COLINÉARITÉ (A3, H, J1)
(08 / 05 / 2020, © Monfort, Dicostat2005, 2005-2020)

La notion de **colinéarité** s'associe, en **Statistique**, à un « défaut » particulier dans un modèle, notamment linéaire. Sa définition est d'origine algébrique.

Ce défaut peut provenir notamment d'**identités** (**équation comptable**, etc), non prises en compte, reliant les variables exogènes. Il est « congénital » dans certains modèles (**modèle d'analyse de la variance**, **modèle d'analyse de la covariance**) associés à des **plan d'expérience**.

(i) En algèbre linéaire, soit E un **espace vectoriel** sur un corps commutatif **K**. On dit que $(x, y) \in E^2$ est un **couple d'éléments colinéaire(s)** ssi il existe un couple $(\lambda, \mu) \in \mathbf{K}^2 \setminus (0, 0)$, tq :

(1) $\lambda\, x + \mu\, y = 0.$

Par suite, une **matrice** carrée dont deux colonnes (resp lignes) tq x et y sont colinéaires n'est pas de plein **rang**.

(ii) En Statistique, et particulièrement dans l'étude d'un **modèle de régression**, on utilise la définition suivante. Soit $x = \{x_1, ..., x_K\}$ une suite finie de vecteurs de E. On dit que x est une **suite colinéaire** ssi il existe une suite $\{\lambda_1, ..., \lambda_K\} \in \mathbf{R}^K$, non (identiquement) nulle, ie $(\lambda_1, ..., \lambda_K) \neq (0, ..., 0)$, tq :

(2) $\Sigma_{k=1}^{K} \lambda_k . x_k = 0.$

Certains coefficient λ_k peuvent, ici, être nuls, mais le nombre de coefficient non nuls doit être supérieur ou égal à 2. Autrement dit, il existe une suite $\{\lambda_1, ..., \lambda_L\} \in \mathbf{R}^L$, (avec $1 < L < K$) tq :

(2)' $\Sigma_{l=1}^{L} \lambda_l . x_l = 0.$

On écrit souvent « **collinéaire** » au lieu de colinéaire, et l'on dit aussi **multicolinéaire** (voire même multicollinéaire) au lieu de colinéaire.

(iii) Pour un modèle de **régression linéaire** (ou affine), écrit dans l'**espace d'observation** $\mathbf{R}^N$ selon $y = X\, b + u$, dans lequel $y (\in) \mathbf{R}^N$ (ie y est à valeurs dans $\mathbf{R}^N$), $X (\in) M_{NK} (\mathbf{R})$ (avec $K < N$), $E\, u = 0$ et $V\, u = \sigma^2\, I_N$, la **colinéarité stricte** se traduit par l'existence d'une **singularité** originaire des **variables exogènes**. Cette dégénérescence se caractérise par :

(3) $1 < rg\, X < K.$

Autrement dit, la **matrice** X n'est pas de plein rang en colonnes, ou encore, la variété linéaire qui supporte la suite $x = \{x_1, ..., x_K\}$ est de dimension inférieure à K.

Lorsque la colinéarité est stricte, on ne peut que définir des fonctions estimables des coefficients b (cf **estimabilité**) et l'on utilise souvent l'inversion généralisée des matrices (cf **inverse**, **matrice inverse généralisée**).

Lorsque r = rg X < K, on peut écrire X h = 0 pour un certain vecteur h $\in \mathbf{R}^K$, ou encore dét (X' X) = 0. Dans le système des **équations normales** X' y = X' X b, la **matrice** X' X correspond à une **application linéaire** surjective (cf **application surjective**) et elle n'est pas (ordinairement) inversible : b ne peut donc être estimé. On peut néanmoins estimer certaines **formes linéaires** l : b $\mapsto$ l' b de b ssi, en notant A^+ la **matrice pseudo-inverse** de A, on a :

(2) $\quad l' (I_K - X^+ X) = 0 \quad$ (ie l $\perp$ Ker X),

ou ssi il existe $(\alpha_1, ..., \alpha_r)' \in \mathbf{R}^r$ tq :

(3) $\quad l = \Sigma_{i=1}^{r} \alpha_i . v_i$,

où v_i représente un **vecteur propre** associé à une **valeur propre** $\lambda_i > 0$ de X' X ($\forall$ i $\in N_r^*$).

Par suite, le meilleur estimateur linéaire sans biais de l' b est l' X^+ y (cf **estimateur sans biais**). On montre que :

(4)
$$E (l' X^+ y) = l' b,$$
$$V (l' X^+ y) = \sigma^2 l' (X' X)^+ l.$$

(iv) Dans le même cadre (modèle de régression linéaire y = X b + u, avec E u = 0 et V u = $\sigma^2 . I_N$), on dit qu'il y a **colinéarité approchée**, ou **quasi-colinéarité**, entre les variables exogènes ssi :

(5) $\quad$ rg X = K $\quad$ et $\quad$ dét (X' X) $< \infty \quad$ (ie abs (dét (X' X)) << + ∞).

Il existe alors une suite $(\lambda_k)_{k=1,...,K}$ tq le vecteur $z = \Sigma_{k=1}^{K} \lambda_k x_k$ ait une **norme** « petite », eg :

(6) $\quad ||z|| = \max_{n=1}^{N} |z_n| << + \infty$.

Plusieurs **indices de colinéarité** (approchée) permettent de détecter celle-ci. Si l'on note R_X la **matrice des corrélations** empiriques calculées à partir de X, ces indices sont eg :

(a) des valeurs propres quasi-nulles de R_X ;

(b) des **corrélations** deux à deux r_{kl} proches de 1 en valeur absolue ;

(c) des coefficients $c_k^2 = 1 - (1 / r^{kk})$ proches de 1 (où r^{kl} désigne l'élément courant de l'inverse R_X^{-1}).

(v) En cas de quasi-colinéarité, la **dispersion** $V\,\hat{b} = \sigma^2 (X'\,X)^{-1}$ de l'estimateur $\hat{b}$ du paramètre b, estimé par la **méthode des moindres carrés ordinaires**, est importante, ce qui peut s'apprécier de plusieurs façons.

Pour estimer b, plusieurs méthodes sont possibles :

(a) **régression sur composantes principales** ;

(b) **estimateur de HOERL-KENNARD** ;

(c) **modification** de la **spécification** du modèle, par élimination, ou transformation de certaines variables exogènes.

On peut aussi :

(a) soit estimer une forme linéaire $l : b \mapsto l'\,b$ dans laquelle l est combinaison linéaire des vecteurs propres associés aux plus grandes valeurs propres de X' X (régression sur composantes principales précédente). Un **test préliminaire** (ie préliminaire à l'estimation) portant sur les valeurs propres aide à déterminer les composantes à retenir ;

(b) soit intégrer au modèle des **informations** supplémentaires : observations nouvelles, lorsque c'est possible, ou **théorie bayésienne** ;

(c) soit, le cas échéant, éliminer les observations (X_n , y_n) qui contribuent le plus à réduire le **rang** de X' X.

(vi) Dans le **modèle non linéaire** $y = F(b) + u$, avec $E\,u = 0$ et $V\,u = \sigma^2 . I_N$, la **colinéarité stricte (locale)** se traduit par l'existence d'un **voisinage** $V \subset \mathbf{R}^Q$ de b (ou d'un estimateur $\tilde{b}$ de b) tq $rg\,D\,F(b) < Q$, $\forall\, b \in V$.

Une définition de la **colinéarité approchée (locale)** peut aussi être donnée. Le traitement suit la même démarche que dans le cas du modèle linéaire, en étudiant le **modèle linéarisé** du modèle initial.