Fiche de Biostatistique

Analyses de la variance

D. Chessel & A.B. Dufour

Résumé

La fiche donne des indications sur les principes de construction des tableaux d'analyse de la variance et des illustrations sur leur usage.

Plan

1. Analyses de la variance à un bloc

Nous avons vu que le problème de la régression simple s'exprimait par l'estimation d'un modèle. $\mathbf{x} = (x_1, ..., x_n)$ est une variable explicative et $\mathbf{y} = (y_1, ..., y_n)$ est une variable à expliquer. Si on suppose simplement que $y_i = ax_i + b + \varepsilon_i$, l'estimation des paramètres par la méthode des moindres carrés s'écrit :

Trouver $\hat{a}$ et $\hat{b}$ qui minimisent $E(a,b) = \sum_{i=1}^{n} p_i (y_i - ax_i - b)^2 = \|\mathbf{y} - a\mathbf{x} - b\mathbf{1}_n\|_{\mathbf{D}}^2$

La solution est donnée par la projection au sens de **D** de **y** sur le $sev(\mathbf{x}, \mathbf{1}_n)$

$$\|\mathbf{y} - m(\mathbf{y})\mathbf{1}_n\|_{\mathbf{D}}^2 = \|\hat{\mathbf{y}} - m(\mathbf{y})\mathbf{1}_n\|_{\mathbf{D}}^2 + E(a,b) \leftrightarrow Totale = Expliquée + Erreur$$

Dans le cas d'une pondération uniforme, si on suppose que y_i est la réalisation d'une variable aléatoire Y_i qui suit une loi normale de moyenne $ax_i + b$ et de variance σ^2, l'estimation de *a* et *b* par le maximum donne le même résultat $\hat{a}$ et $\hat{b}$. De plus la décomposition de la projection s'accompagne des éléments constituants le tableau d'analyse de la variance :

$$\underset{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{\sigma^2} \to \chi^2_{n-1}}{\sum_{i=1}^{n}(y_i - m(\mathbf{y}))^2} = \underset{\frac{\sum_{i=1}^{n}(Ax_i + B - \bar{Y})^2}{\sigma^2} \to \chi^2_1}{\sum_{i=1}^{n}(\hat{y}_i - m(\mathbf{y}))^2} + \underset{\frac{\sum_{i=1}^{n}(Y_i - Ax_i - B)^2}{\sigma^2} \to \chi^2_{n-2}}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Nous avons détaillé les justifications pour une régression simple mais l'ensemble des résultats reste vrai pour un sous-espace de dimension quelconque.

1.1. ANOVA à un facteur

Quinze veaux [1] ont été répartis au hasard en trois lots, alimentés chacun d'une façon différente. Les gains de poids observés au cours d'une même période et exprimés en kg étant les suivants, peut-on admettre qu'il n'y a pas de relation entre l'alimentation et la croissance des veaux ?

Alimentation		
1	2	3
37,7	45,2	48,3
44,6	54,2	44,1
42,1	38,1	56,9
45,1	48,3	42,2
43,2	55,1	54

La variable observée est un vecteur à 18 composantes :

[1] Dagnelie, P. (1981) Théorie et méthodes statistiques. Exercices. Les Presses Agronomiques de Gembloux, Gembloux, 186 p.

```
> gain
 [1] 37.7 44.6 42.1 45.1 43.2 45.2 54.2 38.1 48.3 55.1 48.3 44.1 56.9 42.2
54.0
```

L'espace de projection est défini par les trois indicatrices :

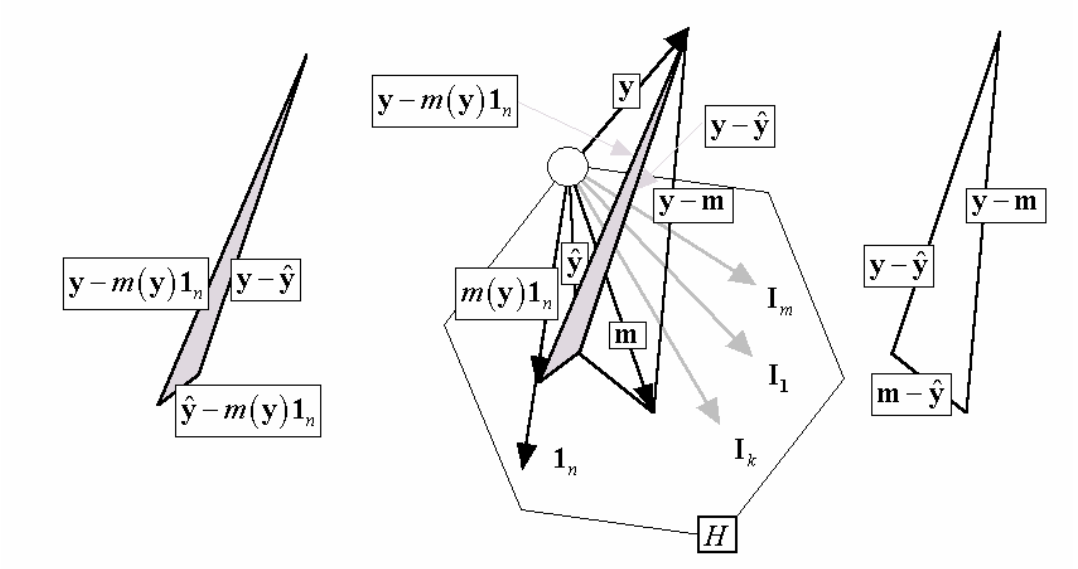
```
> ali
 [1] t1 t1 t1 t1 t1 t2 t2 t2 t2 t2 t3 t3 t3 t3 t3
Levels:  t1 t2 t3

> as.numeric(ali=="t1")
 [1] 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0
> as.numeric(ali=="t2")
 [1] 0 0 0 0 0 1 1 1 1 1 0 0 0 0 0
> as.numeric(ali=="t3")
 [1] 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1
```

Projeter sur un paquet d'indicatrices de classe, c'est simplement calculer les moyennes par classe (fiche Géométrie de l'espace des variables, §3).

```
> tapply(gain,ali,mean)
   t1    t2    t3
42.54 48.18 49.10
```

La variable estimée est obtenue en remplaçant la valeur observée par la moyenne des valeurs de la classe dans laquelle elle se trouve :

```
> predict(lm(gain~ali))
    1     2     3     4     5     6     7     8     9    10    11    12    13
42.54 42.54 42.54 42.54 42.54 48.18 48.18 48.18 48.18 48.18 49.10 49.10 49.10
   14    15
49.10 49.10
```



$\mathbb{R}^n$ se décompose en deux sous-espaces, respectivement H engendré par les indicatrices des classes et son complémentaire orthogonal $H^\perp$. H contient les combinaisons linéaires des indicatrices des classes donc **les variables constantes par classes**. $H^\perp$ contient les variables orthogonales à toutes les indicatrices donc **les variables centrées par classes**.

```
> gain.pred <- predict(lm(gain~ali))
> gain.resi <- gain-gain.pred

> gain.pred
    1     2     3     4     5     6     7     8     9    10    11    12    13
42.54 42.54 42.54 42.54 42.54 48.18 48.18 48.18 48.18 48.18 49.10 49.10 49.10
```

```
   14    15
49.10 49.10
> gain.resi
     1      2      3      4      5      6      7      8      9     10     11
 -4.84   2.06  -0.44   2.56   0.66  -2.98   6.02 -10.08   0.12   6.92  -0.80
    12     13     14     15
 -5.00   7.80  -6.90   4.90
> cor(gain.resi,gain.pred)
[1] 3.625e-16
```

Prédictions et résidus sont non corrélés.

$$\boxed{\mathbf{H_1} : Y_i \to \mathcal{N}\left(\mu_i, \sigma\right) \text{ avec } \mu_i = a_{Cl(i)}}$$

La somme des carrés des écarts au **modèle vrai** suit une loi Khi2 (définition du modèle):

$$\frac{\sum_{i=1}^{n}\left(Y_i - \mu_i\right)^2}{\sigma^2} \to \chi^2_n \text{ donc } \boxed{\frac{\sum_{i=1}^{n}\left(Y_i - a_{Cl(i)}\right)^2}{\sigma^2} \to \chi^2_n}$$

On applique le théorème de Cochran. Si l'hypothèse $\mathbf{H_1}$ est vérifiée :

$$\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{\sigma^2} \to \chi^2_{n-m} \text{ donc } \boxed{\frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}_{Cl(i)}\right)^2}{\sigma^2} \to \chi^2_{n-m}}$$

La somme des carrés des écarts au **modèle estimé** suit une loi Khi2 (espace de projection de dimension *m*).

D'autre part, le vecteur des observations et le vecteur des prédictions se projettent au même endroit sur le vecteur des constantes (théorème des trois perpendiculaires). On a toujours :

$$\sum_{i=1}^{n}\left(y_i - m(\mathbf{y})\right)^2 = \sum_{i=1}^{n}\left(y_i - m\left(\mathbf{y}/Cl(i)\right)\right)^2 + \sum_{i=1}^{n}\left(m\left(\mathbf{y}/Cl(i)\right) - m(\mathbf{y})\right)^2$$

$$\text{SCT=SCR + SCE}$$

$$n\left(\text{Variance totale}\right) = n\left(\text{Variance intra}\right) + n\left(\text{Variance inter}\right)$$

Si la variable **qualitative** n'a pas d'effet sur la variable **y** ($\mu_i = \mu$), la situation est décrite par l'hypothèse nulle :

$$\boxed{\mathbf{H}_0 : Y_i \to \mathcal{N}\left(\mu, \sigma\right)}$$

Que la vraie valeur μ_i soit μ n'enlève rien au résultat précédent :

La décomposition en parties orthogonales (théorème de Cochran) implique donc que :

$$\boxed{\frac{\sum_{i=1}^{n}\left(\overline{Y}_{Cl(i)} - \overline{Y}\right)^2}{\sigma^2} \to \chi^2_{m-1}}$$

et que les deux variables sont indépendantes.

Il reste alors à utiliser :

$$\frac{\frac{\sum_{i=1}^{n}\left(\overline{Y}_{Cl(i)}-\overline{Y}\right)^2}{(m-1)}}{\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}_{Cl(i)}\right)^2}{(n-m)}} \rightarrow F_{m-1,n-m}$$

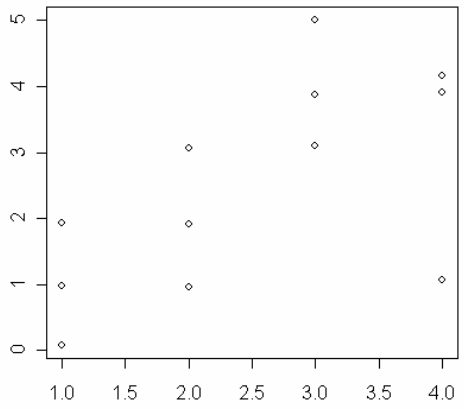
Toute cette information s'exprime dans un **tableau d'analyse de la variance** :

```
> anova(lm(gain~ali))
Analysis of Variance Table

Response: gain
          Df Sum Sq Mean Sq F value Pr(>F)
ali        2    126      63    1.95   0.18
Residuals 12    388      32
```

	DDL	SC	CM	F	Proba
Facteur	m - 1	$\sum_{i=1}^{n}\left(\overline{Y}_{Cl(i)}-\overline{Y}\right)^2 \left(\rightarrow \sigma^2\chi^2_{m-1}\right)$	$CME=\frac{\sum_{i=1}^{n}\left(\overline{Y}_{Cl(i)}-\overline{Y}\right)^2}{(m-1)}$	$\frac{CME}{CMR}\left(\rightarrow F_{m-1,n-m}\right)$	$\mathrm{P}\left(F>F_{obs}\right)$
Résiduelle	$n-m$	$\sum_{i=1}^{n}\left(Y_i-\overline{Y}_{Cl(i)}\right)^2 \left(\rightarrow \sigma^2\chi^2_{n-m}\right)$	$CMR=\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}_{Cl(i)}\right)^2}{(n-m)}$		
Total	$n-1$	$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2 \left(\rightarrow \sigma^2\chi^2_{n-1}\right)$	$CMT=\frac{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}{(n-1)}$		

On peut faire les calculs manuellement sans difficulté, mais c'est maintenant un aspect très secondaire. Exemple :

```
> obs <- c(0,1,2,3,1,2,5,4,3,4,4,1)
> fac <- as.factor(rep(c("1","2","3","4"),rep(3,4)))
> obs
 [1] 0 1 2 3 1 2 5 4 3 4 4 1
> fac
 [1] 1 1 1 2 2 2 3 3 3 4 4 4
Levels:  1 2 3 4
plot(as.numeric(fac),jitter(obs))
```



$$\mathbf{y} = \begin{bmatrix} 0 & 3 & 5 & 4 \\ 1 & 1 & 4 & 4 \\ 2 & 2 & 3 & 1 \end{bmatrix} \quad \hat{\mathbf{y}} = \begin{bmatrix} 1 & 2 & 4 & 3 \\ 1 & 2 & 4 & 3 \\ 1 & 2 & 4 & 3 \end{bmatrix} \quad \mathbf{y} - \hat{\mathbf{y}} = \begin{bmatrix} -1 & 1 & 1 & 1 \\ 0 & -1 & 0 & 1 \\ 1 & 0 & -1 & -2 \end{bmatrix}$$

n = 12, m = 4, SCR=12, SCT=27, SCE=15

```
> anova(lm(obs~fac))
Analysis of Variance Table

Response: obs
          Df Sum Sq Mean Sq F value Pr(>F)
fac        3   15.0     5.0    3.33  0.077 .
Residuals  8   12.0     1.5
```

1.2. Test de linéarité

La situation précédente s'étend de manière très importante au cas où l'hypothèse nulle n'est pas simple.

Supposons que nous ayons deux hypothèses :

$$\boxed{\mathbf{H}_0 : Y_i \to \mathcal{N}(\mu_i, \sigma)} \text{ et } \boxed{\mathbf{H}_1 : Y_i \to \mathcal{N}(\mu_i', \sigma)}$$

Supposons que les deux modèles sont linéaires, c'est-à-dire que l'estimation se fait par projection sur un sous-espace de $\mathbb{R}^n$. Nous connaissons déjà les cas :

$\mu_i = \mu$	Projection sur le vecteur $\mathbf{1}_n$
$\mu_i = a_{Cl(i)}$	Projection sur le sous-espace des indicatrices
$\mu_i = ax_i$	Projection sur le vecteur $\mathbf{x}$
$\mu_i = ax_i + b$	Projection sur le sous-espace engendré par $\mathbf{x}$ et $\mathbf{1}_n$
$\mu_i = a_1 x_i^1 + a_2 x_i^2 + ... + a_p x_i^p + b$	Projection sur le sous-espace $sev(\mathbf{x}_1, ..., \mathbf{x}_p, \mathbf{1}_n)$

Notons H_0 le sous-espace défini par l'hypothèse $\mathbf{H}_0$ et H_1 le sous-espace défini par l'hypothèse $\mathbf{H}_1$.

$$\frac{Y_i - \mu_i}{\sigma} \to N_{0,1} \Rightarrow \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i}{\sigma} \right)^2 \to \chi_n^2 \Rightarrow \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i}{\sigma} \right)^2 \to \chi_{n-\dim(\mathrm{H}_0)}^2 \text{ [1] (Cochran)}$$

$$\frac{Y_i - \mu_i'}{\sigma} \to N_{0,1} \Rightarrow \sum_{i=1}^{n} \left(\frac{Y_i - \mu_i'}{\sigma} \right)^2 \to \chi_n^2 \Rightarrow \sum_{i=1}^{n} \left(\frac{Y_i - \hat{Y}_i'}{\sigma} \right)^2 \to \chi_{n-\dim(\mathrm{H}_1)}^2 \text{ [2] (Cochran)}$$

Quand peut-on comparer les deux modèles, c'est-à-dire tester une hypothèse contre l'autre ? Ceci est possible quand les hypothèses sont **emboîtées** ce qui signifie qu'un des deux espaces est contenu dans l'autre. Supposons en effet que $\mathrm{H}_0 \subseteq \mathrm{H}_1$. Le projeté sur H_0 du projeté sur H_1 d'un vecteur est le projeté sur H_0 de ce vecteur. On a :

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i' \right)^2 + \sum_{i=1}^{n} \left(\hat{Y}_i - \hat{Y}_i' \right)^2$$

Sous l'hypothèse $\mathbf{H}_0$, on a $\sum_{i=1}^{n}\left(\frac{Y_i-\hat{Y}_i}{\sigma}\right)^2 \rightarrow \chi^2_{n-\dim(\mathrm{H}_0)}$ directement. Mais on a aussi :

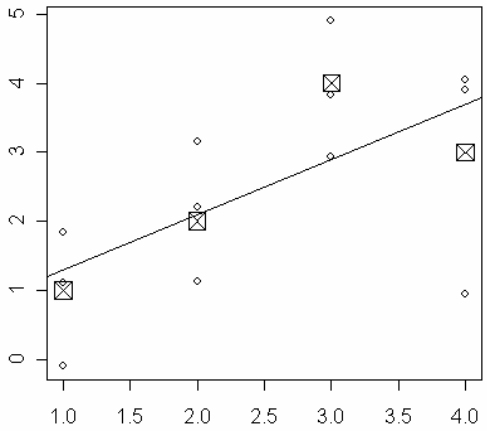
$$\sum_{i=1}^{n}\left(\frac{Y_i-\hat{Y}'_i}{\sigma}\right)^2 \rightarrow \chi^2_{n-\dim(\mathrm{H}_1)} \quad \sum_{i=1}^{n}\left(\frac{\hat{Y}_i-\hat{Y}'_i}{\sigma}\right)^2 \rightarrow \chi^2_{\dim(\mathrm{H}_1)-\dim(\mathrm{H}_0)}$$

et l'indépendance des deux par le théorème de Cochran. Donc le tableau d'analyse de variance :

	DDL	SC	CM	F	Proba
Entre modèles	$ddl1 = \dim(\mathrm{H}_1)-\dim(\mathrm{H}_0)$	$\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{Y}'_i\right)^2 \left(\rightarrow \sigma^2\chi^2_{ddl1}\right)$	$CME=\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\hat{Y}'_i\right)^2}{ddl1}$	$\frac{CME}{CMR}\left(\rightarrow F_{ddl1,ddl2}\right)$	$\mathrm{P}(F>F_{obs})$
Résiduelle modèle 1	$ddl2 = n-\dim(\mathrm{H}_1)$	$\sum_{i=1}^{n}\left(Y_i-\hat{Y}'_i\right)^2 \left(\rightarrow \sigma^2\chi^2_{ddl2}\right)$	$CMR=\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}'_i\right)^2}{ddl2}$		
Résiduelle modèle 0	$ddl3 = n-\dim(\mathrm{H}_0)$	$\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2 \left(\rightarrow \sigma^2\chi^2_{ddl3}\right)$	$CMT=\frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}{ddl3}$		

Exemple.

```
> obs
 [1] 0 1 2 3 1 2 5 4 3 4 4 1
> fac
 [1] 1 1 1 2 2 2 3 3 3 4 4 4
Levels:  1 2 3 4
> lm1 <- lm(obs~fac)
> anova(lm1)
Analysis of Variance Table

Response: obs
          Df Sum Sq Mean Sq F value Pr(>F)
fac        3   15.0     5.0    3.33  0.077 .
Residuals  8   12.0     1.5
---
> points(as.numeric(fac),predict(lm1),pch=7,cex=2)
> abline(lm(obs~as.numeric(fac)))
```



Cercles : les observations. Croix : le modèle de l'analyse de variance. Droite : le modèle de la régression simple.

```
> xfac <- as.numeric(fac)
> obs
 [1] 0 1 2 3 1 2 5 4 3 4 4 1
```

```
> xfac
 [1] 1 1 1 2 2 2 3 3 3 4 4 4

> lm0 <- lm(obs~xfac)

> anova(lm0)
Analysis of Variance Table

Response: obs
          Df Sum Sq Mean Sq F value Pr(>F)
xfac       1   9.60    9.60    5.52  0.041 *
Residuals 10  17.40    1.74
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1

> coefficients(lm0)
(Intercept)        xfac
        0.5         0.8
> abline(lm0)
```

On a le droit de comparer les deux modèles. La variable quantitative **x** est constante par classe. C'est une combinaison d'indicatrices. Le vecteur des constantes est constant par classe (il est constant tout court !) . C'est une combinaison d'indicatrices. Les modèles linéaires du type a**x**+b sont constants par classe et **l'espace de projection de la régression simple est dans l'espace de projection de l'analyse de variance**. On peut donc tester l'hypothèse de linéarité de la relation.

```
> anova(lm0,lm1)
Analysis of Variance Table

Model 1: obs ~ xfac
Model 2: obs ~ fac
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     10       17.4
2      8       12.0  2    5.4     1.8   0.23
```

Le logiciel utilise la même décomposition mais pas les mêmes côtés du triangle. Le résultat est équivalent comme en témoigne l'usage du modèle constant :

```
> lmbase <- lm(obs~1)

> anova(lmbase,lm0)
Analysis of Variance Table

Model 1: obs ~ 1
Model 2: obs ~ xfac
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     11       27.0
2     10       17.4  1    9.6    5.52  0.041 *
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1

> anova(lmbase,lm1)
Analysis of Variance Table

Model 1: obs ~ 1
Model 2: obs ~ fac
  Res.Df Res.Sum Sq Df Sum Sq F value Pr(>F)
1     11         27
2      8         12  3     15    3.33  0.077 .
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1
```

Comparer avec les analyses précédentes.

On retiendra l'architecture fondamentale d'un tableau d'analyse de la variance :

Source de variation		DDL	SC	CM	F	Proba
Facteur	$\hat{Y}-\overline{Y}$	$m-1$	$\left\|\hat{Y}-\overline{Y}\right\|^2$	$CME=\frac{\left\|\hat{Y}-\overline{Y}\right\|^2}{(m-1)}$	$\frac{CME}{CMR}$	$P(F>F_{obs})$
Résiduelle	$Y-\hat{Y}$	$n-m$	$\left\|Y-\hat{Y}\right\|^2$	$CMR=\frac{\left\|Y-\hat{Y}\right\|^2}{(n-m)}$		
Totale	$Y-\overline{Y}$	$n-1$	$\left\|Y-\overline{Y}\right\|^2$	$CMT=\frac{\left\|Y-\overline{Y}\right\|^2}{(n-1)}$		

Il résume ce qui a été dit sur la régression simple et l'analyse de variance à un facteur. L'essentiel est dans la colonne 2 et le reste s'en déduit en pensant l'estimation comme une projection et le degré de liberté comme une dimension d'un sous-espace.

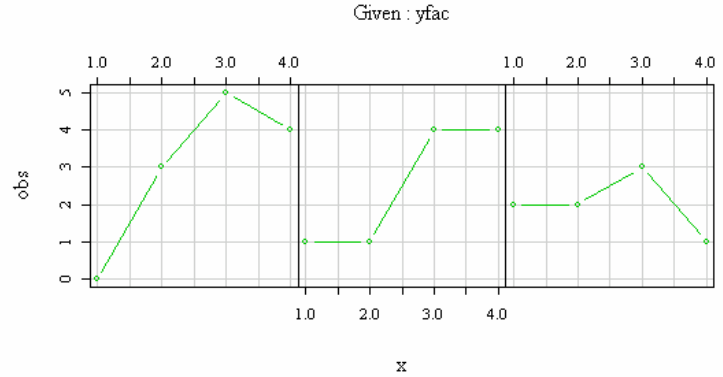
2. Deux facteurs

2.1. Plans complets sans répétitions

Reprenons le même exemple numérique en supposant qu'il s'agit de la mesure de l'abondance d'une espèce dans 3 stations (lignes) à quatre dates d'échantillonnage (colonnes).

$$\mathbf{y}=\begin{bmatrix}0 & 3 & 5 & 4\\ 1 & 1 & 4 & 4\\ 2 & 2 & 3 & 1\end{bmatrix}$$

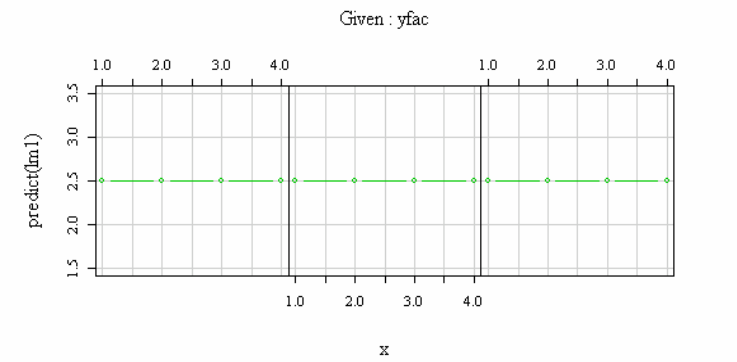
```
> obs
 [1] 0 1 2 3 1 2 5 4 3 4 4 1
> xfac
 [1] 1 1 1 2 2 2 3 3 3 4 4 4
Levels:  1 2 3 4
> yfac
 [1] 1 2 3 1 2 3 1 2 3 1 2 3
Levels:  1 2 3
```




```
> coplot(obs~x|yfac,show=F,col=3,row=1,type="b")
```

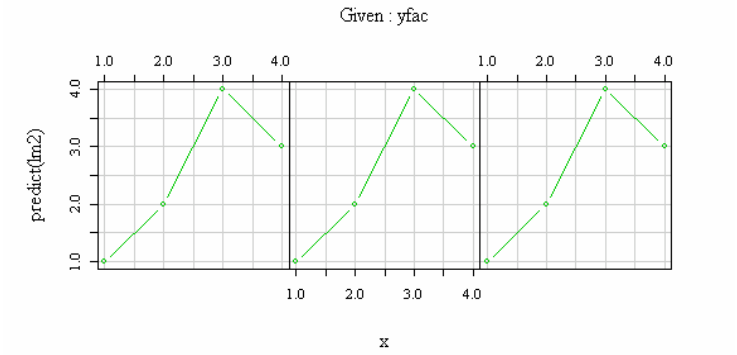
Les données peuvent être écrites soit comme une matrice, soit comme un vecteur. Le modèle le plus simple est : $\hat{y}_{ij}=b \Leftrightarrow \hat{\mathbf{y}}=b\mathbf{1}_n$ [1]

```
> lm1 <- lm(obs~1)
> lm1

Call:
lm(formula = obs ~ 1)

Coefficients:
(Intercept)
        2.5

> mean(obs)
[1] 2.5
```



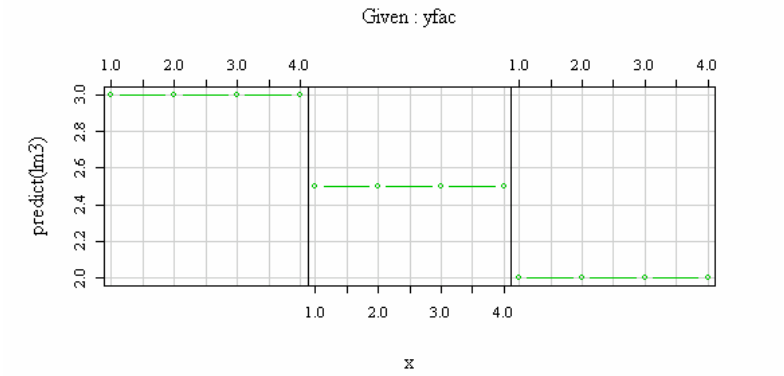

```
> coplot(predict(lm1)~x|yfac,show=F,col=3,row=1,type="b")
```

Le modèle « effet date » s'écrit $\hat{y}_{ij} = c_j \Leftrightarrow \hat{\mathbf{y}} = c_1\mathbf{K}_1 + ... + c_k\mathbf{K}_k$




```
> lm2 <- lm(obs~xfac)
> coplot(predict(lm2)~x|yfac,show=F,col=3,row=1,type="b")
> tapply(obs,xfac,mean)
1 2 3 4
1 2 4 3
```

Le modèle « effet station » s'écrit $\hat{y}_{ij} = d_i \Leftrightarrow \hat{\mathbf{y}} = d_1\mathbf{L}_1 + ... + d_l\mathbf{L}_l$

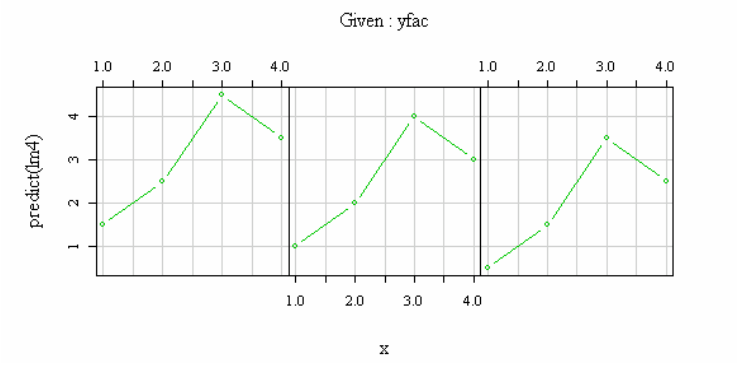



```
> lm3 <- lm(obs~yfac)
> tapply(obs,yfac,mean)
  1   2   3
3.0 2.5 2.0
> coplot(predict(lm3)~x|yfac,show=F,col=3,row=1,type="b")
```

Le modèle « double effet » s'écrit $\hat{y}_{ij} = d_i + c_j \Leftrightarrow \hat{\mathbf{y}} = d_1\mathbf{L}_1 + ... + d_l\mathbf{L}_l + c_1\mathbf{K}_1 + ... + c_k\mathbf{K}_k$.
En fait, ce modèle est mal défini puisqu'il existe une infinité de solutions identiques :

$$\hat{y}_{ij} = d_i + c_j = \left(d_i + z\right) + \left(c_j - z\right)$$

Mais la solution est unique :




```
> lm4 <- lm(obs~yfac+xfac)
> coplot(predict(lm4)~x|yfac,show=F,col=3,row=1,type="b")
```

$$\mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \\ 1 \\ 2 \\ 5 \\ 4 \\ 3 \\ 4 \\ 4 \\ 1 \end{bmatrix} \left[\mathbf{K}_1, \mathbf{K}_2, \mathbf{K}_3, \mathbf{K}_4\right] = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \left[\mathbf{L}_1, \mathbf{L}_2, \mathbf{L}_3\right] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Le sous-espace de projection contient 7 indicatrices de classe (4 pour les colonnes et 3 pour les lignes). Ces 7 vecteurs ne sont pas indépendants et donc ne forment pas une base (la somme des 4 égale la somme des 3). Le sous-espace de projection n'est donc pas défini par un système de vecteurs orthogonaux comme dans une analyse de variance à un facteur ou par un système de vecteurs libres comme dans une régression multiple. Un ajustement est nécessaire.

En général, pour un plan complet à *l* modalités lignes et *k* modalités colonnes, on a *l* + *k* indicatrices. Les *l* premières définissent le sous-espace L et les *k* dernières définissent le sous-espace K. La réunion des deux définit le sous-espace K+L. On veut projeter **y** sur K+L. L'opération n'est pas simple mais le résultat est très simple. Il suffit de voir que $\text{K+L=sev}\left(\mathbf{1}_n\right) \oplus \text{sev}\left(\mathbf{K}_2, ..., \mathbf{K}_k\right) \oplus \text{sev}\left(\mathbf{L}_2, ..., \mathbf{L}_l\right)$. En enlevant une indicatrice par classe, on conserve un système de vecteurs orthogonaux, qui n'est pas orthogonal à $\mathbf{1}_n$. On choisit alors de sortir $\mathbf{1}_n$ mais de centrer les indicatrices conservées. Cette opération est simplement illustrée sur l'exemple en cours :

$$[\mathbf{K}_1,\mathbf{K}_2,\mathbf{K}_3,\mathbf{K}_4,\mathbf{L}_1,\mathbf{L}_2,\mathbf{L}_3]=\begin{bmatrix}1&0&0&0&1&0&0\\1&0&0&0&0&1&0\\1&0&0&0&0&0&1\\0&1&0&0&1&0&0\\0&1&0&0&0&1&0\\0&1&0&0&0&0&1\\0&0&1&0&1&0&0\\0&0&1&0&0&1&0\\0&0&1&0&0&0&1\\0&0&0&1&1&0&0\\0&0&0&1&0&1&0\\0&0&0&1&0&0&1\end{bmatrix}=\begin{bmatrix}1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\end{bmatrix}\overset{\perp}{\oplus}\begin{bmatrix}-1/4&-1/4&-1/4\\-1/4&-1/4&-1/4\\-1/4&-1/4&-1/4\\3/4&-1/4&-1/4\\3/4&-1/4&-1/4\\3/4&-1/4&-1/4\\-1/4&3/4&-1/4\\-1/4&3/4&-1/4\\-1/4&3/4&-1/4\\-1/4&-1/4&3/4\\-1/4&-1/4&3/4\\-1/4&-1/4&3/4\end{bmatrix}\overset{\perp}{\oplus}\begin{bmatrix}-1/3&-1/3\\2/3&-1/3\\-1/3&2/3\\-1/3&-1/3\\2/3&-1/3\\-1/3&2/3\\-1/3&-1/3\\2/3&-1/3\\-1/3&2/3\\-1/3&-1/3\\2/3&-1/3\\-1/3&2/3\end{bmatrix}$$

$$[\mathbf{K}_1,\mathbf{K}_2,\mathbf{K}_3,\mathbf{K}_4,\mathbf{L}_1,\mathbf{L}_2,\mathbf{L}_3]=\mathrm{sev}(\mathbf{1}_n)\overset{\perp}{\oplus}\mathrm{K}_0\overset{\perp}{\oplus}\mathrm{L}_0$$

L'espace de projection a été décomposé en trois parties orthogonales deux à deux. La première contient les variables constantes, la seconde les variables centrées et constante par colonnes, la troisième les variables centrées et constantes par lignes. Vérifier que tout vecteur d'un sous-espace est orthogonal à tous les vecteurs d'un autre. On note $\overline{y_{i\bullet}}$ la moyenne sur la ligne *i*, $\overline{y_{\bullet j}}$ la moyenne sur la colonne *j* et $\overline{y_{\bullet\bullet}}$ la moyenne générale. La projection a alors comme composante de rang *i* :

$$\hat{y}_{ij}=\overline{y_{\bullet\bullet}}+\left(\overline{y_{i\bullet}}-\overline{y_{\bullet\bullet}}\right)+\left(\overline{y_{\bullet j}}-\overline{y_{\bullet\bullet}}\right)=\overline{y_{i\bullet}}+\overline{y_{\bullet j}}-\overline{y_{\bullet\bullet}}$$

```
> predict(lm4)
  1   2   3   4   5   6   7   8   9  10  11  12
1.5 1.0 0.5 2.5 2.0 1.5 4.5 4.0 3.5 3.5 3.0 2.5
> predict(lm2)+predict(lm3)-predict(lm1)
  1   2   3   4   5   6   7   8   9  10  11  12
1.5 1.0 0.5 2.5 2.0 1.5 4.5 4.0 3.5 3.5 3.0 2.5
```

Ce résultat très simple ne tient que pour un plan complet sans répétition. Cette opération permet d'aborder la notion fondamentale de **contrastes**.

$$[\mathbf{K}_1,\mathbf{K}_2,\mathbf{K}_3,\mathbf{K}_4,\mathbf{L}_1,\mathbf{L}_2,\mathbf{L}_3]=\begin{bmatrix}1&0&0&0&1&0&0\\1&0&0&0&0&1&0\\1&0&0&0&0&0&1\\0&1&0&0&1&0&0\\0&1&0&0&0&1&0\\0&1&0&0&0&0&1\\0&0&1&0&1&0&0\\0&0&1&0&0&1&0\\0&0&1&0&0&0&1\\0&0&0&1&1&0&0\\0&0&0&1&0&1&0\\0&0&0&1&0&0&1\end{bmatrix}=\begin{bmatrix}1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\end{bmatrix}\oplus\begin{bmatrix}0&0&0\\0&0&0\\0&0&0\\1&0&0\\1&0&0\\1&0&0\\0&1&0\\0&1&0\\0&1&0\\0&0&1\\0&0&1\\0&0&1\end{bmatrix}\oplus\begin{bmatrix}0&0\\1&0\\0&1\\0&0\\1&0\\0&1\\0&0\\1&0\\0&1\\0&0\\1&0\\0&1\end{bmatrix}$$

Pour faire le calcul explicite du modèle, nous avons utilisé les indicatrices centrées. Pour faire la calcul numérique, on utilise les indicatrices brutes en enlevant dans chaque famille un vecteur en trop. La famille des 6 vecteurs qui définit le sous-espace de projection et en forme une autre base définit les coefficients du modèle. L'estimation est une combinaison linéaire de ces six vecteurs. Le vecteur $\mathbf{1}_n$ est bien connu. Les autres

ont la propriété de former une base du complémentaire orthogonal de $\mathbf{1}_n$ dans le sous-espace de projection.

```
> coefficients (lm4)
(Intercept)       yfac2       yfac3       xfac2       xfac3       xfac4
        1.5        -0.5        -1.0         1.0         3.0         2.0

> contrasts(xfac)
  2 3 4
1 0 0 0
2 1 0 0
3 0 1 0
4 0 0 1

> contrasts(yfac)
  2 3
1 0 0
2 1 0
3 0 1
```

Cette écriture indique simplement que les indicatrices 2, 3 et 4 de la variable yfac sont isolées pour servir de base avec le vecteur des constantes au sous espace engendré par les indicatrices de yfac. On fait de même pour l'autre espace.

On trouve donc une base du sous-espace K+L et **alors seulement** le modèle est une combinaison linéaire unique des éléments de la base. Les coefficients de cette combinaison linéaire sont les coefficients du modèle. Ainsi à la ligne 2 et à la colonne 3, on peut calculer l'estimation soit par la méthode manuelle :

moyenne ligne 2 + moyenne colonne 3 – moyenne générale = 2.5 + 4 – 2.5 = 4

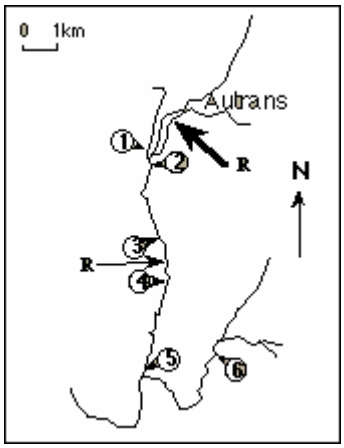
soit par la méthode des contrastes :

1(1.5) + 1(-0.5) +0(-1) + 0(1.0) + 1(3.0) +0(2.0) = 4

2.2. Plans incomplets

Quand une donnée est perdue, le calcul est beaucoup plus compliqué et ne peut se faire à la main. Le principe reste par contre le même.

tem	deb	ph	cond	oxyg	dbo5	ammo	nitra	phos	sta	sai
10	41	8.5	295	110	2.3	0.12	3.4	0.11	1	1
13	62	8.3	325	95	2.3	0.11	3	0.13	1	2
1	25	8.4	315	91	1.6	0.07	6.4	0.03	1	3
3	118	8	325	100	1.6	0.17	1.8	0.19	1	4
11	158	8.3	315	13	7.6	2.85	2.7	1.5	2	1
13	80	7.6	380	20	21	9.8	0.8	3.65	2	2
3	63	8	425	38	36	12.5	2.2	6.5	2	3
3	252	8.3	360	100	9.5	2.52	4.6	1.6	2	4
11	198	8.5	290	113	3.3	0.4	4	0.1	3	1
15	100	7.8	385	46	15	7.9	7.7	4.5	3	2
2	79	8.1	350	84	7.1	2.7	13.2	3.7	3	3
3	315	8.3	370	100	8.7	2.8	4.8	2.85	3	4
12	280	8.6	290	126	3.5	0.45	4	0.73	4	1
16	140	8	360	76	12	4.9	8.4	3.45	4	2
3	85	8.3	330	106	2	0.42	12	1.6	4	3
3	498	8.3	330	100	4.8	1.04	4.4	0.82	4	4
13	322	8.5	285	117	3.6	0.48	4.6	0.84	5	1
15	160	8.4	345	91	1.7	0.22	10	1.74	5	2
2	72	8.6	305	91	1.6	0.1	9.5	1.25	5	3
2	390	8.2	330	100	1.7	0.56	5	0.6	5	4
11	303	8.5	245	100	1.7	0.05	2.7	0.16	6	1
13	310	8.2	285	82	8.5	0.59	3.7	0.6	6	2
4	181	8.6	270	105	2.8	0.1	3.66	0.43	6	3
3	480	8.2	290	100	1.3	0.04	2.2	0.13	6	4

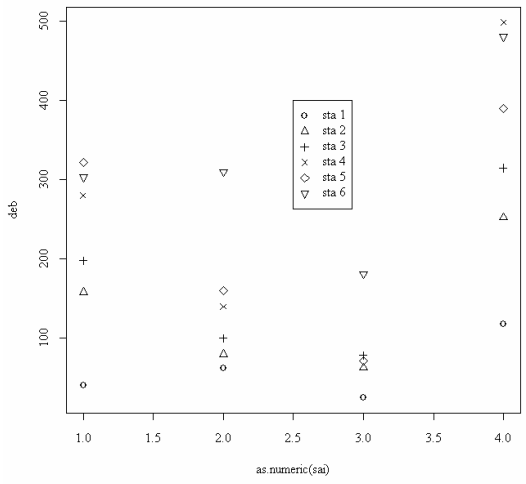


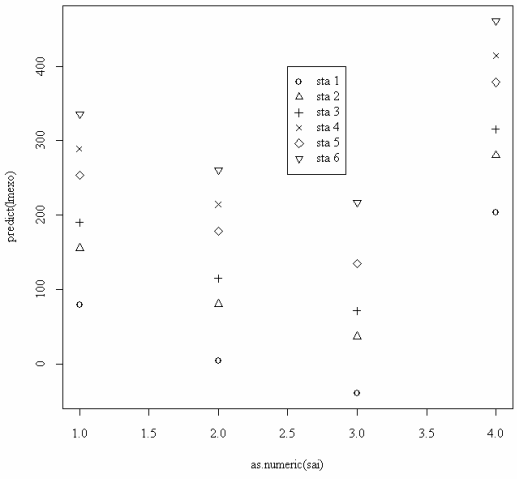
```
> meaudret <- read.table("meaudret.txt",h=T)
```

On a mesuré 9 variables, respectivement 1- Température (°C), 2- Débit (l/s), 3- pH, 4- Conductivité (mmho/cm), 5- Oxygène (% saturation), 6- DBO5 (mg/l oxygène), 7- Ammoniaque (mg/l), 8- Nitrates (mg/l) et 9- Orthophosphates (mg/l) dans 6 stations d'un réseau hydrographique à 4 saisons (printemps, été, automne, hiver).

Supposons que le relevé de la station 4 à la date 3 soit perdu. Estimer les données manquantes.

```
> exo <- meaudret[-15,]
> sta <- as.factor(exo$sta)
> sai <- as.factor(exo$sai)
> deb <- exo$deb
> sta
 [1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 5 5 5 5 6 6 6 6
Levels:  1 2 3 4 5 6
> sai
 [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 4 1 2 3 4 1 2 3 4
Levels:  1 2 3 4
> deb
 [1]  41  62  25 118 158  80  63 252 198 100  79 315 280 140 498 322 160  72
390
[20] 303 310 181 480

> lmexo <- lm(deb~sta+sai)
> lmexo

Call:
lm(formula = deb ~ sta + sai)

Coefficients:
(Intercept)         sta2         sta3         sta4         sta5         sta6
       78.6         76.8        111.5        210.7        174.5        257.0
       sai2         sai3         sai4
      -75.0       -118.5        125.2
> plot(as.numeric(sai),predict(lmexo),pch=as.numeric(sta))
> legend(2.5,400,paste("sta",1:6),pch=1:6)
> plot(as.numeric(sai),deb,pch=as.numeric(sta))
> legend(2.5,400,paste("sta",1:6),pch=1:6)
```





Pointer le valeur manquante.

```
> 78.6+210.7-118.5
[1] 170.8
> tapply(deb,sta,mean)
    1     2     3     4     5     6
 61.5 138.3 173.0 306.0 236.0 318.5
```

```
> tapply(deb,sai,mean)
    1     2     3     4
217.0 142.0  84.0 342.2
```

Discuter en fonction de la carte des stations. Répéter l'exercice pour les autres variables.

2.3. Plans non orthogonaux

On en arrive à la notion délicate d'effets partiels. Quand il manque des données dans une case ou au contraire quand chaque case contient plusieurs données en nombre inégal, on dit que le plan n'est pas orthogonal. Quand le plan est complet avec une valeur par case, ou plusieurs valeurs en nombre égal, on dit que le plan est orthogonal. Le terme employé renvoie directement à la géométrie des variables.

Prenons l'exemple d'un plan 2x2 à 2 répétitions :

numéro	A1	A2	B1	B2		numéro	Constante	A	B
1	1	0	1	0		1	1	-1/2	-1/2
2	1	0	1	0		2	1	-1/2	-1/2
3	1	0	0	1		3	1	-1/2	1/2
4	1	0	0	1	⇒	4	1	-1/2	1/2
5	0	1	1	0		5	1	1/2	-1/2
6	0	1	1	0		6	1	1/2	-1/2
7	0	1	0	1		7	1	1/2	1/2
8	0	1	0	1		8	1	1/2	1/2

Le sous-espace de projection est de dimension 3 et on en connaît une base orthogonale. La projection se fait directement en calculant des moyennes et le tableau d'analyse de variance se calcule directement avec des sommes de carrés. Il suffit qu'il manque une valeur pour que cet édifice soit détruit :

numéro	A1	A2	B1	B2		numéro	Constante	A	B
1	1	0	1	0		1	1	-3/7	-3/7
2	1	0	1	0		2	1	-3/7	-3/7
3	1	0	0	1		3	1	-3/7	4/7
4	1	0	0	1	⇒	4	1	-3/7	4/7
5	0	1	1	0		5	1	4/7	-3/7
6	0	1	1	0		6	1	4/7	-3/7
7	0	1	0	1		7	1	4/7	4/7
*	*	*	*	*		*	*	*	*

Il peut s'agir d'une simple perturbation numérique, par exemple quand on a environ 10 mesures par case (parfois 8, parfois 11, …) et dans ce cas le logiciel réglera le problème. Il peut s'agir d'un problème fondamental qui va jusqu'à la confusion des facteurs. Nous nous placerons dans le premier cas.

Que le plan soit orthogonal ou qu'il soit numériquement perturbé, se pose la question du rôle respectif des facteurs. La réponse est fort différente dans les deux cas. La même question est en jeu dans la régression multiple.

Remarque préalable : deux modèles identiques et deux tests différents.

```
> lma <- lm(deb~sai)
> lmb <- lm(deb~-1+sai)
> anova(lma)
Analysis of Variance Table

Response: deb
          Df Sum Sq Mean Sq F value Pr(>F)
sai        3 210434   70145    6.17 0.0041 **
Residuals 19 215957   11366

> anova(lmb)
Analysis of Variance Table

Response: deb
          Df  Sum Sq Mean Sq F value  Pr(>F)
sai        4 1141266  285317    25.1 2.3e-07 ***
Residuals 19  215957   11366

> predict(lma)
    1     2     3     4     5     6     7     8     9    10    11    12    13
217.0 142.0  84.0 342.2 217.0 142.0  84.0 342.2 217.0 142.0  84.0 342.2 217.0
   14    15    16    17    18    19    20    21    22    23
142.0 342.2 217.0 142.0  84.0 342.2 217.0 142.0  84.0 342.2
> predict(lmb)
    1     2     3     4     5     6     7     8     9    10    11    12    13
217.0 142.0  84.0 342.2 217.0 142.0  84.0 342.2 217.0 142.0  84.0 342.2 217.0
   14    15    16    17    18    19    20    21    22    23
142.0 342.2 217.0 142.0  84.0 342.2 217.0 142.0  84.0 342.2

> coefficients(lma)
(Intercept)        sai2        sai3        sai4
      217.0       -75.0      -133.0       125.2
> coefficients(lmb)
 sai1  sai2  sai3  sai4
217.0 142.0  84.0 342.2
```

Quand on estime un modèle à deux facteurs, on projette sur un sous-espace du type A + B. Quand on estime le modèle sur le seul facteur A, on projette sur A. Quand on estime sur le seul facteur B, on projette sur B. Quand on cherche à mesurer la part de A et de B dans le modèle mixte, intervient la géométrie respective des sous-espaces A et B dans le sous-espace A+B.

Un modèle est un vecteur de A+B, donc un vecteur de type $\mathbf{a}+\mathbf{b}$ où $\mathbf{a}\in \mathrm{A}, \mathbf{b}\in \mathrm{B}$. L'écriture n'est pas unique à cause du vecteur $\mathbf{1}_n$. Mais il suffit de savoir que $\mathrm{A+B}\supseteq \mathrm{A}$.

On est alors dans la stratégie des modèles emboîtés et on a le tableau d'analyse de variance :

Source de variation		DDL	SC	CM	F	Proba
Facteur A	$\hat{Y}_A - \overline{Y}$	m_A -1	$\left\|\hat{Y}_A - \overline{Y}\right\|^2$	$CMA = \frac{\left\|\hat{Y}_A - \overline{Y}\right\|^2}{m_A - 1}$	$\frac{CMA}{CMR}$	
Facteur B sachant A	$\hat{Y}_{A+B} - \hat{Y}_A$	$m_B - 1$	$\left\|\hat{Y}_{A+B} - \hat{Y}_A\right\|^2$	$CMB = \frac{\left\|\hat{Y}_{A+B} - \overline{Y}\right\|^2}{m_B - 1}$	$\frac{CMB}{CMR}$	
Résiduelle	$Y - \hat{Y}_{A+B}$	$n - m_A - m_B + 1$	$\left\|Y - \hat{Y}_{A+B}\right\|^2$	$CMR = \frac{\left\|Y - \hat{Y}_{A+B}\right\|^2}{\left(n - m_A - m_B + 1\right)}$		
Totale	$Y - \overline{Y}$	$n-1$	$\left\|Y - \overline{Y}\right\|^2$			

```
> sai
 [1] 1 2 3 4 1 2 3 4 1 2 3 4 1 2 4 1 2 3 4 1 2 3 4
Levels:  1 2 3 4
> sta
 [1] 1 1 1 1 2 2 2 2 3 3 3 3 4 4 4 5 5 5 5 6 6 6 6
Levels:  1 2 3 4 5 6
> deb
 [1]  41  62  25 118 158  80  63 252 198 100  79 315 280 140 498 322 160  72
390
[20] 303 310 181 480
> lm2 <- lm(deb~sta+sai)
> anova(lm2)
Analysis of Variance Table

Response: deb
          Df Sum Sq Mean Sq F value  Pr(>F)
sta        5 189927   37985    11.8 0.00013 ***
sai        3 191331   63777    19.8 2.6e-05 ***
Residuals 14  45134    3224

> lm3 <- lm(deb~sai+sta)
> anova(lm3)
Analysis of Variance Table

Response: deb
          Df Sum Sq Mean Sq F value  Pr(>F)
sai        3 210434   70145    21.8 1.5e-05 ***
sta        5 170823   34165    10.6 0.00023 ***
Residuals 14  45134    3224
```

Les deux modèles sont identiques et les résiduelles sont égales. Par contre, **parce que le plan est incomplet**, le sous-espace B sachant A (la partie de A+B formé de vecteurs orthogonaux à A) n'est pas exactement la partie de B composée des vecteurs orthogonaux à $\mathbf{1}_n$. Cette difficulté provient de la nécessité de décomposer la variation totale en vecteurs orthogonaux (théorème de Cochran). Vérifier qu'elle disparaît dans le plan complet sans répétitions.

La même question se retrouve en régression multiple. Considérons l'exemple suivant. Aux élections européennes de 1984, le candidat de l'extrême droite avait obtenu pour la première fois un score important. Il est calculé par région administrative (n = 21) dans la colonne score du tableau ci-dessous. Dans les mêmes régions, on connaît le taux de population immigrée de l'époque (immi, en 1/100), le taux de chômage (chom, en 1/100) et le taux d'urbanisation (urba, en 1/100). On veut expliquer le score avec les variables socio-économiques.

score	immi	chom	urba
14.5	13.3	7.1	93.6
10.7	5.4	9.5	62.4
10.8	4.6	9.7	60.7
8.9	3.3	11	69.1
9.3	5.1	7.8	62.9
7.6	1.7	9.8	53.4
10.1	5.4	8.6	57.9
9.1	4.8	11.8	86.4
12.4	8	9.2	72.4
12.5	8.1	7.4	73.2
12	7.4	8.2	58.8
6.8	1.4	9.6	60.1
6.8	0.7	9.4	55.6
6.7	1.7	10	50.5
8.3	4.6	9.5	64.6
8.1	4.8	8.5	59.3
4.8	2.7	6.9	50.9
12.9	9.1	7.5	76.9
7.4	4.6	8.3	58.2
13.2	6.5	11.4	70.7
19	8.2	10.5	89.6

```
> elec <- read.table("euro84.txt",h=T)
> cor(elec)
        score     imm     chom   urba
score 1.00000  0.8142  0.04734 0.7687
imm   0.81419  1.0000 -0.35454 0.7616
chom  0.04734 -0.3545  1.00000 0.1306
urba  0.76865  0.7616  0.13061 1.0000


> anova(lm(score~imm+urba,data=elec))
Analysis of Variance Table

Response: score
          Df Sum Sq Mean Sq F value  Pr(>F)
imm        1  143.7   143.7   41.93 4.3e-06 ***
urba       1   11.4    11.4    3.32   0.085 .
Residuals 18   61.7     3.4

> anova(lm(score~urba+imm,data=elec))
Analysis of Variance Table

Response: score
          Df Sum Sq Mean Sq F value Pr(>F)
urba       1  128.1   128.1   37.37  9e-06 ***
imm        1   27.0    27.0    7.88  0.012 *
Residuals 18   61.7     3.4
```

Les deux tableaux d'analyse de la variance sont du type :

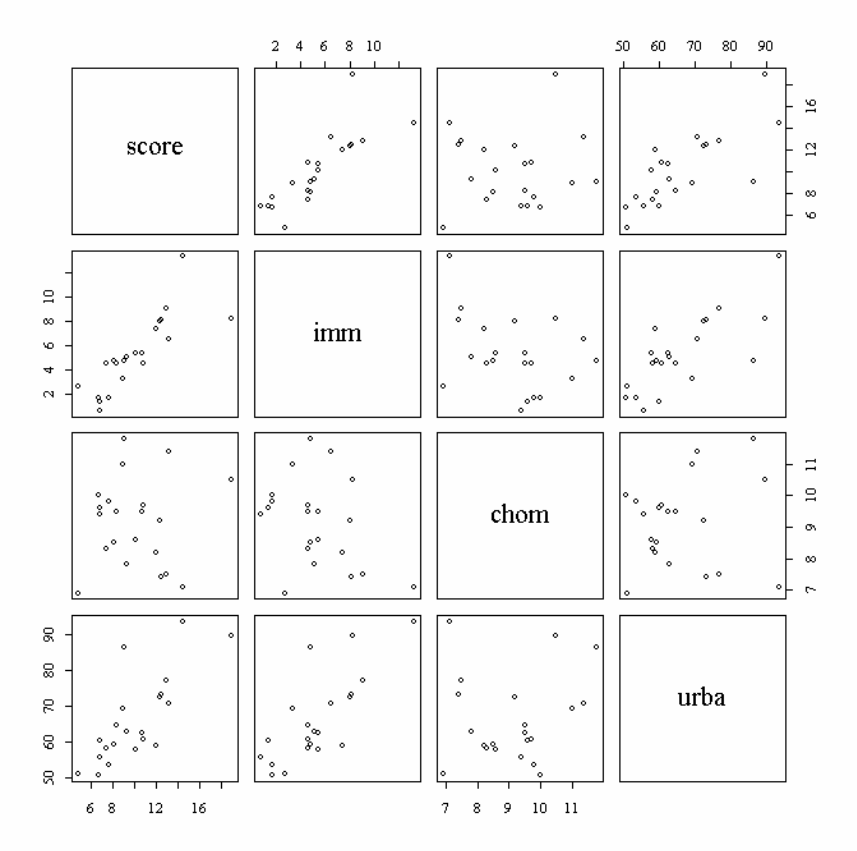
Source de variation		DDL	SC	CM	F	Proba
Facteur x	$\hat{Y}_x - \overline{Y}$	1	$\left\|\hat{Y}_x - \overline{Y}\right\|^2$	$CMA = \left\|\hat{Y}_x - \overline{Y}\right\|^2$	$\frac{CMx}{CMR}$	
Facteur z sachant x	$\hat{Y}_{x+z} - \hat{Y}_A$	1	$\left\|\hat{Y}_{x+z} - \hat{Y}_x\right\|^2$	$CMB = \left\|\hat{Y}_{x+z} - \overline{Y}\right\|^2$	$\frac{CMz}{CMR}$	
Résiduelle	$Y - \hat{Y}_{x+z}$	$n-3$	$\left\|Y - \hat{Y}_{x+z}\right\|^2$	$CMR = \frac{\left\|Y - \hat{Y}_{x+z}\right\|^2}{(n-3)}$		
Totale	$Y - \overline{Y}$	$n-1$	$\left\|Y - \overline{Y}\right\|^2$			

On vérifie le contenu théorique du tableau d'analyse de la variance dans le premier cas.

```
> ybar <- rep(mean(elec$score),21)
> ybar
 [1] 10.09 10.09 10.09 10.09 10.09 10.09 10.09 10.09 10.09 10.09 10.09 10.09
[13] 10.09 10.09 10.09 10.09 10.09 10.09 10.09 10.09 10.09
> yx <- predict(lm(elec$score~elec$imm))
> sum((yx-ybar)^2)
[1] 143.7
> yxz <- predict(lm(elec$score~elec$imm+elec$urba))
> sum((yxz-yx)^2)
[1] 11.39
> y <- elec$score
> sum((y-yxz)^2)
[1] 61.7
> sum((yx-ybar)*(yxz-yx))
[1] -8.563e-15
> sum((yx-ybar)*(y-yxz))
[1] -1.884e-14
> sum((yxz-yx)*(y-yxz))
[1] 4.493e-15
> sum((yxz-yx)^2+(y-yxz)^2+(yx-ybar)^2)
[1] 216.8
> sum((y-ybar)^2)
[1] 216.8
```

L'orthogonalité (somme des produits nulle) est fondamentale. Elle assure la décomposition.

```
> pairs(elec)
```



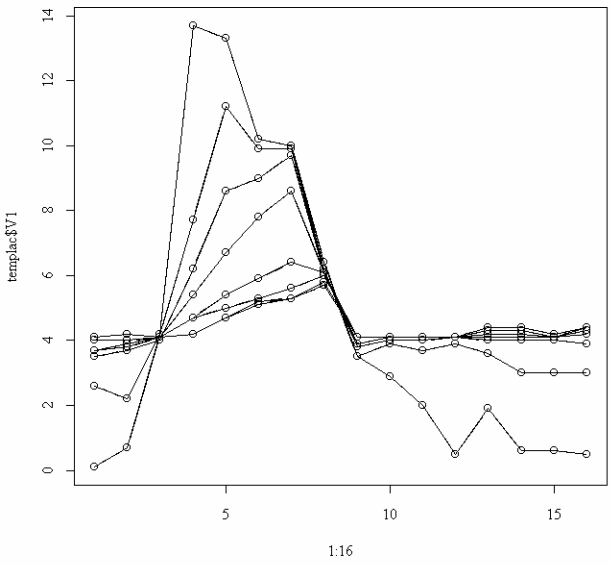
Les deux explicatives sont corrélées et c'est la source de difficulté principale. Lorsque les deux variables explicatives sont de corrélation nulle les deux ordres d'introduction des variables donnent les mêmes résultats. L'exemple est choisi pour montrer qu'entre le résultat statistique et l'interprétation qui en est faîte une part irréductible est laissée à l'utilisateur. On devra choisir entre « le vote d'extrême droite est un phénomène lié à l'urbanisation et légèrement amplifié par l'immigration » ou « le vote d'extrême droite est un phénomène lié à l'immigration ». Si on oublie de préciser que les deux facteurs sont liés, ou ce qui est fréquent, si on l'ignore, l'inconséquence du discours tenu est garantie.

2.4. Notion d'interaction

On sait maintenant que l'analyse de la variance est basée sur la décomposition d'un vecteur (la variable à prédire centrée) en somme de vecteurs orthogonaux. Le principe très général a ses limites. Il s'étend de lui-même dans des décompositions plus complexes qui ont souvent un intérêt expérimental considérable.

Considérons la température de l'eau dans un lac d'altitude[2] à 16 dates (lignes) et 8 profondeurs (colonnes. V1 = -2.5 m, V2 = -5 m, V3 = -7.5 m , …, V7 = -17.5 m, V8 = -19 m) :

```
> templac
     V1   V2  V3  V4  V5  V6  V7  V8
1   0.1  2.6 3.5 3.7 3.7 3.7 4.0 4.1
2   0.7  2.2 3.7 3.8 3.8 3.9 4.0 4.2
3   4.2  4.2 4.0 4.1 4.1 4.1 4.1 4.1
4  13.7  7.7 6.2 5.4 4.7 4.7 4.2 4.2
5  13.3 11.2 8.6 6.7 5.4 5.0 4.7 4.7
6  10.2  9.9 9.0 7.8 5.9 5.3 5.2 5.1
7  10.0  9.9 9.7 8.6 6.4 5.6 5.3 5.3
8   6.4  6.2 6.1 6.1 6.1 6.0 5.8 5.7
9   3.5  3.5 3.8 3.9 4.1 4.1 4.1 4.1
10  2.9  3.9 4.0 4.1 4.1 4.1 4.1 4.1
11  2.0  3.7 4.0 4.1 4.1 4.1 4.1 4.1
12  0.5  3.9 4.1 4.1 4.1 4.1 4.1 4.1
13  1.9  3.6 4.0 4.1 4.2 4.2 4.3 4.4
14  0.6  3.0 4.0 4.1 4.2 4.2 4.3 4.4
15  0.6  3.0 4.0 4.1 4.1 4.1 4.1 4.2
16  0.5  3.0 3.9 4.2 4.3 4.3 4.4 4.4

> plot(1:16,templac$V1,type="n")
> apply(templac,2,lines,x=1:16,type="b")
```



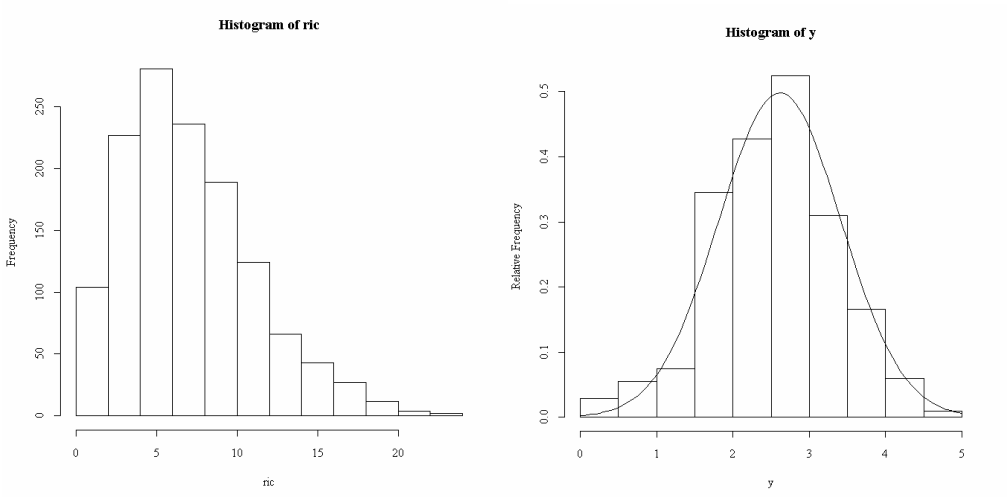
L'évolution de la température en fonction du temps dépend de la profondeur. On dit qu'il y a interaction entre le temps et la profondeur sur la température. **Il y a interaction entre deux facteurs sur une variable lorsque l'effet d'un facteur sur la variable est modifié par la valeur de l'autre facteur**. L'analyse de la variance propose une stratégie pour tester la présence d'une interaction. On aborde ici le cas de deux facteurs qualitatifs.

```
> ric
   [1]  5  3  5  3  4  1  5  3  2  3  1  2  6  4  6  2 11  2  9  4 12  5 12  8
  [25] 15  7  9 16 18 11 10  4  9  8 12 15 13 24 12 18  8 17 16 13  9  8 11 21
…
 [625]  8  8  9  7 10  8  6  6  6  8  7  5  5  6  7  6  4  5  5  7  9  7  7  5
…
```

[2] Chacornac, J.M. (1986) Lacs d'altitude : Métabolisme oligotrophique et approche typologique des écosystèmes. Thèse de doctorat. Universté Lyon 1. 1-214.

```
[1057] 16  9 11 17  6 15 10 14 12 19 12 13 15 19  5  6  9 10 15  6 19  7  9 11
…
[1273]  9  7  8  6  5  4  5  4  3  5  4  4  4  5  4  7  5  6  5  7  7  7  4  7
[1297]  4  7  4  4  6  7  4  5  4  7  2  1  5  5  3  5  4  4  7
```

On a compté le nombre d'espèces d'Oiseaux au cours de 1315 relevés ornithologiques régulièrement répartis sur les 52 semaines de l'année, dans 14 stations , le matin ou le soir [3]. On a un facteur sta (numéro de la station à 14 modalité), un facteur sem (numéro de la semaine à 52 modalités) et un facteur heu (matin, soir, 2 modalités).



La richesse (nombre d'espèces) a une distribution dissymétrique qui nuit à la qualité des modèles linéaires (trop grande influence des fortes valeurs). La variable à expliquer sera sa racine carrée :

```
> y <- sqrt(ric)
> hist(y,pro=T)
> w0 <- seq(0,5,le=100)
> lines(w0,dnorm(w0,mean(y),sqrt(var(y))))
```

La richesse dépend fortement de l'heure :

```
> anova(lm(y~heu))
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value Pr(>F)
heu         1    111     111     200 <2e-16 ***
Residuals 1313    731       1
```

La richesse dépend fortement de la station :

```
> anova(lm(y~sta))
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value Pr(>F)
sta        13    136      10    19.2 <2e-16 ***
Residuals 1301    707       1
```

[3] Convention d'étude n° 228/92 du Parc National des Ecrins.

La richesse dépend fortement de la semaine :

```
> anova(lm(y~sem))
Analysis of Variance Table

Response: y
            Df Sum Sq Mean Sq F value Pr(>F)
sem         51    218       4    8.62 <2e-16 ***
Residuals 1263    625    0.49
```

Comment tester une interaction ? Nous avons trois facteurs contrôlés et pour un couple de deux d'entre elles de nombreuses répétitions. Prenons le couple heu-sem. La première variable a deux modalités et la seconde 52. Il y a donc 104 possibilités heure-semaine. Une indicatrice d'un couple de modalités est exactement le produit terme à terme des indicatrices de chacune des deux modalités. Si A est le sous-espace engendré par les indicatrices de la première, si B est le sous-espace engendré par les indicatrices de la seconde, on connaît déjà A+B engendré par la réunion des deux paquets d'indicatrices et on peut définir AxB le sous-espace engendré par l'ensemble des produits terme à terme des indicatrices des deux paquets.

```
> table(heu,sem)
       sem
heu     1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
  matin 9 12 13 14 14 12 14 14 13 13 13 14 14 11 14 13 14 14 13 12 12 13 13 14
  soir  9 13 13 14 14 13 14 14 14 13 13 14 14 12 12 13 13 13 13 12 12 14 11 13
       sem
heu     25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47
48
  matin 13 13 12 14 12 14 13 12 12 11 13 14 13 14 11 14 13 14 12 12 12 13 14
12
  soir  14 11 12 14 12  9  9 13 11 10 13 13 12 14 10 13 11 12 12 12 14 12 11
11
       sem
heu     49 50 51 52
  matin 13 14 14 12
  soir  12 14 11 11
```

Ici il y a 104 indicatrices dans AxB car toutes les combinaisons sont représentées. Ce n'est pas vrai pour le couple semaine-station :

```
> table (sta,sem)
    sem
sta  1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27
28
  1  0 2 2 2 2 2 2 2 2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
2
…
  6  0 1 1 2 2 0 2 2 2  2  0  2  2  0  2  0  2  2  0  0  2  2  0  2  2  2  2
2
…
  14 2 2 2 2 2 2 2 2 2  2  2  2  2  2  2  2  2  2  2  2  2  2  1  2  2  2  2
2
    sem
sta  29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
  1   2  2  2  2  2  2  2  2  2  2  0  2  2  2  2  2  2  2  2  2  2  2  2  2
  2   2  2  2  2  2  2  2  2  2  2  0  2  2  2  2  2  2  2  2  2  2  2  2  2
  3   2  2  1  2  1  2  2  2  2  2  1  2  1  2  2  2  1  2  1  1  1  2  1  1
  4   2  2  2  2  2  2  2  2  2  2  0  2  2  2  0  0  2  2  2  2  0  2  2  0
…
  13  2  2  1  2  1  1  1  2  2  2  2  1  1  1  2  2  2  2  2  2  2  2  2  2
  14  2  2  2  2  2  2  2  1  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
> sum(table(sem,sta)==0)
[1] 41
```

Il y a 41 couples semaine-station sans visite. Il y a donc 52 x 14 – 41 = 687 modalités semaine-station. On notera $n_{AB} = \dim(\text{AxB})$.

Dans AxB, on trouve toutes les variables constantes par couple de valeurs sur A et B, en particulier les variables constantes par modalité de A, les variables constantes par modalité de B et les combinaisons linéaires des deux :

$$\left.\begin{array}{l} \text{AxB} \supseteq \text{A} \\ \text{AxB} \supseteq \text{B} \end{array}\right\} \Rightarrow \text{AxB} \supseteq \text{A+B}$$

Les sous-espaces A, A+B et AxB sont emboîtés, donc la décomposition en éléments orthogonaux est possible et on peut construire un tableau d'analyse de la variance du type :

Source de variation		DDL	SC	CM	F	Proba
Facteur A	$\hat{Y}_A - \overline{Y}$	m_A -1	$\left\|\hat{Y}_A - \overline{Y}\right\|^2$	$CMA = \frac{\left\|\hat{Y}_A - \overline{Y}\right\|^2}{(m_A - 1)}$	$\frac{CMA}{CMR}$	
Facteur B sachant A	$\hat{Y}_{A+B} - \hat{Y}_A$	$m_B - 1$	$\left\|\hat{Y}_{A+B} - \hat{Y}_A\right\|^2$	$CMB = \frac{\left\|\hat{Y}_{A+B} - \overline{Y}\right\|^2}{(m_B - 1)}$	$\frac{CMB}{CMR}$	
Interaction	$\hat{Y}_{AxB} - \hat{Y}_{A+B}$	$m_{AxB} - m_A - m_B + 1$	$\left\|\hat{Y}_{AxB} - \hat{Y}_{A+B}\right\|^2$	$CMI = \frac{\left\|\hat{Y}_{AxB} - \hat{Y}_{A+B}\right\|^2}{(m_{AxB} - m_A - m_B - 1)}$	$\frac{CMI}{CMR}$	
Résiduelle	$Y - \hat{Y}_{AxB}$	$n - m_{AxB}$	$\left\|Y - \hat{Y}_{AxB}\right\|^2$	$CMR = \frac{\left\|Y - \hat{Y}_{AxB}\right\|^2}{(n - m_{AxB})}$		
Totale	$Y - \overline{Y}$	$n - 1$	$\left\|Y - \overline{Y}\right\|^2$			

```
> anova(lm(y~sem*heu))
Analysis of Variance Table

Response: y
            Df Sum Sq Mean Sq F value Pr(>F)
sem         51    218       4   10.46 <2e-16 ***
heu          1    112     112  275.12 <2e-16 ***
sem:heu     51     19    0.37    0.91   0.66
Residuals 1211    494    0.41

> anova(lm(y~sem*sta))
Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value Pr(>F)
sem        51  217.7     4.3   11.03 <2e-16 ***
sta        13  135.5    10.4   26.95 <2e-16 ***
sem:sta   622  246.6     0.4    1.02   0.38
Residuals 628  243.0     0.4

> anova(lm(y~sta*heu))
Analysis of Variance Table

Response: y
            Df Sum Sq Mean Sq F value  Pr(>F)
sta         13    136      10   23.19 < 2e-16 ***
heu          1    110     110  244.05 < 2e-16 ***
sta:heu     13     17       1    2.93 0.00033 ***
Residuals 1287    580 0.45056
```

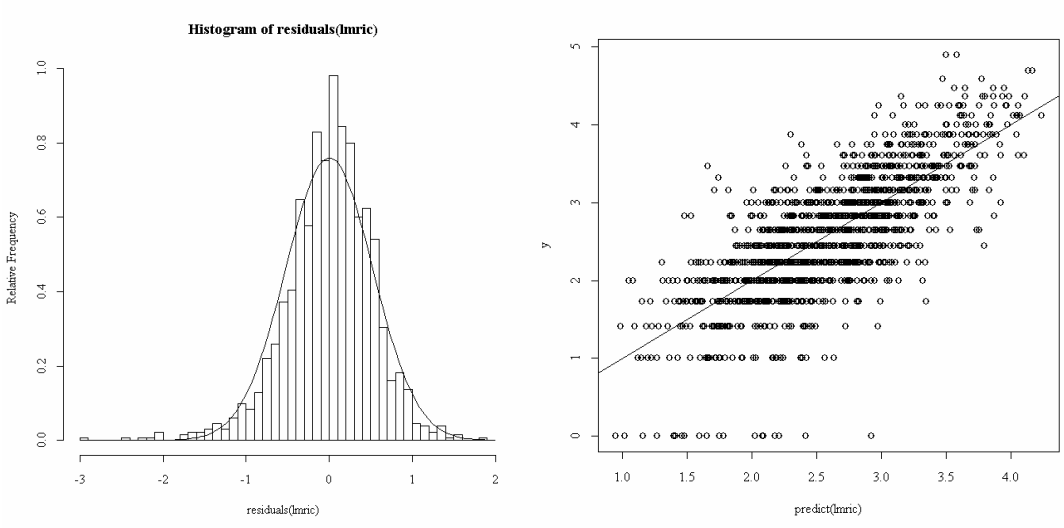
On admet une interaction station-heure et on construit le modèle :

```
> lmric <- lm(y~sem+heu*sta)
```

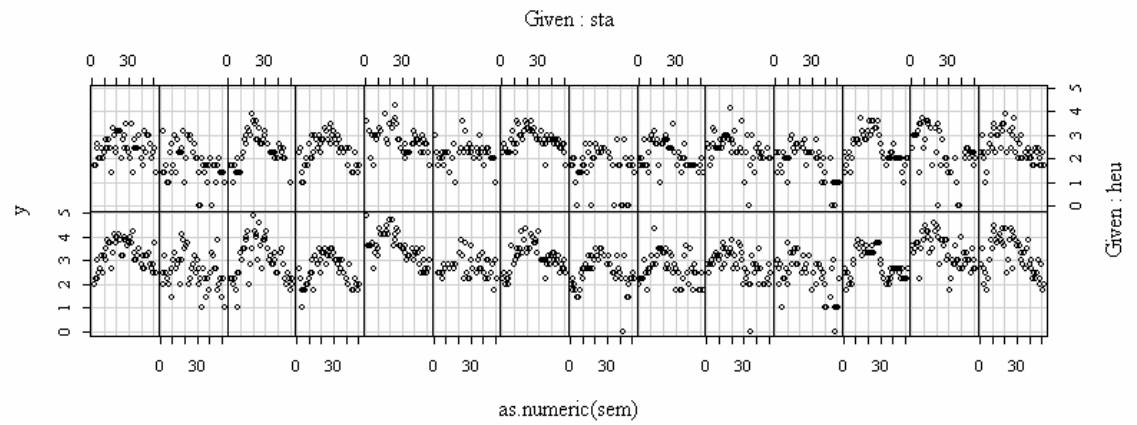
On étudie ses résidus et ses prédictions (commenter).

```
> hist(residuals(lmric),pro=T,nclass=50)
```

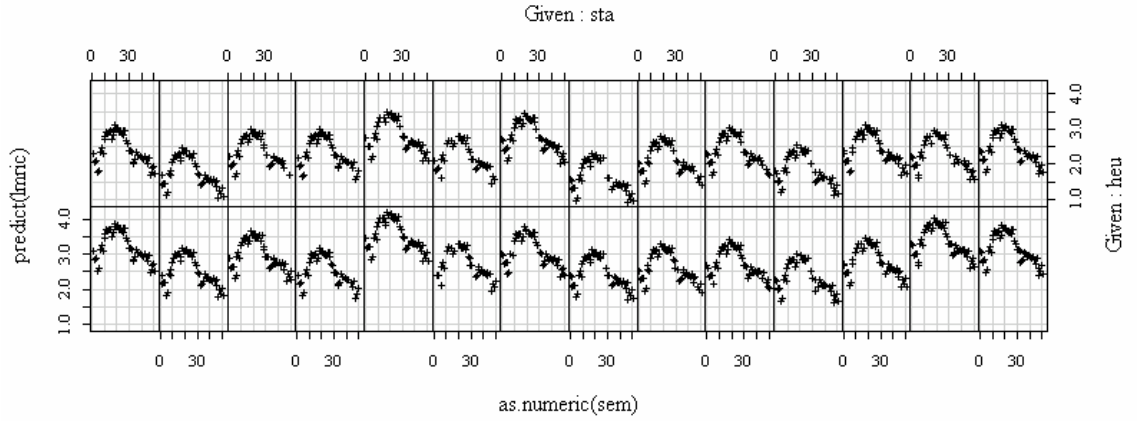
```
> w0 <- seq(-3,2,le=100)
> lines (w0,dnorm(w0,0,sqrt(var(residuals(lmric)))))
> plot(predict(lmric),y)
> abline(0,1)
```




```
> coplot(y~as.numeric(sem)|sta*heu,show=F)
```




```
> coplot(predict(lmric)~as.numeric(sem)|sta*heu,show=F,pch="+")
```



On a choisi de tracer la même courbe temporelle décalée d'une constante qui dépend du couple heure-station. Le tableau d'analyse de la variance se complique encore **mais reste basé sur le même principe** :

```
> anova(lmric)
Analysis of Variance Table

Response: y
            Df Sum Sq Mean Sq F value  Pr(>F)
sem         51    218       4   14.57 < 2e-16 ***
heu          1    112     112  383.05 < 2e-16 ***
sta         13    134      10   35.22 < 2e-16 ***
heu:sta     13     17       1    4.36 3.1e-07 ***
Residuals 1236    362 2.9e-01
```

Les espaces emboîtés sont maintenant A, A+B, A+B+C, A+BxC.

3. L'analyse de la covariance

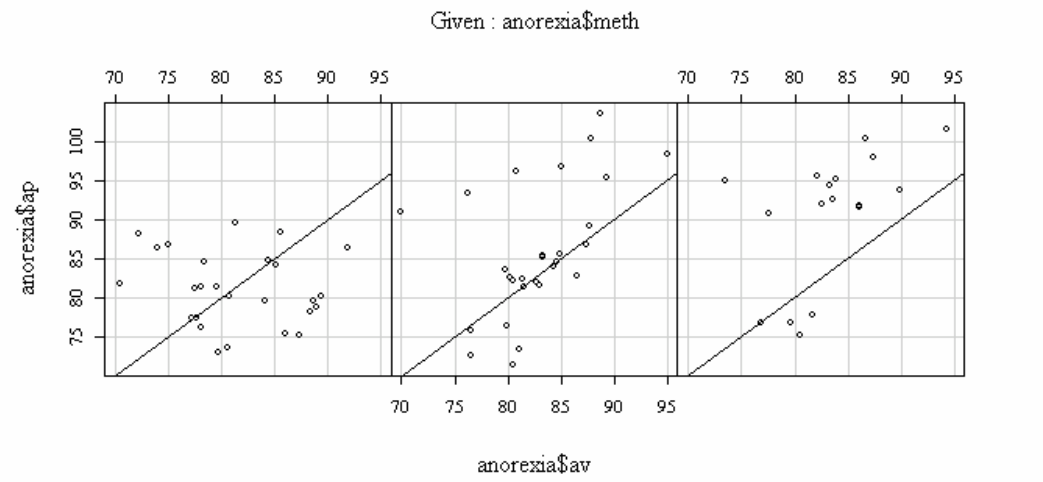
C'est le cas particulier de l'analyse de l'interaction entre une variable qualitative et une variable quantitative. Prenons l'exemple proposé par B. Everitt [4].

Traitement de psychologie comprtementale		Traitement standard		Thérapie familiale	
Avant	Après	Avant	Après	Avant	Après
80.5	82.2	80.7	80.2	83.8	95.2
84.9	85.6	89.4	80.1	83.3	94.3
81.5	81.4	91.8	86.4	86	91.5
82.6	81.9	74	86.3	82.5	91.9
79.9	76.4	78.1	76.1	86.7	100.3
88.7	103.6	88.3	78.1	79.6	76.7
94.9	98.4	87.3	75.1	76.9	76.8
76.3	93.4	75.1	86.7	94.2	101.6
81	73.4	80.6	73.5	73.4	94.9
80.5	82.1	78.4	84.6	80.5	75.2
85	96.7	77.6	77.4	81.6	77.8
89.2	95.3	88.7	79.5	82.1	95.5
81.3	82.4	81.3	89.6	77.6	90.7
76.5	72.5	78.1	81.4	83.5	92.5
70	90.9	70.5	81.8	89.9	93.8
80.4	71.3	77.3	77.3	86	91.7
83.3	85.4	85.2	84.2	87.3	98
83	81.6	86	75.4		
87.7	89.1	84.1	79.5		
84.2	83.9	79.7	73		
86.4	82.7	85.5	88.3		
76.5	75.7	84.4	84.7		
80.2	82.6	79.6	81.4		
87.8	100.4	77.5	81.2		
83.3	85.2	72.3	88.2		
79.7	83.6	89	78.8		
84.5	84.6				
80.8	96.2				
87.4	86.7				

On a le poids (supposons qu'il s'agit de livres et non de kg comme indiqué dans l'ouvrage) avant et après traitement de jeunes filles soignées pour l'anorexie (trois méthodes). « It is instructive to look at the three scatterplots of after/before » disent les auteurs.

[4] in Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J. & Ostrowski, E. (1994) A handbook of small data sets. Chapman & Hall, London. 1-458. p. 229 n° 285 Anorexia data.

```
> coplot(anorexia$ap~anorexia$av|anorexia$meth,row=1,show=F,panel=f1)
> f1
function(x,y,col="black",pch=1,...){
points(x,y)
abline(0,1)}
```



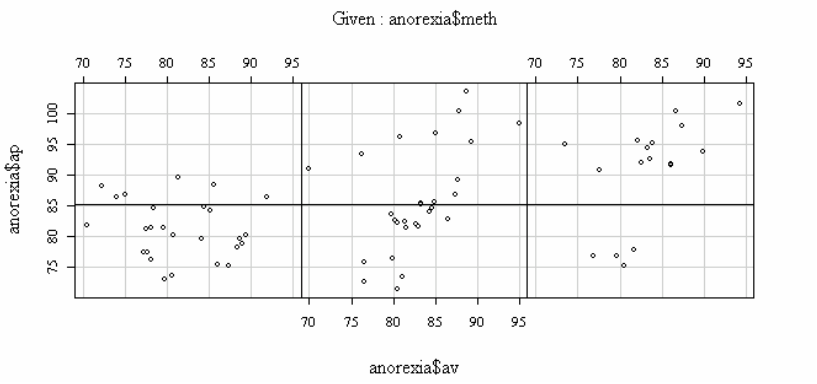
On a utilise le modèle **y** = **x** et les données. Interpréter directement.

La variable av (explicative **x**) est le poids avant traitement. La variable ap (à prédire **y**) est le poids après traitement. La variable meth (explicative qualitative **z**) est la méthode de traitement à trois modalités. Passons en revue quelques modèles. A est le sous-espace vectoriel engendré par **x**. Il est de dimension 1. $\mathbf{1}_n$ est le vecteur habituel des constantes. B est le sous-espace engendré par les indicatrices de **z**. Il est de dimension 3. Le complémentaire de $\mathbf{1}_n$ dans B est de dimension 2. Le sous-espace AxB contient les trois variables issues du produit terme à terme de **x** par les indicatrices de **z**. **Mais, contrairement au cas de deux facteurs qualitatifs, ces vecteurs ne suffisent pas à définir un espace contenant les précédents, ce qui est indispensable**. Il faut y rajouter les indicatrices des classes. Les espaces ont alors une constitution très simple :

$$
\begin{array}{cccccc}
 & & & & & \\
1 & x_{11} & 1 & 0 & 0 & 1\ 0\ 0\ x_{11}\ 0\ 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots \\
1 & x_{1m_1} & 1 & 0 & 0 & 1\ 0\ 0\ x_{1m_1}\ 0\ 0 \\
1 & x_{21} & 0 & 1 & 0 & 0\ 1\ 0\ 0\ x_{21}\ 0 \\
\mathbf{1}_n \rightarrow \vdots & \text{A} \rightarrow \vdots & \text{B} \rightarrow \vdots & \vdots & \vdots & \text{AxB} \rightarrow \vdots\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots \\
1 & x_{2m_2} & 0 & 1 & 0 & 0\ 1\ 0\ 0\ x_{2m_2}\ 0 \\
1 & x_{31} & 0 & 0 & 1 & 0\ 0\ 1\ 0\ 0\ x_{31} \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots\ \vdots\ \vdots\ \vdots\ \vdots\ \vdots \\
1 & x_{3m_3} & 0 & 0 & 1 & 0\ 0\ 1\ 0\ 0\ x_{3m_3}
\end{array}
$$

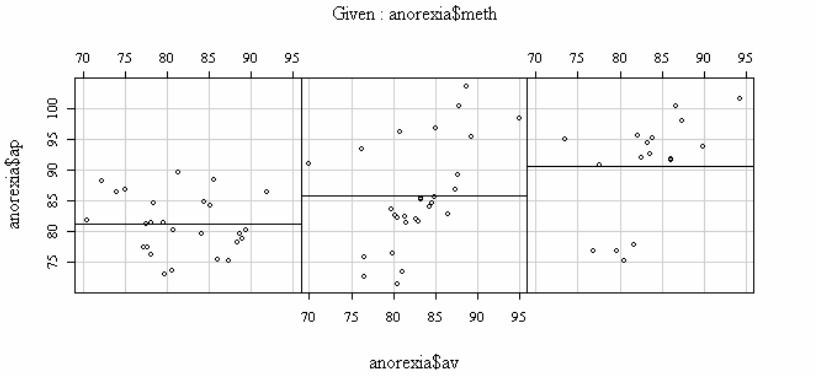
Le modèle $\mathbf{1}_n$:

```
> m0 <- mean(anorexia$ap)
> coplot(anorexia$ap~anorexia$av|anorexia$meth,row=1,show=F,panel=f2)
> f2
function(x,y,col="black",pch=1,...){
points(x,y)
abline(h=m0)}
```



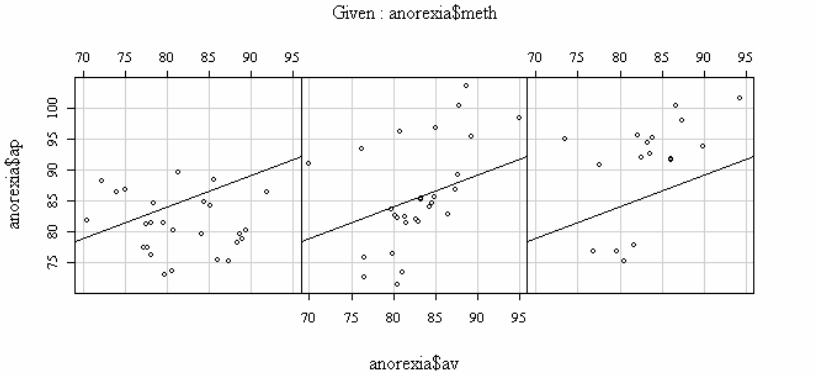
Le modèle B :

```
> coplot(anorexia$ap~anorexia$av|anorexia$meth,row=1,show=F,panel=f2)
> f2
function(x,y,col="black",pch=1,...){
points(x,y)
abline(h=mean(y))}
```



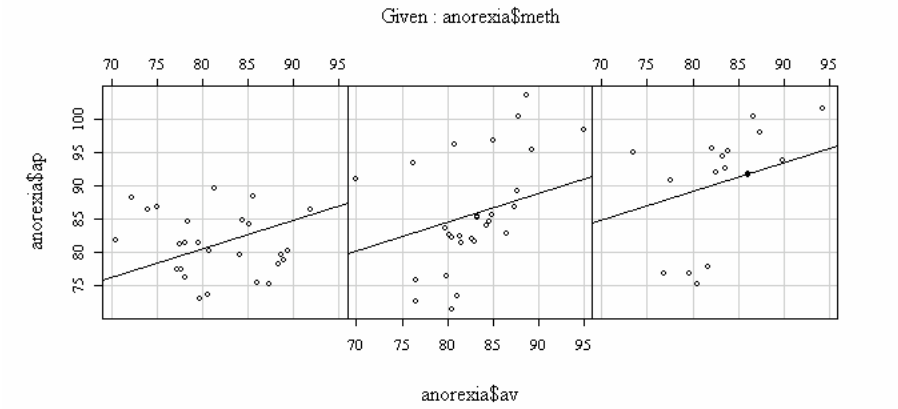
Le modèle A :

```
> w0 <- coefficients(lm(ap~av,data=anorexia))
> w0
(Intercept)          av
    42.7006      0.5154
> coplot(anorexia$ap~anorexia$av|anorexia$meth,row=1,show=F,panel=f2)
> f2
function(x,y,col="black",pch=1,...){
points(x,y)
abline(w0)}
```



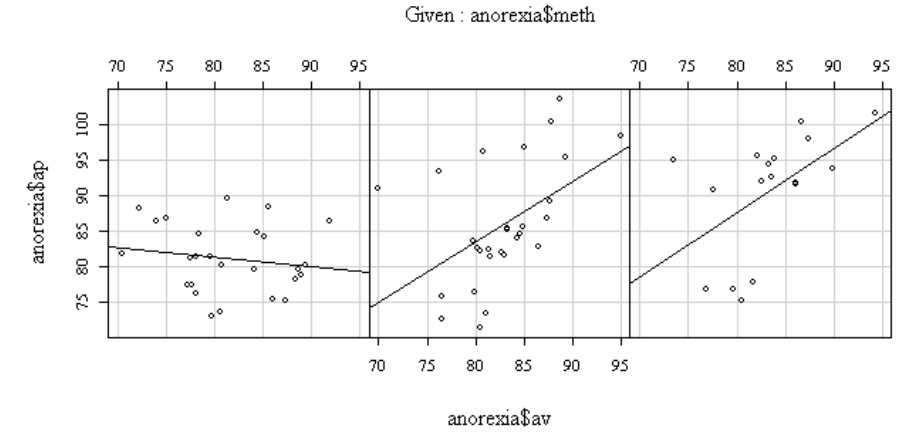
Le modèle A+B :

```
> w0 <- coefficients(lm(ap~-1+av+meth,data=anorexia))
> w0
     av   methc  methcb  methft
 0.4345 45.6740 49.7711 54.3342
> coplot(anorexia$ap~anorexia$av|anorexia$meth,row=1,show=F,panel=f2)
> f2
function(x,y,col="black",pch=1,...){
points(x,y)
abline(mean(y)-w0[1]*mean(x),w0[1])}
```



Le modèle A+B+AxB :

```
> coplot(anorexia$ap~anorexia$av|anorexia$meth,row=1,show=F,panel=f2)
> f2
function(x,y,col="black",pch=1,...){
points(x,y)
abline(lm(y~x)) }
```



Pour choisir :

```
> lmano <- lm(ap~av*meth,data=anorexia)
> lmano

Call:
lm(formula = ap ~ av * meth, data = anorexia)

Coefficients:
(Intercept)           av       methcb       methft    av.methcb    av.methft
     92.051       -0.134      -76.474      -77.232        0.982        1.043

> contrasts(anorexia$meth)
   cb ft
c   0  0
cb  1  0
ft  0  1
```

```
> anova(lmano)
Analysis of Variance Table

Response: ap
          Df Sum Sq Mean Sq F value  Pr(>F)
av         1    507     507   11.75 0.00105 **
meth       2    766     383    8.89 0.00038 ***
av:meth    2    466     233    5.41 0.00667 **
Residuals 66   2845      43
```

La théorie générale a engendré la tableau de l'analyse de covariance :

Source de variation		DDL	SC	CM	F	Proba
Facteur A	$\hat{Y}_A - \bar{Y}$	1	$\left\|\hat{Y}_A - \bar{Y}\right\|^2$	$CMA = \left\|\hat{Y}_A - \bar{Y}\right\|^2$	$\frac{CMA}{CMR}$	
Facteur B sachant A	$\hat{Y}_{A+B} - \hat{Y}_A$	$m_B - 1$	$\left\|\hat{Y}_{A+B} - \hat{Y}_A\right\|^2$	$CMB = \frac{\left\|\hat{Y}_{A+B} - \bar{Y}\right\|^2}{(m_B - 1)}$	$\frac{CMB}{CMR}$	
Interaction	$\hat{Y}_{AxB} - \hat{Y}_{A+B}$	$m_B - 1$	$\left\|\hat{Y}_{AxB} - \hat{Y}_{A+B}\right\|^2$	$CMI = \frac{\left\|\hat{Y}_{AxB} - \hat{Y}_{A+B}\right\|^2}{(m_B - 1)}$	$\frac{CMI}{CMR}$	
Résiduelle	$Y - \hat{Y}_{AxB}$	$n - 2m_B$	$\left\|Y - \hat{Y}_{AxB}\right\|^2$	$CMR = \frac{\left\|Y - \hat{Y}_{AxB}\right\|^2}{(n - 2m_B)}$		
Totale	$Y - \bar{Y}$	$n - 1$	$\left\|Y - \bar{Y}\right\|^2$			

Les deux facteurs peuvent cependant être introduits dans l'autre ordre :

```
> lmano1 <- lm(ap~meth*av,data=anorexia)
> anova(lmano1)
Analysis of Variance Table

Response: ap
          Df Sum Sq Mean Sq F value  Pr(>F)
meth       2    919     459   10.66 9.7e-05 ***
av         1    354     354    8.21  0.0056 **
meth:av    2    466     233    5.41  0.0067 **
Residuals 66   2845      43
---
Signif. codes:  0  `***'  0.001  `**'  0.01  `*'  0.05  `.'  0.1  ` '  1
```

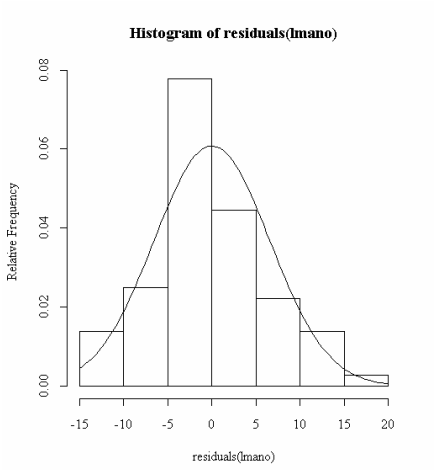
Les résultats sont cohérents parce que les explicatives ne sont pas corrélées :

```
> anova(lm(av~meth,data=anorexia))
Analysis of Variance Table

Response: av
          Df Sum Sq Mean Sq F value Pr(>F)
meth       2     33      16     0.6   0.55
Residuals 69   1874      27
```

Fondamentalement, les résultats obtenus par les trois méthodes sont très différents. La vérification de la normalité des résidus doit achever toujours l'analyse :

```
> hist(residuals(lmano),proba=T)
> w0 <- seq(-15,20,le=100)
> lines(w0,dnorm(w0,0,sqrt(43)))
```

Pourquoi 43 ?



4. Perspectives

En conclusion, on notera que nous n'avons en rien épuisé la question. On retiendra les principes fondamentaux.

a) Analyser des données, ce n'est pas utiliser une technique, mais construire progressivement un modèle acceptable des données.
b) Les techniques basées sur l'analyse de variance supposent la normalité des résidus. Une observation *y* est la réalisation d'une variable aléatoire normale de moyenne μ et de variance σ^2. μ est une fonction des variables de contrôle et σ^2 est une constante à estimer.

c) Les modèles sont définis par des sous-espaces vectoriels engendrés par les variables quantitatives ou les modalités des variables qualitatives ou une combinaison des deux. Une série de sous-espaces emboîtés définit un tableau d'analyse de la variance. Les sous-espaces définissent les estimations des modèles par projection. Ce langage est une nécessité. Donnons un exemple[5]. « *Un aspect important de toute planification est de n'attribuer à un effet que ce qui lui est réellement dû. Si tel n'est pas le cas, les conclusions que nous tirerons de l'analyse statistique seront erronées : on dit qu'il y a confusion d'effets. Deux effets sont confondus (effets principaux ou interactions d'un ordre quelconque) si les espaces vectoriels qui leur correspondent ont en commun un sous-espace de dimension non nulle. Il y a confusion totale si les deux espaces coïncident et confusion partielle si un espace est contenu dans l'autre.* » L'organisation des travaux expérimentaux s'appuie sur la théorie *des plans d'expériences.*

d) Nous n'avons parler que d'effets fixes, c'est-à-dire de variables prédictives connues sans erreur. La théorie se complique grandement quand un prédicteur devient lui-même une variable aléatoire.

5 Tomassone, R., Dervin, C. & Masson, J.P. (1993) Biométrie Modélisation de phénomènes biologiques. Masson, Paris. 1-553. Notion de confusion p. 266.

e) Lorsque la variabilité intrinsèque des résultats n'est pas normale (binomiale pour des fréquences, poissonienne pour des dénombrements, exponentielle pour des temps d'attente), on entre dans la sphère des *modèles linéaires généralisés*.

f) Si l'analyse statistique propose des outils, l'interprétation reste en dernier ressort dans le champ de l'expérience.