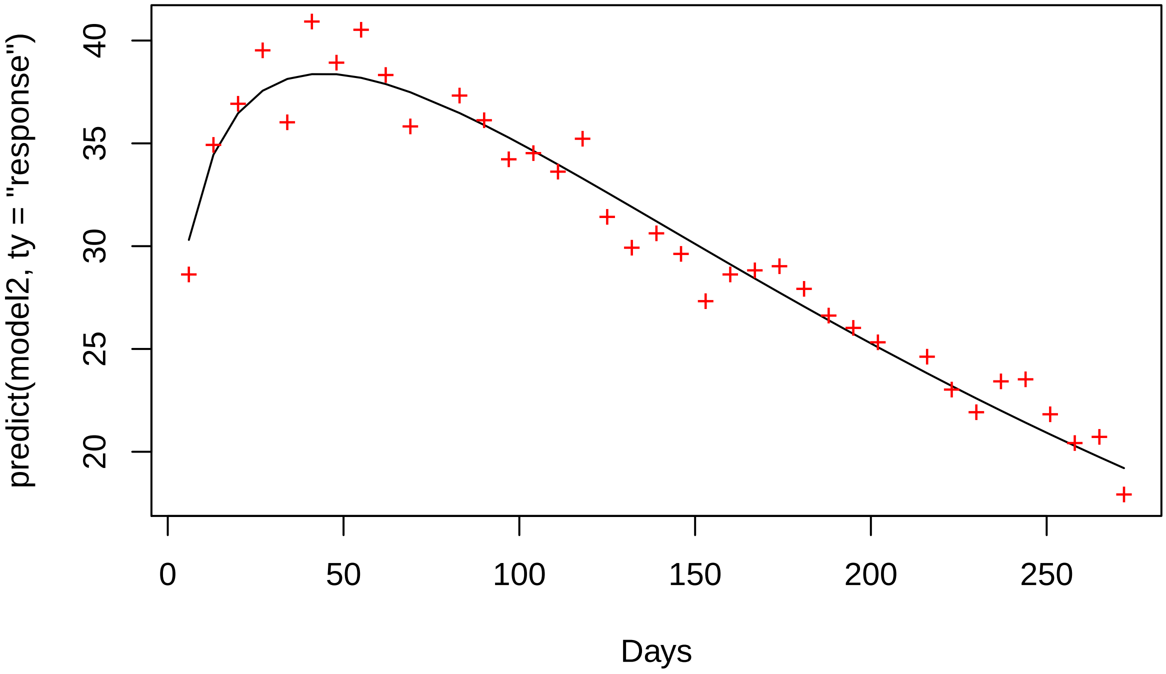
With the residual analysis we see an slightly pattern between residuals vs fitted values. We also see that the left tail of the residuals is not on the straight line which corresponds to the Normal. In order to see if we can find a better model we next fit two new models, that require the Generalized Linear Modelling theory.

Second model: Normal for the response and days and log(days) as explanatory

```
model2<-glm(PROD~Days+I(log(Days)),family=gaussian(link="log"),data=diaryprod)
summary(model2)
```

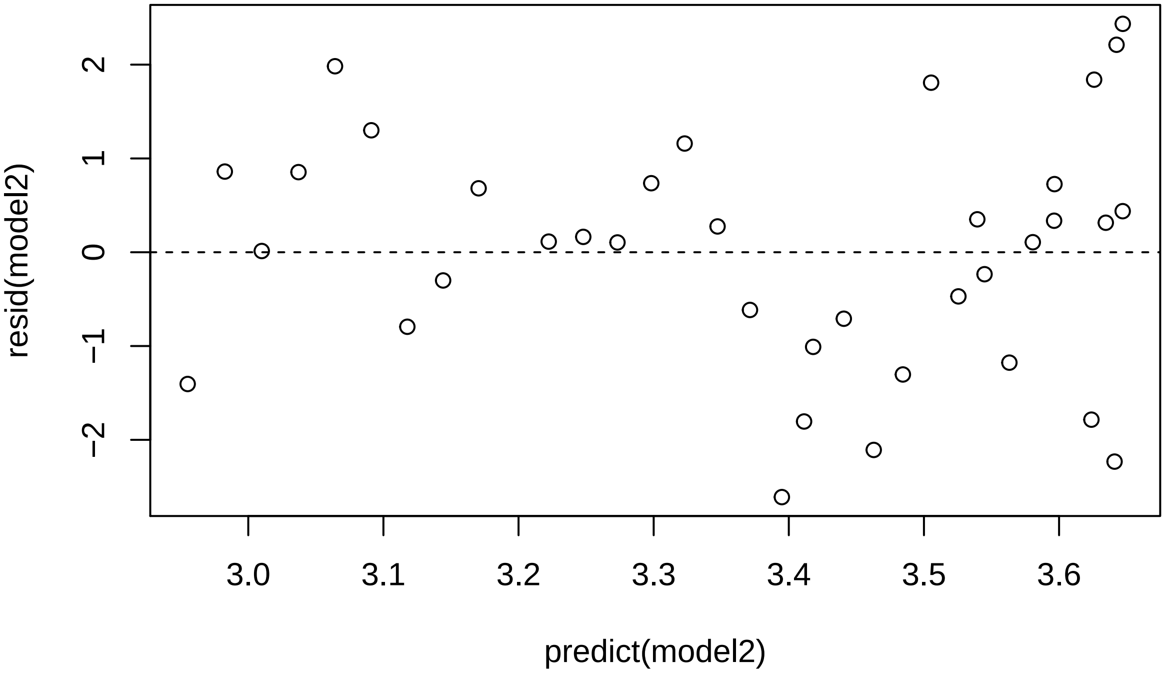
```
##
## Call:
## glm(formula = PROD ~ Days + I(log(Days)), family = gaussian(link = "log"),
##     data = diaryprod)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6105  -0.7946   0.1135   0.7359   2.4350
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.0661759  0.0614394   49.91  < 2e-16 ***
## Days          -0.0047024  0.0002331  -20.18  < 2e-16 ***
## I(log(Days))   0.2083624  0.0189088   11.02 9.17e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 1.727785)
##
##     Null deviance: 1466.752  on 36  degrees of freedom
## Residual deviance:   58.745  on 34  degrees of freedom
## AIC: 130.11
##
## Number of Fisher Scoring iterations: 4
```

```
with(diaryprod,plot(Days,predict(model2,ty="response"),ty="l",ylim=c(min(PROD),max(PROD))))
with(diaryprod,points(Days,PROD,col="red",pch="+"))
```

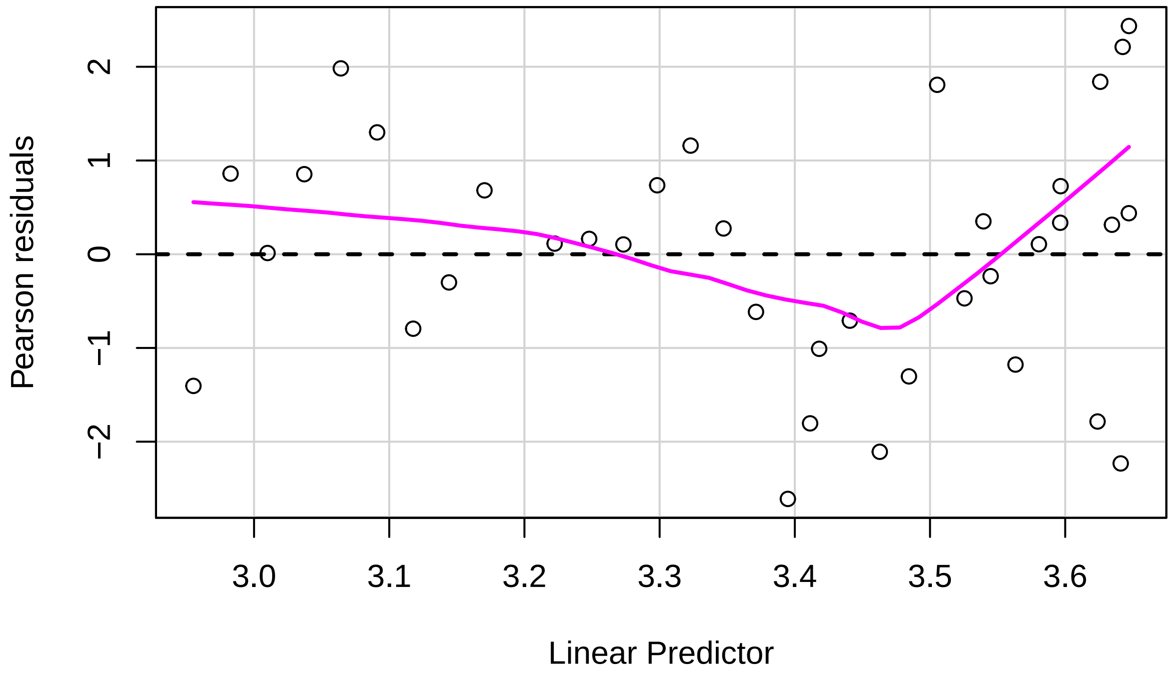


We do the residual analysis

```
plot(predict(model2),resid(model2))
abline(h=0,lty=2)
```



```r
residualPlot(model2)
```



We do not find very clear differences with the previous model. This is because the log-normal and the normal distribution just differ in the tail and they are quite similar.

The two models fitted consider a constant variance. Next, we are going to assume that the variance changes as a quadratic funtion of the mean. This will be done by assuming that the response distribution is the Gamma distribution.

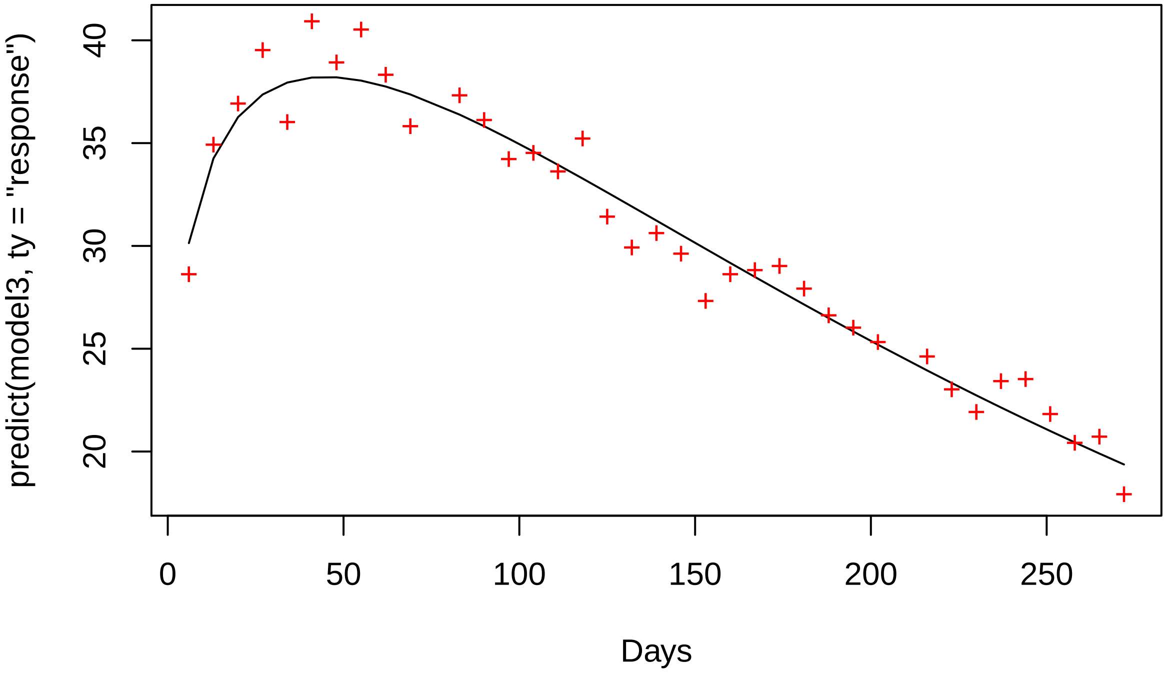
Third model: Gamma for the response and days and log(days) as explanatory

```r
model3<-glm(PROD~Days+I(log(Days)),family=Gamma(link="log"),data=diaryprod)
summary(model3)
```

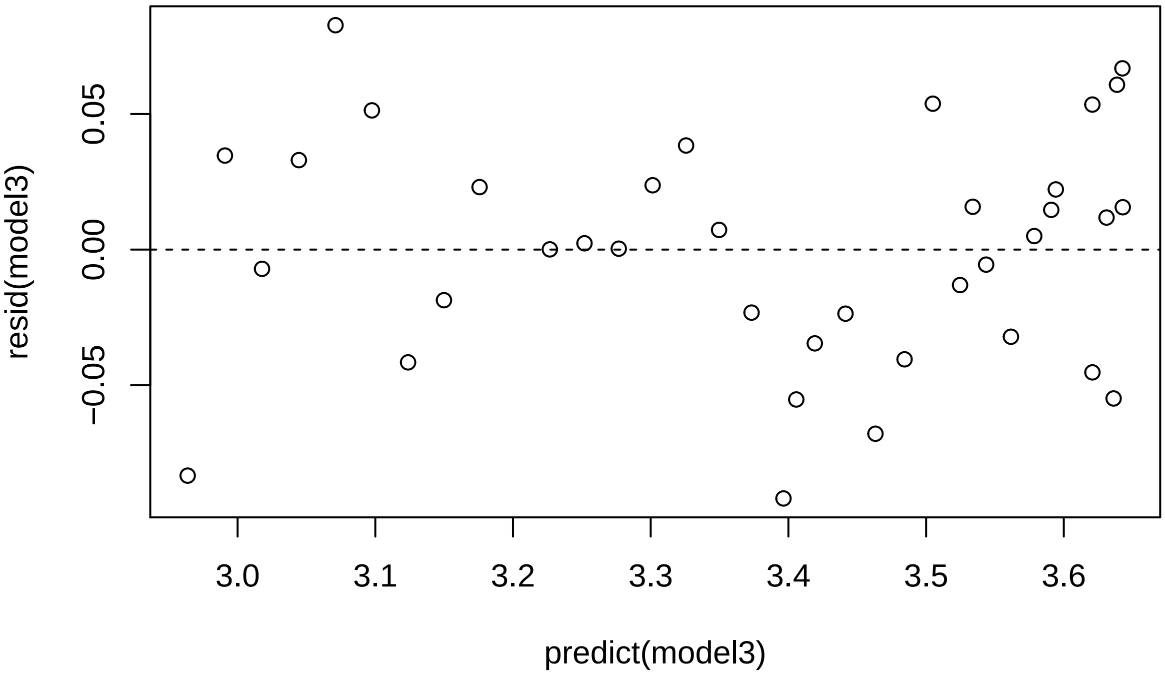
```
##
## Call:
## glm(formula = PROD ~ Days + I(log(Days)), family = Gamma(link = "log"),
##     data = diaryprod)
##
## Deviance Residuals:
##       Min         1Q     Median         3Q        Max
## -0.091783  -0.032143   0.002326   0.023722   0.082757
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.0612870  0.0635561   48.17  < 2e-16 ***
## Days         -0.0046406  0.0002168  -21.41  < 2e-16 ***
## I(log(Days))  0.2077641  0.0191725   10.84 1.43e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.001903793)
```

```
##
##     Null deviance: 1.701490  on 36  degrees of freedom
## Residual deviance: 0.065097  on 34  degrees of freedom
## AIC: 128.7
##
## Number of Fisher Scoring iterations: 3
```

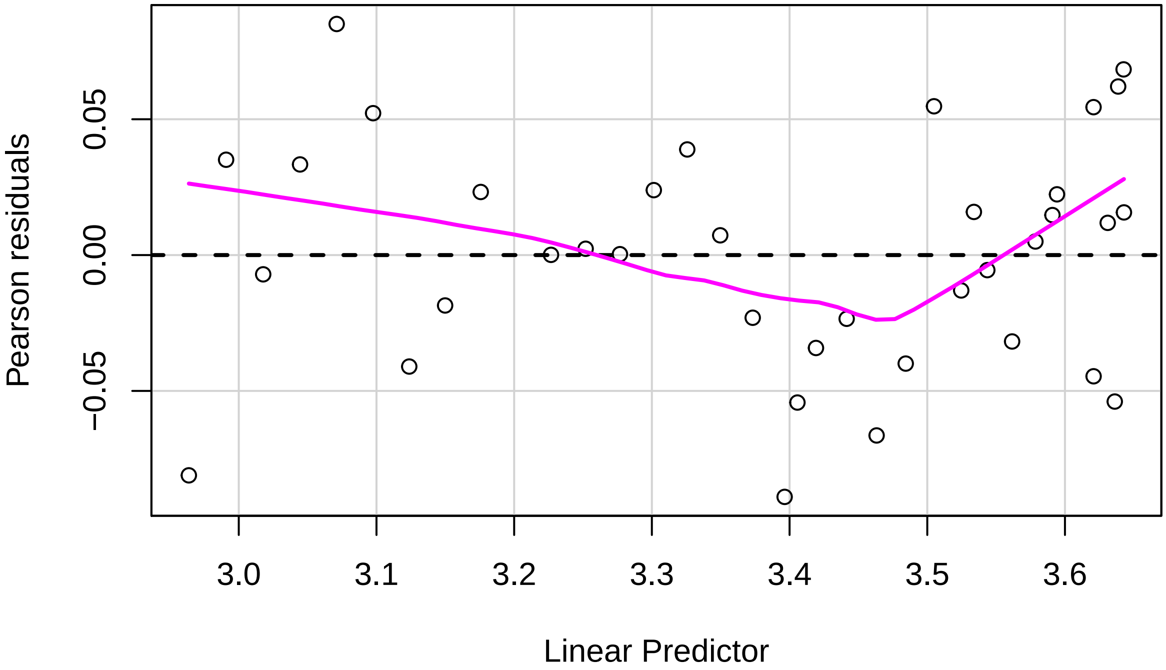
```
with(diaryprod,plot(Days,predict(model3,ty="response"),ty="l",ylim=c(min(PROD),max(PROD))))
with(diaryprod,points(Days,PROD,col="red",pch="+"))
```



```
plot(predict(model3),resid(model3))
abline(h=0,lty=2)
```



```
residualPlot(model3)
```

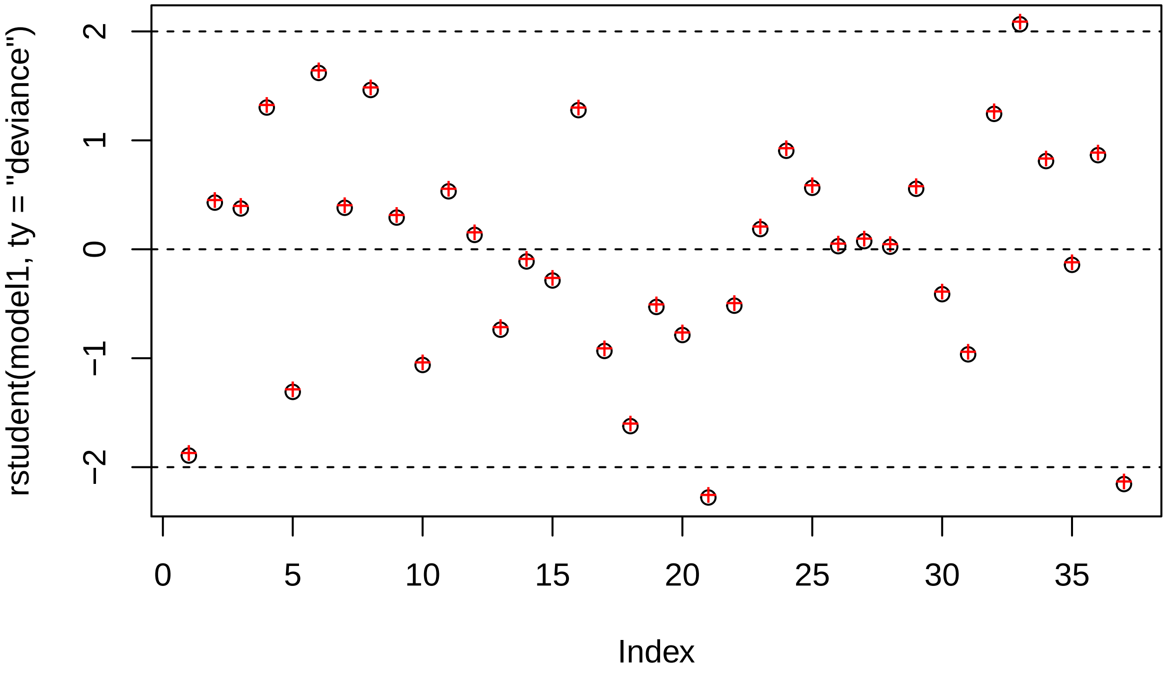


For this model we also see that both explanatory variables are significat and that the #Akaike Information Criteia# is smaller which means that this model is better than model two. The residual values are also smaller than in the second model.

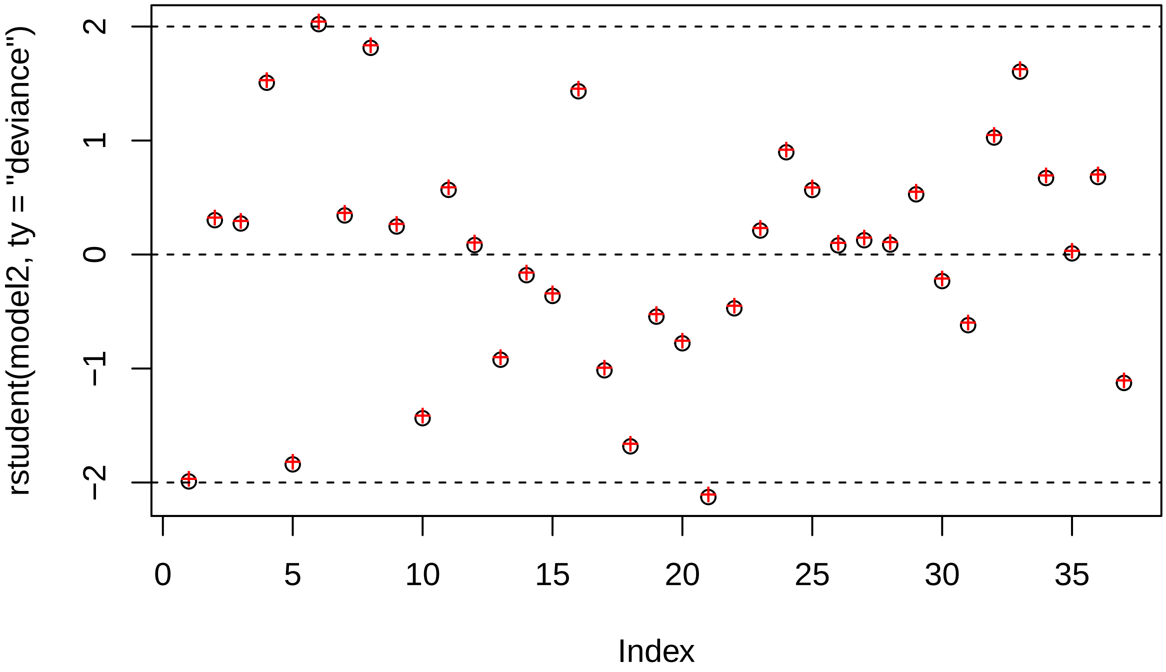
Comparive study of the three models cosidered

```
c(model1=summary(model1)$r.squared,model2=1-model2$dev/model2$null.dev,model3=1-model3$dev/model3$null.
##    model1    model2    model3
## 0.9627809 0.9599491 0.9617409
```

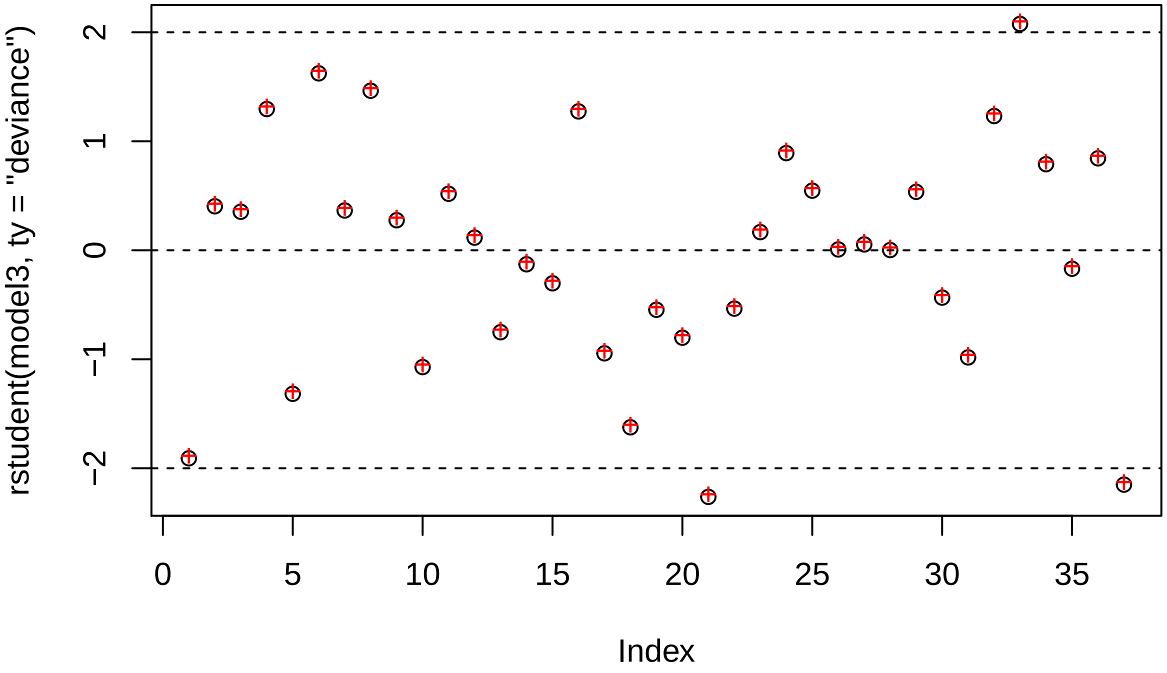
```
plot(rstudent(model1,ty="deviance"))
points(rstudent(model1,ty="pearson"),col="red",pch="+")
abline(h=c(-2,0,2),lty=2)
```



```
plot(rstudent(model2,ty="deviance"))
points(rstudent(model2,ty="pearson"),col="red",pch="+")
abline(h=c(-2,0,2),lty=2)
```



```
plot(rstudent(model3,ty="deviance"))
points(rstudent(model3,ty="pearson"),col="red",pch="+")
abline(h=c(-2,0,2),lty=2)
```


rstudent(model3, ty = "deviance")
Index