ML2 TP ANOVA

```
library(lmtest)
library(car)
```

Exercice 2: ANOVA à un facteur

Le fichier “chemical.txt” contient les observations de concentration chimique dans le sang **conc** (en ng/ml) pour un groupe de 10 patients après administration orale de 4 doses dfférentes **dose** (25,50,100,200 mg) d’un médicament (almitrine bismesylate). On veut étudier l’influence du traitement sur les caractéristiques chimiques du sang.

```
data=data.frame(read.csv("C:/Users/Philippine/Documents/Cours/Maths/M1/S2/Data_Mining/chemical.txt", sep=" ", h=F, skip=1, col.names = c("index","conc","dose")))[,2:3]
data
```

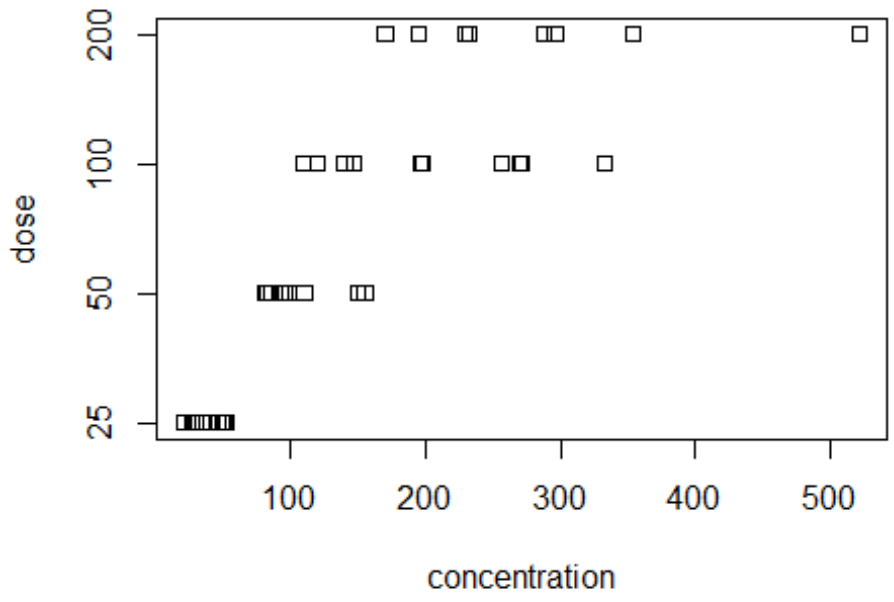
```
##    conc dose
## 1    34   25
## 2    92   50
## 3   256  100
## 4   229  200
## 5    46   25
## 6   150   50
## 7   271  100
## 8   232  200
## 9    50   25
## 10   81   50
## 11  270  100
## 12  288  200
## 13   49   25
## 14  155   50
## 15  120  100
## 16  195  200
## 17   21   25
## 18   85   50
## 19  333  100
## 20  354  200
## 21   52   25
## 22   95   50
## 23  198  100
## 24  288  200
## 25   30   25
## 26   95   50
## 27  109  100
## 28  288  200
## 29   29   25
```

```
## 30   82   50
## 31  140  100
## 32  170  200
## 33   27   25
## 34  110   50
## 35  147  100
## 36  522  200
## 37   51   25
## 38   99   50
## 39  196  100
## 40  296  200
```

```
Y=data$conc
X=as.factor(data$dose)
```

1. Représenter les données à l'aide de boîtes à moustaches. Commenter. Les hypothèses d'une analyse de variance semble-t'elle vérifées ?

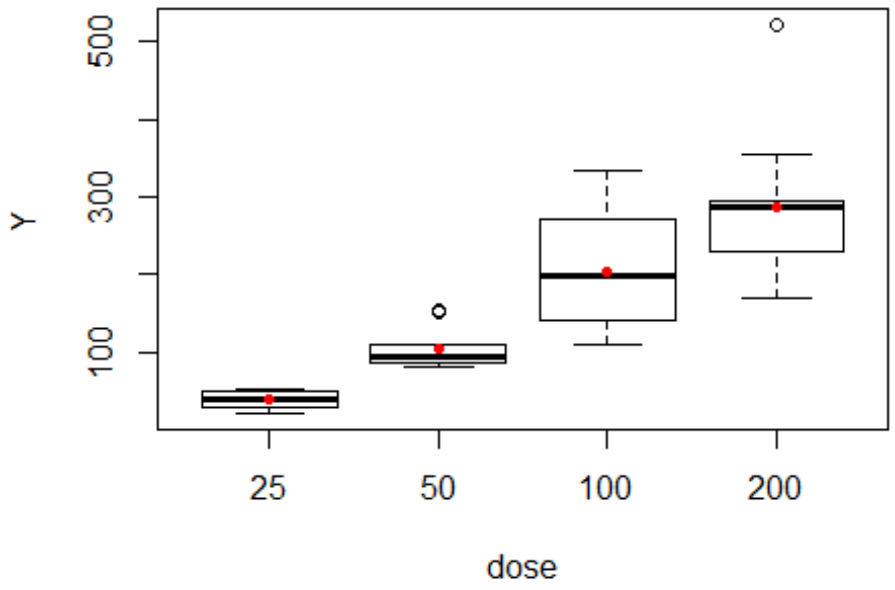
```
stripchart(Y~X, xlab="concentration", ylab="dose")
```



```r
moy=by(Y,X,mean);moy
```

```
## X: 25
## [1] 38.9
## ------------------------------------------------------------
## X: 50
## [1] 104.4
## ------------------------------------------------------------
## X: 100
## [1] 204
## ------------------------------------------------------------
## X: 200
## [1] 286.2
```

```r
boxplot(Y~X, xlab="dose")
points(1:4,moy,col='red',pch=20)
```



Commentaire : Les diagrammes en boite montrent des différences de moyenne (points rouges) entre les différentes modalités du facteur **dose**. De plus, la répartition des concentration en fonction des modalités est très différente: la modalité 100mg est celle qui connait les plus grands écarts de concentration mais aussi les mieux répartis. Les patients doivent réagir très différemment les uns des autres quand on leur administre une dose de 100mg de produit. En revanche, lorsque l'on administre une dose de 25mg ou de 50mg, les patients semblent réagir de manière plus homogène. Enfin, la dernière modalité possède la plus forte moyenne et un outlier ayant une forte valeur de concentration.

Est-ce une erreur de mesure ou a-t-on à faire à un patient très atypique ? Influence-t-t-il beaucoup la moyenne de cette modalité ? Il faudrait évaluer si cette individu est vraiment considéré comme outlier (distance de cook). Il en va de même pour la modalité 50mg.

Pour réaliser une analyse de variance (à un facteur), il faut que la variable cible Y soit quantitative et que la variable explicative soit un facteur à q modalités.

De plus, pour une analayse de variance, on s'appui sur l'hypothèse de linéarité et de normalité du modèle. Il faut donc vérifier que les résidus du modèle Yik=yik+eik soit gaussiens, centrés, de même variance et indépendants.

Dans cette étude, nous avons bien Y (conc) quantitative et la variable **dose** est un facteur à 4 modalités (25,50,100,200).Nous vérifions dans la deuxième question les hypothèses citées.

2. On cherche à vérifier si l'hypothèse d'homoscédasticité des modalités a lieu. Pour cela, effectuer le test de Bartlett.

```
mod1=aov(Y~X)
res1=mod1$residuals
shapiro.test(res1) #test de normalité des résidus

##
##  Shapiro-Wilk normality test
##
## data:  res1
## W = 0.87814, p-value = 0.0004691

dwtest(mod1)       #test d'indépendance des résidus

##
##  Durbin-Watson test
##
## data:  mod1
## DW = 1.9093, p-value = 0.4521
## alternative hypothesis: true autocorrelation is greater than 0

bartlett.test(res1~X) #test homoscédasticité des résidus

##
##  Bartlett test of homogeneity of variances
##
## data:  res1 by X
## Bartlett's K-squared = 33.69, df = 3, p-value = 2.303e-07
```

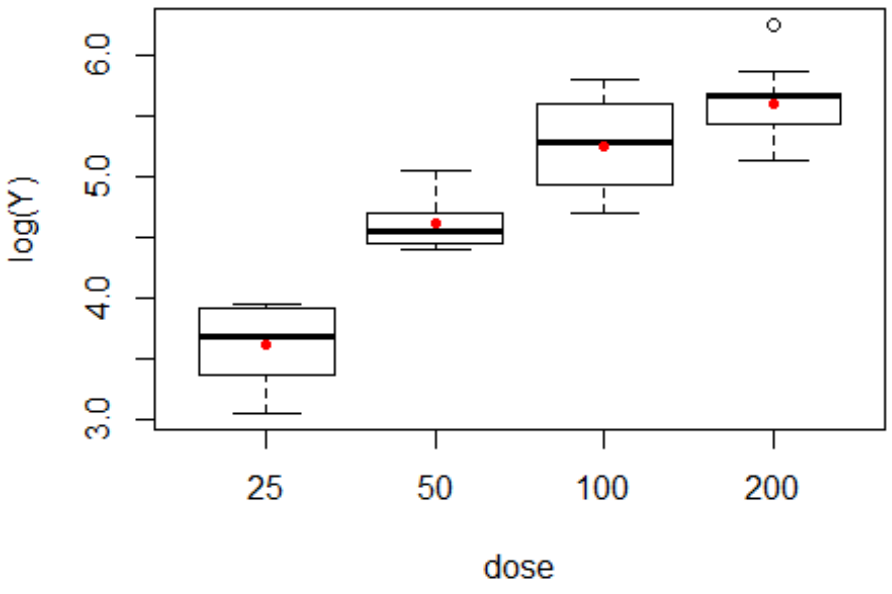
Le test de Bartlett nous indique une non homoscédasticité des résidus: les résidus n'ont pas tous même variance. L'hypothèse d'homoscédasticité n'est donc pas vérifiée avec ce modèle. (On voit aussi que l'hypothèse de normalité est rejetée mais que celle d'indépendance ne l'est pas).

3. Pour éviter ce phénomène, on propose de s'intéresser à une transformation logarithmique de la variable **conc** (Y) . Représenter les boîtes à moustaches des données transformées. Vérifier la procédure de stabilisation avec le test de Bartlett sur **log(Y)** ainsi que la normalité.

```
moy_log=by(log(Y),X,mean);moy_log
```

```
## X: 25
## [1] 3.615077
## ------------------------------------------------------------
## X: 50
## [1] 4.622302
## ------------------------------------------------------------
## X: 100
## [1] 5.253116
## ------------------------------------------------------------
## X: 200
## [1] 5.609546
```

```
boxplot(log(Y)~X, xlab="dose")
points(1:4,moy_log,col='red',pch=20)
```



```r
mod2=aov(log(Y)~X)
res2=mod2$residuals
bartlett.test(res2~X)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  res2 by X
## Bartlett's K-squared = 2.1285, df = 3, p-value = 0.5462
```

```r
shapiro.test(res2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res2
## W = 0.97653, p-value = 0.5628
```

Les nouveaux diagrammes permettent d'avoir une meilleure vision de la répartition des concentrations mesurées en fonction de la dose administrée. Les diagrammes des 2 premières modalités montrent une répartition plus ou moins homogène de la concertation autour de la médiane de la modalité. On remarque toujours une différence de moyenne entre chaque modalité (même si la présence d'un outlier influence peut-être encore la moyenne de la modalité et que celle-ci ne soit en faite pas si différente de la moyenne de la modalité 100) .

On a maintenant homoscédasticité des résidus grâce au changement logarithmique effectué sur Y. On a également la normalité des résidus. Ce nouveau modèle répond donc aux hypothèses de l'analyse de variance, c'est à dire la normalité, l'homoscédasticité et l'indépendance des résidus. (On les suppose centrés par construction)

4. Réaliser l'analyse de variance (phase d'estimation). Retrouver manuellement les résultats en construisant la matrice 'Xind' adaptée.

```r
mod2$coefficients
```

```
## (Intercept)         X50        X100        X200
##    3.615077    1.007226    1.638039    1.994470
```

```r
y1=mod2$coefficients[1];y1 # moyenne de la modalité 1
```

```
## (Intercept)
##    3.615077
```

```r
y2=mod2$coefficients[2]+y1;y2 # moyenne de la modalité 2
```

```
##      X50
## 4.622302
```

```r
y3=mod2$coefficients[3]+y1;y3 # moyenne de la modalité 3
```

```
##     X100
## 5.253116
```

```r
y4=mod2$coefficients[4]+y1;y4 # moyenne de la modalité 4
```

```
##     X200
## 5.609546
```

```r
y_bar=(y1+y2+y3+y4)/4;y_bar # moyenne générale
```

```
## (Intercept)
##     4.77501
```

```r
anova(mod2) #on trouve SCM = 22.9380
```

```
## Analysis of Variance Table
##
## Response: log(Y)
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X          3 22.9380  7.6460  74.723 1.575e-15 ***
## Residuals 36  3.6837  0.1023
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
            # SCR=3.6837
            # SCM=22.9380
            # sigma2_est=0.1023
            # F=74.723
            # p-val=1.575e-15
```

```r
Xind=as.matrix(data.frame(model.matrix(~X-1))); Xind
```

```
##    X25 X50 X100 X200
## 1    1   0    0    0
## 2    0   1    0    0
## 3    0   0    1    0
## 4    0   0    0    1
## 5    1   0    0    0
## 6    0   1    0    0
## 7    0   0    1    0
## 8    0   0    0    1
## 9    1   0    0    0
## 10   0   1    0    0
## 11   0   0    1    0
## 12   0   0    0    1
## 13   1   0    0    0
## 14   0   1    0    0
## 15   0   0    1    0
## 16   0   0    0    1
## 17   1   0    0    0
```

```
## 18   0   1    0    0
## 19   0   0    1    0
## 20   0   0    0    1
## 21   1   0    0    0
## 22   0   1    0    0
## 23   0   0    1    0
## 24   0   0    0    1
## 25   1   0    0    0
## 26   0   1    0    0
## 27   0   0    1    0
## 28   0   0    0    1
## 29   1   0    0    0
## 30   0   1    0    0
## 31   0   0    1    0
## 32   0   0    0    1
## 33   1   0    0    0
## 34   0   1    0    0
## 35   0   0    1    0
## 36   0   0    0    1
## 37   1   0    0    0
## 38   0   1    0    0
## 39   0   0    1    0
## 40   0   0    0    1
```

```
#Estimation des paramètres à la main (on utilise log(Y))
Beta_chap=solve(t(Xind)%*%Xind)%*%t(Xind)%*%log(Y);Beta_chap #donne les valeu
r des moyennes par modalité
```

```
##          [,1]
## X25  3.615077
## X50  4.622302
## X100 5.253116
## X200 5.609546
```

```
Yc=Xind%*%Beta_chap      #Y chapeau
SCR=sum((log(Y)-Yc)**2);SCR
```

```
## [1] 3.683706
```

```
sigma_est=SCR/36;sigma_est
```

```
## [1] 0.1023252
```

```
SCM=sum((Yc-mean(log(Y)))**2);SCM
```

```
## [1] 22.93801
```

```
testF=(SCM/3)/(SCR/36);testF
```

```
## [1] 74.7226
```

```
p_val=1-pf(testF,3,36);p_val  # 3 = nb de modalité -1
                              # 36 = dimension de Y - nombre de modalité de X
```

```
## [1] 1.554312e-15
```

5. Construire le tableau d'analyse de variance et interpréter les résultats. Conclure quant à l'effet du traitement.

On va construire à la main le tableau que nous donne la commande anova(mod2)

```
tableau_anova=data.frame(Source=c("X","Residuals"),Df=c(3,36),Sum_Sq=c(22.938
,3.684),Mean_Sq=c(7.646,0.102),F_value=c(74.72,"NA"),pval=c(1.58e-15,"NA"));
tableau_anova
```

```
##      Source Df Sum_Sq Mean_Sq F_value     pval
## 1         X  3 22.938   7.646   74.72 1.58e-15
## 2 Residuals 36  3.684   0.102      NA       NA
```

Interprétation : La p-value du test de Fisher nous conduit à rejeter l'hypothèse nulle d'égalité des moyennes entre modalité (On a une grande valeur de SCM et une petite valeur de SCR, cela était donc prévisible). La dose de médicament injectée a donc vraisemblablement une influence sur la concentration dans le sang car les moyennes sont significativement différentes.

6. On veut à présent comparer plus précisément les effets de la dose sur la concentration selon la quantité de médicament prescrite. Interpréter les résultats de coef(mod2) puis ceux de coef(mod3) où mod3=lm(concent~-1+dose). Comparer deux à deux les effets selon la dose, à l'aide de la méthode de Bonferroni puis de la méthode de Tukey.

```
coef(mod2)
```

```
## (Intercept)         X50        X100        X200
##    3.615077    1.007226    1.638039    1.994470
```

```
mod3=lm(log(Y)~-1+X);coef(mod3)
```

```
##      X25      X50     X100     X200
## 3.615077 4.622302 5.253116 5.609546
```

```
LSD_Bonf=qt(1-(0.05/6)/2,36)*sqrt(8*sigma_est/40);LSD_Bonf
```

```
## [1] 0.3994085
```

```r
pairwise.t.test(log(Y),X,p.adj="bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  log(Y) and X
##
##     25      50      100
## 50  1.7e-07 -       -
## 100 8.9e-13 0.00054 -
## 200 2.7e-15 2.7e-07 0.10478
##
## P value adjustment method: bonferroni
```

```r
LSD_Tukey=qtukey(0.95,4,36)*sqrt(4*sigma_est/40);LSD_Tukey
```

```
## [1] 0.3852824
```

```r
TukeyHSD(aov(mod2),'X')
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mod2)
##
## $X
##              diff         lwr      upr     p adj
## 50-25   1.0072256  0.62194312 1.392508 0.0000002
## 100-25  1.6380390  1.25275656 2.023321 0.0000000
## 200-25  1.9944696  1.60918713 2.379752 0.0000000
## 100-50  0.6308134  0.24553100 1.016096 0.0005012
## 200-50  0.9872440  0.60196157 1.372526 0.0000003
## 200-100 0.3564306 -0.02885186 0.741713 0.0783512
```

```
plot(TukeyHSD(mod2,'X'))
```

Les coefficients affichés par la commande coef(mod2) représentent les moyennes des modalités, en prenant la première modalité comme modalité de référence. On a donc, comme vu précédemment, y1=3.615, y2=4.622, y3=5.253, y4=5.609. Les Yik sont donc égaux à yik+eik.

Les valeurs observées de la concentration en fonction de la dose de médicament prescrite peuvent donc être considérées comme étant égales à la moyenne des valeurs de concentration observée pour cette modalité + un terme d'erreur gaussien.

Ce 3ème modèle (mod3) correspond à ce que l'on fait à la main, les coefficients ici sont directement les moyennes de chaque modalité. On retrouve d'ailleurs les mêmes résultats. (Ouf!)

Avec la méthode de Bonferroni, 2 moyennes sont significativement différentes si elles diffèrent de plus de 0,399. on a donc y1!=y2,y3,y4; y2!=y1,y3,y4 et y3!=y1,y2 mais y3 et y4 sont significativement égales.(Ce qu'on peut voir plus facilement grâce à la commande pairwise)

Avec la méthode de Tukey, 2 moyennes sont significativement différentes si elles diffèrent de plus de 0,385. On voit donc grâce à la commande plot(TukeyHSD(mod2,'X')) que là aussi y3 et y4 ne sont pas significativement différentes. Il semblerait donc que les différentes doses aient une influence particulière sur la concentration mais que les doses 100mg et 200mg aient le même effet (si l'on se fie à ces indicateurs)

7. Tester avec la méthode de Tukey une différence significative entre une dose de 25mg et 50mg .

On a vu à la question précédente que LSD_Tukey donne une p-val de 1.7e-07 pour l'hypothèse nulle "y1=y2" donc y1 et y2 ne sont pas significativement égales. Il y a une différence significative entre y1 et y2, c'est à dire entre une dose de 25mg et une dose de 50mg.

8. On veut tester à l'aide de la méthode des contrastes l'égalité de la différence des effets entre 25mg et 50mg et entre 50mg et 100mg . On utilise alors la fonction lht de la library car. Même question entre 25mg et 50mg et entre 200mg et 100mg.

```
library(car)
lht(mod3,c(1,-2,1,0))
```

```
## Linear hypothesis test
##
## Hypothesis:
## X25 - 2 X50  + X100 = 0
##
## Model 1: restricted model
## Model 2: log(Y) ~ -1 + X
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     37 3.9198
## 2     36 3.6837  1   0.23614 2.3078 0.1375
```

```
lht(mod3,c(1,-1,1,-1))
```

```
## Linear hypothesis test
##
## Hypothesis:
## X25 - X50  + X100 - X200 = 0
##
## Model 1: restricted model
## Model 2: log(Y) ~ -1 + X
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     37 8.3326
## 2     36 3.6837  1    4.6489 45.433 7.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
lht(mod2,c(0,2,-1,0)) #on retrouve bien les mêmes résultats, que ce soit avec X ou X-1
```

```
## Linear hypothesis test
##
## Hypothesis:
## 2 X50 - X100 = 0
##
## Model 1: restricted model
## Model 2: log(Y) ~ X
##
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     37 3.9198
## 2     36 3.6837  1   0.23614 2.3078 0.1375
```

```r
lht(mod2,c(0,1,-1,1))
```

```
## Linear hypothesis test
##
## Hypothesis:
## X50 - X100  + X200 = 0
##
## Model 1: restricted model
## Model 2: log(Y) ~ X
##
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     37 8.3326
## 2     36 3.6837  1    4.6489 45.433 7.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La p-value du 1er test nous conduit à ne pas rejeter l'hypothèse nulle H0:"y1-y2=y2-y3". Cela signifie que les différences d'effet entre 25mg et 50mg et 50mg et 100mg sont significatvement égales.

La p-value du 2ème test nous conduit cette fois-ci à rejeter l'hypothèse nulle H0:"y1-y2=y4-y3". Cela signifie que les différences d'effet entre 25mg et 50mg et 100mg et 200mg sont significativement différentes. Ces résultats sont assez logiques car nous avons mis en avant dans les questions prédédentes que y1, y2 et y3 étaient significativement différentes entre elles mais que y3 et y4 ne pouvaient pas être considérées comme significativement différentes. L'effet entre 100mg et 200mg est donc très faible comparé à celui entre 25mg et 50mg.

Exercice 3: ANOVA à deux facteurs

1. Déterminer les moyennes des modalités et des interactions et représenter graphiquement les effets moyens et les interactions. Quelles sont les conjectures envisageables quant aux résultats de l'analyse?

```
data(warpbreaks)
donnees=warpbreaks
model=lm(breaks~wool+tension + wool:tension, data=donnees)
moy_wool=by(donnees$breaks,donnees$wool,mean);moy_wool
```

```
## donnees$wool: A
## [1] 31.03704
## ------------------------------------------------------------
## donnees$wool: B
## [1] 25.25926
```

```
moy_tension=by(donnees$breaks,donnees$tension,mean);moy_tension
```

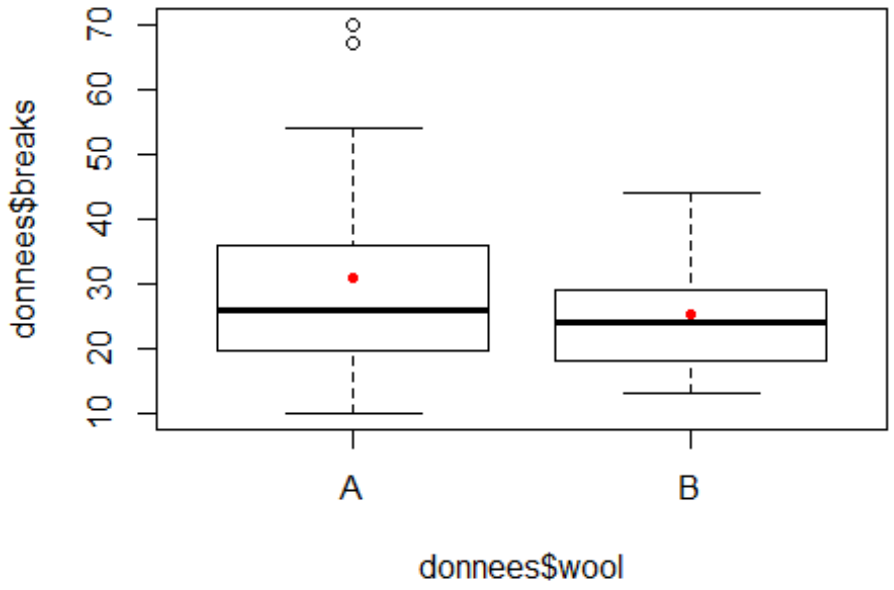
```
## donnees$tension: L
## [1] 36.38889
## ------------------------------------------------------------
## donnees$tension: M
## [1] 26.38889
## ------------------------------------------------------------
## donnees$tension: H
## [1] 21.66667
```

```
by(donnees$breaks,list(donnees$wool,donnees$tension),mean)
```

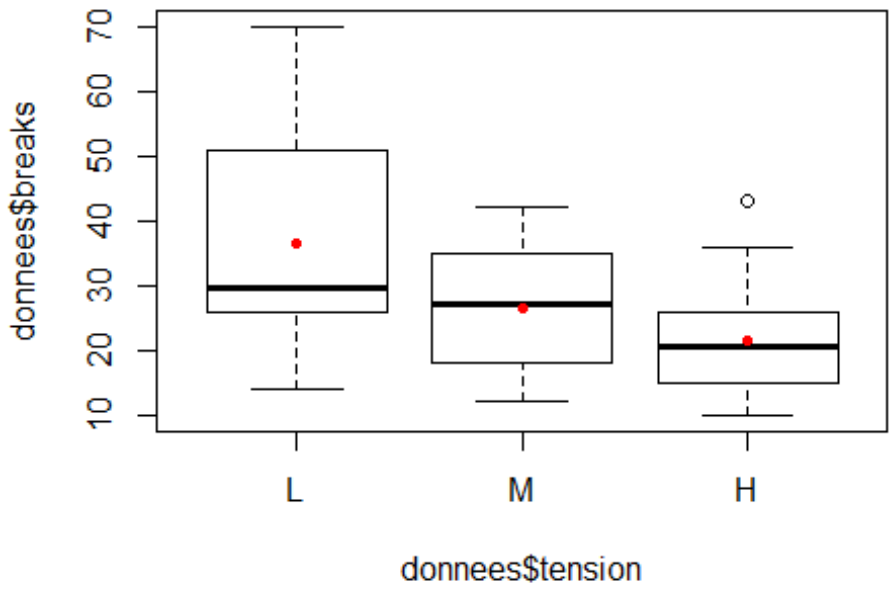
```
## : A
## : L
## [1] 44.55556
## ------------------------------------------------------------
## : B
## : L
## [1] 28.22222
## ------------------------------------------------------------
## : A
## : M
## [1] 24
## ------------------------------------------------------------
## : B
## : M
## [1] 28.77778
## ------------------------------------------------------------
## : A
## : H
## [1] 24.55556
## ------------------------------------------------------------
```

```
## : B
## : H
## [1] 18.77778
```

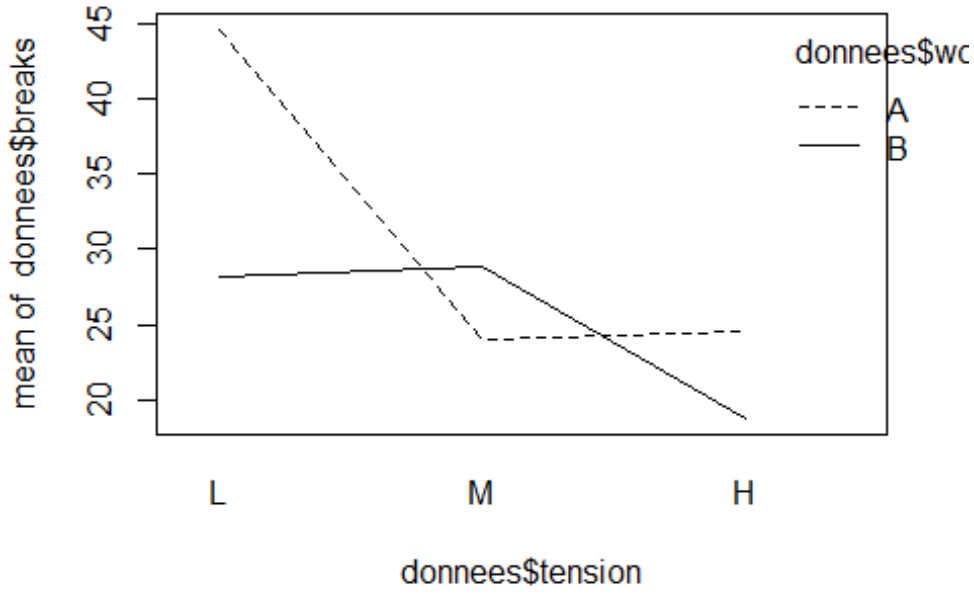
```r
boxplot(donnees$breaks~donnees$wool)
points(1:2,moy_wool,col='red',pch=20)
```



```r
boxplot(donnees$breaks~donnees$tension)
points(1:3,moy_tension,col='red',pch=20)
```




```r
interaction.plot(donnees$tension,donnees$wool,donnees$breaks)
```



Interprétation: La moyenne pour la modalité A de Wool est 31.04 et celle pour la modalité B est 25.26. On voit également grâce au boxplot que ces moyennes sont assez proches et que les valeurs de breaks sont plus étendues pour la modalité A que pour la modalité B. On note la présence de 2 outliers pour la modalité A qui augmentent certainement la moyenne de la moyenne de la modalité A. Graphiquement parlant (et en supprimant les 2 points outliers), on aurait donc tendance à dire que le facteur **wool** seul n'a pas d'effet significatif sur la variable **breaks** (en terme de différence de moyenne : SCM faible, en terme de dispersion des données et SCR fort, ce qui implique un non rejet de l'hypothèse nulle d'égalité des moyennes).

Les moyennes pour les modalités L,M et H du facteur **tension** sont respectivement 36.39, 26.39, 21.67. Les diagrammes en boite mettent en évidence des écarts de moyenne entre les 3 modalités. La répartition des valeurs de breaks pour les modalités M et H semble être homogène autour de la moyenne/médiane alors que c'est beaucoup plus hétérogène pour la modalité L. De plus, les écarts entre moyennes ont l'air assez faibles entre les modalités M et H, donc on pourrait penser que même s'il y a des écarts de moyennes, la répartition très hétérogène des valeurs de breaks pour chaque modalité donne un SCR élevé. On se retrouve donc comme dans le cas précédent avec un SCM faible et un SCR fort donc une statistique de test F faible et donc un rejet de l'hypothèse nulle.

Cependant, toutes ces conjectures sont purement graphique, on ne peut donc en réalité rien conclure quant à l'effet réel de chacun des facteurs sur la variable **breaks**. Si l'on se base sur les différences de moyennes uniquement, on aurait tendance à dire que le facteur **tension** a plus d'influence que le facteur **wool**.

Les moyennes pour les modalités des deux facteurs combinés sont : yAL=44.56, yBL=28.22, yAM=24, yBM=28.78, yAH=24.56 et yBH=18.78. Le plot des interactions montre des interactions entre les modalités des deux facteurs. Pour la modalité L de **tension** il y a des fortes différences de moyennes entre les 2 modalités de **wool**. On note de même des interactions plus faible pour les modalités H et M. Ce graphique nous indique donc qu'il y a des interactions entre les deux facteurs et donc que même si individuellement, les facteurs semblent n'avoir aucun effet, ils peuvent avoir un effet à travers leurs interactions.

2. Vérifier les hypothèses du modèle gaussien. Conclure.

```
residus=model$residuals
shapiro.test(residus)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residus
## W = 0.98686, p-value = 0.8162
```

```r
bartlett.test(residus,donnees$wool)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residus and donnees$wool
## Bartlett's K-squared = 4.8197, df = 1, p-value = 0.02814
```

```r
bartlett.test(residus,donnees$tension)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residus and donnees$tension
## Bartlett's K-squared = 6.9535, df = 2, p-value = 0.03091
```

```r
bartlett.test(residus,donnees$wool:donnees$tension)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residus and donnees$wool:donnees$tension
## Bartlett's K-squared = 12.977, df = 5, p-value = 0.0236
```

```r
dwtest(model)
```

```
##
##  Durbin-Watson test
##
## data:  model
## DW = 2.2376, p-value = 0.575
## alternative hypothesis: true autocorrelation is greater than 0
```

On remarque que ce modèle vérifie bien les hypothèses de normalité et d'indépendance des résidus mais pour l'hypothèse d'homoscédasticité cela est plus ambigu au risque 5% (on rejetterait l'hypothèse d'homoscédasticité ici). Regardons ce qu'il se passe si l'on considère log(breaks).

3. On propose le changement de variable lbreaks=log(breaks). Reprendre la question précédente.

```r
model2=lm(log(breaks)~wool+tension + wool:tension, data=donnees)
residus2=model2$residuals
shapiro.test(residus2)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residus2
## W = 0.97292, p-value = 0.2583
```

```r
bartlett.test(residus2,donnees$wool)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residus2 and donnees$wool
## Bartlett's K-squared = 1.643, df = 1, p-value = 0.1999
```

```r
bartlett.test(residus2,donnees$tension)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residus2 and donnees$tension
## Bartlett's K-squared = 0.23213, df = 2, p-value = 0.8904
```

```r
bartlett.test(residus2,donnees$wool:donnees$tension)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  residus2 and donnees$wool:donnees$tension
## Bartlett's K-squared = 2.8778, df = 5, p-value = 0.7188
```

```r
dwtest(model2)
```

```
##
##  Durbin-Watson test
##
## data:  model2
## DW = 2.06, p-value = 0.3167
## alternative hypothesis: true autocorrelation is greater than 0
```

On a maintenant vérification de tous les postulats d'un modèle linéaire. On va donc pouvoir faire une analyse de variance à deux facteurs (en prenant en compte les interactions).

4. Interpréter l'analyse de variance de votre modèle: Expliquez pourquoi la présence d'interactions fortes compliquent l'interprétation des effets des facteurs.

```r
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: log(breaks)
##              Df Sum Sq Mean Sq F value   Pr(>F)
## wool          1 0.3125 0.31253  2.2344 0.141511
## tension       2 2.1762 1.08808  7.7792 0.001185 **
## wool:tension  2 0.9131 0.45657  3.2642 0.046863 *
## Residuals    48 6.7138 0.13987
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Interprétation : Cette anova nous montre qu'au risque 5% on doit rejeter l'hypothèse nulle de la nullité des coefficients d'interactions entre les deux facteurs (même si la p-value est très ambiguë au risque 5%). Il semble donc y avoir des interactions significatives entre les facteurs ayant donc un effet sur la variable **breaks**.

Si l'on considérait un risque 1%, on ne rejetterait pas l'hypothèse nulle et donc on interpréterait les effets des facteurs (individuellement) de la manière suivante:

1. La p-value du facteur **wool** est supérieure à 0.01 donc on ne rejette pas l'hypothèse nulle de nullité du coefficient d'effet principal du facteur **wool**. En des termes plus clairs, le facteur **wool** n'a significativement pas d'effet sur la variable **breaks**.

2. A l'inverse du facteur **wool**, la p-value du facteur **tension** (0.0012) indique qu'il faut rejeter l'hypothèse nulle. Le facteur **tension** a donc un effet significatif sur la variable **breaks**.

Or, comme nous considérons un risque 5% ici, on ne peut plus du tout conclure la même chose. Comme il existe des interactions significatives entre les deux facteurs, les effets principaux (affichés ici) de chaque facteur en sont affectés et donc les résultats que nous avons ici sont ininterprétables. Pour pouvoir interpréter et quantifier les effets principaux des facteurs, il faudrait réaliser une anova à un facteur (tension par exemple) pour chaque modalité de wool fixée.

5. Le facteur **wool** présente une influence à travers son interaction avec le facteur **tension**. Comparer le modèle complet avec interaction avec le modèle sans le facteur **wool** en utilisant la commande "anova(mod1,mod2)". Que fait anova(mod1,mod2)?

```
model3=lm(log(breaks)~tension, data=donnees)
anova(model2,model3)
```

```
## Analysis of Variance Table
##
## Model 1: log(breaks) ~ wool + tension + wool:tension
## Model 2: log(breaks) ~ tension
##   Res.Df    RSS Df Sum of Sq     F  Pr(>F)
## 1     48 6.7138
## 2     51 7.9395 -3   -1.2257 2.921 0.04339 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La commande anova(model2,model3) permet de comparer les deux modèles en terme de qualité explicative (modélisation) des données. Cette commande permet de tester H0:"Le modèle 1 est meilleur que le modèle 2" vs H1:"Le modèle 2 est meilleur que le modèle 1".

Ici, au risque 5% la p-value est une fois de plus ambigue. Si l'on suit strictement la règle de décision alors on rejette l'hypothèse nulle, le modèle sans le facteur **wool** semble "meilleur" ou du moins explique au moins autant les données que le modèle complet. Il semblerait donc que même si le facteur **wool** a un effet sur **breaks** à travers son interaction avec le facteur **tension**, si ce facteur est exclus de l'analyse, cela ne change pas grand-chose.

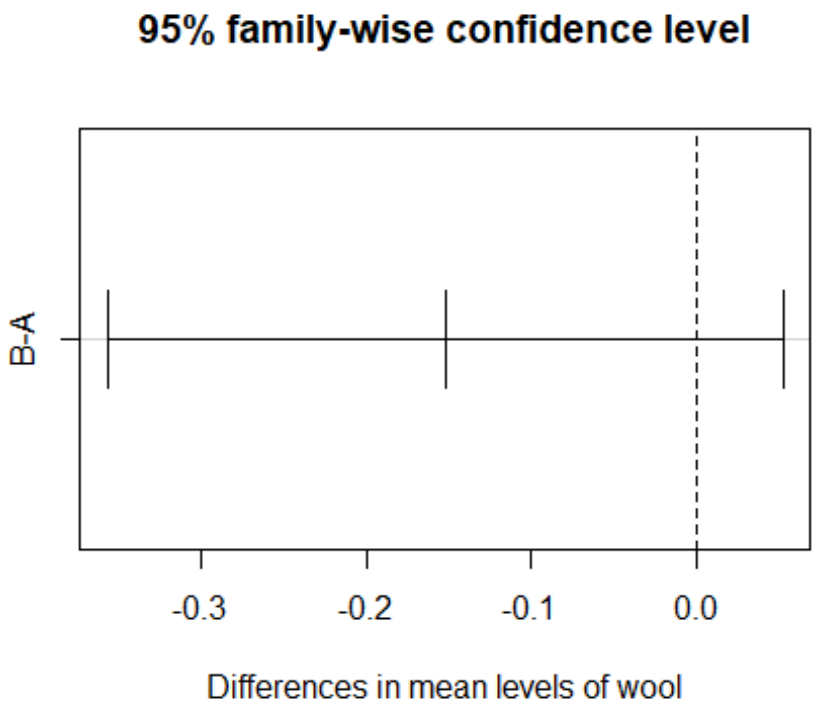
Si l'on se place au risque 1%, dans ce cas, on ne rejette pas l'hypothèse nulle et le facteur **wool** apporte bel et bien de l'information.

6. Comparer les modalités de **wool** puis **tension** deux à deux avec le test de Tukey. Les représenter graphiquement. Pourquoi ces comparaisons ne sont pas pertinentes voir erronées ici?

```
TukeyHSD(aov(model2),'wool')
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = model2)
##
## $wool
##           diff        lwr        upr     p adj
## B-A -0.1521536 -0.3568127 0.05250558 0.1415114
```

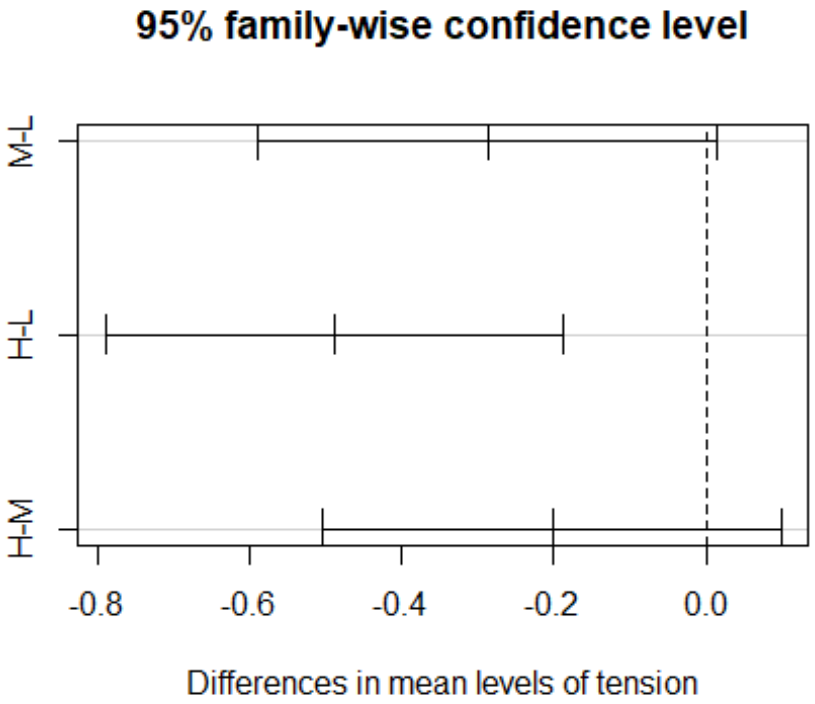
```
plot(TukeyHSD(aov(model2),'wool'))
```



```r
TukeyHSD(aov(model2),'tension')

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = model2)
##
## $tension
##           diff        lwr         upr     p adj
## M-L -0.2871237 -0.5886239  0.01437636 0.0649432
## H-L -0.4892747 -0.7907748 -0.18777463 0.0007948
## H-M -0.2021510 -0.5036511  0.09934912 0.2465032

plot(TukeyHSD(aov(model2),'tension'))
```



On voit ici que pour le facteur **wool**, on ne rejette pas H0 (au risque 5%), donc les moyennes des deux modalités ne sont pas significativement différentes. Le facteur **wool** n'a pas d'effet sur la variable **breaks**.

Pour le facteur **tension**, seules les modalités H et L présentent une différence significative de moyenne (au risque 5%). Les moyennes des modalités M et L et des modalités H et M sont significativement (et respectivement) égales d'après cet indicateur.

Or, ces comparaisons ne sont pertinentes que dans le cas où l'on a pas rejettée l'hypothèse de nullité des coefficients d'interaction entre facteurs. Or précédemment, nous avons rejeté au risque 5% cette hypothèse. Il y a donc des interactions entre les 2 facteurs et on ne peut plus interpréter et quantifier les effets principaux de chaque facteur de cette manière.

7. Comparer les interactions deux à deux.

```
TukeyHSD(aov(model2),'wool:tension')
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = model2)
##
## $`wool:tension`
##                   diff        lwr         upr     p adj
## B:L-A:L -0.4355668365 -0.9588143  0.08768059 0.1534713
## A:M-A:L -0.6011957092 -1.1244431 -0.07794828 0.0157632
## B:M-A:L -0.4086186238 -0.9318661  0.11462881 0.2071300
## A:H-A:L -0.6003226799 -1.1235701 -0.07707525 0.0159794
## B:H-A:L -0.8137936293 -1.3370411 -0.29054620 0.0004035
## A:M-B:L -0.1656288727 -0.6888763  0.35761856 0.9341161
## B:M-B:L  0.0269482127 -0.4962992  0.55019564 0.9999876
## A:H-B:L -0.1647558434 -0.6880033  0.35849159 0.9354994
## B:H-B:L -0.3782267929 -0.9014742  0.14502064 0.2822984
## B:M-A:M  0.1925770854 -0.3306703  0.71582452 0.8820529
## A:H-A:M  0.0008730293 -0.5223744  0.52412046 1.0000000
## B:H-A:M -0.2125979201 -0.7358454  0.31064951 0.8318529
## A:H-B:M -0.1917040562 -0.7149515  0.33154337 0.8840239
## B:H-B:M -0.4051750056 -0.9284224  0.11807242 0.2148594
## B:H-A:H -0.2134709494 -0.7367184  0.30977648 0.8294547
```

```
plot(TukeyHSD(aov(model2),'wool:tension'))
```

Si l'on compare 2 à 2 les interactions, on se rend compte que pour la modalité A de **wool** il y a une différence de moyenne significative pour les modalités M et L de **tension**, pareil pour les modalités H et L. De plus yBH est significtivement différente de yAL. Pour toutes les autres combinaisonq de modalités des deux facteurs, on ne peut pas considérer leurs moyennes comme significativement différentes. Ainsi l'effet des interacions des facteurs passe par les combinaisons explicitées ici.

8. Du fait des interactions, on peut comparer les modalités de **tension** conditionnellement à celles de **wool**. Que font les fonctions suivantes? Pourquoi? Interpéter.

   1.warpbreaks.wool=split(warpbreaks,wool)
   2.lbreaksA.aov=aov(log(breaks)~tension,data=warpbreaks.wool$A)
   3.lbreaksA.HSD <-TukeyHSD(lbreaksA.aov)

```
warpbreaks.wool=split(donnees,donnees$wool);warpbreaks.wool
```

```
## $A
##    breaks wool tension
## 1      26    A       L
## 2      30    A       L
## 3      54    A       L
## 4      25    A       L
## 5      70    A       L
## 6      52    A       L
```

```
## 7      51    A       L
## 8      26    A       L
## 9      67    A       L
## 10     18    A       M
## 11     21    A       M
## 12     29    A       M
## 13     17    A       M
## 14     12    A       M
## 15     18    A       M
## 16     35    A       M
## 17     30    A       M
## 18     36    A       M
## 19     36    A       H
## 20     21    A       H
## 21     24    A       H
## 22     18    A       H
## 23     10    A       H
## 24     43    A       H
## 25     28    A       H
## 26     15    A       H
## 27     26    A       H
##
## $B
##    breaks wool tension
## 28     27    B       L
## 29     14    B       L
## 30     29    B       L
## 31     19    B       L
## 32     29    B       L
## 33     31    B       L
## 34     41    B       L
## 35     20    B       L
## 36     44    B       L
## 37     42    B       M
## 38     26    B       M
## 39     19    B       M
## 40     16    B       M
## 41     39    B       M
## 42     28    B       M
## 43     21    B       M
## 44     39    B       M
## 45     29    B       M
## 46     20    B       H
## 47     21    B       H
## 48     24    B       H
## 49     17    B       H
## 50     13    B       H
## 51     15    B       H
## 52     15    B       H
```

```
## 53     16    B       H
## 54     28    B       H
```

Cette commande permet de séparer le dataframe des données suivant les modalités du facteur choisi. Ici le facteur choisi étant **wool**, le dataframe a été séparé en 2. Un dataframe ne contenant que les enregistrements ayant pour modalité A (wool) mais toutes les autres modalités de **tension**. Un deuxième dataframe qui lui contient tous les enregistrements ayant B comme modalité. Cette commande est très utile car elle va permettre de réaliser une anova à un facteur (le facteur **tension** ici) en fixant une modalité du deuxième facteur (wool). On va donc pouvoir quantifier les effets principaux de chaque facteur conditionnellement à la modalité fixée de l'autre.

```
lbreaksA.aov=aov(log(breaks)~tension,data=warpbreaks.wool$A)
summary(lbreaksA.aov)
```

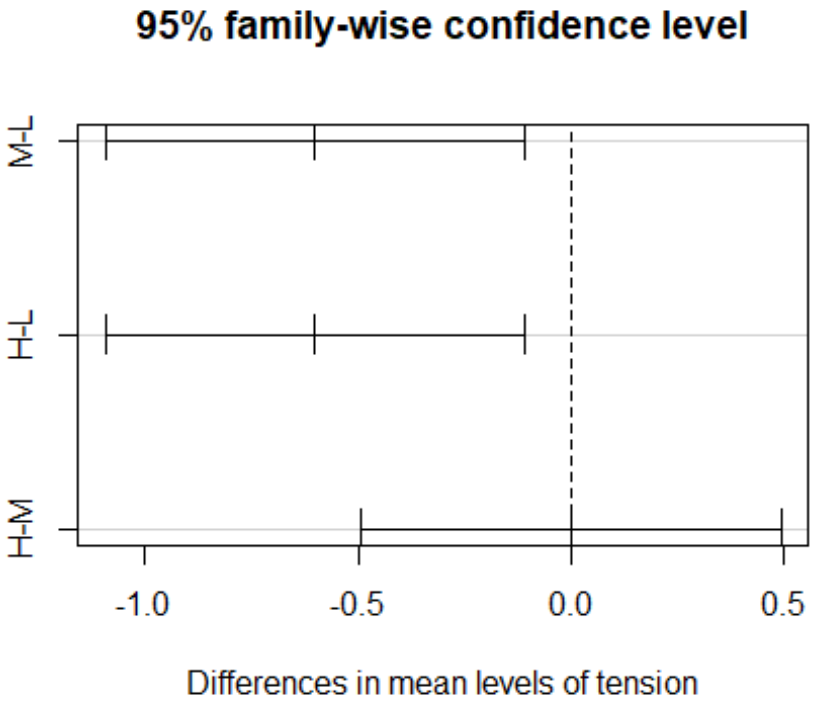
```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## tension      2  2.165  1.0827   6.194 0.00678 **
## Residuals   24  4.195  0.1748
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cette commande réalise l'analyse de variance à un facteur (tension) en ne prenant que les enregistrements ayant A comme modalité pour **wool**. On peut alors interpréter les résultats conditionnellment à A. Le résumé de l'anova nous indique qu'on rejette l'hypothèse nulle d'égalité des moyennes des modalités de **tension**, conditionnellement à A. Le facteur **tension**|wool=A a donc un effet sur la variable **breaks**. Nous allons maintenant regarder quelles moyennes sont significativement différentes.

```
lbreaksA.HSD=TukeyHSD(lbreaksA.aov);lbreaksA.HSD
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = log(breaks) ~ tension, data = warpbreaks.wool$A)
##
## $tension
##              diff        lwr        upr     p adj
## M-L -0.6011957092 -1.0933930 -0.1089984 0.0146660
## H-L -0.6003226799 -1.0925200 -0.1081254 0.0148170
## H-M  0.0008730293 -0.4913243  0.4930703 0.9999892
```

```
plot(lbreaksA.HSD)
```



On voit donc grâce à cette commande que yL est significativement différente de yM et yH mais que yM et yH ne peuvent pas être considérées comme significativement différentes, toujours conditionnellement à wool=A.

Interprétation: On remarque tout d'abord que ce graphique ne correspond plus du tout au graphique de la question 6 où seules yH et yL étaient significativement différentes. C'est normal, car on prend ici en compte les interactions des facteurs. Ce graphique confirme ce qui a été dit à la question précédente: le facteur **tension** a donc un effet significatif sur la variable **breaks** à travers ces différences de moyennes, conditionnellement à wool=A.

On peut donc prédire la réaction d'une laine en terme d'usure grâce à ces résultats: si l'on a 2 laines de type A et que l'on applique à l'une d'elle une tension H et à l'autre une tension M, alors on sait qu'il n'y aura normalement aucune différence particulière pour les mesures de **breaks** enregistrées. En revanche si l'on applique une tension H à la première et une tension L à la seconde, on doit s'attendre à une différence dans les mesures de **breaks** observées. Sachant que yAL est plus élevée que yAH, on doit s'attendre à une valeur de **breaks** plus élevée pour la laine subissant la tension L que pour la laine subissant la tension H.

9. Même question que la question 8 mais avec les commandes suivantes. On procédera de préférence en utilisant le modèle complet pour optimiser l'estimation de la variance résiduelle:
    1. lbreaks.cond<-emmeans(lbreaks.aov,~tension|wool)
    2. pairs(lbreaks.cond); cld(lbreaks.cond)

```
library(emmeans)

library(multcomp)

lbreaks.aov=aov(log(breaks)~tension+wool+tension:wool,data=donnees)
lbreaks.cond=emmeans(lbreaks.aov,~tension|wool)
pairs(lbreaks.cond)

## wool = A:
##  contrast  estimate    SE df t.ratio p.value
##  L - M     0.601196 0.176 48  3.410  0.0037
##  L - H     0.600323 0.176 48  3.405  0.0038
##  M - H    -0.000873 0.176 48 -0.005  1.0000
##
## wool = B:
##  contrast  estimate    SE df t.ratio p.value
##  L - M    -0.026948 0.176 48 -0.153  0.9872
##  L - H     0.378227 0.176 48  2.145  0.0914
##  M - H     0.405175 0.176 48  2.298  0.0657
##
## Results are given on the log (not the response) scale.
## P value adjustment: tukey method for comparing a family of 3 estimates
```

La commande pairs(lbreaks.cond) permet de réaliser les anova à un facteur en fixant les modalités du deuxième facteur. Elle permet donc de quantifier et d'interpréter les effets principaux de chaque facteur lorsqu'il y a des interactions significatives entre facteurs.

On retrouve ici que pour la modalité A fixée de **wool**, les moyennes des modalités M et H sont significativement égales alors que yL!=yH et yL!=yM.

Pour la modalité B fixée de **wool**, on remarque qu'on ne rejette aucune hypothèse nulle. Les moyennes des 3 modalités L, M et H sont toutes significativement égales. Le facteur **tension** n'a donc aucune influence sur la variable **breaks** conditionnellement à wool=B.

Le critère utilisé ici étant le critère de Tukey.

```
cld(lbreaks.cond)

## wool = A:
##  tension emmean    SE df lower.CL upper.CL .group
##  M         3.12 0.125 48     2.87     3.37  1
##  H         3.12 0.125 48     2.87     3.37  1
##  L         3.72 0.125 48     3.47     3.97   2
##
```

```
## wool = B:
##  tension emmean    SE df lower.CL upper.CL .group
##  H         2.90 0.125 48     2.65     3.15  1
##  L         3.28 0.125 48     3.03     3.53  1
##  M         3.31 0.125 48     3.06     3.56  1
##
## Results are given on the log (not the response) scale.
## Confidence level used: 0.95
## P value adjustment: tukey method for comparing a family of 3 estimates
## significance level used: alpha = 0.05
```

Cette commande donne les groupes auquels appartiennent les différentes modalités d'un facteur en fonction de leur différence de moyenne (quantifié par la méthode de Tukey), comme il avait été vu dans le cours sur l'anova à un facteur avec les différentes espèces de fleurs.

Ici, conditionnellement à wool=A, l'écart de moyenne entre la modalité M et la modalité H du facteur **tension** n'est pas assez grand pour placer la modalité M dans un groupe différent de celui de H (groupe 1 par défaut) alors que l'écart de moyenne entre la modalité M et L est suffisament grand pour placer la modalité L dans un second groupe (groupe 2).

Conditionnellement à wool=B, toutes les modalités sont plaçées dans le même groupe. En effet, par défaut la modalité H est plaçée dans le groupe 1, puis la modalité M n'ayant pas une moyenne significativement différente de celle de H, elle est aussi plaçée dans le groupe 1. Et ce, de même pour la modalité L dont la moyenne n'est pas significativement différente de celle de M.

Conclusion: Cette analyse de variance à 2 facteurs à mis en évidence des interactions significatives entre les 2 facteurs **wool** et **tension**. Ainsi, pour quantifier et interpréter les effets principaux de chaque facteur, nous avons fixé tour après tour les modalités de **wool** pour étudier les effets du facteur **tension** sur la variable **breaks**.

Ce qui ressort de cette analyse est que si le type de laine est A, alors on doit/peut s'attendre à une différence de réaction de la laine à l'usure suivant la tension appliquée, lorsque l'on compare deux laines de même type (A en l'occurence) et qu'on leur applique des tensions différentes. Les différences auront lieu si on applique une tension L et H ou un tension L et M mais pas lorsqu'on applique une tension H et M.

Si le type de laine est B, alors d'après cette étude, il n'y aura aucune différence de réaction de la laine à l'usure suivant les différentes tensions appliquées car l'étude n'a montré aucune différence significative de moyenne entre les modalités de **tension**|wool=B.