Les différents types de sommes de carrés dans le modèle linéaire et leur mise en œuvre avec SAS

Alain Baccini
Institut de Mathématiques de Toulouse
Novembre 2010

Le but de cette note est de préciser la signification des différents types de sommes de carrés (ainsi que les "philosophies" sous-jacentes) que l'on trouve dans la plupart des logiciels de statistique, en particulier SAS*, notamment dans le contexte de l'analyse de variance (ANOVA). Pour illustrer notre propos, nous nous placerons en ANOVA à deux facteurs croisés.*

Cette note constitue l'annexe B du cours **Le Modèle Linéaire Gaussien Général***, en libre accès sur ce même site.*

1 Introduction

Considérons un modèle d'analyse de variance à deux facteurs croisés dans lequel :

- le premier facteur, noté F_1, possède J niveaux ($J \geq 2$) qui seront indicés par j ;
- le second facteur, noté F_2, possède K niveaux ($K \geq 2$) qui seront indicés par k ;
- au croisement du niveau j de F_1 et du niveau k de F_2, on réalise n_{jk} observations ($n_{jk} \geq 1$) d'une v.a.r. Y (le plan d'expérience est donc complet, pas nécessairement équilibré) ;
- chaque observation est notée y_{ijk} ($i = 1, \ldots, n_{jk}$; $j = 1, \ldots, J$; $k = 1, \ldots, K$) ;
- on pose : $n_{j+} = \sum_{k=1}^{K} n_{jk}$ (effectif marginal du niveau j de F_1) ; $n_{+k} = \sum_{j=1}^{J} n_{jk}$ (effectif marginal du niveau k de F_2) ; $n = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk}$ (nombre total d'observations).

Introduisons les différentes moyennes partielles empiriques des observations y_{ijk} :

$$\begin{aligned}
\overline{y}_{\bullet jk} &= \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk} \ ; \\
\overline{y}_{\bullet j\bullet} &= \frac{1}{n_{j+}} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} y_{ijk} \ ; \\
\overline{y}_{\bullet\bullet k} &= \frac{1}{n_{+k}} \sum_{j=1}^{J} \sum_{i=1}^{n_{jk}} y_{ijk} \ ; \\
\overline{y}_{\bullet\bullet\bullet} &= \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} y_{ijk} \ .
\end{aligned}$$

Considérons maintenant la somme des carrés totale du modèle, quantité à $n-1$ degrés de liberté :

$$SST = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} (y_{ijk} - \overline{y}_{\bullet\bullet\bullet})^2.$$

Nous allons tout d'abord expliciter la décomposition de la quantité SST.

2 Décomposition de la somme totale des carrés

Remarquons tout d'abord l'égalité suivante :

$$y_{ijk} - \overline{y}_{\bullet\bullet\bullet} = (y_{ijk} - \overline{y}_{\bullet jk}) + (\overline{y}_{\bullet j\bullet} - \overline{y}_{\bullet\bullet\bullet}) + (\overline{y}_{\bullet\bullet k} - \overline{y}_{\bullet\bullet\bullet}) + (\overline{y}_{\bullet jk} - \overline{y}_{\bullet j\bullet} - \overline{y}_{\bullet\bullet k} + \overline{y}_{\bullet\bullet\bullet}).$$

En élevant au carré, il vient :

$$\begin{aligned}(y_{ijk}-\overline{y}_{\bullet\bullet\bullet})^2 &= (y_{ijk}-\overline{y}_{\bullet jk})^2+(\overline{y}_{\bullet j\bullet}-\overline{y}_{\bullet\bullet\bullet})^2+(\overline{y}_{\bullet\bullet k}-\overline{y}_{\bullet\bullet\bullet})^2\\ &+ (\overline{y}_{\bullet jk}-\overline{y}_{\bullet j\bullet}-\overline{y}_{\bullet\bullet k}+\overline{y}_{\bullet\bullet\bullet})^2+\sum_{\ell=1}^{6}DP_\ell(ijk),\end{aligned}$$

où les $DP_\ell(ijk)$ $(\ell=1,\ldots,6)$ représentent les doubles produits du développement de ce carré.

Par triple sommation,on obtient :

$$\begin{aligned}SST &= \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk}-\overline{y}_{\bullet jk})^2+\sum_{j=1}^{J}n_{j+}(\overline{y}_{\bullet j\bullet}-\overline{y}_{\bullet\bullet\bullet})^2+\sum_{k=1}^{K}n_{+k}(\overline{y}_{\bullet\bullet k}-\overline{y}_{\bullet\bullet\bullet})^2\\ &+ \sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\overline{y}_{\bullet jk}-\overline{y}_{\bullet j\bullet}-\overline{y}_{\bullet\bullet k}+\overline{y}_{\bullet\bullet\bullet})^2+\sum_{\ell=1}^{6}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}DP_\ell(ijk).\end{aligned}$$

Pour détailler les doubles produits, remarquons tout d'abord que si les quantités x_{jk} sont des réels indépendants de i, on peut écrire :

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}x_{jk}(y_{ijk}-\overline{y}_{\bullet jk})=\sum_{j=1}^{J}\sum_{k=1}^{K}x_{jk}[\sum_{i=1}^{n_{jk}}(y_{ijk}-\overline{y}_{\bullet jk})]=0.$$

Il s'ensuit que les sommes des trois premiers doubles produits (ceux dans lesquels la quantité $y_{ijk}-\overline{y}_{\bullet jk}$ est en facteur) sont nulles. Par contre, les trois autres ne sont, en général, pas nulles. La quatrième s'écrit :

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}DP_4(ijk)=2\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\overline{y}_{\bullet j\bullet}-\overline{y}_{\bullet\bullet\bullet})(\overline{y}_{\bullet\bullet k}-\overline{y}_{\bullet\bullet\bullet})=SDP_1.$$

Les sommes des deux derniers doubles produits peuvent être regroupées dans l'expression suivante :

$$\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}[DP_5(ijk)+DP_6(ijk)]=SDP_2$$

$$=2\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}(\overline{y}_{\bullet j\bullet}+\overline{y}_{\bullet\bullet k}-2\overline{y}_{\bullet\bullet\bullet})(\overline{y}_{\bullet jk}-\overline{y}_{\bullet j\bullet}-\overline{y}_{\bullet\bullet k}+\overline{y}_{\bullet\bullet\bullet}).$$

Enfin, un développement des parenthèses figurant dans les expressions SDP_1 et SDP_2 ci-dessus définies permet d'obtenir (les calculs sont simples, mais assez fastidieux) :

$$SDP_1=2(\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}\overline{y}_{\bullet j\bullet}\overline{y}_{\bullet\bullet k}-n\overline{y}^2_{\bullet\bullet\bullet});$$

$$SDP_2=4(n\overline{y}^2_{\bullet\bullet\bullet}-\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}\overline{y}_{\bullet j\bullet}\overline{y}_{\bullet\bullet k}).$$

Finalement, la somme de tous les doubles produits figurant dans SST s'écrit :

$$SDP=2(n\overline{y}^2_{\bullet\bullet\bullet}-\sum_{j=1}^{J}\sum_{k=1}^{K}n_{jk}\overline{y}_{\bullet j\bullet}\overline{y}_{\bullet\bullet k}).$$

Explicitons maintenant les sommes de carrés en introduisant les quantités suivantes :

– somme des carrés due au facteur F_1 (quantité à $J-1$ degrés de liberté) :

$$SSF_1=\sum_{j=1}^{J}n_{j+}(\overline{y}_{\bullet j\bullet}-\overline{y}_{\bullet\bullet\bullet})^2\ ;$$

– somme des carrés due au facteur F_2 (quantité à $K-1$ degrés de liberté) :

$$SSF_2 = \sum_{k=1}^{K} n_{+k}(\overline{y}_{\bullet\bullet k} - \overline{y}_{\bullet\bullet\bullet})^2 \ ;$$

– somme des carrés due aux interactions (quantité à $(J-1)(K-1)$ degrés de liberté) :

$$SSF_{1*2} = \sum_{j=1}^{J}\sum_{k=1}^{K} n_{jk}(\overline{y}_{\bullet jk} - \overline{y}_{\bullet j\bullet} - \overline{y}_{\bullet\bullet k} + \overline{y}_{\bullet\bullet\bullet})^2 \ ;$$

– somme des carrés due aux erreurs (ou résiduelle ; quantité à $n-JK$ degrés de liberté) :

$$SSE = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk} - \overline{y}_{\bullet jk})^2 \ .$$

On peut finalement réécrire la somme des carrés totale sous la forme :

$$SST = SSF_1 + SSF_2 + SSF_{1*2} + SSE + SDP.$$

Remarque 1 *Dans le cas particulier d'un plan équilibré ($n_{jk} = n_0 \ , \forall (j,k)$), on vérifie sans difficulté que $SDP = 0$. La décomposition de SST est alors d'interprétation évidente. Par contre, ce n'est pas le cas avec les plans déséquilibrés pour lesquels la quantité SDP est en général non nulle.*

Lorsque la quantité SDP est non nulle, il n'est pas possible de l'affecter à une unique source de variation (F_1, F_2 ou $F_1 * F_2$). Ceci explique les difficultés rencontrées pour spécifier les sources de variation dans un modèle relatif à un plan déséquilibré. Pour cette raison, on a recours à d'autres raisonnements pour spécifier ces sources, ce qui explique l'existence de plusieurs types de sommes de carrés, selon la philosophie choisie.

3 Exemple

On considère le jeu de données ci-dessous, dans lequel la variable réponse quantitative Y est expliquée par deux facteurs croisés, F_1 à deux niveaux et F_2 à trois niveaux. Il y a au total 18 observations dans un plan complet déséquilibré (il s'agit d'un exemple fictif, dans lequel les observations de Y ont été choisies pour faciliter les calculs "à la main", de façon à permettre un certain contrôle des résultats fournis par le logiciel `SAS`).

Outre les valeurs initiales des y_{ijk}, le tableau ci-dessous donne toutes les sommes et moyennes partielles (dans chaque cellule, chaque ligne, chaque colonne), ainsi que la somme et la moyenne globales.

	niveau 1 de F_2	niveau 2 de F_2	niveau 3 de F_2	sommes	moyennes
niveau 1 de F_1	10 14 18	36 40 44 48	82 86		
sommes	**42**	**168**	**168**	**378**	
moyennes	*14*	*42*	*84*		*42*
niveau 2 de F_1	22 26	24 28 32	60 64 68 72		
sommes	**48**	**84**	**264**	**396**	
moyennes	*24*	*28*	*66*		*44*
sommes	**90**	**252**	**432**	**774**	
moyennes	*18*	*36*	*72*		*43*

À partir du tableau ci-dessus, on peut calculer facilement les expressions suivantes :

$$SSF_1 = 18 \ ; \quad SSF_2 = 8514 \ ; \quad SSF_{1*2} = 1050 \ ; \quad SSE = 240 \ ; \quad SDP = -180.$$

On en déduit : $SST = 9642$. Dans le modèle "complet" (également appelé modèle "plein" et comportant un effet général, les effets de F_1, les effets de F_2 et les effets d'interactions), la somme des carrés relative au modèle vaudra donc : $9642 - 240 = 9402$.

4 Traitement des données avec SAS

4.1 Traitement initial

Nous avons utilisé la procédure `GLM` du logiciel `SAS` pour traiter ces données selon un modèle d'analyse de variance à deux facteurs croisés, avec interactions. À la suite de la commande `model`, nous avons rajouté les options `ss1, ss2, ss3` et `ss4` pour obtenir les sommes de carrés de type I, de type II, de type III et de type IV (ces dernières uniquement dans le premier traitement, puisqu'elles ne se distinguent des précédentes que dans certains plans incomplets).

En supposant les données contenues dans le fichier `deseq.don`, le programme SAS est le suivant :

```
data deseq;
infile 'deseq.don';
input f1 f2 y;
run;
proc glm data=deseq;
class f1 f2;
model y = f1 f2 f1*f2 / ss1 ss2 ss3 ss4;
run;
quit;
```

En voici les résultats.

```
                          The GLM Procedure

Dependent Variable: y

                                   Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F

Model                       5    9402.000000    1880.400000     94.02   <.0001

Error                      12     240.000000      20.000000

Corrected Total            17    9642.000000


              R-Square     Coeff Var      Root MSE        y Mean

              0.975109      10.40032      4.472136      43.00000


Source                     DF      Type I SS    Mean Square   F Value   Pr > F

f1                          1      18.000000      18.000000      0.90   0.3615
f2                          2    8801.112108    4400.556054    220.03   <.0001
f1*f2                       2     582.887892     291.443946     14.57   0.0006


Source                     DF     Type II SS    Mean Square   F Value   Pr > F

f1                          1     305.112108     305.112108     15.26   0.0021
f2                          2    8801.112108    4400.556054    220.03   <.0001
f1*f2                       2     582.887892     291.443946     14.57   0.0006


Source                     DF    Type III SS    Mean Square   F Value   Pr > F

f1                          1     223.384615     223.384615     11.17   0.0059
f2                          2    8664.968610    4332.484305    216.62   <.0001
f1*f2                       2     582.887892     291.443946     14.57   0.0006
```

```
Source                     DF     Type IV SS    Mean Square   F Value   Pr > F

f1                          1     223.384615     223.384615     11.17   0.0059
f2                          2    8664.968610    4332.484305    216.62   <.0001
f1*f2                       2     582.887892     291.443946     14.57   0.0006
```

On constate tout d'abord que le modèle est très significatif et que le coefficient R^2 est très proche de 1 : on a donc un très bon ajustement du modèle aux données.

Ensuite, on voit que les sommes de carrés de type IV sont identiques à celles de type III : c'est normal, puisque les sommes de type IV ne sont différentes des sommes de type III que dans le cas de certains plans incomplets. Nous ne les ferons donc plus figurer dans les traitements qui vont suivre, mais on pourra trouver des précisions sur les sommes de type IV dans Milliken & Johnson (1984) ou dans Azaïs (1994).

Nous allons maintenant détailler la façon d'obtenir les autres sommes de carrés figurant ci-dessus.

Remarque 2 *Insistons encore ici sur le fait que, dans un plan complet équilibré, les quatre types de sommes sont toujours identiques.*

4.2 Somme des carrés relative aux interactions

Le principe de calcul est très simple et assez naturel : on fait la différence entre la somme des carrés relative aux erreurs dans le modèle additif (sans interactions) et celle, également relative aux erreurs, dans le modèle complet (avec interactions). Le résultat obtenu représente donc la somme des carrés relative aux interactions. On constate que c'est la même, quel que soit le type de somme (ici : 582.89).

Contrôlons ce résultat en mettant en œuvre le modèle additif.

```
proc glm data=deseq;
class f1 f2;
model y = f1 f2 / ss1 ss2 ss3;
run;
quit;
```

On obtient les résultats suivants.

```
Dependent Variable: y

Model: y = f1 f2
                                  Sum of
Source                     DF        Squares    Mean Square   F Value   Pr > F

Model                       3    8819.112108    2939.704036     50.01   <.0001

Error                      14     822.887892      58.777707

Corrected Total            17    9642.000000


             R-Square     Coeff Var      Root MSE        y Mean

             0.914656      17.82945      7.666662      43.00000
```

On voit qu'en faisant la différence entre 822.89 et 240, on retrouve bien la quantité 582.89.

4.3 Somme des carrés relative au facteur F_2

Le principe général reste le même : on fait la différence entre la somme des carrés relative aux erreurs dans un modèle où on a enlevé F_2 et un modèle de référence, dans lequel il figure. Le problème est que le modèle de référence varie : c'est le modèle complet (effets de F_1, de F_2 et des

interactions) dans le cas des sommes de type III et c'est le modèle additif (effets de F_1 et de F_2 seulement) dans le cas des sommes de type I et de type II.

On a donc besoin des résultats de deux modèles : le modèle avec F_1 et les interactions, et le modèle avec seulement F_1. Mettons ces deux modèles en œuvre.

```
proc glm data=deseq;
class f1 f2;
model y = f1 f1*f2 / ss1 ss2 ss3;
run;
proc glm data=deseq;
class f1 f2;
model y = f1 / ss1 ss2 ss3;
run;
quit;
```

Étudions en les résultats.

```
                          The GLM Procedure

Dependent Variable: y

Model: y = f1 f1*f2
                                   Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F

Model                       5    9402.000000    1880.400000     94.02   <.0001

Error                      12     240.000000      20.000000

Corrected Total            17    9642.000000


              R-Square     Coeff Var      Root MSE        y Mean

              0.975109      10.40032      4.472136      43.00000


Source                     DF      Type I SS    Mean Square   F Value   Pr > F

f1                          1      18.000000      18.000000      0.90   0.3615
f1*f2                       4    9384.000000    2346.000000    117.30   <.0001


Source                     DF     Type II SS    Mean Square   F Value   Pr > F

f1                          1      18.000000      18.000000      0.90   0.3615
f1*f2                       4    9384.000000    2346.000000    117.30   <.0001


Source                     DF    Type III SS    Mean Square   F Value   Pr > F

f1                          1     223.384615     223.384615     11.17   0.0059
f1*f2                       4    9384.000000    2346.000000    117.30   <.0001


Dependent Variable: y

Model: y = f1
                                   Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F

Model                       1      18.000000      18.000000      0.03   0.8648
```

```
Error                      16     9624.000000      601.500000

Corrected Total            17     9642.000000
```

Pour les sommes de type I et de type II, on obtient : $9624 - 822.89 = 8801.11$: on retrouve bien le résultat de la première sortie.

Par contre, on ne peut pas retrouver ici la somme des carrés de type III pour F_2, puisque la somme des carrés relative aux erreurs dans le modèle avec F_1 et les interactions reste la même que dans le modèle complet : 240. C'est **un des paradoxes de SAS**, qui permet de déclarer un modèle avec un seul facteur et les interactions, mais ne le traite pas comme tel, puisqu'il rajoute l'effet du facteur enlevé, F_2, dans les effets d'interactions : ceux-ci se trouvent ainsi avoir 4 degrés de liberté (2 pour F_2 et 2 pour les interactions) et une somme de carrés égale à 9384, que l'on obtient en additionnant 8801.11 et 582.89, sommes relatives respectivement à F_2 et aux interactions dans le modèle complet, si l'on considère les sommes de type I ou II. Il faudra donc avoir recours à un artifice pour retrouver la somme de type III relative à F_2 (voir le point 4.5).

4.4 Somme des carrés relative au facteur F_1

Tout d'abord, remarquons que ces sommes sont toutes différentes, selon le type I, II ou III considéré. Cela provient de ce que les philosophies sont ici toutes trois différentes.

Type III

Le type III conserve la même philosophie : le modèle de référence étant le modèle complet, on calcule la différence entre la somme des carrés relatives aux erreurs dans le modèle avec les effets de F_2 et ceux des interactions et la même somme dans le modèle complet. On se heurte encore à la même difficulté : il n'est pas possible d'obtenir directement la première somme des carrés. Nous aurons donc recours, là encore, au même artifice que précédemment (voir encore le point 4.5).

Type II

Pour les sommes de type II, on doit faire la différence entre la somme des carrés relative aux erreurs dans le modèle avec les seuls effets de F_2 et la même somme dans le modèle additif. Mettons en œuvre le modèle avec le seul facteur F_2.

```
proc glm data=deseq;
class f1 f2;
model y = f2 / ss1 ss2 ss3;
run;
```

En voici les résultats.

```
                          The GLM Procedure

Dependent Variable: y

Model: y = f2

                                  Sum of
Source                     DF        Squares    Mean Square   F Value   Pr > F

Model                       2     8514.000000    4257.000000     56.61   <.0001

Error                      15     1128.000000      75.200000

Corrected Total            17     9642.000000
```

On obtient $1128 - 822.89 = 305.11$, qui est, dans le modèle complet, la somme des carrés de type II relative à F_1.

Type I

Pour le type I, il convient de préciser la philosophie globale de cette approche. Elle suppose en effet que les effets déclarés dans la commande `model y = f1 f2 f1*f2` sont ordonnés. Autrement dit, les premiers effets à prendre en compte sont ceux de F_1, puis ceux de F_2, enfin ceux des interactions. Par conséquent, pour calculer la somme des carrés relative à l'un de ces trois effets, on considère la différence des sommes des carrés relatives aux erreurs dans deux modèles : "le plus grand modèle" ne contenant pas cet effet (ici, le modèle constant) et "le plus petit modèle" le contenant (ici, le modèle avec seulement F_1). Dans le modèle constant, la somme des carrés relative aux erreurs est la somme des carrés totale à savoir 9642. Le modèle ne comportant que F_1 a déjà été étudié en 4.3 (somme des carrés relative aux erreurs dans ce modèle : 9624). D'où la somme des carrés relative à F_1 pour le type I : $9642 - 9624 = 18$. On peut remarquer qu'il s'agit en fait de la somme des carrés relative à F_1 telle que nous l'avons définie au paragraphe 2 (SSF_1) et dont la valeur a été donnée au paragraphe 3.

Autre illustration de la philosophie de type I

De façon à bien comprendre la philosophie des sommes de type I, déclarons maintenant le modèle complet en commençant par F_2 :

```
proc glm data=deseq;
class f1 f2;
model y = f2 f1 f1*f2 / ss1 ss2 ss3;
run;
```

En voici les résultats.

```
                          The GLM Procedure

Dependent Variable: y

                                   Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F

Model                       5    9402.000000    1880.400000     94.02   <.0001

Error                      12     240.000000      20.000000

Corrected Total            17    9642.000000


              R-Square     Coeff Var      Root MSE        y Mean

              0.975109      10.40032      4.472136      43.00000


Source                     DF      Type I SS    Mean Square   F Value   Pr > F

f2                          2    8514.000000    4257.000000    212.85   <.0001
f1                          1     305.112108     305.112108     15.26   0.0021
f1*f2                       2     582.887892     291.443946     14.57   0.0006


Source                     DF     Type II SS    Mean Square   F Value   Pr > F

f2                          2    8801.112108    4400.556054    220.03   <.0001
f1                          1     305.112108     305.112108     15.26   0.0021
f1*f2                       2     582.887892     291.443946     14.57   0.0006


Source                     DF    Type III SS    Mean Square   F Value   Pr > F
```

```
f2                          2    8664.968610    4332.484305    216.62   <.0001
f1                          1     223.384615     223.384615     11.17   0.0059
f1*f2                       2     582.887892     291.443946     14.57   0.0006
```

Mise à part l'inversion de l'ordre des facteurs F_1 et F_2, les sommes de type II et de type III sont identiques à ce qu'elles étaient dans le modèle initial. Par contre, il n'en va pas de même pour les sommes de type I : seule celle relative aux interactions est inchangée ; celle relative à F_1 a augmenté (elle est passée de 18 à 305.11), tandis que celle relative à F_2 a baissé (elle est passée de 8801.11 à 8514). Remarquons en passant que la somme relative à F_2 (8514) est maintenant égale à celle définie au paragraphe 2 (SSF_2) et donnée au paragraphe 3.

Remarque 3 *En additionnant les sommes des carrés relatives aux différents effets dans le type I, on retrouve, quel que soit l'ordre de déclaration des facteurs, la somme des carrés relative au modèle complet, à savoir* 9402. *Ceci est une propriété générale, et seules les sommes de type I possèdent cette propriété, comme on peut le constater en faisant les additions : pour le type II, on trouve* 9689.11 *; pour le type III,* 9471.24.

4.5 Retour sur les sommes de type III

Revenons maintenant aux sommes de type III et essayons de retrouver les sommes de carrés relatives à chacun des deux facteurs F_1 et F_2. Pour cela, il faut passer par l'intermédiaire d'un modèle de régression sur indicatrices, mais pas n'importe quelles indicatrices !

Introduction d'indicatrices

Nous allons introduire les indicatrices des niveaux de F_1, celles des niveaux de F_2, et celles des cellules obtenues par croisement de F_1 et de F_2, comme cela se fait dans le cadre du *paramétrage* dit *centré* de l'ANOVA (paramétrage associé à un effet moyen général, des effets principaux pour chaque niveau de chaque facteur, leur somme étant nulle, et des effets d'interactions doublement centrés).

Pour F_1, c'est en fait la différence entre l'indicatrice du niveau 1 et celle du niveau 2 qui doit intervenir (une seule variable car un seul degré de liberté) ; nous la noterons L_1. Pour F_2, on doit utiliser deux variables (deux degrés de liberté) : la différence entre l'indicatrice du niveau 1 et celle du niveau 3 et la différence entre l'indicatrice du niveau 2 et celle du niveau 3 ; nous les noterons respectivement C_1 et C_2. Enfin, pour les cellules, c'est le produit des précédentes qui seront utilisés (il n'y en a que deux) : $LC_1 = L_1 \times C_1$; $LC_2 = L_1 \times C_2$.

Voici un programme SAS permettant de créer ces différences d'indicatrices :

```
data indi1;
set deseq;
L1 = 0;
if f1 = 1 then L1 =  1;
if f1 = 2 then L1 = -1;
C1 = 0;
if f2 = 1 then C1 =  1;
if f2 = 3 then C1 = -1;
C2 = 0;
if f2 = 2 then C2 =  1;
if f2 = 3 then C2 = -1;
LC1 = L1 * C1;
LC2 = L1 * C2;
run;
```

En voici les résultats :

```
Obs    f1    f2     y    L1    C1    C2    LC1    LC2

 1      1     1    10     1     1     0      1      0
 2      1     1    14     1     1     0      1      0
```

3	1	1	18	1	1	0	1	0
4	1	2	36	1	0	1	0	1
5	1	2	40	1	0	1	0	1
6	1	2	44	1	0	1	0	1
7	1	2	48	1	0	1	0	1
8	1	3	82	1	-1	-1	-1	-1
9	1	3	86	1	-1	-1	-1	-1
10	2	1	22	-1	1	0	-1	0
11	2	1	26	-1	1	0	-1	0
12	2	2	24	-1	0	1	0	-1
13	2	2	28	-1	0	1	0	-1
14	2	2	32	-1	0	1	0	-1
15	2	3	60	-1	-1	-1	1	1
16	2	3	64	-1	-1	-1	1	1
17	2	3	68	-1	-1	-1	1	1
18	2	3	72	-1	-1	-1	1	1

Régression sur les indicatrices

Faisons maintenant la régression de Y sur l'ensemble de ces indicatrices :

```
proc glm data=indi1;
model y = L1 C1 C2 LC1 LC2 / ss1 ss2 ss3;
run;
```

Les résultats sont les suivants :

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9402.000000	1880.400000	94.02	<.0001
Error	12	240.000000	20.000000		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.975109	10.40032	4.472136	43.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
L1	1	18.000000	18.000000	0.90	0.3615
C1	1	8388.765957	8388.765957	419.44	<.0001
C2	1	412.346150	412.346150	20.62	0.0007
LC1	1	480.760233	480.760233	24.04	0.0004
LC2	1	102.127660	102.127660	5.11	0.0432

Source	DF	Type II SS	Mean Square	F Value	Pr > F
L1	1	223.384615	223.384615	11.17	0.0059
C1	1	4443.428571	4443.428571	222.17	<.0001
C2	1	588.255319	588.255319	29.41	0.0002
LC1	1	579.428571	579.428571	28.97	0.0002
LC2	1	102.127660	102.127660	5.11	0.0432

Source	DF	Type III SS	Mean Square	F Value	Pr > F
L1	1	223.384615	223.384615	11.17	0.0059
C1	1	4443.428571	4443.428571	222.17	<.0001
C2	1	588.255319	588.255319	29.41	0.0002
LC1	1	579.428571	579.428571	28.97	0.0002
LC2	1	102.127660	102.127660	5.11	0.0432

On remarquera tout d'abord que l'on obtient exactement les mêmes résultats généraux qu'avec le modèle complet d'ANOVA (sommes des carrés relatives au modèle et aux erreurs, degrés de liberté, coefficient R^2...), ce qui est logique.

On retrouve également des résultats identiques pour les sommes de carrés de type I : 18 pour L_1 (donc pour F_1) ; $8388.77+412.35 = 8801.12$ pour C_1+C_2 (donc pour F_2) ; $480.76+102.13 = 582.89$ pour $LC_1 + LC_2$ (donc pour les interactions).

Par contre, il n'en va pas de même pour les sommes de type II et de type III. Tout d'abord, on remarque qu'elles sont maintenant identiques, ce qui s'explique par le fait que le seul modèle par rapport auquel on peut, dans ce cadre, se référer est le modèle complet (la notion de modèle additif n'a plus de sens dans le cadre d'une régression). Ensuite, ces sommes s'obtiennent toujours en faisant la différence des sommes de carrés relatives aux erreurs au sein de deux modèles : le modèle dans lequel on enlève seulement l'effet considéré (L_1, C_1...) et le modèle complet considéré ci-dessus. Nous laissons le soin au lecteur de vérifier ce résultat.

Somme des carrés de type III relative au facteur F_2

Refaisons maintenant la régression de Y sur les seules indicatrices L_1, LC_1 et LC_2, autrement dit sur le facteur F_1 et sur les interactions.

```
proc glm data=indi1;
model y = L1 LC1 LC2 / ss1 ss2 ss3;
run;
```

En voici les résultats :

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	737.031390	245.677130	0.39	0.7646
Error	14	8904.968610	636.069186		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.076440	58.65212	25.22041	43.00000

En faisant la différence $8904.97 - 240 = 8664.97$, on retrouve maintenant la somme des carrés de type III relative à F_2 dans le modèle complet d'ANOVA.

Somme des carrés de type III relative au facteur F_1

Dans la même optique, faisons maintenant la régression de Y sur les indicatrices C_1, C_2, LC_1 et LC_2 (autrement dit, sur F_2 et sur les interactions).

```
proc glm data=indi1;
model y = C1 C2 LC1 LC2 / ss1 ss2 ss3;
run;
```

On obtient :

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	9178.615385	2294.653846	64.38	<.0001
Error	13	463.384615	35.644970		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.951941	13.88451	5.970341	43.00000

La différence $463.38 - 240 = 223.38$ redonne la somme des carrés de type III relative à F_1.

Encore un paradoxe de SAS !

Les indicatrices utilisées ci-dessus, ou plutôt les différences d'indicatrices (indicatrice d'un niveau moins indicatrice du dernier niveau), apparaissent naturellement dans le paramétrage centré du modèle d'ANOVA.

On peut se demander ce qu'il se passe si on remplace ces indicatrices par celles qui apparaissent naturellement dans le *paramétrage* du modèle d'ANOVA *réalisé par SAS*. Il s'agit simplement des indicatrice des niveaux des facteurs, à l'exception du dernier, et des produits de ces indicatrices pour les interactions.

Voici encore un programme permettant d'obtenir ces indicatrices :

```
data indi2;
set deseq;
LS1 = 0;
if f1 = 1 then LS1 = 1;
CS1 = 0;
if f2 = 1 then CS1 = 1;
CS2 = 0;
if f2 = 2 then CS2 = 1;
LCS1 = LS1 * CS1;
LCS2 = LS1 * CS2;
run;
```

Et voici la table SAS obtenue :

Obs	f1	f2	y	LS1	CS1	CS2	LCS1	LCS2
1	1	1	10	1	1	0	1	0
2	1	1	14	1	1	0	1	0
3	1	1	18	1	1	0	1	0
4	1	2	36	1	0	1	0	1
5	1	2	40	1	0	1	0	1
6	1	2	44	1	0	1	0	1
7	1	2	48	1	0	1	0	1
8	1	3	82	1	0	0	0	0
9	1	3	86	1	0	0	0	0
10	2	1	22	0	1	0	0	0
11	2	1	26	0	1	0	0	0
12	2	2	24	0	0	1	0	0
13	2	2	28	0	0	1	0	0
14	2	2	32	0	0	1	0	0
15	2	3	60	0	0	0	0	0
16	2	3	64	0	0	0	0	0
17	2	3	68	0	0	0	0	0
18	2	3	72	0	0	0	0	0

Faisons maintenant la régression de Y sur les cinq indicatrices ci-dessus.

```
proc glm data=indi2;
model y = LS1 CS1 CS2 LCS1 LCS2 / ss1 ss2 ss3;
run;
```

En voici les résultats :

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9402.000000	1880.400000	94.02	<.0001
Error	12	240.000000	20.000000		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.975109	10.40032	4.472136	43.00000

Source	DF	Type I SS	Mean Square	F Value	Pr > F
LS1	1	18.000000	18.000000	0.90	0.3615
CS1	1	4324.500000	4324.500000	216.22	<.0001
CS2	1	4476.612108	4476.612108	223.83	<.0001
LCS1	1	570.887892	570.887892	28.54	0.0002
LCS2	1	12.000000	12.000000	0.60	0.4536

Source	DF	Type II SS	Mean Square	F Value	Pr > F
LS1	1	432.000000	432.000000	21.60	0.0006
CS1	1	2352.000000	2352.000000	117.60	<.0001
CS2	1	2475.428571	2475.428571	123.77	<.0001
LCS1	1	495.157895	495.157895	24.76	0.0003
LCS2	1	12.000000	12.000000	0.60	0.4536

Source	DF	Type III SS	Mean Square	F Value	Pr > F
LS1	1	432.000000	432.000000	21.60	0.0006
CS1	1	2352.000000	2352.000000	117.60	<.0001
CS2	1	2475.428571	2475.428571	123.77	<.0001
LCS1	1	495.157895	495.157895	24.76	0.0003
LCS2	1	12.000000	12.000000	0.60	0.4536

Encore une fois, les résultats généraux n'ont pas changé. De même, les sommes de carrés de type I permettent de retrouver celles du modèle complet en ANOVA : 18 pour LS_1, c'est-à-dire pour F_1 ; $4324.50 + 4476.61 = 8801.11$ pour $CS_1 + CS_2$, c'est-à-dire pour F_2 ; $570.89 + 12.00 = 582.89$ pour $LCS_1 + LCS_2$, c'est-à-dire pour les interactions.

Par contre, les sommes de carrés de type II et de type III, encore une fois égales, ne redonnent pas les résultats de l'ANOVA avec le modèle complet et sont différentes de ce qu'elles étaient avec la première série d'indicatrices.

Intéressons nous maintenant au modèle de régression de Y sur les indicatrices LS_1, LCS_1 et LCS_2.

```
proc glm data=indi2;
model y = LS1 LCS1 LCS2 / ss1 ss2 ss3;
run;
```

Voici les résultats :

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	5898.000000	1966.000000	7.35	0.0034
Error	14	3744.000000	267.428571		
Corrected Total	17	9642.000000			

R-Square	Coeff Var	Root MSE	y Mean
0.611699	38.03080	16.35324	43.00000

On pourra vérifier que la somme des carrés relative aux erreurs dans ce modèle (3744) ne permet pas de retrouver la somme des carrés de type III relative au facteur F_2 dans le modèle complet d'ANOVA. Ainsi, l'usage d'indicatrices permet de retrouver ces sommes, mais uniquement les indicatrices utilisées dans le paramétrage centré, celles utilisées dans le paramétrage SAS ne le permettant pas.

4.6 Cas particulier du modèle additif

Revenons sur le modèle additif, déjà traité en 4.2. Voici les sommes de carrés que l'on obtient dans ce modèle :

Source	DF	Type I SS	Mean Square	F Value	Pr > F
f1	1	18.000000	18.000000	0.31	0.5887
f2	2	8801.112108	4400.556054	74.87	<.0001

Source	DF	Type II SS	Mean Square	F Value	Pr > F
f1	1	305.112108	305.112108	5.19	0.0389
f2	2	8801.112108	4400.556054	74.87	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f1	1	305.112108	305.112108	5.19	0.0389
f2	2	8801.112108	4400.556054	74.87	<.0001

On remarque que les sommes de carrés de type II et de type III sont égales dans ce cas. Ce résultat est général et s'explique par le fait que le modèle de référence est ici le modèle additif, que ce soit pour le type II ou pour le type III.

5 Quelle philosophie suivre ?

À l'issue de cette étude, on peut légitimement se poser la question : quelles sommes de carrés utiliser ? On se doute que la réponse n'est pas univoque et qu'elle est liée à la fois au type de données dont on dispose et à la philosophie que l'on souhaite suivre.

Tout d'abord, nous laisserons de côté les sommes de type IV qui ne concernent que les plans incomplets et déséquilibrés que nous n'avons pas envisagés ici (encore faut-il signaler que le type IV est parfois critiqué dans le contexte des plans incomplets déséquilibrés).

Ensuite, les sommes de type I sont spécifiques des modèles dans lesquels il existe un ordre naturel entre les facteurs. Pour des données de ce type, ce sont bien sûr ces sommes qu'il faut

considérer. Dans les autres cas, il n'est pas courant de les utiliser (même si elles ont de bonnes propriétés, comme on l'a signalé).

Remarque 4 *Il convient de distinguer ce qu'on a appelé ici "ordre" entre les facteurs (on considère que l'un est plus important que l'autre, les interactions ayant nécessairement un moindre niveau d'importance) et ce qu'on appelle habituellement facteur* hiérarchisé *(la définition des niveaux du facteur hiérarchisé dépend du niveau de l'autre facteur dans lequel on se trouve ; de tels facteurs sont aussi ordonnés, mais de façon plus "structurelle"), situation plus particulière. Dans la procédure* `GLM` *de SAS, il est possible de faire un traitement spécifique pour des facteurs dont l'un est hiérarchisé à l'autre. C'est d'ailleurs dans ce contexte que les sommes de type I prennent tout leur sens.*

Reste donc le choix entre les sommes de type II et de type III pour les cas standards, mais déséquilibrés. Il est à noter que ce choix ne se pose que dans le cadre des modèles avec interactions, les deux types étant équivalents pour les modèles additifs. D'une façon générale, il est préconisé d'utiliser les sommes de type III de préférence à celles de type II. En particulier, on remarquera que SAS ne fournit par défaut que les sommes de type I et de type III.

Terminons cette discussion par la remarque ci-dessous dans laquelle on va préciser un peu plus les choses.

Remarque 5 *La discussion sur le choix des sommes de carrés à utiliser dans la pratique est l'occasion de revenir sur la pratique des tests relatifs aux différents effets dans un modèle complexe comme une ANOVA à au moins deux facteurs. Considérons encore, pour simplifier, une ANOVA à deux facteurs croisés.*

On peut préconiser la démarche consistant à tester en premier lieu les interactions, puis à passer au modèle additif si elles ne sont pas significatives. Cette démarche, assez naturelle, n'est pas la seule utilisée dans la pratique statistique. De plus, elle a le défaut suivant : elle conduit, lors des tests des effets principaux de chacun des deux facteurs, à prendre en compte dans le numérateur de l'estimateur de la variance (donc dans le dénominateur de la statistique de Fisher) les sommes de carrés, certes faibles mais non nulles, relatives aux interactions. Cela peut conduire à un biais dans la statistique de Fisher, donc dans la décision relative aux effets principaux.

D'où une autre démarche, tout aussi courante, qui consiste à tester chaque facteur au sein du modèle complet (avec interactions) et qu'on appelle souvent "non pooling"*, autrement dit non regroupement (des sommes de carrés des différents effets dans le numérateur de l'estimateur de la variance). Dans ce contexte, les sommes de type II ne sont pas justifiées (en fait, il n'y a pas de contexte dans lequel elles soient réellement justifiées).*

Dans la pratique, on peut envisager de mener en parallèle les deux démarches ci-dessus. Lorsqu'elles conduisent à la même décision, il n'y a pas de problème. En cas de décisions contradictoires, il convient d'être très prudent et d'étudier en détails les deux modèles en présence, en particulier en utilisant un critère de choix de modèle.

En guise de conclusion générale, **nous préconisons d'utiliser systématiquement les sommes de type III**. S'il existe un ordre entre les facteurs, notamment dans le cas de facteurs hiérarchisés, on devra aussi considérer les sommes de type I, voire les privilégier en cas de contradiction dans les décisions issues des tests. Si on est en présence d'un plan incomplet et déséquilibré, on devra considérer, en plus des sommes de type III, les sommes de type IV. Enfin, les sommes de type II sont déconseillées dans tous les cas.

Références

• J.-M. Azaïs, "Analyse de variance non orthogonale ; l'exemple de SAS/GLM", Revue de Statistique Apliquée, 42 (2), 27-41, 1994.

• G.A. Milliken & D.E. Johnson, "Analysis of messy data", Volume I : designed experiments, Van Nostrand Reinhold, 1984.