Analyses de variance et covariance

Résumé

Introduction au modèle linéaire et modèle linéaire général : analyse de variance et covariance.
Retour au plan du cours.

1 Introduction

Les techniques dites d'*analyse de variance* sont des outils entrant dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par une ou plusieurs variables qualitatives. L'objectif essentiel est alors de comparer les moyennes empiriques de la variable quantitative observées pour différentes catégories d'unités statistiques. Ces catégories sont définies par l'observation des variables qualitatives ou *facteurs* prenant différentes modalités ou encore de variables quantitatives découpées en classes ou *niveaux*. Une combinaison de niveaux définit une *cellule*, *groupe* ou *traitement*.

Il s'agit donc de savoir si un facteur ou une combinaison de facteurs (*interaction*) a un *effet* sur la variable quantitative en vue, par exemple, de déterminer des conditions optimales de production ou de fabrication, une dose optimale de médicaments.... Ces techniques apparaissent aussi comme des cas particuliers de la régression linéaire multiple en associant à chaque modalité une *variable indicatrice* (dummy variable) et en cherchant à expliquer une variable quantitative par ces variables indicatrices. L'appellation "analyse de variance" vient de ce que les tests statistiques sont bâtis sur des comparaisons de sommes de carrés de variations.

L'analyse de variance est souvent utilisée pour analyser des données issue d'une *planification d'expérience* au cours de laquelle l'expérimentateur a la possibilité de contrôler *a priori* les niveaux des facteurs avec pour objectif d'obtenir le maximum de précision au moindre coût. Ceci conduit en particulier à construire des facteurs orthogonaux deux à deux (variables explicatives non linéairement corrélées) afin de minimiser la variance des estimateurs (cf. chapitre précédent). On distingue le cas particulier important où les cellules ont le même effectif, on parle alors de *plan orthogonal* ou *équirépété* ou *équilibré* (balanced), qui conduit à des simplifications importantes de l'analyse de variance associée. On appelle plan *complet* un dispositif dans lequel toutes les combinaisons de niveaux ont été expérimentées. On distingue entre des modèles fixes, aléatoires ou mixtes selon le caractère déterministe (contrôlé) ou non des facteurs par exemple si les modalités résultent d'un choix aléatoire parmi un grand nombre de possibles. Seuls les modèles fixes sont considérés.

L'analyse de covariance considère une situation plus générale dans laquelle les variables explicatives sont à la fois quantitatives, appelées covariables, et qualitatives ou facteurs. L'objectif est alors de comparer, non plus des moyennes par cellules, mais les paramètres des différents modèles de régressions estimées pour chaque combinaison de niveau. Ce type de modèle est introduit en fin de chapitre.

Les spécificités de la planification d'expérience ne seront qu'abordées dans ce chapitre. Les applications en sont surtout développées en milieu industriel : contrôle de qualité, optimisation des processus de production, ou en agronomie pour la sélection de variétés, la comparaison d'engrais, d'insecticides.... La bibliographie est abondante à ce sujet.

2 Modèle à un facteur

Cette situation est un cas particulier d'étude de relations entre deux variables statistiques : une quantitative Y admettant une densité et une qualitative T ou facteur qui engendre une partition ou classification de l'échantillon en J groupes, cellules ou classes indicées par j. L'objectif est de comparer les distributions de Y pour chacune des classes en particulier les valeurs des moyennes et variances.

Un préalable descriptif consiste à réaliser un graphique constitué de boites à moustaches parallèles : une pour chaque modalité. Cette représentation donne une première appréciation de la comparaison des distributions (moyenne, variance) internes à chaque groupe.

2.1 Modèles

Pour chaque niveau j de T, on observe n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de la variable Y et où $n = \sum_{j=1}^{J} n_j$ est la taille de l'échantillon ($n > J$). On suppose qu'à l'intérieur de chaque cellule, les observations sont indépendantes équidis-

tribuées de moyenne μ_j et de variance *homogène* $\sigma_j^2 = \sigma^2$. Ceci s'écrit :

$$y_{ij} = \mu_j + \varepsilon_{ij}$$

où les ε_{ij} sont i.i.d. suivant une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests. Cette dernière hypothèse n'étant pas la plus sensible. Les espérances μ_j ainsi que le paramètre de nuisance σ^2 sont les paramètres inconnus à estimer.

On note respectivement :

$$\begin{aligned} \bar{y}_{.j} &= \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}, \\ s_j^2 &= \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2, \\ \bar{y}_{..} &= \frac{1}{n}\sum_{i=1}^{n_j}\sum_{j=1}^{J} y_{ij}, \end{aligned}$$

les moyennes et variances empiriques de chaque cellule, la moyenne générale de l'échantillon.

Les paramètres μ_j sont estimés sans biais par les moyennes $\bar{y}_{.j}$ et comme le modèle s'écrit alors :

$$y_{ij} = \bar{y}_{.j} + (y_{ij} - \bar{y}_{.j}),$$

l'estimation des erreurs est $e_{ij} = (y_{ij} - \bar{y}_{.j})$ tandis que les valeurs prédites sont $\widehat{y}_{ij} = \bar{y}_{.j}$.

Sous l'hypothèse d'homogénéité des variances, la meilleure estimation sans biais de σ^2 est

$$s^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2}{n - J} = \frac{1}{n - J}[(n-1)s_1^2 + \cdots + (n_J - 1)s_J^2]$$

qui s'écrit donc comme une moyenne pondérée des variances empiriques de chaque groupe.

Notons $\mathbf{y}$ le vecteur des observations $[y_{ij}|i = 1, n_j; j = 1, J]'$ mis en colonne, $\mathbf{u} = [\varepsilon_{ij}|i = 1, n_j; j = 1, J]'$ le vecteur des erreurs, $\mathbf{1}_j$ les variables indicatrices des niveaux et $\mathbf{1}$ la colonne de 1s. Le $i$ème élément d'une variable indicatrice (dummy variable) $\mathbf{1}_j$ prend la valeur 1 si la $i$ème observation y_i est associée au $j$ème et 0 sinon.

Comme dans le cas de la régression linéaire multiple, le modèle consiste à écrire que l'espérance de la variable Y appartient au sous-espace linéaire engendré par les variables explicatives, ici les variables indicatrices :

$$\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{1}_1 + \cdots + \beta_J\mathbf{1}_J + \mathbf{u}.$$

La matrice $\mathbf{X}$ alors construite n'est pas de plein rang $p + 1$ mais de rang p. La matrice $\mathbf{X}'\mathbf{X}$ n'est pas inversible et le modèle admet une infinité de solutions. Nous disons que les paramètres β_j ne sont pas *estimables* ou identifiables. En revanche, certaines fonctions (combinaisons linéaires) de ces paramètres sont estimables et appelées *contrastes*.

Dans le cas du modèle d'analyse de variance à *un* facteur, la solution la plus simple adoptée consiste à considérer un sous-ensemble des indicatrices ou de combinaisons des indicatrices de façon à aboutir à une matrice inversible. Ceci conduit à considérer différents modèles associés à différentes paramétrisation. *Attention*, les paramètres β_j ainsi que la matrice $\mathbf{X}$ prennent à chaque fois des significations différentes.

Un premier modèle (cell means model) s'écrit comme celui d'une régression linéaire multiple sans terme constant avec $\boldsymbol{\beta} = [\mu_1, \ldots, \mu_J]'$ le vecteur des paramètres :

$$\begin{aligned} \mathbf{y} &= \beta_1\mathbf{1}_1 + \cdots + \beta_J\mathbf{1}_J + \mathbf{u} \\ \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \end{aligned}$$

Les calculs se présentent simplement (cf. exo 1) mais les tests découlant de ce modèle conduiraient à étudier la nullité des paramètres alors que nous sommes intéressés par tester l'égalité des moyennes.

Une autre paramétrisation, considérant cette fois le vecteur $\boldsymbol{\beta} = [\mu_J, \mu_1 - \mu_J, \ldots, \mu_{J-1} - \mu_J]'$ conduit à écrire le modèle (base cell model) de régression avec terme constant :

$$\mathbf{y} = \beta_0\mathbf{1} + \beta_1\mathbf{1}_1 + \cdots + \beta_{J-1}\mathbf{1}_{J-1} + \mathbf{u}.$$

C'est celle de SAS alors que Systat considère des paramètres d'effet différentiel $\mu_j - \mu_.$ par rapport à l'effet moyen $\mu_. = 1/J\sum_{j=1}^{J}\mu_j$. Ce dernier est en-

core un modèle (group effect model) de régression linéaire avec terme constant mais dont les variables explicatives sont des différences d'indicatrices et avec $\boldsymbol{\beta} = [\mu_., \mu_1 - \mu_., \ldots, \mu_{J-1} - \mu_.]'$:

$$\mathbf{y} = \beta_0 \mathbf{1} + \beta_1(\mathbf{1}_1 - \mathbf{1}_J) + \cdots + \beta_{J-1}(\mathbf{1}_{J-1} - \mathbf{1}_J) + \mathbf{u}.$$

2.2 Test

On désigne les différentes sommes des carrés des variations par :

$$\begin{aligned}
\text{SST} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{..})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}^2 - n\bar{y}_{..}^2, \\
\text{SSW} &= \sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2 = \sum_{j=1}^{J}\sum_{i=1}^{n_j} y_{ij}^2 - \sum_{j=1}^{J} n_j \bar{y}_{.j}^2, \\
\text{SSB} &= \sum_{j=1}^{J} n_j(\bar{y}_{.j} - \bar{y}_{..})^2 = \sum_{j=1}^{J} n_j \bar{y}_{.j}^2 - n\bar{y}_{..}^2,
\end{aligned}$$

où "T" signifie totale, "W" (within) intra ou résiduelle, "B" (between) inter ou expliquée par la partition. Il est facile de vérifier que SST=SSB+SSW.

On considère alors l'hypothèse

$$H_0 \ : \mu_1 = \cdots = \mu_J,$$

qui revient à dire que la moyenne est indépendante du niveau ou encore que le facteur n'a pas d'effet, contre l'hypothèse

$$H_1 \ : \exists (j,k) \text{ tel que } \mu_j \neq \mu_k$$

qui revient à reconnaître un effet ou une influence du facteur sur la variable Y.

Dans les modèles précédents, l'étude de cette hypothèse revient à comparer par un test de Fisher un modèle complet (les moyennes sont différentes) avec un modèle réduit supposant la nullité des paramètres β_j et donc l'égalité des moyennes à celle de la dernière cellule ou à la moyenne générale.

Les résultats nécessaires à la construction du test qui en découle sont résumés dans la table d'analyse de la variance :

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	*F*
Modèle (inter)	$J-1$	SSB	MSB=SSB/$(J-1)$	MSB/MSW
Erreur (intra)	$n-J$	SSW	MSW=SSW/$(n-J)$	
Total	$n-1$	SST		

Pratiquement, un programme de régression usuel permet de construire estimation et test de la nullité des β_j sauf pour le premier modèle qui doit tester l'égalité au lieu de la nullité des paramètres.

Dans le cas de deux classes $(J = 2)$ on retrouve un test équivalent au test de Student de comparaison des moyennes de deux échantillons indépendants.

2.3 Comparaisons multiples

Si l'hypothèse nulle est rejetée, la question suivante consiste à rechercher quelles sont les groupes ou cellules qui possèdent des moyennes significativement différentes. De nombreux tests et procédures ont été proposés dans la littérature pour répondre à cette question.

Une procédure naïve consiste à exprimer, pour chaque paire j et l de groupes, un intervalle de confiance au niveau $100(1-\alpha)\%$ de la différence $(\mu_j - \mu_l)$:

$$(\bar{y}_{.j} - \bar{y}_{.l}) \pm t_{\alpha/2;(n-J)} s \left[\frac{1}{n_j} + \frac{1}{n_l}\right]^{1/2}.$$

Si, pour un couple (j,l) fixé a priori, cet intervalle inclut 0, les moyennes ne sont pas jugées significativement différentes au niveau α. L'orthogonalité des facteurs rendant les tests indépendants justifierait cette procédure mais elle ne peut être systématisée. En effet, si J est grand, il y a un total de $J(J-1)/2$ comparaisons à considérer et on peut s'attendre à ce que, sur le simple fait du hasard, $0,05 \times J(J-1)/2$ paires de moyennes soient jugées significativement différentes même si le test global accepte l'égalité des moyennes.

D'autres procédures visent à corriger cette démarche afin de contrôler globalement le niveau des comparaisons. Certaines proposent des intervalles plus conservatifs (plus grands) en ajustant le niveau $\alpha' < \alpha$ définissant les valeurs critiques $t_{\alpha'/2;(n-J)}$ (Bonferroni $\alpha' = \alpha/(J(J-1)/2)$, Sidak). Dans le

même esprit, la méthode de Scheffe, la plus conservative, projette l'ellipsoïde de confiance des moyennes des μ_i en intervalles de confiance des différences ou de toute combinaison linéaire de celles-ci (contrastes).

D'autres procédures définissent des intervalles studentisés fournissant des valeurs critiques spécifiques qui sont tabulées ou calculées par le logiciel. Certaines de ces méthodes ou certaines présentations graphiques des résultats sont uniquement adaptées au cas équirépété (Tukey) tandis que d'autres sont adaptées à des classes présentant des effectifs différents (GT2, Gabriel).

En résumé, pour comparer toutes les moyennes dans le cas équirépété, les méthodes de Tukey ou Scheffe sont utilisées, celle de Bonferroni convient encore au cas déséquilibré. Pour comparer les moyennes à celle d'une classe ou traitement témoin, la méthode de Bonferroni ($\alpha' = \alpha/(J(J-1)/2)$) est encore utilisée tandis que Dunnet remplace Tukey dans le cas équirépété.

2.4 Homogénéité de la variance

Une hypothèse importante du modèle induit par l'analyse de variance est l'homogénéité des variances de chaque groupe. Conjointement à l'estimation du modèle et en supposant la normalité, il peut être instructif de contrôler cette homogénéité par un test de l'hypothèse

$$H_0 : \sigma_1^2 = \cdots = \sigma_J^2.$$

Bartlett a proposé le test suivant. Posons

$$M = \sum_{j=1}^{J} (n_j - 1) \ln(s^2/s_j^2)$$

et

$$c = \frac{1}{3(J-1)} \left(\sum_{j=1}^{J} (\frac{1}{n_j - 1}) - 1/\sum_{j=1}^{J} (n_j - 1) \right).$$

Sous H_0 et pour de grands échantillons, la statistique $M/(c+1)$ suit un χ^2 à $(J-1)$ degrés de liberté. Dans les mêmes conditions, une approximation peut être fournie par la statistique

$$F = \frac{dM}{(J-1)(d/f - M)},$$

avec

$$f = (1-c) + 2/d \text{ et } d = (J+1)/c^2,$$

qui suit une loi de Fisher à $(J-1)$ et d degrés de liberté.

Néanmoins ce test n'est pas robuste à la violation de l'hypothèse de normalité. C'est pourquoi il lui est préféré la méthode de Levene qui considère les variables :

$$Z_{ij} = |y_{ij} - \bar{y}_{.j}|$$

sur lesquelles est calculée une analyse de variance. Malgré que les Z_{ij} ne soient ni indépendantes ni identiquement distribuées suivant une loi normale, la statistique de Fisher issue de l'ANOVA fournit un test raisonnable de l'homoscédasticité.

Le graphique représentant le nuage des résidus ou les boites à moustaches en fonction des niveaux du facteur complète très utilement le diagnostic. En cas d'hétéroscédasticité et comme en régression, une transformation de la variable à expliquer Y ($\sqrt{Y}, \ln(Y), 1/Y \ldots$) permet de limiter les dégâts.

2.5 Tests non paramétriques

Lorsque l'hypothèse de normalité n'est pas satisfaite et que la taille trop petite de l'échantillon ne permet pas de supposer des propriétés asymptotiques, une procédure non-paramétrique peut encore être mise en œuvre. Elles sont des alternatives plausibles au test de Fisher pour tester l'égalité des moyennes.

La procédure la plus utilisée est la construction du test de Kruskal-Wallis basée sur les rangs. Toutes les observations sont ordonnées selon les valeurs y_{ij} qui sont remplacées par leur rang r_{ij}, les ex æquo sont remplacés par leur rang moyen. On montre que la statistique de ce test, construite sur la somme des rangs à l'intérieur de chaque groupe, suit asymptotiquement une loi du χ^2 à $(J-1)$ degrés de liberté.

Une autre procédure, utilisant cette fois des rangs normalisés ($a_{ij} = r_{ij}/(n+1)$) conduit à une autre statistique utilisée dans le test de van der Waerden.

3 Modèle à deux facteurs

La considération de deux (ou plus) facteurs explicatifs, dans un modèle d'analyse de variance, engendre plusieurs complications. La première concerne la notion d'*interaction* entre variables explicatives. D'autres seront introduites dans la section suivante. Cette section décrit le cas de deux facteurs explicatifs croisés c'est-à-dire dont les niveaux d'un facteur ne sont pas conditionnés par ceux de l'autre. Les niveaux du premier facteur sont notés par un indice j variant de 1 à J, ceux du deuxième par un indice k variant de 1 à K.

Pour chaque combinaison, on observe un même nombre $n_{jk} = c > 1$ de répétitions ce qui nous place dans le cas particulier d'un plan **équilibré** ou **équirépété**. Ceci introduit des simplifications importantes dans les estimations des paramètres ainsi que dans la décomposition des variances. Le cas plus général est évoqué dans la section suivante.

3.1 Modèle complet

On peut commencer par écrire un modèle de variance à un facteur présentant $J \times K$ niveaux (j, k) :

$$y_{ijk} = \mu_{jk} + \varepsilon_{ijk} \text{ où } \left\{ \begin{array}{rcl} j & = & 1, \ldots, J \, ; \\ k & = & 1, \ldots, K \, ; \\ i & = & 1, \ldots, n_{jk} = c \, ; \end{array} \right.$$

en supposant que les termes d'erreur ε_{ijk} sont mutuellement indépendants et de même loi. Chacun des paramètres μ_{jk} est estimé sans biais par la moyenne

$$\bar{y}_{.jk} = \frac{1}{c} \sum_{i=1}^{c} y_{ijk}.$$

Définissons également les moyennes suivantes :

$$\begin{aligned} \bar{y}_{.j.} &= \frac{1}{K} \sum_{k=1}^{K} \bar{y}_{.jk}, \\ \bar{y}_{..k} &= \frac{1}{J} \sum_{j=1}^{J} \bar{y}_{.jk}, \\ \bar{y}_{...} &= \frac{1}{J} \sum_{j=1}^{J} \bar{y}_{.j.} = \frac{1}{K} \sum_{k=1}^{K} \bar{y}_{..k}. \end{aligned}$$

qui n'ont de sens que dans le cas équirépété. La même convention du point en indice est également utilisée pour exprimer les moyennes des paramètres μ_{ijk}.

Les moyennes de chaque cellule sont alors décomposées en plusieurs termes afin de faire apparaître l'influence de chaque facteur :

Terme	*Paramètre*	*Estimation*
Moyenne générale	$\mu_{..}$	$\bar{y}_{...}$
Effet niveau j du 1er facteur	$\alpha_j = \mu_{j.} - \mu_{..}$	$\bar{y}_{.j.} - \bar{y}_{...}$
Effet niveau k du 2ème facteur	$\beta_k = \mu_{.k} - \mu_{..}$	$\bar{y}_{..k} - \bar{y}_{...}$
Effet de l'*interaction*	$\gamma_{jk} = \mu_{jk} - \mu_{j.} - \mu_{.k} + \mu_{..}$	$\bar{y}_{.jk} - \bar{y}_{.j.} - \bar{y}_{..k} + y_{...}$

Avec les notations du tableau ci-dessus, on appelle $\mu_{..}$ l'effet général, $\mu_{j.}$ l'effet du niveau j du premier facteur, α_j l'effet différentiel du niveau j du premier facteur (même chose avec $\mu_{.k}$ et β_k pour le 2ème facteur), γ_{jk} l'effet d'interaction des niveaux j et k.

Un modèle d'analyse de variance à deux facteurs s'écrit alors :

$$y_{ijk} = \mu_{..} + \alpha_j + \beta_k + \gamma_{jk} + \varepsilon_{ijk} \text{ où } \left\{ \begin{array}{rcl} j & = & 1, \ldots, J \, ; \\ k & = & 1, \ldots, K \, ; \\ i & = & 1, \ldots, n_{jk} = c \, ; \end{array} \right.$$

avec les contraintes

$$\sum_{j=1}^{J} \alpha_j = \sum_{k=1}^{K} \beta_k = 0 \, ; \, \forall k, \sum_{j=1}^{J} \gamma_{jk} = 0 \, ; \, \forall j, \sum_{k=1}^{K} \gamma_{jk} = 0$$

qui découlent de la définition des effets et assurent l'unicité de la solution.

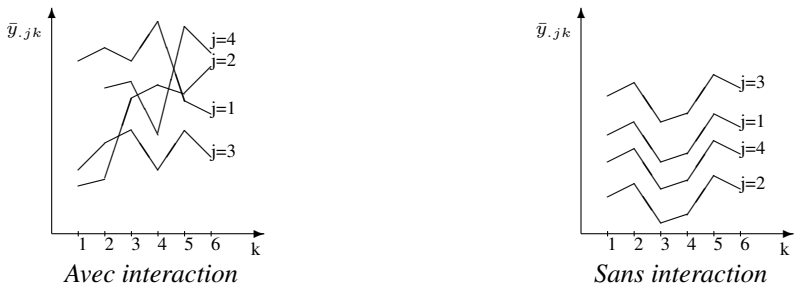

FIGURE 1 – Moyennes de la variable Y pour chaque niveau d'un facteur en fonction des niveaux de l'autre facteur.

3.2 Interaction

Lorsque les paramètres d'interaction γ_{jk} sont tous nuls, le modèle est dit *additif* ce qui correspond à une situation très particulière. Elle intervient lorsque

$$\bar{y}_{.jk} - \bar{y}_{..k} = \bar{y}_{.j.} - y_{...} \quad \forall j = 1, \dots J \; ; \; \forall k = 1, \dots, K$$

ce qui signifie que les écarts relatifs du premier facteur sont indépendants du niveau k du 2ème facteur (et vice versa).

Graphiquement, cela se traduit dans la figure 3.1 qui illustre les comportements des moyennes des cellules de modèles avec ou sans interaction (additif). Chaque ligne est appelée un *profil*, et la présence d'interactions se caractérise par le croisement de ces profils tandis que le parallélisme indique l'absence d'interactions. La question est évidemment de tester si des croisements observés sont jugés significatifs.

Attention, un manque de parallélisme peut aussi être dû à la présence d'une relation non-linéaire entre la variable Y et l'un des facteurs.

3.3 Modèles de régression

Comme dans le cas du modèle à un facteur, l'analyse d'un plan à deux facteurs se ramène à l'estimation et l'étude de modèles de régression sur variables indicatrices. En plus de celles des niveaux des deux facteurs $\{\mathbf{1}_1^1, \dots, \mathbf{1}_J^1\}$, et $\{\mathbf{1}_1^2, \dots, \mathbf{1}_K^2\}$, la prise en compte de l'interaction nécessite de considérer les indicatrices de chaque cellule ou traitement obtenues par produit des indicatrices des niveaux associés :

$$\mathbf{1}_{jk}^{1\times 2} = \mathbf{1}_j^1 \times \mathbf{1}_k^2 \; ; j = 1 \dots, J \; ; k = 1, \dots, K.$$

Le modèle s'écrit alors avec une autre paramétrisation :

$$\mathbf{y} = \beta_0\mathbf{1} + \beta_{1,1}\mathbf{1}_1^1 + \cdots + \beta_{1,J}\mathbf{1}_J^1 + \beta_{2,1}\mathbf{1}_1^2 + \cdots + \beta_{2,K}\mathbf{1}_K^2 + \beta_{1\times 2,1}\mathbf{1}_1^{1\times 2} + \cdots + \beta_{1\times 2,JK}\mathbf{1}_{J,K}^{1\times 2} + \mathbf{u},$$

il comporte $1 + I + J + I \times J$ paramètres mais les colonnes de $\mathbf{X}$ sont soumises à de nombreuses combinaisons linéaires : une par paquet de $\mathbf{1}_j^1$ ou de $\mathbf{1}_k^2$ et une pour chaque paquet de $\mathbf{1}_{jk}^{1\times 2}$ à j ou k fixé. La matrice $\mathbf{X}'\mathbf{X}$ n'est pas inversible. Différentes approches sont proposées pour résoudre ce problème d'identifiabilité des paramètres.

- Supprimer une des indicatrices : en fonction de la base d'indicatrices choisie, différents modèles et donc différentes paramétrisations sont considérées.
- Ajouter une contrainte sur les paramètres afin de rendre unique la solution.
- Chercher une solution à partir d'un *inverse généralisé* [1] de la matrice $\mathbf{X}'\mathbf{X}$.

Dans le cas du modèle d'analyse de variance à *un* facteur, seule la première solution est couramment employée. Les autres, plus générales, le sont dans le cas de plusieurs facteurs et justifiées par des plannifications plus complexes ; différents inverses généralisés permettant de reconstruire les solutions avec contraintes ou par élimination d'une variable indicatrice. Les différents modèles considérés par les logiciels conduisent alors à des tests équivalents mais attention, la matrice $\mathbf{X}$ et le vecteur β prennent des significations différentes.

3.4 Stratégie de test

Une première décomposition de la variance associée au test général de nullité de tous les paramètres est proposée dans les logiciels mais celle-ci ne présente que peu d'intérêt. On considère ensuite les sommes de carrés spécifiques

1. On dit que la matrice $\mathbf{A}^-$ est inverse généralisé de la matrice carrée $\mathbf{A}$ si elle vérifie : $\mathbf{A}^-\mathbf{A}\mathbf{A}^- = \mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}^-$.

au cas équirépété :

$$\begin{aligned}
\text{SST} &= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{...})^2 &&= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2 - cJK\bar{y}_{...}^2,\\
\text{SS1} &= cK\sum_{j=1}^{J}(\bar{y}_{.j.}-\bar{y}_{...})^2 &&= cK\sum_{j=1}^{J} y_{.j.}^2 - cJK\bar{y}_{...}^2,\\
\text{SS2} &= cJ\sum_{k=1}^{K}(\bar{y}_{..k}-\bar{y}_{...})^2 &&= cJ\sum_{k=1}^{K} y_{..k}^2 - cJK\bar{y}_{...}^2,\\
\text{SSI} &= c\sum_{j=1}^{J}\sum_{k=1}^{K}(\bar{y}_{.jk}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})^2 &&= c\sum_{j=1}^{J}\sum_{k=1}^{K}\bar{y}_{.jk}^2 - cK\sum_{j=1}^{J} y_{.j.}^2 -\\
& && \quad -cJ\sum_{k=1}^{K} y_{..k}^2 + cJK\bar{y}_{...}^2,\\
\text{SSE} &= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K}(y_{ijk}-\bar{y}_{.jk})^2 &&= \sum_{i=1}^{c}\sum_{j=1}^{J}\sum_{k=1}^{K} y_{ijk}^2 - c\sum_{j=1}^{J}\sum_{k=1}^{K}\bar{y}_{.jk}^2.
\end{aligned}$$

Dans ce cas, il est facile de montrer que tous les "doubles produits" des décompositions s'annulent (théorème de Pythagore) et

$$\text{SST} = \text{SS1} + \text{SS2} + \text{SSI} + \text{SSE}.$$

On parle alors de plans *orthogonaux* et les trois hypothèses suivantes (associées à des regroupements de contrastes) peuvent être considérées de façon indépendante :

$$\begin{aligned}
H_{03} &: \quad \gamma_{11} = \cdots = \gamma_{JK} = 0, && \text{pas d'effet d'interaction.}\\
H_{02} &: \quad \beta_1 = \cdots = \beta_K = 0, && \text{et } H_{03}\text{, pas d'effet du 2ème facteur}\\
H_{01} &: \quad \alpha_1 = \cdots = \alpha_J = 0, && \text{et } H_{03}\text{, pas d'effet du 1er facteur}
\end{aligned}$$

Elles sont évaluées dans la table ci-dessous qui présente l'unique décomposition de la variance dans le cas équirépété [2].

2. Les options SS1,SS2, SS3, SS4 de SAS fournissent ainsi les mêmes résultats.

Source de variation	*d.d.l.*	*Somme des carrés*	*Variance*	*F*
1er facteur	$J-1$	SS1	MS1=SS1/$(J-1)$	MS1/MSE
2ème facteur	$K-1$	SS2	MS2=SS2/$(K-1)$	MS2/MSE
Interaction	$(J-1)(K-1)$	SSI	MSI=$\frac{\text{SSI}}{(J-1)(K-1)}$	MSI/MSE
Erreur	$JK(c-1)$	SSE	MSE=SSE/$JK(c-1)$	
Total	$cJK-1$	SST		

Différentes stratégies de test sont suivies dans la littérature mais la plus couramment pratiquée consiste à comparer le modèle complet avec chacun des sous-modèles :

– Évaluer H_{03} de présence ou absence des termes d'interaction. Il existe des modèles intermédiaires de structuration de l'interaction mais le cas le plus simple du "tout ou rien" est retenu. Deux possibilités se présentent alors.

1. Si l'interaction est significativement présente alors les deux facteurs sont influents ne serait-ce que par l'interaction. Il n'y a pas lieu de tester leur présence par H_{01} et H_{02}. Néanmoins il est d'usage de comparer les différentes statistiques F de test afin d'apprécier les rapports d'influence entre les effets principaux et l'interaction.
2. Si l'interaction n'est pas significativement présente, il reste alors à tester l'effet de chaque facteur. Certains auteurs ré-estiment le modèle *additif* sans paramètre d'interaction (cf. remarque ci-dessous). Cela est déconseillé pour se protéger contre un manque possible de puissance du test de l'interaction. En effet, une faible interaction non décelée fausse l'estimation s^2 de σ_2. Il est donc préférable de conserver le modèle complet et de tester l'influence des facteurs par la nullité des α_j et β_j à partir des statistiques de la table ci-dessus.

Remarques

1. Si, compte tenu de connaissances *a priori* liées à un problème spécifique, l'interaction est éliminée du modèle, on est donc conduit à estimer un modèle additif plus simple (sans paramètres γ_{jk}). Dans ce cas, le nombre de paramètres à estimer et ainsi le nombre de degrés de liberté, la somme de carrés SSE et donc l'estimation $s^2 = MSE$ de σ^2 ne sont plus les valeurs

fournies par la table d'analyse de variance ci-dessus. On distingue donc le cas d'un modèle *a priori* additif d'un modèle dans lequel l'hypothèse de nullité des interactions est acceptée.

2. D'autres tests plus spécifiques sont construits en considérant des combinaisons linéaires des paramètres (contrastes) ou en effectuant des comparaisons multiples comme dans le cas à un facteur (Bonferroni, Tukey, Scheffe...).
3. Les tests d'homogénéité des variances se traitent comme dans le cas du modèle à un facteur en considérant les JK combinaisons possibles.

4 Problèmes spécifiques

Certaines contraintes expérimentales peuvent induire des spécificités dans la planification et ainsi, par conséquence, dans le modèle d'analyse de variance associé. Un exposé détaillé des situations possibles sort du cadre de ce cours de 2ème cycle. Nous nous contenterons de citer ici certains problèmes courants en soulignant les difficultés occasionnées et quelques éléments de solution.

4.1 Facteur bloc

Les facteurs peuvent jouer des rôles différents. Certains sont contrôlés par l'expérimentateur qui sait en fixer précisément le niveau, d'autres, appelés *blocs*, sont des sources de variation propres aux procédés expérimentaux mais dont il faut tenir compte dans l'analyse car source d'hétérogénéité. L'exemple le plus typique concerne l'expérimentation agronomique en plein champ dans laquelle il est impossible de garantir l'homogénéité des conditions climatiques, hydrométriques ou encore de fertilité. Chaque champ ou bloc est donc découpé en parcelles "identiques" qui recevront chacune un traitement.

Dans d'autres situations, certaines mesures ne sont pas indépendantes, par exemple, lorsqu'elles sont réalisées sur les mêmes individus dans le cas de *mesures répétées*. Il est alors indispensable d'introduire un facteur bloc rendant compte de la structure particulière de l'expérimentation.

L'objectif est de séparer pour contrôler "au mieux" les sources de variation. Une "randomisation", ou tirage au sort, est réalisé à l'intérieur de chaque bloc afin de répartir "au hasard", dans l'espace, dans le temps, l'expérimentation des traitements ou combinaisons des autres facteurs.

4.2 Plan sans répétition

Si une seule expérience ou mesure est réalisée par cellule ou traitement, les composantes d'interaction et résiduelles sont confondues. Aucune hypothèse n'est testable dans le cadre général précédent. Il est néanmoins possible de se placer dans le cadre du modèle additif afin de tester l'influence de chaque facteur sous l'hypothèse implicite de non interaction.

4.3 Plans déséquilibrés, incomplets

Le cas de plans déséquilibrés, c'est-à-dire dans lesquels le nombre d'observations n'est pas le même dans chaque cellule ou pour chaque traitement, nécessite une attention particulière, surtout si, en plus, des cellules sont vides. Différents problèmes surgissent alors :

- les moyennes $\bar{y}_{.j.}$ ou $\bar{y}_{..k}$ définissant les estimateurs n'ont plus de sens,
- les "doubles produits" des décompositions des sommes de carrés ne se simplifient plus, il n'y a plus "orthogonalité",
- en conséquence, les hypothèses précédentes ou ensembles de contrastes ne peuvent plus être considérés de manière indépendante.

Néanmoins, l'approche générale par modèle linéaire des indicatrices reste valide. La solution obtenue par inverse généralisé :

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$$

n'est pas unique mais est utilisée pour construire des fonctions estimables des éléments de $\mathbf{b}$: $k'\mathbf{b}$ où k est un vecteur définissant un contraste. Plusieurs contrastes linéairement indépendants étant regroupés dans une matrice $\mathbf{K}$, l'hypothèse associée : $\mathbf{K}'\mathbf{b} = 0$ est alors testable en considérant la somme des carrés

$$SSK = (\mathbf{K}'\mathbf{b})'[\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-}\mathbf{K}]^{-1}(\mathbf{K}'\mathbf{b})$$

avec rang$(\mathbf{K})$ pour nombre de degrés de liberté.

Cette procédure "à la main" de construction des tests étant assez lourde, SAS propose l'étude d'hypothèses "standards" à travers quatre options. La première (SS1) fait intervenir l'ordre d'introduction des facteurs et est plus particulièrement adaptée aux modèles hiérarchisés, par exemple polynômiaux. La troisième (SS3) est conseillée dans les cas où les inégalités d'effectifs n'ont pas de signification particulière, ne sont pas dépendantes des niveaux des facteurs.

Les deux autres options (SS2, SS4) ne sont guère utilisées, SS4, prévue pour les plans incomplets peut fournir des résultats étranges. En pratique standard, SS1 et SS3 sont comparées afin de s'assurer ou non de l'équirépartition puis les résultats de SS3 sont interprétés comme dans le cas équirépété.

4.4 Modèles à plus de deux facteurs

La prise en compte de plus de deux facteurs dans un modèle d'analyse de variance n'introduit pas de problème théorique fondamentalement nouveau. Seule la multiplication des indices et l'explosion combinatoire du nombre d'interactions à considérer compliquent la mise en œuvre pratique d'autant que beaucoup d'expérimentations sont nécessaires si la réalisation d'un plan complet est visée. Dans le cas contraire, tous les niveaux d'interaction ne sont pas testables et, comme dans le cas sans répétition, il faudra considérer des modèles moins ambitieux en supposant implicitement des hypothèses sur l'absence d'interactions d'ordres élevés. Si les facteurs sont très nombreux, il est courant de limiter chacun à 2 (ou 3 pour un modèle quadratique) niveaux et de ne considérer que certaines combinaisons deux à deux de facteurs. On parle alors de plans *fractionnaires*.

4.5 Facteurs hiérarchisés

Certains facteurs ou blocs peuvent par ailleurs être hiérarchisés ou emboîtés : les niveaux de certains facteurs sont conditionnés par d'autres facteurs. La composante d'interaction se confond alors avec la composante relative au facteur subordonné. Le modèle d'analyse de variance adapté à cette situation est dit *hiérarchisé*. Dans SAS, une syntaxe particulière permet de définir la structure.

5 Analyse de covariance

L'*analyse de covariance* se situe encore dans le cadre général du modèle linéaire et où une variable quantitative est expliquée par plusieurs variables à la fois quantitatives et qualitatives. Dans les cas les plus complexes, ont peut avoir plusieurs facteurs (variables qualitatives) avec une structure croisée ou hiérarchique ainsi que plusieurs variables quantitatives intervenant de manière linéaire ou polynômiale. Le principe général est toujours d'estimer des modèles "*intra-groupes*" et de faire apparaître (tester) des effets différentiels "*inter-groupes*" des paramètres des régressions. Ainsi, dans le cas plus simple où seulement une variable parmi les explicatives est quantitative, nous sommes amenés à tester l'hétérogénéité des constantes et celle des pentes (interaction) entre différents modèles de régression linéaire.

5.1 Modèle

Le modèle est explicité dans le cas élémentaire où une variable quantitative Y est expliquée par une variable qualitative T à J niveaux et une variable quantitative, appelée encore covariable, X. Pour chaque niveau j de T, on observe n_j valeurs $x_{1j}, \ldots, x_{n_j j}$ de X et n_j valeurs $y_{1j}, \ldots, y_{n_j j}$ de Y ; $n = \sum_{j=1}^{J} n_j$ est la taille de l'échantillon.

En pratique, avant de lancer une procédure de modélisation et tests, une démarche exploratoire s'appuyant sur une représentation en couleur (une par modalité j de T) du nuage de points croisant Y et X et associant les droites de régression permet de se faire une idée sur les effets respectifs des variables : parallélisme des droites, étirement, imbrication des sous-nuages.

On suppose que les moyennes conditionnelles $E[Y|T]$, c'est-à-dire calculées à l'intérieur de chaque cellule, sont dans le sous-espace vectoriel engendré par les variables explicatives quantitatives, ici X. Ceci s'écrit :

$$y_{ij} = \beta_{0j} + \beta_{1j} x_{ij} + \varepsilon_{ij} \; ; \quad j = 1, \ldots, J; \quad i = 1, \cdots, n_j$$

où les ε_{ij} sont i.i.d. suivant une loi centrée de variance σ^2 qui sera supposée $\mathcal{N}(0, \sigma^2)$ pour la construction des tests.

Notons $\mathbf{y}$ le vecteur des observations $[y_{ij}|i = 1, n_j; j = 1, J]'$ mis en colonne, $\mathbf{x}$ le vecteur $[x_{ij}|i = 1, n_j; j = 1, J]'$, $\mathbf{u} = [\varepsilon_{ij}|i = 1, n_j; j = 1, J]'$ le vecteur des erreurs, $\mathbf{1}_j$ les variables indicatrices des niveaux et $\mathbf{1}$ la colonne de 1s. On note encore $\mathbf{x}.\mathbf{1}_j$ le produit terme à terme des deux vecteurs, c'est-à-dire le vecteur contenant les observations de $\mathbf{X}$ sur les individus prenant le niveau j de T et des zéros ailleurs.

La résolution simultanée des J modèles de régression est alors obtenue en considérant globalement le modèle :

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

dans lequel $\mathbf{X}$ est la matrice $n \times 2J$ constituée des blocs $[\mathbf{1}_j|\mathbf{x}.\mathbf{1}_j]$; $j =$

$1, \ldots, J$. L'estimation de ce modèle global conduit, par bloc, à estimer les modèles de régression dans chacune des cellules.

Comme pour l'analyse de variance, les logiciels opèrent une reparamétrisation faisant apparaître des effets différentiels par rapport au dernier niveau (SAS/GLM, SAS/INSIGHT) ou par rapport à un effet moyen (Systat), afin d'obtenir directement les bonnes hypothèses dans les tests. Ainsi, dans le premier cas, on considère la matrice de même rang (sans la Jème indicatrice)

$$\mathbf{X} = [\mathbf{1}|\mathbf{x}|\mathbf{1}_1|\cdots|\mathbf{1}_{J-1}|\mathbf{x}.\mathbf{1}_1|\cdots|\mathbf{x}.\mathbf{1}_{J-1}]$$

associée aux modèles :

$$y_{ij} = \beta_{0J}+(\beta_{0j}-\beta_{0J})+\beta_{1J}x_{ij}+(\beta_{1j}-\beta_{1J})x_{ij}+\varepsilon_{ij}\,; \quad j = 1, \ldots, J-1; i = 1, \ldots, n_j.$$

5.2 Tests

Différentes hypothèses sont alors testées en comparant le modèle complet

$$\begin{aligned}\mathbf{y} = \beta_{0J}\mathbf{1} \quad &+ \quad (\beta_{01}-\beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1}-\beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \\ &+ \quad (\beta_{11}-\beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1}-\beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u}\end{aligned}$$

à chacun des modèles réduits :

(i) $\mathbf{y} = \beta_{0J}\mathbf{1} + (\beta_{01}-\beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1}-\beta_{0J})\mathbf{1}_{J-1} + \beta_{1J}\mathbf{x} + \mathbf{u}$

(ii) $\mathbf{y} = \beta_{0J}\mathbf{1} + (\beta_{01}-\beta_{0J})\mathbf{1}_1 + \cdots + (\beta_{0J-1}-\beta_{0J})\mathbf{1}_{J-1} + +(\beta_{1j}-\beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1}-\beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u}$

(iii) $\mathbf{y} = \beta_{0J}\mathbf{1} + \beta_{1J}\mathbf{x} + (\beta_{1j}-\beta_{1J})\mathbf{x}.\mathbf{1}_1 + \cdots + (\beta_{1J-1}-\beta_{1J})\mathbf{x}.\mathbf{1}_{J-1} + \mathbf{u}$

par un test de Fisher. Ceci revient à considérer les hypothèses suivantes :

- H_0^i : pas d'interaction, $\beta_{11} = \cdots = \beta_{1J}$, les droites partagent la même pente β_{1J},
- H_0^{ii} : β_{1J}=0,
- H_0^{iii} :$\beta_{01} = \cdots = \beta_{0J}$, les droites partagent la même constante à l'origine β_{0J}.

On commence donc par évaluer i), si le test n'est pas significatif, on regarde ii) qui, s'il n'est pas non plus significatif, conduit à l'absence d'effet de la variable X. De même, toujours si i) n'est pas significatif, on s'intéresse à iii) pour juger de l'effet du facteur T.

5.3 Cas général

Ce cadre théorique et les outils informatiques (SAS/GLM) permettent de considérer des modèles beaucoup plus complexes incluant plusieurs facteurs, plusieurs variables quantitatives, voire des polynômes de celles-ci, ainsi que les diverses interactions entre qualitatives et quantitatives. Le choix du "bon" modèle devient vite complexe d'autant que la stratégie dépend, comme pour la régression linéaire multiple, de l'objectif visé :

descriptif : des outils multidimensionnels descriptifs (ACP, AFD, AFCM...) s'avèrent souvent plus efficaces pour sélectionner, en première approche, un sous-ensemble de variables explicatives avant d'opérer une modélisation,

explicatif : de la prudence est requise d'autant que les hypothèses ne peuvent être évaluées de façon indépendante surtout si, en plus, des cellules sont déséquilibrées ou vides,

prédictif : la recherche d'un modèle efficace, donc parcimonieux, peut conduire à négliger des interactions ou effets principaux lorsqu'une faible amélioration du R^2 le justifie et même si le test correspondant apparaît comme significatif. L'utilisation du C_p est possible mais en général ce critère n'est pas calculé et d'utilisation délicate pour définir ce qu'est le "vrai" modèle de référence. En revanche, le PRESS donne des indications pertinentes.

6 Exemple

6.1 Les données

Les données, extraites de Jobson (1991), sont issues d'une étude marketing visant à étudier l'impact de différentes campagnes publicitaires sur les ventes de différents aliments. Un échantillon ou "panel" de familles a été constitué en tenant compte du lieu d'habitation ainsi que de la constitution de la famille. Chaque semaine, chacune de ces familles ont rempli un questionnaire décrivant les achats réalisés.

Nous nous limitons ici à l'étude de l'impact sur la consommation de lait de quatre campagnes diffusées sur des chaînes locales de télévision. Quatre villes, une par campagne publicitaire, ont été choisies dans cinq différentes régions géographiques. Les consommations en lait par chacune des six familles par ville alors été mesurées (en dollars) après deux mois de campagne.

Les données se présentent sous la forme d'un tableau à 6 variables : la région

géographique, les 4 consommations pour chacune des villes ou campagnes publicitaires diffusées, la taille de la famille.

6.2 Analyse de variance à un facteur

Une première étude s'intéresse à l'effet du simple facteur "type de campagne publicitaire". On suppose implicitement que les familles ont été désignées aléatoirement indépendamment de l'appartenance géographique ou de leur taille. La procédure SAS/ANOVA est utilisée dans le programme suivant. Elle plus particulièrement adaptée aux situations équilibrées comme c'est le cas pour cet exemple. Le cas déséquilibré ne pose pas de problème majeur pour un modèle à un facteur mais pour deux facteurs ou plus, un message signale que les résultats sont fournis sous la responsabilité de l'utilisateur. Dans ce cas, la procédure plus générale SAS/GLM doit être utilisée.

Après une réorganisation des données permettant de construire une nouvelle variable décrivant le facteur "pub" ainsi que la variable unique consommation, le programme suivant a été exécuté :

```
title;
options pagesize=66 linesize=110 nonumber nodate;
proc anova data=sasuser.milkcc;
class pub;
model consom=pub;
means pub/bon scheffe tukey;
run;
```

SAS/ANOVA estime les paramètres du modèle d'analyse de variance à un facteur puis présente ensuite les résultats des tests de comparaison multiple demandés en option. Cette procédure signale explicitement que des problèmes peuvent apparaître si certains tests, spécifiques au cas équirépété, sont utilisés hors de leur contexte. Différentes options de présentation des résultats sont proposées : tests avec niveau paramétrable (5% par défaut) de significativité, intervalles de confiance des différences ou des moyennes.

Dans cet exemple, une des trois procédures de tests utilisée ne conclut pas aux mêmes résultats. Les tests de Scheffe acceptent tous l'hypothèse H_0 d'égalité des différentes moyennes. on retrouve ainsi le caractère conservatif de cette procédure.

La procédure SAS/NPAR1WAY a ensuite été exécutée pour obtenir les résultats des test non-paramétriques.

```
proc npar1way data=sasuser.milkcc;
class pub;
var consom;
run;
```

Les résultats sont encore "mitigés".

Source	DF (1)	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	4585.68048667(2)	1528.56016222(5)	3.16(7)	0.0275(8)
Error	116	56187.44398000(3)	484.37451707(6)		
Corrected Total	119	60773.12446667(4)			

R-Square	C.V.	Root MSE	CONSOM Mean
0.075456(12)	54.05283(11)	22.00851011(9)	40.71666667(10)

(1)	degrés de liberté pour le calcul des moyennes et la sélection de la loi de Fisher du test global
(2)	SSB
(3)	SSW
(4)	SST=SSW+SSB
(5)	SSB/DF
(6)	s^2 =MSE=SSW/DF est l'estimation de σ_u^2
(7)	Statistique F du test de Fisher du modèle global
(8)	$P(f_{p;n-p-1} > F)$; H_0 est rejetée au niveau α si $P < \alpha$
(9)	s =racine de MSE
(10)	moyenne empirique de la variable à expliquée
(11)	Coefficient de variation $100\times$ (9)/(10)
(12)	Coefficient de détermination R^2

```
Tukey's Studentized Range (HSD)
Alpha= 0.05  df= 116  MSE= 484.3745
Minimum Significant Difference= 14.813
Means with the same letter are not significantly different.
```

Tukey Grouping	Mean	N	PUB
A	51.030	30	4
A			
B A	39.647	30	2
B A			
B A	37.239	30	1
B			
B	34.951	30	3

Test non-paramétrique

```
Wilcoxon Scores (Rank Sums) for Variable CONSOM
Classified by Variable PUB
                  Sum of      Expected      Std Dev         Mean
PUB         N      Scores      Under H0      Under H0        Score
1          30   1675.00000        1815.0    164.999427   55.8333333
2          30   1781.50000        1815.0    164.999427   59.3833333
3          30   1562.50000        1815.0    164.999427   52.0833333
4          30   2241.00000        1815.0    164.999427   74.7000000

Kruskal-Wallis Test (Chi-Square Approximation)
CHISQ =  7.3266    DF =  3    Prob > CHISQ = 0.0622
```

6.3 Modèle à deux facteurs

Une étude graphique préalable des interactions est toujours instructive :

```
proc means data=sasuser.milkcc mean stderr;
class pub region;
var consom;
output out=cellmoy mean=moycons;
run;
symbol i=join v=dot cv=black ;
symbol2 i=join v=% cv=black h=2;
symbol3 i=join v='"' cv=black h=2;
symbol4 i=join v=# cv=black h=2;
symbol5 i=join v=$ cv=black h=2;%$
proc gplot data=cellmoy;
plot moycons*region=pub;
run;
goptions reset=all; quit;
```

Nous sommes dans le cas équirépété, la procédure SAS/ANOVA reste valide mais SAS/GLM, plus générale, est utilisée et fournit dans ce cas les mêmes résultats. Cette procédure adaptée aux situations complexes fournit également d'autres options (contrastes, estimation des paramètres...).

```
title;
options pagesize=66 linesize=110 nonumber nodate;
proc glm data=sasuser.milkcc;
class pub region;
model consom= pub region pub*region;
```

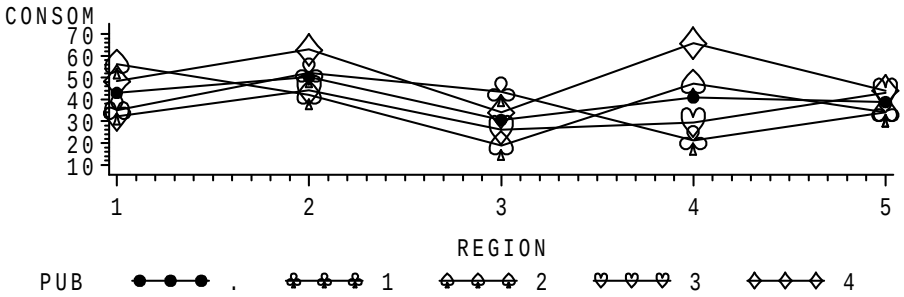

FIGURE 2 – Profil moyen et profils de la consommation moyenne de chaque région en fonction du type de campagne.

```
run;
```

```
General Linear Models Procedure
 (0)
Source             DF      Sum of Squares      Mean Square    F Value    Pr > F
Model              19      18391.10933333     967.95312281     2.28      0.0045
Error              100     42382.01513333     423.82015133
Corrected Total    119     60773.12446667

     R-Square         C.V.          Root MSE          CONSOM Mean
     0.302619     50.56134     20.58689271          40.71666667

Source             DF      Type III SS         Mean Square    F Value    Pr > F
                   (1)                              (5)          (6)        (7)
PUB                 3      4585.68048667(2)   1528.56016222     3.61      0.0160
REGION              4      4867.51141667(3)   1216.87785417     2.87      0.0268
PUB*REGION         12      8937.91743000(4)    744.82645250     1.76      0.0658
```

(0)	Tableau associé au test global de nullité de tous les paramètres.
(1)	Degrés de liberté pour le calcul des moyennes et sélection de la loi de Fisher.
(2)	SS1
(3)	SS2
(4)	SSI
(5)	SS1,2,I/DF
(6)	Statistique F pour chacun des tests
(7)	$P(f_{p;n-p-1} > F)$; H_i est rejetée au niveau α si $P < \alpha$

6.4 Analyse de covariance

La variable “taille” est quantitative. On s’intéresse à différents modèles de régression visant à expliquer la consommation en fonction de la taille de la famille conditionnellement au type de campagne publicitaire.

```
proc glm data=sasuser.milk;
class pub;
model consom=pub taille pub*taille;
run;
```

Les résultats ci-dessous conduiraient à conclure à une forte influence de la taille mais à l’absence d’influence du type de campagne. Les droites de régression ne semblent pas significativement différentes.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
PUB	3	227.1807	75.7269	0.57	0.6377 (1)
TAILLE	1	40926.0157	40926.0157	306.57	0.0001 (2)
TAILLE*PUB	3	309.8451	103.2817	0.77	0.5111 (3)

(1)	Test de la significativité des différences des termes constants.
(2)	Test de l’influence du facteur quantitatif.
(3)	Test de la significativité des différences des pentes (interaction).

Néanmoins, compte tenu des résultats précédents (analyse de variance), le même calcul est effectué pour chaque région :

```
proc glm data=sasuser.milk;
by region;
class pub;
model consom=pub taille pub*taille;
run;
```

Région	Source	DF	Type III SS	Mean Square	F Value	Pr > F
1	PUB	3	72.02974	24.00991	4.62	0.0164
	TAILLE	1	7178.32142	7178.32142	1380.25	0.0001
	TAILLE*PUB	3	217.37048	72.45683	13.93	0.0001
2	PUB	3	231.73422	77.24474	30.36	0.0001
	TAILLE	1	8655.25201	8655.25201	3402.34	0.0001
	TAILLE*PUB	3	50.15069	16.71690	6.57	0.0042
3	PUB	3	79.54688	26.51563	6.01	0.0061
	TAILLE	1	6993.30160	6993.30160	1585.35	0.0001
	TAILLE*PUB	3	173.19305	57.73102	13.09	0.0001
4	PUB	3	415.66664	138.55555	15.23	0.0001
	TAILLE	1	9743.37830	9743.37830	1071.32	0.0001
	TAILLE*PUB	3	361.39556	120.46519	13.25	0.0001
5	PUB	3	15.35494	5.11831	0.79	0.5168
	TAILLE	1	8513.28516	8513.28516	1314.71	0.0001
	TAILLE*PUB	3	52.75119	17.58373	2.72	0.0793

Il apparaît alors qu’à l’intérieur de chaque région (sauf région 5), les campagnes de publicité ont un effet tant sur la constante que sur la pente.

Ceci incite donc à se méfier des *interactions* et encourage à toujours conserver le facteur bloc dans une analyse. Une approche complète, considérant *a priori* toutes les variables (3 facteurs), est ici nécessaire (cf. TP).