Analyse de la variance à plusieurs facteurs : plans factoriels

Auteur : Anne-Laure FOUGÈRES

L'analyse de variance à plusieurs facteurs ("factorial design" en anglais) répond aux mêmes objectifs que l'analyse de variance à un facteur, dans le cas où plusieurs variables indépendantes sont à étudier. Il s'agit ici encore de comparer les différentes parts de variabilité d'une variable dépendante associées aux différentes sources de variation.

Dans ce chapitre, toutes les cellules (on appelle "cellule" toute combinaison des différents niveaux des variables indépendantes) contiennent des sujets *différents*. Si les mêmes sujets contribuent à plusieurs cellules, i.e. donnent plusieurs scores dans l'étude, on parle d'analyse à mesures répétées, et ce cadre sera étudié ultérieurement.

1. Présentation d'un jeu de données

On considère comme exemple une expérience visant à étudier l'effet de l'âge et du sexe de 36 personnes sur l'offre d'achat ($\times 100$ euros) qui leur est faite par des garagistes pour une voiture usagée (la même voiture pour tous).

2. Description de l'expérience

• unités expérimentales : les sujets. Chaque sujet va voir un garagiste différent, on peut donc considérer que les observations sont indépendantes.
• deux facteurs. Ce sont des "facteurs de classification" (par opposition à des "facteurs expérimentaux") i.e des facteurs qui sont des caractéristiques inhérentes aux sujets. Il n'y a donc pas de randomisation possible sur les facteurs. De plus, ce sont des "facteurs fixes". Le premier, "sexe", a 2 modalités (homme/femme), tandis que le deuxième, "âge", a 3 modalités (jeune/moyen/avancé). Enfin, ce sont des facteurs croisés, il y a 6 combinaisons possibles de sexe et d'âge.
• plan balancé : 6 sujets par cellule.
• randomisation : sujets de chaque groupe choisis au hasard ; garagistes choisis au hasard ;

3. Modèle balancé

$$Y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk},$$

où : • Y_{ijk} = valeur de la variable dépendante pour l'unité k de la modalité i du premier facteur (A) et de la modalité j du deuxième facteur (B) ;
• μ = moyenne globale de la variable dépendante ;
• τ_i = effet de la modalité i du facteur A sur la variable dépendante ;
• β_j = effet de la modalité j du facteur B sur la variable dépendante ;
• $(\tau\beta)_{ij}$ = effet de l'*interaction* entre les facteurs A et B pour les unités

des modalités (i, j) ;
• ϵ_{ijk} = terme d'erreur aléatoire, tels que l'on ait indépendance des ϵ_{ijk}, et que $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$;
• $i = 1, 2, \cdots, a$; $j = 1, 2, \cdots, b$; $k = 1, 2, \cdots, c$.
• $\sum_{i=1}^{a} \tau_i = 0, \sum_{j=1}^{b} \beta_j = 0, \sum_{i=1}^{a} (\tau\beta)_{ij} = 0$ pour tout j, et $\sum_{j=1}^{b} (\tau\beta)_{ij} = 0$ pour tout i.

Le terme d'interaction permet de tenir compte du fait que l'effet d'un des facteurs peut dépendre des modalités de l'autre facteur. Par exemple, si l'écart entre les offres faites aux hommes et aux femmes varie selon l'âge. Sans interaction, les tracés des moyennes des hommes et des femmes devraient être globalement parallèles.

Attention : si l'effet de l'interaction est significatif, alors on ne peut pas énoncer de résultats concernant l'effet du sexe sur les offres d'achat sans faire référence à l'âge des sujets, et réciproquement.

4. Estimation

On fait appel aux estimateurs des moindres carrés suivants :

$$\hat{\mu} = \bar{Y}_{...}, \quad \hat{\tau}_i = \bar{Y}_{i..} - \bar{Y}_{...}, \quad \hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}, \quad \widehat{(\tau\beta)}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...} \ ,$$

où

$$\bar{Y}_{ij.} = \frac{1}{c} \sum_{k=1}^{c} Y_{ijk} \ , \bar{Y}_{i..} = \frac{1}{b} \sum_{j=1}^{b} \bar{Y}_{ij.} \ , \bar{Y}_{.j.} = \frac{1}{a} \sum_{i=1}^{a} \bar{Y}_{ij.} \ , \text{ et } \bar{Y}_{...} = \frac{1}{a} \sum_{i=1}^{a} \bar{Y}_{i..}.$$

Valeur prédite associée à l'observation (ijk) :

$$\widehat{Y}_{ijk} = \hat{\mu} + \hat{\tau}_i + \hat{\beta}_j + \widehat{(\tau\beta)}_{ij} = \bar{Y}_{ij.}.$$

5. Analyse de variance

$$\text{SST} = \text{SSA} + \text{SSB} + \text{SSAB} + \text{SSE},$$

où SST est la somme des carrés totale, SSA (resp. SSB) la somme des carrés due au facteur A (resp. B), SSAB la somme des carrés due aux interactions entre les 2 facteurs A et B, et SSE est la somme des carrés des erreurs. Ceci s'écrit mathématiquement :

$$\sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (Y_{ijk} - \bar{Y}_{...})^2 = bc \sum_{i=1}^{a} (\bar{Y}_{i..} - \bar{Y}_{...})^2 + ac \sum_{j=1}^{b} (\bar{Y}_{.j.} - \bar{Y}_{...})^2$$

$$+ c \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...})^2 + \sum_{i=1}^{a} \sum_{j=1}^{b} \sum_{k=1}^{c} (Y_{ijk} - \bar{Y}_{ij.})^2$$

Le tableau d'analyse de variance à deux facteurs se présente globalement de la même façon que celui à un facteur ; on a plus précisément, avec Statistica :

Effet	dl effet	MC effet	dl erreur	MC erreur	F	niveau p
Facteur A	$a-1$	MSA	$ab(c-1)$	MSE	$F_1 = \frac{\text{MSA}}{\text{MSE}}$	$P(Z_1 > F_1)^\star$
Facteur B	$b-1$	MSB	$ab(c-1)$	MSE	$F_2 = \frac{\text{MSB}}{\text{MSE}}$	$P(Z_2 > F_2)^\star$
Interaction AB	$(a-1)(b-1)$	MSAB	$ab(c-1)$	MSE	$F_3 = \frac{\text{MSAB}}{\text{MSE}}$	$P(Z_3 > F_3)^\star$

$\star$ la variable Z_1 (resp. Z_2, resp. Z_3) suit une loi de Fisher, à $a-1$ (resp. $b-1$, resp. $(a-1)(b-1)$) et $ab(c-1)$ degrés de liberté.

Le premier test à examiner est celui des interactions, sur la dernière ligne du tableau. Il s'agit du test de $[H_0 : (\tau\beta)_{ij} = 0 \text{ pour tout } i, j]$ contre $[H_1 : (\tau\beta)_{ij} \neq 0$ pour au moins un $(i, j)]$.

1. *Si l'hypothèse H_0 d'interaction nulle est rejetée* (i.e le niveau p est très petit, par exemple inférieur à 0.05), alors cela signifie que l'effet du facteur A est lié aux modalités du facteur B. Dans ce cas, on ne peut pas interpréter les tests effectués sur les effets simples (lignes 1 et 2 du tableau) : au lieu de cela, on compare les valeurs prises par la variable dépendante pour chaque combinaison de modalités. On peut utiliser un test de comparaisons multiples pour comparer deux à deux les différentes combinaisons de modalités.

2. *Si l'hypothèse H_0 d'interaction nulle est acceptée*, alors on peut regarder :

• le test $[H_0 : \tau_i = 0 \text{ pour tout } i]$ contre $[H_1 : \tau_i \neq 0 \text{ pour au moins un } i]$ effectué sur la première ligne ;
• le test $[H_0 : \beta_j = 0 \text{ pour tout } j]$ contre $[H_1 : \beta_j \neq 0 \text{ pour au moins un } j]$ effectué sur la deuxième ligne.