TP3 : modèle ANOVA

DJEBALI Wissam

1 mars 2018

Modèle ANOVA : Régression linéaire avec variables qualitatives

Packages R : MASS

L'analyse de la variance, c'est la régression quand les variables prédictives sont qualitatives plus précisément, il s'agit des tests qu'on effectue dans ce cadre.

L'analyse de la covariance, c'est la régression quand certaines des variables prédictives sont qualitatives et d'autres quantitatives (on appelle alors ces dernières covariables).

La manière la plus simple de voir l'analyse de la variance, c'est comme une généralisation du test de Student : elle permet de voir si la moyenne d'une variable quantitative est la même dans différents groupes ou, en d'autres termes, si une variable quantitative dépend d'une variable qualitative.

Une manière plus générale de voir l'analyse de la variance (et c'est le point de vue adopté par la commande "anova" sous R), c'est comme un test comparant plusieurs modèles.

L'ANOVA s'applique dès que :

_ on veut monter une expérimentation

_ on veut montrer l'effet de variables qualitatives sur une variable quantitative

ANOVA à un facteur (One way ANOVA)

Ici on va étudier les données bee.

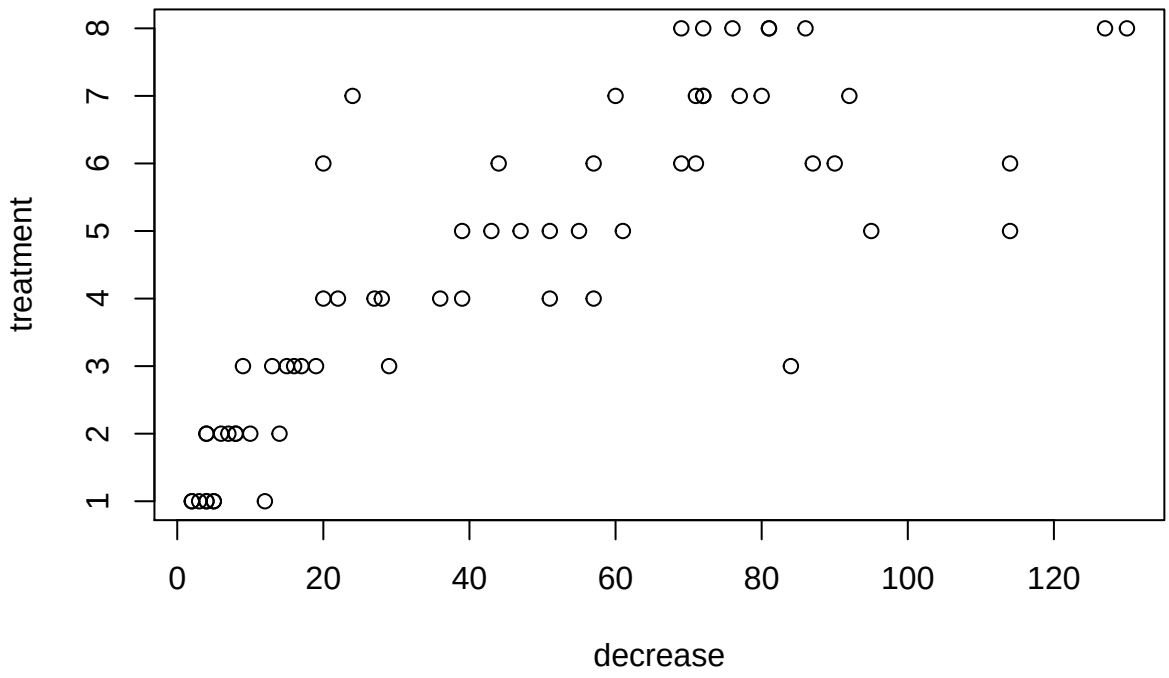
Les groupes seront données par les différentes modalités de traitement :

- groupe traitement A
- groupe traitement B
- groupe traitement C
- groupe traitement D
- groupe traitement E
- groupe traitement F
- groupe traitement G
- groupe traitement H

```
bee<-data.frame(OrchardSprays)
bee<-bee[,c(1,4)]

# Visualisation des donées bee
# bee$treatment est une variable qualitative en facteur
plot(bee)

# On transfrome la variable treatment en variable auantitative
bee2<-bee
bee2$treatment<-as.numeric(bee2$treatment)
# Visualisation des donées bee
# bee$treatment est une variable quantitative en numérique maintenant
plot(bee2)
```



Le changement d'interprétation de la variable treatment en variable numérique n'a rien changé au lien entre les variables, donc autant garder la variable en facteur

On a toujours un lien linéaire entre decrease et treatment, vu lorque la variable treatment était sous forme de facteur.

Etude de la moyenne des groupes(différentes modalités de bee$treatment)

```
reg<-lm(bee$decrease~bee$treatment)
summary(reg)
```

```
##
## Call:
## lm(formula = bee$decrease ~ bee$treatment)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.000  -9.500  -1.625   3.812  58.750
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.625      7.253   0.638  0.52631
## bee$treatmentB    3.000     10.258   0.292  0.77101
## bee$treatmentC   20.625     10.258   2.011  0.04918 *
## bee$treatmentD   30.375     10.258   2.961  0.00449 **
## bee$treatmentE   58.500     10.258   5.703 4.60e-07 ***
## bee$treatmentF   64.375     10.258   6.276 5.39e-08 ***
```

```
## bee$treatmentG   63.875     10.258   6.227 6.48e-08 ***
## bee$treatmentH   85.625     10.258   8.347 2.08e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.52 on 56 degrees of freedom
## Multiple R-squared:  0.7044, Adjusted R-squared:  0.6674
## F-statistic: 19.06 on 7 and 56 DF,  p-value: 9.499e-13
```

```
# Anova sur la régression linéaire
anov<-anova(reg)
anov
```

```
## Analysis of Variance Table
##
## Response: bee$decrease
##               Df Sum Sq Mean Sq F value    Pr(>F)
## bee$treatment  7  56160  8022.9  19.062 9.499e-13 ***
## Residuals     56  23570   420.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lors de l'analyse des moyennes, on a que la p-value est faible (<0.05) donc on rejète H_0 = *"les moyennes des groupes sont égales"*, donc la variable *treatment* a bien une influence sur la variable *decrease*.

Test des moyennes deux à deux

```
# 5) Test deux à deux
A<-bee[which(bee$treatment=="A"),]
B<-bee[which(bee$treatment=="B"),]
C<-bee[which(bee$treatment=="C"),]
D<-bee[which(bee$treatment=="D"),]

# A contre B
t.test(A$decrease-B$decrease)
```

```
##
##  One Sample t-test
##
## data:  A$decrease - B$decrease
## t = -1.5067, df = 7, p-value = 0.1756
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -7.708088  1.708088
## sample estimates:
## mean of x
##        -3
```

```
# La p-value est supérieur à 0.05 donc on ne rejète pas
# H0 = A et B ont leur moyenne proche
```

Etude de la variance des groupes(différentes modalités de bee$treatment)

```
# Test de barlett ou Test d'homogénéité
bartlett.test(bee$decrease~bee$treatment)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  bee$decrease by bee$treatment
## Bartlett's K-squared = 42.031, df = 7, p-value = 5.128e-07
```

```
# Le Test de Barlett n'est pas très robuste si les données ne sont pas de loi normale, mieux vaut utili
```

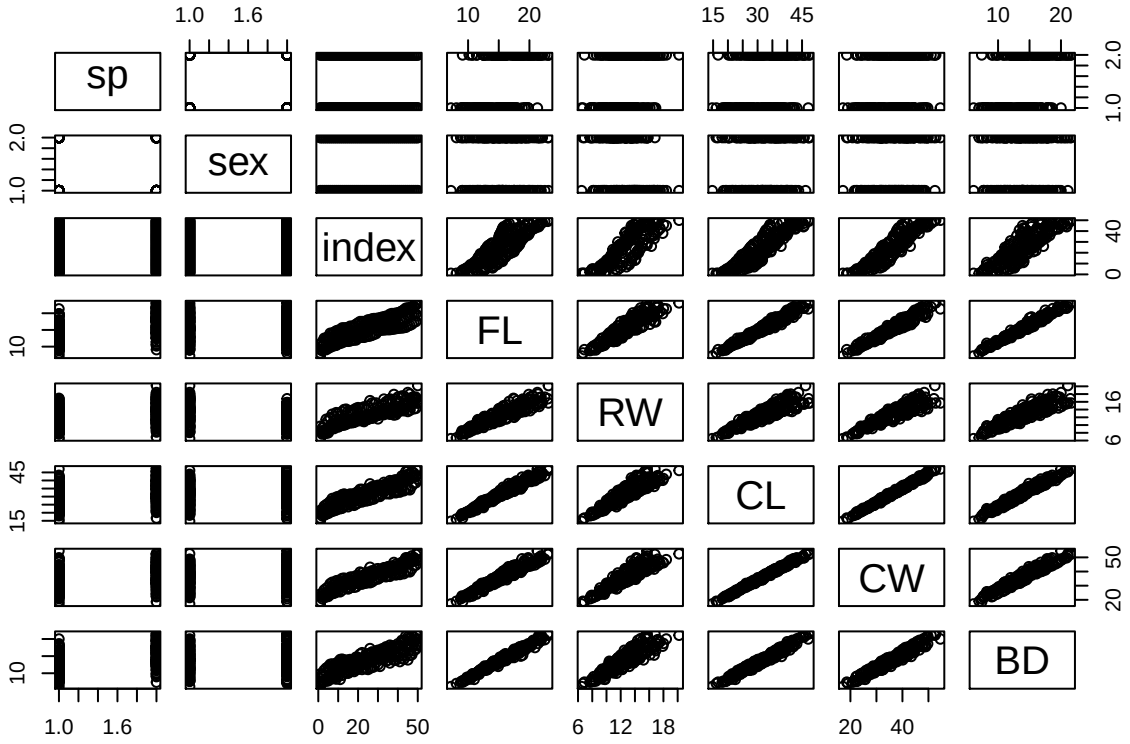
Lors de l'analyse des variances, on a que la p-value est faible on rejète H_0 =*les variances de chaque groupe sont homogènes*

ANOVA à facteurs multiples

Ici on va étudier les données crabs, on va étudier la variable body depth *BD* en fonction du facteur sex *Sex* et espèce *Sp*

```
crab<-crabs

plot(crab)
```



Préquel : Analyse ANOVA à un facteur avec les différentes variables qualitatives

```
# Espèce
reg_sp<-lm(crab$BD~crab$sp)
summary(reg_sp)
```

```
##
```

```
## Call:
## lm(formula = crab$BD ~ crab$sp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.0780 -2.1830  0.0695  2.3170  7.4170
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.5830     0.3110  40.460  < 2e-16 ***
## crab$spO      2.8950     0.4398   6.582 4.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.11 on 198 degrees of freedom
## Multiple R-squared:  0.1795, Adjusted R-squared:  0.1754
## F-statistic: 43.33 on 1 and 198 DF,  p-value: 4.06e-10
```

```
# L'espèce O influe le plus sur la variable BD, contrairement à l'espèce B

anov_sp<-anova(reg_sp)
anov_sp
```

```
## Analysis of Variance Table
##
## Response: crab$BD
##            Df  Sum Sq Mean Sq F value   Pr(>F)
## crab$sp     1  419.05  419.05  43.327 4.06e-10 ***
## Residuals 198 1915.03    9.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# on a que la p-value est faible (<0.05) donc on rejète H0=les moyennes de groupes sont égales. Donc l'


# Sex
reg_sex<-lm(crab$BD~crab$sex)
summary(reg_sex)
```

```
##
## Call:
## lm(formula = crab$BD ~ crab$sex)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.624 -2.449  0.076  2.463  7.376
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.7240     0.3420  40.134   <2e-16 ***
## crab$sexM     0.6130     0.4836   1.268    0.206
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.42 on 198 degrees of freedom
```

```
## Multiple R-squared:  0.00805,    Adjusted R-squared:  0.00304
## F-statistic: 1.607 on 1 and 198 DF,  p-value: 0.2064
```

```
# Le sexe M influe le plus sur la variable BD, contrairement au sexe F

anov_sex<-anova(reg_sex)
anov_sex
```

```
## Analysis of Variance Table
##
## Response: crab$BD
##            Df  Sum Sq Mean Sq F value Pr(>F)
## crab$sex    1   18.79  18.788  1.6068 0.2064
## Residuals 198 2315.30  11.693
```

```
# On a que la p-value est est de 0.20 (>0.05) donc on ne rejète pas HO=les moyennes des groupes sont ég
```

La **p value (Pr(>F))** de *anov_sp* est faible <0.05 alors que la **p value (Pr(>F))** de *anov_sex* est > 0.05.

On a donc l'impression que seul l'espèce influe sur la var BD.

ANOVA à deux facteurs

```
reg_crab<-lm(BD~sex+sp+sex:sp,data=crab)
summary(reg_crab)
```

```
##
## Call:
## lm(formula = BD ~ sex + sp + sex:sp, data = crab)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.924 -2.224  0.059  2.250  6.650
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.8160     0.4349  27.167  < 2e-16 ***
## sexM          1.5340     0.6151   2.494   0.0135 *
## spO           3.8160     0.6151   6.204 3.21e-09 ***
## sexM:spO     -1.8420     0.8699  -2.118   0.0355 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.075 on 196 degrees of freedom
## Multiple R-squared:  0.2058, Adjusted R-squared:  0.1936
## F-statistic: 16.93 on 3 and 196 DF,  p-value: 8.131e-10
```

```
anov_crab<-anova(reg_crab)
anov_crab
```

```
## Analysis of Variance Table
##
## Response: BD
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## sex         1   18.79   18.79  1.9864   0.16030
## sp          1  419.05  419.05 44.3050 2.751e-10 ***
## sex:sp      1   42.41   42.41  4.4841   0.03547 *
```

```
## Residuals 196 1853.83    9.46
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La **p-value (Pr(>F))** de l'intéraction des variables sex et sp est <0.05 (sex :sp 0.03547).

Donc on rejète H_0 =*"les moyennes des groupes sont égales"*, donc l'intéraction des variables qualitatives *sex* et *sp* a bien une influence sur la variable *BD*.

D'où la variable sex a bien aussi une influence sur la variable *BD*, même si celle-ci est légère, chose que l'on ne pouvait pas voir précédemment.