

REVUE DE STATISTIQUE APPLIQUÉE

J.-M. AZAÏS

**Analyse de variance non orthogonale.
L'exemple de SAS/GLM**

Revue de statistique appliquée, tome 42, n° 2 (1994), p. 27-41

<http://www.numdam.org/item?id=RSA_1994__42_2_27_0>

© Société française de statistique, 1994, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

ANALYSE DE VARIANCE NON ORTHOGONALE L'EXEMPLE DE SAS/GLM

J.-M. Azaïs

*Laboratoire de Statistique et Probabilités, URA CNRS 745
Université Paul Sabatier, 118, route de Narbonne, F31062 Toulouse Cedex.*

RÉSUMÉ

En utilisant le vocabulaire de la procédure GLM de SAS pour dénommer les différentes sommes de carrés, nous présentons les différentes décompositions possibles en analyse de variance non orthogonale. Nous explicitons les hypothèses linéaires sous-jacentes et nous sommes donc en mesure de préciser dans quelle situation chaque décomposition doit être employée. Le cas le plus important est celui du modèle à deux facteurs croisés avec interaction. Le modèle additif ainsi que les modèles hiérarchiques sont également traités.

Mots-clés : *Analyse de variance, tests d'hypothèses, fonctions estimables, logiciels statistiques.*

SUMMARY

Using the vocabulary of SAS GLM procedure, we present the possible decompositions in unbalanced multiway analysis of variance. Outlining the corresponding hypotheses, permits to explain when each one has to be used. The most important case, is the two way model with interaction, but the additive model and nested models are also considered.

Key-words : *Unbalanced analysis of variance, tests of hypotheses, estimable functions, statistical software.*

1. Introduction

Y a t'il encore matière à faire un article sur l'analyse de variance? Un lecteur non averti pourrait penser qu'il s'agit d'un sujet dont la théorie (le modèle linéaire statistique) est bien connue et dont la pratique est devenue classique. Cette opinion se révélerait fausse sur les deux points. D'abord on peut citer des articles récents consacrés aux aspects théoriques et épistémologiques de l'analyse de variance (Speed 1987, Tuckey 1991). Ensuite, comme nous allons le voir dans cet article, la mise en application pratique et la compréhension profonde des algorithmes des grands logiciels statistiques amènent à certaines réflexions fondamentales.

Un des points les plus débattus concerne la définition et le test d'effets principaux dans un modèle avec interaction, (Searle 1987) une fois que le test de cette même interaction a été déclaré non significatif. A-t'on le droit de se poser une telle

question et quelle est l'hypothèse linéaire correspondante ? Il nous a semblé nécessaire de faire le point sur ce sujet en montrant pourquoi la réponse ne pouvait pas être unique et en donnant des interprétations nouvelles de certaines décompositions. Comme nous allons le voir tout dépend de la philosophie que l'on a d'un test statistique : il s'agit de savoir si on peut considérer le modèle additif comme un modèle réel ou si on doit toujours le considérer comme une approximation, autrement dit comme un modèle inadéquat.

Les éléments de réponse à ces questions ne se trouvent pas toujours dans les monographies car beaucoup d'auteurs rechignent à écrire noir sur blanc certains choix arbitraires qui relèvent plus d'une norme que d'une réelle démarche scientifique. Ces choix sont pourtant nécessaires quand on veut passer à l'analyse proprement dite de cas concrets. C'est donc vers les notices des logiciels qu'il faut se tourner pour y lire la pratique statistique.

2. Bref rappel des formules principales de l'analyse de variance

Dans tout ce qui suit, nous entendrons par analyse de variance, l'explication d'une variable quantitative Y par une fonction d'un certain nombre de variables qualitatives ou facteurs avec des erreurs indépendantes centrées équidistribuées (et éventuellement Gaussiennes). C'est un modèle linéaire dans le sens où la variable $Y \in \mathbb{R}^n$ a une espérance qui varie dans un sous espace $[H_1]$ de dimension r_1 et une variance égale à $\sigma^2 I_n$ (I_n est la matrice identité de taille n). On sait que les estimateurs optimaux sont ceux des moindres carrés et, dans le cas où il n'y a qu'un facteur, ce sont des moyennes ordinaires. Dans le cas le plus général, ils sont obtenus par solution matricielle des équations normales.

Soit H_0 l'hypothèse nulle : $E(Y) \in [H_0]$ (espace de dimension r_0). Le test uniformément plus puissant de cette hypothèse parmi les tests invariants est basé sur la statistique (Coursol 1980) :

$$\hat{F} = \frac{\|\mathbf{P}_{[H_1]/[H_0]} Y\|^2 / (r_1 - r_0)}{\hat{\sigma}^2}$$

avec

$$\hat{\sigma}^2 = \|Y - \mathbf{P}_{[H_1]} Y\|^2 / (n - r_1)$$

où $[H_1]/[H_0]$ est l'orthogonal de $[H_0]$ dans $[H_1]$ et $\mathbf{P}_E$ désigne le projecteur orthogonal sur l'espace E .

Nous allons d'abord étudier en détail le modèle à deux facteurs croisés avec interaction car il contient dans un cadre simple toutes les difficultés conceptuelles. Nous examinerons rapidement les autres cas en fin d'article.

Nous dirons que le facteur A est plus fin que le facteur B si l'égalité des niveaux du facteur A implique celle des niveaux du facteur B . Dans notre cas particulier l'interaction (représentée par le facteur produit) est plus fine que les effets principaux ; chaque effet principal est plus fin que l'effet «moyenne générale».

Nous utiliserons les conventions classiques « $\cdot$ » et « $+$ » pour désigner respectivement la moyenne et la somme sur l'indice remplacé. Par exemple si on dispose de I valeurs, $Y_1 \dots Y_I$, on a :

$$Y. = (1/I)Y_+ = (1/I) \sum_{i=1,I} Y_i$$

Nous serons parfois amenés à citer quelques commandes de SAS (ou d'autres logiciels). Pour les distinguer du reste du texte nous les placerons entre les symboles $\langle \rangle$.

3. Trois philosophies différentes dans le cas de deux facteurs croisés avec interaction

3.1. L'approche de GLIM

GLIM est un logiciel de modélisation et non pas un logiciel de modèle linéaire. C'est pour cette raison qu'il propose une approche extrémiste quand il est appliqué à de l'analyse de variance. Cette approche a peu d'intérêt pratique dans la mesure où elle se démarque nettement des autres, mais elle est simple et constitue une bonne introduction.

On s'interdit de tester la significativité d'un effet dans un modèle s'il existe un autre effet plus fin que lui dans le modèle. Par définition une interaction est plus fine qu'un effet principal. En conséquence, dans une analyse à deux facteurs, on testera d'abord la présence de l'interaction. Si elle est déclarée significative les deux facteurs sont pertinents et on s'arrête là. Si elle est non significative, on teste les effets principaux dans le modèle additif.

Cette approche se trouve en contradiction avec une pratique qui tend à faire l'unanimité sur le plan statistique : le «non-pooling» ou «non regroupement» : on ne regroupe pas les différentes sommes de carrés des effets non significatifs avec la somme des carrés résiduelle.

3.2. Le «Non-regroupement» point commun entre l'approche de GENSTAT et l'approche américaine

La critique de l'approche simple 3.1 est basée sur l'argument méta-statistique suivant : Une fonction de deux variables qualitatives n'a «aucune chance» d'être strictement additive. Le modèle additif n'est qu'une approximation. Il faut donc, en testant un effet principal, prendre une estimation de la variance σ^2 qui ne dépende pas de la présence d'une faible interaction, non détectée par le test.

Rappelons que le dénominateur du test de Fisher est simplement un estimateur de σ^2 . Dans l'approche précédente, quand on teste un effet principal, on incorpore l'interaction dans l'espace résiduel et donc dans l'estimation de σ^2 . Or cette interaction, bien qu'ayant été déclarée non significative, peut exister et biaiser l'estimation. Pour éviter cela on ne regroupe pas («non-pooling») l'interaction avec la résiduelle et on utilise l'estimation de σ^2 du modèle interactif.

3.3. L'approche de Genstat et de Splus

Elle est basée sur la philosophie de Nelder : les hypothèses testées sont définies par une succession de modèles. Dans le modèle à deux facteurs, on ajuste d'abord le premier, ensuite de manière additive le second et enfin l'interaction. Les différences de sommes de carrés définissent les numérateurs du test de Fisher. Pour respecter le principe du «non-regroupement», le dénominateur est calculé à partir de la résiduelle du modèle le plus grand : le modèle avec toutes les interactions. On trouve cette approche dans la directive <ANOVA> de Genstat pour le cas équilibré et dans la macro <AUNBALANCED> pour le cas non équilibré. Dans la philosophie de SAS, il s'agit d'une décomposition de type I (cf 4.2). Dans Splus, la directive <AOV> donne la même décomposition et par l'intermédiaire de la directive <Drop> on obtient ce qui sera ensuite appelé la décomposition de type II (cf 4.3).

3.4. L'approche de SAS, BMDP et SYSTAT

Elle repose sur les mêmes prémices que l'approche précédente : le modèle additif est forcément inadéquat. Mais maintenant on va, non seulement en tirer des conséquences pour l'estimation de la variance σ^2 , mais encore pour la définition des effets principaux dans le modèle complet.

Si la démarche qui consiste à définir un effet principal en présence d'une interaction qu'il compose peut paraître étrange, remarquons qu'elle est couramment employée dans toute la partie de la statistique qui s'occupe des plans factoriels (John 1971).

Dans l'exposé de cette dernière approche nous allons considérer trois niveaux de complexité.

– *Le cas équirépété : chaque combinaison des deux facteurs est répétée le même nombre r de fois.*

– *Le cas non équirépété mais avec toutes les «cellules» observées. La modalité i du premier facteur et la modalité j du second facteur sont observées r_{ij} fois avec $r_{ij} > 0$.*

– *Le cas de «cellules» non observées : $r_{ij} = 0$ pour certaines valeurs.*

4. Analyse de variance à deux facteurs croisés quand toutes les cellules sont observées ($r_{ij} > 0$).

4.1. Décomposition de type III

On a mesuré les variables Y_{ijk} : $k^{\text{ème}}$ observation de la modalité i du premier facteur (que nous appellerons A) avec la modalité j du second facteur (que nous appellerons B),

$$i = 1, \dots, I; j = 1, \dots, J; k = 1, \dots, r_{ij} \text{ avec } \forall i, \forall j, r_{ij} > 0.$$

Il est bien connu que la présentation des formules est plus simple en travaillant au niveau du facteur croisé (i, j) plutôt qu'en introduisant dès le début les effets principaux. On pose donc le modèle d'analyse de variance à **un** facteur (i, j) .

$$E(Y_{ijk}) = \theta_{ij} \quad (1)$$

dont l'estimateur est $\hat{\theta}_{ij} = Y_{ij}$. Le modèle (1) est régulier, toute fonction est donc estimable. On définit ensuite les effets principaux, la moyenne générale et l'interaction par les formules classiques :

- moyenne générale $\mu = \theta_{..}$
- effet du premier facteur (A) $\alpha_i = \theta_{i.} - \theta_{..}$
- effet du second facteur (B) $\beta_j = \theta_{.j} - \theta_{..}$
- effet de l'interaction (A*B) $\gamma_{ij} = \theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..}$

qui seront estimés par les mêmes quantités avec un chapeau. La décomposition ci-dessus sera appelée : décomposition (D).

Cas équiréparté ($r_{ij} = r$)

Les tests de la nullité des effets γ, α, β issus de la décomposition (D) définissent respectivement le test de l'interaction et les tests sans «regroupement» des effets A et B. La seule alternative existante est le test «avec regroupement» de A et B dans le modèle additif, suivant la philosophie GLIM.

Cas non équiréparté

Commençons par une remarque très importante. Il n'y a qu'un seul test de l'interaction. Quelle que soit la situation, l'hypothèse «la fonction θ_{ij} est additive» ne peut s'exprimer que comme la nullité de tous les 2×2 contrastes :

$$\text{pour tout } i, j, i', j' : \theta_{ij} - \theta_{i'j} - \theta_{ij'} + \theta_{i'j'} = 0.$$

Par contre il y a plusieurs tests de la présence d'effets principaux car **leur définition n'est pas unique.**

La décomposition (D) est la décomposition de type III de SAS qui est équivalente dans notre cas au «Weighted squares of means» de Yates (1934). Le test de l'effet A se traduit par le test de la nullité des paramètres α de la décomposition (D), c'est-à-dire :

$$\theta_{i.} \text{ est constant.}$$

Ce test est également proposé comme choix standard par BMDPAV2 et par SYSTAT.

4.2. Décomposition de type I

Si les données ne sont pas équiréparties, il convient de savoir pourquoi et quel sens ont alors les effectifs r_{ij} . Souvent ils n'en ont aucun, car ils sont conséquence de données manquantes, de parties de l'expérience qui n'ont pu être conduites, ou d'une planification fantaisiste. Dans ce cas, il vaut mieux utiliser la décomposition de type III. Mais dans les cas où les effectifs traduisent l'importance réelle des différentes cellules, on calculera les marges du tableau θ_{ij} en utilisant les poids r_{ij} . On testera donc :

$$\rho_i = \sum_j \frac{r_{ij}}{r_{i+}} \theta_{ij} \text{ ne dépend pas de } i. \quad (2)$$

Dans ce modèle la combinaison des pondérations implique :

$$\hat{\rho}_i = Y_{i..}$$

et il est facile d'en déduire que le numérateur du test de Fisher de l'hypothèse (2) est le même que celui du test de l'effet A dans une analyse de variance à un seul facteur. Ce test est obtenu dans SAS comme le test de type I de A dans un modèle où A est placé en premier.

Dans SAS un test de type I est un test ajusté pour tous les termes précédents dans l'écriture du modèle. Dans le modèle $A_1, A_2, \dots, A_n$, le test de type I de l'effet A_i est le test de l'hypothèse nulle : «la projection de $E(Y)$ sur l'orthogonal de $A_1 + A_2 + \dots + A_{i-1}$ dans $A_1 + \dots + A_i$ est nulle». En conséquence dans le modèle $A, B, A * B$, l'effet A n'est ajusté (dans les sommes de carrés de type I) que pour la moyenne générale (ce qui est implicite).

4.3. Décomposition de type II

Les tests de type II sont les tests de type I où l'effet considéré est placé en dernier par rapport aux effets qui ne sont pas plus fins que lui. En conséquence le test de type II de A est maintenant ajusté pour la moyenne générale et pour B (mais pas pour $A * B$ car $A * B$ est plus fin que A). Ce test a même numérateur que le test de A dans le modèle additif. Ils ne diffèrent que par le «non-regroupement» pour l'estimation de σ^2 .

L'hypothèse testée ne peut s'exprimer simplement en fonction des réponses cellulaires θ_{ij} du modèle (1). En effet, les estimateurs du modèle additif n'ont pas, dans le cas général, une expression simple sous forme de moyennes, ils dépendent de la solution matricielle des équations normales (ces hypothèses peuvent être obtenues dans SAS/GLM par l'option `< /e2 >` (fonctions estimables de type 2) dans la ligne `<model>`). En conséquence, en considérant le modèle additif comme inadéquat, on ne sait pas très bien ce que l'on teste avec les sommes de carrés de type II. Cependant on ne possède pas de contre-exemple flagrant où l'hypothèse testée perdrait tout sens.

4.4. Autres types de décompositions

Il peut exister une pondération p_{ij} de l'importance des différentes cellules qui soit distincte, à la fois des effectifs (analyse de type I) et de la pondération uniforme ($p_{ij} = \text{cste}$; analyse de type III). Dans ce cas, on définit les effets principaux comme des moyennes utilisant ces pondérations. Pour tester l'effet A on testera l'hypothèse :

$$\rho_i = \sum_j \frac{p_{ij}}{p_{i+}} \theta_{ij} \text{ ne dépend pas de } i \quad (3)$$

Ce type d'analyse n'est pas disponible de manière standard dans les logiciels statistiques. Une utilisation astucieuse et laborieuse de la directive <contrast> de SAS permet cependant de répondre à la question.

NB : Dans le cas où toutes les cellules sont observées la décomposition de type IV qui sera définie à la section 5 est strictement équivalente à la décomposition de type III.

Exemple : considérons le cas le plus simple de deux facteurs à deux niveaux et la table d'effectifs :

n	1
1	1

La non présence d'effet ligne se traduit respectivement, dans les analyses de type I (avec ligne en premier) II et III, par la nullité du contraste suivant sur θ_{ij} .

$$(1/(2n+2)) \begin{bmatrix} 2n & 2 \\ -(n+1) & -(n+1) \end{bmatrix}; (1/(3n+1)) \begin{bmatrix} 2n & n+1 \\ 2n & -(n+1) \end{bmatrix};$$

$$(1/2) \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$$

Quand n vaut 1 toutes ces quantités sont égales. Quand n tend vers l'infini, leur limite vaut :

$$(1/2) \begin{bmatrix} 2 & 0 \\ -1 & -1 \end{bmatrix}; (1/3) \begin{bmatrix} 2 & 1 \\ -2 & -1 \end{bmatrix}; (1/2) \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}.$$

5. Analyse de variance à deux facteurs croisés quand certaines cellules sont absentes ($r_{ij} = 0$)

Remarquons d'abord que le problème de définition des effets principaux est mal posé dans la mesure où on peut avoir des réponses arbitrairement grandes ou

petites dans les cellules non observées. Il n'est donc pas possible d'estimer les moyennes marginales (θ_i , par exemple) définies en 4.1. Cela n'empêche cependant pas l'existence de réponses pratiques qui exigent toutefois une propriété minimale du modèle, la connexité : c'est-à-dire que le modèle additif soit estimable. Dans tout ce qui suit nous supposons que cette condition est vérifiée.

Remarquons encore que le test de l'interaction ne teste pas la totalité de l'espace d'interaction de la table complète, mais simplement une sous partie qui correspond à l'orthogonal de l'espace des fonctions additives, dans l'espace des fonctions des cellules observées. Si ce test est toujours unique, il n'en est pas de même du test des effets principaux.

5.1. Décomposition de type II

Sa définition est la même que dans le cas où toutes les cellules sont observées. Elle ne diffère donc du modèle additif que par l'utilisation d'un estimateur de la variance σ^2 sans «regroupement». Cette approche est conseillée par certains auteurs (Searle 1987) pour éviter les méthodes trop complexes. En effet tous les paramètres étant définis grâce à la connexité, elle pose très peu de problèmes conceptuels.

Cependant, comme pour toute analyse de type II, il est difficile d'exprimer simplement les hypothèses testées en fonction des paramètres θ_{ij} . (SAS/GLM permet de voir les formes de ces hypothèses dans chaque cas particulier en affichant les fonctions estimables de type II par l'option `</e2>` dans la ligne `<model>`).

Dans le cas où certaines cellules sont absentes, la décomposition (D) ne peut être utilisée car des moyennes associées à des niveaux différents du facteur A (par exemple) sont calculées sur des ensembles de niveaux différents du facteur B. Elles ne peuvent donc être comparées. Nous allons présenter deux décompositions (type III et IV) qui, dans le cas où toutes les cellules sont observées, coïncident avec la décomposition vue en 4.1 mais qui sont différentes de la décomposition(D) dans le cas général.

5.2. Décomposition de type IV

C'est l'analyse la plus séduisante *a priori*, mais nous allons voir que sa mise en œuvre par SAS pose certains problèmes. Considérons une table croisée 4×5 de la forme :

où les cases hachurées correspondent aux cellules non observées. Pour comparer les niveaux 1 et 2 du facteur ligne, on le fera dans la sous-table obtenue avec les niveaux 3, 4 et 5 du facteur colonne, ce qui donne une sous-table complète. Par contre pour comparer les niveaux 2, 3 et 4, on utilisera la sous-table complète obtenue avec les niveaux 1, 2 et 3 du facteur colonne. De manière générale on compare deux niveaux dans la plus grande sous-table complète qui les contient.

Cette approche présente un inconvénient majeur : il y a $I \times (I - 1)/2$ comparaisons entre les I lignes qui correspondent chacune à la nullité d'un contraste de type IV. La nullité de tous ces contrastes définit un espace qui est en général de dimension plus importante que $I - 1$ qui est la taille que l'on s'attend à trouver pour l'espace de l'effet principal ligne.

L'algorithme SAS ne considère que les $I - 1$ comparaisons **avec le dernier niveau**. Ces comparaisons sont libres car elles portent sur des lignes différentes. On a donc bien défini un espace de taille requise, mais il dépend de l'ordre de codage des différents niveaux du facteur. On peut le vérifier dans l'exemple du Tableau 1 où le F de l'effet ligne peut passer de 528 à zéro à la suite d'une numérotation à l'envers des niveaux.

Comment trouver la $i^{\text{ème}}$ fonction de type IV selon SAS, associée à l'effet ligne (par exemple)? On considère le nombre k de colonnes communes à la ligne i et la ligne I . On place les coefficients $1/k$ sur la ligne i et les colonnes communes et les coefficients $-1/k$ sur la ligne I et les colonnes communes. La combinaison linéaire correspondante appliquée aux θ_{ij} définit la $i^{\text{ème}}$ fonction estimable de type IV. La nullité conjointe de ces $(I - 1)$ fonctions estimables définit l'hypothèse de type IV d'absence d'effet ligne. Voici deux exemples sur des tables 3×3 où sont représentées les deux fonctions estimables, l'une en *italique* et l'autre en **gras**.

	<i>1</i>	
1		
-1	<i>-1</i>	

	<i>1/2</i>	<i>1/2</i>
1		
-1	<i>-1/2</i>	<i>-1/2</i>

Dans le premier cas l'absence d'effet ligne se traduit par :

$$\theta_{12} = \theta_{32}; \theta_{21} = \theta_{31}.$$

On voit bien que si l'on applique cette méthode à la table donnée en introduction du paragraphe, les contrastes de type IV de comparaison des lignes 2, 3 et 4 correspondront bien à l'idée de se ramener à une sous-table complète. Mais cela n'est plus vrai pour le troisième contraste qui compare les niveaux 1 et 4.

L'inconvénient majeur des contrastes (ou fonctions estimables) de type IV peut être deviné sur les diagrammes ci-dessus : si l'on renverse par exemple l'ordre des lignes du premier tableau en miroir, les fonctions estimables porteront maintenant respectivement sur la deuxième et la troisième colonne. C'est sur cette considération qu'est construit le contre-exemple du tableau 1. De part et d'autre on a représenté

TABLEAU 1
*Variation de l'analyse de type IV de SAS quand on change
l'ordre de numérotation d'un facteur.*
Les parties communes ont été placées en position médiane.

data toto;	data toto;
input trait bloc reponse;	input trait bloc reponse;
cards;	cards;
1 2 0	4 2 0
1 2 1	4 2 1
1 2 0	4 2 0
1 3 4	4 3 4
1 3 4	4 3 4
1 4 0	4 4 0
2 1 0	3 1 0
2 3 16	3 3 16
2 3 16	3 3 16
2 4 20	3 4 20
2 4 19	3 4 19
3 1 12	2 1 12
3 4 10	2 4 10
3 4 9	2 4 9
4 1 12	1 1 12
4 3 4	1 3 4
;	;
run;	run;
proc glm;	proc glm;
class trait bloc;	class trait bloc;
model reponse = trait bloc / ss3 ss4;	model reponse = trait bloc / ss3 ss4;
run;	run;

Dependent Variable: REPONSE

Source	DF	Sum of Squares	Mean Square	F Value
Model	9	781.27083333	86.80787037	312.51
Error	6	1.66666667	0.27777778	
Corrected Total	15	782.93750000		

Source	DF	Type III SS	Mean Square	F Value	Type III SS	Mean Square	F Value
TRAIT	3	215.23592571	71.74530857	258.28	215.23592571	71.74530857	258.28
BLOC	3	35.74529289	11.91509763	42.89	35.74529289	11.91509763	42.89
TRAIT*BLOC	3	272.73410405	90.91136802	327.28	272.73410405	90.91136802	327.28

Source	DF	Type IV SS	Mean Square	F Value	Type IV SS	Mean Square	F Value
TRAIT	3*	0.00000000	0.00000000	0.00	440.18333333	146.72777778	528.22
BLOC	3*	102.12083333	34.04027778	122.55	102.12083333	34.04027778	122.55
TRAIT*BLOC	3	272.73410405	90.91136802	327.28	272.73410405	90.91136802	327.28

The SAS System 10:50 Thursday, February 13, 19

General Linear Models Procedure

* NOTE: Other Type IV Testable Hypotheses exist which may yield different SS.

deux analyses identiques d'un même corpus de données. La seule différence est que dans un des cas le facteur trait est numéroté de manière croissante et dans le second de manière décroissante. Les parties communes ont été placées en position médiane.

En conclusion l'utilisation de sommes de carrés de type IV est dangereuse dès qu'elles ne sont pas uniques.

Remarque : Notre définition des fonctions estimables de type IV n'est pas complète puisque nous n'avons pas précisé ce qui se passe dans le cas où $k = 0$. La réponse n'est pas claire. Cela n'est expliqué nulle part dans la documentation SAS et le comportement de l'algorithme dans ce cas est difficile à comprendre. De toutes façons, l'exemple montre que l'on ne peut se satisfaire d'une utilisation aveugle des sommes de carrés de type IV surtout quand elles ne sont pas uniques (on peut

obtenir les fonctions estimables correspondantes par l'option `< /e4 >` dans la ligne `<model>`).

5.3. Décomposition de type III

Comme on le vérifie dans l'exemple du tableau 1, les sommes de carrés de type III sont indépendantes de l'ordre. On peut les définir de la façon suivante (Goodnight 1978) :

On considère l'espace $\mathcal{E}$ des fonctions réelles définies sur les cellules observées. Cet espace est de dimension inférieure à IJ . On définit $\mathcal{L}$ et $\mathcal{C}$ comme les sous espaces vectoriels de $\mathcal{E}$ composés de fonctions constantes respectivement par lignes et par colonnes. On définit l'espace $\mathcal{A}$ des fonction additives $\mathcal{A} = \mathcal{L} + \mathcal{C}$. Comme la table est connexe, $\dim(\mathcal{L}) = I$, $\dim(\mathcal{C}) = J$, $\dim(\mathcal{A}) = I + J - 1$. On considère la métrique qui donne des poids égaux à toutes les cellules observées et des poids nuls aux autres. Pour cette métrique, on définit $\mathcal{J}$: l'espace de l'interaction (de la table donnée) comme étant l'orthogonal de $\mathcal{A}$ dans $\mathcal{E}$. Enfin on définit pour l'effet ligne

$$\mathcal{H} = \mathcal{J}^\perp \cap \mathcal{C}^\perp = \mathcal{A} \cap \mathcal{C}^\perp.$$

L'espace $\mathcal{H}$, qui est de dimension $I - 1$, définit l'espace des fonctions estimables de type III associées à l'effet ligne : à tout élément h de $\mathcal{H}$ on associe la fonction estimable de type III

$$\sum h_{ij} \theta_{ij}.$$

La nullité conjointe de toutes ces fonctions définit l'hypothèse de type III. Par construction ces fonctions ne dépendent

- ni d'un effet colonne
- ni d'une interaction (du sous espace particulier associé à la table) rajoutée aux données.

Interprétation

L'intérêt de la première propriété est évident : on évite la confusion entre les deux facteurs; c'est une propriété minimale.

L'intérêt de la seconde l'est moins, mais dans le cas où les effectifs r_{ij} non nuls sont égaux (ou presque égaux), le résultat suivant qui semble nouveau fournit une interprétation.

Proposition : Soit N l'ensemble des cellules non vides, $\langle \cdot \rangle$ et $\|\cdot\|$, le produit scalaire et la norme sur $\mathcal{E}$:

$$\langle f, g \rangle = \sum_{i,j \in N} f_{ij} \cdot g_{ij},$$

soit h un élément de $\mathcal{H}$,

alors h est de norme minimale parmi toutes les fonctions ν de $\mathcal{E}$ telles que

$$\sum_{i,j \in N} h_{ij} \theta_{ij} = \sum_{i,j \in N} \nu_{ij} \theta_{ij} \quad (4)$$

si la réponse θ_{ij} est une fonction additive : $\theta_{ij} = a_i + b_j$.

En particulier, si les effectifs non nuls r_{ij} sont tous égaux, la fonction estimable de type III :

$$\sum_{i,j \in N} h_{ij} \hat{\theta}_{ij}$$

est de variance minimale parmi toutes les fonctions estimables ayant la même espérance sous le modèle additif.

Remarquons que l'expression (4) est indépendante des valeurs b_j .

Démonstration. Soit v appartenant à $\mathcal{E}$ et satisfaisant (4). Les contraintes impliquent que

$$v = h + k \text{ où } k \in \mathcal{J}.$$

Cette somme est orthogonale. En utilisant le théorème de Pythagore :

$$\|v\|^2 = \|h\|^2 + \|k\|^2$$

d'où la minimalité de la norme de h .

Quand tous les effectifs sont égaux ($r_{ij} = r$) la variance de la fonction estimable associée à v vaut : $(\sigma^2/r)\|v\|^2$. Si la norme de h est minimale, la variance de la fonction estimable associée l'est également. $\square$

Remarque : l'interprétation ci-dessus constitue une réponse aux critiques de Milliken et Johnson (1984 p. 185).

Pour trouver numériquement les fonctions de type III, on orthogonalise les fonctions de type IV. Considérons par exemple la table :

Une des fonctions de type IV vaut

	1	
	-1	

Si on lui ajoute la quantité suivante qui appartient à l'espace de l'interaction (qui est de dimension 1)

	-1/3	1/3
1/3		-1/3
-1/3	1/3	

On obtient la fonction suivante dont on peut vérifier qu'elle appartient à $\mathcal{H}$.

		2	1
1/3	1		-1
	-1	-2	

NB : Il est bien-sûr clair que dans le cas de dispositifs contenant très peu de cellules manquantes, on aura intérêt à estimer les quelques données manquantes pour utiliser les résultats de la partie 4.

6. Autres modèles

6.1. Le modèle hiérarchique à deux facteurs

Dans ce modèle on ne considère que deux effets : l'effet principal du premier facteur (noté A) et l'effet hiérarchisé du second facteur (noté $B(A)$) qui regroupe l'effet principal de B et l'interaction de la décomposition précédente. Il n'y a qu'une décomposition possible : la somme des carrés du modèle complet (1) ajustée pour l'effet de A teste la pertinence de l'effet hiérarchisé B ; la somme de carrés de A ajustée pour la moyenne générale teste la pertinence de A . La seule question est l'estimation de σ^2 dans ce dernier test : la plupart des logiciels proposent par défaut un estimateur sans «regroupement».

6.2. Le modèle additif

Bien que notre philosophie soit de penser que ce modèle ne devrait pas être employé car il n'est qu'une approximation, nous ne pouvons éviter d'en parler dans la mesure où quand r_{ij} est constant et égal à 1 on ne peut utiliser un autre modèle. On pose donc dans le modèle (1)

$$\theta_{ij} = a_i + b_j.$$

Deux décompositions sont possibles :

– La décomposition de type II qui est la seule raisonnable (comme il n'y a pas d'effet plus fin qu'un autre, les décompositions de type III et IV sont sans objet et sont définies par convention comme égales à celle de type II). Chacun des deux facteurs est ajusté pour l'autre. Pour le premier facteur par exemple on teste :

$$a_i = \text{Cte}; \text{ quand } i \text{ varie.}$$

– La décomposition de type I donne des hypothèses bizarres. Pour le premier facteur, on teste (Searle 1987 p121) :

$$a_i + \left(\sum n_{ij} b_j \right) / n_{i+} = \text{Cte}; \text{ quand } i \text{ varie.}$$

6.3. Modèles à plus de deux facteurs

Quand le modèle n'a que des facteurs croisés, on applique les principes de la partie 4 en définissant les moyennes marginales successives du tableau des moyennes associées au facteur produit.

Quand un modèle comprend deux facteurs hiérarchisés A et B , on définit l'effet de A et l'effet de B hiérarchisé à A . Le problème est d'abord de savoir si le facteur B est pertinent. On teste donc si on peut remplacer, dans le modèle complet, le facteur croisé $A * B$ par A seul :

Si oui on enlève l'effet hiérarchisé

Si non on pose le modèle avec $A * B$ seul.

Dans les deux cas on ne conserve qu'un seul facteur : A ou $A * B$ et on peut se ramener au cas précédent où tous les facteurs sont croisés. Pour respecter le principe du «non-regroupement», l'effet hiérarchisé n'est pas incorporé dans la résiduelle dans le cas où l'on conserve A seul.

7. Conclusion

Pour un dispositif déséquilibré et dans le cas d'un modèle non strictement additif, la définition d'un effet principal est délicate. Plusieurs réponses sont possibles. SAS a le mérite de poser le problème et de donner plusieurs réponses. Cela est

déroutant pour le débutant, mais cela ne masque pas les difficultés par des choix arbitraires. Fort heureusement dans beaucoup de cas, les différents types d'analyses donnent des résultats concordants et il n'y a pas lieu d'étudier plus finement le problème. Dans les autres cas, il faudra déterminer avec soin la décomposition qui correspond le mieux au problème pratique étudié.

Références

- COURSOL J. (1980), *Technique statistique des modèles linéaires*. Cimpa, Nice.
- GOODNIGHT J.H. (1978), *Test of hypotheses in fixed effects linear models*. SAS technical Report R-101.
- JOHN P.M.W. (1971), *Statistical design and analysis of experiments*. Mac Millan Compagny, New York, 1971.
- MILLIKEN G.A., JOHNSON D.E. (1984), *Analysis of messy data vol 1*. Van Nostrand Reinhold, N.Y., USA.
- SEARLE S. (1987), *Linear models for unbalanced data*. Wiley, New York.
- SPEED T.D. (1987), *What is an analysis of variance ?* The Annals of Statistics, 15 (3), 885-910.
- TUCKEY J.W. (1991), *The philosophy of multiple comparisons*. Statistical Science, 6(1), 100-116.
- YATES F. (1934), *The analysis of multiple classifications with unequal numbers in the different classes*. J. Amer. Stat. Assoc, 29, 51-66.

Références des logiciels

- BMDP : DIXON W.J., BROWN M.B., ENGELMAN L., JENNRICH R.T. (1990). *Bmdp Statistical Software manual*. University of California Press, Berkeley. U.S.A.
- GENSTAT : PAYNE R.W., LANE P.W., AINSLEY A.E., BICKNELL K.E., DIGBY P.G.N., HARDING S.A., LEECH P.K., SIMPSON H.R., TODD A.D., VERRIER P.J., WHITE R.P., GOWER J.C., TUNNICLIFFE-WILSON G., PATERSON L.J. (1987). *Genstat 5 Reference Manual*. Oxford University Press, Oxford, U.K.
- GLIM : Payne C.D. (1987). *The GLIM system*. N.A.G., Oxford. U.K.
- SAS : SAS Institute Inc (1989), *SAS/STAT² User's guide, version 6, Fourth edition, vol 1 & 2*, Cary, NC : SAS Institute INC. U.S.A.
- Spplus : BECKER R.A., CHAMBERS J.M., WILKS A.R. (1988). *The new S language*. Wadsworth, Pacific Grove, California. U.S.A.
- SYSTAT : WILKINSON L. (1990). *SYSTAT : The System for Statistics*. EVANSTON IL. : SYSTAT, INC. U.S.A.