Linear Model: Testing Pates

In a paté tasting, each person involved has to taste and give a score to 5 different pates, which were randomly assigned. The explanatory varirables correspond to:

- The person that has tasted the patés.
- Paté code.

Both are factors. And several response variables: scores of color, smell, texture, taste and order.

The marks were between 0 and 10, and the order is the preferred order of the taster where 1 means the preferred by the taster.

For each one of the variables that have been scored,

(1) Find the appropriate linear model assuming that conditions for the ANOVA are satisfied. Perform the corresponding tests and interpret them, including multiple comparisons. Is it necessary to apply any kind of data transformation?

```r
library(car)
library(emmeans)
library(tables)
library(RcmdrMisc)

dd<-read.csv2("PATE.csv")
head(dd)
```

```
##   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
```

```r
dim(dd)
```

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

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

Descriptive statistics

First of all, we are going to perform descriptive statistics:

```r
summary(dd)
```

```
##       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
```

1

```
##  Median :6.00    Median :3
##  Mean   :5.95    Mean   :3
##  3rd Qu.:7.00    3rd Qu.:4
##  Max.   :8.00    Max.   :5
```

Let us transform variables PER (person) and PATE as factors.

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

COLOR

We will start off analyzing the variable **COLOR**: we compute the number of observations and the mean and sd of the color scores:

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

		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:

(a) Paté 140 is the one with the smallest score.
(b) Patés 372 and 525 are the ones with the largest scores.
(c) The variances do not seem to differ much, but the patés with the largest scores seem to have the variance smaller.

In what follows we show the eight scores of color for each pate:

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

				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 people 3 and 7 tend to give higher scores

Modelization

We will do two models to compare both: the **TWO-WAY without interaction anova model**:

$$Color_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Where α_i is the paté and β_j is the person. The matrix would be of the type:

$$
\begin{pmatrix}
1 & 1 & 0 & \cdots & 0 \\
1 & 1 & 0 & \cdots & 1 \\
1 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & 0 & 1 & \cdots & 1
\end{pmatrix}
\begin{pmatrix}
\mu \\
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_i \\
\vdots \\
\beta_1 \\
\beta_2 \\
\vdots \\
\beta_i \\
\vdots
\end{pmatrix}
$$

And the **ONE-WAY anova model**:
$$Color_{ij} = \mu + e_{ij}$$

```
summary(one_model<-lm(color~pate, dd))
```

```
##
## Call:
## lm(formula = color ~ pate, data = dd)
##
## 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(one_model)
```

```
## 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
```

```
Anova(one_model)
```

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

From this ONE-way ANOVA model we see:

(a) There are significant differences between the patés, as a consequence of the fact that the **Omnibus test** is significative and that some of the paraemters are significatively different from zero.
(b) The baseline pate (113) does not seem to be different from pate 140, nor from pate 220.
(c) The pate explains 33% of the variability observed in the color variables.
(d) From the ANOVA table we see that the paté is significative, thus it has an influence on the color score.
(e) 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 paté. Thus, there is more unexplained variability than explained variability.

```
summary(two_model<-lm(color~pate+per, dd))
```

```
##
## Call:
## lm(formula = color ~ pate + per, data = dd)
##
## 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
```

```
anova(two_model)
```

```
## Analysis of Variance Table
```

```
##
## Response: color
##           Df Sum Sq Mean Sq F value  Pr(>F)
## pate       4  24.15  6.0375  5.9005 0.00142 **
## per        7  19.60  2.8000  2.7365 0.02692 *
## Residuals 28  28.65  1.0232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

See that the residuals in the anova table in model1 are 48.25, and in model2: 28.65. The effects of person that we do not count in model1 are translated to the error (randomness, residuals).

From model2 we deduce the following things:

(a) Person 3 is clearly different from person 1 (baseline). The rest of the people do not differ significatively from person 1.
(b) The two factors now explain 60% of the variability in the color variable.
(c) Looking at the Anova table we see that both factors are significative.
(d) 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 paté).
(e) Looking at the residual versus fitted plot we see 5 inclined lines that correspond to each one of the pates (és perquè les dades són enteres, que pot comportar complicacions en la normalitat, però fem-ne la QQplot per poder suposar normalitat). 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.
(f) The Normality of the residuals can now be assumed.
(g) We do not see patterns in the plot.

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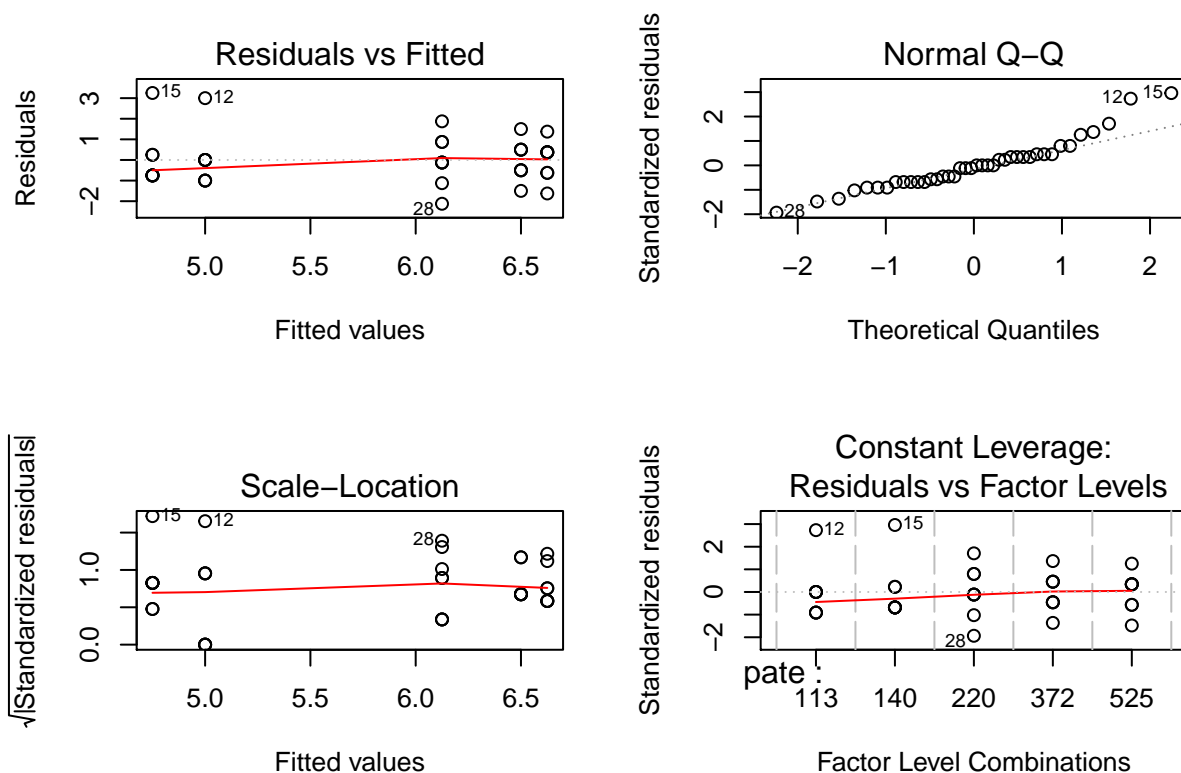
```
par(oldpar)
```

Plotting the residuals versus predicted we observe: (a) There are just 5 predicted values, one for each catherogy (paté 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. (b) There are two patés that show a larger variability on the residuals. (c) The residuals do not seem to follow an standarized Normal distribution. (d) By plotting the standarized residuals the homocedasticity property is dobutful.

(2) Compare the results if we assume that the person has a significant influence upon the score.

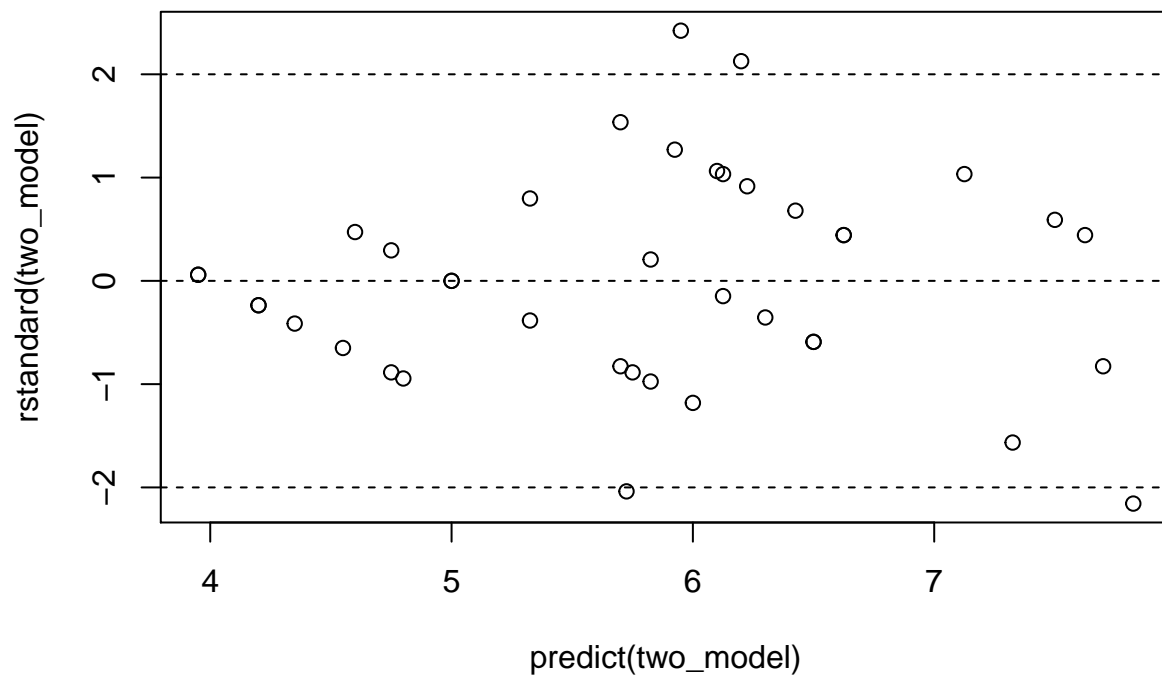We are going to assume a **TWO-way anova without interaction**:

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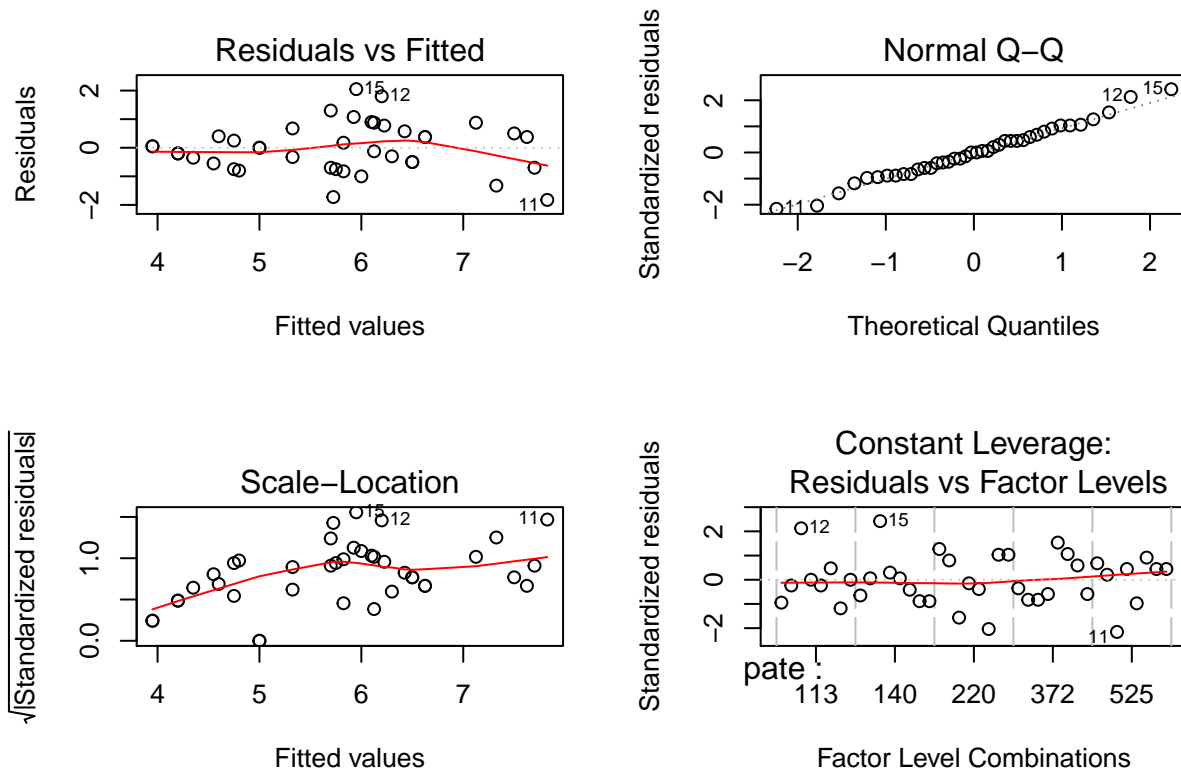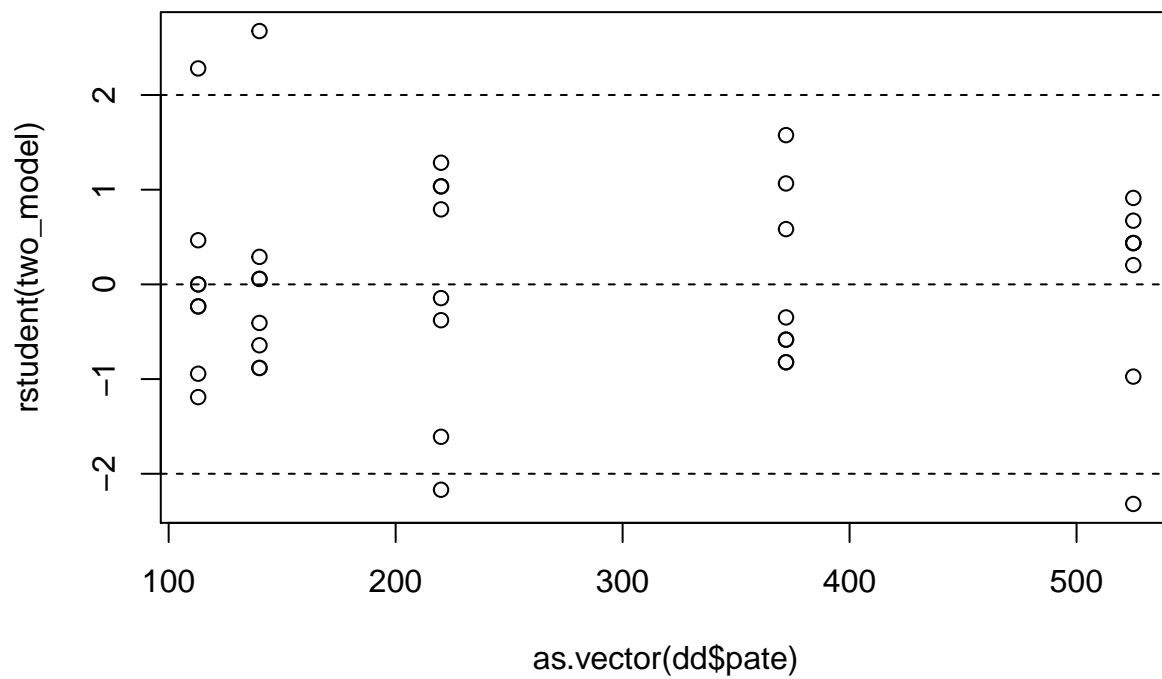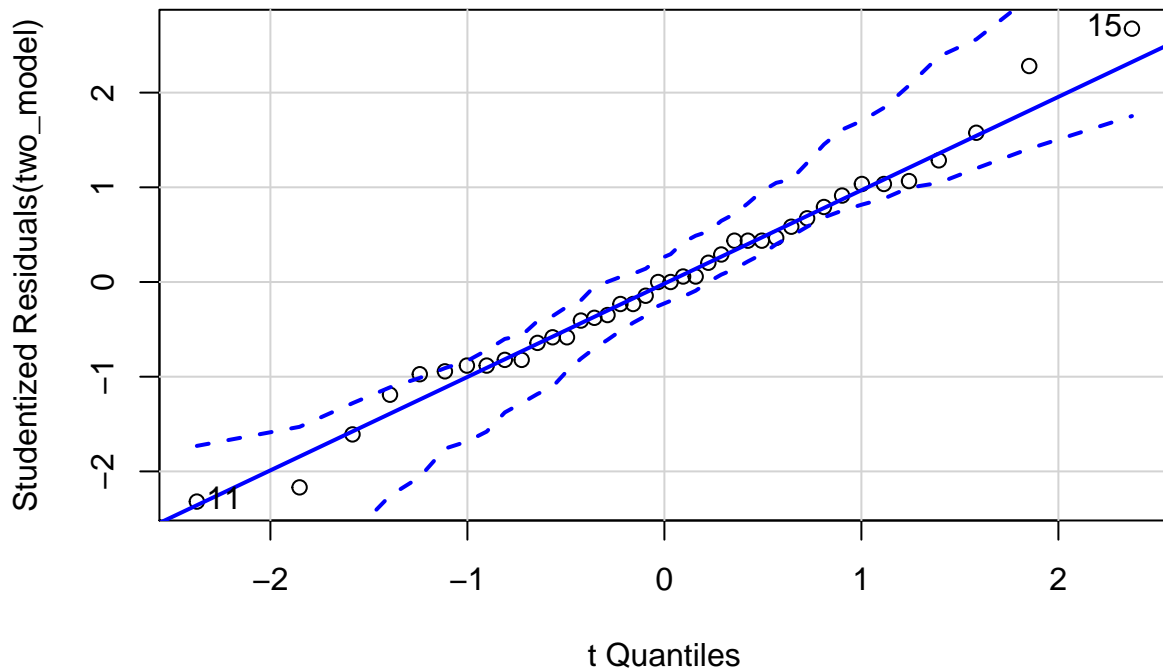


```
par(oldpar)
```

Some additional plots:

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

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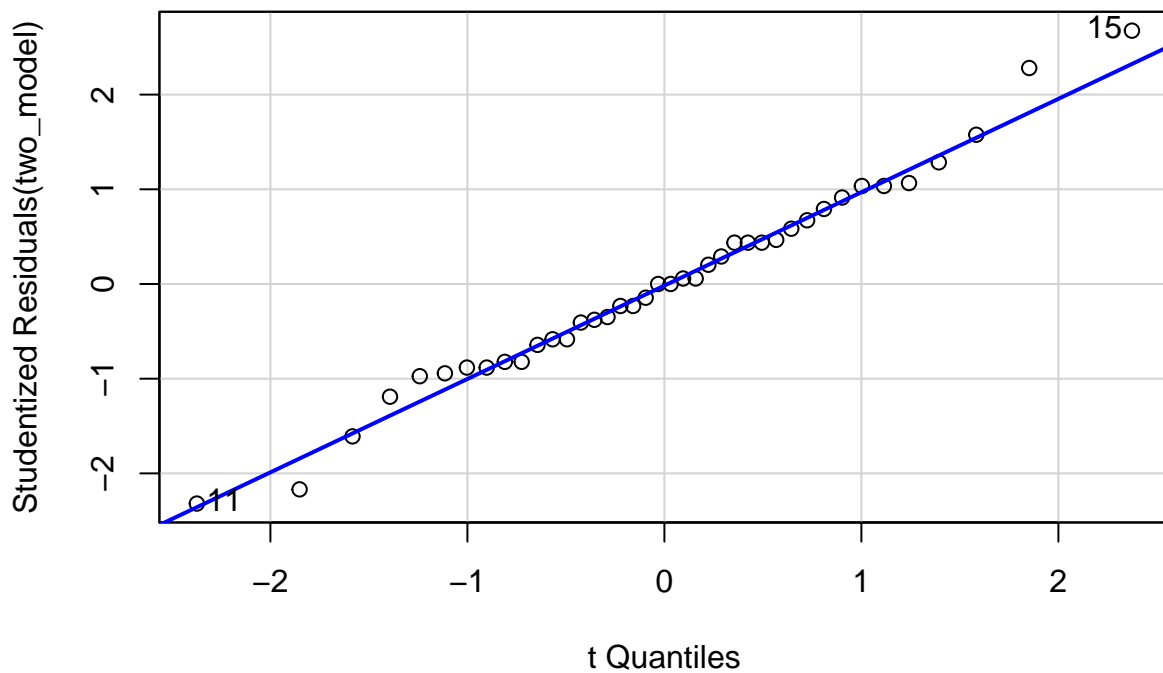


```
qqPlot(two_model)
```

7

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


[1] 11 15

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

Per determinar de quin paté és millor el color, hem de fer les comparacions múltiples, farem comparacions del model1 i del model2 (tot i que model1 és una mica pitjor perquè els residus (errors) són més grans, les

persones estan a l'atzar): Conclusió de l'anova: veiem que hi ha efecte del paté i efecte de persona. Quines diferències hi ha entre els pates? (cal fer 10 tests simultanis).

```
(emm1<-emmeans(one_model,~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
```

```
(emm2<-emmeans(two_model,~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 Remember: levels with the same letter are not statistically different.

Si el bo és el gran, hem de dir que ho faci a la inversa, reversed=T, (A el millor tractament, b el següent, etc).

```
CLD(emm1,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
```

En els models 1 i 2, les estimacions son exactament les mateixes, canvien les desviacions tipus.

```
CLD(emm2,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
```

El 525 i el 375 no els podem distingir. Tots els que ha sortit amb A son els millors. AB vol dir que no els podem distingir ni dels A ni dels B. A i B no són grups, és simplement per poder distingir els patés. Els millors patés, segons l'estadística del color, són de la marca A: 525 i 372.

```
CLD(emmeans(one_model,~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
```

SMELL

Modelization

With **TWO-WAY without interaction anova model**:

$$Smell_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

Where α_i is el paté i β_j is la persona.

```
summary(model<-lm(smell~pate+per, dd))
```

```
##
## Call:
## lm(formula = smell ~ pate + per, data = dd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9750 -0.4125  0.0250  0.6062  1.7000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.5000     0.5323  10.332 4.66e-11 ***
## pate140      -0.7500     0.4860  -1.543  0.13398
## pate220       0.8750     0.4860   1.801  0.08256 .
## pate372       1.5000     0.4860   3.087  0.00453 **
## pate525       0.8750     0.4860   1.801  0.08256 .
## per2         -1.4000     0.6147  -2.278  0.03059 *
## per3          0.8000     0.6147   1.301  0.20371
## per4         -0.4000     0.6147  -0.651  0.52053
## per5         -0.6000     0.6147  -0.976  0.33738
## per6         -2.2000     0.6147  -3.579  0.00128 **
## per7          0.6000     0.6147   0.976  0.33738
## per8         -0.8000     0.6147  -1.301  0.20371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.9719 on 28 degrees of freedom
## Multiple R-squared:  0.6924, Adjusted R-squared:  0.5716
## F-statistic: 5.731 on 11 and 28 DF,  p-value: 9.034e-05
```

```
anova(two_model)
```

```
## Analysis of Variance Table
##
## Response: color
##           Df Sum Sq Mean Sq F value  Pr(>F)
## pate       4  24.15  6.0375  5.9005 0.00142 **
## per        7  19.60  2.8000  2.7365 0.02692 *
## Residuals 28  28.65  1.0232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```
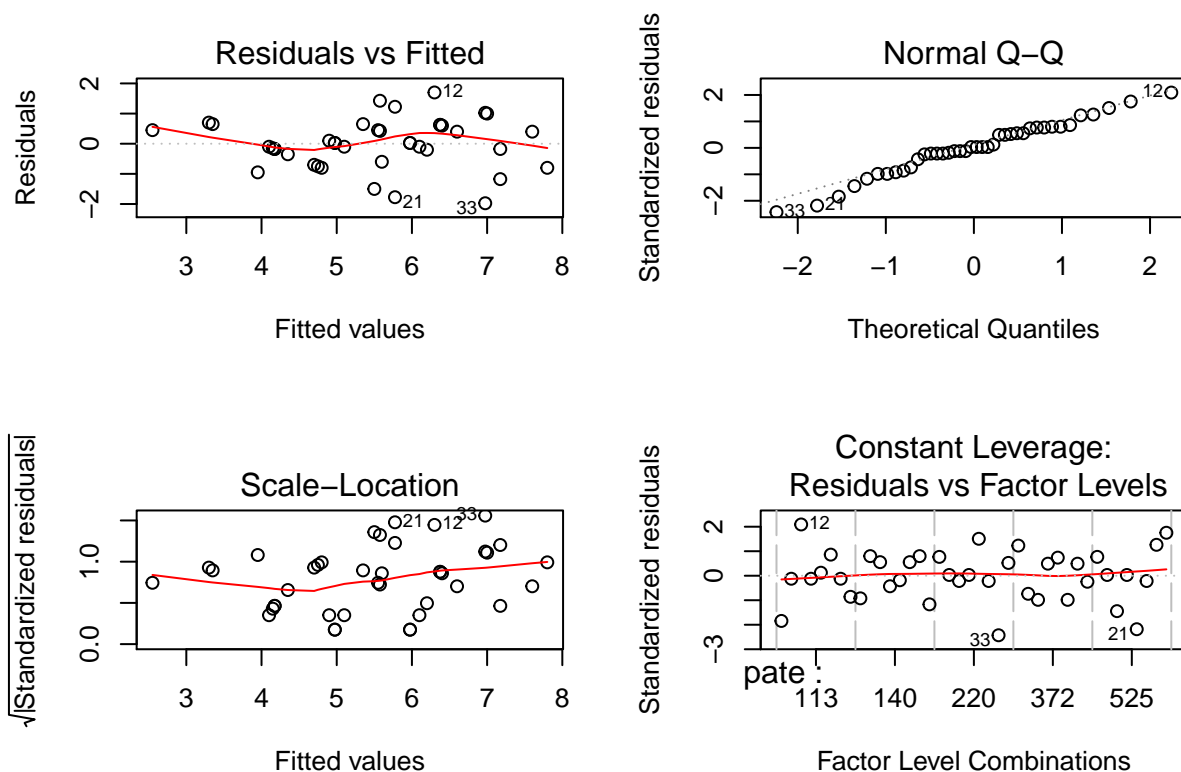
From this TWO-way ANOVA model we see:

(a) The **Omnibus test** is significative. The baseline paté (113) does not seem to be different in smell from the paté 140, 220 and 525.
(b) The two factors explain 69% of variability.
(c) Looking at the anova table, we see that both factors are significative.

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



```
par(oldpar)
```

```
CLD(emmeans(model,~pate),Letters=letters, reversed=T, alpha=0.05)
```

```
##  pate emmean        SE df lower.CL upper.CL .group
##  372   6.500 0.3436282 28 5.796109 7.203891  a
```

11

```
## 220    5.875 0.3436282 28 5.171109 6.578891  ab
## 525    5.875 0.3436282 28 5.171109 6.578891  ab
## 113    5.000 0.3436282 28 4.296109 5.703891   bc
## 140    4.250 0.3436282 28 3.546109 4.953891    c
##
## 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
```

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

```
##  pate emmean        SE df lower.CL upper.CL .group
## 372    6.500 0.3436282 28 5.796109 7.203891  a
## 220    5.875 0.3436282 28 5.171109 6.578891  ab
## 525    5.875 0.3436282 28 5.171109 6.578891  ab
## 113    5.000 0.3436282 28 4.296109 5.703891  ab
## 140    4.250 0.3436282 28 3.546109 4.953891   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 = 0.01
```

Podriem fer l'anova de la variable ordre (igual que la variable color)? Quan ja tenim col.locats 4 patés, el cinquè ja es fix. L'ordre no compleix la independència. Per tant, l'anova amb ORDRE no es pot fer perquè no compleix el requisit de independència, però hi ha un fet curiós: quan hi ha ordres, hi ha el test no paràmetric de Friedman (*No paràmetric* vol dir que no suposem cap distribució de les dades.), que pot fer l'aproximació de l'anova de l'ORDRE del paté, tenint en compte les persones. Però, d'entrada, no podem plantejar l'anova de l'ORDRE perquè no hi ha indepedència: a continuació veiem que el **test Omnibus** falla.

```r
summary(model<-lm(order~pate+per, dd))
```

```
##
## Call:
## lm(formula = order ~ pate + per, data = dd)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3750 -1.0312 -0.0625  1.0625  2.2500
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.375e+00  8.393e-01   4.021 0.000397 ***
## pate140      3.750e-01  7.662e-01   0.489 0.628348
## pate220     -2.500e-01  7.662e-01  -0.326 0.746633
## pate372     -1.375e+00  7.662e-01  -1.795 0.083524 .
## pate525     -6.250e-01  7.662e-01  -0.816 0.421546
## per2         1.434e-15  9.692e-01   0.000 1.000000
## per3         1.423e-15  9.692e-01   0.000 1.000000
## per4         1.303e-15  9.692e-01   0.000 1.000000
## per5         1.223e-15  9.692e-01   0.000 1.000000
## per6         1.123e-15  9.692e-01   0.000 1.000000
## per7         1.151e-15  9.692e-01   0.000 1.000000
## per8         8.426e-16  9.692e-01   0.000 1.000000
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.532 on 28 degrees of freedom
## Multiple R-squared:  0.1781, Adjusted R-squared:  -0.1448
## F-statistic: 0.5517 on 11 and 28 DF,  p-value: 0.8504
```

anova(two_model)

```
## Analysis of Variance Table
##
## Response: color
##           Df Sum Sq Mean Sq F value  Pr(>F)
## pate       4  24.15  6.0375  5.9005 0.00142 **
## per        7  19.60  2.8000  2.7365 0.02692 *
## Residuals 28  28.65  1.0232
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

oldpar<-**par**(mfrow=**c**(2,2))
plot(model, ask=F)
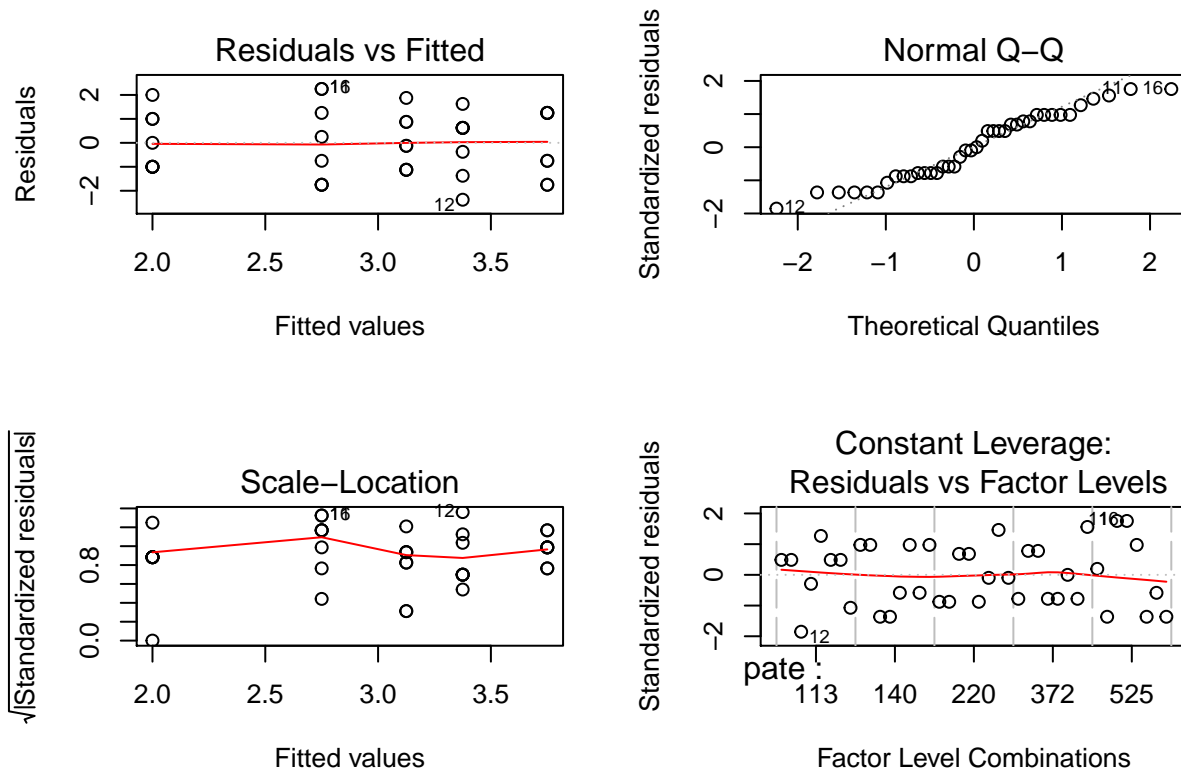
par(oldpar)

Ens podriem preguntar: *hi ha alguna interacció entre persona i paté?* Hem suposat fins ara que totes les persones puntuaven similar: totes reconeixien que el millor és el mateix i el pitjor el mateix, etc.

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.