MODÈLE D'ANALYSE DE LA VARIANCE (J3, L)
(08 / 01 / 2020, © Monfort, Dicostat2005, 2005-2020)

La **théorie des plans d'expérience** et les méthodes d'**analyse de la variance** (cf **décomposition de la variance**) conduisent à un **modèle** important, appelé **modèle d'analyse de la variance**. On en présente deux aspects qui précisent les rapports entre le **modèle de régression multiple** et le formalisme des modèles pouvant être associés à un **plan d'expérience**.

(i) **Premier aspect**. Soit K+1 **vars**, dont K **variables exogènes** ξ_1 ,..., ξ_K , constituant un vecteur ξ, et une **variable endogène** η.

On appelle **modèle d'analyse de la variance (non linéaire)** un modèle de régression multiple, qui s'exprime (dans l'**espace de variables**) sous la forme (ici non linéaire pr à b) :

(1) $\eta = f(\xi , b) + \varepsilon$, avec $b \in \mathbf{R}^Q$ et $E\,\varepsilon = 0$,

dans laquelle les **va** ξ_k ($k \in N_K^*$) sont à valeurs dans $N_1 = \{0, 1\}$ et la va η à valeurs dans **R**.

Si $X \in M_{NK}$ ($\{0,1\}$) désigne la **matrice d'observation** constituée de N observations (disposées en lignes) de ξ et y le **vecteur aléatoire** à valeurs dans $\mathbf{R}^N$ constitué des N observations correspondantes de η (les « résultats » de l'**expérience**), l'équation (1) est « observée » selon :

(2) $y_n = f(X_n , b) + u_n$,avec $b \in \mathbf{R}^Q$,

où X_n est la n-ième ligne de X, y_n la n-ième coordonnée de y et u_n la n-ième **copie** de la **perturbation aléatoire** ε. On a donc $X_n = (x_{n1} ,..., x_{nK})$ et $x_{nk} \in \{0,1\}$, $\forall\ (n, k) \in N_N^* \times N_K^*$. Les équations (2) s'écrivent aussi sous la forme compacte (dans l'**espace des observations**) :

(2) $y = F(b) + u$.

Le plus souvent, le modèle (1) est supposé linéaire (ie f est bilinéaire, avec Q = K et F = X), d'où la forme usuelle :

(3) $y = X\,b + u$ ou $y = \sum_{k=1}^{K} b_k\, x_k + u$, avec $X = [x_1 ,..., x_K]$.

(ii) On appelle parfois même **plan d'expérience** un **modèle de régression** (3) tq la **matrice** X ne contient que des 0 ou des 1 et rg X < K.

Une telle matrice X, qui n'est pas de plein **rang** en colonne (ie rg X' X < K), caractérise, en effet, la plupart de tels plans.

(iii) Un **modèle d'analyse de la variance** est donc un modèle de régression multiple particulier (souvent linéaire), dans lequel les variables exogènes sont qualitatives :

ici, ce sont des **variables indicatrices**, ie des **variables dichotomiques**. Il ne vérifie donc pas une hypothèse usuelle du modèle linéaire (ie rg X = K).

Cette terminologie vient notamment de l'association souvent faite entre N_K^* et une **suite** de « classes » $\{C_1, ..., C_K\}$ (eg une **partition** de **R**) définie en posant, $\forall$ (n, k) $\in N_N^* \times N_K^*$:

$$(4) \qquad x_{nk} = \begin{cases} 1 \text{ si } y_n \in C_k , \\ 0 \text{ sinon.} \end{cases}$$

(iv) En pratique, N_K^* est mis en bijection avec le nombre K des **traitements** mis en oeuvre dans le plan d'expérience associé au modèle : K est le nombre de « combinaisons » des **facteurs expérimentaux** F_h qui agissent sur la variable η, avec $h \in N_H^*$. Si I_h désigne le nombre de **niveaux** $i_h \in \{1,..., I_h\}$ de F_h , $\forall$ $h \in N_H^*$, alors $K = \prod_{h=1}^{H} I_h$.

On note, par commodité, $I = (i(1), ..., i(H))$ la **suite** d'**indices** $(i_1, ..., i_H)$, n(I) l'entier $n_I = n_{i(1) \dots i(H)} \in \{1,..., N_I\}$) et $N_I = N_{i(1) \dots i(H)} \geq 0$ le nombre d'observations y_n (alors notées $y_{I,n(I)}$) auxquelles on applique conjointement les niveaux i_1 de F_1 ,..., et i_H de F_H .

On pose alors :

$$(5) \qquad \begin{aligned} &\mathcal{T}_h = \{1,..., I_h\} && \text{(ensemble des niveaux de } F_h), \ \forall\ h \in N_H^*, \\ &\mathcal{T} = \prod_{h=1}^{H} \mathcal{T}_h && \text{(ensemble des traitements } I = (i_1, ..., i_H)), \\ &N_I = N_{i(1) \dots i(H)} , && n_I = n_{i(1) \dots i(H)} \in \{1,..., N_I\}, \\ &N = \sum_{I \in \mathcal{T}_e} N_I && \text{(nombre d'}\textbf{unités expérimentales}\text{)}, \end{aligned}$$

où $\mathcal{T}_e \subset \mathcal{T}$ désigne l'ensemble des traitements $\mathcal{T}$ effectivement appliqués. On note alors, conformément, $y_{I,nI} = y_{i(1)\dots\ i(H)\ n\ i(1)\dots\ i(H)}$ la **mesure** (ou **observation**) de la variable η effectuée sur l'unité n_I soumise au traitement I.

Ceci conduit à un second aspect.

(v) **Second aspect**. Celui-ci relie le modèle d'analyse de la variance à la **modélisation** d'un plan d'expérience. Soit F_h ($h \in N_H^*$) une suite (finie) de facteurs susceptibles d'influer sur une **variable d'intérêt** η. On note, comme précédemment, $y_{I,nI}$ le résultat observé sur l'unité n_I auquel on applique simultanément les niveaux i_h des F_h . On peut associer à ce plan d'expérience une représentation de la forme :

$$(6) \qquad \begin{aligned} y_{I,n(I)} = \ & b^0 \\ & + (b^1_{i(1)} + ... + b^H_{i(H)}) \\ & + (b^{12}_{i(1)i(2)} + ... + b^{H-1,H}_{i(H-1)i(H)}) + ... + b^{1\dots H}_{i(1)\dots i(H)} \end{aligned}$$

$$+ \ldots$$

$$+ b^{1\ldots H}_{i(1)\ldots i(H)}$$

$$+ u_{I,n(I)} ,$$

avec $I = (i_1 ,..., i_H) \in \mathcal{J}_e$, $\mathcal{J}_h = \{1,..., I_h\}$, $\mathcal{J} = \Pi_{h=1}^{H} \mathcal{J}_h$ et $n_I \in \{1,..., N_I\}$, $\forall\ I \in \mathcal{J}_e$ (ensemble des traitements effectifs).

La représentation (6) est appelée **modèle (linéaire) d'analyse de la variance, non équilibré et avec intéractions** (de tous ordres) associé au plan d'expérience considéré, que l'on peut noter ($y_{I,n(I)}$, $n_I \in \{1,..., N_I\}$, $I \in \mathcal{J}_e$).

Une terminologie courante est la suivante (cf **plan d'expérience**) :

(a) un H-uple I est appelé **traitement**, ou parfois **case**, ou **cellule** ;

(b) N_I est le **nombre de répétitions du traitement** (ou nombre d'observations dans la « case ») I ;

(c) le terme « constant » b^0 est appelé **niveau général** de η (cf **constante**) ;

(d) les termes $b^h_{i(h)}$ ($i_h \in I_h$, $\forall\ h \in N_H^*$) sont les **effets différentiels** propres aux niveaux de chaque facteur F_h ;

(e) les termes $b^{hk}_{i(h)i(k)}$ (avec $h < k$) sont les **intéractions (d'ordre 2)** des couples de facteurs (F_h , F_k), etc.

Selon le type d'expérience étudié, les **effets factoriels** $b^h_{i(h)}$, $b^{hk}_{i(h)i(k)}$, etc, peuvent être aléatoires ou non (**effets fixes**).

(vi) Ainsi (**typologie de C. EISENHART**) :

(a) lorsque les paramètres b sont « certains », on dit que les facteurs F_h exercent des **effets « fixes »** (ie **non aléatoires**) et l'équation (6) correspond à un **modèle de régression linéaire** usuel dans lequel X comporte une **singularité**. On l'appelle **modèle de type I**. En effet, ce modèle se ramène à un **modèle de régression multiple** (linéaire) en introduisant des **variables indicatrices** ad hoc, ie :

$z_0 = e_N$ (vecteur « **constante** », ou premier vecteur bissecteur $(1 ,..., 1)' \in \mathbf{R}^N$),

$$z_1 = \text{vecteur tq } z_{i(h),\,n} = \begin{cases} 1 \text{ si l'observation n est soumise au niveau } i_h \text{ de } F_h , \\ 0 \text{ sinon,} \end{cases}$$

$$z_2 = \text{vecteur tq } z_{i(h)\,i(k)\,,\,n} = \begin{cases} 1 \text{ si n est soumise aux niveaux } i_h \text{ de } F_h \text{ et } i_k \text{ de } F_k , \\ 0 \text{ sinon,} \end{cases}$$

etc,

ce qui introduit :

(7) $1 + C_H^1 + ... + C_H^H = 2^H = K$

variables (ou **paramètres** correspondants).

On aboutit ainsi à un **modèle de régression** usuel de la forme :

(8) $y = X\,b + u, \qquad \text{avec } X \in M_{NK}(\{0,1\}) \text{ et } b \in \mathbf{R}^K .$

Cependant, le paramètre b de (8) n'est pas estimable car (cf **estimabilité**, **singularité**) :

(9) $\operatorname{rg} X < K.$

On doit donc se restreindre aux seules fonctions (linéaires) estimables de b (de la forme $b \mapsto L\,b$), ie on doit introduire des relations entre les coefficients b, donc une **contrainte** (vectorielle, en général) sur b. Le **théorème de FRISH-VAUGH** peut s'appliquer à cette situation (modèles de type I) ;

(b) lorsque les paramètres b sont stochastiques, on dit que les facteurs F_h exercent sur η des **effets aléatoires** : on parle alors de **modèle de type II**, ou de **modèle à variance composée**, ou encore de **modèle d'analyse des composantes de la variance**.

(c) enfin, si certains facteurs sont à effets fixes et d'autres à effets aléatoires, on parle de **modèle mixte d'analyse de la variance**. Un tel modèle peut être considéré comme cas particulier du précédent (en supposant eg que la **variance** de certains effets sont nuls). On le qualifie aussi de **modèle de type III**.

(vii) A titre d'exemples de modèles d'analyse de la variance, on peut citer :

(a) le modèle d'analyse de la variance à un facteur (non équilibré) ;

(b) le modèle d'analyse de la variance à deux facteurs (avec intéraction) ;

(c) le **modèle de C. EISENHART**, de la forme :

(10) $y_{I,n(I)} = m_I + u_{I,n(I)}\,, \quad \forall\, n_I \in \{1,..., N_I\}, \qquad \forall\, I \in \mathcal{T}_e\,,$

qui correspond à l'action de H facteurs F_h . Ce modèle peut se ramener à la forme (8) comme suit :

(11) $y = \Sigma_{I \in \mathcal{T}e}\ m_I\,.\,f_I + u,$

où y, les f_I ($I \in \mathcal{T}_e$) et u sont à valeurs dans $\mathbf{R}^N$ (avec $N = \Sigma_{I \in \mathcal{T}e}\ N_I$), où f_I est un vecteur dont toutes les coordonnées sont nulles, sauf celles relatives aux N_I

observations n_l soumises à l'action conjointe des niveaux i_h des facteurs F_h ($\forall$ h $\in$ N_H*), lesquelles valent alors 1. La forme (11) est bien du type (8) avec Card $\mathcal{T}_e$ = $\Pi_{h=1}^{H} I_h$ = K, X = $(f_i)_{l \in \mathcal{T}e}$ et b = $(m_l)_{l \in \mathcal{T}e}$.

Lorsque les effets sont aléatoires et indépendants (ou simplement non corrélés) entre eux, la forme (6) du modèle permet de définir la **variabilité** (ou « variance », ou **forme quadratique** de **dispersion**) totale S_T^2, que l'on peut décomposer selon la **formule d'analyse de la variance** générale suivante :

(12) $S_T^2 = \Sigma_{h=1}^{H} S_h^2 + \Sigma_{h=1}^{H} \Sigma_{k=h+1}^{K} S_{hk}^2 + ... + S_{1...H}^2 + S_u^2$,

dans laquelle S_h^2 représente la **variabilité propre** du facteur F_h , S_{hk}^2 la **variabilité conjointe** (ou l'**effet conjugué**) dû(e) à l'**intéraction** des facteurs F_h et F_k ,..., et $S_{1...H}^2$ la **variabilité commune** à l'ensemble des **intéractions factorielles**, S_u^2 étant la **variabilité résiduelle** des perturbations aléatoires $u_{l,n(l)}$.

Les **degrés de liberté** resp associés aux S_h^2 sont alors $(I_h - 1)$, ceux des S_{hk}^2 sont $(I_h - 1)(I_k - 1)$,..., ceux de $S_{1...H}^2$ sont $\Pi_{h=1}^{H} (I_h - 1)$ et ceux de S_u^2 sont au nombre de :

(13) $d_U = (N - 1) - \Sigma_{H=1}^{H} (I_h - 1) - \Sigma_{h=1}^{H} \Sigma_{k=1(k>h)}^{K} (I_h - 1)(I_k - 1) - ... - \Pi_{h=1}^{H} (I_h - 1)$.

Ainsi, S_T^2 possède N - 1 degrés de liberté, avec $N = \Sigma_{l \in \mathcal{T}e} N_l$.

(viii) L'**estimation** d'un modèle d'analyse de la variance s'effectue à l'aide des méthodes applicables aux modèles de régression, compte tenu des **singularités** exprimées dans {(8),(9)} : eg **méthode des moindres carrés ordinaires** ou **méthode du maximum de vraisemblance**, avec **contrainte sur les paramètres**.

Souvent, en pratique, notamment en matière de **tests** relatifs à b, ou encore pour estimer b par la méthode du maximum de vraisemblance, on admet l'hypothèse **stochastique** de **normalité** :

(14) $u_{l,n(l)} \sim \mathcal{N}_1 (0, \sigma_u^2)$, $\forall n_l \in \{1,..., N_l\}$, $\forall l \in \mathcal{T}_e$,

selon laquelle les **effets « extérieurs » au plan d'expérience**, ou **effets « extérieurs » au modèle**, (eg facteurs non pris en compte ou incontrôlables, erreurs de mesure, etc), mais qui peuvent influencer η, sont normalement distribués, homoscédastiques et non corrélés (ou indépendants) (cf **homoscédasticité**, **indépendance**, **corrélation**).

Sous cette hypothèse, et en notant avec une tilde (˜) les estimateurs des termes figurant dans (12), on a eg :

(15) $f_h = \{(S_h^2)^{\sim} / (I_h - 1)\} / \{(S_u^2)^{\sim} / d_u\} \sim \mathcal{F} (I_h - 1, d_u)$ (**loi de FISHER**), $\forall$ h $\in$ N_H*,

ce qui permet un **test** de l'**hypothèse** « nulle » (cf **hypothèse nulle**) :

(16) H_0^1 : $b_1 = 0$ (ie $b_{1,i(1)} = ... = b_{H,i(H)} = 0$).

De même, $\forall$ (h, k) $\in (N_H^*)^2_<$:

(17) $f_{hk} = \{(S_{hk}^2)^\sim / ((I_h - 1)(I_k - 1))\} / \{(S_u^2)^\sim / d_u\} \sim \mathcal{F}((I_h - 1)(I_k - 1), d_u)$,

ce qui permet de tester l'hypothèse d'absence d'intéraction au second ordre :

(18) $H_0^{12} : b^{hk}_{i(h)i(k)} = 0, \quad \forall (i_h , i_k) \in I_h \times I_k , \quad \forall (h, k) \in (N_H^*)^2_<$.

(ix) Le modèle d'analyse de la variance non linéaire (forme (1)) n'est guère utilisé tel quel, notamment en raison de la souplesse d'utilisation des variables indicatrices : celles-ci permettent, en effet, de coder des liaisons non linéaires (cf **codage**).

On peut noter, cependant, l'existence de modèles de type multiplicatif, eg (cas de deux facteurs non équilibrés) :

(a) avec perturbation additive :

(19) $y_{ij,n(i,j)} = b^0 . b^1_i . b^2_j + u_{ij,n(i,j)}$;

(b) avec perturbation multiplicative :

(20) $y_{ij,n(i,j)} = b^0 . b^1_i . b^2_j (1 + u_{ij,n(i,j)})$,

d'où Log $y_{ij,n(i,j)}$ = Log b^0 + Log b^1_i + Log b^2_j + Log $(1 + u_{ij,n(i,j)})$, avec l'approximation éventuelle Log $(1 + u_{ij,n(i,j)}) \# u_{ij,n(i,j)}$;

(c) avec perturbation exponentielle :

(21) $y_{ij,n(i,j)} = b^0 . b^1_i . b^2_j \exp(u_{ij,n(i,j)})$,

d'où la forme Log $y_{ij,n(i,j)}$ = Log b^0 + Log b^1_i + Log b^2_j + $u_{ij,n(i,j)}$.

(x) Le modèle d'analyse de la variance est généralement robuste pr à l'hypothèse de normalité (tq (14)), ou pr à l'hypothèse d'homoscédasticité des perturbations (σ_u^2 ne dépend pas des observations n_l , $\forall$ l) (cf **robustesse**).

Il est cependant en général peu robuste pr à l'**hypothèse de sphéricité** (ie de **dispersion** diagonale, ou de non corrélation, des perturbations) (cf **loi sphérique**) : dans ce cas, les estimateurs sont toujours des **estimateurs sans biais**, mais les **tests** ou les **régions de confiance** relatifs aux paramètres sont biaisés.

Lorsque les perturbations sont hétérocédastiques, ou lorsqu'elles sont corrélées, on utilise généralement la **méthode des moindres carrés généralisés** (compte tenu des contraintes sur les paramètres), ou encore une **méthode non paramétrique** ou une **méthode affranchie**.

(xi) Le modèle d'analyse de la variance peut s'étendre :

(a) à un modèle multivarié (ie comportant plusieurs variables endogènes η_1 ,..., η_G), eg pour prendre en compte plusieurs types de **mesures** (« résultats » d'observation) effectuées sur les unités expérimentales. Les formules précédentes se généralisent directement ;

(b) à des variables qualitatives plus générales que les indicatrices précédentes. On procède alors à un **codage** des modalités de chaque **variable qualitative** à l'aide de variables indicatrices du type précédent : chaque modalité de chaque variable qualitative est codée à l'aide d'une indicatrice spécifique, ce qui constitue un « **codage disjonctif complet** ».