STT-4100 / STT-7230

Planification des expÃ©riences

Notes de cours

Automne 2015

Professeur : Lajmi LAKHAL–CHAIEB
lakhal@mat.ulaval.ca

Notes de cours crÃ©Ã©es par

Louis-Paul RIVEST

Emmanuelle RENY-NOLIN

et rÃ©visÃ©es par

Lajmi LAKHAL–CHAIEB

DÃ©partement de mathÃ©matiques et de statistique


UNIVERSITÉ
LAVAL

Table des matières

1 Introduction et rappels

Ce chapitre vise à présenter les notions de base de la planification d'une expérience, et à réviser certains concepts déjà abordés lors des cours de première et deuxième année qui seront utiles à notre apprentissage.

Rappelons d'abord que l'analyse statistique qui servira à répondre aux questions de recherche est intrinsèquement liée à la façon dont les données ont été récoltées, c'est-à-dire au plan de l'expérience. Le choix des variables à mesurer, les modalités des facteurs testées, le mode de sélection des unités expérimentales, le nombre d'observations, le protocole d'application des traitements et l'ordre de collecte sont tous des aspects qui auront une influence sur la façon dont les résultats devront être analysés. Il est donc primordial d'étudier les méthodes d'analyse en profondeur afin de comprendre les impacts de ces choix et de bien conseiller un chercheur avant la collecte de ses données.

1.1 Objectif d'une expérience

Il va de soi que nous nous attarderons aux expérimentations dont les conditions peuvent être contrôlées, du moins partiellement. Il existe cependant plusieurs niveaux de contrôle, et l'importance de l'expérience détermine les paramètres sur lesquels il faudra insister.

Les *expériences exploratoires* sont des projets-pilotes visant par exemple à sélectionner des modalités de traitements (comme des doses de médicaments ou des variétés de céréales) ou à tester un protocole d'observation. On leur alloue souvent des ressources limitées. Dans le domaine médical, c'est la première phase des essais cliniques, souvent réalisée sur des animaux ou sur un nombre restreint de patients.

Les *expériences principales* sont au coeur de la recherche et bénéficient de beaucoup d'attention (i.e. de budget et de temps). Ces expériences sont réalisées dans des conditions contrôlées de façon très stricte, et par conséquent sont souvent menées en laboratoire ou dans un serre ou un hôpital, tout comme les expériences préliminaires.

Finalement, les *expériences confirmatoires* visent à vérifier dans des conditions les plus réelles possibles les conclusions de l'étude principale. Dépendant du domaine d'étude, il s'agira d'une expérience en champ à grande échelle, ou en forêt, ou sur une grande population de patients faisant usage d'un médicament. Ces expériences sont souvent conduites sur de longues périodes, ou sur de vastes territoires. (Dagnélie [4], p. 27)

1.2 Étapes de planification d'une expérience

La connaissance du sujet d'étude est fondamentale dans l'élaboration d'un protocole expérimental rigoureux. Si le contexte ne nous permet pas de faire une revue de littérature des études déjà effectuées sur des questions similaires, il est primordial d'impliquer étroitement le chercheur dans la planification. Le meilleur moment pour poser des questions, c'est avant d'aller sur le terrain...

1. Énoncer très clairement la question de recherche. On peut avoir plusieurs objectifs, d'importances égales ou différentes.

2. Identifier une variable réponse qui servira d'indicateur pour répondre à la question.

3. Déterminer les facteurs à étudier et leurs modalités, les effets d'autres variables à contrôler, les interactions à prendre en compte.
   - Si les facteurs sont qualitatifs avec peu de modalités (ex : race de bovins, variété de basilic), on dira qu'il s'agit de facteurs fixes et les niveaux à comparer font partie de l'objectif même de l'étude. Si le nombre de modalités possibles est très grand, il faut faire un choix, souvent aléatoire, ce qui conditionnera le statut du facteur (ex : les cliniques médicales où seront sélectionnés les patients, les opérateurs de la machinerie).
   - Si les facteurs sont quantitatifs (ex : doses de médicaments, humidité du sol), on peut choisir quelques valeurs réparties uniformément ou progressivement sur le domaine de la variable. L'aspect numérique de ces variables pourra être pris en compte dans l'analyse pour localiser des extrema ou vérifier des tendances paramétriques (polynomiales, exponentielles, ...etc.).
   - Si le nombre de combinaisons de modalités est trop élevé pour les ressources disponibles, il est aussi possible de conduire des expériences factorielles dites *fractionnaires*, où seulement un sous-ensemble des traitements est étudié. Nous porterons dans ce cours une attention plus grande aux expériences factorielles complètes.

4. Identifier les unités expérimentales (celles sur lesquelles le traitement sera appliqué) et les unités d'observation (celles sur lesquelles la mesure sera prise). Ces unités peuvent prendre des formes très variées selon le domaine d'étude et le type de facteur pris en compte. Une unité expérimentale peut être une parcelle de forêt ou un seul arbre ou encore une feuille. Si le traitement est appliqué simultanément à un groupe d'individus (ex : une méthode d'enseignement, un épandage

aérien d'engrais), le groupe entier constitue l'unité expérimentale. Si un individu reçoit plusieurs traitements, on considérera chaque combinaison personne-traitement comme une unité expérimentale. Il est en général souhaitable d'avoir le plus grand nombre d'unités expérimentales possible, mais des contraintes pratiques peuvent nuire à cet objectif : le coût, le temps, ou le type de facteur ou de variable réponse.

5. Choisir un plan d'expérience approprié, i.e. déterminer comment les traitements seront alloués aux unités expérimentales. C'est ici qu'on prend en considération les principes de randomisation, de blocage et de répétition dont nous discuterons plus loin.

6. Déterminer le nombre d'unités expérimentales dans chaque traitement. Généralement, ce calcul de taille d'échantillons se base sur une puissance souhaitée pour une différence de traitements, dans la limite du budget disponible.

7. Établir le modèle statistique correspondant au plan. Au cours de l'analyse, les facteurs non significatifs seront généralement conservés dans le modèle, car c'est le plan de l'expérience qui dicte les sources de variation. On ne cherchera pas à prédire avec précision la valeur de la variable réponse, mais à bien étudier les différences de moyennes engendrées par les combinaisons de traitements.

8. Collecter les données de la façon la plus objective possible. Il est très important de s'en tenir au plan.

9. Valider le modèle et obtenir des conclusions pertinentes à la question de recherche. Chaque modèle statistique s'accompagne de postulats, par exemple la normalité des erreurs ou l'homogénéité des variances d'un traitement à l'autre. C'est ici qu'intervient un test d'adéquation du modèle en analysant les résidus.

10. Faire de l'inférence sur les bonnes populations (intervalles de confiance, tests d'hypothèses) et interpréter les conclusions de façon claire pour le scientifique.

1.3 Principes de base de la planification d'expériences

Randomisation : allocation aléatoire des traitements aux unités expérimentales et/ou de l'ordre dans lequel les unités recevront les traitements. Peut être effectuée à l'aide de tables (ex. Fisher et Yates) ou de procédures informatiques (ex. Proc PLAN de SAS ou librairie `blockrand` de R). La randomisation permet de minimiser la dépendance entre les données, d'éliminer les effets indésirables de facteurs

non pris en compte dans l'étude, de réduire les erreurs systématiques résultant de l'allocation des traitements (comme les caractéristiques communes des premiers individus arrivés ou des parcelles en bordure de champ), ou de l'usure des instruments, de l'apprentissage des sujets, pouvant affecter la variable réponse.

Blocage : séparation des unités expérimentales non homogènes en groupes homogènes appelés blocs. C'est une restriction à la randomisation, mais cela permet d'augmenter la précision des tests en partitionnant l'erreur aléatoire en deux : la variabilité entre les blocs, et l'erreur résiduelle à l'intérieur des blocs. La randomisation complète n'est souhaitable que si les unités sont très semblables. Les blocs sont en général inclus comme un facteur aléatoire dans le modèle statistique, d'où la réduction de l'erreur résiduelle et le gain de puissance.

Réplication : application non simultanée du même traitement à plusieurs unités expérimentales. Permet d'estimer la variance de l'erreur expérimentale, i.e. la variabilité entre les unités expérimentales. Il faut faire attention aux pseudo-réplications, i.e. à de multiples mesures prises sur des unités d'observations provenant de la même unité expérimentale. Par exemple, si un engrais est mélangé à la terre d'un plant de tomates, l'unité expérimentale est le plant. Même si on prend des mesures sur 10 tomates du même plant, il ne s'agit que d'une réplication pour le traitement "engrais".

Nombre d'observations : Comme dans toute situation d'inférence statistique, la puissance des tests (ou la précision des intervalles de confiance) est affectée par le nombre d'observations, le seuil des tests (ou le niveau de confiance des intervalles) et par la variabilité naturelle des mesures. Il peut être parfois nécessaire de conduire des expériences pilotes pour estimer l'ordre de grandeur de la variance des observations, afin de déterminer le bon nombre d'observations permettant d'atteindre une puissance souhaitée pour un scénario en particulier.

Les calculs de puissance deviennent rapidement très compliqués lorsque le plan d'expérience est élaboré. On ramène souvent le plan à un facteur (le plus important) pour estimer la puissance des tests et déterminer les tailles d'échantillons.

Mesures objectives : Autant que possible, on mesure la variable réponse à l'aide d'un appareil ou d'un instrument ne dépendant pas de l'observateur. Lorsque la variable à mesurer est une évaluation visuelle faite par un expert, il est essentiel

que ce dernier ne sache pas quel traitement a été appliqué à l'unité expérimentale qu'il évalue. Si l'unité expérimentale est un individu ayant reçu un médicament ou goûté un produit par exemple, elle doit ignorer la catégorie de traitement auquel elle appartient. Lorsque le sujet et l'observateur ne connaissent pas le traitement, on parle d'étude à double-insu ou en double-aveugle.

On tentera de garder invariantes les conditions expérimentales, i.e. les facteurs pouvant influencer la réponse mais non contrôlés par l'étude, pour éliminer le plus possible les variables confondantes.

Mesures limitrophes : Toutes les observations mesurées ne sont pas nécessairement équivalentes et non nuisibles dans l'analyse. En effet, un dispositif expérimental est souvent plus petit que le territoire réel dans lequel évolue la population. Par exemple, des plantes situées sur des parcelles de champ adjacentes peuvent s'affecter mutuellement. Le rendement réel des parcelles respectives est mieux représenté par les plantes centrales que par l'ensemble des plantes. Il vaut parfois mieux laisser une bordure dont les observations seront exclues de l'analyse. Il en va de même pour des individus qui feraient des tests de goût : il faut leur donner du temps (ou du pain !) entre deux "traitements"...

Nature et lien des facteurs entre eux : un facteur est-il fixe ou aléatoire ? Deux facteurs sont-ils croisés ou emboîtés ? Certaines variables doivent parfois être mesurées et incluses dans le modèle statistique même si elles ne font pas partie de l'étude comme telle, car elles influencent la valeur de la variable réponse. Par exemple, le poids d'un individu peut influencer sa pression artérielle, ou la quantité de pluie tombée peut influencer le rendement d'une parcelle de maïs. Ces questions seront récurrentes dans le cours.

1.4 Plans d'expériences considérés connus

Pour la suite de ces notes, nous considérerons que les schémas d'expériences suivants ne vous sont pas inconnus. Nous reviendrons sur certains d'entre eux, parfois en les abordant sous un angle un peu différent de l'analyse classique des cours d'introduction. Les exercices à la fin du chapitre vous permettront de revoir certains d'entre eux à travers des mises en situation concrètes.

Expériences simples

- comparaison d'une moyenne avec une constante
- comparaison de deux moyennes sur des populations indépendantes avec variances égales
- comparaison de deux moyennes sur des populations indépendantes avec variances inégales
- comparaison de deux moyennes sur des données appariées

Plan complètement aléatoire

- analyse de la variance (anova) à un facteur fixe
- plan factoriel à plusieurs facteurs fixes avec interactions
- anova à un facteur aléatoire
- anova à plusieurs facteurs aléatoires avec interactions
- modèle mixte (au moins un facteur fixe et un facteur aléatoire)

Plan avec des restrictions à la randomisation

- Plans à blocs aléatoires complets
- Plan en carré latin

1.5 Loi normale multivariée

1.5.1 Rappels sur les lois univariées normale, chi-deux, t et de Fisher

Soit Z une variable aléatoire réelle de fonction de densité

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}.$$

On dit que Z suit une loi normale standard et on écrit $Z \sim \mathcal{N}(0,1)$. On a alors $E(Z) = 0$ et $\text{Var}(Z) = 1$ et la fonction génératrice des moments de Z est égale à $M_Z(t) = E(e^{Zt}) = \exp\{t^2/2\}$ pour tout nombre réel t.

Soient μ et σ deux nombres réels et posons $X = \mu + \sigma Z$. On dit alors X suit une loi normale de moyenne μ et de variance σ^2 et on écrit $X \sim \mathcal{N}(\mu, \sigma^2)$. La fonction génératrice des moments de X est égale à $M_X(t) = E(e^{tX}) = \exp\{\mu t + \sigma^2 t^2/2\}$ pour tout nombre réel t.

Soit $\{Z_1, \cdots, Z_n\}$ un échantillon de n variables indépendantes identiquement distribuées selon $\mathcal{N}(0,1)$. Pour $i = 1, \cdots, n$, on a alors Z_i^2 suit une loi de chi-deux à 1 degrés de liberté et on écrit $Z_i^2 \sim \chi_1^2$. En plus, $Y_n = Z_1^2 + \cdots + Z_n^n$ suit une loi de chi-deux à n degrés de liberté et on écrit $Y_n \sim \chi_n^2$.

Soient maintenant $Z \sim \mathcal{N}(0,1)$ et $Y_n \sim \chi_n^2$ deux variables aléatoires indépendantes et posons

$$U_n = \frac{Z}{\sqrt{Y_n/n}}.$$

On a alors U_n suit une loi t de $student$ à n degrés de liberté et on écrit $U_n \sim t_n$.

Finalement, soient $Y_n \sim \chi_n^2$ et $Y_m \sim \chi_m^2$ deux variables aléatoires indépendantes et posons

$$V_{n,m} = \frac{Y_n/n}{Y_m/m}.$$

On a alors $V_{n,m}$ suit une loi F de Fisher à n et m degrés de liberté et on écrit $V_{n,m} \sim F_{n,m}$.

D'autres propriétés et caractéristiques de ces distributions se trouvent dans un bon nombre de livres de statistique mathématique [25].

1.5.2 Loi normale multivariée

Dans ce qui suit, les symboles en gras désignent des vecteurs et des matrices.

Soient n variables aléatoires $\{X_1, \cdots, X_n\}$. On dit que le vecteur aléatoire $\mathbf{X} = (X_1, \cdots, X_n)^\top$ suit une loi normale multivariée si et seulement si toute combinaison linéaire des X_is suit une loi normale univariée. Autrement dit, si et seulement si $\mathbf{a}^\top \mathbf{X} = a_1 X_1 + \cdots + a_n X_n$ suit une loi normale univariée pour tout vecteur $\mathbf{a} = (a_1, \cdots, a_n)^\top$.

Une loi normale multivariée est caractérisée par un vecteur de moyennes $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix}$ où $\mu_i = E(X_i)$ pour tout $i = 1, \cdots, n$ et une matrice de variance-covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{2n} & \cdots & \sigma_n^2 \end{pmatrix}$$

où $\sigma_i^2 = \mathrm{Var}(X_i)$ pour tout $i = 1, \cdots, n$ et $\sigma_{ij} = \mathrm{Cov}(X_i, X_j)$ pour tous $i \neq j$.

Sous forme matricielle, on écrit $\boldsymbol{\mu} = E(\mathbf{X})$, $\boldsymbol{\Sigma} = E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^\top] = E(\mathbf{X}\mathbf{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}^\top$ et $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ où l'indice n indique que la loi normale est de dimension n.

Pour être une matrice de variance covariance, $\boldsymbol{\Sigma}$ doit satisfaire deux conditions. Primo, elle doit être sysmétrique puisque $\text{Var}(X_i, X_j) = \text{Var}(X_j, X_i)$ pour tous $i \neq j$. Secondo, elle doit être semi-définie positive. En effet, soit $\mathbf{a} = (a_1, \cdots, a_n)$ un vecteur de nombres réels et posons $Y = \mathbf{a}^\top\mathbf{X} = a_1X_1 + \cdots a_nX_n$. On a alors $\text{Var}(Y) = \mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a}$. Cette variance soit être positive ou nulle. Ainsi, une matrice $\boldsymbol{\Sigma}$ est dite semi-définie positive si et seulement si $\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a} \geq 0$ pour tout vecteur $\mathbf{a} = (a_1, \cdots, a_n)^\top$ de nombres réels. La matrice est dite définie positive si l'égalité $\mathbf{a}^\top\boldsymbol{\Sigma}\mathbf{a} = 0$ n'est possible que si $\mathbf{a} = (0, \cdots, 0)^\top$ est un vecteur nul.

Toute matrice définie positive est inversible (son déterminant est strictement positif) et son inverse est également définie positive.

Par exemple, en dimension $n = 2$, on vérifie que $\mathbf{A_1}$ n'est pas semi-définie positive, que $\mathbf{A_2}$ et $\mathbf{A_3}$ le sont et que seule $\mathbf{A_3}$ est définie positive, où

$$\mathbf{A_1} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \mathbf{A_2} = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} \text{ et } \mathbf{A_3} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

Soit $\{X_1, \cdots, X_n\}$ un échantillon d'observation indépendantes et identiquement distribuées provenant d'une distribution $\mathcal{N}_1(\mu, \sigma^2)$ alors le vecteur $\mathbf{X} = (X_1, \cdots, X_n)^\top$ suit une loi normale multivariée $\mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ avec $\boldsymbol{\mu} = (\mu, \cdots, \mu)^\top$ et $\boldsymbol{\Sigma} = \sigma^2\mathbf{I_n}$ où $\mathbf{I_n}$ est la matrice identité de dimension n.

Dans ce qui suit, on présente plusieurs propriétés de la loi normale multivariée.

Propriété 1 : Fonction génératrice des moments

Si $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, alors pour $\mathbf{t} \in \mathcal{R}^n$,

$$M_{\mathbf{X}} = E[\exp(\mathbf{t}^\top\mathbf{X})] = \exp(\mathbf{t}^\top\boldsymbol{\mu} + \mathbf{t}^\top\boldsymbol{\Sigma}\mathbf{t}/2)$$

La preuve de ce résultat est basée sur la décomposition spectrale de $\boldsymbol{\Sigma}$ *et ne sera pas présentée dans ce cours.*

Propriété 2 : Transformation linéaire

Soient $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{A}$ une matrice $m \times n$ et $\mathbf{b}$ un vecteur de dimension m. On a alors $\mathbf{AX+b} \sim \mathcal{N}_m(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top)$.

Preuve : il suffit de calculer la fonction génératrice des moments de $\mathbf{AX+b}$.

Propriété 3 : Standardisation

Soit $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Si $\boldsymbol{\Sigma}$ est définie-positive, alors on a

$$\mathbf{Z} = (\boldsymbol{\Sigma}^{-1/2})^\top (\mathbf{X} - \boldsymbol{\mu}) \sim \mathcal{N}_n(\mathbf{0}, \mathbf{I_n})$$

Preuve : Cas particulier de la propriété précédente.

Propriété 4 : Fonction de densité

Soit $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. La fonction de densité de $\mathbf{X}$ est

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}}.$$

Preuve : On pose $\boldsymbol{Z} = (\boldsymbol{\Sigma}^{-1/2})^\top (\boldsymbol{X} - \boldsymbol{\mu})$. D'après ce qui précéde, les composantes de $\boldsymbol{Z}$ sont indépendantes et suivent chacune $\mathcal{N}_1(0,1)$. La fonction de densité du vecteur aléatoire $\boldsymbol{Z}$ est donc donnée par

$$f_{\boldsymbol{Z}}(\boldsymbol{z}) = f_{Z_1}(z_1) \times \cdots \times f_{Z_n}(z_n) = \frac{1}{\sqrt{2\pi}} e^{-z_1^2/2} \times \cdots \times \frac{1}{\sqrt{2\pi}} e^{-z_n^2/2} = \exp(-\boldsymbol{z}^\top \boldsymbol{z}/2)/(2\pi)^{n/2}.$$

On obtient l'expression de $f_{\boldsymbol{X}}(\boldsymbol{x})$ en effectuant un changement de variables $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{Z}$, pour lequel le jacobien est $|\boldsymbol{\Sigma}^{-1/2}| = 1/|\boldsymbol{\Sigma}|^{1/2}$.

Propriété 5 : Fonction de répartition et calcul de probabilités

En dimension $n = 1$, la fonction de répartition de la loi normale n'a pas d'expression explicite et on calcule les probabilités du type $P(a \leq X \leq b)$ à l'aide de tables ou logiciels statistiques. Lorsque $n > 1$, ces calculs de probabilités se compliquent. On calcule les probabilités du type $P(a_1 \leq X_1 \leq b_1, \cdots, a_n \leq X_n \leq b_n)$ à l'aide de logiciels statistiques. Par exemple, en R, les fonctions *pmvnorm* et *pmnorm* des librairies mvtnorm et mnormt, respectivement, effectuent ces calculs. Cependant, c'est de plus en plus difficile de calculer ces probabilités avec une bonne précision lorsque n augmente [6].

Propriété 6 : Lois marginales et conditionnelles

Soit $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$. Si $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ se décompose en deux sous vecteurs aléatoires $\mathbf{X}_1$ et $\mathbf{X}_2$ de dimensions respectives n_1 et n_2 $(n_1 + n_2 = n)$ et si $\boldsymbol{\mu}$ et $\boldsymbol{\Sigma}$ se décomposent selon

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \text{ et } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}$$

alors on a

(i) $\mathbf{X}_1 \sim \mathcal{N}_{n_1}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ et $\mathbf{X}_2 \sim \mathcal{N}_{n_2}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$

(ii) $\mathbf{X}_1$ et $\mathbf{X}_2$ sont indépendants si et seulement si $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}^\top = \mathbf{0}$.

(iii) Si $\boldsymbol{\Sigma}_{11}$ est définie positive alors la distribution conditionnelle de $\mathbf{X}_2$ sachant $\mathbf{X}_1 = \mathbf{x}_1$ est $\mathcal{N}_{n_2}\{\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\}$

Preuve : Les points **(i)** *et* **(ii)** *se démontrent en appliquant la propriété 2 et en utilisant la fonction génératrice des moments, respectivement.*

Pour démontrer **(iii)**, *on pose*

$$\boldsymbol{Y} = \begin{pmatrix} \mathbf{I}_{n_1} & \boldsymbol{0} \\ -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}$$

avec $\boldsymbol{Y}_1 = \boldsymbol{X}_1$ *et* $\boldsymbol{Y}_2 = \boldsymbol{X}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$. *D'après la propriété 2,*

$$\boldsymbol{Y} \sim \mathcal{N}_n \left[\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \end{pmatrix} \right]$$

D'après les points **(i)** *et* **(ii)**, $\boldsymbol{Y}_2$ *et* $\boldsymbol{Y}_1 = \boldsymbol{X}_1$ *sont indépendants et* $\boldsymbol{Y}_2 \sim \mathcal{N}_{n_2}(\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$. *On en déduit la distribution conditionnelle de* $\boldsymbol{X}_2 = \boldsymbol{Y}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{X}_1$ *sachant* $\boldsymbol{X}_1 = \boldsymbol{x}_1$.

Propriété 7 : Forme quadratique

Si $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ et $\boldsymbol{\Sigma}$ est définie positive alors la forme quadratique

$$Q = (\mathbf{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) \sim \chi^2_n$$

Preuve : Il suffit d'écrire $Q = \boldsymbol{Z}^\top \boldsymbol{Z} = \sum_{i=1}^n Z_i^2$ *avec* $\boldsymbol{Z} = (\boldsymbol{\Sigma}^{-1/2})^\top(\boldsymbol{X} - \boldsymbol{\mu})$.

Exemple : Analyse de la variance à un facteur aléatoire

Le modèle d'analyse de la variance à un facteur aléatoire s'écrit $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, $i = 1, \cdots, I$ et $j = 1, \cdots, n_i$ où I est le nombre de traitements, n_i la taille du $i^{\text{ème}}$ échantillon, $\tau_i \sim \mathcal{N}_1(0, \sigma_\tau^2)$ et $\epsilon_{ij} \sim \mathcal{N}_1(0, \sigma^2)$. Soit $\mathbf{Y}_i = (y_{i1}, \cdots, y_{in_i})^\top$ représentant les données du $i^{\text{ème}}$ échantillon. Sous forme matricielle, on écrit

$$\mathbf{Y}_i = \mu \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \tau_i \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$

Ces données sont dépendantes puisque $\text{Cov}(y_{ij}, y_{ik}) = \sigma^2_\tau$ si $j \neq k$. En appliquant la propriété 2, on montre que

$$\mathbf{Y}_i \sim \mathcal{N}_{n_i}\left[\mu \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \sigma^2_\tau & \sigma^2_\tau & \cdots & \sigma^2_\tau \\ \sigma^2_\tau & \sigma^2 + \sigma^2_\tau & \cdots & \sigma^2_\tau \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2_\tau & \sigma^2_\tau & \cdots & \sigma^2 + \sigma^2_\tau \end{pmatrix}\right].$$

1.6 Exercices

1. Vrai ou Faux ?

(a) Une pseudo-répétition est une duplication d'une ligne dans un jeu de données, faisant croire que la mesure a été prise plusieurs fois.

(b) Un facteur est considéré aléatoire quand il a plus de cinq modalités.

(c) La randomisation dans l'attribution des traitements permet de réduire la présence de variables confondantes.

(d) Plus il y a d'observations indépendantes dans les I échantillons d'une analyse de la variance à un facteur, plus la statistique F risque d'être élevée.

(e) Dans un modèle de régression linéaire simple, on suppose généralement que les observations proviennent d'une loi normale de moyenne 0 et de variance σ^2.

(f) On dispose du poids de 12 personnes avant et après un régime. La différence de poids moyens aura une variance supérieure si on considère les 24 données en deux groupes indépendants de variances égales (avant-après) que si on considère 12 paires de données (test-t apparié). La statistique T sera plus significative dans le test apparié.

(g) Une étude à double insu signifie que le chercheur ET le statisticien sont tenus dans l'ignorance par rapport aux traitements reçus par les unités échantillonnales.

(h) Pour comparer les moyennes de 4 niveaux d'un facteur avec une analyse de la variance, il faut au moins une observation par traitement.

(i) Un modèle d'analyse de la variance à deux facteurs fixes doit obligatoirement contenir un terme d'interaction.

(j) Si on a trois traitements à appliquer à des souris et que trois techniciens travaillent dans notre laboratoire, il est important que chaque technicien ne fasse l'application que d'un seul traitement.

2. Identifier des problèmes potentiels pouvant découler des situations suivantes.

(a) À la sortie d'un magasin, une personne en fauteuil roulant effectue un sondage auprès des clients sur les investissements publics pour améliorer l'accès des commerces aux handicapés.

(b) Question de sondage : *À quelle fréquence visitez-vous la bibliothèque municipale ?* Jamais ○ Parfois ○ Assez souvent ○ Très souvent ○

(c) Le Parti Pourlagloire envoie 2000 questionnaires par la poste pour connaître les intentions de vote des citoyens. Parmi les 35 questionnaires retournés, 80% indiquent une intention de vote pour le Parti Pourlagloire.

(d) Titre d'un article : Il y a autant de personnes plus intelligentes que la moyenne que de personnes moins intelligentes que la moyenne.

(e) Une compagnie pharmaceutique fait un appel de volontaires pour une étude clinique sur un médicament contre l'apnée du sommeil, par des publicités télévisées. Les 10 premiers participants répondants aux critères recevront le nouveau traitement développé par la compagnie, les 10 suivants des somnifères standards, et les 10 derniers un placebo.

(f) Un industriel veut savoir si le choix de l'opérateur d'une machine influe sur la qualité de la production, et ce pour les trois types de machines de son usine. Il choisit donc quatre employés au hasard, et les fait travailler sur chaque type de machine à deux reprises dans un ordre randomisé. Il mesure la qualité de la production pour chaque répétition, ce qui lui donne 24 mesures continues. Le statisticien analyse les données à l'aide d'un modèle d'anova à deux effets fixes avec interaction. Puisque l'interaction n'est pas significative, il peut interpréter le test comparant les moyennes entre opérateurs.

(g) Une étudiante en sociologie veut analyser la perception des étudiants du coût des études universitaires en fonction de diverses caractéristiques personnelles (sexe, âge, revenu, etc). Elle prépare un questionnaire de 50 questions à choix multiples, qu'elle distribue sur le campus dans diverses salles d'attente et points de service. Après un mois, elle récolte 9 questionnaires remplis et procède à l'analyse.

(h) Un kinésiologue veut savoir si la différence de force entre la main gauche et la main droite des droitiers est importante. Il sélectionne aléatoirement 20 adultes droitiers et mesure la force de leurs deux mains avec un appareil. Il ajuste ensuite un modèle de régression linéaire simple avec les paires de points recueillis, et vérifie si la pente de cette droite diffère de 1.

3. Le but d'une expérience est de comparer 3 médicaments expérimentaux avec le médicament actuellement sur le marché pour le traitement de l'hypertension. Au-delà du test de Fisher de l'analyse de la variance, toutes les comparaisons $\mu_{traitement\ i} - \mu_{actuel}$ seront testées pour savoir si elles diffèrent de 0. On suppose que la variance des mesures dans un groupe est σ^2. Vous disposez de 40 patients pour faire votre étude.

(a) Jacques propose de répartir les 40 unités également dans les 4 groupes de traitement. Marie suggère plutôt de donner davantage d'importance au groupe de référence, en lui accordant 19 patients, et 7 patients par groupe de traitement. Laquelle des deux répartitions est la plus efficace pour les comparaisons de cette étude ?

(b) Quelle est la répartition optimale pour que les comparaisons avec la moyenne des mesures dans le traitement actuel soient le plus précises possible ?

4. Un agriculteur veut savoir si le rendement moyen d'une nouvelle variété de blé est différent du rendement moyen de la variété qu'il utilise habituellement. Il dispose de 16 parcelles de champ aux propriétés similaires. Il en choisit 6 au hasard pour semer la nouvelle variété et les 10 autres sont utilisées pour la variété standard. À la fin de la saison, il obtient les rendements suivants en tonnes par hectare. (Exemple tiré de Mead [12].)

Variété	Rendement						Moyenne	Variance
Nouvelle	2.5	2.1	2.4	2.0	2.6	2.3	$\overline{x}_1 = 2.3$	$s_1^2 = 0.054$
Standard	2.2	1.9	1.8	2.1	2.1		$\overline{x}_2 = 2.0$	$s_2^2 = 0.047$
	1.7	2.3	2.0	1.7	2.2			

Proposez une analyse statistique de ces données qui permette de répondre à la question de l'agriculteur. Donnez l'équation du modèle et ses postulats, les hypothèses statistiques testées, et réalisez l'analyse avec le logiciel SAS. Énoncez clairement vos conclusions.

5. Une compagnie d'assurances doit décider si elle augmentera les primes de ses clients l'année prochaine. On décide d'examiner d'abord les réclamations des deux dernières années (année 1 et année 2) par un échantillonnage de 50 clients. Si la moyenne des réclamations augmente de façon significative de l'année 1 à l'année 2, la compagnie envisagera une hausse des primes pour l'année 3. Sinon, elle conservera les primes actuelles.

Client	an 1	an 2		an 1	an 2		an 1	an 2		an 1	an 2		an 1	an 2
1	777	372	11	576	607	21	990	896	31	938	851	41	1116	610
2	1323	773	12	184	484	22	705	498	32	661	193	42	777	628
3	1259	255	13	707	747	23	715	1282	33	277	975	43	218	448
4	394	908	14	684	635	24	806	651	34	564	717	44	1350	869
5	973	1186	15	917	713	25	1204	708	35	1012	755	45	686	1209
6	963	628	16	762	999	26	939	1435	36	943	776	46	421	637
7	459	1021	17	945	862	27	306	318	37	544	791	47	606	202
8	856	280	18	1000	1001	28	750	1152	38	1047	1130	48	591	340
9	913	392	19	98	883	29	336	921	39	231	1181	49	482	1003
10	1132	346	20	136	1354	30	368	516	40	319	958	50	960	1255

Proposez une analyse statistique de ces données qui réponde aux besoins de la compagnie d'assurances. Donnez l'équation du modèle et ses postulats, les hypothèses statistiques testées, et réalisez l'analyse avec le logiciel SAS. Énoncez clairement vos conclusions.

6. La nouvelle propriétaire d'un magasin d'aliments naturels veut savoir s'il existe des différences importantes dans les ventes moyennes entre les six jours de la semaine où le magasin est ouvert, soit du lundi au samedi. Elle utilise les valeurs des ventes des quinze dernières semaines. Voici un aperçu de ses données, arrondies au dollar le plus près.

Jour	Semaine				Moyenne	Écart type
i	1	2	...	15	$\overline{x}_i$	s_i
1. Lundi	1879	1335	...	947	1727	801
2. Mardi	610	1966	...	1637	1588	800
3. Mercredi	1373	860	...	2099	1225	596
4. Jeudi	2424	1071	...	2630	2517	1102
5. Vendredi	5052	4361	...	3922	4299	580
6. Samedi	4354	6280	...	6017	5761	874

Les données complètes sont disponibles sur le site du cours. Proposez une analyse statistique de ces données permettant de répondre aux questions suivantes :

(a) Y a-t-il une différence significative entre les ventes moyennes des six jours de la semaine ?

(b) Si oui, quels sont les jours où les ventes moyennes sont les plus élevées ? Les plus basses ?

Donnez l'équation du modèle et ses postulats, les hypothèses statistiques testées, et réalisez l'analyse avec le logiciel SAS. Énoncez clairement vos conclusions.

7. Dans le *Guide de consommation de carburant 2006* publié par le Ministère des Ressources naturelles du Canada, on retrouve des données sur la cylindrée et la consommation d'essence en ville de plusieurs modèles de véhicules. En voici un échantillonnage aléatoire de 22 modèles.

Modèle	Cylindrée (litres)	Consom. (l/100 km)
Hyundai ACCENT	1.6	8.3
Kia RIO	1.6	7.4
Mini COOPER S	1.6	10.0
Acura RSX	2.0	10.4
Ford FOCUS	2.0	9.2
Saturn ION	2.0	10.1
Mazda 6 TURBO	2.3	12.5
Volks NEW BEETLE	2.5	10.4
Chrysler SEBRING	2.7	11.1
Honda ACCORD	3.0	11.5
Toyota CAMRY	3.0	11.5

Modèle	Cylindrée (litres)	Consom. (l/100 km)
Volks TOUAREG	3.2	14.6
Chevrolet IMPALA	3.5	11.3
Nissan ALTIMA	3.5	11.0
Audi S4 CABRIOLET	4.2	15.8
Ford MUSTANG	4.6	13.8
Ford EXPLORER 4X4	4.6	16.6
Dodge DAKOTA	4.7	15.6
Jeep COMMANDER 4X4	4.7	15.6
Pontiac GRAND PRIX	5.3	12.9
Ferrari SCAGLIETTI	5.7	22.9
Rolls Royce PHANTOM	6.7	18.8

Proposez une analyse statistique de ces données permettant de répondre aux questions suivantes :

(a) Y a-t-il une relation linéaire significative entre la cylindrée d'une voiture et sa consommation d'essence ?

(b) Quelle proportion de la variabilité de la consommation d'essence est-elle expliquée par la cylindrée ?

(c) Les voitures ayant une cylindrée de 5.3 litres consomment en moyenne combien de litres aux 100 km ? Pouvez-vous donner un intervalle de confiance à 95% pour cette moyenne ?

8. Trois instruments peuvent être utilisés pour mesurer une dimension particulière sur une pièce industrielle. Vingt pièces sont sélectionnées au hasard dans la production pour comparer les instruments. L'opérateur mesurera chaque pièce deux fois avec chaque instrument, et toutes les mesures seront prises dans un ordre complètement aléatoire. Voici un aperçu des données récoltées.
(Exemple inspiré de Montgomery[16], p. 497)

Pièce	Instrument 1		Instrument 2		Instrument 3	
1	21	20	20	20	19	21
2	24	23	24	24	23	24
3	20	21	19	21	20	22
...	...	...	...	...	...	...
20	19	19	18	17	19	17

Proposez une analyse statistique de ces données permettant de répondre aux questions suivantes :

(a) Les pièces produites dans cette usine ont-elles des mesures très variables ?

(b) Les trois instruments donnent-ils lieu à des mesures qui diffèrent statistiquement en moyenne ?

9. Soit $\mathbf{X} = (X_1, X_2, X_3)^\top$ un vecteur aléatoire qui suit une loi nomale de dimension 3 de moyenne $\boldsymbol{\mu} = (1, -1, 0)^\top$ et de matrice de variance-covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} 9 & 1/2 & 2 \\ 1/2 & 1 & 0 \\ 2 & 0 & 4 \end{pmatrix}$$

(a) Donner la loi conjointe de $Y_1 = 2X_1 - X_3$ et $Y_2 = X_1 + X_2 + X_3$.

(b) Donner la loi conditionnelle de X_2 sachant $X_1 = X_3 = 1/2$.

(c) Calculer la corrélation conditionnelle de la paire (X_1, X_3) sachant $X_2 = -1$.

10. Soit $\mathbf{X} = (X_1, X_2, X_3)^\top$ un vecteur aléatoire qui suit une loi nomale de dimension 3 de moyenne $\boldsymbol{\mu} = (1, -1, 0)^\top$ et de matrice de variance-covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} 7 & 0 & 0 \\ 0 & 1 & -1/2 \\ 0 & -1/2 & 4 \end{pmatrix}$$

(a) Dire les quelles parmi les paires suivantes de variables aléatoires sont indépendantes : $(X_1, X_2), (X_1, X_3)$ et (X_2, X_3).

(b) Quelle est la loi de $(X_2, X_3)^\top$?

(c) En déduire la loi conditionnelle de $(X_2, X_3)^\top$ sachant $X_1 = -100$.

11. Soit $\mathbf{X} = (X_1, X_2)^\top$ un vecteur aléatoire qui suit une loi nomale de dimension 2 de moyenne $\boldsymbol{\mu} = (0, 0)^\top$ et de matrice de variance-covariance

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

où ρ est un réel compris entre -1 et 1.

Donner la loi de $X_1^2 - 2\rho X_1 X_2 + X_2^2$.

2 Le modèle linéaire

Nous verrons dans ce chapitre que les modèles d'analyse de variance à effets fixes peuvent être paramétrisés comme des modèles de régression, où les variables explicatives sont des variables indicatrices représentant l'appartenance aux diverses modalités des facteurs. Le modèle linéaire constitue un cadre plus général pour étudier les effets de plusieurs variables sur une mesure continue distribuée normalement.

2.1 Rappel du modèle de régression linéaire multiple

Sous forme de n équations :

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i \qquad (i = 1, ..., n) \quad$ où $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d.

Sous forme matricielle :

$$\begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & ... & x_{k1} \\ 1 & x_{12} & x_{22} & ... & x_{k2} \\ ... & ... & ... & ... & ... \\ 1 & x_{1n} & x_{2n} & ... & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}$$

$$Y \quad = \quad X \quad \beta \quad + \quad \varepsilon$$

où $\varepsilon \sim N_n(0, \sigma^2 I)$, i.e. une loi normale multivariée de dimension n de moyenne 0 et de variance σ^2 et sans covariance entre les observations.

Estimation de β par les moindres carrés

L'estimateur de β est le vecteur $\hat{\beta} = (\hat{\beta}_0, ..., \hat{\beta}_k)^T$ qui minimise la somme des carrés des erreurs, soit

$$\sum_i \varepsilon_i^2 = \varepsilon^T \varepsilon = (Y - X\beta)^T (Y - X\beta).$$

La solution de ce problème d'optimisation est relativement simple. En fait $\hat{\beta}$ est la solution des équations normales (obtenues en dérivant $(Y - X\beta)^T (Y - X\beta)$ par rapport

à β),

$$X^T X\beta = X^T Y.$$

Lorsque la matrice $X^T X$ est inversible, il y a une solution unique à ce système d'équations : $\hat{\beta} = (X^T X)^{-1} X^T Y$. De plus, la distribution de l'estimateur $\hat{\beta}$ est

$$\hat{\beta} \sim N_{k+1}(\beta, \sigma^2 (X^T X)^{-1})$$

où N_{k+1} signifie une distribution normale multivariée, en dimension $k+1$.

Combinaisons linéaires des β_j

En vertu de la propriété des transformations linéaires de vecteurs suivant une loi normale multivariée, si C est une matrice de constantes $d \times (k+1)$ et que $\hat{\beta}$ suit la normale multivariée précédente alors

$$C\hat{\beta} \sim N_d(C\beta, C\sigma^2 (X^T X)^{-1} C^T)$$

Tests d'hypothèses concernant $C\beta$

Soit C une matrice de constantes $d \times (k+1)$ de rang d et c un vecteur de constantes $d \times 1$. Supposons que l'on veuille tester les hypothèses

$$H_0 : C\beta = c$$
$$H_1 : C\beta \neq c$$

Méthode 1 pour la construction d'une statistique F : forme quadratique

Une première façon de construire une statistique de Fisher pour tester H_0 ci-dessus consiste à utiliser directement la forme quadratique des estimations centrées et réduites en comparaison de la variance résiduelle. On rejette l'hypothèse nulle au seuil α si

$$\frac{(C\hat{\beta} - c)^T (C(X^T X)^{-1} C^T)^{-1} (C\hat{\beta} - c)/d}{\hat{\sigma}^2} > F_{d,n-k-1,1-\alpha}, \tag{1}$$

où $\hat{\sigma}^2 = \dfrac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{n-k-1}$ est une estimation sans biais de la variance résiduelle.

Pour démontrer que la loi de cette statistique est bien une Fisher, on utilise les propriétés des formes quadratiques des vecteurs suivant des lois normales multivariées et on déduit que $(C\hat{\beta}-c)^T(C(X^TX)^{-1}C^T\sigma^2)^{-1}(C\hat{\beta}-c)$ suit une loi χ^2_d. En fait le numérateur de (1) peut s'interpréter comme la somme de carrés pour l'hypothèse H_0, $(C\hat{\beta}-c)^T(C(X^TX)^{-1}C^T)^{-1}(C\hat{\beta}-c)$, divisée par d, ses degrés de liberté.

On peut ensuite démontrer que $\frac{(n-k-1)\hat{\sigma}^2}{\sigma^2}$ suit une χ^2_{n-k-1} et qu'elle est indépendante du numérateur de la statistique F. Notons que sous forme matricielle, la somme de carrés résiduels s'écrit

$$SSE = (n-k-1)\hat{\sigma}^2 = (Y-X\hat{\beta})^T(Y-X\hat{\beta}) = Y^T\{I-X(X'X)^{-1}X^T\}Y$$

Méthode 2 pour la construction d'une statistique F : différence de deux SSE

Il existe une deuxième façon de construire la somme de carrés pour tester $H_0 : C\beta = c$, souvent appelée le test général de la régression. Il suffit de faire la différence entre deux sommes de carrés résiduelles, celle pour le modèle restreint où l'estimateur $\hat{\beta}_r$ satisfait $C\hat{\beta}_r = c$ et celle pour le modèle complet où $C\hat{\beta}_c \neq c$. Une deuxième façon d'écrire la somme de carrés pour H_0 est

$$\begin{aligned} SSE_{restreint} - SSE_{complet} &= (Y-X\hat{\beta}_r)^T(Y-X\hat{\beta}_r) - (Y-X\hat{\beta}_c)^T(Y-X\hat{\beta}_c) \\ &= (C\hat{\beta}-c)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta}-c) \end{aligned} \tag{2}$$

On obtient la même somme de carrés au numérateur qu'avec la première façon, on utilisera évidemment le même dénominateur pour obtenir la même statistique de Fisher.

Quelques cas particuliers

1. Test sur UNE des variables explicatives de la régression multiple
   Si $C = (0,\ldots,0,1,0,\ldots,0)$ où le 1 correspond à la j^e variable et $c = 0$, alors $C\beta - c = \beta_j - 0$ et le test porte sur la signification de la variable explicative X_j. La somme de carrés pour tester $H_0 : \beta_j = 0$ peut s'écrire

   $$SS(X_j|O, X_1,\ldots,X_{j-1},X_{j+1},\ldots,X_k) = \hat{\beta}_j[C(X^TX)^{-1}C^T]^{-1}\hat{\beta}_j = \hat{\beta}_j^2/(X^TX)^{jj},$$

   où $(X^TX)^{jj}$ est le terme $(j+1, j+1)$ de $(X^TX)^{-1}$.

La notation "$X_j|O, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k$" signifie que l'ordonnée à l'origine et les $k-1$ autres variables explicatives sont dans le modèle lorsque la somme de carrés mesurant l'importance relative de X_j est calculée.

2. Test sur l'unique variable explicative en régression linéaire simple

   Si $k = 1$, il n'y a qu'une variable explicative dans le modèle, la matrice X a seulement deux colonnes et l'estimateur de la pente de la régression s'écrit $\hat{\beta}_1 = S_{xy}/S_{xx} = \sum(y_i - \overline{y})x_i / \sum(x_i - \overline{x})^2$. De plus,

$$(X^TX)^{-1} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} = \frac{1}{nS_{xx}} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}.$$

   La somme de carrés pour tester $H_0 : \beta_1 = 0$ s'écrit

$$SS(X|O) = \hat{\beta}_1^2/(X^TX)^{11} = \hat{\beta}^2\, S_{xx}$$

   où $(X^TX)^{11}$ est le terme $(2,2)$ de $(X^TX)^{-1}$ et vaut $1/S_{xx}$.

   En général, la somme de carrés pour l'importance X_j dans une régression linéaire simple, $SS(X_j|O)$, est différente de la somme de carrés dans une régression multiple, $SS(X_j|O, X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_k)$.

3. Cas $k = 2$ avec deux variables explicatives orthogonales

   Les deux variables explicatives sont orthogonales si $\sum x_{1i}x_{2i} = n\overline{x}_1\overline{x}_2$, c'est-à-dire si le coefficient de corrélation de Pearson entre les deux variables est exactement 0. On a alors,

$$\begin{aligned}(X^TX)^{-1} &= \begin{pmatrix} n & n\overline{x}_1 & n\overline{x}_2 \\ n\overline{x}_1 & \sum x_{1i}^2 & n\overline{x}_1\overline{x}_2 \\ n\overline{x}_2 & n\overline{x}_1\overline{x}_2 & \sum x_{2i}^2 \end{pmatrix}^{-1} \\ &= \begin{pmatrix} \frac{1}{n} + \frac{\overline{x}_1^2}{(n-1)s_1^2} + \frac{\overline{x}_2^2}{(n-1)s_2^2} & -\frac{\overline{x}_1}{(n-1)s_1^2} & -\frac{\overline{x}_2}{(n-1)s_2^2} \\ -\frac{\overline{x}_1}{(n-1)s_1^2} & \frac{1}{(n-1)s_1^2} & 0 \\ -\frac{\overline{x}_2}{(n-1)s_2^2} & 0 & \frac{1}{(n-1)s_2^2} \end{pmatrix},\end{aligned}$$

   où $s_j^2 = \sum(x_{ji} - \overline{x}_j)^2/(n-1)$ sont pour $j = 1, 2$ les variances des deux variables explicatives. Dans ce cas $SS(X_2|O, X_1) = \hat{\beta}_2^2(n-1)s_2^2 = SS(X_2|O)$. Ainsi lorsque les deux variables explicatives sont orthogonales, le fait d'inclure ou non une variable pour construire la somme de carrés de l'autre variable ne change

pas le résultat. Les deux sommes de carrés $SS(X_2|O, X_1)$ et $SS(X_2|O)$ sont identiques. Ce résultat demeure vrai dans le cas multivarié lorsque X_1 et X_2 sont des matrices associées à k_1 et k_2 variables explicatives dans la mesure où les variables explicatives du premier groupe sont toutes orthogonales aux variables explicatives du deuxième groupe.

2.2 ANOVA à un facteur avec le modèle linéaire

Considérons I populations indépendantes. Pour $i = 1, 2, \cdots, I$, on dispose d'un échantillon $\{y_{i1}, y_{i2}, \cdots, y_{in_i}\}$ de taille n_i issu d'une loi $N(\mu_i, \sigma^2)$. Notons que σ^2 ne varie pas d'un échantillon à l'autre. Soit $n = n_1 + n_2 + \cdots + n_I$ la taille totale de l'ensemble des échantillons. À partir de ces échantillons, on voudrait tester les hypothèses :

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \cdots = \mu_I \\ H_1 &: \text{les moyennes } \mu_i \text{ ne sont pas toutes égales} \end{aligned}$$

On peut décrire le modèle de deux façons : directement à partir des moyennes de traitements, ce qu'on appelle le μ-modèle, ou en décomposant les déviations des traitements par rapport à la moyenne globale, ce qu'on appelle le modèle avec effets.

2.2.1 Le μ-modèle

Le modèle d'ANOVA à un facteur peut être énoncé en fonction des moyennes de traitements :

$$y_{ij} = \mu_i + \varepsilon_{ij} \qquad i = 1, \ldots, I; \quad j = 1, \ldots n_i. \tag{3}$$

Ce modèle peut être vu comme un modèle linéaire. On pose $\beta = (\mu_1, \ldots, \mu_I)'$ le vecteur des moyennes théoriques, Y est le vecteur $n \times 1$ des observations et X est la matrice $n \times I$ suivante :

$$\begin{bmatrix} y_{11} \\ \ldots \\ y_{1n_1} \\ y_{21} \\ \ldots \\ y_{2n_2} \\ \ldots \\ y_{In_I} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \ldots \\ \mu_I \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \ldots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \ldots \\ \varepsilon_{2n_2} \\ \ldots \\ \varepsilon_{In_I} \end{bmatrix}$$

$$Y \quad = \quad X \quad \beta \quad + \quad \varepsilon$$

On a ainsi :

$$X^T X = \begin{bmatrix} n_1 & 0 & \ldots & 0 \\ 0 & n_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \ldots & n_I \end{bmatrix}, \quad X^T Y = \begin{bmatrix} n_1 \overline{y}_{1.} \\ n_2 \overline{y}_{2.} \\ \vdots \\ n_I \overline{y}_{I.} \end{bmatrix} \quad \text{et} \quad \hat{\beta} = \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \ldots \\ \hat{\mu}_I \end{bmatrix} = \begin{bmatrix} \overline{y}_{1.} \\ \overline{y}_{2.} \\ \vdots \\ \overline{y}_{I.} \end{bmatrix}$$

On peut interpréter ce modèle comme un ajustement d'une droite de pente nulle dans chaque traitement, et où les paramètres sont les ordonnées à l'origine. L'estimation des moindres carrés des paramètres est donc obtenue de la même façon qu'avec le modèle linéaire, en résolvant les équations normales : $\hat{\beta} = (X^T X)^{-1} X^T Y$.

Test d'égalité des moyennes

L'hypothèse d'homogénéité des moyennes s'écrit $H_0 : C\beta = 0$, où la matrice C est une matrice $(I-1) \times I$ de contrastes (combinaisons linéaires des paramètres dont la somme des coefficients est nulle). Plusieurs choix sont disponibles, dont les deux suivants :

$$C = \begin{pmatrix} 1 & 0 & \dots & 0 & -1 \\ 0 & 1 & \dots & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \dots & 1 & -1 \end{pmatrix} \quad \text{ou} \quad C = \begin{pmatrix} 1-1/I & -1/I & \dots & -1/I & -1/I \\ -1/I & 1-1/I & \dots & -1/I & -1/I \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1/I & -1/I & \dots & 1-1/I & -1/I \end{pmatrix}.$$

La première matrice C compare toutes les moyennes à la dernière alors que la deuxième compare successivement les $I-1$ premières moyennes à la moyenne globale.

Pour construire la somme de carrés pour tester H_0, la forme quadratique mettant en jeu la matrice C peut être utilisée, mais la méthode (2) de la différence entre deux sommes de carrés résiduelles est la plus simple. La somme de carrés résiduelle pour le modèle réduit, où toutes les ordonnées à l'origine sont égales et donc estimées par la moyenne globale, est la somme de carrés totale

$$SSE_{restreint} = \sum_{ij} (y_{ij} - \overline{y}_{..})^2$$

Pour le modèle complet, où les ordonnées à l'origine sont distinctes et donc estimées par les moyennes de traitement, on obtient

$$SSE_{complet} = \sum_{ij} (y_{ij} - \overline{y}_{i.})^2$$

En vertu de la décomposition de la somme de carrés totale de l'ANOVA à un facteur, on obtient

$$SS(\mu_1 = \ldots = \mu_I | O) = SSE_{restreint} - SSE_{complet} = \sum_i n_i (\overline{y}_{i.} - \overline{y}_{..})^2.$$

On appelle communément cette quantité la somme de carrés associée à la régression, ou la somme des carrés associée aux traitements. La statistique de Fisher sera construite en la divisant par ses degrés de liberté $(I-1)$, puis par $\hat{\sigma}^2$.

2.2.2 Le modèle avec effets

On écrit souvent le modèle d'ANOVA à un facteur en termes d'une moyenne globale μ et d'effets τ_i de chaque traitement de la façon suivante,

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \qquad i = 1, \ldots, I; \quad j = 1, \ldots n_i. \tag{4}$$

On est donc en présence d'un modèle linéaire avec

$$X = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & 0 & 0 & \ldots & 1 \end{pmatrix} \text{ et } \beta = \begin{pmatrix} \mu \\ \tau_1 \\ \vdots \\ \tau_I \end{pmatrix}$$

Notons cependant que la matrice X de dimension $n \times (I+1)$ n'est pas de plein rang. En effet son rang est I car la somme des I dernières colonnes est égale à la première. Puisque X est singulière, la matrice X^TX n'est pas inversible et la formule standard $\hat{\beta} = (X^TX)^{-1}X^TY$ pour estimer β ne s'applique pas car $(X^TX)^{-1}$ n'existe pas. On voit ci-dessous que la première colonne de la matrice X^TX est la somme des colonnes suivantes.

$$X^TX = \begin{bmatrix} N & n_1 & n_2 & \ldots & n_I \\ n_1 & n_1 & 0 & \ldots & 0 \\ n_2 & 0 & n_2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ n_I & 0 & 0 & \ldots & n_I \end{bmatrix}$$

Imposition de contraintes

La singularité de la matrice X vient du fait que le modèle (4) est surparamétré car plusieurs valeurs des paramètres donnent des moyennes identiques pour les I traitements. En d'autres termes, il existe une infinité de solutions aux I équations normales car elles contiennent $I+1$ inconnues.

On ajoute souvent une contrainte pour faire en sorte que les paramètres soient bien définis. Deux types de contraintes sont utilisés en général :

1. La somme de certains paramètres est nulle, par exemple $\sum_{i=1}^{I} \tau_i = 0$.
2. Des paramètres sont fixés à 0, par exemple $\tau_I = 0$.

Avec la contrainte 1, on peut exprimer un paramètre en fonction des autres, par exemple $\tau_I = \mu - \tau_1 - ... - \tau_{I-1}$. On peut donc retirer τ_I du modèle, de même qu'avec la contrainte 2 où il est fixé à 0. L'imposition de contraintes ramène la dimension de la matrice X à $n \times I$. Les deux matrices X correspondantes sont les suivantes, où la dernière colonne est retirée pour l'estimation.

$$X_{contr.\ 1} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & -1 & -1 & \dots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & \dots & -1 \end{pmatrix} \text{ et } X_{contr.\ 2} = \begin{pmatrix} 1 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & \dots & 0 \\ 1 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \dots & 0 \end{pmatrix}$$

En fait les modèles obtenus avec les deux contraintes sont des reparamétrisations du μ-modèle. Pour la première contrainte, $\mu = \sum \mu_i / I$ et $\tau_i = \mu_i - \mu$ alors que pour la deuxième, $\mu = \mu_I$ et $\tau_i = \mu_i - \mu_I$. Les estimations des paramètres et leur interprétation sont évidemment affectées par le choix de la contrainte, sauf pour certaines combinaisons linéaires appelées *fonctions estimables*, sur lesquelles nous reviendrons plus loin.

Inverses généralisés

Une façon plus générale de traiter un modèle linéaire avec une matrice X singulière est de faire appel aux *inverses généralisés* de $X'X$ pour résoudre les équations normales définissant l'estimateur des moindres carrés de β. C'est l'approche retenue par SAS pour ses procédures GLM et MIXED. Les estimations obtenues sont équivalentes à celles issues de l'imposition de contraintes linéaires.

- Définition : On appelle *inverse généralisé* de la matrice $A_{p\times q}$ une matrice $A^-_{q\times p}$ qui vérifie la relation suivante :
$$AA^-A \;=\; A$$

- Il existe plusieurs façons de calculer un inverse généralisé. En voici une parmi tant d'autres.
Si $A_{p\times q}$ peut être partitionnée en sous-matrices de la façon suivante :
$A_{p\times q} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$ où rang(A_{11})=rang(A) et que A_{11}^{-1} existe,
alors la matrice $A^-_{q\times p} = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}$ est un inverse généralisé de $A_{p\times q}$.

- Preuve : $AA^-A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$

  Pour que $AA^-A = A$, il faut donc que $A_{21}A_{11}^{-1}A_{12} = A_{22}$.

  Puisque A est singulière et que rang(A)=rang(A_{11}), alors il existe une matrice K telle que $[A_{21}\ A_{22}] = K\ [A_{11}\ A_{12}]$.

  Il suit que :
$$\begin{aligned} A_{21} &= K\ A_{11} \\ A_{22} &= K\ A_{12} \end{aligned}$$

  De la première ligne on tire $K = A_{21}A_{11}^{-1}$.

  De la deuxième ligne on tire $A_{22} = KA_{12} = A_{21}A_{11}^{-1}A_{12}$.
  C'est ce qu'il nous fallait démontrer.

- Exercice : Trouver un inverse généralisé de $A = \begin{bmatrix} 1 & 2 & 3 & 0 \\ 4 & 5 & 6 & 1 \\ 5 & 7 & 9 & 1 \end{bmatrix}$.

- Propriétés des inverses généralisés : Soit $(X'X)^-$ un inverse généralisé de $X'X$.
  (a) $(X'X)^-$ est non unique ; il existe une infinité de choix pour $(X'X)^-$.
  (b) $(X'X)^-$ n'est pas nécessairement symétrique.
  (c) $[(X'X)^-]'$ est aussi un inverse généralisé de $(X'X)$.
  (d) $H = X(X'X)^-X'$, la matrice de projection sur l'espace-colonne de X, est invariante par rapport au choix de $(X'X)^-$.
  (e) $H = X(X'X)^-X'$ est symétrique et idempotente, de rang égal au rang de X.
  (f) $HX = X(X'X)^-X'X = X$.

- Propriétés de $\hat{\beta}$:

  Lorsque $X'X$ est inversible, $\hat{\beta} = (X'X)^{-1}X'Y$ a les propriétés suivantes :
  — $E(\hat{\beta}) = \beta$
  — Var($\hat{\beta}$)=$\sigma^2(X'X)^{-1}$
  — BLUE (Best Linear Unbiased Estimator) : $t'\hat{\beta}$ est l'estimateur de variance minimale de $t'\beta$ parmi tous les estimateurs sans biais fonctions linéaires des y_i.

  Lorsque $X'X$ n'est pas inversible, $\hat{\beta} = (X'X)^-X'Y$ a les propriétés suivantes :
  — $E(\hat{\beta}) = (X'X)^-X'X\beta$
  — Var($\hat{\beta}$)=$\sigma^2(X'X)^-X'X[(X'X)^-]'$
  — BLUE (Best Linear Unbiased Estimator) :
  Si $t'\beta$ est une fonction estimable (i.e. qui ne dépend pas du choix de $(X'X)^-$), alors $t'\hat{\beta}$ est l'estimateur de variance minimale de $t'\beta$ parmi tous les estimateurs sans biais fonctions linéaires des y_i.

Fonctions estimables

Nous avons établi que les estimations des paramètres varient en fonction de la contrainte imposée. Par exemple, l'estimation de τ_1 est $\hat{\tau}_1 = \overline{y}_{1\bullet} - \sum \overline{y}_{i\bullet}/I$ sous la contrainte d'une somme nulle et $\hat{\tau}_1 = \overline{y}_{1\bullet} - \overline{y}_{I\bullet}$ sous la contrainte $\tau_I = 0$. Pour cette raison, on dira que τ_1 n'est pas une fonction estimable.

Définition. Une fonction $L^T\beta$ des paramètres d'un modèle linéaire est estimable si son estimation $L^T\hat{\beta}$ est toujours la même, peu importe les contraintes imposées aux paramètres pour obtenir une matrice X de plein rang.

Si X est de plein rang toutes les fonctions des paramètres sont estimables. En général $L^T\beta$ est estimable si L appartient à l'espace vectoriel généré par les lignes de X. En d'autres termes, pour que $L\beta$ soit estimable, il faut et il suffit que L soit une combinaison linéaire des lignes de X, i.e. que l'on puisse écrire $L = aX$ pour un vecteur de coefficients réels $a = (a_1, a_2, ..., a_N)$.
Remarques : 1) La dimension du vecteur a peut être réduite à I coefficients, le nombre de lignes indépendantes de X (une par cellule de traitement).
2) On aurait pu écrire que L doit être une combinaison linéaire des lignes de $X'X$, car ces matrices ont le même espace-ligne.

La matrice des variables explicatives X est singulière car $X[1, -1, -1, \ldots, -1]^T = 0$. En d'autres mots la somme des I dernières colonnes de X est égale à la première colonne. Une fonction estimable des paramètres s'écrit $L\beta$ où $L[1, -1, -1, \ldots, -1]^T = 0$. La forme générale du vecteur L est $L = (l_1, l_2, \ldots, l_I, l_1 - \sum_2^I l_i)$.

Exemples pour le modèle équilibré à un facteur :

- La moyenne pour la modalité 1 du facteur A, $\mu + \tau_1$, est estimable ; on prend $L = (1, 1, 0, \ldots, 0)$, qui est la première ligne de la matrice X. Le vecteur $a = (1, 0, 0, ..., 0)$ remplit l'équation $L = aX$, la forme générale est respectée. Peu importe la contrainte sur les paramètres, $\mu + \tau_1$ est toujours estimé par $\overline{y}_{1\bullet}$.
- La différence de moyennes entre les modalités 1 et 2 du facteur A, $\tau_1 - \tau_2$, est estimable ; on prend $L = (0, 1, -1, 0, \ldots, 0)$.
- Le paramètre μ n'est pas estimable car $L = (1, 0, \ldots, 0)$ n'est pas de la forme $L = (l_1, l_2, \ldots, l_I, l_1 - \sum_2^I l_i)$.

— L'hypothèse de l'homogénéité des moyennes s'écrit $H_0 : L\beta = 0$ pour tout $L = (0, l_2, \ldots, l_I, -\sum_2^I l_i)$. L sera alors une matrice de dimensions $(I-1) \times (I+1)$. En fait l'hypothèse nulle d'homogénéité est vraie si tous les contrastes sont nuls.

Inférence sur les fonctions estimables

L'inférence présentée ici se base sur le fait que les observations sont présumées issues d'une loi normale. On estime le paramètre de variance σ^2 par le MSE, et les degrés de liberté (ν) du quantile de la loi de Student dans les formules ci-dessous sont ceux associés à l'erreur.

Intervalle de confiance de niveau $1-\alpha$ pour une fonction estimable $L\beta$

$$L\hat{\beta} \ \pm \ t_{\alpha/2,\nu}\sqrt{L(X^TX)^-L^T \ \ MSE}$$

Test d'hypothèses de seuil α sur une fonction estimable $L\beta$

$$\begin{aligned} H_0 : \ & L\beta \ = \ 0 \\ H_1 : \ & L\beta \ \neq \ 0 \end{aligned}$$

Si H_0 est vraie, $F_0 \ = \ (L\hat{\beta})^T[L(X^TX)^-L^T \ \ MSE]^{-1}(L\hat{\beta}) \ \sim \ F_{1,\nu}$.

On rejettera H_0 si $F_0 \ > \ F_{1-\alpha,1,\nu}$.

Remarque : Il est possible de tester plusieurs fonctions estimables simultanément avec un seuil global α. L est alors une matrice $p \times k$ de coefficients réels pour tester p fonctions, et la loi de Fisher aura p et ν degrés de liberté.

Logiciel SAS : Procédures GLM et MIXED

Énoncé MODEL, option E

Permet d'afficher la forme générale des fonctions estimables. Utile pour déterminer les coefficients à inscrire pour les énoncés CONTRAST et ESTIMATE.

Énoncé ESTIMATE :

Permet d'estimer des fonctions linéaires des paramètres (en autant qu'elles soient estimables) en spécifiant soi-même le vecteur L. On obtient l'estimation ponctuelle et son erreur-type. Pour calculer l'intervalle de confiance, il faut spécifier l'option CLPARM dans l'énoncé MODEL. Dans la procédure MIXED, on peut inclure les paramètres aléatoires dans la fonction spécifiée.

Énoncé CONTRAST :

Permet de faire des tests d'hypothèses sur une ou plusieurs fonctions estimables.

Exemple : Fonction estimable dans SAS, analyse de variance à un facteur de 5 niveaux

Tentons d'estimer la moyenne globale de la variable réponse, μ.

```
proc glm data=aov;
class a;
model y = a /e;
estimate 'essai1' intercept 1  / e;
estimate 'essai2' intercept 1  a 0 0 0 0 0 / e;
run;
```

```
                General Form of Estimable Functions

Effect          Coefficients
  Intercept       L1
  A         1     L2
  A         2     L3
  A         3     L4
  A         4     L5
  A         5     L1-L2-L3-L4-L5
```

```
Coefficients for Estimate essai1

     Intercept                 1
     A         1             0.2
     A         2             0.2
     A         3             0.2
     A         4             0.2
     A         5             0.2

Coefficients for Estimate essai2

     Intercept                 1
     A         1               0
     A         2               0
     A         3               0
     A         4               0
     A         5               0
```

Discussion :

— Dans la commande *estimate* pour `essai1`, les coefficients des paramètres τ_i ne sont pas précisés. SAS, par défaut, accorde le coefficient $1/I$ à chaque paramètre τ_i.

— L'estimation de `essai1` est la moyenne des moyennes, $\sum \bar{y}_{i\bullet}/I$ qui est évidemment différente de $\bar{y}_{\bullet\bullet}$, la moyenne globale, lorsque les n_i sont différents (ce qui n'est pas le cas ici).

— Dans la commande *estimate* pour `essai2`, tous les coefficients de la combinaison linéaire sont spéficiés. Or la fonction n'est pas estimable. Le log de SAS porte la mention

```
essai2 is not estimable
```

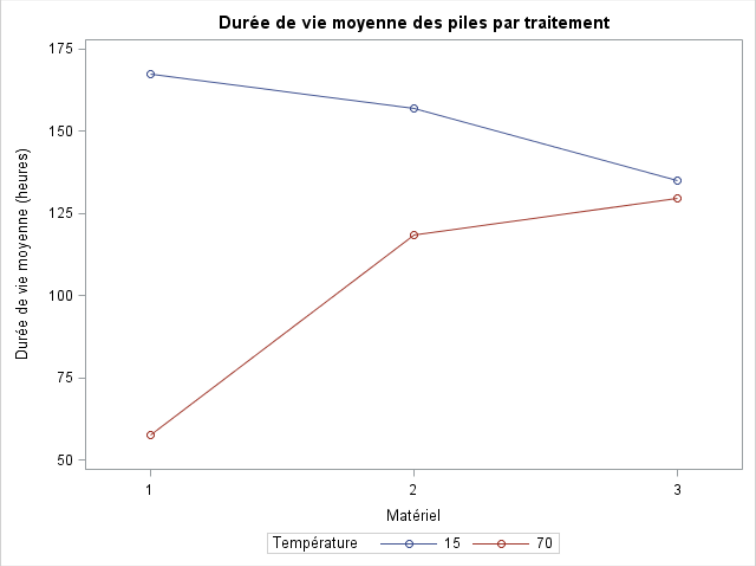
et aucune estimation n'apparaît en sortie.

2.3 ANOVA à deux facteurs avec le modèle linéaire

2.3.1 Un exemple

Considérons un exemple de plan complètement aléatoire équilibré à deux facteurs fixes, A=*température* et B=*matériel* , respectivement à 2 et 3 niveaux chacun. La variable réponse est la durée de vie de piles en heures. Voici les données. (Inspiré de Montgomery [16], p. 165).

Température	**Matériel 1**		**Matériel 2**		**Matériel 3**	
15°F	155	180	188	126	110	160
70°F	40	75	122	115	120	139



Puisque les deux courbes de moyennes ne sont pas parallèles, on pourrait s'attendre à une interaction significative. La différence entre les moyennes pour les matériaux ne semble pas majeure, par contre la différence de moyennes entre les deux températures apparaît beaucoup plus importante. Mais il faut conduire des tests d'hypothèses pour vérifier objectivement si ces constats sont statistiquement significatifs.

2.3.2 Le μ-modèle

Le modèle ANOVA à deux facteurs peut s'écrire en fonction des moyennes par traitement

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, J, \quad k = 1, \cdots, n_{ij}, \tag{5}$$

où $\varepsilon_{ijk} \sim N(0, \sigma^2)$, μ_{ij} est la moyenne pour le traitement (i, j) et n_{ij} la taille de l'échantillon pour ce traitement. Il est utilisé pour analyser les données provenant d'expériences complètement randomisées à deux facteurs, A et B, avec respectivement I et J modalités.

On considère dans ce chapitre le cas où les n_{ij} sont égaux. Notre objectif est d'estimer certaines fonctions des paramètres de ce modèle et de construire des tests pour les hypothèses

(i) de l'absence d'interaction entre les deux facteurs, et
(ii) d'homogénéité des modalités des facteurs individuels.

Exemple.

Le μ-modèle pour cette expérience, exprimé sous forme matricielle, est

$$Y = \begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{131} \\ y_{132} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{231} \\ y_{232} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{13} \\ \mu_{21} \\ \mu_{22} \\ \mu_{23} \end{bmatrix} + \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ \varepsilon_{121} \\ \varepsilon_{122} \\ \varepsilon_{131} \\ \varepsilon_{132} \\ \varepsilon_{211} \\ \varepsilon_{212} \\ \varepsilon_{221} \\ \varepsilon_{222} \\ \varepsilon_{231} \\ \varepsilon_{232} \end{bmatrix} = X\beta + \varepsilon$$

On trouve

$$X^TX = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}, \; X^TY = \begin{bmatrix} y_{11\bullet} \\ y_{12\bullet} \\ y_{13\bullet} \\ y_{21\bullet} \\ y_{22\bullet} \\ y_{23\bullet} \end{bmatrix} \text{ et } \hat{\beta} = \begin{bmatrix} \overline{y_{11\bullet}} \\ \overline{y_{12\bullet}} \\ \overline{y_{13\bullet}} \\ \overline{y_{21\bullet}} \\ \overline{y_{22\bullet}} \\ \overline{y_{23\bullet}} \end{bmatrix} = \begin{bmatrix} 167.5 \\ 157.0 \\ 135.0 \\ 57.5 \\ 118.5 \\ 129.5 \end{bmatrix}$$

La variance des estimateurs peut être obtenue de façon matricielle :

$$Var(\hat{\beta}) = Var[(X^TX)^{-1}X^TY] = (X^TX)^{-1}X^T \; Var(Y) \; X[(X^TX)^{-1}]^T = \sigma^2(X^TX)^{-1}$$

Pour chaque $\hat{\beta}_k = \overline{y_{ij\bullet}}$, on estime l'erreur-type par $\sqrt{\frac{MSE}{n_{ij}}}$. Dans cet exemple,

$$\hat{\sigma}^2 = MSE = \frac{1}{6}[(155 - 167.5)^2 + \ldots + (139 - 129.5)^2] = 717.0$$

Dans SAS, un tel modèle s'ajuste en incluant seulement le terme d'interaction dans l'énoncé MODEL, sans ordonnée à l'origine. Cela revient à faire une anova à un facteur, à $I \times J = 6$ modalités. Le test de la table d'anova est plus ou moins pertinent ici, car il vérifie si toutes les moyennes de cellules sont simultanément nulles. La table permet toutefois d'identifier la valeur du MSE.

```
proc glm data=batterie;
class temp materiel ;
model duree=temp*materiel/noint xpx solution;
run;quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	210098.0000	35016.3333	48.84	<.0001
Error	6	4302.0000	717.0000		
Uncorrected Total	12	214400.0000			

L'option XPX permet de faire afficher la matrice $X^T X$, et SOLUTION donne les estimations des paramètres avec leurs erreurs-types, présenté ci-dessous.

Parameter	Estimate	Standard Error	t Value	Pr > |t|
temp*materiel 15 1	167.5000000	18.93409623	8.85	0.0001
temp*materiel 15 2	157.0000000	18.93409623	8.29	0.0002
temp*materiel 15 3	135.0000000	18.93409623	7.13	0.0004
temp*materiel 70 1	57.5000000	18.93409623	3.04	0.0229
temp*materiel 70 2	118.5000000	18.93409623	6.26	0.0008
temp*materiel 70 3	129.5000000	18.93409623	6.84	0.0005

Test sur l'interaction $A * B$

On peut utiliser la première méthode de construction des sommes de carrés en régression (avec une matrice de contrastes) pour construire une statistique de test pour nos hypothèses, dans la mesure où on peut écrire ces dernières en fonction des moyennes μ_{ij}.

Pour la signification de l'interaction, on veut savoir si la différence entre deux modalités d'un facteur est constante quand l'autre facteur varie. L'hypothèse d'intérêt s'exprime de la façon suivante :

$$H_0^{inter} : (\mu_{ij} - \mu_{i'j}) - (\mu_{ij'} - \mu_{i'j'}) = 0$$

pour $i, i' = 1, \ldots, I; j, j' = 1, \ldots, J$; avec $(I-1)(J-1)$ degrés de liberté

Exemple.

L'hypothèse d'absence d'interaction est

$$\begin{aligned} H_0^{inter}: \quad (\mu_{11}-\mu_{21})-(\mu_{12}-\mu_{22}) &= 0 \\ (\mu_{11}-\mu_{21})-(\mu_{13}-\mu_{23}) &= 0 \end{aligned}$$

ce qui est équivalent à

$$H_0^{inter}: \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{pmatrix} \beta = C_{A*B}\beta = 0,$$

Puisque

$$C_{A*B}\hat{\beta} = \begin{pmatrix} 71.5 \\ 104.5 \end{pmatrix} \text{ et } C_{A*B}(X^TX)^{-1}C_{A*B}^T = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

La somme de carrés pour l'interaction est donc

$$\begin{aligned} SS(A*B|O,A,B) &= (C_{A*B}\hat{\beta})^T(C_{A*B}(X^TX)^{-1}C_{A*B}^T)^{-1}C_{A*B}\hat{\beta} \\ &= \begin{pmatrix} 71.5 & 104.5 \end{pmatrix}\begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix}\begin{pmatrix} 71.5 \\ 104.5 \end{pmatrix} \\ &= 5707.167 \end{aligned}$$

Cette somme de carrés sera divisée par $(I-1)(J-1)=2$ degrés de liberté pour obtenir le carré moyen, qui sera à son tour divisé par le MSE pour faire une statistique de Fisher si les conditions du modèle sont respectées.

Dans SAS, un énoncé CONTRAST permet de préciser les coefficients de la matrice C_{A*B} pour obtenir le test de Fisher.

```
contrast "interaction" temp*materiel 1 -1  0 -1 1 0,
                       temp*materiel 1  0 -1 -1 0 1;
```

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
interaction	2	5707.166667	2853.583333	3.98	0.0794

Test sur les facteurs simples A et B

— $H_0^A : \sum_j \frac{\mu_{ij}}{J} = \sum_j \frac{\mu_{i'j}}{J}, \qquad i, i' = 1, \ldots, I$ avec $(I-1)$ degrés de liberté

— $H_0^B : \sum_i \frac{\mu_{ij}}{I} = \sum_i \frac{\mu_{ij'}}{I}, \qquad i, i' = 1, \ldots, I$ avec $(J-1)$ degrés de liberté

Ces tests devraient être conduits seulement lorsque l'interaction est non significative.

Exemple.

$$\begin{array}{rccc} H_0^A: & \mu_{11}+\mu_{12}+\mu_{13} & = & \mu_{21}+\mu_{22}+\mu_{23} \\ H_0^B: & \mu_{11}+\mu_{21} & = & \mu_{12}+\mu_{22} \\ & \mu_{11}+\mu_{21} & = & \mu_{13}+\mu_{23} \end{array}$$

Les sommes de carrés pour les effets simples se calculent de la même façon que pour le test sur l'interaction, mais en utilisant les matrices

$$C_A = \begin{pmatrix} 1 & 1 & 1 & -1 & -1 & -1 \end{pmatrix} \text{ et } C_B = \begin{pmatrix} 1 & -1 & 0 & 1 & -1 & 0 \\ 1 & 0 & -1 & 1 & 0 & -1 \end{pmatrix}.$$

Faisons les calculs avec des énoncés CONTRAST de GLM :

```
contrast "facteur A" temp*materiel 1  1  1 -1 -1 -1;
contrast "facteur B" temp*materiel 1 -1  0  1 -1  0,
                     temp*materiel 1  0 -1  1  0 -1;
```

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
facteur A	1	7905.333333	7905.333333	11.03	0.0160
facteur B	2	1410.500000	705.250000	0.98	0.4271

On obtient $SS(A|O, B, A*B) = 7905.33$ et $SS(B|O, A, A*B) = 1410.5$.

Les notations $A|O, B, A*B$ et $B|O, A, A*B$ insistent sur le fait que ces deux sommes de carrés s'interprètent comme si elles avaient été calculées à partir d'un modèle complet qui comprenait un terme d'interaction et des effets simples. Dans la notation SAS, ces sommes de carrés sont dites de type 3 car lors de l'évaluation de la somme de carrés pour une composante toutes les autres composantes sont gardées dans le modèle.

2.3.3 Le modèle avec effets

Dans le modèle ANOVA à plusieurs facteurs dit "avec effets", on sépare les sources de variabilité en une moyenne globale et des déviations (effets) dûes aux facteurs simples et à leurs interactions. On utilise l'écriture suivante

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}, \tag{6}$$

où $\quad i = 1, \ldots, I, \quad j = 1, \ldots, J, \quad k = 1, \ldots, n_{ij}$

On est donc en présence de $(I+1)\times(J+1)$ paramètres ; la dimension de l'ensemble des fonctions estimables est $I \times J$, le nombre de moyennes μ_{ij} différentes dans le modèle (voir (5)). Pour pouvoir estimer les paramètres il faut donc imposer $I+J+1$ contraintes. Peu importe le type de contraintes retenues, le modèle s'écrit alors avec une moyenne globale μ, $I-1$ effets simples τ, $J-1$ effets simples β et $(I-1)(J-1)$ paramètres d'interaction $\tau\beta$.

1) Contraintes où des sous-groupes de paramètres somment à zéro.

Elles impliquent

$$\begin{aligned} \tau_I &= -\textstyle\sum_{i=1}^{I-1} \tau_i & \tau\beta_{Ij} &= -\textstyle\sum_{i=1}^{I-1} \tau\beta_{ij} \\ \beta_J &= -\textstyle\sum_{j=1}^{J-1} \beta_j & \tau\beta_{iJ} &= -\textstyle\sum_{j=1}^{J-1} \tau\beta_{ij} \\ & & \tau\beta_{IJ} &= -\textstyle\sum_{i=1}^{I-1}\sum_{j=1}^{J-1} \tau\beta_{ij} \end{aligned}$$

Les estimateurs des paramètres sous ces contraintes sont

$$\begin{aligned} \hat{\mu} &= \overline{y_{\bullet\bullet\bullet}} \\ \hat{\tau}_i &= \overline{y_{i\bullet\bullet}} - \overline{y_{\bullet\bullet\bullet}} \\ \hat{\beta}_j &= \overline{y_{\bullet j\bullet}} - \overline{y_{\bullet\bullet\bullet}} \\ \hat{\tau\beta}_{ij} &= \overline{y_{ij\bullet}} - \overline{y_{i\bullet\bullet}} - \overline{y_{\bullet j\bullet}} + \overline{y_{\bullet\bullet\bullet}} \end{aligned}$$

2) Contraintes où certains paramètres sont fixés à zéro.

On pose alors

$$\tau_I = \beta_J = \tau\beta_{iJ} = \tau\beta_{Ij} = 0,$$

pour $i = 1, \ldots, I$ et $j = 1, \ldots, J$.

Ce second type de contrainte est celui imposé par le calcul d'un inverse généralisé de X^TX.

Les estimateurs des paramètres sous ces contraintes sont

$$\begin{aligned}
\hat{\mu} &= \overline{y_{IJ\bullet}} \\
\hat{\tau}_i &= \overline{y_{iJ\bullet}} - \overline{y_{IJ\bullet}} \\
\hat{\beta}_j &= \overline{y_{Ij\bullet}} - \overline{y_{IJ\bullet}} \\
\hat{\tau\beta}_{ij} &= \overline{y_{ij\bullet}} - \overline{y_{iJ\bullet}} - \overline{y_{Ij\bullet}} + \overline{y_{IJ\bullet}}
\end{aligned}$$

Exemple.

Reprenons l'exemple présenté en début de section, avec un facteur A=*température* et un facteur B=*matériel* , respectivement à 2 et 3 niveaux chacun. Les paramètres du modèle linéaire réduit sont $\beta = (\mu, \tau_1, \beta_1, \beta_2, \tau\beta_{11}, \tau\beta_{12})^T$ et la matrice X du modèle linéaire avec la première contrainte (somme à 0), est

$$X = \left(\begin{array}{c|c|cc|cc}
1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & 0 & 1 & 0 & 1 \\
1 & 1 & -1 & -1 & -1 & -1 \\
1 & 1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & 0 & -1 & 0 \\
1 & -1 & 1 & 0 & -1 & 0 \\
1 & -1 & 0 & 1 & 0 & -1 \\
1 & -1 & 0 & 1 & 0 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1 & 1 & 1
\end{array}\right) = \left(\begin{array}{c|c|c|c} X_O & X_\tau & X_\beta & X_{\tau\beta} \end{array}\right)$$

On identifie bien dans la matrice X la première colonne X_O pour l'ordonnée à l'origine, la deuxième colonne X_τ pour l'effet du facteur A, la matrice 12×2 X_β pour l'effet du facteur B et finalement la matrice 12×2 $X_{\tau\beta}$ pour l'interaction. On note que $X_{\tau\beta}$ est obtenu en faisant le produit terme à terme des deux colonnes de X_β par X_τ. C'est comme la définition des contrastes d'interaction.

Notons que les quatre composantes de la matrice X sont orthogonales entre elles, car le plan est équilibré. Par exemple, $X_\tau^T X_{\tau\beta} = 0$. Ainsi,

$$X^T X = \begin{pmatrix} X_O^T X_O & 0 & 0 & 0 \\ 0 & X_\tau^T X_\tau & 0 & 0 \\ 0 & 0 & X_\beta^T X_\beta & 0 \\ 0 & 0 & 0 & X_{\tau\beta}^T X_{\tau\beta} \end{pmatrix} = \begin{pmatrix} 12 & 0 & 0 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 4 & 0 & 0 \\ 0 & 0 & 4 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 4 \\ 0 & 0 & 0 & 0 & 4 & 8 \end{pmatrix}$$

La formule (2.1) pour la somme de carrés résiduelles donne

$$SSE = Y^T[I - X_O(X_O^T X_O)^{-1}X_O^T - X_\tau(X_\tau^T X_\tau)^{-1}X_\tau^T - X_\beta(X_\beta^T X_\beta)^{-1}X_\beta^T - X_{\tau\beta}(X_{\tau\beta}^T X_{\tau\beta})^{-1}X_{\tau\beta}^T]Y.$$

Cette décomposition est équivalente à la décomposition standard de la somme de carrés totale dans une ANOVA à deux facteurs.

Somme de carrés totale :

$$SST = Y^T[I - X_O(X_O^T X_O)^{-1}X_O^T]Y = \sum_i \sum_j \sum_k (y_{ijk} - \overline{y}_{\bullet\bullet\bullet})^2$$

Somme de carrés pour l'effet de A :

$$SSA = Y^T X_\tau(X_\tau^T X_\tau)^{-1}X_\tau^T Y = nJ \sum_i (\overline{y}_{i\bullet\bullet} - \overline{y}_{\bullet\bullet\bullet})^2$$

Somme de carrés pour l'effet de B :

$$SSB = Y^T X_\beta(X_\beta^T X_\beta)^{-1}X_\beta^T Y = nI \sum_j (\overline{y}_{\bullet j\bullet} - \overline{y}_{\bullet\bullet\bullet})^2$$

Somme de carrés pour l'interaction $A * B$:

$$SSAB = Y^T X_{\tau\beta}(X_{\tau\beta}^T X_{\tau\beta})^{-1}X_{\tau\beta}^T Y = n \sum_i \sum_j (\overline{y}_{ij\bullet} - \overline{y}_{i\bullet\bullet} - \overline{y}_{\bullet j\bullet} + \overline{y}_{\bullet\bullet\bullet})^2$$

Dans cet exemple les tailles d'échantillons n_{ij} sont toutes égales à deux. L'utilisation de la régression linéaire pour traiter ce modèle redonne des résultats déjà obtenus dans le cours d'analyse de variance concernant l'analyse de variance à deux facteurs. Les sommes de carrés sont définies de façon unique peu importe les composantes laissées dans le modèle lors du calcul de la somme de carrés pour un effet. Par exemple,

$$SS(A|O, B, A * B) = SS(A|O, B) = SS(A|O)$$

Exemple.

```
proc glm data=batterie ;
class temp materiel ;
model duree = temp|materiel/ e inverse solution;
run;
```

L'algorithme de construction de l'inverse généralisé implanté dans GLM élimine les rangées de la matrice qui sont des combinaisons linéaires des précédentes. Il s'agit donc du dernier paramètre pour chaque groupe de paramètres, ce qui revient à le fixer à 0.

X'X Generalized Inverse (g2)

	Intercept	temp 15	temp 70	materiel 1	materiel 2	materiel 3	Dummy001	Dummy002	Dummy003	Dummy004	Dummy005	Dummy006	duree
Intercept	0.5	-0.5	0	-0.5	-0.5	0	0.5	0.5	0	0	0	0	129.5
temp 15	-0.5	1	0	0.5	0.5	0	-1	-1	0	0	0	0	5.5
temp 70	0	0	0	0	0	0	0	0	0	0	0	0	0
materiel 1	-0.5	0.5	0	1	0.5	0	-1	-0.5	0	0	0	0	-72
materiel 2	-0.5	0.5	0	0.5	1	0	-0.5	-1	0	0	0	0	-11
materiel 3	0	0	0	0	0	0	0	0	0	0	0	0	0
Dummy001	0.5	-1	0	-1	-0.5	0	2	1	0	0	0	0	104.5
Dummy002	0.5	-1	0	-0.5	-1	0	1	2	0	0	0	0	33
Dummy003	0	0	0	0	0	0	0	0	0	0	0	0	0
Dummy004	0	0	0	0	0	0	0	0	0	0	0	0	0
Dummy005	0	0	0	0	0	0	0	0	0	0	0	0	0
Dummy006	0	0	0	0	0	0	0	0	0	0	0	0	0
duree	129.5	5.5	0	-72	-11	0	104.5	33	0	0	0	0	4302

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	129.5000000	B	18.93409623	6.84	0.0005
temp 15	5.5000000	B	26.77685568	0.21	0.8441
temp 70	0.0000000	B	.	.	.
materiel 1	-72.0000000	B	26.77685568	-2.69	0.0361
materiel 2	-11.0000000	B	26.77685568	-0.41	0.6955
materiel 3	0.0000000	B	.	.	.
temp*materiel 15 1	104.5000000	B	37.86819246	2.76	0.0329
temp*materiel 15 2	33.0000000	B	37.86819246	0.87	0.4170
temp*materiel 15 3	0.0000000	B	.	.	.
temp*materiel 70 1	0.0000000	B	.	.	.
temp*materiel 70 2	0.0000000	B	.	.	.
temp*materiel 70 3	0.0000000	B	.	.	.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

General Form of Estimable Functions	
Effect	Coefficients
Intercept	L1
temp 15	L2
temp 70	L1-L2
materiel 1	L4
materiel 2	L5
materiel 3	L1-L4-L5
temp*materiel 15 1	L7
temp*materiel 15 2	L8
temp*materiel 15 3	L2-L7-L8
temp*materiel 70 1	L4-L7
temp*materiel 70 2	L5-L8
temp*materiel 70 3	L1-L2-L4-L5+L7+L8

Remarques :

- Les estimations des paramètres sont celles obtenues avec le dernier paramètre de chaque sous-groupe fixé à 0. On obtient par exemple $\hat{\mu} = \overline{y_{23\bullet}} = 129.5$ et $\hat{\tau}_1 = \overline{y_{13\bullet}} - \overline{y_{23\bullet}} = 135.0 - 129.5 = 5.5$.
- Toutes les fonctions estimables sont obtenues en donnant des valeurs aux 6 quantités $\text{L}1$, $\text{L}2$, $\text{L}4$, $\text{L}5$, $\text{L}7$, $\text{L}8$. On peut donc construire au maximum 6 fonctions estimables indépendantes.
- Les combinaisons linéaires qui lient les lignes (ou les colonnes) de X^TX sont identifiées par les combinaisons des L_i. Par exemple, la 3^e ligne de X^TX est égale à la première ligne moins la deuxième.
- Les paramètres eux-mêmes $(\mu, \tau_i, \beta_j, \tau\beta_{ij})$ ne sont pas estimables car aucune série de valeurs des L_i ne permet de créer la combinaison linéaire souhaitée.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	15023.00000	3004.60000	4.19	0.0551
Error	6	4302.00000	717.00000		
Corrected Total	11	19325.00000			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temp	1	7905.333333	7905.333333	11.03	0.0160
materiel	2	1410.500000	705.250000	0.98	0.4271
temp*materiel	2	5707.166667	2853.583333	3.98	0.0794

- On obtient les mêmes sommes de carrés qu'avec le μ-modèle.
- La somme $SSA + SSB + SSAB + SSE$ donne la somme des carrés totale : $19\,325$.
- On aurait pu obtenir ces sommes de carrés en faisant la différence de deux SS résiduelles : celle du modèle excluant le terme concerné et celle du modèle contenant le terme. Voir les détails ci-dessous.
- Lorsque tous les autres termes sont inclus dans le modèle lors du calcul d'une somme de carrés, on parle de SS de type III. Il existe 4 types de sommes de carrés dont nous discuterons au prochain chapitre. Elles sont toutes égales dans le cas balancé.

Calcul des sommes de carrés par soustration de deux SSE

Facteur inclus	Énoncé MODEL	SSE
O	`duree =`	19 325.00
$O,\ A$	`duree = temp`	11 419.67
$O,\ B$	`duree = materiel`	17 914.50
$O,\ A,\ B$	`duree = temp materiel`	10 009.17
$O,\ A,\ B,\ A*B$	`duree = temp|materiel`	4 302.00

$$\begin{aligned} SSA &= SS(A|O) \\ &= SSE(O) - SSE(A,O) \\ &= 19\,325.00 - 11\,419.67 \\ &= 7\,905.33 \end{aligned}$$

$$\begin{aligned} SSA &= SS(A|O,B) \\ &= SSE(O,B) - SSE(A,O,B) \\ &= 17\,914.50 - 10\,009.17 \\ &= 7\,905.33 \end{aligned}$$

$$\begin{aligned} SSB &= SS(B|O) \\ &= SSE(O) - SSE(B,O) \\ &= 19\,325.00 - 17\,914.50 \\ &= 1\,410.50 \end{aligned}$$

$$\begin{aligned} SSB &= SS(B|O,A) \\ &= SSE(O,A) - SSE(B,O,A) \\ &= 11\,419.67 - 10\,009.17 \\ &= 1\,410.50 \end{aligned}$$

$$\begin{aligned} SSAB &= SS(AB|O,A,B) \\ &= SSE(O,A,B) - SSE(AB,O,A,B) \\ &= 10\,009.17 - 4\,302.00 \\ &= 5\,707.17 \end{aligned}$$

2.4 Exercices

1. Vrai ou Faux ?

(a) Un plan d'expérience est dit *équilibré* si $n_{ij} = n$ pour tout (i, j), donc où $n_{i\bullet} = n_{\bullet j}$ pour tout (i, j).

(b) Peu importe la contrainte utilisée pour résoudre les équations normales du modèle d'analyse de variance, les fonctions linéaires des paramètres sont toujours estimées de façon unique.

(c) Soit G^-, un inverse généralisé de G, i.e. que $GG^-G = G$. Il suit toujours que G est symétrique.

(d) Lorsqu'on modélise un plan complètement aléatoire à l'aide du μ-modèle, les estimateurs des paramètres obtenus par la méthode des moindres carrés sont uniques et sans biais.

2. Calculer un inverse généralisé de la matrice suivante. Vérifier qu'il s'agit bien d'un inverse généralisé.

$$A = \begin{bmatrix} 4 & 3 & 2 & 1 \\ 8 & 6 & 4 & 2 \\ 12 & 9 & 6 & 3 \end{bmatrix}$$

3. En vous basant sur la définition et les propriétés des inverses généralisés de $X'X$, répondez aux questions suivantes :

(a) Si $(X'X)^-$ est un inverse généralisé de $X'X$, montrez que $[(X'X)^-]'$ est aussi un inverse généralisé de $X'X$.

(b) La matrice $H = X(X'X)^-X'$ est la matrice de projection sur l'espace-colonne de X.

(i) Montrer que $HX = X$
(Indice : utiliser le résultat de (a), puis appliquer le lemme suivant :
Si $PX'X = QX'X$, alors $PX' = QX'$).

(ii) Montrer que H est symétrique, i.e. que $H' = H$.

(iii) Montrer que H est idempotente, i.e. que $HH = H$.

4. Reprenons le modèle avec effets à deux facteurs fixes croisés. Supposons que le facteur A comporte $I = 2$ niveaux, et le facteur B ait $J = 3$ niveaux. Les fonctions suivantes des paramètres sont-elles estimables ?

(a) $\beta_1 - \beta_2 + \beta_3$

(b) $2\mu + \beta_1 + \beta_2$

(c) $\mu + \tau_1 + \beta_2 + \tau\beta_{12}$

5. On considère le modèle avec effets pour une analyse de variance à plusieurs effets fixes où on inclut toutes les interactions. Ce modèle surparamétrisé à k paramètres s'écrit sous forme matricielle comme suit :

$$Y = X\beta + \varepsilon \quad \text{où} \quad \begin{cases} Y & = \text{matrice } N \times 1 \text{ des observations}; \\ X & = \text{matrice } N \times k \text{ de variables indicatrices}; \\ \beta & = \text{matrice } k \times 1 \text{ des paramètres}; \\ \varepsilon & = \text{matrice } N \times 1 \text{ des erreurs aléatoires.} \end{cases}$$

On suppose que $\varepsilon \sim NMV(\mathbf{0}, \sigma^2 I)$, où $\mathbf{0}$ est un vecteur de 0 et I est la matrice identité de dimensions $N \times N$.

(a) Trouver $\hat{\beta}$, l'estimateur des moindres carrés du vecteur de paramètres β en minimisant la somme des carrés des erreurs, i.e. $\varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$.
Rappels : Si A est une matrice carrée $k \times k$ et b un vecteur $k \times 1$, alors $\dfrac{\partial Ab}{\partial b} = A^T$ et $\dfrac{\partial b^T Ab}{\partial b} = Ab + A^T b$

(b) Déterminer l'espérance de l'estimateur $\hat{\beta}$ trouvé en (a).

(c) Déterminer la variance de l'estimateur $\hat{\beta}$ trouvé en (a).

6. Les données ci-dessous représentent la durée de vie de batteries de deux marques différentes, utilisées dans trois types d'appareils électroniques (Montgomery [16], p. 201). Utilisez la procédure GLM de SAS pour répondre aux questions.

	Appareil		
Marque	Radio	Caméra	DVD
A	8.6	7.9	5.4
	8.2	8.4	5.7
B	9.4	8.5	5.8
	8.8	8.9	5.9

(a) Estimez les paramètres du μ-modèle.

(b) Utilisez un énoncé "contrast" pour calculer la somme des carrés associée à l'interaction en vous basant sur le μ-modèle.

(c) Estimez les paramètres du modèle avec effets. Quelle est la contrainte utilisée ?

(d) Calculez la somme des carrés associée à l'interaction en vous basant sur le modèle avec effets.

(e) Analysez les effets des facteurs principaux en tenant compte du résultat du test sur l'interaction.

(f) Réalisez une analyse des résidus pour vous assurer que les postulats du modèle sont respectés.

3 Étude des plans déséquilibrés

Lorsque le nombre de répétitions dans chaque combinaison de traitements n'est pas constant, les plans ne sont plus orthogonaux. Cela signifie qu'une corrélation peut être introduite entre les facteurs du simple fait du déséquilibre des tailles d'échantillons. Les formules usuelles des sommes de carrés associées aux facteurs ne peuvent plus être additionnées pour obtenir la somme de carrés totale. De plus, les espérances des moyennes échantillonnales pondérées dépendent des autres paramètres.

Il existe plusieurs schémas de déséquilibre, qui n'ont pas tous la même cause ni le même effet. La raison de l'inégalité aura un impact sur la façon de spécifier les hypothèses de recherche (et par conséquent sur les sommes de carrés calculées). Est-il préférable de comparer des moyennes pondérées ou non pondérées des niveaux de traitement ? Bruno Scherrer ([19]) distingue quatre structures de déséquilibre, que nous reprenons ci-dessous en les comparant avec le plan équilibré (le plan 1 ci-dessous). Les points représentent des observations. La valeur de r est celle du coefficient de corrélation échantillonnal que l'on obtiendrait si les points représentaient des couples de variables (A, B) ayant comme valeurs les modalités 1, 2 et 3 des deux facteurs.

1. Plan orthogonal équilibré : interprétation simplifiée, puissance optimale. À privilégier dans les expériences contrôlées.

Facteur A	$r = 0$		
3	• • •	• • •	• • •
2	• • •	• • •	• • •
1	• • •	• • •	• • •
Facteur $B \Rightarrow$	1	2	3

2. Plan orthogonal partiellement équilibré (en A) : utile si le déséquilibre en B reflète la structure de la population.

Facteur A	$r = 0$		
3	• •	• • • • •	• • •
2	• •	• • • • •	• • •
1	• •	• • • • •	• • •
Facteur $B \Rightarrow$	1	2	3

3. Plan orthogonal déséquilibré : facteur de proportionnalité d'une ligne à l'autre (ou d'une colonne à l'autre). Comme le précédent, utile si le déséquilibre est structurel. Peut parfois provenir d'une contrainte de coût.

Facteur A	$r = 0$		
3	•	••	• • •
2	••	• • ••	• • • • ••
1	•	••	• • •
Facteur $B \Rightarrow$	1	2	3

4. Plan non orthogonal déséquilibré : peut être causé par des données manquantes aléatoirement ou des contraintes pratiques.

Facteur A	$r = 0.5$		
3	•	••	• • • • •
2	••	• • • ••	••
1	• • • • •	••	•
Facteur $B \Rightarrow$	1	2	3

5. Plan non orthogonal incomplet : certaines combinaisons de traitements peuvent être inexistantes ou difficiles à inclure dans l'expérience. Les manques peuvent aussi être fortuits. À traiter avec circonspection, car toutes les combinaisons de moyennes ne seront pas nécessairement estimables.

Facteur A	$r = 0.64$		
3		•	• • ••
2	••	• • • ••	• • •
1	• • •	••	
Facteur $B \Rightarrow$	1	2	3

Dans ce chapitre, nous nous concentrerons sur le modèle à deux facteurs fixes avec interaction. Les modèles à plus de deux facteurs sont une généralisation aisée qui ne ferait qu'alourdir le propos.

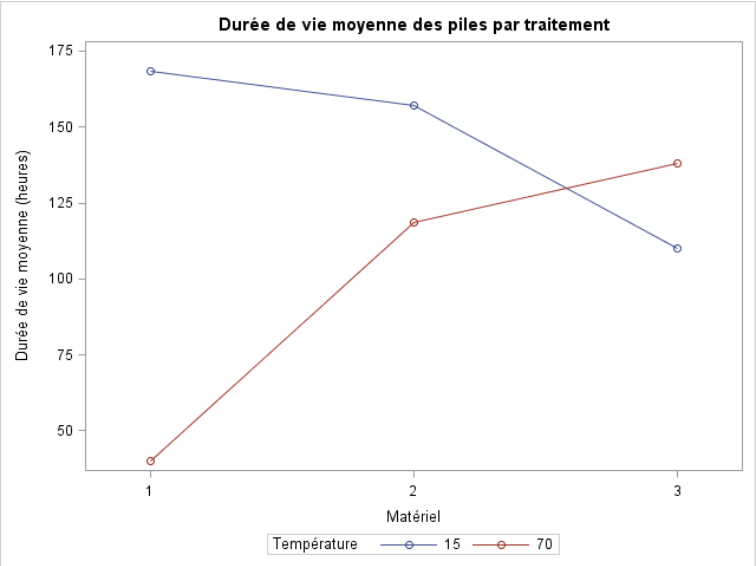
$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \epsilon_{ijk}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, J, \quad k = 1, \cdots, n_{ij}, \quad (7)$$

On supposera d'abord que toutes les tailles d'échantillons sont strictement positives. Le cas des plans avec des cellules vides sera étudié à la section 3.6.

3.1 Un exemple

Reprenons l'exemple du chapitre précédent, mais en modifiant légèrement le nombre d'observations par combinaison de traitements. Les n_{ij} varient maintenant entre 1 et 3.

Température	**Matériel 1**	**Matériel 2**	**Matériel 3**	$\overline{y}_{i\bullet\bullet}$
15°F	170 155 180	188 126	110	*154.83*
$(\overline{y}_{ij\bullet})$	*(168.3)*	*(157.0)*	*(110.0)*	
70°F	40	122 115	120 139 155	*115.17*
$(\overline{y}_{ij\bullet})$	*(40.0)*	*(118.5)*	*(138.0)*	
$\overline{y}_{\bullet j\bullet}$	*136.25*	*137.75*	*131.00*	$\overline{y}_{\bullet\bullet\bullet}$ =*135.00*



Ajustons le modèle complet avec effets (A, B et $A*B$).

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}, \quad \text{où} \quad \begin{cases} i = 1, 2 \\ j = 1, 2, 3 \\ k = 1 \text{ à } 3 \end{cases} \quad \text{et} \quad \varepsilon_{ijk} \sim N(0, \sigma^2)$$

La matrice X du modèle surparamétré pour ces données est la matrice 12×12 suivante

$$X_{\text{surparamétré}} = \left(\begin{array}{c|cc|ccc|ccccc} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

Les paramètres du modèle linéaire réduit sont $\beta = (\mu, \tau_1, \beta_1, \beta_2, \tau\beta_{11}, \tau\beta_{12})$. La matrice X du modèle linéaire réduit, avec la contrainte où tous les paramètres d'un même type somment à 0, est

$$X_{\text{contrainte}} = \left(\begin{array}{c|c|cc|cc} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{array}\right), X^T X = \begin{pmatrix} 12 & 0 & 0 & 0 & 4 & 2 \\ 0 & 12 & 4 & 2 & 0 & 0 \\ 0 & 4 & 8 & 4 & 0 & -2 \\ 0 & 2 & 4 & 8 & -2 & -2 \\ 4 & 0 & 0 & -2 & 8 & 4 \\ 2 & 0 & -2 & -2 & 4 & 8 \end{pmatrix}$$

Remarquons que les colonnes de la matrice X pour l'interaction sont obtenues en faisant le produit d'une colonne pour l'effet de A par une colonne pour B. On note que les quatre composantes de la matrice X (les quatres groupes de colonnes associées à μ, τ, β et $\tau\beta$ respectivement) ne sont pas orthogonales. En effet, la matrice $X^T X$ n'est pas diagonale ni bloc-diagonale.

Lorsqu'on soumet le programme suivant dans SAS, on obtient entre autres la table ANOVA du modèle qui compare les six moyennes entre elles :

```
proc glm data=batterie2 ;
class temp materiel ;
model duree = temp|materiel;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	14522.83333	2904.56667	6.06	0.0243
Error	6	2877.16667	479.52778		
Corrected Total	11	17400.00000			

Si on veut étudier les effets des facteurs et de leur interaction, on remarque que SAS peut fournir trois tableaux de tests de Fisher, présentant des sommes de carrés de type I, II et III, dont les valeurs ne sont plus identiques (comme c'était le cas avec un plan équilibré).

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	1	4720.333333	4720.333333	9.84	0.0201
materiel	2	555.791667	277.895833	0.58	0.5887
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

Source	DF	Type II SS	Mean Square	F Value	Pr > F
temp	1	5175.625000	5175.625000	10.79	0.0167
materiel	2	555.791667	277.895833	0.58	0.5887
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temp	1	5256.734848	5256.734848	10.96	0.0162
materiel	2	1934.308333	967.154167	2.02	0.2138
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

On remarque aussi que la décomposition de la variabilité totale en sommes de carrés n'est plus toujours orthogonale. En d'autres termes, l'équation $SSA+SSB+SSAB+SSE = SST$ n'est plus nécessairement vraie, et donc ces sommes de carrés ne sont plus toujours indépendantes...

Type de SS	$SSA + SSB + SSAB + SSE$
Type I	17 400
Type II	17 855
Type III	19 315
SST	17 400

De plus, les sommes de carrés calculées avec les formules classiques ne sont plus exactement celles que l'on retrouve dans les tableaux...

$$
\begin{aligned}
SSA &= \textstyle\sum_i \sum_j \sum_k (\overline{y}_{i\bullet\bullet} - \overline{y}_{\bullet\bullet\bullet})^2 &&= 4\,720.33 \\
SSB &= \textstyle\sum_i \sum_j \sum_k (\overline{y}_{\bullet j\bullet} - \overline{y}_{\bullet\bullet\bullet})^2 &&= 100.50 \\
SSAB &= \textstyle\sum_i \sum_j \sum_k (\overline{y}_{ij\bullet} - \overline{y}_{i\bullet\bullet} - \overline{y}_{\bullet j\bullet} + \overline{y}_{\bullet\bullet\bullet})^2 &&= 10\,118.50 \\
SSE &= \textstyle\sum_i \sum_j \sum_k (y_{ijk} - \overline{y}_{ij\bullet})^2 &&= 2\,877.17 \\
SST &= \textstyle\sum_i \sum_j \sum_k (y_{ijk} - \overline{y}_{\bullet\bullet\bullet})^2 &&= 17\,400.00
\end{aligned}
$$

Alors que représentent les sommes de carrés de type I, II et III, et quelles hypothèses permettent-elles de tester ? Sont-elles toutes des façons valides de mesurer les effets des traitements ? Sinon, comment faire pour identifier le bon test de comparaison de moyennes ?

3.2 Les sommes de carrés de type I à III

SAS distingue trois types de sommes de carrés pour les plans déséquilibrés (la procédure GLM produit aussi un quatrième type utile pour les plans à cellules vides, et équivalent au type III si tous les $n_{ij} > 0$) :

1. Les sommes de carrés de type I, aussi appelées sommes de carrés séquentielles, tiennent compte des composantes spécifiées avant celle à l'étude. Les sommes de carrés de type I dépendent donc de l'ordre dans lequel les facteurs sont inclus dans le modèle. Le test sur le premier facteur revient à comparer les moyennes pondérées, donc sans corriger pour le niveau de l'autre facteur.

2. Les sommes de carrés de type II, calculées en excluant les composantes contenant l'effet testé. En l'absence d'interaction, le test sur les facteurs simples comparent les moyennes "pondérées par l'effectif efficace" (Scherrer, [19], p. 113), i.e. des poids basés sur les moyennes harmoniques visant à réduire la variance de l'estimation de l'effet des traitements.

3. Les sommes de carrés de type III, calculées en gardant toutes les autres composantes du modèle pour évaluer l'impact d'un effet. Ces tests donnent la même importance à toutes les combinaisons de traitement, sans égard aux différences entre les n_{ij}.

Voici un tableau des sommes de carrés utilisées au numérateur de la statistique F pour tester chacun des facteurs, pour les types I, II et III. Le dénominateur de la statistique F est toujours le MSE lorsque tous les effets sont fixes.

Facteur	Deg. lib.	Type I	Type II	Type III
A	$a-1$	$SS(A\|O)$	$SS(A\|O,B)$	$SS(A\|O,B,AB)$
B	$b-1$	$SS(B\|O,A)$	$SS(B\|O,A)$	$SS(B\|O,A,AB)$
AB	$(a-1)(b-1)$	$SS(AB\|O,A,B)$	$SS(AB\|O,A,B)$	$SS(AB\|O,A,B)$

Remarques :

- $SS(B|O,A) = SSE(O,A) - SSE(O,A,B)$.
- On utilise parfois la notation de réduction de l'erreur, avec la lettre R :

$$SS(B|O,A) = R(\beta|\mu,\tau) = R(\beta_1,...,\beta_J|\mu,\tau_1,...,\tau_I)$$

- Nous reviendrons plus en détails sur les hypothèses testées par chaque type de sommes de carrés en termes de moyennes comparées.

3.3 Fonctions estimables dans le modèle avec effets

La plupart des logiciels pour traiter des données de plan d'expérience utilisent des modèles avec effets. C'est donc en fonction des paramètres $\mu, \tau_i, \beta_j, \tau\beta_{ij}$ que sont exprimées les hypothèses testées. C'est aussi l'écriture en fonction des effets (ou paramètres) qui permet de déterminer ce qu'une statistique estime vraiment. L'objectif de cette section est de bien comprendre le rôle des fonctions estimables dans des plans à deux facteurs déséquilibrés. Nous verrons ensuite quelles hypothèses sont testées par chacun des trois types de sommes de carrés.

Cas du modèle à deux facteurs avec $n_{ij} > 0$

- Rappelons qu'une fonction estimable est une combinaison linéaire des paramètres invariante par rapport au choix de la contrainte retenue pour l'estimation des paramètres. Ce n'est évidemment pas le cas des paramètres eux-mêmes.
- Le vecteur β pour ce modèle compte $1+I+J+IJ = (I+1)(J+1)$ paramètres.
- L'ensemble des fonctions estimables est de dimension IJ (le nombre de cellules de traitement).
- Pour que $L\beta$ soit estimable, il faut que L appartienne à l'espace vectoriel engendré par les lignes de X, i.e. que $L = aX$ pour un vecteur a.
- Ainsi, le vecteur L doit satisfaire $I+J+1$ équations : $LS = 0$ où S est la matrice $(I+1)(J+1) \times (I+J+1)$ des singularités de X telle que $XS = 0$, où X est la matrice des variables explicatives du modèle surparamétrisé (sans contrainte).

Considérons l'exemple abordé précédemment.

Température	**Matériel 1**			**Matériel 2**		**Matériel 3**		
$15°F$	170	155	180	188	126	110		
$70°F$	40			122	115	120	139	155

La matrice X pour ces données est la matrice 12×12 suivante

$$X_{\text{surparamétré}} = \left(\begin{array}{c|cc|ccc|ccccc} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right) = \left(\begin{array}{c|c|c|c} X_O & X_\tau & X_\beta & X_{\tau\beta} \end{array}\right)$$

— La somme des colonnes de X_τ donne la première colonne X_O. On peut dire que $X_{C_1} - X_{C_2} - X_{C_3} = 0$. C'est la première singularité de X.
— La somme des colonnes de X_β donne la première colonne X_O. On peut dire que $X_{C_1} - X_{C_4} - X_{C_5} - X_{C_6} = 0$. C'est la seconde singularité de X.
— La somme des trois premières colonnes de $X_{\tau\beta}$ donne la deuxième colonne de X. On peut dire que $X_{C_2} - X_{C_7} - X_{C_8} - X_{C_9} = 0$.
— Etc.
— La matrice S qui caractérise les 6 singularités de X est

$$S^T = \left(\begin{array}{c|cc|ccc|ccccc} 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \end{array}\right)$$

L'ensemble des fonctions estimables est donné par l'ensemble des $L\beta$ tel que $LS = 0$ (ou $S^T L^T = 0$). On trouve donc

$$\begin{array}{rlllll} L = & (l_1, l_2, l_3 & , l_4, l_5, l_6 & , l_7, l_8, l_9 & , l_{10} & , l_{11} & , l_{12}) \\ = & (l_1, l_2, l_1 - l_2 & , l_4, l_5, l_1 - l_4 - l_5 & , l_7, l_8, l_2 - l_7 - l_8 & , l_4 - l_7 & , l_5 - l_8 & , l_1 - l_2 - l_4 + l_7 - l_5 + l_8) \end{array}$$

Toutes les fonctions estimables des paramètres sont obtenues en donnant des valeurs numériques aux paramètres libres $l_1, l_2, l_4, l_5, l_7, l_8$. Par exemple, si on prend $l_1 = l_2 = l_4 = l_7 = 1$ et $l_5 = l_8 = 0$ on a $L\beta = \mu + \tau_1 + \beta_1 + \tau\beta_{11}$, la moyenne du traitement (1,1). Pour estimer la moyenne du traitement (2,3) on prend $l_1 = 1$ et $l_2 = l_4 = l_5 = l_7 = l_8 = 0$.

Une autre façon de trouver l'ensemble des fonctions estimables est de considérer l'espace vectoriel généré par les lignes de X. En fait X a 6 lignes différentes ou indépendantes (une pour chaque cellule de traitement) et un élément de cet espace vectoriel s'écrit

$$L = aX_{lignes} = \begin{pmatrix} a_1 & a_2 & a_3 & a_4 & a_5 & a_6 \end{pmatrix} \left(\begin{array}{c|cc|ccc|ccccc} 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right).$$

La forme générale du vecteur ligne L définissant une fonction estimable est donc

$$L = \left(\sum a_i, a_1 + a_2 + a_3, a_4 + a_5 + a_6, a_1 + a_4, a_2 + a_5, a_3 + a_6, a_1, a_2, a_3, a_4, a_5, a_6\right).$$

On retrouve le résultat donné plus haut en termes des l_i en posant $l_1 = \sum a_i$, $l_2 = a_1 + a_2 + a_3$, $l_4 = a_1 + a_4$, $l_5 = a_2 + a_5$, $l_7 = a_1$ et $l_8 = a_2$.

Exemple.

Reprenons la procédure GLM pour ajuster le modèle à deux effets fixes, et utilisons des énoncés `estimate` pour investiguer quelques combinaisons linéaires des paramètres.

```
proc glm data=batterie2 ;
class temp materiel ;
model duree = temp materiel temp*materiel / e ;
estimate 'mu11a'             temp 1 0 materiel 1 0 0 / e;
estimate 'mu11b' intercept 1 temp 1 0 materiel 1 0 0 / e;
estimate 'mu11c' intercept 1 temp 1 0 materiel 1 0 0 temp*materiel 1 0 0 0 0 0/e;
run;quit;
```

L'option e de l'énoncé `model` permet de faire afficher la forme générale des fonctions estimables. On retrouve le vecteur L développé plus tôt.

General Form of Estimable Functions	
Effect	**Coefficients**
Intercept	L1
temp 15	L2
temp 70	L1-L2
materiel 1	L4
materiel 2	L5
materiel 3	L1-L4-L5
temp*materiel 15 1	L7
temp*materiel 15 2	L8
temp*materiel 15 3	L2-L7-L8
temp*materiel 70 1	L4-L7
temp*materiel 70 2	L5-L8
temp*materiel 70 3	L1-L2-L4-L5+L7+L8

Examinons les trois énoncés différents qui tentent d'estimer la moyenne du traitement (1,1), soit μ_{11}. L'option e de l'énoncé `estimate` nous permet de voir par quelles valeurs les coefficients manquants ont été remplacés.

Coefficients for Estimate mu11a	
	Row 1
Intercept	0
temp 15	1
temp 70	0
materiel 1	1
materiel 2	0
materiel 3	0
temp*materiel 15 1	0.8333333333
temp*materiel 15 2	0.3333333333
temp*materiel 15 3	0.3333333333
temp*materiel 70 1	0.5
temp*materiel 70 2	0
temp*materiel 70 3	0

Coefficients for Estimate mu11b	
	Row 1
Intercept	1
temp 15	1
temp 70	0
materiel 1	1
materiel 2	0
materiel 3	0
temp*materiel 15 1	0.8333333333
temp*materiel 15 2	0.3333333333
temp*materiel 15 3	0.3333333333
temp*materiel 70 1	0.5
temp*materiel 70 2	0
temp*materiel 70 3	0

Coefficients for Estimate mu11c	
	Row 1
Intercept	1
temp 15	1
temp 70	0
materiel 1	1
materiel 2	0
materiel 3	0
temp*materiel 15 1	1
temp*materiel 15 2	0
temp*materiel 15 3	0
temp*materiel 70 1	0
temp*materiel 70 2	0
temp*materiel 70 3	0

En vertu de l'interprétation que SAS a fait des deux premières fonctions à estimer, elles ne sont pas estimables. SAS a comblé lui-même les valeurs des coefficients manquants.

- `mu11a` ne satisfait pas la contrainte $L3 = L1 - L2$.
- `mu11b` ne satisfait pas la contrainte $L9 = L2 - L7 - L8$.
- `mu11c` est estimable, et son estimateur est $\overline{y_{11\bullet}}$.

```
NOTE: mu11a is not estimable.
NOTE: mu11b is not estimable.
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
mu11c	168.333333	12.6428870	13.31	<.0001

3.4 Comparaison de moyennes pour les modalités d'un facteur

Dans un plan déséquilibré à deux facteurs, il y a plusieurs façons de définir la moyenne d'un niveau d'un facteur, par exemple la moyenne de y à la modalité i du facteur A. On peut entre autres choisir de donner la même importance à toutes les observations ou de donner un poids égal à chaque cellule de traitement, sans tenir compte des tailles d'échantillons. On peut aussi utiliser n'importe quelle autre pondération, si on lui trouve une pertinence, bien entendu !

- Moyenne pondérée par les tailles d'échantillons :

$$\mu_i^{(1)} = \frac{\sum_{j=1}^{b} n_{ij}\mu_{ij}}{n_{i\bullet}} = \mu + \tau_i + \frac{\sum_j n_{ij}(\beta_j + \tau\beta_{ij})}{\sum_j n_{ij}}$$

  — Estimée par la moyenne arithmétique de toutes les observations du niveau i de A :

$$\overline{y_{i\bullet\bullet}} = \frac{\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} y_{ijk}}{\sum_{j=1}^{J} n_{ij}} = \frac{\sum_{j=1}^{J} n_{ij}\overline{y_{ij\bullet}}}{n_{i\bullet}}$$

  — Chaque *observation* a le même poids.

  — Estimateur non biaisé de $\mu_i^{(1)}$.

  — $Var(\overline{y_{i\bullet\bullet}}) = \dfrac{\sigma^2}{n_{i\bullet}}$

  — Ce sont ces moyennes qui sont comparées par le test de type I sur le facteur principal (le premier facteur spécifié dans le modèle, le facteur A).

  — Pour les tests de comparaisons multiples, ces moyennes sont calculées par l'énoncé MEANS, avec l'option `t` ou `lsd`.

  — Dans notre exemple, ces moyennes pour les niveaux du facteur A sont : $\overline{y}_{1\bullet\bullet} = 154.83$ et $\overline{y}_{2\bullet\bullet} = 115.17$ (voir détails dans l'exemple plus loin).

- Moyenne non pondérée (ou équipondérée) :

$$\mu_i^{(2)} = \frac{\sum_{j=1}^{J} \mu_{ij}}{J} = \mu + \tau_i + \frac{\sum_j (\beta_j + \tau\beta_{ij})}{J}$$

— Estimée par la moyenne arithmétique des moyennes des cellules du niveau i de A :

$$\hat{\mu}_i^{(2)} = \frac{\sum_{j=1}^{J} \overline{y_{ij\bullet}}}{J}$$

— Chaque *cellule* a le même poids.

— Estimateur non biaisé de $\mu_i^{(2)}$.

— $Var(\hat{\mu}_i^{(2)}) = \frac{\sigma^2}{J^2} \sum_{j=1}^{J} \frac{1}{n_{ij}}$

— Ce sont ces moyennes qui sont comparées par le test de type III sur les facteurs simples (A et B).

— Pour les tests de comparaisons multiples, ces moyennes sont calculées par l'énoncé LSMEANS. Les LSMEANS s'écrivent en fonction des paramètres du modèle et leurs estimations dépendent du modèle. Ainsi les $\hat{\mu}_i^{(2)}$ obtenues avec un modèle sans interaction (modèle additif) ne sont pas les mêmes que celles obtenues avec le modèle complet.

— Dans notre exemple, ces moyennes pour les niveaux du facteur A sont :

$$\hat{\mu}_1^{(2)} = \frac{168.3 + 157.0 + 110.0}{3} = 145.11 \text{ et } \hat{\mu}_2^{(2)} = \frac{40.0 + 118.5 + 138.0}{3} = 98.83$$

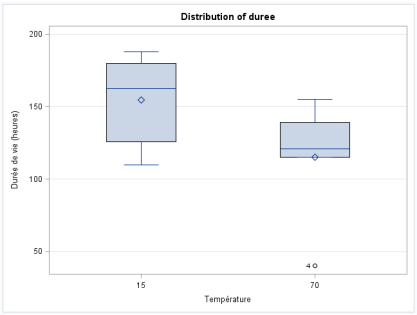
Pour des données observationnelles, les n_{ij} caractérisent la population et on voudra travailler avec les $\hat{\mu}_i^{(1)}$, soit les moyennes pondérées et l'analyse de type I. Dans le cas contraire, i.e. quand le déséquilibre n'est pas structurel mais aléatoire ou dû à des contraintes de coût ou d'éthique par exemple, on va préférer les $\hat{\mu}_i^{(2)}$, soit les moyennes non pondérées et l'analyse de type III.

Exemple.

Température	Matériel 1	Matériel 2	Matériel 3	$\overline{y}_{i\bullet\bullet}$
15°F	170 155 180	188 126	110	*154.83*
$(\overline{y}_{ij\bullet})$	*(168.3)*	*(157.0)*	*(110.0)*	
70°F	40	122 115	120 139 155	*115.17*
$(\overline{y}_{ij\bullet})$	*(40.0)*	*(118.5)*	*(138.0)*	
$\overline{y}_{\bullet j\bullet}$	*136.25*	*137.75*	*131.00*	$\overline{y}_{\bullet\bullet\bullet}$ =*135.00*

```
proc glm data=batterie2 ;
class temp materiel ;
model duree = temp|materiel;
means   temp / lsd ;
lsmeans temp / e stderr pdiff;
run;quit;
```

Résultats de l'énoncé MEANS (moyennes pondérées)



t Tests (LSD) for duree

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	6
Error Mean Square	479.5278
Critical Value of t	2.44691
Least Significant Difference	30.936

Means with the same letter are not significantly different.

t Grouping	Mean	N	temp
A	154.83	6	15
B	115.17	6	70

Notons au passage que la valeur observée d'une statistique t qui comparerait ces moyennes vaudrait :

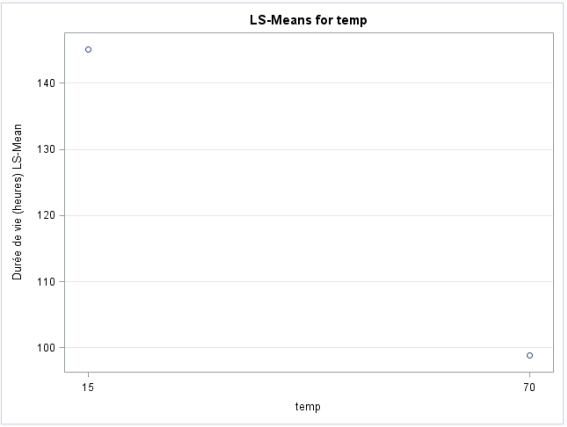
$$t_{obs} = \frac{\overline{y}_{1\bullet\bullet} - \overline{y}_{2\bullet\bullet}}{\sqrt{MSE\left(\frac{1}{n_{1\bullet}} + \frac{1}{n_{2\bullet}}\right)}} = 3.14$$

Il s'agit de la racine carrée de la statistique observée F pour le test de type I sur le facteur température.

Résultats de l'énoncé LSMEANS (moyennes non pondérées)

The GLM Procedure
Least Squares Means

Coefficients for temp Least Square Means		
	temp Level	
Effect	15	70
Intercept	1	1
temp 15	1	0
temp 70	0	1
materiel 1	0.33333333	0.33333333
materiel 2	0.33333333	0.33333333
materiel 3	0.33333333	0.33333333
temp*materiel 15 1	0.33333333	0
temp*materiel 15 2	0.33333333	0
temp*materiel 15 3	0.33333333	0
temp*materiel 70 1	0	0.33333333
temp*materiel 70 2	0	0.33333333
temp*materiel 70 3	0	0.33333333



temp	duree LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1=LSMean2 Pr > |t|
15	145.111111	9.883399	<.0001	0.0162
70	98.833333	9.883399	<.0001	

Dans l'exemple les tailles d'échantillons sont de 1, 2 et 3 pour $i = 1$ et $i = 2$, ainsi les `lsmeans` pour A ont la même variance. Remarquons que le seuil observé, 0.0162 est le même que celui du test F de type III sur le facteur température.

3.5 Hypothèses testées par les sommes de carrés de type I à III

Puisque les sommes de carrés ne sont plus égales dans le cas déséquilibré, il faut savoir quelles hypothèses sont testées par chacune d'entre elles. Le tableau ci-dessous, tiré de Littell et al. [11], résume les hypothèses pour un modèle à deux facteurs fixes, en termes des μ_{ij}. Les fonctions estimables nous aideront à interpréter les tests basés sur les sommes de carrés de types I, II et III, en fonction du modèle avec effets.

Type de SS	Hypothèse nulle
	Effet de A
I	$\sum_{j=1}^{J} \frac{n_{1j}}{n_{1\bullet}}\mu_{1j} = \ldots = \sum_{j=1}^{J} \frac{n_{Ij}}{n_{I\bullet}}\mu_{Ij}$
II	$\sum_{j=1}^{J} n_{1j}\mu_{1j} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{1j}n_{ij}}{n_{\bullet j}}\mu_{ij}$, ... , $\sum_{j=1}^{J} n_{Ij}\mu_{Ij} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{Ij}n_{ij}}{n_{\bullet j}}\mu_{ij}$
III, IV	$\overline{\mu_{1\bullet}} = \ldots = \overline{\mu_{I\bullet}}$ où $\overline{\mu_{i\bullet}} = \frac{1}{J}\sum_{j=1}^{J}\mu_{ij}$
	Effet de B
I, II	$\sum_{i=1}^{I} n_{i1}\mu_{i1} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{i1}n_{ij}}{n_{i\bullet}}\mu_{ij}$, ... , $\sum_{i=1}^{I} n_{iJ}\mu_{iJ} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{iJ}n_{ij}}{n_{i\bullet}}\mu_{ij}$
III, IV	$\overline{\mu_{\bullet 1}} = \ldots = \overline{\mu_{\bullet J}}$ où $\overline{\mu_{\bullet j}} = \frac{1}{I}\sum_{i=1}^{I}\mu_{ij}$
	Effet de A*B
I, II, III, IV	$\mu_{ij} - \mu_{i'j} = \mu_{ij'} - \mu_{i'j'}$

Nous nous baserons sur notre exemple pour illustrer les différentes hypothèses.

Exemple.

```
proc glm data=batterie2 ;
class temp materiel ;
model duree = temp|materiel / e1 e2 e3;
```

Coefficients de type I

Effect	temp	materiel	temp*materiel
Intercept	0	0	0
temp 15	L2	0	0
temp 70	-L2	0	0
materiel 1	0.3333*L2	L4	0
materiel 2	0	L5	0
materiel 3	-0.3333*L2	-L4-L5	0
temp*materiel 15 1	0.5*L2	0.6*L4-0.075*L5	L7
temp*materiel 15 2	0.3333*L2	-0.2*L4+0.4*L5	L8
temp*materiel 15 3	0.1667*L2	-0.4*L4-0.325*L5	-L7-L8
temp*materiel 70 1	-0.1667*L2	0.4*L4+0.075*L5	-L7
temp*materiel 70 2	-0.3333*L2	0.2*L4+0.6*L5	-L8
temp*materiel 70 3	-0.5*L2	-0.6*L4-0.675*L5	L7+L8

de type II

temp	materiel	temp*materiel
0	0	0
L2	0	0
-L2	0	0
0	L4	0
0	L5	0
0	-L4-L5	0
0.3*L2	0.6*L4-0.075*L5	L7
0.4*L2	-0.2*L4+0.4*L5	L8
0.3*L2	-0.4*L4-0.325*L5	-L7-L8
-0.3*L2	0.4*L4+0.075*L5	-L7
-0.4*L2	0.2*L4+0.6*L5	-L8
-0.3*L2	-0.6*L4-0.675*L5	L7+L8

de type III

temp	materiel	temp*materiel
0	0	0
L2	0	0
-L2	0	0
0	L4	0
0	L5	0
0	-L4-L5	0
0.3333*L2	0.5*L4	L7
0.3333*L2	0.5*L5	L8
0.3333*L2	-0.5*L4-0.5*L5	-L7-L8
-0.3333*L2	0.5*L4	-L7
-0.3333*L2	0.5*L5	-L8
-0.3333*L2	-0.5*L4-0.5*L5	L7+L8

Rappelons que les tests d'hypothèses sont de la forme $H_0 : L\beta = 0$ contre l'alternative bilatérale, où β est le vecteur des 12 paramètres du modèle avec effets. Le test est basé sur la statistique F_0, obtenue par différence de SSE au numérateur, ou comme suit :

$$F_0 \;=\; (L\hat{\beta})^T[L(X^TX)^-L^T \;\; MSE]^{-1}(L\hat{\beta}) \;\sim\; F_{p,\nu} \qquad (\text{si } H_0 \text{ est vraie})$$

.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	1	4720.333333	4720.333333	9.84	0.0201
materiel	2	555.791667	277.895833	0.58	0.5887
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

Source	DF	Type II SS	Mean Square	F Value	Pr > F
temp	1	5175.625000	5175.625000	10.79	0.0167
materiel	2	555.791667	277.895833	0.58	0.5887
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temp	1	5256.734848	5256.734848	10.96	0.0162
materiel	2	1934.308333	967.154167	2.02	0.2138
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

• Test sur l'interaction

Ce test est le même pour tous les types de sommes de carrés. L'absence d'interaction se traduit par le parallélisme des courbes de moyennes, ce qui signifie que l'effet d'un facteur est le même pour tous les niveaux d'un autre facteur. SAS indique qu'il y a deux coefficients libres dans le vecteur L : $L7$ et $L8$. On peut donc tester simultanément deux équations indépendantes. C'est le nombre de degrés de liberté.

$\mu-$modèle :	H_0 :	$\mu_{11} - \mu_{21} = \mu_{12} - \mu_{22}$	(voir tableau
		$\mu_{11} - \mu_{21} = \mu_{13} - \mu_{23}$	début de section)
modèle avec effets :	H_0 :	$\tau\beta_{11} - \tau\beta_{21} - \tau\beta_{12} + \tau\beta_{22} = 0$	$(L7, L8) = (1, -1)$
		$\tau\beta_{11} - \tau\beta_{21} - \tau\beta_{13} + \tau\beta_{23} = 0$	$(L7, L8) = (1, 0)$

Le seuil observé du test F est de 0.0134, bien en deçà du seuil nominal habituel de 0.05.Les courbes de moyennes n'étant pas du tout parallèles dans notre exemple (voir le graphique en début de chapitre), on pouvait se douter que l'interaction serait significative.

Normalement, on ne devrait interpréter les tests sur les facteurs simples que si l'interaction n'est pas significative, autrement l'effet de A varie selon les niveaux de B (et vice versa), et un effet global ne peut être évalué. Il faut théoriquement fixer les niveaux d'un facteur pour tester l'effet de l'autre. Nous interpréterons quand même les résultats ci-dessous, mais dans un but pédagogique seulement.

• Test sur le facteur A (température)

Sommes de carrés de type I

Selon le tableau des hypothèses en fonction des μ_{ij}, la somme de carrés de type I pour l'effet de A vérifie l'égalité suivante :

$$H_0 : \frac{n_{11}\mu_{11} + n_{12}\mu_{12} + n_{13}\mu_{13}}{n_{1\bullet}} = \frac{n_{21}\mu_{21} + n_{22}\mu_{22} + n_{23}\mu_{23}}{n_{2\bullet}}$$

Bref, ce test évalue si les moyennes *pondérées* de tous les niveaux de A sont égales. Toutes les observations ont le même poids, et on ne corrige pas pour le facteur matériel. Cela serait approprié si le déséquilibre dans l'échantillon était proportionnel à la production d'une usine, par exemple.

En remplaçant les n_{ij} de notre exemple dans l'équation, on obtient l'hypothèse nulle suivante :

$$H_0 : \frac{3\mu_{11} + 2\mu_{12} + \mu_{13}}{6} = \frac{\mu_{21} + 2\mu_{22} + 3\mu_{23}}{6}$$

En termes du modèle avec effets, il suffit de poser $L2 = 1$ dans le vecteur L défini dans le tableau fourni par SAS pour la forme des fonctions estimables pour les sommes de carrés de type I, ou de remplacer chaque μ_{ij} par son équivalent $\mu + \tau_i + \beta_j + \tau\beta_{ij}$ dans l'équation ci-dessus.

Le test est significatif (seuil observé 0.0201) dans notre exemple, bien que son interprétation soit inappropriée à cause de l'interaction.

Sommes de carrés de type II

La pondération des moyennes de cellules est déterminée de façon à minimiser la variance des effets des traitements (la différence entre deux niveaux d'un facteur). Cette pondération est pertinente pour tester le facteur A dans les cas de fort déséquilibre, d'absence d'interaction, d'égalité des variances, lorsqu'on veut corriger pour le facteur B.

Pour comprendre l'hypothèse nulle associée à la somme de carrés de type II pour A, on peut utiliser la construction suivante :

1. On soustrait des moyennes de traitement l'effet simple de B, i.e. la moyenne pondérée des cellules de ce niveau de B et on obtient :

$$\mu_{ij}^{-B} = \mu_{ij} - \sum_i \frac{n_{ij}}{n_{\cdot j}} \mu_{ij}$$

   Par exemple, pour le traitement (1,1), cette moyenne est :

$$\mu_{11}^{-B} = \mu_{11} - \frac{n_{11}\mu_{11} + n_{21}\mu_{21}}{n_{\cdot 1}} = \mu_{11} - \frac{3\mu_{11} + \mu_{21}}{4}$$

2. On calcule la somme de carrés pour l'hypothèse qui compare les μ_{ij}^{-B} pondérées :

$$H_0 : \sum_j n_{ij}\mu_{ij}^{-B} = 0 \qquad \text{pour tout } i$$

En réexprimant, sous forme des paramètres du modèle avec effets, l'hypothèse testée par cette somme de carrés présentée plus tôt, on retrouve les résultats de la sortie SAS. Dans notre exemple, la somme de carrés de type II pour A teste l'hypothèse

$$\begin{aligned}
H_0 &: n_{11}\mu_{11}^{-B} + n_{12}\mu_{12}^{-B} + n_{13}\mu_{13}^{-B} = 0 \\
H_0 &: \sum_{j=1}^{3} n_{1j}\mu_{1j} = \sum_{i=1}^{2}\sum_{j=1}^{3} \frac{n_{1j}n_{ij}}{n_{\bullet j}}\mu_{ij} \\
H_0 &: n_{1\bullet}\tau_1 + \sum_j n_{1j}\tau\beta_{1j} - \sum_j\sum_i \frac{n_{1j}n_{ij}}{n_{\bullet j}}(\tau_i + \tau\beta_{ij}) = 0 \\
H_0 &: (\tau_1 - \tau_2)\sum_j \frac{n_{1j}n_{2j}}{n_{\bullet j}} + \sum_j \frac{n_{1j}n_{2j}}{n_{\bullet j}}(\tau\beta_{1j} - \tau\beta_{2j}) = 0.
\end{aligned}$$

En remplaçant les n_{ij} par les tailles d'échantillon de l'exemple on obtient,

$$(5/2)\tau_1 - (5/2)\tau_2 + (3/4)(\tau\beta_{11} - \tau\beta_{21}) + (\tau\beta_{12} - \tau\beta_{22}) + (3/4)(\tau\beta_{13} - \tau\beta_{23}) = 0$$

On peut réécrire cette hypothèse comme $H_0 : L\beta = 0$ où

$$L = (0, 5/2, -5/2, 0, 0, 0, 3/4, 1, 3/4, -3/4, -1, -3/4)$$

est proportionnel au vecteur

$$(0, 1, -1, 0, 0, 0, 0.3, 0.4, 0.3, -0.3, -0.4, -0.3)$$

associé à la somme de carrés de type II pour A dans le tableau SAS des fonctions estimables lorsqu'on pose L2=1.

Sommes de carrés de type III

$\mu-$modèle : $H_0 : \quad \mu_{11} + \mu_{12} + \mu_{13} = \mu_{21} + \mu_{22} + \mu_{23}$

mod. avec effets : $H_0 : \quad \tau_1 - \tau_2 + \dfrac{\tau\beta_{11} + \tau\beta_{12} + \tau\beta_{13} - \tau\beta_{21} - \tau\beta_{22} - \tau\beta_{23}}{3} = 0$
$(L2 = 1)$

On remarque que les sommes de carrés de type III testent des hypothèses qui comparent des moyennes non pondérées pour les tailles d'échantillon dans chacun des traitements. C'est la plus appropriée lorsque le déséquilibre est fortuit.

Rappelons que l'interprétation des tests sur les facteurs simples est inappropriée lorsque l'interaction est significative, comme c'est le cas ici. On devrait normalement fixer les niveaux de matériel et faire trois tests de comparaison des niveaux de température. Cela est possible avec l'option `slice=` de l'énoncé LSMEANS.

```
lsmeans temp*materiel / slice=materiel ;
```

temp*materiel Effect Sliced by materiel for duree					
materiel	DF	Sum of Squares	Mean Square	F Value	Pr > F
1	1	12352	12352	25.76	0.0023
2	1	1482.250000	1482.250000	3.09	0.1292
3	1	588.000000	588.000000	1.23	0.3106

• Test sur le facteur B (matériel)

Sommes de carrés de types I et II

Il s'agit de la comparaison de moyennes pondérées par "l'effectif efficace". Il y a deux coefficients libres dans le vecteur L : $L4$ et $L5$. On peut donc tester simultanément deux équations indépendantes.

Selon le tableau des hypothèses en fonction des μ_{ij}, on devrait tester les trois équations suivantes (dont seulement deux sont indépendantes) :

$$n_{11}\mu_{11}+n_{21}\mu_{21} = \frac{n_{11}}{n_{1\bullet}}\left[n_{11}\mu_{11}+n_{12}\mu_{12}+n_{13}\mu_{13}\right]+\frac{n_{21}}{n_{2\bullet}}\left[n_{21}\mu_{21}+n_{22}\mu_{22}+n_{23}\mu_{23}\right]$$

$$n_{12}\mu_{12}+n_{22}\mu_{22} = \frac{n_{12}}{n_{1\bullet}}\left[n_{11}\mu_{11}+n_{12}\mu_{12}+n_{13}\mu_{13}\right]+\frac{n_{22}}{n_{2\bullet}}\left[n_{21}\mu_{21}+n_{22}\mu_{22}+n_{23}\mu_{23}\right]$$

$$n_{13}\mu_{13}+n_{23}\mu_{23} = \frac{n_{13}}{n_{1\bullet}}\left[n_{11}\mu_{11}+n_{12}\mu_{12}+n_{13}\mu_{13}\right]+\frac{n_{23}}{n_{2\bullet}}\left[n_{21}\mu_{21}+n_{22}\mu_{22}+n_{23}\mu_{23}\right]$$

Si on remplace les valeurs de n_{ij} de notre exemple dans les deux premières équations et qu'on garde les coefficients entiers, on obtient l'hypothèse nulle suivante :

$$\begin{aligned} H_0: \quad & 9\mu_{11}+5\mu_{21} = 6\mu_{12}+3\mu_{13}+2\mu_{22}+3\mu_{23} \\ & 4\mu_{12}+4\mu_{22} = 3\mu_{11}+\mu_{13}+\mu_{21}+3\mu_{23} \end{aligned}$$

En remplaçant les μ_{ij} par $\mu+\tau_i+\beta_j+\tau\beta_{ij}$, on obtient la forme suivante (qui respecte les contraintes de $(L4, L5)$ spécifiées dans la sortie SAS) :

$$\begin{aligned} H_0: \quad & 14\beta_1-8\beta_2-6\beta_3+9\tau\beta_{11}-6\tau\beta_{12}-3\tau\beta_{13}+5\tau\beta_{21}-2\tau\beta_{22}-3\tau\beta_{23}=0 \\ & (L4, L5) = (14, -8) \\ & -4\beta_1+8\beta_2-4\beta_3-3\tau\beta_{11}+4\tau\beta_{12}-\tau\beta_{13}-\tau\beta_{21}+4\tau\beta_{22}-3\tau\beta_{23}=0 \\ & (L4, L5) = (-4, 8) \end{aligned}$$

Remarques :

— On aurait pu spécifier d'autres constantes pour $(L4, L5)$ et on aurait obtenu la même somme de carrés et donc testé la même hypothèse. Par exemple, on aurait pu choisir les deux équations correspondant à $(L4, L5) = (1, 0)$ et $(L4, L5) = (1, -1)$, la somme de carrés pour l'effet de matériel aurait quand même été 555.79.
— Cette construction de l'hypothèse met en lumière que l'hypothèse nulle testée par $SS_2(B)$ ne dépend pas des paramètres τ_i associés à l'effet simple de A.

Sommes de carrés de type III

$$\mu-\text{modèle :} \quad H_0 : \quad \mu_{11} + \mu_{21} = \mu_{12} + \mu_{22}$$
$$\mu_{11} + \mu_{21} = \mu_{13} + \mu_{23}$$

$$\text{mod. effets :} \quad H_0 : \quad \beta_1 - \beta_2 + \frac{\tau\beta_{11} + \tau\beta_{21} - \tau\beta_{12} - \tau\beta_{22}}{2} = 0 \quad (L4, L5) = (1, -1)$$
$$\beta_1 - \beta_3 + \frac{\tau\beta_{11} + \tau\beta_{21}}{2} - \tau\beta_{13} - 3\tau\beta_{23} = 0 \quad (L4, L5) = (1, 0)$$

On remarque que les sommes de carrés de type III testent des hypothèses qui comparent des moyennes non pondérées pour les tailles d'échantillon dans chacun des traitements.

Quelle somme de carrés choisir pour tester les effets ?

En quelques mots, si les différences entre les n_{ij} sont dues aux aléas de l'expérience, ou à des contraintes physiques ou économiques n'ayant pas de lien avec la répartition naturelle de la population, on utilisera de préférence les sommes de carrés de type III car les hypothèses testées ne dépendent pas des tailles d'échantillon. Si on traite des données d'observation, obtenues en échantillonnant une vraie population, les n_{ij} sont associées à l'importance relative des différentes cellules dans la population. Ceci justifie l'utilisation de la somme de carrés de type I pour A.

Exemple.

On échantillonne des citoyens d'une ville et on crée des cellules en croisant l'âge (en I modalités) et le niveau d'éducation (en J modalités). La variable d'intérêt est le salaire annuel. La somme de carrés de type I pour âge compare les salaires selon les classes d'âge sans tenir compte de la scolarité. Chaque individu a le même poids dans le calcul.

La somme de carrés de type III compare les salaires des classes d'âge en corrigeant pour les différences de niveaux d'éducation entre les classes d'âge. Chaque niveau d'éducation a le même poids dans le calcul.

Dans la prochaine section, nous traiterons de plans d'expérience avec des cellules vides, c'est-à-dire avec des n_{ij} égaux à 0. C'est dans ce contexte que la notion de fonction estimable est vraiment utile.

3.6 Schéma avec des tailles d'échantillon n_{ij} nulles

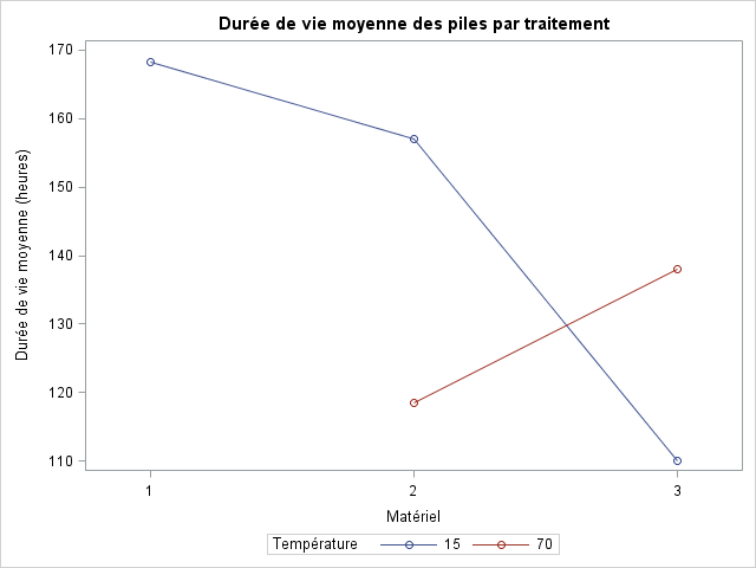
On travaille avec le modèle à deux facteurs avec interaction, où quelques tailles d'échantillons sont nulles :

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \epsilon_{ijk}, \quad i = 1, \cdots, I, \quad j = 1, \cdots, J, \quad k = 1, \cdots, n_{ij}, \quad (8)$$

Reprenons l'exemple avec un facteur A (température) à 2 modalités et un facteur B (matériel) à 3 modalités. Les tailles d'échantillon varient de 0 à 3. Les données sont

Température	**Matériel 1**	**Matériel 2**	**Matériel 3**	$\overline{y}_{i\bullet\bullet}$
15°F	170 155 180	188 126	110	*154.83*
$(\overline{y}_{ij\bullet})$	*(168.3)*	*(157.0)*	*(110.0)*	
70°F		122 115	120 139 155	*130.20*
$(\overline{y}_{ij\bullet})$		*(118.5)*	*(138.0)*	
$\overline{y}_{\bullet j\bullet}$	*168.3*	*137.75*	*131.00*	$\overline{y}_{\bullet\bullet\bullet}$ =*143.64*

Puisque n_{21} est nulle, μ_{21} n'est pas estimable. Dans un premier temps on peut traiter ces données selon un plan d'analyse de variance à un facteur ayant 5 modalités et tester l'hypothèse d'homogénéité des moyennes dans ce modèle. Si elle est rejetée, on peut poursuivre l'analyse et examiner dans quelle mesure les facteurs interagissent. Pour ce faire on identifie des contrastes associés à l'interaction ou à l'effet simple des deux facteurs.



3.6.1 Le μ-modèle

Interaction $A * B$ (température*matériel)

Lorsque toutes les tailles d'échantillon étaient positives, on a écrit l'hypothèse d'absence d'interaction (à deux degrés de liberté) comme

$$\begin{aligned} H_0^{A*B}: \quad & \mu_{11} - \mu_{21} = \mu_{12} - \mu_{22} \\ & \mu_{11} - \mu_{21} = \mu_{13} - \mu_{23} \\ & \mu_{12} - \mu_{22} = \mu_{13} - \mu_{23} \end{aligned}$$

ou bien

$$H_0^{A*B}: \begin{pmatrix} 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & -1 & -1 & 0 & 1 \end{pmatrix} \beta = L_{A*B}\beta = 0,$$

où $\beta = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23})^T$. Puisque μ_{21} n'est pas estimable, on ne peut tester l'interaction qu'en ne considérant que les deux dernières modalités du facteur B. Un test d'interaction partiel considère seulement la troisième équation de l'hypothèse nulle complète

$$H_0^{A*B,partiel}: \mu_{12} - \mu_{22} = \mu_{13} + \mu_{23}$$

ou bien

$$H_0^{A*B,partiel}: \begin{pmatrix} 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix} \beta = 0.$$

Les contrastes d'interaction comprenant μ_{21}, comme $\mu_{11} - \mu_{21} - \mu_{13} + \mu_{23}$, ne sont pas estimables. De plus, dans la spécification de ce contraste, il faudra retirer le 0 associé à μ_{21}, car cette moyenne ne fait pas partie du vecteur des paramètres.

Si $H_0^{A*B,partiel}$ n'est pas rejetée, il faut faire l'hypothèse que $\mu_{11} - \mu_{21} = \mu_{13} + \mu_{23}$ pour pouvoir déclarer que les deux facteurs n'interagissent pas. *Des informations supplémentaires sont requises pour conclure que les facteurs n'interagissent pas*. Si l'hypothèse d'absence d'interaction est plausible, on peut énoncer des contrastes qui comparent les modalités d'un facteur entre elles.

Facteur A (température)

On peut comparer les deux modalités de la température pour les matériaux 2 et 3 avec

$$L_{A(mat2-3)} = \begin{pmatrix} 0 & 1 & 1 & 0 & -1 & -1 \end{pmatrix}.$$

ou bien pour le matériel 3 avec

$$L_{A(mat3)} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & -1 \end{pmatrix}$$

Facteur B (matériel)

On peut comparer les matériaux 1 et 2 avec le matériel 3 lorsque $i = 1$ (température $= 15°F$), et les matériaux 2 et 3 entre eux lorsque $i = 2$ (température $= 70°F$). Ceci donne

$$L_{B(version1)} = \begin{pmatrix} 1 & 1 & -2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix}.$$

Un autre choix tout aussi judicieux pour comparer les matériaux est

$$L_{B(version2)} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 1 & -1 \end{pmatrix}.$$

On peut facilement tester ces hypothèses $H_0 : L\beta = 0$ avec un μ-modèle pour les 5 moyennes estimables. On élimine dans le code SAS la 4^e colonne des matrices L ci-dessus car elle correspond à μ_{21}.

Exemple.

Comparaison des 5 moyennes entre elles

```
proc glm data=batterie3;
class temp  materiel ;
model duree=temp*materiel;
run;quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	4677.378788	1169.344697	2.44	0.1579
Error	6	2877.166667	479.527778		
Corrected Total	10	7554.545455			

Normalement, l'analyse s'arrêterait ici car aucune différence significative n'est détectée entre les 5 moyennes... Poursuivons tout de même pour comprendre comment nous aurions procédé dans le cas d'un test significatif.

Analyse avec le μ-modèle sur SAS

```
proc glm data=batterie3;
class temp  materiel ;
model duree=temp*materiel/noint;
contrast 'A*B partiel' temp*materiel 0 1 -1   -1  1;
contrast 'A(mat2-3)'   temp*materiel 0 1  1   -1 -1;
contrast 'A(mat3)'     temp*materiel 0 0  1    0 -1;
contrast 'B(version1)' temp*materiel 1 1 -2    0  0,
                       temp*materiel 0 0  0    1 -1;
contrast 'B(version2)' temp*materiel 1 0 -1    0  0,
                       temp*materiel 0 1 -1    1 -1;
run;quit;
```

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
A*B partiel	1	1895.250000	1895.250000	3.95	0.0940
A(mat2-3)	1	47.250000	47.250000	0.10	0.7642
A(mat3)	1	588.000000	588.000000	1.23	0.3106
B(version1)	2	2751.840230	1375.920115	2.87	0.1335
B(version2)	2	2718.859649	1359.429825	2.83	0.1359

Tous les tests sont non significatifs au seuil de 5%.

3.6.2 Le modèle avec effets

Le traitement de schémas avec des cellules vides à l'aide d'un modèle avec effets est plus problématique. En effet les contraintes de type "somme à 0", ou "dernière composante égale à 0", ne sont pas suffisantes pour obtenir un modèle où tous les paramètres sont estimables. La construction d'une matrice de variables explicatives pour ce problème est complexe. La procédure GLM de SAS traite ce type de données en créant un quatrième type de somme de carrés.

Nous étudierons le cas de notre exemple, puisque la construction des sommes de carrés de type IV dépend de la position et du nombre de cellules vides.

Exemple.

```
proc glm data=batterie3 ;
class temp materiel ;
model duree= temp|materiel/ e e4 ss4;
run;quit;
```

General Form of Estimable Functions	
Effect	Coefficients
Intercept	L1
temp 15	L2
temp 70	L1-L2
materiel 1	L4
materiel 2	L5
materiel 3	L1-L4-L5
temp*materiel 15 1	L4
temp*materiel 15 2	L8
temp*materiel 15 3	L2-L4-L8
temp*materiel 70 2	L5-L8
temp*materiel 70 3	L1-L2-L5+L8

Type IV Estimable Functions			
	Coefficients		
Effect	temp	materiel	temp*materiel
Intercept	0	0	0
temp 15	L2	0	0
temp 70	-L2	0	0
materiel 1	0	L4	0
materiel 2	0	L5	0
materiel 3	0	-L4-L5	0
temp*materiel 15 1	0	L4	0
temp*materiel 15 2	0.5*L2	0.5*L5	L8
temp*materiel 15 3	0.5*L2	-L4-0.5*L5	-L8
temp*materiel 70 2	-0.5*L2	0.5*L5	-L8
temp*materiel 70 3	-0.5*L2	-0.5*L5	L8

Source	DF		Type IV SS	Mean Square	F Value	Pr > F
temp	1	*	47.250000	47.250000	0.10	0.7642
materiel	2	*	2718.859649	1359.429825	2.83	0.1359
temp*materiel	1		1895.250000	1895.250000	3.95	0.0940

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

L'énoncé e4 écrit la forme générale des hypothèses testées par ces sommes de carrés en fonction des paramètres du modèle. Les sommes de carrés de type IV sont les mêmes que celles associées aux contrastes `"A (mat2-3)"`, `"B(version2)"` et `"A*B partiel"` que nous avons soumis à partir du μ-modèle.

Rappel des contrastes du μ-modèle correspondants :

$L_{A(mat2-3)} = \begin{pmatrix} 0 & 1 & 1 & 0 & -1 & -1 \end{pmatrix}$

$L_{B(version2)} = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0.5 & -0.5 & 0 & 0.5 & -0.5 \end{pmatrix}$

$L_{A*B(partiel)} = \begin{pmatrix} 0 & 1 & -1 & 0 & -1 & 1 \end{pmatrix}$

Facteur A = température (aux matériels 2 et 3)

H_0 modèle avec effets : $\tau_1 - \tau_2 + \frac{\tau\beta_{12} + \tau\beta_{13}}{2} - \frac{\tau\beta_{22} + \tau\beta_{23}}{2} = 0$

H_0 $\mu-$modèle : $\frac{1}{2}(\mu_{12} + \mu_{13} - \mu_{22} - \mu_{23}) = 0$

Facteur B = matériel

H_0 modèle avec effets : $\beta_1 - \beta_3 + \tau\beta_{11} - \tau\beta_{13} = 0$

$\beta_2 - \beta_3 + \frac{\tau\beta_{12} + \tau\beta_{22}}{2} - \frac{\tau\beta_{13} + \tau\beta_{23}}{2} = 0$

H_0 $\mu-$modèle : $\mu_{11} - \mu_{13} = 0$

$\frac{1}{2}(\mu_{12} + \mu_{22} - \mu_{13} - \mu_{23}) = 0$

Facteur A*B (test partiel)

H_0 modèle avec effets : $\tau\beta_{12} - \tau\beta_{13} - \tau\beta_{22} + \tau\beta_{23} = 0$

H_0 $\mu-$modèle : $\mu_{12} - \mu_{13} - \mu_{22} + \mu_{23} = 0$

Calcul des moyennes : Les énoncés pour estimer les moyennes sont

```
lsmeans temp / e stderr;
means   temp ;
```

The GLM Procedure
Least Squares Means

Coefficients for temp Least Square Means		
	temp Level	
Effect	15	70
Intercept	1	1
temp 15	1	0
temp 70	0	1
materiel 1	0.33333333	0.33333333
materiel 2	0.33333333	0.33333333
materiel 3	0.33333333	0.33333333
temp*materiel 15 1	0.33333333	0
temp*materiel 15 2	0.33333333	0
temp*materiel 15 3	0.33333333	0
temp*materiel 70 2	0	0.5
temp*materiel 70 3	0	0.5

temp	duree LSMEAN	Standard Error	Pr > |t|
15	145.111111	9.883399	<.0001
70	Non-est	.	.

Level of temp	N	duree	
		Mean	Std Dev
15	6	154.833333	30.9994624
70	5	130.200000	16.5438810

Il y a une cellule manquante à la modalité 2 de A ainsi la moyenne de A n'est pas estimable pour cette modalité. Avec l'énoncé `lsmeans`, SAS interprète cette moyenne comme une fonction des paramètres $L^T\beta$ où L n'appartient pas à l'ensemble des fonctions estimables (en effet le coefficient de β_1 (0.33) n'est pas égal à celui de $\tau\beta_{11}$ (0), tel que requis dans la forme générale des fonctions estimables (coefficient L4)). L'énoncé `means`, quant à lui, permet toujours de calculer des estimations des moyennes marginales, car il calcule des moyennes pondérées de toutes les observations par niveau.

Pour que les valeurs moyennes aux 2 modalités de A soient estimables, il faut simplifier le modèle. On considère maintenant le modèle sans interaction.

```
proc glm data=batterie3 ;
 class temp materiel ;
 model duree= temp materiel/ e e4 ss4;
 lsmeans temp / e stderr ;
 means   temp ;
 run;quit;
```

General Form of Estimable Functions	
Effect	Coefficients
Intercept	L1
temp 15	L2
temp 70	L1-L2
materiel 1	L4
materiel 2	L5
materiel 3	L1-L4-L5

Type IV Estimable Functions		
	Coefficients	
Effect	temp	materiel
Intercept	0	0
temp 15	L2	0
temp 70	-L2	0
materiel 1	0	L4
materiel 2	0	L5
materiel 3	0	-L4-L5

Énoncé `lsmeans`

The GLM Procedure
Least Squares Means

Coefficients for temp Least Square Means		
	temp Level	
Effect	15	70
Intercept	1	1
temp 15	1	0
temp 70	0	1
materiel 1	0.33333333	0.33333333
materiel 2	0.33333333	0.33333333
materiel 3	0.33333333	0.33333333

temp	duree LSMEAN	Standard Error	Pr > |t|
15	149.861111	11.435187	<.0001
70	139.861111	13.989079	<.0001

Énoncé `means`

Level of temp	N	duree	
		Mean	Std Dev
15	6	154.833333	30.9994624
70	5	130.200000	16.5438810

Les problèmes d'estimabilité disparaissent en enlevant l'interaction. Les moyennes aux deux modalités de A sont maintenant des fonctions estimables. L'estimation à la modalité 1 de A est un peu différente de celle obtenue avec le modèle qui inclut une interaction. C'est normal, les `lsmeans` dépendent du modèle sous-jacent. Le modèle sans interaction est plus simple. L'estimation de la moyenne à la modalité 2 de A est une fonction compliquée des moyennes des 5 traitements. Son erreur-type est un peu plus grande que celle pour la modalité 1.

Les moyennes pondérées $\bar{y}_{i\bullet\bullet}$ ne dépendent pas du modèle choisi. On obtient les mêmes valeurs avec les deux modèles.

Cet exemple démontre les limites du μ-modèle : il ne permet pas de spécifier facilement un modèle sans interaction lorsque certaines cellules du plan d'expérience sont vides. Le modèle avec effets est plus flexible ; il couvre un ensemble de modèles plus grand que le μ-modèle.

Discussion

Tous les tests effectués et les estimations calculées dans un modèle linéaire pour une expérience factorielle peuvent s'écrire en fonction des paramètres du modèle avec effets. Dans la plupart des analyses il n'est pas vraiment nécessaire de se référer à ces paramètres. Cependant, pour des plans complexes avec des cellules vides, ils sont la référence ultime pour bien comprendre le fonctionnement du modèle linéaire sous-jacent et les hypothèses testées.

3.7 Exercices

1. Vrai ou Faux ?

(a) Dans un plan d'expérience à 2 facteurs fixes avec interaction, la somme des carrés utilisée au numérateur de la statistique F pour tester l'interaction est la même pour les types I, II et III.

(b) Dans un plan équilibré, le nombre de paramètres dans le modèle avec effets est égal au nombre de combinaisons de traitements dans l'expérience.

(c) Pour un plan à deux facteurs, les sommes de carrés de type IV sont utilisées dans le cas où au moins un des $n_{ij} = 0$. Si $n_{ij} > 0$ pour tout (i, j), elles sont égales aux sommes de carrés de type III.

(d) Il est préférable d'utiliser les sommes de carrés de type I quand les données manquantes sont dues à des problèmes lors de la collecte des observations.

(e) On ne peut pas conduire de test sur l'interaction quand le plan d'expérience à deux facteurs fixes contient au moins une cellule vide.

(f) On ne peut pas conduire de test sur l'interaction quand le plan d'expérience à deux facteurs fixes contient une seule observation par cellule.

2. Dans l'exemple traité au début de ce chapitre, on donne les sommes de carrés de types I, II et III pour tester les effets des facteurs température, matériel et leur interaction.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	1	4720.333333	4720.333333	9.84	0.0201
materiel	2	555.791667	277.895833	0.58	0.5887
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

Source	DF	Type II SS	Mean Square	F Value	Pr > F
temp	1	5175.625000	5175.625000	10.79	0.0167
materiel	2	555.791667	277.895833	0.58	0.5887
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temp	1	5256.734848	5256.734848	10.96	0.0162
materiel	2	1934.308333	967.154167	2.02	0.2138
temp*materiel	2	9246.708333	4623.354167	9.64	0.0134

(a) Retrouvez ces neuf sommes de carrés en faisant la différence de deux termes d'erreur, i.e. de sommes de carrés associées à l'erreur de deux modèles différents. (Prenez pour acquis que $SSE(O, A, A*B) = 4\,811.48$ et que $SSE(O, B, A*B) = 8\,133.90$).

(b) Recalculez la statistique F de type III pour l'interaction, et écrivez précisément la probabilité ayant servi au calcul du seuil observé 0.0134.

(c) La statistique F de type I pour le facteur température (facteur A) est-elle la même que celle d'une analyse de variance à un facteur ? Discutez.

3. En utilisant la notation $SS(\bullet|\bullet)$, donner le tableau des sommes de carrés de types I, II et III utilisées pour tester les facteurs dans un plan à TROIS facteurs fixes (A à I niveaux, B à J niveaux et C à K niveaux) avec interactions doubles et triple. En d'autres termes, ajoutez une ligne au tableau suivant pour chaque facteur.

Facteur	SS type I	SS type II	SS type III

4. On veut comparer l'efficacité des divers modes de transport dans les villes de Montréal, Québec et Sherbrooke. On a mesuré le temps nécessaire (en minutes) pour se rendre au travail de plusieurs personnes choisies au hasard dans chaque ville, en tenant compte du mode de transport utilisé. On considère ici la portion des données correspondant aux gens qui habitent à une distance de 8 à 10 km de leur lieu de travail. Voici un résumé des observations :

	Montréal	Québec	Sherbrooke
Transport en commun	$\overline{y_{11\bullet}} = 25$ $n_{11} = 100$	$\overline{y_{12\bullet}} = 30$ $n_{12} = 20$	$\overline{y_{13\bullet}} = 29$ $n_{13} = 30$
Voiture	$\overline{y_{21\bullet}} = 28$ $n_{21} = 75$	$\overline{y_{22\bullet}} = 15$ $n_{22} = 75$	$\overline{y_{23\bullet}} = 10$ $n_{23} = 63$
Vélo	$\overline{y_{31\bullet}} = 28$ $n_{31} = 25$	$\overline{y_{32\bullet}} = 26$ $n_{32} = 5$	$\overline{y_{33\bullet}} = 29$ $n_{33} = 7$

En analysant ces données selon un plan complètement aléatoire à deux facteurs fixes avec interaction, on a obtenu une estimation de la variance des observations de $\hat{\sigma}^2 = 64$ minutes^2.

(a) Calculer l'estimation de la moyenne pondérée du temps de transport à Québec. Calculer l'erreur-type associée à cette estimation.

(b) Calculer l'estimation de la moyenne non pondérée du temps de transport à Québec. Calculer l'erreur-type associée à cette estimation.

(c) En vous inspirant de cet exemple, dites dans quelle condition il serait préférable de calculer des moyennes pondérées.

5. Vous trouverez sur le site web un fichier contenant les résultats partiels d'une expérience sur les stratégies de recherche de nourriture des fourmis au Sierra Nevada Aquatic Research Laboratory (source : UCLA Statistics Data sets). La sélection naturelle veut que les colonies optimisent leur gain net, i.e. maximisent l'entrée de nourriture en minimisant les pertes de travailleuses.

Les fourmis ont été sélectionnées au hasard dans 4 colonies (notées 5, 6, 7 et 8) à différentes distances du monticule d'entrée du nid (1 m, 4 m et 7 m). On a mesuré la masse en mg de plusieurs fourmis pour chaque combinaison de facteurs. On veut savoir si la masse moyenne des fourmis varie en fonction de la colonie et de la distance au nid.

On propose le modèle d'analyse à deux facteurs fixes suivant :

$$y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$$

τ_i : effet du niveau i de la colonie $(i = 1, 2, 3, 4)$
β_j : effet du niveau j de la distance $(j = 1, 2, 3)$
$\tau\beta_{ij}$: effet de l'interaction du niveau i de la colonie et du niveau j de la distance
$\varepsilon_{ijk} \sim N(0, \sigma^2)$ $i.i.d.$ $(k = 1, ..., n_{ij})$

(a) Donnez les dimensions des matrices Y, X, β et ε pour le modèle avec effets (surparamétrisé) ajusté à ces données. Donnez également les dimensions pour les matrices du modèle avec effets après réduction, i.e. après élimination des paramètres dépendants linéairement des autres.

(b) Importez les données dans SAS. Calculez les moyennes par cellule de traitement à l'aide de la procédure MEANS et tracez les courbes d'interaction. Avez-vous l'impression que l'interaction sera significative ? Sinon, les facteurs ont-ils l'air d'avoir un effet important ?

(c) **(i)** Calculez la moyenne des trois niveaux du facteur distance avec les énoncés *means* et *lsmeans* de PROC GLM.

(ii) Donnez les combinaisons linéaires des moyennes ayant mené aux 2 estimations obtenues en **(i)** pour la distance de 7 m. Autrement dit, refaire les calculs de SAS.

(d) En supposant que le facteur colonie est spécifié le premier dans le modèle, écrire la somme des carrés de type I servant de numérateur à la statistique F pour tester le facteur distance avec la notation de réduction de l'erreur ($SS(\bullet|\bullet)$). Cette $SS_I(Distance)$ peut être calculée en faisant la différence de 2 sommes de carrés. Lesquelles ? Quelles sont leurs valeurs ?

(e) Les fonctions suivantes sont-elles estimables ? Si oui, donnez leur estimation (la valeur de cette fonction pour cet exemple) et l'erreur-type qui y est associée (la formule en fonction du MSE et la valeur).

(i) μ

(ii) $\beta_2 - \beta_3$

(iii) $\tau_1 + \beta_1 + \tau\beta_{11} - (\tau_2 + \beta_2 + \tau\beta_{22})$

(iv) $\beta_1 + \beta_2 - 2\beta_3 + \tau\beta_{41} + \tau\beta_{42} - 2\tau\beta_{43}$

(v) $\mu + \tau_2$

(f) Analysez ces données selon le plan proposé. Interprétez clairement les conclusions. Faites la validation des postulats du modèle.

6. Un forestier planifie une expérience pour comparer l'exactitude d'appareils GPS. Les caractéristiques du paysage peuvent affecter la validité des mesures, il propose donc d'étudier deux facteurs fixes (sans interaction) : le type de GPS (3 types) et l'environnement (2 niveaux : dégagé et encombré). La variable réponse est une mesure de l'erreur commise à chaque utilisation de son appareil. Il collecte ses données selon un plan complètement aléatoire, et les modélise comme suit :

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} = \mu + \tau_i + \beta_j + \varepsilon_{ijk}$$

μ_{ij} : espérance de l'erreur commise avec le type i de GPS et l'environnement j
τ_i : effet du niveau i du type de GPS $(i = 1, 2, 3)$
β_j : effet du niveau j de l'environnement $(j = 1, 2)$
$\varepsilon_{ijk} \sim N(0, \sigma^2)$ $i.i.d.$ $(k = 1, ..., n_{ij})$

Voici les sorties SAS obtenues avec le programme suivant :

```
proc glm data=test;
class type envir;
model y= type envir/e solution inverse ;
run;quit;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	34.7722156	11.5907385	2.46	0.0851
Error	26	122.4539618	4.7097678		
Corrected Total	29	157.2261774			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
type	2	17.53042890	8.76521445	1.86	0.1756
envir	1	15.30882483	15.30882483	3.25	0.0830

X'X Generalized Inverse (g2)

	Intercept	type1	type2	type3	envir1	envir2	y
Intercept	0.1678	-0.1413	-0.1338	0	-0.0582	0	-2.2610
type I	-0.1413	0.2339	0.1434	0	-0.0035	0	1.9800
type II	-0.1338	0.1434	0.2294	0	-0.0210	0	1.5436
type III	0	0	0	0	0	0	0
envir 1	-0.0582	-0.0035	-0.0210	0	0.1359	0	1.4424
envir 2	0	0	0	0	0	0	0
y	-2.2609	1.9800	1.5436	0	1.4424	0	122.4540

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		-2.260939547 B	0.88903743	-2.54	0.0173
type	1	1.979953361 B	1.04948392	1.89	0.0704
type	2	1.543614608 B	1.03953598	1.48	0.1496
type	3	0.000000000 B	.	.	.
envir	1	1.442395610 B	0.80004278	1.80	0.0830
envir	2	0.000000000 B	.	.	.

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable

General Form of Estimable Functions

Effect		Coefficients
Intercept		L1
type	1	L2
type	2	L3
type	3	L1-L2-L3
envir	1	L5
envir	2	L1-L5

(a) En utilisant le μ-modèle, la différence entre la moyenne non pondérée du type I de GPS et celle du type II de GPS s'écrit $\frac{\mu_{11}+\mu_{12}}{2} - \frac{\mu_{21}+\mu_{22}}{2}$. Exprimer cette différence en fonction des paramètres du modèle avec effets.

(b) La quantité trouvée en (a) est-elle une fonction estimable ?

(c) Calculer une estimation de la quantité trouvée en (a), qu'elle soit unique ou non.

(d) Calculer l'erreur-type de l'estimation calculée en (c).

(e) Calculer un intervalle de confiance à 95% pour la quantité trouvée en (a).

7. On réalise une expérience complètement randomisée à deux facteurs fixes (A avec 3 modalités) et B (avec 4 modalités) comprenant $n = 5$ répétitions pour chaque traitement. Le modèle à l'étude est $y_{ijk} = \mu + \tau_i + \beta_j + \tau\beta_{ij} + \varepsilon_{ijk}$ avec $i = 1, 2, 3,\ j = 1, 2, 3, 4$ et $k = 1, ..., 5$. On veut réaliser une analyse préliminaire des résultats après avoir récolté les données pour 6 des 12 traitements à l'étude. Les résultats préliminaires (moyennes et variances des 5 mesures obtenues pour 6 traitements) sont :

A \ B	j=1	j=2	j=3	j=4
i=1	$\bar{y}_{ij}$=2, s_{ij}^2=1.1	ND	$\bar{y}_{ij}$=2.5, s_{ij}^2=1.3	$\bar{y}_{ij}$=3, s_{ij}^2=1.8
i=2	ND	$\bar{y}_{ij}$=4, s_{ij}^2=0.9	ND	ND
i=3	$\bar{y}_{ij}$=5.4, s_{ij}^2=1.2	ND	$\bar{y}_{ij}$=6.5, s_{ij}^2=1.5	ND

(a) Donner une estimation de la variance résiduelle σ^2. Combien de degrés de liberté y sont associés ?

(b) Donner un contraste d'interaction estimable et tester si ce dernier est nul.

(c) Donner un contraste estimable qui compare des modalités de A entre elles et tester s'il est nul.

(d) La moyenne de y à la modalité 1 de A est-elle estimable ?

4 Analyse de la covariance

4.1 Introduction et procédure

Qu'est-ce que l'analyse de la covariance, ou ANCOVA ?

- C'est un modèle linéaire dans lequel on retrouve des variables explicatives discrètes (appelées facteurs) et continues (appelées covariables). En ce sens, c'est un mélange de l'analyse de variance et de la régression. Nous considérerons le cas simple à un facteur et une covariable.

- On peut voir ce modèle comme une comparaison des droites de régression dans chaque niveau du facteur.

- On peut aussi voir ce modèle comme une comparaison de moyennes ajustées pour la covariable.

- Idéalement, la covariable ne devrait pas être directement influencée par le traitement.

- L'ajout d'une covariable dans un modèle d'anova permet aussi de réduire la composante de variabilité associée à l'erreur, et donc augmente la puissance des tests. Cela est utile quand l'hétérogénéité des unités expérimentales ne peut être réduite par blocage, parce que la covariable ne peut être contrôlée ou discrétisée.

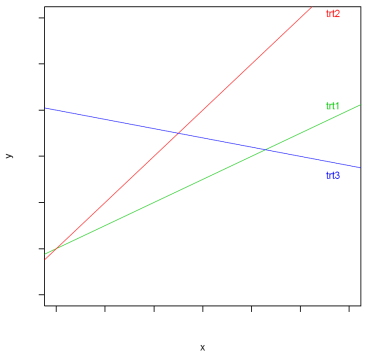
Exemples :

- Y : taux de cholestérol
  Facteurs : médicaments et régime alimentaire
  X : âge
- Y : résultat à un examen
  Facteur : méthode d'enseignement
  X : temps d'étude
- Y : volume de bois utilisable sur un arbre
  Facteurs : espèces, régions
  X : Diamètre à hauteur de poitrine (ou la taille de l'arbre)
- Y : rendement de blé
  Facteur : variété de blé

X : humidité du grain avant l'ensemencement

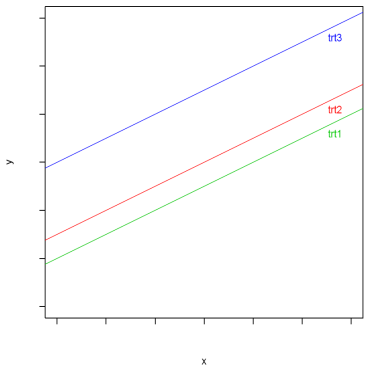
On ajustera un modèle avec une droite de régression par niveau du facteur, et on cherchera à simplifier ce modèle le plus possible en le comparant statistiquement à d'autres modèles plus parcimonieux.

Modèle 1 : une droite par traitement



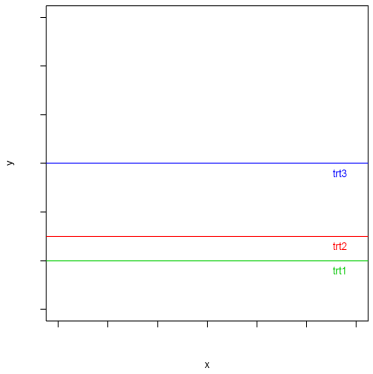
$$y_{ij} = \mu_i + \beta_i x_{ij} + \varepsilon_{ij} \quad \left\{ \begin{array}{l} i = 1, ..., a \\ j = 1, ..., n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{array} \right.$$

Modèle 2 : une droite par traitement, mais avec pente commune
(parfois le seul modèle présenté dans la littérature sur l'ANCOVA)



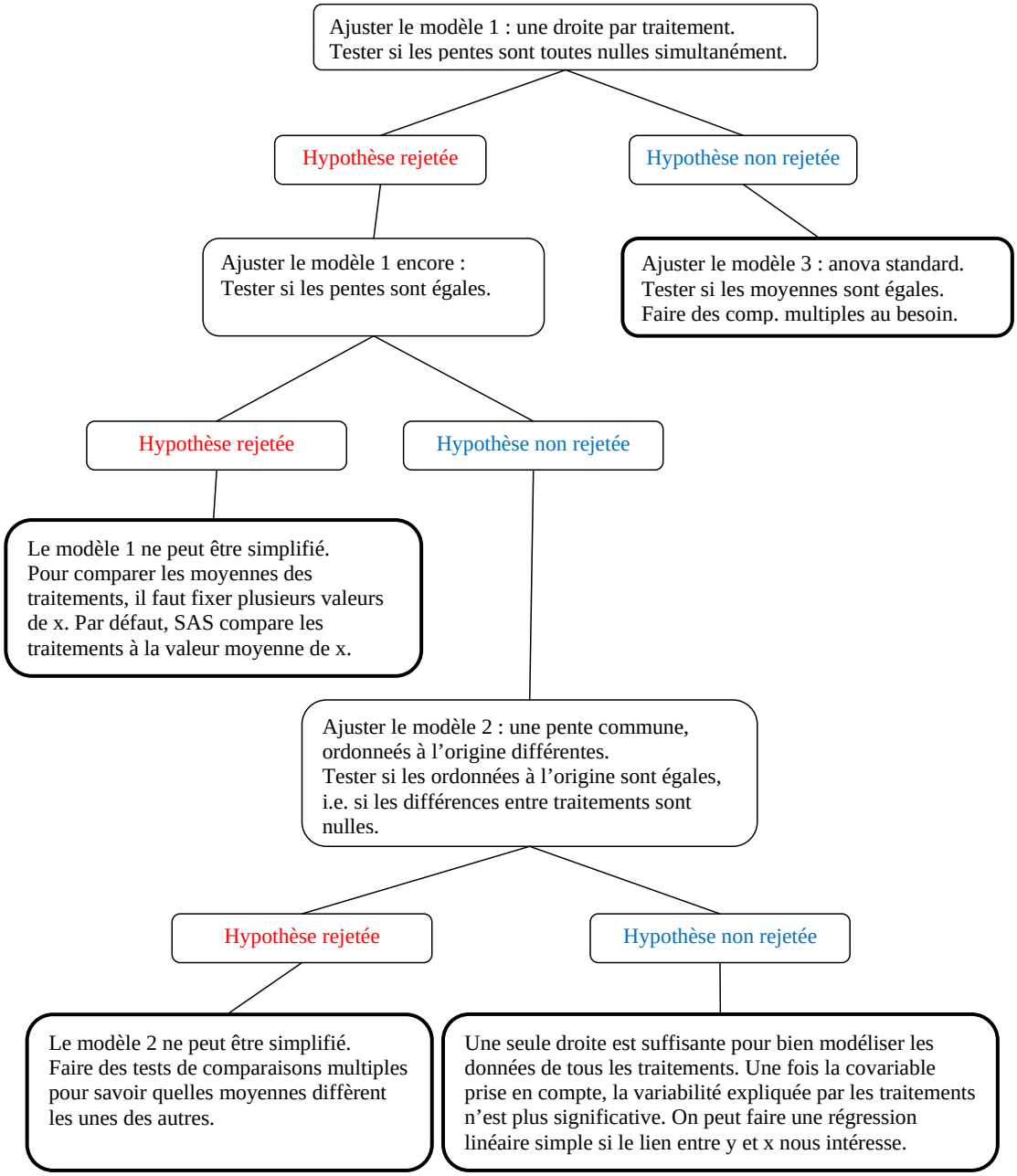
$$y_{ij} = \mu_i + \beta x_{ij} + \varepsilon_{ij} \quad \left\{ \begin{array}{l} i = 1, ..., a \\ j = 1, ..., n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{array} \right.$$

Modèle 3 : analyse de variance standard, sans covariable



$$y_{ij} = \mu_i + \varepsilon_{ij} \quad \left.\begin{array}{l} i = 1, ..., a \\ j = 1, ..., n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{array}\right.$$

Procédure pour l'analyse de la covariance dans un modèle à un facteur fixe.



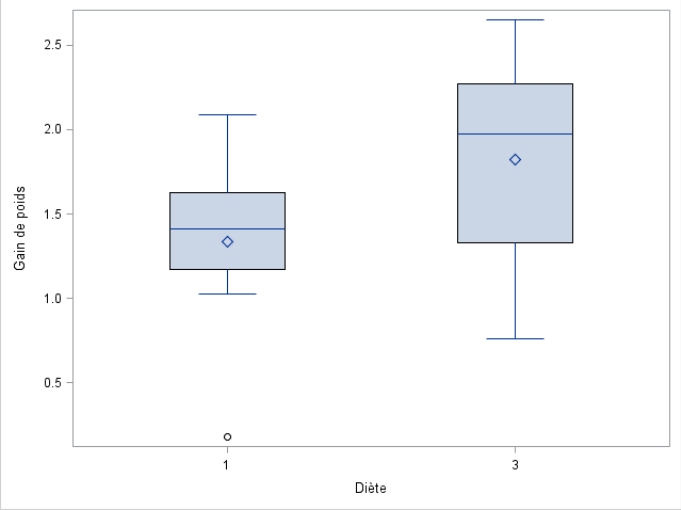
4.2 Exemple d'application

On veut comparer le gain de poids moyen quotidien de boeufs nourris pendant 160 jours selon deux diètes différentes. Deux groupes de 8 boeufs sont constitués (`trt`=1 et `trt`=3), et on mesure le poids initial des bêtes (variable `cov`) en plus de leur gain de poids (variable `rep`). Voici les observations et les moyennes de chaque groupe, extrait de l'exemple 5.3 de SAS System for Mixed Models ([10]).

Obs	trt	cov	rep
1	1	338	1.03
2	1	403	1.31
3	1	394	1.59
4	1	499	2.09
5	1	371	1.66
6	1	395	1.42
7	1	414	1.41
8	1	315	0.18
9	3	444	1.82
10	3	450	2.13
11	3	482	2.33
12	3	391	2.21
13	3	486	2.65
14	3	316	1.58
15	3	309	1.08
16	3	308	0.76

N Obs	Variable	Label	Mean	Std Dev
16	rep	Gain de poids	1.5781250	0.6347306
	cov	Poids initial	394.6875000	65.3010656

Diète	N Obs	Variable	Label	Mean	Std Dev
1	8	rep	Gain de poids	1.3362500	0.5582098
		cov	Poids initial	391.1250000	55.2227632
3	8	rep	Gain de poids	1.8200000	0.6465292
		cov	Poids initial	398.2500000	77.8400374



Le but de l'analyse est de vérifier l'impact du traitement sur le gain de poids moyen des animaux. On pourrait donc utiliser l'analyse de variance à un facteur fixe. Par contre, puisqu'on dispose d'information sur le poids initial des bêtes, il serait intéressant de l'utiliser.

S'il existe un lien entre le poids initial et le gain de poids moyen, l'ajout du poids initial dans le modèle fera diminuer l'erreur résiduelle pour la comparaison des moyennes et augmentera donc la précision des tests.

On suppose une relation linéaire entre cette covariable et la variable réponse, que l'on pourra vérifier à l'aide de graphiques et de tests d'hypothèses. On voudra vérifier si cette relation est la même pour tous les traitements. Nous commencerons par ajuster un modèle où les droites sont distinctes pour les deux traitements. Nous réduirons ce modèle autant que possible par la suite.

Remarques :

1. Si les groupes ont tous la même valeur moyenne pour la covariable ($\overline{x_{1\bullet}} \approx \overline{x_{2\bullet}} \approx ... \approx \overline{x_{I\bullet}}$), alors l'ajout d'une covariable dans le modèle n'aura pas d'effet sur la comparaison des moyennes de Y.

2. La covariable ne devrait pas être influencée par le traitement, i.e. que le facteur principal ne devrait pas avoir d'effet sur sa valeur. Autrement, on a confusion entre le traitement et la covariable, et le traitement expliquerait la même portion de variabilité de Y que la covariable, ce qui fausserait la significativité des tests.

3. Malgré la remarque précédente, on ne devrait pas vérifier statistiquement l'hypothèse d'absence du facteur sur X (par exemple par une anova sur la variable X), car l'ANCOVA est utile quand les traitements ont des moyennes de X différentes accidentellement.

4.3 Modèle 1 : une droite par traitement

$$\left.\begin{array}{rcccccll} y_{ij} & = & \mu_i & + & \beta_i x_{ij} & + & \varepsilon_{ij} & (\mu - \text{modèle}) \\ & = & \mu + \tau_i & + & (\beta + \gamma_i) x_{ij} & + & \varepsilon_{ij} & (\text{modèle avec effets}) \end{array}\right\} \begin{array}{l} i = 1, ..., I \\ j = 1, ..., n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{array}$$

Estimation des paramètres :

L'estimation des paramètres est plus intuitive pour le μ-modèle, car il s'agit d'une droite de régression dans chaque traitement paramétrisée avec uen pente et une ordonnée à l'origine. Pour un i fixé, les estimateurs de β_i et μ_i sont ceux (sans biais) de la régression linéaire simple :

$$\hat{\beta}_i = \frac{\sum_{j=1}^{n_i} (x_{ij} - \overline{x_{i\bullet}})(y_{ij} - \overline{y_{i\bullet}})}{\sum_{j=1}^{n_i} (x_{ij} - \overline{x_{i\bullet}})^2}$$

$$\hat{\mu}_i = \overline{y_{i\bullet}} - \hat{\beta}_i \overline{x_{i\bullet}}$$

Le modèle avec effets étant surparamétrisé, l'estimation des paramètres dépend de la contrainte utilisée pour résoudre les équations normales.

Première hypothèse à tester sur la covariable : Nullité de toutes les pentes

$$\begin{array}{rcl} H_0 & : & \beta_1 = ... = \beta_I = 0 \\ H_1 & : & \text{au moins une des pentes } \neq 0 \end{array}$$

- On compare en fait les modèles 1 (une droite par traitement) et 3 (absence de covariable). Par le test général de la régression, la statistique du test est la suivante :

$$F_0 = \frac{(SSE_3 - SSE_1)/I}{SSE_1/N - 2I} \sim F_{I,N-2I} \text{ si } H_0 \text{ est vraie.}$$

- Si H_0 est rejetée, on teste la prochaine hypothèse, celle de l'égalité des pentes. Sinon, on ajuste le modèle 3, i.e. l'anova à un facteur standard.

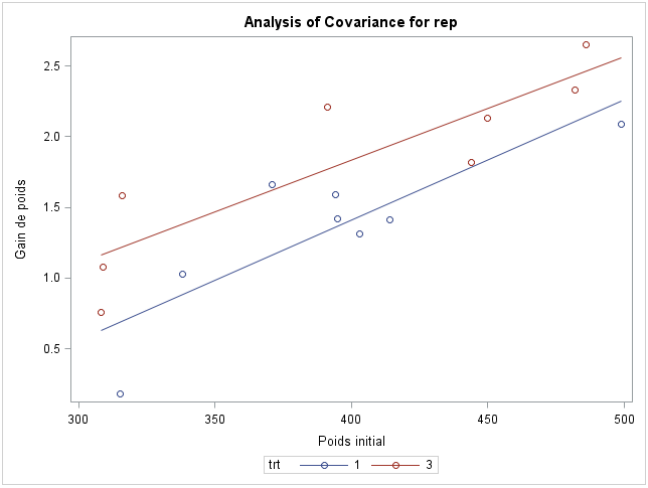
Exemple : Utilisons le μ-modèle pour tester la nullité simultanée des pentes. L'option `noint` nous assure qu'aucune ordonnée à l'origine de référence n'est paramétrisée, et l'absence de la variable `cov` seule empêche la création d'une pente de référence.

```
proc glm data=bouffe;
 class trt ;
 model rep = trt  cov*trt / noint solution;
 run;quit;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	2	0.85186967	0.42593483	3.95	0.0482
cov*trt	2	3.81228106	1.90614053	17.66	0.0003

Parameter	Estimate	Standard Error	t Value	Pr > |t|
trt 1	-1.989509179	0.88701705	-2.24	0.0446
trt 3	-1.092778537	0.64576170	-1.69	0.1164
cov*trt 1	0.008503060	0.00224834	3.78	0.0026
cov*trt 3	0.007313945	0.00159506	4.59	0.0006

Le 2^e test de la table d'anova indique que les pentes ne sont pas toutes nulles simultanément, et donc que la présence de la covariable est justifiée.



Les équations des droites du modèle 1 sont estimées par :

$$\hat{y}_1 = -1.99 + 0.0085\, x \qquad \text{et} \qquad \hat{y}_2 = -1.09 + 0.0073\, x$$

Deuxième hypothèse à tester sur la covariable (si la première est rejetée) : Les pentes sont-elles toutes égales ?

Rappel du modèle 1 :

$$\left.\begin{array}{rcccccl} y_{ij} & = & \mu_i & + & \beta_i x_{ij} & + & \varepsilon_{ij} \quad (\mu - \text{modèle}) \\ & = & \mu + \tau_i & + & (\beta + \gamma_i)x_{ij} & + & \varepsilon_{ij} \quad (\text{modèle avec effets}) \end{array}\right\} \begin{array}{l} i = 1, ..., I \\ j = 1, ..., n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{array}$$

$$\left\{\begin{array}{lcl} H_0 & : & \beta_1 = ... = \beta_I \\ H_1 & : & \exists (i,j) \text{ t.q. } \beta_i \neq \beta_j \end{array}\right\} \text{ ou } \left\{\begin{array}{lcl} H_0 & : & \gamma_1 = ... = \gamma_I = 0 \\ H_1 & : & \text{au moins un des } \gamma_i \neq 0 \end{array}\right\}$$

- On compare ici les modèles 1 (une droite par traitement) et 2 (pentes égales). La statistique du test est la suivante :

$$F_0 = \frac{(SSE_2 - SSE_1)/I - 1}{SSE_1/N - 2I} \sim F_{I-1,N-2I} \text{ si } H_0 \text{ est vraie.}$$

- Si H_0 est rejetée, le modèle 1 ne peut pas être simplifié. Voir l'hypothèse sur les traitements et la comparaison des moyennes ajustées ci-dessous. Si H_0 n'est pas rejetée, on ajuste le modèle 2.

Exemple :

Nous utilisons ici le modèle avec effets, pour vérifier que les différences entre les pentes sont nulles (i.e. que les pentes sont toutes égales). L'ordonnée à l'origine de référence (μ) est incluse par défaut, et l'ajout de cov dans le modèle introduit une pente de référence.

```
proc glm data=bouffe;
 class trt ;
 model rep = trt cov cov*trt / solution;
 run; quit;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	1	0.07208155	0.07208155	0.67	0.4297
cov	1	3.55251589	3.55251589	32.92	<.0001
cov*trt	1	0.02007868	0.02007868	0.19	0.6739

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1.092778537	B	0.64576170	-1.69	0.1164
trt 1	-0.896730641	B	1.09718159	-0.82	0.4297
trt 3	0.000000000	B	.	.	.
cov	0.007313945	B	0.00159506	4.59	0.0006
cov*trt 1	0.001189115	B	0.00275667	0.43	0.6739
cov*trt 3	0.000000000	B	.	.	.

Le test sur `cov*trt` indique que les pentes ne diffèrent pas de façon significative. On pourra donc passer au modèle 2, plus simple.

Les équations des droites du modèle 1 sont estimées par les mêmes valeurs qu'avec le μ-modèle, mais avec une droite de référence (celle de la diète 3) :

$$\hat{y}_1 = (-1.093 - 0.897) + (0.0073 + 0.0012)\, x = -1.99 + 0.0085\, x$$

et

$$\hat{y}_2 = -1.09 + 0.0073\, x$$

Hypothèse sur les traitements (dans le cas où les pentes sont différentes) :

Si le modèle 1 est retenu, la différence entre les valeurs prédites de Y d'un traitement à l'autre ne sera pas la même pour toutes les valeurs de X. Le test global sur le traitement fourni par SAS vérifie l'égalité des ordonnées à l'origine, i.e. la valeur moyenne de Y lorsque $X = 0$.

$$\left.\begin{array}{ll} H_0: & \mu_1 = ... = \mu_I \\ H_1: & \exists (i,j) \text{ t.q. } \mu_i \neq \mu_j \end{array}\right\} \text{ ou } \left\{\begin{array}{ll} H_0: & \tau_1 = ... = \tau_I = 0 \\ H_1: & \text{au moins un des } \tau_i \neq 0 \end{array}\right.$$

Remarques :

- Si H_0 est rejetée, c'est que les valeurs des ordonnées à l'origine sont différentes pour au moins deux traitements. Si cela est pertinent pour le chercheur, on peut procéder à des comparaisons multiples au point $X = 0$ (lorsque le facteur a plus de 2 modalités). Si H_0 n'est pas rejetée, c'est que les valeurs moyennes des ordonnées à l'origine ne diffèrent pas significativement.

- Pour faire un test global de comparaison des moyennes de Y au point $X = x^*$, il faut transformer la variable X en $X - x^*$. Le test comparera les valeurs de Y au point $X - x^* = 0$.

Exemple :

Dans notre exemple, lorsqu'on applique le modèle 1, le seuil observé est de 0.4297 pour le test qui compare le gain de poids moyen avec les deux diètes lorsque $X = 0$, i.e. pour un poids initial de 0 kg. Ce n'est pas significatif, mais on voit que l'interprétation de ce test n'est pas pertinente dans le contexte de ces données.

Il pourrait être pertinent de comparer les valeurs de Y pour un poids initial donné, comme 350 kg ou 400 kg. Il faudrait alors se créer une nouvelle covariable $X2 = X - 350$ ou $X3 = X - 400$ et réajuster le modèle 1. Rappelons toutefois que dans notre exemple le modèle 1 peut être simplifié, car les pentes ne diffèrent pas significativement.

Comparaison de moyennes ajustées (avec des droites de pentes différentes) :

Dans le cas du modèle 1, la différence estimée entre deux moyennes de Y ne sera pas la même pour toutes les valeurs de X. Il faut donc fixer plusieurs valeurs de X et comparer les traitements entre eux. Dans ce modèle, l'espérance de la moyenne échantillonnale du traitement i se calcule comme suit :

$$E(\overline{y_{i\bullet}}) = E\left[\frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}\right] = E\left[\frac{\sum_{j=1}^{n_i}(\mu_i + \beta_i x_{ij} + \varepsilon_{ij})}{n_i}\right]$$

$$= \frac{E(n_i\mu_i) + E(\beta_i \sum_{j=1}^{n_i} x_{ij}) + \sum_{j=1}^{n_i} E(\varepsilon_{ij})}{n_i}$$

$$= \mu_i + \beta_i \overline{x_{i\bullet}}$$

La valeur $\overline{y_{1\bullet}}$ correspond à une covariable égale à $\overline{x_{1\bullet}}$;
La valeur $\overline{y_{2\bullet}}$ correspond à une covariable égale à $\overline{x_{2\bullet}}$;
et ainsi de suite.

On veut comparer les valeurs moyennes de Y *ajustées pour* X, i.e. pour une même valeur de X, disons x^*. Les moyennes ajustées correspondent à la valeur de Y sur chaque droite correspondant à $X = x^*$, et sont calculées comme suit :

$$\overline{y_{i\bullet}}\big|_{X=x^*} = \overline{y_{i\bullet}} - \hat{\beta}_i(\overline{x_{i\bullet}} - x^*)$$

On peut montrer que l'espérance et la variance d'une moyenne ajustée sont :

$$E(\overline{y_{i\bullet}}\big|_{X=x^*}) = \mu_i + \beta_i x^* \qquad V(\overline{y_{i\bullet}}\big|_{X=x^*}) = \sigma^2\left[\frac{1}{n_i} + \frac{(\overline{x_{i\bullet}} - x^*)^2}{\sum_{j=1}^{n_i}(x_{ij} - \overline{x_{i\bullet}})^2}\right]$$

On remarque que la variance est minimale lorsque $x^* = \overline{x_{i\bullet}}$.

Puisque les moyennes ajustées sont des combinaisons linéaires des observations, elles suivent aussi une loi normale, et on en déduit l'intervalle de confiance de la différence entre deux moyennes ajustées :

$$\left(\overline{y_{i\bullet}}|_{X=x^*} - \overline{y_{k\bullet}}|_{X=x^*}\right) \pm t_{\alpha/2,N-2I}\sqrt{MSE\left[\frac{1}{n_i}+\frac{1}{n_k}+\frac{(\overline{x_{i\bullet}}-x^*)^2}{\sum_{j=1}^{n_i}(x_{ij}-\overline{x_{i\bullet}})^2}+\frac{(\overline{x_{k\bullet}}-x^*)^2}{\sum_{j=1}^{n_k}(x_{kj}-\overline{x_{k\bullet}})^2}\right]}$$

L'erreur-type sera minimale au point-milieu de $\overline{x_{i\bullet}}$ et $\overline{x_{k\bullet}}$.

Exemple :
Les énoncés `lsmeans`, `estimate` et `contrast` permettent l'évaluation de moyennes ajustées en précisant la valeur de la covariable. Dans `lsmeans`, l'option `at` permet de spécifier la valeur de X où calculer les moyennes et l'option `cl` fait afficher les limites de confiance. Pour obtenir ces limites avec l'énoncé `estimate`, il faut ajouter l'option `clparm` dans l'énoncé `model`.

Voici quelques exemples, où vous pourrez vérifier l'application des formules ci-dessus pour une moyenne ou une différence de moyennes ajustées.

```
proc glm data=bouffe;
 class trt ;
 model rep = trt cov cov*trt/ clparm ;
 lsmeans  trt / stderr at cov = 330;
 estimate "moy aj. trt1"      intercept 1 trt 1  0  cov 330 cov*trt 330    0;
 estimate "moy aj. trt3"      intercept 1 trt 0  1  cov 330 cov*trt   0  330;
 estimate "moy aj. trt1-trt3" intercept 0 trt 1 -1  cov   0 cov*trt 330 -330;
 run;quit;
```

Rappel :

Diète	N Obs	Variable	Label	Mean	Std Dev
1	8	rep	Gain de poids	1.3362500	0.5582098
		cov	Poids initial	391.1250000	55.2227632
3	8	rep	Gain de poids	1.8200000	0.6465292
		cov	Poids initial	398.2500000	77.8400374

Mentionnons que le $MSE = 0.1079$.

Least Squares Means at cov=330

trt	rep LSMEAN	Standard Error	Pr > |t|
1	0.81650048	0.17993194	0.0007
3	1.32082326	0.15918448	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
moy aj. trt1	0.81650048	0.17993194	4.54	0.0007	0.42446246	1.20853851
moy aj. trt3	1.32082326	0.15918448	8.30	<.0001	0.97399007	1.66765646
moy aj. trt1-trt3	-0.50432278	0.24023989	-2.10	0.0576	-1.02776053	0.01911497

Estimation du gain de poids moyen d'un animal pesant 330 kg et ayant suivi la diète 1, et son erreur-type :

$$\overline{y_{1\bullet}}\big|_{X=330} = 1.336 - 0.0085(391.125 - 330) = 0.8165$$

$$\text{err.} - \text{type} = \sqrt{0.1079\left[\frac{1}{8} + \frac{(391.125 - 330)^2}{21\,346.9}\right]} = 0.1799$$

Erreur-type de la différence de gain de poids moyen d'animaux pesant 330 kg et ayant suivi la diète 1 et la diète 3 respectivement :

$$\sqrt{0.1079\left[\frac{1}{8} + \frac{1}{8} + \frac{(391.125 - 330)^2}{21\,346.9} + \frac{(398.25 - 330)^2}{42\,413.5}\right]} = 0.2402$$

4.4 Modèle 2 : une droite par traitement, mais avec pente commune

$$\left.\begin{array}{rcccccl} y_{ij} & = & \mu_i & + & \beta x_{ij} & + & \varepsilon_{ij} \quad (\mu - \text{modèle}) \\ & = & \mu + \tau_i & + & \beta x_{ij} & + & \varepsilon_{ij} \quad (\text{modèle avec effets}) \end{array}\right\} \begin{array}{l} i = 1, ..., I \\ j = 1, ..., n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{array}$$

Estimation des paramètres :

$$\hat{\beta} = \frac{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_{i\bullet}})(y_{ij} - \overline{y_{i\bullet}})}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_{i\bullet}})^2}$$

$$\hat{\mu}_i = \overline{y_{i\bullet}} - \hat{\beta}\, \overline{x_{i\bullet}}$$

L'estimateur de β est le même pour tous les traitements, puisque la pente est la même. De plus, il reste le même pour le modèle avec effets et le μ-modèle. Pour la partie représentant le traitement, le μ-modèle donne lieu à des estimateurs uniques qui représentent les ordonnées à l'origine des droites dans chaque groupe. Le modèle avec effets étant surparamétrisé, l'estimation des paramètres dépend de la contrainte utilisée pour résoudre les équations normales.

Hypothèse à tester sur la covariable : la pente unique est-elle nulle ?

$$\begin{array}{rcl} H_0 & : & \beta = 0 \\ H_1 & : & \beta \neq 0 \end{array}$$

- On compare ici les modèles 2 et 3. La statistique du test est la suivante :

$$F_0 = \frac{(SSE_3 - SSE_2)/1}{SSE_2/N - I - 1} \sim F_{1,N-I-1} \text{ si } H_0 \text{ est vraie.}$$

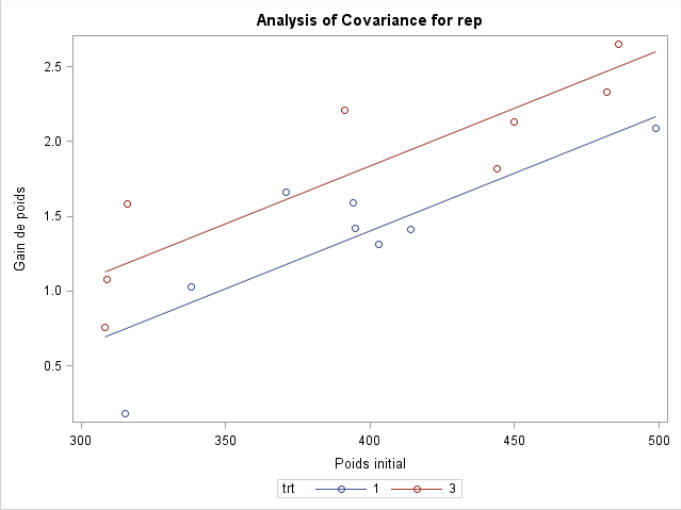
- Si H_0 est rejetée, le modèle 2 ne peut pas être simplifié. Voir le test sur le traitement et la comparaison des moyennes plus loin. Si H_0 n'est pas rejetée, on ajuste le modèle 3, i.e. une anova standard sans covariable. Ceci devrait arriver assez rarement, car la procédure dicte de faire un test sur la nullité de toutes les pentes d'entrée de jeu. Si cette hypothèse est rejetée au début de l'analyse, il y a peu de chance que la pente unique soit déclarée non significative ici.

Exemple :

```
proc glm data=bouffe;
 class trt ;
 model rep = trt cov /solution;
 run;quit;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	1	0.73314828	0.73314828	7.25	0.0185
cov	1	3.79220239	3.79220239	37.49	<.0001

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-1.251327348	B	0.51406230	-2.43	0.0301
trt 1	-0.428801582	B	0.15927559	-2.69	0.0185
trt 3	0.000000000	B	.	.	.
cov	0.007712059		0.00125954	6.12	<.0001



Le test sur la covariable est significatif, donc la pente commune n'est pas nulle et on conserve le modèle 2, tel qu'attendu.

Les équations des droites du modèle 2 (avec effets) sont estimées avec une ordonnée à l'origine de référence (celle de la diète 3), et une pente commune :

$\hat{y}_1 = (-1.251 - 0.429) + 0.0077\, x = -1.68 + 0.0077\, x$ et

$\hat{y}_2 = -1.251 + 0.0077\, x$

Hypothèse sur les traitements (lorsque la pente unique est non nulle) :

Puisque dans le modèle 2 les pentes sont les mêmes pour tous les traitements, la différence entre les valeurs prédites de Y d'un traitement à l'autre est la même pour toutes les valeurs de X. On fait le test sur les ordonnées à l'origine, mais il est valide pour toutes les valeurs de X.

$$\left.\begin{array}{lcl} H_0 & : & \mu_1 = ... = \mu_I \\ H_1 & : & \exists (i,j) \text{ t.q. } \mu_i \neq \mu_j \end{array}\right\} \text{ ou } \left\{\begin{array}{lcl} H_0 & : & \tau_1 = ... = \tau_I = 0 \\ H_1 & : & \text{ au moins un des } \tau_i \neq 0 \end{array}\right.$$

Si H_0 est rejetée, c'est que les valeurs des ordonnées à l'origine sont différentes pour au moins deux traitements. Si cela est pertinent, on peut procéder à des comparaisons multiples.

Exemple :

Le seuil observé du test sur l'égalité des ordonnées à l'origine est de 0.0185. On considère que les moyennes de Y pour une même valeur de X sont différentes. Ainsi, pour deux boeufs ayant un même poids initial, le choix de la diète 1 occasionne un gain de poids inférieur à la diète 3 en moyenne.

Comparaison de moyennes ajustées (avec des droites parallèles) :

Dans le modèle 2, la différence entre deux moyennes est la même pour toutes les valeurs de X. Il faut tout de même comparer des moyennes ajustées, pour comparer les traitements entre eux à une même valeur de X. Dans ce modèle, l'espérance de la moyenne échantillonnale du traitement i est :

$$E(\overline{y_{i\bullet}}) \quad = \mu_i + \beta\, \overline{x_{i\bullet}}$$

La valeur $\overline{y_{1\bullet}}$ correspond à une covariable égale à $\overline{x_{1\bullet}}$;
La valeur $\overline{y_{2\bullet}}$ correspond à une covariable égale à $\overline{x_{2\bullet}}$;
et ainsi de suite.

On veut comparer les valeurs moyennes de Y *ajustées pour* X, i.e. pour une même valeur de X, disons x^*. Les moyennes ajustées correspondent à la valeur de Y sur chaque droite correspondant à $X = x^*$, et sont calculées comme suit :

$$\overline{y_{i\bullet}}\big|_{X=x^*} = \overline{y_{i\bullet}} - \hat{\beta}\,(\overline{x_{i\bullet}} - x^*)$$

On peut montrer que l'espérance et la variance d'une moyenne ajustée sont :

$$E(\overline{y_{i\bullet}}|_{X=x^*}) = \mu_i + \beta x^* \qquad V(\overline{y_{i\bullet}}|_{X=x^*}) = \sigma^2 \left[\frac{1}{n_i} + \frac{(\overline{x_{i\bullet}} - x^*)^2}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_{i\bullet}})^2} \right]$$

Puisque les moyennes ajustées sont des combinaisons linéaires des observations, elles suivent aussi une loi normale, et on en déduit l'intervalle de confiance de la différence entre deux moyennes ajustées pour le modèle 2 :

$$\left(\overline{y_{i\bullet}}|_{X=x^*} - \overline{y_{k\bullet}}|_{X=x^*} \right) \pm t_{\alpha/2,N-I-1} \sqrt{MSE \left[\frac{1}{n_i} + \frac{1}{n_k} + \frac{(\overline{x_{i\bullet}} - \overline{x_{k\bullet}})^2}{\sum_{i=1}^{I} \sum_{j=1}^{n_i} (x_{ij} - \overline{x_{i\bullet}})^2} \right]}$$

Remarquons que l'erreur-type de la différence de deux moyennes ajustées ne dépend pas de la valeur de x^*.

Exemple :

```
proc glm data=bouffe;
 class trt ;
 model rep = trt cov / solution;
 lsmeans trt / stderr pdiff; *(qui correspond à la valeur moyenne de la covariable);
 lsmeans trt / stderr pdiff at cov=330;
 lsmeans trt / stderr pdiff at cov=460;
 estimate "moy.aj. trt1,cov=330"    intercept 1 trt 1  0 cov 330;
 estimate "moy.aj. trt1-trt3"       intercept 0 trt 1 -1        ;
 run;quit;
```

			H0:LSMEAN=0	H0:LSMean1=LSMean2
trt	rep LSMEAN	Standard Error	Pr > |t|	Pr > |t|
1	1.36372421	0.11253543	<.0001	0.0185
3	1.79252579	0.11253543	<.0001	

Lorsqu'aucune valeur de la covariable n'est précisée, l'estimation se fait à la valeur moyenne de X, dans notre cas à $X = 394.69$.

Least Squares Means at cov=330

trt	rep LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1=LSMean2 Pr > |t|
1	0.86485042	0.13627725	<.0001	0.0185
3	1.29365200	0.14154107	<.0001	

Least Squares Means at cov=460

trt	rep LSMEAN	Standard Error	H0:LSMEAN=0 Pr > |t|	H0:LSMean1=LSMean2 Pr > |t|
1	1.86741804	0.14202056	<.0001	0.0185
3	2.29621962	0.13672353	<.0001	

Parameter	Estimate	Standard Error	t Value	Pr > |t|
moy.aj. trt1,cov=330	0.86485042	0.13627725	6.35	<.0001
moy.aj. trt1-trt3	-0.42880158	0.15927559	-2.69	0.0185

Rappel :

Diète	N Obs	Variable	Label	Mean	Std Dev
1	8	rep	Gain de poids	1.3362500	0.5582098
		cov	Poids initial	391.1250000	55.2227632
3	8	rep	Gain de poids	1.8200000	0.6465292
		cov	Poids initial	398.2500000	77.8400374

Mentionnons que le $MSE = 0.1012$.

Estimation du gain de poids moyen d'un animal pesant 330 kg et ayant suivi la diète 1, et son erreur-type :

$$\overline{y_{1\bullet}}\big|_{X=330} = 1.336 - 0.0077(391.125 - 330) = 0.865$$

$$\text{err.} - \text{type} = \sqrt{0.1012\left[\frac{1}{8} + \frac{(391.125 - 330)^2}{63\,760.4}\right]} = 0.1363$$

Erreur-type de la différence de gain de poids moyen d'animaux pesant 330 kg et ayant suivi la diète 1 et la diète 3 respectivement :

$$\sqrt{0.1012\left[\frac{1}{8} + \frac{1}{8} + \frac{(391.125 - 398.25)^2}{63\,760.4}\right]} = 0.1593$$

4.5 Modèle 3 : analyse de variance standard, sans covariable

$$\left.\begin{array}{rcccll} y_{ij} & = & \mu_i & + & \varepsilon_{ij} & (\mu - \text{modèle}) \\ & = & \mu + \tau_i & + & \varepsilon_{ij} & (\text{modèle avec effets}) \end{array}\right\} \begin{array}{l} i = 1, ..., I \\ j = 1, ..., n_i \\ \varepsilon_{ij} \sim N(0, \sigma^2) \end{array}$$

Estimation des paramètres :

$$\hat{\mu}_i \quad = \overline{y_{i\bullet}}$$

Le μ-modèle donne lieu à des estimateurs uniques. Le modèle avec effets étant surparamétrisé, l'estimation des paramètres dépend de la contrainte utilisée pour résoudre les équations normales.

Hypothèse sur les traitements :

$$\left\{\begin{array}{lcl} H_0 & : & \mu_1 = ... = \mu_I \\ H_1 & : & \exists (i,j) \text{ t.q. } \mu_i \neq \mu_j \end{array}\right\} \text{ ou } \left\{\begin{array}{lcl} H_0 & : & \tau_1 = ... = \tau_I = 0 \\ H_1 & : & \text{au moins un des } \tau_i \neq 0 \end{array}\right.$$

On conduit un test F de la façon habituelle.

Comparaison de moyennes :

Si le test global F s'avère significatif, c'est qu'il existe au moins deux moyennes qui diffèrent de façon significative au seuil spécifié. Il est donc justifié de procéder à des comparaisons multiples pour voir où se situent les différences importantes.

Exemple :

Nous avons vu précédemment que la covariable est significative dans notre modèle 2. Il n'est donc pas approprié d'utiliser seulement l'analyse de variance standard. Or, nous allons ajuster l'anova simplement pour voir ce qui se serait produit si nous avions omis de mesurer la covariable dans la planification de l'expérience.

```
proc glm data=bouffe ;
 class trt;
 model rep=trt;
 run;quit;
```

Source	DF	Type III SS	Mean Square	F Value	Pr > F
trt	1	0.93605625	0.93605625	2.57	0.1315

Remarques :

- La différence entre les deux diètes n'est pas significative lorsqu'on ne considère pas le poids initial des bêtes. Elle l'était lorsqu'on considérait la présence de la covariable.
- Pourtant, la différence entre les poids moyens n'était pas si grande : $\overline{x}_{1\bullet} = 391.125$ kg et $\overline{x}_{3\bullet} = 398.25$ kg. On voit encore l'importance de ne pas se fier aux estimations ponctuelles uniquement.
- De plus, les moyennes non ajustées diffèrent encore plus que les moyennes ajustées, malgré que leur différence soit non significative.

  Cela signifie que la covariable a réduit de beaucoup l'erreur résiduelle afin de rendre le test significatif, et que la pente était importante.

4.6 Exercices

1. Vrai ou Faux ?

(a) On considère une analyse de la covariance avec une droite dans chacun des deux traitements, où les deux pentes sont égales. La différence entre deux moyennes ajustées à $X = x^*$ est toujours inférieure à la différence entre les deux moyennes non ajustées.

(b) Un modèle qui inclut une covariable aura une somme des carrés associée à l'erreur inférieure au même modèle sans la covariable.

(c) L'estimation ponctuelle de la moyenne de Y au point $X = x^*$ est la même avec le modèle 1 d'ancova qu'avec une régression linéaire simple ajustée séparément dans chaque traitement.

(d) L'erreur-type de la moyenne de Y au point $X = x^*$ est la même avec le modèle 1 d'ancova qu'avec une régression linéaire simple ajustée séparément dans chaque traitement.

(e) Avec le modèle 1, la précision de l'estimation de la moyenne de Y au point $X = x^*$ est maximale lorsque $x^* = \overline{x_{\bullet\bullet}}$.

(f) Dans SAS, en ajustant le modèle 1, on peut faire un test global de comparaison des I moyennes ajustées au point $X = x^*$ en utilisant l'option `at cov=x*` dans l'énoncé `model` de PROC GLM.

2. On considère le modèle 2 d'ancova, où les pentes entre Y et la covariable sont égales dans tous les traitements.

(a) Montrez que la variance d'une moyenne ajustée de Y au point $X = x^*$ dépend de x^*.

(b) Montrez que la variance de la différence entre deux moyennes ajustées de Y au point $X = x^*$, disons $\overline{y_{1\bullet}}|_{X=x^*} - \overline{y_{2\bullet}}|_{X=x^*}$, ne dépend pas de x^*.

3. Le modèle 2 qui s'ajuste le mieux aux données utilisées en exemple dans ce chapitre admet certains postulats nécessaires pour l'inférence. Nommez ces postulats, et vérifiez s'ils semblent respectés dans cet exemple.

4. On considère le modèle d'analyse de covariance avec effets

$$y_{ij} = \mu + \tau_i + \beta_1 x_{ij} + \varepsilon_{ij}$$

avec $i = 1, .., 3$ et $j = 1, 2, 3$ et $\varepsilon_{ij} \sim N(0, \sigma^2)$. Le vecteur des paramètres linéaires est $\beta = (\mu, \tau_1, \tau_2, \tau_3, \beta_1)^T$.

(a) Quelles sont les dimensions de la matrice X du modèle ?

(b) Quel est le rang de la matrice X (e.g. la dimension de l'espace des fonctions estimables) de ce modèle linéaire ? (On peut supposer que les x_{ij} sont tous différents.)

(c) Exprimez $E(\overline{y_{1\bullet}} - \overline{y_{2\bullet}})$ comme une fonction $L\beta$ des paramètres du modèle.

(d) Donnez la fonction des paramètres $L\beta$ qui mesure la différence de moyennes entre les traitements 1 et 3 à $x = 30$. Donnez un estimateur non biaisé de cette différence, écrit en termes des y_{ij} et des x_{ij}.

5. Un kinésiologue souhaite comparer l'efficacité de trois programmes d'exercices pour réduire le rythme cardiaque après effort. 24 jeunes hommes ont été assignés aléatoirement à l'un des trois programmes, qu'ils ont dû suivre rigoureusement pendant 8 semaines. On a ensuite fait courir les hommes pendant 6 minutes et mesuré leur rythme cardiaque. Puisque les hommes n'avaient pas nécessairement la même forme physique au début de l'étude, on a mesuré leur rythme cardiaque au repos avant le début de l'entraînement.

Voici les données, issues de [15] (vous les trouverez aussi sur le site du cours).

Programme 1		**Programme 2**		**Programme 3**	
rc_fin	rc_init	rc_fin	rc_init	rc_fin	rc_init
118	56	148	60	153	56
138	59	159	62	150	58
142	62	162	65	158	61
147	68	157	66	152	64
160	71	169	73	160	72
166	76	164	75	154	75
165	83	179	84	155	82
171	87	177	88	164	86

(a) Ajustez un modèle d'analyse de la covariance à ces données. Simplifiez le modèle autant que possible en vous basant sur des tests d'hypothèses appropriés. Écrivez clairement l'équation générale du modèle final que vous ajustez.

(b) Quelles sont les équations particulières des trois droites ajustées ?

(c) Pour chacune des valeurs de rythme cardiaque initial ci-dessous,
- calculez les moyennes ajustées du rythme cardiaque final dans chaque programme ;
- donnez l'erreur-type associée à chaque moyenne ajustée ;
- faites un test global pour vérifier si les programmes ont un effet différent.

- faites des tests de comparaisons multiples si cela est pertinent.

(i) $x = 60$ pulsations / minute ;

(ii) $x = 85$ pulsations / minute ;

(d) Un statisticien propose de réaliser une analyse de variance sur les données appariées, i.e. en utilisant la différence $rc_{fin} - rc_{init}$ comme variable réponse. Ce modèle mènera-t-il aux mêmes conclusions que le modèle ajusté en (a) ? À quel modèle de régression avec rc_{fin} en variable réponse ce modèle correspond-il ?

5 Le modèle mixte

5.1 Spécification du modèle

Nous avons étudié le modèle linéaire à effets fixes, spécifié sous forme matricielle de la façon suivante :

$$Y = X\beta + \varepsilon \quad \text{où} \quad \varepsilon \sim N_n(0, \sigma^2 I_n).$$

Ce modèle est fort utile, mais peut s'avérer parfois trop restrictif pour certains plans d'expériences. On peut souhaiter une plus grande flexibilité de la matrice de variances-covariances de Y, par exemple en ajoutant des effets aléatoires au modèle ou en supposant des variances non homogènes pour certains groupes de traitements. C'est le modèle mixte qui nous permettra de modéliser un phénomène dans ces conditions. On le spécifie comme suit :

$$Y = X\beta + ZU + \varepsilon$$

où

$$\begin{array}{rcl} Y & = & \text{matrice } N \times 1 \text{ des observations} \\ X & = & \text{matrice } N \times p \text{ d'indicatrices ou de variables continues pour les effets fixes} \\ \beta & = & \text{matrice } p \times 1 \text{ des paramètres des effets fixes} \\ Z & = & \text{matrice } N \times k \text{ d'indicatrices ou de variables continues pour les effets aléatoires} \\ U & = & \text{matrice } k \times 1 \text{ des paramètres des effets aléatoires} \\ \varepsilon & = & \text{matrice } N \times 1 \text{ des erreurs aléatoires résiduelles} \end{array}$$

On suppose que :

- les vecteurs U et ε suivent une loi normale multivariée ;
- l'espérance de ces vecteurs est nulle : $E\begin{bmatrix} U \\ \varepsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$;
- la variance-covariance de ces vecteurs est la suivante : $Var\begin{bmatrix} U \\ \varepsilon \end{bmatrix} = \begin{bmatrix} G & 0 \\ 0 & R \end{bmatrix}$.

Les sous-matrices G et R peuvent prendre différentes formes, auxquelles on référera comme la *structure de covariance*.

En vertu des propriétés d'additivité et de transformation linéaire de la loi normale multidimensionnelle, la distribution du vecteur de la variable dépendante est

$$Y \sim N_N(X\beta, ZGZ^T + R).$$

Sa densité est donnée par :

$$f(Y|\beta, G, R) = L(\beta, G, R) = \frac{\exp\{-(Y - X\beta)^T V^{-1} (Y - X\beta)/2\}}{(2\pi)^{N/2}|V|^{1/2}}. \qquad (9)$$

Cette densité est la vraisemblance à maximiser pour estimer les paramètres.

Exemple de modèle mixte : le plan à blocs aléatoires complets

Considérons un plan dans lequel un facteur fixe à I modalités, où chaque traitement est répété une fois dans J blocs composés d'unités expérimentales homogènes. On appelle ce schéma un plan à blocs aléatoires complets.

Si y_{ij} est la mesure de la variable réponse sur l'unité expérimentale ayant reçu le traitement i dans le bloc j, on peut écrire ce modèle avec effets comme suit :

$$y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij} \left\} \begin{array}{l} i = 1, ..., I \\ j = 1, ..., J \end{array} \right. , \quad \text{où}$$

- μ et τ_i sont les paramètres des effets fixes tels que la moyenne pour le i^e traitement est $\mu_i = \mu + \tau_i$;
- b_j est l'effet aléatoire du j^e bloc ;
- ε_{ij} est l'erreur aléatoire associée à l'unité expérimentale du i^e traitement dans le j^e bloc ;
- $b_j \sim N(0, \sigma_b^2)$ i.i.d. ;
- $\varepsilon_{ij} \sim N(0, \sigma^2)$ i.i.d. ;
- b_j et ε_{ij} sont indépendants.

Sous forme matricielle, on peut écrire ce modèle comme suit :

$$\begin{bmatrix} Y_{11} \\ Y_{21} \\ . \\ . \\ Y_{I1} \\ . \\ . \\ . \\ Y_{1J} \\ Y_{2J} \\ . \\ . \\ Y_{IJ} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & . & . & 0 \\ 1 & 0 & 1 & & & 0 \\ . & . & & . & & . \\ . & . & & & . & . \\ 1 & 0 & 0 & . & . & 1 \\ . & . & & . & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ 1 & 1 & . & . & . & 0 \\ 1 & 0 & 1 & & & 0 \\ . & . & & . & & . \\ . & . & & & . & . \\ 1 & 0 & . & . & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \tau_1 \\ \tau_2 \\ . \\ . \\ \tau_I \end{bmatrix} + \begin{bmatrix} 1 & . & . & . & 0 \\ 1 & . & . & . & 0 \\ . & & . & & . \\ . & & & . & . \\ 1 & . & . & . & 0 \\ . & . & & & . \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & 1 \\ 0 & . & . & . & 1 \\ . & & . & & . \\ . & & & . & . \\ 0 & . & . & . & 1 \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ b_J \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ . \\ . \\ \varepsilon_{I1} \\ . \\ . \\ . \\ \varepsilon_{1J} \\ \varepsilon_{2J} \\ . \\ . \\ \varepsilon_{IJ} \end{bmatrix}$$

On suppose que :

- $U \sim N_J(0, \sigma_b^2 I_J)$;
- $\varepsilon \sim N_N(0, \sigma^2 I_N)$;
- La matrice de variance-covariance des observations est composée de J blocs de dimensions $I \times I$ et de 0. Elle s'écrit comme suit :

$$V_{N\times N} = ZGZ^T + R$$

$$= \begin{bmatrix} \sigma_b^2+\sigma^2 & \sigma_b^2 & ... & \sigma_b^2 & & 0 & 0 & \dots & 0 \\ \sigma_b^2 & \sigma_b^2+\sigma^2 & ... & \sigma_b^2 & & 0 & 0 & \dots & 0 \\ ... & ... & ... & ... & & & & \dots & \\ \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2+\sigma^2 & & 0 & 0 & \dots & 0 \\ & & & & . & & & & \\ & & & & . & & & & \\ & & & & . & & & & \\ 0 & 0 & \dots & 0 & & \sigma_b^2+\sigma^2 & \sigma_b^2 & ... & \sigma_b^2 \\ 0 & 0 & \dots & 0 & & \sigma_b^2 & \sigma_b^2+\sigma^2 & ... & \sigma_b^2 \\ & & \dots & & & ... & ... & ... & ... \\ 0 & 0 & \dots & 0 & & \sigma_b^2 & \sigma_b^2 & ... & \sigma_b^2+\sigma^2 \end{bmatrix}$$

μ-modèle ou modèle avec effets ?

Dans un modèle linéaire mixte, l'écriture de la partie fixe du modèle peut se faire de deux façons, avec un μ-modèle ou un modèle avec effets. Les logiciels tels SAS optent pour la deuxième paramétrisation. On est en présence, comme dans l'exemple précédent, d'une matrice X singulière.

La notion d'estimabilité est donc pertinente pour les modèles mixtes. Une combinaison linéaire $L\beta$ des paramères fixes est estimable si L appartient à l'espace vectoriel généré par les lignes de X. On peut se ramener à un modèle de plein rang en imposant des contraintes de type "somme à 0" ou "dernier paramètre fixé à 0". Dans la procédure MIXED de SAS, l'option `/e` de l'énoncé `model` permet de faire sortir la forme générale des fonctions estimables.

Dans un modèle mixte, les effets aléatoires sont traités d'une façon particulière ; ils n'interviennent pas dans l'écriture de la partie fixe du modèle. La notion d'estimabilité ne s'applique pas aux effets aléatoires. On peut "prédire" toutes les variables U_i apparaissant dans l'écriture générale du modèle.

5.2 Estimation des paramètres

Cette étape est plus complexe pour le modèle mixte que pour le modèle linéaire à effets fixes : il faut estimer β, mais aussi les variances/covariances apparaissant dans G et R.

5.2.1 Estimation des paramètres de covariance dans G et R

Maximum de vraisemblance (ML)

On peut appliquer la méthode du maximum de vraisemblance (ML) et supposer sans perte de généralité que la matrice X est réduite à une matrice de plein rang. La log-vraisemblance à maximiser est obtenue en prenant le logarithme de $f(Y)$:

$$l(\beta, G, R) = c - \frac{1}{2} log(|V|) - \frac{1}{2}(Y - X\beta)^T \, V^{-1} \, (Y - X\beta)$$

En égalant à 0 les dérivées de cette fonction par rapport à β et à chacun des paramètres de covariance, on peut résoudre un système d'équations et obtenir les estimateurs du maximum de vraisemblance. Une procédure souvent utilisée est de résoudre numériquement ces équations pour les composantes de variance de G et R (avec un algorithme

comme Newton-Raphson ou ses variantes), puis d'insérer ces estimations dans l'équation de la dérivée par rapport à β (Jiang [9]) :

$$\frac{\partial l}{\partial \beta} = X^T V^{-1} Y - X^T V^{-1} X\beta = 0$$

Notons cependant que les estimateurs de G et R qui maximisent la vraisemblance sont biaisés, même pour des modèles simples.

Maximum de vraisemblance restreint (REML)

Les estimateurs du maximum de vraisemblance restreint (REML) sont souvent utilisés pour contourner ce problème. Dans Milliken et Johnson ([13], p. 392) et Jiang ([9], p. 13), ces derniers sont définis comme les valeurs qui maximisent la vraisemblance "restreinte"

$$L_R(G, R) = \frac{\exp\{-(A^T Y)^T \{A^T V A\}^{-1} (A^T Y)/2\}}{(2\pi)^{(N-p)/2} |A^T V A|^{1/2}},$$

où p est le nombre de paramètres fixes (incluant l'ordonnée à l'origine) et A est une matrice $N \times (N-p)$ de rang $N-p$ et telle que $A^T X = 0$. Notons que les colonnes de A sont fixes, elle ne dépendent pas de G ni de R. $A^T Y$ est une transformation visant à "éliminer" les effets fixes de la vraisemblance. En fait la vraisemblance REML n'est rien d'autre que la densité de $A^T Y$, soit une $N(0, A^T V A)$. Le résultat de la maximisation ne dépend pas du choix de la matrice A (Jiang, [9]).

C'est la méthode utilisée par défaut par la procédure MIXED de SAS.

Autres méthodes

Il existe d'autres méthodes pour estimer G et R, comme la méthode des moments. On peut estimer les composantes de la variance en résolvant le système d'équations qui pose les sommes de carrés égales à leurs espérances : on obtient ainsi les estimateurs de la méthode des moments (chapitre 19 de Milliken et Johnson[13]). Les estimations sont des combinaisons linéaires de sommes de carrés dont on peut estimer la variance. Il est important de savoir que cette méthode peut donner lieu à des estimations hors de l'espace de définition des paramètres, comme par exemple des variances négatives, ce qui n'est pas le cas des estimateurs du maximum de vraisemblance ou REML.

Algorithmes itératifs

La méthode de maximum de vraisemblance et celle du maximum de vraisemblance restreint s'appuient sur des algorithmes itératifs pour estimer G et R (Newton-Raphson dans PROC MIXED). Elles donnent des estimateurs qui ont de meilleures propriétés asymptotiques que les estimateurs de la méthode des moments. Les variances échantillonnales de ces estimations sont obtenues à l'aide de la matrice d'information de Fisher pour G et R.

5.2.2 Estimation des paramètres dans β et U

La méthode des moindres carrés ordinaires (OLS), utilisée dans le modèle linéaire avec effets fixes seulement, proposait d'estimer β en minimisant la somme des carrés des erreurs, représentée par la fonction

$$(Y - X\beta)^T (Y - X\beta).$$

Pour tenir compte des composantes de la variance dans G et R, on remplace cette fonction par la suivante, dont la minimisation porte le nom de *moindres carrés généralisés* (GLS) :

$$(Y - X\beta)^T \, V^{-1} \, (Y - X\beta)$$

De plus, on insère une estimation des paramètres de G et de R (issus par exemple du ML ou du REML) dans cette fonction (le V devient $\hat{V}$). La solution aux équations normales pour le vecteur β est donc

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} \, X^T \hat{V}^{-1} Y$$

Il s'agit donc du même résultat que celui obtenu en minimisant la vraisemblance. Notons que lorsque le modèle est surparamétrisé, on peut remplacer les inverses impliquant X par des inverses généralisés :

$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-} \, X^T \hat{V}^{-1} Y$$

Peu importe la méthode utilisée pour estimer G et R, la distribution asymptotique de $\hat{\beta}$ est toujours la même :

$$\hat{\beta} \simeq N_p(\beta, [X^T V^{-1} X]^{-1}).$$

Un estimateur de la matrice de variances-covariances de $\hat{\beta}$ est donné par

$$v(\hat{\beta}) = [X^T(\hat{V})^{-1}X]^{-1} = [X^T(Z\hat{G}Z^T + \hat{R})^{-1}X]^{-1}.$$

Une procédure similaire permet d'estimer les paramètres associés aux effets aléatoires, dans le vecteur U :

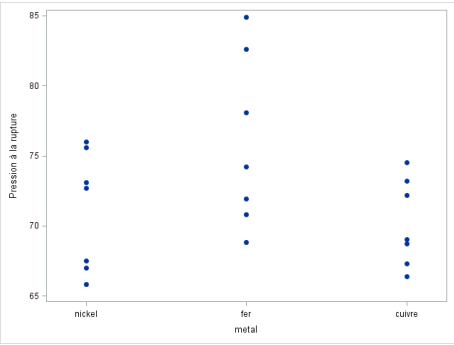
$$\hat{U} = \hat{G}Z^T\,\hat{V}^{-1}\,(Y - X\hat{\beta})$$

L'intérêt de ces estimations est limité, car les facteurs aléatoires sont présents dans le modèle pour qu'on tienne compte de la variabilité dont ils sont responsables. On veut pouvoir inférer sur toute la population des modalités de ces facteurs (inférence large) et non seulement sur les modalités présentes dans les données (inférence étroite). Utiliser les estimations des paramètres des effets aléatoires (par exemple dans une combinaison linéaire) reviendrait à réduire la portée de nos conclusions aux modalités concernées.

Exemple : Plan à blocs aléatoires complets

Considérons les données ci-dessous (Wackerly [25], exemple 13.49) pour illustrer l'estimation des paramètres d'un PBAC à partir de la procédure MIXED de SAS.

On veut comparer trois agents liants métalliques (nickel, fer et cuivre) du point de vue de la pression nécessaire pour rompre un lien entre deux sections d'un matériau composite. On dispose de sept lingots d'un matériau composite séparés en six morceaux, qui seront liés deux par deux avec chaque type de métal. Voici les observations :



Lingot	Cuivre	Fer	Nickel
1	72.2	71.9	67.0
2	66.4	68.8	67.5
3	74.5	82.6	76.0
4	67.3	78.1	72.7
5	73.2	74.2	73.1
6	68.7	70.8	65.8
7	69.0	84.9	75.6
Moy. $\overline{y_{i\bullet}}$	70.2	75.9	71.1
É-type s_i	3.11	6.14	4.26

Si on soumet les commandes ci-dessous à SAS, il ajustera le modèle mixte correspondant au plan à blocs aléatoires complets. Notons que seuls les facteurs fixes apparaissent dans l'énoncé *model*. On présente quelques éléments de la sortie.

```
proc mixed data=metal;
class  lingot metal;
model  press = metal /solution;
random lingot        /solution;
run;
```

Dimensions	
Covariance Parameters	2
Columns in X	4
Columns in Z	7
Subjects	1
Max Obs Per Subject	21

Covariance Parameter Estimates	
Cov Parm	Estimate
lingot	11.4478
Residual	10.3716

Solution for Fixed Effects						
Effect	metal	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept		71.1000	1.7655	6	40.27	<.0001
metal	cuivre	-0.9143	1.7214	12	-0.53	0.6050
metal	fer	4.8000	1.7214	12	2.79	0.0164
metal	nickel	0	.	.	.	.

Solution for Random Effects						
Effect	lingot	Estimate	Std Err Pred	DF	t Value	Pr > |t|
lingot	1	-1.5580	1.9777	12	-0.79	0.4461
lingot	2	-3.7086	1.9777	12	-1.88	0.0853
lingot	3	4.0743	1.9777	12	2.06	0.0618
lingot	4	0.2341	1.9777	12	0.12	0.9077
lingot	5	0.8485	1.9777	12	0.43	0.6755
lingot	6	-3.0429	1.9777	12	-1.54	0.1498
lingot	7	3.1527	1.9777	12	1.59	0.1369

Les estimations du paramètre β utilisent la contrainte "dernière composante à 0" (en raison de l'algorithme de formation de l'inverse généralisé). L'estimation de μ, l'ordonnée à l'origine, est donc $\overline{y}_{3\bullet} = 71.1$, et les autres paramètres fixes sont des déviations des moyennes par rapport à ce μ de référence.

Notons que $\overline{y}_{\bullet 1} - \overline{y}_{\bullet 2} = 2.80 \neq -1.556 + 3.709 = 2.153$. Ces estimations diffèrent de celles du modèle où on aurait considéré lingot comme un facteur fixe.

5.3 Inférence statistique

5.3.1 Inférence sur les paramètres des effets fixes

Un seul paramètre β_j

La variance de $\hat{\beta}_j$, le j^e élément de $\hat{\beta}$, est $v(\hat{\beta})_{jj}$, l'élément (j,j) de $v(\hat{\beta})$. Dans un modèle linéaire standard, on a $v(\hat{\beta})_{jj} = \hat{\sigma}^2[(X'X)^{-1}]_{jj}$. Cette variance est égale à une constante multipliée par une variable aléatoire χ^2_{N-p} où p est le nombre de paramètres fixes du modèle. Ainsi

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{v(\hat{\beta})_{jj}}} \sim t_{N-p}. \tag{10}$$

Pour un modèle linéaire standard, la distribution (10) est exacte. Dans un modèle mixte, la loi de $(\hat{\beta}_j - \beta_j)/\sqrt{v(\hat{\beta})_{jj}}$ est complexe. On l'approxime souvent à l'aide d'une distribution t, mais avec des degrés de liberté estimés.

Une méthode pour estimer les degrés de liberté de cette distribution approximative est la suivante (détaillée à la section 22.3.2 de Milliken et Johnson [13]) :

i) Connaissant la matrice de variances-covariances de $\hat{R}$ et $\hat{G}$, on estime la variance de $v(\hat{\beta})_{jj}$ par linéarisation ;
ii) On utilise la méthode de Satterthwaite pour estimer les degrés de liberté m de $v(\hat{\beta})_{jj}$;
iii) On utilise (10) avec $N - p$ remplacé par m.

Une combinaison linéaire des paramètres de β

On peut tester des hypothèses ou construire des intervalles de confiance sur des combinaisons linéaires des effets fixes, s'il s'agit d'une fonction estimable. On écrit l'hypothèse de la façon suivante :

$$H_0 : L\beta = 0$$

où L peut être un vecteur ou une matrice de coefficients.

Si L est un vecteur, on contruit la statistique suivante pour tester H_0 :

$$T = \frac{L\hat{\beta}}{\sqrt{L\, v(\hat{\beta})\, L^T}} \approx t_m$$

Cette statistique suit une loi exacte de Student (sous H_0) dans le cas équilibré et dans quelques autres cas. En général, T suit une loi de Student approximative, et ses degrés de liberté doivent être approximés. Nous les noterons m.

Si L est une matrice (donc on teste plusieurs combinaisons linéaires simultanément), on construit la statistique F suivante pour tester H_0 :

$$F = \frac{(L\hat{\beta})^T\, (L\, v(\hat{\beta})\, L^T)^{-1}\, L\hat{\beta}}{rang(L)} \approx F_{rang(L),m}$$

Comme pour t, cette statistique suit une loi exacte de Fisher (sous H_0) dans certains cas seulement. En général, F suit une loi de Fisher approximative, et ses degrés de liberté au dénominateur doivent être approximés. La loi approximative est donc $F_{rang(L),m}$.

La deuxième méthode de construction d'une somme de carrés pour les modèles linéaires, qui s'appuie sur une différence de sommes de carrés résiduelles, ne se généralise pas au modèle mixte. En effet, les estimateurs des paramètres de variance pour le modèle complet et le modèle restreint ne sont pas identiques, ce qui rend impossible la comparaison de sommes de carrés résiduelles pour les deux modèles.

5.3.2 Inférence sur les paramètres des effets aléatoires

On peut étendre l'inférence à des combinaisons linéaires impliquant des modalités particulières des effets aléatoires, mais tel que discuté précédemment, cela réduit la portée de l'inférence. Il est néanmoins possible de le faire en spécifiant un vecteur ou une matrice L de coefficients pour tester l'hypothèse suivante (et dont la partie fixe correspond à une fonction estimable) :

$$H_0 : L \begin{bmatrix} \beta \\ U \end{bmatrix} = 0$$

On utiliserait la statistique F ci-dessous, suivant une loi de Fisher asymptotique :

$$F = \frac{\begin{bmatrix} \hat{\beta} \\ \hat{U} \end{bmatrix}^T L^T (L\, v(\hat{\beta}, \hat{U}) L^T)^{-1} L \begin{bmatrix} \hat{\beta} \\ \hat{U} \end{bmatrix}}{rang(L)} \approx F_{rang(L),m}$$

Le modèle mixte est très flexible car il distingue bien les deux types d'effets : fixe ou aléatoire. Le prix à payer pour cette flexibilité est que l'inférence doit se faire à l'aide de méthodes approximatives dont la légitimité repose sur des résultats asymptotiques. Il n'y a pas de façon unique de calculer le seuil observé pour un test ni de déterminer les degrés de liberté pour les statistiques des tests.

Exemple

On peut faire afficher la matrice L utilisée pour les tests de type III sur les facteurs fixes avec l'option e3 de l'énoncé `model`. La matrice $v(\hat{\beta}) = (X^T\hat{V}^{-1}X)^-$, i.e. l'estimation de la variance-covariance des paramètres des effets fixes s'obtient avec l'option `covb`.

```
proc mixed data=metal;
class  lingot metal;
model  press = metal / e3 covb;
random lingot ;
estimate "cuivre-fer tous lingots"      metal 1 -1 0;
estimate "cuivre-fer tous lingots idem"     metal 1 -1 0 | lingot 0 0 0 0 0 0 0 ;
estimate "cuivre lingots 1-7"  intercept 7  metal 7  0 0 | lingot 1 1 1 1 1 1 1/divisor=7 ;
estimate "cuivre tous lingots" intercept 7  metal 7  0 0 | lingot 0 0 0 0 0 0 0/divisor=7 ;
run;
```

Type 3 Coefficients for metal

Effect	metal	Row1	Row2
Intercept			
metal	cuivre	1	
metal	fer		1
metal	nickel	-1	-1

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
metal	2	12	6.36	0.0131

Covariance Matrix for Fixed Effects

Row	Effect	metal	Col1	Col2	Col3	Col4
1	Intercept		3.1171	-1.4817	-1.4817	
2	metal	cuivre	-1.4817	2.9633	1.4817	
3	metal	fer	-1.4817	1.4817	2.9633	
4	metal	nickel				

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > |t|
cuivre-fer tous lingots	-5.7143	1.7214	12	-3.32	0.0061
cuivre-fer tous lingots idem	-5.7143	1.7214	12	-3.32	0.0061
cuivre lingots 1-7	70.1857	1.2172	12	57.66	<.0001
cuivre tous lingots	70.1857	1.7655	12	39.75	<.0001

Estimation et comparaison de moyennes

Estimation d'une moyenne de traitement

La moyenne des observations du traitement i s'écrit en fonction des effets du modèle :

$$\overline{y_{i\bullet}} = \mu + \tau_i + \overline{b_\bullet} + \overline{\varepsilon_{i\bullet}}$$

La variance de cette moyenne est :

$$\begin{aligned} Var(\overline{y_{i\bullet}}) &= Var(\mu + \tau_i + \overline{b_\bullet} + \overline{\varepsilon_{i\bullet}}) \\ &= Var(\overline{b_\bullet}) + Var(\overline{\varepsilon_{i\bullet}}) \\ &= \frac{\sigma_b^2}{J} + \frac{\sigma^2}{J} \end{aligned}$$

Comparaison de deux moyennes de traitements

La différence entre la moyenne des observations du traitement i et la moyenne des observations du traitement k s'écrit en fonction des effets du modèle :

$$\begin{aligned} \overline{y_{i\bullet}} - \overline{y_{k\bullet}} &= \mu + \tau_i + \overline{b_\bullet} + \overline{\varepsilon_{i\bullet}} - (\mu + \tau_k + \overline{b_\bullet} + \overline{\varepsilon_{k\bullet}}) \\ &= \tau_i + \overline{\varepsilon_{i\bullet}} - \tau_k - \overline{\varepsilon_{k\bullet}} \end{aligned}$$

La variance de cette différence est :

$$\begin{aligned} Var(\overline{y_{i\bullet}} - \overline{y_{k\bullet}}) &= Var(\tau_i - \tau_k + \overline{\varepsilon_{i\bullet}} - \overline{\varepsilon_{k\bullet}}) \\ &= Var(\overline{\varepsilon_{i\bullet}}) + Var(\overline{\varepsilon_{k\bullet}}) \\ &= \frac{\sigma^2}{J} + \frac{\sigma^2}{J} \\ &= \frac{2\sigma^2}{J} \end{aligned}$$

Exemple

```
proc mixed data=metal;
class   lingot metal;
model   press = metal ;
random  lingot ;
lsmeans metal / pdiff;
run;
```

Least Squares Means

Effect	metal	Estimate	Standard Error	DF	t Value	Pr > |t|
metal	cuivre	70.1857	1.7655	12	39.75	<.0001
metal	fer	75.9000	1.7655	12	42.99	<.0001
metal	nickel	71.1000	1.7655	12	40.27	<.0001

Differences of Least Squares Means

Effect	metal	_metal	Estimate	Standard Error	DF	t Value	Pr > |t|
metal	cuivre	fer	-5.7143	1.7214	12	-3.32	0.0061
metal	cuivre	nickel	-0.9143	1.7214	12	-0.53	0.6050
metal	fer	nickel	4.8000	1.7214	12	2.79	0.0164

Pour le métal 3, on obtient une estimation de $\overline{y}_{3.} = 71.1$ avec une erreur-type de 1.765 comme dans la sortie contenant les estimations des paramètres fixes du modèle.

5.3.3 Inférence sur les paramètres de covariance

Notons d'entrée de jeu qu'éliminer les paramètres de covariance non significatifs n'est pas un objectif de l'analyse dans une expérience planifiée. On s'intéresse surtout aux effets fixes du modèle et on souhaite tenir compte de la structure de covariance basée sur le plan d'expérience. Il est quand même possible de mesurer l'ampleur de la variabilité due aux facteurs aléatoires inclus dans le modèle.

On peut tester des hypothèses sur les paramètres de covariance de deux façons : avec un test de Wald (basé sur la normalité asymptotique des estimateurs du maximum de vraisemblance, non-valide sur la frontière du domaine de définition des paramètres) et avec un test du rapport de vraisemblances (basé sur un mélange de lois du χ^2 lorsque l'hypothèse porte sur la frontière du domaine du paramètre).

Test de Wald

Supposons que θ est le paramètre de covariance sur lequel on veut inférer. Pour tester l'hypothèse bilatérale

$$H_0 : \theta = c,$$

le test de Wald calcule la statistique suivante à partir de l'estimateur du maximum de vraisemblance $\hat{\theta}$:

$$Z = \frac{\hat{\theta} - c}{\sqrt{v(\hat{\theta})}} \approx N(0,1)$$

L'estimation de l'erreur-type $\sqrt{v(\hat{\theta})}$ est obtenue avec l'inverse de la matrice d'information de Fisher (l'opposé de la dérivée seconde de la log-vraisemblance). Ce test est asymptotique, i.e. qu'il est valide pour de grands échantillons seulement. On l'obtient avec l'option `covtest` dans l'énoncé `proc mixed` de la procédure MIXED.

Il est important de noter que lorsque c vaut 0 (lorsqu'on veut tester la significativité d'un paramètre de covariance), la loi normale n'est plus du tout un bon modèle pour $\hat{\theta}$, car les variances ne peuvent pas être négatives (Verbeke et Molenberghs [23], p. 64). Il faut utiliser une autre procédure pour tester ces hypothèses.

Test du rapport de vraisemblances

Une alternative est le test du rapport de vraisemblances, qui permet de comparer deux modèles dont l'un est un cas particulier de l'autre. Pour tester l'hypothèse

$$H_0 : \theta = c,$$

on soustrait les log-vraisemblances sous H_0 et sans contrainte :

$$U = [-2 \ \text{log-vrais}(\hat{\theta}_{H_0})] \ - \ [-2 \ \text{log-vrais}(\hat{\theta})] \quad \sim \chi^2_1 \quad \text{sous} \ H_0$$

Cependant, si l'objectif est de tester la significativité d'un paramètre de covariance, par exemple $H_0 : \theta = 0$, il faut ajuster la distribution de la statistique U, car 0 se situe sur la frontière de la distribution du paramètre. La statistique U suit dans ce cas une distribution qui est souvent un mélange de distributions χ^2 (dont les probabilités ne correspondent pas à une combinaison linéaire des probabilités des chi^2 impliquées). Le seuil observé est calculé comme suit (Self-Liang [21] et Verbeke-Molenberghs [23] p. 69) :

$$\text{p-value} = \begin{cases} 1 & \text{si } U_{obs} = 0 \\ 0.5\, P(\chi^2_1 > U_{obs}) & \text{si } U_{obs} > 0 \end{cases}$$

Ceci correspond à un mélange 50-50 entre une χ^2_0 et une χ^2_1. Pour tester des hypothèses multiples (sur un groupe de paramètres), se référer à Verbeke-Molenberghs [23].

Tests F basés sur les espérances des carrés moyens

Pour certains plans d'expérience simples, il est possible d'établir une statistique de Fisher à partir de l'espérance des carrés moyens (utile lorsque les estimateurs des paramètres de covariance peuvent être exprimés comme une combinaison linéaire des carrés moyens). Lorsque le modèle se complique, cette méthode est moins appropriée. La procédure GLM utilise cette approche.

Exemple

```
proc mixed data=metal covtest cl;
class  lingot metal;
model  pres = metal;
random lingot;
run;
```

Test de Wald (à éviter) : $H_0 : \sigma_b^2 = 0$ (non signif) et $H_0 : \sigma^2 = 0$ (!)

Covariance Parameter Estimates							
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
lingot	11.4478	8.7204	1.31	0.0946	0.05	3.8811	121.55
Residual	10.3716	4.2342	2.45	0.0072	0.05	5.3332	28.2618

Comme il n'est pas raisonnable de supposer que $\hat{\sigma}_b^2 \sim N(0, V(\hat{\sigma}_b^2))$ sous H_0, il n'est pas non plus raisonnable d'interpréter la statistique Z présentée dans le tableau ci-dessus, ni le seuil observé qui lui est associé (0.0946).

Cependant, l'option `covtest` permet d'afficher l'erreur-type estimée de $\hat{\sigma}_b^2$ et l'option `cl` calcule un intervalle de confiance pour σ_b^2 basé sur la loi du χ^2 avec une estimation des degrés de liberté par la méthode de Satterthwaite. Cet intervalle de confiance est abordé plus en détails à la section suivante.

Test du rapport de vraisemblances : $H_0 : \sigma_b^2 = 0$ (signif).

Fit Statistics	
-2 Res Log Likelihood	112.4
AIC (smaller is better)	114.4
AICC (smaller is better)	114.7
BIC (smaller is better)	115.3

Fit Statistics	
-2 Res Log Likelihood	107.8
AIC (smaller is better)	111.8
AICC (smaller is better)	112.6
BIC (smaller is better)	111.7

$$\begin{aligned} U_{obs} &= -2\,\text{log-vrais.(sans lingot)} - [-2\,\text{log-vrais.(avec lingot)}] \\ &= 112.4 - 107.8 = 4.6 \end{aligned}$$

$$\begin{aligned} \text{p-value} &= 0.5 \times P(\chi_1^2 > 4.6) \\ &= 0.5 \times 0.032 = 0.016 \end{aligned}$$

Test F basé sur les espérances des carrés moyens : $H_0 : \sigma_b^2 = 0$ (signif)

```
proc glm data=metal;
class  lingot  metal;
model  press = metal lingot;
random lingot;
run;quit;
```

Source	Type III Expected Mean Square
metal	Var(Error) + Q(metal)
lingot	Var(Error) + 3 Var(lingot)

The GLM Procedure
Tests of Hypotheses for Mixed Model Analysis of Variance

Dependent Variable: press Pression à la rupture

Source	DF	Type III SS	Mean Square	F Value	Pr > F
metal	2	131.900952	65.950476	6.36	0.0131
lingot	6	268.289524	44.714921	4.31	0.0151
Error: MS(Error)	12	124.459048	10.371587		

La statistique $F_{obs} = 4.31$ est construite en divisant $MS(lingot)$ par MSE. Le seuil observé pour $H_0 : \sigma_b^2 = 0$ est 0.0151, très près de la valeur du test du rapport de vraisemblances. Notons que dans le cas du plan à blocs aléatoires, il s'agit du même test que si lingot avait été un facteur fixe.

5.4 Détermination des degrés de liberté par la méthode de Satterthwaite

Dans un modèle statistique, si $\hat{\theta}$ estime le paramètre θ et si on dispose d'une estimation de la variance échantillonnale $v(\hat{\theta})$, on peut souvent approximer la distribution échantillonnale de $\hat{\theta}$ par une $N(\theta, v(\hat{\theta}))$ et l'intervalle de confiance à $100(1-\alpha)\%$ pour θ est $\hat{\theta} \pm z_{\alpha/2}\sqrt{v(\hat{\theta})}$. Cette méthode ne convient pas lorsqu'on estime des composantes de la variance associées à des facteurs aléatoires ayant peu de modalités, disons moins de 30. En effet, ces estimateurs ont souvent une distribution asymétrique qui est mal approximée par une loi normale.

On utilise plutôt la distribution χ^2 pour approximer la distribution échantillonnale des composantes de variance. Satterthwaite [20] a suggéré une méthode pour estimer les degrés de liberté de la χ^2 à partir de $\hat{\theta}$ et de $v(\hat{\theta})$. On en discute aussi dans Welch [26].

Cette méthode est utile lorsqu'on travaille avec des modèles mixtes relativement complexes. Elle est implantée dans la procédure MIXED de SAS. Elle est basée sur les deux résultats ci-dessous.

Résultat 1 : Si la variable aléatoire X est distribuée selon une $c\chi^2_m$ où c est une constante, alors le nombre de degrés de liberté est égal à

$$m = 2E(X)^2/Var(X).$$

Preuve :

- Si $Y \sim \chi^2_m$, alors $E(Y) = m$ et $Var(Y) = 2m$.
- L'approximation de Satterthwaite consiste à approximer la distribution de $X = cY$ par une loi nommée $c\chi^2_m$. On a donc $E(X) = E(cY) = cm$ et $Var(X) = Var(cY) = 2c^2m$.
- On peut donc en déduire que $m = 2E(X)^2/Var(X)$ et $c = E(X)/m$.

En particulier si $\hat{\theta}$ est une estimation de variance échantillonnale $v(\hat{\theta})$, alors l'approximation de Satterthwaite de la distribution de $\hat{\theta}$ est

$$\hat{\theta} \sim \frac{\theta}{m}\chi^2_m \text{ avec } m = \frac{2\hat{\theta}^2}{v(\hat{\theta})}.$$

Résultat 2 : Si $\hat{\theta}$ estime le paramètre θ et si $v(\hat{\theta})$ est une estimation de sa variance échantillonnale, alors un intervalle de confiance de niveau $1-\alpha$ pour θ construit à l'aide de l'approximation de Satterthwaite est

$$\left[\frac{m\hat{\theta}}{\chi^2_{m,\alpha/2}}, \frac{m\hat{\theta}}{\chi^2_{m,1-\alpha/2}}\right] \qquad \text{où} \quad m = \frac{2\hat{\theta}^2}{v(\hat{\theta})}.$$

Exemple : Plan à blocs aléatoires complets

Le modèle s'écrit $y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij}$ où $b_j \sim N(0, \sigma_b^2)$ et $\varepsilon_{ij} \sim N(0, \sigma^2)$, avec $i = 1, \ldots I$, où I est le nombre de modalités du facteur fixe A et $j = 1, \ldots, J$ où J est le nombre de blocs.

L'estimateur REML de la variance associée au bloc est dans ce cas le même que celui de la méthode des moments et peut s'écrire

$$\hat{\sigma}_b^2 = \frac{MSB - MSE}{I}.$$

Puisque $\dfrac{(J-1)MSB}{\sigma^2 + I\sigma_b^2} \sim \chi^2_{J-1}$, et que $\dfrac{(I-1)(J-1)MSE}{\sigma^2} \sim \chi^2_{(I-1)(J-1)}$, on a

$$Var(MSB) = \frac{2(\sigma^2 + I\sigma_b^2)^2}{J-1} \qquad \text{et} \qquad Var(MSE) = \frac{2(\sigma^2)^2}{(I-1)(J-1)}.$$

Puisque ces deux variables aléatoires sont indépendantes, il suit que

$$\begin{aligned} Var(\hat{\sigma}_b^2) &= \frac{1}{I^2}\left(\frac{2(\sigma^2 + I\sigma_b^2)^2}{J-1} + \frac{2(\sigma^2)^2}{(I-1)(J-1)}\right) \\ &= \frac{2\sigma_b^4}{J-1} + \frac{4\sigma_b^2\sigma^2}{I(J-1)} + \frac{2\sigma^4}{I(I-1)(J-1)}. \end{aligned}$$

Donc les degrés de liberté associés à $\hat{\sigma}_b^2$ sont

$$m = \frac{2\sigma_b^4}{Var(\hat{\sigma}_b^2)} = \frac{J-1}{1 + \frac{2}{I}\left(\frac{\sigma^2}{\sigma_b^2}\right) + \left(\frac{\sigma^2}{\sigma_b^2}\right)^2\left[\frac{1}{I(I-1)}\right]}.$$

Ainsi m est inférieur ou égal à $J-1$. En fait, m tend vers $J-1$ lorsque I est grand (on a un grand nombre d'observations pour chaque bloc), ou lorsque σ_b^2 est grand par rapport à σ^2.

Exemple de calcul avec SAS.

On fait d'abord appel à la procédure GLM pour le calcul du MSB et pour le test F présenté dans sa table d'anova.

```
proc glm data=metal;
 class  lingot;
 model  y = metal lingot;
 random lingot;
run;
```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	400.1904762	50.0238095	4.82	0.0076
Error	12	124.4590476	10.3715873		
Corrected Total	20	524.6495238			

Source	DF	Type III SS	Mean Square	F Value	Pr > F
metal	2	131.9009524	65.9504762	6.36	0.0131
lingot	6	268.2895238	44.7149206	4.31	0.0151

On a

$$\begin{aligned}
\hat{\sigma}_b^2 &= \frac{MSB - MSE}{I} = \frac{44.71 - 10.37}{3} = 11.45 \\
v(\hat{\sigma}_b^2) &= \frac{1}{I^2}\left(\frac{2(\sigma^2 + I\sigma_b^2)^2}{J-1} + \frac{2(\sigma^2)^2}{(I-1)(J-1)}\right) \\
&= \frac{1}{9}\left(\frac{2(10.37 + 3 \times 11.45)^2}{6} + \frac{2(10.37)^2}{12}\right) \\
&= 76.05 = (8.72)^2
\end{aligned}$$

Ces statistiques sont calculées directement avec l'option `covtest` de PROC MIXED.

On utilise l'option `covtest` pour estimer la variance de l'estimateur REML de σ_b^2, et l'option `cl` pour obtenir l'intervalle de confiance pour σ_b^2 basé sur la loi du χ^2.

```
proc mixed  data=metal  covtest cl;
 class  lingot;
 model  y = metal;
 random lingot ;
run;
```

Covariance Parameter Estimates							
Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
lingot	11.4478	8.7204	1.31	0.0946	0.05	3.8811	121.55
Residual	10.3716	4.2342	2.45	0.0072	0.05	5.3332	28.2618

Si on teste (à tort !) l'hypothèse $H_0 : \sigma_b^2 = 0$ avec la statistique $Z = \hat{\sigma}_b^2/\sqrt{v(\hat{\sigma}_b^2)}$, on trouve $z_{obs} = 1.31$ pour un seuil observé unilatéral de 0.0946 (beaucoup plus élevé que celui du test de la table ANOVA : 0.0151. À titre de comparaison, le test du rapport de vraisemblance donne lieu à un p-value de 0.016). Tel que discuté à la section précédente, ce test de Wald est trompeur car la loi normale n'est pas une bonne approximation de la loi de $\hat{\sigma}_b^2$, qui est plutôt étirée vers la droite. Dans plan à blocs aléatoires, la valeur maximale de la statistique Z est $\sqrt{(J-1)/2}$ (car $m = 2 * z_{obs}^2 \leq J-1$). Ce maximum est toujours inférieur à 1.645 si $J \leq 6$. Ainsi, tester $H_0 : \sigma_b^2 = 0$ contre $H_1 : \sigma_b^2 > 0$ avec cette statistique entraîne l'acceptation automatique de H_0 au seuil de 5% si $J \leq 6$!

Le nombre de degrés de liberté associés à $\hat{\sigma}_b^2$ est estimé par

$$m = 2z_{obs}^2 = 2 \times \left(\frac{11.45}{8.72}\right)^2 = 3.45.$$

L'intervalle de confiance pour σ_b^2 à 95% obtenu avec la méthode de Satterthwaite est

$$\left[\frac{m \times \hat{\sigma}_b^2}{\chi^2_{3.45,0.025}}, \frac{m \times \hat{\sigma}_b^2}{\chi^2_{3.45,0.975}}\right] = \left[\frac{3.45 \times 11.45}{10.17}, \frac{3.45 \times 11.45}{0.325}\right] = [3.88,\ 121.55]$$

De la même manière, on peut vérifier que SAS calcule les degrés de liberté pour l'estimation de σ^2 par $m = 2 \times 2.45^2 = 12.00$.

L'approximation de Satterthwaite est couramment utilisée pour construire des intervalles de confiance pour des estimations de variance. Dans un modèle linéaire mixte, elle permet aussi d'approximer les degrés de liberté des estimateurs de variance des paramètres fixes du modèle et ainsi de calculer des intervalles de confiance pour ces paramètres.

5.5 Analyse des résidus

Bien entendu, notre modèle s'est basé sur des postulats que nous devons vérifier pour valider l'analyse. On vérifie les postulats à l'aide des résidus.

Proc MIXED calcule deux sortes de résidus, liés aux moyennes marginales et aux moyennes conditionnelles :

	Moyennes	Résidus	Option
marginales	$E(Y) = X\beta$	$\hat{\varepsilon}_{marg} = Y - X\hat{\beta}$	$outpm =$
conditionnelles	$E(Y\|U) = X\beta + ZU$	$\hat{\varepsilon}_{cond} = Y - X\hat{\beta} - Z\hat{U}$	$outp =$

Puisque l'inférence porte sur toute la population des effets aléatoires, on utilisera les résidus marginaux. Par contre, comment vérifier que la structure de V est respectée ?

On utilise la décomposition de Cholesky pour calculer une sorte de "racine carrée" de la matrice $\hat{V}$, que l'on exprimera comme $(\hat{V}^{1/2})'\hat{V}^{1/2} = V$. Si on définit les *résidus standardisés* (scaled residuals) par

$$\hat{\varepsilon}_{scaled} = \hat{V}^{-1/2}\hat{\varepsilon}_{marg},$$

on sait que ces résidus auront approximativement une moyenne de 0, une variance unitaire, et ne seront pas corrélés. (Remarque : Ceci est vrai malgré que $\hat{V}$ ne soit pas exactement la matrice de covariance de $\hat{\varepsilon}_{marg}$.)

Dans SAS, ces résidus se calculent avec les options `outpm=` et `vciry`.

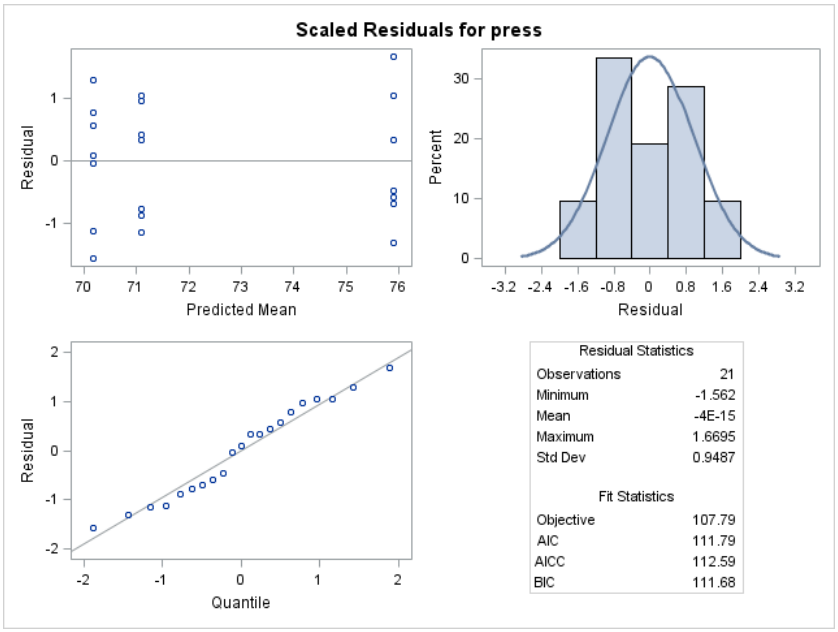
Exemple

```
proc mixed data=metal;
class  lingot  metal;
model  press = metal / outpm=res vciry;
random lingot;
run;

proc print data=res;
run;
```

Le jeu de données créé par `outpm` contient des intervalles de confiance sur les valeurs prédites pour la moyenne, les résidus classiques, les résidus standardisés (qui nous intéressent ici), et les valeurs de la variable dépendante standardisées. SAS fournit automatiquement un tableau de 3 figures pour l'analyse graphique des résidus standardisés.

Obs	lingot	metal	press	Pred	StdErrPred	DF	Alpha	Lower	Upper	Resid	ScaledResid	ScaledDep
1	1	nickel	67.0	71.1636	1.38052	12	0.05	68.1557	74.1715	-4.16359	-1.21482	19.5488
2	1	fer	71.9	75.9636	1.38052	12	0.05	72.9557	78.9715	-4.06359	-0.97670	17.6498
3	1	cuivre	72.2	70.2493	1.38052	12	0.05	67.2414	73.2572	1.95070	0.97455	15.1581
4	2	nickel	67.5	71.1636	1.38052	12	0.05	68.1557	74.1715	-3.66359	-0.65481	12.0647



Structure de covariance

On voit que les résidus standardisés ont une dispersion similaire dans tous les traitements. On peut donc conclure que $\sigma^2 + \sigma_b^2$ est relativement constant d'un métal à l'autre.

Normalité des résidus

On utilise les diagnostics habituels pour vérifier la normalité des résidus. L'histogramme fourni par PROC MIXED est relativement symétrique, le diagramme quantile-quantile est à peu près linéaire.

On peut aussi conduire des tests de normalité avec la procédure UNIVARIATE. Les résultats non significatifs dans notre exemple, donc la normalité des observations n'est pas rejetée.

```
proc univariate data=res plot normal;
var  ScaledResid;
run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.957663	Pr < W	0.4704
Kolmogorov-Smirnov	D	0.117008	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.057256	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.341746	Pr > A-Sq	>0.2500

5.6 Exercices

1. Vrai ou Faux ?

(a) Un facteur est dit aléatoire quand le nombre d'observations dans chaque modalité n'est pas déterminé d'avance.

(b) L'ajout de facteurs aléatoires dans un modèle d'analyse de la variance vise à préserver l'hypothèse d'indépendance entre les observations.

(c) Le but de la formation de blocs aléatoires est d'isoler une partie de la variabilité de y qui serait autrement confondue avec l'erreur résiduelle.

(d) On retrouvera 20 degrés de liberté à l'erreur dans un plan d'expérience à 4 blocs aléatoires complets visant à comparer 5 traitements.

(e) L'inférence sur les paramètres de covariance porte sur les paramètres de la matrice U.

(f) L'algorithme d'estimation REML (maximum de vraisemblance restreint) estime d'abord les paramètres du vecteur β en transformant les observations pour "éliminer" les paramètres de covariance, puis utilise ces estimations $\hat{\beta}$ pour les injecter dans l'estimation des paramètres de covariance.

2. **(a)** Montrez que la matrice de variance-covariance des observations Y dans un modèle mixte est donnée par $V = ZGZ' + R$.

(b) Dans un plan à blocs aléatoires complets (un facteur à I modalités répété une fois dans J blocs), quelle est la variance d'une observation y_{ij} ?

(c) Dans un PBAC, quelle est la variance de la moyenne des observations du traitement i ($\overline{y}_{i\bullet}$) ?

(d) Dans un PBAC, quelle est la covariance entre deux observations d'un même bloc (y_{ij} et $y_{i'j}$) ?

(e) Dans un PBAC, quelle est la covariance entre deux observations d'un même traitement mais de blocs différents (y_{ij} et $y_{ij'}$) ?

3. Un chercheur en éducation veut évaluer la part de variabilité due à l'enseignant dans la réussite des élèves de quatrième secondaire aux épreuves uniformes de mathématiques. Pour son projet pilote, il sélectionne aléatoirement quatre enseignants affectés à des classes de milieu social similaire et note les moyennes obtenues par chacun de leurs groupes. Voici les observations récoltées :

Professeur	Aline	Bertrand	Claudette	Denis
Groupe 1	75	67	62	71
Groupe 2	76	63		68
Groupe 3	68			

On propose d'utiliser un modèle d'ANOVA à un facteur aléatoire

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij}, \qquad i = 1, 2, 3, j = 1, n_i,$$

où on suppose que τ_i et ε_{ij} sont des variables aléatoires indépendantes suivant des lois normales de moyennes nulles et de variances égales à σ_τ^2 et σ^2, respectivement.

(a) Quelles sont les dimensions des matrices du modèle exprimé sous forme matricielle ? Explicitez les matrices X, β, Z et U.

(b) Quelle est la distribution théorique du vecteur d'observations $Y = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{31}, y_{41}, y_{42})^T$? Détaillez le vecteur de l'espérance et la matrice de variances-covariances.

(c) Donnez une estimation ponctuelle de σ_τ^2, et son erreur-type associée.

(d) Utilisez le test du rapport de vraisemblance pour évaluer l'ampleur de l'impact du professeur dans la note moyenne des élèves.

4. On considère une expérience pour comparer les $I = 3$ modalités d'un facteur A à l'aide de $J = 3$ blocs. Le modèle est

$$y_{ij} = \mu + \tau_i + b_j + \varepsilon_{ij} \text{ avec } i = 1, 2, 3, \quad j = 1, 2, 3$$

où $b_j \sim N(0, \sigma_b^2)$ et $\varepsilon_{ij} \sim N(0, \sigma^2)$ sont des variables aléatoires indépendantes. Le schéma contient deux cellules vides :

A \ B	j=1	j=2	j=3
i=1	ND	ND	
i=2			
i=3			

(a) Ordonnons le vecteur d'observations selon le bloc. Quelles sont les dimensions des matrices du modèle exprimé sous forme matricielle ? Explicitez les matrices X, β, Z et U.

(b) Quelle est la distribution théorique de $Y = (y_{21}, y_{31}, y_{22}, y_{32}, y_{13}, y_{23}, y_{33})^T$? Détaillez le vecteur de l'espérance et la matrice de variances-covariances.

(c) Quelle est la distribution théorique de $\overline{Y}_{2\bullet}$?

(d) Quelle serait la distribution théorique de $\overline{Y}_{2\bullet}$ si on enlevait le bloc du modèle ? L'estimation de μ_2 serait-elle plus précise ou moins précise ?

(e) Selon le modèle avec le facteur bloc, quelle est la distribution conjointe des deux contrastes suivants :

$$\begin{bmatrix} y_{21} - y_{22} - y_{31} + y_{32} \\ y_{21} - y_{23} - y_{31} + y_{33} \end{bmatrix}$$

Pensez à définir une matrice de coefficients A et à appliquer les propriétés de la loi normale multivariée sur AY.

5. On a obtenu, lors de l'ajustement d'un modèle statistique, une estimation $\hat{\theta}$ d'un paramètre de covariance θ qui vaut $\hat{\theta} = 1.4$ avec une estimation de variance $v(\hat{\theta}) = 0.09$.

(a) Calculez un intervalle de confiance à 95% pour θ en utilisant

(i) une approximation normale pour la distribution de $\hat{\theta}$;

(ii) une modification de la loi khi-deux $\theta\ \chi^2_m/m$ comme distribution de $\hat{\theta}$ où m est déterminé selon la méthode de Satterthwaite.

(b) Refaites les deux calculs précédents si $v(\hat{\theta}) = 0.81$ et si $v(\hat{\theta}) = 0.001$. Placez les résultats des calculs dans un tableau.

(c) Dans quelles conditions les intervalles de confiance calculés selon les deux méthodes donnent-ils des résultats à peu près identiques ?

6. Considérons 2 échantillons indépendants, de taille n_1 et n_2. On suppose que chaque groupe d'observations est issu d'une loi normale $N(\mu_i, \sigma_i^2)$. Dénotons par $\overline{Y}_{i\bullet}$ et S_i^2 la moyenne et la variance empiriques de chaque échantillon.

(a) Comment peut-on définir la loi de S_1^2 ?

(b) Quelles sont l'espérance et la variance de S_1^2 ?

(c) Lorsqu'on veut estimer la différence des moyennes $\mu_1 - \mu_2$, on utilise $\overline{Y}_{1\bullet} - \overline{Y}_{2\bullet}$. Quelle est l'estimation de la variance de cette différence ?

(d) Utiliser l'approximation de Satterthwaite pour trouver les degrés de liberté de la loi de Student utilisée pour construire un intervalle de confiance pour $\mu_1 - \mu_2$.

(e) Cette formule vous rappelle-t-elle quelque chose ?

6 Les modèles hiérarchiques

Jusqu'à maintenant, nous avons étudié des plans d'expérience à un facteur (fixe ou aléatoire), des plans factoriels à deux ou plusieurs facteurs croisés (fixes ou aléatoires). Des facteurs sont dits croisés lorsqu'on combine toutes les modalités de chaque facteur avec les modalités des autres facteurs pour collecter des observations.

6.1 Caractérisation du modèle hiérarchique

Pour qu'un modèle soit considéré comme hiérarchique, il faut que les modalités d'un des facteurs (disons B) changent en fonction du niveau d'un autre facteur (disons A). On dira alors que B est $emboîté$ dans A, ou $niché$ dans A, et on notera cet état de fait par $B(A)$.

Voici quelques remarques sur les modèles hiérarchiques en général :

- Les facteurs A et B peuvent être fixes ou aléatoires.
- A et B ne peuvent pas être croisés, i.e. qu'on ne peut pas avoir l'interaction $A * B$ dans le modèle si B est emboîté dans A.
- D'autres facteurs (fixes ou aléatoires) peuvent être présents dans le modèle. Ils peuvent être simples, nichés ou croisés.

Nous présentons dans les pages suivantes trois exemples pour illustrer des situations dans lesquelles les modèles hiérarchiques peuvent s'appliquer. Nous analyserons en détails les données des exemples 1 et 3 dans le reste du chapitre.

Exemple 1

On veut comparer l'efficacité des insecticides disponibles sur le marché pour enrayer un type de parasite particulier présent dans l'herbe. Quatre fournisseurs produisent les insecticides les plus utilisés. La compagnie A fabrique 3 produits ; la compagnie B en fabrique 2 autres ; la compagnie C fabrique 2 produits et la compagnie D fabrique 4 produits. On souhaite comparer les quatre compagnies et leurs onze produits.

On remplit 33 contenants de verre avec de la terre, de l'herbe et 400 insectes. On applique chaque insecticide sur 3 contenants choisis de façon aléatoire. On compte ensuite le nombre d'insectes vivants dans chaque pot. Voici les observations (données tirées de Milliken et Johnson [13], page 627) :

Produit	Cie A	Cie B	Cie C	Cie D
1	151	140	96	79
	135	152	108	74
	137	133	94	73
2	118	151	84	67
	132	132	87	78
	135	139	82	63
3	131			90
	137			81
	121			96
4				83
				89
				94

Notons que les produits sont emboîtés dans les compagnies, car les niveaux du facteur Produit changent à chaque niveau du facteur Compagnie. En d'autres termes, nous ne pouvons pas considérer que le produit 1 de la compagnie A correspond au produit 1 des autres compagnies.

De plus, il est évident que l'interaction entre deux facteurs emboîtés n'a pas de sens. Une interaction entre A et B signifie que la différence entre les modalités de B n'est pas la même d'une modalité à l'autre de A (et vice versa). Dans un modèle hiérarchique, les modalités de B ne sont pas les mêmes d'une modalité à l'autre de A ; il n'est donc pas question de voir si l'effet de B est constant pour tous les niveaux de A.

On a ici la structure de traitements suivante :

Facteur	C	$P(C)$	$V(PC)$
Statut (F ou A)			
Degrés de liberté			

En fait on peut présenter ces données comme provenant d'un modèle d'analyse de variance à un facteur, Produit, à 11 modalités. Parmi les 10 degrés de liberté associés aux différences inter-produit, on peut en distinguer 3 qui sont associés aux différences entre les compagnies et 7 pour les différences entre les produits d'une même compagnie.

Le modèle avec effets pour l'exemple 1 s'écrirait comme suit :

$$y_{ijk} = \mu + \tau_i + \gamma_{j(i)} + \varepsilon_{ijk} \left\} \begin{array}{l} i = 1, ..., 4 \\ j(i) = 1, ..., p_i \\ k = 1, 2, 3 \end{array} \right.$$

- μ est un paramètre de référence ;
- τ_i est l'effet de la compagnie i ;
- $\gamma_{j(i)}$ est l'effet du produit j de la compagnie i ;
- ε_{ijk} est l'erreur aléatoire associée à la k^e unité expérimentale sur laquelle on a appliqué l'insecticide j de la compagnie i ;
- $\varepsilon_{ijk} \sim N(0, \sigma^2)$ i.i.d..

Exercice : Écrire ce modèle sous forme matricielle.

Exemple 2

On étudie le temps d'attente pour recevoir des soins à l'urgence dans les hôpitaux du Canada. On voudra comparer la performance des provinces par rapport à cet indicateur. Dans chaque province, on choisit 3 villes au hasard parmi les villes de $50\,000$ habitants et plus. Dans chaque ville, on choisit au hasard deux hôpitaux sur lesquels une mesure globale du temps d'attente est prise.

Notons que les villes sont emboîtées dans les provinces, car les modalités du facteur ville changent à chaque niveau du facteur province. En d'autres termes, nous ne pouvons pas considérer que la ville 1 de Colombie-Britannique correspond à la ville 1 du Québec. Il en va de même pour les hôpitaux, nichés dans les villes. Encore une fois, il n'y a pas d'interaction entre les facteurs nichés, car cela n'aurait pas de sens de vérifier si la différence entre Québec et Montréal est la même dans toutes les provinces !

On a ici la structure de traitements suivante :

Facteur	
Statut (F ou A)	
Degrés de liberté	

Le modèle pour l'exemple 2 s'écrirait comme suit :

$$y_{ijk} = \mu + \tau_i + b_{j(i)} + \varepsilon_{ijk} \left\{ \begin{array}{l} i = 1, ..., 10 \\ j(i) = 1, 2, 3 \ \ \forall i \\ k = 1, 2 \end{array} \right.$$

- μ est un paramètre de référence ;
- τ_i est l'effet de la province i ;
- $b_{j(i)}$ est l'effet aléatoire de la ville j dans la province i ;
- ε_{ijk} est l'erreur aléatoire associée au k^e hôpital dans la ville j dans la province i ;
- $b_{j(i)} \sim N(0, \sigma_b^2)$ i.i.d. ;
- $\varepsilon_{ijk} \sim N(0, \sigma^2)$ i.i.d..

Exercice : Écrire ce modèle sous forme matricielle.

Exemple 3

On veut comparer la sensation de confort chez les hommes et les femmes soumis à trois températures ambiantes. On dispose de 9 chambres, dont 3 choisies au hasard seront amenées à une température de $15°$C, 3 autres à $20°$C, et 3 autres à $25°$C. On placera tour à tour dans chaque chambre deux hommes et deux femmes choisis au hasard parmi les 36 participants. Après un certain temps passé (seul) dans la chambre, les sujets donneront une mesure de leur confort (par rapport à la température) sur une échelle de 1 à 15. Cet exemple est traité à la page 147 de Milliken et Johnson [13].

Il est important de remarquer que ce plan comporte deux types d'unités expérimentales, qui sont de tailles différentes. La température est appliquée sur une chambre (premier type d'unité), tandis que le sexe est relatif à une personne (deuxième type d'unité).

On a ici la structure de traitements suivante :

Facteur	
Statut (F ou A)	
Degrés de liberté	

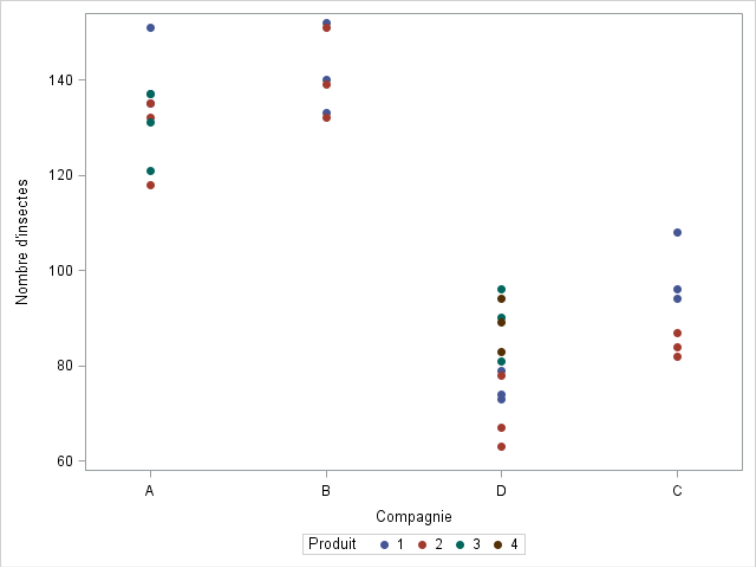
Le modèle pour l'exemple 3 s'écrirait comme suit :

$$y_{ijkl} = \mu + \tau_i + b_{j(i)} + \beta_k + \tau\beta_{ik} + \varepsilon_{ijkl} \left\{ \begin{array}{l} i = 1,2,3 \\ j(i) = 1,2,3 \ \ \forall i \\ k = 1,2 \\ l = 1,2 \end{array} \right.$$

- μ est un paramètre de référence ;
- τ_i est l'effet de la température i ;
- $b_{j(i)}$ est l'effet aléatoire de la chambre j dans la température i ;
- β_k est l'effet du sexe k ;
- $\tau\beta_{ik}$ est l'interaction entre l'effet de la température i et du sexe k ;
- ε_{ijkl} est l'erreur aléatoire associée à la l^e personne du sexe k dans la chambre j de température i ;
- $b_{j(i)} \sim N(0, \sigma_b^2)$ i.i.d. ;

- $\varepsilon_{ijkl} \sim N(0, \sigma^2)$ i.i.d..

6.2 Analyse du schéma de l'exemple 1 sur les insecticides



6.2.1 Spécification du modèle dans SAS

Aucun facteur aléatoire ne fait partie du modèle (sauf l'erreur aléatoire, bien entendu !). On peut donc utiliser les procédures GLM ou MIXED, avec le même énoncé `model`, et on obtiendra les mêmes résultats. Puisque nous étudions le cadre général des modèles mixtes, la deuxième option sera favorisée dans tout le chapitre.

Les facteurs emboîtés sont précisés à l'aide de parenthèses.

```
proc glm data=insect;
class cie prod;
model Nmoust = cie prod(cie);
run;quit;

proc mixed data=insect;
class cie prod;
model Nmoust = cie prod(cie);
run;
```

6.2.2 Estimation des paramètres de covariance

Puisqu'il n'y a aucun facteur aléatoire dans ce modèle qui ajouterait des corrélations entre les observations, le seul paramètre de variance à estimer est le σ^2. Par défaut, les paramètres de covariance sont estimés par maximum de vraisemblance restreint (REML) avec PROC MIXED. Les options `covtest` et `cl` permettent l'affichage de bornes de confiance et une base pour l'estimation des degrés de liberté. La procédure GLM aurait affiché une table d'anova et le MSE aurait donné la même estimation ponctuelle.

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
Residual	57.2727	17.2684	3.32	0.0005	0.05	34.2571	114.73

6.2.3 Estimation des paramètres des effets fixes

Comme dans les autres modèles, l'estimation des paramètres des effets fixes du modèle hiérarchique dépend des contraintes utilisées pour résoudre les équations normales. On obtient une estimation avec l'option `solution` de l'énoncé `model`. Cependant ces estimations sont d'un intérêt limité.

Par contre, on est souvent intéressé à comparer les moyennes des différents traitements, il est bon de les regarder pour se forger une intuition. On peut calculer les moyennes échantillonnales avec la procédure MEANS, mais pour un calcul des erreurs-types cohérent avec le modèle, on fait appel à l'énoncé `lsmeans` de MIXED.

```
lsmeans cie;
lsmeans prod(cie);
```

Least Squares Means							
Effect	Compagnie	Produit	Estimate	Standard Error	DF	t Value	Pr > |t|
cie	A		133.00	2.5226	22	52.72	<.0001
cie	B		141.17	3.0896	22	45.69	<.0001
cie	C		91.8333	3.0896	22	29.72	<.0001
cie	D		80.5833	2.1847	22	36.89	<.0001
prod(cie)	A	1	141.00	4.3693	22	32.27	<.0001
prod(cie)	A	2	128.33	4.3693	22	29.37	<.0001
prod(cie)	A	3	129.67	4.3693	22	29.68	<.0001
prod(cie)	B	1	141.67	4.3693	22	32.42	<.0001
prod(cie)	B	2	140.67	4.3693	22	32.19	<.0001
prod(cie)	C	1	99.3333	4.3693	22	22.73	<.0001
prod(cie)	C	2	84.3333	4.3693	22	19.30	<.0001
prod(cie)	D	1	75.3333	4.3693	22	17.24	<.0001
prod(cie)	D	2	69.3333	4.3693	22	15.87	<.0001
prod(cie)	D	3	89.0000	4.3693	22	20.37	<.0001
prod(cie)	D	4	88.6667	4.3693	22	20.29	<.0001

On remarque que les erreurs-types des moyennes par produit sont égales. Puisque chaque moyenne est calculée à partir des trois observations indépendantes dans chaque cellule, on a que

$$\text{Erreur-type}(\overline{y_{ij\bullet}}) = \sqrt{\frac{\hat{\sigma}^2}{3}} = \sqrt{\frac{57.27}{3}} = 4.369$$

Un calcul similaire permet de retrouver les valeurs des erreurs-types pour les moyennes par compagnie, qui dépendent du nombre d'observations :

$$\text{Erreur-type}(\overline{y_{i\bullet\bullet}}) = \sqrt{\frac{\hat{\sigma}^2}{n_{i\bullet}}}$$

6.2.4 Tests sur les facteurs fixes

Pour comparer les 11 insecticides entre eux, on peut utiliser une anova à un facteur.

```
proc mixed data=insect;
class cie prod;
model Nmoust = cie*prod;
run;
```

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
cie*prod	10	22	42.45	<.0001

On obtient une $F_{10,22}$ de 42.45, et un seuil observé $< 10^{-4}$. Il y a donc des différences significatives entre certains insecticides. On pourrait poursuivre l'analyse en faisant des comparaisons multiples des 11 insecticides entre eux.

On peut aussi décomposer la variabilité en deux sources : la compagnie et les produits(compagnie), soit à l'aide de contrastes orthogonaux basés sur le modèle ci-dessus, ou préférablement en précisant le modèle hiérarchique directement avec `model Nmoust = cie prod(cie)`. On obtient alors le tableau suivant :

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
cie	3	22	132.78	<.0001
prod(cie)	7	22	3.74	0.0081

Ces tests sont basés sur une matrice L de coefficients pour définir les combinaisons linéaires $L\beta$ qui seront testées à l'aide de la statistique F ci-dessous :

$$F = \frac{(L\hat{\beta})^T \left(L\, v(\hat{\beta})\, L^T\right)^{-1} L\hat{\beta}}{rang(L)} \approx F_{rang(L),m}$$

Notons que le vecteur de paramètres aléatoires U ne fait pas partie de ce modèle, et que $v(\hat{\beta}) = \hat{\sigma}^2(X'X)^-$.

- Comparaison des compagnies :

Type 3 Coefficients for cie					
Effect	Compagnie	Produit	Row1	Row2	Row3
Intercept					
cie	A		1		
cie	B			1	
cie	C				1
cie	D		-1	-1	-1
prod(cie)	A	1	0.3333		
prod(cie)	A	2	0.3333		
prod(cie)	A	3	0.3333		
prod(cie)	B	1		0.5	
prod(cie)	B	2		0.5	
prod(cie)	C	1			0.5
prod(cie)	C	2			0.5
prod(cie)	D	1	-0.25	-0.25	-0.25
prod(cie)	D	2	-0.25	-0.25	-0.25
prod(cie)	D	3	-0.25	-0.25	-0.25
prod(cie)	D	4	-0.25	-0.25	-0.25

L'hypothèse testée peut s'écrire :
H_0 : $\overline{\mu_{A\bullet}} = \overline{\mu_{D\bullet}}$
$\overline{\mu_{B\bullet}} = \overline{\mu_{D\bullet}}$
$\overline{\mu_{C\bullet}} = \overline{\mu_{D\bullet}}$

Globalement, on conclut que l'hypothèse $H_0 : \overline{\mu_{A\bullet}} = \overline{\mu_{B\bullet}} = \overline{\mu_{C\bullet}} = \overline{\mu_{D\bullet}}$ est rejetée ($F = 132.78$, seuil observé $<.0001$).

Le nombre moyen de moustiques tués par les produits des quatre compagnies diffèrent. On pourra effectuer des comparaisons multiples entre les compagnies.

Comparaisons multiples

Le test sur le facteur compagnie est significatif, c'est-à-dire que certaines compagnies ont une meilleure performance pour l'élimination de moustiques avec leurs produits en général. On peut vérifier (tableau `Differences of Least Squares Means`) que toutes les compagnies diffèrent deux à deux au seuil de 5% sauf les compagnies A et B, la compagnie D étant la meilleure et la compagnie B étant la moins performante.

```
lsmeans cie / pdiff;
```

Least Squares Means						
Effect	Compagnie	Estimate	Standard Error	DF	t Value	Pr > |t|
cie	A	133.00	2.5226	22	52.72	<.0001
cie	B	141.17	3.0896	22	45.69	<.0001
cie	C	91.8333	3.0896	22	29.72	<.0001
cie	D	80.5833	2.1847	22	36.89	<.0001

Differences of Least Squares Means							
Effect	Compagnie	Compagnie	Estimate	Standard Error	DF	t Value	Pr > |t|
cie	A	B	-8.1667	3.9886	22	-2.05	0.0527
cie	A	C	41.1667	3.9886	22	10.32	<.0001
cie	A	D	52.4167	3.3371	22	15.71	<.0001
cie	B	C	49.3333	4.3693	22	11.29	<.0001
cie	B	D	60.5833	3.7839	22	16.01	<.0001
cie	C	D	11.2500	3.7839	22	2.97	0.0070

- Comparaison des produits au sein d'une même compagnie :

L'hypothèse testée peut s'écrire :

$$
\begin{aligned}
H_0: \quad & \overline{\mu_{1(A)}} = \overline{\mu_{3(A)}} \\
& \overline{\mu_{2(A)}} = \overline{\mu_{3(A)}} \\
& \overline{\mu_{1(B)}} = \overline{\mu_{2(B)}} \\
& \overline{\mu_{1(C)}} = \overline{\mu_{2(C)}} \\
& \overline{\mu_{1(D)}} = \overline{\mu_{4(D)}} \\
& \overline{\mu_{2(D)}} = \overline{\mu_{4(D)}} \\
& \overline{\mu_{3(D)}} = \overline{\mu_{4(D)}}
\end{aligned}
$$

On conclut que H_0 est rejetée ($F = 3.74$, seuil observé=0.0081).

Type 3 Coefficients for prod(cie)									
Effect	Compagnie	Produit	Row1	Row2	Row3	Row4	Row5	Row6	Row7
Intercept									
cie	A								
cie	B								
cie	C								
cie	D								
prod(cie)	A	1	1						
prod(cie)	A	2		1					
prod(cie)	A	3	-1	-1					
prod(cie)	B	1			1				
prod(cie)	B	2			-1				
prod(cie)	C	1				1			
prod(cie)	C	2				-1			
prod(cie)	D	1					1		
prod(cie)	D	2						1	
prod(cie)	D	3							1
prod(cie)	D	4					-1	-1	-1

Les produits d'une même compagnie diffèrent entre eux pour au moins une com-

pagnie. On pourra effectuer des comparaisons multiples au sein de chaque compagnie, après avoir vérifié quelles compagnies génèrent des différences.

Comparaisons multiples

Le test sur le facteur emboîté produit(compagnie) étant significatif, cela veut dire qu'il existe au moins une compagnie pour laquelle les produits diffèrent. On fait donc un test pour comparer les produits au sein de chaque compagnie, à l'aide de l'option `slice` de `lsmeans`. On voit dans le tableau ci-dessous que les produits sont hétérogènes dans les compagnies C et D seulement.

```
lsmeans prod(cie)/slice=cie pdiff;
```

Tests of Effect Slices

Effect	Compagnie	Num DF	Den DF	F Value	Pr > F
prod(cie)	A	2	22	2.54	0.1019
prod(cie)	B	1	22	0.03	0.8729
prod(cie)	C	1	22	5.89	0.0238
prod(cie)	D	3	22	5.07	0.0081

Les produits diffèrent dans les compagnies C et D. Puisque la compagnie C n'a que deux produits, inutile de réaliser des tests de comparaisons deux à deux. On doit donc regarder les tests de comparaisons multiples pour les produits de la compagnie D seulement. On peut voir lesquels diffèrent les uns des autres.

Differences of Least Squares Means

Effect	Compagnie	Produit	Compagnie	Produit	Estimate	Standard Error	DF	t Value	Pr > |t|
prod(cie)	D	1	D	2	6.0000	6.1791	22	0.97	0.3421
prod(cie)	D	1	D	3	-13.6667	6.1791	22	-2.21	0.0377
prod(cie)	D	1	D	4	-13.3333	6.1791	22	-2.16	0.0421
prod(cie)	D	2	D	3	-19.6667	6.1791	22	-3.18	0.0043
prod(cie)	D	2	D	4	-19.3333	6.1791	22	-3.13	0.0049
prod(cie)	D	3	D	4	0.3333	6.1791	22	0.05	0.9575

- Comparaisons des produits de compagnies différentes

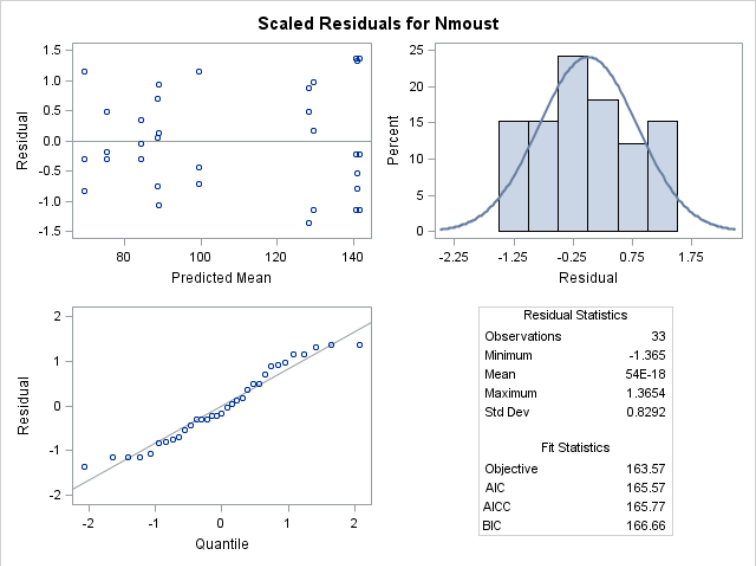
Puisque les produits sont un facteur fixe et que dès la planification de l'expérience nous avions prévu de les comparer même s'ils ne sont pas issus d'une même compagnie, il est pertinent de comparer tous les produits entre eux, en interprétant le reste du tableau `Differences of Least Squares Means`.

6.2.5 Analyse des résidus

Puisqu'aucun facteur aléatoire ne figure dans le modèle, les résidus marginaux et conditionnels sont équivalents. On peut donc utiliser l'option `outp` ou l'option `outpm` pour le calcul des résidus. Prenons la deuxième option, puisque c'est celle qui sera favorisée tout au long de notre étude sur les modèles mixtes, plus généraux.

```
proc mixed data=insect;
class cie prod;
model Nmoust= cie prod(cie) / outpm=out vciry;
run;

ods select TestsForNormality;
proc univariate data=out normal;
var ScaledResid;
run;
```

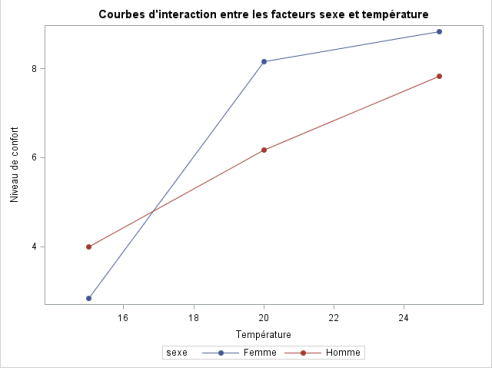


Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.948533	Pr < W	0.1205
Kolmogorov-Smirnov	D	0.099287	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.062322	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.458397	Pr > A-Sq	>0.2500

6.3 Analyse du schéma de l'exemple 3 sur la sensation de confort

Voici les observations (tirées de Milliken et Johnson [13] p. 147).

Température	Sexe	Chambre 1	Chambre 2	Chambre 3
15°C	H	5	5	4
		4	4	2
	F	1	5	1
		2	5	3
20°C	H	8	6	5
		8	3	7
	F	10	8	8
		7	8	8
25°C	H	12	8	6
		8	7	6
	F	11	8	6
		13	8	7



Rappel du modèle : une corrélation (σ_b^2) est introduite entre les observations d'une même chambre. La matrice V est bloc-diagonale avec un bloc par chambre.

$$y_{ijkl} = \mu + \tau_i + b_{j(i)} + \beta_k + \tau\beta_{ik} + \varepsilon_{ijkl} \left\{ \begin{array}{l} i = 1,2,3 \\ j(i) = 1,2,3 \ \ \forall i \\ k = 1,2 \\ l = 1,2 \end{array} \right.$$

- y_{ijkl} est une mesure du confort entre 1 et 15 (1=froid ; 8=confortable ; 15=chaud) ;
- τ_i est l'effet de la température i ;
- $b_{j(i)}$ est l'effet aléatoire de la chambre j dans la température i ;
- β_k est l'effet du sexe k (1=Homme, 2=Femme) ;
- $b_{j(i)} \sim N(0, \sigma_b^2)$ i.i.d. et $\varepsilon_{ijkl} \sim N(0, \sigma^2)$ i.i.d..

6.3.1 Spécification du modèle dans SAS

```
proc mixed data=confort covtest cl;
class   temp chambre sexe;
model   confort = temp sexe temp*sexe ;
random  chambre(temp);
run;
```

6.3.2 Estimation et tests des paramètres de covariance

Les estimations des paramètres de covariance sont calculées en maximisant la vraisemblance REML. Dans cet exemple avec données balancées ce sont aussi les estimateurs des moments déduits à partir des espérances des sommes de carrés moyens.

Avec l'option `covtest`, PROC MIXED fait le test unilatéral de Wald basé sur la normalité asymptotique des paramètres. Nous avons déjà discuté au chapitre 5 du fait que ce postulat est inadéquat pour les paramètres de covariance. Si on veut vraiment faire de l'inférence sur ces paramètres, il est préférable d'interpréter les intervalles de confiance basés sur l'approximation de Satterthwaite, obtenus avec l'option `cl`.

Covariance Parameter Estimates

Cov Parm	Estimate	Standard Error	Z Value	Pr > Z	Alpha	Lower	Upper
chambre(temp)	2.3576	1.6042	1.47	0.0708	0.05	0.8710	17.3238
Residual	1.6528	0.4771	3.46	0.0003	0.05	1.0077	3.1986

Pour évaluer l'importance des paramètres de covariance, il est aussi possible d'utiliser des tests de rapport de vraisemblances pour comparer deux modèles dont l'un est un cas particulier de l'autre. Dans le cas qui nous occupe, on pourrait comparer le modèle contenant le facteur chambre(temp) et le modèle ne le contenant pas. La différence entre les deux statistiques sera comparée à un quantile de la loi du khi-deux à un degré de liberté.

On peut également considérer les critères d'ajustement pour comparer les modèles.

```
proc mixed data=confort;
class  temp chambre sexe;
model  confort = temp sexe temp*sexe;
random chambre(temp);
run;
```

Fit Statistics	
-2 Res Log Likelihood	122.4
AIC (smaller is better)	126.4
AICC (smaller is better)	126.8
BIC (smaller is better)	126.8

```
proc mixed data=confort ;
class temp sexe;
model confort = temp sexe temp*sexe;
run;
```

Fit Statistics	
-2 Res Log Likelihood	133.8
AIC (smaller is better)	135.8
AICC (smaller is better)	135.9
BIC (smaller is better)	137.2

La statistique du rapport de vraisemblance est $U_{obs} = 133.8 - 122.4 = 11.4$. Puisqu'elle est non nulle, le seuil observé pour $H_0 : \sigma_b^2 = 0$ sera calculé ainsi :

$$0.5 \times P(\chi_1^2 > 11.4) = 0.0004.$$

Les indices d'ajustement AIC et BIC suggèrent également d'inclure un effet chambre dans le modèle.

Remarque importante : que les tests sur σ_b^2 soient significatifs ou pas, en général on laisse chambre dans le modèle, car c'est l'unité expérimentale pour le facteur température. On veut conserver une structure de corrélation où les observations provenant d'une même chambre sont dépendantes. Lorsqu'on trie les données par chambre, on peut faire afficher la matrice des variances-covariances estimées des observations avec l'option `v` de l'énoncé `random`. En voici un aperçu pour les 8 premières observations :

Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8
1	4.0104	2.3576	2.3576	2.3576				
2	2.3576	4.0104	2.3576	2.3576				
3	2.3576	2.3576	4.0104	2.3576				
4	2.3576	2.3576	2.3576	4.0104				
5					4.0104	2.3576	2.3576	2.3576
6					2.3576	4.0104	2.3576	2.3576
7					2.3576	2.3576	4.0104	2.3576
8					2.3576	2.3576	2.3576	4.0104

6.3.3 Estimation des paramètres des effets fixes

On peut toujours estimer les paramètres du vecteur β avec l'option `solution` dans l'énoncé `model`. Or, ce qui nous intéresse davantage est l'estimation des moyennes qui seront comparées par des tests. Il faut utiliser l'énoncé `lsmeans` si on veut que les erreurs-types soient calculées correctement.

```
lsmeans temp sexe temp*sexe;
```

Least Squares Means

Effect	sexe	Température	Estimate	Standard Error	DF	t Value	Pr > |t|
temp		15	3.4167	0.9610	6	3.56	0.0120
temp		20	7.1667	0.9610	6	7.46	0.0003
temp		25	8.3333	0.9610	6	8.67	0.0001
sexe	Femme		6.6111	0.5948	24	11.11	<.0001
sexe	Homme		6.0000	0.5948	24	10.09	<.0001
temp*sexe	Femme	15	2.8333	1.0302	24	2.75	0.0111
temp*sexe	Homme	15	4.0000	1.0302	24	3.88	0.0007
temp*sexe	Femme	20	8.1667	1.0302	24	7.93	<.0001
temp*sexe	Homme	20	6.1667	1.0302	24	5.99	<.0001
temp*sexe	Femme	25	8.8333	1.0302	24	8.57	<.0001
temp*sexe	Homme	25	7.8333	1.0302	24	7.60	<.0001

6.3.4 Tests sur les facteurs fixes

La matrice L qui permet de tester l'hypothèse $H_0 : L\beta = 0$ est affichée ci-dessous pour le test sur l'interaction. Les paramètres sur les effets aléatoires ne sont pas impliqués dans les hypothèses sur les effets fixes, à moins qu'on ne le fasse explicitement dans un énoncé `estimate`. Les tests sont faits à partir de la statistique suivante :

$$F = \frac{\hat{\beta}^T L^T [L(X^T \hat{V}^{-1} X)^{-1} L^T]^{-1} L\hat{\beta}}{rang(L)},$$

où $\hat{V}$ est une estimation de la matrice de variances-covariances du vecteur des données.

Type 3 Coefficients for temp*sexe

Effect	sexe	Température	Row1	Row2
Intercept				
temp		15		
temp		20		
temp		25		
sexe	Femme			
sexe	Homme			
temp*sexe	Femme	15	1	
temp*sexe	Homme	15	-1	
temp*sexe	Femme	20		1
temp*sexe	Homme	20		-1
temp*sexe	Femme	25	-1	-1
temp*sexe	Homme	25	1	1

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
temp	2	6	7.15	0.0259
sexe	1	24	2.03	0.1667
temp*sexe	2	24	4.76	0.0182

Il va de soi que les tests sur les facteurs simples peuvent être interprétés seulement si le test sur l'interaction est non significatif. Dans le cas qui nous occupe, il faudra utiliser l'option `slice` de l'énoncé `lsmeans` et les comparaisons multiples pour comparer le niveau de confort selon les températures et le sexe, car le test sur `temp*sexe` a un seuil observé de 0.0182. Il faut fixer les niveaux d'un facteur pour tester l'autre.

- Étude du facteur température

```
lsmeans temp*sexe/ slice=sexe pdiff;
```

Tests of Effect Slices					
Effect	sexe	Num DF	Den DF	F Value	Pr > F
temp*sexe	Femme	2	24	10.19	0.0006
temp*sexe	Homme	2	24	3.48	0.0471

Que concluez-vous ? Quelles lignes du tableau `Differences of Least Squares Means` interpréterez-vous ?

Differences of Least Squares Means									
Effect	sexe	Température	_sexe	Température	Estimate	Standard Error	DF	t Value	Pr > |t|
temp*sexe	Femme	15	Homme	15	-1.1667	0.7422	24	-1.57	0.1291
temp*sexe	Femme	15	Femme	20	-5.3333	1.4569	24	-3.66	0.0012
temp*sexe	Femme	15	Homme	20	-3.3333	1.4569	24	-2.29	0.0313
temp*sexe	Femme	15	Femme	25	-6.0000	1.4569	24	-4.12	0.0004
temp*sexe	Femme	15	Homme	25	-5.0000	1.4569	24	-3.43	0.0022
temp*sexe	Homme	15	Femme	20	-4.1667	1.4569	24	-2.86	0.0086
temp*sexe	Homme	15	Homme	20	-2.1667	1.4569	24	-1.49	0.1500
temp*sexe	Homme	15	Femme	25	-4.8333	1.4569	24	-3.32	0.0029
temp*sexe	Homme	15	Homme	25	-3.8333	1.4569	24	-2.63	0.0146
temp*sexe	Femme	20	Homme	20	2.0000	0.7422	24	2.69	0.0127
temp*sexe	Femme	20	Femme	25	-0.6667	1.4569	24	-0.46	0.6514
temp*sexe	Femme	20	Homme	25	0.3333	1.4569	24	0.23	0.8210
temp*sexe	Homme	20	Femme	25	-2.6667	1.4569	24	-1.83	0.0797
temp*sexe	Homme	20	Homme	25	-1.6667	1.4569	24	-1.14	0.2639
temp*sexe	Femme	25	Homme	25	1.0000	0.7422	24	1.35	0.1905

- Conclusions chez les hommes ?

Température :	$15°C$	$20°C$	$25°C$
Confort moyen :	4.00	6.17	7.83

- Conclusions chez les femmes ?

Température :	$15°C$	$20°C$	$25°C$
Confort moyen :	2.83	8.17	8.83

- Étude du facteur sexe

```
lsmeans temp*sexe / slice=temp pdiff;
```

Tests of Effect Slices					
Effect	Température	Num DF	Den DF	F Value	Pr > F
temp*sexe	15	1	24	2.47	0.1291
temp*sexe	20	1	24	7.26	0.0127
temp*sexe	25	1	24	1.82	0.1905

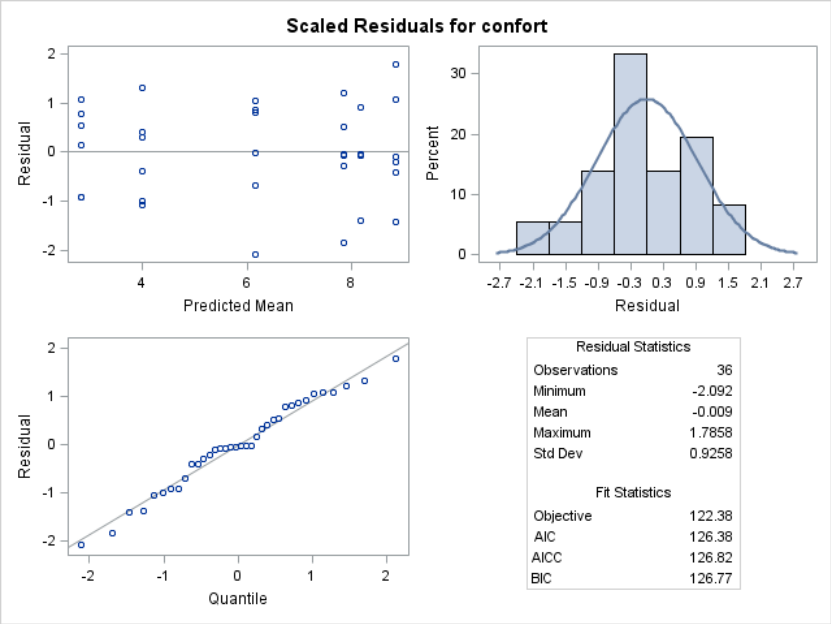
Que concluez-vous ?

6.3.5 Analyse des résidus

Dans SAS la commande `outp=` produit des résidus du type $r_i = y_i - X_i\hat{\beta} - Z_i\hat{\gamma}$.Les effets aléatoires sont soustraits, et on dit que ces résidus sont *conditionnels*.

La commande `outpm=` donne quant à elle les résidus $r_i = y_i - X_i\hat{\beta}$ dont la matrice de variance-covariance est approximativement V. L'option `vciry` fait en sorte que le vecteur des résidus soit prémultiplié par $\hat{V}^{-1/2}$, l'inverse de la racine carrée de la matrice $\hat{V}$ obtenue avec la décomposition de Cholesky. Cette transformation produit les résidus `ScaledResid` approximativement indépendants de loi $N(0,1)$.

```
proc mixed data=confort;
class  temp chambre sexe;
model  confort= temp sexe temp*sexe / outpm=out vciry;
random chambre(temp);
run;
```



```
ods select TestsForNormality;
proc univariate data=out normal;
var ScaledResid;
run;
```

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.976919	Pr < W	0.6408
Kolmogorov-Smirnov	D	0.09842	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.057006	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.329624	Pr > A-Sq	>0.2500

6.4 Exercices

1. Vrai ou Faux ?

(a) Si le facteur A est emboîté dans B, on utilisera la notation $B(A)$.

(b) Si le facteur A est emboîté dans B, alors l'interaction $A * B$ doit absolument faire partie du modèle.

(c) Un facteur emboîté dans un facteur fixe est en général aléatoire.

(d) Un facteur emboîté dans un facteur aléatoire est en général aléatoire.

2. Dans chacune des situations ci-dessous, déterminez les facteurs en cause, leur statut fixe ou aléatoire, les degrés de liberté qui y sont associés (au numérateur), et les relations qui les unissent (croisés ou emboîtés).

(a) Un fabriquant de fer forgé subit souvent des pertes matérielles en raison de la présence d'impuretés dans les solutions de trempage pré-galvanisation, et souhaite déterminer la source principale de ces impuretés. Ces liquides proviennent de deux fournisseurs différents. Pour chacun d'eux, il choisira 4 barils de liquide au hasard dans lesquels il prélèvera 3 échantillons différents pour évaluer la quantité d'impuretés.

(b) On veut tester l'effet de la caféine selon le sexe dans une étude à double insu. Dix hommes et dix femmes boivent une tasse de café régulier, et dix hommes et dix femmes boivent un tasse de café décaféiné. Aucun d'entre eux ne connaît la sorte de café qui lui est assignée. Un observateur externe mesure au cours de l'heure qui suit certains effets physiques caractéristiques de l'éveil et de l'excitation. (Idéalement, on aurait pris des mesures avant l'expérience pour que chaque participant soit son propre contrôle. Ce n'est pas fait ici.).

(c) On veut vérifier l'efficacité d'un test sur la salive permettant de détecter trois doses d'alcool. Dans un environnement supervisé, trois groupes relativement similiaires de 5 personnes sont formés, et une des trois quantités d'alcool (ajustée selon leur poids) leur est administrée. Une demi-heure après la consommation, les individus fournissent deux échantillons de salive. Ces échantillons sont utilisés pour mesurer la présence (et la concentration) d'alcool à l'aide d'un indicateur compris entre 0 et 20. le modèle doit permettre d'isoler la variabilité entre les personnes.

(d) Dans le schéma précédent, supposez maintenant que le premier échantillon de salive est testé avec le dispositif "Alco-net" et le second échantillon est testé avec le dispositif "Drink or Drive". En plus de comparer les résultats de

l'indicateur en fonction de la dose d'alcool, on veut maintenant comparer la performance des tests et savoir si les deux performent aussi bien selon la dose.

3. On veut comparer l'efficacité de 4 stratégies de marketing (journal, télévision, radio et affiche dans la rue) pour publiciser 2 types de soda (Coke et Pepsi). Pour chaque type de soda, on engage 3 agences de publicité. Chaque agence emploie les 4 méthodes de marketing dans 4 villes différentes (une dans chaque ville). On mesure après un mois les augmentations relatives des ventes dans chaque ville.

(a) Donnez les facteurs de ce dispositif expérimental, ainsi que leurs degrés de liberté et leur nature (fixe ou aléatoire).

(b) Écrivez le modèle avec effets. Définissez clairement chaque terme du modèle et les postulats.

(c) La forme matricielle du modèle avec effets s'écrit

$$Y = X\beta + ZU + \varepsilon.$$

Donnez les dimensions de chaque matrice.

4. Une compagnie possède deux usines de fabrication de pièces de métal sur lesquelles elle souhaite évaluer la qualité de la finition de la surface. Chaque usine est équipée de quatre machines, que l'on voudra comparer. Plusieurs opérateurs travaillent sur chaque machine, mais à cause de la disposition des machines dans les usines, les opérateurs ne peuvent pas travailler sur deux machines différentes.

On choisit donc au hasard trois opérateurs sur chaque machine, et on prend deux mesures de la qualité de la finition pour chacun d'eux. On désire que le modèle tienne compte de la variabilité entre les opérateurs. Voici les observations.

Usine 1				
Opér.	M.1	M.2	M.3	M.4
1	79	92	88	36
	62	99	75	53
2	94	85	53	40
	74	79	56	56
3	46	76	46	62
	57	68	57	47

Usine 2				
Opér.	M.1	M.2	M.3	M.4
1	67	82	68	53
	52	70	66	61
2	88	90	72	42
	59	87	84	49
3	68	84	59	54
	63	79	62	57

(a) Proposez un modèle pour l'analyse de ces données. Précisez les postulats associés.

(b) **(i)** Estimez les paramètres de covariance par maximum de vraisemblance restreint.

(ii) Faites un test sur le paramètre de covariance associé à l'opérateur. Énoncez clairement les hypothèses et les conclusions tirées.

(c) Existe-t-il des différences significatives de qualité de finition moyenne entre les deux usines ?

(d) Donnez l'équation de la variance (en fonction du modèle) de la moyenne des observations de l'usine 1. Faites calculer par SAS une estimation de cette moyenne et de l'erreur-type qui lui est associée.

(e) Existe-t-il des différences significatives entre les machines d'une même usine ? Si oui, cela est-il vrai dans les deux usines ?

(f) Testez l'hypothèse que, dans l'usine 1, la qualité moyenne du fini des machines 1 et 2 ne diffère pas de la qualité moyenne des machines 3 et 4. Autrement dit, faites un test bilatéral pour $H_0 : \frac{\mu_{1(1)}+\mu_{2(1)}}{2} = \frac{\mu_{3(1)}+\mu_{4(1)}}{2}$.

(g) Exprimez la matrice de variance-covariance des 6 premières observations, i.e. des observations sur la première machine de la première usine. Déterminez sa forme générale et faites calculer par SAS les valeurs numériques estimées.

(h) Considérons maintenant que plusieurs machines sont utilisées dans chaque usine pour fabriquer le produit d'intérêt. On suppose ici que les données ci-dessus ont été obtenues en choisissant dans chaque usine quatre machines au hasard parmi toutes les machines fonctionnelles.

(i) Proposez un modèle pour l'analyse de ces données, avec le nouveau plan d'expérience.

(ii) Donnez les estimations des paramètres de covariance par la méthode du maximum de vraisemblance restreint. Ces estimations changent-elles par rapport au premier modèle ?

(iii) Donnez une estimation (ainsi que l'erreur-type associée) pour la qualité moyenne des plaques de métal dans l'usine 1.

(iv) Exprimez la matrice de variance-covariance des 6 premières observations, i.e. des observations sur la première machine de la première usine. Déterminez sa forme générale et faites calculer les valeurs numériques estimées.

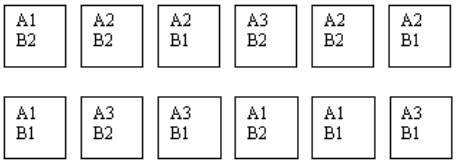
7 Plans à randomisation restreinte : split-plots et autres schémas

7.1 Restriction à la randomisation

Les plans à randomisation restreinte réfèrent à tous les plans d'expérience où les unités expérimentales ne sont pas allouées de façon complètement aléatoire dans chaque combinaison de traitements. Le plan à blocs aléatoires complets est l'exemple le plus simple d'une restriction à la randomisation : chaque bloc doit contenir toutes les combinaisons de traitements. Les figures ci-dessous présentent deux répartitions différentes des unités expérimentales dans une expérience à deux facteurs fixes :

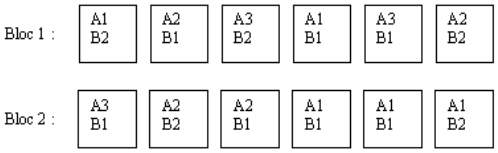
- A à a=3 modalités : A1, A2 et A3
- B à b=2 modalités : B1 et B2
- On observe n=2 répétitions de chaque combinaison de traitements.

On peut imaginer que A est une variété de céréales et que B est une dose d'azote. On applique ces traitements sur des portions de champs et on mesure le rendement dans chaque portion.



Plan complètement aléatoire

Source	Statut	DL	Exemple
A	F	a-1	2
B	F	b-1	1
A*B	F	(a-1)(b-1)	2
Erreur	A	ab(n-1)	6
Total		abn-1	11



Plan à blocs aléatoires complets

Source	Statut	DL	Exemple
Bloc	A	n-1	1
A	F	a-1	2
B	F	b-1	1
A*B	F	(a-1)(b-1)	2
Erreur	A	(ab-1)(n-1)	5
Total		abn-1	11

7.2 Split-plots, etc.

Nous étudierons dans ce document les plans avec deux ou trois restrictions à la randomisation à des niveaux différents. De plus, les différents traitements ne seront pas appliqués à des unités expérimentales de même taille. Ces plans sont appelés généralement des *split-plots*, mais on trouve aussi les appellations suivantes : *plans à parcelles partagées*, *plans en tiroirs*, *plans en unités subdivisées*.

Les grandes unités expérimentales sur lesquelles est appliqué le premier facteur sont appelées *parcelles principales* ; les petites unités expérimentales, imbriquées dans les grandes, sur lesquelles est appliqué le second facteur sont appelées *sous-parcelles*. On peut aussi avoir des *sous-sous-parcelles*, imbriquées dans les sous-parcelles, si un troisième facteur est appliqué avec une restriction à la randomisation. On parlera alors de *split-split-plot*.

Plusieurs variations sur le même thème sont possibles.

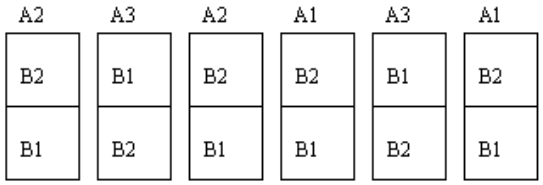
- On peut concevoir un plan où deux facteurs croisés sont appliqués de façon aléatoire aux parcelles principales et un autre facteur est appliqué aux sous-parcelles. On parlera alors d'un *split-plot avec plan complètement aléatoire à deux facteurs au premier niveau*.
- Les facteurs au premier niveau pourraient être disposés en blocs aléatoires complets, et le plan s'appellerait *split-plot avec plan à blocs aléatoires complets au premier niveau*.
- On peut penser à un plan où un facteur est appliqué de façon complètement aléatoire aux parcelles principales, puis deux facteurs croisés sont appliqués aux sous-parcelles. Il s'agira d'un *split-plot avec plan complètement aléatoire à un facteur au premier niveau et avec plan complètement aléatoire à deux facteurs au second niveau*.
- Etc.

Tous ces schémas ont un point en commun : les tests effectués sur le facteur appliqué aux sous-parcelles est plus puissant que ceux effectués sur le facteur appliqué en parcelle principale, car le nombre de répétitions est souvent plus élevé. Les degrés de liberté à l'erreur sont donc souvent plus élevés pour les sous-parcelles.

Un plan en split-plot peut être adopté soit pour accroître la précision des comparaisons entre sous-parcelles, (par rapport aux comparaisons entre parcelles principales), soit parce que certains traitements ne peuvent pas être appliqués à de petites parcelles.

Exemple 1 : Split-plot, PCA au premier niveau

On peut répartir les parcelles principales de façon complètement aléatoire dans les niveaux du facteur A, à la manière d'un PCA. Ensuite, les parcelles principales sont divisées en petites parcelles, et on attribue les niveaux de B de façon aléatoire à chaque sous-parcelle à l'intérieur d'une grande parcelle. Voici un exemple de cette randomisation à deux niveaux :



Split-plot avec PCA au premier niveau

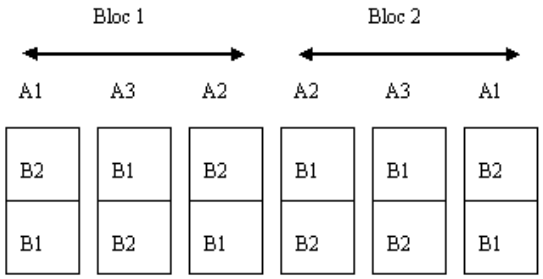
Source	Statut	DL	Exemple
A	F	a-1	2
Parcelle(A)	A	a(n-1)	3
B	F	b-1	1
A*B	F	(a-1)(b-1)	2
Sous-parcelle	A	a(b-1)(n-1)	3
Total		abn-1	11

L'exemple 3 du chapitre précédent sur les plans hiérarchisés est un split-plot avec PCA au premier niveau. L'attribution des traitements température et sexe n'est pas totalement randomisée, et le facteur température est appliqué sur une unité expérimentale plus grande (la parcelle=la chambre) que le facteur sexe (la sous-parcelle=la personne).

Si on considère le facteur température appliqué sur les chambres (le premier niveau), on voit qu'il s'agit d'un plan complètement aléatoire. Si on faisait la moyenne des observations pour chaque chambre et qu'on appliquait le modèle d'anova à un facteur fixe sur les chambres (sans le facteur sexe), le test F sur le facteur température serait exactement le même que le test F issu du split-plot complet.

Exemple 2 : Split-plot, PBAC au premier niveau

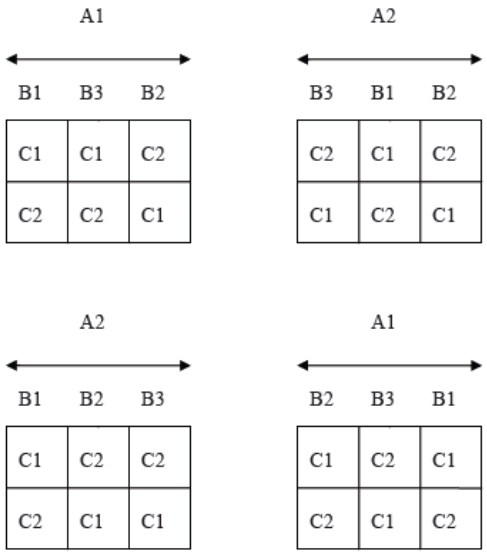
Il est aussi possible de partager les parcelles principales en blocs, et de les allouer à un niveau de A de sorte que toutes les modalités de A se retrouvent une et une seule fois dans chaque bloc. On aurait alors un plan à blocs aléatoires complets pour l'attribution des parcelles principales. Ensuite, chaque grande parcelle est subdivisée en sous-parcelles auxquelles on attribue un niveau de B. Voici un exemple de cette randomisation à deux niveaux :



Split-plot avec PBAC au premier niveau

Source	Statut	DL	Exemple
Bloc	A	k-1	1
A	F	a-1	2
Parcelle	A	(a-1)(k-1)	2
B	F	b-1	1
A*B	F	(a-1)(b-1)	2
Sous-parcelle	A	a(k-1)(b-1)	3
Total		abk-1	11

Exemple 3 : Split-split-plot, PCA au premier niveau

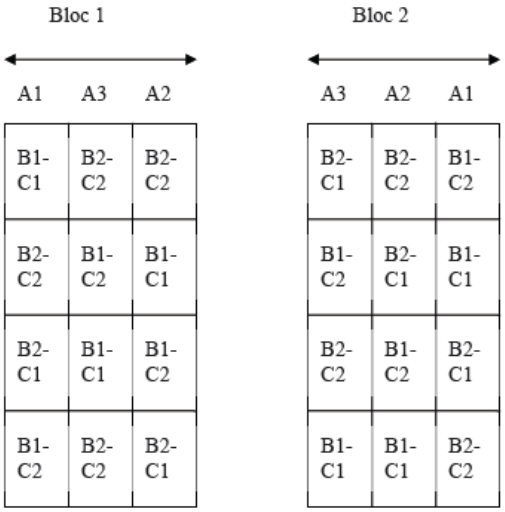


Split-split-plot avec PCA au premier niveau

Source	Statut	DL	Exemple
A	F	a-1	1
Parcelle(A)	A	a(k-1)	2
B	F	b-1	2
A*B	F	(a-1)(b-1)	2
Sous-parcelle	A	a(k-1)(b-1)	4
C	F	c-1	1
A*C	F	(a-1)(c-1)	1
B*C	F	(b-1)(c-1)	2
A*B*C	F	(a-1)(b-1)(c-1)	2
Sous-sous-parcelle	A	ab(k-1)(c-1)	6
Total		abck-1	23

k = nombre de parcelles dans chaque niveau de A = 2.

Exemple 4 : Split-plot, PBAC au premier niveau et PCA 2 facteurs au second niveau



Split -plot avec PBAC au premier niveau et PCA (2 facteurs) au second niveau

Source	Statut	DL	Exemple
Bloc	A	k-1	1
A	F	a-1	2
Parcelle	A	(a-1)(k-1)	2
B	F	b-1	1
C	F	c-1	1
B*C	F	(b-1)(c-1)	1
A*B	F	(a-1)(b-1)	2
A*C	F	(a-1)(c-1)	2
A*B*C	F	(a-1)(b-1)(c-1)	2
Sous-parcelle	A	a(k-1)(bc-1)	9
Total		abck-1	23

k = nombre de blocs = 2.

7.3 Exemple d'application dans un contexte industriel

Considérons un exemple tiré du livre de Douglas C. Montgomery [16], p. 540.

Un manufacturier de papier souhaite comparer l'effet de trois méthodes de préparation de la pulpe à papier (contenant diverses concentrations de bois dur) et de quatre températures de cuisson sur la résistance à la traction du papier. Il sait que les conditions de fabrication sont légèrement variables d'une journée à l'autre, alors il répétera son expérience sur 3 journées.

À chaque journée d'expérimentation, on fait un lot de chaque préparation. Chaque lot est séparé en quatre sous-lots qui seront cuits à 200, 225, 250 et 275 $^{\circ}C$ respectivement. Nous reproduisons les données ici.

	Jour 1			Jour 2			Jour 3		
Température	P1	P2	P3	P1	P2	P3	P1	P2	P3
200	30	34	29	28	31	31	31	35	32
225	35	41	26	32	36	30	37	40	34
250	37	38	33	40	42	32	41	39	39
275	36	42	36	41	40	40	40	44	45

On a le schéma d'analyse de variance suivant :

Facteur	Définition	Statut	DL
J	bloc jour	A	2
P	prép. appliquée à la parcelle	F	2
JP	parcelle principale = un lot	A	4
T	temp. appliquée à la sous-parcelle	F	3
TP	interaction temp.*prép.	F	6
$Erreur$	sous-parcelle = un sous-lot	A	18
Total			35

Il s'agit ici d'un split-plot à blocs aléatoires complets au premier niveau.

Le modèle s'écrirait de façon plus formelle comme suit :

$$y_{ijk} = \mu + b_j + \tau_i + b\tau_{ij} + \beta_k + \tau\beta_{ik} + \varepsilon_{ijk} \left\{ \begin{array}{l} i = 1,2,3 \\ j = 1,2,3 \\ k = 1,2,3,4 \end{array} \right. , \quad \text{où}$$

- μ est un paramètre de référence ;
- b_j est l'effet aléatoire du jour j ;
- τ_i est l'effet fixe de la préparation i ;
- $b\tau_{ij}$ est l'effet aléatoire jour*préparation, qui identifie la parcelle ;
- β_k est l'effet fixe de la température k ;
- $\tau\beta_{ik}$ est l'effet fixe préparation*température ;
- ε_{ijk} est l'erreur aléatoire de la sous-parcelle ijk ;
- $b_j \sim N(0, \sigma_b^2)$ i.i.d. ;
- $b\tau_{ij} \sim N(0, \sigma_{b\tau}^2)$ i.i.d. ;
- $\varepsilon_{ijk} \sim N(0, \sigma^2)$ i.i.d..

7.3.1 Spécification du modèle dans SAS

```
proc mixed data=papier;
class  jour prep temp;
model  resistance = prep temp temp*prep;
random jour jour*prep ;
run;
```

7.3.2 Structure de covariance

Notre modèle contient 2 facteurs aléatoires en plus de l'erreur résiduelle : le jour et l'interaction jour*préparation. Par conséquent, toutes les observations ayant été collectées le même jour sont corrélées positivement (σ_b^2), et toutes les observations venant du même jour et de la même préparation sont corrélées encore plus fortement ($\sigma_b^2 + \sigma_{b\tau}^2$). Les observations venant de jours différents sont non corrélées.

Si on ordonne les observations par jour puis par préparation, la matrice de variance-covariance des 36 observations (V) sera bloc-diagonale à deux niveaux : trois blocs 12×12 se placent sur la diagonale principale, et chacun de ces blocs est formé de 3 blocs 4×4 sur sa diagonale.

On obtient les estimations REML suivantes pour σ^2, σ_b^2 et $\sigma_{b\tau}^2$:

Covariance Parameter Estimates

Cov Parm	Estimate	Std Error	Z Value	Pr Z	Alpha	Lower	Upper
jour	2.4757	3.2754	0.76	0.2249	0.05	0.5236	1102.73
jour*prep	1.2743	1.6371	0.78	0.2182	0.05	0.2768	408.69
Residual	3.9722	1.3241	3.00	0.0013	0.05	2.2679	8.6869

Voici l'estimation d'un bloc des douze observations d'une journée (option `v` de l'énoncé `random`).

Estimated V Matrix for Subject 1

Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7	Col8	Col9	Col10	Col11	Col12
1	7.72	3.75	3.75	3.75	2.48	2.48	2.48	2.48	2.48	2.48	2.48	2.48
2	3.75	7.72	3.75	3.75	2.48	2.48	2.48	2.48	2.48	2.48	2.48	2.48
3	3.75	3.75	7.72	3.75	2.48	2.48	2.48	2.48	2.48	2.48	2.48	2.48
4	3.75	3.75	3.75	7.72	2.48	2.48	2.48	2.48	2.48	2.48	2.48	2.48
5	2.48	2.48	2.48	2.48	7.72	3.75	3.75	3.75	2.48	2.48	2.48	2.48
6	2.48	2.48	2.48	2.48	3.75	7.72	3.75	3.75	2.48	2.48	2.48	2.48
7	2.48	2.48	2.48	2.48	3.75	3.75	7.72	3.75	2.48	2.48	2.48	2.48
8	2.48	2.48	2.48	2.48	3.75	3.75	3.75	7.72	2.48	2.48	2.48	2.48
9	2.48	2.48	2.48	2.48	2.48	2.48	2.48	2.48	7.72	3.75	3.75	3.75
10	2.48	2.48	2.48	2.48	2.48	2.48	2.48	2.48	3.75	7.72	3.75	3.75
11	2.48	2.48	2.48	2.48	2.48	2.48	2.48	2.48	3.75	3.75	7.72	3.75
12	2.48	2.48	2.48	2.48	2.48	2.48	2.48	2.48	3.75	3.75	3.75	7.72

Tests du rapport de vraisemblances

On peut réaliser un test du rapport des vraisemblances en ajustant un modèle sans le paramètre testé et en comparant la différence des -2*log-vraisemblances avec le quantile d'une loi de χ^2_1.

```
model resistance = prep  temp  temp*prep;
random jour  jour*prep ;
                    Fit Statistics
         -2 Res Log Likelihood           122.3
         AIC (smaller is better)         128.3
         AICC (smaller is better)        129.5
         BIC (smaller is better)         125.6


model resistance = prep  temp  temp*prep;
random jour ;
                    Fit Statistics
         -2 Res Log Likelihood           123.6
         AIC (smaller is better)         127.6
         AICC (smaller is better)        128.1
         BIC (smaller is better)         125.8


model resistance= prep  temp  temp*prep;
random  jour*prep ;
                   Fit Statistics
         -2 Res Log Likelihood           123.8
         AIC (smaller is better)         127.8
         AICC (smaller is better)        128.3
         BIC (smaller is better)         128.2
```

Que concluez-vous ?

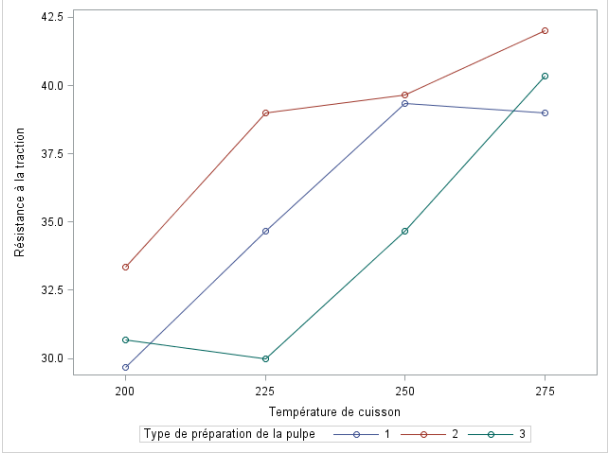
(Rappelons que ces tests nous indiquent la part de variabilité de chaque facteur aléatoire dans le modèle. Il n'est pas recommandé de supprimer des sources de variabilité, qu'elles soient significatives ou non, car on veut conserver la structure du plan d'expérience.)

7.3.3 Estimation des paramètres des effets fixes

On peut estimer les paramètres du vecteur β avec l'option `solution` dans l'énoncé `model`. Or, ce qui nous intéresse davantage est l'estimation des moyennes, obtenues avec l'énoncé `lsmeans`. Les erreurs-types tiennent compte de la présence d'effets aléatoires dans le modèle.

```
lsmeans temp prep temp*prep;
```

Effect	prep	temp	Estimate	Std Error	DF	t Value	Pr > |t|
temp		200	31.2222	1.1867	18	26.31	<.0001
temp		225	34.5556	1.1867	18	29.12	<.0001
temp		250	37.8889	1.1867	18	31.93	<.0001
temp		275	40.4444	1.1867	18	34.08	<.0001
prep	1		35.6667	1.2574	4	28.37	<.0001
prep	2		38.5000	1.2574	4	30.62	<.0001
prep	3		33.9167	1.2574	4	26.97	<.0001
prep*temp	1	200	29.6667	1.6044	18	18.49	<.0001
prep*temp	1	225	34.6667	1.6044	18	21.61	<.0001
prep*temp	1	250	39.3333	1.6044	18	24.52	<.0001
prep*temp	1	275	39.0000	1.6044	18	24.31	<.0001
prep*temp	2	200	33.3333	1.6044	18	20.78	<.0001
prep*temp	2	225	39.0000	1.6044	18	24.31	<.0001
prep*temp	2	250	39.6667	1.6044	18	24.72	<.0001
prep*temp	2	275	42.0000	1.6044	18	26.18	<.0001
prep*temp	3	200	30.6667	1.6044	18	19.11	<.0001
prep*temp	3	225	30.0000	1.6044	18	18.70	<.0001
prep*temp	3	250	34.6667	1.6044	18	21.61	<.0001
prep*temp	3	275	40.3333	1.6044	18	25.14	<.0001



7.3.4 Tests sur les facteurs fixes

La matrice L' qui permet de tester l'hypothèse $H_0 : L\beta = 0$ est affichée ci-dessous pour le test sur l'interaction. Les paramètres sur les effets aléatoires ne sont pas impliqués dans les hypothèses sur les effets fixes, à moins qu'on ne le fasse explicitement dans un énoncé `estimate`. Les tests sont faits à partir de la statistique suivante :

$$F = \frac{\hat{\beta}^T L^T \{L(X^T \hat{V}^{-1} X)^{-1} L^T\}^{-1} L\hat{\beta}}{rang(L)}$$

Type 3 Coefficients for prep*temp

Effect	prep	temp	Row1	Row2	Row3	Row4	Row5	Row6
Intercept								
prep	1							
prep	2							
prep	3							
temp		200						
temp		225						
temp		250						
temp		275						
prep*temp	1	200	1					
prep*temp	1	225		1				
prep*temp	1	250			1			
prep*temp	1	275	-1	-1	-1			
prep*temp	2	200				1		
prep*temp	2	225					1	
prep*temp	2	250						1
prep*temp	2	275				-1	-1	-1
prep*temp	3	200	-1			-1		
prep*temp	3	225		-1			-1	
prep*temp	3	250			-1			-1
prep*temp	3	275	1	1	1	1	1	1

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
prep	2	4	7.08	0.0485
temp	3	18	36.43	<.0001
prep*temp	6	18	3.15	0.0271

7.3.5 Comparaisons multiples

Puisque l'interaction `prep*temp` est significative, il faut fixer les niveaux d'un facteur pour tester l'autre. Cela se fait avec l'option `slice` de l'énoncé `lsmeans`.

Étude du facteur température

```
lsmeans temp*prep/slice=prep pdiff;
```

Tests of Effect Slices

Effect	prep	temp	Num DF	Den DF	F Value	Pr > F
prep*temp	1		3	18	15.50	<.0001
prep*temp	2		3	18	10.21	0.0004
prep*temp	3		3	18	17.03	<.0001

On voit que l'effet de la température est significatif pour toutes les préparations.

- Pour la préparation 1, les températures $250°C$ et $275°C$ créent une résistance similaire :

Differences of Least Squares Means

Effect	prep	temp	_prep	_temp	Estimate	Std Err	DF	t Value	Pr > |t|
prep*temp	1	200	1	225	-5.000	1.627	18	-3.07	0.0066
prep*temp	1	200	1	250	-9.667	1.627	18	-5.94	<.0001
prep*temp	1	200	1	275	-9.333	1.627	18	-5.74	<.0001
prep*temp	1	225	1	250	-4.667	1.627	18	-2.87	0.0102
prep*temp	1	225	1	275	-4.333	1.627	18	-2.66	0.0159
prep*temp	1	250	1	275	0.333	1.627	18	0.20	0.8400

Température :	$200°C$	$225°C$	$275°C$	$250°C$
Résistance moyenne :	29.7	34.7	39.0	39.3

- Pour la préparation 2, les températures $225°C$, $250°C$ et $275°C$ créent une résistance similaire :

Effect	prep	temp	_prep	_temp	Estimate	Std Err	DF	t Value	Pr > |t|
prep*temp	2	200	2	225	-5.667	1.627	18	-3.48	0.0027
prep*temp	2	200	2	250	-6.333	1.627	18	-3.89	0.0011
prep*temp	2	200	2	275	-8.667	1.627	18	-5.33	<.0001
prep*temp	2	225	2	250	-0.667	1.627	18	-0.41	0.6869
prep*temp	2	225	2	275	-3.000	1.627	18	-1.84	0.0818
prep*temp	2	250	2	275	-2.333	1.627	18	-1.43	0.1688

Température :	$200°C$	$225°C$	$250°C$	$275°C$
Résistance moyenne :	33.3	39.0	39.7	42.0

- Pour la préparation 3, les températures $200°C$ et $225°C$ créent une résistance similaire :

Effect	prep	temp	_prep	_temp	Estimate	Std Err	DF	t Value	Pr > |t|
prep*temp	3	200	3	225	0.667	1.627	18	0.41	0.6869
prep*temp	3	200	3	250	-4.000	1.627	18	-2.46	0.0243
prep*temp	3	200	3	275	-9.667	1.627	18	-5.94	<.0001
prep*temp	3	225	3	250	-4.667	1.627	18	-2.87	0.0102
prep*temp	3	225	3	275	-10.333	1.627	18	-6.35	<.0001
prep*temp	3	250	3	275	-5.667	1.627	18	-3.48	0.0027

Température :	$225°C$	$200°C$	$250°C$	$275°C$
Résistance moyenne :	30.0	30.7	34.7	40.3

Étude du facteur préparation

```
lsmeans temp*prep/slice=temp pdiff;
```

Tests of Effect Slices

Effect	prep	temp	Num DF	Den DF	F Value	Pr > F
prep*temp		200	2	18	2.05	0.1572
prep*temp		225	2	18	11.58	0.0006
prep*temp		250	2	18	4.47	0.0266
prep*temp		275	2	18	1.29	0.2990

On voit que les préparations diffèrent pour les températures $225°C$ et $250°C$ seulement.

- À $225°C$, toutes les préparations ont une résistance statistiquement différente :

Differences of Least Squares Means

Effect	prep	temp	_prep	_temp	Estimate	Std Err	DF	t Value	Pr > |t|
prep*temp	1	225	2	225	-4.333	1.870	18	-2.32	0.0325
prep*temp	1	225	3	225	4.667	1.870	18	2.50	0.0225
prep*temp	2	225	3	225	9.000	1.870	18	4.81	0.0001

Préparation :	3	1	2
Résistance moyenne :	30.0	34.7	39.0

- À $250°C$, les préparations 1 et 2 créent une résistance similaire :

Differences of Least Squares Means

Effect	prep	temp	_prep	_temp	Estimate	Std Err	DF	t Value	Pr > |t|
prep*temp	1	250	2	250	-0.333	1.870	18	-0.18	0.8605
prep*temp	1	250	3	250	4.667	1.870	18	2.50	0.0225
prep*temp	2	250	3	250	5.000	1.870	18	2.67	0.0155

Préparation :	3	1	2
Résistance moyenne :	34.7	39.3	39.7

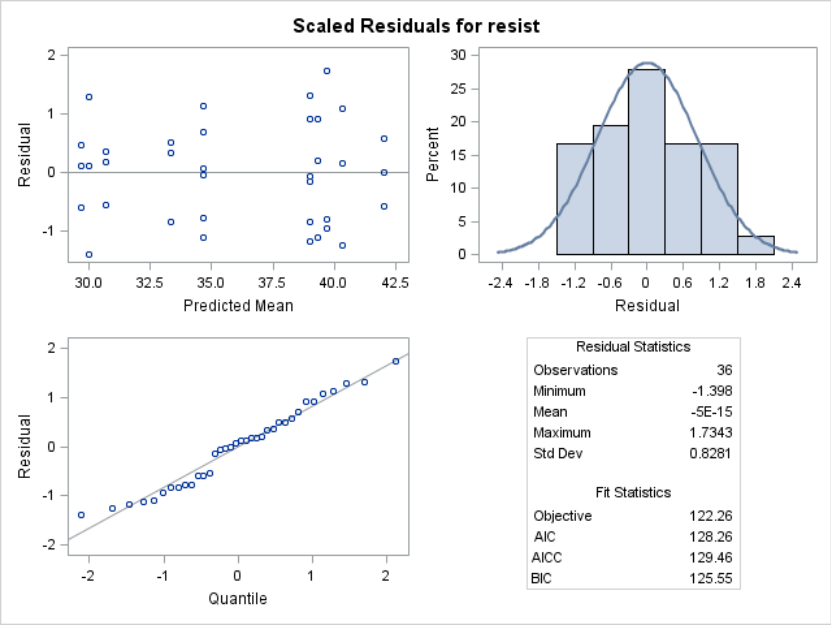
On peut aussi vouloir comparer deux à deux les résistances moyennes de diverses combinaisons de température et de préparation. On utilise alors les autres lignes du tableau `Differences of Least Squares Means`.

7.3.6 Analyse des résidus

On utilise les résidus normés (*scaled residuals*) pour faire l'analyse des résidus. La normalité est respectée, et on n'observe pas de patron particulier dans le graphique des résidus en fonction des valeurs prédites.

Tests for Normality

Test		--Statistic---		-----p Value------
Shapiro-Wilk	W	0.967027	Pr < W	0.3500
Kolmogorov-Smirnov	D	0.108199	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.060483	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.387538	Pr > A-Sq	>0.2500



7.4 Exercices

1. Vrai ou Faux ?

(a) Le terme split-plot réfère au fait que les parcelles sont séparées dans des blocs aléatoires.

(b) Dans un split-plot, le test sur le facteur appliqué à la parcelle est en général plus précis que le test sur le facteur appliqué à la sous-parcelle.

(c) Dans un split-plot, on peut appliquer le facteur A sur les parcelles de façon complètement aléatoire.

(d) Dans un split-plot, on peut appliquer le facteur B sur les sous-parcelles de façon complètement aléatoire.

(e) Un split-plot à deux facteurs fixes contient en général deux ou trois paramètres de covariance.

2. Une étude veut évaluer la performance des élèves du secondaire en mathématiques et en français selon le programme d'étude (régulier, sport-étude et international). Cinq écoles offrant les 3 programmes sont recrutées pour l'étude. Trois étudiants de chaque école participante, un par programme d'étude, sont sélectionnés au hasard parmi les finissants. Chaque étudiant sélectionné passe deux examens (un en français et un en mathématiques) dans un ordre aléatoire. Les données sont les notes des 15 étudiants recrutés aux deux examens.

(a) Donnez les facteurs, leurs modalités, leur nature (fixe ou aléatoire) et leur relation (croisé ou emboité).

(b) Nommer le schéma expérimental utilisé pour planifier cette expérience et déterminez les degrés de liberté utilisés pour tester les effets de chaque composante fixe (au dénominateur, estimés par la méthode "containment" ou d'inclusion).

(c) Écrivez le modèle mixte pour les notes obtenues (associer un indice, i , j ou k, à chaque facteur de l'expérience et écrire un modèle pour y en fonction d'effets fixes et aléatoires).

(d) Écrivez les énoncés de la procédure MIXED qui vous permettraient d'ajuster de modèle aux données.

3. Un fabriquant de patins veut comparer l'usure (y) des lames de patins fabriquées avec trois sortes d'acier (1, 2 et 3). Il réalise une expérience en recrutant des joueurs de hockey : 4 joueurs d'avant, 4 défenseurs et 4 gardiens de but. Il donne à chaque joueur trois paires de patins (une pour chaque sorte d'acier) et des instructions à

suivre au cours des trois prochains mois (par exemple un joueur doit utiliser l'acier 2 le premier mois, l'acier 3 le deuxième mois et l'acier 1 le troisième mois) qui assurent la randomisation des traitements. Chaque joueur contribue pour 3 mesures d'usure à l'expérience, une pour chaque sorte d'acier.

(a) Nommez ce schéma expérimental. Donner les facteurs et leurs modalités, leurs natures (fixe ou aléatoire) et leurs relations (croisés ou emboités).

(b) Dressez le tableau des facteurs pour cette expérience, et donnez les degrés de liberté pour chaque composante (au dénominateur, estimés par la méthode "containment" ou d'inclusion).

(c) Comment aurait-il fallu planifier cette expérience pour pouvoir l'analyser selon un schéma complètement randomisé à deux facteurs ?

(d) Refaites (a) et (b) si parmi les quatre joueurs recrutés à chaque position on retrouve deux adolescents et deux adultes. Incluez le facteur « Âge » du joueur dans le modèle.

4. On compare le rendement (en tonnes par acre) de 3 variétés de luzerne en fonction de la date de récolte. On dispose de 6 champs, qui sont séparés en trois parcelles qui recevront chacune une des 3 variétés de luzerne. Chaque parcelle est séparée en quatre sous-parcelles, dans lesquelles la récolte sera effectuée à 4 dates différentes. On prend une mesure de rendement par sous-parcelle. Voici le tableau des rendements moyens par combinaison de traitements.

Variété/Date	20 août	5 septembre	20 septembre	5 octobre	Moyenne $\overline{y}_{i\bullet\bullet}$
Cossack	1.77	1.64	1.58	1.30	1.57
Ladak	1.88	1.82	1.66	1.31	1.67
Ranger	1.70	1.61	1.48	1.41	1.55
Moyenne $\overline{y}_{\bullet\bullet k}$	1.78	1.34	1.57	1.69	1.60

(a) Proposez un modèle correspondant à cette expérience.

(b) Écrivez les énoncés de la procédure MIXED permettant d'ajuster votre modèle.

(c) Voici les résultats des tests sur les effets fixes. Que pouvez-vous en déduire ?

```
    Type 3 Tests of Fixed Effects
Effect    Num DF  Den DF  F Value   Pr > F
Var            2      10     0.65   0.5412
Date           3      45    23.39   <.0001
Var*Date       6      45     1.25   0.2973
```

Differences of Least Squares Means

Effect	Var	_Var	Estimate	Std Error	DF	t Value	Pr > |t|
Var	Cossack	Ladak	-0.09458	0.1065	10	-0.89	0.3956
Var	Cossack	Ranger	0.01917	0.1065	10	0.18	0.8608
Var	Ladak	Ranger	0.1137	0.1065	10	1.07	0.3108

Differences of Least Squares Means

Effect	Date	_Date	Estimate	Std Error	DF	t Value	Pr > |t|
Date	Aout20	Oct5	0.09000	0.05575	45	1.61	0.1134
Date	Aout20	Sept20	0.2067	0.05575	45	3.71	0.0006
Date	Aout20	Sept5	0.4406	0.05575	45	7.90	<.0001
Date	Oct5	Sept20	0.1167	0.05575	45	2.09	0.0420
Date	Oct5	Sept5	0.3506	0.05575	45	6.29	<.0001
Date	Sept20	Sept5	0.2339	0.05575	45	4.20	0.0001

(d) Donnez l'équation de la variance (en fonction des composantes du modèle) de la moyenne des observations de la variété Ladak. Pouvez-vous estimer cette variance en vous basant sur l'estimation des composantes de variance du tableau ci-dessous ?

Covariance Parameter Estimates

Cov Parm	Estimate
Bloc	0.05781
Var*Bloc	0.02707
Residual	0.02797

8 L'analyse de variance à mesures répétées

En médecine, en psychologie, en biologie, il est parfois utile de mesurer les fluctuations d'une caractéristique d'intérêt pendant une certaine période de temps. Les sujets, qui peuvent être séparés en groupes soumis à divers traitements, sont donc étudiés à plusieurs reprises au cours d'une semaine, d'une année ou d'une décennie. Ce genre d'expérience donne lieu à ce que l'on appelle des *données longitudinales*, ou des $mesures\ répétées$. Le but de ces études est de déterminer si les traitements administrés aux unités expérimentales ont une influence sur la variable mesurée, et si cette dernière évolue dans le temps ou reste constante. Comme dans toute analyse statistique, le défi consiste à distinguer la véritable information des variations dues à l'unicité des individus.

L'analyse de variance standard exige une condition impossible à remplir d'emblée avec un schéma longitudinal : l'indépendance des observations. En effet, les mesures prises sur deux individus sont indépendantes, ce qui n'est pas le cas pour les mesures prises sur un même individu à des temps différents. Il faudra tenir compte de cette dépendance dans l'analyse. Un schéma où l'individu est considéré comme un bloc pourrait être envisagé, mais nous verrons qu'il existe d'autres structures de covariance parfois mieux adaptées aux données temporelles.

Un autre problème fréquemment rencontré dans l'analyse des mesures répétées est la présence de données manquantes. En médecine par exemple, les patients peuvent entrer ou sortir de l'étude à tout moment, ou encore ne pas se présenter à un test en cours d'expérience.

En revanche, le plan d'expérience à mesures répétées comporte des avantages non négligeables. Tout d'abord, le nombre de sujets nécessaires est moins élevé que si on changeait d'unité à chaque temps. Ceci fait grandement diminuer les coûts de l'entreprise, surtout lorsqu'il s'agit d'humains ou d'animaux de laboratoire. Les mesures prises sur un seul individu sont moins variables que celles prises sur des sujets différents. Il en résulte un important gain de précision. Enfin, l'équivalence des sujets par rapport au traitement (hypothèse nécessaire à l'analyse comparative) est évidente puisque chaque individu est son propre contrôle, c'est-à-dire son propre point de comparaison avec l'état initial.

Trois analyses possibles

Il existe trois manières de traiter l'analyse de variance à mesures répétées. La première considère un modèle univarié calqué sur le split-plot. Le traitement constitue le facteur principal dans lequel sont emboîtés les sujets (qui tiennent lieu de parcelle). Le temps est le facteur appliqué à la sous-parcelle (la sous-parcelle est ici une combinaison de sujet et temps). Il faut noter qu'on ne peut pas randomiser les modalités du facteur temps, ainsi les deux schémas ne sont pas équivalents. Les méthodes d'analyse du schéma split-plot s'appliquent aux mesures répétées seulement si on peut considérer que les différences entre paires d'observations prises sur un même individu ont la même variance. Cette condition porte le nom de Huynh et Feldt, et certaines structures de matrices de covariances la remplissent (dont les formes HF et CS pour *compound symmetry*) présentées en annexe. Si la condition de Huynh et Feldt n'est pas vérifiée, les tests basés sur la structure split-plot sont souvent libéraux (ils rejettent l'hypothèse nulle trop souvent). C'est pourquoi on préférera un cadre plus général à cette approche trop limitative. (Le chapitre 26 du livre de Milliken et Johnson [13] présente des analyses de modèles à mesures répétées à l'aide du schéma split-plot.)

La seconde approche utilise une modélisation multivariée, où les p mesures prises sur un individu constituent un *vecteur* d'observations. Aucune hypothèse n'est faite concernant la matrice de variances covariances des observations faites sur le même sujet. Cette matrice, de dimension $p \times p$, fait intervenir $p \times (p+1)/2$ paramètres. La méthode du rapport de vraisemblance est utilisée pour construire des tests pour évaluer l'interaction TRAITEMENT×TEMPS et pour tester si les modalités de TEMPS sont homogènes. Ils font intervenir des statistiques de Lawley-Hotelling, Pillai ou Wilks, dans des tests de MANOVA (multivariate analysis of variance). Cette analyse, appliquable par la procédure GLM, est peu puissante à cause du grand nombre de paramètres à estimer. De plus, elle nous force à éliminer tous les sujets ayant une donnée manquante ou plus, ce qui est plutôt prohibitif, car les données manquantes sont assez fréquentes lorsqu'on travaille avec des sujets vivants. Etant donné que la même matrice de variances-covariances doit être utilisée pour tous les individus, les temps d'observation doivent forcément être identiques pour tous les sujets de l'expérience. L'approche multivariée est traitée dans la section 27.2 du livre de Milliken et Johnson [13].

Enfin, la troisième possibilité est d'ajuster un modèle mixte aux données. Cette méthode comporte plusieurs avantages par rapport aux deux premières. Elle permet d'incorporer à

l'analyse des sujets ayant des données manquantes (sans estimer ces dernières), d'avoir des temps d'observation différents d'un sujet à l'autre tout en modélisant la dépendance entre les mesures prises sur le même sujet. En effet, cette dernière étape est cruciale dans l'analyse, car les tests sur les effets fixes sont grandement influencés par le choix de la structure de covariance. C'est la seule option qui permette une véritable modélisation des liens unissant les observations basée sur les données. C'est aussi la seule méthode qui sera abordée dans ce chapitre. La section 27.4 de Milliken et Johnson [13] traite de cette approche.

Exemple : concentration sanguine de potassium chez les chiens

Les résultats d'une étude vétérinaire ont été présentés dans Biometrics [7]. 36 chiens sans race particulière ont été répartis de façon aléatoire en quatre groupes de pré-traitement, avant de subir une occlusion de l'artère coronaire. Nous utiliserons les données sur 32 chiens afin d'équilibrer le plan pour simplifier la présentation.

groupe 1 : groupe témoin, aucun pré-traitement subi ;
groupe 2 : interruption du réseau nerveux 3 semaines avant l'occlusion ;
groupe 3 : interruption du réseau nerveux immédiatement avant l'occlusion ;
groupe 4 : dilatation du lit vasculaire 3 semaines avant l'occlusion.

Les expérimentateurs ont mesuré 7 fois la concentration de potassium dans le sang de leurs cobayes, soit 1, 3, 5, 7, 9, 11 et 13 minutes après l'occlusion de l'artère coronaire.

Les objectifs de cette expérience sont de déterminer si la concentration de potassium varie dans le temps après l'occlusion, et si les différents pré-traitements ont une influence sur la manière dont les chiens réagissent.

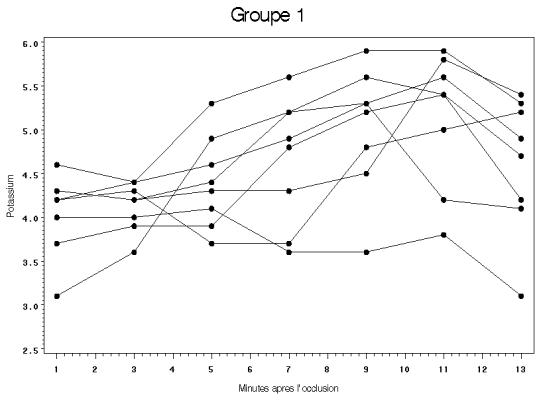
On trouve à page suivante les observations brutes et les profils individuels de la concentration de potassium chez les chiens des quatre groupes.

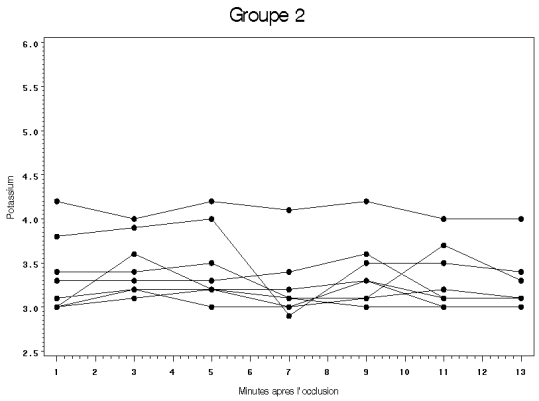
Groupe 1	Temps après l'occlusion						
Chien	1	3	5	7	9	11	13
1	4.0	4.0	4.1	3.6	3.6	3.8	3.1
2	4.2	4.3	3.7	3.7	4.8	5.0	5.2
3	4.3	4.2	4.3	4.3	4.5	5.8	5.4
4	4.2	4.4	4.6	4.9	5.3	5.6	4.9
5	4.6	4.4	5.3	5.6	5.9	5.9	5.3
6	3.1	3.6	4.9	5.2	5.3	4.2	4.1
7	3.7	3.9	3.9	4.8	5.2	5.4	4.2
8	4.3	4.2	4.4	5.2	5.6	5.4	4.7

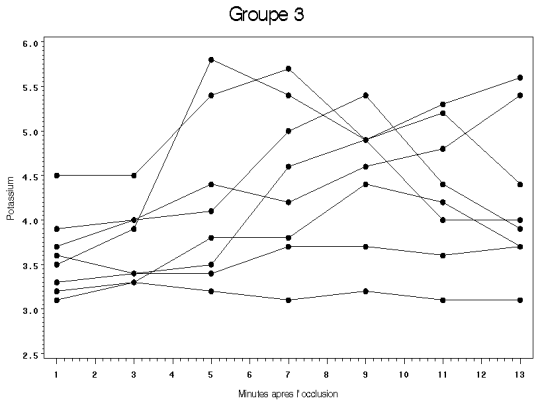
Groupe 2	Temps après l'occlusion						
Chien	1	3	5	7	9	11	13
9	3.4	3.4	3.5	3.1	3.1	3.7	3.3
10	3.0	3.2	3.0	3.0	3.1	3.2	3.1
11	3.0	3.1	3.2	3.0	3.3	3.0	3.0
12	3.1	3.2	3.2	3.2	3.3	3.1	3.1
13	3.8	3.9	4.0	2.9	3.5	3.5	3.4
14	3.0	3.6	3.2	3.1	3.0	3.0	3.0
15	3.3	3.3	3.3	3.4	3.6	3.1	3.1
16	4.2	4.0	4.2	4.1	4.2	4.0	4.0

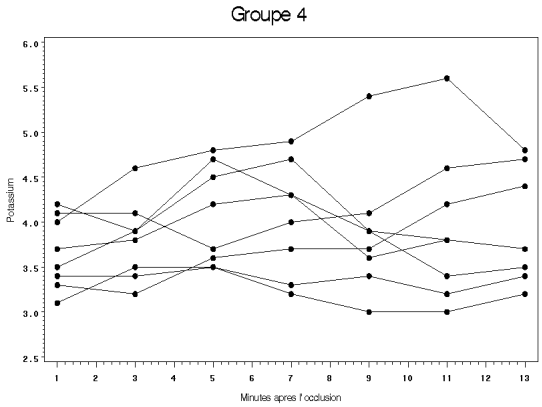
Groupe 3	Temps après l'occlusion						
Chien	1	3	5	7	9	11	13
17	3.2	3.3	3.8	3.8	4.4	4.2	3.7
18	3.3	3.4	3.4	3.7	3.7	3.6	3.7
19	3.1	3.3	3.2	3.1	3.2	3.1	3.1
20	3.6	3.4	3.5	4.6	4.9	5.2	4.4
21	4.5	4.5	5.4	5.7	4.9	4.0	4.0
22	3.7	4.0	4.4	4.2	4.6	4.8	5.4
23	3.5	3.9	5.8	5.4	4.9	5.3	5.6
24	3.9	4.0	4.1	5.0	5.4	4.4	3.9

Groupe 4	Temps après l'occlusion						
Chien	1	3	5	7	9	11	13
25	3.1	3.5	3.5	3.2	3.0	3.0	3.2
26	3.3	3.2	3.6	3.7	3.7	4.2	4.4
27	3.5	3.9	4.7	4.3	3.9	3.4	3.5
28	3.4	3.4	3.5	3.3	3.4	3.2	3.4
29	3.7	3.8	4.2	4.3	3.6	3.8	3.7
30	4.0	4.6	4.8	4.9	5.4	5.6	4.8
31	4.2	3.9	4.5	4.7	3.9	3.8	3.7
32	4.1	4.1	3.7	4.0	4.1	4.6	4.7

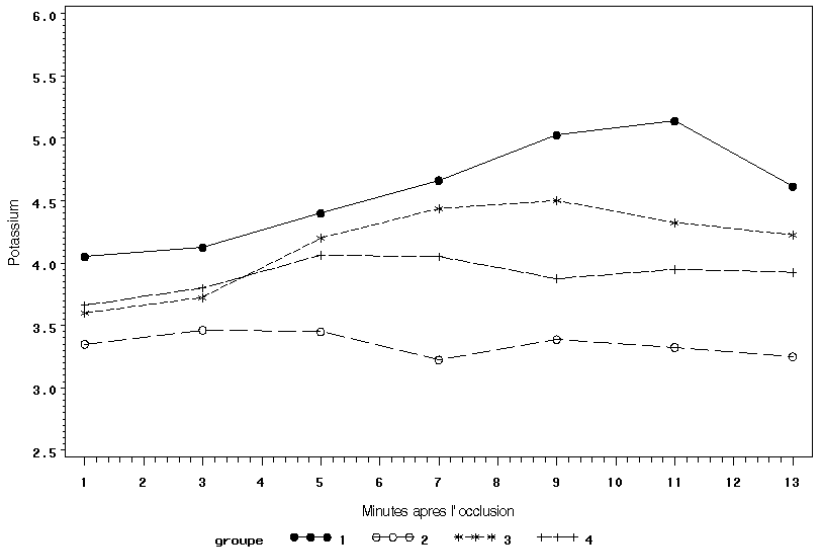








La figure ci-dessous montre le profil moyen dans chacun des groupes. L'analyse de variance nous permettra de vérifier si les différences observées sont significatives.



8.1 Procédure d'ajustement de modèle

Le modèle linéaire mixte ($Y = X\beta + ZU + \varepsilon$) permet une structure très souple pour la matrice de variance-covariance des observations ($V = ZGZ^T + R$). Pour fixer les idées, nous considérerons un modèle ayant un facteur fixe A à r modalités, n sujets subissant chaque modalité, un facteur fixe T à p niveaux répété sur les sujets (souvent le temps, mais qui pourrait aussi être une répétition spatiale sur une même parcelle), et l'interaction entre le traitement et le temps.

Les rn sujets sont en fait un facteur "aléatoire", dans le sens où on tient compte de la variabilité due à l'individu sans s'intéresser aux modalités précises de ce facteur (mais bien à une population plus large d'individus). Or, dans le logiciel SAS, le cadre particulier des mesures répétées nous permet de ne pas spécifier le sujet explicitement dans un énoncé $random$; il ne participe donc pas au terme ZU. On le spécifie plutôt dans un énoncé $repeated$, qui nous permet de spécifier une forme particulière pour la matrice R. Nous verrons les détails plus loin. Il est évidemment possible de considérer toutes sortes d'effets aléatoires dans le modèle, mais ce n'est pas le cas qui nous occupe. Nous omettrons donc la partie ZU dans le modèle, et la variance des observations deviendra $V = R$.

$$
\begin{array}{ccccccc}
Y & = & & X & & \beta & + \quad \varepsilon \\
\begin{bmatrix} y_{111} \\ y_{112} \\ . \\ y_{11p} \\ ... \\ ... \\ y_{rn1} \\ y_{rn2} \\ . \\ y_{rnp} \end{bmatrix} & = & \left[\begin{array}{c|c|c|c} 1 & 1 \; .. \; 0 & 1 & 0 \; .. \; 0 \\ 1 & 1 \; .. \; 0 & 0 & 1 \; .. \; 0 \\ . & . \; . \; . \; . & . & . \; . \; . \; . \\ 1 & 1 \; .. \; 0 & 0 & 0 \; .. \; 1 \\ .. & .. \; .. \; .. \; .. & .. & .. \; .. \; .. \; .. \\ .. & .. \; .. \; .. \; .. & .. & .. \; .. \; .. \; .. \\ 1 & 0 \; .. \; 1 & 1 & 0 \; .. \; 0 \\ 1 & 0 \; .. \; 1 & 0 & 1 \; .. \; 0 \\ . & . \; . \; . \; . & . & . \; . \; . \; . \\ 1 & 0 \; .. \; 1 & 0 & 0 \; .. \; 1 \end{array}\right] & \begin{bmatrix} \mu \\ A_1 \\ ... \\ A_r \\ T_1 \\ T_2 \\ ... \\ T_p \\ AT_{11} \\ AT_{12} \\ ... \\ AT_{rp} \end{bmatrix} & + \begin{bmatrix} \varepsilon_{111} \\ \varepsilon_{112} \\ . \\ \varepsilon_{11p} \\ ... \\ ... \\ \varepsilon_{rn1} \\ \varepsilon_{rn2} \\ . \\ \varepsilon_{rnp} \end{bmatrix}
\end{array}
$$

On suppose que la matrice R, correspondant à la fois à $Var(Y)$ et à $Var(\varepsilon)$, a une forme bloc-diagonale, avec un bloc de covariances $[R_{ij}]$ pour les p mesures prises sur le j^e sujet du groupe i :

$$R = \begin{bmatrix} [R_{11}] & & & 0 \\ & [R_{12}] & & \\ & & \cdots & \\ 0 & & & [R_{rn}] \end{bmatrix}$$

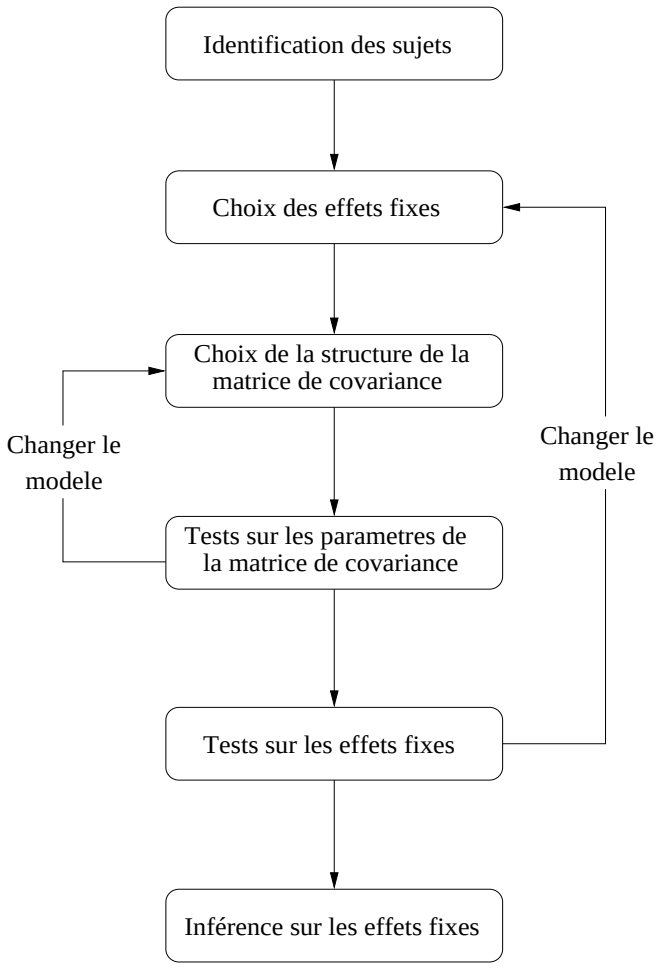
Les blocs R_{11} à R_{rn}, de dimension $p \times p$, ont tous la même structure. Les paramètres de covariance qui les composent peuvent être identiques, ou varier selon le groupe de traitement, par exemple. On peut modéliser toute structure faisant partie de la liste présentée en annexe. Le modèle s'ajuste donc mieux aux données que lorsque la variance des observations a une structure rigide ne pouvant être modifiée.

Exercice : Dans l'exemple sur les chiens, quelles sont les dimensions des matrices du modèle ?

Étapes de l'analyse

Les étapes à suivre pour ajuster un modèle approprié ressemblent plus souvent à une série d'itérations qu'à une suite unidirectionnelle de vérifications. La figure de la page suivante illustre bien le cheminement approprié. Elle est tirée de l'article de Wolfinger et Chang [28].

Une fois le sujet identifié, il faut sélectionner les effets fixes à évaluer, ainsi que leurs interactions. L'étape suivante consiste à choisir les structures des matrices de covariances des facteurs aléatoires et du facteur intra-sujets, notées respectivement G et R. Pour ce faire, on peut mettre à contribution les connaissances *a priori* fournies par la théorie scientifique, considérer les matrices échantillonnales et les graphiques de résidus. Ce choix peut ensuite être validé par des techniques plus formelles, tels des tests du rapport des vraisemblances pour comparer des structures emboîtées, et des critères d'information basés sur les valeurs de la vraisemblance permettant de comparer plusieurs structures de covariance pour des modèles contenant les mêmes facteurs fixes.



Une fois la structure de covariance fixée, on regarde les tests sur les effets fixes. PROC MIXED calcule une statistique F pour chacun des effets fixes, ainsi que le seuil observé qui lui est associé. On peut procéder à des inférences plus spécifiques sur les effets fixes à l'aide de contrastes ou de combinaisons linéaires des modalités des effets à estimer. Attention : les degrés de liberté sont estimés ; plusieurs méthodes d'estimation sont possibles.

8.2 Le programme SAS

Puisqu'aucun facteur aléatoire autre que le sujet n'est considéré, la description du modèle se fait via les énoncés *model* et *repeated*.

L'énoncé `model` identifie la variable réponse et les variables représentant les effets fixes, dont les composantes se trouvent dans le vecteur β. SAS construit ainsi la matrice X.

L'énoncé `repeated` sert à identifier les sujets sur lesquels sont répétées les mesures de la variable réponse, ainsi que la forme des blocs de R. Par exemple, l'ajustement du modèle avec une covariance de type $compound\ symmetry$ (cs) peut se faire à l'aide des commandes suivantes :

```
proc mixed data=chien;
class    groupe temps chien;
model    potassium = groupe  temps  groupe*temps;
repeated temps / subject = chien(groupe)  type = cs;
```

Les options de l'instruction `repeated` sont cruciales dans la définition du modèle ajusté. Les options `subject=`, `group=` et `type=` méritent une attention particulière.

subject=*sujet* : désigne le facteur qui distingue les individus de l'étude. Rappelons que les observations sur deux individus différents sont considérées indépendantes. La covariance des mesures prises sur un sujet est un bloc de la matrice R.

group=*effet* : définit un facteur générant l'hétérogénéité des blocs de la matrice R. Tous les sujets ayant la même valeur du groupe ont la même matrice de covariances. Tous les blocs de R conservent néanmoins la même structure de base, mais la valeur des paramètres varie quand le groupe change. Dans l'exemple ci-dessus, on aurait pu faire varier l'estimation des paramètres dans chaque groupe de pré-traitement avec l'option `group=groupe`.

type=*structure_de_covariance* : précise la structure générale des blocs de la matrice R. Voir en annexe pour une liste des structures fréquentes.

8.3 Choix de la structure de la covariance

On pourrait être tenté de choisir la structure de covariance la plus souple, où tous les paramètres sont estimés séparément. Il faut appliquer le principe de parcimonie, sans tomber dans l'autre extrême où des paramètres importants sont négligés. Une surparamétrisation peut mener à une estimation inefficace des profils de la variable réponse, et possiblement à des erreurs-types trop grandes. Par contre, une forme trop restrictive pourrait invalider les inférences sur l'espérance de la variable réponse si la structure ne convient pas réellement aux données, par exemple en sous-estimant les erreurs- types. Une analyse avec PROC MIXED se compose donc de plusieurs essais dont il faut choisir le plus pertinent.

Le choix d'une structure peut se faire de différentes façons. Tout d'abord, il est bon d'avoir une idée *a priori* de la covariance théorique. On peut regarder les matrices de covariances échantillonnales et noter les tendances qui s'y dessinent. Or, des critères formels existent afin d'éclairer notre décision.

Regardons l'allure de la matrice des covariances échantillonnales, tous chiens confondus.

```
proc corr data=chienmulti cov ;
var t1 t3 t5 t7 t9 t11 t13;
run;
```

Covariance Matrix

	t1	t3	t5	t7	t9	t11	t13
t1	0.24	0.19	0.21	0.26	0.27	0.31	0.25
t3	0.19	0.19	0.22	0.25	0.26	0.28	0.23
t5	0.21	0.22	0.51	0.52	0.44	0.40	0.37
t7	0.26	0.25	0.52	0.73	0.67	0.59	0.47
t9	0.27	0.26	0.44	0.67	0.78	0.72	0.55
t11	0.31	0.28	0.40	0.59	0.73	0.90	0.71
t13	0.25	0.23	0.37	0.47	0.55	0.71	0.68

1) Critères d'information

Cinq critères d'information calculés par PROC MIXED (avec l'option `ic`) peuvent être utilisés comme échelle de comparaison entre des modèles ayant les mêmes effets fixes mais des structures de covariances différentes. Ces critères sont basés sur la valeur maximale de la log-vraisemblance (l_R). On optera pour les structures fournissant les plus petites valeurs de ces critères (ce qui peut parfois mener à des conclusions différentes d'un critère à l'autre). Par exemple, le critère de Schwarz pénalise davantage les modèles avec plusieurs paramètres de covariance que celui d'Akaike. Dans le tableau ci-dessous, q est le nombre de paramètres de covariance, rn est le nombre total de sujets dans l'étude, $N = nrp$ est le nombre total d'observations, et x est le rang de X. Dans notre exemple, rn=32, N = 224 et x=28.

Critère d'ajustement	**Formule**
Akaike(AIC)	$-2l_R + 2q$
Akaike corrigé (AICC)	$-2l_R + \frac{2q(N-x)}{N-x-q-1}$
Schwarz Bayesian Criterion(BIC)	$-2l_R + q\ln(rn)$
Hannan et Quinn(HQIC)	$-2l_R + 2q\ln(\ln(rn))$
Bozdogan(CAIC)	$-2l_R + q(\ln(rn)+1)$

Exemple.

Nous présentons les valeurs des deux critères les plus fréquemment utilisés (AICC et BIC), pour quelques modèles de covariance.

Modèle	**Structure**	**Nb. par.**	$-2l_R$	**AICC**	**BIC**
1	UN	28	211.1	276.8	308.1
2	UN (group=groupe)	112	-32.8	496.1	355.3
3	VC	1	421.3	423.3	424.8
4	VC (group=groupe)	4	400.5	408.7	414.4
5	CS	2	318.3	322.4	325.3
6	CS (group=groupe)	8	276.6	293.3	304.3
7	HF	8	303.4	319.4	331.1
7	CSH	8	291.5	308.3	319.3
8	CSH (group=groupe)	32	243.2	320.2	354.1
9	AR(1)	2	260.3	264.3	267.2
10	AR(1) (group=groupe)	8	237.7	254.5	265.4
11	ARH(1)	8	236.4	253.2	264.1
12	ARH(1) (group=groupe)	32	204.4	281.4	315.3

Quel modèle est-il favorisé par les critères d'information ?

2) Rapport de vraisemblance pour structures emboîtées

Pour des structures de covariance emboîtées, i.e. dont l'une est un cas particulier de l'autre, il est également possible de construire des tests du rapport de vraisemblances en soustrayant les valeurs correspondantes de -2 fois la log-vraisemblance (identifiée dans la sortie SAS par *-2 Res Log Likelihood*). On compare la différence obtenue avec une distribution du χ^2 dont les degrés de liberté correspondent à la différence du nombre de paramètres à estimer entre les deux structures (Wolfinger-Chang [28]).

Si le résultat est significatif, alors le test est en faveur du modèle le plus élaboré des deux. Si le test est non significatif, c'est que la perte d'ajustement due aux contraintes sur les paramètres en passant à une matrice plus simple n'est pas significative. La forme la moins complexe s'adapte bien aux données.

Exemple

Tous les modèles sont des cas particuliers de l'absence de structure ; le modèle UN peut donc être comparé avec tous les autres (sans regroupement) par ce test.

Modèle complexe	Modèle simple	χ^2_{obs}	Deg. lib.	p-value	Choix
UN	VC	210.2	27	.000	UN
UN	CS	107.2	26	.000	UN
UN	CSH	80.4	20	.000	UN
UN	AR(1)	49.2	26	.004	UN
UN	ARH(1)	25.3	20	.190	ARH(1)

Le modèle UN (group=groupe) peut être comparé avec tous les autres (avec regroupement).

Modèle complexe	Modèle simple	χ^2_{obs}	Deg. lib.	p-value	Choix
UN (groupe)	VC (groupe)	433.3	108	.000	UN (groupe)
UN (groupe)	CS (groupe)	309.4	104	.000	UN (groupe)
UN (groupe)	CSH (groupe)	276.0	80	.000	UN (groupe)
UN (groupe)	AR(1) (groupe)	270.5	104	.000	UN (groupe)
UN (groupe)	ARH(1) (groupe)	240.1	80	.000	UN (groupe)

Les modèles de même structure peuvent être comparés selon leur regroupement. Par exemple, on pourrait comparer les deux paires de structures suivantes :

Modèle complexe	Modèle simple	χ^2_{obs}	Deg. lib.	p-value	Choix
AR(1) (groupe)	AR(1)	22.6	6	.001	AR(1) (groupe)
ARH(1) (groupe)	ARH(1)	32.0	24	.127	ARH(1)

Les modèles de même structure peuvent être comparés selon que leurs variances sont identiques ou hétérogènes. Par exemple, les comparaisons suivantes pourraient être effectuées :

Modèle complexe	Modèle simple	χ^2_{obs}	Deg. lib.	p-value	Choix
ARH(1)	AR(1)	23.9	6	.001	ARH(1)
ARH(1) (groupe)	AR(1) (groupe)	33.3	24	.098	AR(1) (groupe)

8.3.1 Estimation ponctuelle des paramètres de covariance

Voici l'estimation des paramètres de covariances pour la forme ARH(1) :

Cov Parm	Subject	Estimate	Std Error	Z Val	Pr Z
Var(1)	chien(groupe)	0.1963	0.04865	4.03	<.0001
Var(2)	chien(groupe)	0.1640	0.04117	3.99	<.0001
Var(3)	chien(groupe)	0.4852	0.1251	3.88	<.0001
Var(4)	chien(groupe)	0.4749	0.1204	3.94	<.0001
Var(5)	chien(groupe)	0.4232	0.1055	4.01	<.0001
Var(6)	chien(groupe)	0.4842	0.1162	4.17	<.0001
Var(7)	chien(groupe)	0.4220	0.1008	4.18	<.0001
ARH(1)	chien(groupe)	0.8007	0.03985	20.10	<.0001

On obtient par conséquent l'estimation suivante de la matrice des covariances entre les 7 mesures de concentration sanguine de potassium prises sur un chien du premier groupe (groupe témoin). Les autres chiens ont la même matrice de covariances.

Estimated R Matrix for chien(groupe) 1 1

Row	Col1	Col2	Col3	Col4	Col5	Col6	Col7
1	0.1963	0.1437	0.1979	0.1568	0.1185	0.1015	0.0759
2	0.1437	0.1640	0.2259	0.1790	0.1353	0.1159	0.0866
3	0.1979	0.2259	0.4852	0.3844	0.2906	0.2489	0.1860
4	0.1568	0.1790	0.3844	0.4749	0.3590	0.3075	0.2298
5	0.1185	0.1353	0.2906	0.3590	0.4232	0.3625	0.2710
6	0.1015	0.1159	0.2489	0.3075	0.3625	0.4842	0.3620
7	0.0759	0.0866	0.1860	0.2298	0.2710	0.3620	0.4220

8.4 Tests sur les effets fixes

Une fois la structure de la matrice des covariances déterminée, on peut vérifier le niveau de signification des effets fixes.

```
model    potassium   = groupe  temps  groupe*temps;
repeated temps       / subject=chien(groupe)  type=arh(1);
lsmeans  groupe*temps/ slice=groupe slice=temps pdiff;
```

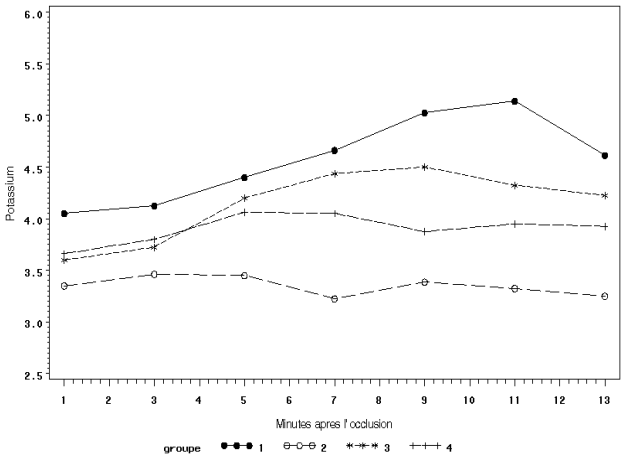
```
             Type 3 Tests of Fixed Effects
                    Num     Den
   Effect            DF      DF    F Value    Pr > F
   groupe             3      28       8.70    0.0003
   temps              6     168       4.63    0.0002
   groupe*temps      18     168       1.77    0.0331
```

La méthode *between-within* utilisée par défaut pour estimer les degrés de liberté donne les mêmes degrés que l'analyse split-plot avec la méthode *containement*. Est-ce judicieux sachant la forte dépendance positive entre les données ? Le seuil observé pour l'interaction, obtenu avec une F à 18 et 168 degrés de liberté, est de 0.0331.

La méthode de Satterthwaite donne 81.9 degrés de liberté au dénominateur, résultant en un seuil observé de 0.0440. La méthode d'estimation Kenward-Roger donne une statistique observée $F_{(18,110)} = 1.66$ (seuil observé de 0.0568). Les tests d'interaction MANOVA ont quant à eux des seuils observés allant de 0.065 à 0.08. L'interaction est donc marginalement significative. On pourra fixer les niveaux du groupe pour comparer les temps, et vice versa.

Tests of Effect Slices

Effect	groupe	temps	Num DF	Den DF	F Val	Pr > F
groupe*temps	1		6	168	4.71	0.0002
groupe*temps	2		6	168	0.85	0.5368
groupe*temps	3		6	168	3.42	0.0032
groupe*temps	4		6	168	0.96	0.4572
groupe*temps		1	3	168	3.42	0.0187
groupe*temps		3	3	168	3.63	0.0142
groupe*temps		5	3	168	2.77	0.0436
groupe*temps		7	3	168	6.73	0.0003
groupe*temps		9	3	168	9.68	<.0001
groupe*temps		11	3	168	9.48	<.0001
groupe*temps		13	3	168	6.28	0.0005



Quelques constatations jusqu'à maintenant :

- L'effet temporel est significatif dans les groupes 1 (témoin) et 3 (interruption du système nerveux avant l'occlusion) seulement.
- Il existe des différences significatives entre certains groupes de pré-traitement dès l'occlusion de l'artère, et durant toute la durée d'observation. Il faudra voir si les différences sont les mêmes au cours du temps.

8.5 Comparaisons multiples

Comparaison des temps dans les groupes 1 et 3

D'après l'analyse de la page précédente, on peut comparer les temps deux à deux dans les groupes 1 et 3. Il faut isoler les lignes qui nous concernent dans le tableau `Differences of Least Squares Means` (qui contient 378 lignes de comparaisons des 28 cellules !). Nous l'avons fait ci-dessous pour le groupe 1.

Differences of Least Squares Means

Effect	groupe	temps	_groupe	_temps	Estimate	Std Err	DF	t Value	Pr > |t|
groupe*temps	1	1	1	3	-0.07500	0.09550	168	-0.79	0.4333
groupe*temps	1	1	1	5	-0.3500	0.1890	168	-1.85	0.0658
groupe*temps	1	1	1	7	-0.6125	0.2115	168	-2.90	0.0043
groupe*temps	1	1	1	9	-0.9750	0.2187	168	-4.46	<.0001
groupe*temps	1	1	1	11	-1.0875	0.2443	168	-4.45	<.0001
groupe*temps	1	1	1	13	-0.5625	0.2415	168	-2.33	0.0210
groupe*temps	1	3	1	5	-0.2750	0.1571	168	-1.75	0.0819
groupe*temps	1	3	1	7	-0.5375	0.1874	168	-2.87	0.0047
groupe*temps	1	3	1	9	-0.9000	0.1990	168	-4.52	<.0001
groupe*temps	1	3	1	11	-1.0125	0.2282	168	-4.44	<.0001
groupe*temps	1	3	1	13	-0.4875	0.2272	168	-2.15	0.0333
groupe*temps	1	5	1	7	-0.2625	0.1547	168	-1.70	0.0915
groupe*temps	1	5	1	9	-0.6250	0.2023	168	-3.09	0.0023
groupe*temps	1	5	1	11	-0.7375	0.2428	168	-3.04	0.0028
groupe*temps	1	5	1	13	-0.2125	0.2586	168	-0.82	0.4125
groupe*temps	1	7	1	9	-0.3625	0.1501	168	-2.42	0.0168
groupe*temps	1	7	1	11	-0.4750	0.2074	168	-2.29	0.0233
groupe*temps	1	7	1	13	0.05000	0.2338	168	0.21	0.8309
groupe*temps	1	9	1	11	-0.1125	0.1510	168	-0.74	0.4573
groupe*temps	1	11	1	13	0.5250	0.1509	168	3.48	0.0006

On préférera une synthèse graphique de ce tableau (même s'il est déjà considérablement réduit).

Groupe 1 :

Temps :	1	3	5	13	7	9	11
Moyenne :	4.05	4.13	4.40	4.61	4.66	5.02	5.14

En faisant de même pour le groupe 3, on obtient le résultat suivant :

Groupe 3 :

Temps :	1	3	5	13	11	7	9
Moyenne :	3.60	3.73	4.20	4.23	4.32	4.44	4.50

Remarques sur le calcul des erreurs-types

• **Erreur-type de la moyenne au temps k dans le groupe i**

L'estimation de la moyenne fait intervenir les 8 observations (indépendantes) prises au temps k dans le groupe i. On calcule la variance comme suit :

$$Var(\overline{y_{i\bullet k}}) = Var\left(\frac{\sum_{j=1}^{8} y_{ijk}}{8}\right) = Var\left(\frac{\sum_{j=1}^{8} \mu + \tau_i + \beta_k + \tau\beta_{ik} + \varepsilon_{ijk}}{8}\right)$$
$$= Var\left(\frac{\sum_{j=1}^{8} \varepsilon_{ijk}}{8}\right) = \frac{\sum_{j=1}^{8} Var(\varepsilon_{ijk})}{64} = \frac{\sigma_k^2}{8}$$

Exemple

Dans le groupe 2 ($i = 2$), à 5 minutes ($k = 3$), le calcul se fait comme suit :

$$\widehat{Var}(\overline{y_{2\bullet 3}}) = \frac{\hat{\sigma}_3^2}{8} = \frac{0.4852}{8} = 0.06065 = 0.2463^2$$

Least Squares Means

Effect	groupe	temps	Estimate	Std Err	DF	t Val	Pr > |t|
groupe*temps	2	5	3.4500	0.2463	168	14.01	<.0001

• **Erreur-type de la différence de moyennes entre deux temps dans le groupe i**

$$\begin{aligned} Var(\overline{y_{i\bullet k}} - \overline{y_{i\bullet k'}}) &= Var(\overline{y_{i\bullet k}}) + Var(\overline{y_{i\bullet k'}}) - 2\,Cov\left(\frac{\sum_{j=1}^{8} y_{ijk}}{8}, \frac{\sum_{j=1}^{8} y_{ijk'}}{8}\right) \\ &= \frac{\sigma_k^2)}{8} + \frac{\sigma_{k'}^2}{8} - \frac{2\sum_{j=1}^{8} Cov(y_{ijk}, y_{ijk'})}{8 \times 8} \\ &= \frac{\sigma_k^2}{8} + \frac{\sigma_{k'}^2}{8} - \frac{2\sigma_k\sigma_{k'}\rho^{|k-k'|}}{8} \end{aligned}$$

La covariance entre y_{ijk} et $y_{ijk'}$ est constante pour tous les couples (i, j), car nous avons modélisé des sous-blocs de R identiques pour tous les chiens (nous n'avons pas estimé les paramètres différemment d'un groupe à l'autre avec l'option `group`).

Exemple

Comparons les temps $k = 1$ (1 min) et $k' = 3$ (5 min) dans le groupe 2 ($i = 2$).

$$\begin{aligned} \widehat{Var}(\overline{y_{2\bullet 1}} - \overline{y_{2\bullet 3}}) &= \frac{\hat\sigma_1^2}{8} + \frac{\hat\sigma_3^2}{8} - \frac{2\hat\sigma_1\hat\sigma_3\hat\rho^2}{8} \\ &= \frac{0.1963}{8} + \frac{0.4852}{8} - \frac{2\sqrt{0.1963 \times 0.4852} \times .8007^2}{8} \\ &= 0.0357 = 0.189^2 \end{aligned}$$

Differences of Least Squares Means

Effect	groupe	temps	_groupe	_temps	Estimate	Std Err	DF	t Val	Pr > |t|
groupe*temps	2	1	2	5	-0.1000	0.1890	168	-0.53	0.5974

Comparaisons des groupes pour chaque temps

Si on retourne aux tests sur les effets fixes, on s'aperçoit qu'il est préférable de fixer les temps pour comparer les groupes, en raison de l'interaction significative. On remarque également qu'il existe des différences entre les groupes pour tous les temps (tableau `Tests of Effects Slices`).

Comme précédemment, il faut isoler les lignes qui nous concernent dans le tableau `Differences of Least Squares Means`. Nous l'avons fait ci-dessous pour le temps 1.

Differences of Least Squares Means

Effect	groupe	temps	_groupe	_temps	Estimate	Std Err	DF	t Value	Pr > |t|
groupe*temps	1	1	2	1	0.7000	0.2215	168	3.16	0.0019
groupe*temps	1	1	3	1	0.4500	0.2215	168	2.03	0.0438
groupe*temps	1	1	4	1	0.3875	0.2215	168	1.75	0.0821
groupe*temps	2	1	3	1	-0.2500	0.2215	168	-1.13	0.2607
groupe*temps	2	1	4	1	-0.3125	0.2215	168	-1.41	0.1602
groupe*temps	3	1	4	1	-0.06250	0.2215	168	-0.28	0.7782

Temps 1 :

Groupe :	2	3	4	1
Moyenne :	3.35	3.60	3.66	4.05

Temps 3 :

Groupe :	2	3	4	1
Moyenne :	3.46	3.73	3.80	4.13

Temps 5 :

Groupe :	2	4	3	1
Moyenne :	3.45	4.06	4.20	4.40

Temps 7 :

Groupe :	2	4	3	1
Moyenne :	3.23	4.05	4.44	4.66

Temps 9 :

Groupe :	2	4	3	1
Moyenne :	3.39	3.87	4.50	5.03

Temps 11 :

Groupe :	2	4	3	1
Moyenne :	3.33	3.95	4.33	5.14

Temps 13 :

Groupe :	2	4	3	1
Moyenne :	3.25	3.93	4.23	4.61

Remarques sur le calcul des erreurs-types

• **Erreur-type de la différence de moyennes entre 2 groupes au temps** k

$$\begin{aligned} Var(\overline{y_{i\bullet k}} - \overline{y_{i'\bullet k}}) &= Var(\overline{y_{i\bullet k}}) + Var(\overline{y_{i'\bullet k}}) - 2\,Cov\left(\frac{\sum_{j=1}^{8} y_{ijk}}{8}, \frac{\sum_{j=1}^{8} y_{i'jk}}{8}\right) \\ &= \frac{\sigma_k^2}{8} + \frac{\sigma_k^2}{8} - \frac{2\sum_{j=1}^{8} Cov(y_{ijk}, y_{i'jk})}{8 \times 8} \\ &= \frac{\sigma_k^2}{4} \end{aligned}$$

y_{ijk} et $y_{i'jk}$ sont des mesures prises sur des chiens différents, car elles réfèrent à des groupes différents. Leur covariance est donc nulle.

Exemple

La variance de la différence des moyennes des groupes 2 et 4 au temps=5 minutes se calcule comme suit ($i = 2$, $i' = 4$ et $k = 3$) :

$$\widehat{Var}(\overline{y_{2\bullet 3}} - \overline{y_{4\bullet 3}}) = \frac{\hat{\sigma}_3^2}{4} = \frac{0.4852}{4} = 0.1213 = 0.3483^2$$

Differences of Least Squares Means

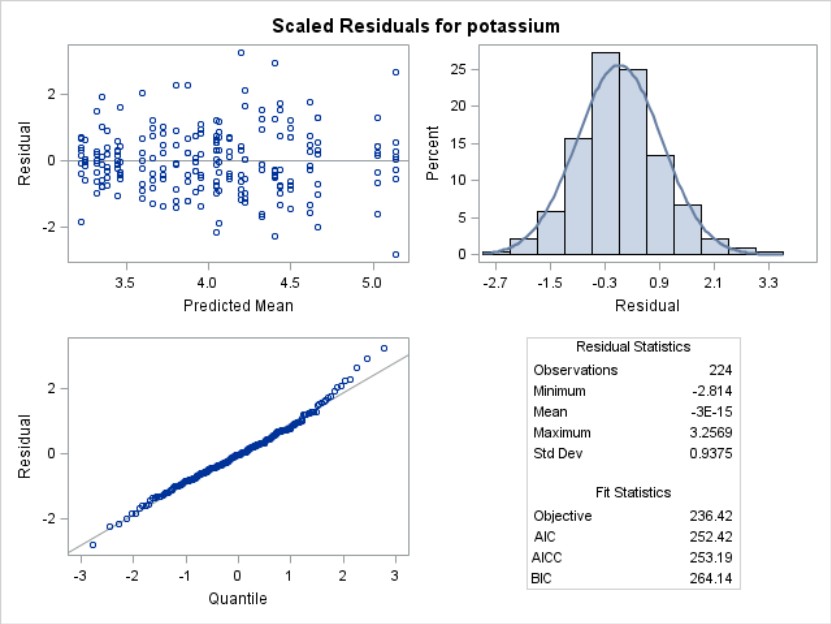
Effect	groupe	temps	_groupe	_temps	Estimate	Std Err	DF	t Val	Pr > |t|
groupe*temps	2	5	4	5	-0.6125	0.3483	168	-1.76	0.0805

8.6 Analyse des résidus

```
proc univariate data=out_pm normal;
var ScaledResid;
run;
```

Tests for Normality

Test	--Statistic---		-----p Value-----	
Shapiro-Wilk	W	0.988371	Pr < W	0.0663
Kolmogorov-Smirnov	D	0.053843	Pr > D	0.1121
Cramer-von Mises	W-Sq	0.094761	Pr > W-Sq	0.1347
Anderson-Darling	A-Sq	0.635197	Pr > A-Sq	0.0974



8.7 Annexe : Structures de covariance dans PROC MIXED

Vous trouverez aux pages suivantes une liste (non exhaustive !) de quelques structures de covariance offertes par PROC MIXED afin de modéliser les matrices G et R. La structure spécifiée détermine la forme des $blocs$ des matrices bloc-diagonales.

Dans ces matrices, p est la dimension des blocs (égale au nombre de mesures répétées lorsqu'on a un bloc par sujet), q est un nombre défini par l'utilisateur, w est le nombre de facteurs aléatoires, $1(A)$ est une variable indicatrice valant 1 lorsque l'événement A est réalisé et 0 sinon.

Les structures tenant compte de la dépendance spatiale entre les observations doivent être suivies de `(coord)`, i.e. la liste des c variables représentant les coordonnées de la position de chaque observation. Ces variables seront utilisées pour calculer la distance euclidienne d_{ij} entre les observations i et j. Ces structures sont très utiles lorsque les temps ne sont pas également espacés, par exemple.

Structure	Description	Nb. param.	Élément (i,j)	Exemple
AR(1)	Autorégressive(1)	2	$\sigma^2\rho^{\lvert i-j\rvert}$	$\sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$
ARH(1)	AR(1) Hétérogène	$p+1$	$\sigma_i\sigma_j\rho^{\lvert i-j\rvert}$	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix}$
ARMA(1,1)	Moy. Mobile Autorégressive	3	$\sigma^2[\gamma\rho^{\lvert i-j\rvert-1}1(i\neq j)+1(i=j)]$	$\sigma^2 \begin{bmatrix} 1 & \gamma & \gamma\rho & \gamma\rho^2 \\ \gamma & 1 & \gamma & \gamma\rho \\ \gamma\rho & \gamma & 1 & \gamma \\ \gamma\rho^2 & \gamma\rho & \gamma & 1 \end{bmatrix}$
CS	Compound Symmetry	2	$\sigma_1+\sigma^2 1(i=j)$	$\begin{bmatrix} \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2+\sigma_1 \end{bmatrix}$
CSH	CS Hétérogène	$p+1$	$\sigma_i\sigma_j[\rho 1(i\neq j)+1(i=j)]$	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{bmatrix}$
HF	Huynh-Feldt	$p+1$	$(\sigma_i^2+\sigma_j^2)/2+\lambda 1(i\neq j)$	$\begin{bmatrix} \sigma_1^2 & \frac{\sigma_1^2+\sigma_2^2}{2}-\lambda & \frac{\sigma_1^2+\sigma_3^2}{2}-\lambda & \frac{\sigma_1^2+\sigma_4^2}{2}-\lambda \\ \frac{\sigma_2^2+\sigma_1^2}{2}-\lambda & \sigma_2^2 & \frac{\sigma_2^2+\sigma_3^2}{2}-\lambda & \frac{\sigma_2^2+\sigma_4^2}{2}-\lambda \\ \frac{\sigma_3^2+\sigma_1^2}{2}-\lambda & \frac{\sigma_3^2+\sigma_2^2}{2}-\lambda & \sigma_3^2 & \frac{\sigma_3^2+\sigma_4^2}{2}-\lambda \\ \frac{\sigma_4^2+\sigma_1^2}{2}-\lambda & \frac{\sigma_4^2+\sigma_2^2}{2}-\lambda & \frac{\sigma_4^2+\sigma_3^2}{2}-\lambda & \sigma_4^2 \end{bmatrix}$
VC	Composantes de la variance	w	$\sigma_k^2 1(i=j)$, i corresp. au k^e effet	$\begin{bmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{bmatrix}$

Structure	**Description**	**Nb. param.**	**Élément** (i,j)	**Exemple**
TOEP	Toeplitz	p	$\sigma_{\lvert i-j\rvert+1}$	$\begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{bmatrix}$
TOEP(q)	Toeplitz à q bandes	q	$\sigma_{\lvert i-j\rvert+1}1(\lvert i-j\rvert < q)$	$\begin{bmatrix} \sigma^2 & \sigma_1 & 0 & 0 \\ \sigma_1 & \sigma^2 & \sigma_1 & 0 \\ 0 & \sigma_1 & \sigma^2 & \sigma_1 \\ 0 & 0 & \sigma_1 & \sigma^2 \end{bmatrix}$
TOEPH	TOEP Hétérogène	$2p-1$	$\sigma_i\sigma_j\rho_{\lvert i-j\rvert}$	$\begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 & \sigma_1\sigma_4\rho_3 \\ \sigma_2\sigma_1\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_1 & \sigma_2\sigma_4\rho_2 \\ \sigma_3\sigma_1\rho_2 & \sigma_3\sigma_2\rho_1 & \sigma_3^2 & \sigma_3\sigma_4\rho_1 \\ \sigma_4\sigma_1\rho_3 & \sigma_4\sigma_2\rho_2 & \sigma_4\sigma_3\rho_1 & \sigma_4^2 \end{bmatrix}$
TOEPH(q)	TOEP Hétéro. à q bandes	$p+q-1$	$\sigma_i\sigma_j\rho_{\lvert i-j\rvert}1(\lvert i-j\rvert < q)$	
UN	Sans structure	$\dfrac{p(p+1)}{2}$	σ_{ij}	$\begin{bmatrix} \sigma_1^2 & \sigma_{21} & \sigma_{31} & \sigma_{41} \\ \sigma_{21} & \sigma_2^2 & \sigma_{32} & \sigma_{42} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{43} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$
UN(q)	À q bandes	$\frac{q}{2}(2p-q+1)$	$\sigma_{ij}1(\lvert i-j\rvert < q)$	
SP(POW)(coord)	Puissance	2	$\sigma^2\rho^{d_{ij}}$	$\sigma^2 \begin{bmatrix} 1 & \rho^{d_{12}} & \rho^{d_{13}} & \rho^{d_{14}} \\ \rho^{d_{21}} & 1 & \rho^{d_{23}} & \rho^{d_{24}} \\ \rho^{d_{31}} & \rho^{d_{32}} & 1 & \rho^{d_{34}} \\ \rho^{d_{41}} & \rho^{d_{42}} & \rho^{d_{43}} & 1 \end{bmatrix}$

8.8 Exercices

1. Vrai ou Faux ?

(a) L'analyse de variance à mesures répétées ne s'applique qu'à des expériences longitudinales, i.e. des observations prises à différents temps sur un même individu.

(b) L'ajustement de la structure de covariance CS (coumpound symmetry) peut être comparé avec celui de la structure auto-régressive (AR(1)) par un test du rapport de vraisemblances.

(c) La log-vraisemblance d'un modèle ajusté avec une structure de covariance UN est toujours supérieure à celle d'un modèle AR(1) sur les mêmes données.

(d) Dans un plan équilibré, le critère d'information de Schwarz sera plus élevé que le critère d'Akaike si le nombre de sujets dans l'étude dépasse 8.

(e) Un analyse de données longitudinales de patients séparés dans des groupes de traitements avec un modèle split-plot est équivalent à l'utilisation de la structure de covariance compound symmetry (CS) pour modéliser la dépendance entre les temps.

(f) La structure de covariance a un effet sur les erreurs-types des moyennes, mais pas sur les tests sur les effets fixes.

2. Pour vérifier l'ajustement d'une structure de covariance, on peut comparer les valeurs des critères d'information (AIC, AICC, BIC, etc.) de plusieurs modèles et sélectionner celui qui a la plus petite valeur. Nommer deux autres moyens de choisir un modèle.

3. Retour sur l'exemple de ce chapitre sur les chiens dans 4 groupes de traitement avec structure de covariance ARH(1).

(a) Donnez l'expression de la variance de la moyenne de toutes les observations prises au temps k. Calculez une estimation de cette variance pour le temps $=$ 5 minutes ($k = 3$).

(b) Quelle est la variance de la différence de moyennes entre les temps k et k' ? Calculez une estimation de cette variance pour les temps $=$ 1 et 5 minutes ($k = 1$ et $k' = 3$).

(c) Quelle est la variance de la moyenne de toutes les observations du groupe i ? Estimez cette variance pour le groupe 2.

(d) Quelle est la variance de la différence de moyennes entre les groupes i et i' Estimez cette variance pour la différence entre les groupes 2 et 4.

4. On veut comparer l'évolution temporelle de la concentration dans le sang de trois médicaments pour traiter l'hypercholestérolémie. En tout 15 sujets (5 par traitement) sont recrutés et la concentration de médicament y est mesurée 0, 2, 4 et 6 minutes après l'injection. On veut traiter ces données à l'aide d'un modèle à mesures répétées qui suppose que les 4 mesures prises sur un sujet $(y_{ij1}, y_{ij2}, y_{ij3}, y_{ij4})$ suivent une distribution normale de dimension 4 avec une matrice de variances-covariances dont la structure est à déterminer.

(a) Quelles sont les dimensions des matrices Y, X et β du modèle d'analyse de la variance ?

(b) Quel est le rang de la matrice X ?

(c) Quelles sont les dimensions de la matrice de covariance R de toutes les observations de l'étude ? Combien de blocs contient-elle ?

(d) On pense ajuster les structures de covariances suivantes pour les blocs de la matrice R :

$$UN(2),\ AR(1),\ ARH(1),\ CS,\ CSH,\ TOEP,\ TOEPH(2)$$

Les paramètres de chacune de ces structures seront aussi estimés différemment dans chaque groupe de traitement. Écrivez dans chaque cas le nombre de paramètres impliqués.

(e) Supposons que les blocs de R ont la forme auto-régressive d'ordre 1.

(i) Donnez les lois marginales de y_{111} et $\overline{y}_{1\bullet 1}$. Interprétez ces quantités.

(ii) Donnez la loi marginale de $\overline{y}_{\bullet\bullet 3} - \overline{y}_{\bullet\bullet 1}$.

(f) Considérons pour simplifier la notation que $Y = (y_1, y_2, y_3, y_4)^T$ est le vecteur des observations prises sur un seul sujet. Supposons que les blocs de R ont la forme CS (compound symmetry). On veut modéliser les différences avec la mesure au temps 0. Les variables mesurées sur un sujet deviennent $x_1 = y_2 - y_1$, $x_2 = y_3 - y_1$, $x_3 = y_4 - y_1$; et les données sont des vecteurs de différences 3×1 pour les 15 individus.

(i) Exprimer le vecteur $X = (x_1, x_2, x_3)^T$ comme une transformation linéaire de Y.

(ii) Exprimer la matrice de variances-covariances de $X = (x_1, x_2, x_3)$ en fonction des paramètres σ^2 et σ_1 de R.

(iii) La matrice de variances-covariances de X a-t-elle la forme CS ?

(iv) Quelle est la corrélation entre deux observations x_i et x_k sur un même individu ?

5. On étudie l'inflammation d'une veine due à des piqûres intraveineuses de trois médicaments faites sur des lapins. On pique un lapin sur une oreille et on prend la température des deux oreilles après 0, 30, 60 et 90 minutes. Une grande différence de température est une indication de la présence d'inflammation ; c'est donc la variable réponse considérée dans l'expérience. Quinze lapins sont distribués aléatoirement aux trois traitements. Vous trouverez les données sur le site du cours.

(a) Écrivez le modèle d'analyse de variance à mesures répétées (sous forme matricielle) où l'on considère le temps comme un facteur catégorique, sans rien spécifier pour l'instant sur la matrice de variance-covariance.

(b) Choisissez une des structures de covariance suivantes :
- sans structure
- compound symmetry
- compound symmetry à variances hétérogènes
- auto-régressif d'ordre 1
- auto-régressif d'ordre 1 à variances hétérogènes

(i) en vous basant sur le critère d'information d'Akaike corrigé ;

(ii) en vous basant sur le critère d'information de Schwarz ;

(iii) en vous basant sur le test du rapport des vraisemblances.

(c) Écrivez la forme générale de la matrice de variance-covariance sélectionnée en (b) (celle qui revient le plus souvent). Donnez les estimations des paramètres.

(d) Faites l'analyse des effets fixes en utilisant la structure de covariance sélectionnée en (b).

9 Synthèse des plans d'expérience et introduction à la consultation statistique

Dans les chapitres précédents, nous nous sommes attardés à l'étude de schémas d'expériences déjà conçues, et non à la planification d'une expérience comme telle à partir d'objectifs généraux ou spécifiques. De plus, nous avons traité les schémas un à la fois. Il est maintenant temps de faire une synthèse de ces apprentissages, afin de mettre en lumière les spécificités de chaque schéma et de choisir une bonne allocation des ressources pour répondre aux questions du chercheur le plus efficacement possible.

Et comme il s'agit d'une brève introduction à la consultation statistique, nous vous proposerons plusieurs mises en situation provenant de domaines divers où vous pourrez vous exercer à sélectionner le bon plan d'expérience. Mais auparavant, nous vous présentons quelques éléments qu'un statisticien devrait garder en tête lorsqu'il travaille en collaboration avec des chercheurs pour valider des hypothèses scientifiques.

9.1 Poser les bonnes questions

Un chercheur peut faire appel à un statisticien à plusieurs moments : au début de la planification de l'expérience, en cours de collecte de données lorsqu'un imprévu survient, au moment de l'analyse des données, ou pour interpréter une analyse déjà réalisée par quelqu'un d'autre. Le succès de votre collaboration dépendra bien sûr de votre compétence à bien répondre à ses questions, mais surtout de votre capacité à lui poser les bonnes questions. Plusieurs éléments de cette section sont inspirés du document de l'American Statistical Association [2].

Si vous intervenez lors de la planification de l'expérience, vous serez appelés à réfléchir aux points suivants :

- la question de recherche, i.e. le but ou les buts de l'expérience (description, estimation, prédiction, inférence) ;
- le nombre et le type de variable réponse (discrète ou continue) ;
- le nombre et le type de variables explicatives ;
- le nombre de facteurs en cause, leur statut (fixe ou aléatoire) et leurs relations (simples, croisés ou emboîtés) ;
- les façons possibles d'appliquer les traitements aux unités expérimentales (taille des unités, biais de sélection) ;

- les façons possibles de collecter les observations (nombre d'intervenants, instruments de mesure, études à l'aveugle, calendrier de collecte) ;
- quelle différence entre deux moyennes est-elle considérée importante, quelle taille d'effet veut-on détecter ;
- a-t-on une idée de la variabilité des mesures ;
- les tailles d'échantillon minimales pour atteindre une puissance suffisante (versus le budget disponible ou l'accessibilité des unités expérimentales) ;
- la façon d'entrer les observations dans un chiffrier, ou de coder la base de données selon le format de collecte ;
- comment les autres études sur des sujets similaires organisaient la collecte (si vous n'êtes pas familier avec le domaine d'étude, une revue de littérature peut être utile) ;
- etc.

Si vous intervenez après la collecte des données, vous devriez poser assez de questions sur les items de la liste précédente pour être en mesure de refaire l'expérience de manière identique si elle vous était confiée. Si les données ne permettent pas de répondre à la question de recherche pour diverses raisons (méthodologie statistique, déviation au protocole expérimental, données manquantes, etc.), il sera de votre responsabilité d'en aviser le chercheur et de lui dire quelles informations peuvent être tirées des données qu'il vous a soumises.

Plusieurs des questions que vous devrez poser pourront sembler plus ou moins pertinentes à l'analyse statistique du point de vue du chercheur. Vous devrez peut-être le convaincre que pour choisir le bon modèle, vous devez saisir toutes les subtilités de la collecte de données. Il faut prévenir ce qu'on appelle parfois les erreurs de type III, soit l'apport de bonnes réponses aux mauvaises questions...

9.2 Éthique et autres considérations

Derrière les bonnes intentions des chercheurs et des statisticiens en général, il y a parfois des questions scientifiques ayant de grands impacts sur les communautés, il y a parfois beaucoup d'argent en jeu, et le statisticien doit être au fait des pressions qui pourraient lui être soumises ou des conséquences d'un manque de rigueur dans sa démarche.

Sans faire une liste exhaustive de tous les défis auxquels il pourrait faire face, notons quelques sujets sensibles auxquels le statisticien devrait réfléchir :

- La quête du p-value inférieur à 5% : la signification scientifique vs statistique
- La tendance des scientifiques à publier surtout des études significatives
- La pression de publication chez les universitaires
- L'importance de la répétition d'expériences pour confirmer les connaissances acquises et la difficulté de trouver du financement pour des études confirmatoires
- La relation consultant-statisticien/chercheur-scientifique : et si on vous demandait de manipuler des données, d'éliminer des valeurs "extrêmes", de trouver une façon de rendre les résultats concluants...
- Les minuscules échantillons avec 1000 mesures, les questionnaires à 100 questions posées dans le même ordre à peu de répondants
- Les formulations trompeuses : une étude a démontré que...
- La révision par les pairs, les méta-analyses et les initiatives comme Cochrane pour mettre en commun les connaissances
- L'équité des points de vue dans les médias, le déséquilibre entre la science et l'opinion
- etc.

Pour enrichir vos discussions, nous recommandons la lecture des articles suivants :

- Ethical Guidelines for Statistical Practice, ASA, 1999 [1]
- Statistics and Ethics : Some Advice for Young Statisticians, Vardeman S.B. et Morris, M.D., ASA, 2003 [22]
- Why Most Published Research Findings Are False, Ioannidis, J.P.A. (2005), PLoS Med 2(8), e124. [8]

9.3 Randomisation à l'aide de la procédure PLAN de SAS

La procédure PLAN permet de construire des plans d'expériences factorielles, contenant des facteurs croisés, emboîtés ou des blocs aléatoires. Voici les énoncés de base de la procédure, où i, j, k sont le nombre de modalités des facteurs. Les commandes entre crochets sont optionnelles.

```
PROC PLAN < SEED=germe >;
   FACTORS facteur=i < type_de_selection > facteur=j < type_de_selection >;
   OUTPUT OUT= sortie < DATA=jeu_contenant_observations > ;
   TREATMENTS facteur=k < type_de_selection >;
   RUN;
```

Il y a 5 façons de générer les combinaisons de traitements à chaque niveau du facteur précédent (que l'on appelle le type de sélection) dont les principales sont `random` (ordre aléatoire, par défaut) et `ordered` (ordre habituel croissant).

Voici quelques exemples d'application pour randomiser l'allocation des traitements dans différents plans d'expérience.

1. Plan complètement aléatoire : un facteur fixe à 2 niveaux, 3 répétitions par niveau. La procédure randomise une colonne d'un jeu de données.

```
data test;
    unite =0;
    do  A =1,2;
    do rep=1,2,3;
     unite =unite+1;
     output;end;end;
   run;
proc print data=test;
   run;
```

Obs	unite	A	rep
1	1	1	1
2	2	1	2
3	3	1	3
4	4	2	1
5	5	2	2
6	6	2	3

```
proc plan ;
  factors unite=6;
  output data=test out=sortie;
  run;
proc print data=sortie;
  run;
```

Obs	unite	A	rep
1	1	1	1
2	6	1	2
3	4	1	3
4	2	2	1
5	5	2	2
6	3	2	3

2. Plan complètement aléatoire : 2 facteurs fixes à 2 et 3 niveaux, 3 répétitions par combinaison de traitements.

```
data test2;
   unite=0;
   do A=1,2;
   do B=1,2,3;
   do rep=1,2,3;
   unite=unite+1;
   output;
   end;end;end;
   run;
proc print data=test2;
   run;

proc plan ;
   factors unite=18;
   output data=test2 out=sortie2(drop=rep);
   run;
proc sort data=sortie2;
   by unite;
proc print data=sortie2;
   run;
```

Obs	unite	A	B	rep
1	1	1	1	1
2	2	1	1	2
3	3	1	1	3
4	4	1	2	1
5	5	1	2	2
6	6	1	2	3
7	7	1	3	1
8	8	1	3	2
9	9	1	3	3
10	10	2	1	1
11	11	2	1	2
12	12	2	1	3
13	13	2	2	1
14	14	2	2	2
15	15	2	2	3
16	16	2	3	1
17	17	2	3	2
18	18	2	3	3

Obs	unite	A	B
1	1	2	1
2	2	2	3
3	3	1	3
4	4	1	1
5	5	2	3
6	6	2	1
7	7	1	1
8	8	2	2
9	9	2	2
10	10	2	1
11	11	1	3
12	12	2	2
13	13	1	2
14	14	1	1
15	15	1	2
16	16	2	3
17	17	1	2
18	18	1	3

3. Plan à blocs aléatoires complets : 3 blocs, un facteur A à 5 niveaux.

```
 proc plan ordered;
      factors bloc=3 cellule=5;
      treatments A=5 random;
      output out=pbac;
      run;

   proc print data=pbac;
      run;
```

Plot Factors			
Factor	Select	Levels	Order
bloc	3	3	Ordered
cellule	5	5	Ordered

Treatment Factors			
Factor	Select	Levels	Order
A	5	5	Random

bloc	cellule					A				
1	1	2	3	4	5	2	5	4	1	3
2	1	2	3	4	5	2	4	3	1	5
3	1	2	3	4	5	2	4	3	5	1

Obs	bloc	cellule	A
1	1	1	2
2	1	2	5
3	1	3	4
4	1	4	1
5	1	5	3
6	2	1	2
7	2	2	4
8	2	3	3
9	2	4	1
10	2	5	5
11	3	1	2
12	3	2	4
13	3	3	3
14	3	4	5
15	3	5	1

4. Plan hiérarchique : 3 facteurs emboîtés. 2 plants dans chaque pot, 4 pots dans chaque serre.

```
proc plan ;
factors serres=3 pots=4 plants=2 ;
output out=hierarchique;
run;

proc print data=hierarchique;
run;
```

serres	pots	plants	
3	1	1	2
	3	1	2
	4	1	2
	2	1	2
1	1	1	2
	4	1	2
	2	2	1
	3	1	2
2	2	2	1
	3	2	1
	4	2	1
	1	1	2

5. Split-plot avec PBAC au premier niveau : 3 blocs, 4 parcelles par bloc où on assigne un des 4 niveaux du facteur A, et 3 niveaux de B appliqués aux sous-parcelles.

```
proc plan seed=37277;
factors Bloc=3 ordered A=4 B=3;
run;
```

Bloc	A	B		
1	4	3	2	1
	3	3	1	2
	2	2	3	1
	1	2	3	1
2	1	2	3	1
	3	2	1	3
	4	2	1	3
	2	3	2	1
3	4	3	1	2
	1	3	1	2
	2	1	3	2
	3	2	3	1

9.4 Exercices

1. Pour chacune des situations expérimentales, proposez un modèle, précisez si chaque facteur est fixe ou aléatoire, et donnez les degrés de liberté associés à chacun d'eux dans une table d'anova (les tailles d'échantillon sont égales). Nommez le modèle statistique.

(a) Le groupe Nautilus veut proposer deux programmes de conditionnement physique aux usagers. On conduit d'abord une expérience pour s'assurer de leur efficacité et de la constance des résultats entre les centres. Trois centres sont choisis au hasard pour l'étude. Dans chacun d'eux, on sélectionne au hasard 10 clients ayant à peu près les mêmes caractéristiques physiques dont 5 suivront le programme d'exercices 1 et 5 autres suivront le programme d'exercices 2. On prend diverses mesures continues sur les participants à la fin d'un mois d'entraînement.

(b) On veut mesurer l'onctuosité de 5 variétés de maïs selon la méthode de cuisson utilisée. On divise un champ en 5 parcelles. Sur chacune de ces parcelles, on sème une des 5 variétés de maïs, allouées aléatoirement. À la fin de l'été, quelques épis de chaque parcelle seront cuits au BBQ, quelques-uns à la vapeur, et d'autres à l'eau bouillante. On prendra une mesure globale de l'onctuosité pour chaque méthode de cuisson. On répète cette expérience sur deux autres champs différents et les résultats sont analysés globalement.

(c) On veut vérifier la fiabilité d'un appareil sensé mesurer la concentration de glucose dans un certain type de sérum. On choisit de comparer trois niveaux de concentration standard. Un même technicien conduira l'expérience sur quatre jours. On planifie deux essais chaque jour. À chaque essai, le technicien prend la mesure de glucose sur un échantillon de chaque concentration, dans un ordre aléatoire. L'appareil retourne une mesure de la concentration pour chaque éprouvette. La variable d'intérêt est l'erreur relative sur chaque mesure prise.

(d) Un manufacturier de produits médicaux fabrique des veines artificielles. Ces dernières sont produites en extrayant des rondins à l'intérieur de cylindres de résine. Il soupçonne que les imperfections des tubes résultant de cette opération (pouvant mener au rejet par le patient) sont causées par la trop forte pression utilisée pour l'extraction. Puisqu'il sait que les différents lots de cylindres de résine ne sont pas homogènes, il sélectionne au hasard 6 cylindres reçus de son fournisseur lors d'envois différents. Chaque cylindre est coupé en 3 parties. Chaque partie servira à fabriquer une veine artificielle en appliquant une pression

faible, moyenne ou élevée (répartition aléatoire). On mesure la qualité de chaque veine produite.

(e) Une étudiante en psychologie veut comparer l'effet de deux anti-dépresseurs par rapport à un placebo. Elle souhaite aussi savoir si le traitement médical a un effet différent selon que le patient suit une psychothérapie ou non après le diagnostic de dépression. Elle fait un appel pour des volontaires ; 24 personnes acceptent de participer. On alloue aléatoirement à chacun d'eux un des traitements médicaux (les patients ne seront pas informés de ce qu'ils prennent). On choisit au hasard s'ils suivront une thérapie ou non. On s'assure par contre d'obtenir un plan équilibré, i.e. d'avoir le même nombre de patients pour chaque combinaison de traitements. Suggérer une méthode d'analyse pour les deux méthodes de collecte des données :

(i) Une rencontre avec l'étudiante au début et à la fin de l'étude permettra d'évaluer l'état de santé mentale des patients.

(ii) Une rencontre avec l'étudiante à 5 moments différents durant l'étude permettra d'évaluer l'état de santé mentale des patients.

(f) On veut comparer la production laitière des vaches selon diverses formules alimentaires. On sélectionne aléatoirement trois fermes québécoises parmi celles qui font l'élevage de vaches laitières. On veut comparer la nourriture pure à celle contenant des hormones ajoutées. De plus, on veut savoir s'il est préférable de donner la nourriture sous forme de grains ou de foin. On dispose dans chaque ferme de huit vaches pour l'étude, dont deux subiront le régime de grains sans hormones, deux seront nourries de grains avec hormones, deux avec du foin pur, et deux avec du foin avec ajout d'hormones. On mesure après un mois de ce régime la quantité moyenne de lait fournie par chaque vache pendant une semaine.

(g) Un boulanger cherche à optimiser sa recette de biscuits. Il prépare neuf grands bols de farine. Trois d'entre eux sont choisis au hasard et on y ajoute du sel. Trois autres bols sont choisis et on y ajoute de la poudre à pâte. Dans les trois derniers bols, on ajoute plutôt du bicarbonate de soude. Le contenu de chaque bol est ensuite bien mélangé et séparé en quatre portions. Deux d'entre elles seront mélangées avec du beurre, tandis que les deux autres recevront de la graisse végétale. Dans chaque sous-portion, on ajoute du sucre, des pépites de chocolat, et on étend le mélange sur une plaque qui ira au four à 350 degrés. On mesure à la fin de la cuisson diverses qualités des biscuits.

(h) On veut comparer le bourgeonnement des trembles entre la rive sud et la rive nord du fleuve St-Laurent. De chaque côté du fleuve, on choisit cinq parcelles de $10\,000$ m^2 dans des forêts mixtes situées à peu près à une même latitude. Dans chaque parcelle, on choisit 10 trembles ayant un diamètre à hauteur de poitrine d'environ 30 cm. Toute l'expérience est conduite la même journée au cours du printemps, et on évalue le développement des bourgeons sur chaque arbre.

(i) Le ministère de l'éducation conduit un projet-pilote pour tester des méthodes d'enseignement et d'évaluation pour l'étude de l'anglais au primaire. On choisit au hasard quatre écoles primaires de la province de Québec. Dans chacune d'entre elles, un groupe de 27 élèves recevra la méthode d'enseignement A, et un autre groupe de 27 élèves recevra la méthode B. On s'assure que les deux méthodes soient enseignées par la même personne dans une école. Après deux mois, on sépare chaque classe en trois groupes de 9 élèves. Chaque sous-groupe recevra une des trois évaluations proposées par le ministère. La variable réponse sera la note sur 100 obtenue par chaque élève à l'évaluation.

2. Pour chacun des plans d'expérience suivants, donnez le tableau des facteurs du modèle, leur statut si possible, et leurs degrés de liberté respectifs. Prenez pour acquis que les facteurs A, B et C sont fixes et que les blocs sont aléatoires.

(a)

Plan 1

A2-B1	A3-B2	A2-B2	A3-B2	A1-B1	A3-B1
A2-B2	A1-B1	A3-B1	A2-B1	A3-B2	A1-B2
A1-B2	A3-B1	A1-B2	A1-B1	A2-B2	A2-B1

(b)

Plan 2

Bloc 1:	A2	A3	A1	A4
Bloc 2:	A4	A2	A3	A1
Bloc 3:	A1	A4	A2	A3

(c)

Plan 3

A2-B2 (x,y)	A1-B2 (x,y)	A2-B1 (x,y)	A1-B1 (x,y)
A2-B1 (x,y)	A1-B2 (x,y)	A1-B1 (x,y)	A2-B2 (x,y)
A1-B1 (x,y)	A2-B2 (x,y)	A2-B1 (x,y)	A1-B2 (x,y)

(d)

Plan 4

	Bloc-col. 1:	Bloc-col. 2:	Bloc-col. 3:	Bloc-col. 4:
Bloc-ligne 1:	A2	A1	A4	A3
Bloc-ligne 2:	A4	A3	A2	A1
Bloc-ligne 3:	A3	A2	A1	A4
Bloc-ligne 4:	A1	A4	A3	A2

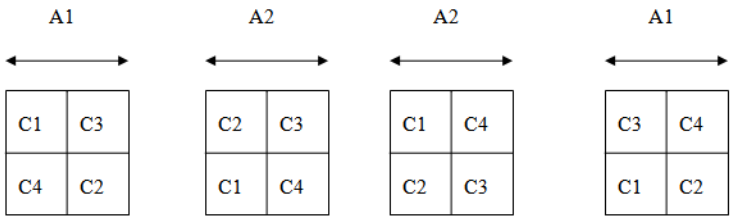
(e)

Plan 5

A1-B2	A2-B4	A1-B2	A1-B1
A2-B3	A2-B4	A1-B1	A2-B3
A1-B1	A2-B3	A1-B2	A2-B4

(f)

Plan 6



(g)

Plan 7

Bloc 1				Bloc 2			
A1-B2	A2-B2	A1-B1	A2-B1	A2-B1	A2-B2	A1-B1	A1-B2
C1	C2	C2	C1	C1	C2	C2	C1
C2	C1	C1	C2	C2	C1	C1	C2

(h)

Plan 8

B1			B2		B3		
A3	A1	A2	A4	A5	A7	A6	A8
C1	C2	C2	C1	C2	C1	C2	C2
C2	C2	C1	C1	C1	C2	C1	C2
C2	C1	C1	C2	C1	C2	C1	C1
C1	C1	C2	C2	C2	C1	C2	C1

10 Bibliographie

Références

[1] American Statistical Association (1999). *Ethical Guidelines for Statistical Practice*, (Lien web)

[2] American Statistical Association (2003). *When you consult a statistician... What to expect*, (Lien web)

[3] Crowder, M.J. et Hand D.J. (1990). *Analysis of Repeated Measures*, Chapman and Hall, London. (Catalogue bibliothèque)

[4] Dagnelie, P. (2003). *Principes d'expérimentation. Planification des expériences et analyse de leurs résultats*, Presses agronomiques de Gembloux, Belgique. (Catalogue bibliothèque)

[5] Everitt, B.S. (1995). The analysis of repeated measures : a practical review with examples, *Statistician*, **44**(1), 113-135.

[6] Genz, A. (2004). Numerical computation of rectangular bivariate and trivariate normal and t-probabilities, *Statistics and Computing*, **14**, 251–260.

[7] Grizzle, J. E. and Allen, D. M. (1969). Analysis of growth and dose response curves, *Biometrics*, Vol. 25, No 2, 357-381.

[8] Ioannidis, J.P.A. (2005). Why Most Published Research Findings Are False, PLoS Med, vol. 2 no.8, e124. (Lien web)

[9] Jiang, J. (2007). Linear and Generalized Linear Mixed Models and Their Applications, Springer, New York. (Accès bibliothèque)

[10] Littell, R. C., Milliken, G. A., Stroup, W. W., et Wolfinger, R. D. (2002). *SAS System for Mixed Models*, SAS Institute Inc., Cary. (Catalogue bibliothèque)

[11] Littell et al. (2002). *SAS for Linear Models*, SAS Institute Inc., Cary. (Catalogue bibliothèque)

[12] Mead, R., Curnow, R.N. et Hasted, A.M. (2003). *Statistical Methods in Agriculture and Experimental Biology*, 3^e edition, Chapman & Hall, New York. (Catalogue bibliothèque)

[13] Milliken, G.A. et Johnson, D.E. (2009). *Analysis of Messy Data. Vol. 1 : Designed Experiments*, 2^e edition, CRC Press. (Accès bibliothèque)

[14] Milliken, G.A. et Johnson, D.E. (1989). *Analysis of Messy Data. Vol. 2 : Nonreplicated Experiments*, Chapman & Hall, New York. (Catalogue bibliothèque)

[15] Milliken, G.A. et Johnson, D.E. (2002). *Analysis of Messy Data. Vol. 3 : Analysis of Covariance*, CRC Press. (Accès bibliothèque)

[16] Montgomery, D.C. (2005). *Design and Analysis of Experiments*, 6^e edition, John Wiley & Sons, 2005. (Catalogue bibliothèque)

[17] Philippeau, G. (1989). *Théorie des plans d'expérience. Application à l'agronomie*, Service des études statistiques de l'I.T.C.F.. (Catalogue bibliothèque)

[18] SAS Institute, Aide en ligne de SAS 9.3, support.sas.com/documentation/onlinedoc/base/, Cary, North Carolina.

[19] Scherrer, B. (2009). *Biostatistique. Vol. 2*, 2^e edition, Gaëtan Morin Éditeur, Montréal. (Catalogue bibliothèque)

[20] Satterthwaite, F. E. (1946). *An approximate distribution of estimates of variance components.*, Biometrics Bulletin 2 : 110-114, doi :10.2307/3002019. (Lien JSTOR)

[21] Self, S. et Liang, K.Y. (1987). *Asymptotic Properlies of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions*, Journal of the American Statistical Association, Vol. 82, No. 398, pp. 605-610

[22] Vardeman, S.B et Morris, M.D. (2003). *Statistics and Ethics : Some Advice for Young Statisticians.*, American Statistical Association, vol. 57, no. 1, (Lien web)

[23] Verbeke, G. et Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data*, Springer Series in Statistics, Springer-Verlag, New York.

[24] Verbeke, G. et Molenberghs, G. (1997). *Linear Mixed Models in Practice. A SAS-Oriented Approach*, Lectures Notes in Statistics, **126**, Springer-Verlag, New York. (Catalogue bibliothèque)

[25] Wackerly, D.D., Mendenhall, W., Scheaffer, R.L. (2007). *Mathematical Statistics With Applications*, 7^e edition, Thomson-Brooks/Cole, Belmont. (Catalogue bibliothèque)

[26] Welch, B. L. (1947). *The generalization of Student's problem when several different population variances are involved.*, Biometrika 34 : 28-35 (Lien JSTOR)

[27] Wolfinger, R. D. (1993). Covariance structure selection in general mixed models, *Communications in Statistics : Simulation and Computation*, Texas, **22**(4), 1079-1106.

[28] Wolfinger, R. D. et Chang, M. (1995). *Comparing the SAS GLM and MIXED Procedures for Repeated Measures*, SAS Institute Inc., Cary.