# Pate-tasting

*Jordi Valero and Marta Pérez-Casany*

*31 de octubre de 2018*

**ONE/TWO WAY ANOVA**

**We want to compare 5 different PATES.**

**Five response variables are recorded.**

```
library(car)
```

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

```
library(HH)
```

```
## Loading required package: lattice
## Loading required package: grid
## Loading required package: latticeExtra
## Loading required package: RColorBrewer
## Loading required package: multcomp
## Loading required package: mvtnorm
## Loading required package: survival
## Loading required package: TH.data
## Loading required package: MASS
##
## Attaching package: 'TH.data'
## The following object is masked from 'package:MASS':
##
##     geyser
## Loading required package: gridExtra
##
## Attaching package: 'HH'
## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
library(doBy)
library(car)
library(emmeans)
```

```
##
## Attaching package: 'emmeans'
```

```
## The following object is masked from 'package:HH':
##
##     as.glht
## The following object is masked from 'package:multcomp':
##
##     cld
```

```
library(tables)
```

```
## Loading required package: Hmisc
## Loading required package: Formula
## Loading required package: ggplot2
##
## Attaching package: 'ggplot2'
## The following object is masked from 'package:latticeExtra':
##
##     layer
##
## Attaching package: 'Hmisc'
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

**Loading the data and printing the top part**

```
setwd("G:/PiE2/2018")
patedata<-read.csv2("./Dades/PATE.csv")
head(patedata)
```

```
##   per pate color smell text taste order
## 1   1  525     7     7    7     6     3
## 2   1  113     4     4    5     7     4
## 3   1  220     7     7    7     7     2
## 4   1  372     6     8    7     8     1
## 5   1  140     4     4    5     4     5
## 6   2  525     6     5    6     6     1
```

```
dim(patedata)
```

```
## [1] 40  7
```

**The data set has observations of 40 rows and 7 columns.**

**We perform descriptive statistics (EDA)**

**Variables: color, smell, texture and taste take values from zero to ten.**

**Variable: order takes values from one to five, where one means the prefered.**

```
summary(patedata)
```

```
##       per             pate         color         smell          text
##  Min.   :1.00   Min.   :113   Min.   :4.0   Min.   :3.0   Min.   :3.000
##  1st Qu.:2.75   1st Qu.:140   1st Qu.:5.0   1st Qu.:4.0   1st Qu.:5.000
##  Median :4.50   Median :220   Median :6.0   Median :5.5   Median :6.000
##  Mean   :4.50   Mean   :274   Mean   :5.8   Mean   :5.5   Mean   :6.075
##  3rd Qu.:6.25   3rd Qu.:372   3rd Qu.:7.0   3rd Qu.:7.0   3rd Qu.:7.000
##  Max.   :8.00   Max.   :525   Max.   :8.0   Max.   :8.0   Max.   :9.000
##      taste          order
##  Min.   :3.00   Min.   :1
##  1st Qu.:5.00   1st Qu.:2
##  Median :6.00   Median :3
##  Mean   :5.95   Mean   :3
##  3rd Qu.:7.00   3rd Qu.:4
##  Max.   :8.00   Max.   :5
```

**First let us transform variables PER (person) and PATE as factors.**

```
patedata$per<-as.factor(patedata$per)
patedata$pate<-as.factor(patedata$pate)
```

names(patedata)

### ANALYSIS OF THE VARIABLE COLOR

# Descriptiva

**In what follows for each pate we compute the number of observations and the mean and sd of the color punctuation**

```
tabular(pate~color*((n=1)+mean+sd),patedata)
```

| | color | | |
|---|---|---|---|
| pate | n | mean | sd |
| 113 | 8 | 5.000 | 1.3093 |
| 140 | 8 | 4.750 | 1.3887 |
| 220 | 8 | 6.125 | 1.2464 |
| 372 | 8 | 6.500 | 0.9258 |
| 525 | 8 | 6.625 | 0.9161 |

In this table we see that:

1) pate 140 is the one with the smaller puctuation
2) pates 372 and 525 are the ones with the larger puntuation.
3) the variances do not seem to differ much, but the pates with the larger puctuation seem to have the smaller variance.

**In what follows we show the eight puctuations of color for each pate**

```
tabular(pate~mean*color*per,patedata)
```

| | | | | mean color per | | | | |
|---|---|---|---|---|---|---|---|---|
| pate | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 113 | 4 | 4 | 8 | 5 | 4 | 5 | 5 | 5 |
| 140 | 4 | 4 | 8 | 5 | 4 | 4 | 5 | 4 |
| 220 | 7 | 6 | 6 | 6 | 5 | 4 | 8 | 7 |
| 372 | 6 | 5 | 7 | 6 | 7 | 7 | 8 | 6 |
| 525 | 7 | 6 | 6 | 7 | 5 | 7 | 8 | 7 |

From this table it seems that persons 3 and 7 are the ones that tend to give greater puctuations.

## MODELIZATION

### we start comparing the pates by means of the ONE-WAY anova model

```
model1<-lm(color~pate,patedata)

summary(model1)
```

```
##
## Call:
## lm(formula = color ~ pate, data = patedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1250 -0.7500 -0.0625  0.4063  3.2500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.0000     0.4151  12.045 5.27e-14 ***
## pate140      -0.2500     0.5871  -0.426  0.67283
## pate220       1.1250     0.5871   1.916  0.06352 .
## pate372       1.5000     0.5871   2.555  0.01512 *
## pate525       1.6250     0.5871   2.768  0.00895 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.174 on 35 degrees of freedom
## Multiple R-squared:  0.3336, Adjusted R-squared:  0.2574
## F-statistic:  4.38 on 4 and 35 DF,  p-value: 0.005632
```

```
Anova(model1, ty=3)
```

```
## Anova Table (Type III tests)
##
## Response: color
##             Sum Sq Df  F value    Pr(>F)
## (Intercept) 200.00  1 145.0777 5.267e-14 ***
## pate         24.15  4   4.3795  0.005632 **
## Residuals    48.25 35
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(model1)
```

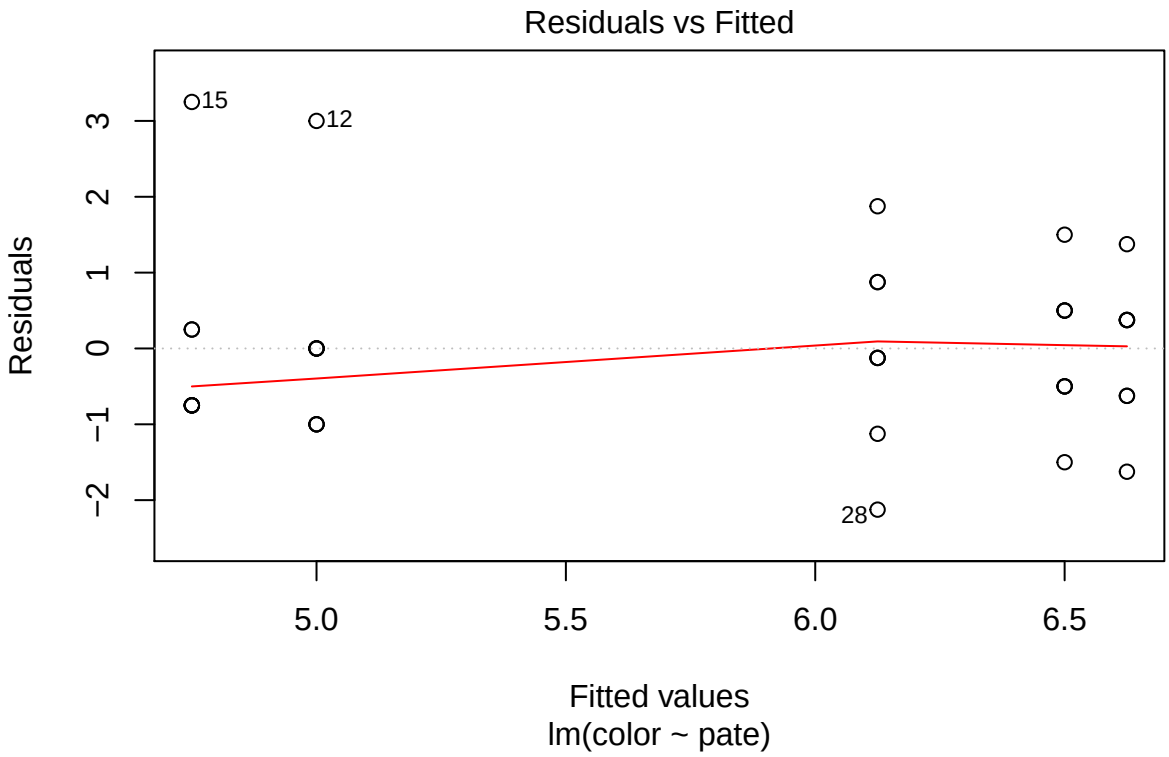
```
## Analysis of Variance Table
##
## Response: color
##           Df Sum Sq Mean Sq F value   Pr(>F)
## pate       4  24.15  6.0375  4.3795 0.005632 **
## Residuals 35  48.25  1.3786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

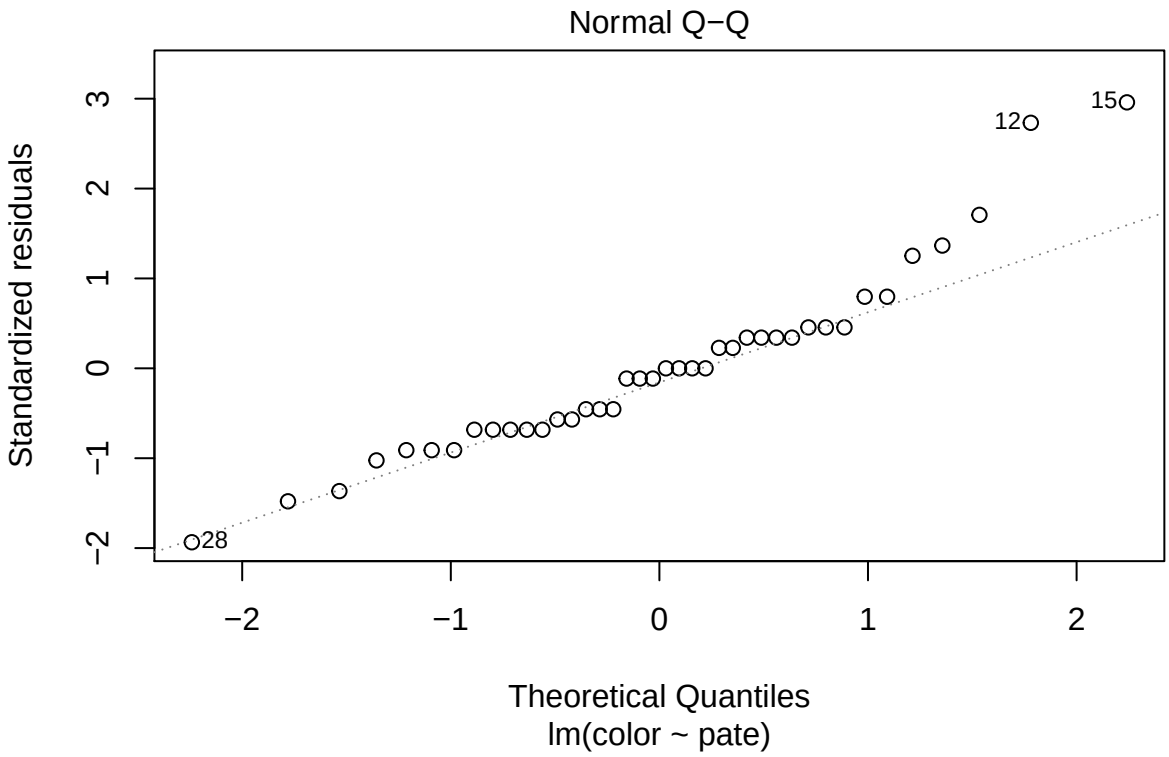
From this ONE-way ANOVa model we see:

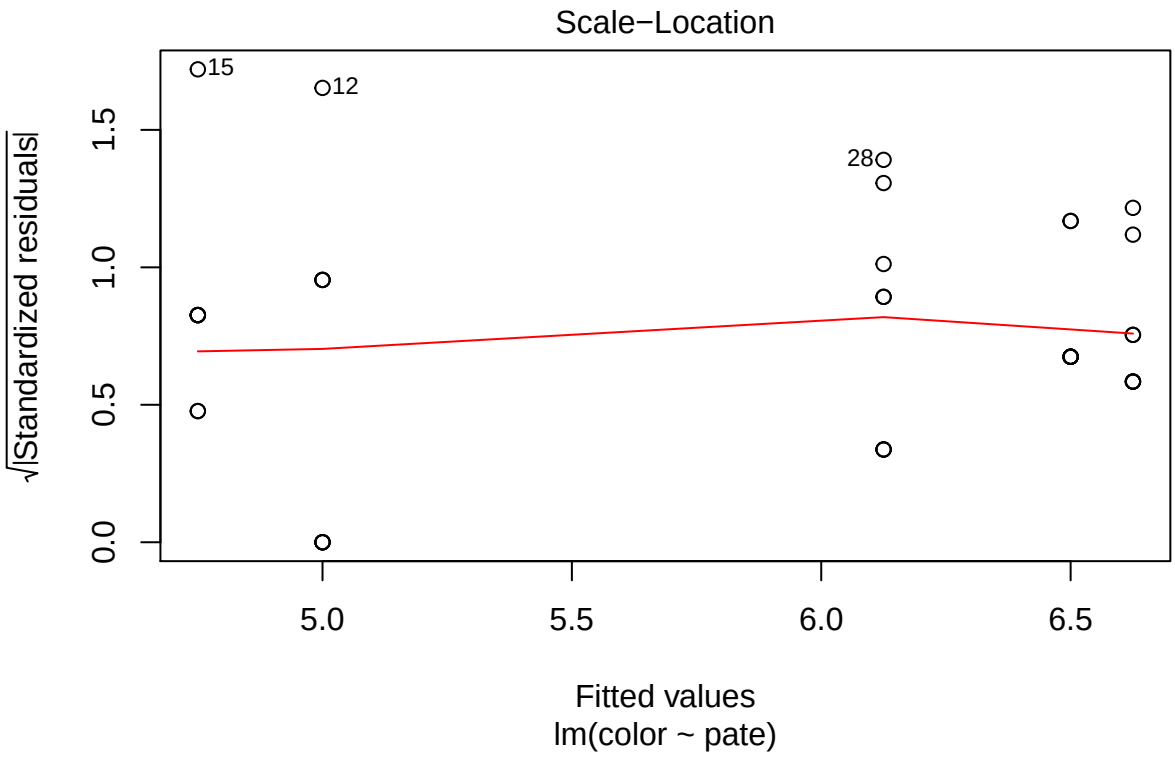
1) there are significant differences between the pates, as a consequence of the fact that the omnibus test is significative and that some of the paraemters are significatively different from zero.
2) The baseline pate (113) doesn't seem to be different from pate 140, neither from pate 220.
3) The pate explains 33% of the variability observed in the color variables.
4) From the ANOVa table we see that the pate is significative, thus it has an influence on the color puctuation.
5) From the anova table we see that the Sum of Squares that corresponds to the residuals is larger than the one that corresponds to the pate. Thus, there is more unexplained variability than explained variability.

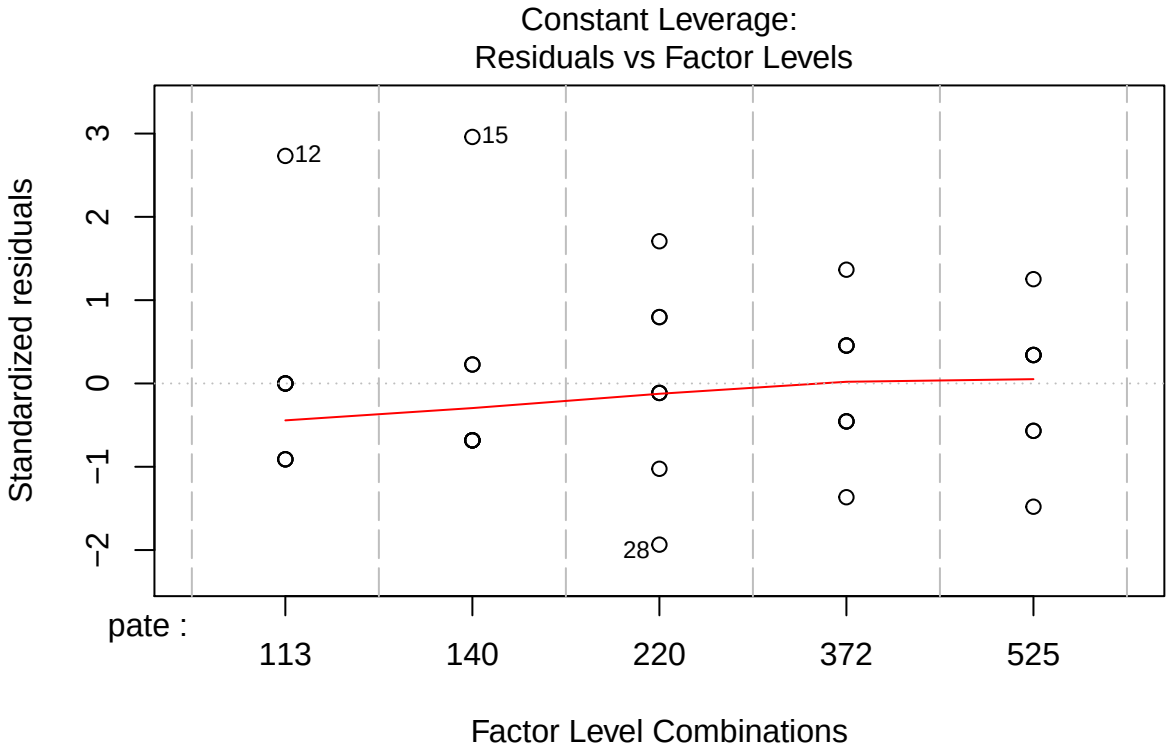
**In what follows we perform the residual analysis**

```r
plot(model1, ask=F)
```


Residuals vs Fitted
Residuals
Fitted values
lm(color ~ pate)
15
12
28
3
2
1
0
−1
−2
5.0
5.5
6.0
6.5


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(color ~ pate)
28
12
15
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
15
12
28
5.0
5.5
6.0
6.5
Fitted values
lm(color ~ pate)




Plotting the residuals versus predicted we observe:

1) there are just 5 predicted values, one for each catherogy (pate type) that are equal to the sample mean of each pate. We do not see eight residuals for each pate because some of them are equal.
2) There are two pates that show a larger variability on the residuals.
3) The residuals do not seem to follow an standarized Normal distribution.
4) By plotting the standarized residuals the homocedasticity property is dobutful.

**Next we are going to assume that the person has a significant influence on the puctuation.**

**We are going to assume a TWO-way anova without interaction**

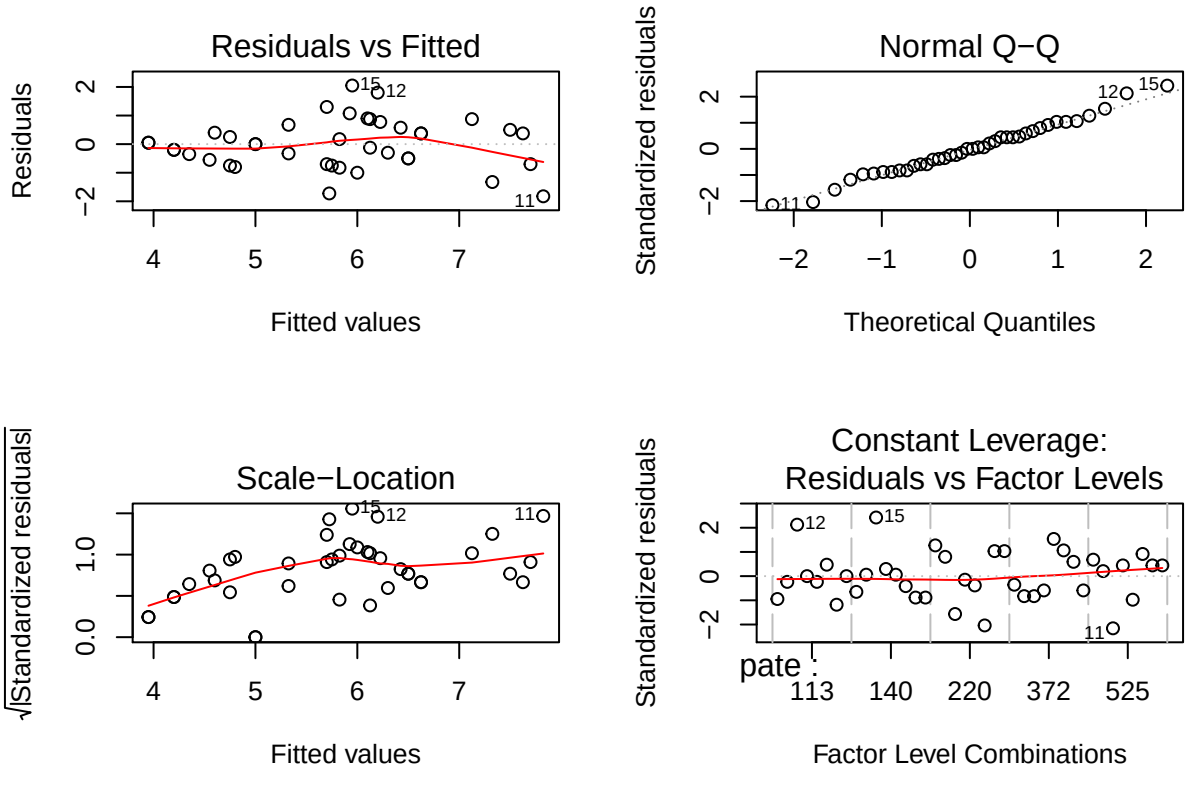
```
model2<-lm(color~pate+per, patedata)

summary(model2)
```

```
##
## Call:
## lm(formula = color ~ pate + per, data = patedata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.8250 -0.5875  0.0000  0.5188  2.0500
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.8000     0.5540   8.664 2.07e-09 ***
## pate140      -0.2500     0.5058  -0.494  0.62495
## pate220       1.1250     0.5058   2.224  0.03436 *
## pate372       1.5000     0.5058   2.966  0.00611 **
## pate525       1.6250     0.5058   3.213  0.00330 **
## per2         -0.6000     0.6398  -0.938  0.35634
## per3          1.4000     0.6398   2.188  0.03715 *
## per4          0.2000     0.6398   0.313  0.75689
## per5         -0.6000     0.6398  -0.938  0.35634
## per6         -0.2000     0.6398  -0.313  0.75689
## per7          1.2000     0.6398   1.876  0.07116 .
## per8          0.2000     0.6398   0.313  0.75689
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.012 on 28 degrees of freedom
## Multiple R-squared:  0.6043, Adjusted R-squared:  0.4488
## F-statistic: 3.887 on 11 and 28 DF,  p-value: 0.001785
```

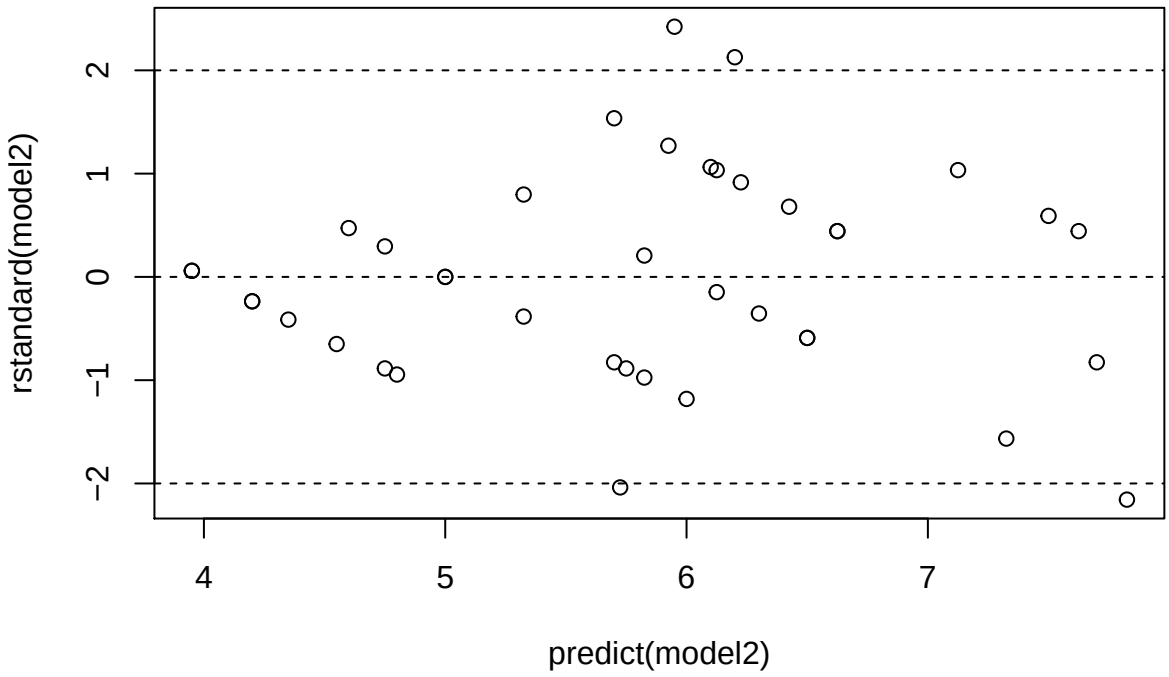
```
oldpar <- par( mfrow=c(2,2))
plot(model2,ask=F)
```



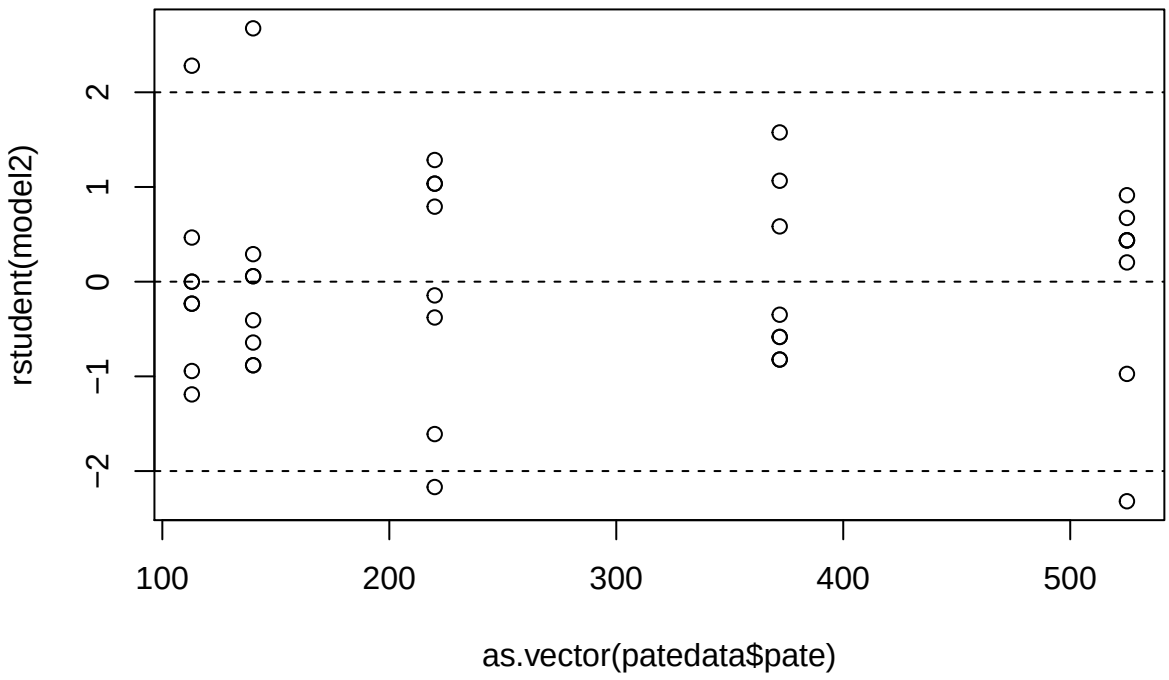

```
par(oldpar)
```

**We do some additional plots**

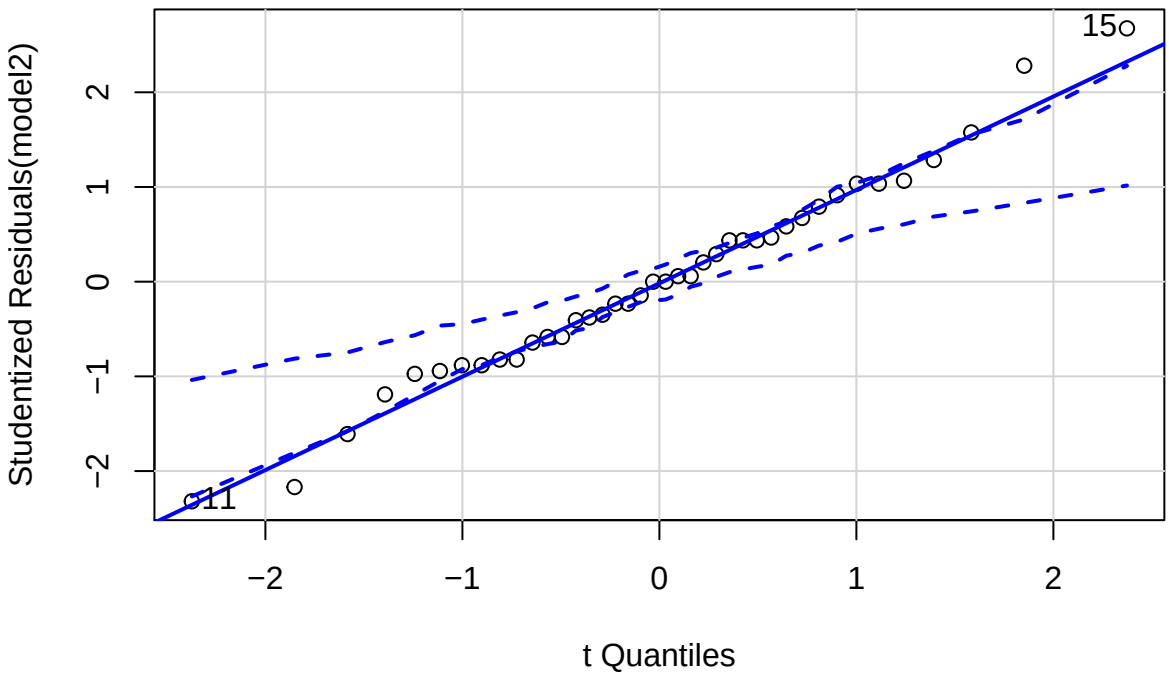
```
plot(predict(model2),rstandard(model2))
abline(h=c(-2,0,2),lty=2)
```




```
plot(as.vector(patedata$pate),rstudent(model2))
abline(h=c(-2,0,2),lty=2)
```

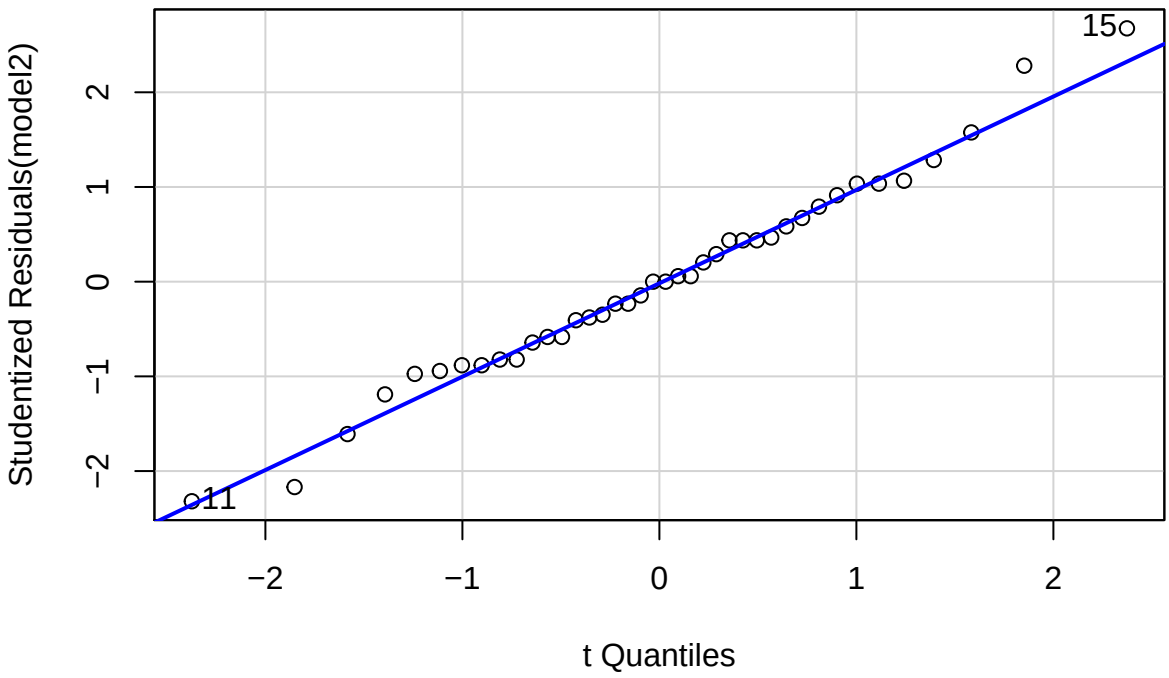



```
qqPlot(model2)
```




```
## [1] 11 15
```

```
qqPlot(model2,simulate=F,envelope=F)
```




```
## [1] 11 15
```

From model2 we deduce the following things:

1) Person 3 is clearly different from person 1 (baseline). The rest of the persons do not differ significatively from person 1.
2) The two factors now explain 60% of the variability in the color variable.
3) Looking at the Anova table we see that both factor are significative.
4) Looking at the anova table, we see that there is still a lot of variability that remains unexplained (approximattly the same variability explained by the pate).
5) Looking at the residual versus fitted plot we see 5 inclined lines that correspond to each one of the pates. In each line there should be 8 circles, one for each person, or less in the case that two or more persons have punctuated equaly a given pate.
6) The Normality of the residuals can now be assumed.
7) We do not see patterns in the plot

We think that this model can be accepted as a good model to explain the differences in the color. Nevertheless it should be interesting to investigate which other features may have an important influence in the color that have not been taken into account.

Important to observe that it is not possible to consider a TWO-way anova with interaction because in this case we will not have degrees of freedom for the residual sum of squares.

**Let us compute confidence levels for the marginal means (means of each pate)**

**The standard error estimation will change depending on the model considered.**

```
emmeans(model1,~pate)
```

```
##  pate emmean        SE df lower.CL upper.CL
##  113   5.000 0.4151162 35 4.157269 5.842731
##  140   4.750 0.4151162 35 3.907269 5.592731
##  220   6.125 0.4151162 35 5.282269 6.967731
##  372   6.500 0.4151162 35 5.657269 7.342731
##  525   6.625 0.4151162 35 5.782269 7.467731
##
## Confidence level used: 0.95
```

```
emmeans(model2,~pate)
```

```
##  pate emmean        SE df lower.CL upper.CL
##  113   5.000 0.3576336 28 4.267421 5.732579
##  140   4.750 0.3576336 28 4.017421 5.482579
##  220   6.125 0.3576336 28 5.392421 6.857579
##  372   6.500 0.3576336 28 5.767421 7.232579
##  525   6.625 0.3576336 28 5.892421 7.357579
##
## Results are averaged over the levels of: per
## Confidence level used: 0.95
```

**Let us perform the multiple comparisons by means of the Tukey methods**

**Remind: levels with the same letter are not statistically different**

```
CLD(emmeans(model1,~pate),Letters=letters,reversed=T)
```

```
##  pate emmean        SE df lower.CL upper.CL .group
##  525   6.625 0.4151162 35 5.782269 7.467731  a
##  372   6.500 0.4151162 35 5.657269 7.342731  a
##  220   6.125 0.4151162 35 5.282269 6.967731  ab
##  113   5.000 0.4151162 35 4.157269 5.842731  ab
##  140   4.750 0.4151162 35 3.907269 5.592731   b
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 5 estimates
## significance level used: alpha = 0.05
```

```
CLD(emmeans(model2,~pate),Letters=letters,reversed=T)
```

```
##  pate emmean        SE df lower.CL upper.CL .group
##  525   6.625 0.3576336 28 5.892421 7.357579  a
##  372   6.500 0.3576336 28 5.767421 7.232579  a
##  220   6.125 0.3576336 28 5.392421 6.857579  ab
##  113   5.000 0.3576336 28 4.267421 5.732579   b
##  140   4.750 0.3576336 28 4.017421 5.482579   b
##
## Results are averaged over the levels of: per
```

```
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 5 estimates
## significance level used: alpha = 0.05
```

**We do the same with model 1 and a significance level of** 0.01

```
CLD(emmeans(model1,~pate),Letters=letters,reversed=T,alpha=0.01)
```

```
##  pate emmean        SE df lower.CL upper.CL .group
##  525   6.625 0.4151162 35 5.782269 7.467731  a
##  372   6.500 0.4151162 35 5.657269 7.342731  a
##  220   6.125 0.4151162 35 5.282269 6.967731  a
##  113   5.000 0.4151162 35 4.157269 5.842731  a
##  140   4.750 0.4151162 35 3.907269 5.592731  a
##
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 5 estimates
## significance level used: alpha = 0.01
```

In this case, given that we require a larger value to declare significative a difference, we do not see differences statistically significant.