ER01

Analyse de l'influence de l'origine géographique de tournesols sur leur teneur en acide gras

Alice Joffard, Antonin Lambilliotte, Vincent Michielini, Laureline Pinault

Lyon, December 2, 2016

1 Description des données

Les données étudiées représentent la teneur en acide gras, en pourcentage de masse, de tournesols de différentes origines (Afrique du sud, Hongrie et Maroc), mesurée avec deux testeurs différents. Le jeu de donnée présente trois points par combinaison des deux facteurs origine et testeur.Le but principal est de savoir quelle est l'influence de l'origine des tournesols sur leur teneur en acide gras.

2 Analyse des données

On commence par charger les données sur RStudio à partir du fichier "tournesol.txt" :

```
data=read.table("tournesol.txt",sep="\t",h=T,dec=".")
teneur=data$teneur
testeur=data$testeur
origine=data$origine
```

2.1 Influence de l'origine

On commence par représenter les données à l'aide d'un graphique montrant l'influence de l'origine des tournesols sur leur teneur en acide gras, indépendamment du testeur utilisé pour la mesure :

```
ggplot(data,aes(x=origine, y=teneur))+geom_point()
```

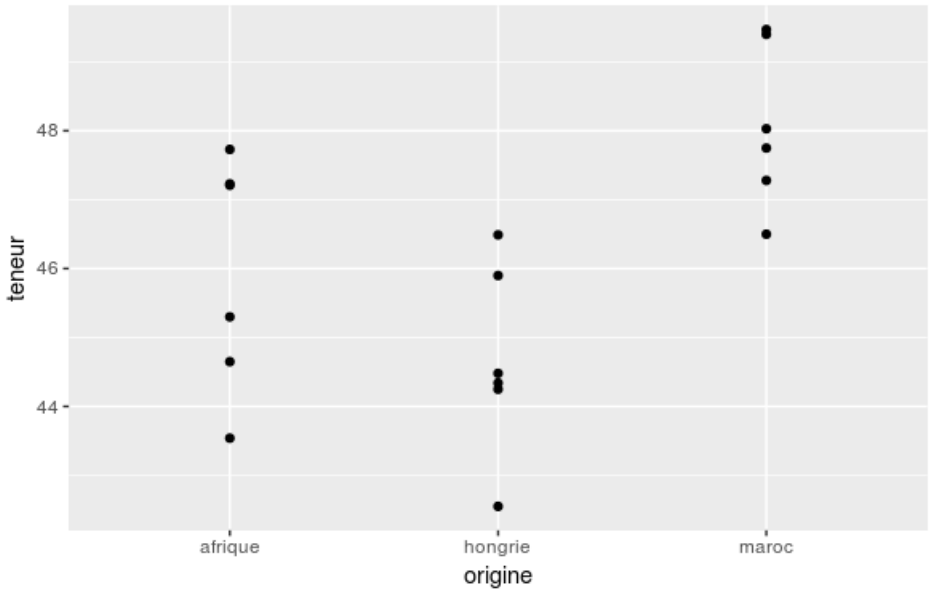

Figure 1: Teneur en acide gras des tournesols en fonction de leur origine

Visuellement, on voit une différence de la teneur en acide gras des tournesols selon leur origine : les tournesols hongrois paraissent moins gras que les africains, qui paraissent eux même moins gras que les marocains. On effectue une régression linéaire de la teneur en acide gras en fonction de l'origine :

```
lm2=lm(teneur~origine)
summary(lm2)
plot(lm2)
```

On obtient les résultats suivants :

```
Call:
lm(formula = teneur ~ origine)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4033 -0.7546 -0.2550  1.2817  1.8217

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     45.9433     0.5857  78.437   <2e-16 ***
originehongrie  -1.2750     0.8284  -1.539   0.1446
originemaroc     2.1283     0.8284   2.569   0.0214 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.435 on 15 degrees of freedom
Multiple R-squared:  0.5347,Adjusted R-squared:  0.4726
F-statistic: 8.617 on 2 and 15 DF,  p-value: 0.003224
```

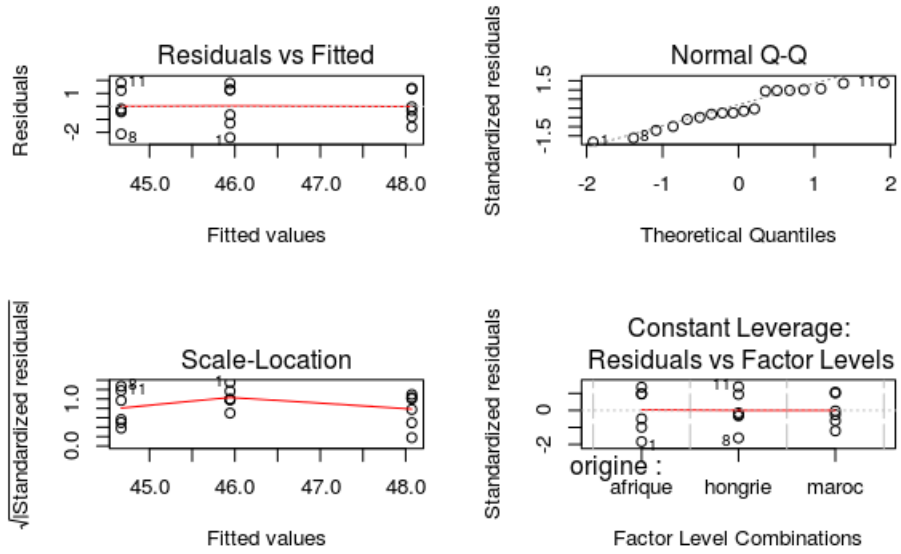

Figure 2: Graphiques du modèle linéaire de la teneur en acide gras en fonction de l'origine

Le modèle prend pour origine par défaut l'Afrique du sud et compare les autres origines à celle-ci, sans prendre en compte le testeur utilisé. On se rend compte que le Maroc est significativement différent de l'Afrique du Sud, contrairement à la Hongrie. Les graphiques du modèles sont convenables, les résidus semblent suivre une loi normale et les mesures semblent indépendantes. Cependant, le modèle ne prend pas en compte l'effet éventuel du testeur, qui doit être pris en compte.

2.2 Influence du testeur

On représente à présent les données à l'aide d'un graphique montrant l'influence du testeur utilisé pour la mesure de la teneur en acide gras, indépendamment de l'origine des tournesols :

```
ggplot(data,aes(x=testeur, y=teneur))+geom_point()
```

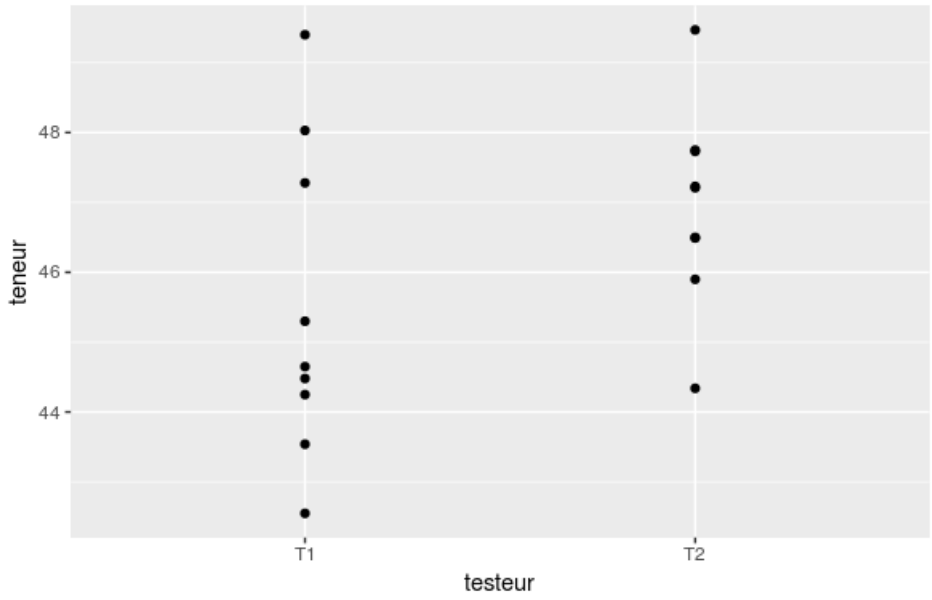

Figure 3: Teneur en acide gras des tournesols en fonction du testeur utilisé

Visuellement, on ne constate pas de différence flagrante entre les deux testeurs, origines confondues. On effectue une régression linéaire de la teneur en acide gras en fonction du testeur :

```
lm1=lm(teneur~testeur)
summary(lm1)
plot(lm1)
```

On obtient les résultats suivants :

```
Call:
lm(formula = teneur ~ testeur)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9478 -1.0478 -0.3278  0.7872  3.9022

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.4978     0.6278  72.468   <2e-16 ***
testeurT2     1.4600     0.8879   1.644     0.12
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.884 on 16 degrees of freedom
Multiple R-squared:  0.1446,Adjusted R-squared:  0.0911
F-statistic: 2.704 on 1 and 16 DF,  p-value: 0.1196
```

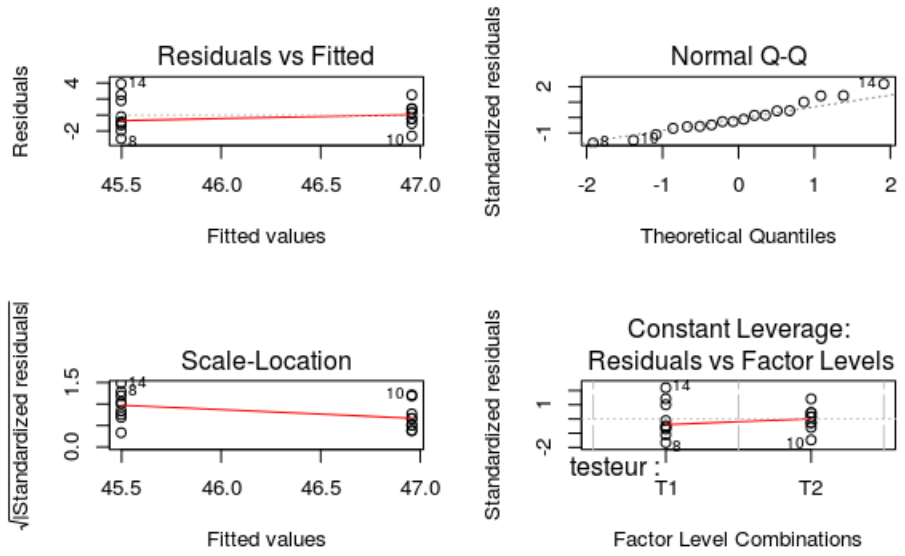

Figure 4: Graphiques du modèle linéaire de la teneur en acide gras en fonction du testeur utilisé

Le modèle prend pour testeur par défaut T1 et compare T2 avec T1, sans prendre en compte l'origine des tournesols. Le summary montre qu'il n'y a a priori pas d'influence du testeur sur la teneur. Les graphiques du modèles sont convenables, les résidus semblent suivre une loi normale et les mesures semblent indépendantes. Cependant, le modèle ne prend cette fois-ci pas en compte l'effet de l'origine. Il faut prendre les deux en compte en même temps.

2.3 Influence simultanée du testeur et de l'origine

On représente à présent les données à l'aide d'un interaction plot réalisé manuellement avec ggplot, qui va montrer l'influence des deux facteurs simultanément. On regroupe les 3 points de chaque combinaison des deux paramètres par moyenne avec tapply et on créé ainsi un nouveau jeu de données que l'on représente graphiquement en utilisant le testeur comme abcisse et l'origine en couleur.

```
newdata=as.data.frame.table(with(data, tapply(teneur,
```

```
list(testeur, origine), mean)),
dnn=c("newtesteur",'neworigine','newteneur'))
newteneur=newdata$Freq
newtesteur=newdata$Var1
neworigine=newdata$Var2
ggplot(newdata,aes(x=newtesteur,y=newteneur,group=neworigine,
col=neworigine))+geom_line()+ylab("Teneur moyenne")
+xlab("Testeur")+labs(colour = "Origine")
```

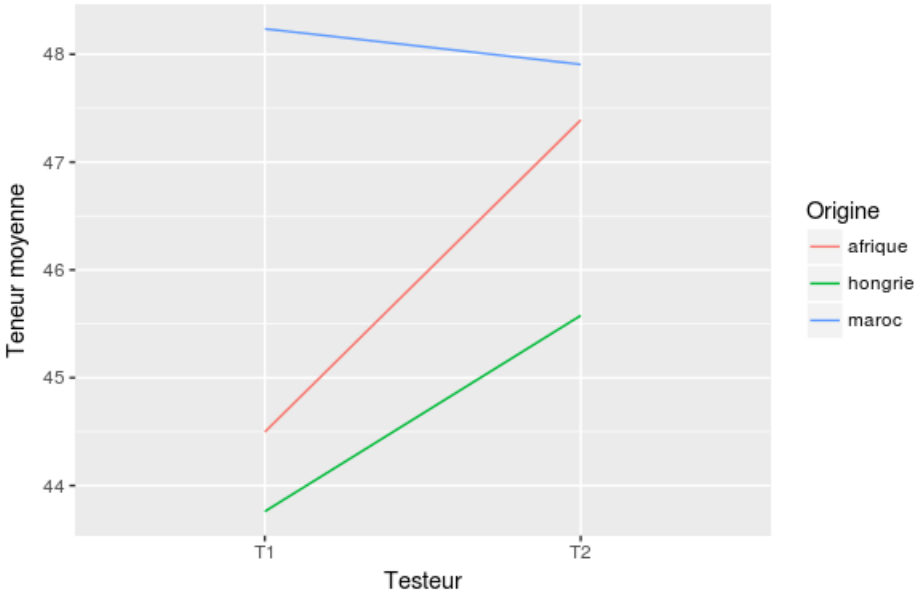

Figure 5: Teneur moyenne en acide gras des tournesols en fonction de leur origine et du testeur utilisé

On peut voir sur ce graphique des différences évidentes à la fois entre les pays, entre les testeurs contrairement à ce que stipulait la précédente analyse, et on constate également que l'effet du testeur est différent selon l'origine, ce qui peut expliquer le fait que l'effet du testeur ne soit pas significatif, et nous pousse à créer un modèle linéaire avec intéraction entre les deux facteurs :

```
lm3=lm(teneur~testeur*origine)
summary(lm3)
plot(lm3)
```

On obtient les résultats suivants :

```
Call:
```

```
lm(formula = teneur ~ testeur * origine)

Residuals:
     Min       1Q   Median       3Q      Max
-1.40667 -0.76917 -0.00167  0.66250  1.56333

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)               44.4967     0.6057  73.465  < 2e-16 ***
testeurT2                  2.8933     0.8566   3.378 0.005490 **
originehongrie            -0.7367     0.8566  -0.860 0.406633
originemaroc               3.7400     0.8566   4.366 0.000918 ***
testeurT2:originehongrie  -1.0767     1.2114  -0.889 0.391581
testeurT2:originemaroc    -3.2233     1.2114  -2.661 0.020757 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.049 on 12 degrees of freedom
Multiple R-squared:  0.801,Adjusted R-squared:  0.718
F-statistic: 9.658 on 5 and 12 DF,  p-value: 0.0006888
```

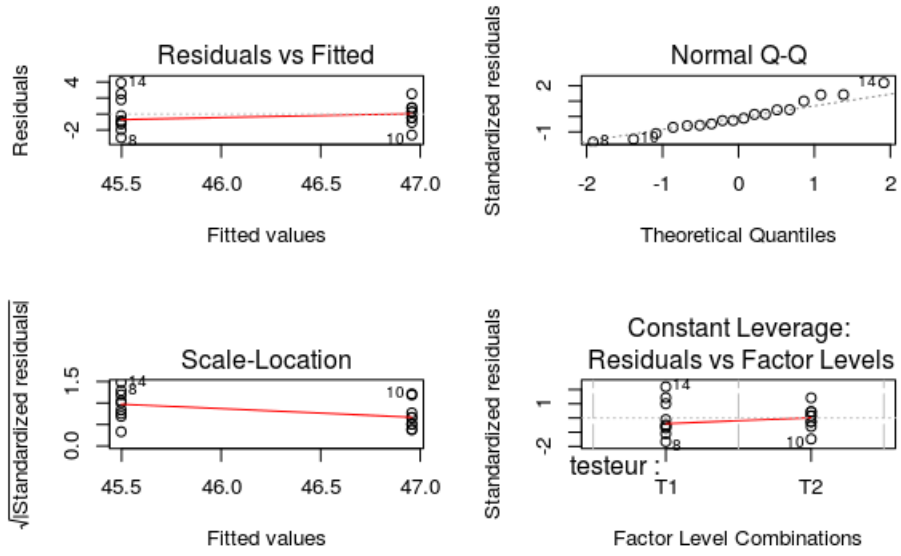

Figure 6: Graphiques du modèle linéaire de la teneur en acide gras en fonction de l'origine et du testeur utilisé

Le modèle prend pour testeur par défaut T1 et pour origine par défaut l'Afrique du sud, et compare toutes les combinaisons possibles de paramètres avec celle ci. Le summary montre qu'il y a une influence du testeur en Afrique du Sud, qu'il y a une différence entre Afrique du Sud et Maroc pour le testeur T1 mais pas entre Afrique du Sud et Hongrie, tout comme pour le testeur T2. Les graphiques du modèles sont convenables, les résidus semblent suivre une loi normale et les mesures semblent indépendantes. On peut finalement garder ce modèle pour étudier l'influence de l'origine des tournesols sur leur teneur en acide gras, qui montre que quel que soit le testeur utilisé, la teneur en acide gras des tournesols marocains est plus élevée que celle des tournesols hongrois et africains, qui eux sont similaires.