Analyse de la Variance en R (6)

October 17, 2018

Greta Laage,Luc Adjengue

Contents

1 Introduction

L'analyse de variance étudie si les valeurs d'une variable numérique sont différentes en fonction de son groupe d'appartenance. On utilise la fonction **aov()** pour faire une analyse de variance.

Il faut dans un premier temps charger les données. Celles-ci doivent être sous la forme d'un tableau dont une des colonnes contient la variable expliquée et les autres colonnes contiennent les facteurs.

On veut analyser l'influence de 3 types de régime et de 4 types d'exercices physiques ainsi que leur interaction sur la perte de poids.

```
In [4]: # Chargement des données
        d <- read.csv("6_data.csv", header = TRUE, sep = ";",dec = ",")
```

```
In [9]: # affichier les premières lignes du tableau de données
        head(d)
```

Regime	Exercice_physique	Perte_poids
R1	EX1	7
R1	EX1	12
R1	EX1	5
R1	EX1	8
R1	EX1	12
R1	EX1	6

2 Un seul facteur, premier exemple

On étudie l'influence du régime sur la perte de poids.

2.1 Définir et ajuster le modèle

```
In [10]: # Modèle d'analyse de la variance
         fit = aov(Perte_poids ~ Regime,data = d)
```

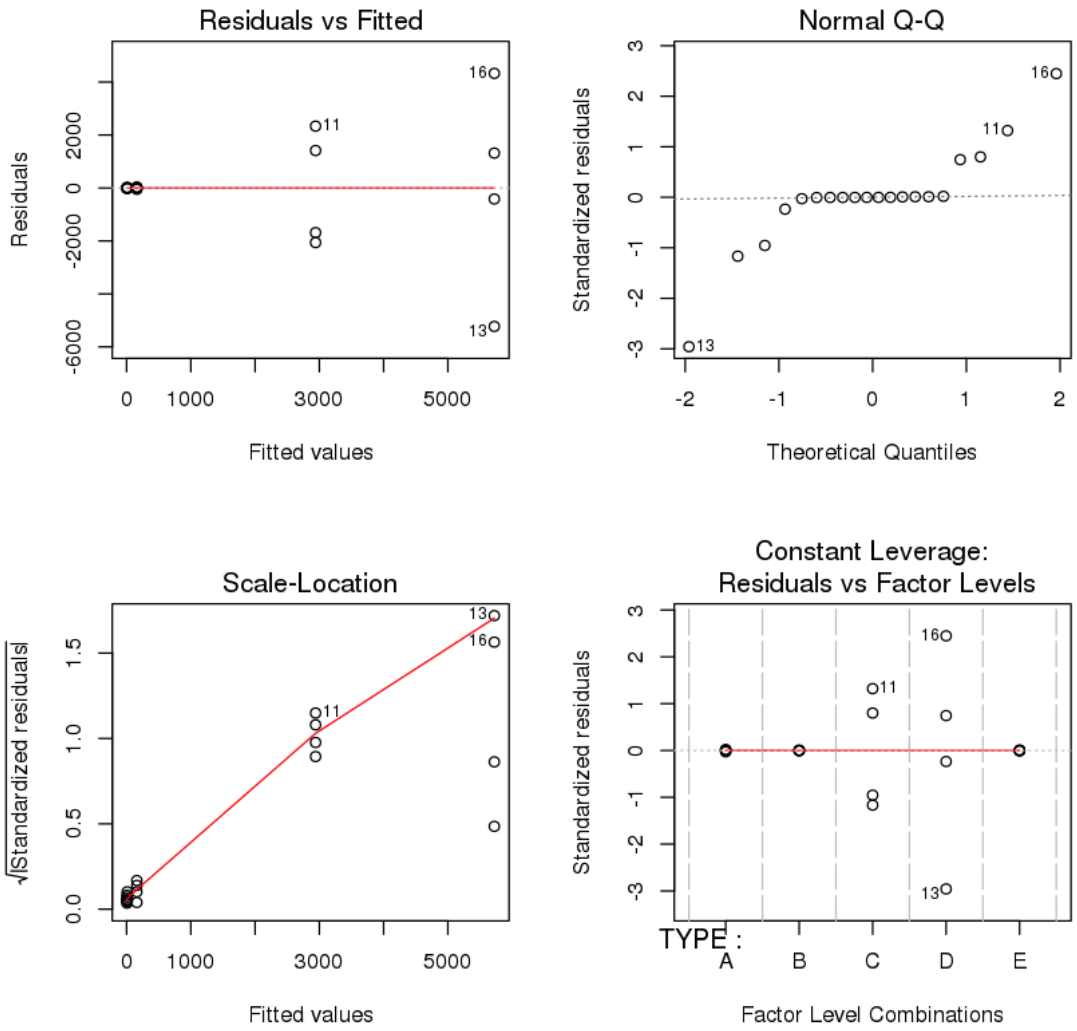
```
In [11]: # Résultats du modèle
         summary(fit)
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
Regime        2   61.7   30.88   1.446  0.243
Residuals    69 1473.8   21.36
```

La valeur-p n'est pas inférieure à 5% donc on peut conclure qu'au seuil de 5%, il n'y pas de différence significative entre les traitements.

2.2 Diagramme des résidus

```
In [4]: par(mfrow=c(2,2))
        plot(fit)
```



2.3 Méthode de Tukey pour déterminer les paires de moyennes significativement différentes

On utilise la fonction **TukeyHSD()** qui prend en paramètre le modèle ANOVA ajusté pour effectuer les comparaisons des paires de moyennes.

```
In [7]: TukeyHSD(fit)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = DUREE ~ TYPE, data = d)
```

```
$TYPE
        diff        lwr       upr     p adj
B-A  -153.50  -4610.737  4303.737 0.9999674
C-A  2782.00  -1675.237  7239.237 0.3454736
D-A  5563.25   1106.013 10020.487 0.0115524
E-A  -149.00  -4606.237  4308.237 0.9999710
C-B  2935.50  -1521.737  7392.737 0.2974817
D-B  5716.75   1259.513 10173.987 0.0093981
E-B     4.50  -4452.737  4461.737 1.0000000
D-C  2781.25  -1675.987  7238.487 0.3457196
E-C -2931.00  -7388.237  1526.237 0.2988208
E-D -5712.25 -10169.487 -1255.013 0.0094552
```

La fonction **TukeyHSD()** ci-dessus renvoie les valeurs suivantes:

- **diff** : différence entre les moyennes de deux groupes
- **lwr,upr** : les deux valeurs extrêmes de l'intervalle de confiance à 95% (par défaut)
- **p adj** : la valeur p après ajustement pour les comparaisons multiples

Observations: Avec les valeurs données ci-dessus, on peut dire que seules les paires D,A; D,B et E,D sont significativement différentes car leur valeur-p est petite.

3 Un facteur, deuxième exemple

On étudie l'influence du type d'exercice physique sur la perte de poids.

```
In [12]: fit2 = aov(Perte_poids ~ Exercice_physique, data = d )

In [13]: summary(fit2)

                  Df Sum Sq Mean Sq F value Pr(>F)
Exercice_physique  3  721.2  240.39   20.07  2e-09 ***
Residuals         68  814.3   11.98
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1


In [14]: TukeyHSD(fit2)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Perte_poids ~ Exercice_physique, data = d)

$Exercice_physique
```

```
             diff        lwr        upr     p adj
EX2-EX1  7.4444444   4.406397 10.4824917 0.0000001
EX3-EX1  4.1666667   1.128619  7.2047139 0.0031649
EX4-EX1 -0.1666667  -3.204714  2.8713806 0.9989136
EX3-EX2 -3.2777778  -6.315825 -0.2397306 0.0294946
EX4-EX2 -7.6111111 -10.649158 -4.5730639 0.0000000
EX4-EX3 -4.3333333  -7.371381 -1.2952861 0.0019967
```

4 Plusieurs facteurs

On peut faire l'analyse de variance avec soit la fonction **lm()**, de la même manière que pour la régression linéaire multiple puis en utilisant la fonction **anova()**.

On peut sinon utiliser à nouveau la fonction **aov()**

```
In [15]: # première méthode
         fit3 = lm(Perte_poids ~ Exercice_physique + Regime, data = d)
```

```
In [16]: anova(fit3)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Exercice_physique	3	721.1667	240.38889	21.081608	1.087026e-09
Regime	2	61.7500	30.87500	2.707674	7.410179e-02
Residuals	66	752.5833	11.40278	NA	NA

```
In [20]: # seconde méthode
         fit4 = aov(Perte_poids ~ Exercice_physique + Regime, data = d)
```

```
In [18]: anova(fit4)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Exercice_physique	3	721.1667	240.38889	21.081608	1.087026e-09
Regime	2	61.7500	30.87500	2.707674	7.410179e-02
Residuals	66	752.5833	11.40278	NA	NA