.

Le Modèle linéaires et ses généralisations

Master de Statistiques
Yaoudé, Janvier 2005

Xavier Guyon
Samos - Université Paris 1

Ces notes sont celles du cours "Modèle linéaire et ses généralisations" donné au Master de statistiques de l'Université de Yaoundé, Cameroun. Une partie complémentaire portant sur le planification expérimentale à été donnée par Michel Mdoumbe. Le logiciel retenu pour traiter les exercices était `R`*. Le plan de ce cours est le suivant :*

Chapitre 1

Le modèle linéaire

1.1 Le modèle linéaire standard

1.1.1 Introduction

Un *modèle de régression* explique la valeur espérée $E(y)$ d'une variable réelle y à partir de conditions x observables et d'un paramètre inconnu β. *Le modèle est linéaire* si $E(y)$ est linéaire en β. y est la variable à expliquer (ou variable endogène, variable dépendante), x la variable explicative x (variable exogène, variable indépendante). Le fait que y est réelle est important pour la définition du modèle linéaire. Par contre x peut être quantitative (un vecteur de $\mathbf{R}^p$), qualitative (une variable de classe) ou mixte, des composantes de x étant quantitatives, d'autres étant qualitatives.

Lorsque $x \in \mathbf{R}^p$, on dit que x est un *régresseur* et on parle de modèle de *régression linéaire :* par exemple y est une vitesse de circulation coronarienne mesurée par effet Doppler, et $x = (X, T)$ où X est le poids de l'individu et T son taux de cholestérol. Si x est qualitative, on parle de modèle d'*analyse de la variance* : par exemple y est un rendement à l'hectare d'une culture et $x = (X, Z)$ croise deux facteurs qualitatifs, X un mode de culture et Z l'apport ou non d'un engrais azoté. Enfin, si x est mixte, on parle d'*analyse de la covariance :* par exemple, y est une durée de survie à un cancer du sein et $x = (T, X)$ où T est le traitement appliqué et X l'âge d'apparition du cancer.

L'ajustement statistique, ou estimation du modèle, se fait sur la base d'observations individuelles $\{(y_i, x_i), i = 1, n\}$, l'indice individuel i étant remplacé par t lorsque les données temporelles.

1.1.2 Le modèle linéaire standard

Supposons que les observations sont $\{(x_i, y_i), i = 1, n\}$ et que $x_i \in \mathbf{R}^p$. *Le modèle linéaire standard* traduit la dépendance linéaire de l'espérance en $\beta = {}^t(\beta_1, \beta_2, \cdots, \beta_p)$, un paramètre inconnu non-contraint de $\mathbf{R}^p$:

$$y_i = {}^t x_i \beta + \varepsilon_i, i = 1, n \tag{1.1}$$

On fait les hypothèses suivantes (H_ε) sur les résidus (ε_i) :

$$(H_\varepsilon) : \begin{cases} \text{(i) } E(\varepsilon_i \mid x_i) = 0, \\ \text{(ii) } Var(\varepsilon_i) = \sigma^2 \text{ et} \\ \text{(iii) } Cov(\varepsilon_i, \varepsilon_j) = 0 \text{ si } i \neq j \end{cases}$$

(i) les résidus sont *centrés* conditionnellement à x ; (ii) ils sont de *variances finies*, toutes égales à σ^2 ; (iii) ils sont *décorrélés.* On dit alors que ε est un *bruit blanc* (noté BB).

Lues sur les variables endogènes $(y_i, i = 1, n)$, ces propriétés s'écrivent :

$$\text{(i) } E(y_i) = {}^t x_i \beta, \text{ (ii) } Var(y_i) = \sigma^2, \text{ (iii) } Cov(y_i, y_j) = 0 \text{ si } i \neq j$$

Les paramètres du modèle sont β et σ^2. Ajuster le modèle, c'est estimer ces paramètres.

Ecriture matricielle du modèle

Notons $Y = {}^t(y_1, y_2, \cdots, y_n) \in \mathbf{R}^n$ le vecteur des observations, $\varepsilon = {}^t(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n)$ celui des résidus, $X = (x_{ij})_{j=1,p}^{i=1,n} = (X_1, X_2, \cdots, X_p)$ la matrice exogène $n \times p$. La i-ème ligne de X n'est autre que ${}^t x_i$, la transposée de la i-ème condition exogène $x_i \in \mathbf{R}^p$. X est appelée la matrice du *dispositif expérimental*. La k-ième colonne $X_k \in \mathbf{R}^n$ de X correspond aux n réalisations de la k-ième variable exogène x_{ik}, $i = 1, n$: X_k est le k-ième vecteur exogène.

L'écriture matricielle du modèle (1.1) est :

$$(ML1) : Y = X\beta + \varepsilon \ , \ E(\varepsilon) = 0 \text{ et } Cov(\varepsilon) = \sigma^2 I_n \tag{1.2}$$

$E(\varepsilon) = 0$ est le vecteur des espérances des n coordonnées de ε ; $Cov(\varepsilon)$ est la matrice $n \times n$ de terme (i, j) égal à $Cov(\varepsilon_i, \varepsilon_j)$. La linéarité de l'espérance $E(Y)$ en β s'écrit : $E(Y) = X\beta = \sum_1^p \beta_k X_k$.

Identifiabilité du paramètre β

La paramétrisation $E(Y) = X\beta$ est *identifiable,* ou encore propre, si la décomposition $E(Y) = X\beta$ sur les vecteurs exogènes $X_1, X_2, \cdots, X_p$ est unique. Dans ce cas, β_k s'interprète comme la coordonnée de $E(Y)$ sur la k-ième exogène X_k. Cette condition équivaut au fait que $X_1, X_2, \cdots, X_p$ sont linéairement indépendants dans $\mathbf{R}^n$, ou encore que X est de rang plein égal à p. Le sous-espace vectoriel $\mathcal{E}_X$ de $\mathbf{R}^n$ engendré par $X_1, X_2, \cdots, X_p$ est alors de dimension p ; c'est l'espace auquel appartient la moyenne $E(Y)$.

Non-identifiabilité. Sans identifiabilité, la représentation $E(Y) = X\beta$ en β n'est pas unique et β n'est ni interprétable, ni estimable. Pour rendre une paramétrisation identifiable, il suffit de sélectionner une sous-famille de régresseurs linéairement indépendants qui engendrent l'espace de la moyenne $\mathcal{E}_X$.

Paramétrisations équivalentes. (Z, γ), où Z est une matrice $n \times p$ et $\gamma \in \mathbf{R}^p$, est une paramétrisation équivalente à (X, β) si $E(Y) = X\beta \equiv Z\gamma$. Les modèles associés à l'une ou à l'autre des paramétrisations sont identiques. Seule l'interprétation des paramètres change. Si M est une matrice $p \times p$ régulière connue, les paramétrisations (X, β) et (Z, γ), avec $Z = XM$ et $\beta = M\gamma$, sont équivalentes.

1.1.3 Exemples de modèles linéaires

Example 1 *Régression affine simple*

Ce modèle fait dépendre de façon affine $E(y_i)$ en fonction d'une exogène x_i réelle :

$$E(y_i) = \beta_0 + \beta_1 x_i, \ i = 1, n$$

Notons $\mathbf{1}$ le vecteur de $\mathbf{R}^n$ dont toutes les coordonnées valent 1, $\mathbf{x} = {}^t(x_1, x_2, \cdots, x_n)$. Le modèle s'écrit $E(Y) = X\beta$ avec $X = (\mathbf{1}, \mathbf{x})$ et $\beta = {}^t(\beta_0, \beta_1)$. Deux variables exogènes, $\mathbf{1}$ et $\mathbf{x}$, expliquent $E(Y)$. Le paramètre β est de dimension 2. β_0 est l'intercepte ou ordonnée à l'origine, β_1 la pente en x. La paramétrisation est propre dès que deux valeurs x_i sont différentes. Une paramétrisation équivalente est $E(y_i) = \beta'_0 + \beta_1 x_i$ avec $\beta'_0 = \beta_0 - \beta_1 \overline{x}$ où $\overline{x} = \frac{1}{n} \sum x_i$.

La régression linéaire sur x, $E(y_i) = \beta_1 x_i$, $i = 1, n$, ne comporte qu'un seul paramètre β_1.

Example 2 *Régression affine multiple*

C'est la généralisation du modèle précédent au cas de p exogènes $x_{i1}, x_{i2}, \cdots, x_{ip}$:

$$E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \ i = 1, n \tag{1.3}$$

Ce modèle dépend de $(p + 1)$ paramètres et admet pour matrice $X = (\mathbf{1}, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p)$. Le modèle est identifiable si les $(p + 1)$ colonnes de X sont linéairement indépendantes.

Example 3 *Modèle de courbe de croissance*

Si Y_t dépend d'une variable $t \in \mathbf{R}$, une façon de modéliser $f(t) = E(Y_t)$ est de décomposer $f(t)$ sur une base de fonctions connues $x_1(t), x_2(t), \cdots, x_p(t)$:

$$f(t) = E(Y_t) = \sum_{k=1}^{p} \beta_k x_k(t)$$

Si les observations ont lieu en $t_1, t_2, \cdots, t_n$, la matrice exogène vaut $X = (X_{ik})$, avec $X_{ik} = x_k(t_i)$, $i = 1, n$ et $k = 1, p$. Par exemple, le modèle polynomial de degré 2 correspond au choix $x_1(t) \equiv 1$, $x_2(t) = t$ et $x_3(t) = t^2$: $E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2$, $X = (\mathbf{1}, \mathbf{t}, \mathbf{t}^2)$. Le modèle est identifiable si 3 valeurs t_i sont distinctes.

Example 4 *Analyse de la variance à un facteur*

C'est la situation où l'exogène $x \in A = \{a_1, a_2, \cdots, a_p\}$ est une variable de classe à p modalités. Supposons que pour $x = a_i$ on dispose de $n_i > 0$ observations y_{ik}, $k = 1, n_i$, ceci pour $i = 1, p$. Le modèle d'analyse de la variance s'écrit :

$$E(y_{ik}) = m_i,\ k = 1, n_i\ ;\ i = 1, p$$

Le nombre total d'observations est $n = \sum_{i=1}^{p} n_i$, le paramètre $\beta = {}^t(m_1, m_2, \cdots, m_p)$. La matrice $X = n \times p$ est constituée uniquement de 0 et de 1 :

$$X = \begin{pmatrix} \mathbf{1}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & \mathbf{1}_{n_p} \end{pmatrix}$$

Ce modèle est identifiable. Cette structure particulière de X conduit à une résolution simplifiée de l'estimation qui sera étudiée au chapitre 2.

Example 5 *Modèle d'analyse de la covariance*

Certains régresseurs sont quantitatifs, d'autres qualitatifs. Par exemple, pour :

$$E(y_{ik}) = a_i + b_i x_{ik},\ k = 1, n_i; i = 1, p$$

i repère la modalité qualitative (une CSP) et x_{ik} est une variable réelle (le revenu). Ce modèle dépend de $2p$ paramètres $\beta = {}^t({}^t\mathbf{a}, {}^t\mathbf{b})$, $\mathbf{a} = {}^t(a_1, a_2, \cdots, a_p)$, $\mathbf{b} = {}^t(b_1, b_2, \cdots, b_p)$. Notant $y = (y_{11}, y_{12}, \cdots, y_{1n_1}\ ;\ \cdots\ ;\ y_{p1}, y_{p2}, \cdots, y_{pn_p})$ et $\mathbf{x}_i = {}^t(x_{1i}, x_{2i}, \cdots, x_{n_i,i})$, on a :

$$X = (X_a, X_b)\text{, avec } X_a = \begin{pmatrix} \mathbf{1}_{n_1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{1}_{n_p} \end{pmatrix},\ X_b = \begin{pmatrix} \mathbf{x}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{x}_p \end{pmatrix}$$

Le modèle est identifiable si pour chaque i, il existe deux modalités exogènes différentes.

Example 6 *Modélisation exogène d'une série temporelle*

Une série temporelle présentant une tendance T et une saisonnalité trimestrielle S est modélisée par :

$$E(y_t) = T(t) + S(t) \text{ avec par exemple : } \begin{cases} T(t) = a + b f_1(t) + c f_2(t) \\ S(t) = \beta_{[t]} \end{cases}$$

$[t] \in \{1, 2, 3, 4\}$ repère le numéro du trimestre de l'instant t : il y a 4 paramètres de saisonnalités. La tendance a été décomposée sur trois fonctions $\mathbf{1}$, f_1 et f_2. Le nombre de paramètres apparents est 7 : $\beta = {}^t(a, b, c, \beta_1, \beta_2, \beta_3, \beta_4)$. Mais sous cette forme β *n'est*

pas identifiable : en effet les régresseurs $\mathbf{1}_n$, S_1, S_2, S_3 et S_4 ($S_k(i) = 1$ si i est congru à k modulo 4, $S_k(i) = 0$ sinon, $k = 1, 4$) sont liés puisque $\mathbf{1}_n = S_1 + S_2 + S_3 + S_4$. Une paramétrisation propre s'obtiendra par exemple en supprimant le régresseur $\mathbf{1}_n$ (et le paramètre a associé) : le modèle est de dimension 6 et non 7 comme aurait pu le laisser croire la paramétrisation initiale. Une autre paramétrisation s'obtient en recentrant les effets saisonniers, les paramètres β étant contraints par $\beta_1 + \beta_2 + \beta_3 + \beta_4 = 0$ et le paramètre a étant maintenu.

1.2 L'estimation par moindres carrés ordinaires

On supposera toujours par la suite que le modèle est identifiable. L'estimation de β par moindres carrés ordinaires (MCO) est une valeur $\widehat{\beta}$ qui minimise la somme de carrés résiduelle :

$$SCR(\beta) = \|Y - X\beta\|^2 = \sum_{i=1}^{n} (y_i - {}^t x_i \beta)^2$$

La fonction $\beta \mapsto SCR(\beta)$ est strictement convexe. En effet, la matrice $\frac{\partial^2 SCR(\beta)}{\partial \beta^2} = 2 \times {}^tXX$ est définie positive (notée dp[1]) puisque X est de rang plein : $\widehat{\beta}$ est donc unique, annulant le gradient de $SCR(\beta)$, $\frac{\partial}{\partial \beta} SCR(\widehat{\beta}) = 0$:

$$\sum_i x_i (y_i - {}^t x_i \widehat{\beta}) = 0, \text{ ou encore } {}^tXX\widehat{\beta} = {}^tXY$$

${}^tXX$ étant inversible, l'estimation des MCO (notée EMCO) est :

$$\widehat{\beta} = ({}^tXX)^{-1} \, {}^tXY \tag{1.4}$$

Proposition 1 *Estimation des MCO*

(a) $\widehat{\beta}$ estime sans biais β : $E(\widehat{\beta}) = \beta$. Sa variance vaut : $Var(\widehat{\beta}) = \sigma^2({}^tXX)^{-1}$.

(b) Théorème de Gauss-Markov : parmi les estimateurs linéaires et sans biais de β, l'EMCO $\widehat{\beta}$ est l'estimateur de moindre variance[2] (en anglais, BLUE pour Best Linear Unbiased Estimator).

(c) $\widehat{\sigma}^2 = \frac{SCR(\widehat{\beta})}{n-p} = \frac{1}{n-p} \left\| Y - X\widehat{\beta} \right\|^2$ estime sans biais σ^2.

Preuve :

(a) $\widehat{\beta}$ est unique, donnée par (1.4). Posant $A = ({}^tXX)^{-1} \, {}^tX$, $\widehat{\beta} = AY$ et donc $E(\widehat{\beta}) = AE(Y) = AX\beta = \beta$: $\widehat{\beta}$ est sans biais. Sa variance vaut (cf. § 14.1) :

$$Var(\widehat{\beta}) = AVar(Y){}^tA = \sigma^2({}^tXX)^{-1} \times ({}^tXX) \times ({}^tXX)^{-1} = \sigma^2({}^tXX)^{-1}$$

(b) Soit $\beta^* = CY$ un autre estimateur linéaire et sans biais de β ; la condition sans biais se traduit par $E(CY) = CX\beta = \beta$, soit $CX = I_p$. Ecrivant $C = A + M$, on vérifie que $M{}^tA = A{}^tM = 0$, et donc :

$$Var(\beta^*) = \sigma^2 \, {}^tCC = \sigma^2\{M{}^tM + {}^tAA\} = Var(\widehat{\beta}) + \sigma^2 M{}^tM$$

${}^tMM$ étant sdp, $\widehat{\beta}$ est de moindre variance que β^*.

(c) *Interprétation géométrique de $\widehat{Y} = X\widehat{\beta}$ et estimation de σ^2.*

[1] M réelle et symétrique, de dimension $p \times p$, est semi-définie positive (sdp) si pour tout $u \in \mathbf{R}^p$, ${}^tuMu \geq 0$. Elle est définie positive (dp) si ${}^tuMu > 0$ pour tout $u \neq 0$.

Une caractérisation de la sdp (resp. de la dp) est la suivante : notons $M(k) = (M_{ij})_{1 \leq i,j \leq k}$ la matrice $k \times k$ extraite, $k = 1, p$; alors M est sdp (resp. dp) $\Longleftrightarrow$ pour $k = 1, p$, $\det\{M(k)\} \geq 0$ (resp. > 0).

[2] Deux estimateurs sans biais $\widehat{\theta}$ et θ^* de $\theta \in \mathbf{R}^p$ peuvent être comparés au moyen de leurs variances : $\widehat{\theta}$ est de moindre variance que θ^* si $\Delta = Var(\theta^*) - Var(\widehat{\theta})$ est sdp. Si $p = 1$, ceci équivaut à $Var(\theta^*) \geq Var(\widehat{\theta})$.

$E(Y) = X\beta$ appartient à $\mathcal{E}_X$, l'espace de la moyenne de Y. $\widehat{\beta}$ étant l'unique valeur minimisant $\|Y - X\beta\|^2$, $\widehat{Y} = X\widehat{\beta}$ est la projection orthogonale de Y (vecteur de $\mathbf{R}^n$) sur $\mathcal{E}_X$ (sous-espace de dimension p). Soit P la projection orthogonale de $\mathbf{R}^n$ sur $\mathcal{E}_X$:

$$PY = X\widehat{\beta} \text{ , soit } P = X({}^tXX)^{-1}\,{}^tX$$

$Q = I_n - P$ est la projection orthogonale sur $\mathcal{E}_X^{\perp}$, le sous-espace de $\mathbf{R}^n$ orthogonal à $\mathcal{E}_X$. Comme $QX = 0$, $QY = Q\varepsilon$. Soit $\mathcal{B}$ une base orthonormale de $\mathbf{R}^n$ constituée d'une base de $\mathcal{E}_X$ complétée par une base de $\mathcal{E}_X^{\perp}$: $\mathcal{B} = \{f_1, f_2, \cdots, f_p \mid f_{p+1}, \cdots, f_n\}$. Dans $\mathcal{B}$, $e = Q\varepsilon = {}^t(0, 0, \cdots, 0 \mid e_{p+1}, \cdots, e_n)$. On en déduit :

$$Var(Q\varepsilon) = \sigma^2 Q\,{}^tQ = \sigma^2 \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{pmatrix}$$

Donc, $E(SCR(\hat{\beta})) = \|Q\varepsilon\|^2 = E(\sum_{p+1}^n e_j^2) = (n-p)\sigma^2 : \widehat{\sigma}^2$ estime sans biais de σ^2. .□

Pour un changement de paramètre régulier $\theta = M\beta$, M étant connue, l'EMCO de θ est $\widehat{\theta} = M\widehat{\beta}$.

Example 7 *Estimation d'une droite de régression*

Considérons la régression affine : $E(y_i) = a + bx_i, i = 1, n$. Notons : $\overline{x} = \frac{1}{n}\sum_i x_i$ et $Var(x) = (\frac{1}{n}\sum_i x_i^2) - \overline{x}^2 = \frac{1}{n}\sum_i (x_i - \overline{x})^2$ la moyenne et la variance empirique des x, $Cov(x, y) = \frac{1}{n}\sum_i x_i y_i - \overline{xy} = \frac{1}{n}\sum_i (x_i - \overline{x})(y_i - \overline{y})$ et $\rho(x, y) = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}$ la covariance et la corrélation empirique de x et y. $X = (\mathbf{1}, \mathbf{x})$ est de rang 2 si deux x_i sont différents. On a alors :

$${}^tXX = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix},\ ({}^tXX)^{-1} = \frac{1}{n^2 Var(x)} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}$$

$$\widehat{b} = \frac{Cov(x, y)}{Var(x)},\ \widehat{a} = \overline{y} - \widehat{b}\overline{x},\ \text{ et } \widehat{\sigma}^2 = Var(y)\{1 - \rho(x, y)^2\}$$

La droite de régression estimée de y en x , est :

$$\widehat{y}_i = \overline{y} + \frac{Cov(x, y)}{Var(x)}(x_i - \overline{x}) = \overline{y} + \frac{\rho(x, y) \times \sigma(y)}{\sigma(x)}(x_i - \overline{x})$$

La somme des carrés résiduels associée vaut

$$SCR = nVar(y)\{1 - \rho^2(x, y)\}$$

Cette somme est nulle si les n-points (x_i, y_i) se situent sur une même droite : on retrouve là le fait que la corrélation entre deux variables X et Y vaut ± 1 si il existe une liaison affine presque sûre entre les deux variables, $+1$ si cette liaison est positive, -1 sinon.

La variance résiduelle de la régression est estimée par $\frac{SCR}{n-2}$.

Puisque $Var(\widehat{b}) = \sigma^2\{nVar(x)\}^{-1}$, à n fixé, la précision sur b est d'autant meilleure que les x sont dispersés. Les deux estimateurs $\widehat{a}$ et $\widehat{b}$ sont non-corrélés si $\overline{x} = 0$.

Pour le modèle linéaire $E(y_i) = bx_i, i = 1, n$, $\widehat{b} = \sum x_i y_i / \sum x_i^2$ et $Var(\widehat{b}) = \sigma^2\{\sum x_i^2\}^{-1}$. La précision s'améliore si la dispersion de x autour de 0 augmente.

1.3 Le modèle linéaire gaussien

Les seules hypothèses utilisées jusqu'à maintenant portent sur les espérances, les variances et les covariances des variables (y_i). Pour aller plus loin dans l'étude statistique, par exemple tester une sous-hypothèse, valider un modèle, construire un intervalle de confiance sur un paramètre, il faut ajouter une hypothèse qui précise la loi des y_i. Les modèles

gaussiens permettent de répondre à ces questions : un modèle linéaire est gaussien si pour $i = 1, n$:

$$y_i = {}^t x_i\beta + \varepsilon_i,\ i = 1, n \text{ où les } \varepsilon_i \text{ sont } i.i.d.\ \mathcal{N}(0, \sigma^2)$$

Les n observations sont gaussiennes et indépendantes[3]. Vectoriellement :

$$Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n) \tag{1.5}$$

où $\mathcal{N}_n(\mu, \Sigma)$ est la loi gaussienne sur $\mathbf{R}^n$ de moyenne $\mu \in \mathbf{R}^n$ et de covariance Σ, une matrice $n \times n$ (cf. § 14.2). (1.5) résume les trois propriétés suivantes :

(i) l'espérance de Y est $X\beta$;
(ii) la variance de Y est $\sigma^2 I_n$;
(iii) Y est une variable aléatoire vectorielle gaussienne.

1.3.1 Lois des estimateurs $(\widehat{\beta}, \widehat{\sigma}^2)$

Proposition 2 *Propriétés de l'estimateur des MCO pour le modèle linéaire gaussien (1.5).*

(a) $\widehat{\beta}$, l'estimateur des MCO, est aussi l'estimateur du maximum de vraisemblance (EMV).

(b) Parmi les estimateurs sans biais, l'EMCO est l'estimateur de moindre variance (en anglais, BUE pour Best Unbiased Estimator).

(c) $\widehat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2({}^t XX)^{-1})$ et $\widehat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$; $\widehat{\beta}$ et $\widehat{\sigma}^2$ sont indépendants.

Preuve :

(a) Les observations $\{y_i, i = 1, n\}$ étant gaussiennes indépendantes, la log-vraisemblance vaut :

$$l_n(\beta, \sigma^2) = -\frac{n}{2}\{\log 2\pi + \log \sigma^2\} - \frac{1}{2\sigma^2}\sum_1^n (y_i - {}^t x_i\beta)^2$$

A σ fixé, l'estimateur des MCO maximise la vraisemblance : c'est l'EMV.

Posons $SCR = SCR(\widehat{\beta})$. Un calcul direct de la dérivée en σ^2 montre que l'EMV de σ^2 est $\widehat{\sigma}^2_{MV} = \frac{1}{n}SCR$, d'espérance $E(\widehat{\sigma}^2_{MV}) = \frac{n-p}{n}\sigma^2$. Cet estimateur est biaisé mais converge en espérance vers σ^2 si $n \longrightarrow \infty$: $\widehat{\sigma}^2_{MV}$ est asymptotiquement non-biaisé.

(b) est une conséquence directe de l'inégalité de Cramer-Rao (cf. Annexes), la matrice d'information de Fischer valant :

$$E(-\frac{\partial^2 l_n(\beta, \sigma^2)}{\partial^2 \beta}) = \frac{1}{\sigma^2}\sum_1^n {}^t x_i x_i = \frac{1}{\sigma^2}\,{}^t XX$$

(c) $\widehat{\beta} = AY$ étant une transformée linéaire d'une variable vectorielle gaussienne, $\widehat{\beta}$ est gaussienne (cf. § 14.2). Sa moyenne et sa variance ont été identifiées. On a donc le résultat annoncé pour la loi de $\widehat{\beta}$. D'autre part $(n-p)\widehat{\sigma}^2 = \sum_{p+1}^n e_j^2$. Puisque $e = Q\varepsilon$ est une variable gaussienne centrée de covariance

$$\sigma^2 \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{n-p} \end{pmatrix}$$

les variables (e_j) sont normales centrées et indépendantes, de variance σ^2. La variable $(n-p)\widehat{\sigma}^2$ suit donc une loi $\sigma^2\chi^2_{n-p}$ (pour la définition de la loi du χ^2 à p degré de liberté (ddl), cf. § 14.3).

Reste à établir l'indépendance de $\widehat{\beta}$ et $\widehat{\sigma}^2$. On a le résultat plus fort suivant : $e = Q\varepsilon$ et $\widehat{\beta}$ sont indépendants. $(e, \widehat{\beta})$ étant un vecteur gaussien, il suffit de vérifier que $Cov(e, \widehat{\beta}) = 0$. Puisque e est centrée et que $QX = 0$, on a :

$$Cov(e, \widehat{\beta}) = E[Q\varepsilon\ {}^t\{({}^t XX)^{-1}\ {}^t X\varepsilon\}] = \sigma^2 QX({}^t XX)^{-1} = 0$$

$\widehat{\sigma}^2$ étant une fonction de e qui est indépendant de $\widehat{\beta}$, $\widehat{\sigma}^2$ et $\widehat{\beta}$ sont indépendants. $\ldots\ldots\square$

[3] Pour un vecteur gaussien, la non-corrélation des coordonnées équivaut à l'indépendance.

1.3.2 Le test de Student

Test sur une coordonnée de β

On veut tester, pour a un réel connu : $(H_0) : \beta_k = a$ contre $(H_1) : \beta_k \neq a$.

La variance de $\widehat{\beta}_k$ valant $\sigma^2(\widehat{\beta}_k) = \sigma^2\{({}^tXX)^{-1}\}_{kk}$, $N = \frac{\widehat{\beta}_k - a}{\sigma(\widehat{\beta}_k)}$ suit sous (H_0) une loi normale réduite, loi notée $\mathcal{N}(0,1)$. La variance σ^2 étant inconnue, on la remplace par son estimation $\widehat{\sigma}^2$.

Proposition 3 *Sous (H_0), $T = \frac{\widehat{\beta}_k - a}{\widehat{\sigma}(\widehat{\beta}_k)}$ suit une loi de Student à $(n-p)$ d.d.l.*

Preuve :
$T = \frac{\widehat{\beta}_k - a}{\sigma(\widehat{\beta}_k)} / \frac{\widehat{\sigma}(\widehat{\beta}_k)}{\sigma(\widehat{\beta}_k)}$. Puisque $\widehat{\sigma}^2$ est indépendant de $\widehat{\beta}$, T est, sous (H_0), le quotient d'une variable $\mathcal{N}(0,1)$ et d'une variable $\sqrt{\frac{\chi^2_{n-p}}{n-p}}$, les deux variables étant indépendantes. T suit, sous (H_0), une loi de Student à $(n-p)$ degré de liberté (pour la définition de la loi de Student, voir le § 14.3). ..... □

La région de rejet de (H_0) pour l'alternative *bilatérale* (H_1) au niveau α est :

$$\mathcal{R}_\alpha = \{|T| \geq t(n-p, \frac{1}{2}\alpha)\}$$

$t(q,\alpha)$ est le quantile de la loi de Student T_q à q ddl défini par $P(T_q \geq t(q,\alpha)) = \alpha$. Si l'alternative est *unilatérale* $\{\beta_k < a\}$, la région de rejet est $\mathcal{R}_\alpha = \{T \leq -t(n-p,\alpha)\}$. L'intervalle de confiance bilatéral pour β_k au niveau $(1-\alpha)$ est :

$$[\widehat{\beta}_k - \widehat{\sigma}\sqrt{[({}^tXX)^{-1}]_{kk}} \times t(n-p, \tfrac{1}{2}\alpha), \widehat{\beta}_k + \widehat{\sigma}\sqrt{[({}^tXX)^{-1}]_{kk}} \times t(n-p, \tfrac{1}{2}\alpha)]$$

Test sur une combinaison linéaire de β

Soient $b \in \mathbf{R}^p$, $b \neq 0$, et $a \in \mathbf{R}$ donnés. Pour tester :

$$(H_0) : {}^tb\beta = a \text{ contre } (H_1) : {}^tb\beta \neq a$$

on utilise la statistique T suivante qui suit une loi de Student à $(n-p)$ ddl sous (H_0) :

$$T = \frac{{}^tb\widehat{\beta} - a}{\widehat{\sigma}\sqrt{{}^tb({}^tXX)^{-1}b}}$$

Test sur la variance résiduelle σ^2

Pour $\sigma_0 > 0$ une valeur connue, $S^2 = (n-p)\frac{\widehat{\sigma}^2}{\sigma_0^2}$ suit une loi du χ^2_{n-p}. La région de rejet du test de $(H_0) : \sigma = \sigma_0$ contre $(H_1) : \sigma \neq \sigma_0$ est donc :

$$\mathcal{R}_\alpha = \{S^2 \geq q(n-p, \frac{\alpha}{2})\} \cup \{S^2 \leq q(n-p, 1-\frac{\alpha}{2})\}$$

Le quantile $q(m,\alpha)$ est caractérisé par $P(\chi^2_m > q(m,\alpha)) = \alpha$. Pour l'alternative unilatérale $(H_1') : \sigma > \sigma_0$, la région de rejet est $\{S^2 \geq q(n-p,\alpha)\}$. L'intervalle de confiance bilatéral pour σ^2 au niveau $1-\alpha$ est :

$$[\frac{n-p}{q(n-p, \frac{\alpha}{2})}\widehat{\sigma}^2, \frac{n-p}{q(n-p, 1-\frac{\alpha}{2})}\widehat{\sigma}^2]$$

1.3.3 Le test de Fisher

Sous-modèle linéaire.
Le test de Fisher est le test central pour l'étude des modèles linéaires gaussiens. Il généralise le test de Student au cas d'une sous-hypothèse linéaire générale. Considérons deux modèles linéaires gaussiens pour l'observation Y,

$$\begin{array}{lll}(\Omega) & : & Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n),\ \beta \in \mathbf{R}^p \text{ et} \\ (\omega) & : & Y \sim \mathcal{N}_n(Z\gamma, \sigma^2 I_n),\ \gamma \in \mathbf{R}^q\end{array}$$

On dit que (ω) est un *sous-modèle linéaire* de (Ω) si son espace de la moyenne $\mathcal{E}_Z$ est un sous-espace vectoriel de $\mathcal{E}_X$. Deux situations conduisent à la définition d'un sous-modèle :

• on spécifie $\beta = H\gamma$ à partir d'un paramètre γ de plus petite dimension q, H étant une matrice $p \times q$ connue de rang q (Z vaut alors XH). Par exemple, dans une analyse de la variance où la condition i est repérée par une condition x_i réelle, la moyenne μ_i peut être spécifiée de façon affine par une régression (ω) :

$$(\Omega) : E(y_{ij}) = \mu_i,\ j = 1, n_i; i = 1, p\,;\ \text{ et } (\omega) : \mu_i = a + bx_i, i = 1, p$$

Les paramètres sont $\beta = {}^t(\mu_1, \mu_2, \cdots, \mu_n)$, $\gamma = {}^t(a, b)$, $q = 2$, et $H = (\mathbf{1}, \mathbf{x})$.

•• on impose $r = p - q$ contraintes linéaires sur β, ces contraintes étant linéairement indépendantes. Par exemple $(\omega) : \beta_1 = \beta_2$ et $\beta_1 + 2\beta_3 = 0$ $(q = 2)$.

La statistique de Fisher.
La construction de la statistique de Fisher permettant de tester (ω) dans (Ω) est naturelle : on estime Y sous l'un et l'autre modèle, $\widehat{Y}_\Omega = X\widehat{\beta}_\Omega$ et $\widehat{Y}_\omega = X\widehat{\beta}_\omega$; puis on évalue l'écart $\left\|\widehat{Y}_\Omega - \widehat{Y}_\omega\right\|^2$. Si, relativement à $\widehat{\sigma}^2_\Omega$, cet écart est grand, on rejette (ω). La statistique de Fisher est :

$$F = \frac{\frac{1}{p-q}\left\|\widehat{Y}_\Omega - \widehat{Y}_\omega\right\|^2}{\widehat{\sigma}^2_\Omega}$$

Proposition 4 *La statistique F vaut encore :*

$$F = \frac{(SCR(\omega) - SCR(\Omega))/(p-q)}{SCR(\Omega)/(n-p)}$$

Sous (ω), F suit une loi de Fisher à $(p-q)$ et $(n-p)$ ddl (pour la définition de la loi de Fisher, cf. § 14.3). La région de rejet de (ω) au niveau α est :

$$R_\alpha = \{F \geq f(p-q, n-p; \alpha)\}$$

où $f(r, s; \alpha)$ est le α-quantile de la loi de Fisher, $P(F_{r,s} \geq f(r, s; \alpha)) = \alpha$.

Preuve :
L'expression de F découle des identités $\left\|\widehat{Y}_\Omega - \widehat{Y}_\omega\right\|^2 = SCR(\omega) - SCR(\Omega)$ et $\widehat{\sigma}^2_\Omega = SCR(\Omega)/(n-p)$. Pour identifier la loi de la statistique, commençons par démontrer le résultat suivant :

Lemma 5 *Soient $Z \sim \mathcal{N}_m(\mu, \sigma^2 I_m)$ une variable gaussienne de $\mathbf{R}^m$, $\mathcal{E}$ un sous-espace vectoriel de $\mathbf{R}^m$ de dimension p, $0 < p < m$, P et $(I - P)$ les projections orthogonales sur $\mathcal{E}$ et $\mathcal{E}^\perp$, et $Z = PZ + (I-P)Z$. Alors :*

(a) PZ et $(I-P)Z$ sont deux vecteurs gaussiens indépendants, $PZ \sim \mathcal{N}_p(\mu_1, \sigma^2 I_p)$ avec $\mu_1 = P\mu$, et $(I-P)Z \sim \mathcal{N}_{m-p}(\mu_2, \sigma^2 I_{m-p})$ avec $\mu_2 = (I-P)\mu$.

(b) $\|PZ\|^2$ et $\|(I-P)Z\|^2$ sont des $\sigma^2\chi'^2$ décentrés (cf. Annexes) indépendants respectivement à p et $m-p$ ddl.

Preuve du lemme :
D'une part la transformée linéaire d'une gaussienne est une gaussienne ; d'autre part, puisque $Cov(PZ,(I-P)Z) = \sigma^2 P\,{}^t(I-P) = \sigma^2 P(I-P) = 0$, les deux vecteurs gaussiens PZ et $(I-P)Z$ sont indépendants. ..... □

Suite de la preuve : on applique le lemme à $Z = (I-P_\omega)Y$ et à $P = (P_\Omega - P_\omega)$, obtenant la décomposition $(I-P_\omega)Y = (I-P_\Omega)Y + (P_\Omega - P_\omega)Y$ dans $\mathbf{R}^m$ où $m = n-q$. Il suffit alors de constater que :

(i) $SCR(\Omega) = \|(I-P_\Omega)Y\|^2$ est un $\sigma^2\chi^2(n-p)$ centré (en effet $\mu_1 = 0$).

(ii) $SCR(\omega) - SCR(\Omega) = \|(P_\Omega - P_\omega)Y\|^2$ est une $\sigma^2\chi^2(p-q)$ centré ($E(\widehat{Y}_\Omega) = E(Y)$ sous (ω)) indépendant de $SCR(\Omega)$. Sous l'alternative, ce χ^2 est décentré, de paramètre de non centralité $\lambda^2(\beta) = \frac{1}{\sigma^2}\|X\beta - P_\omega X\beta\|^2$. ..... □

Commentaires sur le test de Fisher.

(1) Le nombre de ddl du numérateur $(p-q)$ correspond à la "chute" de dimension entre (Ω) et (ω). $(n-p)$ est le nombre de ddl résiduels pour estimer la variance dans (Ω).

(2) Si $q = p-1$, (ω) peut s'écrire à partir d'une contrainte linéaire ${}^t b\beta = 0$. La statistique de Fisher $F_{1,n-p}$ n'est autre que le carré de la statistique de Student, $F = T^2_{n-p}$.

(3) *Sécurité dans la décision.* Si on rejette (ω), c'est toujours avec sécurité, le risque d'erreur (ou risque de première espèce du test, ou probabilité de rejeter à tort (ω)) valant α. *Au contraire*, si on retient (ω), il faudra examiner la puissance du test pour prendre la décision avec sécurité. La fonction puissance du test est définie par $P(\beta) = P_\beta(\mathcal{R}_\alpha)$, $\beta \in (\Omega)$. Pour le test de Fisher, la puissance s'obtient à partir de la distribution d'une loi de Fisher décentrée $F'(p-q, n-p; \lambda^2(\beta))$ (pour la définition, cf. § 14.3) de paramètre de non-centralité $\lambda^2(\beta) = \frac{1}{\sigma^2}\|X\beta - P_\omega(X\beta)\|^2$ ($\lambda^2(\beta) = 0$ sous (ω)). La puissance, définie sur l'ensemble des alternatives $\beta \in \Omega\backslash\omega$, vaut :

$$\Pr(\beta) = P\{F(p-q, n-p; \lambda^2(\beta)) \geq f(p-q, n-p; \alpha)\}$$

Deux types de table (Hartley-Pearson et Fox) permettent d'évaluer $P(\beta)$. Ces tables se trouvent dans les ouvrages spécialisés.

(4) *Exemple de calcul de paramètre de non centralité*

(i) Analyse de la variance à 1-facteur :

$(\Omega) : y_{ij} = m_i + \varepsilon_{ij}$, $j = 1, n_i$t, $i = 1, p$ et $(\omega) : m_i = a + bx_i$ pour $i = 1, p$ et x_i une variable de controle réel. Un calcul simple donne $\lambda^2(m) = \frac{1}{\sigma^2}\sum n_i(m_i - \overline{m})^2$ où $\overline{m}$ est la moyenne pondérée des (m_i).

(ii) Regression affine : notons (ω) la sous hypothèse $b = 0$. Alors $\lambda^2(a,b) = \frac{1}{\sigma^2}\sum(x_i - \overline{x})^2$.

1.3.4 Test de Wald, ellipsoide de confiance sur β

Test de Wald

Soient p et q deux entiers, $0 < q < p$, C une matrice $(p-q)\times p$ de rang $(p-q)$ et $c \in \mathbf{R}^{p-q}$ des quantités connues. On veut tester :

$$(\omega) : C\beta = c \text{ contre } (\Omega) : C\beta \neq c$$

ESous (ω), $(C\widehat{\beta} - c) \sim \mathcal{N}_{p-q}(0, \sigma^2 C({}^tXX)^{-1}\,{}^tC)$. Utilisant le lemme donné ci-dessous, on a :

$$W = \frac{{}^t(C\widehat{\beta} - c)\{C({}^tXX)^{-1}\,{}^tC)\}^{-1}\,(C\widehat{\beta} - c)}{(p-q)\,\widehat{\sigma}^2} \underset{(\omega)}{\sim} F(p-q, n-p)$$

W est la *statistique de Wald*. La statistique de Wald présente deux avantages en comparaison de la statistique de Fischer : elle fait intervenir uniquement l'estimation dans (Ω) et elle est valable pour une sous-hypothèse affine $c \neq 0$.

Lemma 6 *Si $Z \sim \mathcal{N}_m(0, V)$ et si V est inversible, alors ${}^tZV^{-1}Z \sim \chi^2_m$.*

Preuve :
V^{-1} comme V est une symétrique, dp : il existe donc A de même taille telle que $V^{-1} = {}^tAA$ ([4]). Donc, ${}^tZV^{-1}Z = \|AZ\|^2$. Le résultat est alors une conséquence du fait que $AZ \sim \mathcal{N}_m(0, I)$ puisque, en effet, $Var(AZ) = AV\,{}^tA = I_m$.....□

Ellipsoide de confiance sur β

Puisque $(\widehat{\beta}-\beta) \sim \mathcal{N}_p(0, \sigma^2({}^tXX)^{-1})$, le lemme précédent montre que $\frac{{}^t(\widehat{\beta}-\beta){}^tXX(\widehat{\beta}-\beta)}{\sigma^2} \sim \chi^2_p$ sous (Ω). Ainsi, un ellipsoide de confiance sur β est

$$E_{1-\alpha}(\beta) = \{\left\|X(\widehat{\beta}-\beta)\right\|^2 \leq \sigma^2 q(n-p;\alpha)\}$$

1.3.5 Exemples d'utilisation du test de Fisher

Example 8 *Test de rupture de modèle*

Considérons une série temporelle modélisée sur deux périodes par :

$$(\Omega) : \left\{ \begin{array}{l} \textit{Période 1 : } Y_1 = \mathcal{N}_n(X_1\beta(1), \sigma_1^2 I_n) \text{ avec } \beta(1) \in \mathbf{R}^p,\ n \text{ observations} \\ \textit{Période 2 : } Y_2 = \mathcal{N}_m(X_2\beta(2), \sigma_2^2 I_m) \text{ avec } \beta(2) \in \mathbf{R}^p,\ m \text{ observations} \end{array} \right.$$

les deux séries étant indépendantes. Notant $Y = {}^t({}^tY_1, {}^tY_2)$, le modèle s'écrit :

$$E(Y) = \begin{pmatrix} X_1 & \mathbf{0} \\ \mathbf{0} & X_2 \end{pmatrix} \begin{pmatrix} \beta(1) \\ \beta(2) \end{pmatrix}$$

La sous-hypothèse d'absence de rupture entre les deux périodes se traduit par :

$$(\omega) : \sigma_1^2 = \sigma_2^2 \text{ et } \beta(1) = \beta(2)$$

Le test de (ω) se fait en deux temps :

• On teste d'abord l'égalité des variances $\sigma_1^2 = \sigma_2^2$ en utilisant la statistique :

$$G = \frac{SCR_\Omega(1)/n-p}{SCR_\Omega(2)/m-p} \underset{(\omega)}{\sim} F(n-p, m-p)$$

en spécifiant la région de rejet selon que l'alternative est unilatérale ou bilatérale.

•• Si on retient $\sigma_1^2 = \sigma_2^2$, la variance commune σ^2 est estimée par $\frac{1}{n+m-2p}SCR_\Omega$, où $SCR_\Omega = SCR_\Omega(1) + SCR_\Omega(2)$. On teste alors $\beta(1) = \beta(2)$. Cette égalité définit un sous-modèle linéaire (ω) de dimension p. La statistique de Fisher vaut :

$$F = \frac{\frac{1}{p}\{SCR(\omega) - SCR(\Omega)\}}{\frac{1}{n+m-2p}SCR(\Omega)} \underset{(\omega)}{\sim} F(p, n+m-2p)$$

Example 9 *Non-significativité des exogènes et coefficient de corrélation multiple R^2_Ω*

[4] Une matrice M symétrique réelle étant diagonalisable sur $\mathbf{R}$ admet la décomposition spectrale $M = {}^tPDP$: P, la matrice dont les colonnes sont les vecteurs propres, peut être choisie orthogonale ; D est la matrice diagonale des valeurs propres. Si M est sdp, les valeurs propres sont ≥ 0. Il suffit alors de prendre $A = D^{\frac{1}{2}}P$ pour obtenir $M = {}^tAA$.

Considérons le modèle de régression multiple (1.3) :

$$(\Omega) : E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},\ i = 1, n$$

Les variables exogènes x sont non-significatives si pour tout i, $E(y_i) = \beta_0$:

$$(\omega_0) : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

La statistique du test de (ω_0) est :

$$F = \frac{\{SCR(\omega_0) - SCR(\Omega)\}/p}{SCR(\Omega)/(n-p)} \underset{(\omega_0)}{\sim} F(p, n-p)$$

Dans (ω_0), $\widehat{\beta}_0 = \overline{y}$ et $\widehat{Y}_{\omega_0} = \overline{y}\mathbf{1}$. Définissons :

$$SCT = \left\|Y - \widehat{Y}_{\omega_0}\right\|^2,\ SCE(\Omega) = \left\|\widehat{Y}_\Omega - \widehat{Y}_{\omega_0}\right\|^2 \text{ et } R^2_\Omega = \frac{\left\|\widehat{Y}_\Omega - \widehat{Y}_{\omega_0}\right\|^2}{\left\|Y - \widehat{Y}_{\omega_0}\right\|^2} \doteq \frac{SCE(\Omega)}{SCT}$$

SCT est la somme des carrés totale, $SCE(\Omega)$ la somme des carrés expliquée par (Ω). R^2_Ω est toujours ≤ 1, d'autant plus proche de 1 que (Ω) explique bien y. Quelque soit le modèle $(\omega_0) \subseteq (\omega) \subseteq (\Omega)$, on a toujours : $SCT = SCE(\omega) + SCR(\omega)$.

Proposition 7 *Coefficient de corrélation multiple et test F*

(1) R^2_Ω est appelé le coefficient de corrélation multiple entre Y et $(X_1, X_2, \cdots, X_p)$. C'est le carré de la corrélation entre Y et $\widehat{Y}_\Omega$. R^2_Ω est une mesure de la qualité de l'ajustement de Y par (Ω).

(2) La statistique F du test de (ω_0) dans (Ω) s'écrit $F = \frac{n-(p+1)}{p} \times \frac{R^2_\Omega}{1-R^2_\Omega}$.

Preuve :

(1) $(\widehat{Y}_\Omega - \widehat{Y}_{\omega_0})$ étant la projection orthogonale de $(Y - \widehat{Y}_{\omega_0})$ sur $\mathcal{E}_X$, on a :

$$R^2_\Omega = \frac{\left\langle \widehat{Y}_\Omega - \widehat{Y}_{\omega_0}, Y - \widehat{Y}_{\omega_0} \right\rangle}{\left\|Y - \widehat{Y}_{\omega_0}\right\|^2} = \frac{\left\langle \widehat{Y}_\Omega - \widehat{Y}_{\omega_0}, Y - \widehat{Y}_{\omega_0} \right\rangle^2}{\left\|Y - \widehat{Y}_{\omega_0}\right\|^2 \times \left\|\widehat{Y}_\Omega - \widehat{Y}_{\omega_0}\right\|^2} = \rho^2(Y, \widehat{Y}_\Omega)$$

(2) Le calcul de F à partir de R^2_Ω résulte d'un calcul direct. ..... □

Plus généralement, la statistique de Fisher du test de (ω) dans (Ω) s'explicite à partir des deux coefficients de corrélation multiples R^2_ω et R^2_Ω :

$$F = \frac{n - \dim(\Omega)}{\dim(\Omega) - \dim(\omega)} \times \frac{R^2_\Omega - R^2_\omega}{1 - R^2_\Omega}$$

On retrouve la forme particulière du test de (ω_0) dans (Ω) puisque $R^2_{\omega_0} = 0$.

Example 10 *Test de non-corrélation pour un couple gaussien*

La remarque précédente permet de construire un test de non-corrélation pour un couple gaussien. La question est la suivante : soit $Z = (X, Y)$ un couple gaussien et $\{z_i, i = 1, n\}$ un n-échantillon de Z. Comment tester que $\rho(X, Y) = 0$? Commençons par rappeler le résultat important suivant concernant la loi conditionnelle pour un couple gaussien (cf. § 14.2) : soit $Z = (X, Y)$ un couple gaussien,

$$Z \sim \mathcal{N}_2\{(m_1, m_2); \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}\},\ \sigma_1 \text{ et } \sigma_2 > 0 \text{ et } |\rho| < 1$$

La loi de Y conditionnelle à $(X = x)$ est : $\mathcal{L}(Y \mid X = x) \sim \mathcal{N}(a + bx, \sigma^2)$:

- l'espérance conditionnelle est affine en x[5] ;
- la variance conditionnelle est indépendante de x : $\sigma^2 = \sigma_2^2(1 - \rho^2)$;
- la loi conditionnelle est gaussienne.

Y conditionnelle à X suit donc le modèle linéaire$\widehat{\rho}$ gaussien $(\Omega) : Y = a + bX + \varepsilon$. Dans ce modèle, $\rho = 0$ équivaut à $(\omega_0) : b = 0$. On peut donc tester $\rho = 0$ en utilisant la statistique de Fisher F du test de $(\omega_0) : b = 0$,

$$F = (n-2)\frac{\widehat{\rho}^2}{1-\widehat{\rho}^2} \underset{\rho=0}{\sim} F(1, n-2)$$

$\widehat{\rho}^2 = \frac{\{\sum_i (x_i - \overline{x})(y_i - \overline{y})\}^2}{\sum_i (x_i - \overline{x})^2 \sum_i (y_i - \overline{y})^2}$ est la corrélation empirique entre les (y_i) et les (x_i). Le test de non-corrélation pour un couple gaussien repose donc sur une loi de Fischer $F(1, n-2)$, ou encore, en prenant sa racine carrée, sur une loi de Student T_{n-2} :

$$T = \sqrt{n-2}\frac{\widehat{\rho}}{\sqrt{1-\widehat{\rho}^2}}$$

Les 2 ddl perdus $(n-2)$ s'expliquent par l'estimation des deux paramètres auxiliaires a et b.

Remarque : si $\rho \neq 0$, on obtient un intervalle de confiance asymptotique (n grand) sur ρ de la façon suivante. La distribution de $\widehat{\rho}$ étant fortement dissymétrique (en effet $\widehat{\rho}$ est contraint à appartenir à $[-1, +1]$), et $Var(\widehat{\rho}) \cong \frac{(1-\rho^2)^2}{n}$ dépendant de ρ, on considère la transformation

$$z = \frac{1}{2}\log\frac{1+\widehat{\rho}}{1-\widehat{\rho}}$$

On alors, $z \sim \mathcal{N}(\frac{1}{2}\log\frac{1+\rho}{1-\rho} + \frac{\rho}{2(n-1)}, \frac{1}{n-3})$, l'approximation étant bonne à partir de $n = 20$. On en déduit alors un intervalle de confiance sur ρ en considérant l'application inverse permettant de remonter de z à $\widehat{\rho}$.

1.3.6 Prédiction dans un modèle linéaire gaussien

Supposons que $(\Omega) : Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n), \beta \in \mathbf{R}^p$ est estimé sur la base de n observations $\{(x_i, y_i), i = 1, n\}$, d'EMCO $\widehat{\beta}$ et $\widehat{\sigma}^2$. Soit $x = {}^t(x_1, x_2, \cdots, x_p)$ une nouvelle condition exogène : comment prédire au mieux $E(y_x) = {}^t\beta x$, y_x étant la réponse associée à la condition x ? On supposera que y_x est indépendante des $\{y_i, i = 1, n\}$.

Proposition 8 *La prédiction sans biais et de variance minimum de $E(y_x) = {}^t\beta x$ est ${}^t\widehat{\beta}x$. L'erreur de prédiction est $\sigma^2\{\,{}^tx({}^tXX)^{-1}x\}$. L'intervalle de confiance symétrique pour ${}^t\beta x$ au niveau $(1-\alpha)$ est $[{}^t\widehat{\beta}x \pm \widehat{\sigma}\sqrt{\{\,{}^tx({}^tXX)^{-1}x\}} \times t(n-p, \frac{\alpha}{2})]$.*

Preuve :
${}^t\widehat{\beta}x$ estime sans biais ${}^t\beta x$ et $Var({}^t\widehat{\beta}x) = {}^txVar(\widehat{\beta})x = \sigma^2\,{}^tx({}^tXX)^{-1}x$. Soit ${}^tcY$ une autre prédiction linéaire. Elle est sans biais si ${}^tcX = {}^tx$. Soit $P = X({}^tXX)^{-1}\,{}^tX$ le projecteur orthogonal de $\mathbf{R}^n$ sur $\mathcal{E}_X$. Au sens des matrices sdp, $P \preceq I_n$: en effet, $\forall u \in \mathbf{R}^n$, ${}^tu(I_n - P)u = \|(I_n - P)u\|^2 \geq 0$. L'optimalité de ${}^t\widehat{\beta}x$ résulte de la minoration :

$$Var({}^tcY) = \sigma^2\,{}^tcc = \sigma^2\,{}^tcI_nc \geq \sigma^2\,{}^tcX({}^tXX)^{-1}\,{}^tXc = Var({}^t\widehat{\beta}x)$$

L'erreur de prédiction et l'intervalle de confiance sur ${}^t\beta x$ s'obtiennent alors à partir de la loi de ${}^t\widehat{\beta}x$. ..... □

Remarquons que si on doit propser une prédiction pour y sachant x, on proposera encore ${}^t\widehat{\beta}x = \widehat{y}$. Mais, cette fois ci, il faut ajouter σ^2 à l'erreur de prédiction puisque en effet $y - \widehat{y} = {}^t(\beta - \widehat{\beta}) + \varepsilon$ où ε est indépendant de $\widehat{\beta}$.

[5] C'est la droite des moindres carrés : $E(Y \mid x) = a + bx = E(Y) - b(x - E(X))$ avec $b = \rho\frac{\sigma_2}{\sigma_1} = \frac{Cov(x,y)}{Var(x)}$.

1.4 Exercices sur le modèle linéaire

Sauf mention du contraire, les résidus des modèles considérés sont i.i.d. gaussiens. Des exercices et données associées viennent du livre de Bernard Prum : *Modèle linéaire : comparaison de groupes et régression,* Les éditions de l'INSERM (1996)

Exercise 1 *Croissance d'une colonies de bactéries*

On mesure la taille de colonies bactériennes sur des boîtes de Petri x jour après l'ensemencement. On a de bonnes raisons de penser que y, le logarithme du rayon de ces colonies, croît linéairement avec x. Le tableau suivant donne les mesures faites :

x	y	x	y
1	6.68	11	17.01
1	8.94	12	15.51
2	5.91	12	16.00
2	13.22	12	12.39
2	9.90	13	14.48
3	6.34	15	14.62
3	4.51	16	13.58
4	8.72	18	10.74
8	13.63	20	20.89
8	12.34	20	16.92

On en déduit : $\sum x = 183$, $\sum x^2 = 2503$, $\sum y = 242.51$, $\sum y^2 = 3296.91$ et $\sum xy = 2639.82$.

(1) Estimer la droite de régression affine : $y = ax + b$ ainsi que la variance résiduelle. Quelle est la corrélation entre x et y. En donner un intervalle de confiance.

(2) Quelles sont les variances des estimations de a et b. Donner les intervalles de confiance pour a et b ainsi que l'ellipsoide de confiance pour (a, b).

(3) Avec quelle précision prédira-t-on : (a) $E(y)$ pour x donné ; (b) y pour x donné. Représentez graphiquement ces deux zones de confiance en fonction de $x > 0$.

(4) Représenter graphiquement les données $\{(x_i, y_i), i = 1, 20\}$. Estimer "graphiquement" la droite de regression. Faites vos remarques sur le modèle retenu et les estimations obtenues. Utilisant le résultat de (3)-b, tester que la $4^{ème}$ données $(x, y) = (2, 13.22)$ est aberrante.

(5) **T.P. :**Retrouver ces résultats numériques en utilisant **R** (estimations, ecarts types, SCR...)

Exercise 2 *Cycle biologique circadien (cycle sur 24 heures)*

L'activité de la faune microbienne d'un marais d'Afrique équatoriale dépend de T l'heure de la journée, en particulier à cause des variations de lumière et de température. Trois journées de suite, on mesure à chaque heure cette activitée Y en quantité de méthane dégagée. Le tableau suivant donne les résultats (en cm^3 de méthane),

Heure	J1	J2	J3
1	156.4	139.8	148.0
2	53.8	103.8	99.5
3	51.6	152.8	123.2
4	97.8	150.5	72.5
5	105.7	113.1	95.2
6	97.7	110.5	84.4
7	69.9	81.1	85.4
8	74.1	52.8	122.3
9	103.9	89.9	45.4
10	137.1	106.6	112.6
11	141.1	131.6	80.0
12	115.7	75.4	118.1

Heure	J1	J2	J3
13	82.1	119.0	144.0
14	208.8	124.8	168.9
15	215.1	148.0	209.2
16	207.7	225.5	150.7
17	167.4	168.9	205.7
18	146.3	146.4	211.6
19	199.8	203.8	230.0
20	143.4	173.9	168.6
21	118.9	160.4	175.5
22	111.2	129.4	119.3
23	143.9	172.9	90.4
24	61.0	68.7	80.1

L'heure 1 correspond au coucher du soleil (19h) et 1 à 12 aux heures de la nuit ; l'heure 13 correspond au lever du soleil (7h) et 13 à 24 aux heures du jour.

(1) On suppose que le cycle circadien, périodique, se traduit par une variation sinusoïdale autour de la moyenne μ :

$$(H_2) : y_T = \mu + a\ \sin(\omega T) + \varepsilon_T,\ T = 1, 72$$

où $\omega = \frac{2\pi}{24}$. Notant $s_T = \sin(\omega T)$, les données donnent : $\sum s = 0$, $\sum s^2 = 36$, $\sum y = 9300.4$, $\sum sy = -1687.07$. Estimer cette régression. On obtient $SCR(H_2) = 77683.53$. Donner un I.C. pour a.

T.P. : déclarer ce modèle sous **R** et retrouver les résultats annoncés (estimùation des coefficient, leurs variances et covariances, SCR....).

(2) Un spécialiste prétend que les cycles nocturne et diurne sont différents et suggère :

$$(H_3) : \begin{cases} y_T = m + b\sin(\omega T) + \varepsilon_T, \text{ les heures la nuit} \\ y_T = m + c\sin(\omega T) + \varepsilon_T, \text{ les heures le jour} \end{cases}$$

On trouve $SCR(H_3) = 68015.79$. Que conclure ?

retrouvez les résultats annoncés en utilisant **R**. Donner les estimations, variances et covariances des 3 paramètres. Représentez graphiquement les données ainsi que les deux ajustements.

(3) On propose de modéliser le cycle périodique par le modèle d'analyse de la variance,

$$(H_{24}) : y_{T,j} = m_T + \varepsilon_{T,j},\ T = 1, 24 \text{ et } j = 1, 3$$

Utilisant **R**, estimer ce modèle et tester la validité de (H_3).

(4) Proposer un modèle permettant de tester qu'il n'y a pas de différence significatives entre les 3 jours. Effectuer ce test.

Exercise 3 *Nombre de lymphocytes T4*

On mesure le nombre de lymphocytes T4 chez des patients sida avéré traités selon trois thérapies, appelées simplement 1, 2 et 3.

(1) On commence par se demander si les espérances m_1, m_2 et m_3 du nombre Y de lymphocytes T4 sont les mêmes ou non pour les trois thérapies On donne :

	thérapie 1	thérapie 2	thérapie 3
n	20	30	25
SY	10 970	15 866	12 616
SY^2	6 697 908	9 550 638	7 130 500

Poser le test (Anova à 1 facteur) et conclure. Estimer la variance résiduelle.

(2) On convient qu'il faut prendre en compte le temps T écoulé depuis l'instant où s'est avéré le sida. Les données sont :

	thérapie 1	thérapie 2	thérapie 3
n	20	30	25
ST	2 080	2 681	2 062
ST^2	390 810	415 793	248 140
SY	10 970	15 866	12 616
SYT	798 350	970 261	799 507
SY^2	6 697 908	9 550 638	7 130 500

On considère d'abord globalement les trois thérapies et on modélise Y par (H_2) : $E(Y) = a + bT$. Comparer au modèle de la première question et commenter. Tester $b = 0$ contre $b < 0$.

(3) La question que l'on se pose en fait est l'influence de la thérapie sur la croissance du taux de lymphocytes. On considère donc le modèle (H_4) :

$$(H_4) : E(Y_{i,T}) = a + b_i T \text{ pour } i = 1, 2, 3$$

Expliquez pourquoi a a été choisi identique pour les trois thérapies. Estimer les 4 paramètres ainsi que la variance résiduelle. Tester $(H_2) : b_1 = b_2 = b_3$ contre (H_4). Pour un patient traité avec la thérapie 1, on a observé $T = 211$ et $Y = 343$. Estimer le résidu relatif à ce patient.

Remarque : ici, on ne dispose pas des données brutes, mais seuleument de statistiques résumant ces données brutes. On ne peut pas utiliser le logiciel **R**.

Exercise 4 *Régression sur deux variables explicatives.*

Le modèle (ω) expliquant y à partir de deux variables x et z est estimé par :

$$(\omega) : \widehat{y}_t = \underset{(121.6)}{1.56} + \underset{(0.32)}{8.68}\, x_t + \underset{(0.24)}{0.74}\, z_t,\ t = 1, 265\,;\ R^2_\omega = 0.839$$

Les écarts types estimés sont entre parenthèses. Les paramètres sont, dans l'ordre, (a_0, a_1, a_2).

(1) Tester $a_2 = 0$ contre $a_2 > 0$. Tester $a_2 = 1$ contre $a_2 \neq 1$.

(2) On donne : $\widehat{Cov}(\widehat{a}_0, \widehat{a}_2) = 20.4$. Tester $a_0 = a_2$ contre $a_0 > a_2$.

(3) Les deux variables x et z sont-telles globalement significatives ?

(4) Deux nouveaux régresseurs s et w sont introduits, conduisant au modèle affine (Ω) en x, z, s, w. On trouve $R^2_\Omega = 0.842$. (Ω) est-il significativement différent de (ω) ?

Exercise 5 *Test de Wald de l'hypothèse : $\beta = \beta(0)$*

Soit le modèle $E(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$, $i = 1, 24$. Tester "$\beta_1 = \beta_2 = 1$ et $\beta_3 = -2$" sachant que la matrice des sommes de produits des variables *recentrées* y, x_1, x_2, x_3 est :

$$S = \begin{pmatrix} 60 & * & * & * \\ 7 & 10 & * & * \\ -7 & 10 & 30 & * \\ -26 & 5 & 15 & 20 \end{pmatrix}$$

Exercise 6 *Changement de structure et effet trimestriel*

L'estimation d'un modèle (ω) a été obtenue pour des données trimestrielles allant de $1958-I$ (1^{er} trimestre de 1958) à $1976-IV$ et 3 variables explicatives x_1, x_2 et x_3 (écarts types estimés entre (.)) :

$$(\omega) : \widehat{y}_t = \underset{(1.2)}{2.2} + \underset{(0.005)}{0.104} x_{1t} - \underset{(0.242)}{3.48}\, x_{2t} + \underset{(0.15)}{0.34} x_{3t}$$

On donne $SCE(\omega) = \left\| \widehat{Y}_\omega - \overline{y}1 \right\|^2 = 109.6$ et $SCR(\omega) = 18.48$.

(1) Calculer R^2_ω et tester la significativité globale de (ω).

(2) $1968 - III$ est un trimestre possible de début de changement de structure. On observe sur chacune des deux périodes (avant et après cette date) : $SCR(\omega_1) = 9.32$ et $SCR(\omega_2) = 7.46$. La variabilité résiduelle est-elle la même sur les deux périodes ? Si oui, tester l'absence de changement de structure.

(3) On suppose qu'il n'y a pas de changement. Soit $(\mathcal{S})$ le sur-modèle de (ω) incluant un effet trimestriel. On observe : $SCE(\mathcal{S}) = 111.2$. La saisonnalité trimestrielle est-elle significative ?

Chapitre 2

L'analyse de la variance

Un modèle linéaire expliquant y est appelé modèle d'analyse de la variance quand la variable explicative x est *qualitative* : $x \in E = \{1, 2, \cdots, p\}$, un ensemble de p classes. Notant encore i la modalité de x, le modèle d'*analyse de la variance à un facteur* s'écrit :

$$Y_{i,k} = m_i + \varepsilon_{ik}, k = 1, n_i; i = 1, p \tag{2.1}$$

ε est un BB de variance σ^2. La modalité i est répétée n_i fois ; $k = 1, n_i$ est l'indice de répétition. Le nombre d'observations est $n = \sum_{i=1}^{p} n_i$ et le paramètre $m = {}^t(m_1, m_2, \cdots, m_p) \in \mathbf{R}^p$. Le modèle est gaussien si les résidus sont gaussiens. L'indice i peut être interprété comme un indice de *population* : par exemple, E est l'ensemble des CSP, $Y_{i,k}$ le salaire de l'individu k de la catégorie i.

La spécificité de la moyenne et des décompositions des sommes de carrés justifie cette étude particulière de l'analyse de la variance. Numériquement, les inversions $({}^tXX)^{-1}$ sont inutiles et les logiciels proposent une procédure spécifique distincte de celle du modèle linéaire général.

Pour $E = I \times J = \{1, 2, \cdots, I\} \times \{1, 2, \cdots, J\}$, et $x = (i, j)$, on parle l'*analyse de la variance à deux facteurs*. Par exemple $i \in \{Pri, Sec, Tech, Sup\}$ est le niveau d'étude et $j \in \{M, F\}$ le sexe de l'individu observé.

L'analyse de la variance à 3 facteurs est associée à $x = (i, j, k)$, $x \in E = I \times J \times K = \{1, 2, \cdots, I\} \times \{1, 2, \cdots, J\} \times \{1, 2, \cdots, K\}$. Et ainsi de suite.

Pour les modèles à 2 facteurs ou plus, on se limitera à l'étude des *dispositifs équilibrés* pour lesquels le nombre de répétitions est le même quelle que soit la modalité. Par exemple, le modèle équilibré à deux facteurs s'écrit :

$$Y_{i,j;k} = m_{ij} + \varepsilon_{i,j;k} \ , \ k = 1, n \ ; \ i = 1, I \text{ et } j = 1, J \tag{2.2}$$

L'étude spécifique des modèles déséquilibrés est plus délicate. Elle peut toujours être effectuée en utilisant les outils habituels du modèle linéaire. Les tests sont donnés pour des résidus i.i.d. et gaussiens.

2.1 L'analyse de la variance à un facteur

$$(\Omega) : Y_{i,k} = m_i + \varepsilon_{ik} \ , \ k = 1, n_i; \ i = 1, p$$

On supposera chaque $n_i > 0$ et $n = \sum_i n_i > p$.

2.1.1 Estimations de m et σ^2

$Y = ((Y_{i,k}, k = 1, n_i); i = 1, p)) \in \mathbf{R}^n$ est le vecteur des observations. Le vecteur des paramètres est $m = {}^t(m_1, m_2, \cdots, m_p) \in \mathbf{R}^p$. Le modèle est identifiable. La somme des

carrés résiduels vaut :

$$SCR(m) = \|Y - E_\Omega(Y)\|^2 = \sum_{i=1}^{p}\sum_{k=1}^{n_i}(Y_{i,k} - m_i)^2$$

Notation : si $(Z_\alpha, \alpha \in \Lambda)$ est une famille finie de réels, on note $Z_\bullet$ leur moyenne arithmétique, $Z_\bullet = \frac{1}{|\Lambda|}(\sum_\alpha Z_\alpha)$ ($|\Lambda|$ est le cardinal de Λ).

On vérifie facilement que l'EMCO de m et celle de σ^2 qui en est déduite sont :

$$\widehat{m}_i = Y_{i\bullet},\ i = 1, p \text{ et } \widehat{\sigma}^2 = \frac{SCR(\Omega)}{n-p} \text{ où } SCR(\Omega) = \sum_{i=1,p}\sum_{k=1,n_i}(Y_{ik} - Y_{i\bullet})^2$$

2.1.2 Test d'identité des p populations

$$(\omega) : m_1 = m_2 = \cdots = m_p = m$$

Notant $Y_{\bullet\bullet}$ la moyenne arithmétique des observations (Y_{ij}), l'EMCO dans (ω) est :

$$\widehat{m}(\omega) = Y_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{p}\sum_{k=1}^{n_i} Y_{ik} \text{ et } SCR(\omega) = \sum_{i=1}^{p}\sum_{k=1}^{n_i}(Y_{ik} - Y_{\bullet\bullet})^2$$

Décomposition de l'analyse de la variance.
La variance sur la population totale vaut : $V = SCR(\omega) = SCT = \sum_{i,k}(Y_{ik} - Y_{\bullet\bullet})^2$. Un calcul direct donne la décomposition fondamentale de l'analyse de la variance :

$$V = B + W \text{ , avec } \begin{cases} B = \sum_i n_i(Y_{i\bullet} - Y_{\bullet\bullet})^2 = SCR(\omega) - SCR(\Omega) \\ W = SCR(\Omega) = \sum_{i,k}(Y_{ik} - Y_{i\bullet})^2 \end{cases}$$

Ce qui se lit : *la variance totale V est la somme de la variance interpopulation B et de la variance intrapopulation W*[1].

Le tableau d'analyse de la variance à un facteur est le suivant :

Source de variation	*Somme de carrés*	*ddl*	*Carré moyen*
Effet I	$B = \sum_i n_i(Y_{i\bullet} - Y_{\bullet\bullet})^2$	$p-1$	$\frac{B}{p-1}$
Résiduelle	$W = \sum_{i,k}(Y_{ik} - Y_{i\bullet})^2$	$n-p$	$\widehat{\sigma}^2_\Omega = \frac{W}{n-p}$
Totale	$V = \sum_{i,k}(Y_{ik} - Y_{\bullet\bullet})^2$	$n-1$	****

La statistique de Fisher du test de (ω) est :

$$F = \frac{n-p}{p-1}\frac{\sum_i n_i(Y_{i\bullet} - Y_{\bullet\bullet})^2}{\sum_{i,k}(Y_{ik} - Y_{i\bullet})^2} = \frac{n-p}{p-1}\frac{B}{W} \underset{(\omega)}{\sim} F(p-1, n-p)$$

Test de Bartlett d'identité des p variances résiduelles

Supposons (Ω^*) que chaque population ait sa propre variance σ_i^2, c'est à dire que $Var(\varepsilon_{ij}) = \sigma_i^2$, $i = 1, p$. Raisonablement, il faut commencer par tester (Ω) : $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_p^2$ dans (Ω^*). Après quoi, si on conclut à l'homogénéité des variances résiduelles, on effectue le test désiré sur la moyenne.

Pour tester (Ω) dans (Ω^*), on utilise le test asymptotique du rapport de vraisemblance (cf. Annexes). Il est facile de voir que, à une constante universelle près, la log vraisemblance sous (Ω^*) vaut $l_n(\Omega^*) = -\frac{1}{2}\sum_i n_i \log \widehat{\sigma_i^2}$, où $\widehat{\sigma_i^2} = \frac{SCR(i)}{n_i}$ est l'EMV de σ_i^2, alors que sous

[1] B pour "*Between*" et W pour "*Within*".

(Ω), $l_n(\Omega) = -\frac{n}{2}\log\widehat{\sigma^2}$ où $\widehat{\sigma^2} = \frac{\sum SCR(i)}{n}$ est l'EMV sous (Ω). On en déduit que, pour $n_i \to \infty$, $i = 1, p$, et puisque $\dim(\Omega^*) - \dim(\Omega) = (p-1)$:

$$n\log\widehat{\sigma^2} - \sum_{i=1,p} n_i \log\widehat{\sigma_i^2} \underset{(\Omega)}{\sim} \chi_{p-1}$$

Bartlett a donné une forme améliorée de cette approximation asymptotique :

$$B = \frac{(n-p)\log\widehat{\sigma^2} - \sum_{i=1,p}(n_i-1)\log\widehat{\sigma_i^2}}{1 + \frac{1}{3(p-1)}[\sum_{i=1,p}\frac{1}{n_i-1} - \frac{1}{n-p}]} \underset{(\Omega)}{\sim} \chi^2_{p-1}$$

Pour les analyses à 2 ou plus de 2 facteurs, je n'ai pas traité le cas déséquilibré, signalant seulement que cela se résoud en utilisant le modèle linéaire, et que le cas des dispositifs orthogonaux à deux facteurs ($n_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}}$ pour tout i, j) présente les même propriétés que celles du modèle équilibré. De toute façon, tu as ici des partie importantes et propre à développer avec la planification expérimentale

Je pense qu'il serait aussi important de développer les modèles à effets aléatoires et les modèles mixtes. Je pense que les médecins autant que les agronomes en ont besoin.

2.2 Analyse de la variance à 2 facteurs

2.2.1 Le modèle complet équilibré

C'est le modèle (2.2) avec un nombre fixe $n \geq 2$ de répétitions : $\forall i, j$, $n_{ij} = n$:

$$(\Omega) : Y_{i,j;k} = m_{ij} + \varepsilon_{i,j;k}\ ,\ k = 1, n\,;\ i = 1, I\ ,\ j = 1, J$$

i est le premier facteur, j le deuxième facteur. Ce modèle, de dimension paramétrique $I \times J$, a pour paramètres $(m_{ij}, (i,j) \in I \times J)$. Posant $t = (i,j)$ et considérant le modèle à un facteur t, l'EMCO de (m_{ij}) et la $SCR(\Omega)$ associée sont :

$$\widehat{m}_{ij} = Y_{ij\bullet},\ SCR(\Omega) = \sum_{ijk}(Y_{ijk} - Y_{ij\bullet})^2 \text{ et } \widehat{\sigma}^2_\Omega = \frac{SCR(\Omega)}{IJ(n-1)}$$

Décomposition de l'espace de la moyenne

On peut décomposer le paramètre $m = (m_{ij})$ de la façon suivante :

$$\begin{array}{rcl} m_{ij} & = & \mu + a_i + b_j + c_{ij}, \text{ avec} \\ \forall i, j & : & a_\bullet = b_\bullet = c_{i\bullet} = c_{\bullet j} = 0 \end{array} \tag{2.3}$$

Les contraintes imposées sur a, b, c font que cette décomposition est unique, la correspondance entre les paramètres (m_{ij}) et $\theta = (\mu, a, b, c)$ étant bijective :

$$\forall i, j : \left\{ \begin{array}{c} \mu = m_{\bullet\bullet},\ a_i = m_{i\bullet} - m_{\bullet\bullet},\ b_j = m_{\bullet j} - m_{\bullet\bullet} \\ c_{ij} = m_{ij} - m_{i\bullet} - m_{\bullet j} + m_{\bullet\bullet} \end{array} \right.$$

L'interprétation des paramètres (μ, a, b, c) est la suivante :

(1) μ est la moyenne générale ;
(2) Les (a_i) sont les effets principaux (centrés) du facteur i ;
(3) Les (b_j) sont les effets principaux (centrés) du facteur j ;
(4) Les (c_{ij}) sont les interactions entre les facteurs I et J.

Les espaces associés à μ, a, b, c sont respectivement de dimensions 1, $I-1$, $J-1$ et $(I-1)(J-1)$. Pour obtenir la dimension de l'espace des interactions, il faut remarquer que les $(I+J)$ relations $\{\forall i, j : c_{i\bullet} = c_{\bullet j} = 0\}$ sont contraintes par $\sum_i c_{i\bullet} = \sum_j c_{\bullet j}$. L'identité :

$$IJ = 1 + (I-1) + (J-1) + (I-1)(J-1)$$

traduit que la dimension de (m_{ij}) est égale à celle de $\theta = (\mu, a, b, c)$.

Estimation des effets principaux et des interactions

$m \mapsto \theta$ étant bijective, l'estimation des MCO de θ s'obtient en remplaçant (m_{ij}) par $(\widehat{m}_{ij})$ dans cette transformation :

$$\widehat{\mu} = Y_{\bullet\bullet\bullet},\ \widehat{a}_i = Y_{i\bullet\bullet} - Y_{\bullet\bullet\bullet},\ \widehat{b}_j = Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet},\ \widehat{c}_{ij} = Y_{ij\bullet} - Y_{i\bullet\bullet} - Y_{\bullet j\bullet} + Y_{\bullet\bullet\bullet}$$

L'autre propriété importante est la *décomposition* des sous espace (μ), (a), (b) et (c) est *orthogonale,* espaces respectivement de dimension 1, $I-1$, $J-1$ et $(I-1)(J-1)$. En particulier, on en déduit la décomposition de la somme des carrés totale SCT :

$$SCT = \sum_{ijk}(Y_{ijk} - Y_{\bullet\bullet\bullet})^2 = nJ\sum_i \widehat{a}_i^2 + nI\sum_j \widehat{b}_j^2 + n\sum_{ij}\widehat{c}_{ij}^2 + SCR(\Omega)$$

La première somme $SCE(I)$ mesure l'effet du facteur I, la deuxième $SCR(J)$ associée à J, la troisième $SCR(I\times J)$ à l'interaction $I\times J$. La $SCR(\Omega) = \sum_{ijk}(Y_{ijk} - Y_{ij\bullet})^2$ ne dépend pas du choix de paramétrisation de (Ω). Ces différentes SC sont des $\sigma^2\chi^2$ indépendants, de d.d.l. la dimension de l'espace associé. On en déduit le tableau d'analyse de la variance à deux facteurs :

Source	*Somme des carrés*	*ddl*	*Carré moyen*
Effet I	$nJ\sum_i \widehat{a}_i^2$	$I-1$	$\frac{nJ}{I-1}\sum_i \widehat{a}_i^2$
Effet J	$nI\sum_j \widehat{b}_j^2$	$J-1$	$\frac{nI}{J-1}\sum_j \widehat{b}_j^2$
Interaction $I\times J$	$n\sum_{ij}\widehat{c}_{ij}^2$	$(I-1)(J-1)$	$\frac{n}{(I-1)(J-1)}\sum_{ij}\widehat{c}_{ij}^2$
Résiduelle	$\sum_{ijk}(Y_{ijk} - Y_{ij\bullet})^2$	$(n-1)IJ$	$\widehat{\sigma}^2(\Omega) = \frac{\sum_{ijk}(Y_{ijk}-Y_{ij\bullet})^2}{(n-1)IJ}$
SCT	$\sum_{ijk}(Y_{ijk} - Y_{\bullet\bullet\bullet})^2$	$nIJ-1$	*****

Pour le produit scalaire canonique de $\mathbf{R}^{I\times J\times n}$, (2.3) s'interprète comme une *somme directe orthogonale* d'espaces vectoriels :

$$\mathbf{R}^{I\times J\times n} = C \oplus M_I \oplus M_J \oplus M_{I\times J} \tag{2.4}$$

C est l'espace des fonctions constantes (de dimension 1), $C\oplus M_I$ l'espace des fonctions ne dépendant que de i (de dimension I), $C\oplus M_J$ l'espace des fonctions ne dépendant que de j (de dimension J) ; l'espace des interactions $M_{I\times J}$ est l'orthogonal de $C\oplus M_I\oplus M_J$ dans $\mathbf{R}^{I\times J}$; il est de dimension $I\times J - (I+J-1) = (I-1)(J-1)$. Remarquons au passage que les seul sous modèle ayant un sens intrinsèque sont C, $C\oplus M_I$ et $C\oplus M_J$.

La décomposition (2.4) permet de définir les sous-modèles intrinsèques C, $C\oplus M_I$ et $C\oplus M_J$, le sous-modèle additif $C\oplus M_I\oplus M_J$ correspondant à l'annulation des interactions c ; il est de dimension $I+J-1$.

2.2.2 Le modèle additif $(\mathcal{A})$: $E(Y_{ijk}) = \mu + a_i + b_j$

(2.4) étant une décomposition orthogonale, l'EMCO des paramètres μ, a et b coïncide avec celle de ces paramètres dans le modèle complet, et la nouvelle $SCR(\mathcal{A})$ n'est autre que $SCR(\Omega)$ à laquelle on ajoute $SCR(I\times J)$. On obtient :

$$\widehat{m}_{ij}(\mathcal{A}) = \widehat{Y}_{ijk}(\mathcal{A}) = \widehat{\mu} + \widehat{a}_i + \widehat{b}_j = Y_{i\bullet\bullet} + Y_{\bullet j\bullet} - Y_{\bullet\bullet\bullet},\ SCR(\mathcal{A}) = SCR(\Omega) + n\sum_{ij}\widehat{c}_{ij}^2$$

L'interprétation de $SCR(\mathcal{A})$ est la suivante : à la $SCR(\Omega)$ du modèle complet (Ω) s'ajoute la somme des carrés correspondant aux interactions. $SCR(\mathcal{A})$ est à $nIJ-(I+J-1)$ ddl.

Le test d'additivité des facteurs I et J repose sur la statistique :

$$F = \frac{(n-1)IJ}{(I-1)(J-1)} \times \frac{n\sum_{ij}\widehat{c}_{ij}^2}{SCR(\Omega)} \underset{(\mathcal{A})}{\sim} F((I-1)(J-1),(n-1)IJ)$$

2.2.3 Absence d'effet J : $(\mathcal{I}) : m_{ij} = m_i = \mu + a_i$

On a les inclusions : $(\mathcal{I}) \subset (\mathcal{A}) \subset (\Omega)$. $(\mathcal{I})$ est estimé par :

$$\widehat{m}_{ij}(\mathcal{I}) = \widehat{\mu} + \widehat{a}_i = Y_{i\bullet\bullet},\ SCR(\mathcal{I}) = nI\sum_j \widehat{b}_j^2 + n\sum_{ij}\widehat{c}_{ij}^2 + SCR(\Omega)$$

A $SCR(\mathcal{A})$ s'ajoute la somme des carrés due au facteur J. Ce modèle peut aussi s'écrire directement comme un modèle au seul facteur I, (j,k) étant l'indice de répétition :

$$(\mathcal{I}) : \forall (i,j,k),\ E(Y_{ijk}) = m_i$$

On en déduit une deuxième expression équivalente de $SCR(\mathcal{I})$:

$$SCR(\mathcal{I}) = \sum_{ijk}(Y_{ijk} - Y_{i\bullet\bullet})^2$$

Suivant que l'alternative est (Ω) ou $(\mathcal{A})$, la statistique du test de $(\mathcal{I})$ est F_1 ou F_2 :

$$\begin{aligned} F_1 &= \frac{(n-1)IJ}{I(J-1)} \times \frac{nI\sum_j \widehat{b}_j^2 + n\sum_{ij}\widehat{c}_{ij}^2}{SCR(\Omega)} \underset{(\mathcal{I})}{\sim} F(I(J-1),(n-1)IJ) \\ F_2 &= \frac{(n-1)IJ + I + J - 1}{(J-1)} \times \frac{nI\sum_j \widehat{b}_j^2}{SCR(\mathcal{A})} \underset{(\mathcal{I})}{\sim} F(J-1,(n-1)IJ + I + J - 1) \end{aligned}$$

2.2.4 Dispositif sans répétition

Supposons qu'on ne dispose que d'une observation par modalité (i,j) :

$$Y_{ij} = m_{ij} + \varepsilon_{ij},\ i = 1,I \text{ et } j = 1,J$$

Le modèle complet (Ω) n'est plus estimable puisqu'il y a IJ observations pour $IJ + 1$ paramètres $\{(m_{ij}), \sigma^2\}$. Par contre le modèle additif $(\mathcal{A})$: $m_{ij} = \mu + a_i + b_j$ l'est. L'estimation des MCO de $(\mathcal{A})$ reste inchangée par rapport à celle du modèle en prenant $n = 1$:

$$\begin{aligned} \widehat{m}_{ij} &= \widehat{\mu} + \widehat{a}_i + \widehat{b}_j = Y_{i\bullet} + Y_{\bullet j} - Y_{\bullet\bullet} \\ SCR(\mathcal{A}) &= \sum_{i,j}(Y_{ij} - Y_{i\bullet} - Y_{\bullet j} + Y_{\bullet\bullet})^2,\ \widehat{\sigma}^2 = \frac{SCR(A)}{(I-1)(J-1)} \end{aligned}$$

D'autres modèles intermédiaires $(\mathcal{B})$, $(\mathcal{A}) \subset (\mathcal{B}) \subset (\Omega)$, sont estimables. Il faut que $\dim \mathcal{B} < IJ$. Par exemple, si le facteur j est repéré par une variable $z_j \in \mathbf{R}$, on peut proposer une modélisation régressive de l'interaction c_{ij} :

$$(\mathcal{B}) : m_{ij} = \mu + a_i + b_j + c_i(z_j - \overline{z})$$

Puisque $\sum_j (z_j - \overline{z}) = 0$, les contraintes sur l'interaction sont satisfaites dès que $c_\bullet = 0$. Cette contrainte assure que la nouvelle interaction appartient encore à l'espace général $M_{I\times J}$: l'estimation de μ, a et b reste inchangée. Celle de $c = (c_i)$ s'obtient en régressant $(\widehat{c}_{ij})_j$ sur $(c_i(z_j - \overline{z}))_j$, c.a.d. en minimisant

$$\Delta(c) = \sum_{ij}\{\widehat{c}_{ij} - c_i(z_j - \overline{z})\}^2, \text{ où } \widehat{c}_{ij} = Y_{ij} - Y_{i\bullet} + Y_{\bullet j} - Y_{\bullet\bullet}$$

On obtient $\widehat{c}_i = \sum_j \widehat{c}_{ij}(z_j - \overline{z})\{\sum_j (z_j - \overline{z})^2\}^{-1}$. $SCR(\mathcal{B})$ est à $IJ - 2I - J + 2$ ddl. Le test de non-interaction dans $(\mathcal{B})$ repose sur la statistique :

$$F = \frac{\sum_{ij}\widehat{c}_i^2(z_j - \overline{z})^2/(I-1)}{\widehat{\sigma}_{\mathcal{B}}^2} \underset{(\mathcal{A})}{\sim} F(I-1, IJ - 2I - J + 2)$$

2.3 Analyse de la variance à trois facteurs

Le modèle d'analyse de la variance à 3 facteurs, équilibré et avec répétitions, est :

$$(\Omega) : E(Y_{ijk,l}) = m_{ijk},\ i = 1, I, j = 1, J, k = 1, K \text{ et } l = 1, L, \text{ avec } L \geq 2$$

L'indice de répétition est l. (Ω) est de dimension IJK. Les EMCO sont $\widehat{m}_{ijk} = Y_{ijk,\bullet}$, $SCR(\Omega) = \sum_{ijkl}(Y_{ijkl} - Y_{ijk\bullet})^2$, $\widehat{\sigma}^2_\Omega = SCR(\Omega)/(L-1)IJK$.

La décomposition en effets principaux et interactions étend celle étudiée pour 2 facteurs :

$$m_{ijk} = \mu + a_i + b_j + c_k + (ab)_{ij} + (ac)_{ik} + (bc)_{jk} + (abc)_{ijk} \tag{2.5}$$

Il y a 3 effets principaux a, b et c (de dimensions $I-1$, $J-1$ et $K-1$), 3 interactions d'ordre 2, $(ab), (bc)$ et (ac) (de dimensions $(I-1)(J-1)$, $(J-1)(K-1)$ et $(I-1)(K-1)$) et une interaction d'ordre 3, (abc), de dimension $(I-1)(J-1)(K-1)$. Ces paramètres sont définis de façon unique sous les contraintes :

$$\begin{aligned} a_\bullet &= b_\bullet = c_\bullet = 0 \\ (ab)_{i\bullet} &= (ab)_{\bullet j} = (ac)_{i\bullet} = (ac)_{\bullet k} = (bc)_{j\bullet} = (bc)_{\bullet k} = 0 \\ (abc)_{ij\bullet} &= (abc)_{i\bullet k} = (abc)_{\bullet jk} = 0 \end{aligned}$$

$(m_{ijk}) \mapsto \theta = (\mu, a, b, c, (ab), (ac), (bc), (abc))$ étant bijective, l'EMCO de θ s'obtient en remplaçant (m_{ijk}) par $(\widehat{m}_{ijk})$ dans cette correspondance.

La décomposition (2.5) permet de définir des sous-modèles (ω) dont l'estimation se déduira de celle de θ en conservant les composantes de θ spécifiant (ω). Cette propriété résulte de l'orthogonalité de la décomposition (2.5).

Examinons l'exemple du modèle (ω) sans interaction d'ordre 3 et avec (ab) comme unique interaction d'ordre 2, modèle que l'on peut écrire :

$$(\omega) : (abc) = (ac) = (bc) = 0, \text{ ou } (\omega) : a + b + c + (ab), \text{ ou } (\omega) : (a, b) + c$$

(ω) est le modèle additif $(a, b) + c$. Sa dimension s'obtient soit en comptant les paramètres contraints, ici $1 + (I-1) + (J-1) + (K-1) + (I-1)(J-1)$, soit en décomptant du modèle complet les paramètres éliminés, ici $IJK - \{(I-1)(J-1) + (I-1) + (J-1)\}(K-1)$.

La somme des carrés résiduels d'un sous-modèle s'obtient en ajoutant à $SCR(\Omega)$ les SC associées aux effets principaux ou interactions qui n'apparaissent plus dans (ω). Le test de (ω) dans le modèle complet repose sur la statistique de Fischer $F(\dim(\Omega) - \dim(\omega), IJKL - \dim(\Omega))$. Pour le modèle (ω) examiné ci-dessus, cette statistique sera à $(IJ-1)(K-1)$ et $IJK(L-1)$ ddl.

2.4 Exercices : Analyse de la Variance et modèle linéaire

Exercise 7 *Identité de 8 portées de cochons*

Le tableau ci-dessous donne les poids en livres à la naissance de 8 portées de cochons :

1	2.0	2.8	3.3	3.2	4.4	3.6	1.9	3.3	2.8	1.1
2	3.5	2.8	3.2	3.5	2.3	2.4	2.0	1.6		
3	3.3	2.6	3.6	3.1	3.2	3.3	2.9	3.4	3.2	3.2
4	3.2	3.3	3.2	2.9	3.3	2.5	2.6	2.8		
5	2.6	2.6	2.9	2.0	2.0	2.1				
6	3.1	2.9	3.1	2.5						
7	2.6	2.2	2.2	2.5	1.2	1.2				
8	2.5	2.4	3.0	1.5						

T.P. : Utiliser **R** pour répondre aux questions suivantes.

(1) Tester l'égalité des poids moyens des 8 portées.

(2) Les portées sont issues de deux géniteurs A et B : 1, 3 et 4 proviennent de A, les 5 autres de B. Tester l'identité des géniteurs.

(3) Tester l'identité des 4 portées les plus importantes {1, 2, 3 et 4} et des 4 portées les moins importantes {5,6,7 et 8}.

Exercise 8 *Infections hépatiques*

(1) On mesure y le nombre de monocytes par mm^3 chez 200 sujets présentant trois infections hépatiques différentes, notées H1, H2 et H3. Le tableau ci-contre donne le nombre de sujets, les sommes SY et SY^2 par groupe :

	H1	H2	H3
n	50	90	60
SY	26 003	87 886	43 469
SY^2	14 303 809	90 336 042	34 279 059

Y a-t-il une différence significative entre les trois infections ?

(2) Il s'avère que H2 recouvre deux maladies différentes, disons H2a et H2b. Le tableau ci dessous est relatif aux mêmes données, mais on y a séparé les deux maladies :

	H1	H2a	H2b	H3
n	50	40	50	60
SY	26 003	36 630	51 256	43 469
SY^2	14 303 809	36 572 940	53 763 102	34 279 059

Pour le critère "nombre de monocytes par mm^3", y a-t-il une différence significative entre H2a et H2b.

Exercise 9 *Vitesse d'apprentissage pour des rats de laboratoire*

On mesure le temps y nécessaire à des rats de laboratoire pour faire fonctionner un dispositif leur donnant de la nourriture. Certains rats ont déjà appris à faire fonctionner un appareillage analogue (rats dit P=prédressés), d'autre non. Certains rats recoivent une substance accroissant leur capacité (rats dit D=dopés), d'autres non. Voici les mesures (en minutes).

non D - non P	12.73	16.79	12.63	17.83	15.21	16.44
non D - P	14.21	11.94	13.22	14.74	13.03	11.72
D - non P	14.91	11.44	16.70	12.93	14.39	15.00
D - P	12.38	10.73	11.19	13.61	12.22	10.00

Quel modèle retenir entre (H_1) : pas d'effet P ni D ; (H_P) : seul effet P ; (H_D) : seul effet D ; $(\mathcal{A})$: modèle P + D et (H_4) : modèle complet.

On donne : $SC(D) = 9.38$, $SC(P) = 32.22$, $SCR(A) = 56.66$ et $SCR(H_4) = 42.34$.

T.P. : utiliser **R** pour dresser la table d'analyse de la variance. Retrouver les sommes de carrés annoncées.

Exercise 10 *Vitesse coronarienne : régression où analyse de la variance ?*

On mesure chez $n = 20$ patients le poids x_i, le taux de cholestérol t_i et la vitesse de circulation coronarienne y_i mesurée par effet Dopler. Le but est de proposer des modèles expliquant y à partir de x et t.

(1) On propose deux modèles de régression affine :

$$(H_3) : E(y) = m + ax + bt \text{ et } (H_2) : (H_3) \text{ avec } b = 0$$

On trouve $SCR(H_3) = 22.63$ et $SCR(H_2) = 34.46$. Tester la non influence du taux de cholestérol dans (H_3).

(2) Doutant que les relations entre x, t et y soient linéaires, on adopte une démarche moins paramétrique en regroupant les individus par classes :

– de poids : 40 à 50kg, 51 à 60, 61 à 71, 71 à 80 et >80kg
– de taux de cholestérol : ≤ 1.8, 1.9 à 2.1, 2.2 à 2.6 et ≥ 2.7

Il se trouve qu'un individu et un seul se trouve dans chacune des 20 cases obtenues, et les données observées sont :

Cholestérol	≤ 1.8	1.9 à 2.1	2.2 à 2.6	≥ 2.7
Poids ≤ 50	77.88	77.41	76.52	75.09
51 à 60	77.00	72.42	72.09	71.96
61 à 70	68.63	68.53	70.60	68.62
71 à 80	66.28	66.34	67.88	64.22
> 80	61.06	62.34	59.58	55.61

On se place dans le modèle additif :

$$(\mathcal{A}) : E(y_{i,j}) = m + a_i + b_j,\ i = 1,5 \text{ et } j = 1,4$$

Tester au niveau 5% (H_x) : *pas d'effet poids* ; (H_t) : *pas d'effet cholestérol* (on a obtenu : $SCR(poids) = 691.48$, $SCR(chol.) = 28.05$, $SCR(\mathcal{A}) = 27.74$ et $SCT(\mathcal{A}) = 747.27$; retrouvez ces statistiques en utilisant **R**). Comparez les modèles (H_3) et ($\mathcal{A}$).

T.P. : utiliser **R** pour dresser la table d'analyse de la variance.

(3) On étend le modèle ($\mathcal{A}$) en ($\mathcal{A}^+$) avec une interaction modélisée par :

$$(ab)_{ij} = \theta_j(x_i - \overline{x}) \text{ et } \overline{\theta} = 0$$

On trouve $SCR((\mathcal{A}^+) = 22.65$. Tester la validité du modèle additif.

Exercise 11 *Cancer du sein : un modèle d'analyse de la covariance*

On étudie la durée de survie y de femmes atteintes d'un cancer du sein ceci pour trois types de traitements A, B et C. Le tableau suivant figure également l'âge x d'apparition du cancer :

Trait.	A	Trait.	B	Trait.	C
Age	Survie	Age	Survie	Age	Survie
32.7	6.5	33.3	8.5	30.3	11.9
37.2	8.8	40.4	5.6	31.7	5.6
37.3	10.0	41.6	9.1	31.9	7.9
39.8	8.7	43.4	7.4	33.9	9.0
42.6	8.4	44.5	4.1	36.2	8.7
44.2	4.1	46.5	5.9	39.9	9.8
45.4	6.1	47.8	7.7	41.4	9.5
47.0	5.6	47.9	6.4	42.6	7.6
47.4	3.7	49.2	5.8	43.3	7.7
47.6	8.9	52.3	6.3	43.6	5.2
49.3	6.4	52.8	5.7	43.6	8.5
50.2	5.2	52.8	3.3	44.1	7.4
50.4	7.4	53.0	2.7	44.5	5.1
51.4	4.0	55.2	4.0	45.9	5.7
51.8	7.0	56.1	3.2	46.5	7.3
52.0	6.8	56.4	4.3	48.8	4.6
53.5	4.6	56.5	3.8	49.0	6.8
53.6	4.7	56.6	1.5	49.2	5.8
55.8	4.7			50.4	8.6
56.4	4.7			50.7	5.1
58.7	4.3			52.7	6.5
59.4	3.8				
63.3	2.1				

(1) Calculer la moyenne de survie dans chaque groupe. Sans tenir compte de l'âge d'apparition du cancer, tester l'existence (H_T) d'un effet traitement (on obtient : $SC(Trait) = 55.13$ et $SCR(A) = 88.33$, $SCR(B) = 73.91$ et $SCR(C) = 73.65$).

(2) On soupçonne un lien entre l'âge d'apparition et la durée de survie et on envisage les deux modèles :

$$(H_2) : E(y) = m + ax \text{ et } (H_4) : E(y) = \begin{cases} m_A + ax \text{ pour le trait. } A \\ m_B + ax \text{ pour le trait. } B \\ m_C + ax \text{ pour le trait. } C \end{cases}$$

Tester (H_2) contre (H_4) (on donne $SCR(H_2) = 121.36$, $SCR(H_4) = 113.37$). Expliquer le résultat des questions (1) et (2) (constater qu'une Anova sur les âges montre une différence significative sur les âges x, d'où une confusion).

(3) Expliquer comment voir estimeriez le modèle (H_6) où on fait dépendre la pente a du traitement A, B ou C. Comment tester (H_2) dans (H_6) ?

T.P. : utiliser **R** pour retrouver les SC annoncées, mais aussi les estimations des paramètres et leur ecarts types.

Exercise 12 *Modélisation d'un coût de maintenance*

Le tableau suivant donne les coûts de maintenance d'un appareil en fonction de son âge x,

x	.5	.5	1	1	1	4	4	4	4.5	4.5	4.5	5	5	5	5.5	6	6
y	16	18	98	47	55	49	72	68	62	105	103	89	152	120	99	76	137

T.P. : utiliser **R** pour répondre aux questions suivantes.

(1) Effectuer l'analyse de la variance du modèle (Ω) à un facteur *qualitatif* "Age".

(2) On considère les deux sous-modèles de régression expliqués par x_i :

$$(H_3) : E(Y_{ij}) = a + bx_i + cx_i^2,\ i = 1, 7\,; \quad (H_1) : c = 0$$

Tester (H_3) dans (Ω) et (H_2) dans (H_3).

(3) Soit (H_2^{-3}) le modèle (H_2) avec la modélisation spécifique de la donnée $n°3$, $E(Y_{2,1}) = m$ ((H_2^{-3}) traduit que la donnée $n°3$ est aberrante). Les paramètres de (H_2^{-3}) sont a, b,m. Tester que la donnée $n°3$ n'est pas aberrante.

Exercise 13 *Influence de la densité d'un semis sur le rendement*

Des parcelles de $52,5$ m^2 sont ensemencées en orge avec des densités de semis différentes. On a relevé le nombre de plantes semées, et trois rendements différents pour chaque densité,

Densité / m^2	Plantes levées	Les 3 rendements
100	96	21.1 20.0 19.7
200	162	21.3 21.6 22.2
300	292	22.0 21.4 23.6
400	388	22.4 22.0 23.5
500	488	21.8 21.6 23.5

T.P. : répondre aux questions suivantes.

(1) Effectuer la régression du nombre de plantes levées sur le nombre de plantes semées. Peut-on admettre que le nombre de plantes levées est proportionnel au nombre de plantes semées.

(2) Les rendements en riz dépendent-ils de la densité de semis ?

(3) Estimer le modèle : "le rendement est une fonction affine de la densité de semis".

Chapitre 3

Asymptotique du modèle linéaire

L'objectif est l'étude asymptotique (n grand) du modèle linéaire :

$$(\Omega) : y_i = {}^t x_i \beta + e_i,\ i = 1, n \tag{3.1}$$

Nous allons répondre aux questions suivantes :

(i) Les estimateurs des MCO sont-ils convergents : $(\widehat{\beta}, \widehat{\sigma}^2) \xrightarrow{\Pr} (\beta, \sigma^2)$ *si* $n \to \infty$ *?*
(ii) Si oui, à quelle vitesse ? Quelle est la loi approximative de $\{(\widehat{\beta}, \widehat{\sigma}^2) - (\beta, \sigma^2)\}$ *?*
(iii) Delta méthode : loi, pour n grand, d'une transformée non-linéaire $F(\widehat{\beta})$ *de* $\widehat{\beta}$ *?*
(iv) Comment tester une sous-hypothèse non-linéaire pour n grand ?

On supposera que les résidus sont i.i.d. et centrés, de variance σ^2, sans donner plus de précision sur leur loi commune. L'écriture matricielle de (3.1) est $Y = X\beta + e$, $E(e) = 0$ et $Cov(e) = \sigma^2 I_n$. Si nécessaire, on notera X_n et $e(n)$ pour X et e afin d'expliciter leurs dépendances en n.

Deux types de convergence sont considérés :

- la convergence en probabilité, notée $\xrightarrow{\Pr}$;
- la convergence en loi, notée $\xrightarrow{Loi}$.

Leurs définitions et les résultats principaux liés à ces modes de convergence sont données dans l'annexe.

Les réponses dépendent du comportement de la matrice exogène X_n pour n grand et de conditions de moments sur les résidus. Pour la première question (convergence des MCO), il est suffisant de supposer que les résidus sont non corrélés. Pour les autres, l'indépendance des résidus est nécessaire mais non leur caractère gaussien.

3.1 Convergence des MCO de β

Rappelons que si (Z_n) est une suite de variables aléatoires vectorielles de $\mathbf{R}^k$ dont l'espérance converge vers m et dont la matrice de variance tend vers 0 (c'est à dire que tous ses termes tendent vers 0), alors $Z_n \xrightarrow{\Pr} m$. L'EMCO de β étant sans biais, on obtient la condition suffisante de convergence suivante :

Proposition 9 *Considérons le modèle (3.1) à résidus bruit blanc centré de variance finie. Alors* $\widehat{\beta}_n \xrightarrow{\Pr} \beta$ *dès que* $({}^t X_n X_n)^{-1} \longrightarrow 0$.

Example 11 *Dispositif expérimental ergodique*

Si $\frac{1}{n} {}^t X_n X_n \longrightarrow Q$ et si Q est définie positive, les MCO sont convergents. Ces deux conditions sont satisfaites si les (x_i) sont les réalisations i.i.d. d'une loi Z sur $\mathbf{R}^p$ de carré

intégrable telle que $Q = E(Z^t Z)$ est dp. En effet la loi faible des grands nombres dit que :

$$\frac{1}{n}\,{}^t X_n X_n = \frac{1}{n}\sum_1^n x_i^t x_i \longrightarrow E(Z\,{}^t Z)$$

Example 12 *Convergence des MCO pour une régression affine.*

La condition précédente est suffisante mais elle n'est pas nécessaire. Si les (x_i) sont bornés, il suffit que $nVar(x) = \sum_1^n (x_i - \overline{x}(n))^2 \to \infty$.

En effet $\Delta_n = ({}^t X_n X_n)^{-1} = \frac{1}{\sum_1^n (x_i - \overline{x}(n))^2} \begin{pmatrix} \frac{1}{n}\sum_1^n x_i^2 & -\overline{x}(n) \\ -\overline{x}(n) & 1 \end{pmatrix}$.

Example 13 *Analyse de la variance à un facteur*

La convergence des MCO est assurée dès que chaque $n_i \to \infty$.

3.1.1 Convergence gaussienne de $\widehat{\beta}_n$

On suppose que les résidus sont i.i.d. et admettent un moment absolu d'ordre 3 :

$$(H_e) : \text{les } (e_i) \text{ sont i.i.d. et } (E(|e_1|^3) < \infty$$

et que la suite des dispositifs (X_n) vérifie :

$$(H_X) : (i)\ \frac{1}{n}\,{}^t X_n X_n \longrightarrow Q \text{ où } Q \text{ est dp, et } (ii)\ \sum_1^n \|x_i\|^3 = O(n^{\frac{3}{2}})$$

$(H_X - (i))$ assure la convergence des MCO. $(H_X - (ii))$ va permettre de vérifier la condition du théorème central limite (TCL) de Lyapunov (cf. Annexes), et assurera la normalité asymptotique de $\widehat{\beta}_n$. $(H_X - ii)$ est satisfaite si les (x_i) sont bornés.

Proposition 10 *Considérons le modèle linéaire (3.1) à résidus i.i.d., centrés et de variance σ^2. Alors, sous les conditions (H_e) et (H_X) :*

$$\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{Loi} \mathcal{N}_p(0, \sigma^2 Q^{-1})$$

Dans la pratique, pour n grand, $\frac{\sigma^2}{n}Q^{-1}$ est remplacée par $\widehat{\sigma}^2({}^t X_n X_n)^{-1}$: la loi approximative de l'EMCO est $(\widehat{\beta}_n - \beta) \sim \mathcal{N}_p(0, \widehat{\sigma}^2({}^t X_n X_n)^{-1})$.

Preuve :
L'espérance et la variance de $\widehat{\beta}_n$ ont déjà été identifiées. Il suffit donc de démontrer la convergence gaussienne. Soit $e(n)$ le vecteur des résidus :

$$\sqrt{n}(\widehat{\beta}_n - \beta) = A_n Z_n, \text{ avec } A_n = (\frac{1}{n}\,{}^t X_n X_n)^{-1} \text{ et } Z_n = n^{-\frac{1}{2}}({}^t X_n e(n))$$

Puisque A_n converge vers Q^{-1} (continuité de l'inversion matricielle), il suffit d'établir la convergence gaussienne multidimensionnelle de Z_n. Elle équivaut à la convergence gaussienne de toute suite (réelle) $({}^t a Z_n)$, où $a \in \mathbf{R}^p$ est non-nul. Or ${}^t a Z_n = n^{-\frac{1}{2}} S_n$, avec :

$$S_n = \sum_1^n z_i \text{ , et } z_i = ({}^t a x_i) e_i$$

Les variables z_i sont indépendantes, centrées, admettant un moment absolu d'ordre 3. D'autre part, $s_n^2 = Var(S_n) = \sigma^2\,{}^t a({}^t X_n X_n)^{-1} a$ est équivalente lorsque $n \longrightarrow \infty$ à cn. D'après (H_X), $c > 0$, et donc $\frac{\kappa_n}{s_n} \longrightarrow 0$ où $\kappa_n^3 = \sum_1^3 E(|z_i|^3)$. Le TCL de Lyapunov assure donc la convergence gaussienne de Z_n. ..... □

3.2 Convergence de $\widehat{\sigma}^2$ et de $(\widehat{\beta}, \widehat{\sigma}^2)$

3.2.1 Convergence de $\widehat{\sigma}^2$

La variance résiduelle est estimée par $\widehat{\sigma}_n^2 = \frac{{}^tY(I-P)Y}{n-p} = \frac{{}^te(I-P)e}{n-p}$, où $p = \dim(\Omega)$ et P est l'opérateur de projection orthogonale sur l'espace de la moyenne $\mathcal{E}_X$. La deuxième égalité résulte de $(I-P)Y = (I-P)e$. Notons $S_n = {}^te(I-P)e$:

$$n^{-\frac{1}{2}} S_n = n^{-\frac{1}{2}}\, {}^tee - {}^tZ_n(n^{-\frac{1}{2}}\, {}^tX_nX_n)^{-1}Z_n, \text{ avec } Z_n = n^{-\frac{1}{2}}\,({}^tX_ne)$$

Proposition 11 *Convergence de l'estimation $\widehat{\sigma}_n^2$ de σ^2.*

(i) Sous (H_X) et (H_e), $\widehat{\sigma}^2 \xrightarrow{\Pr} \sigma^2$.

(ii) Si de plus $0 < Var(e_1^2) < +\infty$, alors $\sqrt{n}(\widehat{\sigma}^2 - \sigma^2) \xrightarrow{Loi} \mathcal{N}(0, Var(e_1^2))$.

Preuve :

(i) résulte de (a) : ${}^tZ_n({}^tX_nX_n)^{-1}Z_n \xrightarrow{\Pr} 0$ (Z_n converge en loi vers une gaussienne et $(n^{-\frac{1}{2}}\, {}^tX_nX_n)^{-1} \longrightarrow 0$) et de (b) : l'équivalence asymptotique [1] entre $\frac{1}{n-p}\, {}^tee$ et $\frac{1}{n}\, {}^tee$.

(ii) résulte d'une part du TLC standard pour les variables i.i.d. (e_i^2) (la condition $E(e_1^4) < \infty$ est bien satisfaite), et d'autre part du fait que ${}^tZ_n(n^{-\frac{1}{2}}\, {}^tX_nX_n)^{-1}Z_n \xrightarrow{\Pr} 0$ et de l'équivalence asymptotique entre $\frac{1}{n-p}\, {}^tee$ et $\frac{1}{n}\, {}^tee$. ..... □

3.3 Distribution asymptotique de $F(\widehat{\beta}_n)$: la delta méthode

Considérons une déformation $F : \mathbf{R}^p \longrightarrow \mathbf{R}^q$ (linéaire ou non) de classe $\mathcal{C}^1$ sur un voisinage de β. Si $\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{Loi} \mathcal{N}_p(0, \sigma^2 Q^{-1})$ et si $JF_\beta = (\frac{\partial F_i}{\partial x_j}(\beta))$ est la matrice jacobienne $p \times q$ de F en β, on a le résultat suivant (cf. Annexes, delta méthode) :

$$\sqrt{n}\{F(\widehat{\beta}_n) - F(\beta)\} \xrightarrow{Loi} \mathcal{N}_q(0, \sigma^2 JF_\beta Q^{-1}\, {}^tJF_\beta)$$

Si $F(\beta) = A\beta$ est linéaire, $JF_\beta = A$ et on retrouve un résultat connu : $A\widehat{\beta}_n$ est de variance $AVar(\widehat{\beta}_n)^tA$, gaussienne si $\widehat{\beta}_n$ l'est. Si F est non-linéaire, le résultat asymptotique se maintient en retenant l'application linéaire tangente JF_β de F en β.

Je pense qu'ici il est bien venu de parler de transformation stabilisant la variance : par exemple, $Arc\sin$ pour stabiliser la variance d'une binomiale, $\sqrt{\cdot}$ pour stabiliser la variance d'une loi de Poisson, $\frac{1}{2}\log\frac{1+\widehat{\rho}}{1-\widehat{\rho}}$ pour stabiliser la variance de l'estima teur empirique d'une corrélation, log pour stabiliser la variance de l'estimation de $\widehat{\sigma^2}$, etc ....

La Delta méthode est également utile pour les tests de sous-hypothèses non-linéaires.

3.4 Exercices : TCL, asymptotique du modèle linéaire

Exercise 14 *Modèle linéaire de Bernoulli*

Soit $y_1, y_2, \cdots, y_n$ un échantillon d'une loi de Bernoulli de paramètre p, $0 < p < 1$. On peut interpréter y comme suivant le modèle linéaire à résidus Bernoulli recentrés :

$$y_i = p + e_i,\ i = 1, n$$

(1) Montrer que l'estimateur des MCO et du MV de p coïncident. Déterminer la loi asymptotique de cet estimateur.

(2) On a $Var(e_i) = \sigma^2 = p(1-p)$. Proposer deux estimateurs de σ^2. Les comparer.

[1] Deux suites (U_n) et (V_n) sont asymptotiquement équivalentes si $U_n - V_n \xrightarrow{\Pr} 0$.

Exercise 15 *Estimation du maximum d'une courbe de croissance*

On considère le modèle de régression quadratique :

$$y_i = P(x_i) + e_i,\ i = 1, 30, \text{ où } P(x) = a + bx + cx^2$$

les résidus étant i.i.d. gaussiens, de variance σ^2. Le dispositif (x_i) donne, notant $\beta = (a, b, c)$:

$$({}^tXX)^{-1} = \frac{1}{30}\begin{pmatrix} 2.2 & -0.5 & 1 \\ * & 4 & -0.7 \\ * & * & 1.2 \end{pmatrix}$$

et l'estimation des MCO est : $\widehat{y}_i = 12.4 + 1.7x_i + 1.4x_i^2$, et $SCR = 54.3$.

(1) Tester $c = 0$ contre $c > 0$.

(2) Tester (H_0) : P ne s'annule jamais, contre $(H_1) = (H_0)^c$.

(3) Soit x^* le point où P atteint son extremum et $m^* = P(x^*)$. Estimer (x^*, m^*) et déterminer la loi approximative de cet estimateur.

Exercise 16 *Test d'une sous hypothèse non linéaire*

On considère le modèle de régression

$$y_t = ax_t + bz_t + e_t \text{ pour } t = 1, 22$$

On observe $({}^tXX)^{-1} = \begin{pmatrix} 0.2 & -0.05 \\ -0.05 & 0.2 \end{pmatrix}$, les MCO donnent : $\widehat{a} = 1.2$, $\widehat{b} = 1.7$, $SCR = 30$.

(1) Donner la matrice de covariance estimée de ${}^t(\widehat{a}, \widehat{b})$. Tester $a = b$ contre $a > b$.

(2) Donner la loi approximative du ratio $\widehat{r} = \frac{\widehat{a}}{\widehat{b}}$ (on supposera que $n = 22$ est assez grand pour appliquer la delta méthode). Tester $r = 1$ contre $r > 1$.

Exercise 17 *Transformation stabilisant la variance*

(1) Soit Y une variable aléatoire telle que $Var(Y) = \sigma^2 E(Y)$ avec $m = E(Y)$ "grand". Vérifier que $Z = \sqrt{Y}$ est approximativement $\mathcal{N}(\sqrt{m}, \frac{\sigma^2}{4})$. Donner un exemple de telle loi.

(2) Si $Var(Y) = \sigma^2\{E(Y)\}^2$ (Y est à coefficient de variation $\kappa = \frac{E(Y)}{\sigma(Y)}$ constant) et si $m = E(Y)$ est grand, $\log(Y)$ est approximativement $\mathcal{N}(\log m - \sigma^2, \sigma^2)$. Donner un exemple de telle loi.

(3) Montrer que la transformation $z = \frac{1}{2}\log\frac{1+R}{1-R}$ stabilise la variance de $R \sim \mathcal{N}(0, \frac{(1-R^2)^2}{n})$.

Exercise 18 *Régression à résidus de loi exponentielle*

Soit $y_i = a + e_i$, $i = 1, n$, une régression constante à résidus i.i.d. de loi exponentielle $\mathcal{E}xp(\theta^{-1})$, $\theta > 0$ ($E(e_1) = \theta$, $Var(e_1) = \theta^2$).

(1) θ est connu. Déterminer l'EMCO de a et sa loi asymptotique, ainsi que la loi asymptotique de l'estimateur de la variance résiduelle σ^2 (ici égale à θ^2).

(2) Déterminer la log-vraisemblance des y. En déduire que l'EMV de a et de θ sont :

$$\widehat{a} = y_{(1)} \text{ et } \widehat{\theta} = \overline{y} - y_{(1)}, \text{ où } y_{(1)} = \inf_{i=1,n} y_i$$

(2.1) Déterminer la loi de $\widehat{a}$. Calculer $E(\widehat{a})$ et $E(\widehat{\theta})$. Comment débiaiser ces estimateurs ?
(2.2) Déterminer la loi asymptotique de $\widehat{\theta}$. Comparer cet estimateur à celui obtenu en (1) (montrer que $\sqrt{n}(y_{(1)} - a) \xrightarrow{\Pr} 0$; en déduire que $\sqrt{n}\widehat{\theta}$ est équivalent à $\sqrt{n}\overline{y}$).

Exercise 19 *Loi des grand nombre et Théorème Central Limite*

Soit (X_i) une suite de variables réelles i.i.d. de moyenne m, de variance σ^2 et admettant un moment d'ordre 4. On considère les trois quantités :

$$U_n = \frac{1}{n}\sum_{i=1,n} X_i^2,\ V_n = \frac{1}{n}\sum_{i=1,n-1} X_i X_{i-1} \text{ et } Q_n = \frac{V_n}{U_n}$$

(1) Vérifier que ces trois quantités convergent en probabilité lorsque $n \longrightarrow \infty$ et identifier leurs limites.

(2) Ecrire les trois TCL associés. Quelle est la covariance limite entre $\sqrt{n}U_n$ et $\sqrt{n}V_n$.

(3) *Applications* : (i) $X_1 \sim \mathcal{B}er(p)$; (ii) $X_1 \sim \mathcal{U}nif[-\theta, \theta]$; (iii) $X_1 \sim \mathcal{E}xp(\theta)$, $\theta > 0$.

Chapitre 4

Résidus non-sphériques : les MCG

Les résidus $\varepsilon \in \mathbf{R}^n$ d'un modèle sont dits *sphériques* si leur matrice de covariance $\Sigma = Cov(\varepsilon) = \sigma^2 I_n$. C'est l'hypothèse qui a été faite auparavant. Dans le cas contraire, on dit que les résidus sont *non-sphériques*. Ce chapitre leur est consacré :

$$(\Omega) : Y = X\beta + \varepsilon,\ E(\varepsilon) = 0,\ \Sigma = Cov(\varepsilon) = \sigma^2 R \neq cI_n \tag{4.1}$$

Deux situations classiques conduisent à la non-sphéricité :

(1) Les résidus sont décorrélés mais leurs variances sont inégales : $Cov(\varepsilon) = Diag(\sigma_i^2)$ est diagonale et pour un $i \neq j$, $\sigma_i^2 \neq \sigma_j^2$: il y a *hétérocédasticité* des résidus.

(2) Les résidus sont corrélés : pour un couple (i, j), $Cov(\varepsilon_i, \varepsilon_j) \neq 0$. Par exemple, ε a une structure de processus.

On supposera toujours que β est identifiable ; $\mathcal{E}_X$ dénote l'espace de la moyenne.

4.1 Σ est connu à un facteur près : les MCG

4.1.1 Réduction de (Ω) à un modèle sphérique $(\widetilde{\Omega})$

Supposons que $\Sigma = \sigma^2 R$ est inversible et connue à un facteur près σ^2.

Décomposition spectrale de R.

R étant symétrique réelle et dp, elle admet la décomposition spectrale $R = OD^tO$: O est la matrice dont les colonnes sont les vecteurs propres de R, vecteurs choisis orthonormés $(O\,{}^tO = {}^tOO = I_n)$; D est la matrice diagonale des valeurs propres > 0.

Cette décomposition permet de définir une racine carrée de R, $R^{\frac{1}{2}} = OD^{\frac{1}{2}}$, $D^{\frac{1}{2}}$ étant la matrice diagonale des racines carrées des valeurs propres. Posant $R^{-\frac{1}{2}} = [R^{\frac{1}{2}}]^{-1}$, on définit le modèle $(\widetilde{\Omega}) \equiv R^{-\frac{1}{2}}(\Omega)$:

$$(\widetilde{\Omega}) : \widetilde{Y} = \widetilde{X}\beta + \widetilde{\varepsilon}, \text{ avec } \widetilde{U} = R^{-\frac{1}{2}}U \text{ pour } U = Y, X, \varepsilon$$

Proposition 12 *$(\widetilde{\Omega})$ est un modèle observable à résidus sphériques.*

En effet, $Cov(\widetilde{\varepsilon}) = R^{-\frac{1}{2}} cov(\varepsilon)\, {}^tR^{-\frac{1}{2}} = \sigma^2 I_n$. $(\widetilde{\Omega})$ est donc un *modèle linéaire standard,* observable puisque R est connue. $(\widetilde{\Omega})$ est gaussien si (Ω) l'est. La bonne estimation de β dans (Ω) est celle des MCO dans $(\widetilde{\Omega})$: nous l'appellerons *estimation des moindres carrés généralisés* (MCG). Si R est diagonale, on parle d'estimation par moindres carrés pondérés (MCP).

4.1.2 L'estimation des MCG

Calculant l'EMCO dans $(\widetilde{\Omega})$ et remarquant que $\widetilde{X} = R^{-\frac{1}{2}}X$, on obtient :

Proposition 13 *L'estimation des MCG de β vaut :*

$$\widehat{\beta}_{MCG(\Omega)} = \widehat{\beta}_{MCO(\widetilde{\Omega})} = ({}^t\widetilde{X}\widetilde{X})^{-1}\ {}^t\widetilde{X}\widetilde{Y} = ({}^tXR^{-1}X)^{-1}\ {}^tXR^{-1}Y \tag{4.2}$$

Interprétation géométrique.

R^{-1} étant symétrique et dp, $\|Z\|^2_{R^{-1}} = {}^tZR^{-1}Z$ définit une norme sur $\mathbf{R}^n$ (la norme euclidienne $\|.\|_2$ correspond à $R = I_n$), deux vecteurs u et v étant R^{-1}-orthogonaux si ${}^tuR^{-1}v = 0$.

L'estimation des MCG minimise $SCR(\beta) = \|Y - X\beta\|^2_{R^{-1}} = \left\|\widetilde{Y} - \widetilde{X}\widetilde{\beta}\right\|^2_2 : \widehat{Y}_\Omega = X\widehat{\beta}_{MCG}$ est donc la projection R^{-1}-orthogonale de Y sur $\mathcal{E}_X$. Cet opérateur de projection P_R s'écrit :

$$P_R = X({}^tXR^{-1}X)^{-1}\ {}^tXR^{-1}$$

Proposition 14 *Propriété de l'estimation des MCG*

(1) $\widehat{\beta} = ({}^tXR^{-1}X)^{-1}\ {}^tXR^{-1}Y$ estime sans biais β ; $Var(\widehat{\beta}) = ({}^tX\Sigma^{-1}X)^{-1}$.

(2) Gauss-Markov : $\widehat{\beta}$ est le meilleur estimateur linéaire et sans biais de β.

(3) $\widehat{Y} = X\widehat{\beta}$ est la projection R^{-1}-orthogonale de Y sur l'espace $\mathcal{E}_X$.

(4) $\widehat{\sigma}^2 = \frac{1}{n-p}SCR(\Omega)$ estime sans biais σ^2, avec $SCR(\Omega) = \left\|Y - X\widehat{\beta}\right\|^2_{R^{-1}}$.

(5) Si Y est gaussienne : (i) les MCG sont optimaux parmi les estimateurs sans biais ; pour β, ils coïncident avec le MV ; (ii) $\widehat{\beta}$ est gaussienne ; (iii) $\widehat{\sigma}^2 \sim \frac{\sigma^2}{n-p}\chi^2_{n-p}$; (iv) $\widehat{\beta}$ et $\widehat{\sigma}^2$ sont indépendants.

Preuve :
La densité des observations gaussiennes y est :

$$L_n(y;\beta,\sigma^2) = \{2\pi\sigma^2\}^{-n/2}\det(R)^{-\frac{1}{2}}\exp\{-\frac{1}{2\sigma^2}\ {}^t(Y - X\beta)R^{-1}(Y - X\beta)\}$$

Minimiser ${}^t(Y - X\beta)R^{-1}(Y - X\beta) = \|Y - X\beta\|^2_{R^{-1}}$ en β revient à maximiser la vraisemblance : MCG et MV coïncident pour l'estimation de β. Les autres propriétés découlent de la réduction $(\Omega) \mapsto (\widetilde{\Omega})$ et des propriétés du modèle linéaire standard $(\widetilde{\Omega})$. .....□

Remarques.

(1) Soit (ω) une sous-hypothèse linéaire sur $E(Y)$; la loi de la statistique de Fischer est inchangée *à condition* de définir les SCR pour la norme $\|.\|^2_{R^{-1}}$.

(2) $\widetilde{\sigma}^2 = \frac{1}{n-p}\left\|Y - X\widehat{\beta}\right\|^2_2$($\widehat{\beta}$ est l'EMCG) est biaisé de σ^2 si les résidus sont non-sphériques. De même, la statistique F associée aux SCR calculées pour la norme $\|.\|_2$ ne suit pas une loi de Fisher.

(3) Il y a deux façons de calculer l'estimation par MCG : (i) soit on réduit le modèle (Ω) et on effectue les MCO dans $(\widetilde{\Omega})$, $R^{-\frac{1}{2}}$ s'obtenant à partir de la décomposition spectrale de R ; (ii) soit on utilise les formules matricielles donnant l'EMCG et les SCR sont calculées pour la norme $\|.\|_{R^{-1}}$.

(4) La décomposition spectrale de R peut être coûteuse si n est grand. Dans certain cas, la transformation $(\Omega) \mapsto (\widetilde{\Omega})$ s'explicite facilement. C'est le cas des modèles à résidus AR (cf. infra).

(5) L'estimation $\widehat{\beta}_{MCO} = ({}^tXX)^{-1}\,{}^tXY$ reste une bonne estimation, sans biais. Sa variance vaut :

$$Var(\widehat{\beta}_{MCO}) = \sigma^2({}^tXX)^{-1}\,{}^tXRX\ ({}^tXX)^{-1} \geq Var(\widehat{\beta}_{MCG})$$

4.1.3 Exemple : les moindres carrés pondérés

C'est la méthode d'estimation associée à la situation R diagonale et connue :

$$(\Omega) : \left\{ \begin{array}{l} y_i = {}^t x_i\beta + \varepsilon_i, i = 1, n \text{ avec } E(\varepsilon_i) = 0 \\ E(\varepsilon_i\varepsilon_j) = 0 \text{ si } i \neq j,\ Var(\varepsilon_i) = \sigma^2 a_i^2, \text{ avec } a_i > 0 \text{ connu} \end{array} \right.$$

$(\widetilde{\Omega})$ s'obtient en multipliant chaque équation de (Ω) par a_i^{-1} :

$$(\widetilde{\Omega}) : \widetilde{y}_i = \frac{y_i}{a_i} = {}^t(\frac{x_i}{a_i})\beta + \widetilde{\varepsilon}_i,\ i = 1, n$$

Pour la régression : $y_i = ax_i + \varepsilon_i$, $i = 1, n$, $x_i \in \mathbf{R}$, l'estimateur des MCP de a est :

$$\widehat{a} = \{\sum_i \frac{x_i y_i}{a_i^2}\}/\{\sum_i \frac{x_i^2}{a_i^2}\}, \text{ avec } Var(\widehat{a}) = \sigma^2/\{\sum_i \frac{x_i^2}{a_i^2}\}$$

Si $Var(y_i) = \sigma^2 x_i^2 > 0$, l'EMCP est $\widehat{a} = \frac{1}{n}\sum_i \frac{y_i}{x_i}$: c'est l'estimateur du ratio de a.

4.2 Σ est inconnue : MCO et MCQG

4.2.1 Propriétés des MCO

Commençons par récapituler les propriétés des MCO.

• *Ce que permettent les MCO.*

(i) l'estimateur des MCO, $\beta^* = ({}^tXX)^{-1}\,{}^tXY$ est *sans biais*. Il présente l'avantage de ne pas nécessiter la connaissance de Σ, la covariance des résidus.

Proposition 15 *Propriété des MCO*
(1) β^* estime sans biais β et $Var(\beta^*) = \sigma^2\Delta$, où $\Delta = ({}^tXX)^{-1}\,{}^tX\Sigma X\ ({}^tXX)^{-1}$.
(2) Si le modèle est gaussien, $\beta^* \sim \mathcal{N}_p(\beta, \sigma^2\Delta)$.

(ii) On va voir que, sous de bonnes conditions, l'EMCO est convergente.

• *Ce que ne permettent pas les MCO.*

(i) L'estimateur $\sigma^{*2} = \frac{1}{n-p}\|Y - X\beta^*\|_2^2$ déduit des MCO est biaisé ; en effet, $\|Y - X\beta^*\|_2^2$ $=^t \varepsilon(I - P)\varepsilon = Trace\{(I - P)\varepsilon^t\varepsilon\}$, où P est le projecteur orthogonal sur $\mathcal{E}_X$. Il s'en suit que $E\|Y - X\beta^*\|_2^2 = Trace\{(I - P)\Sigma\} \neq (n - p)\sigma^2$ si les résidus sont non-sphériques.

(ii) La statistique F associée aux MCO et aux SCR calculées pour la norme euclidienne $\|.\|_2$ ne suit pas une loi de Fischer en situation non-sphérique.

(iii) β^* n'est pas indépendante de l'estimation $\widehat{\sigma}^2$ déduite des MCG : en effet, $Y^* = X\beta^*$ est la projection orthogonale de Y sur $\mathcal{E}_X$ et $\widehat{\sigma}^2 = \frac{1}{n-p}\left\|Y - \widehat{Y}\right\|_{R^{-1}}^2$. Mais Y^* et $(Y - \widehat{Y})$ ne sont plus orthogonaux puisque $\widehat{Y}$ n'est plus la projection orthogonale de Y sur $\mathcal{E}_X$. Il n'est pas possible de construire simplement un test de Student ou de Fischer sur la base de β^* (l'EMCO) et de l'estimation sans biais $\widehat{\sigma}^2$ (MCG).

(iv) L'estimation des MCO n'est pas optimale.

4.2.2 Convergence des MCO

Sous de bonnes conditions asymptotiques sur X et sur Σ, l'estimation des MCO est convergente. Si A est une matrice symétrique sdp, on notera $\lambda_M(A)$ (resp. $\lambda_m(A)$) la plus grande (resp. la plus petite) valeur propre de A.

Proposition 16 *L'estimateur β^* des MCO est convergent si :*
(i) $\lambda_M(\Sigma)$ est bornée en n.
(ii) $\lambda_m({}^tXX) \longrightarrow \infty$ pour $n \to \infty$.

Preuve :
β^* étant sans biais, il suffit de vérifier que la trace de $Var(\beta^*)$ tend vers 0. Or, utilisant l'identité Trace(AB) =Trace(BA), on a pour $A = ({}^tXX)^{-1}\,{}^tX$ et $B = \Sigma X({}^tXX)^{-1}$:

$$\begin{aligned}\text{Trace}(Var(\beta^*)) &= \text{Trace}\{({}^tXX)^{-1}\,{}^tX\Sigma X({}^tXX)^{-1}\} = \text{Trace}\{\Sigma X({}^tXX)^{-2}\,{}^tX\}\\ &\leq \lambda_M(\Sigma)\text{Trace}\{X({}^tXX)^{-2}\,{}^tX\} \text{ (cf. lemme ci-dessous)}\\ &= \lambda_M(\Sigma)\text{Trace}\{({}^tXX)^{-1}\} \leq p\lambda_M(\Sigma)/\lambda_m({}^tXX) \underset{n\to\infty}{\longrightarrow} 0\end{aligned}$$

Lemma 17 *Si Σ et V sont deux matrices réelles $n \times n$, symétriques et dp, alors :*
$$Trace(\Sigma V) \leq \lambda_M(\Sigma)Trace(V).$$

Preuve du lemme :
Soit ${}^tO\Sigma O = D$, O orthogonale et D diagonale, la décomposition spectrale de Σ. Posant $Q = {}^tOVO$, $Trace(\Sigma V) = Trace({}^tO\Sigma O\,{}^tOVO) = Trace(DQ)$. On obtient :
$$Trace(\Sigma V) = \textstyle\sum_i d_{ii}q_{ii} \leq (\max_i d_{ii})\sum_i q_{ii} = \lambda_M(\Sigma)Trace(Q)$$
Q et V ayant même trace, on en déduit le résultat annoncé. ..... □

Cette convergence des MCO est importante : elle permet en effet de se rapprocher des MCG lorsque Σ est de forme paramétrique connue.

Nous allons voir que la condition (i) sur Σ est vérifiée pour la classe des processus *stationnaires au second ordre à covariance sommable.*

4.2.3 Résidus stationnaires à covariance sommable

Definition 1 $(Z_t)_{t\in\mathbf{N}}$ *est un processus réel stationnaire au second ordre si :*
(i) pour tout $t \geq 0$, $E(Z_t)$ existe et est constant : $E(Z_t) = m$.
(ii) pour tout $s, t \geq 0$, $Cov(Z_s, Z_t)$ existe et ne dépend que de $|t - s|$:
$$\forall s, t \in \mathbf{N},\ Cov(Z_t, Z_s) = \gamma(|t - s|) < \infty$$

La covariance est sommable si de plus $\sum_{h\geq 0}|\gamma(h)| < \infty$. (ε_t) est un résidu stationnaire au second ordre si les variables ε_t sont centrées.

Un premier exemple de résidu stationnaire au second ordre est donné par un bruit blanc de carré intégrable.

Example 14 *Processus $AR(1)$*

L'étude des modèles Auto-Régressifs (AR) (plus généralement celle des $ARMA$) est un domaine important de la modélisation statistique, celui des *séries chronologiques.* Nous n'aborderons pas ce sujet ici, nous contentant de présenter le modèle $AR(1)$.

$\varepsilon = (\varepsilon_t)_{t\geq 0}$ est un résidu $AR(1)$ stationnaire au second ordre s'il est centré et si sa fonction de covariance est donnée, pour un paramètre $|\rho| < 1$, par :
$$\forall h \in \mathbf{Z},\ \gamma(h) = \sigma^2\rho^{|h|}$$

Un tel processus existe : en effet, considérons $\eta = (\eta_t)_{t\in\mathbf{Z}}$ un bruit blanc de carré intégrable, $\sigma_\eta^2 = Var(\eta_t)$, et $|\rho| < 1$. On montre facilement que les $\varepsilon_t = \sum_{s\geq 0}\rho^s\eta_{t-s}$ sont bien définies dans L^2 (l'espace des variables de carré intégrable) et on vérifie que ε est un $AR(1)$ de paramètre ρ et de variance $\sigma^2 = \sigma_\eta^2\sum_{s\geq 0}\rho^{2s} = \frac{\sigma_\eta^2}{1-\rho^2}$.

Une conséquence de cette construction montre que ε vérifie l'*équation d'auto-régression* : $\forall t \geq 0$, $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$. D'où la terminologie $AR(1)$: régressé sur son passé, ε_t ne dépend que de $\rho\varepsilon_{t-1}$ au bruit blanc d'innovation près η_t.

Les processus $AR(1)$ sont clairement à covariance sommable.

Proposition 18 *Si ε est un processus au second ordre tel que $\sup_h\{\sum_i|\gamma_{i,i+h}|\} < \infty$, alors la condition (i) de la proposition 4.2.2 est satisfaite.*

Preuve : Il suffit de remarquer que si $\Sigma = (\gamma_{i,j})$ est une matrice $n \times n$ diagonalisable, et si λ est une valeur propre, alors $|\lambda| \leq \sup_i\sum_j|\gamma_{i,j}|$. ..... □

4.2.4 Cas où Σ est de forme paramétrique connue : les MCQG

Une covariance $AR(1)$ s'exprime facilement à partir du paramètres inconnu ρ. Plus généralement, si $\Sigma = \Sigma(\phi)$ est une forme paramétrique connue pour Σ, $\phi \in \mathbf{R}^q$ étant un paramètre inconnu, l'estimation des *moindres carrés quasi généralisés* (MCQG) est définie par l'algorithme suivant :

Algorithme des MCQG

(1) Estimer β par β^*, l'estimation des MCO.
(2) On obtient les résidus estimés par MCO : $\varepsilon^* = Y - X\beta^*$.
(3) Sur la base de ε^*, estimer ϕ^* par une "bonne" méthode.
(4) Σ est alors estimée par $\Sigma^* = \Sigma(\phi^*)$.
(5) L'estimateur $\widehat{\beta}_{MCQG}$ des MCQG n'est autre que l'EMCG pour Σ^* :

$$\widehat{\beta}_{MCQG} = \widehat{\beta}_{MCG}(\Sigma(\phi^*))$$

Il est naturel de penser que si l'estimation de ϕ est convergente[1], les MCQG vont s'approcher des MCG. Sous de bonnes conditions, ce résultat est vrai ([2]) :

$$\sqrt{n}(\widehat{\beta}_{MCQG} - \beta) \xrightarrow{loi} \mathcal{N}_p(0, \lim_{n\to\infty} n({}^tX\Sigma^{-1}X)^{-1})$$

La matrice de covariance asymptotique est celle des MCG. Les MCQG sont donc asymptotiquement sans biais et efficaces dans la classe des estimateurs linéaires sans biais.

Les MCQG pour des résidus $AR(1)$

Examinons cette procédure lorsque le résidu est un $AR(1)$ stationnaire :

$$\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t, \text{ où } |\rho| < 1 \text{ et } \eta \text{ bruit blanc} \tag{4.3}$$

ρ, la corrélation à distance 1, étant estimée par la corrélation empirique :

$$\widehat{\rho} = \frac{\sum_{t=1}^{n-1} \varepsilon_t\varepsilon_{t+1}}{\sum_{t=1}^{n} \varepsilon_t^2}$$

Si η est un bruit blanc gaussien, on a le résultat :

Proposition 19 *Si les résidus sont $AR(1)$, gaussiens et stationnaires, alors : $\widehat{\rho} \xrightarrow{\Pr} \rho$.*

Cette estimation $\widehat{\rho}$ suppose que les résidus soient observés. Comme ils ne le sont pas, on estime ρ sur la base des résidus ε^* des MCO : si les MCO pour β sont convergents, à nouveau $\rho^* = \widehat{\rho}(\varepsilon^*) \xrightarrow{\Pr} \rho$. Les MCQG seront alors les MCG pour $\Sigma^* = \Sigma(\rho^*)$.
Prédiction : supposons que le modèle (**??**) est observé aux instants $t = 1, n$ et que ρ connue. Pour une nouvelle condition exogène x_{n+1}, comment prédire $E(y_{n+1})$? On obtient la réponse en passant par le modèle réduit et en utilisant le résultat de prédiction pour le modèle linéaire standard.

Proposition 20 *(1) La prédiction linéaire optimale et sans biais de $E(y_{n+1})$ est $\widehat{y}_{n+1} = {}^tx_{n+1}\widehat{\beta} + \widehat{\varepsilon}_{n+1}$, où :*
(i) $\widehat{\beta}$ est l'estimation des MCG (des MCQG si ρ est estimé) de β sur la base des n premières observations, et,
(ii) $\widehat{\varepsilon}_{n+1} = \rho\widehat{\varepsilon}_n$ où $\widehat{\varepsilon}_n = y_n - {}^tx_n\widehat{\beta}$.
(2) La prédiction optimale à l'horizon $h > 0$ est : $\widehat{y}_{n+h} = {}^tx_{n+h}\widehat{\beta} + \rho^h\widehat{\varepsilon}_n$.

[1] Par exemple, on s'assure que les MCO de β sont convergents, puis on utilise une méthode convergente (MV, ou méthode de moment) pour estimer le paramètre du modèle de ε.

[2] Amemiya, *General least squares with estimated auto-covariance matrix*, 1973, Econometrica 41, 723-732.

Preuve :
Comme la transformation de $y_t - \rho y_{t-1}$ fait passer à un modèle linéaire standard, la prédiction optimale pour $\widetilde{y}_{n+1}$ est :

$$\widehat{\widetilde{y}}_{n+1} = \widehat{y}_{n+1} - \rho y_n = {}^t(x_{n+1} - \rho x_n)\widehat{\beta}$$

Il suffit alors de résoudre en $\widehat{y}_{n+1}$ et d'observer que $\widehat{\varepsilon}_n = (y_n - {}^t x_n \widehat{\beta})$ pour obtenir la prédiction optimale à l'horizon 1. Le résultat à l'horizon $h > 0$ s'obtient par récurrence ; en particulier, la prédiction optimale du résidu ε_{n+h} est $\rho^h \widehat{\varepsilon}_n$. .....□

4.3 Régressions empilées ou modèle S.U.R.

On parle de régressions empilées pour la donnée d'un ensemble de M *régressions simultanées* expliquant M variables endogènes. Les paramètres de chaque régression sont autonomes (d'où le qualificatif *unrelated*) mais les résidus entre les régressions sont corrélés (d'où le qualificatif *seemingly*[3]). Donnons un exemple.

4.3.1 Un exemple : consommations simultanées de gaz et d'électricité

On s'intéresse aux consommations *simultanées* d'électricité Y_{1i} et de gaz Y_{2i} de n foyers : il y a $M = 2$ endogènes à expliquer, Y_1 et Y_2, chacune s'expliquant à partir d'exogènes, x_1 pour Y_1, x_2 pour Y_2 :

$$(\Omega) : \left\{ \begin{array}{l} Y_{1t} = {}^t x_{1t}\beta_1 + \varepsilon_{1t} \\ Y_{2t} = {}^t x_{2t}\beta_2 + \varepsilon_{2t} \end{array} \right. , t = 1, n$$

Les deux régressions *semblent* sans relations, mais :
(i) d'une part les deux consommations d'un même foyer sont corrélées ;
(ii) les 2 consommations sont probablement de variances différentes.

Les deux régressions ne sont donc qu'*apparemment* sans relations puisque, posant $\varepsilon_t = {}^t(\varepsilon_{1t}, \varepsilon_{2t})$:

$$\text{les } (\varepsilon_t) \text{ sont i.i.d. et } Var(\varepsilon_t) = \left(\begin{array}{cc} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{array} \right) = \Gamma$$

Pour l'observation totale $Y = {}^t({}^t Y_1, {}^t Y_2)$ de dimension $2n$, la covariance de Y est une matrice $2n \times 2n$ qui admet une structure particulière :

$$\Sigma = Cov(Y) = Cov(\varepsilon) = \Gamma \otimes I_n$$

où $A \otimes B$ est le produit de Kronecker de A et B : si $A = (a_{ij})$ est une matrice $p \times q$ et $B = (b_{kl})$ une matrice $r \times s$, $A \otimes B$ est la matrice $pr \times qs$ caractérisée par sa décomposition en $p \times q$ blocs, chacun de taille $r \times s$:

$$A \otimes B = \left(\begin{array}{cccc} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{array} \right)$$

$\Sigma = Cov(\varepsilon) = \Gamma \otimes I_n$ n'est pas sphérique si $\sigma_{12} \neq 0$ ou si $\sigma_1^2 \neq \sigma_2^2$. Cependant, Σ ayant une forme paramétrique connue $\Sigma(\sigma_1^2, \sigma_2^2, \sigma_{12})$, les MCG ou les MCQG sont possibles.

[3] S.U.R. signifie Seemingly Unrelated Regressions. Ces modèles ont été introduits par A. Zellner , *An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias*, Jour. American Stat. Ass., 1962, 57, 348-368.

4.3.2 Le modèle S.U.R. général

Supposons que les observations sont temporelles. A chaque instant, on observe M endogènes Y_j ; chaque Y_j suit son propre modèle de régression mais les résidus instantanés de ces régressions corrélés. Le modèle s'écrit :

$$(\Omega) : \left\{ \begin{array}{c} (\Omega_j) : Y_{jt} = {}^t x_{tj}\beta_j + \varepsilon_{tj},\ t = 1, n,\ \beta_j \in \mathbf{R}^{p_j},\ j = 1, M \\ \varepsilon(t) = {}^t(\varepsilon_{t1}, \varepsilon_{t2}, \cdots, \varepsilon_{tM}) \text{ sont i.i.d centrés, } Var(\varepsilon(t)) = \Gamma \end{array} \right.$$

Γ traduit la simultanéité des équations. L'écriture matricielle globale des $M \times n$ observations $Y = {}^t({}^tY(1), {}^tY(2), \cdots, {}^tY(M))$ est :

$$Y = \begin{pmatrix} X_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & X_2 & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{0} & X_M \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{pmatrix} + \varepsilon, \text{ où } \Sigma = Cov(\varepsilon) = \Gamma \otimes I_n$$

Trois situations se présentent :

1^{er} Cas : homogénéité et non-corrélation des ε.

$\Sigma = \sigma^2 I_{nM}$: les régressions sont décorrélées et les résidus sont de même variance σ^2. β_j est estimé dans (Ω_j)[4] et σ^2 l'est à partir des $SCR(\Omega_j)$ cumulées ; notant $p = \sum_j p_j$, on obtient :

$$\widehat{\beta}_j = ({}^tX_jX_j)^{-1}\ {}^tX_jY(j),\ \widehat{\sigma}^2 = \frac{1}{nM - p} SCR(\Omega) \text{ où } SCR(\Omega) = \sum_{j=1,M} SCR(\Omega_j)$$

$2^{ème}$ Cas : hétérocédasticité.

Σ est diagonale mais elle n'est pas proportionnelle à l'identité. Les M régressions sont décorrélées mais avec des variances résiduelles propres : il y a *hétérocédasticité* des endogènes. L'estimation des β_j reste inchangée. Par contre, chaque variance résiduelle est estimée séparément dans (Ω_j) par $\widehat{\sigma}_j^2 = SCR(\Omega_j)/(n - p_j)$.

Test d'homogénéité des résidus : (ω) $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_M^2 = \sigma^2$.

Dans le cas gaussien, la statistique du rapport de vraisemblance (RV) est bien adaptée à ce test (cf. § 14.8) : la log-vraisemblance d'un modèle linéaire gaussien (M) valant, à une constante près, $-\frac{n}{2}\log\widehat{\sigma}^2(M)$, la statistique du RV s'écrit :

$$\Delta = 2\log\frac{L_n(\Omega)}{L_n(\omega)} = 2(l_n(\Omega) - l_n(\omega)) = -\ n \sum_{j=1,M} \log\frac{\widehat{\sigma}^2}{\widehat{\sigma}_j^2}$$

Si les estimations de chaque σ_j^2 sont convergentes, Δ suit, sous (ω), un χ^2_{M-1}. Une meilleure approximation de la loi limite est obtenue en prenant les estimations débiaisées (encore notées $\widehat{\sigma}_j^2$) des variances résiduelles et $\Delta^* = \sum_{j=1}^M (n - p_j)\log\frac{\widehat{\sigma}^2}{\widehat{\sigma}_j^2}$.

Remarques.

(1) Ce résultat se maintient pour la comparaison de M régressions indépendantes observées sur des échantillons de tailles différentes $n_j, j = 1, M$. La statistique du test vaut alors :

$$\Delta^* = \sum_{j=1,M} (n_j - p_j)\log\frac{\widehat{\sigma}^2}{\widehat{\sigma}_j^2}$$

(2) En présence de $M = 2$ endogènes et si le modèle est gaussien, on dispose d'un test exact non-asymptotique : si $\widehat{\sigma}_i^2$ est l'estimation sans biais de σ_i^2, $F = \frac{\widehat{\sigma}_1^2}{\widehat{\sigma}_2^2} \sim F(n_1 - p_1, n_2 - p_2)$ si $\sigma_1^2 = \sigma_2^2$.

[4] Dans tout système de régressions empilées, β_j peut être estimée séparément dans (Ω_j). Mais ici, comme dans le deuxième cas, cette estimation est optimale.

3ème Cas : Cas général, $\Sigma = \Gamma \otimes I_n = \sigma^2 R$**,** $\Gamma = \sigma^2 \Lambda$.

• Si Γ est connue, R l'est aussi, d'inverse $R^{-1} = \Lambda^{-1} \otimes I_n$[5]. L'EMCG est :

$$\begin{aligned} \widehat{\beta} &= {}^t(\widehat{\beta}_1, \cdots, \widehat{\beta}_M) = [{}^t X(\Lambda^{-1} \otimes I_n)^{-1} X]^{-1}\ {}^t X(\Lambda^{-1} \otimes I_n)^{-1} Y \\ \widehat{\sigma}^2 &= \frac{1}{Mn - p} \left\| Y - X\widehat{\beta} \right\|^2_{\Lambda^{-1} \otimes I_n} \end{aligned}$$

L'inversion de R n'utilise que l'inversion de Λ, matrice de petite dimension.

• Si Λ est inconnue, on estime chaque β_j par MCO dans (Ω_j) ; on en déduit les résidus estimés $\widehat{\varepsilon}_t$ et l'estimation $\widehat{\Gamma} = \frac{1}{n} \sum_i \widehat{\varepsilon}_t\ {}^t\widehat{\varepsilon}_t$. On utilise alors les MCQG pour $\widehat{\Gamma}$.

4.4 Exercices : Résidus non sphériques et MCG

Exercise 20 *MCP pour un modèle de variance résiduelle $\sigma^2 z_i$, z_i connus*

On considère modèle expliqué par les exogènes $x_i \in \mathbf{R}$ et $z_i > 0$:

$$y_i = a x_i + e_i,\ Var(e_i) = \sigma^2 z_i,\ i = 1, n,\ Cov(e_i, e_j) = 0 \text{ si } i \neq j$$

(1) Déterminer l'EMCO et l'EMCP de a, leurs variances, et l'estimation sans biais de σ^2. Tester $a = 0$ contre $a > 0$. Examiner les deux cas particuliers (i) $z_i = x_i^2$; (ii) $z_i = x_i$.

(2) Vérifier que l'estimation de σ^2 déduite des MCO de façon standard est biaisée (examiner le cas $n = 2$).

Exercise 21 *Régression poissonienne*

Le nombre de pannes $Y_i \in \mathbf{N}$ d'un équipement d'âge x_i est une variable de Poisson $\mathcal{P}(a x_i)$ de paramètre $a x_i$. On dispose de n observations indépendantes (y_i, x_i).

(1) Ecrire le modèle linéaire associé. Donner l'EMCO de a et sa variance.

(2) Donner l'EMCP de a. Le comparer à l'EMCO. Quelle est l'EMV de a ?

Exercise 22 *Transformation normalisante, MCO et MCG*

Soit le modèle à $2n$ observations à résidus centrés et décorrélés pour des t différents :

$$(\Omega) : \left\{ \begin{array}{l} y_{1t} = a x_{1t} + e_{1t} \\ y_{2t} = b x_{2t} + e_{2t} \end{array} \right. , a \text{ et } b \in \mathbf{R},\ t = 1, n,\ Cov \begin{pmatrix} e_{1t} \\ e_{2t} \end{pmatrix} = \sigma^2 \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

(1) Déterminer c tel que y_{1t} soit décorrélée de $y'_{2t} = y_{1t} + c y_{2t}$. Pour ce c, déterminer d tel que $\{\widetilde{y}_t = (y_{1t}, \widetilde{y}_{2t} = d y'_{2t}),\ t = 1, n\}$ définisse un modèle sphérique $(\widetilde{\Omega})$.

(2) Expliciter l'écriture matricielle de (Ω) et de $(\widetilde{\Omega})$. En déduire l'EMCO et l'EMCG de (a, b). Comparer leurs variances.

Exercise 23 *Test d'homogénéité des variances résiduelles*

On considère le modèle d'analyse de la covariance :

$$(\Omega) : y_{ij} = a_i + b_i x_{ij} + e_{ij},\ j = 1, n_i,\ i = 1, 4,\ e_{ij} \sim \mathcal{N}(0, \sigma_i^2), \text{ indépendants}$$

On note (Ω_H) le sous-modèle avec homogénéité des 4 variances. On observe :

n_i	20	18	20	22
$SCR(\Omega_i)$	27	22	110	18

(1) Tester l'homogénéité des 4 variances résiduelles.

(2) Tester l'homogénéité $(\Omega_{1,2,4})$ des variances résiduelles des groupes $1, 2$ et 4. Tester l'homogénéité du groupe 3 avec les trois autres groupes.

(3) $(\omega_{1,2,4})$ est le sous-modèle de $(\Omega_{1,2,4})$ avec égalité des pentes : $b_1 = b_2 = b_4$. L'EMCO donne $SCR(\omega_{1,2,4}) = 84$. Tester $(\omega_{1,2,4})$.

[5] Si A et B sont des matrices carrées inversibles, l'inverse du produit de Kronecker $A \otimes B$ est $A^{-1} \otimes B^{-1}$.

Exercise 24 *Un système de deux régressions empilées*

Tester $a = b$ dans le modèle de régressions empilées,

$$(\Omega) : \left\{ \begin{array}{l} y_{1t} = ax_{1t} + e_{1t} \\ y_{2t} = bx_{2t} + e_{2t} \end{array} \right. , a \text{ et } b \in \mathbf{R}, \; t = 1, 20, \text{ avec } Cov \left(\begin{array}{c} e_{1t} \\ e_{2t} \end{array} \right) = \left(\begin{array}{cc} 1 & 1 \\ 1 & 2 \end{array} \right)$$

sachant qu'on a observé la matrice des moments croisés d'ordre 2 :

	y_1	y_2	x_1	x_2
y_1	10	-1	1	-1
y_2		15	5	1
x_1			11	1
X_2				1

Chapitre 5

Choix et validation de modèle

Voilà quelques questions auxquelles nous allons chercher à répondre :

- Comment choisir entre plusieurs modèles ?
  - Comment valider un modèle sélectionné.
  - Si le modèle n'est pas valide, quelles sont les erreurs de spécification ?
  - Comment remédier à ces erreurs et proposer un modèle alternatif ?

Par souci de simplicité, nous supposerons le modèle à résidus gaussiens et sphériques. Si la covariance résiduelle $\Sigma = \sigma^2 R$ n'est pas sphérique, mais si R est connue (resp. estimable), les résultats que nous allons donner pour la situation sphérique ($\Sigma = \sigma^2 I$) se maintiendront à condition d'utiliser les estimations des MCG en remplacement des MCO (resp. des MCQG) et de calculer les sommes des carrés résiduels pour la norme $\|.\|_{R^{-1}}$ (resp. $\|.\|_{\widehat{R}^{-1}}$) en remplacement de la norme euclidienne $\|.\|_2$.

5.1 Choix de modèle

5.1.1 Critère de parcimonie

M régresseurs $\{X_1, X_2, \cdots, X_M\}$ sont à notre disposition pour expliquer Y : comment choisir "convenablement" un sous-ensemble $P \subseteq \{1, 2, \cdots, M\}$ de régresseurs $\{X_i, i \in P\}$ expliquant Y ? Ce choix doit répondre à deux objectifs contradictoires :

1. P doit être de taille assez réduite pour que le modèle soit facilement interprétable.
2. P doit être assez grand pour que l'ajustement de Y soit correct.

Le choix doit donc être *parcimonieux* (peu de paramètres) mais fournissant un *bon ajustement* (des paramètres en nombre suffisant). Présentons quelques critères classiques de choix de modèle.

La statistique C_p de Mallows ; le graphique de $p \mapsto \widehat{\sigma}^2_p(p)$

Si le modèle considéré (P) admet p paramètres, c'est la statistique

$$C_p = \frac{SCR(P)}{s^2} - (n - 2p)$$

où $SCR(P)$ est la somme des carrés résiduels dans (P), s^2 est l'estimation de la variance dans le modèle incluant toutes les variables explicatives, et n est le nombre d'observations. Si (P) est sans biais, $E(SCR(P)) = (n-p)\sigma^2$, de même $E(s^2) = \sigma^2$ et donc $E(C_p) \approx p$: sous un modèle sans biais et si n est assez grand, $C_p \approx p$; au contraire, si le modèle est biaisé, C_p prendra des valeurs plus grandes que p. On trace donc sur un graphique

les points (p, C_p) et on conclut en retenant une petite valeur de p où C_p est voisin de p. Remarquons que du fait du caractère aléatoire, des valeurs de C_p peuvent être observée sous la droite $C_p = p$.

Il est aussi instructif de tracer la courbe $(p, \widehat{\sigma}^2_p)$ où $\widehat{\sigma}^2_p = \frac{SCR(P)}{n-p}$ est l'estimation de la variance résiduelle dans le modèle P. Le courbe s'approche de l'asymptote $\widehat{\sigma}^2_p = \sigma^2$ pour P grand, et on pourra retenir un modèle utilisant moins de variables dès que $\widehat{\sigma}^2_p$ sera proche du niveau de cette asymptote σ^2.

La log-vraisemblance pénalisée : critères AIC, BIC

La qualité de l'ajustement d'un modèle peut être évaluée par la log-vraisemblance $l_n(P)$ sous (P). Là aussi, $l_n(P) \leq l_n(Q)$ si $P \subseteq Q$, et il faut pénaliser l_n par la dimension du modèle. Deux critères sont classiquement utilisés :

(1) Le *critère de AIC* de *Akaike* (Akaike Information Critérion) : P maximise $AIC(P) = 2l_n(P) - 2\dim(P)$.
(2) Le *critère BIC* (*Bayesian I.C.*) : P maximise $BIC(P) = 2l_n(P) - (\log n)\dim(P)$.

Ces critères sont très généraux et s'appliquent chaque fois que l'on sait expliciter la vraisemblance du modèle : modèle linéaire gaussien, modèles de régression Logit binaires ou polytomiques, modèles log-linéaire de table de contingence, les régressions non-linéaires gaussiennes, les modèles paramétriques de durée.

Pour un modèle linéaire gaussien, la log-vraisemblance vaut, à une constante près :

$$l_n(P) = -\frac{n}{2}\log(\widehat{\sigma}^2(P)) \text{ où } \widehat{\sigma}^2(P) = \frac{SCR(P)}{n},$$

et le critère AIC minimisera

$$AIC(P) = n\log(SCR(P)) + 2 \times \dim(P)$$

5.2 Validation d'un modèle linéaire gaussien

Une fois retenu un ensemble de régresseurs X expliquant Y, il faut tester la validité de ce choix $(\Omega) : Y \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$. Cette validité est à trois niveaux :

(1) $E(Y) = X\beta$: validité de la spécification de la moyenne.
(2) $Cov(Y) = \sigma^2 I_n$: validité de la covariance sphérique.
(3) Y est gaussienne : validité du modèle de loi gaussienne.

Ecrivant $Y = X\beta + \varepsilon$, la validation se fait sur la base des résidus estimés :

$$\widehat{\varepsilon}_\Omega = (I - P)Y \sim \mathcal{N}_n(0, \sigma^2(I - P))$$

Pour la décomposition orthogonale $\mathbf{R}^n = \mathcal{E}_X \oplus \mathcal{E}_X^\perp$, $\widehat{\varepsilon}_\Omega \in \mathcal{E}_X^\perp$ peut est représenté par ses $(n-p)$ coordonnées $(\widehat{e}_{p+1}, \widehat{e}_{p+2}, \cdots, \widehat{e}_n)$ dans une base orthonormale de $\mathcal{E}_X^\perp$[1]. La validité du modèle (Ω) se traduit par :

$$(\Omega) : \widehat{e}_{p+1}, \widehat{e}_{p+2}, \cdots, \widehat{e}_n \text{ est une suite } i.i.d. \text{ gaussienne}$$

En pratique, on utilise un *test approché* (Ω_A) s'appuyant sur le vecteur $\widehat{\varepsilon}$ écrit dans la base canonique de $\mathbf{R}^n$, et on teste :

$$(\Omega_A) : \widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \cdots, \widehat{\varepsilon}_p, \widehat{\varepsilon}_{p+1}, \cdots, \widehat{\varepsilon}_n \text{ sont } i.i.d. \text{ gaussiennes}$$

[1] Par exemple celle associée aux $n - p$ vecteurs propres associés à la valeur propre 0 de P.

5.2.1 La paramétrisation de la moyenne est-elle bonne ?

Une mauvaise paramétrisation de la moyenne peut avoir plusieurs origines : (a) trop de régresseurs ont été retenus : on dit qu'il y a surparamétrisation ; (b) il manque des régresseurs : il y a sousparamétrisation ; (c) plus généralement, certains régresseurs ont été oubliés, d'autres sont superflus.

Surparamétrisation et sousparamétrisation

(a) *Surparamétrisation* : pas d'oubli de variables mais certaines sont superflues.

$(\Omega) : E(Y) = (X(1), X(2))\beta$ avec $\beta = {}^t({}^t\beta(1), {}^t\beta(2))$ (modèle de travail)
$(\omega) : E(Y) = X(1)\beta(1)$ (vrai modèle, inconnu).

X est de dimension $n \times (p+q)$, le vrai paramètre est $\beta = {}^t({}^t\beta(1), {}^t\mathbf{0})$. $X(2)$ correspond aux q régresseurs superflus.

L'estimation des MCO dans (Ω) vérifie $\widehat{\beta}_\Omega \sim \mathcal{N}_{p+q}({}^t({}^t\beta(1), {}^t\mathbf{0}), \sigma^2({}^tXX)^{-1})$: $\widehat{\beta}_\Omega(1)$ estime sans biais $\beta(1)$ mais sa variance est supérieure à celle des MCO dans (ω). En effet, $Var(\widehat{\beta}_\Omega(1)) \succeq Var(\widehat{\beta}_\omega(1))$ résulte des deux constatations suivantes : (i) pour $X = (X(1), X(2))$, $\{({}^tXX)^{-1}\}_{11} = A^{-1}$, $A = {}^tX(1)X(1) - C$ où $C = {}^tX(1)X(2)\{{}^tX(2)X(2)\}^{-1}$ ${}^tX(2)X(1)$; (ii) si A et B sont deux matrices symétriques régulières de même taille $(A \succeq B) \Leftrightarrow (A^{-1} \preceq B^{-1})$.

D'autre part, $\widehat{\sigma}^2(\Omega) = \frac{1}{n-(p+q)} SCR(\Omega)$ estime aussi sans biais σ^2, mais $Var(\widehat{\sigma}^2(\Omega)) = \frac{2\sigma^2}{(n-(p+q)} \geq Var(\widehat{\sigma}^2(\omega)) = \frac{\sigma^2}{n-p}$.

La surparamétrisation ne biaise pas les estimateurs mais elle en diminue la précision.

(b) *Sousparamétrisation* : oubli régresseurs.

$(\omega) : E(Y) = X(1)\beta(1)$ (modèle de travail)
$(\Omega) : E(Y) = (X(1), X(2))\beta$, $\beta = {}^t({}^t\beta(1), {}^t\beta(2))$, avec $\beta(2) \neq \mathbf{0}$ (vrai modèle).

Les régresseurs oubliés sont $X(2)$: l'EMCO de $\beta(1)$ dans (ω) est biaisée, de biais $b = E(\widehat{\beta}(1)) - \beta(1) = {}^t\{X(1)X(1)\}^{-1}\ {}^tX(1)X(2)\beta(2)$, et l'eqm de $\widehat{\beta}(1)$ est

$$eqm(\widehat{\beta}(1)) = b^2 + Var(\widehat{\beta}(1))$$

Ce biais est acceptable si l'estimation dans (ω) diminue fortement la variance : comme en situation de presque colinéarité, on peut avoir intérêt à supprimer certains régresseurs.

(c) *Le cas général* : des régresseurs ont été oubliés, d'autres sont superflus. Ceci conduit à une estimation biaisée de β et de la variance résiduelle.

Tests de validation de la moyenne : $(\Omega)\, E(Y) = X\beta$.

(a) Si σ^2 *est connue*, $\|\widehat{\varepsilon}\|^2 \sim \sigma^2\chi^2_{n-p}$ sous (Ω), $\sim \sigma^2\chi'^2_{n-p}(\delta^2)$ sinon. Le paramètre de non-centralité $\delta^2 = \|E(Y) - X\beta\|^2$ dépend de l'alternative à (Ω).

La probabilité $P\{\chi'^2_{n-p}(\delta^2) > a\}$ étant une fonction croissante de δ^2 (cf. § 14.3), on prendra comme région de rejet de (Ω) : $R_\alpha = \{\|\widehat{\varepsilon}\|^2 > \sigma^2 q(n-p;\alpha)\}$.

Si σ^2 est inconnue mais est préalablement estimée par $\widehat{\sigma}^2_P \sim \sigma^2\chi^2_d$ indépendamment des observations donnant l'estimation de β, le test repose sur la statistique $F = \frac{d}{n-p}\frac{\|\widehat{\varepsilon}\|^2}{\widehat{\sigma}^2_P}$. F suit une loi $F(n-p, d)$ sous (Ω) ; sinon F suit une loi Fischer décentrée.

(b) *Valeurs aberrantes.* Certaines observations y_i peuvent ne pas répondre au modèle $E(y_i) = {}^tx_i\beta$. Pour les détecter, on examine les résidus réduits $r_i = \widehat{\varepsilon}_i/\widehat{\sigma}(\widehat{\varepsilon}_i)$: sous (Ω), ils suivent une loi de Student à $n-p$ ddl. La représentation graphique des points $\{(i, r_i), i = 1, n\}$ permet de détecter ces valeurs aberrantes.

Si on détecte $q > 0$ valeurs aberrantes $\{y_i, i \in I_0\}$, on peut corriger le modèle en attribuant un paramètre à chaque observation aberrante :

$$(\Omega_{I_0}) : E(y_i) = \mu_i \text{ si } i \in I_0, \text{ et } E(y_i) = {}^t x_i\beta \text{ sinon}$$

Si le modèle initial est de dimension p, le modèle étendu (Ω_{I_0}) est de dimension $p+q$. Le test d'absence de valeurs aberrantes repose alors sur la statistique de Fischer de $(\Omega_{I_0}) \subset (\Omega)$:

$$F = \frac{q}{n-(p+q)} \frac{SCR(\Omega) - SCR(\Omega_{I_0})}{SCR(\Omega_{I_0})}$$

Plus généralement, si les données aberrantes $\{y_i, i \in I_0\}$ suivent un modèle spécifique, on intégrera cette information dans un modèle étendu $(\Omega^*_{I_0})$: $E(Y_{I_0}) = Z\gamma$ et $E(Y_{I_0^c}) = X_{I_0^c}\beta$.

(c) Une troisième façon de tester la validité de (Ω) est de *plonger* (Ω) *dans un surmodèle* (Ω_E) peu contestable. On testera alors (Ω) dans (Ω_E). Par exemple, si une analyse de la variance dépend d'une modalité i associée à une variable $x_i \in \mathbf{R}$:

$$(\Omega) : E(y_{ij}) = a + bx_i,\ j = 1, n_i,\ i = 1, p$$

le modèle (Ω_E) naturel est le modèle complet d'analyse de la variance, $E(y_{ij}) = m_i$.

5.2.2 La covariance des résidus est-elle sphérique ?

L'*hétérocédasticité* ou la *corrélation des résidus* constituent les deux alternatives classiques à la sphéricité des résidus.

Homogénéité des résidus.

Supposons que l'ensemble $I = \{1, 2, \cdots, n\}$ des n observations est divisé en r classes connues $I_1 \cup I_2 \cup \cdots \cup I_r$, la variance étant constante sur chaque classe : $Var(y_i) = \sigma_k^2$ si $i \in I_k$, $k = 1, r$. L'homogénéité des variances se traduit par :

$$(H_0) : \sigma_1^2 = \sigma_2^2 = \cdots = \sigma_r^2 = \sigma^2$$

Si le modèle linéaire est gaussien et si le dispositif des X assure la convergence de $\widehat{\sigma}_k^2$ pour chaque classe, on utilisera la statistique du RV :

$$\Delta = \sum_{k=1,r} n_k \log \frac{\widehat{\sigma}^2}{\widehat{\sigma}_k^2} \underset{(H_0)}{\sim} \chi^2_{r-1}$$

Le *test de Goldfeld-Quandt* examine le cas particulier d'une partition en 2 groupes, $I = I_1 \cup I_2$; les paramètres de la moyenne étant estimés séparément sur chaque groupe, le quotient des estimations des variances résiduelles suivra *exactement* une loi de Fischer à $(n_1 - p, n_2 - p)$ ddl si $\sigma_1^2 = \sigma_2^2$ si le modèle linéaire est gaussien. La partition peut être construite afin d'avoir un test plus puissant : par exemple, si on pense que $Var(y_i) = \sigma^2 x_i^2$ où x_i est observable, on range les observations par valeurs croissantes de x_i^2, I_1 correspondant aux premières observations et I_2 aux dernières. Si nécessaire, on peut intercaler une zone de séparation entre I_1 et I_2.

Non-corrélation des résidus

On va présenter deux tests de non-corrélation des résidus : l'un est *paramétrique*[2], c'est le test de Durbin-Watson ; l'autre est *non-paramétrique*, c'est le test des séquences[3].

[2] Un test est paramétrique si le calcul de sa loi utilise explicitement la loi du modèle sous-jacent : le test de Student d'égalité des moyennes est paramétrique, utilisant le caractère gaussien des deux populations.

[3] Un test est non-paramétrique si le calcul de sa loi ne nécessite pas la connaissance explicite de la loi du modèle sous-jacent. Par exemple, le test de signe pour la comparaison de deux moyennes est un test non-paramétrique.

Le test des séquences Supposons que l'on veuille tester le caractère *i.i.d.* d'une suite à valeur dans E :

$$(H_0) : \varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n \text{ sont des v.a. i.i.d. à valeur dans } E$$

Si $E = A \cup B$ est une partition fixée de E, on résume la suite des ε par la suite (s_i) des signes + ou de −, $s_i = +$ si $\varepsilon_i \in A$, ou $s_i = -$ sinon. Par exemple, si ε est une suite de réels de somme nulle, un choix naturel de partition de E est $A = \mathbf{R}^+$, $s_i = signe(\varepsilon_i)$. Soit n_+ (resp. n_-) le nombre de signes + (resp. de signes −), $n = n_+ + n_-$. Une *séquence* de la suite $(s_1, s_2, \cdots, s_n)$ est une suite maximale de signes successifs constants. On note alors R le nombre de séquences. Par exemple pour la suite de $n = 18$ signes :

$$++---+-+++++-+++--$$

on a : $n_+ = 10$, $n_- = 8$ et $R = 8$ séquences. A (n_+, n_-) fixés et sous (H_0), les n_+ signes + sont placés au hasard parmi les n positions. Il en découle que la loi sous (H_0) de la statistique $R \in \mathbf{N}$ dépend uniquement de (n_+, n_-) : la loi de R est libre de celle de ε. On dispose de tables statistiques pour $n_+ \leq 20$ et $n_- \leq 20$.

Si R est faible, la liaison entre les ε successifs est positive : il y a association ; pour R élevé, la liaison est négative : on dit qu'il y a intrication. Les tables donnent $r_-(\alpha)$ (resp. $r_+(\alpha)$) tel que $r_-(\alpha)$ (resp. $r_+(\alpha)$) est le plus grand k vérifiant $P_{H_0}(R \leq k) \leq \alpha$ (resp. le plus petit k vérifiant $P_{H_0}(R \geq k) \leq \alpha$). Pour n grand, R est approximativement normale de moyenne et de variance fonction de n_+ et n_- (cf. Tables statistiques).

Test de non-corrélation d'un couple gaussien On peut également utiliser le test (cf. Chapitre 1) de non corrélation pour un couple gaussien pour lequel on dispose d'un échantillon. Pour ce faire, on groupe par 2 (sans recouvrement) la suite des résidus (estimés) : par exemple, si $n = 2m$, on, sous l'hypothèse i.i.d. gaussiens, a m-paires indépendantes et non corrélées :

$$((\varepsilon_1, \varepsilon_2), (\varepsilon_3, \varepsilon_4), (\varepsilon_5, \varepsilon_6), \cdots, (\varepsilon_{n-1}, \varepsilon_n))$$

Si n est impair, on perd la dernière observation.

5.2.3 Les observations sont-elles gaussiennes ?

Le test de Jarque-Bera, le test d'ajustement du χ^2 et le test de Kolmogorov-Smirmov sont des tests asymptotiques, utilisables pour n grand. D'autres méthodes, telle la droite de Henry, sont descriptives. Décrivons le test de Jarque-Bera et la méthode de la droite de Henry.

Le test de Jarque-Bera

La statistique est construite sur la constatation suivante : une loi gaussienne est de *coefficient d'asymétrie* $\kappa_3 = E(X - \mu)^3/\sigma^3$ nul (*skewness*, $\kappa_3 = 0$ si X symétrique) et de *coefficient d'aplatissement* $\kappa_4 = E(X - \mu)^4/\sigma^4 = 3$ (*kurtosis*). Jarque et Bera testent l'hypothèse plus générale : $(H_0) : \kappa_3 = 0$ et $\kappa_4 = 3$.

Soient S et K les estimateurs empiriques de κ_3 et κ_4 issus des résidus estimés par MCO. La statistique :

$$JB = \frac{n-p}{6}\{S^2 + \frac{1}{4}(K-3)^2\}$$

prend une valeur faible sous l'hypothèse de normalité. Jarque et Bera ont montré que la loi asymptotique de JB est, sous $(H0)$, un χ^2 à 2 ddl.

Le test d'ajustement du *Chi* 2

Rappelons la construction générale de ce test. On veut tester :

$$(H_0) : Y_1, Y_2, \cdots, Y_n \text{ est une suite de variables } i.i.d. \text{ de loi } P_\theta.$$

où θ est un paramètre de dimension $p \geq 0$. Si $p > 0$, on dit que la loi est incomplètement spécifiée ; si la loi est complètement spécifiée, $p = 0$.

(i) Estimer θ par $\widehat{\theta}$, l'estimateur du maximum de vraisemblance.

(ii) Soit E l'espace d'état de Y ; on divise E en K classes C_1,C_2, $\cdots$, C_K ;

(iii) on note n_k le nombre de $Y_i, i = 1, n$, tombant dans C_k et $\widehat{p}_k = P(Y_{\widehat{\theta}} \in C_k)$, $k = 1, K$.

(iii) on calcule le $\Delta = \sum_{k=1}^{K} \frac{(n_k - n\widehat{p}_k)^2}{n\widehat{p}_k}$.

Δ est la *distance du* χ^2 entre la distribution empire observée et la distribution théorique estimée. Le résultat théorique dit que, sous (H_0), *et si tous les* $n\widehat{p}_k \geq 5$, alors Δ suit une loi du *Chi* 2 à $(K - p - 1)$ degré de liberté. Ainsi, la région de rejet de (H_0) au niveau α est

$$\mathcal{R}_\alpha = \{\Delta \geq q(\alpha; K - p - 1)\}$$

$q(\alpha; K - p - 1)$ étant le α-quantile d'un chi 2 à $(K - p - 1)$ ddl.

Revenons au cas du modèle linéaire gaussien et du test :

$$(H_0) : \widehat{\varepsilon}_1, \widehat{\varepsilon}_2, \cdots, \widehat{\varepsilon}_n \text{ est } i.i.d. \ \mathcal{N}(0, \sigma^2)$$

Ici, $E = \mathbf{R}$ et $\widehat{\varepsilon}$ est centré car le régresseur constant est retenu dans l'ensemble des variables explicatives. On choisit des classes adjacentes $C_1 =]-\infty, c_1]$,$C_2 =]c_1, c_2]$, $\cdots$, $C_p =]c_{p-1}, +\infty[$; l'estimation du MV de σ^2 est $\widehat{\sigma}^2 = \frac{1}{n}\sum_1^n \varepsilon_i^2$, et, sous les conditions $n\widehat{p}_k \geq 5$, le test d'ajustement reposera sur une loi du chi 2 à $(K - 2)$ ddl. Il est recommandé de choisir un découpage en classe d'effectifs estimés tous égaux, $\widehat{p}_k = K^{-1}$, avec la contrainte $\frac{n}{K} \geq 5$; un tel choix assure une bonne puissance au test.

5.2.4 Détection d'erreurs de spécification : méthodes graphiques

Les méthodes graphiques sont très utiles pour juger de l'adéquation d'un modèle. Elles permettent aussi de détecter l'origine des erreurs de spécification et nous guident sur le choix de modèles alternatifs plausibles. Ces méthodes sont basées sur la représentation graphique de points $\{(u_i, \widehat{\varepsilon}_i), i = 1, n\}$, où u_i est soit l'indice individuel i (ou le temps t), soit x_i, l'une des variables exogènes, soit l'endogène y_i. Tous les logiciels offrent la possibilité (ordre `PLOT`) de telles représentations graphiques.

5.3 Exercices : Choix et validation de modèle

Exercise 25 *Validation de modèle*

(Ω) : $y_t = a + bx_{1t} + cx_{2t} + dx_{3t} + \varepsilon_t$ est estimé sur la base de $n = 25$ observations. Les valeurs x_{1t} et les résidus $\widehat{\varepsilon}_t$ estimés par MCO sont donnés ci-dessous :

t	1	2	3	4	5	6	7	8	9	10	11	12
x_{1t}	0.5	.25	0	1	2	2.5	2.2	2	1.8	1.7	1.5	0.5
$\widehat{\varepsilon}_t$	0.1	0.7	0.7	−0.3	0.1	0.5	0.2	0.1	−0.2	−0.1	**−1.4**	0.7

$t = 13$	14	15	16	17	18	19	20	21	22	23	24	25
0.7	0.6	0.5	1	0.7	1.2	2.2	2.3	1.1	1.4	1.6	1.3	1.4
0.3	0.4	0.3	0.1	0.6	−0.2	0.3	0.3	−0.1	−0.9	−0.9	−0.6	−0.7

On trouve : $SCR(\Omega) = \sum_{i=1}^{25} \widehat{\varepsilon}_t^2 = 7.34$; $\sum_{i=1}^{25} \widehat{\varepsilon}_t^3 = -3.43$; $\sum_{i=1}^{25} \widehat{\varepsilon}_t^4 = 6.67$; $\sum_{i=1}^{24} \widehat{\varepsilon}_t \widehat{\varepsilon}_{t+1} = 3.65$.

(1) Tester la normalité des résidus en utilisant la statistique de Jarque-Bera.

(2) Représenter $\{(t, \widehat{\varepsilon}_t); t = 1, 25\}$ et les barrières de niveau $\pm 2\widehat{\sigma}$. Conséquence ?

(3) Soit $(\Omega^{(11)})$ le modèle avec son propre paramètre en $t = 11$:

$$(\Omega^{(11)}) : E(y_t) = a + bx_{1t} + cx_{2t} + dx_{3t} \text{ pour } t \neq 11, \text{ et } E(y_{11}) = m$$

On observe : $SCR(\Omega^{(11)}) = 5.94$. Tester que la donnée en $t = 11$ n'est pas aberrante.

(4) Tester l'indépendance des (ε_t) en utilisant le test des séquences.

(5) Représenter les points $\{(x_{1t}, \widehat{\varepsilon}_t); t = 1, 25\}$. A quelle erreur de spécification cette représentation fait-elle penser ? Comment améliorer le modèle ?

Exercise 26 *Choix d'un modèle pour le cycle biologique circadien*

T.P. : Utiliser R pour choisir un des modèles parmi les suivant :

On se place dans le contexte de l'exercice 3 (feuille *Modèle linéaire*) sur la modélisation du cycle journalier d'activité de la faune microbienne d'un marais d'Afrique équatoriale. Entre (H_2), (H_3), (H_{24}) et (H_9), le modèle étendant (H_3) en introduisant un effet jour, et pour les critères $\overline{R}^2$, C_p, s^2 et AIC, quel modèle retenir. Comparer ces résultats avec ceux des tests déjà effectués.

Exercise 27 *Choix du modèle d'analyse de traitements pour le cancer du sein*

Même question pour les modèles (H_T), (H_2), (H_4) et (H_6) étudiés dans l'exercice d'étude de durée de vie pour un cancer du sein (Analyse de la variance).

Exercise 28 *Graphiques d'invalidation d'un modèle*

Imaginer les représentations graphiques planes $\{(u_i, \widehat{\varepsilon}_i), i = 1, n\}$ correspondant aux erreurs de spécification suivantes (chaque fois, faire un choix adapté de la variable abscisse u_i) :

- Corrélation entre les résidus successifs (positive, ou négative).
- Non-homogénéité de la variance résiduelle (par exemple, la variance croît avec x_i).
- Détection de valeurs aberrantes.
- Oubli d'un effet linéaire exogène dans l'explication de $E(Y)$.
- Oubli d'un effet quadratique exogène dans l'explication de $E(Y)$.

Chapitre 6

Régressions logistique et polytomique

6.1 Introduction

On étudie la modélisation de données endogènes *catégorielles* Y à partir de conditions exogènes x. On parle de régression logistique si Y est *binaire*, de régression polytomique si Y est polytomique (à plus de deux états). L'endogène Y est par exemple le type de transport utilisé par un salarié (transport individuel ou transport en commun), l'état de santé d'un individu (sain ou malade), le statut d'occupation d'un logement (propriétaire ou locataire), la présence ou non d'une plante sur une parcelle agricole, etc.

La variable exogène x à état dans E peut être qualitative ($E = A = \{a_1, a_2, \cdots, a_K\}$), quantitative ($E = \mathbf{R}^p$), ou mixte $E = A \times \mathbf{R}^p$. Par exemple, pour expliquer le type de transport utilisé, on peut faire intervenir l'éloignement résidence-lieu de travail, le sexe du salarié, son salaire, un facteur de mobilité dans le travail : deux composantes de x sont quantitatives, deux sont qualitatives. Commençons par examiner le cas d'une réponse Y binaire.

Sans contrainte et en codant par 0 et 1 les deux états, Y sous la condition x est une variable de *Bernoulli* $\mathcal{B}(\pi(x))$ caractérisée par la probabilité :

$$P(Y = 1 \mid x) = \pi(x)$$

Un modèle spécifiera la forme fonctionnelle de $\pi(x)$. Sans contrainte sur $\pi(x)$, le modèle est *complet* ou *saturé*, noté $(\mathcal{S})$.

6.1.1 Structure des données

On supposera que les *observations* sont *indépendantes*. Il y a deux structures de données :

• *les données individuelles :* $\{(y_i, x_i), i = 1, n\}$, le dispositif des x est $\{x_1, x_2, \cdots, x_n\}$.

• *les données répétées* : $\{((y_{it}, x_t), i = 1, n_t), t = 1, T\}$. Sous x_t, il y a $n_t \geq 1$ observations y_{it}. Le nombre total d'observations est $n = \sum_{t=1}^{T} n_t$; le dispositif expérimental est $\{(x_1, n_1), (x_2, n_2), \cdots, (x_T, n_T)\}$.

L'étude statistique du modèle saturé $(\mathcal{S})$ n'est possible qu'en cas de données répétées. Si on note $y_t = \sum_{i=1}^{n_t} y_{it}$ le nombre de fois où 1 est réalisé sous la condition x_t, les données répétées sont résumées par le tableau :

Condition x	x_1	x_2	$\cdots$	x_i	$\cdots$	x_T
Effectif total	n_1	n_2	$\cdots$	n_i	$\cdots$	n_T
Effectif des $Y_{it} = 1$	y_1	y_2	$\cdots$	y_i	$\cdots$	y_T

6.1.2 Modèle Logit, Probit et autres

Pour fixer les idées, supposons que l'espace d'état de x est $E = \mathbf{R}^p$. Soit $F : \mathbf{R} \to [0,1]$ une fonction connue. Une façon de modéliser $\pi(.)$ est d'écrire :

$$\pi(x) = F({}^t\beta x)$$

où $\beta \in \mathbf{R}^p$ est un paramètre inconnu. Généralement, on choisit pour F une fonction de répartition[1] (notée fdr). Le modèle est linéaire en β à travers F, une fonction non-linéaire. F est la *fonction de lien* du modèle (*link function*) .

Example 15 *La régression logistique ou modèle Logit*

Elle correspond au choix $F = \Lambda(u) = \frac{e^u}{1+e^u} = (1+e^{-u})^{-1}$, la fdr logistique. L'étude du modèle Logit est le thème central de ce chapitre.

Example 16 *Seuillage d'une variable quantitative : le modèle Probit*

Le modèle Probit correspond au choix $F = \Phi$, la fdr gaussienne réduite. Une justification de ce choix est la suivante : considérons une observation binaire y qui est le résultat d'un *seuillage* au niveau s d'une variable latente $y^* \in \mathbf{R}$ (y^* n'est pas observée), $y = \mathbf{1}\{y^* < s\}$, où $\mathbf{1}(A) = 1$ si A est réalisé, 0 sinon, est l'indicatrice de l'événement A. Si y^* suit un modèle (Ω^*) linéaire et gaussien, y suit un modèle Probit (Ω) :

$$(\Omega^*) \quad : \quad y = \mathbf{1}\{y^* < s\}\,, \text{avec } y^* = {}^t\gamma x + \varepsilon,\ \varepsilon \sim \mathcal{N}(0,\sigma^2)$$

$$(\Omega) \quad : \quad \pi(x) = P(Y=1 \mid x) = P(Y^* < s) = \Phi(\frac{s - {}^t\gamma x}{\sigma}) = \Phi({}^t\beta x^*)$$

Le paramètre β^* du modèle Probit est :
(i) ${}^t\beta = {}^t(\frac{1}{\sigma}, \frac{-{}^t\gamma}{\sigma})$ et $x^* = {}^t(s,x)$ si le seuil s est connu ;
(ii) ${}^t\beta = {}^t(\frac{s}{\sigma}, \frac{-{}^t\gamma}{\sigma})$ et $x^* = {}^t(1,x)$ si le seuil s est inconnu.

Comment choisir F ? Approche graphique

Si les données sont répétées et si $x_t \in \mathbf{R}$, les T points $\{z_t = (x_t, \frac{r_t}{n_t}), t = 1,T\}$ apportent une information sur la fonction de lien F à choisir : existence ou non d'un centre de symétrie, différences des comportements aux asymptotes $\pi = 0$ ou $\pi = 1$, etc.

Identifiabilité du modèle

On s'assurera d'abord que le dispositif des x rend β identifiable c'est-à-dire, si $\beta \neq \beta'$, les lois de $\{(Y_i \mid x_i), i = 1,n\}$ associées à β et β' doivent être différentes. Si F est strictement croissante, l'identifiabilité équivaut à l'existence d'un x_i tel que ${}^t x_i\beta \neq {}^t x_i\beta'$. Si X est la matrice $n \times p$ des conditions exogènes, ${}^t X = {}^t({}^t x_1, {}^t x_2, \cdots, {}^t x_n)$, l'identifiabilité équivaut, comme pour le modèle linéaire, au fait $rang(X) = p$. On supposera par la suite que cette condition est satisfaite.

Codage d'une variable exogène qualitative

Si une composante z de x est qualitative, $z \in \{a_1, a_2, \cdots, a_k\}$, z peut être codée dans $\mathbf{R}^{k-1}$ en identifiant a_l au l-ième vecteur de la base canonique de $\mathbf{R}^{k-1}$, $l = 1, k-1$, et a_k à 0. Si x est une variable mixte, on peut coder sa partie qualitative et x peut encore être interprétée comme un élément de $\mathbf{R}^{p^*}$. Comme pour un modèle linéaire, aux situations $x \in E$ quantitatif, qualitatif ou mixte, correspondent les terminologies modèle

[1] Une fonction de répartition F est caractérisée par trois conditions : (i) $F \in [0,1]$ et est croissante ; (ii) F est continue à droite ; (iii) $\lim_{x\to-\infty} F(x) = 0$ et $\lim_{x\to+\infty} F(x) = 1$.

de régression, analyse de la variance et analyse de la covariance. Par exemple, si $x \in I \times J$ est qualitatif à deux facteurs :

$$\pi(i,j) = F(\mu_{i,j})$$

$\mu_{i,j} \in \mathbf{R}$ peut être décomposé comme dans une analyse de la variance à deux facteurs. Pour une condition mixte $(x,i) \in \mathbf{R}^p \times I$, on écrira :

$$\pi(x,i) = F(\mu_i(x))$$

sans restriction pour $\mu_i(x)$ pour le modèle complet. Le sous-modèle additif s'écrira $\mu_i(x) = \alpha_i + \mu(x)$, avec l'éventuelle spécification $\mu(x) = {}^t\beta x$ pour $\mu(x)$.

6.2 La régression logistique

6.2.1 Le modèle Logit

La *distribution logistique* est associée à la fdr $\Lambda(u) = (1+e^{-u})^{-1} = \frac{e^u}{1+e^u}$. Comme la fdr Φ gaussienne réduite, Λ est symétrique, $\Lambda(-u) = 1 - \Lambda(u)$, $(0, \frac{1}{2})$ étant centre de symétrie pour le graphe de Λ. La distribution logistique est centrée, de variance $\frac{\pi^2}{3}$, de densité $f(u) = \Lambda'(u) = \Lambda(u)(1-\Lambda(u)) = \frac{e^u}{(1+e^u)^2}$.

La fonction *Logit* : $[0,1] \to \overline{\mathbf{R}}$ est la fonction réciproque de Λ,

$$Logit(y) = \log \frac{y}{1-y}$$

Definition 2 *Supposons que $x \in \mathbf{R}^p$. Le modèle Logit est défini par :*

$$\pi(x) = P(Y = 1 \mid x) = \Lambda({}^t\beta x) \text{ ou } Logit(\pi(x)) = {}^t\beta x \tag{6.1}$$

Le quotient $OR(x) = \frac{\pi(x)}{1-\pi(x)} = \frac{P(Y=1|x)}{P(Y=0|x)}$ est appelé le *rapport des chances* sous x (*Odd ratio*) : le modèle Logit traduit que le log du rapport des chances suit le modèle linéaire ${}^t\beta x$. L'interprétation de β est la suivante : si la i-ème coordonnée de x est augmentée de 1, donnant x_i^+, le rapport des chances est multiplié par le facteur $\exp \beta_i$: $\frac{OR(x_i^+)}{OR(x)} = e^{\beta_i}$. Ou encore, $\frac{\partial Logit(\pi(x))}{\partial x_i} = \beta_i$.

6.2.2 Estimation du modèle Logit

Approche graphique : le Logit empirique

Pour des données répétées, la fonction *Logit* empirique est définie par :

$$x_t \to z_t = \log \frac{y_t + \frac{1}{2}}{(n_t - y_t) + \frac{1}{2}},\ t = 1,T \text{ avec } y_t = \sum y_{it}.$$

La constante $\frac{1}{2}$ interdit au numérateur et au dénominateur de prendre la valeur 0. La représentation du nuage des points $\{(x_t, z_t)\}$ donne une approche empirique de la modélisation : si $x \in \mathbf{R}$ et si le "nuage" $\{(x_t, z_t), t = 1,T\}$ est proche d'une droite, on optera pour le modèle Logit affine ${}^t\beta x = \alpha + \beta x$, une estimation graphique de α et β étant possible. Si $\{(x_t, z_t), t = 1,T\}$ évoque un nuage quadratique en x_t, on préférera le modèle $Logit\{\pi(x)\} = \alpha + \beta x + \gamma x^2$.

Estimation par le maximum de vraisemblance

La densité de la loi de Bernoulli $Y \sim \mathcal{B}er(\pi)$ est, pour $y \in \{0,1\}$:

$$P(Y=y) = \pi^y(1-\pi)^{1-y} = (1-\pi)\exp\{y \times Logit(\pi)\}$$

En données individuelles $\{(y_i, x_i), i=1,n\}$, la vraisemblance vaut, notant $\pi_i = \Lambda({}^t\beta x_i)$:

$$L_n(\beta) = \textstyle\prod_{i=1}^n \pi_i{}^{y_i}(1-\pi_i)^{1-y_i} = \prod_i (1-\pi_i)\exp\{y_i \times Logit(\pi_i)\}$$

Notant $\langle a,b\rangle = {}^t ab$ le produit scalaire sur $\mathbf{R}^p$, la log-vraisemblance l_n vaut :

$$l_n(\beta) = \langle T_x(y), \beta\rangle + \psi_x(\beta), \text{ avec :} \tag{6.2}$$

$$T_x(y) = \sum_{i=1}^n y_i\, x_i, \text{ et } \psi_x(\beta) = -\sum_{i=1}^n \log(1+e^{\langle\beta, x_i\rangle}) \tag{6.3}$$

Pour des données répétées, $T_x(y) = \sum_{t=1}^T y_t x_t$, $\psi_x(\beta) = -\sum_{t=1}^T n_t \ln(1+e^{\langle\beta,x_t\rangle})$.

Le modèle Logit appartient à une *famille exponentielle* (cf. § 14.6) : $T_x(y)$ est la statistique canonique associée à β ; $\psi_x(\beta)$ est la constante de normalisation faisant de L_n une densité de probabilité. A cette constante près, la log-vraisemblance est *linéaire* en β.

Pour obtenir le comportement asymptotique de l'estimation du MV, on utilisera la condition suivante sur les (x_i) :

($\mathbf{H}(\mathcal{X})$) **:** les (x_i) sont les réalisations i.i.d. d'une loi λ sur F et F est compact.

Proposition 21 ***Estimation du MV du modèle*** *LOGIT*

(1) *La log-vraisemblance $\beta \to l_n(\beta)$ est strictement concave.*

(2) *L'information de Fischer $I_n(\beta)$ est inversible si β est identifiable ; elle vaut :*

$$I_n(\beta) = -\psi^{(2)}_{\beta^2}(\beta) = \sum_{i=1,n} x_i \times^t x_i \pi_i(1-\pi_i), \text{ avec } \pi_i = \Lambda({}^t\beta x_i)$$

(3) *Si $I(\beta) = \int_F x \times {}^t x\ \pi(x)(1-\pi(x))\lambda(dx)$ est régulière, si l'espace des paramètres Θ est compact, alors :*

$$\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow{Loi} \mathcal{N}(0, I(\beta)^{-1})$$

Preuve :

(1 et 2) $[\log(1+e^u)]' = \Lambda(u)$, et $[\log(1+e^u)]'' = \Lambda'(u) = \Lambda(u)(1-\Lambda(u))$. Donc :

$$\frac{\partial^2 l_n}{\partial\beta^2}(\beta) = \psi^{(2)}_{\beta^2}(\beta) = -\sum_{i=1,n} x_i \times^t x_i \pi_i(1-\pi_i)$$

Soit $a = \inf\{\pi_i(1-\pi_i), i=1,n\}$; a est > 0, et $\sum_{i=1,n} x_i \times^t x_i \pi_i(1-\pi_i) \le a\,{}^t XX$ au sens des matrices sdp. X étant de rang plein, la matrice hessienne est définie négative ; d'où (1). Etant indépendante des y_i, $I_n(\beta) = \frac{\partial^2 l_n}{\partial\beta^2}(\beta)$. On en déduit (2).

(3) Admis.

. . . . . □

Commentaires :

(i) La concavité a une conséquence numérique importante : un algorithme itératif pour l'optimisation de l_n convergera bien vers la valeur $\widehat{\beta}_n$: (1) il n'y a pas de risque de converger vers un maximum local non-global ; (2) la convergence de l'algorithme ne dépend pas du point d'initialisation de l'algorithme[2].

[2] Ce n'est plus le cas si la log-vraisemblance n'est pas concave. On verra que pour une régression non-linéaire (cf. Ch. 11), la somme des carrés résiduels $SCR(\beta)$ peut présenter des maximums locaux non-globaux ; on est donc amené à initialiser l'algorithme avec une "bonne" estimation initiale, ce qui est une question délicate.

(ii) Si l'espace d'état des exogènes F est un sous-ensemble de $\mathbf{R}^p$, $I(\beta)$ est inversible dès que λ charge positivement p ouverts autour de p points $\{x_k, k = 1, p\}$, ces p points étant linéairement indépendants dans E.

(iii) En situation de données répétées, $I_n(\beta) = \sum_1^T c_n(t)\, x_t \times {}^t x_t\, \pi_t(1-\pi_t)$, avec $c_n(t) = \frac{n_t}{n}$. Si $\liminf_n c_n(t) \geq c > 0$ pour tout t, $I_n(\beta)$ est régulière pour n assez grand dès que $\sum_1^T x_k\, {}^t x_k$ est régulière.

(iv) Dans la pratique, on utilise l'approximation $(\widehat{\beta}_n - \beta) \sim \mathcal{N}(0, I_n(\widehat{\beta}_n)^{-1})$.

Estimation de $\pi(x)$, prédiction de $(Y \mid x)$ et calibration d'une observation y.

Pour une nouvelle condition x, $\pi(x)$ est estimée par $\widehat{\pi}_x = \pi(x, \widehat{\beta})$. L'intervalle de confiance (la loi) asymptotique pour $\pi(x)$ (de $\pi(x, \widehat{\beta})$) s'obtient en calculant la variance de $\widehat{\pi}_x$. Puisque $\Lambda' = \Lambda(1 - \Lambda)$, la delta-méthode (cf. Annexe, §4, Propriété 7) donne :

$$Var(\widehat{\pi}_x) \approx \pi_x^2(1 - \pi_x)^2\ {}^t x Var(\widehat{\beta}) x$$

Il y a plusieurs façons de *prédire* $(Y \mid x)$: (i) prendre le résultat de la simulation d'une loi de Bernoulli $\mathcal{B}(\widehat{\pi}_x)$; (ii) choisir un seuil $\alpha \in]0, 1[$ et prendre $\widehat{y}_x = \mathbf{1}\{\widehat{\pi}_x \geq \alpha\}$. Un bon choix de α est $\alpha^* = Arg \min\{\sum (y_i - \widehat{y}_i(\alpha))^2, 0 < \alpha < 1\}$.

Le problème inverse de la prédiction est celui de la *calibration* : quelle condition exogène x_0 assure une probabilité $\pi_0 = P(Y = 1 \mid x_0)$ donnée ? Pour un modèle Logit affine sur $\mathbf{R}$, on choisira $\widehat{x}_0$ t.q. $\widehat{\alpha} + \widehat{\beta}\widehat{x}_0 = Logit(\pi_0)$, ou tout autre choix dans l'intervalle de confiance[3] :

$$\{x \in \mathbf{R} :\mid (\widehat{\alpha} + \widehat{\beta}x) - Logit(\pi_0) \mid \leq 2\widehat{\sigma}(\widehat{\alpha} + \widehat{\beta}x)\}$$

Analogies et différences avec la régression linéaire

Notons $y = {}^t(y_i, i = 1, n)$, $m(\beta) = E_\beta(Y) = {}^t(\pi(x_i, \beta), i = 1, n)$. L'EMV pour un modèle exponentiel étant caractérisée par l'égalité $T_x(y) = E_{\widehat{\beta}}(T(Y))$ (cf. § 14.6), l'équation de l'EMV vérifie ${}^t X y = {}^t X m(\widehat{\beta}_n)$. Cette équation est équivalente aux équations normales ${}^t X y = {}^t X X \widehat{\beta}$ du modèle linéaire. Les conditions d'identifiabilité pour β et l'expression de la matrice de covariance de $\widehat{\beta}_n$ s'écrivent de façon analogue pour les deux modèles :

$$Var(\widehat{\beta}_n) = \{-\frac{\partial^2 l_n}{\partial \beta^2}(\beta)\}^{-1} = \{{}^t X D^{-1} X\}^{-1} \text{ où } D = VarY = Diag(\pi_i(1 - \pi_i))_{i=1,n}.$$

Par contre, du fait de la non-linéarité de la fonction $\beta \to m(\beta)$, la recherche de l'EMV nécessite de recourir à un algorithme d'optimisation itératif et l'expression de l'EMV n'est plus explicite. D'autre part, le résultat sur la loi de $\widehat{\beta}_n$ et le test de sous-modèle ne sont valables qu'asymptotiquement pour n assez grand. Pour la régression *linéaire gaussienne*, la normalité de l'EMCO et le test de Fischer sont valables pour tout n.

6.2.3 Validation et test de sous-modèle

Le résultat général sur le test du rapport de vraisemblance (test du RV) s'applique dès que la sous-hypothèse est régulière de classe $\mathcal{C}^2$ (cf. § 14.8, condition **(C5)**) : si $\mathcal{M}_1 \subset \mathcal{M}_2$ sont deux modèles emboîtés de dimensions $p_1 < p_2$, si $l_n(\mathcal{M})$ est la log-vraisemblance calculée à la valeur $\widehat{\beta}_n(\mathcal{M})$, l'EMV de β sous $\mathcal{M}$, alors, sous $\mathcal{M}_1$:

$$2\{l_n(\mathcal{M}_2) - l_n(\mathcal{M}_1)\} \xrightarrow{loi} \chi^2(p_2 - p_1)$$

Test de la non-influence de x.

Pour le modèle $Logit(\pi(x)) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$, le test de $\beta_1 = \cdots = \beta_p = 0$ repose sur un $\chi^2(p)$. Le test de nullité d'une coordonnée de β reposera sur un χ_1^2.

[3] $\widehat{\sigma}(\widehat{\alpha} + \widehat{\beta}x)$ dépendant de x, la zone de confiance est délimitée par deux hyperboles.

Test de sous-hypothèse.

Considérons une sous-hypothèse régulière $(H_0) : \beta = r(\alpha)$, $\alpha \in \mathbf{R}^q$, où r est de classe $\mathcal{C}^2$ et $\frac{\partial r}{\partial \alpha}(\alpha)$ est de rang plein. La statistique du test du RV suit asymptotiquement, sous (H_0), un $\chi^2(p-q)$.

Test de validation de modèle en situation de données répétées.

Le modèle saturé $(\mathcal{S})$ est de dimension T. Il est estimé par :

$$(\mathcal{S}) : \widehat{\pi}_t = \frac{y_t}{n_t}, \text{ où } \pi_t = P(Y = 1 \mid x_t),\ t = 1, T$$

Sa log-vraisemblance vaut $l_n(\mathcal{S}) = \sum_1^T \{y_t \log \widehat{\pi}_t + (n_t - y_t)\log(1 - \widehat{\pi}_t)\}$. Un sous-modèle $(\mathcal{M})$ de dimension p est donc validé dans $(\mathcal{S})$ à partir de la statistique :

$$2(l_n(\mathcal{S}) - l_n(\mathcal{M})) \overset{loi}{\rightsquigarrow} \chi^2(T-p) \text{ sous } (\mathcal{M})$$

Ce résultat s'applique dès que pour chaque t, n_t est assez grand. En particulier, la statistique du test de non-influence de x suit une loi du χ^2_{T-1}.

On peut également utiliser le test du χ^2 d'indépendance de x et de y : la distance du χ^2, notée Δ^2, vaut, en posant $n_{t0} = n_t - y_t, n_{t1} = y_t, n_{+j} = \sum_t n_{tj}$:

$$\Delta^2 = n \sum\sum \frac{(n_{tj} - \frac{n_t}{n} n_{+j})^2}{n_t n_{+j}}, t = 1, T \text{ et } j = 0, 1$$

Asymptotiquement, Δ^2 suit une loi du $\chi^2(T-1)$ si x est non-influente.

Indicateurs de la qualité d'ajustement de $\mathcal{M} \subset \mathcal{S}$.

Pour des données répétées, un indicateur de la qualité de $\mathcal{M}$ est

$$I = \frac{l_n(\mathcal{M}) - l_n(\mathcal{M}_0)}{l_n(\mathcal{S}) - l_n(\mathcal{M}_0)}$$

Ici, $(\mathcal{M}_0)$ est le modèle "x est sans influence". $0 \leq I \leq 1$ et I sera d'autant plus proche de 1 que $\mathcal{M}$ explique bien le modèle.

S'il n'y a pas de répétitions, notant $\widehat{y}_i = \widehat{\pi}_i$ et en utilisant l'analogie avec le coefficient de corrélation multiple du modèle linéaire, on peut définir :

$$R^2_{\mathcal{M}} = 1 - \frac{\|y - \widehat{y}(\mathcal{M})\|^2}{\|y - \widehat{y}(\mathcal{M}_0)\|^2} = \frac{\sum_i (\widehat{\pi}_i(\mathcal{M}) - \overline{y})^2}{\sum_i (y_i - \overline{y})^2}$$

Le coefficient pénalisé par la dimension de $(\mathcal{M})$ est : $1 - \overline{R}^2 = \frac{n}{n-p}(1 - R^2)$.

Critères de choix de modèle du type AIC.

$$XIC(\mathcal{M}) = -2 \log l_n(\mathcal{M}) + c(n) d(\mathcal{M})$$

Le critère AIC correspond à une vitesse de pénalisation constante $c(n) = 2$; le critère BIC à $c(n) = \log n$.

6.2.4 Résidus et validation de modèle

Pour des données répétées, les résidus estimés réduits sont :

$$\varepsilon_t = \varepsilon_t(\widehat{\pi}_t) = \frac{y_t - n_t \widehat{\pi}_t}{\sqrt{n_t \widehat{\pi}_t (1 - \widehat{\pi}_t)}}, t = 1, T$$

Pour n_t grand, $\varepsilon_t \sim \mathcal{N}(0, 1)$ si le modèle est valide puisque y_t suit une loi binomiale $\mathcal{B}(n_t, \pi_t)$: de fortes déviations de ε_t par rapport à la loi $\mathcal{N}(0, 1)$ incitent à corriger le modèle : on détectera ainsi des données aberrantes, l'oubli ou la mauvaise prise en compte d'exogènes, etc.

6.2.5 Pourquoi choisir une modélisation Logit ?

Plusieurs raisons justifient le choix du modèle Logit.

(1) *C'est la famille exponentielle canoniquement associée à la loi de Bernoulli.*
Le modèle Logit est associé à l'écriture suivante de la loi de Bernoulli :

$$P(Y = y \mid x) = (1 - \pi_x) \exp\{y \times Logit(\pi_x)\}.$$

Dans la terminologie de McCullagh et Nelder, le modèle Logit est un *modèle linéaire généralisé canonique*. Le modèle Probit n'appartient pas lui à une famille exponentielle.

(2) *Analyse discriminante et modèle Logit.*
Supposons que l'on est en présence de deux populations gaussiennes $\mathcal{P}_0$ et $\mathcal{P}_1$, repérables par $x \in \mathbf{R}^p$, associée à $y = 0$ et à $y = 1$, en proportions $\{p_0, p_1\}$ et se différenciant par leur moyenne $\mu_0 \neq \mu_1$:

$$\mathcal{L}(X \mid Y = i) \sim \mathcal{N}(\mu_i, \Sigma),\ i = 0, 1$$

Si on observe $x \in \mathbf{R}^p$, l'analyse discriminante (cf. Saporta) affectera x à $\mathcal{P}_1$ avec la probabilité

$$Logit\, P(Y = 1 \mid x) = \alpha + \beta x$$

où $\beta = \Sigma^{-1}(\mu_1 - \mu_0)$, $\alpha = \log \frac{p_0}{p_1} + \frac{1}{2}\{{}^t\mu_0\Sigma^{-1}\mu_0 - {}^t\mu_1\Sigma^{-1}\mu_1\}$. En effet, la loi du mélange gaussien est $p(x) = c\sum_{i=0}^{1} p_i \exp -\frac{1}{2}{}^t(x - \mu_i)\Sigma^{-1}(x - \mu_i)$. Un calcul direct donne :

$$\begin{aligned} P(Y &= 1 \mid x) = \frac{P(X = x \mid Y = 1)P(Y = 1)}{p(x)} \\ &= \{1 + \frac{p_0}{p_1} \exp[-\frac{1}{2}({}^t\mu_0\Sigma^{-1}\mu_0 - {}^t\mu_1\Sigma^{-1}\mu_1) + {}^t(\mu_0 - \mu_1)\Sigma^{-1}x]\}^{-1} \\ &= \{1 + \exp(\alpha + \beta x)\}^{-1} \end{aligned}$$

6.3 Autres modélisations de données binaires

6.3.1 Le modèle Probit

Le modèle Probit correspond au choix de fonction de lien $F = \Phi$, où Φ est la fdr gaussienne réduite. Il s'introduit naturellement si y est issue du seuillage d'une variable latente y^* qui suit (Ω^*), un modèle linéaire gaussien :

$$\begin{aligned} (\Omega) &: \quad y = \mathbf{1}\{y^* < s\}\,, \text{avec } (\Omega^*) : y^* = {}^t\gamma x + \varepsilon \\ (\Omega) &: \quad \pi(x) = P(Y = 1 \mid x) = \Phi(\frac{s - {}^t\gamma x}{\sigma}) = \Phi({}^t\beta x^*) \end{aligned}$$

Le paramètre β^* est :
(i) ${}^t\beta = {}^t(\frac{1}{\sigma}, \frac{-{}^t\gamma}{\sigma})$ pour $x^* = {}^t(s, x)$ si le seuil s est connu ;
(ii) ${}^t\beta = {}^t(\frac{s}{\sigma}, \frac{-{}^t\gamma}{\sigma})$ pour $x^* = {}^t(1, x)$ si le seuil s est inconnu.

Si $X = (\mathbf{1}, \mathbf{x})$ est de rang plein, β est identifiable dans le cas (i) : $\frac{1}{\sigma}$ (donc σ) et $\frac{\gamma}{\sigma}$ (donc γ) sont identifiables. Par contre, seuls les paramètres $\frac{s}{\sigma}$ et $\frac{\gamma}{\sigma}$ sont identifiables si s est inconnu.

Proximité des modèles Probit et Logit.
A une homothétie près sur x, les modèles Logit et Probit sont proches l'un de l'autre : après réduction de la variance pour Λ ($Var(\Lambda) = \frac{\pi^2}{3}$), on a :

$$\sup_{x \in \mathbb{R}} \mid \Phi(x) - \Lambda(\frac{\pi}{\sqrt{3}}x) \mid \simeq 0.023$$

Donc, si on effectue les ajustements Probit $\Phi({}^t\widehat{\gamma}x)$ et Logit $\Lambda({}^t\widehat{\beta}x)$ sur les mêmes données et si l'un des modèles est valide, $\widehat{\gamma}$ sera voisin de $\frac{\pi}{\sqrt{3}}\widehat{\beta} \sim 1.81\widehat{\beta}$.

Concavité de la log*-vraisemblance et asymptotique de l'EMV.*
Bien que n'appartenant pas à une famille exponentielle, la log-vraisemblance du modèle Probit est concave. Pour justifier ce résultat, revenons à une fdr générale F deux fois dérivable et de densité $f = F'$. Notons $F_i = F({}^t\beta x_i)$. La log-vraisemblance s'écrit :

$$l_n(\beta) = \sum_{i=1,n} \{y_i \log F_i + (1 - y_i) \log(1 - F_i)\}$$

La concavité de l_n est assurée par celle de $\log F$ et de $\log(1 - F)$. Une CS est :

$$(i)\ f'F - f^2 \leq 0 \text{ et } (ii)\ f'(1 - F) + f^2 \geq 0$$

Quelques manipulations montrent que ces conditions sont satisfaites pour $F = \Phi$.

Calcul de la matrice d'information.

Le i-ème terme de la dérivée seconde de l_n vaut, notant $f_i = f({}^t\beta x_i)$:

$$a_i = \left\{ (\frac{f_i'}{F_i} - \frac{f_i^2}{F_i^2})y_i - (\frac{f_i'}{1 - F_i} + \frac{f_i^2}{(1 - F_i)^2})(1 - y_i) \right\} x_i \times^t x_i$$

La matrice d'information de Fischer est donc :

$$I_n(\beta) = -E_\beta(l^{(2)}_{\beta^2}) = \sum_{i=1,n} \frac{f_i^2}{F_i(1 - F_i)} x_i \times {}^t x_i \tag{6.4}$$

Proposition 22 *Estimation du MV pour le modèle Probit*

(1) *La vraisemblance du modèle Probit est strictement concave.*

(2) *La matrice d'information de Fischer est donnée par (6.4) avec $f = \varphi$, la densité gaussienne réduite. Elle est inversible si β est identifiable.*

(3) *Sous $(H(\mathcal{X}))$, et pour n grand : $(\widehat{\beta}_n - \beta) \sim \mathcal{N}_p(0, I(\widehat{\beta}_n)^{-1})$.*

6.3.2 La distribution complémentaire log-log

C'est la fdr G définie par [4] :

$$G : \mathbf{R} \to [0, 1],\ G(u) = 1 - e^{-e^u}, u \in \mathbf{R}$$

La fonction réciproque est : $G^{-1}(y) = \log\{-\log(1 - y)\}$, $y \in [0, 1]$. La courbe de réponse du modèle affine est associé est : $\pi(x) = 1 - \exp[-\exp(\alpha + \beta x)]$.

L'interprétation de β est la suivante : pour deux conditions exogènes x_1 et x_2, on a :

$$(1 - \pi(x_2)) = (1 - \pi(x_1))^{\exp[\beta(x_2 - x_1)]}$$

La probabilité en x_2 est celle en x_1 élevée à la puissance $\exp \beta(x_1 - x_2)$. G se différentie fondamentalement des distributions Logit et Probit sur deux aspects :

(1) $\{(u, G(u)), u \in R\}$ n'a pas de centre de symétrie ;

(2) les *queues de la distribution* en $-\infty$ et en $+\infty$ ont des comportements différents l'un de l'autre : G s'approche plus rapidement de $G = 1$ que de $G = 0$.

[4] G est la distribution de valeur extrême de *Gumbel*. Elle a pour moyenne 0.577 et pour variance $\pi^2/6$.

6.4 Modèles de régression polytomique

Un modèle polytomique explique une variable endogène Y à valeur *catégorielle* à partir d'une condition exogène $x \in E$. La variable x peut être quantitative, qualitative ou mixte. Y prend $K \geq 2$ modalités qualitatives. Les modèles binaires (Logit, Probit ou autres, $K = 2$) ont fait l'objet du chapitre 8.

Y peut être, par exemple, le type de transport utilisé par un salarié de la région (métro, bus, transport individuel ; $K = 3$), le vote pour un candidat choisi parmi K, le degré de gravité d'un goitre (codé de 1 à 5) ou des appréciations (échelonnées de 1 à 10) que des dégustateurs attribuent à un produit alimentaire. Dans les deux derniers exemples, il existe un ordre implicite entre les modalités et on parle de données *ordinales*. Sinon, comme dans les deux premiers exemples ou pour des modalités de couleurs, les données sont *nominales*. La modélisation pourra prendre en compte cette classification des données.

Y peut être le croisement de deux (ou plus) données binaires (ou polytomiques) Y_1 et Y_2 : par exemple Y_1 repère un problème de souffle et Y_2 un problème de toux : Y est bivariée à 4 modalités, $Y \in \{0, 1\}^2$.

L'espace $F = \{a_1, a_2, \cdots, a_K\}$ des états de Y sera codé $\{0, 1, 2, \cdots, K-1\}$. 0 sera un état de référence. Ce choix arbitraire est sans conséquence sur la généralité des modèles que nous allons décrire.

6.4.1 Estimation d'une régression polytomique

Modéliser $(Y \mid x) \in F$, c'est choisir une forme paramétrique pour les probabilités :

$$\pi_k(x, \beta) = P(Y = k \mid x),\ \beta \in \mathbf{R}^p,\ x \in E,\ k = 1, K-1$$

Comme en situation binaire, il est inutile de modéliser $\pi_0(x, \beta) = 1 - \sum_1^{K-1} \pi_k(x, \beta)$ qui est connue à partir des autres π_k.

Que les données soient individuelles $\{(y_i, x_i), i = 1, n\}$ ou répétées, $\{(y_{it}, x_t),\ i = 1, n_t;\ t = 1, T\}$, on supposera que les observations sont *indépendantes*. On s'assurera d'abord de l'identifiabilité des paramètres du modèle, c'est-à-dire de l'injectivité de la correspondance de $\mathbf{R}^p$ (espace des β) dans $\mathbf{R}^{n(K-1)}$ (espace des lois $((Y_i \mid x_i),\ i = 1, n)$) :

$$\beta \to ((\pi_k(x_i, \beta); k = 1, K-1);\ i = 1, n)$$

L'estimation par le MV se fait de façon standard. Sous de bonnes conditions de régularité du modèle (ce sera le cas si le modèle est exponentiel) et d'ergodicité du dispositif expérimental $\mathcal{X} = \{x_1, x_2, \cdots, x_n\}$ pour $n \longrightarrow \infty$, on a les résultats classiques : variance s'exprimant à partir de l'information de Fischer, normalité asymptotique de l'EMV, test asymptotique de RV suivant un χ^2, test de validation de modèle. Si la log-vraisemblance est strictement concave (c'est le cas d'un modèle exponentiel identifiable), les méthodes itératives d'optimisation convergeront vers l'unique EMV quelle que soit la valeur retenue pour l'initialisation de l'algorithme.

6.4.2 Le modèle saturé ou modèle multinomial

Le *modèle complet* $(\mathcal{S})$ pour des données répétées est :

$$(\mathcal{S}) : \pi_{kt} = P(Y_{it} = k \mid x_t)\,;\ i = 1, n_t\,;\ t = 1, T\,;\ k = 1, K-1$$

$(\mathcal{S})$ dépend de $T(K-1)$ paramètres (π_{0t} est donnée par $\pi_{0t} = 1 - \sum_1^{K-1} \pi_{kt}$). Soient : $n_{kt} = \sum_i \mathbf{1}(y_{it} = k)$ l'effectif des observations $(y_{it})_i$ prenant la modalité k ($\sum_k n_{kt} = n_t$), $n = \sum_{t=1}^T n_t$, $\pi(t) = (\pi_{kt}, k = 0, K-1)$ et $\pi = (\pi(t), t = 1, T)$.

Pour chaque t, la variable de comptage $\mathbf{N}_t = (n_{kt},\, k = 0, K-1)$ suit une loi multinomiale $\mathcal{M}(n_t, \pi(t))$ ([5]). Ces variables sont indépendantes pour des t différentes. La log-vraisemblance s'écrit :

$$l_n((n_{kt}), \pi) = h((n_t)) + g((n_{kt})) + \sum_{t,k} n_{kt} \log \pi_{kt}$$

Le modèle multinomial appartient à une famille exponentielle, les EMV $\widehat{\pi}_{kt} = \frac{n_{kt}}{n_t}$ vérifiant :

(1) $\widehat{\pi}(t)$ estime sans biais $\pi(t)$; ces T estimations sont indépendantes.
(2) $Var(\ \widehat{\pi}(t)) = \frac{1}{n_t}\Sigma_t$ où $\Sigma_t(k,k) = \pi_{kt}(1-\pi_{kt})$ et $\Sigma_t(k,l) = -\pi_{kt}\pi_{lt}$ si $l \neq k$.
(3) Si $n_t \to \infty$, $\sqrt{n_t}(\widehat{\pi}(t) - \pi(t)) \xrightarrow{Loi} \mathcal{N}_K(0, \Sigma_t)$.

Du fait de la contrainte $\sum_k \widehat{\pi}_{kt} = 1$, Σ_t est une matrice singulière de rang $K-1$.

6.4.3 Le modèle logistique polytomique

Le cas de données nominales

Posons $a_k = \log \pi_k$. Le modèle logistique polytomique nominal spécifie les rapports $\frac{\pi_k}{\pi_0}$:

$$a_k - a_0 = \log \frac{\pi_k}{\pi_0} = {}^t x(k)\beta_k,\ k \geq 1 \tag{6.5}$$

A chaque modalité $k \geq 1$ est associée une condition exogène $x(k)$ et un paramètre β_k, tous les deux dans $\mathbf{R}^{p(k)}$. Les $x(k)$ pour des k différents peuvent coïncider ou non ; les dimensions $p(k)$ peuvent être égales ou non. Notons $\beta = (\beta_k, k \geq 1)$; sous la contrainte $\beta_0 = 0$, le paramètre β est identifiable[6].

Definition 3 *Le modèle logistique polytomique est le modèle (6.5). Il s'écrit aussi :*

$$\pi_k(x) = P(Y = k \mid x, \beta) = \frac{\exp\, {}^t x(k)\beta_k}{1 + \sum_{l=1}^{K-1} \exp\, {}^t x(l)\beta_l},\ k = 1, K-1$$

Des extensions s'obtiennent en remplaçant ${}^t x(k)\beta_k$ par d'autres fonctions $a_k(x, \beta)$.

Example 17 *Choix de transport : auto, bus ou métro*

Codons ces trois modalités respectivement par $\{0, 1, 2\}$. Si z est une caractéristique socio-économique de l'individu et $x(k)$ une caractéristique liée au type de transport, on peut choisir de modéliser a_k par $a_k(x, \beta) = \log \pi_k = \alpha + {}^t x(k)\beta + {}^t z\gamma$. Dans ce modèle ternaire, seules interviennent les différences $a_k(x, \beta) - a_0(x, \beta)$, $k = 1, 2$, les paramètres latents α, γ disparaissant :

$$\log \frac{\pi_k(x, \beta)}{\pi_0(x, \beta)} = {}^t\{x(k) - x(0)\}\beta,\ k = 1, 2$$

Example 18 *Modèle à* log*-ratio constant (Proportional odd model)*

Réécrivant ${}^t x(k)\beta_k = \alpha_k + {}^t x(k)\gamma_k$, le modèle à log-ratio constant est défini par l'égalité des γ_k : $\forall k \neq 0, \gamma_k = \gamma$. C'est un modèle additif $(k) + (x)$.

[5] La loi multinomiale généralise à $K > 2$ la loi binomiale. Pour $K = 2$, elle s'écrit $(N, n-N)$ où N est la loi binomiale $\mathcal{B}(n, \pi_1)$. La loi multinomiale $\mathcal{M}(n, \pi)$ de paramètres n, un entier ≥ 2 et $\pi = (\pi_1, \pi_2, \cdots, \pi_K)$, K probabilités de somme 1, est définie sur $\{0, 1, \cdots, n\}^K$ par : $P(\mathbf{N} = m) = \frac{n!}{\prod_{k=1}^K m_k!} \prod_{k=1,K} \pi_k^{m_k}$.

[6] Pour $k = 0$, le choix implicite $\beta_0 = 0$ ne limite pas la généralité du modèle. En effet, toute probabilité $\pi_k = a_k\{\sum_0^{K-1} a_l\}^{-1}$ est invariante par homothétie sur les a. Le choix $a_0 = 1$ correspond à $\beta_0 = 0$.

Le cas de données ordinales : modèle logistique cumul é

Un modèle peut prendre en compte la structure d'ordre. Le modèle *Logit cumulé* modélise le rapport des chances des événements complémentaires $(Y \leq k)$ et $(Y > k)$, $k = 0, K-2$:

$$\log \frac{P(Y \leq k \mid x)}{P(Y > k \mid x)} = {}^t x(k)\theta_k$$

Le sous-modèle à log-ratio constant traduit le parallélisme des différents hyperplans :

$$(\omega) : \ {}^t x\theta_k = \alpha_k + {}^t x\gamma \text{ et } \dim(\omega) = (K-1) + p \text{ si } \gamma \in \mathbf{R}^p$$

Sauf l'identité des modèles saturés, il n'y a pas d'emboîtement entre les modèles ordinaux et les modèles nominaux. On choisira entre l'un ou l'autre sur la base d'un critère de type *AIC* ou d'une procédure de test.

6.5 Exercices : régression logistique ou polytomique

Utiliser `R` pour traiter les différents exercices qui suivent.

Exercise 29 *Modélisation de données de mortalité infantile.*

Le tableau ci-dessous[7] donne les fréquences r/n de mortalité infantile dans une population à risque en fonction de trois facteurs $x = (a, b, c)$ concernant la mère :

(a) $a = 0$ si la mère fume au plus 5 cigarettes par jour, $a = 1$ sinon ;

(b) $b = 0$ si la mère a moins de 30 ans, $b = 1$ sinon ;

(c) $c = 0$ si la grossesse dure entre 197 et 260 jours, 1 sinon (durée de grossesse normale).

cig : a	0	1	0	1	0	1	0	1
âge : b	0	0	1	1	0	0	1	1
gest. : c	0	0	0	0	1	1	1	1
effectif r	50	9	41	4	24	6	14	1
total n	365	49	188	15	4036	465	1508	125

(1) Quel est la dimension du modèle saturé $(\mathcal{S})$. Calculer sa log-vraisemblance.

(2) Expliquer pourquoi $(\mathcal{S})$ peut être reparamétré en $a, b, c \in \{0, 1\}$ de la façon suivante :

$$(\mathcal{S}) : Logit\{\pi(a,b,c)\} = \beta_1 + \beta_2 a + \beta_3 b + \beta_4 c + \beta_5 ab + \beta_6 ac + \beta_7 bc + \beta_8 abc$$

(3) Effectuer la régression logistique descendante en respectant la hiérarchie des interactions. Tester chacun des modèles dans $(\mathcal{S})$. Quel modèle retenir ? Interpréter les résultats.

Exercise 30 *Différents choix pour la fonction de lien F.*

Le tableau ci-dessous[8] donne le nombre de scarabées morts après 5 heures d'exposition à l'oxyde de carbone pour 8 doses x de concentration en oxyde de carbone :

$z = \log(\text{dose})$	6.91	7.24	7.55	7.84	8.11	8.37	8.61	8.84
effectif r	6	13	18	28	52	53	61	60
total n	59	60	62	56	63	59	62	60

[7] Wermuth N., 1976, *Exploratory analyses of multidimensional contingency tables,* Proc. 9th Int. Biometrics Conference, V.I., *279-95.*

[8] Bliss C.I., 1935, *The calculation of dosage-mortality curve*, Ann. Appl. Biol. 22, 134-167.

(1) Représenter $\{(\log(x_t), \frac{r_t}{n_t}), t = 1, 8\}$. Quelle fonction de lien F privilégier ?

(2) Effectuer l'ajustement affine sur z pour les 3 fonctions de lien Logit, Probit, complémentaire-log-log. Quel modèle retenir ? Tester la validité de chacun des modèles dans le modèle saturé. x est-elle influente ?

(3) Pour les 3 modèles, calculer l'indicateur de qualité $Q_{\mathcal{M}} = \sum_1^8 n_t(r_t - \widehat{r}_t(\mathcal{M}))$.

****

Données catégorielles polytomiques.

****

Exercise 31 *Symptômes respiratoires des mineurs de charbon en fonction de l'âge x .*

Le tableau ci-dessous[9] donne les fréquences croisées de deux symptômes respiratoires chez les mineurs de charbon en fonction de l'âge x. Il y a 9 classes d'âge. Le premier symptôme S est à deux modalités : manque de souffle ($S = 1$) ou non ($S = 0$). Le deuxième, T, est aussi à deux modalités : présence de toux ($T = 1$) ou non ($T = 0$). Il y a 4 modalités (S, T), $F = \{oui, non\}^2$.

Manque de souffle	oui	oui	non	non	
Problème de toux	oui	non	oui	non	Total
20-24	9	7	95	1841	1952
25-29	23	9	105	1654	1791
30-34	54	19	177	1863	2113
35-39	121	48	257	2357	2783
40-44	169	54	273	1778	2274
45-49	269	88	324	1712	2393
50-54	404	117	245	1324	2090
55-59	406	152	225	967	1750
60-64	372	106	132	526	1136
Total	1827	600	1833	14022	18282

(1) Etudier les deux modèles Logit marginaux en S et en T.

(2) Analyser le modèle Logit-polytomique $Y = (S, T)$ à quatre modalités, x étant pris

(i) comme une variable de classe ;

(ii) comme variable réelle de régression (le milieu de la classe).

Tester si le modèle est à hasard proportionnel.

Exercise 32 *Influence de l'apport d'iode sur le développement de goitres*

Une recherche menée en Afrique sahélienne par l'Institut Santé et Développement (*Y. Le Roux, Rhône-Poulenc*) a eu pour objectif d'étudier l'influence de l'apport d'iode dans l'eau de forages sur le développement des goitres .

Il y a 5 niveaux de goitres $\{G1, G2, G3, G4, G5\}$, classés du plus bénin $G1$ (inexistant) au plus grave $G5$ (goitre irréversible) et 4 facteurs : X_1, le lieu (3 villages dont un témoin, $\{V_1, V_2, V_3\}$) ; X_2, le sexe $\{H, F\}$; X_3, l'apport d'iode (oui ou non) ; X_4, deux dates successives d'examen $\{-, +\}$.

Le nombre de croisements total possibles pour les 4 facteurs est de 24, mais seuls 12 sont observés. Les données sont les suivantes :

[9] Ashford et Sowden, *Multivariate Probit Analysis*, Biometrics 26, 535-546, 1970.

Village	Sexe	Iode	Jour	G1	G2	G3	G4	G5	Total
1	H	non	−	106	12	46	11	0	175
1	H	non	+	60	31	46	15	0	152
1	F	non	−	77	21	71	65	11	245
1	F	non	+	46	28	63	65	11	213
2	H	non	−	127	27	45	12	1	212
2	H	oui	+	145	28	19	1	1	194
2	F	non	−	69	21	65	50	2	207
2	F	oui	+	76	40	41	13	2	172
3	H	non	−	91	8	14	6	0	119
3	H	oui	+	94	14	10	0	0	118
3	F	non	−	42	18	45	34	4	143
3	F	oui	+	50	29	38	13	3	133

Les cinq effectifs égaux à 0 ont été portés à la valeur 0.1.

(1) Vérifier que les modalités $V \times I$ ainsi que $I \times J$ ne sont pas toutes observées et que seule l'interaction d'ordre 3, $V \times S \times J$, est identifiable. En déduire que pour les croisements observés, le modèle identifiable maximal est :

$$(\mathcal{S}) : V + S + I + J + V * S + V * J + S * I + S * J + V * S * J$$

(2) Analyser les modèles logistiques nominaux : $\log \frac{\pi(i|x)}{\pi(5|x)} = m_i(V, S, I, J)$, $i = 1, 4$:

(i) $(\mathcal{S})$: maximal ; (ii) $(\mathcal{S}^{-V})$: pas d'effet village ; (iii) $(\mathcal{A}) : V + S + I + J$ (additif)
(iv) $(\Omega) : S + I + J + S * I$; (v) $(\omega) : S + I + J$.

Montrer que leurs dimensions respectives sont 56, 24, 24, 20 et 16. Tester $(\mathcal{S}^{-V})$ dans $(\mathcal{S})$ et (ω) dans $(\mathcal{S})$. Quel modèle retenir sur la base du critère AIC ?

(3) Etudier les modèles Logit cumulés.

Chapitre 7

Modèle log-linéaire de table de contingence

Les modèles log-*linéaires* permettent de dégager des propriétés d'indépendance et/ou d'indépendance conditionnelle pour une loi discrète π sur un espace produit E fini. Si $E = I \times J$, π est une loi à deux facteurs (X, Y), si $E = I \times J \times K$, π est une loi à trois facteurs (X, Y, Z), et ainsi de suite. Par exemple, un modèle à 3 facteurs croise *Souffle* $\times$ *Age* $\times$ *Fumeur*, un autre à 4 facteurs croisera $CSP \times Etudes \times Sexe \times Etat\ Civil$. Dans ces modèles, tous les facteurs ont le même statut : il n'y a ni facteur à expliquer (ou endogène), ni facteur explicatif (ou exogène).

Un n-échantillon de π est résumé par les effectifs $(n_i, i \in I)$ d'apparition de chaque état : $(n_{ij}, (i,j) \in I \times J)$, $(n_{ijk}, (i,j,k) \in I \times J \times K)$ sont respectivement appelées des *tables de contingence* à deux facteurs, à trois facteurs.

Les modèles graphiques donnent une approche générale pour les tables de contingence avec un nombre de facteurs quelconque. Nous ne les aborderons pas ici.

7.1 Modèle *log*-linéaire

7.1.1 Un modèle linéaire pour $\log \pi$

Soit $\pi = (\pi_l)$ une loi *positive* sur $E = \{0, 1, 2, \cdots, L-1\}$:

$$\pi \textit{ est positive} : \forall l \in E, \pi_l > 0$$

Le modèle *complet*, sans contrainte sur π autres que celle d'être une probabilité ($\pi_i > 0$ et $\sum_E \pi_i = 1$), est de dimension $L-1$. La loi π peut être reparamétrée à partir des $(L-1)$ réels $\theta_l \in \mathbf{R}$ non-contraints : où $\mu(\theta) = -\log(1 + \sum_1^{L-1} \exp(\theta_l))$[1].

$$\theta_l = \log \frac{\pi_l}{\pi_0}, \text{ ou encore } \log \pi_l = \mu(\theta) + \theta_l,\ l = 1, L-1$$

est la *représentation* log-*linéaire* de π. La correspondance $\pi \leftrightarrow \theta$ est bijective, θ donnant en retour la loi π :

$$\pi_l = \frac{\exp \theta_l}{1 + \sum_{k=1}^{L-1} \exp \theta_k} \text{ pour } l = 0, L-1,$$

Les probabilités π_l sont caractérisées par leur proportionnalité à $\exp \theta_l$, avec $\theta_0 = 0$ ([2]). Pour le modèle complet, les $(L-1)$ paramètres θ_l sont libres.

[1] Il est important de noter que $\mu(\theta)$ n'est pas un paramètre du modèle : contrairement à une moyenne générale μ dans un modèle d'analyse de la variance, μ s'explicite ici à partir des autres paramètres θ.

[2] Une telle contrainte s'impose si on veut que le modèle soit identifiable : en effet, pour $c \in \mathbf{R}$, les paramètres $\{\theta_l, l = 0, L-1\}$ et $\{\theta_l + c, l = 0, L-1\}$ donnent la même loi π.

Soit $C(X) \in \mathbf{R}^{L-1}$ le *codage* de $X \in E$ ainsi défini : pour $l > 0$, $C_l(x) = \mathbf{1}(x = l)$ (c.a.d. 1 si $x = l$, et 0 sinon), 0 étant codé par le vecteur $\mathbf{0}$. X appartient à la famille *exponentielle* :

$$\pi_\theta(x) = \exp\{-\Psi(\theta) + \langle \theta, C(x) \rangle\} \tag{7.1}$$

avec $\Psi(\theta) = -\mu(\theta) = \log(1 + \sum_{k=1}^{L-1} \exp \theta_k)$ et $\langle \theta, C(x) \rangle = \sum_1^{L-1} \theta_l \mathbf{1}(x = l)$. Si $x = (x_1, x_2, \cdots, x_n)$ est un n-échantillon de X, si $N(x) = \sum_{i=1,n} C(x_i) = {}^t(n_1, n_2, \cdots, n_L)$ est le vecteur des effectifs d'apparition des états $l = 1, L-1$, la densité en x est :

$$f_\theta(x) = f_\theta(x_1, x_2, \cdots, x_n) = \exp\{-n\Psi(\theta) + \langle \theta, N(x) \rangle\}$$

7.1.2 Loi multinomiale et test de sous-modèle

Dans l'écriture précédente, $N = {}^t(N_1, N_2, \cdots, N_{L-1})$ est le vecteur des effectifs observés des états autres que 0. L'effectif des 0 est $N_0 = n - \sum_{l=1}^{L-1} N_l$. La variable $\mathbf{N} = (N_0, N_1, N_2, \cdots, N_{L-1})$ sur $\{0, 1, 2, \cdots, n\}^L$ est la loi *multinomiale* $\mathcal{M}(n; \pi_0, \pi_1, \cdots, \pi_{L-1})$ (cf. ch. 9, §9.2).

Soit $(\mathcal{M})$ un sous-modèle $\pi = r(\xi)$, où r est un changement de paramètre régulier[3] de $\mathbf{R}^m$ dans $\mathbf{R}^{L-1}$. Si $\widehat{\xi}$ est l'estimation du MV dans $(\mathcal{M})$, et $\widehat{\pi}$ celle du modèle complet, alors, asymptotiquement :

$$2(l_n(\widehat{\pi}) - l_n(\widehat{\xi})) \underset{\mathcal{M}}{\sim} \chi^2(L - m - 1)$$

Ce test est à rapprocher du test du χ^2 de spécification. Sous $(\mathcal{M})$, les deux statistiques sont asymptotiquement équivalentes.

Pour voir pourquoi la modélisation log-linéaire en θ s'accommode bien de la structure produit de E, nous allons d'abord examiner le cas où $E = I \times J$ est à deux facteurs, puis celui de 3 facteurs, avant de présenter le cadre général des modèles graphiques.

7.2 Tables de contingence à deux facteurs $E = I \times J$

I dénote le premier facteur (resp. J le second facteur) ainsi que l'ensemble des états $I = \{0, 1, \cdots, I-1\}$ du premier facteur (resp. $J = \{0, 1, \cdots, J-1\}$). La loi π une *positive* sur $E = I \times J$. On choisit $\mathbf{0} = (0, 0)$ comme état de référence de E. La représentation $\theta = (\theta_{ij} = \log \frac{\pi_{ij}}{\pi_{00}}, (i, j) \neq (0, 0))$ de π reste valable. Sans contrainte, θ parcourt tout $\mathbf{R}^{IJ-1}$. Imposer à θ d'appartenir à un sous-espace de $\mathbf{R}^{IJ-1}$ définit un sous-modèle.

7.2.1 Décomposition de θ en effets principaux et interactions

Définissons :

- Les $(I-1)$ *effets principaux* α : $\alpha_i = \theta_{i0}, i = 1, I-1$
- Les $(J-1)$ *effets principaux* β : $\beta_j = \theta_{0j}, j = 1, J-1$
- Les $(I-1)\times(J-1)$ *interactions* $(\alpha\beta)$: $(\alpha\beta)_{ij} = \theta_{ij} - \alpha_i - \beta_j = \theta_{ij} - \theta_{i0} - \theta_{0j}, i \times j \neq 0$.

$\delta = (\alpha, \beta, (\alpha\beta)) \in \mathbf{R}^{IJ-1}$ est un nouveau paramètrage[4] attribuant implicitement la valeur 0 aux paramètres α_0, β_0, et $(\alpha\beta)_{ij}$ dès que $i \times j = 0$. θ admet la décomposition :

$$\theta_{ij} = \alpha_i + \beta_j + (\alpha\beta)_{ij}, (i, j) \in I \times J$$

[3] r est de classe $\mathcal{C}^2$ et de jacobien $r'^{(1)}_\xi(\xi)$ de rang m.

[4] D'autres choix sont possibles. Par exemple, le choix de l'analyse de la variance est : $\alpha'_i = \theta_{i.} - \theta_{..}$, $\beta'_j = \theta_{.j} - \theta_{..}$, $(\alpha\beta)'_{ij} = \theta_{ij} - \theta_{i.} - \theta_{.j} + \theta_{..}$, avec $\theta_{i.} = \frac{1}{J}\sum_{j=0}^{J-1} \theta_{ij}$, $\theta_{.j} = \frac{1}{I}\sum_{i=0}^{I-1} \theta_{ij}$ et $\theta_{..} = \frac{1}{IJ}\sum_{i,j} \theta_{ij}$. Les notions d'interaction d'ordre 2 ou de modèle additif sont intrinsèques, indépendantes de ce choix : en effet $(\alpha\beta) \equiv 0 \Longleftrightarrow (\alpha\beta)' \equiv 0$.

La représentation log-linéaire des π_{ij} est :

$$\log \pi_{ij} = \mu(\delta) + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \text{ avec } \mu(\delta) = -\log\{1 + \sum_{(i,j)\neq(0,0)} \exp\theta_{ij}\}$$

La correspondance $\theta \leftrightarrow \delta$ et la paramétrisation en δ permet de dégager des sous-modèles traduisant des hypothèses d'indépendance conditionnelle.

La vraisemblance d'un n-échantillon (n_{ij}) s'écrit :

$$f_\delta(\mathbf{n}) = \exp\{\langle\alpha, n_{\circ+}\rangle + \langle\beta, n_{+\circ}\rangle + \langle(\alpha\beta), n_{\circ\circ}\rangle + n\mu(\delta)\}$$

où $\langle\alpha, n_{\circ+}\rangle = \sum_{i=1}^{I-1} \alpha_i n_{i+}$[5], $\langle\beta, n_{+\circ}\rangle = \sum_{j=1}^{J-1} \beta_j n_{+j}$, $\langle(\alpha\beta), n_{\circ\circ}\rangle = \sum_{i=1}^{I-1}\sum_{,j=1}^{J-1}(\alpha\beta)_{ij} n_{ij}$.

7.2.2 Modèle log-linéaire additif : indépendance de I et J

Ecrivant le modèle complet $\log\pi = a + b + a * b$ (a pour i, b pour j), le sous-modèle additif $(\mathcal{A})$ $\log : \pi = a + b$ annule les interactions $a * b$,

$$(\mathcal{A}) : (\alpha\beta)_{ij} = 0, i = 1, I-1,\ j = 1, J-1$$

$(\mathcal{A})$, de dimension $(I + J - 2)$[6], traduit l'indépendance de I et J : $\forall i, j$, $\pi_{ij} = \mu_i\nu_j$. Les estimations du MV sont $\widehat{\mu}_i = \frac{n_{i+}}{n}$ et $\widehat{\nu}_j = \frac{n_{+j}}{n}$. Pour n grand :

$$\sqrt{n}(\widehat{\mu} - \mu) \overset{loi}{\rightsquigarrow} \mathcal{N}_I(0, \Sigma(\mu)) \text{ et } \sqrt{n}(\widehat{\nu} - \nu) \overset{loi}{\rightsquigarrow} \mathcal{N}_J(0, \Sigma(\nu))$$

avec $\Sigma(\mu)_{ii} = \mu_i(1 - \mu_i)$ et $\Sigma(\mu)_{ij} = -\mu_i\mu_j$, et une forme analogue pour $\Sigma(\nu)$.

La loi asymptotique de $(\widehat{\mu}, \widehat{\nu})$ est gaussienne. Dans un cadre général, la covariance s'obtient en remarquant que $(n - n_{i+} - n_{+j} + n_{ij}, n_{i+} - n_{ij}, n_{+j} - n_{ij}, n_{ij})$ est une variable multinomiale $\mathcal{M}_4(n; (1-\mu_i)(1-\nu_j), \mu_i(1-\nu_j), (1-\mu_i)\nu_j, \mu_i\nu_j)$. Dans le cas particulier présent, cette covariance est nulle puisque $n_{i+} = \sum_k \mathbf{1}(X_k = i)$ est indépendant de $n_{+j} = \sum_k \mathbf{1}(Y_k = j)$: $\widehat{\mu}$ et $\widehat{\nu}$ sont indépendantes.

Le test asymptotique du RV pour $(\mathcal{A})$ repose sur la statistique

$$\Lambda = 2\{l_n(\widehat{\pi}) - l_n(\widehat{\mu}, \widehat{\nu})\} = 2\{l_n(\widehat{\theta}) - l_n(\widehat{\alpha}, \widehat{\beta})\} \overset{loi \text{ si } (\mathcal{A})}{\rightsquigarrow} \chi^2((I-1)(J-1))$$

Le RV vaut $\prod_{ij}\{\frac{n_{ij}}{n}\}^{n_{ij}} / \prod_{ij}\{\frac{\widehat{n}_{ij}}{n}\}^{n_{ij}}$, avec $\widehat{n}_{ij} = n\widehat{\mu}_i\widehat{\nu}_j = \frac{n_{i+}n_{+j}}{n}$. La statistique du test du chi 2 d'indépendance est asymptotiquement équivalente[7] à Λ :

$$\chi^2 = 2 \times \{\sum_{ij} n_{ij} \log \frac{n_{ij}}{\widehat{n}_{ij}}\} \text{ et } \Delta = \sum_{i,j} \frac{(n_{ij} - \widehat{n}_{ij})^2}{\widehat{n}_{ij}}$$

7.3 Tables de contingence à trois facteurs $E = I \times J \times K$

Soient $I = \{0, 1, \cdots, I-1\}$, $J = \{0, 1, \cdots, J-1\}$, $K = \{0, 1, \cdots, K-1\}$, les espaces d'état des 3 facteurs, $E = I \times J \times K$. Une probabilité générale π dépend de $(IJK - 1)$ paramètres. Prenant $\mathbf{0} = (0, 0, 0)$ comme état de référence, π est en bijection avec le paramètre θ de $\mathbf{R}^{IJK-1}$, $\theta = (\theta_{ijk} = \log\frac{\pi_{ijk}}{\pi_0}, (i, j, k) \neq 0)$.

[5] Le remplacement d'un indice par + signifie qu'on a effectué la sommation relativement à cet indice : par exemple, $n_{i+} = \sum_{j=0}^{J-1} n_{ij}$.

[6] $(\mathcal{A})$ est aussi le modèle : $\pi_{ij} = \mu_i\nu_j$. Il y a $(I-1)$ paramètres libres μ_i et $(J-1)$ paramètres libres ν_j.

[7] Ce résultat s'obtient en effectuant un développement limité pour ces statistiques autour de (μ_i, ν_j), la limite des rapports $(\frac{n_{i+}}{n}, \frac{n_{+j}}{n})$), ainsi que pour les produits $\mu_i\nu_j$ valeurs-limites de $\frac{n_{ij}}{n}$.

7.3.1 Décomposition de θ : effets principaux, interactions

Définissons les effets principaux, interactions d'ordre 2 et interactions d'ordre 3 :

- *effets principaux* : $\alpha_i = \theta_{i00}$, $\beta_j = \theta_{0j0}$ et $\gamma_k = \theta_{00k}$.
- *interactions d'ordre 2* : $(\alpha\beta)_{ij} = \theta_{ij0} - \theta_{i00} - \theta_{0j0}$; de façon analogue $(\alpha\gamma)_{ik}$ et $(\beta\gamma)_{jk}$
- *interactions d'ordre 3* : $(\alpha\beta\gamma)_{ijk} = \theta_{ijk} - \theta_{ij0} - \theta_{0jk} - \theta_{i0k} + \theta_{i00} + \theta_{0j0} + \theta_{00k}$.

Chacun des effets $\alpha_i, \beta_j, \gamma_k, (\alpha\beta)_{ij}, \cdots, (\alpha\beta\gamma)_{ijk}$ vaut 0 dès que l'un des indices i, j ou k vaut 0. Il y a $(I-1)$ effets principaux i, $(J-1)$ effets j, $(K-1)$ effets k, $(I-1)(J-1)$ interactions $i \times j$, $(I-1)(K-1)$ interactions $i \times k$, $(J-1)(K-1)$ interactions $j \times k$, et $(I-1)(J-1)(K-1)$ interactions d'ordre 3.

Le nouveau paramètre $\delta = (\alpha, \beta, \gamma, (\alpha\beta), (\alpha\gamma), (\beta\gamma), (\alpha\beta\gamma))$ est de dimension $(IJK - 1)$. La correspondance $\theta \leftrightarrow \delta$ est bijective :

$$\log \pi_{ijk} = \mu(\theta) + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk}$$

La paramétrisation en δ permet de dégager des sous-modèles[8]. Dans le modèle complet, l'EMV est $\widehat{\pi}_{ijk} = \frac{n_{ijk}}{n}$ et les statistiques canoniques du modèle exponentiel sont :

- n_{i++}, n_{+j+}et n_{++k} pour α_i, β_j et γ_k
- n_{ij+}, n_{+jk} et n_{i+k} pour $(\alpha\beta)_{ij}, (\beta\gamma)_{jk}$ et $(\alpha\gamma)_{ik}$
- n_{ijk} pour $(\alpha\beta\gamma)_{ijk}$.

7.3.2 Sous-modèles et interprétations

Intéressons nous aux *sous-modèles hiérarchiques*[9].

Pas d'interaction d'ordre 3 : $\log \pi = a + b + c + a*b + b*c + c*a$

Revenant à la forme exponentielle de π, ce modèle correspond à $\pi_{ijk} = \lambda_{ij}\mu_{jk}\nu_{ik}$. Il n'y a pas d'interprétation particulière en termes d'indépendance ou d'indépendance conditionnelle.

Toute l'information est contenue dans les statistiques $(n_{ij+}, n_{+jk}, n_{i+k})$. θ est de dimension $\{IJ + IK + JK - I - J - K\}$ et les équations du MV sont :

$$n\lambda_{ij}(\widehat{\theta}) = n_{ij+}, n\mu_{jk}(\widehat{\theta}) = n_{+jk}, \nu_{ik}(\widehat{\theta}) = n_{i+k}$$

La résolution de ces équations n'est pas explicite, mais l'existence et l'unicité de $\widehat{\theta}$ est garantie par la log-concavité de la vraisemblance.

Absence d'une interaction d'ordre 2 : $\log \pi = a + b + c + a*b + a*c$

De dimension $I(J + K - 1) - 1$, ce modèle traduit que, conditionnellement à a, les caractères b et c sont indépendants, ce que l'on note : $(b \perp c \mid a)$. La loi π admet la décomposition :

$$\pi_{ijk} = \lambda_i \mu_{j|i} \nu_{k|i}$$

λ_i est la probabilité marginale en i, $\mu_{j|i}$ est la probabilité conditionnelle de j à i fixé et $\nu_{k|i}$ celle de k à i fixé. Chacune des probabilités de cette décomposition s'estime à partir des statistiques suffisantes n_{i++}, n_{ij+} et n_{i+k}. L'EMV de π est $\widehat{\pi}_{ijk} = \frac{n_{ij+} n_{i+k}}{n \times n_{i++}}$.

[8] Là aussi, on pourrait choisir une autre définition des effets principaux α', β', γ', des interactions d'ordre 2 $(\alpha\beta)', \ldots$ et de l'interaction d'ordre 3 $(\alpha\beta\gamma)'$. Ce choix est secondaire dans la mesure où : (i) la notion d'interaction d'ordre 3 est intrinsèque ; (ii) en absence d'interaction d'ordre 3, chaque interaction d'ordre 2 est intrinsèque, par exemple : $\{(\alpha\beta) \equiv 0$ et $(\alpha\beta\gamma) \equiv 0\} \Leftrightarrow \{(\alpha\beta)' \equiv 0$ et $(\alpha\beta\gamma)' \equiv 0\}$; (iii) enfin, l'additivité est intrinsèque.

[9] $(\mathcal{H})$ est hiérarchique chaque fois que l'annulation d'une interaction entraîne celle de toute sur-interaction : par exemple, dans un modèle à 3 facteurs, si $(\alpha\beta) \equiv 0$, alors $(\alpha\beta\gamma) \equiv 0$.

Une seule interaction d'ordre 2 : $\log \pi = a + b + c + a * b$

De dimension $\{IJ + K - 2\}$, ce modèle additif $(a * b) + c$ traduit l'indépendance entre (a, b) et c : $(a, b) \perp c$. La décomposition de π est : $\pi_{ijk} = \lambda_{ij}\mu_k$. Les statistiques suffisantes sont n_{ij+} et n_{++k} et l'EMV $\widehat{\pi}_{ijk} = \frac{n_{ij+}n_{++k}}{n^2}$.

Modèle additif $(\mathcal{A})$ **:** $\log \pi = a + b + c$

De dimension $I+J+K-3$, ce modèle traduit l'indépendance des trois caractères, $\pi_{ijk} = \lambda_i\mu_j\nu_k$. Les statistiques suffisantes sont $n_{i++}, n_{+j+}, n_{++k}$, l'EMV $\widehat{\pi}_{ijk} = \frac{n_{i++}n_{+j+}n_{++k}}{n^3}$.

Modélisation exogène d'un facteur quantitatif

Si i est repérée par $x_i \in \mathbf{R}$, on peut proposer des modèles exogènes spécifiant l'effet principal et/ou une interaction relative à x. Par exemple, une modélisation régressive de α_i et des $(\alpha\beta)_{ij}$ s'écrit :

$$(\mathcal{P}_x) : \log \pi_{ijk} = \mu(\theta) + \{\delta \times (x_i - x_0) + \beta_j + (x_i - x_0)\nu_j\} + \gamma_k$$

$\theta = (\delta, (\beta_j), (\nu_j), (\gamma_k))$. $(\mathcal{P}_x)$, de dimension $\{2J + K - 2\}$, appartient à la famille exponentielle de statistiques canoniques $\sum_{i=1,I-1} n_{i++}x_i$ et $(\sum_{i=1,I-1} n_{ij+}x_i)_j$ pour $(\delta, (\nu_j)_j)$.

7.4 Modèle log-linéaire et modèle Logit

Pour fixer les idées, considérons une table de contingence à trois facteurs $A \times B \times C$, A étant binaire, $A = \{0, 1\}$. Le *modèle explicatif* $(A \mid b, c)$ est caractérisé par les probabilités $\mu(b, c) = P(A = 1 \mid b, c)$. Un modèle log-linéaire pour $A \times B \times C$ *induit* un modèle *Logit*. Ce dernier s'obtient en calculant le rapport des chances par déconditionnement :

$$\log \frac{P(A = 1 \mid b, c)}{P(A = 0 \mid b, c)} = \log \frac{P(A = 1, b, c)}{P(A = 0, b, c)} = \theta_{1bc} - \theta_{0bc}$$

Par exemple, le modèle complet induit le modèle Logit complet :

$$Logit\{\mu(b, c)\} = \alpha_1 + (\alpha\beta)_{1b} + (\alpha\gamma)_{1c} + (\alpha\beta\gamma)_{1bc}$$

avec $\alpha = \alpha_1$, $(\alpha\beta)_{1b} = \beta'_b$, $(\alpha\gamma)_{1c} = \gamma'_c$ et $(\alpha\beta\gamma)_{1bc} = (\beta\gamma)'_{bc}$. On a la correspondance :

Modèle log-linéaire	Logit associé $(A \mid B, C)$
$a + b + c$ *ou* $a + bc$	α
$ab + c$ *ou* $ab + bc$	$\alpha + \beta'_b$
$ab + ac + bc$	$\alpha + \beta'_b + \gamma'_c$
abc	$\alpha + \beta'_b + \gamma'_c + (\beta\gamma)'_{bc}$

On définit de façon analogue les modèles *Logit*-polytomiques conditionnels si A est polytomique.

7.5 Observations de tables sur différents groupes

Supposons qu'on observe X, à K facteurs, sur différents groupes $G_1, G_2, \cdots, G_L$, chaque groupe étant homogène pour $X \in I$. Deux situations se présentent :

(i) les tirages sont effectués au hasard dans $G = \cup_{l=1}^{L} G_l$: on peut alors considérer que l'on observe un nouveau caractère $Y = (X, G)$ à $(K+1)$ facteurs, G étant à L modalités. On applique alors les méthodes habituelles pour une table de contingence à $(K + 1)$ facteurs.

Le modèle complet est de dimension $IL-1$. Le test d'homogénéité des L groupes reposera sur un χ^2 à $I(L-1)$ ddl.

(ii) si au contraire on fixe au préalable les effectifs $n(1)$, $n(2)$, $\cdots$, $n(L)$ de l'échantillon dans chacun des groupes, le sondage est *exogène*, le groupe jouant le rôle de facteur exogène. On étudie alors L tables de contingence autonomes. Chaque table est modélisée par des probabilités $\{\pi_i(l), i \in I\}, l \in G$. Le modèle complet est de dimension $L(I-1)$. Si les observations sont indépendantes de groupe à groupe, la log-vraisemblance sera la somme des log-vraisemblances sur chaque groupe et l'étude statistique se fera de façon standard. Par exemple, l'identité des L groupes s'obtiendra à partir d'une statistique du χ^2 à $(L-1)(I-1)$ degré de liberté.

7.6 Exercices : modèle log-linéaire et table de contingence

T.P. : Utiliser le logiciel `R` pour répondre aux exercices suivants.

Exercise 33 *Influence d'une vaccination sur la poliomyélite.*

Les 174 observations suivantes[10] croisent les trois caractères p : être atteint ou non par la poliomyélite, a : la classe d'âge, v : être vacciné ou non,

a	0-4	0-4	5-9	5-9	10-14	10-14	15-19	15-19	20-39	20-39	>40	>40
v	oui	non	oui	non	oui	non	oui	non	oui	non	oui	non
$\overset{(p)}{non}$	20	10	15	3	3	3	7	1	12	7	1	3
$\overset{(p)}{oui}$	14	24	12	15	2	2	4	6	3	5	0	2

(1) *Sondage endogène :* l'âge est un facteur au même titre que v et p.

(i) Identifier le modèle complet $(\mathcal{S})$ et calculer sa log-vraisemblance.

(ii) Tester (ω), l'indépendance entre la maladie et la vaccination à âge fixé.

(2) *Sondage exogène* : on considère que l'âge est un facteur exogène et que les effectifs $n_1 = 68$, $n_2 = 45, \cdots$, $n_6 = 6$ de l'échantillon dans les différentes classes d'âge ont été préalablement fixés. On dispose donc de 6 tables de contingences indépendantes en $v \times p$.

(i) Soit $(\mathcal{S}')$ le nouveau modèle complet et (ω') celui traduisant, pour chaque classe d'âge, l'indépendance de v et p. Tester (ω').

(ii) On considère que x est un facteur quantitatif, les 5 premières classes étant repérées par leur centre, la dernière par 45. Codant v, p en $\{0,1\}$, une table 2×2 dépend de 3 paramètres. Ecrire $(\mathcal{S}')$ sous forme régressive en $v \times p$. On considère le sous-modèle $(\mathcal{S}'_{\mathcal{R}})$:

$$(\mathcal{S}'_{\mathcal{R}}) : \log \pi(v, p \mid x) = \mu + a \times v \times x + b \times p \times x + c \times v \times p \times x$$

Interpréter $(\mathcal{S}'_{\mathcal{R}})$. Construire le test de $(\mathcal{S}_{\mathcal{R}})$ dans $(\mathcal{S})$.

Exercise 34 *Table à trois facteurs : Souffle×Age×Fumeur.*

Le tableau ci-dessous[11] donne la fréquence des problèmes de souffle (s) en fonction de l'âge (a) et du fait de fumer ou non (f) :

Age (a)	Fumeur (f)	S. normal	S. anormal
<40	non	577	34
<40	oui	682	57
>40	non	164	4
>40	oui	245	74

[10] Chin et altri, 1961, *The influence of salk vaccination on the epidemic pattern of the spread of the virus in the community*, Amer. J. Hyg. 73, 67-94.

[11] Forthofer et Lehnen, *Lifetime learning Publications*, Public program Analysis, Belmont CA 94002, 1981.

(i) Tester l'indépendance des trois caractères.

(ii) Sélectionner un modèle par une procédure ascendante.

(iii) Considérant les deux groupes d'âge comme deux populations distinctes et le sondage comme exogène, tester l'homogénéité des deux populations.

(iv) Etudier le modèle logistique $(s \mid a, f)$ expliquant les problèmes de souffle.

Exercise 35 *Le chômage affecte t-il identiquement les deux membres d'un couple ?*

$X_i \in \{0, 1, 2\}$ est le nombre de chômeur d'un couple i. L'observation d'un 1000-échantillon de X sur une population homogène donne :

Valeurs de X	0	1	2
effectifs	810	170	20

(1) Donner l'EMV des paramètres π_0, π_1 et π_2 de la loi π de X. Quelle est la loi asymptotique de cet estimateur ?

Soit (ω) la sous-hypothèse : *le chômage touche de façon identique et indépendante les deux membres du couple.*

(2) Ecrire la loi de X sous (ω) en fonction de θ ainsi que la contrainte non-linéaire sur π caractérisant (ω). En déduire un test de (ω).

(3) Estimer θ par le MV. Tester (ω) en utilisant le test du RV.

(4) Tester (ω) en utilisant le test du χ^2 d'ajustement.

Chapitre 8

Annexes

8.1 Variable aléatoire à valeur dans $\mathbf{R}^n$

Soient $X \in \mathbf{R}^p$ et $Y \in \mathbf{R}^q$ deux variables aléatoires vectorielles (v.a.v.) de carré intégrable, $A : \mathbf{R}^p \longrightarrow \mathbf{R}^{p'}$ et $B : \mathbf{R}^q \longrightarrow \mathbf{R}^{q'}$ deux applications linéaires. On a les propriétés suivantes :

- $E(AX) = AE(X)$ (linéarité de l'espérance).
- $Cov(AX, BY) = ACov(X, Y)\ {}^tB$ (bilinéarité de la covariance).
- $Var(AX) = AVar(X){}^tA$, et si $a \in \mathbf{R}^p$, $Var({}^taX) = {}^taVar(X)a$.

Loi de $Y = \Phi(X)$ image d'une variable X à densité f_X.

Soient U et V deux ouverts de $\mathbf{R}^p$, $\Phi : U \longrightarrow V$ une bijection continûment dérivable sur U ($\Phi \in \mathcal{C}^1(U)$) telle que $\Phi^{-1} \in \mathcal{C}^1(V)$, où $\Phi^{-1} : V \longrightarrow U$ est l'application réciproque de Φ. Presque sûrement, le jacobien $J\Phi(x) = \left(\frac{\partial \Phi_i}{\partial x_j}(x)\right)_{i,j=1,p}$ de Φ est inversible sur U.

Soit X une v.a.v. à valeur dans $U \subset \mathbf{R}^p$ de densité f_X. Alors, la v.a.v. image $Y = \Phi(X)$ admet une densité f_Y concentrée sur V donnée par :

$$f_Y(y) = \{|\det J\Phi(x)|\}^{-1} f_X(x), \text{ avec } x = \Phi^{-1}(y)$$

Exemple 1 : si A est une bijection linéaire sur $\mathbf{R}^p$, on a pour $\varepsilon = AX$ et $V = AU$:

$$f_\varepsilon(y) = \{\det A\}^{-1} f_X(A^{-1}y)$$

Si ε est un BB gaussien, MV et MC différeront pour l'estimation des paramètres de X dès que $\det A \neq 1$. En effet :

$$\log f_X(x) = \log(|\det A|) - \frac{1}{2\sigma_\varepsilon^2} \sum_1^p \varepsilon_i^2(x, A) - \frac{p}{2} \log(2\pi)$$

Exemple 2 : $X > 0$ est une loi log-normale de paramètre (m, σ^2) si $\log X$ est une loi normale $\mathcal{N}(m, \sigma^2)$. La densité de X est concentrée sur$]0, +\infty[$, de densité

$$f(x) = \mathbf{1}(x > 0) \frac{1}{\sqrt{2\pi}x} \exp - \frac{(\log x - m)^2}{2\sigma^2}$$

8.2 Loi gaussienne multimensionnelle

• La densité de $X \sim \mathcal{N}(m, \sigma^2)$, la gaussienne réelle de moyenne m et de variance $\sigma^2 > 0$ est $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp - \frac{1}{2} \frac{(x-m)^2}{\sigma^2}$. La fdr Φ *normale réduite* $\mathcal{N}(0, 1)$ vérifie $\Phi(-x) = 1 - \Phi(x)$; elle est tabulée pour $x \geq 0$.

• La *loi gaussienne multidimensionnelle* X à valeur dans $\mathbf{R}^p$, de moyenne $m \in \mathbf{R}^p$ et de matrice de covariance régulière $\Sigma = p \times p$ admet pour densité :

$$f(x) = (2\pi)^{-\frac{p}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp -\frac{1}{2}\ {}^t(x-m)\Sigma^{-1}(x-m)$$

Cette loi est caractérisée par sa moyenne et sa variance.

Soit $X = {}^t({}^tX(1), {}^tX(2))$ une décomposition de X en deux composantes de dimensions respectives p_1 et p_2. Notons $m(i)$, Σ_{ij}, $i,j = 1,2$ les décompositions de m et de Σ associées. On a les propriétés :

(1) Si $A : \mathbf{R}^p \longrightarrow \mathbf{R}^q$ est linéaire, alors $AX \sim \mathcal{N}_q(Am, A\Sigma {}^tA)$.

(2) Lois marginales : $X(1) \sim \mathcal{N}_{p_1}(m(1), \Sigma_{11})$ et $X(2) \sim \mathcal{N}_{p_2}(m(2), \Sigma_{22})$.

(3) $X(1)$ et $X(2)$ sont indépendantes si et seulement si elles sont non-corrélées : $\Sigma_{12} = 0$.

(4) La loi de $X(1)$ conditionnelle à $X(2) = x_2$ est gaussienne :

$$\mathcal{L}oi(X(1) \mid X(2) = x_2) \sim \mathcal{N}_{p_1}(m(1) + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - m(2)), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$$

L'espérance conditionnelle est affine en x_2 ; la variance est indépendante de x_2.

• *Couple* $(X,Y) \sim \mathcal{N}_2(m, \Sigma)$ où $Var(X) = \sigma_1^2$, $Var(Y) = \sigma_2^2$, $\rho = Corr(X,Y)$:

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp -\frac{1}{2(1-\rho^2)}\{\widetilde{x}^2 - 2\rho\widetilde{x}\widetilde{y} + \widetilde{y}^2\} \text{ où } \widetilde{z} = \frac{z - E(z)}{\sigma(z)} \\ \mathcal{L}oi(Y \mid X = x) &\sim \mathcal{N}(m_2 + \rho\sigma_2\widetilde{x}, \sigma_2^2(1-\rho^2)) \text{ et } E(Y \mid X = x) = m_2 + \rho\sigma_2\widetilde{x} \end{aligned}$$

X et Y sont indépendantes si et seulement si elles sont non-corrélées. L'espérance conditionnelle définit la droite de régression de Y en $X = x$.

8.3 Lois déduites : Chi2, Student et Fischer

• La *loi du Chi*2 à p degré de liberté (noté χ_p^2) est la loi de la somme :

$$\chi_p^2 \sim Z = Y_1^2 + Y_2^2 + \cdots + Y_p^2 \text{ , où les } (Y_i) \text{ sont } i.i.d.\ \mathcal{N}(0,1)$$

Cette loi est d'espérance p, de variance $2p$. Le carré d'une loi gaussienne réduite suivant une loi $\Gamma(\frac{1}{2}, \frac{1}{2})$, la densité d'un χ_p^2 est celle de la loi $\Gamma(\frac{p}{2}, \frac{1}{2})$. Les quantiles $q(p,\alpha)$ définis par $P(\chi_p^2 > q(p,\alpha)) = \alpha$ sont tabulés. On a les deux propriétés suivantes :

(1) Si $X \sim \mathcal{N}_p(0, \Sigma)$ et si Σ est inversible, alors ${}^tX\Sigma^{-1}X \sim \chi_p^2$. Cette propriété est à la base du test de Wald.

(2) Soit $X = (X_i)_{i=1,n}$ un n-échantillon d'une loi $\mathcal{N}(m, \sigma^2)$, $\overline{X}$ la moyenne arithmétique des X, et $S^2 = \frac{1}{(n-1)} \sum_1^n (X_i - \overline{X})^2$ la variance empirique de l'échantillon. Alors :

(i) $\overline{X} \sim \mathcal{N}(m, \frac{\sigma^2}{n})$; (ii) $S^2 \sim \chi_{n-1}^2$; (iii) $\overline{X}$ et S^2 sont indépendants.

Cette propriété conduit à la définition de la loi de Student et des tests gaussiens.

• *Loi de Student* à p ddl : c'est la loi d'un quotient

$$T_p = \frac{X}{\sqrt{Z/p}} \text{ où } X \sim \mathcal{N}(0,1) \text{ est indépendant de } Z, \text{ un } \chi_p^2$$

La distribution de T_p est symétrique, tabulée pour $x > 0$. Pour n grand ($n > 60$), la distribution de Student s'approche de celle de la variable normale réduite.

• La *loi de Fischer* à p et q ddl est la loi du quotient

$$F_{p,q} = \{\chi_p^2/p\}/\{\chi_q^2/q\}$$

où les deux χ^2 sont indépendants ; p est le nombre de ddl du numérateur, q du dénominateur. La distribution des lois de Fischer est tabulée à partir de ses quantiles. Notons que $(F_{p,q})^{-1}$ est une loi de Fischer $F_{q,p}$: si $f(p,q;\alpha)$ est le α-quantile de $F_{p,q}$ ($P(F_{p,q} > f(p,q;\alpha)) = \alpha$), alors $P(F_{q,p} > \{f(p,q;\alpha)\}^{-1}) = 1-\alpha$, c'est-à-dire que $f(q,p;1-\alpha) = \{f(p,q;\alpha)\}^{-1}$. Le carré d'une loi de Student T_p est la loi $F_{1,p}$.

• *Loi du Chi2 décentrée* : si $Y_i \sim \mathcal{N}(m_i, 1)$ sont p gaussiennes indépendantes, et si $\lambda^2 = \sum_1^p m_i^2$, la loi de la somme des carrés,

$$Z = \sum_1^p Y_i^2 \sim \chi'^2(p, \lambda^2)$$

ne dépend que de p et de λ^2 : c'est la loi du $Chi2$ décentrée, $\chi'^2_p(\lambda^2)$, à p ddl et de paramètre de non-centralité λ^2. A u et p fixés, $P(Z < u)$ est croissante en λ : ces lois permettent donc d'étudier la *fonction puissance* d'un test du $Chi2$. Elles sont tabulées dans des ouvrages spécialisées.

Si le numérateur du Fischer $F_{p,q}$ est remplacé par un $\chi'^2_p(\lambda)$, la loi du quotient, qui ne dépend que de p, q et λ, est la *loi de Fischer décentré* $F'_{p,q}(\lambda)$. Cette distribution intervient dans le calcul de la puissance du test de Fischer. Deux types de tables décrivent ces lois : les tables de Hartley-Pearson donnent la courbe de puissance à niveau α et p fixés ; les tables de Fox donnent, à niveau α et puissance β fixés, la valeur correspondante des λ pour les différents ddl p et q.

8.4 Convergence en probabilité, convergence en loi

Trois types de convergence sont utilisés dans ce livre. Rappelons-en les définitions et propriétés principales (cf. [**?**], [**?**] et [**?**]). Soit (X_n) une suite de v.a.v à valeur dans $\mathbf{R}^p$, X une v.a.v. à valeur dans le même espace, $\|.\|$ la norme euclidienne sur $\mathbf{R}^p$, $|.|$ la valeur absolue sur $\mathbf{R}$.

Convergence en probabilité. (X_n) converge en probabilité vers X si :

$$\forall \varepsilon > 0,\ P(\|X_n - X\| > \varepsilon) \underset{n}{\longrightarrow} 0\,;\ \text{on note : } X_n \xrightarrow{\Pr} X$$

Cette condition équivaut à la convergence en probabilité coordonnée par coordonnée.

Convergence en moyenne quadratique. Supposons que (X_n) et X soient de carré intégrable (chaque coordonnée est de variance finie).

(X_n) converge en moyenne quadratique vers X si $E(\|X_n - X\|^2) \longrightarrow 0 : X_n \xrightarrow{mq} X$.

La convergence en moyenne quadratique entraîne la convergence en probabilité. C'est une conséquence de l'*inégalité de Bienaymé-Tchebichev* qui dit que si X est une v.a. réelle (v.a.r.) de variance finie et si $a > 0$, alors :

$$P(|X - E(X)| \geq a) \leq \frac{Var(X)}{a^2}$$

La convergence en loi. Commençons par définir la *convergence faible* d'une suite (P_n) de probabilités sur les boréliens $\mathcal{B}(\mathbf{R}^p)$ de $\mathbf{R}^p$: (P_n) converge faiblement vers P si $P_n(A) \longrightarrow P(A)$ pour tout borélien A tel que $P(\partial A) = 0$, où ∂A est la frontière topologique de A. Cette convergence est équivalente à la propriété suivante :

$$\forall f : \mathbf{R}^p \longrightarrow \mathbf{R} \text{ continue et bornée : } \int f dP_n \underset{n}{\longrightarrow} \int f dP$$

Sur $\mathbf{R}$, la convergence faible équivaut à :

$$\forall a \text{ réel t.q. } P(\{a\}) = 0 : P_n(]-\infty, a]) \longrightarrow P(]-\infty, a])$$

Si P est diffuse (P ne charge pas les points), il n'est pas nécessaire de préciser la condition $P(\{a\}) = 0$ puisque X ne charge aucun point.

Soit P_X la loi de probabilité d'une v.a.v. X à valeur dans $\mathbf{R}^p$. Pour $(X_n), X$ des v.a.v., notons $P_n = P_{X_n}$ et $P = P_X$. On dira que (X_n) *converge en loi* vers X si P_n converge faiblement vers P. On note : $X_n \xrightarrow{loi} X$.

Dans le cas de v.a.r., notant F_n et F les fonctions de répartition de X_n et de X, la convergence en loi équivaut à :

$$\lim F_n(x) = F(x) \text{ en tout } x \text{ point de continuité de } F \tag{8.1}$$

Si la loi de X est diffuse, il est inutile de préciser la condition de continuité en x.

Convergence en probabilité et convergence en loi.

(1) *Théorème de Slutsky* : Si $X_n \xrightarrow{\Pr} X$ et si $f : \mathbf{R}^p \longrightarrow \mathbf{R}^q$ est continue, alors $f(X_n) \xrightarrow{\Pr} f(X)$.

(2) La convergence en probabilité entraîne la convergence en loi.

(3) *Une CS de convergence en probabilité* : soit (X_n) une suite de v.a.r. de variances finies. Si $E(X_n) \longrightarrow a \in \mathbf{R}$, et $Var(X_n) \longrightarrow 0$, alors $X_n \xrightarrow{\Pr} a$.

(3) Si $f : \mathbf{R}^p \longrightarrow \mathbf{R}^q$ est continue, alors $f(X_n) \xrightarrow{loi} f(X)$ dès que $X_n \xrightarrow{loi} X$.

(4) Pour des v.a. à valeur dans $\mathbf{R}^p$: $X_n \xrightarrow{loi} X$ équivaut à : $\forall a \in \mathbf{R}^p, {}^t aX_n \xrightarrow{loi} {}^t aX$.

(5) Pour des v.a. à valeur dans $\mathbf{R}^p$: si $X_n \xrightarrow{loi} X$ et $Y_n \xrightarrow{loi} a$, alors $X_n + Y_n \xrightarrow{loi} X + a$ et ${}^t X_n Y_n \xrightarrow{loi} {}^t aX$.

(6) *Théorème de Cramer* : soient (A_n), (B_n) deux suites de matrices aléatoires de dimensions respectives $p \times q$ et $q \times r$. Si $A_n \xrightarrow{loi} A$ et $B_n \xrightarrow{\Pr} B$, où B est une matrice constante, alors $A_n B_n \xrightarrow{loi} AB$.

(7) La *Delta-méthode : convergence en loi de $F(X_n)$ pour F non-linéaire.*
Soient (X_n) une suite de v.a.v. à valeur dans $\mathbf{R}^p$, (a_n) une suite positive tendant vers $+\infty$ telle que, pour un $\theta \in \mathbf{R}^p$ on ait : $\sqrt{a_n}(X_n - \theta) \xrightarrow{loi} \mathcal{N}_p(0, \Sigma)$. Soit $F : \mathbf{R}^p \longrightarrow \mathbf{R}^q$ une application de classe $\mathcal{C}^2$ au voisinage de θ. Alors, pour la matrice $p \times q$, $J(\theta) = \left(\frac{\partial F_i}{\partial \theta_j}(\theta)\right)$:

$$\sqrt{a_n}(F(X_n) - F(\theta)) \xrightarrow{loi} \mathcal{N}_q(0, J(\theta)\Sigma\, {}^t J(\theta))$$

Pour $q = 1$, ${}^t J(\theta)$ est le vecteur gradient de F.

8.5 Loi des grands nombres. Théorème central limite

Loi faible des grands nombres dans L^1(1). Soit (X_n) une suite de v.a.r. i.i.d. telle que $E(|X_1|) < \infty$. Alors :

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\Pr} m = E(X_1)$$

Loi faible des grands nombres dans L^2. Soit (X_n) une suite de v.a.r. non-corrélées, de même espérance $E(X_n) = m$, de variances bornées : $\forall n, Var(X_n) \leq \sigma^2 < \infty$. Alors :

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{m.q.} m = E(X_1)$$

[1] La loi forte des grands nombres dit que, sous les mêmes hypothèses, la convergence a lieu presque sûrement, c'est-à-dire partout en dehors d'un ensemble de probabilité 0.

Théorème central limite (T.C.L.). Soit (X_n) une suite de v.a.r. i.i.d., d'espérance μ et de variance finie $\sigma^2 = Var(X_1) > 0$. Posons : $S_n = \sum_{i=1}^n X_i$, $\overline{X}_n = S_n/n$. Alors :

$$\frac{(S_n - n\mu)}{\sigma\sqrt{n}} \xrightarrow{loi} \mathcal{N}(0,1) \text{ ou encore : } \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{loi} \mathcal{N}(0,1)$$

Forme vectorielle du TCL : Soit (X_n) une suite de v.a.v. i.i.d. d'espérance $\mu \in \mathbf{R}^p$ et de covariance Σ régulière ; alors :

$$\frac{(S_n - n\mu)}{\sqrt{n}} \xrightarrow{loi} \mathcal{N}_p(0,\Sigma) \text{ où encore } \Sigma^{-\frac{1}{2}}\frac{(S_n - n\mu)}{\sqrt{n}} \xrightarrow{loi} \mathcal{N}_p(0,I_p)$$

Le théorème central limite de Lyapunov. Soit (X_n) une suite de v.a.r. indépendantes et centrées, telle que pour tout n, $\mu_{3n} = E(|X_n|^3) < \infty$. Soient $s_n^2 = Var(S_n) = \sum_1^n Var(X_i)$ et $\kappa_n^3 = \sum_1^n E(|X_i|^3)$:

$$\text{Si } \frac{\kappa_n}{s_n} \longrightarrow 0, \text{ alors } s_n^{-1} S_n \xrightarrow{loi} \mathcal{N}(0,1)$$

La condition $\frac{\kappa_n}{s_n} \longrightarrow 0$ est vérifiée si les moments d'ordre 3 sont bornés et si la suite $(\frac{s_n}{n})$ est minorée positivement.

Ce résultat est inchangé si on considère *un tableau triangulaire* de v.a.r. $\{\{X_{i,n}, i = 1,n\}, n \in N\}$, avec $S_n = \sum_{i=1}^n X_{i,n}$ et si on remplace les conditions sur $(X_i, i = 1,n)$ par les mêmes conditions sur $\{X_{i,n}, i = 1,n\}$.

8.6 Famille exponentielle de modèles

Concernant cette annexe et les deux suivantes, on pourra consulter [4], [**?**] ou [**?**].

8.6.1 Définitions

Soit $(E,\mathcal{E})$ un espace d'état mesurable, μ une mesure σ-finie[2] et $(P_\theta, \theta \in \Theta)$ un modèle statistique (c.a.d. une famille de lois de probabilité) sur cet espace, Θ étant un ouvert non-vide de $\mathbf{R}^p$.

Definition 4 ***Modèle exponentiel :*** *$(P_\theta, \theta \in \Theta)$ est un modèle exponentiel si pour tout θ, P_θ admet une densité par rapport à μ de la forme :*

$$f_\theta(x) = C(\theta)h(x)\exp\langle T(x), Q(\theta)\rangle \tag{8.2}$$

avec $T(x)$ et $Q(\theta) \in \mathbf{R}^r$, et $\langle u,v\rangle = \sum_i u_i v_i$ le produit scalaire sur $\mathbf{R}^r$.

$T = (T_1, T_2, \cdots, T_r)$ est la statistique canonique du modèle : elle contient toute l'information sur le paramètre θ (statistique exhaustive pour θ). Seules T et Q contribuent à la définition du modèle, $h(.)$ jouant un rôle annexe puisque si la μ est remplacée par $\mu^* = \mu h(.)$, la nouvelle densité est

$$f_\theta^*(x) = C(\theta)\exp\langle T(x), Q(\theta)\rangle \tag{8.3}$$

$C(\theta)$ est la constante de normalisation qui fait de f_θ^* (ou de f_θ) une densité :

$$C(\theta)^{-1} = \int \exp\langle T(x), Q(\theta)\rangle d\mu^*(x)$$

Definition 5 *Variable aléatoire appartenant à une famille exponentielle*

Une variable aléatoire X à valeur dans $(E,\mathcal{E})$ appartient à une famille exponentielle si elle admet une densité du type (8.2) où (8.3).

[2] μ est σ-finie si $E = \bigcup_{n\in\mathbf{N}} A_n$, où $A_n \in \mathcal{E}$ et $\mu(A_n) < \infty$. Pour $(E,\mathcal{E}) = (\mathbf{R}^p, \mathcal{B}(\mathbf{R}^p))$, $\mathbf{R}^p$ muni de sa tribu borélienne, on prendra toujours $\mu = \lambda$, la *mesure de Lebesgue*. Si E est un espace discret (fini, ou dénombrable) et $\mathcal{E} = \mathcal{P}(E)$, on prendra pour μ la *mesure de comptage* définie par $\mu(A) = |A|$, le cardinal de A. Ces deux mesures sont σ-finies.

8.6.2 Exemples

La plupart des modèles classiques sont des modèles exponentiels.

Modèle binomial : la variable binomiale $\mathcal{B}(n,\theta)$, $0<\theta<1$ a pour densité par rapport à la mesure de comptage sur $E=\{0,1,2,\cdots,n\}$,

$$f_\theta(x)=P(X=x)=\binom{n}{x}\theta^x(1-\theta)^{n-x}=(1-\theta)^n\binom{n}{x}\exp\{x\log\frac{\theta}{1-\theta}\},\text{ pour } x\in E$$

On identifie facilement : $T(x)=x$ ($T\in\mathbf{R}$, $r=1$), $Q(\theta)=\log\frac{\theta}{1-\theta}$, $C(\theta)=(1-\theta)^n$ et $h(x)=\binom{n}{x}$. Le modèle binomial est exponentiel. Pour $n=1$, c'est le modèle de Bernoulli de paramètre θ.

Modèle de Poisson : la variable de Poisson $\mathcal{P}(\theta)$, $\theta>0$, sur $E=\mathbf{N}$ a pour densité,

$$f_\theta(x)=e^{-\theta}\frac{\theta^x}{x!}=e^{-\theta}(x!)^{-1}\exp\{x\log\theta\},\ x\in\mathbf{N}$$

C'est un modèle exponentiel avec $T(x)=x$, $Q(\theta)=\log(\theta)$, $C(\theta)=e^{-\theta}$, $h(x)=(x!)^{-1}$.

Loi exponentielle : la loi exponentielle $\mathcal{E}xp(\theta)$, $\theta>0$, a pour densité sur $\mathbf{R}^+$,

$$f_\theta(x)=\theta e^{-\theta x}=\theta\exp\{x\theta\},\ x\geq 0$$

C'est un modèle exponentiel, $T(x)=x$ et $Q(\theta)=\theta$.

Loi Gamma : plus généralement, la loi Gamma $\Gamma(\theta,\kappa)$ ($\kappa=1$ pour la loi exponentielle), θ et $\kappa>0$, a pour densité sur $\mathbf{R}^+$,

$$f_{\theta,\kappa}(x)=\frac{1}{\Gamma(\kappa)}\theta^\kappa x^{\kappa-1}e^{-\theta x}=\frac{1}{\Gamma(\kappa)}\theta^\kappa\exp\{\log(x)(\kappa-1)-x\theta\},\ x\geq 0$$

C'est un modèle exponentiel, avec $T=(\log(x),x)$ et $Q(\theta)=(\kappa-1,\theta)$ et $C(\theta,\kappa)=\Gamma(\kappa)^{-1}\theta^\kappa$ ($\Gamma(\kappa)=\int_0^{+\infty}x^{\kappa-1}e^{-x}dx$).

Modèle gaussien : la loi $\mathcal{N}(m,\sigma^2)$ admet la densité sur $\mathbf{R}$:

$$f(x)=(2\pi\sigma^2)^{-\frac{1}{2}}\exp-\frac{(x-m)^2}{2\sigma^2}=[(2\pi\sigma^2)^{-\frac{1}{2}}\exp-\frac{m^2}{2\sigma^2}]\exp\{-\frac{x^2}{2\sigma^2}+\frac{xm}{\sigma^2}\}$$

C'est un modèle exponentiel de statistique $T(x)=(x,x^2)$ et $Q(\theta)=(\frac{m}{\sigma^2},-\frac{1}{2\sigma^2})$.

Il existe d'autres exemples de modèles exponentiels : le modèle linéaire gaussien, le modèle Logit binaire ou polytomique, le modèle log-linéaire de table de contingence, les modèles linéaires généralisés (régression Gamma, régression poissonienne).

Des exemples de modèles non-exponentiels sont : la loi uniforme sur $[0,\theta]$, la régression gaussienne non-linéaire, le modèle Probit.

8.6.3 Propriétés

Pour le changement de paramètres : $\varphi=Q(\theta)$, la densité de T par rapport à μ^* est :

$$f_T^*(t)=K(\varphi)\exp\langle\varphi,t\rangle$$

$K(\varphi)$ est la constante de normalisation qui fait de f_T^* une densité. Notons :

$$\Phi=\{\varphi\in\mathbf{R}^p\text{ t.q. }K(\varphi)^{-1}=\int_E\exp\langle\varphi,t\rangle dt<+\infty\}\text{ et }\psi(\varphi)=\log K(\varphi)\text{ pour }\varphi\in\Phi$$

Supposons que Φ est d'intérieur non-vide. La variable T admet la densité,

$$f_T^*(t)=\exp\{\langle\varphi,t\rangle+\psi(\varphi)\}$$

Proposition 23 *Propriété de la statistique T.*

(1) Φ est un sous-ensemble convexe de $\mathbf{R}^r$.

(2) $\varphi \mapsto \psi(\varphi)$ est convexe, indéfiniment dérivable sur l'intérieur de Φ.

(3) $E_\varphi(T(X)) = \psi^{(1)}(\varphi)$ et $Var_\varphi(T(X)) = \psi^{(2)}(\varphi)$.

(4) Pour toute observation $x \in E$, $\varphi \mapsto \log p_\varphi(x)$ est concave et $\widehat{\varphi}$, l'estimateur du MV de φ satisfait : $E_{\widehat{\varphi}}(T(X)) = T(x)$.

8.7 Information de Fischer. Inégalité de Cramer-Rao

On garde les notations du paragraphe précédent : $(E, \mathcal{E})$ est l'espace d'état des observations Y, muni d'une mesure σ-finie μ. Soit $(P_\theta, \theta \in \Theta)$ un modèle statistique sur cet espace, Θ étant un ouvert non vide de $\mathbf{R}^p$.

Notations : supposons que pour tout θ, P_θ admet une *densité* $f(., \theta)$ par rapport à μ. Si $\theta \mapsto h(\theta)$ est une fonction réelle de classe $\mathcal{C}^2$ au voisinage de θ, on note $h^{(1)}(\theta) = {}^t(\frac{\partial h}{\partial \theta_i}(\theta), i = 1, p)$ le *gradient* de h en θ, $h^{(2)}(\theta) = (\frac{\partial^2 h}{\partial \theta_i \partial \theta_j}(\theta), i, j = 1, p)$ la matrice *hessienne* $p \times p$ des dérivées secondes.

8.7.1 Densité régulière et information de Fischer

Pour $\theta_0 \in \Theta$ fixé, notons $P_0 = P_{\theta_0}$, E_0 l'espérance sous P_0.

Definition 6 ***La densité f est régulière*** *en θ_0 si :*

(R1) Il existe V_0 voisinage de θ_0 t.q. pour tout $y \in E$, $\theta \mapsto f(y, \theta)$ est de classe $\mathcal{C}^2(V_0)$.

(R2) $(\log f)^{(1)}(Y, \theta_0)$ est centrée, de carré P_{θ_0}-intégrable, vérifiant :

$$E_0\{(\log f)^{(1)}\ {}^t(\log f)^{(1)}(Y, \theta_0)\} = -E_0\{(\log f)^{(2)}(Y, \theta_0)\} = I(\theta_0)$$

(R3) $I(\theta_0)$ est régulière.

Definition 7 ***L'information de Fischer*** *en θ_0 est la matrice $I(\theta_0)$.*

L'information de Fischer est la variance du gradient de $(\log f)$ ou l'espérance de la matrice hessienne de $(\log f)$. L'inversibilité implique que $I(\theta_0)$ est dp. L'information de Fischer est additive : si $Y_1, Y_2, \cdots, Y_n$ sont i.i.d de densité $f(., \theta_0)$, l'information mutuelle est n fois l'information individuelle, $I_n(\theta_0) = n\ I(\theta_0)$.

Example 19 *Modèle exponentiel*

(1) La densité d'un modèle exponentiel $f(y, \theta) = \exp\{\langle \theta, t(y) \rangle + \psi(\theta)\}$ est régulière et l'information de Fischer vaut :

$$I(\theta) = Var_\theta(t(Y)) = \frac{\partial^2 \psi}{\partial \theta^2}(\theta)$$

(2) Pour Y de loi de Bernoulli $\mathcal{B}(p)$, $I(p) = \{p(1-p)\}^{-1} = \{Var_p(Y)\}^{-1}$. Pour la loi binomiale $\mathcal{B}(n, p)$, $I(p) = n\{p(1-p)\}^{-1}$. Pour la loi exponentielle de paramètre $\theta = E_\theta(Y)$, $I(\theta) = \theta^{-2} = \{Var_\theta(Y)\}^{-1}$. Pour la loi gaussienne $\mathcal{N}(m, \sigma^2)$ et $\theta = (m, \sigma^2)$, $I(\theta)$ est diagonale, de coefficients diagonaux σ^{-2} (l'information sur m) et $(2\sigma^4)^{-1}$ (l'information sur σ^2).

8.7.2 Borne de Cramer-Rao

$S \in \mathbf{R}^k$ est une statistique régulière si sa densité est régulière. Soit $\succeq$ la relation d'ordre partiel sur les matrices symétriques $k \times k$: $A \succeq B \Longleftrightarrow A - B$ est sdp. Le résultat suivant donne une borne inférieure pour la variance d'un estimateur sans biais :

Proposition 24 *Considérons un modèle régulier, S un estimateur sans biais et régulier de $g(\theta) \in \mathbf{R}^k$. Alors, notant $Jg(\theta)$ la matrice jacobienne de g :*

$$Var_\theta(S) \succeq Jg(\theta) I(\theta)^{-1} \, {}^t Jg(\theta)$$

Quand un estimateur sans biais atteint la borne de Cramer-Rao, on dit qu'il est *efficace*.

Bibliographie

[1] Agresti, *Categorical data analysis*, J. Wiley, 2ème Edition, 1990

[2] Antoniadis A., Berruyer J. et Carmona R., *Régression non-linéaire et applications*, Economica, 1992

[3] Coursol J., *Technique statistique des modèles linéaires*, Cours du CIMPA, 1980

[4] Dacunha-Castelle D. et Duflo M., *Probabilités et Statistiques*, Tomes 1 et 2, Masson, 2ème Edition, 1994

[5] Draper N et Smith H., *Applied regression analysis,* Wiley, 1981

[6] Guyon X., Statistique et Econométrie : du modèle linéaire aux modèles non-linéaires, Ellipses, 2000

[7] Huet S., Jolivet E. et Messéan A., *La régression non-linéaire : méthodes et applications à la biologie*, INRA, Paris, 1991

[8] McCullagh P. et Nelder J.A., *Generalized Linear Models,* Chapman et Hall, 2ème Edition, 1989

[9] Prum B., *Modèle linéaire : Comparaison de groupes et régression,* Les Editions de l'INSERM, 1996

[10] Saporta G., *Probabilités et analyse des données statistiques,* Technip, 1990

[11] Scheffé H., *The analysis of variance*, J. Wiley, 1970

[12] Tomassone R., Audrain, Lesquoy E. et Millier C., *La régression : nouveaux regards sur une ancienne méthode*, Masson, 1992