MAT7381 Chapitre 6
Analyse de variance

On présente ici quelques cas particuliers d'un modèle linéaire général dans lequel les variables explicatives sont qualitatives. Ce sont des modèles d'analyse de plans d'expérience dans lesquels des valeurs d'une variables sont classées selon les conditions dans lesquelles elles ont été observées, ou le traitement qui a été appliqué. Comme, par exemple, la récolte de céréales dans plusieurs terrains, classée selon le type d'engrais utilisé. Les traitements peuvent être caractérisés par un seul « facteur » (type d'engrais, par exemple); ou par plus d'un, comme, par exemple, le type d'engrais et l'espèce de céréale. On parle alors d'un modèle à un facteur, à deux facteurs, etc.

Analyse de variance à un facteur

6.1 Le modèle

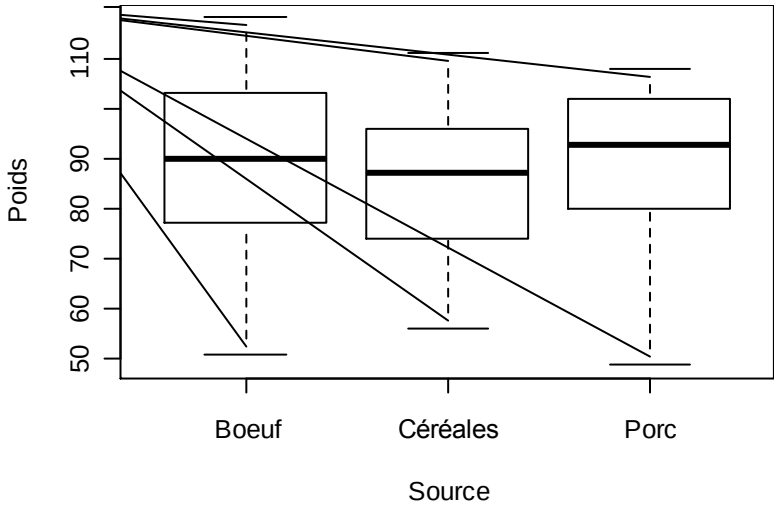
L'analyse de variance à un facteur fournit un test qui généralise le « test de Student » pour comparer deux moyennes: on compare k moyennes ($k \geq 2$) au lieu de 2.

***Exemple* 6.1.1** Le tableau suivant présente les résultats d'une expérience conçue pour déterminer si la source des protéines dans l'alimentation des poussins a un effet sur leur taux de croissance. Un groupe de 60 poussins sont répartis en 3 groupes, chacun recevant une alimentation dont les protéines proviennent d'une source différente : bœuf, céréales, ou porc.

Boeuf	Céréales	Porc	Boeuf	Céréales	Porc
90	107	49	73	98	94
76	95	82	102	74	79
90	97	73	118	56	96
64	80	86	104	111	98
86	98	81	81	95	102
51	74	97	107	88	102
72	74	106	100	82	108
90	67	70	87	77	91
95	89	61	117	86	102
78	58	82	111	92	105

La première question posée est la suivante : y a-t-il des différences entre les différentes sources de protéines (quant à leur effet sur la croissance) ?

Le graphique suivant en donne une première idée :



À première vue, les différences ne semblent pas être très importantes : les écarts entre les médianes semblent mineurs lorsqu'on les compare aux variations de poids. Voici, pour quantifier ces observations, les moyennes et les variances de *y* pour chaque groupe :

	Bœuf	Céréales	Porc	Échantillon complet
Moyennes	89,6	84,9	88,2	87,56667
Variances	313,7263	224,8316	257,6421	260,3514

■

Une étude comme celle décrite au dernier exemple illustre un plan d'expérience classique destiné à déterminer l'effet de différents traitements sur les valeurs d'une variable aléatoire *y*. La démarche est la suivante :

- On répartit un échantillon de *n* individus en *k* groupes ;
- chaque groupe subit un traitement différent ;
- on observe les valeurs de *y* sur toutes les unités de l'échantillon ;
- on décide si les différences entre est groupes sont significative à la lumière des moyennes des groupes et de leurs variance.

Il n'est pas obligatoire, bien que préférable, que les effectifs des groupes soient les mêmes.

***Exemple* 6.1.2** Les données suivantes sont des indices de la distorsion du son produit par des bandes magnétiques [Battacharrya, Gouri K. , Johnson, Richard A. (1977) *Statistical concepts and methods*, Wiley, New York, *p*.502]. Les bandes magnétiques appartiennent à 4 catégories, A, B, C et D, qui diffèrent selon le type d'enduit qui les recouvre. Le but de l'analyse est de déterminer si le type d'enduit a un effet sur la qualité du son, telle que mesurée par l'indice de distorsion.

Indices de distorsion de quatre types de bandes magnétiques

A	B	C	D
10	14	17	12
15	18	16	15
8	21	14	17
12	15	15	15
15		17	16
		15	15
		18	

Nous décrirons le modèle qui servira à l'analyse. Le lecteur peut garder en mémoire le dernier exemple de façon à concrétiser la description.

Les données sont classées en k catégories ($k = 4$ dans l'exemple), la i^e catégorie ayant n_i observations, $i = 1, ..., k$. Soit y_{ij} la j^e observation de la catégorie i (dans l'exemple 6.1.2, l'indice de distorsion de la j^e bande de la catégorie i). Les données peuvent être présentées dans le format suivant :

	T_1	T_2	...	T_k	
	y_{11} y_{12} $\vdots$ y_{1n_1}	y_{21} y_{22} $\vdots$ y_{2n_2}	... ... $\ddots$ ...	y_{k1} y_{k2} $\vdots$ y_{kn_k}	
Moyennes	$\bar{y}_{1\cdot}$	$\bar{y}_{2\cdot}$	...	$\bar{y}_{k\cdot}$	$\bar{y}_{\cdot\cdot}$
Variances	s_1^2	s_2^2	…	s_k^2	S^2

Nous supposons que

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

où les ε_{ij} sont indépendantes et

$$\varepsilon_{ij} \sim \mathcal{N}(0 ; \sigma^2)$$

Les paramètres inconnus du modèle sont $\mu_1, ... ,\mu_k$ et σ^2. Nous acceptons ici comme en régression linéaire une hypothèse d'homoscédasticité : même variance σ^2 pour tous les ε_{ij}. L'objectif usuel d'une analyse de variance est de déterminer s'il y a des différences entre les populations, c'est-à-dire de tester l'hypothèse

$$H_o : \mu_1 = \mu_2 = \ldots = \mu_k$$

L'alternative est la négation de H_o:

H_o: *au moins une des égalités* $\mu_i = \mu_j$ $(i \neq j)$ *n'est pas vérifiée*

6.2 Décomposition des sommes de carrés

Définissons

$$\bar{y}_{i\cdot} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}\ , \qquad \bar{y}_{\cdot\cdot} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{n}\sum_{i=1}^{k} n_i \bar{y}_i$$

La somme des carrés totale, SCT, est

$$\text{SCT} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2$$

Cette somme peut être décomposée en deux sommes de carrés, la somme des carrés expliquée SCE, et la somme des carrés résiduelle, SCR où

$$\text{SCE} = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 \text{ , et } \quad \text{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2$$

Nous pouvons alors démontrer que

$$\text{SCT} = \text{SCE} + \text{SCR}$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_{i=1}^{k} n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i\cdot})^2 .$$

Remarques

- SCT, la somme des carrés *totale*, est une mesure de la dispersion des y_{ij}, indépendamment des catégories. SCT = $(n\text{-}1)S^2$, où S^2 est la variance de toutes les données.

- La dispersion des y_{ij} est en partie attribuable au fait que les traitements sont différents. C'est la partie *expliquée* SCE, une mesure des différences *entre* les différents types de bande magnétique, donc attribuable aux traitements. On remarque que l'écart au carré $(\bar{y}_{i.} - \bar{y}_{..})^2$ entre la moyenne du groupe i est la moyenne globale est pondéré par la taille n_i du groupe i.

- SCR est la partie *résiduelle*, la dispersion entre les bandes *d'un même type*, donc attribuable au fait que ces mesures varient naturellement d'une bande à l'autre, même quand elles sont fabriquées de la même façon (même enduit). ■

Lorsque les μ_i sont égaux, les moyennes échantillonnales $\bar{y}_i$ devraient être assez rapprochées les unes des autres et SCE devrait être petite. Donc nous devrions rejeter H_o si SCE est grand. Mais pour évaluer l'importance de la somme SCE il faudrait la mettre en relation avec la tendance naturelle que les y_{ij} ont à varier, c'est-à-dire, SCR. Le test sera donc fonction du rapport SCE/SCR.

6.3 Propriétés des sommes de carrés

Nous allons tirer les propriétés fondamentales des sommes de carrés à partir d'un cas particulier très simple: le tirage d'un seul échantillon d'une population normale.

Théorème *Soit $y_1, y_2, \ldots, y_n$ n variables aléatoires indépendantes, toutes de loi $\mathcal{N}(\mu\,;\,\sigma^2)$,*

$\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ et $S^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}$. *Alors*

i) $\bar{y} \sim \mathcal{N}(\mu\,;\,\sigma^2/n)$

ii) $(n\text{-}1)S^2/\sigma^2 \sim \chi^2_{n-1}$

iii) $\bar{y}$ et S^2 *sont indépendantes*.

Démonstrations

i) Ce résultat découle du fait que toute combinaison linéaire de variables normales indépendantes est normale (Théorème 1.2.1).

ii) Nous réécrirons la statistique $(n\text{-}1)S^2/\sigma^2$ sous forme matricielle afin d'utiliser les théorèmes du document 3 sur la distribution de formes quadratiques. Le vecteur $\boldsymbol{y} = [y_1; y_2; \ldots; y_n]'$ est de loi normale multidimensionnelle de moyenne $\mu\boldsymbol{e}$ (où $\boldsymbol{e}$ est un vecteur de n « 1 ») et de matrice de covariance $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$. On a, tout d'abord,

$$(n\text{-}1)S^2/\sigma^2 = \sum_{i=1}^n (y_i - \bar{y})^2 = \boldsymbol{y}\mathbf{C}\boldsymbol{y} \text{ où } \mathbf{C} = \mathbf{I}\text{-}\boldsymbol{e}\boldsymbol{e}'/n,$$

car $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2 / n = \boldsymbol{y}'\boldsymbol{y}\text{-}(\boldsymbol{e}'\boldsymbol{y})^2/n = \boldsymbol{y}'\boldsymbol{y}\text{-}(\boldsymbol{e}'\boldsymbol{y})(\boldsymbol{e}'\boldsymbol{y})/n = \boldsymbol{y}'\boldsymbol{y}\text{-}\boldsymbol{y}'(\boldsymbol{e}\boldsymbol{e}/n)'\boldsymbol{y} = \boldsymbol{y}'\mathbf{C}\boldsymbol{y}.$

La statistique $(n\text{-}1)S^2/\sigma^2$ s'écrit donc $\frac{\boldsymbol{y}'\mathbf{C}\boldsymbol{y}}{\sigma^2} = \boldsymbol{z}'\mathbf{C}\boldsymbol{z}$, où $\boldsymbol{z} = \boldsymbol{y}/\sigma \sim \mathcal{N}(\mu\boldsymbol{e}\,;\,\mathbf{I})$. On vérifie aisément que $\mathbf{C}$ est idempotente et que $\mathbf{C}\boldsymbol{\mu} = \mu\mathbf{C}\boldsymbol{e} = \mathbf{0}$. Donc d'après le corollaire du théorème 1.1.3, $\boldsymbol{z}'\mathbf{C}\boldsymbol{z} \sim \chi^2_r$ *où* $r = r(\mathbf{A})$. Il reste à déterminer $r = r(\mathbf{C})$. Puisque $\mathbf{C}$ est idempotente, d'après le théorème 2.2 du document 0, $r(\mathbf{C}) - tr(\mathbf{C}) = tr(\mathbf{I})\text{-}tr(\boldsymbol{e}\boldsymbol{e}'/n) = tr(\mathbf{I})\text{-}tr(\boldsymbol{e}'\boldsymbol{e}/n) = n$ -1.

iii) $\bar{y}$ est fonction de $\boldsymbol{e}'\boldsymbol{y}$ et $(n\text{-}1)S^2/\sigma^2$ est fonction de $\boldsymbol{y}'\mathbf{C}\boldsymbol{y}$. Mais étant donné que $\mathbf{C}$ est idempotente et symétrique, $\boldsymbol{y}'\mathbf{C}\boldsymbol{y} = \boldsymbol{y}'\mathbf{C}\mathbf{C}\boldsymbol{y} = \boldsymbol{y}'\mathbf{C}'\mathbf{C}\boldsymbol{y} = (\mathbf{C}\boldsymbol{y})'(\mathbf{C}\boldsymbol{y})$, ce qui veut dire que $\boldsymbol{y}'\mathbf{C}\boldsymbol{y}$ est fonction de $\mathbf{C}\boldsymbol{y}$. Il suffit donc d'établir l'indépendance de deux fonctions linéaires $\boldsymbol{e}'\boldsymbol{y}$ et $\mathbf{C}\boldsymbol{y}$. D'après le théorème 0.2.2, il suffit de vérifier que $\boldsymbol{e}'\mathbf{C} = \mathbf{0}$, ce qui aisément fait. ■

Nous énonçons ici quelques propriétés des sommes de carrés

1. $\text{SCR}/\sigma^2 \sim \chi^2_{n-k}$ et les $\bar{y}_i$ sont indépendantes de SCR.
2. $\text{SCE}/\sigma^2 \sim \chi^2_{k-1}(\lambda)$ où $\lambda = \dfrac{\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2}{\sigma^2}$ et $\bar{\mu} = \dfrac{1}{n}\sum_{i=1}^{k} n_i\mu_i$. Donc $\lambda = 0$ si et seulement si H_o est vraie.
3. SCR et SCE sont indépendantes.
4. $\text{MCR} = \text{SCR}/(n\text{-}k)$ est un estimateur sans biais de σ^2.

Table d'analyse de variance

Il existe comme avec la régression une façon traditionnelle de présenter les résultats d'une analyse de variance. Nous ajoutons à la table usuelle une colonne d'espérances mathématiques :

Source	*Somme de carrés*	*Degrés de liberté*	*Moyenne des carrés*	*Espérances des moyennes des carrés*
Expliquée	$\text{SCE} = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	k-1	$\text{MCE} = \dfrac{\text{SCE}}{k-1}$	$\sigma^2 + \dfrac{\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2}{k-1}$
Résiduelle	$\text{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	n-k	$\text{MCR} = \dfrac{\text{SCR}}{n-k}$	σ^2
Total	$\text{SCT} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$	n-1	$\text{MCT} = \dfrac{\text{SCT}}{n-1}$	$\sigma^2 + \dfrac{\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2}{n-1}$

Les espérances ci-dessus justifient le choix du rapport F = MCE/MCR, puisque MCE et MCR ont la même espérance si et seulement si H_o est vraie, alors que si H_o est fausse, MCE a tendance à être plus grand que MCR et le rapport F sera en conséquence grand. La distribution de F est connue sous H_o. Nous savons que lorsque H_o est vraie,

$$F = \frac{\text{MCE}}{\text{MCR}} = \frac{\text{SCE}/(k-1)}{SCR/(n-k)} \sim \mathcal{F}_{k\text{-}1;n\text{-}k}$$

Remarquez que l'espérance de MCR d'après la table d'analyse de variance est σ^2 et donc un estimateur sans biais de σ^2 est MCR :

$$\hat{\sigma}^2 = \text{MCR} = \frac{\text{SCR}}{n-k}$$

et donc la statistique F est une mesure réduite des écarts entre les moyennes $\bar{y}_i$:

$$F = \frac{\sum_{i=}^{k} n_i\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 \Big/ (k-1)}{\hat{\sigma}^2}.$$

Exemple numérique

Revenons à l'exemple 6.1.2. Désignons par Q_i la somme des carrés des écarts à l'intérieur de la classe i: $Q_i = \sum_{i=1}^{k}(y_{ij} - \bar{y}_i)^2$. Les résultats sont présentés dans le tableau ci-dessous. La moyenne globale des quatre classes est $\bar{y}$ = 15.

A	B	C	D	*Somme des carrés*
$\bar{y}_1 = 12$	$\bar{y}_2 = 17$	$\bar{y}_3 = 16$	$\bar{y}_4 = 15$	SCE = $\sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = 68$
$Q_1 = 38$	$Q_2 = 30$	$Q_3 = 12$	$Q_4 = 14$	SCR = $\sum_j Q_i = 94$
				SCT = 162

Voici la table d'analyse de variance:

Source	*Somme de carrés*	*dl*	*Moyenne des carrés*	*F*	*Espérances des moyennes des carrés*
Expliquée	SCE = 68	3	MCE = 68/3 = 22,67	$\frac{\text{MCE}}{\text{MCR}} = 4{,}34$	$\sigma^2 + \frac{\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2}{k-1}$
Résiduelle	SCR = 94	18	MCR = 94 /18 = 5,22		σ^2
Total	SCT = 162	21	MCT = 162/21 = 7,71		$\sigma^2 + \frac{\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2}{n-1}$

Le point critique de niveau α = 0,05 pour une $\mathcal{F}_{3,18}$ est $F_{3,18;0,05}$ = 3,16. Puisque F = 4,34 > 3,16, les différences sont significatives.

Voici les résultats d'une analyse de variance faite avec le logiciel R.

```
> anova(lm(y~bande))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value  Pr(>F)
bande      3 68.000  22.667  4.3404 0.01814 *
Residuals 18 94.000   5.222
```

La valeur p (colonne « `Pr(>F)` ») représente la probabilité, sous Ho, qu'une statistique de loi $\mathcal{F}_{3;18}$ prenne une valeur supérieure ou égale à 4,3404 est 0,01814. Ceci montre que les différences entre les bandes sont significatives, du moins au niveau de 2 %. ■

6.4 Estimation des paramètres

Les μ_i sont des paramètres qu'on peut également estimer par la méthode des moindres carrés. Nous choisissons les valeurs μ_i qui minimisent la somme des carrés des écarts entre les observations y_{ij} et leurs moyennes μ_{ij} qui dans ce modèle dépendent de i mais pas de j:

$$Q = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \mu_{ij})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2$$

Nous pouvons minimiser chaque somme $\sum_{j=1}^{n_i}(y_{ij} - \mu_i)^2$ séparément par rapport à μ_i, et il est bien

connu que la valeur $\hat{\mu}_i$ de μ_i qui minimise cette somme de carrés est

$$\hat{\mu}_i = \bar{y}_i$$

Pour déterminer un intervalle de confiance pour μ_i, nous utilisons le fait que $\bar{y}_i \sim \mathcal{N}(\mu_i ; \sigma^2/n_i)$ et que $T = \frac{\bar{y}_i - \mu_i}{\hat{\sigma}/\sqrt{n_i}} \sim t_{n-k}$. Donc un intervalle de confiance est donné par

$$\bar{y}_i - t_{n-k;\alpha/2}\hat{\sigma}_{\bar{y}_i} \le \mu_i \le \bar{y}_i + t_{n-k;\alpha/2}\hat{\sigma}_{\bar{y}_i}$$

où $\hat{\sigma}_{\bar{y}_i} = \frac{\hat{\sigma}}{\sqrt{n_i}}$. Nous avons $\hat{\sigma} = \sqrt{\text{MCR}} = \sqrt{5,22}$ et avec $\alpha = 0,05$, $t_{n-k;\alpha/2} = 2,101$. Un intervalle de confiance pour μ_1 est donné par $[12 - 2,101\frac{\sqrt{5,22}}{\sqrt{5}}; 12 - 2,101\frac{\sqrt{5,22}}{\sqrt{5}}] = [9,85 ; 14,15]$.

Certaines des quantités qui interviennent dans ces calculs sont fournies par le logiciel :

```
> a<-lm(y~bande)
> b<-predict.lm(a,data.frame(bande="1"),se.fit=T)
> b
$fit
[1] 12
$se.fit
[1] 1.021981
```

On a donc l'estimation ($\bar{y}_1$) = 12, et l'écart-type estimé de $\bar{y}_1$ est $\hat{\sigma}_{\bar{y}_1} = \hat{\sigma}/\sqrt{n_1}$ = 1,021981. Le point critique $t_{n-k;\alpha/2}$ = 2,1001 est donné par la commande

```
> talpha<-qt(.975,18)
> talpha
[1] 2.100922
```

L'intervalle de confiance est donc [12-2,1001(1,021981) ; 12+2,1001(1,021981)] = [9,85 ; 14,15].

En fait le calcul peut être entièrement fourni par R :

```
> b<-predict.lm(a,data.frame(bande="A"),se.fit=T,interval="confidence",level=.95)
> b
$fit
     fit      lwr      upr
[1,]  12 9.852898 14.14710
$se.fit
[1] 1.021981
```

On peut obtenir ces résultats pour plusieurs paramètres à la fois (l'intervalle est à 95 % si le niveau n'est pas précisé) :

```
> predict.lm(a,data.frame(bande=c("A","B","C","D")),interval="confidence")
  fit       lwr      upr
1  12  9.852898 14.14710
2  17 14.599467 19.40053
3  16 14.185368 17.81463
4  15 13.039973 16.96003
```

Une interprétation du numérateur de F

Nous pouvons considérer toute somme de carrés expliquée comme une différence de sommes de carrés résiduelles. La somme des carrés résiduelle est, par définition,

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\hat{\mu}_{ij})^2 ,$$

la somme des carrés des écarts entre les observations et leur moyenne estimée. Puisque

$$\hat{\mu}_{ij} = \hat{\mu}_i = \bar{y}_i$$

on a

$$\text{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_i)^2$$

Mais la somme des carrés totale est *elle aussi* une somme de carrés résiduelle : c'est la somme des carrés résiduelle dans le modèle $y_{ij} = \mu + \varepsilon_{ij}$, le modèle dans lequel les moyennes des groupes sont toutes égales. Mais ce modèle est précisément le modèle stipulé par l'hypothèse nulle. On peut donc écrire

$$\text{SCT} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij}-\bar{y})^2 = \text{SCR}_o$$

où la notation SCR_o signifie qu'il s'agit d'une somme de carrés résiduelle sous un ***modèle réduit***, réduit par les contraintes de H_o. La somme des carrés expliquée, qui est la différence SCT - SCR peut alors s'écrire comme

$$\text{SCE} = \text{SCR}_o - \text{SCR}$$

Si ν est le nombre de degrés de liberté de SCR et ν_o est le nombre de degrés de liberté de SCR_o, alors le rapport *F* devient

$$F = \frac{\left[\text{SCR}_o - \text{SCR}\right]/(\nu_o - \nu)}{\text{SCR}/\nu}$$

Cette formule est assez générale: le rapport *F* est toujours de cette forme. La différence SCR_o - SCR, la somme des carrés expliquée, représente la *réduction* d'erreur due à l'introduction du modèle plus complexe. La différence ν_o - ν représente la différence entre le nombre de degrés de liberté de SCR_o et le nombre de degrés de liberté de SCR.

Interprétation de *la statistique F lorsque les n_i sont égaux*

Si $n_1 = \ldots = n_k = r$, la moyenne des carrés expliquée peut s'écrire comme $\text{MCE} = \sum_{i=1}^{k} n_i(\bar{y}_{i.}-\bar{y}_{..})^2$ la statistique *F* peut s'écrire comme $F = \dfrac{r\sum_{i=1}^{k}(\bar{y}_i-\bar{y})^2/(k-1)}{\hat{\sigma}^2} = \dfrac{\sum_{i=1}^{k}(\bar{y}_i-\bar{y})^2/(k-1)}{\hat{\sigma}^2/r} = \dfrac{\sum_{i=1}^{k}(\bar{y}_i-\bar{y})^2/(k-1)}{\hat{\sigma}^2_{\bar{y}_i}}$. Le dénominateur est une estimation de la variance de $\bar{y}_i$. Le numérateur est la variance *échantillonnale* des $\bar{y}_i$, disons $S^2_{\bar{y}_i}$, et donc estime la même chose si et seulement si les moyennes sont égales. Sinon, $S^2_{\bar{y}_i}$ estime $\sigma^2_{\bar{y}_i} + \dfrac{\sum_{i=1}^{k}(\mu_i-\bar{\mu})^2}{k-1}$ et prendra une valeur d'autant plus élevée que H_o est « fausse ».

Une autre paramétrisation du modèle

Une autre façon d'exprimer les paramètres du modèle est

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

où les α_i satisfont

$$\sum_{i=1}^{k} \alpha_i = 0.$$

Puisque $E(y_{ij}) = \mu + \alpha_i$, nous avons forcément la relation $\mu_i = \mu + \alpha_{i.}$, ce qui permet d'exprimer chaque μ_i en fonction de μ et de α_i. Inversement, puisque $\Sigma\mu_i = n\mu + \Sigma\alpha_i = n\mu$, on a $\mu = \Sigma_i\mu_i/n$ et $\alpha_i = \mu_i - \mu$, ce qui permet d'exprimer μ et les α_i en fonction des μ_i.

Les deux modèles sont donc équivalents. L'intention dans cette deuxième paramétrisation est de décomposer la moyenne de la i^e classe en deux parties, l'une, μ, commune à toutes les classes, et l'autre, α_i, propre à la i^e classe. C'est une paramétrisation qui prévoit l'hypothèse qui sera testée, soit

$$\alpha_1 = \alpha_2 = \ldots = \alpha_k = 0.$$

Elle a cependant quelques inconvénients causés par l'introduction de k+1 au lieu de k paramètres: les k α_i et μ. Cet accroissement du nombre de paramètres n'est pas réel, puisqu'on introduit une contrainte qui ramène à k la dimension de l'espace des paramètres. Mais ces manipulations causent des difficultés inutiles.

6.5 Test d'ajustement à une droite

Revenons à la régression linéaire. Nous avons jusqu'ici posé, comme partie du modèle, l'hypothèse que l'espérance $E(y_i)$ est une fonction linéaire $\beta_o + \beta_1 x_i$ des x_i — sans autre évidence que le nuage de points. Dans certains cas, cependant, il est possible de soumettre cette supposition à un test statistique. C'est le cas où une même valeur x_i est accompagnée de plusieurs valeurs de y, $y_{i1}, y_{i2}, \ldots, y_{in_i}$. Soit $x_1, x_2, \ldots x_k$ les valeurs distinctes de x. (Il y a en tout $n = \Sigma n_i$ observations, mais seulement k valeurs distinctes de x.) Le modèle de régression s'écrit

$$\mathcal{M}_o: \; y_{ij} = \beta_o + \beta_1 x_i + \varepsilon_{ij}$$

Mais pour tester cette hypothèse de linéarité, nous commençons par un modèle plus général, dans lequel nous n'imposons pas la linéarité, soit

$$\mathcal{M}: \; y_{ij} = \mu_i + \varepsilon_{ij}$$

Or le modèle $\mathcal{M}$ est le modèle d'analyse de variance introduit dans ce chapitre et le modèle $\mathcal{M}_o$ est le modèle de régression introduit au chapitre 4. Nous pouvons, *dans le modèle* $\mathcal{M}$, tester l'hypothèse que le modèle $\mathcal{M}_o$ s'applique, c'est-à-dire, l'hypothèse linéaire

$$\mu_i = \beta_o + \beta_1 x_i$$

Le rapport F pour tester cette hypothèse aura pour numérateur une somme de carrés expliquée exprimée comme la différence de deux somme de carrés résiduelles, $SCE = SCR_o - SCR$, où SCR est simplement la somme des carrés résiduelle dans le modèle $\mathcal{M}$ et SCR_o est la somme des carrés résiduelle dans le modèle réduit par l'hypothèse nulle, soit le modèle $\mathcal{M}_o$. Nous avons déjà des formules pour ces sommes de carrés:

$$SCR = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

$$\text{SCR}_o = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\beta}_o - \hat{\beta}_1 x_i)^2$$

où $\hat{\beta}_\circ$ et $\hat{\beta}_1$ sont les estimateurs de β_o et β_1 définis au chapitre 4. Quelques manipulations algébriques permettent d'écrire la différence SCR_o - SCR de la manière instructive suivante:

$$\text{SCR}_o - \text{SCR} = \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\beta}_\circ - \hat{\beta}_1 x_i)^2$$

Cette somme de carrés devrait être petite si l'hypothèse $\mu_i = \beta_o + \beta_1 x_i$ est vraie, car $\bar{y}_i$ estime μ_i et $\hat{\beta}_o + \hat{\beta}_1 x_i$ estime $\beta_o + \beta_1 x_i$. Le nombre de degrés de liberté est n-k pour SCR et n-2 pour SCR_o. Donc SCR_o - SCR a k-2 degrés de liberté (ce qui s'explique: la somme a k termes et 2 paramètres estimés). La statistique F est donc

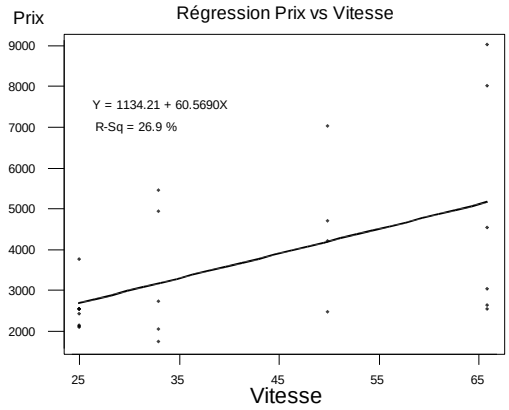
$$F = \frac{\left[\text{SCR}_o - \text{SCR}\right]/(k-2)}{\text{SCR}/(n-k)} \sim \mathcal{F}_{k\text{-}2;n\text{-}k}$$

***Exemple* 6.5.1** On a prélevé les données suivantes sur 24 ordinateurs afin d'analyser la relation entre la vitesse de l'ordinateur et son prix.

ID	*Vitesse (mhz)*	*Prix ($)*	*ID*	*Vitesse (mhz)*	*Prix ($)*
1	25	2 045 $	13	33	4 898 $
2	25	2 069 $	14	33	5 428 $
3	25	2 100 $	15	50	2 432 $
4	25	2 394 $	16	50	4 178 $
5	25	2 499 $	17	50	4 678 $
6	25	2 499 $	18	50	6 995 $
7	25	2 499 $	19	66	2 495 $
8	25	2 515 $	20	66	2 600 $
9	25	3 720 $	21	66	2 999 $
10	33	1 708 $	22	66	4 499 $
11	33	1 999 $	23	66	7 995 $
12	33	2 699 $	24	66	8 999 $

Le graphique suivant montre qu'il y a une certaine relation. Elle est plutôt faible, mais elle existe. La relation est-elle linéaire ?

Figure 6.5.1

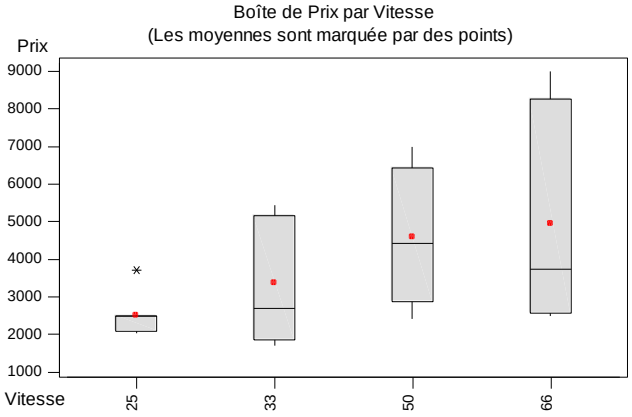


Nous faisons d'abord une analyse descriptive parallèle au raisonnement du test formel qui suivra. Nous commençons par adopter une modèle qui fait le moins d'hypothèses possibles, c'est-à-dire, on suppose seulement que les moyennes des 4 groupes (25 mhz, 33 mhz, 50 mhz, et 66 mhz) sont μ_1, μ_2, μ_3 et μ_4, sans aucune restriction sur les μ_i. On estime ces moyennes par les moyennes échantillonnales, qui sont

$$\bar{y}_1 = 2482{,}222;\ \bar{y}_2 = 3346{,}4;\ \bar{y}_3 = 4570{,}75;\ \bar{y}_4 = 4931{,}167.$$

L'hypothèse de linéarité est l'hypothèse que $\mu_i = \beta_o + \beta_1 x_i$, $i = 1, 2, 3, 4$, c'est-à-dire, que les 4 moyennes se situent sur une droite. Le graphique suivant présente les moyennes $\bar{y}_i$ ainsi qu'une boîte qui résume les données dans chaque classe :

Figure 6.5.2



Le modèle de régression linéaire suppose que $\mu_i = \beta_o + \beta_1 x_i$ et fournit une estimation des coefficients:

$$\text{Prix} = \hat{\beta}_o + \hat{\beta}_1 x_i = 1134 + 60{,}6 x_i$$

Dans ce modèle, l'estimation des moyennes des 4 groupes est:

1134+60,6(25) = 2648,432; 1134+60,6(33) = 3132,984; 1134+60,6(50)

= 4162,657; 1134+60,6(66) = 5131,760.

Nous devons donc comparer les deux séries d'estimation, celles basées sur le modèle d'analyse de variance (les $\bar{y}_i$) et celles basées sur le modèle de régression (les $\hat{\beta}_o + \hat{\beta}_1 x_i$):

Vitesse (i)	25 *mhz*	33 *mhz*	50 *mhz*	66 *mhz*
Effectif (n_i)	9	5	4	6
Modèle d'anova ($\bar{y}_i$)	2482,222	3346,4	4570,75	4931,167
Modèle de régression ($\hat{\beta}_o + \hat{\beta}_1 x_i$)	2648,432	3132,984	4162,657	5131,760

La statistique pour tester l'hypothèse de linéarité est

$$F = \frac{\left[\text{SCR}_o - \text{SCR}\right]/(k-2)}{\text{SCR}/(n-k)}$$

qui suit une loi $\mathcal{F}_{k\text{-}2;n\text{-}k}$ sous l'hypothèse de linéarité. Appliquant l'une des formules de SCR_o-SCR, nous obtenons

$$SCR_o - SCR = \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\beta}_o - \hat{\beta}_1 x_i)^2$$

$$= 9(2482{,}222\text{-}2648{,}432)^2 + 5(3346{,}4\text{-}3132{,}984)^2 + 4(4570{,}75\text{-}4162{,}657)^2 + 6(4931{,}167\text{-}5131{,}760)^2$$

$$= 1\ 383\ 951.$$

Le nombre de degrés de liberté est k-2 = 4 -2 = 2.

Quant à SCR, c'est la somme des carrés des écarts entre les observations et leur moyenne estimée, soit

$$SCR = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2 = \sum_{i=1}^{9}(y_{1j} - \bar{y}_1)^2 + \sum_{i=1}^{5}(y_{2j} - \bar{y}_2)^2 + \sum_{i=1}^{4}(y_{3j} - \bar{y}_3)^2 + \sum_{i=1}^{6}(y_{4j} - \bar{y}_4)^2$$

$$= 2\ 049\ 806 + 11\ 659\ 489 + 10\ 616\ 995 + 41\ 223\ 625 = 65\ 549\ 914$$

Le nombre de degrés de liberté de SCR est n-k = 24-4 = 20.

Donc la valeur de F est

$$F = \frac{\left[SCR_o\text{-}SCR\right]/(k-2)}{SCR/(n-k)} = \frac{\left[66933866 - 65549914\right]/(4-2)}{65549914/(24-4)} = 0{,}21,$$

ce qui, à 2 et 20 degrés de liberté, est non significatif : on ne rejette pas l'hypothèse de linéarité.

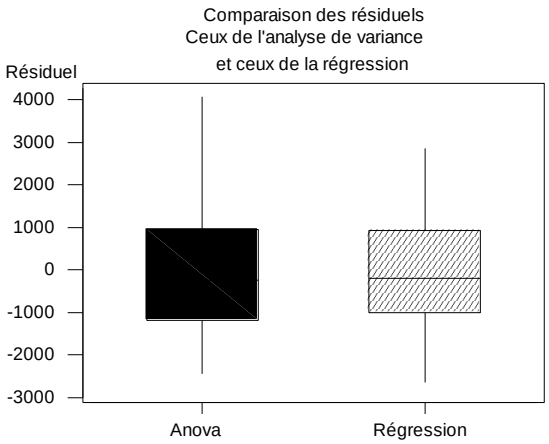
Ceci complète le test.

Remarque. *On peut montrer que*

$$SCR_o - SCR = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\beta}_\circ - \hat{\beta}_1 x_i)^2 - \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2,$$

ce qui signifie que le test est basé sur une comparaison des résiduels sous le modèle de régression avec les résiduels sous un modèle d'analyse de variance. Le modèle de régression étant plus restrictif, la somme des carrés résiduelle est supérieure ou égale à celle d'un modèle d'analyse de variance. Mais la différence ne devrait pas être importante si le modèle de régression est bon. C'est ce qui explique pourquoi cette différence figure au numérateur de la statistique F. Si le modèle de régression est incorrect, les observations s'éloigneraient des estimations $\hat{\beta}_\circ + \hat{\beta}_1 x_i$ *plus que ne le feraient les* $\bar{y}_i$ *; par conséquent les résiduels seraient importants et* SCR_o *serait bien plus grand que* SCR. *Une comparaison visuelle montre pourquoi on ne rejette pas l'hypothèse de linéarité : les résiduels des deux modèles ne sont pas très différents* :

Figure 6.5.3

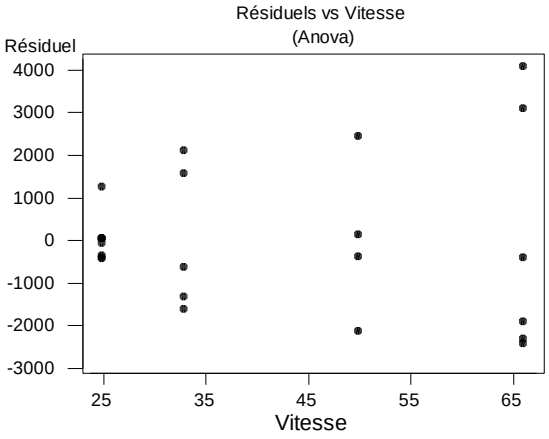


En passant, on peut aussi montrer que

$$\text{SCR}_o \text{ - SCR } = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left[(y_{ij} - \hat{\beta}_\circ - \hat{\beta}_1 x_i) - (y_{ij} - \bar{y}_i) \right]^2$$

Remarque *Si la figure* 5.2 *ne montre pas d'évidence de non linéarité, elle suggère en revanche que les variances varient avec la vitesse. Le graphique suivant le montre encore. On verra maintenant que dans le cas présent il est possible de* tester *l'hypothèse d'homoscédasticité.*

Figure 6.5.4



6.6 Test d'homogénéité de variances

Supposons encore que nous n'ayons de k valeurs distinctes de x : $x_1, \ldots, x_k$, et que pour x_i on avait n_i valeurs $y_{i,1}, \ldots, y_{i,n_i}$ correspondantes, où $n_i > 1$ pour chaque i. Il est alors possible de *tester* l'hypothèse d'homoscédasticité, une hypothèse qui autrement ferait partie du modèle et serait supposée vraie sans démonstration. On commence donc avec un modèle qui ne suppose pas l'homoscédasticité, soit

$$y_{ij} = \mu_i + \varepsilon_{ij}, \text{ où } \varepsilon_{ij} \sim \mathcal{N}(0 ; \sigma_i^2).$$

Homogénéité de variances

Dans ce modèle, on teste l'hypothèse

$$H_o : \sigma_1^2 = \ldots = \sigma_k^2$$

Si λ est le rapport des maximums de vraisemblance, alors si les n_i sont assez grands, la statistique

$$Q = -2 \ln \lambda$$

suit à peu près une loi χ^2 à k-1 degrés de liberté lorsque H_o est vraie.

Il est aisé de vérifier (lorsqu'on remplace les estimateurs du maximum de vraisemblance par les estimateurs sans biais) que

$$Q = (n-k)\ln(s_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(s_i^2)$$

où

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i.)^2}{n_i - 1} \text{ et } s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k} = \text{MCR.}$$

Nous effectuons les calculs avec les données de l'exemple sur les ordinateurs. Il ne faut pas oublier, cependant, que le test est approximatif, puisqu'il est basé sur l'hypothèse que les n_i sont grands, ce qui n'est pas le cas; et également sur l'hypothèse que les données sont normales, ce dont on ne peut pas être sûr.

On a n = 24 et k = 4. Les valeurs des n_i sont 9, 5, 4, et 6.

Les valeurs des s_i^2 sont

256 226 ; 2 914 872 ; 3 538 998 ; et 8 244 725

et s_p^2 = 3277496. Finalement, Q = 16,01646.

Le niveau de signification est $P(\chi_3^2 > 16{,}01646) = 0{,}001$, ce qui est hautement significatif. On peut rejeter l'hypothèse d'homoscédasticité. Cette conclusion est conforme à ce qu'on voit dans le graphique des résiduels, qui semble indiquer clairement que la dispersion des prix augmente avec les vitesses (et le fait, aussi, que les s_i^2 croissent de façon également importante).

6.7 Combinaisons linéaires des moyennes

Il est possible également d'estimer des combinaisons linéaires des μ_i et de tester des hypothèses à propos de ces combinaisons linéaires. Soit $\varphi = \Sigma_i c_i \mu_i$ une combinaison linéaire avec coefficients fixes c_i. Un estimateur sans biais $\hat{\varphi}$ de φ est $\sum c_i \hat{\mu}_i = \sum_i c_i \bar{y}_i$. La distribution de $\hat{\varphi}$ s'obtient facilement: $\hat{\varphi}$ est une fonction linéaire des $\bar{y}_i$, qui à leur tour sont fonctions linéaires des y_i. Donc $\hat{\varphi}$ est normale, de moyenne φ et de variance $\Sigma_i(c_i\sigma)^2/n_i$. La variable

$$Z = \frac{\hat{\varphi} - \varphi}{\sigma\sqrt{\sum c_i^2 / n_i}} \sim \mathcal{N}(0\ ;\ 1)$$

Puisque σ^2 n'est pas connue, nous remplaçons Z par

$$T = \frac{\hat{\varphi} - \varphi}{\hat{\sigma}\sqrt{\sum c_i^2 / n_i}}\ ,$$

et on peut démontrer que

$$T \sim t_{n-k}$$

Ceci nous permet de déterminer un intervalle de confiance pour φ et de tester des hypothèses du genre H_o: $\varphi = \varphi_o$. En particulier, nous pouvons tester des hypothèses à propos de la différence entre 2 des moyennes, ou encore à propos de certaines moyennes particulières.

***Exemple* 6.7.1** Dans le problème traité au début du chapitre, supposons que le traitement A est particulier dans le sens que c'est le traitement employé régulièrement, alors que les trois autres sont des traitements expérimentaux. On veut donc comparer le traitement traditionnel à l'ensemble des traitements expérimentaux, c'est-à-dire, on veut tester l'hypothèse

$$H_o: \mu_1 = (1/3)(\mu_2 + \mu_3 + \mu_4), \text{ ou encore, } H_o: 3\mu_1 - \mu_2 - \mu_3 - \mu_4 = 0.$$

On a

$$T = \frac{3(12) - 17 - 16 - 15}{\sqrt{5,22}\sqrt{\frac{3^2}{5} + \frac{1}{4} + \frac{1}{7} + \frac{1}{6}}} = -3,43$$

Puisque $t_{18;0,05} = 2,101$, on rejette H_o à 5%. On peut donc conclure que l'indice de distorsion des bandes expérimentales est supérieur, en moyenne, à celui des bandes traditionnelles.

Analyse de variance à deux facteurs – facteurs croisés

On présente ici quelques cas particuliers d'un modèle linéaire général dans lequel la variable endogène est exprimée en fonction de deux variables exogènes qualitatives.

6.8 Décomposition des sommes de carrés

Les résultats d'une expérience sont souvent classés selon plus d'un facteur, comme dans l'exemple suivant.

***Exemple* 6.8.1** [Battacharrya, Gouri K., Johnson, Richard A. (1977) *Statistical concepts and methods*, Wiley, New York, *p.*498]. Considérons l'expérience suivante, dont l'objet est de déterminer l'effet de deux hormones sur le poids des cobayes. Les deux hormones et les quantités administrées sont

Hormone A (Œstradiol)	0	0,5 *mg/jour*	
Hormone *B* (Progestérone)	0	0,1 *mg/jour*	10 *mg/jour*

Le plan d'expérience est appelé « plan factoriel » ; les deux facteurs A et B sont « croisés », dans le sens que nous avons des sujets pour chaque combinaison d'un niveau de A et un niveau de B. La variable observée *Y* est le gain de poids durant la période d'observation. Les données sont les suivantes :

	Hormone B (*Progestérone*)		
Hormone A *(Œstradiol)*	0	0,1 *mg/jour*	10 *mg/jour*
0 *mg/jour*	-19 -11 -14	8 -18 -9	7 23 23
0,5 *mg/jour*	-10 -19 -28	-3 -10 -4	32 29 18

■

Le modèle est dit *équilibré* lorsqu'il y a un même nombre r d'observations dans chaque case. Ici $r = 3$. Le modèle s'exprime de la façon suivante. Soit y_{ijk} la k^e observation du niveau i du facteur A et du niveau j du facteur B. On suppose que les y_{ijk} sont indépendantes et que

$$\mathrm{E}(y_{ijk}) = \mu_{ij} \; ; \; \mathrm{Var}(y_{ijk}) = \sigma^2, \; i = 1, \ldots, a; \; j = 1, \ldots, b \tag{6.8.1}$$

On présente également le modèle de la façon équivalente suivante:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ij}, \; \varepsilon_{ijk} \sim \mathcal{N}(0 \; ; \; \sigma^2), \; \varepsilon_{ij} \text{ indépendantes.} \tag{6.8.2}$$

Il est possible de traiter ces données à l'aide d'un modèle d'analyse de variance simple à un facteur de $a \times b$ niveaux, auquel cas on obtient tout de suite une décomposition de la somme des carrés totale:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \bar{y}...)^2 \; = \; r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\bullet} - \bar{y}...)^2 \; + \; \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{ij\bullet})^2$$

où $\bar{y}_{ij\bullet}$ est la moyenne des données de la case i,j, $\bar{y}...$ est la moyenne de toutes les données. Avec les sigles habituels, cette décomposition s'écrit

$$\mathrm{SCT} \; = \; \mathrm{SCE} \; + \; \mathrm{SCR}$$

où SCE a ab-1 degrés de liberté. La moyenne $\mathrm{MCE} = \dfrac{\mathrm{SCE}}{ab-1}$ peut servir au numérateur d'une statistique F pour tester l'hypothèse que les $a \times b$ moyennes sont toutes égales. Le dénominateur est toujours $\mathrm{MCR} = \dfrac{\mathrm{SCR}}{ab(r-1)}$. Mais lorsque les données sont classées selon *deux* facteurs, cette décomposition est insuffisante: lorsqu'on conclut qu'il y a un effet significatif, il importe de savoir s'il s'agit de l'effet de l'un des facteurs, de l'autre, ou d'une interaction entre les deux. La somme SCE, qui mesure les écarts entre les $a \times b$ moyennes, se décompose en trois parties, désignées par SCA, SCB et SCAB

$$r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\bullet}-\bar{y}_{\bullet\bullet})^2 = br\sum_{i=1}^{a}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2 + ar\sum_{j=1}^{b}(\bar{y}_{\bullet j}-\bar{y}_{\bullet\bullet})^2 + r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\bullet}-\bar{y}_{i\bullet}-\bar{y}_{\bullet j}+\bar{y}_{\bullet\bullet})^2$$

SCE = SCA + SCB + SCAB

où $\bar{y}_{i\bullet}$ est la moyenne des données de la i^e ligne et $\bar{y}_{\bullet j}$ est la moyenne des données de la j^e colonne. SCA, qui est la dispersion entre les différents niveaux de *A*, a (*a*-1) degrés de liberté, SCB, la dispersion entre les différents niveaux de *B*, a (*b*-1) degrés de liberté, et SCAB, une mesure des «interactions» entre *A* et *B*, a (*a*-1)(*b*-1) degrés de liberté. Considérons les trois hypothèses suivantes:

H_A: Le facteur A n'a pas d'effet: $\mu_{1\bullet} = ... = \mu_{a\bullet}$, $\mu_{i\bullet} = \Sigma_j \mu_{ij}/b$

H_B: Le facteur B n'a pas d'effet: $\mu_{\bullet 1} = ... = \mu_{\bullet b}$, $\mu_{\bullet j} = \Sigma_i \mu_{ij}/a$

H_{AB}: Aucune interaction entre A et B: $\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu = 0 \ \forall_{ij}$, où $\mu = \Sigma_i \Sigma_j \mu_{ij}/ab$.

Une autre paramétrisation

Les paramètres naturels dans ce contexte sont, à part la variance σ^2, les $a \times b$ moyennes μ_{ij} des groupes. Mais traditionnellement, on désigne plutôt certaines fonctions linéaires des moyennes comme paramètres et l'équation (6.8.2) est écrite plutôt comme

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk},\ i = 1,\ldots,a\ ;\ j = 1,\ldots,b \qquad (6.8.3)$$

À part μ, ces paramètres ne sont pas des moyennes : ce sont des *différences* de moyennes. Le sens de ces paramètres est:

α_i : l'effet du i^e niveau du facteur A
β_j : l'effet du j^e niveau du facteur B
γ_{ij} : l'effet de l'interaction entre le i^e niveau de A et le j^e de B

Formellement, en termes des moyennes, voici comment se définissent les paramètres $\{\mu, \alpha, \beta, \gamma\}$:

- $\mu = \mu$ (même sens dans les deux paramétrisations)
- $\alpha_i = \mu_{i\bullet} - \mu$
- $\beta_j = \mu_{\bullet j} - \mu$
- $\gamma_{ij} = \mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu$

Ces paramètres sont définis en prévision des hypothèses qu'on souhaite normalement tester, soit les trois hypothèses classiques H_A, H_B et H_C. Car celles-ci s'expriment très simplement comme ceci:

H_A: $\alpha_1 = ... = \alpha_a = 0$
H_B: $\beta_1 = ... = \beta_b = 0$
H_{AB}: $\gamma_{ij} = 0$ pour tout *i* et tout *j*.

Remarque *Dans cette paramétrisation, on a* a *alpha,* b *bêta, et ab gamma, ce qui donnerait, à première vue,* $1 + a + b + ab = (a+1)(b+1)$ *paramètres — alors que les* $\{\mu_{ij}\}$ *sont au nombre de ab. On aurait créé de nouveaux paramètres. Il n'en est rien : l'espace paramétrique demeure de dimension ab, tout comme l'espace des moyennes* $\{\mu_{ij}\}$, et ce *grâce à certaines restrictions qu'on impose aux* α, β *et* γ, *soit :* $\Sigma_i \alpha_i = 0$; $\Sigma_j \beta_j = 0$; $\Sigma_i \gamma_{ij} = 0$ *pour tout j* ; $\Sigma_j \gamma_{ij} = 0$ *pour tout i. Ces restrictions découlent de la signification que nous avons donnée aux paramètres. Par exemple, la somme* $\Sigma_i \alpha_i$ *est nécessairement nulle si* $\alpha_i = \mu_{i\bullet} - \mu_{\bullet}$. ■

6.9 Tests d'hypothèses

On peut démontrer que:

1. SCA, SCB, SCAB et SCR sont indépendantes.
2. $SCA/\sigma^2 \sim \chi^2_{a-1}(\lambda_A)$, où $\lambda_A = br\Sigma_i(\mu_{i.} - \mu)^2/\sigma^2$.
3. $SCB/\sigma^2 \sim \chi^2_{b-1}(\lambda_B)$, où $\lambda_B = ar\Sigma_j(\mu_{.j} - \mu)^2/\sigma^2$.
4. $SCAB/\sigma^2 \sim \chi^2_{(a-1)(b-1)}(\lambda_{AB})$, où $\lambda_{AB} = r\Sigma_i\Sigma_j(\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu)^2/\sigma^2$.
5. $SCR/\sigma^2 \sim \chi^2_{ab(r-1)}$ centrale.

Les tests appropriés pour H_A, H_B, et H_{AB} découlent directement de ces propriétés.

Les résultats de l'analyse sont traditionnellement présentés sous la forme d'une ***table d'analyse de variance*** qui, pour une analyse à deux facteurs croisés, prend la forme suivante:

Tableau 6.9.1
Table d'analyse de variance

Source	*Somme de carrés*	*Degrés de liberté*	*Moyenne de carrés*	*Espérances des moyennes de carrés*
Facteur A	$SCA = br\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2$	a-1	$MCA = SCA/(a\text{-}1)$	$\sigma^2 + \frac{br\sum_{i=1}^{a}(\mu_{i\cdot} - \mu)^2}{a-1}$
Facteur B	$SCB = ar\sum_{j=1}^{b}(\bar{y}_{\cdot j} - \bar{y})^2$	b-1	$MCB = SCB/(b\text{-}1)$	$\sigma^2 + \frac{ar\sum_{j=1}^{b}(\mu_{\cdot j} - \mu)^2}{b-1}$
Interactions	$SCAB = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots})^2$	$(a\text{-}1)(b\text{-}1)$	$MCAB = SCAB/(a\text{-}1)(b\text{-}1)$	$\sigma^2 + \frac{r\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu)^2}{(a-1)(b-1)}$
Résiduel	$SCR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk} - \bar{y}_{ij\cdot}\right)^2$	$ab(r\text{-}1)$	$MCR = SCR/ab(r\text{-}1)$	σ^2
Total	$SCT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk} - \bar{y}_{\cdots}\right)^2$	abr-1	$MCT = SCT/(n\text{-}1)$	

Les statistiques pour tester les hypothèses H_A, H_B, et H_{AB} sont

$$F_A = \frac{MCA}{MCR} \qquad F_B = \frac{MCB}{MCR} \qquad F_{AB} = \frac{MCAB}{MCR}$$

et les espérances ci-dessus justifient les régions critiques

$$F_A \sim \mathcal{F}_{a\text{-}1,ab(r\text{-}1);\alpha}\ ,\ F_B \sim \mathcal{F}_{b\text{-}1,ab(r\text{-}1);\alpha} \text{ et } F_{AB} \sim \mathcal{F}_{(a\text{-}1)(b\text{-}1),ab(r\text{-}1);\alpha}\ .$$

Remarque *Les statistiques F peuvent s'exprimer comme un quotient de deux estimateurs de la variance de certaines moyennes. Par exemple,* $F_A = \frac{br\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 \Big/ (a-1)}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 \Big/ (a-1)}{\hat{\sigma}^2 \Big/ br} =$ $\frac{\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})^2 \Big/ (a-1)}{\hat{\sigma}^2_{\bar{y}_{i\cdot\cdot}}}$ *a pour dénominateur un estimateur sans biais de la variance* $\sigma^2_{\bar{y}_i}$ *de* $\bar{y}_{i\cdot\cdot}$ *; et pour*

numérateur la variance échantillonnale des $\bar{y}_{i..}$ *, laquelle est sans biais pour* $\sigma^2_{\bar{y}_i}$ *si et seulement si* H_A *est vraie.* ■

Voici la table d'analyse de variance (par le logiciel **R**) pour les données sur l'effet des deux hormones sur le gain de poids des cobayes :

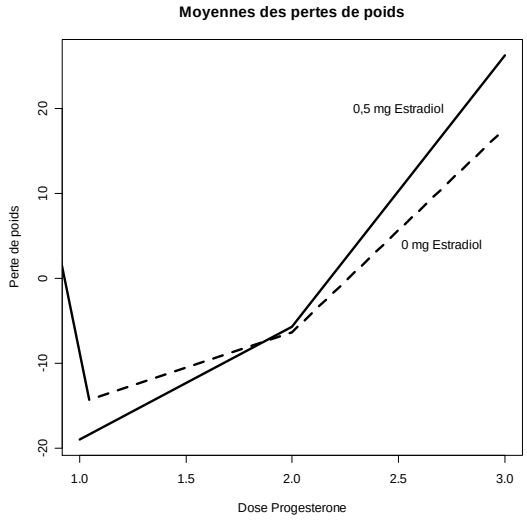
```
> anova(lm(z~oestradiol*progesterone))
                      Df Sum Sq Mean Sq F value    Pr(>F)
oestradiol             1   12.5    12.5  0.1762    0.6821
progesterone           2 4818.8  2409.4 33.9616 1.146e-05 ***
oestradiol:progesterone  2  129.0    64.5  0.9092    0.4289
Residuals             12  851.3    70.9
```

La seule valeur significative est celle pour la progestérone — un niveau de signification inférieur à 0,0005. Le niveau de signification est 0,682 pour l'œstradiol et 0,429 pour les interactions. Donc la progestérone stimule la croissance, l'œstradiol n'a aucun effet. L'effet de la progestérone est le même, quelle que soit la quantité de l'œstradiol consommée.

On peut faire une analyse visuelle de l'interaction à partir du tableau suivant, qui représente les pertes moyennes de poids selon le traitement suivi. Les données dans les cellules sont les moyennes échantillonnales $\bar{y}_{ij}$, qui estiment les μ_{ij}:

	Progestérone			
Œstradiol	0 *mg*	0,1 *mg*	10mg	*Moyenne*
0 mg	-14,66667	-6,333333	17,66667	-1,1111
0,5 mg	-19,00000	-5,666667	26,33333	0,5556
Moyenne	-16,8333	-6,0000	22,0000	-0,2778

Voici une représentation graphique de ces moyennes :



L'interaction existe au niveau de l'échantillon : les droites ne sont pas parallèles. Mais nous venons de conclure que cette absence de parallélisme n'est pas significative, c'est-à-dire qu'il est possible que dans la population les deux lignes soient en fait parallèles.

6.10 Le sens de l'hypothèse de non interaction

L'absence d'interactions est définie formellement par les équations

$$\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} = 0 \text{ pour tout } i, j$$

Mais on peut démontrer que cette formulation est équivalente à

pour toute paire j, j', la différence $\mu_{ij} - \mu_{ij'}$ est indépendante de i

ou encore

pour toute paire i, i', la différence $\mu_{ij} - \mu_{i'j}$ est indépendante de j

Le premier énoncé $\Rightarrow \mu_{ij} = \mu_{i\cdot} + \mu_{\cdot j} - \mu_{\cdot\cdot}$, $\mu_{ij'} = \mu_{i\cdot} + \mu_{\cdot j'} - \mu_{\cdot\cdot}$ et donc $\mu_{ij} - \mu_{ij'} = \mu_{\cdot j} - \mu_{\cdot j'}$, indépendante de i. Réciproquement, si $\mu_{ij} - \mu_{ij'}$ est indépendante de i alors $\mu_{ij} - \mu_{ij'} = \mu_{\cdot j} - \mu_{\cdot j'} \Rightarrow (1/b)\Sigma_{j'}(\mu_{ij} - \mu_{ij'}) = (1/b)\Sigma_{j'}(\mu_{\cdot j} - \mu_{\cdot j'}) \Rightarrow \mu_{ij} - \mu_{i\cdot} = \mu_{\cdot j} - \mu_{\cdot\cdot} \Rightarrow \mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} = 0$. Donc l'absence d'interactions signifie que la différence entre deux niveaux du facteur A est constante: elle ne dépend pas du niveau du facteur B (A et B peuvent, bien sûr, être échangés dans l'énoncé.) L'hypothèse de non interaction est aussi appelée ***hypothèse d'additivité*** puisqu'elle signifie que

$$\mu_{ij} - \mu_{\cdot\cdot} = (\mu_{i\cdot} - \mu_{\cdot\cdot}) + (\mu_{\cdot j} - \mu_{\cdot\cdot}) = \alpha_i + \beta_j$$

C'est donc l'hypothèse que l'effet combiné des deux traitements ($\mu_{ij} - \mu_{\cdot\cdot}$) est la *somme* de deux effets: celui du traitement A ($\mu_{i\cdot} - \mu_{\cdot\cdot}$) et celui du traitement B ($\mu_{\cdot j} - \mu_{\cdot\cdot}$).

6.11 Une hypothèse particulière

L'hypothèse H_B pourrait manquer d'intérêt en présence d'interactions car elle concerne l'égalité de l'effet *moyen* du progestérone, la moyenne étant prise sur les deux niveaux d'œstradiol. Il pourrait être plus intéressant de tester l'hypothèse que la progestérone n'a pas d'effet lorsque le niveau d'œstradiol est fixe, par exemple, en l'absence d'œstradiol, ou en présence de 0,5 mg d'œstradiol. Soit

$H_{A1|B}$: La progestérone n'a pas d'effet en l'absence de l'œstradiol : $\mu_{11} = \mu_{12} = \mu_{13}$
$H_{A2|B}$: La progestérone n'a pas d'effet en présence de 0,5 mg d'œstradiol : $\mu_{21} = \mu_{22} = \mu_{23}$

La somme des carrés du numérateur pour tester la première hypothèse est la dispersion des trois moyennes de la première ligne du tableau :

$$SCA_1|B = 3\{[-14,667-(-1,111)]^2 + [6,333-(-1,111)]^2 + [17,667-(-1,111)]^2\} = 1690,889.$$

La statistique pour tester cette hypothèse est

$$F = \frac{SCA_1|B/2}{MCR} = \frac{1690,889/2}{70,9} = 11,91699,$$

ce qui, à 2 et 12 degrés de liberté, correspond à une p-valeur de 0,0014.

De la même façon, on obtient pour $H_{A2|B}$ la statistique

$$F = \frac{SCA_2|B/2}{MCR} = \frac{3256,889/2}{70,9} = 22,9538,$$

ce qui, à 2 et 12 degrés de liberté correspond à une p-valeur de 0,000079.

Donc les deux hypothèses, $H_{A1|B}$ et $H_{A2|B}$ sont chacune rejetée individuellement. Mais on pourrait aussi tester les deux hypothèses simultanément :

$$H_{A|B} : H_{A1|B} \textit{ et } H_{A2|B} : \mu_{11} = \mu_{12} = \mu_{13} \textit{ et } \mu_{21} = \mu_{22} = \mu_{23}.$$

La statistique F pour tester $H_{A|B}$ est

$$\frac{[SCA_1|B + SCA_2|B]/4}{MCR} = \frac{[1690{,}889 + 3256{,}889]/4}{79{,}9} = \frac{4947{,}778/4}{79{,}9} = \frac{1236{,}944}{79{,}9} = 17{,}43540.$$

À 4 et 12 degrés de liberté, ceci correspond à une p-valeur de 0,0000612.

Il est utile de remarquer ici une autre formulation de l'hypothèse $H_{A|B} : H_{A1|B}$ *et* $H_{A2|B}$. Elle est tout à fait équivalente à H_B *et* H_{AB}. Elle peut donc être testée par la statistique

$$F = \frac{[SCB+SCAB]/(2+2)}{MCR} = \frac{[4818{,}8+129{,}0]/(2+2)}{70{,}9} = 17{,}43540, \text{ la même statistique.}$$

Formulation générale Considérons une classification à deux facteurs, A et B ayant a et b niveaux, respectivement, et r observations par case. Le tableau suivant présente une autre décomposition de la somme des carrés totale.

Tableau 6.11.1
Table d'analyse de variance alternative

Source	*Somme de carrés*	*Degrés de liberté*	*Moyenne de carrés*	*Espérances des moyennes de carrés*
Facteur A	$SCA = br\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2$	a-1	$MCA = \frac{SCA}{a-1}$	$\sigma^2 + \frac{br\sum_{i=1}^{a}(\mu_{i\cdot\cdot} - \mu_{\cdot\cdot\cdot})^2}{a-1}$
Facteur A|B	$SCA\|B = SCB + SCAB = r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2$	$a(b$-1)	$MCA\|B = \frac{SCB+SCAB}{a(b-1)}$	$\sigma^2 + \frac{r\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij\cdot} - \mu_{i\cdot\cdot})^2}{a(b-1)}$
Résiduel	$SCR = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk} - \bar{y}_{ij\cdot}\right)^2$	$ab(r$-1)	$MCR = \frac{SCR}{ab(r-1)}$	σ^2
Total	$SCT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk} - \bar{y}_{\cdot\cdot\cdot}\right)^2$	abr-1	$MCT = \frac{SCT}{n-1}$	

6.12 Estimation dans un modèle restreint par une hypothèse

Si certaines des hypothèses considérées semblent, en vertu des données et des tests, très plausibles, il est tentant de les adopter comme partie du modèle. Ceci permet de simplifier la description du phénomène étudié. Une hypothèse qui simplifie considérablement le modèle est l'hypothèse de non-interaction. Si on l'adopte, le modèle devient

$$E(y_{ijk}) = \mu_{ij} = \mu_{i\cdot} + \mu_{\cdot j} - \mu$$

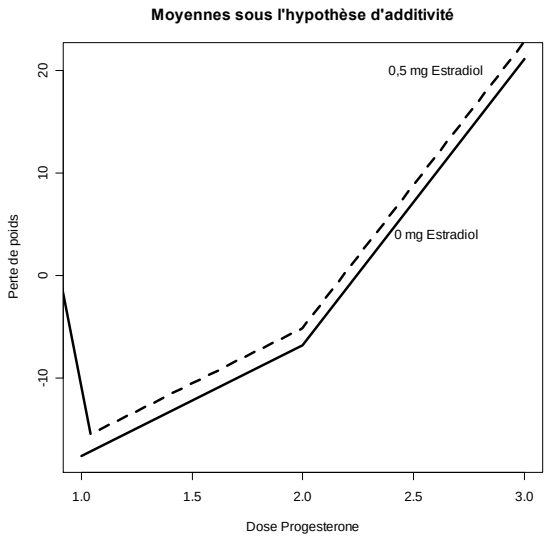
Dans ce cas, les moyennes μ_{ij} sont estimées différemment. On admet facilement (et on peut le justifier formellement) que les moyennes des marges (c'est-à-dire, les $\mu_{i\cdot}$ et les $\mu_{\cdot j}$) sont estimées par les moyennes échantillonnales correspondantes. On a donc le tableau suivant en premier temps:

	Progestérone			
Œstradiol	0 *mg*	0,1 *mg*	10 *mg*	*Moyenne*
0 mg				$\hat{\mu}_{1.}$ = -1,1111
0,5 mg				$\hat{\mu}_{2.}$ = 0,5556
Moyenne	$\hat{\mu}_{.1}$ = -16,8333	$\hat{\mu}_{.2}$ = -6,0000	$\hat{\mu}_{.3}$ = 22,0000	$\hat{\mu}$ = -0,2778

à part σ^2, les paramètres estimés dans le tableau ci-dessus sont en fait les seuls paramètres à estimer, puisque les μ_{ij} sont toutes fonctions des $\mu_{i.}$ et des $\mu_{.j}$. En utilisant la relation $\hat{\mu}_{ij} = \hat{\mu}_{i.} + \hat{\mu}_{.j} - \hat{\mu}$, on obtient les estimations suivantes:

	Progestérone			
Œstradiol	0 *mg*	0,1 *mg*	10 *mg*	*Moyenne*
0 mg	-17,66667	-6,833333	21,16667	$\hat{\mu}_{1.}$ = **-1,1111**
0,5 mg	-16,00000	-5,166667	22,83333	$\hat{\mu}_{2.}$ = **0,5556**
Moyenne	$\hat{\mu}_{.1}$ = **-16,8333**	$\hat{\mu}_{.2}$ = **-6,0000**	$\hat{\mu}_{.3}$ = **22,0000**	$\hat{\mu}$ = **-0,2778**

Voici une représentation graphique de ces moyennes:



Les droites sont parallèles, et c'est précisément le sens de non-interaction.

6.13 Suggestion pour le calcul des espérances

Section omise dans la version courte

6.14 Expression matricielle

Nous présentons ici une expression matricielle du modèle, ainsi qu'une discussion sur l'orthogonalité des effets. Les question auxquelles on répond sont 1) quel est le lien entre l'orthogonalité et l'additivité des sommes de carrés? et 2) quel est le lien entre les effectifs des cases et l'orthogonalité? Pour concrétiser, considérons les données suivantes, classées selon deux facteurs :

Facteur A	B_1		B_2		B_3	
A_1	y_{111}	y_{112}	y_{121}	y_{122}	y_{131}	y_{132}
A_2	y_{211}	y_{212}	y_{221}	y_{222}	y_{231}	y_{232}

Posons $\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij,}$ avec les contraintes $\sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$, de sorte qu'on peut écrire les moyennes comme ceci :

		Facteur B			
		B_1	B_2	B_3	Moyennes
Facteur A	A_1	$\mu+\alpha_1+\beta_1+\gamma_{11}$	$\mu+\alpha_1+\beta_2+\gamma_{12}$	$\mu+\alpha_1-\beta_1-\beta_2-\gamma_{11}-\gamma_{12}$	$\mu+\alpha_1$
	A_2	$\mu-\alpha_1+\beta_1-\gamma_{11}$	$\mu-\alpha_1+\beta_2-\gamma_{12}$	$\mu-\alpha_1-\beta_1-\beta_2+\gamma_{11}+\gamma_{12}$	$\mu-\alpha_1$
	Moyennes	$\mu+\beta_1$	$\mu+\beta_2$	$\mu-\beta_1-\beta_2$	μ

Le modèle d'analyse de variance à deux facteurs peut s'écrire en langage matricielle $E(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta}$ où

$$\begin{bmatrix} y_{111} \\ y_{112} \\ y_{121} \\ y_{122} \\ y_{131} \\ y_{132} \\ y_{211} \\ y_{212} \\ y_{221} \\ y_{222} \\ y_{231} \\ y_{232} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \mu \\ \alpha_1 \\ \beta_1 \\ \beta_2 \\ \gamma_{11} \\ \gamma_{12} \end{bmatrix}, \text{ et } \mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & 0 & 1 & 0 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix}.$$

Partitionnons $\mathbf{X}$ selon les effets principaux A, B et leurs interactions AB :

$\mathbf{X} = [\mathbf{X}_o \mid \mathbf{X}_1 \mid \mathbf{X}_2 \mid \mathbf{X}_3]$,

les partitions ayant, respectivement, 1, 1, 2 et 2 colonnes, correspondant à μ, à α, aux β et aux γ.

Soit

$\mathbf{P}_i$ le projecteur orthogonal sur $\mathcal{C}(\mathbf{X}_i)$, $i = 0, 1, 2, 3$,

$\mathbf{P}_{ij}$ le projecteur sur $\mathcal{C}([\mathbf{X}_i \mid \mathbf{X}_j])$,

$\mathbf{P}_{ij\ell}$ le projecteur sur $\mathcal{C}([\mathbf{X}_i \mid \mathbf{X}_j \mid \mathbf{X}_\ell])$, etc.,

et soit

$$\mathbf{P} = \mathbf{P}_{0123}.$$

Pour généraliser quelque peu, supposons que les matrices $\mathbf{X}_o$, $\mathbf{X}_1$, $\mathbf{X}_2$ et $\mathbf{X}_3$ ont, respectivement, 1, a, b, et $(a-1)(b-1)$ colonnes. Les sommes de carrés au numérateur des statistiques F pour tester les hypothèses habituelles H_A ($\alpha_1 = 0$), H_B ($\beta_1 = \beta_2 = 0$), et H_{AB} ($\gamma_{12} = \gamma_{12} = 0$), sont

$$F_A = \frac{\text{SCA} / (a-1)}{\text{SCR} / (n-ab)}, \qquad F_B = \frac{\text{SCB} / (a-1)}{\text{SCR} / (n-ab)}, \qquad F_{AB} = \frac{\text{SCAB} / [(a-1)(b-1)]}{\text{SCR} / (n-ab)},$$

où

$$\text{SCA} = \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{023})\boldsymbol{y}, \qquad \text{SCB} = \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{013})\boldsymbol{y}, \qquad \text{SCAB} = \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{012})\boldsymbol{y} \text{ et} \qquad \text{SCR} = \boldsymbol{y}'(\mathbf{I}-\mathbf{P})\boldsymbol{y}.$$

Or ces sommes de carrés ne sont pas toujours celles qui figurent dans une table d'analyse de variance. Le but d'une analyse de variance est de décomposer une somme de carrés totale $\text{SCT} = \boldsymbol{y}'(\mathbf{I}-\mathbf{P}_0)\boldsymbol{y}$ d'abord en

sa partie expliquée et sa partie résiduelle,

$$\text{SCT} = \text{SCE} + \text{SCR} = \boldsymbol{y}'(\mathbf{P} - \mathbf{P}_0)\boldsymbol{y} + \boldsymbol{y}'(\mathbf{I} - \mathbf{P})\boldsymbol{y};$$

et ensuite sa partie expliquée en une somme

$$\text{SCE} = \text{SCA} + \text{SCB} + \text{SCAB}.$$

Or les sommes des carrés définies ci-dessus ne satisfont pas toujours cette condition : en général,

$$\text{SCA} + \text{SCB} + \text{SCAB} = \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{023})\boldsymbol{y} + \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{013})\boldsymbol{y} + \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{012})\boldsymbol{y} \neq \boldsymbol{y}'(\mathbf{P} - \mathbf{P}_0)\boldsymbol{y}$$

Pour s'assurer que la somme des sommes de carrés donne bien SCE, certains logiciels les définissent autrement. Procédant successivement, SCA, SCB et SCAB sont définies comme ceci :

$$\text{SCA} = \boldsymbol{y}'(\mathbf{P}_{01}-\mathbf{P}_0)\boldsymbol{y};\ \text{SCB} = \boldsymbol{y}'(\mathbf{P}_{012}-\mathbf{P}_{01})\boldsymbol{y}\ ;\ \text{SCAB} = \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{012})\boldsymbol{y}$$

Mais que testent ces sommes, figurant au numérateur d'une statistique *F*? Généralement, elles ne testent pas ce qu'on prétend tester[1], à moins qu'elles coïncident avec celles définies plus haut, c'est-à-dire, à moins que

$$\boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{023})\boldsymbol{y} = \boldsymbol{y}'(\mathbf{P}_{01}-\mathbf{P}_0)\boldsymbol{y}\ ; \qquad \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{013})\boldsymbol{y} = \boldsymbol{y}'(\mathbf{P}_{012}-\mathbf{P}_{01})\boldsymbol{y}; \qquad \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{012})\boldsymbol{y} = \boldsymbol{y}'(\mathbf{P}-\mathbf{P}_{012})\boldsymbol{y}.$$

Or ces égalités sont vérifiées si et seulement si les colonnes $\mathbf{X}_1$, $\mathbf{X}_2$, et $\mathbf{X}_3$ sont mutuellement *orthogonales*, c'est-à-dire, si $\mathbf{X}_i'\mathbf{X}_j = 0$, $i \neq j$. Ceci découle du fait que dans ce cas, $\mathbf{P}_{0123} = \mathbf{P}_0 + \mathbf{P}_1 + \mathbf{P}_2 + \mathbf{P}_3$. Et quand est-ce que, dans un modèle à deux facteurs, ces matrices sont mutuellement orthogonales? *Quand les données sont équilibrées*, c'est-à-dire, quand chaque case contient le même nombre d'observations. On le vérifie en calculant le produit $\mathbf{X}'\mathbf{X}$, dans l'exemple :

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 12 & 0 & 0 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 4 & 0 & 0 \\ 0 & 0 & 4 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 4 \\ 0 & 0 & 0 & 0 & 4 & 8 \end{bmatrix}.$$

Il suffit d'ajouter quelques observations à certaines des cases pour rompre cette orthogonalité (répétez certaines des lignes de $\mathbf{X}$ un nombre inégal de fois pour voir).

Pourquoi l'orthogonalité est-elle souhaitable? Pour voir ce qui se passe lorsque l'orthogonalité n'est pas vérifiée, considérons le modèle $\text{E}(\boldsymbol{y}) = \mathbf{X}\boldsymbol{\beta} = [\mathbf{X}_1 : \mathbf{X}_2]\begin{bmatrix}\boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2\end{bmatrix}$. La somme des carrés pour tester l'hypothèse $\text{H}_1 : \boldsymbol{\beta}_2 = \mathbf{0}$ est $\boldsymbol{y}'(\mathbf{P}_{12} - \mathbf{P}_1)\boldsymbol{y}$. Si $\mathbf{X}_1$ et $\mathbf{X}_2$ sont mutuellement orthogonaux, $\mathbf{P}_{12} = \mathbf{P}_1+\mathbf{P}_2$ et $\mathbf{P}_{12} - \mathbf{P}_1 = \mathbf{P}_2$ et la statistique devient simplement $\boldsymbol{y}'\mathbf{P}_2\boldsymbol{y}$. Sinon, $\mathbf{P}_{12} - \mathbf{P}_1 = (\mathbf{I}-\mathbf{P}_1)\mathbf{P}_2[\mathbf{P}_2'(\mathbf{I}-\mathbf{P}_1)\mathbf{P}_2]^{-1}\mathbf{P}_2'(\mathbf{I}-\mathbf{P}_1)$. La différence entre $\boldsymbol{y}'\mathbf{P}_2\boldsymbol{y}$ et $\boldsymbol{y}'(\mathbf{I}-\mathbf{P}_1)\mathbf{P}_2[\mathbf{P}_2'(\mathbf{I}-\mathbf{P}_1)\mathbf{P}_2]^{-1}\mathbf{P}_2'(\mathbf{I}-\mathbf{P}_1)\boldsymbol{y}$ peut être importante, et elle l'est d'autant plus que $\mathbf{P}_1\mathbf{P}_2\boldsymbol{y}$ est différent de $\mathbf{0}$.

6.15 Analyse de variance à deux facteurs – facteurs emboîtés

Nous avons considéré à la section 2.4 une décomposition particulière de la somme de carrés totale : les sommes SCB et SCAB sont fusionnées en une somme notée SCA|B. Dans l'exemple présenté, il s'agissait d'un *choix*, une alternative que l'expérimentateur peut adopter ou écarter. Mais il existe des cas

[1] Les formules de SCA et SCB montrent qu'en fait on teste H_A dans un modèle où seul le facteur A est présent; et on teste H_B dans un modèle qui ne comprend que les facteurs A et B, sans interactions.

où la décomposition en 2.4 s'impose, car la décomposition classique n'a pas de sens. C'est le cas des données suivantes qui représentent la teneur en matières grasses (en cg) par 100 g d'orange de 6 différentes variétés employées par un fabricant de produits alimentaires. Chacune des variétés provient de trois pays différents :

		Pays			
		P1	P2	P3	***Moyenne***
Variété	V1	3,5 4,0 3,0 4,5	2,5 4,5 5,5 5,0	3,0 3,0 2,5 3,0	3,66667
	V2	5,0 5,5 4,0 3,5	3,5 3,5 3,0 4,0	4,5 4,0 4,0 5,0	4,12500
	V3	5,0 4,5 5,0 4,5	5,5 6,0 5,0 5,0	5,5 4,5 6,5 5,5	5,20833
	V4	8,5 6,0 9,0 8,5	6,5 7,0 8,0 6,5	7,0 7,0 7,0 7,0	7,33333
	V5	6,0 5,5 3,5 7,0	6,0 8,5 4,5 7,5	6,5 6,5 8,5 7,5	6,45833
	V6	7,0 9,0 8,5 8,5	6,0 7,0 7,0 7,0	11,0 7,0 9,0 8,0	7,91667
	Moyenne	5,79167	5,60417	5,95833	5,78472

L'expérience vise à déterminer s'il y a des différences entre les variétés et entre les pays quant à la teneur en matières grasses. À première vue, il s'agit d'une analyse de variance à deux facteurs. Mais la différence ici vient du fait que les pays P1, P2, et P3 ne sont pas les mêmes pour chaque variété : ils pourraient représenter, par exemple, le Brésil, les États-unis et le Mexique pour la variété V1; l'Inde, la Chine et l'Iran pour la variété V2; etc. Nous avons bien deux facteurs, la variété constituant, disons, le facteur A; et la provenance le facteur B. Mais ils ne sont pas *croisés*, ils sont *emboîtés* : le facteur B (la provenance) est *emboîté* dans A. Une analyse de variance comme celle de la section précédente donnerait les résultats suivants:

```
> anova(lm(y~variete*provenance))
Source        DF        SS        MS         F         P
variete        5   179.642    35.928     38.47     0.000
provenance     2     1.507     0.753      0.81     0.452
Interaction   10    24.326     2.433      2.60     0.012
Error         54    50.437     0.934
Total         71   255.913
```

Cette analyse est correcte en ce qui concerne les variétés : le test pour H_A est le même dans les deux cas. En ce qui concerne la provenance (facteur B) ou les interactions, cette analyse serait fautive, car certaines des moyennes calculées dans le modèle à effets croisés n'ont plus de sens. Comme par exemple $\mu_{.j}$ qui intervient dans le calcul de SCB: ce serait la moyenne des oranges provenant du pays portant le label P_j, ce qui n'est pas sensé. L'hypothèse H_{AB} non plus. Une hypothèse concernant la provenance qui est raisonnable est la suivante :

Il n'y a pas de différence entre les provenances d'une même variété, et ce, pour toutes les variétés.

Formellement, cette hypothèse, que nous désignerons par $H_{A|B}$, s'exprime par

$$H_{A|B}: \ \mu_{i1} = \mu_{i2} = \ldots = \mu_{ib} \quad \text{pour } i = 1, 2, \ldots, a$$

La moyenne des carrés qui figurera au numérateur de la statistique F est

$$\text{MCA|B} = \frac{r\sum_{i=1}^{a}\sum_{j=1}^{b}(\overline{y}_{ij} - \overline{y}_{i.})^2}{a(b-1)} \ ,$$

le nombre de degrés de liberté étant $a(b\text{-}1)$. La somme entre les accolades, $r\sum_{j=1}^{b}(\bar{y}_{ij}.-\bar{y}_{i}..)^2$, mesure les écarts entre les pots à l'intérieur d'un même traitement. Donc la statistique F prendra une valeur élevée s'il y a d'importantes différences entre les pots à l'intérieur d'un même traitement. La statistique F correspondante est

$$F_{A|B} = \frac{MCA|B}{MCR}$$

On peut montrer que SCA|B = SCB + SCAB :

$$r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij}.-\bar{y}_{i}..)^2 = br\sum_{i=1}^{a}(\bar{y}_{i}..-\bar{y})^2 + r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij}.-\bar{y}_{i}..-\bar{y}_{.j}.-\bar{y})^2 ,$$

ce qui explique aussi le nombre de degrés de liberté : $(b\text{-}1)+(a\text{-}1)(b\text{-}1) = a(b\text{-}1)$.

Table d'analyse de variance :

```
> anova(lm(y~variete/provenance))
                   Df  Sum Sq Mean Sq F value  Pr(>F)
variete             5 179.642  35.928 38.4662 < 2e-16 ***
variete:provenance 12  25.833   2.153  2.3048 0.01858 *
Residuals          54  50.438   0.934
```

Paramétrisation
Une façon de paramétrer ce modèle:

$$\mu_{ij} = E(y_{ijk}). \quad = \mu + \alpha_i + \delta_{j(i)}$$

avec la correspondance suivante:

- $\mu = \mu$ (même sens dans les deux paramétrisations)
- $\alpha_i = \mu_{i\bullet} - \mu$
- $\delta_{ij} = (\mu_{\bullet j} - \mu) + (\mu_{ij} - \mu_{i\bullet} - \mu_{\bullet j} + \mu) = \mu_{ij} - \mu_{i\bullet}$

Analyse de variance à deux facteurs avec une observation par cellule

6.15 Analyse de variance à deux facteurs avec une observation par cellule

Dans une analyse de variance à deux facteurs, il arrive qu'on n'ait qu'une seule observation par cellule. Puisque l'estimateur MCR de la variance σ^2 est une mesure de la dispersion des données d'une même cellule, cet estimateur n'existe pas dans ce cas. Il est nécessaire alors d'imposer certaines contraintes aux paramètres afin d'obtenir une estimation de σ^2. Ce problème une généralisation du test d'égalité de deux moyennes avec données appariées.

Le modèle

Considérons les données suivantes.

Exemple [George W. Snedecor et William G. Cochran, Statistical methods, Sixth edition, Iowa State, p. 301] On prend une certaine mesure de la teneur en eau des feuilles des arbres de trois espèces d'agrumes sous trois conditions d'ensoleillement (9 arbres en tout). Voici les données :

Ensoleillement	Orange Shamouti	Pamplemousse	Clémentine	$\bar{y}_{i.}$
Soleil	112	90	123	108,3333
Ombre partielle	86	73	89	82,66667
Ombre	80	62	81	74,33333
$\bar{y}_{.j}$	92,66667	75	97,66667	88,44444

On a donc en tout $a \times b$ observations y_{ij}, et le modèle est

$$y_{ij} = \mu_{ij} + \varepsilon_{ij}\ ,\ \ i = 1, \ldots, a;\ \ j = 1, \ldots, b \quad \text{(ici } a = b = 3\text{)}$$

où les ε_{ij} sont indépendantes, $\varepsilon_{ij} \sim \mathcal{N}(0\ ;\ \sigma^2)$.

La somme des carrés expliquée SCE = $\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2$ se décompose ainsi:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{..})^2 = b\sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2 + a\sum_{j=1}^{b}(\bar{y}_{.j} - \bar{y}_{..})^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}(y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2$$

SCT = SCA + SCB + SCAB.

Dans l'exemple, on a

SCT = $(112\text{-}88{,}44)^2 + (90\text{-}88{,}44)^2 + \ldots + (81\text{-}88{,}44)^2 = 2822{,}222$
SCA = $3[(108{,}33\text{-}68{,}44)^2 + (82{,}67\text{-}68{,}44)^2 + (74{,}33\text{-}68{,}44)^2] = 1884{,}222$
SCB = $3[(92{,}67\text{-}68{,}44)^2 + (75\text{-}68{,}44)^2 + (97{,}67\text{-}68{,}44)^2] = 850{,}8889$
SCAB = SCT – SCA – SCB = 87,11111 ■

Les sommes de carrés SCA, SCB, et SCAB sont indépendantes, et divisées par σ^2 elles suivent chacune une loi χ^2, généralement non centrale. Le problème qui se pose ici, c'est qu'il n'y a pas de somme de carrés résiduelle (dispersion à l'intérieur des cases), et donc pas d'estimateur de variance.

Les espérances des moyennes de carrés

$$\text{MCA} = \text{SCA}/(a\text{-}1),\quad \text{MCB} = \text{SCB}/(b\text{-}1) \text{ et } \text{MCAB} = \text{SCAB}/[(a\text{-}1)(b\text{-}1)]$$

sont présentées dans le tableau suivant:

Moyenne de carrés	*Degrés de liberté*	*Espérance*
MCA	a-1	$\sigma^2 + \dfrac{b\sum_{i=1}^{a}(\mu_{i.} - \mu)^2}{a-1}$
MCB	b-1	$\sigma^2 + \dfrac{a\sum_{j=1}^{b}(\mu_{.j} - \mu)^2}{b-1}$
MCAB	$(a\text{-}1)(b\text{-}1)$	$\sigma^2 + \dfrac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu)^2}{(a-1)(b-1)}$
MCT	ab-1	$\sigma^2 + \dfrac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij} - \mu)^2}{(a-1)(b-1)}$

On voit bien qu'aucun des quotients possibles ne peut servir à tester les hypothèses usuelles H_A, H_B et H_{AB}.

Il sera donc nécessaire d'imposer quelques contraintes supplémentaires. La contrainte normalement imposée est

$$\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu_{\cdot\cdot} = 0 \quad \text{pour } \textit{tout } i, j$$

où

$$\mu_{i\cdot} = \frac{1}{b}\sum_{j=1}^{b}\mu_{ij} \; ; \; \mu_{\cdot j} = \frac{1}{a}\sum_{i=1}^{a}\mu_{ij} \; ; \; \mu_{\cdot\cdot} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}$$

Ce postulat, appelé *hypothèse d'additivité* ou de *non interaction*, que nous ne pouvons pas tester et considérons comme partie du modèle, réduit l'espérance de MCAB à σ^2 et nous permet de tester les hypothèses suivantes:

H_A : Les $\mu_{i\cdot}$ sont égaux : $F_A = \dfrac{\text{MCA}}{\text{MCAB}} \sim \mathcal{F}_{a\text{-}1,(a\text{-}1)(b\text{-}1)}$ lorsque H_A est vraie.

H_B : Les $\mu_{\cdot j}$ sont égaux : $F_B = \dfrac{\text{MCB}}{\text{MCAB}} \sim \mathcal{F}_{b\text{-}1,(a\text{-}1)(b\text{-}1)}$ lorsque H_B est vraie.

Dans l'exemple,

$$F_A = \frac{\text{MCA}}{\text{MCAB}} = \frac{1884,22/3}{87,11/4} = 43,26 \text{ et } F_B = \frac{\text{MCB}}{\text{MCAB}} = \frac{850,89/3}{87,11/4} = 19,536.$$

Les p-valeurs correspondantes sont (à 3 et 4 degrés de liberté dans les deux cas) 0,001953 pour F_A et 0,008625 pour F_B. Les deux facteurs sont donc fortement significatifs.

Voici les résultats produits par le logiciel R :

```
> anova(lm(y~soleil+espece))
Analysis of Variance Table
          Df  Sum Sq Mean Sq F value   Pr(>F)
soleil     2 1884.22  942.11  43.260 0.001953 **
espece     2  850.89  425.44  19.536 0.008625 **
Residuals  4   87.11   21.78
```

Suite omise dans la version courte

Modèle à effets aléatoires

Supposons qu'on veuille déterminer si la consommation d'essence d'une voiture neuve varie d'une voiture à l'autre. On prélève les données suivantes, qui représentent la distance (en milles) parcourue sur un gallon d'essence par 4 voitures. Le nombre d'essais varie d'une voiture à l'autre.

Voiture 1	Voiture 2	Voiture 3	Voiture 4
19	21	21	25
20	22	23	24
21	24	24	26
22	25	27	
26	26		
	27		

Au premier abord, l'allure des données ainsi que la question posée suggèrent une analyse de variance à un facteur. Mais supposons que les 4 voitures sont de fabrication identique et que le but de l'expérience n'est pas — comme il le serait dans une analyse de variance classique — de savoir s'il y a une différence entre les 4 voitures *utilisées pour l'expérience*, mais plutôt de savoir s'il y a une différence entre *les voitures en général*. Les 4 voitures ne constituent donc pas 4 populations dont on veut comparer les moyennes: elles

représentent en fait un échantillon de taille 4 d'une population de voitures de même marque, même modèle. Si y_{ij} représente la j^e observation sur la i^e voiture, nous pouvons considérer un modèle dont l'équation est identique à celle du modèle d'analyse de variance à un facteur, soit

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \quad i = 1, \ldots, k; j = 1, \ldots, n_i$$

où α_i est l'écart entre la moyenne générale μ de toutes les voitures de la population et la moyenne de la i^e voiture tirée dans la population. Puisque les voitures sont tirées au hasard, les α_i ne sont pas des paramètres, mais des variables aléatoires, de même que les ε_{ij}. Nous supposerons que les α_i et les ε_{ij} sont mutuellement indépendantes et que

$$\alpha_i \sim \mathcal{N}(0 ; \sigma_\alpha^2), \qquad \varepsilon_{ij} \sim \mathcal{N}(0; \sigma^2)$$

Le paramètre μ est une constante. Nous appellerons ce modèle le ***modèle à effets aléatoires***, pour le distinguer du modèle discuté précédemment, lequel par contraste s'appellera ***modèle à effets fixes***. Le paramètre σ_α est une mesure de la dispersion *entre* les moyennes des voitures, alors que σ est une mesure de la dispersion entre les différents essais effectués avec la *même* voiture.

Remarque. *Les variables aléatoires α_i sont des constantes une fois les voitures choisies. En d'autres termes, la distribution* conditionnelle *des observations y_{ij}, étant donné $\alpha_1, \ldots, \alpha_k$, est identique à celle qui est stipulée par le modèle à effets fixes.* ■

L'hypothèse que nous voulons tester s'exprime par:

$$H_o : \sigma_\alpha^2 = 0$$

H_o réduit le modèle à

$$y_{ij} = \mu + \varepsilon_{ij}, \quad i = 1, \ldots, k; j = 1, \ldots, n$$

Ce modèle est identique à celui auquel se réduit le modèle à effets fixes lorsqu'on impose à celui-ci l'hypothèse $\mu_1 = \ldots = \mu_k$. Donc sous H_o la statistique F a la même distribution ici que dans le modèle a effets fixes. Nous devrons néanmoins développer les propriétés des sommes de carrés SCR et SCE dans le contexte d'un modèle à effets aléatoires, d'abord pour justifier la région critique $F > F_{k-1,n-k;\alpha}$, ensuite pour permettre le calcul de la fonction de puissance.

Somme de carrés résiduelle. La somme des carrés résiduelle est

$$\text{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\varepsilon_{ij} - \bar{\varepsilon}_{i.})^2$$

Puisque les ε_{ij} sont indépendantes, $\varepsilon_{ij} \sim \mathcal{N}(0 ; \sigma^2)$, on déduit que

$$\frac{\text{SCR}}{\sigma^2} \sim \chi^2_{n-k},$$

exactement comme dans le modèle à effets fixes et on déduit également que MCR = SCR/(n-k) est un estimateur sans biais de σ^2:

$$E(\text{MCR}) = \sigma^2$$

Somme de carrés expliquée. La somme des carrés expliquée est fonction des moyennes échantillonnales

$$\bar{y}_{i.} = \mu + \alpha_i + \bar{\varepsilon}_{i.}$$

Les $\bar{y}_i$ sont indépendantes,

$$\text{Var}(\bar{y}_i) = \text{Var}(\mu + \alpha_i + \bar{\varepsilon}_{i.}) = \text{Var}(\alpha_i) + \text{Var}(\bar{\varepsilon}_{i.}) = \sigma_\alpha^2 + \sigma^2/n_i$$

Si $\bar{\boldsymbol{y}} = [\,\bar{y}_{1.}, \ldots, \bar{y}_{k.}\,]'$, alors

$$\boldsymbol{\mu} = \mathrm{E}(\bar{\boldsymbol{y}}) = \mu\boldsymbol{e} \quad \text{et} \quad \boldsymbol{\Sigma} = \mathrm{Var}(\bar{\boldsymbol{y}}) = \sigma_\alpha^2 \mathbf{I} + \sigma^2 \mathbf{D}^{-1},$$

où $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]'$ et

$$\mathbf{D} = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_k \end{bmatrix}$$

La somme des carrés expliquée

$$\mathrm{SCE} = \sum_{i=1}^{k} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^{k} n_i \bar{y}_{i.}^{\,2} - n\bar{y}_{..}^{\,2}$$

peut s'écrire sous forme matricielle comme

$$\mathrm{SCE} = \bar{\boldsymbol{y}}' \left(\mathbf{D} - \frac{1}{n} \mathbf{D}\mathbf{e}\mathbf{e}'\mathbf{D} \right) \bar{\boldsymbol{y}}$$

Cette somme de carrés ne suit pas en général une loi χ^2. Considérons séparément les cas où les n_i sont égaux et inégaux.

Cas où les effectifs sont égaux

Si $n_1 = \ldots = n_k = r$, alors

$$\boldsymbol{\Sigma} = \mathrm{Var}(\bar{\boldsymbol{y}}) = (\sigma_\alpha^2 + \sigma^2/r)\mathbf{I}$$

et

$$\frac{\mathrm{SCE}}{\sigma^2 + r\sigma_\alpha^2} \sim \chi_{k-1}^2 \text{ centrale}$$

La statistique $F = \mathrm{MCE}/\mathrm{MCR}$ peut s'écrire comme

$$F = \frac{\mathrm{MCE}}{\mathrm{MCR}} = \frac{\sigma^2 + r\sigma_\alpha^2}{\sigma^2} \, \frac{\mathrm{MCE}/(\sigma^2 + r\sigma_\alpha^2)}{\mathrm{MCR}/\sigma^2}$$

Le 2e facteur à droite suit toujours une loi $\mathcal{F}_{k-1,n-k}$; donc la statistique F ci-dessus suit une loi de Fisher si et seulement si le premier facteur vaut 1, c'est-à-dire, si et seulement si $\sigma_\alpha^2 = 0$. Sinon, F a tendance a prendre des valeurs plus grandes que celle d'une variable de loi $\mathcal{F}_{k-1,n-k}$. Ceci justifie l'emploi de la statistique F et de la région critique

$$F = \frac{\mathrm{MCE}}{\mathrm{MCR}} > \mathcal{F}_{k-1,n-k;\alpha}$$

Cas où les effectifs ne sont pas égaux

Lorsque les n_i ne sont pas égaux, le numérateur n'est pas une khi-deux et la statistique n'est donc pas une Fisher. Cependant, le test demeure valide puisque, lorsque $\mathrm{H_o}$ est vraie, SCE/σ^2 suit une loi khi-deux à k-1 degrés de liberté. Le choix de la région critique se justifie par le fait que

$$\mathrm{E(MCE)} = \sigma^2 + \left(n^2 - \sum_{i=1}^{k} n_i^2 \right) \frac{\sigma_\alpha^2}{n(k-1)}$$

laquelle espérance est supérieure à celle du dénominateur [E(MCR) = σ^2], et lui est égale si et seulement si $\sigma_\alpha^2 = 0$.

Exemples. Supposons qu'on veuille savoir si le succès au jeu de pinball est une question de hasard pur, ou si l'adresse du joueur y est pour quelque chose. On choisit k individus et on fait jouer le i^e individu n_i fois. Soit y_{ij} le score du i^e individu au j^e essai. Dans l'équation du modèle, le terme α_i représente « l'aptitude » du i^e joueur, c'est-à-dire, sa moyenne à long terme. Puisque ce joueur a été choisi au hasard, α_i est une variable aléatoire. Cette variable a une certaine variance, σ_α^2, qui mesure les écarts d'aptitude entre les personnes. L'hypothèse que $\sigma_\alpha^2 = 0$ est l'hypothèse qu'il n'y a pas de différence entre les individus, ce qui dans notre contexte signifie que l'aptitude n'entre en rien dans le succès au jeu. Pour un α_i fixe, la variance conditionnelle de y_{ij} est σ^2, qui est donc la dispersion des scores pour un *même* joueur.

Par contre, la variance *inconditionnelle* de y_{ij} est

$$\text{Var}(y_{ij}) = \sigma^2 + \sigma_\alpha^2$$

La composante σ^2 est attribuable au hasard; l'autre, σ_α^2, est attribuable aux différences d'aptitude entre les individus. Remarquez aussi que

$$\sigma_\alpha^2 = \text{Cov}(y_{ij}\,;\, y_{ij'}) \text{ pour tout } j \text{ et } j'$$

En d'autres termes, σ_α^2 est la covariance entre deux scores obtenus par un même joueur (et attribuable uniquement au fait qu'il s'agit du même joueur). Cette covariance est nulle si l'aptitude du joueur ne contribue pas à son succès. Finalement, le quotient

$$\rho = \frac{\sigma_\alpha^2}{\sigma^2 + \sigma_\alpha^2}$$

est une mesure de l'importance de *l'aptitude* au jeu, par opposition au hasard pur. Ce quotient est aussi le coefficient de corrélation entre deux scores obtenus par un même joueur. Il est appelé *coefficient de corrélation interne*. Le coefficient de corrélation interne, aussi appelé coefficient de corrélation *familiale,* tient ses origines de l'application suivante. Une certaine caractéristique y est mesurée sur tous les membres de plusieurs familles (ou les membres de plusieurs portées d'animaux), y_{ij} étant l'observation sur le j^e membre de la i^e famille. σ^2 mesure la dispersion entre les membres d'une même famille (puisque $\sigma^2 = \text{Var}(y_{ij}|\alpha_i)$, alors que σ_α^2 mesure les écarts entre les familles. Le coefficient de corrélation interne ρ pourrait être considéré comme une mesure de la force de l'hérédité ou d'autres facteurs environnementaux propres à la famille.

Estimateurs de paramètres.

À partir des expression des espérances E(MCR) et E(MCE) on obtient des estimateurs sans biais des paramètres σ^2 et σ_α^2:

$$\hat{\sigma}^2 = \text{MCR et } \hat{\sigma}_\alpha^2 = \frac{\text{MCE-MCR}}{n_o} \text{ où } n_o = \frac{n^2 - \sum_{i=1}^{k} n_i^2}{n(k-1)}$$

Remarquez que l'estimateur de σ_α^2 est une différence de moyennes de carrés. Il a donc un inconvénient majeur: il peut prendre des valeurs négatives.

Intervalles de confiance

Il est également possible de déterminer un intervalle de confiance pour ρ lorsque $n_1 = ... = n_k = r$. La statistique F = MCE/MCR suit une loi $\mathcal{F}_{k\text{-}1;n\text{-}k}$ si H_o: $\sigma_\alpha^2 = 0$ est vraie; autrement, c'est

$$\frac{\text{MCE}/(\sigma^2 + r\sigma_\alpha^2)}{\text{MCR}/\sigma^2} = \frac{\sigma^2}{\sigma^2 + r\sigma_\alpha^2} F$$

qui suit une loi $\mathcal{F}_{k\text{-}1;n\text{-}k}$. Alors

$$P\left[F_{k-1;n-k;1-\alpha/2} \le \frac{\sigma^2}{\sigma^2 + r\sigma_\alpha^2} F \le F_{k-1;n-k;\alpha/2} \right] = 1 - \alpha$$

Par la définition de ρ, $\frac{\sigma^2}{\sigma^2 + r\sigma_\alpha^2} = \frac{1-\rho}{1+(r-1)\rho}$. Nous substituons cette expression à l'équation ci-dessus pour avoir

$$P\left[F_{k-1;n-k;1-\alpha/2} \le \frac{1-\rho}{1+(r-1)\rho} F \le F_{k-1;n-k;\alpha/2} \right] = 1 - \alpha$$

Quelques manipulations permettent d'exprimer les inégalités entre parenthèses sous la forme

$$\frac{F - F_2}{F + (r-1)F_2} \le \rho \le \frac{F - F_1}{F + (r-1)F_1}$$

où $F_1 = \mathcal{F}_{k\text{-}1;n\text{-}k;1\text{-}\alpha/2}$ et $F_2 = \mathcal{F}_{k\text{-}1;n\text{-}k;\alpha/2}$.

Analyse de variance à trois facteurs

Section omise dans la version courte

RÉSUMÉ

Analyse de variance à un facteur

Le modèle : $y_{ij} = \mu_i + \varepsilon_{ij}$; $\varepsilon_{ij} \sim \mathcal{N}(0 ; \sigma^2)$, $i = 1,\ldots, k$; $j = 1, \ldots, n_i$

$H_o : \mu_1 = \mu_2 = \ldots = \mu_k$

Table d'analyse de variance

Source	*Somme de carrés*	*Degrés de liberté*	*Moyenne des carrés*	*Espérances des moyennes des carrés*
Expliquée	$\text{SCE} = \sum_{i=1}^{k} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	k-1	$\text{MCE} = \frac{\text{SCE}}{k-1}$	$\sigma^2 + \frac{\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2}{k-1}$
Résiduelle	$\text{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	n-k	$\text{MCR} = \frac{\text{SCR}}{n-k}$	σ^2
Total	$\text{SCT} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2$	n-1	$\text{MCT} = \frac{\text{SCT}}{n-1}$	$\sigma^2 + \frac{\sum_{i=1}^{k} n_i(\mu_i - \bar{\mu})^2}{n-1}$

Test de H_o : $F = \frac{\text{MCE}}{\text{MCR}} = \frac{\text{SCE}/(k-1)}{SCR/(n-k)} \sim \mathcal{F}_{k\text{-}1;n\text{-}k}$ lorsque H_o est vraie

Estimation des paramètres : $\hat{\mu}_i = \bar{y}_i$; $\mathrm{Var}(\hat{\mu}_i) = \frac{\sigma^2}{n_i}$; $\hat{\sigma}^2 = \mathrm{MCR} = \frac{\mathrm{SCR}}{n-k}$

Test d'ajustement à une droite

n observations comportent $k < n$ valeurs distinctes de $x_1, x_2, \ldots x_k$.
Les valeurs y qui correspondent à x_i sont $y_{i1}, y_{i2}, \ldots, y_{in_i}$.

Modèle initial : $\mathcal{M}$: $y_{ij} = \mu_i + \varepsilon_{ij}$; $i = 1, \ldots, k$; $j = 1, \ldots, n_i$

H_o : hypothèse de linéarité

$H_o \Rightarrow \mathcal{M}_o$: $y_{ij} = \beta_o + \beta_1 x_i + \varepsilon_{ij}$

$\mathcal{M}$ est le modèle d'analyse de variance et $\mathcal{M}_o$ est le modèle de régression.
Numérateur du rapport F pour tester H_o SCE = $(SCR_o - SCR)/(k-2)$, où
SCR_o et SCR sont les sommes de carrés résiduelles dans $\mathcal{M}_o$ et $\mathcal{M}$, respectivement, soit :

$$\mathrm{SCR} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 \; ; \; \mathrm{SCR}_o = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \hat{\beta}_o - \hat{\beta}_1 x_i)^2$$

$$\mathrm{SCR}_o - \mathrm{SCR} = \sum_{i=1}^{k} n_i(\bar{y}_i - \hat{\beta}_\circ - \hat{\beta}_1 x_i)^2 \; ; \; F = \frac{\left[\mathrm{SCR}_o - \mathrm{SCR}\right]/(k-2)}{\mathrm{SCR}/(n-k)} \sim \mathcal{F}_{k-2;n-k}$$

Test d'homogénéité de variances

$y_{ij} = \mu_i + \varepsilon_{ij}$, où $\varepsilon_{ij} \sim \mathcal{N}(0 ; \sigma_i^2)$.

$$H_o : \sigma_1^2 = \ldots = \sigma_k^2$$

$s_i^2 = \frac{\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{n_i - 1}$ estime σ_i^2 dans le modèle ; $s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k}$ = MCR estime σ^2 sous H_o.

$Q = (n-k)\ln(s_p^2) - \sum_{i=1}^{k}(n_i - 1)\ln(s_i^2) \sim \chi^2_{k-1}$ à peu près sous H_o.

Combinaisons linéaires des moyennes

Soit $\varphi = \Sigma_i c_i \mu_i$. $\hat{\varphi} = \sum c_i \hat{\mu}_i = \sum_i c_i \bar{y}_i$.

$H_o : \varphi = \varphi_o$. Sous H_o, $Z = \frac{\hat{\varphi} - \varphi_o}{\sigma\sqrt{\sum_{i=1}^{k} c_i^2 / n_i}} \sim \mathcal{N}(0 ; 1)$; $T = \frac{\hat{\varphi} - \varphi_o}{\hat{\sigma}\sqrt{\sum_{i=1}^{k} c_i^2 / n_i}} \sim t_{n-k}$

Analyse de variance à deux facteurs – facteurs croisés

$y_{ijk} = \mu_{ij} + \varepsilon_{ij}$, $\varepsilon_{ijk} \sim \mathcal{N}(0 ; \sigma^2)$, ε_{ij} indépendantes.

H_A: Le facteur A n'a pas d'effet: $\mu_{1.} = \ldots = \mu_{a.}$, $\mu_{i.} = \Sigma_j \mu_{ij}/b$

H_B: Le facteur B n'a pas d'effet: $\mu_{.1} = \ldots = \mu_{.b}$, $\mu_{.j} = \Sigma_i \mu_{ij}/a$

H_{AB}: Aucune interaction entre A et B: $\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu = 0 \; \forall_{ij}$, où $\mu = \Sigma_i \Sigma_j \mu_{ij}/ab$.

Source	*Somme de carrés*	*Degrés de liberté*	*Moyenne de carrés*	*Espérances des moyennes de carrés*
Facteur A	SCA = $br\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot})^2$	a-1	MCA = SCA/(a-1)	$\sigma^2+\frac{br\sum_{i=1}^{a}(\mu_{i\cdot}-\mu)^2}{a-1}$
Facteur B	SCB = $ar\sum_{j=1}^{b}(\bar{y}_{\cdot j\cdot}-\bar{y})^2$	b-1	MCB = SCB/(b-1)	$\sigma^2+\frac{ar\sum_{j=1}^{b}(\mu_{\cdot j\cdot}-\mu)^2}{b-1}$
Interactions	SCAB = $r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot j\cdot}-\bar{y}_{\cdot\cdot\cdot})^2$	(a-1)(b-1)	MCAB = SCAB/(a-1)(b-1)	$\sigma^2+\frac{r\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij}-\mu_{i\cdot}-\mu_{\cdot j}+\mu)^2}{(a-1)(b-1)}$
Résiduel	SCR = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk}-\bar{y}_{ij\cdot}\right)^2$	ab(r-1)	MCR = SCR/ab(r-1)	σ^2
Total	SCT = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk}-\bar{y}_{\cdot\cdot\cdot}\right)^2$	abr-1	MCT = SCT/(n-1)	

1. SCA, SCB, SCAB et SCR sont indépendantes.
2. $SCA/\sigma^2 \sim \chi^2_{a-1}(\lambda_A)$, où $\lambda_A = br\Sigma_i(\mu_{i\cdot} - \mu)^2/\sigma^2$.
3. $SCB/\sigma^2 \sim \chi^2_{b-1}(\lambda_B)$, où $\lambda_B = ar\Sigma_j(\mu_{\cdot j} - \mu)^2/\sigma^2$.
4. $SCAB/\sigma^2 \sim \chi^2_{(a-1)(b-1)}(\lambda_{AB})$, où $\lambda_{AB} = r\Sigma_i\Sigma_j(\mu_{ij} - \mu_{i\cdot} - \mu_{\cdot j} + \mu)^2/\sigma^2$.
5. $SCR/\sigma^2 \sim \chi^2_{ab(r-1)}$ centrale.

Analyse de variance à deux facteurs – facteurs emboîtés

$H_{A|B}$: $\mu_{i1} = \mu_{i2} = \ldots = \mu_{ib}$ pour $i = 1, 2, \ldots, a$; $MCA|B = \frac{r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij} - \bar{y}_{i\cdot})^2}{a(b-1)}$, $F_{A|B} = \frac{MCA|B}{MCR}$

Source	*Somme de carrés*	*Degrés de liberté*	*Moyenne de carrés*	*Espérances des moyennes de carrés*
Facteur A	SCA = $br\sum_{i=1}^{a}(\bar{y}_{i\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot})^2$	a-1	MCA = $\frac{SCA}{a\text{-}1}$	$\sigma^2+\frac{br\sum_{i=1}^{a}(\mu_{i\cdot\cdot}-\mu_{\cdot\cdot\cdot})^2}{a-1}$
Facteur A|B	SCA|B = SCB + SCAB = $r\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot})^2$	a(b-1)	MCA|B = $\frac{SCB+SCAB}{a(b\text{-}1)}$	$\sigma^2+\frac{r\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij\cdot}-\mu_{i\cdot\cdot})^2}{a(b-1)}$
Résiduel	SCR = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk}-\bar{y}_{ij\cdot}\right)^2$	ab(r-1)	MCR = $\frac{SCR}{ab(r-1)}$	σ^2
Total	SCT = $\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{r}\left(y_{ijk}-\bar{y}_{\cdot\cdot\cdot}\right)^2$	abr-1	MCT = $\frac{SCT}{n-1}$	

Deux facteurs avec une observation par cellule

On suppose les interactions nulles

Moyenne de carrés	*Degrés de liberté*	*Espérance*
MCA	a-1	$\sigma^2 + \frac{b\sum_{i=1}^{a}(\mu_i. - \mu)^2}{a-1}$
MCB	b-1	$\sigma^2 + \frac{a\sum_{j=1}^{b}(\mu._j - \mu)^2}{b-1}$
MCAB	(a-1)(b-1)	$\sigma^2 + \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu)^2}{(a-1)(b-1)}$
MCT	ab-1	$\sigma^2 + \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}(\mu_{ij} - \mu)^2}{(a-1)(b-1)}$

H_A : Les $\mu_{i\bullet}$ sont égaux : $F_A = \dfrac{\text{MCA}}{\text{MCAB}} \sim \mathcal{F}_{a\text{-}1,(a\text{-}1)(b\text{-}1)}$ lorsque H_A est vraie.

H_B : Les $\mu_{\bullet j}$ sont égaux : $F_B = \dfrac{\text{MCB}}{\text{MCAB}} \sim \mathcal{F}_{b\text{-}1,(a\text{-}1)(b\text{-}1)}$ lorsque H_B est vraie.

Modèle à effets aléatoires

$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ $i = 1, \ldots, k$; $j = 1, \ldots, n_i$; $\alpha_i \sim \mathcal{N}(0\,;\ \sigma_\alpha^2)$, $\varepsilon_{ij} \sim \mathcal{N}(0; \sigma^2)$;

$H_o : \sigma_\alpha^2 = 0 \Rightarrow y_{ij} = \mu + \varepsilon_{ij}$, $i = 1, \ldots, k$; $j = 1, \ldots, n$; $F = \dfrac{\sum_{i=}^{k} n_i \left(\bar{y}_{i.} - \bar{y}_{..}\right)^2 \Big/ (k-1)}{\hat{\sigma}^2} \sim \mathcal{F}_{k\text{-}1\,;\,n\text{-}k}$ sous H_o.

Annexe

Données illustrant une analyse de variance à trois facteurs

y	Facteur A	Facteur B	Facteur C
131	1	1	1
130	1	1	1
131	1	1	2
125	1	1	2
136	1	1	3
142	1	1	3
150	1	2	1
148	1	2	1
140	1	2	2
143	1	2	2
160	1	2	3
150	1	2	3
157	2	1	1
145	2	1	1
154	2	1	2
142	2	1	2
147	2	1	3
153	2	1	3
151	2	2	1
155	2	2	1
147	2	2	2
147	2	2	2
162	2	2	3
152	2	2	3
134	3	1	1
125	3	1	1
138	3	1	2
138	3	1	2
135	3	1	3
136	3	1	3
138	3	2	1
140	3	2	1
139	3	2	2
138	3	2	2
134	3	2	3
127	3	2	3

Annexe
Options de contrastes

Anova à un facteur

Soit μ_1, μ_2 et μ_3 les moyennes des trois cases et μ la moyenne des trois.

Paramétrisation « `treatment` » : $\mu_i = \mu_1+\alpha_i$, avec $\alpha_1 = 0$. Ici, $\mu_1 = \mu_1$; $\mu_2=\mu_1+\alpha_2$; $\mu_3 = \mu_1+\alpha_3$.

La relation est $\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\ 1 & 1 & 0\\ 1 & 0 & 1\end{bmatrix}\begin{bmatrix}\mu\\ \alpha_2\\ \alpha_3\end{bmatrix}$ ou $\begin{bmatrix}\mu\\ \alpha_2\\ \alpha_3\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\ -1 & 1 & 0\\ -1 & 0 & 1\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}$

```
> options(contrasts=c("contr.treatment","contr.poly"))
> options()$contrasts
[1] "contr.treatment" "contr.poly"
> lm(y~B)
Coefficients:objects()
(Intercept)    B2            B3
-16.83   10.83    38.83
     μ1    α2=μ2−μ1     α3=μ3−μ1
```

Paramétrisation « `sum` »: $\mu_i = \mu+\alpha_i$ avec la contrainte $\Sigma\alpha_i = 0$. Ici, $\mu_1 = \mu+\alpha_1$; $\mu_2 = \mu+\alpha_2$; $\mu_3 = \mu-\alpha_1-\alpha_2$. La relation est $\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix} = \begin{bmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 1 & -1 & -1\end{bmatrix}\begin{bmatrix}\mu\\ \alpha_2\\ \alpha_3\end{bmatrix}$ ou $\begin{bmatrix}\mu\\ \alpha_2\\ \alpha_3\end{bmatrix} = \frac{1}{3}\begin{bmatrix}1 & 1 & 1\\ 2 & -1 & -1\\ -1 & 2 & -1\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}$

```
> options(contrasts=c("contr.sum","contr.poly"))
> options()$contrasts
[1] "contr.sum"  "contr.poly"
> lm(y~B)
Coefficients:
(Intercept)            B1            B2
    -0.2778      -16.5556       -5.7222
        μ      α1=μ1−μ              α2=μ2−μ
```

Paramétrisation « `helmert` » :

$\mu_1 = \mu-\alpha_2-\alpha_3$; $\mu_2 = \mu+\alpha_2-\alpha_3$; $\mu_3 = \mu+2\alpha_3$; ou $\mu = (\mu_1+\mu_2+\mu_3)/3$, $\alpha_2 = (\mu_2-\mu_1)/2$; $\alpha_2 = [\mu_3-(\mu_1+\mu_2)/2]/3$.

La relation est $\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix} = \begin{bmatrix}1 & -1 & -1\\ 1 & 1 & -1\\ 1 & 0 & 2\end{bmatrix}\begin{bmatrix}\mu\\ \alpha_2\\ \alpha_3\end{bmatrix}$ ou $\begin{bmatrix}\mu\\ \alpha_2\\ \alpha_3\end{bmatrix} = \frac{1}{6}\begin{bmatrix}2 & 2 & 2\\ -3 & 3 & 0\\ -1 & -1 & 2\end{bmatrix}\begin{bmatrix}\mu_1\\ \mu_2\\ \mu_3\end{bmatrix}$

```
> options(contrasts=c("contr.helmert","contr.poly"))
> options()$contrasts
[1] "contr.helmert" "contr.poly"
> lm(y~B)
Coefficients:
(Intercept)            B1            B2
    -0.2778        5.4167        11.1389
        μ      (μ2−μ1)/2       [μ3−(μ1+μ2)/2]/3
```

Paramétrisation « `poly` » : On suppose que $\mu_i = \beta_o + \beta_1 x_i + \beta_1 x_i^2$, $x_i = i$, $i = 1, 2, 3$. On aurait donc $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} =$ $\mathbf{G}\boldsymbol{\beta}$, où $\boldsymbol{\beta} = \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \end{bmatrix}$ et $\mathbf{G} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}$. Mais on modifie les colonnes de **G**, sauf la première, de façon à ce qu'elles soient orthogonales à la première et les unes aux autres. Pour ce faire, on remplace la 2e *colonne* $x_2 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ par sa projection sur le complément orthogonal de la première, $\boldsymbol{y}_2 = (\mathbf{I} - \boldsymbol{ee}'/3)\boldsymbol{x}_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$; ensuite on projette la 3e colonne $\boldsymbol{x}_3 = \begin{bmatrix} 1 \\ 4 \\ 9 \end{bmatrix}$ sur le complément orthogonal de l'espace engendré par les deux premières colonnes, ce qui donne $\boldsymbol{y}_3 = \begin{bmatrix} 1/3 \\ -2/3 \\ 1/3 \end{bmatrix}$. Finalement on divise $\boldsymbol{y}_2$ *et* $\boldsymbol{y}_3$ par leur norme, respectivement $2^{½}$ et $(2/3)^{½}$. Finalement, le vecteur $\boldsymbol{\beta}$ est défini par $\boldsymbol{\mu} = \begin{bmatrix} 1 & -1/\sqrt{2} & 1/\sqrt{6} \\ 1 & 0/\sqrt{2} & -2/\sqrt{6} \\ 1 & 1/\sqrt{2} & 1/\sqrt{6} \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \\ \beta_2 \end{bmatrix}$, ou $\boldsymbol{\beta} = \begin{bmatrix} 1 & 1/3 & 1/3 \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 1/\sqrt{6} & \sqrt{2/3} & 1/\sqrt{6} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}$

```
> options(contrasts=c("contr.poly","contr.poly"))
> options()$contrasts
[1] "contr.poly" "contr.poly"
> lm(y~B)
Coefficients:
(Intercept)          B.L          B.Q
    -0.2778      27.4593       7.0083
```

Anova à deux facteurs

Paramétrisation « `treatment` »

Voici les 6 moyennes en termes des paramètres μ, α_2, β_2, β_3, γ_{22}, γ_{23}

	***Hormone B* (*Progestérone*)**		
Hormone A (Estradiol)	0	0,1 *mg/jour*	10 *mg/jour*
0 *mg/jour*	μ	$\mu+\beta_2$	$\mu+\beta_3$
0,5 *mg/jour*	$\mu+\alpha_2$	$\mu+\alpha_2+\beta_2+\gamma_{22}$	$\mu+\alpha_2+\beta_3+\gamma_{23}$

Voici les valeurs estimées des paramètres μ, α_2, β_2, β_3, γ_{22}, γ_{23}

```
> options(contrasts=c("contr.treatment","contr.poly"))
> options()$contrasts
[1] "contr.treatment" "contr.poly"
> lm(y~A*B)
Coefficients:
(Intercept)        A2        B2        B3      A2:B2      A2:B3
    -14.667    -4.333     8.333    32.333     5.000     13.000
```

Paramétrisation « `sum` »

Voici les 6 moyennes en termes des paramètres μ, α_1, β_2, β_3, γ_{11}, γ_{12}

	***Hormone B* (*Progestérone*)**		
Hormone A (Estradiol)	0	0,1 *mg/jour*	10 *mg/jour*
0 *mg/jour*	$\mu+\alpha_1+\beta_1+\gamma_{11}$	$\mu+\alpha_1+\beta_2+\gamma_{12}$	$\mu+\alpha_1-\beta_1-\beta_2-\gamma_{11}-\gamma_{12}$
0,5 *mg/jour*	$\mu-\alpha_1+\beta_1-\gamma_{11}$	$\mu-\alpha_1+\beta_2-\gamma_{12}$	$\mu-\alpha_1-\beta_1-\beta_2+\gamma_{11}+\gamma_{12}$

Voici les 6 valeurs estimées des paramètres μ, α_1, β_2, β_3, γ_{11}, γ_{12}

```
> options(contrasts=c("contr.sum","contr.poly"))
> options()$contrasts
[1] "contr.sum"  "contr.poly"
> lm(y~A*B)
Coefficients:
Intercept)          A1          B1          B2       A1:B1       A1:B2
 -0.2777778  -0.8333333 -16.5555556  -5.7222222   3.0000000   0.5000000
```

Paramétrisation « `helmert` »

Voici les 6 moyennes en termes des paramètres μ, α_1, β_2, β_3, γ_{11}, γ_{12}

	***Hormone B* (*Progestérone*)**		
Hormone A (Estradiol)	0	0,1 *mg/jour*	10 *mg/jour*
0 *mg/jour*	$\mu-\alpha_1-\beta_1-\beta_2+\gamma_{11}+\gamma_{12}$	$\mu-\alpha_1+\beta_1-\beta_2-\gamma_{11}+\gamma_{12}$	$\mu-\alpha_1+2\beta_2-2\gamma_{12}$
0,5 *mg/jour*	$\mu+\alpha_1-\beta_1-\beta_2-\gamma_{11}-\gamma_{12}$	$\mu+\alpha_1+\beta_1-\beta_2+\gamma_{11}-\gamma_{12}$	$\mu+\alpha_1+2\beta_2-+2\gamma_{12}$

Voici les estimations des paramètres μ, α_1, β_2, β_3, γ_{11}, γ_{12}

```
> options(contrasts=c("contr.helmert","contr.poly"))
> options()$contrasts
[1] "contr.helmert" "contr.poly"
> lm(y~A*B)
Coefficients:
(Intercept)    A1        B1      B2 A1:B1 A1:B2
-0.2778 0.83335.4167    11.1389 1.2500  1.7500
```

```
> options(contrasts=c("contr.poly","contr.poly"))
> options()$contrasts
[1] "contr.poly" "contr.poly"
> lm(y~A*B)
Coefficients:
(Intercept)    A.L       B.L     B.Q     A.L:B.L A.L:B.Q
-0.2778 1.178527.4593   7.0083  6.5000  0.8660
```