ANOVA : analyse de variance univariée

Résumé

Le chapitre 3 est consacré aux plans factoriels. Il s'agit de l'appellation appropriée, bien qu'assez peu employée, de l'analyse de variance, appelée par les anglo-saxons "ANalysis Of VAriance" et, pour cette raison, bien connue sous l'acronyme d'ANOVA.

Retour au plan du cours

1 Introduction

L'ANOVA correspond à un modèle linéaire gaussien dans lequel toutes les variables explicatives (les X^j) sont qualitatives. Dans ce contexte, elles sont appelées **facteurs** (d'où le terme de plans factoriels) et leurs modalités sont appelées **niveaux**. Ces niveaux sont supposés choisis, fixés, par l'utilisateur, de sorte que l'on parle souvent de **facteurs contrôlés**. De son côté, la variable aléatoire réponse Y est toujours quantitative et supposée gaussienne.

Seuls seront traités dans ce chapitre les cas de l'analyse de variance à un facteur, à deux facteurs croisés et à trois facteurs croisés. Dans un dernier paragraphe, nous donnerons quelques indications sur les cas plus généraux dont certains seront étudiés au chapitre 4.

Les références bibliographiques du chapitre 3 sont les mêmes que celles du chapitre 2.

Les problèmes abordés dans chacun des paragraphes de ce chapitre seront, à chaque fois, les trois problèmes clés du modèle linéaire gaussien : estimation ponctuelle, estimation par intervalle de confiance et tests. Ils seront traités dans cet ordre, en particulier parce qu'on a besoin de certaines estimations ponctuelles pour construire un intervalle de confiance et pour faire un test. Mais, dans la pratique, on commence en général par faire différents tests pour choisir le modèle le plus adapté aux données considérées, puis on détermine les estimations des paramètres dans le modèle ainsi choisi.

Les paramètres que l'on va utiliser en ANOVA vont représenter des effets particuliers du modèle pris en compte : effet général et effets principaux des niveaux du facteur dans un plan à un seul facteur ; effet général, effets principaux des niveaux de chaque facteur et effets d'interactions dans un plan à deux facteurs... Ces différents effets ne peuvent être pris en compte si on conserve le paramétrage standard du modèle linéaire (par exemple, dans un modèle à deux facteurs, $Y_{ijk} = \beta_{jk} + U_{ijk}$). D'où la nécessité d'utiliser d'autres paramétrages. Il en existe plusieurs et nous en présentons deux dans ce chapitre : le paramétrage dit centré, car il fait intervenir des paramètres centrés, et le paramétrage SAS, utilisé systématiquement dans le logiciel SAS.

Ainsi, pour un plan à deux facteurs croisés, le paramétrage centré consiste à poser : $\beta_{jk} = \mu + \alpha_j^1 + \alpha_k^2 + \gamma_{jk}$. Le paramètre μ représente l'effet général, les paramètres α_j^1 et α_k^2 les effets principaux des deux facteurs et les paramètres γ_{jk} les effets d'interactions. Les α_j^1 sont centrés selon j, les α_k^2 selon k et les γ_{jk} selon j et selon k.

Le paramétrage SAS, tel qu'on le trouve en particulier dans la procédure GLM, consiste, de son côté, à réécrire : $\beta_{jk} = m + a_j^1 + a_k^2 + c_{jk}$. Les paramètres m, a_j^1, a_k^2 et c_{jk} représentent les mêmes notions que celles précisées ci-dessus, mais ils sont définis en se "callant" sur la dernière cellule, d'indice (J, K).

2 Cas d'un seul facteur

Lorsque nécessaire, le facteur considéré sera noté F ; cette notation est certes la même que celle de la statistique du test de Fisher, mais, dans le contexte, il ne devrait pas y avoir de confusion ; de plus, la notation du facteur sera peu utilisée. Par ailleurs, le nombre des niveaux de F sera noté J ($J \geq 2$) et l'indice du niveau courant noté j ($j = 1, \ldots, J$).

Pour chaque niveau j, on réalise n_j observations indépendantes de la v.a.r. (quantitative) à expliquer Y ($n_j \geq 1$), notées y_{ij}, $i = 1, \ldots, n_j$; on pose enfin $n = \sum_{j=1}^{J} n_j$: n est le nombre total d'observations réalisées dans l'expérience.

Si $n_j = n_0, \forall j, j = 1, \ldots, J$, on dit que le plan est **équilibré** ; sinon, on parle de plan **déséquili-bré**. Dans un plan équilibré, n_0 s'appelle le nombre de **répétitions**.

Remarque. — On a utilisé ci-dessus le terme de *plan*. C'est le terme utilisé

dans tout le contexte de l'ANOVA, où l'on parle de plan d'expériences [1] ou de plan factoriel, voire, tout simplement, de plan. En fait, ce terme est d'origine industrielle et, dans un tel environnement, on parle également d'expérience planifiée, ce qui sous-entend, d'ailleurs, que les niveaux du (ou des) facteurs pris en compte sont totalement contrôlés (d'où le terme de facteur contrôlé).

2.1 Écriture initiale du modèle

On commence par écrire le modèle sous la forme :

$$Y_{ij} = \beta_j + U_{ij}.$$

- β_j est le paramètre associé au niveau j du facteur F ; il est inconnu, à estimer ; ce paramètre représente un effet non aléatoire, encore appelé **effet fixe**.
- U_{ij} est la v.a.r. erreur associée à l'observation numéro i du niveau j de F ; on suppose $U_{ij} \sim \mathcal{N}(0, \sigma^2)$, σ^2 étant aussi un paramètre à estimer (il ne dépend pas de j, autrement dit le modèle est homoscédastique) ; par ailleurs, les v.a.r. U_{ij} sont supposées indépendantes (elles sont donc i.i.d.).
- Y_{ij} est la v.a.r. réponse associée à l'observation numéro i du niveau j de F ; on obtient donc $Y_{ij} \sim \mathcal{N}(\beta_j, \sigma^2)$, les Y_{ij} étant indépendantes.

On peut réécrire le modèle sous la forme matricielle

$$Y = \mathbf{X}\beta + U,$$

où Y et U sont des vecteurs de $\mathbb{R}^n$, β est un vecteur de $\mathbb{R}^J$ (ici, $p = J$) et $\mathbf{X}$, appelée **matrice d'incidence**, est une matrice $n \times J$ ne comportant que des 0 et des 1 ; en fait, chaque colonne de $\mathbf{X}$ est l'indicatrice du niveau correspondant de F et nous noterons Z^j l'indicatrice courante. On peut ainsi réécrire :

$$Y = \sum_{j=1}^{J} \beta_j Z^j + U.$$

1. Dans l'expression plan d'expériences, on trouve le terme d'expérience tantôt au singulier et tantôt au pluriel ; nous préférons utiliser le pluriel, d'une part parce que le même plan peut servir à plusieurs expériences, d'autre part parce que le *petit Robert* cite l'expression "Laboratoire d'expériences".

EXEMPLE 1 *Considérons le cas $J = 3$, $n_1 = 2$, $n_2 = 3$, $n_3 = 1$ ($n = 6$). Il vient :*

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Remarque. — Sous la dernière forme donnée ci-dessus, on voit que le modèle est équivalent à un modèle de régression multiple, sans coefficient constant, dont les régresseurs sont les J variables indicatrices Z^j.

Remarque. — On vérifie que les colonnes de $\mathbf{X}$ sont deux à deux orthogonales. On en déduit que $\mathbf{X}'\mathbf{X} = \text{ diag } (n_1 \cdots n_J)$: il s'agit d'une matrice régulière.

2.2 Paramétrage centré

Le paramétrage initial ne permet pas de dissocier d'une part les effets des différents niveaux du facteur F, d'autre part l'effet général (et les choses seront encore plus problématiques en présence de deux facteurs ou plus). D'où la nécessité de réécrire le modèle, le problème étant qu'il existe plusieurs réécritures distinctes (mais, bien sûr, équivalentes).

Dans le paramétrage centré, on pose :

$$\mu = \frac{1}{J}\sum_{j=1}^{J} \beta_j \text{ (moyenne "non pondérée" des } \beta_j) \; ; \; \alpha_j = \beta_j - \mu.$$

On obtient ainsi $\beta_j = \mu + \alpha_j$ et on réécrit le modèle sous la forme :

$$Y_{ij} = \mu + \alpha_j + U_{ij}.$$

On notera la relation $\sum_{j=1}^{J} \alpha_j = 0$.

- Le paramètre μ est appelé l'effet général, ou encore l'effet moyen général.
- Les paramètres α_j ($j = 1, \ldots, J$) sont appelés les effets principaux du facteur F, ou encore les effets différentiels. La littérature statistique anglo-saxonne parle fréquemment de *contrastes*, dans la mesure où il s'agit de paramètres de somme nulle.

– Dans $\mathbb{R}^n$, on peut réécrire le modèle sous la forme suivante :

$$\begin{aligned} Y &= \sum_{j=1}^{J} \beta_j Z^j + U = \mu \mathbb{1}_n + \sum_{j=1}^{J} \alpha_j Z^j + U \\ &= \mu \mathbb{1}_n + \sum_{j=1}^{J-1} \alpha_j Z^j - Z^J \sum_{j=1}^{J-1} \alpha_j + U \\ &= \mu \mathbb{1}_n + \sum_{j=1}^{J-1} \alpha_j (Z^j - Z^J) + U. \end{aligned}$$

On obtient maintenant un modèle de régression linéaire sur les $J-1$ variables $Z^j - Z^J$, avec coefficient constant.

2.2.1 Notation

On notera β_c le vecteur des J paramètres dans ce paramétrage (μ et les α_j, $j = 1, \ldots, J-1$) et $\mathbf{X}_c$ la matrice d'incidence correspondante, de sorte qu'on pourra réécrire $Y = \mathbf{X_c}\beta_c + U$.

EXEMPLE 2 *Dans l'exemple introduit plus haut, $\mathbf{X}_c$ et β_c ont pour expression :*

$$\mathbf{X}_c = (\mathbb{1}_n \ (Z^1 - Z^3) \ (Z^2 - Z^3)) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{pmatrix} ; \ \beta_c = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{pmatrix}.$$

La matrice $\mathbf{X}_c$ est toujours de rang 3, mais ses colonnes ne sont plus orthogonales. Toutefois, elles le seraient dans un plan équilibré.

2.3 Paramétrage SAS

Le principe de ce paramétrage est de se "caller" sur le dernier niveau J du facteur F. On pose ainsi

$$Y_{ij} = m + a_j + U_{ij},$$

avec $m = \beta_J$ et $a_j = \beta_j - \beta_J$, $\forall j = 1, \ldots, J$ (de sorte que $a_J = 0$). On peut alors réécrire :

$$Y = \sum_{j=1}^{J} \beta_j Z^j + U = \beta_J \mathbb{1}_n + \sum_{j=1}^{J} a_j Z^j + U = m \mathbb{1}_n + \sum_{j=1}^{J-1} a_j Z^j + U.$$

On voit qu'il s'agit d'un modèle de régression sur les $J-1$ indicatrices Z^j ($j = 1, \ldots, J-1$), avec coefficient constant. Pour cette raison, le paramètre m est appelé *intercept* dans SAS, comme le coefficient constant d'une régression.

2.3.1 Notation

On notera maintenant β_s le vecteur des J paramètres de ce paramétrage (m et les a_j, $j = 1, \ldots, J-1$) et $\mathbf{X}_s$ la matrice d'incidence correspondante, de sorte qu'on pourra réécrire $Y = \mathbf{X_s}\beta_s + U$.

EXEMPLE 3 *En considérant toujours le même exemple, $\mathbf{X}_s$ et β_s ont pour expression :*

$$\mathbf{X}_s = (\mathbb{1}_n \ Z^1 \ Z^2) = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} ; \ \beta_s = \begin{pmatrix} m \\ a_1 \\ a_2 \end{pmatrix}.$$

La matrice $\mathbf{X}_s$ est encore de rang 3, ses colonnes n'étant pas non plus orthogonales. On notera qu'elles ne le seraient pas davantage dans le cas d'un plan équilibré.

2.4 Estimation des paramètres

En applicant les résultats généraux relatifs à l'estimation dans le modèle linéaire gaussien, on obtient les résultats indiqués ci-dessous.

Vecteur des paramètres dans le paramétrage initial.

$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$, avec $\mathbf{X}'\mathbf{X} = \text{diag}(n_1 \cdots n_J)$ et $\mathbf{X}'y = \begin{pmatrix} n_1\overline{y}_{\bullet 1} \\ \vdots \\ n_J\overline{y}_{\bullet J} \end{pmatrix}$,

où $\overline{y}_{\bullet j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$. On obtient ainsi $\hat{\beta}_j = \overline{y}_{\bullet j}$. De plus $\hat{B} \sim \mathcal{N}_J(\beta, \sigma^2 \text{diag}(\frac{1}{n_1} \cdots \frac{1}{n_J}))$, de sorte que les composantes $\hat{B}_j$ sont indépendantes.

Paramétrage centré.

Il vient : $\hat{\mu} = \frac{1}{J}\sum_{j=1}^{J} \overline{y}_{\bullet j}$ (*attention :* si les effectifs n_j ne sont pas tous égaux, autrement dit si le plan est déséquilibré, $\hat{\mu}$ n'est pas la moyenne générale des observations de Y, notée $\overline{y}_{\bullet\bullet}$). D'autre part, $\hat{\alpha}_j = \overline{y}_{\bullet j} - \hat{\mu}$, de sorte qu'on retrouve $\sum_{j=1}^{J} \hat{\alpha}_j = 0$.

Paramétrage SAS.

On obtient maintenant $\hat{m} = \overline{y}_{\bullet J}$ et $\hat{a}_j = \overline{y}_{\bullet j} - \overline{y}_{\bullet J}$, de sorte qu'on vérifie bien $\hat{a}_J = 0$.

Valeurs prédites.

Elles sont définies par $\hat{y}_{ij} = \hat{\beta}_j = \overline{y}_{\bullet j}$. Elles ne dépendent pas du paramétrage considéré.

Résidus.

On obtient $\hat{u}_{ij} = y_{ij} - \hat{y}_{ij} = y_{ij} - \overline{y}_{\bullet j}$ (même remarque que ci-dessus).

Variance.

Comme pour les valeurs prédites et les résidus, l'estimation de la variance ne dépend pas du paramétrage choisi. Il vient :

$$\hat{\sigma}^2 = \frac{1}{n-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\hat{u}_{ij})^2 = \frac{1}{n-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_{\bullet j})^2.$$

Intervalle de confiance de β_j, de coefficient de sécurité $1-\alpha$.

On obtient l'intervalle de type Student suivant :

$$\overline{y}_{\bullet j} \pm \frac{\hat{\sigma}}{\sqrt{n_j}} t_{n-J}(1 - \frac{\alpha}{2}),$$

où $t_{n-J}(1-\frac{\alpha}{2})$ désigne le quantile d'ordre $1-\frac{\alpha}{2}$ d'une loi de Student à $n-J$ d.d.l. On notera que $\frac{\hat{\sigma}}{\sqrt{n_j}}$ est l'erreur-type de $\hat{B}_j$.

Erreurs-types

Un calcul simple permet de vérifier que l'erreur-type de $\hat{Y}_{ij}$ est encore $\frac{\hat{\sigma}}{\sqrt{n_j}}$, tandis que celle de $\hat{U}_{ij}$ est $\hat{\sigma}\sqrt{\frac{n_j - 1}{n_j}}$. On notera que ces erreurs-types sont constantes dans le cas équilibré.

2.5 Test de l'effet du facteur F

Tester la significativité du facteur F revient à tester la significativité du modèle envisagé. Dans le paramétrage initial du modèle, l'hypothèse nulle se met sous la forme $\{H_0 : \beta_1 = \cdots = \beta_J\}$, l'alternative étant le contraire de H_0.

Dans les autres paramétrages, H_0 est équivalente aux contraintes suivantes :
- dans le paramétrage centré, $\alpha_1 = \cdots = \alpha_J = 0$;
- dans le paramétrage SAS, $a_1 = \cdots = a_J = 0$.

Dans tous les cas, le nombre de contraintes indépendantes, q, imposées par H_0 est égal à $J-1$.

Sous H_0, le modèle s'écrit $Y_{ij} = \mu + U_{ij}$ (il s'agit du modèle constant, ou modèle blanc), et l'on obtient :

$$\hat{\mu}^0 = \overline{y}_{\bullet\bullet} = \frac{1}{n}\sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}$$

(*attention :* si le plan est déséquilibré, $\hat{\mu}^0 \neq \hat{\mu}$).

On prend alors pour expression de la statistique du test de Fisher la quantité

$$f = \frac{\|\hat{y} - \hat{y}^0\|^2}{q\hat{\sigma}^2},$$

$\hat{y}$ et $\hat{y}^0$ désignant, dans $\mathbb{R}^n$, les vecteurs des valeurs prédites respectivement dans le modèle considéré (modèle avec le seul facteur F) et dans le modèle sous H_0 (modèle constant). Il vient (voir le 3.1.4)

$$f = \frac{1}{(J-1)\hat{\sigma}^2}\sum_{j=1}^{J} n_j(\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet})^2, \text{ avec } \hat{\sigma}^2 = \frac{1}{n-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_{\bullet j})^2.$$

On compare enfin cette statistique avec $f_{J-1\,;\,n-J}\,(1-\alpha)$, quantile d'ordre $1-\alpha$ d'une loi de Fisher à $J-1$ et $n-J$ d.d.l.

2.5.1 Synthèse des résultats précédents : le tableau d'analyse de la variance

Il est fréquent de résumer la construction de la statistique du test de Fisher au sein d'un tableau, appelé tableau d'analyse de la variance, qui se présente sous la forme ci-dessous (on notera que la plupart des logiciels fournissent ce tableau).

sources de variation	sommes des carrés	d.d.l.	carrés moyens	valeur de la statistique de Fisher
Facteur F	SSF	$J-1$	$MSF = \frac{SSF}{J-1}$	$\frac{MSF}{MSE}$
Erreur	SSE	$n-J$	$MSE = \frac{SSE}{n-J} = \hat{\sigma}^2$	—
Total	SST	$n-1$	—	—

Les sommes de carrés apparaissant ci-dessus sont définies de la façon suivante :

$$SSF = \sum_{j=1}^{J} n_j(\overline{y}_{\bullet j} - \overline{y}_{\bullet\bullet})^2;$$

on notera que dans le cas d'un plan équilibré (tous les n_j sont égaux à n_0 et, par suite, $\hat{\mu} = \overline{y}_{\bullet\bullet}$), on peut encore écrire : $SSF = n_0\sum_{j=1}^{J}(\hat{\alpha}_j)^2$;

$$SSE = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_{\bullet j})^2;$$

$$SST = SSF + SSE = \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \overline{y}_{\bullet\bullet})^2.$$

2.6 Autres tests

Si l'hypothèse nulle considérée ci-dessus a été rejetée, autrement dit si le modèle à un facteur est significatif, on peut être amené à tester d'autres hypothèses nulles, dans le but de simplifier le modèle. Par exemple, on peut considérer des hypothèses nulles du type : $\beta_j = \beta_{j'}$; $\alpha_j = 0$; $a_j = 0$.

Pour cela, on utilise en général un test de Student (rappelons que, lorsque $q = 1$, le test de Fisher est équivalent à un test de Student). En particulier, le

logiciel SAS propose, pour chaque valeur de j ($j = 1, \ldots, J-1$), le test de Student de l'hypothèse nulle $\{H_0 : \beta_j = \beta_J\}$; cela se fait avec la procédure GLM, au sein de la commande `model`, lorsqu'on rajoute l'option `solution`.

Avec des options spécifiques de la commande `model`, on peut aussi tester la significativité de l'écart entre deux niveaux quelconques du facteur F. Enfin, on peut utiliser la technique de Bonferroni (présentée en 2.2.9) pour construire des intervalles de confiance conjoints pour les écarts $\beta_j - \beta_{j'}$.

2.7 Illustration

2.7.1 Les données

Il s'agit d'un exemple fictif d'analyse de variance à un seul facteur. La variable réponse, en première colonne, prend des valeurs entières comprises entre 10 et 25. Le facteur est à trois niveaux, notés 1,2,3 et figure en seconde colonne. Il y a 9 individus observés, donc 9 lignes dans le fichier des données reproduit ci-après.

```
11 1
13 1
15 2
18 2
21 2
19 3
20 3
22 3
23 3
```

2.7.2 Le programme SAS

Le programme SAS ci-dessous permet de faire les principaux traitements (graphiques compris) dans le cadre d'une ANOVA à un seul facteur. Les commentaires permettent de discerner les différentes phases du programme.

```
* ------------------------------------------------- ;
* options facultatives pour la mise en page des sorties ;
* ------------------------------------------------- ;
options pagesize=64 linesize=76 nodate;
title;
footnote 'ANOVA 1 facteur - Exemple fictif';
* ------------------------------------------------- ;
*                lecture des donnees                ;
*     (le fichier "fic.don" contient les donnees    ;
*     et se trouve dans le repertoire de travail)   ;
* ------------------------------------------------- ;
data fic;
infile 'fic.don';
input y f;
run;
* ------------------------------------------------- ;
*            procedure GLM pour l'ANOVA             ;
* ------------------------------------------------- ;
proc glm data=fic;
class f;
model y = f / ss3;
run;
quit;
* ------------------------------------------------- ;
*        on relance avec l'option "solution"        ;
* ------------------------------------------------- ;
proc glm data=fic;
class f;
model y = f / ss3 solution;
run;
quit;
* ------------------------------------------------- ;
*         on relance avec la commande "output"      ;
*           pour archiver diverses quantites        ;
* ------------------------------------------------- ;
proc glm data=fic noprint;
class f;
model y = f;
output out=sortie p=yy r=uu stdr=erty student=rest;
proc print data=sortie;
run;
quit;
* ------------------------------------------------- ;
*    graphique valeurs predites vs valeurs observees ;
* ------------------------------------------------- ;
```

```
proc gplot data=sortie;
axis1 label=('valeurs observees')
      order=(10 to 25 by 5) length=7cm;
axis2 label=('valeurs' justify=right 'predites')
      order=(10 to 25 by 5) length=7cm;
symbol1 i=none v=dot;
symbol2 i=rl v=none;
plot yy*y y*y / haxis=axis1 vaxis=axis2
                hminor=4 vminor=4
                overlay;
run;
goptions reset=all;
quit;
* ------------------------------------------------------ ;
*                  graphique des residus                 ;
* ------------------------------------------------------ ;
proc gplot data=sortie;
axis1 label=('valeurs predites')
      order=(10 to 25 by 5) length=7cm;
axis2 label=('resisus' justify=right 'studentises')
      order=(-3 to 3 by 1) length=7cm;
symbol v=dot;
plot rest*yy / haxis=axis1 vaxis=axis2
               hminor=4 vminor=0
               vref=-2 vref=0 vref=2;
run;
goptions reset=all;
quit;
```

2.7.3 Les sorties de la procédure GLM

PAGE 1 The GLM Procedure

Class Level Information

Class	Levels	Values
f	3	1 2 3

Number of observations 9

PAGE 2 The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	108.0000000	54.0000000	10.80	0.0103
Error	6	30.0000000	5.0000000		
Corrected Total	8	138.0000000			

R-Square	Coeff Var	Root MSE	y Mean
0.782609	12.42260	2.236068	18.00000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f	2	108.0000000	54.0000000	10.80	0.0103

PAGE 3 The GLM Procedure

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		21.00000000 B	1.11803399	18.78	<.0001
f	1	-9.00000000 B	1.93649167	-4.65	0.0035
f	2	-3.00000000 B	1.70782513	-1.76	0.1295
f	3	0.00000000 B	.	.	.

PAGE 4

Obs	y	f	yy	uu	erty	rest
1	11	1	12	-1	1.58114	-0.63246
2	13	1	12	1	1.58114	0.63246
3	15	2	18	-3	1.82574	-1.64317
4	18	2	18	0	1.82574	0.00000
5	21	2	18	3	1.82574	1.64317
6	19	3	21	-2	1.93649	-1.03280
7	20	3	21	-1	1.93649	-0.51640
8	22	3	21	1	1.93649	0.51640
9	23	3	21	2	1.93649	1.03280

2.7.4 Estimation des paramètres

L'option `solution` de la commande `model` de la procédure GLM permet d'obtenir l'estimation des paramètres du modèle correspondant. Ici, SAS nous fournit les valeurs de $\hat{m}$ (21), de $\hat{a}_1$ (-9) et de $\hat{a}_2$ (-3). On en déduit les estimations des paramètres β_j ($\hat{\beta}_1 = -9 + 21 = 12$; $\hat{\beta}_2 = -3 + 21 = 18$; $\hat{\beta}_3 = 21$), donc des valeurs prédites (`yy`) et des résidus (`uu`).

2.7.5 La commande output

Dans le fichier obtenu avec l'option `out=` de la commande `output` de la procédure GLM (ici, il s'appelle "sortie"), on récupère les valeurs prédites (`p=`), les résidus (`r=`), les erreurs-types des résidus (`stdr=`) et les résidus studentisés (`student=`) (on laisse le soin au lecteur de retrouver toutes ces valeurs).

2.7.6 Les graphiques

Les figures 3.1 et 3.2 donnent les graphiques tels qu'on les obtient en sortie de SAS.

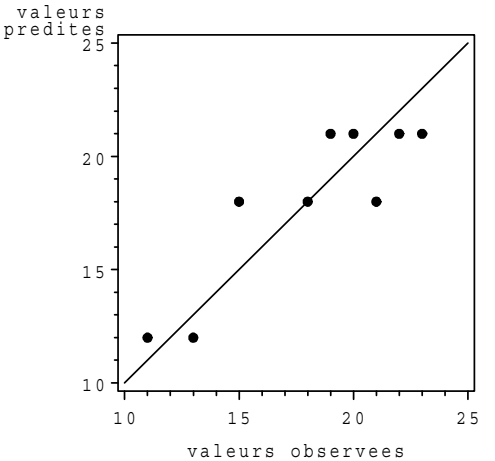

FIGURE 1 – *Graphique valeurs prédites vs valeurs observées.*

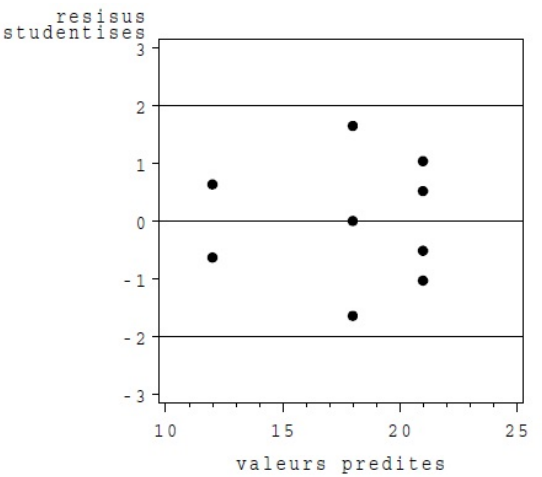

FIGURE 2 – *Graphique des résidus.*

3 Cas de deux facteurs croisés

3.1 Notations

- Le premier facteur est noté F_1 et son nombre de niveaux est noté J ($J \geq 2$) ; ces niveaux seront indicés par j.
- Le second facteur est noté F_2 et son nombre de niveaux est noté K ($K \geq 2$) ; les niveaux de F_2 seront indicés par k.
- Les deux facteurs sont croisés, c'est-à-dire d'une part qu'ils sont symétriques (on peut permuter leurs rôles), d'autre part qu'on réalise des observations pour chacun des croisements (j, k) ; on notera n_{jk} le nombre d'observations réalisées pour le croisement (j, k), le nombre total d'observations étant noté n, de sorte que l'on aura : $n = \sum_{j=1}^{J} \sum_{k=1}^{K} n_{jk}$. Lorsque $n_{jk} = n_0 \ \forall (j, k)$, on dit que le plan est **équilibré** ; sinon, on dit qu'il est **déséquilibré**.
- Chaque croisement (j, k) est en général appelé une *cellule* du plan factoriel.
- Les observations de la variable réponse Y, réalisées au sein de chaque

cellule, seront supposées indépendantes et seront notées y_{ijk}, $i = 1, \ldots, n_{jk}$; elles seront également supposées indépendantes sur l'ensemble des cellules.

Remarque. — On supposera, dans tout ce chapitre, $n_{jk} \geq 1$, autrement dit toute cellule est observée au moins une fois ; on dit, dans ce cas, que le plan est **complet**. Lorsqu'au moins une cellule est telle que $n_{jk} = 0$, on dit que le plan est **incomplet**. Le cas des plans incomplets sera abordé au chapitre 4.

3.2 Écriture initiale du modèle

On écrit chaque v.a.r. sous la forme :

$$Y_{ijk} = \beta_{jk} + U_{ijk}.$$

Matriciellement, cela peut s'écrire :

$$Y = \mathbf{X}\beta + U.$$

Y et U sont deux vecteurs de $\mathbb{R}^n$, β est un vecteur de $\mathbb{R}^{JK}$ (ici, $p = JK$) et $\mathbf{X}$, matrice d'incidence, est de dimension $n \times JK$.

La matrice $\mathbf{X}$ ne contient que des 0 et des 1, ses colonnes étant constituées des indicatrices Z^{jk} des cellules. Elle est de rang JK et ses colonnes sont deux à deux orthogonales. Les éléments β_{jk} du vecteur β sont les paramètres inconnus du modèle, à estimer. Enfin, on suppose toujours $U_{ijk} \sim \mathcal{N}(0, \sigma^2)$, les U_{ijk} étant i.i.d., le paramètre σ^2 étant lui aussi inconnu et à estimer. On a donc $Y_{ijk} \sim \mathcal{N}(\beta_{jk}, \sigma^2)$, les Y_{ijk} étant indépendantes.

EXEMPLE 4 *Considérons le cas très simple $J = 2$, $K = 3$ et $n_0 = 1$ (une seule observation dans chacune des 6 cellules). Dans ce cas, la matrice d'incidence $\mathbf{X}$ est tout simplement égale à la matrice identité $\mathbf{I_6}$.*

3.3 Paramétrage centré

On introduit tout d'abord les quantités suivantes :
- $\beta_{j\bullet} = \frac{1}{K}\sum_{k=1}^{K} \beta_{jk}$; c'est la valeur moyenne des paramètres au niveau j de F_1 ;
- $\beta_{\bullet k} = \frac{1}{J}\sum_{j=1}^{J} \beta_{jk}$; valeur moyenne des paramètres au niveau k de F_2 ;
- $\beta_{\bullet\bullet} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} \beta_{jk}$; valeur moyenne générale.

On notera que les différentes moyennes définies ci-dessus sont toujours des moyennes "non pondé-rées".

On définit ensuite les paramètres centrés de la façon suivante :
- $\mu = \beta_{\bullet\bullet}$: c'est l'*effet général*, ou effet moyen général ;
- $\alpha_j^1 = \beta_{j\bullet} - \beta_{\bullet\bullet}$: *effet principal*, ou différentiel, du niveau j de F_1 ;
- $\alpha_k^2 = \beta_{\bullet k} - \beta_{\bullet\bullet}$: effet principal, ou différentiel, du niveau k de F_2 ;
- $\gamma_{jk} = \beta_{jk} - \mu - \alpha_j^1 - \alpha_k^2 = \beta_{jk} - \beta_{j\bullet} - \beta_{\bullet k} + \beta_{\bullet\bullet}$: *effet d'interaction* des niveaux j de F_1 et k de F_2.

On vérifie, de façon immédiate, les relations de centrage suivantes :

$$\sum_{j=1}^{J} \alpha_j^1 = \sum_{k=1}^{K} \alpha_k^2 = 0\,;\ \sum_{j=1}^{J} \gamma_{jk} = 0,\ \forall k = 1, \ldots, K\,;\ \sum_{k=1}^{K} \gamma_{jk} = 0,\ \forall j = 1, \ldots, J.$$

Finalement, on peut réécrire le modèle sous la forme suivante :

$$Y_{ijk} = \beta_{jk} + U_{ijk} = \mu + \alpha_j^1 + \alpha_k^2 + \gamma_{jk} + U_{ijk}.$$

D'autre part, un raisonnement analogue à celui fait en 3.1.2, mais un peu plus lourd (il nécessite l'usage des indicatrices Z_1^j pour les niveaux de F_1 et Z_2^k pour les niveaux de F_2), nous permet maintenant d'écrire :

$$\begin{aligned} Y = \mu\mathbb{1}_n &+ \sum_{j=1}^{J-1} \alpha_j^1 (Z_1^j - Z_1^J) + \sum_{k=1}^{K-1} \alpha_k^2 (Z_2^k - Z_2^K) \\ &+ \sum_{j=1}^{J-1}\sum_{k=1}^{K-1} \gamma_{jk}(Z_1^j - Z_1^J)(Z_2^k - Z_2^K) + U. \end{aligned}$$

On retrouve ainsi un modèle de régression linéaire avec coefficient constant.

EXEMPLE 5 *Reprenons le cas de l'exemple 4. Sous forme matricielle, on écrira le modèle $Y = \mathbf{X}_c\beta_c + U$, avec :*

$$\mathbf{X}_c = \left(\begin{array}{c|c|cc|cc} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{array}\right) \quad ; \quad \beta_c = \begin{pmatrix} \mu \\ \alpha_1^1 \\ \alpha_1^2 \\ \alpha_2^2 \\ \gamma_{11} \\ \gamma_{12} \end{pmatrix}.$$

Le plan étant équilibré, deux colonnes de $\mathbf{X}_c$ extraites de deux blocs différents sont toujours orthogonales. Par ailleurs, toujours parce que le plan est équilibré, toutes les colonnes, à l'exception de la première, sont centrées.

Remarque. — Le paramétrage centré fait intervenir un paramètre μ, $(J-1)$ paramètres α_j^1 indépendants, $(K-1)$ paramètres α_k^2 indépendants et $(J-1)(K-1)$ paramètres γ_{jk} indépendants. Au total, il y a bien JK paramètres indépendants, comme dans le paramétrage initial en β_{jk}.

Remarque. — Considérons maintenant deux indices distincts j et j', quelconques mais fixés. On peut écrire :

$$\beta_{jk} - \beta_{j'k} = (\alpha_j^1 - \alpha_{j'}^1) + (\gamma_{jk} - \gamma_{j'k}),\ \forall k = 1, \ldots, K.$$

On remarque que le premier terme est indépendant de k et que le second disparaît dans un modèle sans interaction. D'où l'idée de réaliser un graphique avec en abscisses les différents indices k, en ordonnées les valeurs moyennes $\overline{y}_{\bullet jk}$ (estimations des β_{jk}, voir plus loin) et une "courbe" pour chaque indice j (courbes superposées). Si ces courbes sont sensiblement parallèles, on peut négliger les effets d'interactions dans le modèle considéré (voir les figures 3.5 et 3.6).

3.4 Paramétrage SAS

On réécrit maintenant :

$$\begin{aligned}\beta_{jk} &= \beta_{JK} + (\beta_{jK} - \beta_{JK}) + (\beta_{Jk} - \beta_{JK}) + (\beta_{jk} - \beta_{jK} - \beta_{Jk} + \beta_{JK}) \\ &= m + a_j^1 + a_k^2 + c_{jk}.\end{aligned}$$

Les paramètres définis ci-dessus vérifient ainsi les relations :

$$a_J^1 = a_K^2 = 0\ ;\ \ c_{jK} = 0,\ \forall j = 1, \ldots, J\ ;\ \ c_{Jk} = 0,\ \forall k = 1, \ldots, K.$$

Bien sûr, il y a toujours JK paramètres indépendants dans ce nouveau paramétrage.

On peut encore vérifier que l'on obtient maintenant

$$Y = m\mathbb{1}_n + \sum_{j=1}^{J-1} a_j^1 Z_1^j + \sum_{k=1}^{K-1} a_k^2 Z_2^k + \sum_{j=1}^{J-1}\sum_{k=1}^{K-1} c_{jk} Z_1^j Z_2^k + U,$$

m étant toujours appelé *intercept* dans SAS.

EXEMPLE 6 *Reprenons une dernière fois l'exemple 4. Sous forme matricielle, le modèle s'écrit maintenant $Y = \mathbf{X}_s\beta_s + U$, avec :*

$$\mathbf{X}_s = \left(\begin{array}{c|c|cc|cc} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right)\ ;\quad \beta_s = \begin{pmatrix} m \\ a_1^1 \\ a_1^2 \\ a_2^2 \\ c_{11} \\ c_{12} \end{pmatrix}.$$

On notera que les colonnes de $\mathbf{X}_s$ ne sont ni centrées ni orthogonales.

3.5 Estimation des paramètres

3.5.1 Paramétrage initial

À partir des résultats généraux relatifs au modèle linéaire, on déduit :

$$\hat{\beta}_{jk} = \overline{y}_{\bullet jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} y_{ijk}.$$

On sait que l'on a

$$\hat{B}_{jk} \sim \mathcal{N}(\beta_{jk}, \frac{\sigma^2}{n_{jk}}),$$

les différents estimateurs étant indépendants. Enfin, l'erreur-type de $\hat{B}_{jk}$ est $\frac{\hat{\sigma}}{\sqrt{n_{jk}}}$, ce qui permet de construire un intervalle de confiance pour β_{jk}.

3.5.2 Paramétrage centré

Les estimations des paramètres se déduisent dans ce cas de celles ci-dessus. Il vient :

- $\hat{\mu} = \hat{\beta}_{\bullet\bullet} = \dfrac{1}{JK} \displaystyle\sum_{j=1}^{J}\sum_{k=1}^{K} \overline{y}_{\bullet jk}$. Comme dans le cas d'un seul facteur, on notera que $\hat{\mu}$ n'est égal à la moyenne générale des observations de Y que si le plan est équilibré.
- $\hat{\alpha}_j^1 = \hat{\beta}_{j\bullet} - \hat{\beta}_{\bullet\bullet} = \dfrac{1}{K} \displaystyle\sum_{k=1}^{K} \overline{y}_{\bullet jk} - \hat{\mu}$ (les $\hat{\alpha}_j^1$ sont centrés selon j).

– $\hat{\alpha}_k^2 = \hat{\beta}_{\bullet k} - \hat{\beta}_{\bullet\bullet} = \frac{1}{J}\sum_{j=1}^{J} \overline{y}_{\bullet jk} - \hat{\mu}$ (les $\hat{\alpha}_k^2$ sont centrés selon k).

– $\hat{\gamma}_{jk} = \hat{\beta}_{jk} - \hat{\alpha}_j^1 - \hat{\alpha}_k^2 - \hat{\mu} = \hat{\beta}_{jk} - \hat{\beta}_{j\bullet} - \hat{\beta}_{\bullet k} + \hat{\beta}_{\bullet\bullet}$
$= \overline{y}_{\bullet jk} - \frac{1}{K}\sum_{k=1}^{K} \overline{y}_{\bullet jk} - \frac{1}{J}\sum_{j=1}^{J} \overline{y}_{\bullet jk} + \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K} \overline{y}_{\bullet jk}$
(les $\hat{\gamma}_{jk}$ sont centrés selon j et selon k).

3.5.3 Paramétrage SAS

De la même manière, on obtient maintenant :

– $\hat{m} = \overline{y}_{\bullet JK}$;
– $\hat{a}_j^1 = \overline{y}_{\bullet jK} - \overline{y}_{\bullet JK}$ $(\hat{a}_J^1 = 0)$;
– $\hat{a}_k^2 = \overline{y}_{\bullet Jk} - \overline{y}_{\bullet JK}$ $(\hat{a}_K^2 = 0)$;
– $\hat{c}_{jk} = \overline{y}_{\bullet jk} - \overline{y}_{\bullet jK} - \overline{y}_{\bullet Jk} + \overline{y}_{\bullet JK}$
$(\hat{c}_{jk} = 0$, dès que l'un au moins des deux indices est maximum).

3.5.4 Valeurs prédites et résidus

De façon standard, les valeurs prédites $\hat{y}_{ijk}$ valent $\hat{\beta}_{jk}$ $(= \overline{y}_{\bullet jk})$ et les résidus $\hat{u}_{ijk}$ valent $y_{ijk} - \overline{y}_{\bullet jk}$. L'erreur-type de $\hat{y}_{ijk}$ vaut $\frac{\hat{\sigma}}{\sqrt{n_{jk}}}$ et celle de $\hat{u}_{ijk}$ vaut $\hat{\sigma}\sqrt{\frac{n_{jk}-1}{n_{jk}}}$. On notera que ces expressions (indépendantes du paramétrage choisi) ne sont plus valables avec un modèle autre que le modèle complet.

3.5.5 Variance

L'estimation de la variance σ^2 est la suivante :

$$\hat{\sigma}^2 = \frac{1}{n-JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\hat{u}_{ijk})^2 = \frac{1}{n-JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(y_{ijk} - \overline{y}_{\bullet jk})^2.$$

3.6 Tests d'hypothèses

Dans un modèle à deux facteurs croisés, trois hypothèses nulles peuvent être envisagées afin de simplifier le modèle considéré.

3.6.1 Hypothèse H_0 : absence d'effet des interactions

Cette hypothèse peut prendre les différentes formes suivantes, équivalentes :

$$\begin{aligned} H_0 &\iff \gamma_{jk} = 0, \forall (j,k) \iff c_{jk} = 0, \forall (j,k) \\ &\iff \beta_{jk} - \beta_{j'k} \text{est indépendant de} k, \forall (j,j') \\ &\iff \beta_{jk} - \beta_{jk'} \text{est indépendant de} j, \forall (k,k'). \end{aligned}$$

Ainsi, avec le paramétrage centré, H_0 conduit à réécrire le modèle sous la forme :

$$Y_{ijk} = \mu + \alpha_j^1 + \alpha_k^2 + U_{ijk}^0,$$

que l'on appelle le **modèle additif**.

La valeur de la statistique du test de H_0 est

$$f = \frac{1}{(J-1)(K-1)\hat{\sigma}^2}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_{jk}}(\hat{y}_{ijk} - \hat{y}_{ijk}^0)^2,$$

où $\hat{y}_{ijk}^0$ est la valeur prédite de y_{ijk} dans le modèle additif. Cette statistique est à comparer avec le quantile $f_{(J-1)(K-1)\ ;\ n-JK}\ (1-\alpha)$.

3.6.2 Hypothèse H_0' : absence d'effet du facteur F_1

Cette autre hypothèse peut prendre les différentes formes suivantes :

$$\begin{aligned} H_0' &\iff \beta_{1\bullet} = \cdots = \beta_{J\bullet} \iff \alpha_j^1 = 0, \forall j = 1, \ldots, J \\ &\iff a_j^1 = 0, \forall j = 1, \ldots, J. \end{aligned}$$

Convention

Dans tout ce cours, nous ferons les tests de façon hiérarchique, c'est-à-dire que, dans le cas présent, nous ne testerons l'hypothèse H_0' que si, au préalable, l'hypothèse H_0 n'a pas été rejetée ; ainsi, le modèle de référence pour tester H_0' sera le modèle additif. Cette façon de procéder n'est pas universelle, mais elle a le mérite d'être cohérente et simple à interpréter.

Sous H_0', le modèle s'écrit donc :

$$Y_{ijk} = \mu + \alpha_k^2 + U_{ijk}^{0'}.$$

La valeur de la statistique du test est maintenant

$$f' = \frac{1}{(J-1)(\hat{\sigma}^0)^2} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} (\hat{y}^0_{ijk} - \hat{y}^{0'}_{ijk})^2,$$

où $(\hat{\sigma}^0)^2$ désigne l'estimation de σ^2 dans le modèle additif,

$$(\hat{\sigma}^0)^2 = \frac{1}{n-(J+K-1)} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_{jk}} (y_{ijk} - \hat{y}^0_{ijk})^2,$$

et où $\hat{y}^{0'}_{ijk}$ est la valeur prédite de y_{ijk} dans le modèle avec F_2 comme seul facteur. La valeur f' est à comparer au quantile $f_{J-1\ ;\ n-(J+K-1)}\ (1-\alpha)$.

3.6.3 Hypothèse H_0'' : absence d'effet du facteur F_2

Cette dernière hypothèse peut prendre les différentes formes suivantes :

$$\begin{aligned} H_0'' &\iff \beta_{\bullet 1} = \cdots = \beta_{\bullet K} \iff \alpha_k^2 = 0, \forall k = 1, \ldots, K \\ &\iff a_k^2 = 0, \forall k = 1, \ldots, K. \end{aligned}$$

On raisonne alors de façon symétrique par rapport à ce qui a été fait au point précédent.

Remarque. — Le test de Fisher est également utilisé pour tester la significativité du modèle finalement retenu, autrement dit pour tester le modèle constant contre ce modèle.

Remarque. — Si l'on choisit le modèle additif pour un jeu de données, les estimations des paramètres β_{jk} avec le paramétrage centré sont obtenues directement en posant $\hat{\beta}_{jk} = \hat{\mu} + \hat{\alpha}_j^1 + \hat{\alpha}_k^2$, où les estimations $\hat{\mu}$, $\hat{\alpha}_j^1$ et $\hat{\alpha}_k^2$ sont celles obtenues dans le modèle complet. Il suffit donc d'annuler les estimations des paramètres d'interactions pour trouver, à partir du modèle complet, les nouvelles estimations des β_{jk}. Par contre, il en va différemment avec le paramétrage SAS : la matrice d'incidence $\mathbf{X}_s$ doit être modifiée (par suppression des colonnes relatives aux interactions), on doit ensuite calculer $\hat{\beta}_s = (\mathbf{X}_s'\mathbf{X}_s)^{-1}\mathbf{X}_s'y$ et en déduire les nouvelles estimations des paramètres m, a_j^1 et a_k^2. Ainsi, le paramétrage centré apparaît comme plus naturel que le paramétrage SAS (ou que tout autre paramétrage).

3.7 Cas particulier d'un plan équilibré

On dit qu'un plan à deux facteurs croisés est équilibré s'il vérifie $n_{jk} = n_0, \forall (j,k)$. Dans ce cas, $n = n_0 JK$ et diverses écritures se simplifient. On obtient ainsi :

$$\hat{\beta}_{jk} = \overline{y}_{\bullet jk} = \frac{1}{n_0} \sum_{i=1}^{n_0} y_{ijk};$$

$$\hat{\beta}_{j\bullet} = \frac{1}{K} \sum_{k=1}^{K} \overline{y}_{\bullet jk} = \frac{1}{n_0 K} \sum_{k=1}^{K} \sum_{i=1}^{n_0} y_{ijk} = \overline{y}_{\bullet j\bullet}$$

(moyenne de toutes les observations de Y au niveau j du facteur F_1) ;

$$\hat{\beta}_{\bullet K} = \overline{y}_{\bullet\bullet k} \text{ (même chose) ;}$$

$$\hat{\beta}_{\bullet\bullet} = \frac{1}{JK} \sum_{j=1}^{J} \sum_{k=1}^{K} \overline{y}_{\bullet jk} = \frac{1}{n} \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{n_0} y_{ijk} = \overline{y}_{\bullet\bullet\bullet}$$

(moyenne générale de toutes les observations de Y).

Les calculs des statistiques des tests vus précédemment peuvent encore se synthétiser, dans ce cas, sous la forme d'un tableau d'analyse de la variance (voir plus loin).

Les sommes de carrés apparaissant dans ce tableau sont définies de la façon suivante :

TABLE 1 – **Tableau d'analyse de la variance (cas équilibré)**

sources de variation	sommes des carrés	d.d.l.	carrés moyens	valeurs des statistiques de Fisher
F_1	SSF_1	$J-1$	$MSF_1 = \frac{SSF_1}{J-1}$	$\frac{MSF_1}{MSE}$
F_2	SSF_2	$K-1$	$MSF_2 = \frac{SSF_2}{K-1}$	$\frac{MSF_2}{MSE}$
$F_1 * F_2$	SSF_{12}	$(J-1)(K-1)$	$MSF_{12} = \frac{SSF_{12}}{(J-1)(K-1)}$	$\frac{MSF_{12}}{MSE}$
Erreur	SSE	$n-JK$	$MSE = \frac{SSE}{n-JK} = \hat{\sigma}^2$	—
Total	SST	$n-1$	—	—

$$
\begin{aligned}
SSF_1 &= n_0\, K \sum_{j=1}^{J} (\overline{y}_{\bullet j \bullet} - \overline{y}_{\bullet\bullet\bullet})^2 = n_0\, K \sum_{j=1}^{J} (\hat{\alpha}_j^1)^2 \; ; \\
SSF_2 &= n_0\, J \sum_{k=1}^{K} (\overline{y}_{\bullet\bullet k} - \overline{y}_{\bullet\bullet\bullet})^2 = n_0\, J \sum_{k=1}^{K} (\hat{\alpha}_k^2)^2 \; ; \\
SSF_{12} &= n_0 \sum_{j=1}^{J}\sum_{k=1}^{K} (\overline{y}_{\bullet jk} - \overline{y}_{\bullet j \bullet} - \overline{y}_{\bullet\bullet k} + \overline{y}_{\bullet\bullet\bullet})^2 = n_0 \sum_{j=1}^{J}\sum_{k=1}^{K} (\hat{\gamma}_{jk})^2 \; ; \\
SSE &= \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_0} (y_{ijk} - \overline{y}_{\bullet jk})^2 \; ; \\
SST &= \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{n_0} (y_{ijk} - \overline{y}_{\bullet\bullet\bullet})^2 .
\end{aligned}
$$

Remarque. — Dans le tableau ci-dessu, les tests de significativité des facteurs F_1 et F_2 sont faits avec, pour modèle de référence, le modèle *complet* (c'est-à-dire, le modèle comportant tous les effets initialement introduits ; on l'appelle encore le modèle *plein*). Si l'hypothèse de nullité des interactions n'est pas rejetée et qu'on souhaite faire ces tests (significativité de chaque facteur) avec le modèle additif comme référence, au dénominateur de la statistique de Fisher, on doit remplacer MSE par $\frac{SSE + SSF_{12}}{n-J-K+1} = (\hat{\sigma}^0)^2$, estimation de σ^2 dans le modèle additif.

Remarque. — Dans le cas déséquilibré, on ne peut pas construire de tableau analogue. En effet, le développement de SST (la somme des carrés totale) en fonction des autres sommes de carrés fait aussi intervenir dans ce cas les doubles produits (qui sont nuls dans le cas équilibré). Ces doubles produits ne pouvant pas être affectés à un effet spécifique, le tableau d'analyse de la variance n'a plus de raison d'être dans ce cas.

Remarque. — La définition même des sommes de carrés n'est plus très claire dans le cas déséqui-libré. Cela conduit à introduire diverses sommes de carrés (trois), appelées de type I, de type II et de type III (toutes égales dans le cas équilibré). De plus, des sommes de carrés spécifiques au cas des plans incomplets existent également et sont appelées sommes de type IV. On trouvera en Annexe B quelques précisions sur les trois premiers types de sommes de carrés. Sauf indication contraire, il est recommandé d'utiliser les sommes de type III.

3.8 Illustration

3.8.1 Les données

Il s'agit d'un célèbre exemple (fictif) d'analyse de variance à deux facteurs croisés. La variable réponse, en dernière colonne, est le rendement laitier mesuré sur un échantillon de 40 vaches laitières de la même espèce. Il y a deux facteurs contrôlés, tous deux liés à l'alimentation des vaches : la dose, en première colonne, à 2 niveaux (1 = dose faible, 2 = dose forte) ; le régime alimentaire, en deuxième colonne, à 4 niveaux (1 = paille, 2 = foin, 3 = herbe, 4 = aliments ensilés). Pour chaque dose et chaque régime (8 cellules), on a observé le rendement de 5 vaches. On a donc affaire à un plan complet, équilibré, avec 5 répétitions. Les données sont reproduites ci-dessous.

```
1   1    8        2   1    8
1   1   11        2   1    9
```

```
1   1   11      2   1    8
1   1   10      2   1   10
1   1    7      2   1    9
1   2   12      2   2   10
1   2   13      2   2    7
1   2   14      2   2   10
1   2   11      2   2   12
1   2   10      2   2   11
1   3   10      2   3   11
1   3   12      2   3    9
1   3   12      2   3   11
1   3   13      2   3   11
1   3   14      2   3   12
1   4   17      2   4   17
1   4   13      2   4   19
1   4   17      2   4   17
1   4   14      2   4   16
1   4   13      2   4   21
```

3.8.2 Le programme SAS

Le programme ci-dessous réalise la procédure GLM sur les données des vaches laitières, trace les deux graphiques de contrôle du modèle choisi (le modèle complet), puis les deux graphiques des interactions. On trouvera des compléments sur cet exemple en Annexe A.

```
options pagesize=64 linesize=76 nodate;
title;
footnote 'ANOVA 2 facteurs - vaches laitieres';
* ------------------------------------------------ ;
data vach;
infile 'vach.don';
input f1 f2 y;
run;
* ------------------------------------------------ ;
proc glm;
class f1 f2;
model y = f1 f2 f1*f2 / ss3 solution;
output out=sortie p=yy r=uu stdr=erty student=rest;
lsmeans f1 f2 f1*f2 / out=graph;
run;
quit;
* ------------------------------------------------ ;
*          graphiques de controle du modele        ;
* ------------------------------------------------ ;
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'vach1.eps';
goptions colors=(black) hsize=13cm vsize=10cm;
proc gplot data=sortie;
axis1 label=('valeurs observees') order=(6 to 22 by 2)
      minor=none length=7cm;
axis2 label=('valeurs' justify=right 'predites')
      order=(6 to 22 by 2) minor=none length=7cm;
symbol1 v=dot i=none;
symbol2 v=none i=rl;
plot yy*y y*y / haxis=axis1 vaxis=axis2 overlay;
run;
goptions reset=all;
quit;
* ------------------------------------------------ ;
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'vach2.eps';
goptions colors=(black) hsize=13cm vsize=10cm;
proc gplot data=sortie;
axis1 label=('valeurs predites')
      order=(6 to 20 by 2) minor=none length=7cm;
axis2 label=('resisus' justify=right 'studentises')
      order=(-3 to 3 by 1) minor=none length=7cm;
symbol v=dot;
plot rest*yy / haxis=axis1 vaxis=axis2
               vref=-2 vref=0 vref=2;
run;
goptions reset=all;
quit;
* ------------------------------------------------ ;
*              graphiques des interactions         ;
* ------------------------------------------------ ;
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'vach3.eps';
```

```
goptions colors=(black) hsize=13cm vsize=10cm;
proc gplot data=graph;
axis1 label=('premier facteur') order=(1 to 2 by 1)
      minor=none length=6cm;
axis2 label=('moyenne' justify=right 'des effets')
      order=(8 to 20 by 2) minor=none length=6cm;
symbol1 i=join v=dot;
symbol2 i=join v=triangle;
symbol3 i=join v=circle;
symbol4 i=join v=#;
symbol5 i=join v=%;
plot lsmean*f1=f2 / haxis=axis1 vaxis=axis2;
run;
goptions reset=all;
quit;
* ------------------------------------------------ ;
goptions device=psepsf gend='0a'x gaccess=gsasfile;
filename gsasfile 'vach4.eps';
goptions colors=(black) hsize=13cm vsize=10cm;
proc gplot data=graph;
axis1 label=('second facteur') order=(1 to 4 by 1)
      minor=none length=6cm;
axis2 label=('moyenne' justify=right 'des effets')
      order=(8 to 20 by 2) minor=none length=6cm;
symbol1 i=join v=dot;
symbol2 i=join v=triangle;
symbol3 i=join v=circle;
plot lsmean*f2=f1 / haxis=axis1 vaxis=axis2;
run;
goptions reset=all;
quit;
```

3.8.3 Les sorties de la procédure GLM

PAGE 1

The GLM Procedure

Class Level Information

Class	Levels	Values
f1	2	1 2
f2	4	1 2 3 4

Number of observations 40

PAGE 2

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	331.6000000	47.3714286	17.54	<.0001
Error	32	86.4000000	2.7000000		
Corrected Total	39	418.0000000			

R-Square	Coeff Var	Root MSE	y Mean
0.793301	13.69306	1.643168	12.00000

Source	DF	Type III SS	Mean Square	F Value	Pr > F
f1	1	0.4000000	0.4000000	0.15	0.7029
f2	3	290.2000000	96.7333333	35.83	<.0001
f1*f2	3	41.0000000	13.6666667	5.06	0.0056

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		18.00000000 B		0.73484692	24.49	<.0001
f1	1	-3.20000000 B		1.03923048	-3.08	0.0042
f1	2	0.00000000 B		.	.	.
f2	1	-9.20000000 B		1.03923048	-8.85	<.0001
f2	2	-8.00000000 B		1.03923048	-7.70	<.0001
f2	3	-7.20000000 B		1.03923048	-6.93	<.0001
f2	4	0.00000000 B		.	.	.
f1*f2	1	1	3.80000000 B	1.46969385	2.59	0.0145
f1*f2	1	2	5.20000000 B	1.46969385	3.54	0.0013
f1*f2	1	3	4.60000000 B	1.46969385	3.13	0.0037
f1*f2	1	4	0.00000000 B	.	.	.
f1*f2	2	1	0.00000000 B	.	.	.
f1*f2	2	2	0.00000000 B	.	.	.
f1*f2	2	3	0.00000000 B	.	.	.
f1*f2	2	4	0.00000000 B	.	.	.

PAGE 3

Least Squares Means

f1	y LSMEAN

```
1       12.1000000
2       11.9000000


f2         y LSMEAN

1        9.1000000
2       11.0000000
3       11.5000000
4       16.4000000


f1     f2         y LSMEAN

1      1        9.4000000
1      2       12.0000000
1      3       12.2000000
1      4       14.8000000
2      1        8.8000000
2      2       10.0000000
2      3       10.8000000
2      4       18.0000000
```

3.8.4 Commentaires

La commande `lsmeans` de la procédure GLM permet, d'une part, d'obtenir en sortie certaines moyennes des y_{ijk} (ici, selon les niveaux de chaque facteur, puis selon les cellules), d'autre part, de récupérer la table SAS, ici appelée `graph`, qui permet de réaliser les graphiques d'interactions.

3.8.5 Les graphiques

Le programme ci-dessus produit les quatre graphiques 3.3 à 3.6.

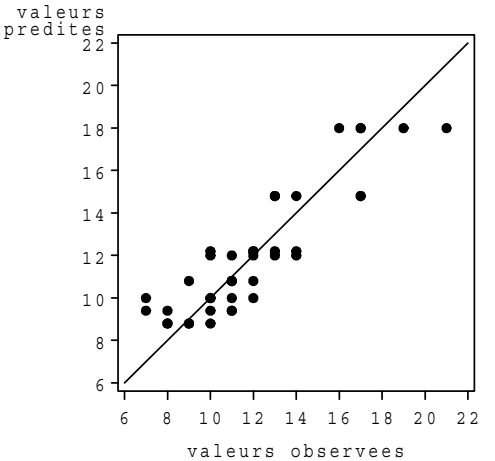

FIGURE 3 – *Graphique valeurs prédites vs valeurs observées.*

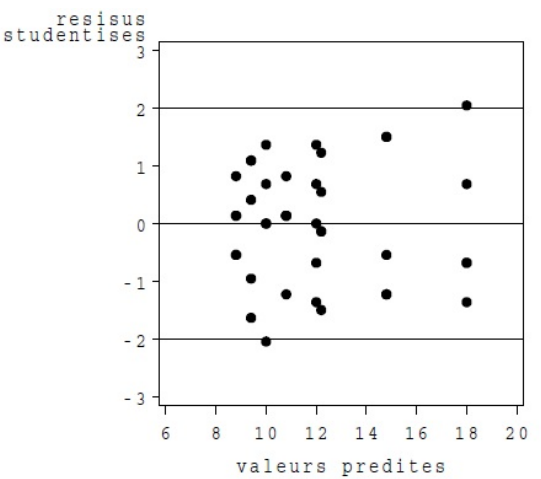

FIGURE 4 – *Graphique des résidus.*

4 Cas de trois facteurs croisés

4.1 Notations

- Les trois facteurs considérés sont notés F_1, F_2 et F_3.
- Le nombre de niveaux de F_1 est noté J, celui de F_2 est noté K et celui de F_3 est noté L ($J \geq 2$; $K \geq 2$; $L \geq 2$).
- Les niveaux de F_1 sont indicés par j, ceux de F_2 par k et ceux de F_3 par ℓ.
- Les trois facteurs étant croisés, on considère les JKL cellules (ou triplets) (j, k, ℓ).

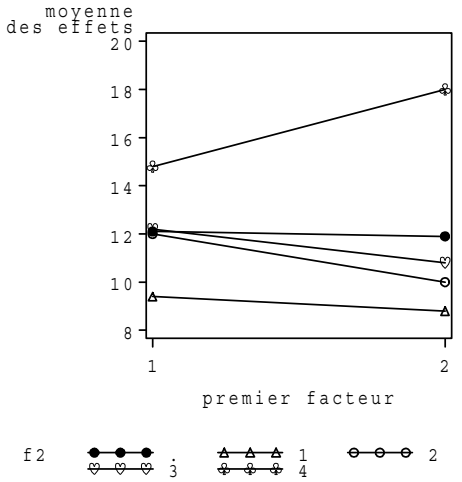

FIGURE 5 – *Premier graphique des interactions.*

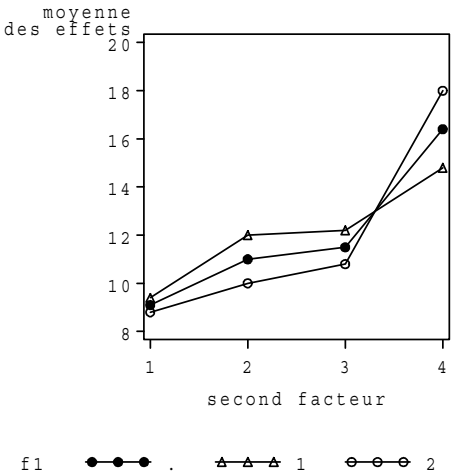

FIGURE 6 – *Second graphique des interactions.*

- Dans chaque cellule, on réalise $n_{jk\ell}$ observations de la variable à expliquer Y et on pose $n = \sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{\ell=1}^{L} n_{jk\ell}$. On suppose toujours que le plan est complet : $\forall\ (j,k,\ell), n_{jk\ell} \geq 1$; de plus, si $\forall\ (j,k,\ell), n_{jk\ell} = n_0$, alors le plan est équilibré.
- On notera $Y_{ijk\ell}$ $(i = 1, \ldots, n_{jk\ell})$ les v.a.r. associées aux observations de Y dans la cellule (j,k,ℓ).

4.2 Modèle

4.2.1 Paramétrage initial

Dans un premier temps, on écrit le modèle sous la forme

$$Y_{ijk\ell} = \beta_{jk\ell} + U_{ijk\ell},$$

avec toujours les mêmes hypothèses sur les v.a.r. $U_{ijk\ell}$ (elles sont i.i.d., $\mathcal{N}(0,\sigma^2)$). Il y a donc JKL paramètres $\beta_{jk\ell}$ indépendants à estimer, en plus de σ^2 (ici, $p = JKL$).

4.2.2 Paramétrage centré

Il fait intervenir toutes les moyennes partielles (non pondérées) des paramètres $\beta_{jk\ell}$. Définissons tout d'abord ces moyennes :

$$\beta_{\bullet k\ell} = \frac{1}{J}\sum_{j=1}^{J}\beta_{jk\ell}\ ;\ \beta_{j\bullet\ell} = \frac{1}{K}\sum_{k=1}^{K}\beta_{jk\ell}\ ;\ \beta_{jk\bullet} = \frac{1}{L}\sum_{\ell=1}^{L}\beta_{jk\ell}\ ;$$

$$\beta_{\bullet\bullet\ell} = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\beta_{jk\ell}\ ;\ \beta_{\bullet k\bullet} = \frac{1}{JL}\sum_{j=1}^{J}\sum_{\ell=1}^{L}\beta_{jk\ell}\ ;\ \beta_{j\bullet\bullet} = \frac{1}{KL}\sum_{k=1}^{K}\sum_{\ell=1}^{L}\beta_{jk\ell}\ ;$$

$$\beta_{\bullet\bullet\bullet} = \frac{1}{JKL}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{\ell=1}^{L}\beta_{jk\ell}.$$

On réécrit alors le modèle sous la forme :

$$Y_{ijk\ell} = \mu + \alpha_j^1 + \alpha_k^2 + \alpha_\ell^3 + \gamma_{jk}^{12} + \gamma_{j\ell}^{13} + \gamma_{k\ell}^{23} + \delta_{jk\ell} + U_{ijk\ell} = \beta_{jk\ell} + U_{ijk\ell}.$$

- Le paramètre μ est l'effet général (1 paramètre). Il est défini par : $\mu = \beta_{\bullet\bullet\bullet}$.

– Les paramètres α_j^1, α_k^2 et α_ℓ^3 sont les effets principaux associés aux différents niveaux de chacun des trois facteurs. Ils sont définis par les relations suivantes :

$$\alpha_j^1 = \beta_{j\bullet\bullet} - \beta_{\bullet\bullet\bullet}\ ;\ \ \alpha_k^2 = \beta_{\bullet k\bullet} - \beta_{\bullet\bullet\bullet}\ ;\ \ \alpha_\ell^3 = \beta_{\bullet\bullet\ell} - \beta_{\bullet\bullet\bullet}.$$

Ils vérifient :

$$\sum_{j=1}^{J} \alpha_j^1 = \sum_{k=1}^{K} \alpha_k^2 = \sum_{\ell=1}^{L} \alpha_\ell^3 = 0.$$

Il y a donc $(J-1)+(K-1)+(L-1)$ paramètres α indépendants.

– Les paramètres γ_{jk}^{12}, $\gamma_{j\ell}^{13}$ et $\gamma_{k\ell}^{23}$ sont les effets d'interactions d'ordre 2 (entre 2 facteurs). Ils sont définis par les relations suivantes :

$$\begin{aligned}
\gamma_{jk}^{12} &= \beta_{jk\bullet} - \beta_{j\bullet\bullet} - \beta_{\bullet k\bullet} + \beta_{\bullet\bullet\bullet}\ ;\\
\gamma_{j\ell}^{13} &= \beta_{j\bullet\ell} - \beta_{j\bullet\bullet} - \beta_{\bullet\bullet\ell} + \beta_{\bullet\bullet\bullet}\ ;\\
\gamma_{k\ell}^{23} &= \beta_{\bullet k\ell} - \beta_{\bullet k\bullet} - \beta_{\bullet\bullet\ell} + \beta_{\bullet\bullet\bullet}.
\end{aligned}$$

Ils vérifient :

$$\sum_{j=1}^{J} \gamma_{jk}^{12} = \sum_{k=1}^{K} \gamma_{jk}^{12} = \sum_{j=1}^{J} \gamma_{j\ell}^{13} = \sum_{\ell=1}^{L} \gamma_{j\ell}^{13} = \sum_{k=1}^{K} \gamma_{k\ell}^{23} = \sum_{\ell=1}^{L} \gamma_{k\ell}^{23} = 0.$$

Il y a donc $(J-1)(K-1)+(J-1)(L-1)+(K-1)(L-1)$ paramètres γ indépendants.

– Les paramètres $\delta_{jk\ell}$ sont les effets d'interactions d'ordre 3 (entre 3 facteurs). Ils sont définis par :

$$\delta_{jk\ell} = \beta_{jk\ell} - \beta_{\bullet k\ell} - \beta_{j\bullet\ell} - \beta_{jk\bullet} + \beta_{\bullet\bullet\ell} + \beta_{\bullet k\bullet} + \beta_{j\bullet\bullet} - \beta_{\bullet\bullet\bullet}.$$

Ils vérifient :

$$\sum_{j=1}^{J} \delta_{jk\ell} = \sum_{k=1}^{K} \delta_{jk\ell} = \sum_{\ell=1}^{L} \delta_{jk\ell} = 0.$$

Il y a donc $(J-1)(K-1)(L-1)$ paramètres δ indépendants.

– Au total, ce nouveau paramétrage fait intervenir

$$\begin{aligned}
1 \ &+\ (J-1)+(K-1)+(L-1)\\
&+\ (J-1)(K-1)+(J-1)(L-1)+(K-1)(L-1)\\
&+\ (J-1)(K-1)(L-1)\\
&=\ JKL
\end{aligned}$$

paramètres indépendants, ce qui est cohérent.

4.2.3 Paramétrage SAS

Le paramétrage SAS utilise toujours, dans ce cas, le principe consistant à prendre le dernier niveau de chaque facteur comme niveau de référence. Dans SAS, on réécrit le modèle sous la forme :

$$Y_{ijk\ell} = m + a_j^1 + a_k^2 + a_\ell^3 + c_{jk}^{12} + c_{j\ell}^{13} + c_{k\ell}^{23} + d_{jk\ell} + U_{ijk\ell} = \beta_{jk\ell} + U_{ijk\ell}.$$

De façon "logique" (compte tenu de ce qui a déjà été fait avec deux facteurs croisés), les paramètres sont définis de la manière suivante :

$$\begin{aligned}
m &= \beta_{JKL}\ ;\\
a_j^1 &= \beta_{jKL} - \beta_{JKL}\ ;\ \ a_k^2 = \beta_{JkL} - \beta_{JKL}\ ;\ \ a_\ell^3 = \beta_{JK\ell} - \beta_{JKL}\ ;\\
(a_J^1 &= a_K^2 = a_L^3 = 0)\ ;\\
c_{jk}^{12} &= \beta_{jkL} - \beta_{jKL} - \beta_{JkL} + \beta_{JKL}\ ;\ \ c_{j\ell}^{13} = \beta_{jK\ell} - \beta_{jKL} - \beta_{JK\ell} + \beta_{JKL}\ ;\\
c_{k\ell}^{23} &= \beta_{Jk\ell} - \beta_{JkL} - \beta_{JK\ell} + \beta_{JKL}\ ;\\
(c_{jK}^{12} &= 0, \forall j\ ;\ \ c_{Jk}^{12} = 0, \forall k\ ;\\
c_{jL}^{13} &= 0, \forall j\ ;\ \ c_{J\ell}^{13} = 0, \forall \ell\ ;\ \ c_{kL}^{23} = 0, \forall k\ ;\ \ c_{K\ell}^{23} = 0, \forall \ell)\ ;\\
d_{jk\ell} &= \beta_{jk\ell} - \beta_{jkL} - \beta_{jK\ell} - \beta_{Jk\ell} + \beta_{JKL} + \beta_{JkL} + \beta_{JK\ell} - \beta_{JKL}\ ;\\
(d_{jk\ell} &= 0,\ \text{dès qu'au moins un des indices } j, k \text{ ou } \ell \text{ est maximum}).
\end{aligned}$$

Il y a toujours un total de JKL paramètres indépendants dans le paramétrage SAS.

4.3 Estimations

Selon le même principe que celui vu dans les cas de un ou de deux facteurs, on obtient, comme estimation de chaque paramètre β_{jkl}, la quantité :

$$\hat{\beta}_{jk\ell} = \overline{y}_{\bullet jk\ell} = \frac{1}{n_{jk\ell}} \sum_{i=1}^{n_{jk\ell}} y_{ijk\ell}.$$

Pour le paramétrage centré, on calcule ensuite toutes les moyennes partielles (toujours non pondérées) de ces estimations, respectivement notées :

$$\hat{\beta}_{\bullet k\ell} \;\; \hat{\beta}_{j\bullet\ell} \;\; \hat{\beta}_{jk\bullet} \;\; \hat{\beta}_{\bullet\bullet\ell} \;\; \hat{\beta}_{\bullet k\bullet} \;\; \hat{\beta}_{j\bullet\bullet} \;\; \hat{\beta}_{\bullet\bullet\bullet}.$$

Les estimations des paramètres μ, α^1_j, α^2_k, α^3_ℓ, γ^{12}_{jk}, $\gamma^{13}_{j\ell}$, $\gamma^{23}_{k\ell}$ et $\delta_{jk\ell}$ s'obtiennent, à partir des estimations précédentes, en utilisant les mêmes formules que celles ayant permis de définir ces paramètres à partir des moyennes partielles des $\beta_{jk\ell}$.

Dans le paramétrage SAS, selon le même principe, on utilise les mêmes formules que celles données plus haut en remplaçant les β par leurs estimations. On obtient ainsi les estimations des paramètres définis par SAS.

4.4 Tests

Là encore, nous préconisons de procéder de façon hiérarchique pour faire les tests (de Fisher) de nullité des différentes catégories de paramètres.

- On commence donc par tester la nullité des effets d'interactions d'ordre 3 :
$$\{H_0 : \delta_{jk\ell} = 0 \ , \ \forall (j,k,\ell)\}.$$
Si cette hypothèse est rejetée, on garde le modèle complet et les tests sont terminés.
- Si, au contraire, l'hypothèse ci-dessus n'est pas rejetée, on prend comme modèle de référence le modèle sans interactions d'ordre 3. Dans ce modèle, on teste successivement la nullité des différents effets d'interactions d'ordre 2 :
$$\begin{aligned} \{H_0^{12} &: \gamma^{12}_{jk} = 0 \ , \ \forall (j,k)\} \ ; \\ \{H_0^{13} &: \gamma^{13}_{j\ell} = 0 \ , \ \forall (j,\ell)\} \ ; \\ \{H_0^{23} &: \gamma^{23}_{k\ell} = 0 \ , \ \forall (k,\ell)\}. \end{aligned}$$
- Si les trois hypothèses sont rejetées, on garde le modèle avec les trois séries d'effets d'interactions d'ordre 2 et les tests sont terminés.
- Si deux hypothèses sont rejetées, on enlève seulement du modèle les effets d'interactions supposés nuls et les tests sont terminés (chacun des trois facteurs doit être conservé car il intervient dans au moins une des séries d'interactions d'ordre deux).
- Si une seule hypothèse est rejetée, on enlève les effets d'interactions supposés nuls et on teste la nullité des effets du facteur n'intervenant pas dans les interactions (il y en a un et un seul). Si cette dernière hypothèse est rejetée, les tests sont terminés. Sinon, on enlève le facteur correspondant et on se retrouve dans un modèle d'ANOVA à deux facteurs (les tests sont également terminés).
- Si aucune hypothèse n'est rejetée, on prend alors le modèle additif (sans aucune interaction) pour référence et on teste séparément la nullité des effets de chaque facteur.

Remarque. — On utilisera encore le test de Fisher pour tester la significativité du modèle retenu.

Remarque. — Dans le cadre d'un plan à trois facteurs, on appelle toujours modèle additif le modèle sans aucune interaction.

5 Généralisation

Conceptuellement, il n'est pas difficile de définir des plans factoriels à quatre facteurs croisés ou plus. Toutefois, leur écriture devient très vite inextricable.

Il faut noter que, dans la pratique, notamment industrielle, il n'est pas rare de trouver de tels plans à au moins quatre facteurs. Toutefois, dans ce genre de situations, on a le plus souvent affaire à des plans incomplets, les plans complets étant trop coûteux à mettre en œuvre. Il convient alors de choisir de façon spécifique les cellules dans lesquelles on réalise les observations, de manière à obtenir un maximum de propriétés statistiques avec un minimum d'observations. L'étude de certains de ces plans factoriels incomplets est abordée dans le chapitre 4.

Enfin, signalons qu'on rencontre également, dans la pratique, des plans à deux ou plusieurs facteurs hiérarchisés (les niveaux d'un facteur sont conditionnés par les niveaux d'un autre facteur). Nous n'abordons pas ce type de plans dans ce cours, mais signalons néanmoins qu'il est très simple de les mettre en œuvre avec la procédure GLM de SAS et que c'est dans ce cas que les sommes de carrés de type I s'avèrent utiles (voir l'Annexe B).