SINGULARITÉ (A3, J, O)

(22 / 03 / 2020, © Monfort, Dicostat2005, 2005-2020)

(i) On parle de **singularité** lorsqu'une **condition de régularité**, ou **hypothèse de régularité**, qui est à la base d'une définition, d'une méthode, d'une propriété ou d'une **procédure** (statistiques), n'est pas vérifiée (cf aussi **procédure statistique**).

Une telle condition peut notamment être :

(a) de nature mathématique : **fonction de vraisemblance** non différentiable, **plan d'expérience** dont la **matrice** n'est pas régulière (cf **matrice singulière**), etc ;

(b) de nature plus « statistique » : variable non gaussienne, loi non symétrique, suite de v_a non indépendantes, etc (cf **loi gaussienne**, **loi symétrique**, **indépendance stochastique**, **suite indépendante**).

(ii) Plus spécifiquement, on parle de **singularité** lorsqu'une **matrice** (**matrice d'observation**, **matrice de covariance**, etc) n'est pas de plein **rang**.

Parmi des exemples classiques :

(a) le **modèle de régression multiple** linéaire (resp non linéaire), écrit dans l'**espace des observations**, $y = X b + u$ (resp $y = F(b) + u$), avec $E u = 0$ et $V u = \Sigma$, peut conduire à deux situations :

(a)₁ s'il existe une **colinéarité entre variables exogènes**, on parle de singularité par aux hypothèses de base de ce type de modèles. Cette **colinéarité** se traduit par :

$$(1) \quad \text{rg } X = L < K < N \quad (\text{resp } \text{rg } D F(b) = M < Q < N).$$

Dans ce cas, le **paramètre** $b \in \mathbf{R}^K$ (resp $b \in \mathbf{R}^Q$) n'est pas estimable (cf **estimabilité**). Par contre, $E y = X b$ est toujours estimable : son estimateur est $(E y)^{\sim} = P y$, où P est la matrice associée au **projecteur** de $\mathbf{R}^N$ sur $\mathbf{R}^L$ (resp $\mathbf{R}^M$), lequel associe à y sa **projection** $P y$ dans le sous-espace engendré par les colonnes de X (resp dans la variété de dimension M définie par l'équation vectorielle $z = F(b)$, $\forall b \in \mathbf{R}^Q$).

Dans le cas linéaire $y = X b + u$, l'**équation estimante** :

$$(2) \quad X' X b = X' y$$

ne possède pas de solution unique en b . On peut traiter le problème :

((a)₁)_a soit en remplaçant b par un sous-vecteur $b_L \in \mathbf{R}^L$ correspondant à L colonnes de X linéairement indépendantes, et en négligeant les $K - L$ autres paramètres (et colonnes). Cette solution dissymétrique est, en général, arbitraire ;

((a)₁)_b soit par **régression sur composantes principales** basée sur X ;

$((a)_1)_c$ soit en rendant b estimable (donc le modèle linéaire identifiable) (cf **modèle identifiable**), en imposant une condition supplémentaire (et symétrique) sur b . Par exemple :

$$(3) \quad R b = r,$$

avec $R \in M_{JK}(\mathbf{R})$, $r \in \mathbf{R}^J$ et $J \geq K - L$.

On montre que le système $\{(2),(3)\}$ admet une solution unique ssi (a) aucune combinaison linéaire non nulle des lignes de R n'est égale à une combinaison linéaire des lignes de X (condition d'existence), et (b) la matrice par blocs $(X' X' / R)$ est de rang $\text{rg}(X' X' / R) = K$ (condition d'unicité). Cette solution unique est :

$$(4) \quad \hat{b} = (X' X + R' R)^{-1} X' y.$$

Lorsque Σ est de la forme $\sigma^2 I_N$ (**matrice scalaire**), on l'appelle parfois **estimateur des moindres carrés ordinaires singulier** de b .

La colinéarité des variables exogènes introduit donc une singularité qui nécessite l'utilisation des notions d'**estimabilité** (du paramètre b) et de **matrice inverse généralisée**.

$(a)_2$ si le **vecteur aléatoire** u (vecteur des **perturbations aléatoires** du modèle) prend ses valeurs dans une variété propre (ie stricte) de $\mathbf{R}^N$, cet autre type de singularité concerne les hypothèses stochastiques habituelles du modèle. Cette **singularité** se traduit par (cf **loi dégénérée**) :

$$(5) \quad \text{rg} \Sigma < N.$$

Ceci peut se produire notamment lorsqu'il existe des contraintes (linéaires) de la forme $L y = h$ (où L et h sont donnés) sur le vecteur y des **observations** de la **variable endogène** η , puisque $V y = V u = \Sigma$ (cf **contrainte sur les variables**) ;

(b) le **modèle d'analyse de la variance** peut s'écrire sous la forme d'un **modèle linéaire** du type précédent, soit :

$$(6) \quad y = X b + u, \quad \text{avec } E u = 0, V u = \sigma^2 \cdot I_N.$$

Le **paramètre** b représente l'ensemble des **effets factoriels** $\beta_{i(h)}^h, \beta_{i(h)j(k)}^{hk}$, etc (en nombre K). En général, b (vecteur des paramètres β précédents) n'est pas estimable et l'on doit se restreindre aux fonctions linéaires de b (cf **estimabilité**).

Autrement dit, si H est le nombre de **facteurs** du modèle, on introduit H relations (non redondantes) sur les paramètres b (eg des relations de la forme $\sum_{h=1}^H \beta_{i(h)}^h = 0$, ie des **contrastes** nuls). Ces relations se traduisent par une **contrainte sur les paramètres** du type $R b = 0$, où $R \in M_{KL}(\{0,1\})$ et où $N = \sum_{l \in \mathcal{I}} N_l$ est le nombre total des observations de la **variable endogène** η .

Pour estimer b par la **méthode des moindres carrés ordinaires** (avec contrainte sur les paramètres), on considère le système à $K + L$ équations et $K + L$ inconnues suivant :

$$(7) \quad \begin{aligned} X' X b + R' \lambda &= X' y, \\ R b &= 0. \end{aligned}$$

ie $A c = d$, en notant A la matrice par blocs $(X' X, R' / R, 0)$, c le vecteur colonne par blocs (b / λ) et d le vecteur colonne par blocs $(X' y, 0)$.

Bien que $\text{rg } X' X < K$, le **rang** de la matrice A du système en (b, λ) précédent vaut cependant $K + L$.

Par suite, le modèle défini par (6) et $R b = 0$ est juste identifié, et possède donc une solution des mco $b^\wedge$ unique (cf **modèle identifié**).

Si les relations linéaires $R b = 0$ représentent des **contrastes**, on dit que seuls les contrastes de ce modèle d'analyse de la variance sont estimables.

(iii) Les exemples précédents montrent aussi les relations entre notions d'**estimabilité** des paramètres et d'**identification** des (équations des) modèles.

On rencontre aussi ce type de situations dans le cadre du **modèle d'interdépendance** (lorsqu'il contient des **identités** reliant des variables) ou celui du **modèle d'analyse de la covariance**, pour lequel les développements précédents sont encore applicables.

En pratique, le choix d'une condition supplémentaire sur b (condition d'estimabilité) n'est pas toujours simple : eg dans les modèle dérivés de plans d'expérience non équilibrés (cf **plan équilibré**).