Codes R pour jeu de données Eucalyptus

Sophie Donnet

20/01/2020

1. Jeu de données

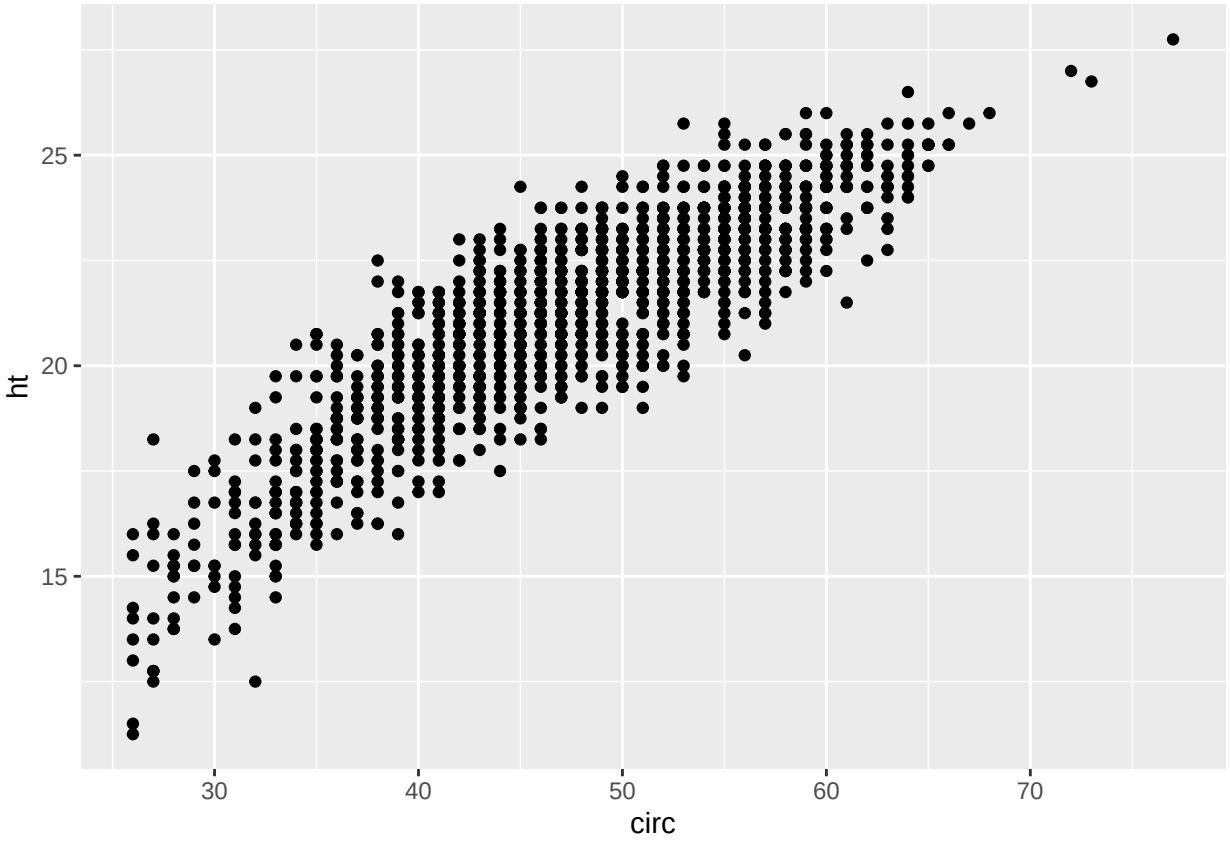
```
euca<-read.table("eucalyptus.txt",header=T,sep=" ")
str(euca)
```

```
## 'data.frame':    1429 obs. of  3 variables:
##  $ ht  : num  18.2 19.8 16.5 18.2 19.5 ...
##  $ circ: int  36 42 33 39 43 34 37 41 27 30 ...
##  $ bloc: Factor w/ 3 levels "A1","A2","A3": 1 1 1 1 1 1 1 1 1 1 ...
```

```
names(euca)
```

```
## [1] "ht"   "circ" "bloc"
```

```
gg_euca <- ggplot(euca, aes(x = circ, y = ht)) + geom_point()
gg_euca
```



2. Modèle linéaire à une variable (chapitre 1)

Estimation

```
reg <- lm(ht~circ,data=euca)
summary(reg)
```

```
##
## Call:
## lm(formula = ht ~ circ, data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7659 -0.7802  0.0557  0.8271  3.6913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.037476   0.179802   50.26   <2e-16 ***
## circ        0.257138   0.003738   68.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 1427 degrees of freedom
## Multiple R-squared:  0.7683, Adjusted R-squared:  0.7682
## F-statistic:  4732 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Objet de sortie

```
names(reg)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

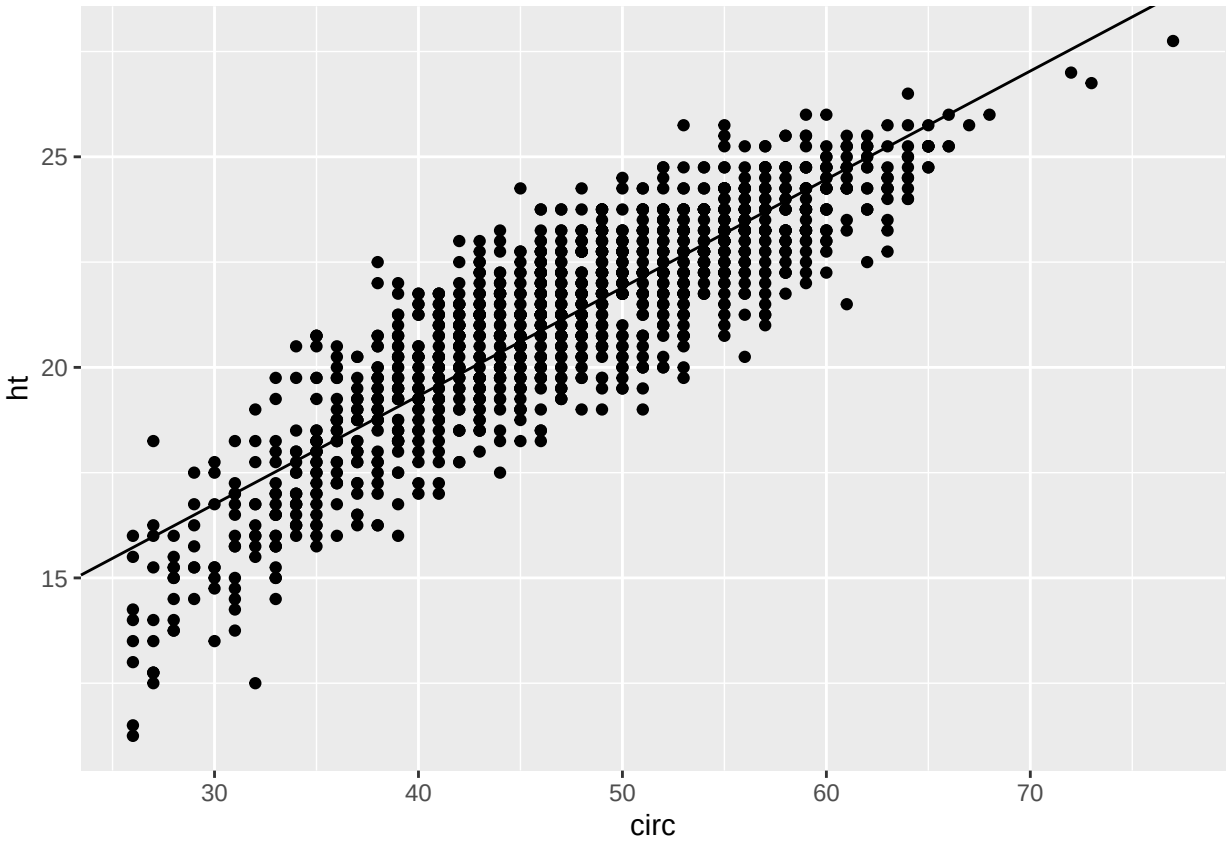
Coefficient de régression

```
coef(reg)
```

```
## (Intercept)        circ
##   9.0374757   0.2571379
```

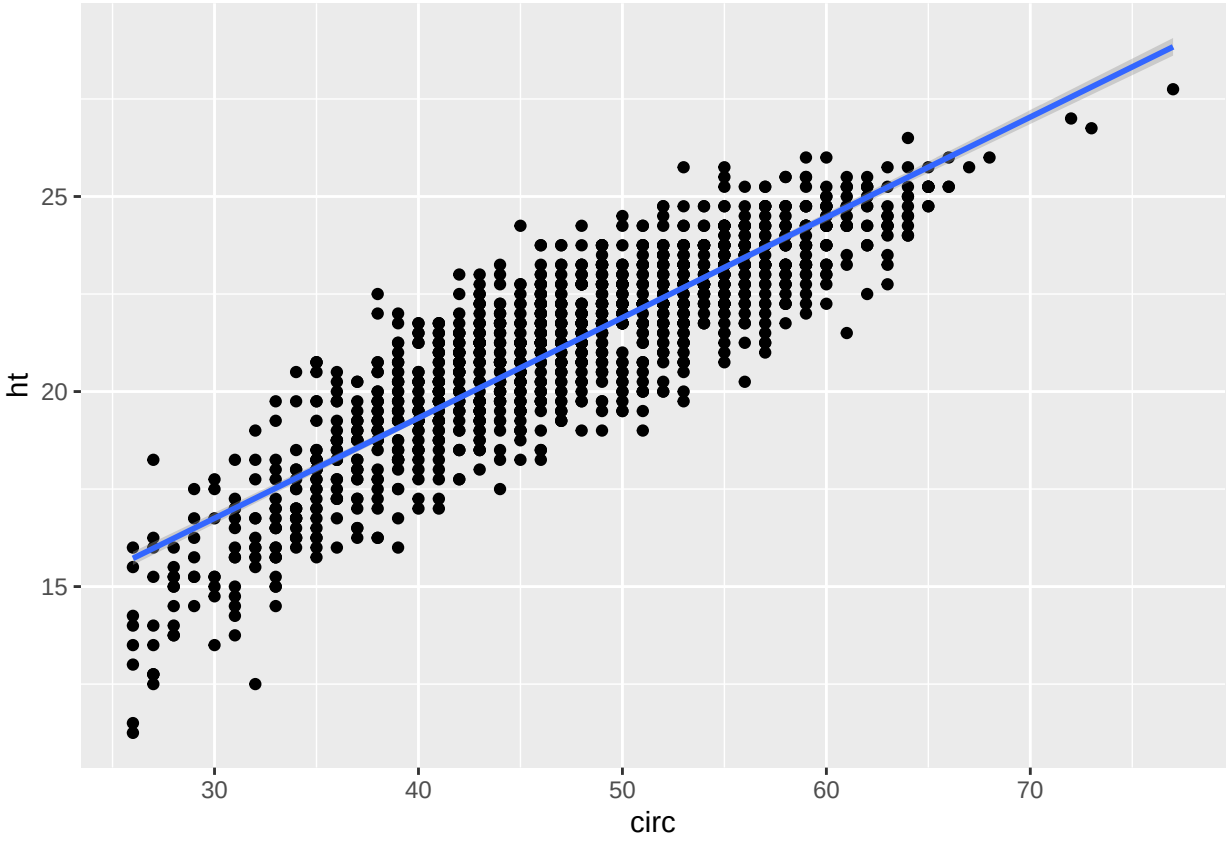
Tracé de la droite de régression : méthode 1

```
gg_euca + geom_abline(intercept=coef(reg)[1], slope = coef(reg)[2])
```



Méthode 2 (par ggplot2)

```
gg_euca + geom_smooth(method='lm')
```



Intervalles de confiances

```
confint(reg)
```

```
##                 2.5 %    97.5 %
## (Intercept) 8.6847719 9.3901795
## circ        0.2498055 0.2644702
```

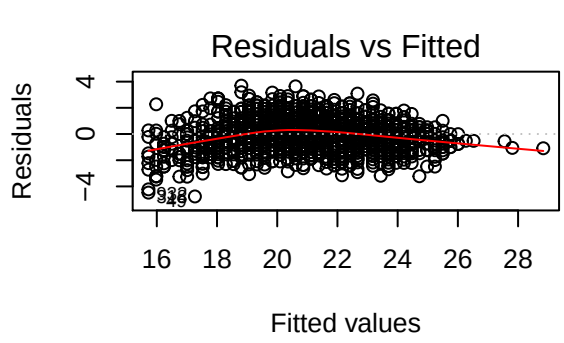
```
confint(reg,level=0.97)
```

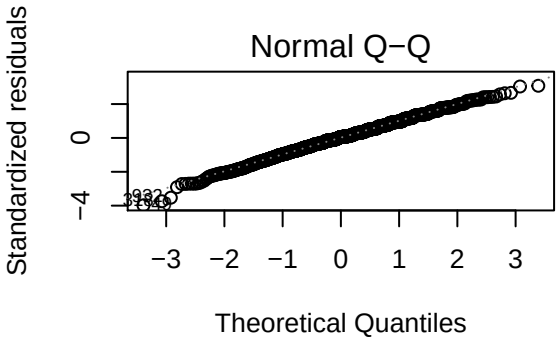
```
##                 1.5 %    98.5 %
## (Intercept) 8.6468993 9.4280520
## circ        0.2490182 0.2652575
```

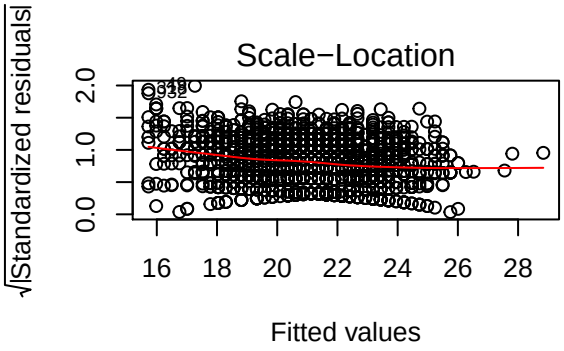
Validation de modèle

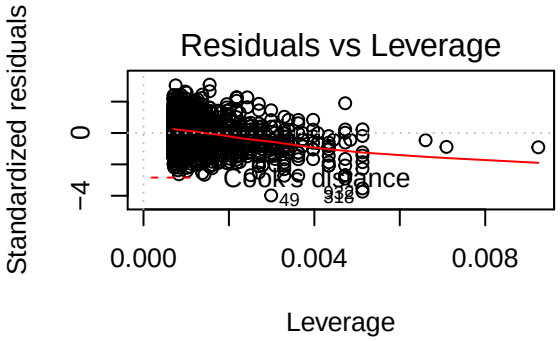
Par la lecture des résidus

```
par(mfrow=c(2,2))
plot(reg)
```









Prédiction

```
xnew=c(46,35,67)
xnew=data.frame(circ=xnew)
predict(reg,new=xnew)
```

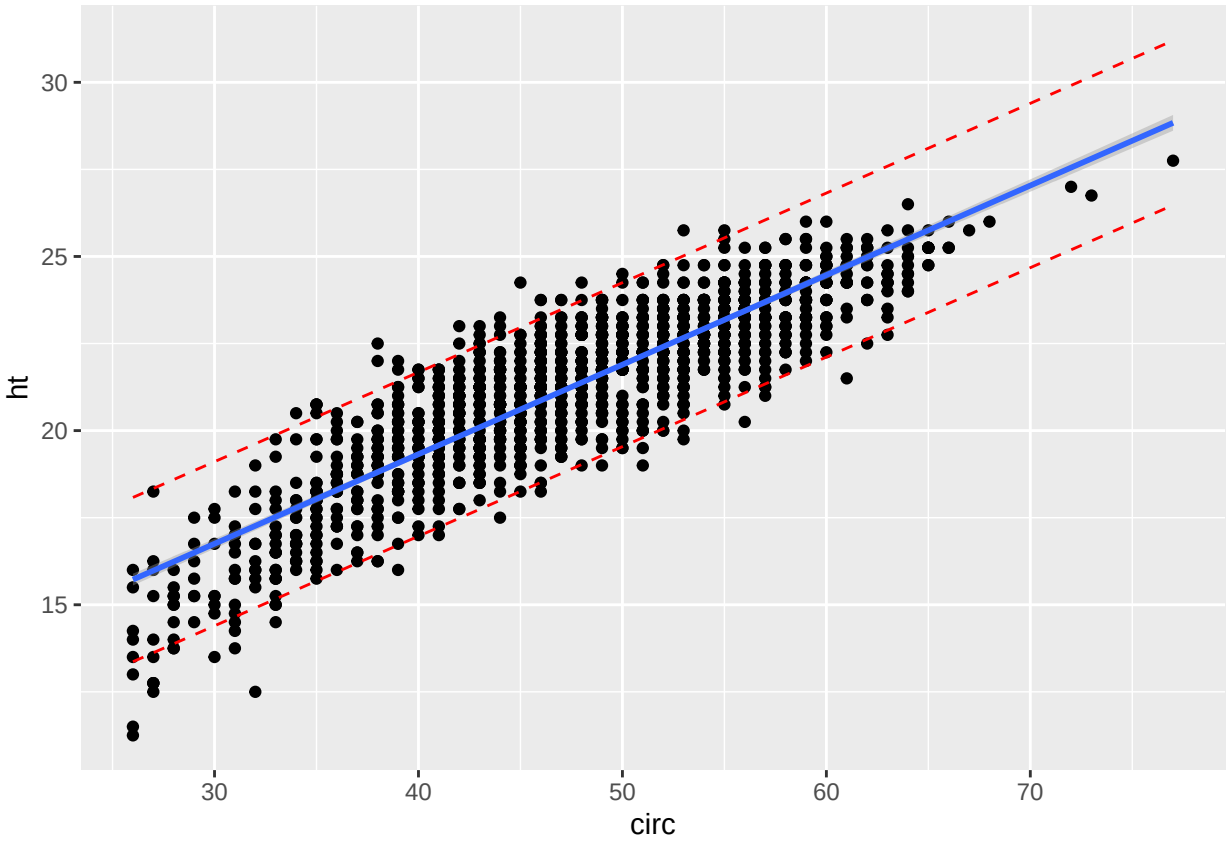
```
##        1        2        3
## 20.86582 18.03730 26.26571
```

```
predict(reg,new=xnew,se.fit=T)
```

```
## $fit
##        1        2        3
## 20.86582 18.03730 26.26571
##
## $se.fit
##          1          2          3
## 0.03212020 0.05600524 0.08001489
##
## $df
## [1] 1427
##
## $residual.scale
## [1] 1.199183
```

```
xnew <- seq(min(euca$circ), max(euca$circ),len = 1000)
xnew <- data.frame(xnew); names(xnew) = 'circ'
ICpred <- as.data.frame(predict(reg,xnew,interval="pred",level=0.95))
res_pred <- cbind(ICpred,xnew)
```

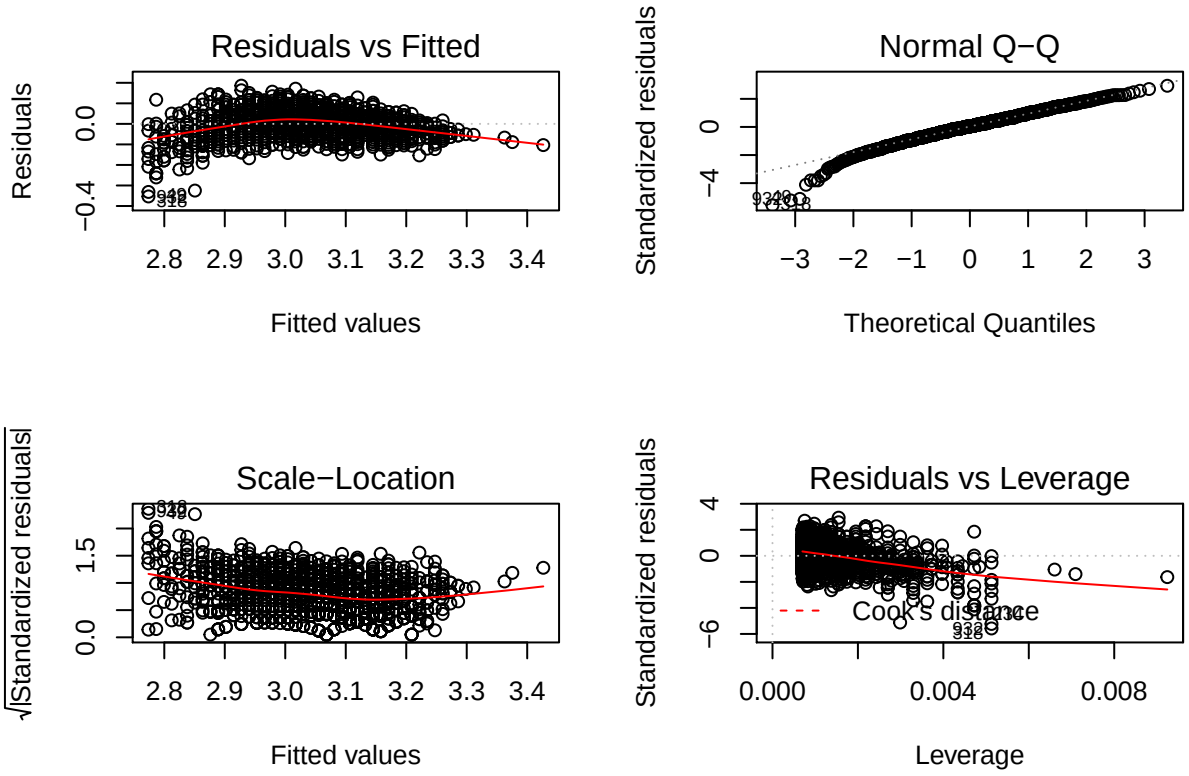
```
ggplot(euca, aes(x = circ, y = ht)) + geom_point() + geom_smooth(method = lm) + geom_line(data=res_pred
```



Passage au log pour

On cherche à améliorer le graphe des résidus

```
reg <- lm(log(ht)~circ,data=euca)
par(mfrow=c(2,2))
plot(reg)
```



#3. Modèle de regression pour plusieurs variables

```
modele0=lm(ht~circ,data=euca)
modele1=lm(ht~circ+I(sqrt(circ)),data=euca)
anova(modele0,modele1)
```

```
## Analysis of Variance Table
##
## Model 1: ht ~ circ
## Model 2: ht ~ circ + I(sqrt(circ))
##   Res.Df    RSS Df Sum of Sq     F    Pr(>F)
## 1   1427 2052.1
## 2   1426 1840.7  1    211.43 163.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(modele1)
```

```
##
## Call:
## lm(formula = ht ~ circ + I(sqrt(circ)), data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1881 -0.6881  0.0427  0.7927  3.7481
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -24.35200    2.61444  -9.314   <2e-16 ***
## circ           -0.48295    0.05793  -8.336   <2e-16 ***
## I(sqrt(circ))   9.98689    0.78033  12.798   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.136 on 1426 degrees of freedom
## Multiple R-squared:  0.7922, Adjusted R-squared:  0.7919
## F-statistic:  2718 on 2 and 1426 DF,  p-value: < 2.2e-16
```

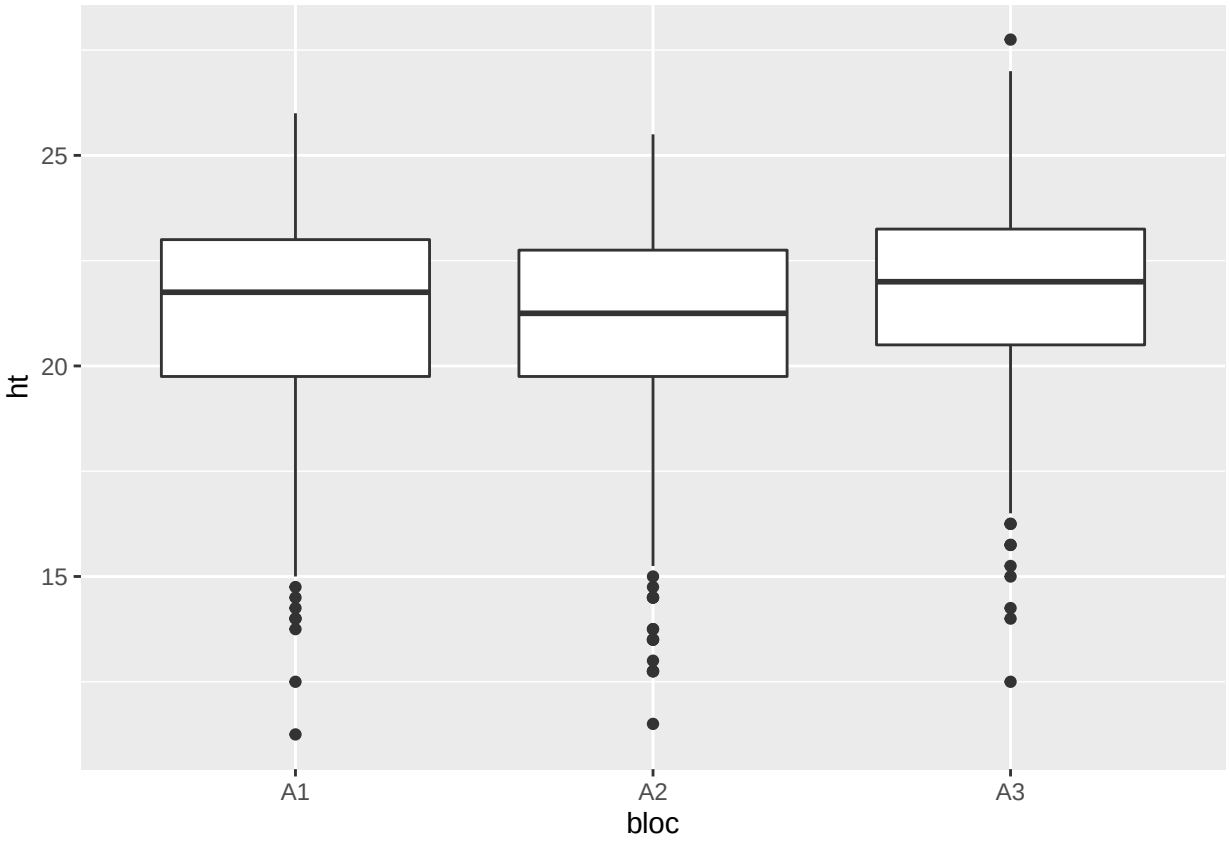
#4. Modèle d'anova à un facteur (Chap 4)

```
is.factor(euca$bloc)
```

```
## [1] TRUE
```

Boxplot des hauteurs par zone

```
library(ggplot2)
p <- ggplot(data=euca, aes(x= bloc, y=ht)) + geom_boxplot()
p
```




```
res_anova <- lm(ht ~ bloc, data = euca)
summary(res_anova)
```

```
##
## Call:
## lm(formula = ht ~ bloc, data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9179 -1.4179  0.3321  1.7905  5.9945
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.1679     0.1078 196.446  < 2e-16 ***
## blocA2       -0.2085     0.1485  -1.404 0.160610
## blocA3        0.5876     0.1760   3.339 0.000863 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.474 on 1426 degrees of freedom
## Multiple R-squared:  0.01487,    Adjusted R-squared:  0.01349
## F-statistic: 10.77 on 2 and 1426 DF,  p-value: 2.288e-05
```

```
test <- anova(res_anova)
test
```

```
## Analysis of Variance Table
##
## Response: ht
##             Df Sum Sq Mean Sq F value    Pr(>F)
## bloc         2  131.7  65.875  10.766 2.288e-05 ***
## Residuals 1426 8725.7   6.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(lsmeans)
```

```
## Loading required package: emmeans
```

```
## The 'lsmeans' package is now basically a front end for 'emmeans'.
## Users are encouraged to switch the rest of the way.
## See help('transition') for more information, including how to
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.
```

4. Modèle d'Ancova

```
mod1 = lm(ht~circ*bloc,data=euca) #le  modèle complet
summary(mod1)
```

```
##
## Call:
## lm(formula = ht ~ circ * bloc, data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7723 -0.7037  0.0539  0.8114  3.3255
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.031e+00  2.895e-01  31.199   <2e-16 ***
## circ         2.576e-01  6.043e-03  42.618   <2e-16 ***
## blocA2      -1.850e-01  4.044e-01  -0.457    0.647
## blocA3       6.165e-01  4.766e-01   1.294    0.196
## circ:blocA2 -1.927e-05  8.454e-03  -0.002    0.998
## circ:blocA3 -6.834e-03  9.802e-03  -0.697    0.486
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.187 on 1423 degrees of freedom
## Multiple R-squared:  0.7736, Adjusted R-squared:  0.7728
## F-statistic: 972.6 on 5 and 1423 DF,  p-value: < 2.2e-16
```

```
mod2 = lm(ht~bloc+circ,data=euca)# modèle sans interaction
anova(mod2,mod1)
```

```
## Analysis of Variance Table
##
## Model 1: ht ~ bloc + circ
## Model 2: ht ~ circ * bloc
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   1425 2005.9
## 2   1423 2005.0  2   0.84752 0.3007 0.7403
```

```
mod3 = lm(ht~ circ,data=euca)# modèle avec seulement circ
anova(mod3,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: ht ~ circ
## Model 2: ht ~ bloc + circ
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1427 2052.1
## 2   1425 2005.9  2    46.188 16.406 9.031e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```