Econométrie des Variables Qualitatives

Emmanuel Duguet

Version 5
2008

TABLE DES MATIÈRES

CHAPITRE 1

Les variables qualitatives explicatives

Les variables qualitatives explicatives sont très nombreuses lorsque l'on étudie les thèmes de l'économie du travail ou de l'innovation. Le but de cette section est d'exposer l'interprétation des coefficients de ces variables dans le cas du modèle linéaire. Ce thème s'étend aux cas où la variable expliquée est qualitative.

Une première utilisation, très répandue, des variables qualitatives consiste à les utiliser sous forme d'indicatrices dans une régression linéaire. Elles servent à indiquer des effets fixes pour indiquer une appartenance à un groupe en général (e.g., région, industrie, catégorie socio professionnelle, niveau de diplôme). Les coefficients de ces variables qualitatives ne s'interprétent plus comme des dérivées par rapport aux variables explicatives, car les dérivées n'existent plus, mais comme un écart moyen par rapport à une modalité de référence. Une seconde utilisation de ces variables qualitatives consiste à découper une variable continue en intervalles puis à examiner la forme de la relation qu'elle entretient avec la variable expliquée. Il s'agit ici d'une approximation par intervalle d'une fonction inconnue.

1.1 Modèle sans terme constant

Nous allons prendre comme exemple introductif une variable qualitative polytomique possédant p modalités. On considère un échantillon de N individus; sans perte de généralité, on suppose que chaque individu appartient à un seul groupe et il y a p groupes différents.[1] Pour sim-

[1] Dans le cas ou des invididus appartiennent à plusieurs groupes dans les données de départ, il est possible de redéfinir la variable qualitative de sorte que tous les individus

plifier l'analyse, on a défini ces groupes de manière à ce qu'ils soient disjoints. On note G_j l'ensemble des indices des individus du groupe j, avec $j = 1, ..., p$. On remarque que $\bigcup_{j=1}^{j=p} G_j = \{1, ..., N\}$. On considère l'estimation d'un modèle linéaire de la forme suivante :

$$y_i = \sum_{j=1}^{p} b_j D_{ji} + u_i,$$
$$\mathrm{E}(u_i) = 0,\ \mathrm{E}\left(u_i^2\right) = \sigma_u^2,\ \mathrm{E}(u_i u_j) = 0\ \forall i \neq j,\ i = 1, ..., N$$

où y_i est la variable expliquée, u_i la perturbation du modèle et les variables D_{ji} sont des variables qualitatives dichotomiques définies par :

$$D_{ji} = \begin{cases} 1 & \text{si } i \in G_j \\ 0 & \text{si } i \notin G_j \end{cases} \quad i = 1, ..., N$$

La modélisation de base consiste donc à remplacer la variable qualitative d'appartenance à un groupe par p variables dichotomiques $(D_{1i}, ..., D_{pi})$ définies par chacune de ses modalités $j \in \{1, ..., p\}$. On remarque les propriétés suivantes des variables dichotomiques, qui montrent que le codage binaire $\{0, 1\}$ est le plus pertinent :

1. $D_{ji}^2 = D_{ji}$ puisque $0^2 = 0$ et $1^2 = 1$;
2. $D_{ji} D_{ki} = 0\ \forall j \neq k$, car un individu i ne peut pas appartenir à deux groupes à la fois;
3. $\sum_{i=1}^{N} D_{ji} = \sum_{i \notin G_j} 0 + \sum_{i \in G_j} 1 = N_j$, le nombre d'individus présents dans le groupe j;
4. $1/N \sum_{i=1}^{N} D_{ji} = N_j/N$, la fraction des individus du groupe j dans la population totale. Dans le cas des variables dichotomiques, la moyenne arithmétique sert donc à calculer des pourcentages.

En utilisant les propriétés de la perturbation, on voit que :

$$\mathrm{E}(y_i|D) = b_j \text{ si } i \in G_j,$$

ainsi les coefficients de régression s'interprétent comme les espérances conditionnelles de la variable expliquée dans le groupe j. Ce n'est pas le cas des variables explicatives quantitatives. On peut également interpréter la différence de deux coefficients comme la différence des espérances conditionnelles entre deux groupes :

$$b_j - b_k = \mathrm{E}(y_i|i \in G_j) - \mathrm{E}(y_i|i \in G_k).$$

appartiennent à un seul groupe.

L'estimation est facilitée en écrivant le modèle individu par individu. On pose :

$$\underset{(1,p)}{D_i} = (D_{1i}, ..., D_{ji}, ..., D_{pi}) \, , \; i = 1, ..., N$$

et l'on écrit le vecteur des paramètres en colonne :

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} .$$

On obtient donc le modèle linéaire suivant :

$$y_i = D_i b + u_i, \; i = 1, ..., N \, .$$

L'estimateur des moindres carrés ordinaires de b est donc défini par :

$$\widehat{b} = \left(\sum_{i=1}^{N} D_i' D_i \right)^{-1} \sum_{i=1}^{N} D_i' \; y_i .$$

La matrice $\sum_{i=1}^{N} D_i' D_i$ est diagonale et donne les nombres d'observations dans chaque groupe. En effet, en utilisant les propriétés 1 et 2 :

$$\underset{(p,p)}{D_i' D_i} = \begin{pmatrix} D_{1i} \\ \\ D_{ji} \\ \\ D_{pi} \end{pmatrix} (D_{i1}, ..., D_{ij}, ..., D_{ip})$$

$$= \begin{pmatrix} D_{1i}^2 & \cdots & D_{1i} D_{ji} & \cdots & D_{1i} D_{pi} \\ \vdots & \ddots & \vdots & & \vdots \\ D_{1i} D_{ji} & \cdots & D_{ji}^2 & \cdots & D_{ji} D_{pi} \\ \vdots & & \vdots & \ddots & \vdots \\ D_{pi} D_{1i} & \cdots & D_{pi} D_{ji} & \cdots & D_{pi}^2 \end{pmatrix}$$

$$= \begin{pmatrix} D_{1i} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & D_{ji} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & D_{pi} \end{pmatrix}$$

en conséquence, en utilisant la propriété 3 :

$$\sum_{i=1}^{N} D_i' D_i = \begin{pmatrix} \sum_{i=1}^{N} D_{1i} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & \sum_{i=1}^{N} D_{ji} & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \sum_{i=1}^{N} D_{pi} \end{pmatrix}$$

$$= \begin{pmatrix} N_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & N_j & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & N_p \end{pmatrix},$$

ce qui implique :

$$\left(\sum_{i=1}^{N} D_i' D_i \right)^{-1} = \begin{pmatrix} 1/N_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 1/N_j & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 1/N_p \end{pmatrix}$$

La seconde partie de l'estimateur des moindres carrés ordinaires est égale à :

$$\sum_{i=1}^{N} D_i' y_i = \begin{pmatrix} \sum_{i=1}^{N} D_{1i} y_i \\ \vdots \\ \sum_{i=1}^{N} D_{ji} y_i \\ \vdots \\ \sum_{i=1}^{N} D_{pi} y_i \end{pmatrix} = \begin{pmatrix} \sum_{i \notin G_1} 0 \times y_i + \sum_{i \in G_1} 1 \times y_i \\ \vdots \\ \sum_{i \notin G_j} 0 \times y_i + \sum_{i \in G_j} 1 \times y_i \\ \vdots \\ \sum_{i \notin G_p} 0 \times y_i + \sum_{i \in G_p} 1 \times y_i \end{pmatrix}$$

Dans l'ensemble on obtient donc les moyennes arithmétiques des p groupes :

$$\widehat{b} = \begin{pmatrix} 1/N_1 \sum_{i \in G_1} y_i \\ \vdots \\ 1/N_j \sum_{i \in G_j} y_i \\ \vdots \\ 1/N_p \sum_{i \in G_p} y_i \end{pmatrix} \triangleq \begin{pmatrix} \overline{y}_1 \\ \vdots \\ \overline{y}_j \\ \vdots \\ \overline{y}_p \end{pmatrix}$$

1.2 Modèle avec un terme constant

Ici il est inutile de refaire les calculs. En effet, les moindres carrés ordinaires reviennent à faire une projection orthogonale du vecteur des observations de la variable expliquée y sur le sous-espace vectoriel engendré par les vecteurs correspondants des variables explicatives, noté $\mathrm{Im}\,(D_1, ..., D_p)$. Ces vecteurs sont linéairement indépendants et forment donc une base de cet espace vectoriel. Pour trouver les coefficients du modèle avec terme constant, il faut avoir en tête les deux éléments suivants :

1. Le terme constant, noté $e_{(N,1)}$ est égal à la somme des vecteurs D_j : $e = \sum_{j=1}^{p} D_j$.

2. La décomposition d'un vecteur y en une base est unique, et les coefficients des moindres carrés ordinaires sont les coordonnées du vecteur y dans la base $(D_1, ..., D_p)$.

La première propriété implique que, dans un modèle avec terme constant, il faut retirer un des vecteur D_j de la liste des variables explicatives pour éviter une multicolinéarité parfaite. La seconde propriété permet de calculer les nouveaux estimateurs des MCO en fonction de $\widehat{b}$. Si on retire la modalité k de la liste des groupes, on estime le modèle :

$$y = c_0 e + c_1 D_1 + ... + c_{k-1} D_{k-1} + c_{k+1} D_{k+1} + ... + c_p D_p + u,$$

après estimation de ce modèle par les moindres carrés ordinaires, on obtient une prévision :

$$\widehat{y} = \widehat{c}_0 e + \widehat{c}_1 D_1 + ... + \widehat{c}_{k-1} D_{k-1} + \widehat{c}_{k+1} D_{k+1} + ... + \widehat{c}_p D_p,$$

en remplaçant la constante par sa valeur, $e = \sum_{j=1}^{p} D_j$, on obtient la formulation équivalente :

$$\begin{aligned}\widehat{y} &= \widehat{c}_0 (D_1 + ... + D_p) + \widehat{c}_1 D_1 + ... + \widehat{c}_{k-1} D_{k-1} + \widehat{c}_{k+1} D_{k+1} + ... + \widehat{c}_p D_p \\ &= (\widehat{c}_0 + \widehat{c}_1) D_1 + ... + (\widehat{c}_0 + \widehat{c}_{k-1}) D_{k-1} + \widehat{c}_0 D_k + (\widehat{c}_0 + \widehat{c}_{k+1}) D_{k+1} + \\ &\qquad ... + (\widehat{c}_0 + \widehat{c}_p) D_p.\end{aligned}$$

La prévision du modèle de départ est égale à :

$$\widehat{y} = \widehat{b}_1 D_1 + ... + \widehat{b}_{k-1} D_{k-1} + \widehat{b}_k D_k + \widehat{b}_{k+1} D_{k+1} + ... + \widehat{b}_p D_p,$$

en utilisant l'unicité de la décomposition en une base, on obtient :

$$\left.\begin{array}{r} \widehat{c}_0 + \widehat{c}_1 = \widehat{b}_1 \\ \vdots \\ \widehat{c}_0 + \widehat{c}_{k-1} = \widehat{b}_{k-1} \\ \widehat{c}_0 = \widehat{b}_k \\ \widehat{c}_0 + \widehat{c}_{k+1} = \widehat{b}_{k+1} \\ \vdots \\ \widehat{c}_0 + \widehat{c}_p = \widehat{b}_p \end{array}\right\} \Leftrightarrow \left\{\begin{array}{l} \widehat{c}_0 = \widehat{b}_k \\ \widehat{c}_1 = \widehat{b}_1 - \widehat{b}_k \\ \vdots \\ \widehat{c}_{k-1} = \widehat{b}_{k-1} - \widehat{b}_k \\ \widehat{c}_{k+1} = \widehat{b}_{k+1} - \widehat{b}_k \\ \vdots \\ \widehat{c}_p = \widehat{b}_p - \widehat{b}_k \end{array}\right.$$

La constante du nouveau modèle représente l'effet de l'indicatrice qui a été enlevée de la régression, et les autres coefficients l'écart entre le coefficient de l'indicatrice courante et de l'indicatrice enlevée. Ainsi l'indicatrice qui a été enlevée correspond à la modalité de référence. C'est la raison pour laquelle il faut indiquer explicitement les modalités des indicatrices enlevées dans les tableaux de régression, elle sont indispensables à l'interprétation.

Remarque 1.1 *Le test de Ficher sur le modèle avec terme constant revient à tester ici l'égalité jointe des moyennes entre tous les groupes. En effet, le test correspond à l'hypothèse nulle* $\mathrm{H}_0 : c_1 = ... = c_p = 0 \Leftrightarrow \mathrm{H}_0 : \mathrm{E}(y_j) - \mathrm{E}(y_k) = 0 \ \forall j \neq k$. *On notera qu'on ne teste pas la nullité du terme constant du modèle* c_0.

Remarque 1.2 *On peut utiliser un simple test de Student pour tester l'égalité des moyennes entre un groupe donné,* k, *et un autre groupe. Il suffit de mettre un terme constant dans le modèle et d'enlever l'indicatrice du groupe dont on tester l'égalité de la moyenne avec les autres groupes.*

1.3 Modèle avec variables explicatives

On introduit maintenant un autre jeu de variables, dont la matrice est notée X, dans le modèle de départ :

$$y_i = \underset{(1,m)}{X_i} \underset{(m,1)}{a} + \underset{(1,p)}{D_i} \underset{(p,1)}{b} + u_i,$$

on a clairement :

$$\mathrm{E}(y_i | X_i, D_{ji} = 1) = X_i a + b_j,$$

de sorte que les coefficients b_j représentent les écarts de moyenne conditionnelle entre deux groupes :

$$\begin{aligned} \mathrm{E}(y_i | X_i, D_{ji} = 1) - \mathrm{E}(y_i | X_i, D_{ki} = 1) &= (X_i a + b_j) - (X_i a + b_k) \\ &= b_j - b_k. \end{aligned}$$

Les résultats de la section précédente sont donc toujours valables. Le terme constant représente le coefficient de l'indicatrice qui a été retirée et les coefficients des autres indicatrices doivent s'interpréter en écart au coefficient de l'indicatrice retirée.

1.4 Modèle avec produits croisés

1.4.1 Cas dichotomique

On peut introduire les produits croisés de manière naturelle à partir du modèle suivant. Considérons que des individus bénéficient d'une mesure d'aide que nous supposerons affectée au hasard (i.e., sans biais de sélection). On note :

$$D_i = \begin{cases} 1 & \text{si l'individu } i \text{ est aidé} \\ 0 & \text{sinon} \end{cases}$$

Une fois cette mesure attribuée, on examine une variable de performance, liée aux objectifs de l'aide, que l'on note y_i. En théorie, pour chaque individu, cette mesure peut prendre deux valeurs :

- y_{0i} : la valeur de y_i si l'individu i n'est pas aidé;
- y_{1i} : la valeur de y_i si l'individu i est aidé.

Ce que l'on cherche à évaluer est l'effet de la mesure, noté :

$$\alpha = \mathrm{E}\left(y_{1i} - y_{0i}\right),$$

qui représente la moyenne des variations de performance associée à la mesure, prise sur l'ensemble des individus. On peut aller plus loin, en introduisant un modèle explicatif des performances potentielles des individus :

$$\begin{aligned} y_{0i} &= a_0 + X_i c_0 + u_{0i} \\ y_{1i} &= a_1 + X_i c_1 + u_{1i}, \end{aligned}$$

où X_i représente les déterminants de la performance. Les coefficients a_1 et a_0 représentent les niveaux moyens de performances en $X_i = 0$, selon que l'on est aidé ou non. Pour obtenir un modèle empiriquement estimable, il faut l'écrire en fonction de quantités observables. Or, on n'observe que y_{0i} lorsque $D_i = 0$ et seulement y_{1i} si $D_i = 1$. La seule variable observable est :

$$y_i = T_i y_{1i} + (1 - T_i)\, y_{0i} = \begin{cases} y_{1i} & \text{si } T_i = 1 \\ y_{0i} & \text{si } T_i = 0 \end{cases}$$

En conséquence, le modèle économétrique s'écrit :

$$\begin{aligned} y_i &= T_i \left(a_1 + X_i c_1 + u_{1i}\right) + \left(1 - T_i\right) \left(a_0 + X_i c_0 + u_{0i}\right) \\ &= a_0 + X_i c_0 + T_i \underbrace{\left(a_1 - a_0\right)}_{a} + T_i X_i \underbrace{\left(c_1 - c_0\right)}_{c} + u_i, \end{aligned}$$

où a est le coefficient de l'aide et c le vecteur des coefficients des variables explicative. La perturbation est égale à :

$$u_i = T_i u_{1i} + \left(1 - T_i\right) u_{0i}.$$

Ce modèle fait apparaître un produit croisé entre la variable d'aide T_i et les variables explicatives de la performance X_i. L'estimation de ce modèle permet d'évaluer l'effet de la mesure car :

$$\begin{aligned} \delta &= \mathrm{E}\left(y_{1i} - y_{0i}\right) \\ &= \mathrm{E}\left(a_1 - a_0 + X\left(c_1 - c_0\right)\right) \\ &= a_1 - a_0 + \mathrm{E}\left(X\right)\left(c_1 - c_0\right) \\ &= a + \mathrm{E}\left(X\right) c \end{aligned}$$

que l'on peut estimer sans biais et de manière convergente par :

$$\widehat{\delta} = \widehat{a} + \overline{X}\ \widehat{c},$$

on remarque que lorsque les variables X sont centrées avant de prendre les produits croisés $\left(\overline{X} = 0\right)$, l'estimateur $\widehat{\delta}$ est obtenu directement par le coefficient de la variable indicatrice d'aide $\widehat{a}$ dans la régression avec produits croisés. On remarque également que ce modèle suppose qu'il existe une hétéroscédasticité par bloc car :

$$\mathrm{V}\left(u_i\right) = \begin{cases} \mathrm{V}\left(u_{0i}\right) & \text{si } T_i = 0 \\ \mathrm{V}\left(u_{1i}\right) & \text{si } T_i = 1 \end{cases}$$

et dans le cas où la mesure d'aide affecte également la variance, $\mathrm{V}\left(u_{0i}\right) \neq \mathrm{V}\left(u_{1i}\right)$, il faut estimer le modèle par les moindres carrés pondérés.

1.4.2 Cas polytomique

On introduit maintenant, en plus des variables explicatives et des indicatrices, les produits croisés des indicatrices et des variables explicatives. On a donc :

$$y_i = D_i b + \left(X_i \otimes D_i\right) c + u_i,$$

avec :

$$\underset{(mp,1)}{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_j \\ \vdots \\ c_p \end{pmatrix},$$

le terme en X_i a été retiré puisque $\sum_{j=1}^{p} X_i D_{ji} = X_i$. L'espérance conditionnelle dans le groupe j devient maintenant :

$$\mathrm{E}\left(y_i | X_i, D_{ji} = 1\right) = X_i c_j + b_j,$$

d'où la différence entre les groupes j et k :

$$\begin{aligned}\gamma_i &\triangleq \mathrm{E}\left(y_i | X_i, D_{ji} = 1\right) - \mathrm{E}\left(y_i | X_i, D_{ki} = 1\right) \\ &= \left(X_i c_j + b_j\right) - \left(X_i c_k + b_k\right) \\ &= X_i \left(c_j - c_k\right) + b_j - b_k,\end{aligned}$$

l'effet varie en fonction des caractéristisques de l'individu i au sein du groupe j. Ce modèle autorise donc une hétérogénéité individuelle au sein de chaque groupe. L'effet moyen est égal à :

$$\begin{aligned}\overline{\gamma} &= \frac{1}{N} \sum_{i=1}^{N} \gamma_i \\ &= \overline{X} \left(c_j - c_k\right) + b_j - b_k.\end{aligned}$$

On peut toutefois estimer directement la partie de l'écart entre les groupes qui ne dépend pas des variables explicatives du modèle, $b_j - b_k$, en utilisant la méthode suivante. On centre la variable X avant de faire les produits croisés; avec cette convention $\overline{X} = 0$ et on obtient directement la différence entre les groupes par $b_j - b_k$. Cette dernière quantité mesure l'écart de moyenne entre les groupes une fois que l'on a éliminé l'effet des variables de X sur ces moyennes.

1.4.3 Cas dichotomique

On considère une variable dichotomique $T_i \in \{0, 1\}$ dont on veut connaître l'effet sur y_i. La variable T_i peut être une caractéristique individuelle ou une mesure de politique économique individuelle. La modalité $T_i = 1$ correspond aux individus qui ont bénéficié de la mesure. La variable y_i est alors une mesure de performance choisie en fonction de l'objectif de politique économique. Les variables explicatives de la performance X_i sont centrées avant de prendre les produits croisés, de sorte que l'on a $\overline{X} = 0$. Le modèle, facilement généralisable, est donné par :

$$\mathrm{E}\left(y_i | X_i, T_i\right) = d + X_i a + T_i b + \left(X_i \otimes T_i\right) c_1, \tag{1.1}$$

où d est le terme constant du modèle. On obtient les espérance suivantes :

$$\begin{aligned}&\mathrm{E}\left(y_i | X_i, T_i = 0\right) = d + X_i a, \\ &\mathrm{E}\left(y_i | X_i, T_i = 1\right) = d + b + X_i \left(a + c_1\right),\end{aligned}$$

d'où l'effet de T_i sur y_i :

$$\gamma_i = \mathrm{E}\left(y_i|X_i, T_i = 1\right) - \mathrm{E}\left(y_i|X_i, T_i = 0\right) = b + X_i c_1,$$

après estimation par les moindres carrés ordinaires on obtient :

$$\widehat{\gamma}_i = \widehat{b} + X_i \widehat{c}_1,$$

d'où l'effet moyen de T_i sur l'échantillon :

$$\widehat{\gamma} = \frac{1}{N}\sum_{i=1}^{N} \widehat{\gamma}_i = \widehat{b} + \left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) \widehat{c}_1 = \widehat{b}.$$

Quand on centre les variables explicatives X_i, on peut donc obtenir directement l'effet moyen de T_i par son coefficient dans la régression (1.1) .

CHAPITRE 2

Les variables qualitatives expliquées

Les bases de données microéconomiques comprennent invariablement des données tronquées, connues seulement par intervalle ou encore de type purement qualitatif. Par exemple, dans l'enquête Innovation du SESSI, on connaît le simple fait d'avoir réalisé une innovation ou encore une appréciation de l'entreprise sur l'importance d'un déterminant de l'innovation. Ce n'est pas toujours un inconvénient, car certains phénomènes ne sont pas quantifiables de façon objective. Dans l'enquête Emploi de l'INSEE on peut connaître le fait qu'un individu travaillle et le nombre d'heures travaillées par les individus qui ont travaillé sur la période d'enquête. Mais on peut préférer créer une variable qualitative de type "pas d'emploi, temps partiel, temps plein" si l'objet de l'étude le justifie. Malgré la perte d'information inévitable quand on passe par exemple, de la valeur exacte d'une variable à sa connaissance par intervalle, il est toujours possible d'étudier ses effets, même si cela comporte certaines limites. Dans cette section, nous donnons quelques exemples de variables qualitatives et leur représentation en économétrie.

2.1 Variables dichotomiques

Une variable dichotomique est une variable qui ne peut prendre que deux modalités exclusives l'une de l'autre, comme “Oui/Non” ou “Inférieur ou égal à/Strictement supérieur à”. Par convention, on code une modalité à 0 et l'autre à 1. Ce n'est pas une obligation, mais cette présentation permet de simplifier la présentation du problème. La variable associée est appelée une indicatrice , dans l'exemple suivant la variable y_i est une

indicatrice d'innovation :

$$y_i = \begin{cases} 1 & \text{si l'entreprise } i \text{ a innové} \\ 0 & \text{sinon} \end{cases}$$

Ce que l'on peut étudier à partir de ce type de variable, ce sont les déterminants de la décision d'innover. Cette décision se traduit en termes statistiques par une probabilité d'innover. Ainsi, on va rechercher quelles sont les variables qui réduisent ou au contraire augmentent la probabilité d'innover. Il faut donc construire un modèle qui nous permette d'estimer l'effet d'un ensemble de variables sur la probabilité qu'une entreprise innove. Pour cela on construit ce que l'on appelle un modèle latent, c'est à dire inobservable. On peut faire une analogie entre ce modèle latent et le modèle théorique qui sert de base à l'étude.

Si l'on pouvait mesurer l'innovation sous forme quantitative, on utiliserait le modèle linéaire standard. Mais, soit pour des raisons de collecte des données soit parce que le phénomène que l'on étudie n'est pas quantifiable, on ne dispose que d'une information qualitative sur celui-ci. Dans notre exemple, on sait juste si l'entreprise a innové ou non. Le modèle latent est le modèle linéaire standard :

$$\pi_i^* = X_i b + u_i, \quad i = 1, ..., N$$

Que représente ce modèle ? La variable endogène π^* est inobservable. On peut l'interpréter ici comme l'espérance de profit associé à l'introduction d'une innovation, compte-tenu d'un effet de remplacement des anciens produits. Les variables explicatives X sont les déterminants de ce gain et le paramètre b mesure l'importance de ces déterminants. Il nous faut maintenant expliquer comment on passe de π^* à y, c'est à dire de ce qui n'est pas observable (π_i^*, u_i) à ce qui l'est (y_i, X_i), car seules ces dernières informations peuvent être utilisées en pratique.

Il est raisonnable de penser que toutes les entreprises cherchent à améliorer leurs produits et leurs procédés de production, même marginalement. Le résultat anticipé de cette activité est justement représentée dans notre modèle par π^*. Mais on n'observera la mise en oeuvre d'une innovation que si cette activité procure des gains significatifs, s'ils dépassent un certain seuil. Notons que ce gain ne sera significatif que si l'innovation l'est également et que cette notion de seuil correspond bien à la définition retenue dans les enquêtes sur l'innovation.[1] Seules sont considérées dans les enquêtes les améliorations significatives de produit

[1] Ainsi dans l'enquête du SESSI sur "l'innovation technologique dans l'industrie", annexée à l'Enquête Annuelle d'Entreprise de 1990, l'innovation de produit est définie sur le questionnaire par : "Un produit est considéré comme comme technologiquement innovant s'il donne lieu à la création d'un nouveau marché ou s'il peut se distinguer substanciellement de produits précédemment fabriqués, d'un point de vue technologique ou par les prestations rendues à l'utilisateur. Ne sont pas concernées

et de procédé. Soit le seuil π_0, qui peut dépendre de chaque industrie, on a :

$$y_i = \begin{cases} 1 & \text{si } \pi_i^* > \pi_0 \\ 0 & \text{si } \pi_i^* \leq \pi_0 \end{cases}$$

Ceci implique que l'on peut maintenant calculer la probabilité d'innover. Elle est égale à :

$$\Pr[y_i = 1] = \Pr[\pi_i^* > \pi_0].$$

Il reste alors à faire une hypothèse sur la distribution conditionnelle de π^* sachant X pour obtenir une forme fonctionnelle précise. Selon l'hypothèse que l'on fait, on obtient un modèle Logit (loi logistique) ou un modèle Probit (loi normale). Cette liste n'est bien sûr pas limitative et chaque hypothèse de distribution mène à un modèle différent. Des tests sont alors nécessaires pour trancher.

2.2 Variables polytomiques ordonnées

Cette fois-ci, la variable qualitative que l'on observe peut prendre plus de deux modalités qui sont ordonnées entre elles[2]. Elles peuvent être définies aussi bien par rapport à une quantité que traduire une appréciation. Par exemple, dans l'enquête Innovation du Ministère de l'Industrie (SESSI) le pourcentage de produits de moins de cinq ans d'âge dans le chiffre d'affaires est donné sous la forme suivante : entre 0 et 10% , de 10% à 30%, de 30% à 70% et plus de 70% . Dans cette même enquête, l'importance de la recherche développement du groupe auquel appartient l'entreprise comme déterminant de l'innovation est donnée sous la forme : "pas du tout", "un peu", "moyennement" et "beaucoup". Dans les deux cas, les modalités traduisent un ordre, qui indique l'intensité de la variable. Le modèle latent représente alors la vraie valeur de la variable, qui n'est pas observable. Cette variable, que l'on cherche à expliquer, est représentée par le modèle latent linéaire :

$$y_i^* = X_i b + u_i, \quad i = 1, ..., N.$$

La variable observable, qualitative, prend maintenant une forme plus

les innovations de nature purement esthétique ou de style (design); en revanche sont concernées, mais isolées, les innovations de conditionnement ou d'emballage.". Pour une présentation de l'enquête et un exemplaire du questionnaire, voir François (1991).

[2] Il existe également des variables qualitatives non ordonnées qui représentent des choix. Les plus connues représentent le choix de mode du transport comme : véhicule individuel , bus , métro.

générale :

$$y_i = \begin{cases} 1 & \text{si} \quad \alpha_0 < y_i^* \leq \alpha_1 \\ 2 & \text{si} \quad \alpha_1 < y_i^* \leq \alpha_2 \\ \vdots & \\ r & \text{si} \quad \alpha_{r-1} < y_i^* \leq \alpha_r \end{cases}$$

Les bornes délimitent les valeurs α_0 et α_r que peut prendre la variable y_i^*. Pour une variable réelle, on adopte la convention $\alpha_0 = \{-\infty\}$ et $\alpha_r = \{+\infty\}$. Plus généralement les bornes peuvent être connues ou inconnues. Pour le pourcentage d'innovation décrit plus haut, elles sont égales à 0, 0.1, 0.3, 0.7 et 1. Dans d'autre cas, les variables sont toujours ordonnées mais on ne connait pas les seuils. C'est le cas quand les personnes intérrogées répondent à une question par "Pas du tout, un peu, moyennement, beaucoup". Pourtant les seuils théoriques existent bien puisque l'on peut ordonner les modalités, on suppose simplement qu'ils sont *constants* au sein d'une population donnée. Dans les deux cas, seuils connus ou inconnus, on peut estimer un modèle pour trouver les déterminants de y_i^*. Cette fois-ci, la probabilité d'observer la modalité j est donnée par :

$$\Pr[y_i = j] = \Pr[\alpha_{j-1} < y_i^* \leq \alpha_j] = \Pr[y_i^* \leq \alpha_j] - \Pr[y_i^* \leq \alpha_{j-1}], \quad j = 1, ..., r.$$

Une fois que l'on a spécifié la loi conditionnelle de y^* sachant X, on peut procéder aux estimations à partir des variables observables (y_i, X_i). Les modèles polytomiques ordonnés peuvent être utilisés pour ce genre de variable endogène. Si la loi des perturbations u_i est normale, on obtient un modèle Probit polytomique ordonné . D'autres hypothèses sur la loi des perturbations u_i donnent d'autres modèles.

2.3 Variables de comptage

Certaines données d'innovation sont discrètes. Ainsi le nombre de brevets n'est pas une donnée quantitative de même nature que les dépenses de recherche et développement. Il s'agit d'une variable qui ne prend que des valeurs entières. Qui plus est, il s'agit du comptage d'événements relativement rares. Sur une année, en France, on compte beaucoup d'entreprises qui ne déposent pas de brevet. Il peut s'agir du résultat d'une décision mais également du simple fait que l'entreprise n'a pas trouvé d'innovation brevetable durant l'année écoulée. La variable expliquée prend ses valeurs dans l'ensemble des entiers naturels $y_i \in \{0, 1, 2, ...\}$.

Ce processus est par nature aléatoire et, comme pour les autres variables, on modélise son espérance mathématique. Ici toutefois, cette espérance est toujours strictement positive et l'on prend donc une forme

exponentielle :

$$\mathrm{E}\left(y_i|\, X_i, b\right) = \exp\left(X_i b + u_i\right) > 0.$$

Cette espérance mathématique est alors supposée être celle d'une loi de Poisson , utilisée pour représenter les variables endogènes discrètes positives ou nulles. Notons bien qu'il y a *deux sources* d'aléas dans cette dernière modélisation. La première vient de l'erreur que l'on fait sur la moyenne de la variable expliquée, représentée par $\exp\left(u_i\right)$, la seconde vient du tirage dans une loi de Poisson dont la moyenne est aléatoire. Dans les modèles usuels, tout l'aléa provient de l'erreur que l'on fait sur la moyenne.

Lorsqu'il n'y a pas de perturbation dans la moyenne ($\mathrm{V}\left(\exp u_i\right) = 0, \forall i$), on parle du modèle de Poisson *homogène*, dans le cas inverse il s'agit du modèle de Poisson *hétérogène*. Notons que l'on peut faire un parallèle entre les données de comptage et les données de durée, car une donnée de comptage donne le nombre d'évènements qui se sont produits pendant une durée donnée. On montre que la loi de Poisson homogène correspond à une loi de durée exponentielle.

2.4 Variables censurées ou tronquées

Une variable censurée ou tronquée est une variable dont on observe la réalisation pour certains individus seulement. La troncature peut provenir soit du processus de collecte des données soit d'une décision prise par ces mêmes individus. C'est ce dernier cas qui nous intéresse.[3] Prenons le cas de l'activité de recherche et développement : une entreprise doit à la fois décider si elle investit ou non dans un programme de recherche et combien elle y investi. Ces deux décisions sont étroitement reliées. Le processus de décision est représenté par une variable latente, qui peut être le critère de décision π^*. Les déterminants de cette décision sont notés X_1. On pose :

$$\pi_i^* = X_{1i} b_1 + u_{1i}, \quad i = 1, ..., N.$$

Cette première variable latente génère une variable qualitative dichotomique :[4]

$$y_i = \begin{cases} 1 & \text{si} \quad \pi_i^* \geq 0 \\ 0 & \text{si} \quad \pi_i^* < 0 \end{cases}$$

Cette indicatrice nous dit si l'entreprise a investi en recherche ou non. Mais elle détermine également s'il est possible d'observer le montant investi en recherche, représenté par une seconde variable r^*. L'investissement en recherche r^*est expliqué par le modèle :

$$r^* = X_{2i} b_2 + u_{2i}, \quad i = 1, ..., N,$$

où X_2 contient les déterminants du montant investi en recherche. On admet de plus que les deux variables latentes, π_i^* et r_i^* sont corrélées entre elles. Cette corrélation provient du fait que l'on obtient généralement r^* en maximisant le profit π^*, ce qui implique que les deux variables sont déterminées simultanément. La variable de recherche observable, notée r est donc donnée par :

$$r_i = \begin{cases} r_i^* & \text{si} \quad \pi_i^* \geq 0 \\ \text{manquant} & \text{si} \quad \pi_i^* < 0 \end{cases}$$

où "manquant" indique une valeur manquante dans la base de données. Lorsque les perturbations u_{1i} et u_{2i} suivent une loi normale bivariée on obtient le modèle *tobit généralisé* de Heckman.

[3] Le cas des censures exogènes correspond au modèle tobit simple, le lecteur pourra trouver une présentation de ce modèle dans Gouriéroux (1989) et Maddala (1983). Le cas que nous présentons ici est celui d'une censure endogène aboutissant au modèle tobit généralisé, étudié à l'origine par Heckman (1976, 1979).

[4] Le seuil peut être mis à 0 sans perte de généralité tant que les variables explicatives contiennent un terme constant.

CHAPITRE 3

Le maximum de vraisemblance

Le maximum de vraisemblance est une méthode d'estimation qui repose sur la distribution conditionnelle des variables que l'on étudie. Intuitivement, elle consiste à estimer un paramètre inconnu en choisissant la valeur de ce paramètre qui maximise la "probabilité" d'observer l'échantillon que l'on observe effectivement. La vraisemblance de l'échantillon est soit la probabilité d'observer l'échantillon (cas discret) soit la densité correspondante (cas continu).

3.1 Définitions et propriétés

Plus généralement, on suppose que la variable expliquée y admet une distribution conditionnelle par rapport aux variables explicatives X dont la densité conditionnelle ou probabilité conditionnelle est notée $f(y|X;\theta)$ où θ est le paramètre que l'on cherche à estimer. On suppose ici que les N observations présentes dans l'échantillon $y = (y_1, ..., y_N)$ sont indépendantes. La vraisemblance de l'échantillon, notée L, est définie par :

$$L(y|X;\theta) = \prod_{i=1}^{N} f(y_i|X_i;\theta).$$

Voici quelques exemples, pour des modèles sans variable explicative.

Exemple 3.1 *Loi normale. Soit un échantillon de variables réelles* $(y_1, ..., y_N)$ *iid selon une loi normale* $N(\theta, \omega)$ *où* ω *est un nombre positif connu. Sa densité est donnée par :*

$$f(y;\theta) = \frac{1}{\sqrt{2\pi\omega}} \exp\left\{-\frac{1}{2\omega}(y-\theta)^2\right\}.$$

La log-vraisemblance de cet échantillon est égale à :

$$\ell\left(y|X;\theta\right)=\sum_{i=1}^{N}\ln f\left(y_i|\theta\right)=-\frac{N}{2}\ln\left(2\pi\omega\right)-\frac{1}{2\omega}\sum_{i=1}^{N}\left(y_i-\theta\right)^2.$$

Exemple 3.2 *Loi de Poisson. Soit un échantillon de variables entières positives* $(y_1,...,y_N)$ *iid selon une loi de Poisson de paramètre* θ *:*

$$f\left(y;\theta\right)=\frac{\exp\left(-\theta\right)\theta^y}{y!}.$$

La log-vraisemblance de cet échantillon est égale à :

$$\ell\left(y|X;\theta\right)=\sum_{i=1}^{N}\ln f\left(y_i|\theta\right)=-N\theta+\ln\left(\theta\right)\sum_{i=1}^{N}y_i-\sum_{i=1}^{N}\ln\left(y_i!\right).$$

Exemple 3.3 *Loi de Bernoulli. Soit un échantillon de variables dichotomiques* $(y_1,...,y_N)$ *iid selon une loi de Bernoulli de paramètre* θ*. Les probabilités sont égales à :*

$$f\left(y;\theta\right)=\theta^y\left(1-\theta\right)^{1-y}=\left\{\begin{array}{ll}\theta & si\ y=1\\ 1-\theta & si\ y=0\end{array}\right..$$

La log-vraisemblance de cet échantillon est égale à :

$$\ell\left(y|X;\theta\right)=\sum_{i=1}^{N}\ln f\left(y_i|\theta\right)=\ln\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^{N}y_i+N\ln\left(1-\theta\right).$$

La méthode du maximum de vraisemblance consiste à estimer θ par $\widehat{\theta}_n$ tel que :

$$L\left(y|X;\widehat{\theta}_n\right)\geq L\left(y|X;\theta\right)\quad\forall\theta\in\Theta,$$

où Θ est l'ensemble des valeurs admissibles du paramètre θ. Cet estimateur est appelé estimateur du maximum de vraisemblance de θ ou en abrégé EMV de θ. Notons ici que cette inégalité est équivalente à :

$$\ln L\left(y|X;\widehat{\theta}_n\right)\geq\ln L\left(y|X;\theta\right)\quad\forall\theta\in\Theta,$$

de sorte que l'on peut maximiser la log-vraisemblance $\ln L$ au lieu de la vraisemblance L. Cette méthode permet de simplifier l'écriture des dérivées de la fonction objectif, car la dérivée d'une somme est plus simple que la dérivée d'un produit.

PROPRIÉTÉ 3.1 *Sous les hypothèses de régularité habituelles (Gouriéroux et Monfort, 1989, ch. VII, p. 192), que nous supposons vérifiées par la*

suite, les estimateurs du maximum de vraisemblance sont convergents, asymptotiquement normaux et asymptotiquement efficaces (i.e., à variance minimale parmi les estimateurs convergents) :

$$\sqrt{N}\left(\widehat{\theta}_N - \theta\right) \underset{N\to+\infty}{\xrightarrow{L}} N\left[0, I_1^{-1}(\theta)\right],$$

où $I_1(\theta)$ est la matrice d'information de Fisher définie par :

$$I_1(\theta) = \underset{X}{\mathrm{E}}\,\underset{y}{\mathrm{V}}\left[\frac{\partial \ln f(y|X,\theta)}{\partial\theta}\right] = \underset{X}{\mathrm{E}}\,\underset{y}{\mathrm{E}}\left[\frac{\partial \ln f(y|X,\theta)}{\partial\theta}\frac{\partial \ln f(y|X,\theta)}{\partial\theta'}\right].$$

De plus, en vertu de l'égalité de la matrice d'information, on peut aussi utiliser la matrice $J_1(\theta)$:

$$J_1(\theta) = \underset{X}{\mathrm{E}}\,\underset{y}{\mathrm{E}}\left[-\frac{\partial^2 \ln f(y|X,\theta)}{\partial\theta\partial\theta'}\right],$$

car on a $I_1(\theta) = J_1(\theta)$. La distribution de $\widehat{\theta}_N$ peut donc être approximée par :

$$\widehat{\theta}_N \overset{A}{\rightsquigarrow} N\left(\theta, \frac{1}{N}\times I_1\left(\widehat{\theta}_N\right)^{-1}\right) = N\left(\theta, I_N\left(\widehat{\theta}_N\right)^{-1}\right).$$

où $\overset{A}{\rightsquigarrow}$ désigne une distribution asymptotique (i.e., utilisable pour de grands échantillons).

Remarque 3.1 *La matrice d'information de l'ensemble de l'échantillon est définie par $I_N(\theta) = N\times I_1(\theta)$, on a donc :*

$$\frac{1}{N} I_1^{-1}(\theta) = \left(N \times I_1(\theta)\right)^{-1} = I_N(\theta)^{-1}.$$

Pour voir d'où vient le résultat de normalité asymptotique, il suffit de partir de la définition de l'estimateur du maximum de vraisemblance. Cette définition est implicite et donnée par la condition du premier ordre :

$$\frac{\partial \ln L\left(y|X,\widehat{\theta}\right)}{\partial\theta} = \sum_{i=1}^{N}\frac{\partial \ln f\left(y_i|X_i,\widehat{\theta}\right)}{\partial\theta} = 0. \tag{3.1}$$

En effectuant un développement limité de $\partial \ln L/\partial\theta$ au voisinage de $\widehat{\theta}$ on obtient :

$$\frac{\partial \ln L\left(y|X,\widehat{\theta}\right)}{\partial\theta} \overset{A}{=} \frac{\partial \ln L(y|X,\theta)}{\partial\theta} + \frac{\partial^2 \ln L(y|X,\theta)}{\partial\theta\partial\theta'}\left(\widehat{\theta}-\theta\right).$$

On remarque ici que ce développement limité devient exact quand $N \to +\infty$. La condition (3.1) implique que :

$$0 \overset{A}{=} \frac{\partial \ln L(y|X,\theta)}{\partial\theta} + \frac{\partial^2 \ln L(y|X,\theta)}{\partial\theta\partial\theta'}\left(\widehat{\theta}-\theta\right),$$

de sorte que l'on peut écrire :

$$\widehat{\theta} - \theta \overset{A}{=} \left[-\frac{\partial^2 \ln L\left(y|X,\theta\right)}{\partial\theta\partial\theta'} \right]^{-1} \frac{\partial \ln L\left(y|X,\theta\right)}{\partial\theta}$$

$$\Leftrightarrow \sqrt{N}\left(\widehat{\theta} - \theta\right) \overset{A}{=} \left[-\frac{1}{N}\frac{\partial^2 \ln L\left(y|X,\theta\right)}{\partial\theta\partial\theta'} \right]^{-1} \frac{1}{\sqrt{N}} \frac{\partial \ln L\left(y|X,\theta\right)}{\partial\theta}$$

La première quantité du membre de droite de l'équation est une moyenne qui converge en probabilité vers l'espérance mathématique correspondante. En appliquant la loi des grands nombres :

$$\begin{aligned} -\frac{1}{N}\frac{\partial^2 \ln L\left(y|X,\theta\right)}{\partial\theta\partial\theta'} &= \frac{1}{N}\sum_{i=1}^{N}\left[-\frac{\partial^2 \ln f\left(y_i|X_i,\theta\right)}{\partial\theta\partial\theta'} \right] \\ &\xrightarrow{p} \underset{X}{\mathrm{E}}\ \underset{y}{\mathrm{E}}\left[-\frac{\partial^2 \ln f\left(y|X,\theta\right)}{\partial\theta\partial\theta'} \right] \overset{\Delta}{=} \mathrm{J}_1\left(\theta\right). \end{aligned}$$

Le second terme du membre de droite de l'équation suit, asymptotiquement, une loi normale. On peut écrire :

$$\begin{aligned} \frac{1}{\sqrt{N}}\frac{\partial \ln L\left(y|X,\theta\right)}{\partial\theta} &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\frac{\partial \ln f\left(y_i|X_i,\theta\right)}{\partial\theta} \\ &= \frac{1}{N}\sum_{i=1}^{N}\sqrt{N}\frac{\partial \ln f\left(y_i|X_i,\theta\right)}{\partial\theta} \\ &= \frac{1}{N}\sum_{i=1}^{N} z_i, \end{aligned}$$

où z_i est la variable dont on cherche la distribution. On a alors, sous les hypothèses usuelles :

$$\sqrt{N}\left(\overline{z} - \mathrm{E}\left(\overline{z}\right)\right) \underset{N\to+\infty}{\xrightarrow{L}} \mathrm{N}\left(0, \mathrm{V}\left(\overline{z}\right)\right), \tag{3.2}$$

où p est le nombre d'éléments de θ. Pour appliquer le théorème de la limite centrale, on a besoin de l'espérance et de la variance de z_i. Pour trouver l'espérance de z_i on utilise la propriété suivante :

PROPRIÉTÉ 3.2 *Soit $f\left(y|X,\theta\right)$ la densité conditionnelle de la variable expliquée. Elle vérifie la propriété suivante :*

$$\underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta}\right] = 0.$$

Preuve :

$$\begin{aligned}
\underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta}\right] &= \underset{y}{\mathrm{E}}\left[\frac{1}{f\left(y|X,\theta\right)}\frac{\partial f\left(y|X,\theta\right)}{\partial\theta}\right] \\
&= \int \frac{1}{f\left(y|X,\theta\right)}\frac{\partial f\left(y|X,\theta\right)}{\partial\theta} f\left(y|X,\theta\right)\mathrm{d}y \\
&= \int \frac{\partial f\left(y|X,\theta\right)}{\partial\theta}\mathrm{d}y \\
&= \frac{\partial}{\partial\theta}\underbrace{\int f\left(y|X,\theta\right)\mathrm{d}y}_{1} \\
&= 0,
\end{aligned}$$

□

On voit que :

$$\mathrm{E}\left(z_i\right) = \sqrt{N}\ \underset{X}{\mathrm{E}}\ \underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta}\right] = 0,$$

Passons maintenant à la variance; nous avons besoin de la quantité suivante :

$$\begin{aligned}
\underset{y}{\mathrm{V}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta}\right] &= \underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta}\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta'}\right] \\
&\quad \underbrace{-\underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta}\right]\underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta'}\right]}_{0} \\
&= \underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta}\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta'}\right]
\end{aligned}$$

Pour calculer la variance de z_i, on utilise la propriété suivante :

PROPRIÉTÉ 3.3 *Soit $f\left(y|X,\theta\right)$ la densité conditionnelle de la variable expliquée. Elle vérifie la propriété suivante :*

$$\underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta}\frac{\partial \ln f\left(y|X,\theta\right)}{\partial\theta'}\right] = \underset{y}{\mathrm{E}}\left[-\frac{\partial^2 \ln f\left(y|X,\theta\right)}{\partial\theta\partial\theta'}\right].$$

Preuve :

On dérive la relation suivante par rapport à θ' :

$$\int \frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta} f\left(y|X,\theta\right) \mathrm{d}y = 0$$

$$\Rightarrow \int \left\{ \frac{\partial^2 \ln f\left(y|X,\theta\right)}{\partial \theta \partial \theta'} f\left(y|X,\theta\right) + \frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta} \frac{\partial f\left(y|X,\theta\right)}{\partial \theta'} \right\} \mathrm{d}y = 0$$

or

$$\frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta} = \frac{1}{f\left(y|X,\theta\right)} \frac{\partial f\left(y|X,\theta\right)}{\partial \theta'}$$
$$\Leftrightarrow \frac{\partial f\left(y|X,\theta\right)}{\partial \theta'} = \frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta} f\left(y|X,\theta\right),$$

en remplaçant dans la relation (3.3), on obtient :

$$\int \left\{ \frac{\partial^2 \ln f\left(y|X,\theta\right)}{\partial \theta \partial \theta'} + \frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta} \frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta'} \right\} f\left(y|X,\theta\right) \mathrm{d}y = 0 \tag{3.3}$$

$$\Leftrightarrow$$

$$\int \frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta} \frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta'} f\left(y|X,\theta\right) \mathrm{d}y =$$
$$- \int \frac{\partial^2 \ln f\left(y|X,\theta\right)}{\partial \theta \partial \theta'} f\left(y|X,\theta\right) \mathrm{d}y \tag{3.4}$$

$$\Leftrightarrow$$

$$\mathop{\mathrm{E}}_{y}\left[\frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta} \frac{\partial \ln f\left(y|X,\theta\right)}{\partial \theta'} \right] = \mathop{\mathrm{E}}_{y}\left[- \frac{\partial^2 \ln f\left(y|X,\theta\right)}{\partial \theta \partial \theta'} \right].$$

□

La variance de z_i est donnée par la formule de la variance totale :

$$\begin{aligned}
\mathrm{V}\left(z_i\right) &= \underset{X}{\mathrm{E}}\ \underset{y}{\mathrm{V}}\left[z_i\right] + \underset{X}{\mathrm{V}}\ \underbrace{\underset{y}{\mathrm{E}}\left[z_i\right]}_{0} \\
&= N\ \underset{X}{\mathrm{E}}\left[\ \underset{y}{\mathrm{V}}\left(\frac{\partial \ln f\left(y_i|X_i,\theta\right)}{\partial\theta}\right)\right] \\
&= N\ \underset{X}{\mathrm{E}}\ \underset{y}{\mathrm{E}}\left[\frac{\partial \ln f\left(y_i|X_i,\theta\right)}{\partial\theta}\frac{\partial \ln f\left(y_i|X_i,\theta\right)}{\partial\theta'}\right] \\
&= N\ \mathrm{I}_1\left(\theta\right),
\end{aligned}$$

donc

$$\begin{aligned}
\mathrm{V}\left(\frac{1}{N}\sum_{i=1}^{N} z_i\right) &= \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{V}\left(z_i\right) \\
&= \frac{N\,\mathrm{V}\left(z_i\right)}{N^2} \\
&= \frac{N^2\,\mathrm{I}_1\left(\theta\right)}{N^2} \\
&= \mathrm{I}_1\left(\theta\right),
\end{aligned}$$

qui est une quantité finie. Globalement, on trouve que :

$$\frac{1}{\sqrt{N}}\frac{\partial \ln L\left(y|X,\theta\right)}{\partial\theta} = \overline{z} \underset{N\to+\infty}{\xrightarrow{L}} \mathrm{N}\left(0,\mathrm{I}_1\left(\theta\right)\right),$$

et l'on déduit de (3.2) que :

$$\sqrt{N}\left(\widehat{\theta}-\theta\right) \overset{A}{=} \mathrm{J}_1\left(\theta\right)^{-1}\overline{z},$$

converge en loi vers une distribution normale d'espérance nulle et de variance :

$$\mathrm{J}_1\left(\theta\right)^{-1}\mathrm{V}\left(\overline{z}\right)\mathrm{J}_1\left(\theta\right)^{-1} = \mathrm{J}_1\left(\theta\right)^{-1}\mathrm{I}_1\left(\theta\right)\mathrm{J}_1\left(\theta\right)^{-1} = \mathrm{J}_1\left(\theta\right)^{-1} = \mathrm{I}_1\left(\theta\right)^{-1}.$$

□

Une dernière propriété est utile, celle de l'invariance fonctionnelle. Elle permet de retrouver l'estimateur du maximum de vraisemblance après un changement de paramètres.

Si l'on effectue un changement de paramètre du type $\tau = h(\theta)$, où h est une fonction inversible choisie par l'économètre, on a :

$$L\left(y|X, h^{-1}(\widehat{\tau}_n)\right) = L\left(y|X, \widehat{\theta}_n\right) \geq L(y|X, \theta) = L\left(y|X, h^{-1}(\tau)\right), \forall \theta \in \Theta$$

donc $\widehat{\tau}_n$ est l'estimateur du maximum de vraisemblance de τ. Il n'est donc pas nécessaire de réestimer le modèle quand on effectue un changement de paramètre.

PROPRIÉTÉ 3.4 (Invariance fonctionnelle)
Soit $\widehat{\theta}_N$ un estimateur du maximum de vraisemblance de θ et $\tau = h(\theta)$ un changement de paramètre. L'estimateur du maximum de vraisemblance de τ est donné par $\widehat{\tau}_N = h\left(\widehat{\theta}_N\right)$.

Cette propriété est très pratique car certaines log-vraisemblances ne sont concaves que par rapport à un changement de paramètres bien précis (e.g., dans le modèle tobit généralisé). On est alors certain que l'optimum ne dépend pas de ce changement de paramètre, et que les algorithmes courants convergent vers ce maximum. On effectue donc toujours les changements de paramètres de ce type quand ils existent. Mais encore faut-il remonter des nouveaux paramètres τ aux paramètres structurels du modèle θ. Pour cela on utilise le théorème de Slutsky.

THÉORÈME 3.1 (Slutsky)
Soit h une fonction de classe C^1 (dérivable et de dérivée première continue), une relation entre deux paramètres $\tau = h(\theta)$, et $\widehat{\theta}_N$ estimateur convergent de θ vérifiant :

$$\sqrt{N}\left(\widehat{\theta}_N - \theta\right) \underset{N\to+\infty}{\xrightarrow{L}} N\left(0, \Omega_{\hat{\theta}}\right)$$

alors,

$$\sqrt{N}\left(h\left(\widehat{\theta}_N\right) - h(\theta)\right) \underset{N\to+\infty}{\xrightarrow{L}} N\left[0, \frac{\partial h}{\partial \theta}(\theta)\, \Omega_{\hat{\theta}} \frac{\partial h}{\partial \theta'}(\theta)\right]$$

Notons bien que ce théorème s'applique même si $\widehat{\theta}_N$ n'est pas un estimateur du maximum de vraisemblance et même si la fonction h n'est pas inversible. Dans la pratique, on estimera la variance asymptotique de $h\left(\widehat{\theta}_n\right)$ par :

$$\widehat{\text{Vas}}\left[h\left(\widehat{\theta}_N\right)\right] = \frac{1}{N} \frac{\partial h}{\partial \theta}\left(\widehat{\theta}_N\right) \widehat{\Omega}_{\hat{\theta}} \frac{\partial h}{\partial \theta'}\left(\widehat{\theta}_N\right),$$

où $\widehat{\Omega}_{\hat{\theta}}$ est un estimateur convergent de $\Omega_{\hat{\theta}}$.

Preuve :

Pour comprendre ce résultat, il suffit de faire un développement limité de $h\left(\widehat{\theta}_N\right)$ au voisinage de θ :

$$h\left(\widehat{\theta}_N\right) \overset{A}{=} h\left(\theta\right) + \frac{\partial h}{\partial \theta}\left(\theta\right)\left(\widehat{\theta}_N - \theta\right)$$
$$\Leftrightarrow \sqrt{N}\left(h\left(\widehat{\theta}_N\right) - h\left(\theta\right)\right) \overset{A}{=} \frac{\partial h}{\partial \theta}\left(\theta\right)\sqrt{N}\left(\widehat{\theta}_N - \theta\right),$$

cette expression est une transformation linéaire de $\sqrt{N}\left(\widehat{\theta}_N - \theta\right)$, en conséquence elle suit asymptotiquement une loi normale. Son espérance mathématique est égale à :

$$\mathrm{E}\left(\sqrt{N}\left(h\left(\widehat{\theta}_N\right) - h\left(\theta\right)\right)\right) = \mathrm{E}\left(\frac{\partial h}{\partial \theta}\left(\theta\right)\sqrt{N}\left(\widehat{\theta}_N - \theta\right)\right)$$
$$= \frac{\partial h}{\partial \theta}\left(\theta\right)\underbrace{\mathrm{E}\left(\sqrt{N}\left(\widehat{\theta}_N - \theta\right)\right)}_{0} = 0,$$

et sa variance est égale à :

$$\mathrm{V}\left(\sqrt{N}\left(h\left(\widehat{\theta}_N\right) - h\left(\theta\right)\right)\right) = \mathrm{V}\left(\frac{\partial h}{\partial \theta}\left(\theta\right)\sqrt{N}\left(\widehat{\theta}_N - \theta\right)\right)$$
$$= \frac{\partial h}{\partial \theta}\left(\theta\right)\underbrace{\mathrm{V}\left(\sqrt{N}\left(\widehat{\theta}_N - \theta\right)\right)}_{\Omega_{\hat{\theta}}}\frac{\partial h}{\partial \theta'}\left(\theta\right).$$

□

3.2 Estimation

Les points candidats à un maximum sont obtenus par la résolution des conditions du premier ordre. En effet, dans les cas usuels les conditions du premier ordre fournissent un maximum local sous réserve de vérification de la condition du second ordre. Il faut alors rechercher numériquement les maxima locaux et prendre celui qui fournit la valeur la plus élevée de la vraisemblance. Toutefois, la plupart des modèles que nous verrons dans ce cours possèdent une log-vraisemblance concave. Dans ce cas particulier, le maximum est unique et donné par les conditions du premier ordre. La log-vraisemblance est égale à :

$$\ell\left(y|X,\theta\right) = \ln L\left(y|X,\theta\right).$$

Quand la solution est unique, on cherche la solution du problème d'optimisation :[1]

$$\widehat{\theta}_N = \arg\max_{\theta} \ell\left(y|X,\theta\right).$$

La condition du premier ordre pour un maximum local est donnée par la nullité du score :

$$\frac{\partial \ell}{\partial \theta}\left(y|X,\widehat{\theta}_N\right) = 0,$$

et la condition du second ordre par un hessien défini négatif :

$$\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\left(y|X,\widehat{\theta}_N\right) \ll 0,$$

où $\ll$ désigne l'infériorité au sens des matrices.

En général l'expression de $\widehat{\theta}_N$ ne peut pas être obtenue directement en fonction des observations, c'est-à-dire sous forme explicite[2], il faut donc recourir à des algorithmes d'optimisation numérique pour effectuer une estimation par le maximum de vraisemblance. Une fois cette valeur obtenue, il nous faut estimer la matrice de covariance de $\widehat{\theta}_N$.

En utilisant la loi des grands nombres, on peut estimer les moments théoriques par les moments empiriques correspondants, soit :

$$\begin{aligned}
\mathrm{I}_1\left(\theta\right) &= \mathop{\mathrm{E}}_{X}\mathop{\mathrm{E}}_{y}\left[\frac{\partial \ln f}{\partial \theta}\left(y|X,\theta\right)\frac{\partial \ln f}{\partial \theta'}\left(y|X,\theta\right)\right] \\
\text{par}\quad \widehat{\mathrm{I}}_1\left(\widehat{\theta}_N\right) &= \frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ln f}{\partial \theta}\left(y_i|X_i,\widehat{\theta}_N\right)\frac{\partial \ln f}{\partial \theta'}\left(y_i|X_i,\widehat{\theta}_N\right), \\
\text{et}\quad \mathrm{J}_1\left(\theta\right) &= \mathop{\mathrm{E}}_{X}\mathop{\mathrm{E}}_{y}\left[-\frac{\partial^2 \ln f}{\partial \theta \partial \theta'}\left(y|X,\theta\right)\right] \\
\text{par}\quad \widehat{\mathrm{J}}_1\left(\widehat{\theta}_N\right) &= -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2 \ln f}{\partial \theta \partial \theta'}\left(y_i|X_i,\widehat{\theta}_N\right).
\end{aligned}$$

Il suffit de prendre l'inverse de l'une de ces deux matrices pour obtenir un estimateur convergent de la matrice de covariance asymptotique de $\sqrt{N}\left(\widehat{\theta}_N - \theta\right)$. Voici quelques exemples, pour des modèles sans variable explicative.

[1] Quand la solution n'est pas unique, on a $\widehat{\theta}_n \in \arg\max_\theta \ell\left(y|x,\theta\right)$ car il y a un ensemble de solutions.

[2] Le modèle linéaire standard est une exception à cette règle.

Exemple 3.4 *Loi normale $N(\theta, \omega)$. On maximise la log-vraisemblance :*

$$\ell\left(y|X,\theta\right) = -\frac{N}{2}\ln\left(2\pi\omega\right) - \frac{1}{2\omega}\sum_{i=1}^{N}\left(y_i-\theta\right)^2,$$

ce qui donne la condition du premier ordre :

$$\frac{\partial \ell}{\partial \theta}\left(y|\widehat{\theta}_N\right) = \frac{1}{\omega}\sum_{i=1}^{N}\left(y_i-\widehat{\theta}_N\right),$$

qui permet d'obtenir l'estimateur du maximum de vraisemblance :

$$\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$$

avec

$$\frac{\partial^2 \ell}{\partial \theta^2}\left(y|\theta\right) = -\frac{N}{\omega} < 0 \quad \forall\theta.$$

Pour trouver la distribution asymptotique de $\widehat{\theta}_N$, on peut utiliser soit $I_N\left(\theta\right)$ soit $J_N\left(\theta\right)$. On a :

$$\begin{aligned} I_N\left(\theta\right) &= \sum_{i=1}^{N}\underset{y}{\mathrm{E}}\left[\frac{\partial \ln f}{\partial\theta}\left(y_i|\theta\right)^2\right] \\ &= \frac{1}{\omega^2}\sum_{i=1}^{N}\underset{y}{\mathrm{E}}\left[\left(y_i-\theta\right)^2\right] \\ &= \frac{N}{\omega}, \end{aligned}$$

de même :

$$J_N\left(\theta\right) = \sum_{i=1}^{N}\underset{y}{\mathrm{E}}\left[-\frac{\partial^2 \ln f}{\partial\theta^2}\left(y_i|\theta\right)\right] = \frac{N}{\omega}.$$

et l'on obtient :

$$\widehat{\theta}_N \overset{A}{\rightsquigarrow} N\left(\theta, \omega/N\right).$$

Exemple 3.5 *Loi de Poisson $P(\theta)$. On maximise la log-vraisemblance :*

$$\ell\left(y|X,\theta\right) = -N\theta + \ln\left(\theta\right)\sum_{i=1}^{N} y_i - \sum_{i=1}^{N}\ln\left(y_i!\right),$$

ce qui donne la condition du premier ordre :

$$\frac{\partial \ell}{\partial \theta}\left(y|\widehat{\theta}_N\right) = -N + \frac{1}{\widehat{\theta}_N}\sum_{i=1}^{N} y_i = 0,$$

qui permet d'obtenir l'estimateur du maximum de vraisemblance :

$$\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$$

avec

$$\frac{\partial^2 \ell}{\partial \theta^2}(y|\theta) = -\frac{1}{\theta^2}\sum_{i=1}^{N} y_i < 0 \quad \forall \theta.$$

Pour trouver la distribution asymptotique de $\widehat{\theta}_N$*, on peut utiliser soit* $I_N(\theta)$ *soit* $J_N(\theta)$*. On a :*

$$\begin{aligned} I_N(\theta) &= \sum_{i=1}^{N} \underset{y}{\mathrm{V}}\left[\frac{\partial \ln f}{\partial \theta}(y_i|\theta)\right] \\ &= \frac{1}{\theta^2}\sum_{i=1}^{N} \underbrace{\underset{y}{\mathrm{V}}[y_i]}_{\theta} \\ &= \frac{N}{\theta}, \end{aligned}$$

de même :

$$\begin{aligned} J_N(\theta) &= \sum_{i=1}^{N} \underset{y}{\mathrm{E}}\left[-\frac{\partial^2 \ln f}{\partial \theta^2}(y_i|\theta)\right] \\ &= \frac{1}{\theta^2}\sum_{i=1}^{N} \underbrace{\underset{y}{\mathrm{E}}(y_i)}_{\theta} \\ &= \frac{N}{\theta}. \end{aligned}$$

et l'on obtient :

$$\widehat{\theta}_N \overset{A}{\rightsquigarrow} N(\theta, \theta/N).$$

Exemple 3.6 *Loi de Bernoulli* $B(\theta)$*. On maximise la log-vraisemblance :*

$$\ell(y|X;\theta) = \ln\left(\frac{\theta}{1-\theta}\right)\sum_{i=1}^{N} y_i + N\ln(1-\theta).$$

ce qui donne la condition du premier ordre :

$$\frac{\partial \ell}{\partial \theta}\left(y|\widehat{\theta}_N\right) = \frac{1}{\widehat{\theta}_N\left(1-\widehat{\theta}_N\right)}\sum_{i=1}^{N} y_i - \frac{N}{1-\widehat{\theta}_N} = 0,$$

qui permet d'obtenir l'estimateur du maximum de vraisemblance :

$$\widehat{\theta}_N = \frac{1}{N}\sum_{i=1}^{N} y_i$$

avec

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \theta^2}(y|\theta) &= -\frac{1-2\theta}{\theta^2(1-\theta)^2}\sum_{i=1}^{N} y_i - \frac{N}{(1-\theta)^2} \\ &= -\frac{(1-\theta)^2\sum_{i=1}^{N} y_i + \theta^2\sum_{i=1}^{N}(1-y_i)}{\theta^2(1-\theta)^2} < 0 \quad \forall\theta,\end{aligned}$$

en utilisant $N = \sum_{i=1}^{N} y_i + \sum_{i=1}^{N}(1-y_i)$. *Pour trouver la distribution asymptotique de* $\widehat{\theta}_N$, *on peut utiliser soit* $I_N(\theta)$ *soit* $J_N(\theta)$. *On a :*

$$\begin{aligned}I_N(\theta) &= \sum_{i=1}^{N} \underset{y}{\mathrm{V}}\left[\frac{\partial \ln f}{\partial \theta}(y_i|\theta)\right] \\ &= \frac{1}{\theta^2(1-\theta)^2}\sum_{i=1}^{N} \underbrace{\underset{y}{\mathrm{V}}[y_i]}_{\theta(1-\theta)} \\ &= \frac{N}{\theta(1-\theta)},\end{aligned}$$

de même :

$$\begin{aligned}J_N(\theta) &= \sum_{i=1}^{N} \underset{y}{\mathrm{E}}\left[-\frac{\partial^2 \ln f}{\partial \theta^2}(y_i|\theta)\right] \\ &= \frac{1}{\theta^2(1-\theta)^2}\sum_{i=1}^{N}\left[(1-2\theta)\underset{y}{\mathrm{E}}(y_i) + \theta^2\right] \\ &= \frac{N}{\theta(1-\theta)}.\end{aligned}$$

et l'on obtient :

$$\widehat{\theta}_N \overset{A}{\rightsquigarrow} N(\theta, \theta(1-\theta)/N).$$

3.3 Les moindres carrés ordinaires

Soit le modèle linéaire standard :

$$y_i = X_i b + u_i, \quad i = 1, ..., N$$

où l'indice i désigne une observation, N la taille de l'échantillon, y_i la variable expliquée, X_i le vecteur des variables explicatives, b le vecteur de leurs coefficients et u_i la perturbation du modèle. On suppose que les perturbations sont indépendantes et identiquement distribuées (iid) selon une loi normale $\mathrm{N}(0, \omega)$. Ces hypothèses impliquent que les variables expliquées y_i, transformations linéaires des u_i, sont indépendantes et suivent des lois normales $\mathrm{N}(X_i b, \omega)$. En conséquence la densité conditionnelle de la i-ème observation est égale à :

$$f\left(y_i|X_i; b, \omega\right) = \frac{1}{\sqrt{2\pi\omega}} \exp\left\{-\frac{\left(y_i - X_i b\right)^2}{2\omega}\right\}.$$

Le paramètre à estimer est donc :

$$\theta = \begin{pmatrix} b \\ \omega \end{pmatrix}.$$

Comme les observations sont indépendantes, la vraisemblance de l'échantillon, notée L, est égale au produit des densités individuelles :

$$\begin{aligned} L\left(y|X;\theta\right) &= \prod_{i=1}^{N} f\left(y_i|X_i;\theta\right) \\ &= \left(2\pi\omega\right)^{-\frac{N}{2}} \exp\left\{-\frac{1}{2\omega}\sum_{i=1}^{N}\left(y_i - X_i b\right)^2\right\}. \end{aligned}$$

La méthode du maximum de vraisemblance consiste à choisir des valeurs de b et ω qui rendent cette densité jointe la plus grande possible. La log-vraisemblance de l'échantillon est égale à :

$$\ell\left(y|X;\theta\right) = -\frac{N}{2}\ln\left(2\pi\omega\right) - \frac{1}{2\omega}\sum_{i=1}^{N}\left(y_i - X_i b\right)^2.$$

On pose :

$$\widehat{\theta}_N = \begin{pmatrix} \widehat{b}_N \\ \widehat{\omega}_N \end{pmatrix}.$$

Les conditions du premier ordre sont données par :

$$\frac{\partial \ell}{\partial \theta}\left(y|X,\widehat{\theta}_n\right) = \begin{bmatrix} \dfrac{\partial \ell}{\partial b}\left(y|X,\widehat{\theta}_N\right) \\ \dfrac{\partial \ell}{\partial \omega}\left(y|X,\widehat{\theta}_N\right) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Leftrightarrow \begin{bmatrix} \dfrac{1}{\widehat{\omega}_N}\displaystyle\sum_{i=1}^{N} X_i'\left(y_i - X_i\widehat{b}_N\right) \\ -\dfrac{N}{2\widehat{\omega}_N} + \dfrac{1}{2\widehat{\omega}_N^2}\displaystyle\sum_{i=1}^{N}\left(y_i - X_i\widehat{b}_N\right)^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

ce qui implique :

$$\widehat{b}_N = \left(\sum_{i=1}^{N} X_i'X_i\right)^{-1}\sum_{i=1}^{N} X_i'y_i \quad \text{et} \quad \widehat{\omega}_N = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - X_i\widehat{b}_N\right)^2.$$

On retrouve l'estimateur des moindres carrés ordinaires de b. Par contre l'estimateur de la variance ne comporte pas de correction pour les degrés de liberté.[3] On estime la matrice de covariance asymptotique à partir de :

$$\mathrm{J}_N(\theta) = \sum_{i=1}^{N}\mathop{\mathrm{E}}_{\theta}\begin{bmatrix} \dfrac{1}{\omega}X_i'X_i & \dfrac{1}{\omega^2}X_i'\left(y_i - X_ib\right) \\ \dfrac{1}{\omega^2}X_i\left(y_i - X_ib\right) & -\dfrac{1}{2\omega^2} + \dfrac{1}{\omega^3}\left(y_i - X_ib\right)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \dfrac{1}{\omega}\sum_{i=1}^{N} X_i'X_i & 0 \\ 0 & \dfrac{N}{2\omega^2} \end{bmatrix}$$

La matrice de covariance asymptotique de $\widehat{\theta}_N$ est donc estimée par :

$$\widehat{\mathrm{Vas}}\left(\widehat{\theta}\right) = \widehat{\mathrm{J}}_N\left(\widehat{\theta}_N\right)^{-1} = \begin{bmatrix} \widehat{\omega}_N\left[\sum_{i=1}^{N} X_i'X_i\right]^{-1} & 0 \\ 0 & \dfrac{2\widehat{\omega}_N^2}{N} \end{bmatrix},$$

elle permet de construire un intervalle de confiance sur la variance. Pour raisonner sur l'écart-type, on utilise l'invariance fonctionnelle et le théorème de Slutsky. L'invariance fonctionnelle garantit que :

$$\widehat{\sigma}_N = \sqrt{\widehat{\omega}_N},$$

[3] Habituellement, à distance finie, on estime ω par $1/(N-p)\sum\left(y_i - x_i\widehat{b}_N\right)^2$, où p est le nombre de variables explicatives (y compris le terme constant). La distribution asymptotique de cet estimateur est la même que celle de $\widehat{\omega}_N$.

est bien l'estimateur du maximum de vraisemblance de $\sigma = \sqrt{\mathrm{V}(u_i)}$. Et le théorème de Slutsky permet de calculer la variance de l'estimateur de l'écart-type. On a une transformation :

$$h(\omega) = \sqrt{\omega} \Rightarrow h'(\omega) = \frac{1}{2\sqrt{\omega}},$$

on peut donc écrire que :

$$\begin{aligned}\widehat{\mathrm{Vas}}(\widehat{\sigma}) &= \frac{1}{2\sqrt{\widehat{\omega}_N}} \frac{2\widehat{\omega}_N^2}{N} \frac{1}{2\sqrt{\widehat{\omega}_N}} \\ &= \frac{\widehat{\omega}_N}{2N}.\end{aligned}$$

CHAPITRE 4

Les algorithmes d'optimisation

4.1 Présentation des algorithmes

Le problème que nous devons résoudre est de trouver une valeur numérique $\widehat{\theta}_N$ qui résout un système d'équations de la forme :[1]

$$s\left(\widehat{\theta}_N\right) = 0$$

où $s(.)$ est une fonction connue des observations et du paramètre θ. Elle dépend aussi bien des données que du modèle postulé (distribution, paramètres). Dans le cas du maximum de vraisemblance $s(\theta)$ est appelé le *score*, dans le cas du pseudo maximum de vraisemblance, on l'appelle le *pseudo score*. Quand la fonction à maximiser est concave, cette condition du premier ordre est suffisante pour un maximum global. Ce sera le cas du modèle Logit, avec lequel nous ferons une des premières applications. La technique que l'on utilise pour parvenir à la valeur $\widehat{\theta}_N$ s'appelle un algorithme. On peut décomposer un algorithme en quatre grandes étapes.

1. *Une valeur initiale. Le choix de la valeur initiale n'est pas problématique quand l'objectif est concave. Dans ce cas, avec un algorithme croissant (voir plus loin), tout point de départ doit mener au maximum. Par contre, quand l'objectif n'est pas concave, ou pour accélérer la procédure quand l'objectif est concave, on prendra un estimateur convergent comme point de départ. Il est en effet possible dans certains cas de trouver un estimateur en deux étapes*

[1] Un système d'équations car $s\left(\widehat{\theta}_N\right)$ est un vecteur.

relativement facile à calculer. Il n'est généralement pas efficace, et c'est la raison pour laquelle on réalise une estimation supplémentaire. On note cette valeur initiale $\theta_{(0)}$.

2. *Une règle d'itération. Une fois la valeur initiale fixée, il faut utiliser une règle qui permette de trouver une nouvelle valeur plus proche du maximum. Le pas de l'itération, défini comme la différence entre deux valeurs successives du paramètre, est déterminé selon différentes méthodes et constitue le coeur de l'algorithme. Nous utiliserons des méthodes dites de gradient et plus particulièrement les algorithmes de Newton-Raphson, du Score, de Berndt-Hall-Hall-Hausman et de Levenberg-Marquardt. La pratique montre qu'ils permettent de traiter la plupart des cas, même difficiles. On résume cette étape par la relation* $\theta_{(p+1)} = \theta_{(p)} + M\left(\theta_{(p)}\right)$ *où* p *est l'itération,* $p \in N$. *La valeur du pas* $M\left(\theta_{(p)}\right)$ *doit dépendre uniquement de la valeur du paramètre à l'étape précédente.*

3. *Une règle d'arrêt de l'algorithme. Au voisinage du maximum la fonction objectif ne doit plus varier, on peut donc baser l'arrêt de la procédure sur la différence entre deux valeurs successives de la fonction objectif. Une seconde condition porte sur le gradient, qui doit être nul (condition du premier ordre). On peut aussi utiliser des variantes, comme l'élasticité de l'objectif aux paramètres du modèle, qui est également nulle à l'optimum et possède l'avantage d'être insensible à un changement d'unité des variables (contrairement au gradient). Dans l'ensemble, tous ces critères sont équivalents à l'optimum.*

4. *Vérifier que l'on a bien atteint un maximum local. Ce problème ne se pose que lorsque l'objectif n'est pas globalement concave. La condition du second ordre pour un optimum local précise que le hessien doit être défini négatif au point en question. Il faudra donc le vérifier systématiquement. Ceci est d'autant plus important que dans le cas du maximum de vraisemblance l'inverse du hessien n'est autre qu'un estimateur convergent de la matrice de covariance du paramètre. En conséquence, si cette propriété n'était pas vérifiée, on obtiendrait un estimateur de la matrice de covariance qui n'est pas défini positif et, pour cette raison, inutilisable.*

4.2 Les méthodes de gradient

Nous voulons maximiser une fonction $\ell(\theta)$, de gradient $s(\theta) = \partial \ell(\theta) / \partial \theta$ et de hessien $H(\theta) = \partial^2 \ell(\theta) / \partial \theta \partial \theta'$. Nous disposons également d'une valeur initiale notée $\theta_{(0)}$. De même, on note $\theta_{(p)}$ la valeur du paramètre

à la p-ième itération. Un algorithme du gradient est une règle d'itération de la forme suivante :

$$\theta_{(p+1)} = \theta_{(p)} + W_{(p)} s\left(\theta_{(p)}\right),$$

où $s\left(\theta_{(p)}\right)$ est le gradient de la fonction que l'on cherche à maximiser et $W_{(p)}$ une matrice qui dépend de l'algorithme particulier que l'on emploie. On vérifie que lorsque l'on a atteint le maximum, $s\left(\widehat{\theta}_n\right) = 0$, et le pas de l'itération est nul. Toutefois, dans la pratique, il peut arriver que le pas d'une itération soit trop fort et dépasse le point qui donne le maximum, on modifie donc la règle précédente en l'écrivant :

$$\theta_{(p+1)} = \theta_{(p)} + \lambda W_{(p)} s\left(\theta_{(p)}\right),$$

où $\lambda \in [0, 1]$. Il est également possible de prendre $\lambda > 1$ au début de l'algorithme pour accélérer la convergence. Le lecteur intéressé par ce dernier cas peut consulter Gouriéroux et Monfort (1989, chap. XIII). La valeur de λ n'est réduite que lorsque $\ell\left(\theta_{(p+1)}\right) < \ell\left(\theta_{(p)}\right)$.

Les algorithmes de gradient sont très employés car il possèdent une propriété intéressante : ils sont *croissants*. Cette propriété signifie que, si $W_{(p)}$ est symétrique et définie positive, alors pour de petits accroissements du pas de l'itération (i.e., λ petit), l'algorithme mène toujours à une valeur supérieure ou égale de l'objectif soit $\ell\left(\theta_{(p+1)}\right) \geq \ell\left(\theta_{(p)}\right)$. Quand la fonction est concave, ceci garantit que l'on parvienne au maximum. Les trois algorithmes de gradient les plus utilisés sont ceux de Newton-Raphson, de Berndt-Hall-Hall-Hausman et du score.

4.2.1 Algorithme de Newton-Raphson

Il consiste à effectuer une approximation quadratique de la fonction à maximiser, en chacun des points de l'itération. Dans ce cas, si le hessien est défini négatif, on obtient le maximum de l'approximation par la condition du premier ordre sur une forme quadratique, dont on peut calculer facilement l'expression analytique parce qu'elle est linéaire. La succession de maxima ainsi obtenue donne la solution du problème. Le développement limité au second ordre de $\ell(\theta)$ au voisinage de $\theta_{(p)}$ est égal à :

$$\ell(\theta) \simeq \ell\left(\theta_{(p)}\right) + s'\left(\theta_{(p)}\right)\left(\theta - \theta_{(p)}\right) + \frac{1}{2}\left(\theta - \theta_{(p)}\right)' H\left(\theta_{(p)}\right)\left(\theta - \theta_{(p)}\right).$$

La maximisation de cette forme quadratique par rapport à θ donne la condition du premier ordre :

$$s\left(\theta_{(p)}\right) + H\left(\theta_{(p)}\right)\left(\theta - \theta_{(p)}\right) = 0 \Leftrightarrow \theta - \theta_{(p)} = -H\left(\theta_{(p)}\right)^{-1} s\left(\theta_{(p)}\right),$$

de plus la dérivée seconde est égale à $H\left(\theta_{(p)}\right)$, qui est définie négative lorsque l'objectif est concave au point $\theta_{(p)}$. Dans ce cas, on a bien un maximum local donné par les conditions du premier ordre. Dans l'ensemble l'itération est donnée par :

$$\theta_{(p+1)} = \theta_{(p)} - \lambda H\left(\theta_{(p)}\right)^{-1} s\left(\theta_{(p)}\right).$$

Cet algorithme présente généralement un pas assez fort dans les premières itérations et, dans l'ensemble, s'avère assez rapide. Tout dépend toutefois si l'on travaille sur des dérivées secondes analytiques ou numériques, car ces dernières augmentent fortement le temps de calcul à chaque itération.

4.2.2 Algorithme de Berndt-Hall-Hall-Hausman

Cet algorithme, justifié dans le cas du maximum de vraisemblance, se base sur l'égalité de la matrice d'information :

$$\mathrm{I}_1(\theta) = \underset{x}{\mathrm{E}}\,\underset{y}{\mathrm{E}}\left[\frac{\partial \ln f(\theta)}{\partial \theta}\frac{\partial \ln f(\theta)}{\partial \theta'}\right] = \underset{x}{\mathrm{E}}\,\underset{y}{\mathrm{E}}\left[-\frac{\partial^2 \ln f(\theta)}{\partial \theta \partial \theta'}\right] = \mathrm{J}_1(\theta).$$

Pour l'estimation, on remplace les moments théoriques par les moments empiriques correspondants, ce qui suggère d'approximer les dérivées secondes par l'opposé des produits croisés des dérivées premières. Pour un échantillon de taille N on peut écrire la fonction objectif et ses dérivées sous la forme :

$$\ell(\theta) = \sum_{i=1}^{N} \ln f(y_i|x_i,\theta),\ s(\theta) = \sum_{i=1}^{N} \frac{\partial \ln f(y_i|x_i,\theta)}{\partial \theta}$$

$$\text{et } H(\theta) = \sum_{i=1}^{N} \frac{\partial^2 f(y_i|x_i,\theta)}{\partial \theta \partial \theta'}$$

La méthode de Berndt-Hall-Hall-Hausman consiste à approximer

$$H(\theta) \quad \text{par} \quad \sum_{i=1}^{N} \frac{\partial \ln f(y_i|x_i,\theta)}{\partial \theta}\frac{\partial \ln f(y_i|x_i,\theta)}{\partial \theta'},$$

ce qui donne la règle d'itération suivante :

$$\theta_{(p+1)} = \theta_{(p)} - \lambda\left[\sum_{i=1}^{N} \frac{\partial \ln f\left(y_i|x_i,\theta_{(p)}\right)}{\partial \theta}\frac{\partial \ln f\left(y_i|x_i,\theta_{(p)}\right)}{\partial \theta'}\right]^{-1} s\left(\theta_{(p)}\right).$$

Cet algorithme ne nécessite que les dérivées au premier ordre et est donc facile à mettre en oeuvre. Toutefois, il implique généralement plus

d'itérations que l'algorithme de Newton-Raphson. Son principal défaut est qu'il ne permet pas de vérifier la négativité du hessien à chaque itération, ce qui peut s'avérer problématique en certains points $\theta_{(p)}$ lorsque l'objectif n'est pas globalement concave.

4.2.3 Algorithme du score

Il s'agit d'un raffinement de l'algorithme de Berndt-Hall-Hall-Hausman où l'on prend l'espérance mathématique des produits croisés du score, égale à l'information de Fisher dans le cas du maximum de vraisemblance, à la place de leurs produits croisés. On approxime :

$$H(\theta) \quad \text{par} \quad \sum_{i=1}^{N} \mathop{\mathrm{E}}_{y} \left[\frac{\partial \ln f(y_i|x_i,\theta)}{\partial \theta} \frac{\partial \ln f(y_i|x_i,\theta)}{\partial \theta'} \right],$$

ce qui donne la règle d'itération :

$$\theta_{(p+1)} = \theta_{(p)} - \lambda \left[\sum_{i=1}^{N} \mathop{\mathrm{E}}_{y} \left[\frac{\partial \ln f\left(y_i|x_i,\theta_{(p)}\right)}{\partial \theta} \frac{\partial \ln f\left(y_i|x_i,\theta_{(p)}\right)}{\partial \theta'} \right] \right]^{-1} s\left(\theta_{(p)}\right).$$

Dans le cas où les dérivées secondes ne dépendent pas de la variable endogène y, cet algorithme est identique à celui de Newton-Raphson (e.g., cas des modèles Logit et de Poisson).

4.2.4 Algorithme de Levenberg-Marquardt

Il s'agit d'une extension de l'algorithme de Newton-Raphson que l'on applique quand l'objectif à maximiser n'est pas globalement concave. Supposons que le hessien au point $\theta_{(p)}$ n'est pas défini négatif, l'algorithme n'est plus nécessairement croissant et l'on n'est plus sûr de parvenir à un maximum local. On pourrait alors penser à utiliser l'algorithme du score ou celui de Berndt-Hall-Hall-Hausman. Mais la pratique montre que l'on aboutit souvent en fait à une valeur propre nulle du paramètre qui pose problème, car elle rend impossible l'inversion nécessaire à l'itération. On utilise donc une modification du hessien qui est définie négative, ce qui garantit que la matrice qui détermine le pas soit définie positive. Plus précisément, on utilise

$$W_p = -\left[H\left(\theta_{(p)}\right) - (1+\alpha)\,\mu_H \mathrm{I}_k\right]^{-1}$$

où I_k est la matrice identité de dimension k (la dimension de θ), μ_H la plus grande valeur propre du hessien (fonction disponible sous SAS-IML), supposée positive en cas de problème (i.e., de non concavité locale de la fonction objectif) et $\alpha > 0$ un paramètre choisi par l'utilisateur.

L'intuition de la méthode est la suivante : puisque le hessien n'est pas défini négatif, on le remplace par une matrice définie négative qui est la plus proche possible du hessien original. Pour cela il suffit de retrancher au hessien la matrice identité multipliée par la plus grande valeur propre du hessien (positive par hypothèse). Toutefois, ceci nous donnerait une valeur propre nulle. Il faut donc retrancher un peu plus que cette quantité; c'est ce que détermine le paramètre α. En général une petite valeur de α suffit et l'algorithme est sensible à de petites variations de ce paramètre (i.e, ce qui fonctionne avec $\alpha = 0.1$ peut ne plus fonctionner du tout avec $\alpha = 0.2$).

Pour déterminer la valeur de α, on utilise le constat suivant : si α est trop élevé, la plus grande valeur propre du hessien tend à croître avec le nombre d'itérations, si la valeur de α est trop petite, elle tend vers 0. Il suffit de se placer entre les deux après quelques essais, et l'on retombe généralement dans une zone de concavité de la fonction objectif. En général, plusieurs estimations sur un même échantillon ne nécessitent qu'une valeur de α, que l'on garde constante pendant tout l'algorithme.

Cet algorithme, qui sert au modèle tobit généralisé et aux moindres carrés non linéaires, deux exemples d'objectifs non concaves, mène à la règle d'itération suivante :

$$\theta_{(p+1)} = \begin{cases} \theta_{(p)} - \lambda \left[H\left(\theta_{(p)}\right) - (1+\alpha)\,\mu_H \mathrm{I}_k \right]^{-1} s\left(\theta_{(p)}\right) & \text{si} \quad \mu_H \geq 0 \\ \theta_{(p)} - \lambda H\left(\theta_{(p)}\right)^{-1} s\left(\theta_{(p)}\right) & \text{si} \quad \mu_H < 0 \end{cases} .$$

4.3 Méthodologie de programmation

Comment s'assurer que le programme que l'on a écrit ne comporte aucune erreur et qu'il soit pratique à utiliser? En prenant un certain nombre de précautions présentées dans cette section. Il y a quatre étapes qu'il faut prendre soin de bien effectuer.

1. *Vérifier la fonction objectif. A la fois par le calcul, mais également en consultant les ouvrages et les articles qui la donnent.*

2. *Vérifier le gradient. Par la même méthode que précédemment, mais également numériquement. Ainsi, on peut détecter des erreurs de recopie aussi bien sur la fonction objectif que sur son gradient. A cette étape, on utilise un algorithme basé sur les dérivées premières, de type Berndt-Hall-Hall-Hausman.*

3. *Vérifier le hessien. On utilise le gradient analytique pour calculer le hessien numérique, afin d'éviter le cumul des erreurs d'approximation. Ce problème est particulièrement sensible ici car les dérivées sont calculées à partir de quantités très petites par définition. On utilise*

un algorithme de Newton-Raphson ou de Levenberg-Marquardt. On peut conserver les dérivées secondes numériques si le calcul des dérivées analytiques est trop complexe, ou pour obtenir de premières estimations.

4. *Paramétrer le programme définitif. Ceci vise à éviter toute intervention sur le programme une fois qu'il a été vérifié, car c'est une source d'erreur potentielle. Par exemple, on peut paramétrer des programmes écrits en SAS-IML par des macro-variables, ce qui permet d'écrire des routines appelées macro-commandes. On peut alors utiliser ces routines sur toutes les bases de données et quel que soit le nombre de variables explicatives.*

CHAPITRE 5

Les variables dichotomiques

5.1 Cas général

On observe une variable dichotomique $y \in \{0,1\}$. Par définition cette variable suit une loi de Bernoulli. A partir d'un modèle théorique on peut écrire que le paramètre de cette loi est égal à p. La probabilité d'une loi de Bernoulli de paramètre p peut s'écrire :

$$\Pr[y=k] = p^k (1-p)^{1-k} = \begin{cases} p^0 (1-p)^{1-0} = 1-p & \text{si } k=0 \\ p^1 (1-p)^{1-1} = p & \text{si } k=1 \end{cases}$$

et son logarithme est égal à :

$$\ln \Pr[y=k] = k \ln p + (1-k) \ln (1-p).$$

On remarque également que l'espérance de y est égale à :

$$\mathrm{E}(y) = 0 \times (1-p) + 1 \times p = p,$$

et que sa variance est égale à :

$$\begin{aligned} \mathrm{V}(y) &= \mathrm{E}\left((y-p)^2\right) \\ &= (0-p)^2 \times (1-p) + (1-p)^2 \times p \\ &= p(1-p)(p+1-p) \\ &= p(1-p). \end{aligned}$$

On considère maintenant un échantillon de N variables aléatoires de Bernoulli $(y_1, ..., y_N)$, indépendantes, et de paramètres $(p_1, ..., p_N)$. Les paramètres des N lois sont différents parce que l'on considère un modèle conditionnel où chaque variable y_i possède une probabilité (i.e., une espérance conditionnelle) qui lui est propre. Ces probabilités dépendent donc de variables explicatives, regroupées dans le vecteur X, et dont les réalisations sont notées $(X_1, ..., X_N)$. Pour bien montrer le caractère conditionnel du modèle, on pose :

$$p_i = p(X_i, \beta), \ i = 1, ..., N$$

où β est le paramètre que l'on cherche à estimer. La log-vraisemblance d'un échantillon $(y_1, ..., y_N)$ est donc égale à :

$$\ell(y|X, \beta) = \sum_{i=1}^{N} y_i \ln p(X_i, \beta) + (1 - y_i) \ln(1 - p(X_i, \beta)).$$

Le score est donc égal à :

$$\frac{\partial \ell}{\partial \beta}(y|X, \beta) = \sum_{i=1}^{N} \frac{\partial p_i}{\partial \beta} \frac{y_i - p_i}{p_i(1 - p_i)},$$

et le hessien à :

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'}(y|X, \beta) = \sum_{i=1}^{N} \frac{\partial^2 p_i}{\partial \beta \partial \beta'} \frac{y_i - p_i}{p_i(1 - p_i)} - \frac{\partial p_i}{\partial \beta} \frac{\partial p_i}{\partial \beta'} \left[\frac{(y_i - p_i)(1 - 2p_i)}{p_i^2 (1 - p_i)^2} + \frac{1}{p_i(1 - p_i)} \right].$$

Pour appliquer l'algorithme du score, on remarque simplement que :

$$\mathrm{E}(y_i|X) = p_i,$$

ce qui permet d'obtenir :

$$\mathop{\mathrm{E}}_{y} \left[\frac{\partial^2 \ell}{\partial \beta \partial \beta'}(y|X, \beta) \right] = -\sum_{i=1}^{N} \frac{\partial p_i}{\partial \beta} \frac{\partial p_i}{\partial \beta'} \frac{1}{p_i(1 - p_i)}.$$

Cette représentation est valable pour tous les modèles dichotomiques; on a juste besoin de l'expression de la probabilité de réalisation d'un évènement en fonction des variables explicatives.

Pour obtenir les modèles usuels, on suppose que la variable qualitative que l'on observe résulte d'un modèle latent qui porte sur une variable

continue, notée y_i^*. Cette variable inobservable est supposée décrite par un modèle linéaire standard donné par :

$$y_i^* = X_i\beta + u_i,$$

où u_i est une perturbation d'espérance nulle, sans perte de généralité tant que le modèle latent contient un terme constant. Pour pouvoir estimer ce modèle par le maximum de vraisemblance, il nous faut écrire la loi de la variable observable conditionnellement aux variables explicatives. Cette variable observable est définie par :

$$y_i = \begin{cases} 1 & \text{si} \quad y_i^* > 0 \\ 0 & \text{si} \quad y_i^* \leq 0 \end{cases}$$

Le fait de prendre la valeur 0 comme seuil de référence n'a aucune incidence sur les estimations tant que le modèle comporte un terme constant, car on peut alors utiliser la variable latente $y_i^* - c$, où c est le seuil de troncature constant. La loi suivie par y_i est une loi de Bernoulli de paramètre $p_i = \Pr[\, y_i^* > 0]$, mais contrairement au cas habituel rencontré en statistique (i.e., modèle marginal), la probabilité est différente pour chaque observation puisqu'elle dépend des variables explicatives (i.e., modèle conditionnel). Le paramètre de la loi de Bernoulli est défini par :

$$p_i = \Pr[\, y_i^* > 0] = \Pr[X_i b + u_i > 0] = \Pr[u_i > -X_i\beta] = 1 - F(-X_i\beta),$$

où F est la fonction de répartition des u_i, $i = 1, ..., N$. La probabilité d'observer une réalisation y est donc donnée par :

$$\Pr[y_i = y] = p_i^y (1 - p_i)^{1-y}, \quad y \in \{0, 1\}.$$

De plus, on a $\mathrm{E}(y_i|X_i, \beta) = p_i$ et $\mathrm{V}(y_i|X_i, \beta) = p_i(1 - p_i)$. Comme les N observations sont supposées indépendantes, la vraisemblance de l'échantillon est donnée par le produit des probabilités individuelles :

$$L(y|X, \beta) = \prod_{i=1}^{N} p_i^{y_i} (1 - p_i)^{1-y_i},$$

d'où la log-vraisemblance :

$$\begin{aligned} \ell(y|X, \beta) &= \sum_{i=1}^{N} y_i \ln(p_i) + (1 - y_i) \ln(1 - p_i) \\ &= \sum_{i=1}^{N} y_i \ln[1 - F(-X_i\beta)] + (1 - y_i) \ln F(-X_i\beta). \end{aligned}$$

La forme spécifique prise par la probabilité p_i dépend directement de l'hypothèse faite sur la loi de la perturbation u. On peut essayer des

distributions différentes lors d'une étude, puis les comparer en effectuant des tests non emboîtés. Les deux lois les plus utilisées en pratique sont les lois normale et logistique. La première définit le modèle Probit, parfois appelé modèle Normit, tandis que la seconde définit le modèle Logit. Ces deux lois sont symétriques, on a donc $F(-z) = 1 - F(z)$, d'où :

$$p_i = F(X_i\beta).$$

Dans le cas du modèle Logit :

$$F(z) = \frac{1}{1+\exp(-z)}.$$

Dans le cas du modèle Probit (ou Normit) :

$$F(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds.$$

On remarque que dans tous les cas, la probabilité p_i est une fonction de la quantité réelle $m_i = X_i\beta$. Cette propriété résulte de la linéarité du modèle latent.

5.2 Le modèle Logit

Dans ce modèle on suppose que la perturbation u suit une loi logistique de fonction de répartition :

$$F(z) = \frac{1}{1+\exp(-z)}.$$

On vérifie directement que $F(-z) = 1 - F(z)$, la loi est symétrique. Sa densité est donnée par :

$$\begin{aligned} f(z) &= \partial F(z)/\partial z \\ &= \frac{\exp(-z)}{[1+\exp(-z)]^2} \\ &= \frac{1}{1+\exp(-z)} \times \frac{\exp(-z)}{1+\exp(-z)} \\ &= F(z)[1-F(z)]. \end{aligned}$$

Cette dernière propriété permet de simplifier toutes les dérivées de la log-vraisemblance. Les moments de cette loi logistique sont donnés par :

$$\mathrm{E}(u) = 0, \quad \mathrm{V}(u) = \frac{\pi^2}{3},$$

elle est donc centrée, mais n'est pas réduite. Il faudra en tenir compte lors de la comparaison des coefficients des modèles Logit et Probit car, dans ce dernier cas, la loi est à la fois centrée et réduite.

Les observations sont indépendantes donc la log-vraisemblance de l'échantillon $(y_1, ..., y_n)$ est donnée par :

$$\ell(y|X,\beta) = \sum_{i=1}^{n} y_i \ln F(m_i) + (1-y_i) \ln [1 - F(m_i)],$$

avec $m_i = X_i\beta$. Le vecteur du score est donné par :

$$\begin{aligned} s(y|X,\beta) &= \frac{\partial \ell}{\partial \beta}(y|X,\beta) \\ &= \sum_{i=1}^{n} X_i' \left[y_i \frac{f(m_i)}{F(m_i)} - (1-y_i) \frac{f(m_i)}{1-F(m_i)} \right] \\ &= \sum_{i=1}^{n} X_i' \left[y_i [1 - F(m_i)] - (1-y_i) F(m_i) \right] \\ &= \sum_{i=1}^{n} X_i' [y_i - F(m_i)]. \end{aligned}$$

On vérifie que son espérance est nulle puisque $\mathrm{E}(y_i|X_i,\beta) = F(m_i)$. Le hessien est donné par :

$$\begin{aligned} H(y|X,\beta) &= \frac{\partial^2 \ell}{\partial b \partial b'}(y|X,\beta) \\ &= \frac{\partial s}{\partial b'}(y|X,\beta) \\ &= -\sum_{i=1}^{n} X_i' X_i f(m_i). \end{aligned}$$

Comme f est une fonction strictement positive, le hessien est bien défini négatif. L'algorithme de Newton-Raphson sera donc croissant. On peut remarquer également que l'information de Fisher du modèle est donnée par :

$$I(y|X,\beta) = \mathrm{E}[-H(y|X,\beta)|X,\beta] = \sum_{i=1}^{n} X_i' X_i f(m_i).$$

On peut également calculer cette quantité en utilisant la variance du

score :

$$\begin{aligned}\mathrm{V}\left[s\left(y|X,\beta\right)|X,\beta\right] &= \sum_{i=1}^{n} X_i' X_i \mathrm{V}\left[y_i - F\left(m_i\right)|X,\beta\right] \\ &= \sum_{i=1}^{n} X_i' X_i F\left(m_i\right)\left(1 - F\left(m_i\right)\right) \\ &= \sum_{i=1}^{n} X_i' X_i f\left(m_i\right).\end{aligned}$$

Donc l'algorithme du score est identique à celui de Newton Raphson. En conséquence, nous utiliserons seulement deux algorithmes pour estimer ce modèle : Berndt-Hall-Hall-Hausman et Newton-Raphson dont les matrices sont données respectivement par :

$$\begin{aligned}W_{B3H}^{-1} &= -\sum_{i=1}^{n} X_i' X_i \left[y_i - F\left(m_i\right)\right]^2, \\ W_{NR}^{-1} &= -\sum_{i=1}^{n} X_i' X_i f\left(m_i\right).\end{aligned}$$

5.3 Le modèle Probit (ou Normit)

Maintenant on suppose que la perturbation u suit une loi normale de fonction de répartition donnée par :

$$F\left(z\right) = \Phi\left(z\right) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{s^2}{2}\right) ds.$$

Comme tous les programmes d'estimation possèdent la fonction de répartition de la loi normale centrée-réduite, nous garderons la notation

$\Phi(z)$. La densité est donnée par :[1]

$$\begin{aligned}\varphi(z) &= \frac{\mathrm{d}}{\mathrm{d}z}\left[\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{s^2}{2}\right)ds\right]\\ &= 1\times\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)-0\\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right).\end{aligned}$$

En dérivant cette densité, on voit directement que :

$$\varphi'(z) = -z\times\varphi(z).$$

Cette dernière relation permet de simplifier l'écriture du hessien et plus généralement, de trouver facilement les dérivées d'un ordre quelconque. Les moments de cette loi normale sont donnés par :

$$\mathrm{E}(u) = 0, \quad \mathrm{V}(u) = 1.$$

Les observations sont indépendantes donc la log-vraisemblance de l'échantillon $y = (y_1, ..., y_n)$ est donnée par :

$$\ell(y|X,\beta) = \sum_{i=1}^{N} y_i \ln\Phi(m_i) + (1-y_i)\ln\left[1-\Phi(m_i)\right].$$

Le vecteur du score est donné par :

$$\begin{aligned}s(y|X,\beta) &= \frac{\partial\ell}{\partial\beta}(y|X,\beta)\\ &= \sum_{i=1}^{N} X_i'\left[y_i\frac{\varphi(m_i)}{\Phi(m_i)} - (1-y_i)\frac{\varphi(m_i)}{1-\Phi(m_i)}\right]\\ &= \sum_{i=1}^{N} X_i'\frac{\varphi(m_i)(y_i-\Phi(m_i))}{\Phi(m_i)(1-\Phi(m_i))}.\end{aligned}$$

On vérifie que son espérance est nulle puisque $\mathrm{E}(y_i|X_i,\beta) = \Phi(m_i)$. Pour simplifier les notations on pose $\varphi(m_i) = \varphi_i$ et $\Phi(m_i) = \Phi_i$. En

[1] On applique le théorème de Leibniz :

$$\frac{\mathrm{d}}{\mathrm{d}z}\int_{a(z)}^{b(z)} f(t)\,\mathrm{d}t = b'(z)\,f(b(z)) - a'(z)\,f(a(z)).$$

utilisant $\varphi_i' = -m_i \times \varphi_i$, le hessien est donné par :

$$\begin{aligned} H(y|X,\beta) &= \frac{\partial^2 \ell}{\partial\beta\partial\beta'}(y|X,\beta) \\ &= \frac{\partial s}{\partial\beta'}(y|X,\beta) \\ &= -\sum_{i=1}^{N} X_i' X_i \left\{ \frac{\varphi_i (y_i - \Phi_i)}{\Phi_i (1-\Phi_i)} \left[m_i + \frac{\varphi_i (1-2\Phi_i)}{\Phi_i (1-\Phi_i)} \right] - \frac{\varphi_i^2}{\Phi_i (1-\Phi_i)} \right\}. \end{aligned}$$

On en déduit que l'information de Fisher du modèle est donnée par :

$$\begin{aligned} \mathrm{I}_1(\beta) &= \mathop{\mathrm{E}}_{X} \mathop{\mathrm{E}}_{y} \left[-H(y|X,\beta) \,|X,\beta \right] \\ &= \mathop{\mathrm{E}}_{X} \mathop{\mathrm{V}}_{y} \left[s(y|X,\beta) \,|X,\beta \right] \\ &= \frac{1}{N} \sum_{i=1}^{N} X_i' X_i \frac{\varphi_i^2}{\Phi_i (1-\Phi_i)}. \end{aligned}$$

On peut également calculer cette quantité en utilisant la variance du score, ce qui est plus pratique ici que de calculer la dérivée seconde. Clairement, l'algorithme du score est différent de l'algorithme de Newton-Raphson. Il nécessite moins de calcul et est également croissant.

On peut donc utiliser trois algorithmes pour estimer ce modèle : Berndt-Hall-Hall-Hausman, le Score et Newton-Raphson dont les matrices sont données respectivement par :

$$\begin{aligned} W_{B3H}^{-1} &= -\sum_{i=1}^{N} X_i' X_i \frac{\varphi_i^2}{\Phi_i^2 (1-\Phi_i)^2} \left[y_i - \Phi_i \right]^2, \\ W_{SC}^{-1} &= -\sum_{i=1}^{N} X_i' X_i \frac{\varphi_i^2}{\Phi_i (1-\Phi_i)}, \\ W_{NR}^{-1} &= \sum_{i=1}^{N} X_i' X_i \left\{ \frac{\varphi_i (y_i - \Phi_i)}{\Phi_i (1-\Phi_i)} \left[m_i + \frac{\varphi_i (1-2\Phi_i)}{\Phi_i (1-\Phi_i)} \right] - \frac{\varphi_i^2}{\Phi_i (1-\Phi_i)} \right\}. \end{aligned}$$

5.4 Interprétation et comparaison des coefficients

5.4.1 Le modèle Probit

En fait, les coefficients du modèle latent ne sont estimés qu'à une constante multiplicative et positive près : l'inverse de l'écart-type de la perturbation du modèle latent. Dans le cas du modèle Probit, le modèle

latent s'écrit :

$$z_i^* = X_i b + v_i, \quad v_i \overset{iid}{\rightsquigarrow} \mathrm{N}\left(0, \sigma^2\right).$$

Le modèle que nous avons estimé s'écrit donc :

$$y_i^* = X_i \beta + u_i, \quad u_i \overset{iid}{\rightsquigarrow} \mathrm{N}\left(0, 1\right)$$

avec $y_i^* = z_i^*/\sigma$, $\beta = b/\sigma$ et $u_i = v_i/\sigma$. Les paramètres b et σ ne sont pas identifiables, seule la fonction $\beta = b/\sigma$ de ces deux paramètres peut être estimée, ou toute fonction monotone de β. En conséquence, tous les coefficients d'un modèle Probit sont implicitement réduits par l'écart-type de la perturbation de la régression, ce qui a un certain nombre de conséquences sur leur interprétation :

1. On ne peut pas comparer les coefficients obtenus sur les régressions de deux variables dichotomiques endogènes différentes, car l'écart-type de la perturbation change avec le modèle latent.

2. Le signe du coefficient β est le même que celui de b car un écart-type est toujours positif.

3. Le ratio de deux coefficients extraits de β est identique au ratio des deux coefficients correspondants de b. On peut donc dire qu'un coefficient est deux fois plus grand qu'un autre.

4. La différence entre deux coefficients extraits de β n'est connue qu'à un facteur multiplicatif positif près, égal à σ^{-1}. On ne peut donc interpréter que le signe de la différence entre deux coefficients, pas la grandeur de l'écart. Par contre, on peut comparer deux écarts tirés de la même régression.

5.4.2 Le modèle Logit

La même interprétation reste valable, à ceci près qu'une loi logistique de paramètres $(0, \phi)$ admet pour espérance 0 et pour variance $\phi^2\pi^2/3$. On peut donc écrire le modèle latent :[2]

$$z_i^* = X_i b + v_i, \quad v_i \overset{iid}{\rightsquigarrow} \Lambda\left(0, \phi\right)$$

[2] La fonction de répartition variable Z suivant une loi logistique de paramètres $\Lambda(\mu, \phi)$ est donnée par :

$$F(z) = \left[1 + \exp\left(-\frac{z-\mu}{\phi}\right)\right]^{-1}.$$

On a :

$$\mathrm{E}(Z) = \mu \text{ et } \mathrm{V}(Z) = \frac{\phi^2\pi^2}{3}.$$

Le modèle que nous avons estimé s'écrit donc :

$$y_i^* = X_i\beta + u_i, \quad u_i \overset{iid}{\rightsquigarrow} \Lambda(0,1)$$

avec $y_i^* = z_i^*/\phi$, $\beta = b/\phi$ et $u_i = v_i/\phi$. L'interprétation des coefficients du modèle Logit se fait donc de la même façon que celle du modèle Probit.

5.4.3 Comparaison des coefficients des modèles Logit et Probit

Dans le modèle Probit, on estime :

$$\beta_{\text{PROBIT}} = \frac{b}{\sqrt{\text{V}(u)}},$$

alors que dans le modèle Logit on estime :

$$\beta_{\text{LOGIT}} = \frac{b}{\phi} \text{ avec } \phi = \frac{\sqrt{3}}{\pi}\sqrt{\text{V}(u)},$$

on en déduit que :

$$\beta_{\text{PROBIT}} \simeq \frac{\sqrt{3}}{\pi}\beta_{\text{LOGIT}}.$$

Dans la pratique on utilisera donc les approximations suivantes :

$$\beta_{\text{PROBIT}} \simeq 0,55 \times \beta_{\text{LOGIT}} \text{ et } \beta_{\text{LOGIT}} \simeq 1,81 \times \beta_{\text{PROBIT}}.$$

On peut imprimer systématiquement les coefficients réduits en fin de programme pour faciliter les comparaisons. La même modification doit être effectuée sur les écarts-types asymptotiques des estimateurs. Il n'en demeure pas moins que les deux modèles sont différents et que la comparaison n'est qu'approximative car la loi logistique n'est qu'une approximation de la loi normale et admet notamment plus de valeurs extrêmes que cette dernière.[3]

5.5 Les aides à l'interprétation

Les coefficients des modèles Logit et Probit ne sont définis qu'à une constante multiplicative près, de sorte qu'ils ne sont pas directement interprétables. La méthode la plus simple pour obtenir des coefficients directement interprétables consiste à calculer l'impact d'une variables explicative directement sur la probabilité. Le cas le plus simple est celui

[3] Le coefficient d'applatissement de la loi logistique est de 1, 2 au lieu de 1 pour la loi normale.

où la variable explicative est binaire; il suffit de comparer les deux états $\{0,1\}$. Dans le cas d'une variables explicative quantitative il faut prendre deux points de référence; par exemple en comparant l'effet du passage du premier au troisième quartile. Deux types de mesures sont utilisées dans la littérature : d'une part, l'effet direct de la variable explicative sur la probabilité ou effet incrémental; d'autre part, l'effet d'une variable explicative sur le ratios des probabilités de réalisation de l'évènement (ou "odds ratio").

5.5.1 Variables explicatives binaires

Sans perte de généralité, considérons le cas d'une seule variable explicative binaire, $X \in \{0,1\}$. La fonction suivante donne les chances que l'évènement $Y = 1$ se réalise par rapport à l'évènement $Y = 0$:

$$R(X) = \frac{\Pr[Y = 1|X]}{\Pr[Y = 0|X]} = \frac{F(\beta_0 + \beta_1 X)}{1 - F(\beta_0 + \beta_1 X)},$$

il s'agit du rapport des probabilités pour une même valeur de X ("the odds function"). Pour voir l'effet de X sur ce ratio, on utilise le "ratio des cotes" ("odds ratio") :

$$\psi_X = \frac{R(1)}{R(0)},$$

soit :

$$\psi_X = \frac{\dfrac{F(\beta_0 + \beta_1)}{1 - F(\beta_0 + \beta_1)}}{\dfrac{F(\beta_0)}{1 - F(\beta_0)}},$$

ce ratio indique la modification des chances d'obtenir l'évènement $Y = 1$ lorsque l'on passe du sous échantillon $X = 0$ au sous échantillon $X = 1$. Le coefficient $\beta = (\beta_0, \beta_1)'$ est généralement estimé avec de nombreuses autres variables de sorte qu'il s'agit d'un effet "toutes choses égales par ailleurs".

Si le modèle comporte plusieurs variables explicatives, on estime généralement un effet au point moyen. Dans ce cas le modèle avec les autres variables explicatives s'écrit :

$$\mathrm{E}(Y) = F(\gamma_0 + \beta_1 X + Z\beta_2),$$

de sorte que le ratio des cotes peut s'écrire :

$$\psi_X = \frac{\dfrac{F\left(\gamma_0 + \beta_1 + \overline{Z}\beta_2\right)}{1 - F\left(\gamma_0 + \beta_1 + \overline{Z}\beta_2\right)}}{\dfrac{F\left(\gamma_0 + \overline{Z}\beta_2\right)}{1 - F\left(\gamma_0 + \overline{Z}\beta_2\right)}},$$

en posant :

$$\beta_0 = \gamma_0 + \overline{Z}\beta_2,$$

on retrouve la même formule que précédemment.

Le cas du modèle Probit s'obtient avec $F = \Phi$. Pour le modèle Logit l'expression se simplifie nettement :

$$\Pr[Y = 1|X] = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)},$$

et

$$\Pr[Y = 0|X] = \frac{1}{1 + \exp(\beta_0 + \beta_1 X)},$$

de sorte que le rapport des probabiltiés est égal à :

$$R(X) = \exp(\beta_0 + \beta_1 X),$$

et que le rapport des cotes est égal à :

$$\psi = \frac{R(1)}{R(0)} = \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1),$$

il suffit de prendre l'exponentielle de ce coefficient pour obtenir l'effet de la variable X sur le rapport des cotes. Ceci permet également de calculer un intervalle de confiance facilement en utilisant le théorème de Slutsky. Il ne s'agit toutefois pas de l'effet de la variable X sur $\Pr[Y = 1|X]$ mais sur le ratio $\Pr[Y = 1|X] / \Pr[Y = 0|X]$. Pour obtenir l'effet incrémental de X sur la probabilité que l'évènement $Y = 1$ se réalise, on doit calculer :

$$\Delta_X = \frac{\Pr[Y = 1|X = 1]}{\Pr[Y = 1|X = 0]} = \frac{F(\beta_0 + \beta_1)}{F(\beta_0)},$$

dans le cas du modèle Logit, on obtient donc :

$$\begin{aligned}
\Delta_X &= \frac{\dfrac{\exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)}}{\dfrac{\exp(\beta_0)}{1 + \exp(\beta_0)}} \\
&= \exp(\beta_1) \times \frac{1 + \exp(\beta_0)}{1 + \exp(\beta_0 + \beta_1)} \\
&= \frac{\exp(\beta_1) + \exp(\beta_0 + \beta_1)}{1 + \exp(\beta_0 + \beta_1)},
\end{aligned}$$

on voit que si $\beta_1 > 0$, alors $\exp(\beta_1) > 1$ et $\Delta_X > 1$.

5.5.2 Variables explicatives quantitatives

Avec des variables quantitatives, il est possible de calculer deux types de quantités : l'effet marginal qui donne la variation de la probabilité que $Y = 1$ and X augmente d'une unité; ou l'effet incrémental, qui mesure la variation de la probabilité que $Y = 1$ quand X passe d'une valeur de référence à une autre. Comme valeurs de référence, on peut prendre les quartiles ou des déciles extrêmes de X. La variation interdécile donne une bonne idée du potentiel d'influence de la variable X, puisqu'elle représente la variation de la probabilité quand on passe des 10% plus petites valeurs de X à ses 10% les plus élevées. Dans le cas d'une loi normale ceci serait équivalent à calculer la variation de la probabilité quand la variable X se situe entre $\overline{X} \pm 1.645 \times \sigma_X$.

L'effet marginal est donné par :

$$\delta_X = \frac{\partial}{\partial X} \Pr\left[Y = 1|X\right] = \beta_1 \times f\left(\beta_0 + \beta_1 X\right),$$

et il varie avec chaque valeur de X. On peut soit faire un graphique sur l'ensemble des valeurs de X, soit prendre un point de référence comme la moyenne ou la médiane. L'effet incrémental du passage de X de a à b est analogue au traitement sur variables qualitatives. On a :

$$\Delta_X = \frac{\Pr\left[Y = 1|X = b\right]}{\Pr\left[Y = 1|X = a\right]} = \frac{F\left(\beta_0 + \beta_1 b\right)}{F\left(\beta_0 + \beta_1 a\right)},$$

dans le cas du modèle Logit :

$$\Delta_X = \frac{\dfrac{\exp\left(\beta_0 + \beta_1 b\right)}{1 + \exp\left(\beta_0 + \beta_1 b\right)}}{\dfrac{\exp\left(\beta_0 + \beta_1 a\right)}{1 + \exp\left(\beta_0 + \beta_1 a\right)}} = \exp\left[\beta_1\left(b - a\right)\right] \times \frac{1 + \exp\left(\beta_0 + \beta_1 a\right)}{1 + \exp\left(\beta_0 + \beta_1 b\right)},$$

dans ce cas, on remarque que le ratio des cotes se simplifie de manière importante avec le modèle Logit :

$$\psi_X = \frac{R\left(b\right)}{R\left(a\right)} = \frac{\exp\left(\beta_0 + \beta_1 b\right)}{\exp\left(\beta_0 + \beta_1 a\right)} = \exp\left[\beta_1\left(b - a\right)\right] = \exp\left(\beta_1\right)^{b-a},$$

cette formule simplifie également le calcul des intervalles de confiance. On peut notamment utiliser les rapports des cotes pour calculer l'effet incrémental d'une variable quantitative et pas seulement d'une variable qualitative.

5.6 Application : la participation des femmes au marché du travail

Cette section présente une version simplifiée d'une équation de participation au marché du travail. La variable que l'on cherche à expliquer est dichotomique : une personne a un emploi ou non au moment de l'enquête. Les données sont issues de l'enquête "Jeunes et Carrières" réalisée par l'INSEE en 1997. L'échantillon comprend des données sur $N = 5425$ couples. On considère la participation des femmes au marché du travail que l'on explique par les déterminants suivants :

1. Age;
2. Nombre d'enfants;
3. Naissance l'année courante;
4. Activité des parents;
5. Nationalité de l'intéressée et de ses parents;
6. Niveau d'éducation (1 = sans diplôme ou première année de CAP; 2 = sans diplôme ou dernière année de CAP; 3 = CAP ou BEP; 4 = Baccalauréat professionnel; 5 = Baccalauréat général ou équivalent; 6 = BTS; 7 = Enseignement supérieur général).
7. Région d'habitation;
8. Les mêmes variables pour le conjoint;

   L'estimation est réalisée sous SAS à partir de la procédure *logistic*. La syntaxe de base, si l'on veut expliquer une variable dichotomique y par les variables explicatives x_1, x_2 et x_3, est la suivante :

SYNTAXE 5.1

```
proc logistic data=tab descending;
model y=x1 x2 x3;
run;
```

Sous cette forme la procédure va chercher les données dans le tableau *tab*. L'option *descending* est très importante car, par défaut, la procédure estime un modèle dichotomique avec une probabilité $p(X_i, \beta) = \Pr[y_i = 0]$ au lieu de $\Pr[y_i = 1]$. Dans le cas du modèle Logit, on a $\Pr[y_i = 0] = F(-X_i\beta)$ de sorte que sans l'option *descending* on estime $-\beta$ au lieu de β. L'option descending permet donc d'imposer $p(X_i, \beta) = \Pr[y_i = 1]$. L'instruction *model* sert à indiquer la variable (dichotomique) expliquée y et la liste des variables explicatives X_1, X_2 et X_3. Par défaut, la procédure *logistic* permet d'estimer le modèle Logit, mais d'autres

modèles sont disponibles. Par exemple, pour estimer un modèle Probit, on utilise la syntaxe suivante :

SYNTAXE 5.2

```
proc logistic data=tab descending;
model y=x1 x2 x3/link=normit;
run;
```

L'option *link=* permet d'indiquer la distribution que l'on souhaite. Le modèle Normit correspond à la fonction de répartition de la loi normale. Pour obtenir le modèle de Weibull, on remplace l'option *link=normit* par *link=cloglog*.

Reprenons notre application. Afin d'estimer le modèle, on entre les commandes :

Programme 5.1

```
proc logistic descending data=tab;
model f_jc97=
f_age f_enf1 f_enf2 f_enf3 f_enf4 f_nai9697
f_mcspmiss f_magricul f_martisan f_mcadre
f_mprofint /*f_employe*/ f_mouvrier f_minactiv
f_pcspmiss f_pagricul f_partisan f_pcadre
f_pprofint f_pemploye /*f_pouvrier*/
/*f_francais*/ f_afrnord f_europ f_autrenat
/*f_pfrancai*/ f_pafrnord f_peurop f_pautrnat
/*f_ndip1*/ f_ndip2 f_ndip3 f_ndip4 f_ndip5 f_ndip6
f_ndip7
h_age
h_mcspmiss h_magricul h_martisan h_mcadre h_mprofint
/*h_employe*/ h_mouvrier h_minactiv
h_pcspmiss h_pagricul h_partisan h_pcadre h_pprofint
h_pemploye /*h_pouvrier*/
/*h_francais*/ h_afrnord h_europ h_autrenat
/*h_pfrancai*/ h_pafrnord h_peurop h_pautrnat
/*h_ndip1*/ h_ndip2 h_ndip3 h_ndip4 h_ndip5 h_ndip6
h_ndip7
/*h_ILEFR97*/ h_CHAMPA97 h_PICARD97 h_HAUTNO97
h_CENTRE97 h_BASSNO97 h_BOURGO97 h_NORDP97
h_LORRAI97 h_ALSA97 h_FRCOM97 h_PAYSLO97
h_BRETA97 h_POITOU97 h_AQUITA97 h_MIDIPY97
h_LIMOUS97 h_RHONEA97 h_AUVER97 h_LANGUE97
h_PROVEN97
```

```
/link=normit;
run;
```

les variables entre commentaires (début /* et fin */) indiquent la modalité de référence. On prend généralement la modalité la plus répandue. Le programme précédent produit la sortie :

Sortie 5.1

The LOGISTIC Procedure

Model Information

Data Set	WORK.TAB
Response Variable	f_jc97
Number of Response Levels	2
Number of Observations	5425
Model	binary probit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	f_jc97	Total Frequency
1	1	3701
2	0	1724

Probability modeled is f_jc97=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates

AIC	6785.321	5798.470
SC	6791.920	6313.174
-2 Log L	6783.321	5642.470

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1140.8513	77	<.0001
Score	1074.0382	77	<.0001
Wald	925.8740	77	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.9183	0.1678	29.9387	<.0001
f_age	1	0.0522	0.00587	79.0265	<.0001
f_enf1	1	-0.2213	0.0665	11.0658	0.0009
f_enf2	1	-0.5780	0.0670	74.5248	<.0001
f_enf3	1	-0.9756	0.0777	157.8460	<.0001
f_enf4	1	-1.6081	0.1101	213.2488	<.0001
f_nai9697	1	-0.2890	0.0622	21.6082	<.0001
F_MCSPMISS	1	-0.2903	0.1115	6.7828	0.0092
F_MAGRICUL	1	0.1182	0.1154	1.0491	0.3057
F_MARTISAN	1	-0.1044	0.0941	1.2298	0.2674
F_MCADRE	1	-0.2240	0.1888	1.4089	0.2352
F_MPROFINT	1	-0.00686	0.0893	0.0059	0.9388
F_MOUVRIER	1	-0.0702	0.0596	1.3843	0.2394
F_MINACTIV	1	-0.1264	0.0521	5.8969	0.0152
F_PCSPMISS	1	-0.1194	0.1081	1.2211	0.2691
F_PAGRICUL	1	0.0425	0.0970	0.1916	0.6616
F_PARTISAN	1	0.00175	0.0727	0.0006	0.9808
F_PCADRE	1	-0.2211	0.0877	6.3645	0.0116
F_PPROFINT	1	-0.0366	0.0686	0.2848	0.5936
F_PEMPLOYE	1	-0.0746	0.0609	1.5025	0.2203
F_AFRNORD	1	-0.7277	0.2178	11.1613	0.0008
F_EUROP	1	-0.0673	0.1735	0.1506	0.6980
F_AUTRENAT	1	-0.7775	0.1747	19.7985	<.0001
F_PAFRNORD	1	-0.1896	0.1184	2.5623	0.1094

F_PEUROP	1	0.00814	0.0972	0.0070	0.9333
F_PAUTRNAT	1	-0.00406	0.0705	0.0033	0.9541
F_NDIP2	1	0.2343	0.0668	12.3118	0.0005
F_NDIP3	1	0.3749	0.0586	40.9611	<.0001
F_NDIP4	1	0.6551	0.0792	68.3899	<.0001
F_NDIP5	1	0.5770	0.0867	44.3076	<.0001
F_NDIP6	1	0.9994	0.0873	131.1030	<.0001
F_NDIP7	1	0.9218	0.0955	93.1190	<.0001
h_age	1	0.00383	0.00587	0.4265	0.5137
H_MCSPMISS	1	-0.1559	0.1092	2.0401	0.1532
H_MAGRICUL	1	-0.1517	0.1079	1.9778	0.1596
H_MARTISAN	1	-0.0929	0.0938	0.9817	0.3218
H_MCADRE	1	0.00537	0.1960	0.0008	0.9781
H_MPROFINT	1	0.0275	0.0887	0.0960	0.7566
H_MOUVRIER	1	0.00255	0.0623	0.0017	0.9674
H_MINACTIV	1	-0.0664	0.0526	1.5936	0.2068
H_PCSPMISS	1	0.0136	0.1063	0.0165	0.8979
H_PAGRICUL	1	0.2244	0.0931	5.8138	0.0159
H_PARTISAN	1	0.1023	0.0718	2.0322	0.1540
H_PCADRE	1	0.0246	0.0899	0.0750	0.7842
H_PPROFINT	1	0.0924	0.0704	1.7231	0.1893
H_PEMPLOYE	1	0.0339	0.0624	0.2957	0.5866
H_AFRNORD	1	-0.0567	0.2017	0.0791	0.7786
H_EUROP	1	0.1085	0.1661	0.4269	0.5135
H_AUTRENAT	1	0.0837	0.1951	0.1839	0.6681
H_PAFRNORD	1	-0.00706	0.1174	0.0036	0.9521
H_PEUROP	1	0.1157	0.1025	1.2733	0.2592
H_PAUTRNAT	1	-0.1482	0.0681	4.7431	0.0294
H_NDIP2	1	0.0360	0.0721	0.2496	0.6174
H_NDIP3	1	0.0432	0.0623	0.4807	0.4881
H_NDIP4	1	0.1063	0.0880	1.4589	0.2271
H_NDIP5	1	-0.1846	0.1099	2.8211	0.0930
H_NDIP6	1	-0.0549	0.0961	0.3268	0.5676
H_NDIP7	1	-0.1936	0.0983	3.8824	0.0488
H_CHAMPA97	1	-0.2672	0.1153	5.3710	0.0205
H_PICARD97	1	-0.1980	0.1167	2.8774	0.0898
H_HAUTNO97	1	-0.1628	0.1068	2.3226	0.1275
H_CENTRE97	1	-0.0725	0.1096	0.4378	0.5082
H_BASSNO97	1	-0.2480	0.1324	3.5110	0.0610
H_BOURGO97	1	-0.2045	0.1145	3.1930	0.0740
H_NORDP97	1	-0.6270	0.0927	45.7330	<.0001
H_LORRAI97	1	-0.2985	0.1197	6.2144	0.0127
H_ALSA97	1	-0.3131	0.1078	8.4422	0.0037
H_FRCOM97	1	-0.0378	0.1199	0.0991	0.7529

H_PAYSLO97	1	-0.1121	0.0993	1.2733	0.2592
H_BRETA97	1	-0.3328	0.1011	10.8330	0.0010
H_POITOU97	1	-0.0622	0.1131	0.3026	0.5822
H_AQUITA97	1	-0.4379	0.1071	16.7092	<.0001
H_MIDIPY97	1	-0.2844	0.1331	4.5682	0.0326
H_LIMOUS97	1	-0.3096	0.1219	6.4570	0.0111
H_RHONEA97	1	-0.2625	0.0872	9.0686	0.0026
H_AUVER97	1	-0.1557	0.1170	1.7717	0.1832
H_LANGUE97	1	-0.6318	0.1229	26.4220	<.0001
H_PROVEN97	1	-0.5341	0.0963	30.7795	<.0001

Association of Predicted Probabilities and Observed Responses

Percent Concordant	76.2	Somers' D	0.527
Percent Discordant	23.5	Gamma	0.528
Percent Tied	0.2	Tau-a	0.228
Pairs	6380524	c	0.763

Les principaux déterminants de la participation des femmes au marché du travail sont le nombre d'enfants (effet négatif) et le niveau d'études (effet positif).

CHAPITRE 6

Les variables polytomiques

6.1 Cas général

Les variables polytomiques correspondent au cas où l'on observe plusieurs modalités en général, qu'elles soient ordonnées ou non. On suppose qu'une variable y_i peut prendre J modalités $y_i \in \{1, 2, ..., J\}$. La probabilité que la variable y_i soit égale à la modalité j est notée :

$$p_{ji} = \Pr[y_i = j],$$

où les probabilités vérifient, pour chaque individu :

$$\sum_{j=1}^{J} p_{ji} = 1, \quad \forall i.$$

On définit également les J variables indicatrices suivantes pour chaque individu :

$$d_{ji} = \begin{cases} 1 & \text{si} \quad y_i = j \\ 0 & \text{sinon} \end{cases} \qquad j = 1, ..., J.$$

Ces variables vérifient :

$$\sum_{j=1}^{J} d_{ji} = 1, \quad \forall i.$$

La log-vraisemblance d'un échantillon $(y_1, ..., y_N)$ s'écrit donc simplement :

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{J} d_{ji} \ln p_{ji}.$$

Le score est donc égal à :

$$\frac{\partial \ell}{\partial \theta} = \sum_{i=1}^{N}\sum_{j=1}^{J} d_{ji}\frac{\partial p_{ji}}{\partial \theta}\frac{1}{p_{ij}},$$

il est d'espérance nulle puisque $\mathrm{E}\left(d_{ji}\right) = p_{ji}$, ce qui implique :

$$\forall i \quad \mathrm{E}\left[\sum_{j=1}^{J} d_{ji}\frac{\partial p_{ji}}{\partial \theta}\frac{1}{p_{ij}}\right] = \sum_{j=1}^{J}\frac{\partial p_{ji}}{\partial \theta} = \frac{\partial}{\partial \theta}\underbrace{\sum_{j=1}^{J} p_{ji}}_{1} = 0.$$

On peut appliquer l'algorithme de Berndt-Hall-Hall-Hausman à partir de la matrice :

$$W_{B3H}^{-1} = -\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{d_{ji}}{p_{ji}^2}\frac{\partial p_{ji}}{\partial \theta}\frac{\partial p_{ji}}{\partial \theta'},$$

Le hessien est donné par :

$$\frac{\partial^2 \ell}{\partial \theta \partial \theta'} = \sum_{i=1}^{N}\sum_{j=1}^{J}\frac{d_{ji}}{p_{ji}}\left[\frac{\partial^2 p_{ji}}{\partial \theta \partial \theta'} - \frac{1}{p_{ji}}\frac{\partial p_{ji}}{\partial \theta}\frac{\partial p_{ji}}{\partial \theta'}\right],$$

ce qui permet d'employer l'algorithme de Newton-Raphson :

$$W_{NR}^{-1} = \sum_{i=1}^{N}\sum_{j=1}^{J}\frac{d_{ji}}{p_{ji}^2}\left[p_{ji}\frac{\partial^2 p_{ji}}{\partial \theta \partial \theta'} - \frac{\partial p_{ji}}{\partial \theta}\frac{\partial p_{ji}}{\partial \theta'}\right],$$

l'algorithme du score peut donc s'obtenir à partir de la matrice suivante :

$$\begin{aligned}\mathop{\mathrm{E}}_{d}\left[\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\right] &= \sum_{i=1}^{N}\sum_{j=1}^{J}\left\{\frac{\partial^2 p_{ji}}{\partial \theta \partial \theta'} - \frac{1}{p_{ji}}\frac{\partial p_{ji}}{\partial \theta}\frac{\partial p_{ji}}{\partial \theta'}\right\} \\ &= -\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{1}{p_{ji}}\frac{\partial p_{ji}}{\partial \theta}\frac{\partial p_{ji}}{\partial \theta'},\end{aligned}$$

car

$$\forall i \quad \sum_{j=1}^{J}\frac{\partial p_{ji}}{\partial \theta} = 0 \Rightarrow \sum_{j=1}^{J}\frac{\partial^2 p_{ji}}{\partial \theta \partial \theta'} = 0.$$

En conséquence :

$$W_{SC}^{-1} = -\sum_{i=1}^{N}\sum_{j=1}^{J}\frac{1}{p_{ji}}\frac{\partial p_{ji}}{\partial \theta}\frac{\partial p_{ji}}{\partial \theta'}.$$

Les dérivées premières des probabilités suffisent donc pour obtenir un algorithme croissant.

6.2 Les variables ordonnées

6.2.1 Cas général

La variable expliquée que l'on observe est qualitative par classe. On suppose qu'il y a J classes numérotées de $j = 1$ à $j = J$. Les données observables sont définies par :

$$y_i = \begin{cases} 1 & \text{si } a_0 < y_i^* \leq a_1 \\ \vdots & \\ j & \text{si } a_{j-1} < y_i^* \leq a_j \\ \vdots & \\ J & \text{si } a_{J-1} < y_i^* \leq a_J \end{cases}$$

Remarque 6.1 *Dans le cas d'une variable latente réelle, on peut poser la convention* $a_0 = \{-\infty\}$ *et* $a_J = \{+\infty\}$.

Remarque 6.2 *Le modèle dichotomique comprend* $J = 2$ *classes et s'obtient comme le cas particulier où* $a_0 = \{-\infty\}$, $a_1 = 0$ *et* $a_2 = \{+\infty\}$.

Les variables polytomiques suivent, par définition, une distribution multinomiale dont les paramètres sont donnés par :

$$\{p_{ji}\}_{j=1,\ldots,J} = \{\Pr[a_{j-1} < y_i^* \leq a_j]\}_{j=1,\ldots,J}.$$

La probabilité associée à cette distribution est simplement :

$$\prod_{j=1}^{J} p_{ji}^{d_{ji}}, \quad \text{où } d_{ji} = \begin{cases} 1 & \text{si } a_{j-1} < y_i^* \leq a_j \\ 0 & \text{sinon} \end{cases}$$

Ces probabilités dépendent de la distribution suivie par la variable latente y_i^*. En supposant que cette perturbation suit une loi de fonction de répartition F qui dépend d'un paramètre θ, on obtient les probabilités suivantes :

$$\begin{aligned} p_{ji} &= \Pr[a_{j-1} < y_i^* \leq a_j] \\ &= \Pr[y_i^* \leq a_j] - \Pr[y_i^* \leq a_{j-1}] \\ &= F(a_j) - F(a_{j-1}). \end{aligned}$$

Remarque 6.3 *Dans le cas particulier où* y^* *est réelle, on a* $F(a_0) = F(-\infty) = 0$ *et* $F(a_J) = F(+\infty) = 1$. *Les probabilités s'écrivent donc* $p_{1i} = F(a_1)$, $p_{ji} = F(a_j) - F(a_{j-1})$ *si* $2 \leq j \leq J-1$ *et* $p_{Ji} = 1 - F(a_{J-1})$.

6.2.2 Le modèle Probit ordonné

Dans ce modèle on suppose que la variable latente y_i^* est générée par un modèle linéaire standard :

$$y_i^* = X_i b + u_i, \quad u_i \overset{\text{iid}}{\rightsquigarrow} \mathrm{N}\left(0, \sigma^2\right).$$

La variable y_i appartient à la classe j dès lors que :

$$\begin{aligned} a_{j-1} < y_i^* \leq a_j \Leftrightarrow a_{j-1} < X_i b + u_i \leq a_j \\ \Leftrightarrow \frac{a_{j-1} - X_i b}{\sigma} < \frac{u_i}{\sigma} \leq \frac{a_j - X_i b}{\sigma} \end{aligned}$$

où u_i/σ suit une loi normale centrée et réduite de fonction de répartition $\Phi(z)$. On en déduit que :

$$\begin{aligned} p_{ji} &= \Pr(y_{ji} = j) \\ &= \Phi\left(\frac{a_{j-1} - X_i b}{\sigma} < \frac{u_i}{\sigma} \leq \frac{a_j - X_i b}{\sigma}\right) \\ &= \Phi\left(\frac{a_j - X_i b}{\sigma}\right) - \Phi\left(\frac{a_{j-1} - X_i b}{\sigma}\right). \end{aligned}$$

Pour simplifier la présentation, on effectue le changement de paramètre suivant :

$$\beta = \frac{b}{\sigma} \text{ et } h = \frac{1}{\sigma},$$

ce qui donne :

$$p_{ji} = \Phi(a_j h - X_i \beta) - \Phi(a_{j-1} h - X_i \beta).$$

Arrivé à ce stade, il faut distinguer le cas où les seuils a_j sont connus ou inconnus. Si les seuils sont connus, les deux paramètres h et β sont identifiables et l'on peut procéder à l'estimation par le maximum de vraisemblance.

Si les seuils sont inconnus, on ne peut pas estimer h et des termes en $a_j h$ apparaissent dans l'expression des probabilités. Ici, on peut remarquer que les quantités inconnues $a_j h$ dépendent uniquement de la classe j. Ceci revient à dire que chaque classe admet un terme constant différent et croissant avec l'ordre de la classe puisque $a_j > a_{j-1} \Rightarrow a_j h > a_{j-1} h$.

Remarque 6.4 *Dans les modèles avec terme constant (le cas usuel) la constante de la première classe est prise comme référence, de sorte que a un terme constant global et les différentes estimations de $a_j h$ pour $j = 2, ..., J-1$ (car $a_J = \{+\infty\}$).*

6.2.2.1 Estimation avec seuils connus

Pour estimer le modèle, on a juste besoin des dérivées des probabilités p_{ij} par rapport aux paramètres $\theta' = (\beta', h)$. Notons que l'on ne peut estimer β et h séparément que lorsque les seuls sont connus. On note $\phi(u)$ la densité de la loi Normale centrée et réduite, on utilise la propriété $\phi'(u) = -u\phi(u)$ et l'on pose la notation :

$$v_{ji} = a_j h - X_i \beta.$$

La forme générale de la probabilité est donc :

$$p_{ji} = \Phi(v_{ji}) - \Phi(v_{j-1i}),$$

et ses dérivées :

$$\frac{\partial p_{ji}}{\partial \beta} = X_i' \left[\phi(v_{j-1i}) - \phi(v_{ji})\right],$$
$$\frac{\partial p_{ji}}{\partial h} = a_j \phi(v_{ji}) - a_{j-1}\phi(v_{j-1i}),$$

Pour calculer les probabilités correspondant aux premières et dernières modalités, il suffit de remarquer que $\phi(-\infty) = \phi(+\infty) = 0$. En conséquence, il suffit d'annuler les deux termes suivants dans les expressions ci-dessus :

$$\phi(v_{0i}) = 0 \quad \text{et} \quad \phi(v_{Ji}) = 0.$$

6.2.2.2 Estimation avec seuils inconnus

On ne peut plus estimer β et h séparément. En posant $\alpha_j = a_j h$, on doit estimer le modèle correspondant aux probabilités :

$$p_{ji} = \Phi(\alpha_j - X_i\beta) - \Phi(\alpha_{j-1} - X_i\beta).$$

Le paramètre à estimer est donc :

$$\theta' = (\beta', \alpha_2, ..., \alpha_{J-1}).$$

et les dérivées de la probabilité sont données par :

$$\frac{\partial p_{ji}}{\partial \beta} = X_i' \left[\phi(v_{j-1i}) - \phi(v_{ji})\right],$$
$$\frac{\partial p_{ji}}{\partial \alpha_j} = \phi(v_{ji}),$$

avec

$$\phi(v_{0i}) = 0 \quad \text{et} \quad \phi(v_{Ji}) = 0.$$

6.3 Les variables non ordonnées

6.3.1 Cas général

Dans ce cas, il n'est plus possible de mettre un ordre sur les modalités de la variable observable. Il peut s'agit, par exemple, d'un pays, d'un mode de transport, d'une couleur etc.. Dans ce cas on remplace l'ordre objectif (i.e. unanime) propre aux variables ordonnées par un ordre subjectif (i.e. propre à chacun) que l'on dérive d'une représentation en termes de comparaisons d'utilités. On considère qu'un individu se trouve confronté à un choix parmi J possibilités et que chaque possibilité $j \in \{1, ..., J\}$ procure une utilité U_j définie par :

$$U_j = V_j(X_j) + \varepsilon_j, \quad j = 1, ..., J$$

où $V_j(X_j)$ est la partie déterministe de l'utilité, c'est-à-dire la partie que l'on peut expliquer, X_j la liste des variables explicatives intervenant dans cette utilité et ε_j la partie aléatoire de l'utilité, indépendante des variables explicatives X_j. Un individu pris au hasard choisit la modalité k si :

$$U_k > U_j, \quad \forall j \neq k,$$

on observe donc le choix k par l'individu i avec la probabilité :

$$\begin{aligned} p_{ki} &= \Pr\left[\bigcap_{j \neq k} (U_{ki} > U_{ji})\right] \\ &= \Pr\left[\bigcap_{j \neq k} (V_k(X_{ki}) + \varepsilon_{ki} > V_j(X_{ji}) + \varepsilon_{ji})\right] \\ &= \Pr\left[\bigcap_{j \neq k} (\varepsilon_{ji} < V_k(X_{ki}) - V_j(X_{ji}) + \varepsilon_{ki})\right]. \end{aligned}$$

La forme particulière de la probabilité p_{ji} dépend de la distribution jointe des $(\varepsilon_{1i}, ..., \varepsilon_{Ji})$ et de la forme retenue pour la partie déterministe des fonctions d'utilité.

6.3.2 Le modèle logistique multinomial

Pour obtenir ce modèle, également appelé "multinomial Logit", on fait l'hypothèse que les $(\varepsilon_{1i}, ..., \varepsilon_{Ji})$ sont indépendamment et identiquement distribués selon une loi de Gompertz de paramètres $(0, 1)$ de fonction de

répartition :[1]

$$F(\varepsilon) = \exp(-\exp(-\varepsilon)).$$

La densité cette distribution est donc égale à :

$$f(\varepsilon) = \exp(-\varepsilon)\exp(-\exp(-\varepsilon)).$$

En notant $V_{ji} = V_j(X_{ji}), \forall j$, on a :

$$\begin{aligned} p_{ki} &= \mathrm{E}(d_{ki} = 1) \\ &= \mathrm{E}\left[\prod_{j\neq k} \Pr[\varepsilon_{ji} < V_{ki} - V_{ji} + \varepsilon_{ki}]\right], \\ &= \int_{-\infty}^{+\infty} \prod_{j\neq k} \Pr[\varepsilon_{ji} < V_{ki} - V_{ji} + \varepsilon_{ki}] f(\varepsilon_{ki}) \mathrm{d}\varepsilon_{ki}. \end{aligned}$$

En développant, on obtient :

$$\begin{aligned} \prod_{j\neq k} \Pr[\varepsilon_{ji} < V_{ki} - V_{ji} + \varepsilon_{ki}] &= \prod_{j\neq k} \exp[-\exp(-(V_{ki} - V_{ji} + \varepsilon_{ki}))] \\ &= \prod_{j\neq k} \exp[-\exp(-\varepsilon_{ki})\exp(V_{ji} - V_{ki})], \end{aligned}$$

ce qui implique :

$$\begin{aligned} p_{ki} = \int_{-\infty}^{+\infty} \prod_{j\neq k} \exp[-\exp(-\varepsilon_{ki})\exp(V_{ji} - V_{ki})]\exp(-\varepsilon_{ki}) \\ \exp(-\exp(-\varepsilon_{ki}))\mathrm{d}\varepsilon_{ki}, \end{aligned}$$

on effectue donc le changement de variable :

$$z = \exp(-\varepsilon_{ki}),$$

ce qui implique :

$$\mathrm{d}\varepsilon_{ki} = -\frac{\mathrm{d}z}{z}, \quad \lim_{\varepsilon_{ki}\to-\infty} z = +\infty \quad \text{et} \quad \lim_{\varepsilon_{ki}\to+\infty} = 0,$$

[1] Une variable Z suit une loi de Gompertz de paramètres (μ, ϕ) si sa fonction de répartition s'écrit :

$$F(z) = 1 - \exp\left\{-\exp\left(\frac{z-\mu}{\phi}\right)\right\}.$$

On obtient :

$$\mathrm{E}(Z) = \mu + 0.5772 \times \beta \text{ et } \mathrm{V}(Z) = \frac{\phi^2\pi^2}{6}.$$

d'où[2]

$$\begin{aligned} p_{ki} &= \int_0^{+\infty} \prod_{j \neq k} \exp\left(-z \exp\left(V_{ji} - V_{ki}\right)\right) \exp\left(-z\right) \mathrm{d}z \\ &= \int_0^{+\infty} \exp\left(-z \sum\nolimits_{j \neq k} \exp\left(V_{ji} - V_{ki}\right)\right) \exp\left(-z\right) \mathrm{d}z \\ &= \int_0^{+\infty} \exp\left\{-z\left(1 + \sum\nolimits_{j \neq k} \exp\left(V_{ji} - V_{ki}\right)\right)\right\} \mathrm{d}z. \end{aligned}$$

La fonction a intégrer est de la forme $\exp\left(-az\right)$ et admet pour primitive $-1/a \exp\left(-az\right)$, on obtient finalement :

$$\begin{aligned} p_{ki} &= \left[-\frac{\exp\left\{-z\left(1 + \sum_{j \neq k} \exp\left(V_{ji} - V_{ki}\right)\right)\right\}}{1 + \sum_{j \neq k} \exp\left(V_{ji} - V_{ki}\right)} \right]_0^{+\infty} \\ &= \frac{1}{1 + \sum_{j \neq k} \exp\left(V_{ji} - V_{ki}\right)}. \end{aligned}$$

Il est possible de réécrire cette probabilité en multipliant son numérateur et son dénominateur par $\exp\left(V_{ki}\right)$, ce qui donne :

$$p_{ki} = \frac{\exp\left(V_{ki}\right)}{\exp\left(V_{ki}\right) + \sum_{j \neq k} \exp\left(V_{ji}\right)} = \frac{\exp\left(V_{ki}\right)}{\sum_{j=1}^{J} \exp\left(V_{ji}\right)}.$$

Pour obtenir le modèle logistique multinomial, on fait l'hypothèse supplémentaire que les utilités sont linéaires par rapport aux paramètres :

$$V_{ji}\left(X_j\right) = X_{ji} b_j, \quad j = 1, ..., J,$$

ce qui donne finalement :

$$p_{ki} = \frac{\exp\left(X_{ki} b_k\right)}{\sum_{j=1}^{J} \exp\left(X_{ji} b_j\right)}, \quad k = 1, ..., J.$$

Comme tous les paramètres du modèle ne sont pas identifiables, on doit imposer une contrainte sur les paramètres b_j. Dans le cas où les variables explicatives sont identiques pour toutes les modalités, on prend une modalité de référence, que l'on note $j = 1$; il s'agit généralement de la modalité la plus répandue. En effet dans ce cas particulier, on a:

$$p_{ki} = \frac{\exp\left(X_i b_k\right)}{\sum_{j=1}^{J} \exp\left(X_i b_j\right)} = \frac{\exp\left(X_i \left(b_k - b_1\right)\right)}{1 + \sum_{j=2}^{J} \exp\left(X_i \left(b_j - b_1\right)\right)},$$

[2] On utilise le fait qu'intervertir les bornes de l'intégrale change son signe.

et tous les effets sont mesurés par rapport à l'utilité que procure le choix le plus courant. On voit clairement que seules les transformations de $b_j - b_1$ peuvent être estimées. On pose $\beta_j = b_j - b_1$, ce qui donne :

$$p_{ki} = \frac{\exp(X_i\beta_k)}{1 + \sum_{j=2}^{J} \exp(X_i\beta_j)}, \quad k = 1, ..., J.$$

On remarque ici que pour $k = 1$:

$$p_{1i} = \frac{1}{1 + \sum_{j=2}^{J} \exp(X_i\beta_j)},$$

parce que $\beta_1 = b_1 - b_1 = 0$. La nullité du coefficient de la première modalité est donc la contrainte identifiante du modèle. On ne calcule donc pas de dérivée par rapport à β_1. Les dérivées de la probabilité sont données simplement par :

$$k = 2, ..., J, \ \frac{\partial p_{ki}}{\partial \beta_k} = X_i' p_{ki}(1 - p_{ki}), \ \left.\frac{\partial p_{ki}}{\partial \beta_j}\right|_{j \neq k} = -X_i' p_{ki} p_{ji}.$$

CHAPITRE 7

Le pseudo maximum de vraisemblance

Dans le cas général, l'estimateur du maximum de vraisemblance n'est convergent et asymptotiquement efficace que si l'hypothèse que l'on fait sur la loi conditionnelle de la variable expliquée y est juste. Sinon, il peut ne pas être convergent. Il existe une famille de distributions pour lesquelles une erreur de spécification de ce type ne remet pas en cause la convergence de l'estimateur du maximum de vraisemblance. Par contre il faut évaluer différemment sa matrice de covariance, c'est à dire changer les statistiques de test. Nous supposerons donc dans cette section que nous ne connaissons *que* l'espérance conditionnelle de la variable expliquée $\mathrm{E}(y|X,\theta)$.

7.1 Le pseudo maximum de vraisemblance à l'ordre 1

7.1.1 La famille exponentielle linéaire à l'ordre 1

DÉFINITION 7.1 *La famille exponentielle linéaire à l'ordre 1 désigne une famille de distributions dont la densité admet la forme suivante :*

$$f(y,m) = \exp\{A(m) + B(y) + C(m) \times y\} \quad où \quad E(y) = m.$$

Cette forme est vérifiée par de nombreuses lois usuelles, dont voici quelques exemples :

Exemple 7.1 *Loi normale.* $y \rightsquigarrow N(m,\omega)$. *Son espérance mathématique*

est égale à $m \in \mathbb{R}$ *et la densité s'écrit :*

$$\begin{aligned}\ln f(y,m) &= \ln\left[\frac{1}{\sqrt{2\pi\omega}}\exp\left(-\frac{1}{2\omega}(y-m)^2\right)\right] \\ &= \underbrace{-\frac{m^2}{2\omega}}_{A(m)}\underbrace{-\frac{1}{2}\ln(2\pi\omega)-\frac{y^2}{2\omega}}_{B(y)} + y\times\underbrace{\frac{m}{\omega}}_{C(m)}.\end{aligned}$$

Exemple 7.2 *Loi de Poisson.* $y \rightsquigarrow P(m)$*. Son espérance mathématique est égale à* $m \in \mathbb{R}^+$ *et les probabilités s'écrivent :*

$$\begin{aligned}\ln f(y,m) &= \ln\left[\frac{\exp(-m)\,m^y}{y!}\right] \\ &= \underbrace{-m}_{A(m)}\underbrace{-\ln y!}_{B(y)} + y\times\underbrace{\ln m}_{C(m)}.\end{aligned}$$

Exemple 7.3 *Loi de Bernoulli.* $y \rightsquigarrow B(m)$*. Son espérance mathématique est égale à* $m \in]0,1[$ *et les probabilités s'écrivent :*

$$\begin{aligned}\ln f(y,m) &= \ln\left[m^y(1-m)^{1-y}\right] \\ &= \underbrace{\ln(1-m)}_{A(m)} + \underbrace{0}_{B(y)} + y\times\underbrace{\ln\frac{m}{1-m}}_{C(m)}.\end{aligned}$$

Exemple 7.4 *Loi Gamma.* $y \rightsquigarrow \Gamma(a,b)$*. Ce cas est un peu plus compliqué que les précédents car l'espérance mathématique n'est pas un paramètre utilisé habituellement avec cette loi. Il faut donc la réécrire en fonction de* $m = ab$*. La forme usuelle de la probabilité est :*

$$f(y) = \frac{y^{a-1}\exp(-y/b)}{b^a\Gamma(a)},\ a,b>0,\ y>0$$

avec :[1]

$$m = ab \Leftrightarrow b = m/a$$

en remplaçant dans la densité, on obtient :

$$\begin{aligned}\ln f(y,m) &= \ln\left[\frac{y^{a-1}\exp(-ya/m)}{(m/a)^a\Gamma(a)}\right] \\ &= \underbrace{-a\ln(m/a)-\ln\Gamma(a)}_{A(m)} + \underbrace{(a-1)\ln y}_{B(y)} + y\times\underbrace{\left(-\frac{a}{m}\right)}_{C(m)}.\end{aligned}$$

[1] Ici on peut prendre soit $b = m/a$ soit $a = m/b$, et l'on prend la première possibilité parce qu'elle mène aux calculs les plus simples.

Exemple 7.5 *Loi Binomiale négative.* $y \rightsquigarrow BN(r,p)$. *Dans ce cas également, l'espérance mathématique n'est pas un paramètre utilisé habituellement avec cette loi. Il faut donc la réécrire en fonction de* $m = r(1-p)/p$. *La forme usuelle de la probabilité est :*

$$f(y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p^r (1-p)^y,\ 0 < p < 1 \text{ et } r > 0,$$

avec :

$$m = r\frac{(1-p)}{p} \Leftrightarrow p = \frac{r}{r+m},$$

en remplaçant dans l'expression de la densité on obtient :

$$\begin{aligned}\ln f(y,m) &= \ln\left[\frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}\left(\frac{r}{r+m}\right)^r\left(\frac{m}{r+m}\right)^y\right] \\ &= \underbrace{r\ln\left(\frac{r}{r+m}\right)}_{A(m)} + \underbrace{\ln\frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)}}_{B(y)} + y \times \underbrace{\ln\left(\frac{m}{r+m}\right)}_{C(m)}.\end{aligned}$$

PROPRIÉTÉ 7.1 *Les lois de la famille exponentielle linéaire vérifient les deux propriétés suivantes :*

1. $\dfrac{\partial A}{\partial m} + m\dfrac{\partial C}{\partial m} = 0.$
2. $V(y) = \left(\dfrac{\partial C}{\partial m}\right)^{-1}$

PREUVE :
Pour démontrer la première propriété, on remarque que :

$$\int f(y,m)\,\mathrm{d}y = 1,$$

en dérivant cette relation par rapport à m, on obtient :

$$\begin{aligned}&\int\left(\frac{\partial A}{\partial m} + \frac{\partial C}{\partial m} \times y\right) f(y,m)\,\mathrm{d}y = 0 \\ &\Leftrightarrow \frac{\partial A}{\partial m}\underbrace{\int f(y,m)\,\mathrm{d}y}_{1} + \frac{\partial C}{\partial m}\underbrace{\int y f(y,m)\,\mathrm{d}y}_{\mathrm{E}(y)} = 0 \\ &\Leftrightarrow \frac{\partial A}{\partial m} + m\frac{\partial C}{\partial m} = 0.\end{aligned}$$

Pour démontrer la seconde propriété, on remarque que :

$$\mathrm{E}(y) = \int y f(y,m)\,\mathrm{d}y = m,$$

en dérivant cette relation par rapport à m, on obtient :

$$\begin{aligned}
&\int \left(\frac{\partial A}{\partial m} + \frac{\partial C}{\partial m} \times y\right) y f(y,m)\,\mathrm{d}y = 1 \\
&\Leftrightarrow \frac{\partial A}{\partial m}\underbrace{\int y f(y,m)\,\mathrm{d}y}_{\mathrm{E}(y)} + \frac{\partial C}{\partial m}\underbrace{\int y^2 f(y,m)\,\mathrm{d}y}_{\mathrm{E}(y^2)} = 1 \\
&\Leftrightarrow \frac{\partial A}{\partial m} m + \frac{\partial C}{\partial m}\left(\mathrm{V}(y) + m^2\right) = 1 \\
&\Leftrightarrow \left(-m\frac{\partial C}{\partial m}\right) m + \frac{\partial C}{\partial m}\left(\mathrm{V}(y) + m^2\right) = 1 \\
&\Leftrightarrow \frac{\partial C}{\partial m}\mathrm{V}(y) = 1 \Leftrightarrow \mathrm{V}(y) = \left(\frac{\partial C}{\partial m}\right)^{-1}.
\end{aligned}$$

□

Vérifions-le sur nos exemples :

Exemple 7.6 *Loi normale* $N(m,\omega)$. *On a* $C(m) = m/\omega$ *donc* $\partial C/\partial m = 1/\omega$ *et* $V(y) = \omega$.

Exemple 7.7 *Loi de Poisson* $P(m)$. *On a* $C(m) = \ln m$ *donc* $\partial C/\partial m = 1/m$ *et* $V(y) = m$.

Exemple 7.8 *Loi de Bernoulli* $B(m)$. *On a* $C(m) = \ln(m/(1-m))$ *donc* $\partial C/\partial m = 1/(m(1-m))$ *et* $V(y) = m(1-m)$.

Exemple 7.9 *Loi Gamma* $\Gamma(a,b)$ *avec* $m = ab$. *On a* $C(m) = -a/m$ *donc* $\partial C/\partial m = a/m^2$ *et* $V(y) = m^2/a = ab^2$.

Exemple 7.10 *Loi Binomiale Négative* $BN(r,p)$ *avec* $m = r(1-p)/p$. *On a* $C(m) = \ln(m/(m+r))$ *donc* $\partial C/\partial m = r/(m(m+r))$ *et* $V(y) = m(m+r)/r = r(1-p)/p^2$.

7.1.2 Estimation

On note θ le paramètre à estimer. Ce paramètre intervient ici dans l'espérance mathématique de la distribution de y. Ainsi, dans le cas d'un modèle linéaire avec variables explicatives X_i, on aurait $m = X_i\theta$. L'espérance peut donc être différente pour chaque observation dans un modèle avec variables explicatives. Plus généralement, on considérera une espérance conditionnelle sous la forme générale :

$$m_i = m\left(X_i, \theta\right).$$

L'estimateur du pseudo maximum de vraisemblance est obtenu en maximisant la pseudo vraisemblance suivante :

$$\begin{aligned}\ell\left(y_i \middle| X_i; \theta\right) &= \sum_{i=1}^{N} \ln f\left(y_i \middle| m\left(X_i, \theta\right)\right) \\ &= \sum_{i=1}^{N} A\left[m\left(X_i, \theta\right)\right] + B\left(y_i\right) + C\left[m\left(X_i, \theta\right)\right] \times y_i.\end{aligned}$$

On remarque que pour la maximisation, on peut négliger les termes qui ne dépendent pas de θ. L'estimateur du pseudo maximum de vraisemblance $\widetilde{\theta}_N$ se définit donc comme :

$$\begin{aligned}\widetilde{\theta}_N &= \arg\max_{\theta} \sum_{i=1}^{N} A\left[m\left(X_i, \theta\right)\right] + B\left(y_i\right) + C\left[m\left(X_i, \theta\right)\right] \times y_i \\ &= \arg\max_{\theta} \sum_{i=1}^{N} A\left[m\left(X_i, \theta\right)\right] + C\left[m\left(X_i, \theta\right)\right] \times y_i.\end{aligned}$$

La condition du premier ordre est donnée par :

$$\begin{aligned}\frac{\partial \ell}{\partial \theta}\left(y_i \middle| X_i; \widetilde{\theta}_N\right) &= 0 \\ &\Leftrightarrow \sum_{i=1}^{N} \frac{\partial m}{\partial \theta}\left[\frac{\partial A}{\partial m} + \frac{\partial C}{\partial m} \times y_i\right] = 0 \\ &\Leftrightarrow \sum_{i=1}^{N} \frac{\partial m}{\partial \theta}\frac{\partial C}{\partial m}\left[y_i - m\left(X_i, \widetilde{\theta}_N\right)\right] = 0 \\ &\Leftrightarrow \sum_{i=1}^{N} \frac{\partial m}{\partial \theta}\Sigma^{-1}\left[y_i - m\left(X_i, \widetilde{\theta}_N\right)\right] = 0,\end{aligned}$$

en utilisant le fait que $\partial A/\partial m + m \times \partial C/\partial m = 0$ et que $\left[\partial C/\partial m\right]^{-1} = \mathrm{V}(y) = \Sigma$. Remarquons ici que Σ est la matrice de covariance de la pseudo loi,

qui peut être différente de la matrice de covariance de la vraie loi, qui est inconnue dans le cas général (i.e., nous n'avons pas fait d'hypothèse sur la variance conditionnelle). La vraie matrice de covariance est notée Ω. On obtient l'équivalent des matrices $\mathrm{I}_N(\theta)$ et $\mathrm{J}_N(\theta)$ de la manière suivante :[2]

$$\begin{aligned}
\mathrm{I}_N(\theta) &= \sum_{i=1}^{N} \underset{0}{\mathrm{E}} \left[\frac{\partial \ln f\left(y_i | m\left(X_i;\theta\right)\right)}{\partial \theta} \frac{\partial \ln f\left(y_i | m\left(X_i;\theta\right)\right)}{\partial \theta'} \right] \\
&= \sum_{i=1}^{N} \underset{0}{\mathrm{E}} \left[\frac{\partial m}{\partial \theta} \Sigma^{-1} \left(y_i - m\right)^2 \Sigma^{-1} \frac{\partial m}{\partial \theta'} \right] \\
&= \sum_{i=1}^{N} \frac{\partial m}{\partial \theta} \Sigma^{-1} \underset{0}{\mathrm{E}} \left[\left(y_i - m\right)^2 \right] \Sigma^{-1} \frac{\partial m}{\partial \theta'} \\
&= \sum_{i=1}^{N} \frac{\partial m}{\partial \theta} \Sigma^{-1} \Omega \Sigma^{-1} \frac{\partial m}{\partial \theta'},
\end{aligned}$$

où $\mathrm{E}_0[.]$ représente l'espérance mathématique par rapport à la *vraie* loi de y. Pour obtenir $\mathrm{J}_N(\theta)$ il faut se rappeler que Σ dépend de m :

$$\begin{aligned}
\mathrm{J}_N(\theta) &= \sum_{i=1}^{N} \underset{0}{\mathrm{E}} \left[-\frac{\partial^2 \ln f\left(y_i | m\left(X_i;\theta\right)\right)}{\partial \theta \partial \theta'} \right] \\
&= -\sum_{i=1}^{N} \underset{0}{\mathrm{E}} \left[\frac{\partial^2 m}{\partial \theta \partial \theta'} \Sigma^{-1} \left(y_i - m\right) \right] \\
&\quad - \sum_{i=1}^{N} \underset{0}{\mathrm{E}} \left[\frac{\partial m}{\partial \theta} \left(\frac{\partial \left(\Sigma^{-1}\right)}{\partial \theta'} \left(y_i - m\right) - \Sigma^{-1} \frac{\partial m}{\partial \theta'} \right) \right] \\
&= \sum_{i=1}^{N} \frac{\partial m}{\partial \theta} \Sigma^{-1} \frac{\partial m}{\partial \theta'}.
\end{aligned}$$

On remarque donc que pour $\Sigma \neq \Omega$, on a :

$$\mathrm{I}_N(\theta) \neq \mathrm{J}_N(\theta).$$

En effectuant un développement limité de la condition du premier

[2] On prend les espérances mathématiques par rapport à la vraie loi car c'est vers ces quantités que convergeront nos statistiques.

ordre au voisinage de $\widetilde{\theta}_N = \theta$, on obtient :

$$\frac{\partial \ell}{\partial \theta}\left(y_i|\, X_i; \widetilde{\theta}_N\right) \overset{pp}{=} \frac{\partial \ell}{\partial \theta}\left(y_i|\, X_i; \theta\right) + \frac{\partial^2 \ell}{\partial \theta \partial \theta'}\left(y_i|\, X_i; \theta\right)\left(\widetilde{\theta}_N - \theta\right)$$

$$\Leftrightarrow \sqrt{N}\left(\widetilde{\theta}_N - \theta\right) \overset{pp}{=} \left[-\frac{1}{N}\frac{\partial^2 \ell}{\partial \theta \partial \theta'}\left(y_i|\, X_i; \theta\right)\right]^{-1} \frac{1}{\sqrt{N}}\frac{\partial \ell}{\partial \theta}\left(y_i|\, X_i; \theta\right),$$

le premier terme converge vers $\mathrm{J}_1(\theta)$ et le second terme est la moyenne de variables aléatoires d'espérance nulle et de variance $\mathrm{I}_1(\theta)$. En conséquence :

$$\sqrt{N}\left(\widetilde{\theta}_N - \theta\right) \overset{A}{\rightsquigarrow} \mathrm{N}\left(0, \mathrm{J}_1^{-1}(\theta)\,\mathrm{I}_1(\theta)\,\mathrm{J}_1^{-1}(\theta)\right),$$

dans la pratique, on utilise :

$$\widetilde{\theta}_N \overset{A}{\rightsquigarrow} \mathrm{N}\left(\theta, \mathrm{J}_N^{-1}(\theta)\,\mathrm{I}_N(\theta)\,\mathrm{J}_N^{-1}(\theta)\right).$$

En effet, on voit que :

$$\begin{aligned}\frac{1}{N}\mathrm{J}_1^{-1}(\theta)\,\mathrm{I}_1(\theta)\,\mathrm{J}_1^{-1}(\theta) &= \left[N\mathrm{J}_1(\theta)\right]^{-1}\left[N\mathrm{I}_1(\theta)\right]\left[N\mathrm{J}_1(\theta)\right]^{-1} \\ &= \mathrm{J}_N^{-1}(\theta)\,\mathrm{I}_N(\theta)\,\mathrm{J}_N^{-1}(\theta).\end{aligned}$$

Reprenons nos trois exemples. Dans tous les cas, la spécification de la moyenne est supposée juste. On a donc $\mathrm{E}(y) = \theta$ et dans tous les cas, on trouve $\widetilde{\theta}_N = 1/N\sum_{i=1}^N y_i$. Pourtant, selon que la loi est normale, de Poisson ou de Bernoulli, les variances asymptotiques sont égales à ω/N, θ/N et $\theta(1-\theta)/N$. Pour calculer les variances asymptotiques du pseudo maximum de vraisemblance, il faut recalculer les matrices $\mathrm{I}_N(\theta)$. On suppose, dans les trois exemples suivants que la vraie variance est égale à ω. La suite montre que, lorsque l'on ne connaît pas la vraie loi avec certitude, il faut mieux utiliser l'estimateur suivant :

$$\widehat{\mathrm{Vas}}\left(\widetilde{\theta}_N\right) = \frac{\hat{\omega}}{N} \quad \text{avec} \quad \hat{\omega} = \frac{1}{N}\sum_{i=1}^N y_i^2.$$

Exemple 7.11 *Pseudo loi normale $N(\theta, \omega)$. La matrice $I_N(\theta)$ est maintenant donnée par :*

$$I_N(\theta) = \sum_{i=1}^N \underset{0}{\mathrm{V}}\left[y_i - \theta\right] = \frac{N}{\omega},$$

donc la variance asymptotique de $\widetilde{\theta}_N$ doit être estimée par :

$$J_N^{-1}(\theta)\, I_N(\theta)\, J_N^{-1}(\theta) = \left(\frac{N}{\omega}\right)^{-1}\frac{N}{\omega}\left(\frac{N}{\omega}\right)^{-1} = \frac{\omega}{N}.$$

Exemple 7.12 *Pseudo loi de Poisson $P(\theta)$. La matrice $I_N(\theta)$ est maintenant donnée par :*

$$I_N(\theta) = \sum_{i=1}^{N} \underset{0}{\mathrm{V}} \left[-1 + \frac{y_i}{\theta} \right] = \frac{N\omega}{\theta^2},$$

donc la variance asymptotique de $\widetilde{\theta}_N$ doit être estimée par :

$$J_N^{-1}(\theta) I_N(\theta) J_N^{-1}(\theta) = \left(\frac{N}{\theta} \right)^{-1} \frac{N\omega}{\theta^2} \left(\frac{N}{\theta} \right)^{-1} = \frac{\omega}{N}.$$

Exemple 7.13 *Pseudo loi de Bernoulli $B(\theta)$. La matrice $I_N(\theta)$ est maintenant donnée par :*

$$I_N(\theta) = \sum_{i=1}^{N} \underset{0}{\mathrm{V}} \left[\frac{y_i}{\theta(1-\theta)} - \frac{1}{1-\theta} \right] = \frac{N\omega}{\theta^2 (1-\theta)^2},$$

donc la variance asymptotique de $\widetilde{\theta}_N$ doit être estimée par :

$$J_N^{-1}(\theta) I_N(\theta) J_N^{-1}(\theta) = \left(\frac{N}{\theta(1-\theta)} \right)^{-1} \frac{N\omega}{\theta^2 (1-\theta)^2} \left(\frac{N}{\theta(1-\theta)} \right)^{-1} = \frac{\omega}{N}.$$

7.1.3 Matrice de covariance robuste à l'hétéroscédasticité de forme inconnue

Le modèle linéaire standard, estimé par le maximum de vraisemblance sous hypothèse de normalité, fournit un estimateur convergent de b même si la distribution de la perturbation u_i n'est pas normale $\mathrm{N}(0, \omega)$. C'est parce que la loi normale appartient à la famille exponentielle linéaire. En effet, elle peut s'écrire sous la forme :

$$f(y, m) = \exp\{A(m) + B(y) + y \times C(m)\}$$

où $m = Xb$ est l'espérance conditionnelle de y. Notons dès maintenant que quelle que soit la valeur du paramètre du second ordre ω, f appartient à la famille exponentielle linéaire. Nous pouvons donc nous en passer et lui donner une valeur quelconque. Ceci tient au fait que le pseudo maximum de vraisemblance à l'ordre 1 ne fait pas d'hypothèse sur la variance conditionnelle de y. Le paramètre ω est alors un paramètre de nuisance que l'on peut fixer arbitrairement; ce n'est plus nécessairement la variance des perturbations, car la vraie loi n'est pas normale dans le cas général. De plus, dans le cas du modèle linéaire, ceci n'affectera pas notre estimateur de b, puisqu'il est déterminé indépendamment de l'estimateur de ω. On peut donc poser $\omega = 1$ sans perte de généralité. On a :

$$A(m) = -\frac{1}{2} \left[m^2 + \ln(2\pi) \right], \quad B(y) = -\frac{y^2}{2}, \quad C(m) = -m.$$

La matrice de covariance de l'estimateur des moindres carrés n'est plus estimée par l'inverse de l'information de Fisher mais par :

$$\widehat{\mathrm{Vas}}\left(\sqrt{N}\left(\widehat{b}_N - b\right)\right) = \widehat{\mathrm{J}}_1^{-1}\left(\widehat{b}_N\right)\widehat{\mathrm{I}}_1\left(\widehat{b}_N\right)\widehat{\mathrm{J}}_1^{-1}\left(\widehat{b}_N\right). \qquad (7.1)$$

La matrice $\mathrm{J}_1(b)$ reste inchangée :

$$\mathrm{J}_1(b) = \underset{X}{\mathrm{E}}\,\underset{0}{\mathrm{E}}\left[-\frac{\partial^2 \ln f}{\partial b \partial b'}\right] = \underset{X}{\mathrm{E}}\left[X'X\right],$$

que l'on peut estimer de manière convergente par :

$$\widehat{\mathrm{J}}_1\left(\widehat{b}_N\right) = \frac{1}{N}\sum_{i=1}^{N} X_i' X_i.$$

La matrice $\mathrm{I}_1(b)$ est égale à :

$$\mathrm{I}_1(b) = \underset{X}{\mathrm{E}}\,\underset{0}{\mathrm{E}}\left[\frac{\partial \ln f}{\partial b}\frac{\partial \ln f}{\partial b'}\right] = \underset{X}{\mathrm{E}}\,\underset{0}{\mathrm{E}}\left[X'X\,u^2\right],$$

avec

$$u = y - Xb,$$

que l'on peut estimer de manière convergente par :

$$\widehat{\mathrm{I}}_1\left(\widehat{b}_N\right) = \frac{1}{N}\sum_{i=1}^{N} X_i' X_i \widehat{u}_i^2 \quad \text{avec} \quad \widehat{u}_i = y_i - X_i \widehat{b}_N.$$

On utilise donc finalement :

$$\widehat{\mathrm{Vas}}\left(\widehat{b}_N\right) = \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1} \sum_{i=1}^{N} X_i' X_i \widehat{u}_i^2 \left(\sum_{i=1}^{N} X_i' X_i\right)^{-1}.$$

La matrice de covariance ainsi obtenue est appelée *matrice de covariance robuste* (White, 1980). Plus précisément elle est robuste aux hypothèses de normalité et d'homoscédasticité des perturbations.[3] On l'utilise donc systématiquement de nos jours.

[3] La variance habituelle de l'estimateur des moindres carrés ordinaires est robuste à l'hypothèse de normalité, seule l'hétéroscédascité pose réellement problème ici.

7.2 Le pseudo maximum de vraisemblance quasi généralisé

7.2.1 La famille exponentielle quasi-généralisée

DÉFINITION 7.2 *La famille exponentielle quasi généralisée désigne une famille de distributions dont la densité admet la forme suivante :*

$$f(y,m,\eta) = \exp\{A(m,\eta) + B(y,\eta) + C(m,\eta) \times y\}$$
$$où \quad \mathrm{E}(y) = m \text{ et } \mathrm{V}(y) = \left[\partial C(m,\eta)/\partial m\right]^{-1}.$$

où η est un paramètre intervenant dans la variance de la pseudo distribution. Il n'est pas forcément égal à la variance de la distribution. Les lois de Poisson et de Bernoulli n'admettent pas de paramètre spécifique intervenant dans la variance, elles n'appartiennent donc pas à la famille exponentielle quasi-généralisée. Par contre, la loi normale appartient à cette famille. Nous introduisons également la loi binomiale négative qui généralise la loi de Poisson. Le lecteur pourra vérifier par lui-même qu'elle appartient à la famille linéaire exponentielle à l'ordre 1.

Exemple 7.14 *Loi normale.* $y \rightsquigarrow N(m,\eta)$. *On a* $\mathrm{E}(y) = m$ *et* $\mathrm{V}(y) = \eta$. *La densité s'écrit :*

$$\begin{aligned}\ln f(y,m,\eta) &= \ln\left[\frac{1}{\sqrt{2\pi\eta}}\exp\left(-\frac{1}{2\eta}(y-m)^2\right)\right] \\ &= \underbrace{-\frac{m^2}{2\eta}}_{A(m,\eta)}\underbrace{-\frac{1}{2}\ln(2\pi\eta) - \frac{y^2}{2\eta}}_{B(y,\eta)} + y \times \underbrace{\frac{m}{\eta}}_{C(m,\eta)}.\end{aligned}$$

Exemple 7.15 *Loi Gamma.* $y \rightsquigarrow \Gamma(\eta, m/\eta)$. *On a* $\mathrm{E}(y) = m$ *et* $\mathrm{V}(y) = m^2/\eta$. *La densité s'écrit :*

$$\begin{aligned}\ln f(y,m,\eta) &= \ln\left[\frac{y^{\eta-1}\exp(-y\eta/m)}{(m/\eta)^{\eta}\Gamma(\eta)}\right] \\ &= \underbrace{-\eta\ln(m/\eta) - \ln\Gamma(\eta)}_{A(m,\eta)}\underbrace{+(\eta-1)\ln y}_{B(y,\eta)} + y \times \underbrace{\left(-\frac{\eta}{m}\right)}_{C(m,\eta)}.\end{aligned}$$

Exemple 7.16 *Loi binomiale négative.* $y \rightsquigarrow BN(m,\eta)$. *On a* $\mathrm{E}(y) = m$ *et* $\mathrm{V}(y) = m(1+\eta m)$. *On remarque que, contrairement à la loi de Poisson, la variance conditionnelle peut être différente de la moyenne conditionnelle. On retrouve la loi de Poisson en prenant la limite quand* $\eta \to 0$. *La densité s'applique, dans le cas de base, aux variables entières*

positives et s'écrit :

$$f(y,m,\eta)=\frac{\Gamma(y+1/\eta)}{\Gamma(y+1)\,\Gamma(1/\eta)}\left(\frac{\eta m}{1+\eta m}\right)^{y}\left(\frac{1}{1+\eta m}\right)^{1/\eta},$$

ce qui implique :

$$\ln f(y,m,\eta)=\underbrace{\ln\left[\left(\frac{1}{1+\eta m}\right)^{1/\eta}\frac{1}{\Gamma(1/\eta)}\right]}_{A(m,\eta)}+\underbrace{\ln\frac{\Gamma(y+1/\eta)}{\Gamma(y+1)}}_{B(y,\eta)}+\underbrace{y\ln\left(\frac{\eta m}{1+\eta m}\right)}_{C(m,\eta)}.$$

7.2.2 Estimation

PROPRIÉTÉ 7.2 *L'estimateur du pseudo maximum de vraisemblance quasi généralisé (PMVQG) $\overline{\theta}_N$ vérifie les trois propriétés suivantes:*

1. $\sqrt{N}\left(\overline{\theta}_N-\theta\right)\overset{A}{\rightsquigarrow}\mathrm{N}\left(0,\mathrm{J}_1^{-1}(\theta)\right)$ avec :

$$J_1(\theta)=\underset{X}{\mathrm{E}}\,\underset{0}{\mathrm{E}}\left[\frac{\partial m}{\partial\theta}\Sigma^{-1}\frac{\partial m}{\partial\theta'}\right].$$

2. Il atteint la borne inférieure des matrices de covariances des estimateurs du pseudo maximum de vraisemblance à l'ordre 1.

3. Si la vraie loi de y appartient à la famille exponentielle linéaire et si les paramètres m et η sont fonctionnellement indépendants, l'estimateur du PMVQG est asymptotiquement équivalent à l'estimateur du maximum de vraisemblance sur la vraie loi.

7.2.3 Les moindres carrés pondérés

Supposons que l'on ait un modèle linéaire hétéroscédastique dont la forme de la variance est connue, donnée par :

$$\mathrm{V}(y_i|X_i)=\omega(X_i,\eta)>0.$$

Dans un premier temps, on applique le pseudo maximum de vraisemblance à l'ordre 1, sans tenir compte de l'hétéroscédasticité des perturbations. Cet estimateur est convergent et sa variance asymptotique est donnée par la relation (7.1). Dans un second temps, on estime le paramètre η. Pour cela, on utilise la relation :

$$\mathrm{E}\left((y_i-m_i)^2\right)=\omega(X_i,\eta),$$

ce qui permet d'écrire la régression :

$$u_i^2 = \omega\left(X_i, \eta\right) + v_i \quad \text{avec} \quad \mathrm{E}\left(v_i\right) = 0.$$

on obtient donc un estimateur convergent de η, noté $\widehat{\eta}$, en remplaçant u_i^2 par $\widehat{u}_i^2$. A partir de cet estimateur, on estime la variance par :

$$\widehat{\omega}_i = \omega\left(X_i, \widehat{\eta}\right).$$

Ensuite on maximise la pseudo vraisemblance quasi généralisée, obtenue en posant $\omega = \widehat{\omega}_i$ dans la pseudo vraisemblance. On note que la variance est différente avec chaque observation et que $\widehat{\omega}_i$ ne dépend que de $\widehat{b}_N$, pas de b . Ceci donne l'estimateur :

$$\begin{aligned}
\overline{b}_N &= \arg\max_b \overline{\ell}\left(y|X, b, \widehat{\omega}\right) \\
&= \arg\max_b -\frac{N}{2}\ln\left(2\pi\right) - \frac{1}{2}\sum_{i=1}^{N}\ln\widehat{\omega}_i - \frac{1}{2}\sum_{i=1}^{N}\frac{1}{\widehat{\omega}_i}\left(y_i - X_i b\right)^2 \\
&= \arg\min_b \sum_{i=1}^{N}\left(\frac{y_i - X_i b}{\sqrt{\widehat{\omega}_i}}\right)^2,
\end{aligned}$$

ce qui correspond à l'estimateur des moindre carrés ordinaires avec une pondération égale à l'inverse de l'écart-type de la perturbation. La variance asymptotique de cet estimateur est donné par :

$$\widehat{\mathrm{Vas}}\left(\overline{b}_N\right) = \left[\widehat{\mathrm{J}}_1\left(\overline{b}_N\right)\right]^{-1} = \left(\sum_{i=1}^{N}\frac{1}{\widehat{\omega}_i}X_i' X_i\right)^{-1}.$$

CHAPITRE 8

Les variables entières

8.1 Le modèle de Poisson

8.1.1 Introduction

La loi de Poisson permet de décrire le nombre de réalisations d'un évènement pendant une période de temps donnée. Soit y_i la variable que l'on souhaite expliquer (e.g., le nombre de brevets). Comme l'espérance mathématique d'une donnée de comptage est toujours strictement positive on l'écrit sous la forme :[1]

$$\mathrm{E}\left(y_i|X_i, b\right) = \exp\left(X_i b\right) > 0,$$

où X_i est le vecteur des variables explicatives et b le paramètre correspondant. On peut réécrire cette relation sous la forme :

$$\ln \mathrm{E}\left(y_i|X_i, b\right) = X_i b,$$

de sorte que

$$\frac{\partial \ln \mathrm{E}\left(y_i|X_i, b\right)}{\partial X_i'} = b.$$

Cette relation fait apparaître que le paramètre b sera le vecteur des élasticités de l'espérance de y par rapport à X dès lors que les variables explicatives seront prises en logarithmes. Mais, contrairement au cas du modèle linéaire, il s'agira pas ici de l'élasticité de y par rapport à X.

[1] Une variable de comptage ne prend que des valeur positives ou nulle, d'où le résultat.

8.1.1.1 Modèle homogène

Le modèle de Poisson homogène s'obtient dès lors que l'on postule que les y_i sont indépendamment et identiquement distribués selon une loi de Poisson de moyenne conditionnelle $m_i = \exp(X_i b)$. La probabilité d'observer une réalisation y_i de la variable de la variable de comptage est donc donnée par :

$$f(y_i) = \frac{\exp(-m_i)\, m_i^{y_i}}{y_i!}, \quad y_i \in \{0, 1, 2, ...\}.$$

Cette hypothèse implique que la variance conditionnelle est égale à la moyenne conditionnelle :

$$\mathrm{V}(y_i|X_i) = \mathrm{E}(y_i|X_i) \quad \forall i.$$

On obtient ce modèle en postulant qu'il n'y a pas de perturbation dans l'expression de l'espérance conditionnelle de y_i. Les pertubations expriment généralement une forme d'hétérogénéité individuelle inobservable, de sorte qu'en leur absence on parle de modèle homogène.

8.1.1.2 Modèle hétérogène

On peut penser que le modèle précédent est insuffisant pour représenter les différences entre les individus, car celles-ci ne s'expriment que par les variables déterministes X_i. On peut penser qu'il existe également des caractéristiques individuelles inobservables, supposées sans corrélation avec les X_i, qui interviennent également dans l'espérance. La moyenne comporte alors une partie déterministe et une partie aléatoire u_i :

$$\ln \mathrm{E}(y_i|X_i, b) = X_i b + u_i,$$

où u_i est une perturbation qui vérifie $\mathrm{E}(\exp u_i) = 1$. La moyenne du processus de Poisson est alors elle-même aléatoire de sorte qu'il y a deux sources d'aléa dans ce modèle : d'une part, un aléa sur la moyenne et, d'autre part, un aléa lié au tirage dans une loi de Poisson de moyenne donnée. On note cette moyenne $\widetilde{m}_i$, définie par :

$$\widetilde{m}_i = \exp(X_i b + u_i) = \exp(X_i b) \exp(u_i) = m_i \exp(u_i).$$

Pour pouvoir écrire la vraisemblance de ce modèle, il faut faire une hypothèse spécifique sur la loi de $\exp(u_i)$. Nous ne suivrons pas cette approche ici car seuls quelques cas peuvent être écrits sous forme explicite; la plupart du temps, il faudrait recourir à l'intégration numérique. Nous prendrons donc une approche par le pseudo maximum de vraisemblance

à l'ordre 1, qui ne nécessite que l'expression de la moyenne conditionnelle de y_i. Ici, cette expression est tout simplement donnée par :

$$\mathrm{E}\left(y_i|X_i,b\right) = \mathrm{E}\left(m_i \exp u_i\right) = m_i \underbrace{\mathrm{E}\left(\exp u_i\right)}_{1} = m_i.$$

Notons ici que l'hypothèse n'implique aucune perte de généralité tant que le modèle contient un terme constant. Si on avait fait l'hypothèse que $\mathrm{E}(\exp u_i) = k$, on aurait trouvé :

$$\mathrm{E}\left(y_i|X_i,b\right) = km_i = k\exp\left(X_i b\right) = \exp\left(-\ln k + X_i b\right),$$

de sorte que le terme $-\ln k$ est absorbé par le terme constant du modèle.

8.1.2 Estimation

L'espérance mathématique d'un modèle hétérogène est la même que celle d'un modèle homogène. De plus, nous avons vu précédemment que la loi de Poisson appartient à la famille exponentielle linéaire. Dans ces conditions, quelle-est la différence entre l'estimateur du maximum de vraisemblance et celui du pseudo maximum de vraisemblance? Réponse : la matrice de covariance asymptotique. Celle du pseudo maximum de vraisemblance est robuste.

8.1.2.1 Maximum de vraisemblance

La log-vraisemblance pour une observation est donnée par :

$$\begin{aligned}\ell_i = \ln\left\{\frac{\exp\left(-m_i\right)m_i^{y_i}}{y_i!}\right\} &= -m_i + y_i \ln m_i - \ln y_i! \\ &= -\exp\left(X_i b\right) + y_i X_i b - \ln y_i!.\end{aligned}$$

On peut également l'écrire comme :

$$\ell\left(y_i,\mu_i\right) = -\exp\left(\mu_i\right) + y_i\mu_i - \ln y_i!,$$

avec $\mu_i = X_i b$. Les dérivées par rapport à μ_i sont égales à :

$$\frac{\partial \ell_i}{\partial \mu_i} = y_i - \exp\left(\mu_i\right) = y_i - \lambda_i, \quad \frac{\partial^2 \ell_i}{\partial \mu_i^2} = -\exp\left(\mu_i\right) = -m_i.$$

On en déduit que le score est égal à

$$\frac{\partial \ell}{\partial b} = \sum_{i=1}^{N} X_i' \left(y_i - m_i\right),$$

et que le hessien est égal à :

$$\frac{\partial^2 \ell}{\partial b \partial b'} = -\sum_{i=1}^{N} X_i' X_i m_i \ll 0.$$

La nullité du score exprime ici encore la condition d'orthogonalité entre les variables explicatives X et le résidu de la régression $\hat{u}_i = y_i - \widehat{m}_i = y_i - \widehat{\mathrm{E}}\left(y_i | X_i, b\right)$. La seule différence avec le cas habituel est que l'espérance mathématique est non linéaire $\widehat{m}_i = \exp\left(X_i \hat{b}\right)$. Comme le hessien est défini négatif, l'estimateur du maximum de vraisemblance $\widehat{b}$ est unique et donné par la condition du premier ordre :

$$\frac{\partial \ell}{\partial b}\left(\widehat{b}\right) = 0 \Leftrightarrow \sum_{i=1}^{N} X_i' \left[y_i - \exp\left(X_i \widehat{b}\right)\right] = 0.$$

La distribution asymptotique de cet estimateur est normale :

$$\sqrt{N}\left(\widehat{b} - b\right) \underset{N \to +\infty}{\xrightarrow{L}} \mathrm{N}\left(0, \mathrm{J}_1^{-1}\left(b\right)\right),$$

avec

$$\mathrm{J}_1\left(b\right) = \underset{X}{\mathrm{E}} \underset{y}{\mathrm{E}} \left[-\frac{\partial^2 \ln f}{\partial b \partial b'}\left(y|X\right)\right].$$

On remarque que la matrice hessienne ne dépend pas de y, ce qui implique qu'elle est égale à son espérance mathématique par rapport à la loi de y. On estime cette matrice par :

$$\widehat{\mathrm{J}_1} = \frac{1}{N} \sum_{i=1}^{N} X_i' X_i \exp\left(X_i \widehat{b}\right).$$

8.1.2.2 Pseudo maximum de vraisemblance

Comme l'espérance mathématique est identique dans les modèles homogène et hétérogène, et comme la loi de Poisson appartient à la famille exponentielle linéaire, seule la matrice de covariance asymptotique est changée. On a :

$$\sqrt{N}\left(\widehat{b} - b\right) \underset{n \to +\infty}{\xrightarrow{L}} \mathrm{N}\left(0, \mathrm{J}_1^{-1}\left(b\right) \mathrm{I}_1\left(b\right) \mathrm{J}_1^{-1}\left(b\right)\right),$$

où $\mathrm{J}_1\left(b\right)$ a été définie dans la section précédente et

$$\mathrm{I}_1\left(b\right) = \underset{X}{\mathrm{E}} \underset{y}{\mathrm{E}} \left[\frac{\partial \ln f}{\partial b}\left(y|X\right) \frac{\partial \ln f}{\partial b'}\left(y|X\right)\right].$$

On estime la matrice $\mathrm{I}_1(b)$ par le moment empirique correspondant :

$$\begin{aligned}\widehat{\mathrm{I}_1} &= \frac{1}{N}\sum_{i=1}^{N} X_i' X_i \left(y_i - \exp\left(X_i\widehat{b}\right)\right)^2 \\ &= \frac{1}{N}\sum_{i=1}^{N} X_i' X_i \hat{u}_i^2.\end{aligned}$$

Par rapport au cas homogène, les écarts-types sont robustes à la distribution de comptage parce que la loi de Poisson appartient à la famille exponentielle linéaire à l'ordre 1.

Examinons maintenant la variance conditionnelle de notre variable expliquée. On suppose que le terme d'hétérogénéité $\varepsilon_i = \exp(u_i)$ suit une loi d'espérance 1 et de variance $\eta_i > 0$. La variance de la variable expliquée est donnée par :

$$\begin{aligned}\mathrm{E}(y_i|X_i) &= \underset{\varepsilon}{\mathrm{E}}\,\underset{y}{\mathrm{V}}(y_i|X_i,\varepsilon_i) + \underset{\varepsilon}{\mathrm{V}}\,\underset{y}{\mathrm{E}}(y_i|X_i,\varepsilon_i) \\ &= \underset{\varepsilon}{\mathrm{E}}\left(\exp(X_i b + u_i)|X_i\right) + \underset{\varepsilon}{\mathrm{V}}\left(\exp(X_i b + u_i)|X_i\right) \\ &= \underset{\varepsilon}{\mathrm{E}}\left(\exp(X_i b)\,\varepsilon_i|X_i\right) + \underset{\varepsilon}{\mathrm{V}}\left(\exp(X_i b)\,\varepsilon_i|X_i\right) \\ &= \exp(X_i b)\underbrace{\underset{\varepsilon}{\mathrm{E}}(\varepsilon_i)}_{1} + \exp(2X_i b)\underbrace{\underset{\varepsilon}{\mathrm{V}}(\varepsilon_i)}_{\eta_i} \\ &= \exp(X_i b)\left(1 + \eta_i \exp(X_i b)\right) \\ &= m_i\left(1 + \eta_i m_i\right).\end{aligned}$$

Ainsi, le modèle possède une variance supérieure à la moyenne et qui croît avec la moyenne.[2] Le maximum de vraisemblance du modèle homogène revient à supposer que $\eta_i = 0, \forall i$; le pseudo maximum de vraisemblance autorise n'importe quel profil de variance du terme d'hétérogénéité.

8.2 Le modèle binomial négatif

Bien que donnant un estimateur convergent, le modèle de Poisson ne donne pas forcément l'estimateur le plus efficace en présence d'hétérogénéité. Plusieurs approches sont possibles pour traiter ce problème. Premièrement, on peut postuler une loi pour le terme d'hétérogénéité et estimer le modèle par le maximum de vraisemblance; deuxièmement, on peut procéder à une estimation par le pseudo maximum de vraisemblance

[2] La relation précédente permet de voir que dans un modèle de Poisson homogène, la variance conditionnelle est toujours égale à l'espérance conditionnelle.

quasi généralisé à condition de choisir une loi ayant un paramètre de variance; troisièmement, on peut procéder à une estimation par le maximum de vraisemblance simulé. Le modèle binomial négatif peut être utilisé avec les deux premières approches.

8.2.1 Estimation par le maximum de vraisemblance

On dit qu'une variable aléatoire $y \in \mathbb{N}$ suit une loi binomiale négative de paramètres (r, p) quand elle admet pour distribution :

$$f(y) = \frac{\Gamma(y+r)}{\Gamma(y+1)\Gamma(r)} p^r (1-p)^y,\ 0 < p < 1 \text{ et } r > 0,$$

$$\text{ceci implique : } \mathrm{E}(y) = \frac{r(1-p)}{p} \quad \text{et} \quad \mathrm{V}(y) = \frac{r(1-p)}{p^2}.$$

Dans notre cas, nous souhaitons obtenir un modèle de Poisson hétérogène, ce qui impose la forme suivante pour les deux premiers moments :

$$\mathrm{E}(y) = m \text{ et } \mathrm{V}(y) = m(1+\eta m),$$

de sorte qu'il faut prendre (r, p) tels que :

$$m = \frac{r(1-p)}{p} \quad \text{et} \quad m(1+\eta m) = \frac{r(1-p)}{p^2}.$$

En divisant la variance par l'espérance, on obtient :

$$1 + \eta m = \frac{1}{p} \Leftrightarrow p = \frac{1}{1+\eta m} \in [0,1] \text{ car } \eta > 0,\ m > 0.$$

En utilisant la définition de l'espérance, on a :

$$r = m\frac{p}{1-p} = \frac{1}{\eta},$$

on peut donc réécrire la densité de la manière suivante :

$$f(y, m, \eta) = \frac{\Gamma(y+1/\eta)}{\Gamma(y+1)\Gamma(1/\eta)} \left(\frac{1}{1+\eta m}\right)^{1/\eta} \left(\frac{\eta m}{1+\eta m}\right)^y,\ \eta > 0, m > 0.$$

On retrouve donc les moments d'un modèle de Poisson hétérogène. Pour obtenir cette distribution, on fait les hypothèses suivantes :

1. Y suit une loi de Poisson d'espérance :

$$\widetilde{m}_i = m_i \varepsilon_i,$$

2. ε_i suit une loi Gamma de paramètres $(1/\eta, \eta)$ dont la densité est donnée par :[3]

$$g(\varepsilon_i) = \frac{\varepsilon_i^{1/\eta-1} \exp(-\varepsilon_i/\eta)}{\eta^{1/\eta}\Gamma(1/\eta)},$$

et de moments

$$\mathrm{E}(\varepsilon_i) = 1,\ \mathrm{V}(\varepsilon_i) = \eta.$$

Plus précisément, dans un modèle standard :

$$m_i = \exp(X_i b) \quad \text{et} \quad \varepsilon_i = \exp(u_i),$$

de sorte que :

$$\widetilde{m}_i = m_i \varepsilon_i = \exp(X_i b + u_i),$$

ce qui donne un modèle de Poisson avec hétérogénéité Log-Gamma car $u_i = \ln \varepsilon_i$.

La loi binomiale négative s'obtient de la manière suivante. La densité du modèle hétérogène, noté f^*, est égale à :

$$f^*(y|X, \varepsilon) = \frac{\exp(-\widetilde{m})\,\widetilde{m}^y}{y!} \text{ avec } \widetilde{m} = m\varepsilon, m = \exp(Xb).$$

Comme la variable aléatoire ε n'est pas observable, on intègre la densité précédente par rapport à cette distribution, afin d'obtenir la densité conditionelle de Y par rapport à X.

$$f(y|X) = \mathop{\mathrm{E}}_{\varepsilon}(f^*(y|X, \varepsilon)) = \int_0^{+\infty} f^*(y|X, u)\, g(\varepsilon)\, \mathrm{d}\varepsilon.$$

On obtient donc l'expression suivante :[4]

$$\begin{aligned} f(y|X) &= \int_0^{+\infty} \frac{\exp(-m\varepsilon)(m\varepsilon)^y}{y!} \frac{\varepsilon^{1/\eta-1}\exp(-\varepsilon/\eta)}{\eta^{1/\eta}\Gamma(1/\eta)} \mathrm{d}\varepsilon \\ &= \frac{m^y}{\Gamma(y+1)\,\eta^{1/\eta}\Gamma(1/\eta)} \int_0^{+\infty} \varepsilon^{y+1/\eta-1} \exp(-\varepsilon(m+1/\eta))\, \mathrm{d}\varepsilon \end{aligned}$$

[3] Dans le cas général, une variable aléatoire ε suit une loi gamma de paramètres (a, b), noté $\varepsilon \rightsquigarrow \Gamma(a, b)$ si elle vérifie :

$$f(\varepsilon) = \frac{\varepsilon^{a-1}e^{-\varepsilon/b}}{b^a\Gamma(a)},\ a, b, \varepsilon > 0,$$
$$\mathrm{E}(\varepsilon) = ab,\ \mathrm{V}(\varepsilon) = ab^2.$$

On rappelle également que :

$$\Gamma(a) = \int_0^{+\infty} x^{a-1}e^{-x}\mathrm{d}x.$$

[4] On utilise $y! = \Gamma(y+1)$.

arrivé à ce stade on effectue le changement de variable :

$$z = (m + 1/\eta)\,\varepsilon,$$

ce qui implique :

$$\varepsilon = \frac{\eta z}{\eta m + 1},\ \mathrm{d}\varepsilon = \frac{\eta \mathrm{d}z}{\eta m + 1},\ \lim_{\varepsilon \to 0} z = 0,\ \lim_{\varepsilon \to +\infty} z = +\infty,$$

d'où l'expression :

$$\begin{aligned} f\left(y|X\right) &= \frac{m^y}{\Gamma\left(y+1\right)\eta^{1/\eta}\Gamma\left(1/\eta\right)} \int_0^{+\infty} \left(\frac{\eta z}{\eta m + 1}\right)^{y+1/\eta-1} \exp\left(-z\right) \frac{\eta}{\eta m + 1} \mathrm{d}z \\ &= \frac{m^y}{\Gamma\left(y+1\right)\eta^{1/\eta}\Gamma\left(1/\eta\right)} \left(\frac{\eta}{\eta m + 1}\right)^{y+1/\eta-1} \frac{\eta}{\eta m + 1} \underbrace{\int_0^{+\infty} z^{y+1/\eta-1} \exp\left(-z\right) \mathrm{d}z}_{\Gamma(y+1/\eta)} \\ &= \frac{\Gamma\left(y+1/\eta\right)}{\Gamma\left(y+1\right)\Gamma\left(1/\eta\right)} \frac{1}{\eta^{1/\eta}} \left(\frac{\eta}{\eta m + 1}\right)^{1/\eta} \left(\frac{\eta m}{\eta m + 1}\right)^{y} \\ &= \frac{\Gamma\left(y+1/\eta\right)}{\Gamma\left(y+1\right)\Gamma\left(1/\eta\right)} \left(\frac{1}{\eta m + 1}\right)^{1/\eta} \left(\frac{\eta m}{\eta m + 1}\right)^{y} \end{aligned}$$

On utiliser cette densité pour estimer les paramètres par le maximum de vraisemblance.

8.2.2 Estimation par le pseudo maximum de vraisemblance quasi généralisé

La loi binomiale négative appartient à la famille exponentielle linéaire à l'ordre 1, ce qui permet d'obtenir un estimateur convergent dit de première étape. Cet estimateur sera utilisé pour estimer η, et l'on pourra ensuite obtenir l'estimateur du pseudo maximum de vraisemblance quasi-généralisé. Tout d'abord, vérifions que la loi binomiale négative appartient à la famille exponentielle linéaire à l'ordre 1 :

$$\ln f\left(y,m\right) = \underbrace{-\frac{1}{\eta}\ln\left(\eta m + 1\right)}_{A(m)} + \underbrace{\ln\frac{\Gamma\left(y+1/\eta\right)}{\Gamma\left(y+1\right)\Gamma\left(1/\eta\right)}}_{B(y)} + y\underbrace{\left(\ln\left(\eta m\right) - \ln\left(\eta m + 1\right)\right)}_{C(m)},$$

on vérifie facilement la variance :

$$\begin{aligned}\mathrm{V}(y) &= \left(\frac{\mathrm{d}C}{\mathrm{d}m}\right)^{-1} \\ &= \left(\frac{1}{m} - \frac{\eta}{\eta m + 1}\right)^{-1} \\ &= \left(\frac{1}{m(\eta m + 1)}\right)^{-1} \\ &= m(1 + \eta m).\end{aligned}$$

8.2.2.1 Estimateur de première étape

Pour cet estimateur on fixe librement la valeur de $\eta > 0$. Par exemple, on fixe une valeur qui simplifie l'expression de la log-vraisemblance, $\eta = 1$, mais l'estimateur obtenu a peu de chance d'être de bonne qualité. Ceci importe toutefois peu, car c'est l'estimateur de seconde étape qui nous intéresse. Si on fixe, par exemple, $\eta = 1$ on obtient un estimateur convergent en optimisant la pseudo log vraisemblance :[5]

$$f(y|X) = \frac{\Gamma(y+1)}{\Gamma(y+1)\Gamma(1)}\left(\frac{m}{m+1}\right)^{y}\left(\frac{1}{m+1}\right)^{1} = \frac{m^y}{(m+1)^{y+1}},$$

d'où une pseudo log vraisemblance à l'ordre 1 :

$$\ell_1(y|X,b) = \sum_{i=1}^{N}\left\{y_i \ln X_i b - (y_i+1)\ln\left(\exp(X_i b)+1\right)\right\}.$$

La maximisation de cette fonction fournit l'estimateur de première étape que l'on note $\widehat{b}$.

8.2.2.2 Estimateur de seconde étape

Il s'agit de l'étape qui fournit le meilleur estimateur. Dans un premier temps, il faut trouver un estimateur convergent de δ. Pour cela, on utilise l'expression de la variance conditionnelle de y_i. On enlève le conditionnement pour ne pas alourdir les notations.

$$\mathrm{V}(y_i) = m_i\left(1 + \frac{m_i}{\delta}\right) \Leftrightarrow \mathrm{E}\left((y_i - m_i)^2\right) = m_i + \eta m_i^2. \tag{8.1}$$

A partir de cette relation, on peut proposer deux estimateurs convergents :

[5] A priori, rien n'empêcherait de prendre l'estimateur du maximum de vraisemblance lui-même, puisque l'estimateur du PMV1 de b est convergent pour toutes les valeurs de η. Cette approche a l'avantage d'être moins arbitraire.

1. Le premier se base sur une réécriture de la relation (8.1) :

$$\mathrm{E}\left(\left(y_i - m_i\right)^2 - m_i\right) = \eta m_i^2,$$

et consiste à régresser $\left(y_i - \widehat{m}_i\right)^2 - \widehat{m}_i$ sur $\widehat{m}_i^2$ par les moindres carrés ordinaires sans terme constant, ce qui donne :

$$\eta_1 = \frac{\sum_{i=1}^N \left[\left(y_i - \widehat{m}_i\right)^2 - \widehat{m}_i\right] \widehat{m}_i^2}{\sum_{i=1}^N \widehat{m}_i^4}, \quad \widehat{m}_i = \exp\left(X\widehat{b}\right).$$

2. Un second estimateur est obtenu en réécrivant la relation (8.1) :

$$\mathrm{E}\left(\frac{1}{m_i^2}\left(y_i - m_i\right)^2 - m_i\right) = \eta \Leftrightarrow \eta = \mathrm{E}\left(\left(\frac{y_i}{m_i} - 1\right)^2 - \frac{1}{m_i}\right),$$

ce qui donne simplement :

$$\eta_2 = \frac{1}{N}\sum_{i=1}^N \left[\left(\frac{y_i}{\widehat{m}_i} - 1\right)^2 - \frac{1}{\widehat{m}_i}\right]$$

La pseudo vraisemblance quasi-généralisée est alors définie par :

$$\ell_{QG} = \sum_{i=1}^N \left\{-\frac{1}{\widehat{\eta}}\ln\left(\widehat{\eta}m_i + 1\right) + \ln\frac{\Gamma\left(y_i + 1/\widehat{\eta}\right)}{\Gamma\left(y_i + 1\right)\Gamma\left(1/\widehat{\eta}\right)} + y_i\left[\ln\left(\widehat{\eta}m_i\right) - \ln\left(\widehat{\eta}m_i + 1\right)\right]\right\}$$

pour l'optimisation, on peut éliminer tous les termes qui de ne dépendent pas de m_i, ce qui donne finalement :

$$\begin{aligned}\widetilde{\ell}_{QG} &= \sum_{i=1}^N \left\{y_i \ln m_i - \left(y_i + \frac{1}{\widehat{\eta}}\right)\ln\left(\widehat{\eta}m_i + 1\right)\right\} \\ &= \sum_{i=1}^N \left\{y_i X_i b - \left(y_i + \frac{1}{\widehat{\eta}}\right)\ln\left(\widehat{\eta}\exp\left(X_i b\right) + 1\right)\right\}.\end{aligned}$$

Le pseudo score, d'espérance nulle, est donné par :

$$\widetilde{s}_{QG} = \sum_{i=1}^N X_i'\left(y_i - \exp\left(X_i b\right)\frac{1 + \widehat{\eta}y_i}{1 + \widehat{\eta}\exp\left(X_i b\right)}\right),$$

et l'estimateur du PMVQG, $\overline{b}$, est défini par :

$$\sum_{i=1}^N X_i'\left(y_i - \exp\left(X_i\overline{b}\right)\frac{1 + \widehat{\eta}y_i}{1 + \widehat{\eta}\exp\left(X_i\overline{b}\right)}\right) = 0.$$

Cet estimateur est convergent, asymptotiquement normal et l'on estime sa matrice de covariance par :

$$\begin{aligned}\widehat{\mathrm{Vas}}\left(\overline{b}\right) &= \left[\widehat{\mathrm{J}_N}\left(\overline{b}\right)\right]^{-1} \\ &= \left[\sum_{i=1}^{N} X_i' X_i \left(\frac{\overline{m}_i}{1+\widehat{\eta}\overline{m}_i}\right)\right]^{-1} \quad \text{avec} \quad \overline{m}_i = \exp\left(X_i\overline{b}\right),\end{aligned}$$

en utilisant $\mathrm{E}\left(y_i|X_i\right) = m_i$.

8.3 Le modèle avec décision

Ce modèle généralise le modèle de Poisson en introduisant une forme explicite d'hétérogénéité. Tous les individus n'ont plus la même probabilité de rencontrer l'évènement étudié. L'observations d'une donnée de comptage fait donc apparaître deux types d'évènements nuls : ceux qui correspondent aux individus qui ne sont pas concernés par l'évènement étudié et ceux qui sont concernés mais qui n'ont pas rencontré l'évènement pendant la période étudiée. Le modèle comporte deux parties. La première partie est une relation de décision relative à l'évènement et se modélise par un modèle pour variable dichotomique. La seconde partie de cette relation donne le nombre d'évènements conditionnellemnet à la réalisation d'au moins un évènement et se modélise par un modèle de comptage. Le modèle latent qui représente la décision est donné par :

$$d_i^* = X_{1i}b_1 + u_i.$$

L'individu i entre dans le processus de comptage lorsque d_i^* lorsque $d_i^* > 0$. On a donc :

$$d_i = \begin{cases} 1 & \text{si } d_i^* > 0 \\ 0 & \text{sinon} \end{cases}.$$

Pour les individus qui sont entrés dans le processus de comptage le nombre de réalisations de l'évènement étudié z_i est distribué selon une loi de comptage $f\left(z_i\right)$. Cette loi est donnée par :

$$f\left(z_i|z_i \geq 1\right) = \Pr\left(d_i^* > 0\right) f\left(z_i\right).$$

On observe donc une réalisation nulle $y_i = 0$ soit lorsque $d_i = 0$ soit lorsque $z_i = 0$. Dans cette version simplifiée du modèle, on suppose que les perturbations u_i sont indépendantes du processus de comptage z_i, de sorte que :

$$\begin{aligned}\Pr\left[y_i = 0\right] &= \Pr\left[\left(d_i = 0\right) \cup \left(d_i = 1 \cap z_i = 0\right)\right] \\ &= \Pr\left(d_i^* \leq 0\right) + \underbrace{\left(1 - \Pr\left(d_i^* \leq 0\right)\right)}_{\Pr\left(d_i^* > 0\right)} \Pr\left(z_i = 0\right).\end{aligned}$$

Pour procéder à une estimation par le maximum de vraisemblance, il faut préciser les distributions suivies par u_i et z_i. Dans le modèle originel, la distribution de u_i est logistique et celle de z_i est de Poisson. On a donc :

$$\Pr\left[d_i^* \leq 0\right] = \frac{1}{1+\lambda_{1i}}, \Pr\left[d_i^* > 0\right] = \frac{\lambda_{1i}}{1+\lambda_{1i}},\ f\left(z_i\right) = \frac{\exp\left(-\lambda_{2i}\right)\lambda_{2i}^{z_i}}{z_i!},$$

avec

$$\lambda_{1i} = \exp\left(X_{1i}b_1\right).$$

La log-vraisemblance du modèle s'écrit donc :

$$\begin{aligned}
\ell_i &= \left(1-d_i\right)\ln\Pr\left[y_i=0\right] + d_i \ln f\left(y_i|y_i \geq 1\right) \\
&= \left(1-d_i\right)\ln\left[\Pr\left(d_i^* \leq 0\right) + \Pr\left(d_i^* > 0\right)\Pr\left(y_i=0\right)\right] \\
&+ d_i \ln\Pr\left(d_i^* > 0\right) f\left(y_i\right) \\
&= \left(1-d_i\right)\ln\left[\frac{1}{1+\lambda_{1i}} + \frac{\lambda_{1i}}{1+\lambda_{1i}}\exp\left(-\lambda_{2i}\right)\right] \\
&+ d_i\left[\ln\frac{\lambda_{1i}}{1+\lambda_{1i}} + \ln\frac{\exp\left(-\lambda_{2i}\right)\lambda_{2i}^{y_i}}{y_i!}\right] \\
&= \left(1-d_i\right)\left[\ln\left(1+\lambda_{1i}\exp\left(-\lambda_{2i}\right)\right) - \ln\left(1+\lambda_{1i}\right)\right] \\
&+ d_i\left[\ln\lambda_{1i} - \ln\left(1+\lambda_{1i}\right) - \lambda_{2i} + y_i\ln\lambda_{2i} - \ln y_i!\right] \\
&= \left(1-d_i\right)\ln\left(1+\lambda_{1i}\exp\left(-\lambda_{2i}\right)\right) \\
&+ d_i\left(\ln\lambda_{1i} - \lambda_{2i} + y_i\ln\lambda_{2i}\right) \\
&- \ln\left(1+\lambda_{1i}\right)
\end{aligned}$$

8.4 Le modèle avec saut

Il s'agit d'un modèle qui permet également de s'écarter de la proportion de réalisations nulles données par la loi de Poisson simple. On considère que le processus qui génère les réalisations nulles diffère de celui qui génère les réalisations positives. La première partie des données $y_i = 0$ est générée par une loi de Poisson de paramètre $\lambda_{1i} = \exp\left(X_{1i}b_1\right)$ et que la partie des données $y_i \geq 1$ est générée par une loi de Poisson de paramètre $\lambda_{2i} = \exp\left(X_{2i}b_2\right)$. La probabilité d'une réalisation nulle est donc égale à :

$$\Pr\left(y_i=0\right) = f_{1i}\left(0\right) = \exp\left(-\lambda_{1i}\right),$$

et celle d'une réalisation strictement positive est égale à :

$$f\left(y_i|y_i>0\right) = \frac{1-f_{1i}\left(0\right)}{1-f_{2i}\left(0\right)}f_2\left(y_i\right),$$

et l'on remarque que l'on a :

$$\begin{aligned} f_1(0) + \sum_{y=1}^{+\infty} \frac{1-f_1(0)}{1-f_2(0)} &= f_1(0) + \frac{1-f_1(0)}{1-f_2(0)} \sum_{y=1}^{+\infty} f_2(y) \\ &= f_1(0) + \frac{1-f_1(0)}{1-f_2(0)} (1-f_2(0)) \\ &= 1. \end{aligned}$$

d'où la log-vraisemblance :

$$\begin{aligned} \ell_i &= (1-d_i) \ln f_{1i}(0) + d_i \ln \frac{1-f_{1i}(0)}{1-f_{2i}(0)} f_2(y_i) \\ &= \underbrace{(1-d_i) \ln f_{1i}(0) + d_i \ln (1-f_{1i}(0))}_{\text{Partie Poissit}} + \underbrace{d_i \ln \frac{f_{2i}(y_i)}{1-f_{2i}(0)}}_{\text{Partie censurée}} \end{aligned}$$

avec $f_{ji}(0) = \exp(-\lambda_{ji})$, $j = 1, 2$.

La log-vraisemblance est séparable en deux parties indépendantes: la partie dichotomique ("Poissit") ne dépend que du paramètre b_1; la partie censurée ne dépend que du paramètre b_2. On peut donc réaliser deux optimisations séparées pour obtenir b_1 et b_2, les estimateurs du maximum de vraisemblance correspondants sont asymptotiquement indépendants.

La log-vraisemblance du modèle Poissit est donnée par :

$$\ell_{1i} = -(1-d_i) \exp(X_{1i} b_1) + d_i \ln [1 - \exp(-\exp(X_{1i} b_1))],$$

et celle du modèle de Poisson censuré par :

$$\ell_{2i} = d_i \{-\exp(X_{2i} b_2) + y_i X_{2i} b_2 - \ln y_i! - \ln [1 - \exp(-\exp X_{2i} b_2)]\}.$$

CHAPITRE 9

Les variables de durée

On rencontre des variables de durée dans de nombreux cas. A l'origine, les modèles ont été développés pour étudier la durée de vie mais d'autres applications ont été mises en oeuvre. En économie, on étudie la durée passée au chômage, dans un emploi ou entre deux emplois, la durée d'un trajet de transport, la durée de vie d'une entreprise ou encore la durée d'un crédit de type "revolving". Or les variables de durée ont des caractéristiques particulières : elles sont strictement positives et souffrent souvent de problèmes de censure. En effet, l'arrêt de la collecte à une date donnée (date d'arrêt de l'alimentation du fichier) fait que des durées commencées n'ont pas eu le temps de se terminer et sont donc censurées. On peut juste affecter une valeur minimale à ces durées observées de manière incomplète. On parle de censure linéaire droite. Inversement, il est possible que l'on commence le fichier à une date où le processus observé a déjà commencé pour certains individus, la durée est alors censurée à gauche. Pour obtenir une bonne estimation, il faut tenir compte de toutes les observations, censurées ou non. En effet, plus une durée est longue plus elle a de chances d'être censurée, de sorte qu'enlever les durées censurées revient à causer un biais de sélection. Par exemple, si l'on étudie la durée du chômage, enlever les données censurées reviendrait à réaliser une étude sans les chômeurs de longue durée, ce qui est difficilement envisageable.

Comme pour les variables aléatoires réelles, on définit la loi d'une variable de durée par sa fonction de répartition. Toutefois, on préfère pour des raisons pratiques, utiliser d'autres concepts plus parlants que la fonction de répartition ou la densité. Cette pratique provient de la démographie et utilise donc des concepts spécifiques comme le taux de mortalité, la probabilité de survie ou l'espérance de vie à la naissance. Nous allons montrer que ces concepts sont rigoureusement équivalents à ceux utilisés dans les autres branches de l'économétrie.

9.1 Terminologie

Considérons une variable aléatoire de durée $T > 0$. Sa fonction de répartition est définie par la probabilité que cette durée soit inférieure à une valeur donnée t :

$$F(t) = \Pr[T \leq t],\ t \in \mathbb{R}^{+*}.$$

Ce concept n'est pas toujours le plus pratique pour l'interprétation. L'économétrie des durées utilise, à la place, le concept de *fonction de survie* $S(t)$ qui donne la probabilité que la durée (de vie) soit supérieure à une valeur donnée t :

$$S(t) = \Pr[T > t] = 1 - F(t),$$

son nom vient de la démographie : elle donne la fraction d'individu d'une génération ayant survécu jusqu'à l'âge t.

La densité de la durée est donnée par :

$$f(t) = \frac{\mathrm{d}F(t)}{\mathrm{d}t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr[t < T \leq t + \Delta t],$$

elle représente l'intensité d'occurence d'une durée exactement égale à t. Cette intensité peut être supérieure à l'unité car il ne s'agit pas d'une probabilité mais d'une densité. La probabilité correspondante se calcule sur un petit intervalle de temps Δt, elle est donnée par :

$$\Pr[t < T \leq t + \Delta t] \simeq f(t)\,\Delta t.$$

Cette densité permet aussi de caractériser la loi de T car on en déduit la fonction de répartition de la manière suivante :

$$F(t) = \int_0^t f(x)\,\mathrm{d}x.$$

Mais la densité est également reliée à la fonction de survie par la relation :

$$f(t) = \frac{\mathrm{d}F(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(1 - S(t)) = -\frac{\mathrm{d}S(t)}{\mathrm{d}t}.$$

La *fonction de hasard* représente une occurence de mortalité instantanée. Comme pour la densité cette occurence peut être supérieure à l'unité. Elle est définie comme la probabilité conditionnelle de sortir (i.e. décéder) à la date t sachant que l'on vécu jusqu'à cette date. En effet, le taux de mortalité à la date t se calcule sur la population survivante à

cette date, et non sur toute la population. On obtient :

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\left[t < T \leq t + \Delta t | \, T > t\right] \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\Pr\left[(t < T \leq t + \Delta t) \cap (T > t)\right]}{\Pr\left[T > t\right]} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{\Pr\left[t < T \leq t + \Delta t\right]}{\Pr\left[T > t\right]} \\ &= \frac{1}{\Pr\left[T > t\right]} \underbrace{\lim_{\Delta t \to 0} \frac{1}{\Delta t} \Pr\left[t < T \leq t + \Delta t\right]}_{f(t)} \\ &= \frac{f(t)}{S(t)}. \end{aligned}$$

Ceci permet également de calculer l'équivalent du taux de mortalité instantané en démographie, sur un intervalle de temps Δt, qui est égal à :

$$\Pr\left[t < T \leq t + \Delta t | \, T > t\right] \simeq h(t) \, \Delta t.$$

Comme la densité et la fonction de répartition, la fonction de hasard caractérise la loi de la durée T :

$$h(t) = \frac{f(t)}{S(t)} = -\frac{1}{S(t)} \frac{\mathrm{d}S(t)}{\mathrm{d}t} = -\frac{\mathrm{d} \ln S(t)}{\mathrm{d}t},$$

La fonction de hasard peut également être employée pour calculer la fonction de survie :

$$\begin{aligned} \int_0^t h(x) \, \mathrm{d}x &= -\int_0^t \frac{\mathrm{d} \ln S(x)}{\mathrm{d}x} \mathrm{d}x \\ &= -\left[\ln S(x)\right]_0^t \\ &= -\ln S(t) + \ln S(0) \\ &= -\ln S(t), \end{aligned}$$

car pour une variable positive $S(0) = \Pr\left[T > 0\right] = 1$. On en déduit que :

$$S(t) = \exp\left\{-\int_0^t h(x) \, \mathrm{d}x\right\},$$

ce qui implique que l'on peut également écrire la densité en fonction du hasard :

$$\begin{aligned} f(t) &= h(t) \, S(t) \\ &= h(t) \exp\left\{-\int_0^t h(x) \, \mathrm{d}x\right\}. \end{aligned}$$

Enfin, remarquons une propriété qui peut s'avérer utile pour calculer l'espérance de la durée (i.e. l'espérance de vie à la naissance) :

$$\begin{aligned}\mathrm{E}(T) &= \int_0^{+\infty} x f(x)\,\mathrm{d}x \\ &= \int_0^{+\infty} -x\frac{\mathrm{d}S(x)}{\mathrm{d}x}\mathrm{d}x,\end{aligned}$$

en intégrant par partie $(u = -x,\ v' = S'(x))$, on obtient :

$$\begin{aligned}\mathrm{E}(T) &= \left[-xS(x)\right]_0^{+\infty} + \int_0^{+\infty} S(x)\,\mathrm{d}x \\ &= \int_0^{+\infty} S(x)\,\mathrm{d}x,\end{aligned}$$

sous l'hypothèse que :

$$\lim_{x\to+\infty} xS(x) = 0,$$

et cette hypothèse est généralement bien vérifiée car les fonctions de survie contiennent souvent des exponentielles. Cette formule peut être utile, mais il faut garder à l'esprit que dans la plupart des cas la méthode la plus simple est celle de la fonction génératrice des moments que nous présenterons plus loin.

9.2 Lois usuelles

A priori, toutes les lois applicables aux variables réelles positives peuvent être utilisées pour modéliser les variables de durée. C'est l'approche employée pour justifier l'utilisation de la loi log-normale. Cependant, les méthodes les plus employées (Weibull, Gamma, Gamma Généralisée, Cox) se basent sur des modèles dits à hasard proportionnels que nous définirons plus loi. Ces modèles possèdent l'avantage de permettre une modélisation directe de la fonction de hasard.

9.2.1 La loi exponentielle

Cette loi, la plus simple, vérifie la propriété forte *d'indépendance temporelle* de la fonction de hasard. Le taux de mortalité (i.e. de sortie) est constant dans le temps :

$$h(t) = h,\ \forall t,$$

cette hypothèse définit ce que l'on appelle un processus de Poisson (qui est également relié à la loi de Poisson dans le cas des données de comptage). En utilisant les propriétés de la section précédente, on retrouve les

différentes manières dont on peut caractériser la distribution :

$$\begin{aligned} S(t) &= \exp\left\{-\int_0^t h\mathrm{d}x\right\} \\ &= \exp\left\{-h\int_0^t \mathrm{d}x\right\} \\ &= \exp\left\{-h\left[x\right]_0^t\right\} \\ &= \exp\left(-ht\right), \\ F(t) &= 1 - S(t) = 1 - \exp\left(-ht\right), \\ f(t) &= h(t)\, S(t) = h\exp\left(-ht\right), \end{aligned}$$

et

$$\begin{aligned} \mathrm{E}(T) &= \int_0^{+\infty} \exp\left(-hx\right)\mathrm{d}x \\ &= \left[-\frac{1}{h}\exp\left(-hx\right)\right]_0^{+\infty} \\ &= \frac{1}{h}. \end{aligned}$$

Cette loi est surtout employée dans modèles d'économie théorique en raison de sa simplicité.

9.2.2 La loi de Weibull

Cette loi généralise la loi exponentielle en autorisant plusieurs type d'évolution de la fonction de hasard dans le temps, résumée dans le graphique 9.1. On remarque que ces évolutions restent toutefois monotones. On a :

$$h(t) = h\alpha t^{\alpha-1},$$

si $\alpha = 1$ on retrouve le modèle exponentiel mais, selon la valeur de α, le hasard peut être aussi bien croissant que décroissant avec la durée.

En utilisant les propriétés de la première section, on obtient les caractéristiques suivantes de la distribution de Weibull :

$$\begin{aligned} S(t) &= \exp\left\{-\int_0^t h(x)\,\mathrm{d}x\right\} \\ &= \exp\left\{-\int_0^t h\alpha x^{\alpha-1}\mathrm{d}x\right\} \\ &= \exp\left\{-h\left[x^{\alpha}\right]_0^t\right\} \\ &= \exp\left(-ht^{\alpha}\right), \end{aligned}$$

Figure 9.1: Fonction de hasard de la loi de Weibull

la fonction de répartition est donnée par :

$$F(t) = 1 - \exp(-ht^{\alpha}),$$

la densité par :

$$\begin{aligned} f(t) &= h(t) S(t) \\ &= h\alpha t^{\alpha-1} \exp(-ht^{\alpha}), \end{aligned}$$

et on peut également calculer l'espérance en utilisant la fonction de survie :

$$\mathrm{E}(T) = \int_0^{+\infty} \exp(-hx^{\alpha})\, \mathrm{d}x,$$

on fait le changement de variable :

$$z = hx^{\alpha} \Leftrightarrow x = \left(\frac{z}{h}\right)^{1/\alpha} \Rightarrow \mathrm{d}x = \frac{1}{\alpha h^{1/\alpha}} z^{1/\alpha - 1} \mathrm{d}z,$$

et les bornes restent inchangées :[1]

$$\begin{aligned}\mathrm{E}(T) &= \frac{1}{\alpha h^{1/\alpha}} \int_0^{+\infty} z^{1/\alpha - 1} \exp(-z)\, \mathrm{d}z \\ &= \frac{\Gamma(1/\alpha)}{\alpha h^{1/\alpha}} \\ &= h^{-1/\alpha} \Gamma(1+\alpha).\end{aligned}$$

Cette loi est une des plus employées dans les applications économétriques.

9.2.3 La loi Gamma généralisée

Cette loi généralise la loi de Weibull, en introduisant un paramètre supplémentaire, qui permet d'obtenir une fonction de hasard non monotone. On la définit par sa densité :

$$f(t) = \frac{\alpha h^{\beta} t^{\alpha\beta - 1} \exp(-h t^{\alpha})}{\Gamma(\beta)}.$$

On retrouve la densité de la loi de Weibull pour $\beta = 1$ et celle de la loi exponentielle pour $\alpha = 1$ et $\beta = 1$. Pour les autres fonctions, nous aurons besoin des fonctions Gamma tronquées.[2] On note :

$$\begin{aligned}\underline{\Gamma}(a, x) &= \int_0^{x} u^{\alpha - 1} e^{-u} \mathrm{d}u, \\ \overline{\Gamma}(a, x) &= \int_x^{+\infty} u^{\alpha - 1} e^{-u} \mathrm{d}u,\end{aligned}$$

et l'on remarque que :

$$\begin{aligned}\lim_{x \to +\infty} \underline{\Gamma}(a, x) &= \Gamma(a), \\ \lim_{x \to 0} \overline{\Gamma}(a, x) &= \Gamma(a), \\ \underline{\Gamma}(a, x) + \overline{\Gamma}(a, x) &= \Gamma(a).\end{aligned}$$

En règle générale on évalue les fonctions Gamma tronquées numériquement. On en déduit la fonction de répartition de la variable de durée de la manière suivante :

$$F(t) = \frac{\alpha h^{\beta}}{\Gamma(\beta)} \int_0^{t} u^{\alpha\beta - 1} \exp(-h u^{\alpha})\, \mathrm{d}u,$$

[1] On rappelle que $\Gamma(x+1) = x\Gamma(x)$.
[2] En anglais : "incomplete Gamma functions".

on effectue le changement de variable $v = hu^{\alpha}$, ce qui implique $u = h^{-1/\alpha}v^{1/\alpha}$ et $\mathrm{d}u = h^{-1/\alpha}\alpha^{-1}v^{1/\alpha-1}\mathrm{d}v$, et les bornes d'intégration deviennent 0 et ht^{α} :

$$\begin{aligned} F(t) &= \frac{\alpha h^{\beta}}{\Gamma(\beta)} \int_0^{ht^{\alpha}} v^{\beta-1/\alpha} h^{1/\alpha-\beta} \exp(-v)\, h^{-1/\alpha}\alpha^{-1}v^{1/\alpha-1} dv \\ &= \frac{1}{\Gamma(\beta)} \int_0^{ht^{\alpha}} v^{\beta-1} \exp(-v)\, \mathrm{d}v \\ &= \frac{\underline{\Gamma}(\beta, ht^{\alpha})}{\Gamma(\beta)}, \end{aligned}$$

on vérifie que lorsque $t \to 0$, $\underline{\Gamma}(\beta, 0) = \Gamma(0) = 0$ de sorte que $F(0) = 0$, et que $\lim_{t\to+\infty} \underline{\Gamma}(\beta, t) = \Gamma(\beta)$, de sorte que $F(t) \to 1$. La fonction de survie est donc donnée par :

$$S(t) = 1 - \frac{\underline{\Gamma}(\beta, ht^{\alpha})}{\Gamma(\beta)} = \frac{\overline{\Gamma}(\beta, ht^{\alpha})}{\Gamma(\beta)}.$$

Ceci ne permet pas d'obtenir de forme explicite pour la fonction de hasard parce que :

$$h(t) = \frac{f(t)}{S(t)} = \frac{\alpha h^{\beta} t^{\alpha\beta-1} \exp(-ht^{\alpha})}{\overline{\Gamma}(\beta, ht^{\alpha})}.$$

Le nom de cette distribution vient du fait qu'elle généralise la loi Gamma $\gamma(\beta, h)$, qui correspond au cas $\alpha = 1$. Pour l'espérance mathématique, le plus simple est de recourir à la fonction génératrice des moments, calculée plus loin.

9.2.4 La loi log-normale

Notons dès maitenant que cette distribution n'est pas reliée aux précédentes. Elle consiste à supposer directement que le logarithme de la variable de durée $\ln T$ suit une loi normale $\mathrm{N}\left(m, \sigma^2\right)$. Sa densité est donc donnée par :

$$f(t) = \frac{1}{\sigma t}\varphi\left(\frac{\ln t - m}{\sigma}\right),$$

où $\varphi(t) = (2\pi)^{-1/2}\exp\left(-t^2/2\right)$ est la densité de loi normale centrée et réduite. Sa fonction de répartition est donc donnée par :

$$F(t) = \int_0^t f(x)\,\mathrm{d}x = \int_0^t \frac{1}{\sigma x}\varphi\left(\frac{\ln x - m}{\sigma}\right)\mathrm{d}x,$$

on remarque alors que $1/(\sigma x)$ est la dérivée par rapport à x de $(\ln x - m)/\sigma$, de sorte que :

$$F(t) = \left[\Phi\left(\frac{\ln x - m}{\sigma}\right)\right]_0^t = \Phi\left(\frac{\ln t - m}{\sigma}\right),$$

car :

$$\lim_{x\to 0} \Phi\left(\frac{\ln x - m}{\sigma}\right) = 0.$$

On en déduit :

$$S(t) = 1 - \Phi\left(\frac{\ln t - m}{\sigma}\right),$$

ainsi que :

$$h(t) = \frac{f(t)}{S(t)} = \frac{1}{\sigma t}\frac{\varphi\left(\frac{\ln t - m}{\sigma}\right)}{1 - \Phi\left(\frac{\ln t - m}{\sigma}\right)}.$$

Ceci implique que le hasard admet obligatoirement un maximum. En effet,

$$\begin{aligned}
\sigma h'(t) &= -\frac{1}{t^2}\frac{\varphi(u)}{1-\Phi(u)} + \frac{1}{t}\frac{\frac{1}{\sigma t}\varphi'(u)(1-\Phi(u)) - \varphi(u)\left(-\frac{1}{\sigma t}\right)\varphi(u)}{(1-\Phi(u))^2} \\
&= \frac{1}{\sigma t^2 (1-\Phi(u))^2} \times \\
&\left(-\sigma(1-\Phi(u))\varphi(u) - u\varphi(u)(1-\Phi(u)) + \varphi(u)^2\right) \\
&= \frac{\varphi(u)}{\sigma t^2 (1-\Phi(u))^2}\left(-\sigma(1-\Phi(u)) - u(1-\Phi(u)) + \varphi(u)\right) \\
&= \frac{\varphi(u)}{\sigma t^2 (1-\Phi(u))^2}\left(-(\sigma+u)(1-\Phi(u)) + \varphi(u)\right),
\end{aligned}$$

on pose :

$$\kappa(u) = \frac{\varphi(u)}{1-\Phi(u)},$$

cette fonction est définie par analogie avec la fonction de hasard mais sur une loi normale centrée réduite. Notons qu'il ne s'agit pas à proprement parler d'une fonction de hasard parce que u peut prendre des valeurs négatives. Cette fonction est strictement croissante, comme le montre le graphique 9.2 :

On peut écrire :

$$h'(t) = \frac{1}{\sigma^2 t^2}\kappa(u)\left(\kappa(u) - (\sigma+u)\right),$$

Figure 9.2: $\varphi(x) / (1 - \Phi(x))$

le hasard atteint son maximum à un point $u = (\ln t - m) / \sigma$ tel que :

$$\kappa(u) = \sigma + u.$$

Ce type de profil est très particulier, car la présence d'un maximum est imposée, et cette hypothèse ne convient pas forcément à tous les processus de durée. Il faut donc être vigilant quand on l'emploie.

9.3 Modélisation en logarithmes

Les variables de durée peuvent toujours être prises en logarithmes, ce qui facilite l'interprétation des résultats quand les variables explicatives sont elles-mêmes en logarithmes ou sous forme d'incatrices. Mais cette modélisation peut également être utilisée pour mieux comprendre les relations entre les différentes loi usuelles, et notamment les loi exponentielles, de Weibull, Gamma et Gamma généralisée.

9.3.1 Rappels

9.3.1.1 Le changement de variable

Nous allons utiliser cette propriété dans toute la section. Supposons que l'on dispose d'une variable de durée T de densité $f_T(t)$ et que l'on effectue un changement de variable $U = g(T)$, la densité de la variable u est donnée par :

$$f_U(u) = \left| \frac{\mathrm{d}g^{-1}(u)}{\mathrm{d}u} \right| f_T\left(g^{-1}(u)\right).$$

9.3.1.2 La loi Gamma

Une variable aléatoire X suit une loi Gamma de paramètres (a, b), notée $\gamma(a, b)$ si sa densité s'écrit :

$$f_X(x) = \frac{b^a x^{a-1} \exp(-bx)}{\Gamma(a)},\ x > 0,\ a > 0,\ b > 0,$$

les deux premiers moments sont égaux à $\mathrm{E}(X) = a/b$ et $\mathrm{V}(X) = a/b^2$. Si $a = 1$, on retrouve la loi exponentielle, notée $\gamma(1, b)$, dont la densité est égale à :

$$f_X(x) = b \exp(-bx),$$

et dont les deux premiers moments sont égaux à $\mathrm{E}(X) = 1/b$ et $\mathrm{V}(X) = 1/b^2$. Si on prend le cas symétrique, une loi Gamma $\gamma(a, 1)$ on obtient la densité :

$$f_X(x) = \frac{x^{a-1} \exp(-x)}{\Gamma(a)},$$

dont les deux premiers moments sont égaux à $\mathrm{E}(X) = \mathrm{V}(X) = a$.

9.3.2 Modèle exponentiel et loi de Gumbel

Posons le modèle en logarithmes suivant :

$$\ln T = -\ln h + U,$$

où U est une variable aléatoire dont on cherche la loi. On sait seulement que la durée T suit une loi exponentielle $\gamma(1, h)$ de densité :

$$f_T(t) = h \exp(-ht).$$

Pour trouver la densité de la loi de U, on remarque que :

$$U = \ln(hT) = g(T),$$

de sorte que :

$$T = \frac{1}{h} \exp(U) = g^{-1}(U)$$

ce qui implique :

$$\frac{\mathrm{d}g^{-1}(u)}{\mathrm{d}u} = \frac{1}{h}\exp(u),$$

d'où la densité :

$$\begin{aligned} f_U(u) &= \frac{1}{h}\exp(u)\, h \exp\left[-h\left(\frac{1}{h}\exp(u)\right)\right] \\ &= \exp(u)\exp(-\exp(u)), \end{aligned}$$

qui n'est autre que la densité d'une loi de Gumbel (i.e. valeur extrême de type I, minimum) d'espérance égale à l'opposée de la constante d'Euler, $\mathrm{E}(U) = -\gamma_E$, avec $\gamma_E \simeq 0,57721$, , et de variance $\pi^2/6$. Pour trouver directement ces résultats on peut utiliser les deux propriétés suivantes de la constante d'Euler :[3]

$$\gamma_E = -\int_0^{+\infty} (\ln x)\, e^{-x}\mathrm{d}x,$$

et

$$\gamma_E^2 + \frac{\pi^2}{6} = \int_0^{+\infty} (\ln x)^2\, e^{-x}\mathrm{d}x.$$

On utilise également les propriétés suivantes de la fonction Gamma :

$$\Gamma(p) = \int_0^{+\infty} x^{p-1}e^{-x}\mathrm{d}x \Rightarrow \Gamma'(p) = \int_0^{+\infty} (\ln x)\, x^{p-1}e^{-x}\mathrm{d}x, \qquad (9.1)$$

car on dérive par rapport à p et non par rapport à x.[4] Ceci implique :

$$\Gamma'(1) = \int_0^{+\infty} (\ln x)\, e^{-x}\mathrm{d}x = -\gamma_E.$$

En dérivant une nouvelle fois la relation (9.1) par rapport à p, on obtient :

$$\Gamma''(p) = \int_0^{+\infty} (\ln x)^2\, x^{p-1}e^{-x}\mathrm{d}x,$$

[3] Pour évaluer la constante d'Euler, on peut utiliser la définition sous forme de série donnée à l'origine par Euler lui-même :

$$\gamma_E = \sum_{k=1}^{+\infty}\left[\frac{1}{k} - \ln\left(1+\frac{1}{k}\right)\right].$$

[4] On a :

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}p}\left(x^{p-1}\right) &= \frac{\mathrm{d}}{\mathrm{d}p}\left(e^{(p-1)\ln x}\right) = (\ln x)\, e^{(p-1)\ln x} \\ &= (\ln x)\, x^{p-1}. \end{aligned}$$

ce qui implique :

$$\Gamma\text{"}\,(1) = \gamma_E^2 + \frac{\pi^2}{6},$$

Nous retrouverons les valeurs $\Gamma'(1)$ et $\Gamma\text{"}\,(1)$ lors de l'étude de la fonction génératrice des moments de la loi de Gumbel. Finalement, on peut réécrire le modèle exponentiel sous la forme :

$$\begin{aligned}\mathrm{E}(\ln T) &= -\ln h + \mathrm{E}(U) \\ &= -(\ln h + \gamma_E),\end{aligned}$$

de sorte qu'en mettant les variables explicatives dans la fonction de hasard, on peut aboutir à un modèle log-linéaire avec une simple correction pour le terme constant du modèle. On remarque également que plus le taux de hasard h est élevé, plus l'espérance de durée est faible.

9.3.3 Modèle exponentiel et loi exponentielle

On peut également définir le modèle exponentiel directement en niveaux et non en logarithmes. C'est l'approche qui est suivie habituellement pour généraliser ce modèle vers les modèles de Weibull, Gamma et Gamma généralisé. De manière cohérente avec la section précédente, on pose :

$$T = g(V) = \frac{V}{h}, \tag{9.2}$$

où V suit une loi exponentielle de paramètre 1, notée $\gamma(1,1)$, de densité :

$$f_V(v) = \exp(-v).$$

La densité de la variable de durée tirée de ce modèle est donnée par la transformation :

$$\begin{aligned}T = \frac{V}{h} \Leftrightarrow V = hT = g^{-1}(T) \\ \Rightarrow \frac{\mathrm{d}g^{-1}(t)}{\mathrm{d}t} = h,\end{aligned}$$

d'où la densité :

$$f_T(t) = \left|\frac{\mathrm{d}g^{-1}(t)}{\mathrm{d}t}\right| f_V\left(g^{-1}(t)\right) = h\exp(-ht),$$

qui correspond à la densité de la loi exponentielle $\gamma(1,h)$.

9.3.4 Modèle de Weibull

Il existe différentes manières de généraliser le modèle exponentiel (9.2). Une première manière consiste à introduire un paramètre d'échelle $\alpha > 0$ dans la définition de la variable de durée :[5]

$$T = g(V) = \left(\frac{V}{h}\right)^{1/\alpha}, \tag{9.3}$$

on retrouve le modèle exponentiel pour $\alpha = 1$. On suppose toujours que V suit une loi exponentielle $\gamma(1,1)$. La loi suivie par T a donc changé puisque l'on a :

$$\begin{aligned} T = \left(\frac{V}{h}\right)^{1/\alpha} &\Leftrightarrow V = hT^{\alpha} = g^{-1}(T) \\ &\Rightarrow \frac{\mathrm{d}g^{-1}(t)}{\mathrm{d}t} = h\alpha t^{\alpha-1}, \end{aligned}$$

de sorte que la densité de T s'écrit :

$$\begin{aligned} f_T(t) &= \left|\frac{\mathrm{d}g^{-1}(t)}{\mathrm{d}t}\right| f_V\left(g^{-1}(t)\right) \\ &= \alpha h t^{\alpha-1} \exp\left(-ht^{\alpha}\right), \end{aligned}$$

qui correspond à la densité d'une variable de Weibull. On remarque qu'en logarithme la relation peut s'écrire :

$$\begin{aligned} \ln T &= \frac{1}{\alpha}(-\ln h + \ln V) \\ &= \frac{1}{\alpha}(-\ln h + U), \end{aligned}$$

de sorte qu'avec nos notations α est un paramètre d'échelle qui porte sur l'ensemble du modèle. L'espérance mathématique correspondante s'écrit :

$$\mathrm{E}(\ln T) = -\frac{1}{\alpha}(\ln h + \gamma_E),$$

car U suit toujours une loi $\gamma(1,1)$ comme dans le modèle exponentiel.

9.3.5 Modèle Gamma

Le modèle Gamma généralise le modèle exponentiel (9.2) en changeant la distribution du terme d'erreur V au lieu d'introduire un paramètre d'échelle. On suppose, comme dans le modèle exponentiel, que :

$$T = g(V) = \frac{V}{h},$$

[5] Les notations utilisées ici sont un peu différentes de celles utilisées habituellement, c'est pour pouvoir simplifier les expressions qui apparaîtront dans la suite du chapitre.

mais cette fois-ci V suit une loi Gamma $\gamma(\beta, 1)$. La densité de V est donc donnée par :

$$f_V(v) = \frac{v^{\beta-1}e^{-v}}{\Gamma(\beta)},\ \beta > 0,\ v > 0,$$

et l'on retrouve le modèle exponentiel en posant $\beta = 1$. En utilisant $g^{-1}(t) = ht$, on obtient :

$$\begin{aligned} f_T(t) &= \left|\frac{\mathrm{d}g^{-1}(t)}{\mathrm{d}t}\right| f_V\left(g^{-1}(t)\right) \\ &= h\frac{(ht)^{\beta-1}e^{-ht}}{\Gamma(\beta)} \\ &= \frac{h^{\beta}t^{\beta-1}e^{-ht}}{\Gamma(\beta)}, \end{aligned}$$

qui correspond à la densité d'une loi Gamma $\gamma(\beta, h)$. Ecrit en espérance le modèle log linéaire donne :

$$\mathrm{E}(\ln T) = -\ln h + \mathrm{E}(\ln V),$$

et nous calculerons plus loin l'espérance de $\ln V$ à partir de sa fonction génératrice des moments.

9.3.6 Modèle Gamma généralisé

Il s'agit d'une troisième généralisation du modèle exponentiel (9.2). Cette fois-ci, nous allons combiner les deux généralisations du modèle de Weibull et du modèle Gamma. On suppose, comme dans le modèle de Weibull, que la variable de durée est définie par la relation :

$$T = g(V) = \left(\frac{V}{h}\right)^{1/\alpha},$$

et, comme dans le modèle Gamma, que V suit une loi Gamma $\gamma(\beta, 1)$ de densité :

$$f_V(v) = \frac{v^{\beta-1}e^{-v}}{\Gamma(\beta)},\ \beta > 0,\ v > 0.$$

On trouve directement la nouvelle densité :

$$\begin{aligned} f_T(t) &= \left|\frac{\mathrm{d}g^{-1}(t)}{\mathrm{d}t}\right| f_V\left(g^{-1}(t)\right) \\ &= h\alpha t^{\alpha-1}\frac{(ht^{\alpha})^{\beta-1}e^{-(ht^{\alpha})}}{\Gamma(\beta)} \\ &= \frac{\alpha h^{\beta}t^{\alpha\beta-1}\exp(-ht^{\alpha})}{\Gamma(\beta)}. \end{aligned}$$

On note cette distribution $\gamma(\beta, h, \alpha)$. On remarque que la distribution exponentielle s'obtient pour $\gamma(1, h, 1)$, la distribution de Weibull $\gamma(1, h, \alpha)$ et la distribution Gamma pour $\gamma(\beta, h, 1)$. Le nom de cette distribution vient du fait qu'elle généralise la loi Gamma puisque, pour $\alpha = 1$, on obtient :

$$f_T(t) = \frac{h^\beta t^{\beta-1} \exp(-ht)}{\Gamma(\beta)},$$

la densité de la loi Gamma à deux paramètres $\gamma(\beta, h)$. Le nom est toutefois trompeur, puisque la loi Gamma généralisée généralise également la loi de Weibull.

Plus généralement, on obtient les cas particuliers suivants :

- $\alpha = 1 : T \rightsquigarrow \gamma(\beta, h, 1)$. Loi Gamma;
- $\beta = 1 : T \rightsquigarrow \gamma(1, h, \alpha)$. Loi de Weibull;
- $\alpha = 1$ et $\beta = 1 :\ T \rightsquigarrow \gamma(1, h, 1)$. Loi exponentielle;
- $\alpha \neq 1$ et $\beta \neq 1 :\ T \rightsquigarrow \gamma(\beta, h, \alpha)$ .Loi Gamma généralisée.

Le modèle log linéaire en espérance s'écrit maintenant :

$$\mathrm{E}(\ln T) = \frac{1}{\alpha}(-\ln h + \mathrm{E}(\ln V)),$$

où $\mathrm{E}(\ln V)$ prend la même valeur que pour le modèle Gamma de la section précédente.

9.3.7 Modèle log-normal

On peut également utiliser la méthode du changement de variable pour le modèle log-normal, mais ici la modélisation ne porte pas sur le hasard mais sur l'espérance mathématique de la variable de durée, quantité qui est décroissante avec le taux de hasard. Dans un modèle avec des variables explicatives, un modèle basé sur l'espérance de la variable de durée implique généralement un changement de signe des coefficients par rapport à un modèle basé sur la fonction de hasard. On pose :

$$\ln T = m + \sigma U$$

où U suit une loi normale centrée et réduite, de sorte que $\ln T$ suit une loi normale $\mathrm{N}(m, \sigma^2)$. On a donc la transformation suivante :

$$T = \exp(m + \sigma U) = g(U) \Leftrightarrow U = \frac{\ln T - m}{\sigma} = g^{-1}(T),$$

ce qui implique :

$$\frac{\mathrm{d}g^{-1}(t)}{\mathrm{d}t} = \frac{1}{\sigma t}.$$

D'autre part, U suit une loi normale de densité $\varphi(u)$, ce qui implique que la densité de T (i.e. de la loi log-normale) est donnée par :

$$f_T(t) = \left|\frac{\mathrm{d}g^{-1}(t)}{\mathrm{d}t}\right| \varphi\left(g^{-1}(t)\right) = \frac{1}{\sigma t}\varphi\left(\frac{\ln t - m}{\sigma}\right).$$

Ici, par construction, l'espérance du modèle log linéaire se réduit à sa plus simple expression :

$$\mathrm{E}(\ln T) = m + \sigma\,\mathrm{E}(U) = m.$$

9.4 Calcul des moments

Lors des estimations, on peut avoir besoin de faire une prévision de la durée moyenne passée dans l'état étudié, ainsi que de calculer la variance pour obtenir un intervalle de confiance à partir de la méthode de Slutsky. La méthode la plus simple avec les distributions qui précèdent est, souvent, d'utiliser les fonctions génératrices des moments.

9.4.1 Fonction génératrice des moments

9.4.1.1 Moments non centrés

La fonction génératrice des moments non centrés d'une variable aléatoire T est définie par :

$$M(s) = \mathrm{E}\left(e^{sT}\right) = \int_0^{+\infty} e^{st} f_T(t)\,\mathrm{d}t,$$

quand cette intégrale converge. On remarque que :

$$M(0) = \mathrm{E}\left(e^0\right) = 1.$$

On vérifie que :

$$M'(s) = \mathrm{E}\left(Te^{sT}\right),$$

et, par récurrence, que :

$$\frac{\mathrm{d}^k M}{\mathrm{d}s^k}(s) = \mathrm{E}\left(T^k e^{sT}\right).$$

En prenant la quantité précédente en $s = 0$, on obtient :

$$\frac{\mathrm{d}^k M}{\mathrm{d}s^k}(0) = \mathrm{E}\left(T^k\right).$$

On peut donc obtenir les moments non centrés par dérivation au lieu de procéder par intégration, ce qui est plus rapide.

9.4.1.2 Moments centrés

De la même manière, on peut obtenir certains moments centrés par la fonction :

$$K(s) = \ln M(s),$$

on voit que :

$$K'(s) = \frac{M'(s)}{M(s)},$$

en prenant la fonction précédente au point $s = 0$ on a :

$$K'(0) = \mathrm{E}(T),$$

en dérivant la fonction K une deuxième fois, on obtient:

$$K"(s) = \frac{M"(s)}{M(s)} - \left(\frac{M'(s)}{M(s)}\right)^2,$$

en la prenant au point $s = 0$:

$$K"(0) = \mathrm{E}\left(T^2\right) - \mathrm{E}(T)^2 = \mathrm{V}(T).$$

9.4.1.3 Moments du logarithme

Enfin, en économétrie des durées, on a souvent besoin des fonctions génératrices des moments du logarithme d'une variable de durée, parce ce que beaucoup de modèles peuvent s'écrire en logarithmes (e.g., Exponentiel, Weibull, Gamma, Gamma généralisé et log-Normal) :

$$M_{\ln T}(s) = \mathrm{E}\left(e^{s \ln T}\right) = \mathrm{E}\left(T^s\right),$$

dont l'intérêt est ici évident. Pour calculer la variance du logarithme d'une variable de durée, on utilsera la fonction correspondante des moments centrés :

$$K_{\ln T}(s) = \ln K_{\ln T}(s) = \ln \mathrm{E}\left(T^s\right).$$

9.4.2 Moments des lois usuelles

9.4.2.1 Loi exponentielle

La densité est donnée par :

$$f(t) = h \exp(-ht),\ t > 0, h > 0,$$

on a donc :

$$\begin{aligned} M(s) &= \mathrm{E}\left(e^{sT}\right) \\ &= h\int_0^{+\infty} e^{-(h-s)t}\mathrm{d}t \\ &= h\left[-\frac{1}{h-s}e^{-(h-s)t}\right]_0^{+\infty} \\ &= \frac{h}{h-s}, \end{aligned}$$

remarquons bien ici que l'on a choisi de mettre dans l'exponentielle un terme en $h-s$ car on utilise cette fonction en $s=0$, ce qui garantit que $h-s>0$ et donc la convergence de l'intégrale.

On en déduit :

$$\begin{aligned} M'(s) &= \frac{h}{(h-s)^2} \Rightarrow \mathrm{E}(T) = M'(0) = \frac{1}{h}, \\ M"(s) &= \frac{2h}{(h-s)^3} \Rightarrow \mathrm{E}\left(T^2\right) = M"(0) = \frac{2}{h^2}, \end{aligned}$$

on pourrait en déduire la variance par la formule classique :

$$\begin{aligned} \mathrm{V}(T) &= \mathrm{E}\left(T^2\right) - \mathrm{E}(T)^2 \\ &= 2/h^2 - (1/h)^2 \\ &= 1/h^2, \end{aligned}$$

mais on peut l'obtenir plus directement par la fonction génératrice des moments centrés. Elle est définie par :

$$K(s) = \ln h - \ln(h-s).$$

On en déduit :

$$\begin{aligned} K'(s) &= \frac{1}{h-s} \Rightarrow K'(0) = \mathrm{E}(T) = \frac{1}{h}, \\ K"(s) &= \frac{1}{(h-s)^2} \Rightarrow K"(0) = \mathrm{V}(T) = \frac{1}{h^2}. \end{aligned}$$

Les fonctions génératrices correspondant au logarithme de la durée $\ln T$ sont celles de la loi de Gumbel données dans la section suivante.

9.4.2.2 Loi de Gumbel

On peut calculer la fonction génératrice de la loi de Gumbel en remarquant qu'une variable de ce type s'obtient comme le logarithme d'une

variable exponentielle $\gamma(1,1)$ et en utilisant la propriété :

$$K_{\ln T}(s) = \mathrm{E}(T^s).$$

Il suffit donc de calculer le moment d'ordre s de la loi exponentielle. En fait, dans ce cas particulier, il n'y a pas de calcul à faire, puisque l'on a :

$$\mathrm{E}(T^s) = \int_0^{+\infty} t^s \exp(-t)\,\mathrm{d}t = \Gamma(1+s),$$

on en déduit la fonction génératrice des moments centrés du logarithme de la loi exponentielle :

$$K_{\ln T}(s) = \ln \Gamma(1+s)$$

ce qui implique :

$$\begin{aligned} K'_{\ln T}(s) &= \frac{\Gamma'(1+s)}{\Gamma(1+s)} \\ &\Rightarrow K'_{\ln T}(0) = \mathrm{E}(\ln T) = \Gamma'(1) = -\gamma_E, \end{aligned}$$

où γ_E est la constante d'Euler. De même, on voit que :

$$K^{"}_{\ln T}(s) = \frac{\Gamma"(1+s)}{\Gamma(1+s)} - \left(\frac{\Gamma'(1+s)}{\Gamma(1+s)}\right)^2,$$

de sorte que :

$$\begin{aligned} K^{"}_{\ln T}(0) &= \mathrm{V}(\ln T) \\ &= \Gamma"(1) - \Gamma'(1)^2 \\ &= \gamma_E^2 + \frac{\pi^2}{6} - (-\gamma_E)^2 \\ &= \frac{\pi^2}{6}. \end{aligned}$$

9.4.2.3 Loi Gamma

La fonction génératrice de la loi Gamma $\gamma(\beta,1)$ est définie par :

$$\begin{aligned} M_T(s) &= \mathrm{E}\left(e^{sT}\right) \\ &= \int_0^{+\infty} e^{st} \frac{t^{\beta-1}e^{-t}}{\Gamma(\beta)}\mathrm{d}t \\ &= \int_0^{+\infty} \frac{t^{\beta-1}e^{-(1-s)t}}{\Gamma(\beta)}\mathrm{d}t, \end{aligned}$$

on effectue donc le changement de variable $x = (1-s)\,t$, de sorte que les bornes sont inchangées et que $\mathrm{d}t = (1-s)^{-1}\mathrm{d}x$. On obtient donc :

$$\begin{aligned} M_T(s) &= \frac{1}{\Gamma(\beta)} \int_0^{+\infty} \left(\frac{x}{1-s}\right)^{\beta-1} e^{-x} \frac{1}{1-s}\mathrm{d}x \\ &= \frac{1}{(1-s)^{\beta}\,\Gamma(\beta)} \underbrace{\int_0^{+\infty} x^{\beta-1} e^{-x}\mathrm{d}x}_{\Gamma(\beta)} \\ &= (1-s)^{-\beta}. \end{aligned}$$

Pour obtenir l'espérance et la variance de la loi Gamma, on utilise donc :

$$K_T(s) = \ln M(s) = -\beta \ln(1-s).$$

La dérivée première donne l'espérance de la distribution :

$$K_T'(s) = \frac{\beta}{1-s} \Rightarrow K_T'(0) = \mathrm{E}(T) = \beta,$$

et la dérivée seconde donne la variance :

$$K_T^{"}(s) = \frac{\beta}{(1-s)^2} \Rightarrow K_T^{"}(0) = \mathrm{V}(T) = \beta.$$

La fonction génératrice du logarithme de cette variable s'obtient par :

$$\begin{aligned} M_{\ln T}(s) &= \mathrm{E}(T^s) \\ &= \int_0^{+\infty} \frac{t^{s+\beta-1} e^{-t}}{\Gamma(\beta)}\mathrm{d}t \\ &= \frac{\Gamma(s+\beta)}{\Gamma(\beta)}, \end{aligned}$$

et pour trouver ses deux premiers moments on utilise :

$$K_{\ln T}(s) = \ln M_{\ln T}(s) = \ln\Gamma(s+\beta) - \ln\Gamma(\beta).$$

Pour le modèle en logarithmes, on utilise donc :

$$K_{\ln T}'(s) = \frac{\Gamma'(s+\beta)}{\Gamma(s+\beta)} \Rightarrow K_{\ln T}'(0) = \mathrm{E}(\ln T) = \frac{\Gamma'(\beta)}{\Gamma(\beta)},$$

ainsi que :

$$\begin{aligned} K_{\ln T}^{"}(s) &= \frac{\Gamma^{"}(s+\beta)}{\Gamma(s+\beta)} - \left(\frac{\Gamma'(s+\beta)}{\Gamma(s+\beta)}\right)^2 \\ &\Rightarrow K_{\ln T}^{"}(0) = \mathrm{V}(\ln T) = \frac{\Gamma^{"}(\beta)}{\Gamma(\beta)} - \left(\frac{\Gamma'(\beta)}{\Gamma(\beta)}\right)^2. \end{aligned}$$

9.4.2.4 Loi de Weibull

Pour trouver la fonction génératrice des moments de la loi de Weibull, il suffit de remarquer que $U = \ln hT^{\alpha}$ suit une loi de Gumbel de fonction génératrice des moments égale à $\Gamma(1+s)$. Ceci implique :

$$\begin{aligned} M_U(s) = \mathrm{E}\left(e^{s\ln(hT^{\alpha})}\right) &= \Gamma(1+s) \\ \Leftrightarrow \mathrm{E}\left(h^s T^{\alpha s}\right) &= \Gamma(1+s) \\ \Leftrightarrow \mathrm{E}\left(T^{\alpha s}\right) &= h^{-s}\Gamma(1+s) \\ \Leftrightarrow \mathrm{E}\left(T^{j}\right) &= h^{-j/\alpha}\Gamma(1+j/\alpha), \end{aligned}$$

avec $j = \alpha s$. On en déduit :

$$\mathrm{E}(T) = h^{-1/\alpha}\Gamma(1+1/\alpha) \text{ et } \mathrm{E}\left(T^2\right) = h^{-2/\alpha}\Gamma(1+2/\alpha),$$

d'où la variance :

$$\begin{aligned} \mathrm{V}(T) &= \mathrm{E}\left(T^2\right) - \mathrm{E}(T)^2 \\ &= h^{-2/\alpha}\left(\Gamma(1+2/\alpha) - \Gamma(1+1/\alpha)^2\right). \end{aligned}$$

On remarque que l'espérance mathématique peut également se simplifier par la formule :

$$\begin{aligned} \Gamma(x) &= (x-1)\Gamma(x-1) \\ \Rightarrow \Gamma(1+1/\alpha) &= \frac{1}{\alpha}\Gamma\left(\frac{1}{\alpha}\right), \end{aligned}$$

ce qui implique :

$$\mathrm{E}(T) = h^{-1/\alpha}\frac{1}{\alpha}\Gamma\left(\frac{1}{\alpha}\right).$$

L'expression des moments du logarithme de la durée est également utilisée dans les applications. D'après ce qui précède :

$$M_{\ln T}(s) = \mathrm{E}(T^s) = h^{-s/\alpha}\Gamma(1+s/\alpha),$$

de sorte que :

$$\begin{aligned} K_{\ln T}(s) &= \ln M_{\ln T}(s) \\ &= -\frac{s}{\alpha}\ln h + \ln\Gamma(1+s/\alpha), \end{aligned}$$

d'où les dérivées :

$$\begin{aligned} K_{\ln T}^{'}(s) &= \frac{1}{\alpha}\left(-\ln h + \frac{\Gamma'(1+s/\alpha)}{\Gamma(1+s/\alpha)}\right) \\ &\Rightarrow K_{\ln T}^{'}(0) = \frac{1}{\alpha}\left(-\ln h + \Gamma'(1)\right) \\ &\Leftrightarrow \mathrm{E}(\ln T) = \left(-\ln h + \Gamma'(1)\right)/\alpha \\ &\Leftrightarrow \mathrm{E}(\ln T) = -\frac{1}{\alpha}\left(\ln h + \gamma_E\right), \end{aligned}$$

et que :

$$\begin{aligned} K_{\ln T}^{"}(s) &= \frac{1}{\alpha^2}\left(\Gamma"(1) - \Gamma'(1)^2\right) \\ &\Leftrightarrow K_{\ln T}^{"}(0) = \frac{1}{\alpha^2}\left(\Gamma"(1) - \Gamma'(1)^2\right) \\ &\Leftrightarrow \mathrm{V}(\ln T) = \frac{\pi^2}{6\alpha^2}. \end{aligned}$$

9.4.2.5 Loi Gamma généralisée

Pour définir le modèle Gamma généralisé, on suppose que $hT^{\alpha} = V$ suit une loi Gamma $\gamma(\beta, 1)$. Les moments de cette loi peuvent être trouvés directement en utilisant :

$$\mathrm{E}\left(V^j\right) = \frac{\Gamma(j+\beta)}{\Gamma(\beta)},$$

ce qui implique :

$$\begin{aligned} \mathrm{E}\left((hT^{\alpha})^j\right) &= \frac{\Gamma(j+\beta)}{\Gamma(\beta)} \\ &\Leftrightarrow \mathrm{E}\left(T^{\alpha j}\right) = h^{-j}\frac{\Gamma(j+\beta)}{\Gamma(\beta)}, \end{aligned}$$

il suffit alors de poser $s = \alpha j$ ($\Leftrightarrow j = s/\alpha$)pour obtenir :

$$\mathrm{E}(T^s) = h^{-s/\alpha}\frac{\Gamma(s/\alpha+\beta)}{\Gamma(\beta)},$$

ce qui permet d'obtenir les moments non centrés :

$$\mathrm{E}(T) = \frac{h^{-1/\alpha}\Gamma(1/\alpha+\beta)}{\Gamma(\beta)} \text{ et } \mathrm{E}\left(T^2\right) = \frac{h^{-2/\alpha}\Gamma(2/\alpha+\beta)}{\Gamma(\beta)},$$

dont on déduit la variance par la formule classique :

$$\begin{aligned} \mathrm{V}(T) &= \mathrm{E}\left(T^2\right) - \mathrm{E}(T)^2 \\ &= h^{-2/\alpha}\left\{\frac{\Gamma(2/\alpha+\beta)}{\Gamma(\beta)} - \left(\frac{\Gamma(1/\alpha+\beta)}{\Gamma(\beta)}\right)^2\right\}. \end{aligned}$$

Pour obtenir les moments du logarithme de la variablel de durée, il suffit de remarquer que la fonction $\mathrm{E}\left(T^s\right)$ est identique à $M_{\ln T}(s)$ de sorte que l'on peut écrire la fonction génératrice des moments centrés :

$$\begin{aligned} K_{\ln T}(s) &= \ln M_{\ln T}(s) \\ &= -\frac{s}{\alpha}\ln h + \ln\Gamma\left(s/\alpha+\beta\right) - \ln\Gamma\left(\beta\right), \end{aligned}$$

d'où l'espérance du logarithme :

$$\begin{aligned} K_{\ln T}^{'}(s) &= -\frac{1}{\alpha}\ln h + \frac{1}{\alpha}\frac{\Gamma'\left(s/\alpha+\beta\right)}{\Gamma\left(s/\alpha+\beta\right)} \\ &\Leftrightarrow K_{\ln T}^{'}(0) = \mathrm{E}\left(\ln T\right) = \frac{1}{\alpha}\left(-\ln h + \frac{\Gamma'\left(\beta\right)}{\Gamma\left(\beta\right)}\right), \end{aligned}$$

et sa variance :

$$\begin{aligned} K_{\ln T}^{"}(s) &= \frac{1}{\alpha^2}\left[\frac{\Gamma"\left(s/\alpha+\beta\right)}{\Gamma\left(s/\alpha+\beta\right)^2} - \left(\frac{\Gamma'\left(s/\alpha+\beta\right)}{\Gamma\left(s/\alpha+\beta\right)}\right)^2\right] \\ &\Leftrightarrow K_{\ln T}^{"}(0) = \mathrm{V}\left(\ln T\right) = \frac{1}{\alpha^2}\left[\frac{\Gamma"\left(\beta\right)}{\Gamma\left(\beta\right)^2} - \left(\frac{\Gamma'\left(\beta\right)}{\Gamma\left(\beta\right)}\right)^2\right]. \end{aligned}$$

9.4.2.6 Loi normale

La loi log-normale n'admet pas de fonction génératrice des moments parce que l'intégrale qui la définit *n'est pas* convergente :

$$K_T(s) = +\infty,$$

mais on peut calculer tous les moments de cette loi en utilisant la fonction génératrice des moments de loi normale. C'est ce qui explique la présence de cette section. Soit une variable aléatoire X suivant une loi normale $\mathrm{N}\left(m,\sigma^2\right)$, sa fonction génératrice des moments peut être obtenue de la manière suivante :

$$\begin{aligned} K_X(s) &= \mathrm{E}\left(e^{sX}\right) \\ &= \int_{-\infty}^{+\infty} e^{sx}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-m}{\sigma}\right)^2}\mathrm{d}x, \end{aligned}$$

on effectue donc le changement de variable $z = (x-m)/\sigma$, ce qui ne change pas la valeur des bornes et implique $x = m+\sigma z$ et $\mathrm{d}x = \sigma\mathrm{d}z$. On

a donc :

$$\begin{aligned} M_X(s) &= \int_{-\infty}^{+\infty} e^{s(m+\sigma z)} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \sigma \mathrm{d}z \\ &= e^{sm} \int_{-\infty}^{+\infty} e^{s\sigma z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \mathrm{d}z \\ &= e^{sm} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(z^2-2s\sigma z\right)} \mathrm{d}z \\ &= e^{sm+s^2\sigma^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(z^2-2s\sigma z+s^2\sigma^2\right)} \mathrm{d}z \\ &= e^{sm+s^2\sigma^2/2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(z-s\sigma)^2} \mathrm{d}z, \end{aligned}$$

et le second terme de l'expression ci-dessus est la somme sur son support de la densité d'une loi normale $\mathrm{N}(s\sigma, 1)$, qui est égale à 1 par définition. On a donc :

$$M_X(s) = \exp\left(sm + s^2\sigma^2/2\right).$$

On peut en déduire tous les moments de la loi normale. Pour obtenir les moments centrés, on prend :

$$\begin{aligned} K_X(s) &= \ln M_X(s) \\ &= sm + s^2\sigma^2/2, \end{aligned}$$

on en déduit :

$$\begin{aligned} K'_X(s) &= m + s\sigma^2 \\ &\Rightarrow K'_X(0) = \mathrm{E}(X) = m, \end{aligned}$$

ainsi que :

$$\begin{aligned} K^{"}_X(s) &= \sigma^2 \\ &\Rightarrow K^{"}_X(0) = \mathrm{V}(X) = \sigma^2. \end{aligned}$$

9.4.2.7 Loi log-normale

Par définition, la fonction génératrice du logarithme de la variable de durée $\ln T$ est identique à celle de la loi normale $\mathrm{N}\left(m, \sigma^2\right)$ donnée par :

$$K_X(s) = K_{\ln T}(s) = \mathrm{E}\left(e^{s \ln T}\right) = \mathrm{E}\left(T^s\right),$$

on en déduit les moments de la loi log-normale :

$$\mathrm{E}\left(T^s\right) = \exp\left(sm + s^2\sigma^2/2\right),$$

ce qui donne pour espérance :

$$\mathrm{E}\left(T\right) = \exp\left(m + \sigma^2/2\right),$$

et pour variance :

$$\begin{aligned}\mathrm{V}\left(T\right) &= \mathrm{E}\left(T^2\right) - \mathrm{E}\left(T\right)^2 \\ &= \exp\left(2m + 2\sigma^2\right) - \exp\left(2m + \sigma^2\right) \\ &= \exp\left(2m + \sigma^2\right)\left(\exp\left(\sigma^2\right) - 1\right).\end{aligned}$$

9.4.3 Résumé

Le tableau suivant résume les hypothèses qu'il faut effectuer pour retrouver chacun des modèles à partir de la relation suivante :

$$\ln T = k_1 + k_2 U,$$

le lecteur notera que les variables explicatives influençant la fonction de hasard ou la durée moyenne se trouvent dans la partie k_1 de ce modèle.

Modèle	Hypothèses
Exponentiel	$k_1 = -\ln h,\ k_2 = 1,\ \exp\left(U\right) \rightsquigarrow \gamma\left(1,1\right)$
Weibull	$k_1 = -\alpha^{-1}\ln h,\ k_2 = \alpha^{-1},\ \exp\left(U\right) \rightsquigarrow \gamma\left(1,1\right)$
Gamma	$k_1 = -\ln h,\ k_2 = 1,\ \exp\left(U\right) \rightsquigarrow \gamma\left(\beta,1\right)$
Gamma généralisé	$k_1 = -\alpha^{-1}\ln h,\ k_2 = \alpha^{-1},\ \exp\left(U\right) \rightsquigarrow \gamma\left(\beta,1\right)$
Log-Normal	$k_1 = m,\ k_2 = \sigma,\ \exp\left(U\right) \rightsquigarrow \mathrm{LN}\left(0,1\right)$

On peut également résumer l'espérance et la variance du terme d'erreur :

Modèle	$\mathrm{E}\left(U\right)$	$\mathrm{V}\left(U\right)$
Exponentiel	$-\gamma_E$	$\pi^2/6$
Weibull	$-\gamma_E$	$\pi^2/6$
Gamma	$\Gamma'\left(\beta\right)/\Gamma\left(\beta\right)$	$\Gamma"\left(\beta\right)/\Gamma\left(\beta\right) - \left(\Gamma'\left(\beta\right)/\Gamma\left(\beta\right)\right)^2$
Gamma généralisé	$\Gamma'\left(\beta\right)/\Gamma\left(\beta\right)$	$\Gamma"\left(\beta\right)/\Gamma\left(\beta\right) - \left(\Gamma'\left(\beta\right)/\Gamma\left(\beta\right)\right)^2$
Log-Normal	0	1

ainsi que de la perturbation complète du modèle :

Modèle	$\mathrm{E}\left(k_2 U\right)$	$\mathrm{V}\left(k_2 U\right)$
Exponentiel	$-\gamma_E$	$\pi^2/6$
Weibull	$-\gamma_E/\alpha$	$\pi^2/\left(6\alpha^2\right)$
Gamma	$\Gamma'\left(\beta\right)/\Gamma\left(\beta\right)$	$\Gamma"\left(\beta\right)/\Gamma\left(\beta\right) - \left(\Gamma'\left(\beta\right)/\Gamma\left(\beta\right)\right)^2$
Gamma généralisé	$\alpha^{-1}\Gamma'\left(\beta\right)/\Gamma\left(\beta\right)$	$\alpha^{-2}\left\{\Gamma"\left(\beta\right)/\Gamma\left(\beta\right) - \left(\Gamma'\left(\beta\right)/\Gamma\left(\beta\right)\right)^2\right\}$
Log-Normal	0	σ^2

9.5 Introduction des variables explicatives

9.5.1 Modèles à hasards proportionnels

Soit $X_i = (X_{1i}, ...; X_{pi})$ un vecteur de p variables explicatives, on dit qu'un modèle est à hasard proportionnel s'il vérifie :

$$h_i(t) = h_0(t) \exp(X_i b),$$

où $h_0(t)$ est une fonction de hasard appelée hasard de base. On remarque qu'avec cette convention, le ratio des hasard de deux individus ne dépend que des variables explicatives, et non du temps :

$$\frac{h_i(t)}{h_j(t)} = \exp((X_i - X_j) b).$$

En prenant la fonction de hasard en logarithmes, on obtient :

$$\ln h_i(t) = \ln h_0(t) + X_i b,$$

de sorte que l'on peut écrire :

$$\frac{\partial \ln h_i(t)}{\partial X_{ki}} = b_k,$$

ceci implique que l'on peut interpréter b_k comme une élasticité quand la variable explicative k est en logarithmes. S'il s'agit d'une indicatrice, le coefficient b_k représente (s'il est proche de 0) l'écart de hasard en pourcentage ($100 b_k$) entre la modalité 1 et la modalité 0 :

$$\begin{aligned} b_k &= \ln h_i(t|X_{ki} = 1) - \ln h_i(t|X_{ki} = 0) \\ &= \ln\left(1 + \frac{h_i(t|X_{ki} = 1) - h_i(t|X_{ki} = 0)}{h_i(t|X_{ki} = 0)}\right) \\ &\simeq \frac{h_i(t|X_{ki} = 1) - h_i(t|X_{ki} = 0)}{h_i(t|X_{ki} = 0)}. \end{aligned}$$

Pour procéder à l'estimation, on aura également besoin de la fonction de survie, à cause des données censurées :

$$\begin{aligned} S_i(t) &= \exp\left\{-\int_0^t h(x)\,\mathrm{d}x\right\} \\ &= \exp\left\{-\exp(X_i b)\int_0^t h_0(x)\,\mathrm{d}x\right\}, \end{aligned}$$

on en déduit la densité :

$$\begin{aligned} f_i(t) &= h_i(t) S_i(t) \\ &= h_0(t) \exp(X_i b) \exp\left\{-\exp(X_i b)\int_0^t h_0(x)\,\mathrm{d}x\right\}. \end{aligned}$$

Les modèles exponentiel, de Weibull et Gamma généralisés sont des modèles à hasard proportionnels. Toutefois dans le dernier cas, on ne peut pas écrire explicitement les fonction de survie et de hasard. Il faut recourir à une intégration numérique.

9.5.2 Le modèle exponentiel

Dans le cas du modèle exponentiel, on a :

$$h_0(t) = h,$$

de sorte que :

$$h_i(t) = h \exp(X_i b),$$

et

$$\int_0^t h_0(x)\,\mathrm{d}x = h \int_0^t \mathrm{d}x = ht,$$

de sorte que :

$$\begin{aligned} S_i(t) &= \exp\{-\exp(X_i b)\, ht\}, \\ f_i(t) &= h \exp(X_i b) \exp\{-\exp(X_i b)\, ht\}. \end{aligned}$$

Dans le cas du modèle de Weibull :

$$h_0(t) = h\alpha t^{\alpha-1},$$

d'où la fonction de hasard individuelle :

$$h_i(t) = h\alpha t^{\alpha-1} \exp(X_i b),$$

et

$$\begin{aligned} \int_0^t h_0(x)\,\mathrm{d}x &= h \int_0^t \alpha x^{\alpha-1}\mathrm{d}x \\ &= h\,[x^\alpha]_0^t \\ &= ht^\alpha, \end{aligned}$$

ce qui implique la fonction de survie :

$$S_i(t) = \exp\{-\exp(X_i b)\, ht^\alpha\},$$

et la densité :

$$f_i(t) = h\alpha t^{\alpha-1} \exp(X_i b) \exp\{-\exp(X_i b)\, ht^\alpha\}.$$

Pour la distribution Gamma, il faut évaluer numériquement les fonction de hasard $h_0(t)$ et de hasard cumulé

$$\Lambda_0(t) = \int_0^t h_0(x)\,\mathrm{d}t,$$

le hasard individuel est donné par la formule habituelle alors que la fonction de survie et la densité sont égales à :

$$\begin{aligned} S_i(t) &= \exp\{-\exp(X_i b)\,\Lambda_0(t)\}, \\ f_i(t) &= h_0(t)\exp(X_i b)\exp\{-\exp(X_i b)\,\Lambda_0(t)\}. \end{aligned}$$

9.6 Ecriture de la vraisemblance

Pour écrire la log vraisemblance, on définit les variables suivantes :

- y_i est la variable de durée observable. Cette durée peut être aussi bien complèté, c'est-à-dire observée jusqu'à son terme, que censurée, c'est-à-dire observée partiellement.
- $d_i \in \{0,1\}$ est une indicatrice de censure. On observe $d_i = 1$ si l'observation i est censurée à droite et $d_i = 0$ si la durée est complète.
- Si la durée n'est pas censurée la vraisemblance de l'individu i est égale à $f_i(y_i)$, sinon elle est égale à $S_i(y_i)$.
- La log vraisemblance avec censure à droite s'écrit toujours :

$$\ell(y|X,\theta) = \sum_{i=1}^{N} (1-d_i)\ln f_i(y_i) + d_i \ln S_i(y_i).$$

- En utilisant $f_i(y_i) = h_i(y_i)\,S_i(y_i)$ on obtient :

$$\begin{aligned} \ell(y|X,\theta) &= \sum_{i=1}^{N} (1-d_i)\left(\ln h_i(y_i) + \ln S_i(y_i)\right) + d_i \ln S_i(y_i) \\ &= \sum_{i=1}^{N} (1-d_i)\ln h_i(y_i) + \ln S_i(y_i) \end{aligned}$$

9.6.1 Modèle exponentiel

Les quantités dont on a besoin sont égales à :

$$\ln h_i(y_i) = \ln h + X_i b,$$

et

$$\begin{aligned}\ln S_i(y_i) &= -\exp(X_i b)\, h y_i \\ &= -\exp(\ln h + X_i b)\, y_i,\end{aligned}$$

on voit que $\ln h$ est le terme constant du modèle de sorte qu'il ne faut pas en mettre dans la liste des variables explicatives. On peut également faire un changement de paramètres :

$$Z_i = (1, X_i) \text{ et } \beta = \begin{pmatrix} \ln h \\ b \end{pmatrix}$$

de sorte que :

$$Z_i \beta = \ln h + X_i b,$$

la log-vraisemblance ce réécrit donc :

$$\ell(y|X,\beta) = \sum_{i=1}^{N} (1 - d_i)(Z_i\beta) - \exp(Z_i\beta)\, y_i,$$

d'où le vecteur du score :

$$\frac{\partial \ell}{\partial \beta}(y|X,\beta) = \sum_{i=1}^{N} Z_i' \left(1 - d_i - \exp(Z_i\beta)\, y_i\right),$$

et le hessien

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'}(y|X,\beta) = -\sum_{i=1}^{N} Z_i' Z_i \exp(Z_i\beta)\, y_i \ll 0.$$

Une fois l'estimation réalisée, on obtient l'estimateur du maximum de vraisemblance de h par :

$$\widehat{h} = \exp\left(\widehat{\beta}_1\right) = g\left(\widehat{\beta}_1\right),$$

et on estime sa variance asymptotique par :

$$\begin{aligned}\widehat{\mathrm{Vas}}\left(\widehat{h}\right) &= \frac{\partial g}{\partial \beta_1}\left(\widehat{\beta}_1\right) \widehat{\mathrm{Vas}}\left(\widehat{\beta}_1\right) \frac{\partial g}{\partial \beta_1'}\left(\widehat{\beta}_1\right) \\ &= \widehat{\mathrm{Vas}}\left(\widehat{\beta}_1\right) \exp\left(2\widehat{\beta}_1\right).\end{aligned}$$

9.6.2 Modèle de Weibull

Les quantités dont on a besoin sont égales à :

$$\ln h_i\left(y_i\right) = \ln h + X_i b + \ln \alpha + (\alpha - 1) \ln y_i,$$

et

$$\begin{aligned} \ln S_i\left(y_i\right) &= -\exp\left(X_i b\right) h y_i^{\alpha} \\ &= -\exp\left(\ln h + X_i b\right) y_i^{\alpha}, \end{aligned}$$

on voit que $\ln h$ est le terme constant du modèle de sorte qu'il ne faut pas en mettre dans la liste des variables explicatives. On peut également faire un changement de paramètres similaire à celui du modèle exponentiel :

$$Z_i = (1, X_i), \ \beta = \begin{pmatrix} \ln h \\ b \end{pmatrix} \text{ et } \theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

de sorte que :

$$\ell\left(y|X,\theta\right) = \sum_{i=1}^{N} (1 - d_i)\left(Z_i\beta + \ln \alpha + (\alpha - 1) \ln y_i\right) - \exp\left(Z_i\beta\right) y_i^{\alpha},$$

pour calculer le vecteur du score, on remarque que :

$$\frac{\mathrm{d}y_i^{\alpha}}{\mathrm{d}\alpha} = \frac{\mathrm{d}}{\mathrm{d}\alpha}\left(e^{\alpha \ln y_i}\right) = y_i^{\alpha} \ln y_i$$

d'où le vecteur du score :

$$\begin{aligned} \frac{\partial \ell}{\partial \alpha}\left(y|X,\theta\right) &= \sum_{i=1}^{N} (1 - d_i)\left(\frac{1}{\alpha} + \ln y_i\right) - \exp\left(Z_i\beta\right) y_i^{\alpha} \ln y_i \\ \frac{\partial \ell}{\partial \beta}\left(y|X,\theta\right) &= \sum_{i=1}^{N} Z_i'\left(1 - d_i - \exp\left(Z_i\beta\right) y_i^{\alpha}\right) \end{aligned}$$

et le hessien

$$\begin{aligned} \frac{\partial^2 \ell}{\partial \alpha^2}\left(y|X,\theta\right) &= -\sum_{i=1}^{N}\left\{\frac{1 - d_i}{\alpha^2} + \exp\left(Z_i\beta\right) y_i^{\alpha}\left(\ln y_i\right)^2\right\} \\ \frac{\partial^2 \ell}{\partial \beta \partial \alpha}\left(y|X,\theta\right) &= -\sum_{i=1}^{N} Z_i' \exp\left(Z_i\beta\right) y_i^{\alpha} \ln y_i \\ \frac{\partial^2 \ell}{\partial \beta \partial \beta'}\left(y|X,\theta\right) &= -\sum_{i=1}^{N} Z_i' Z_i \exp\left(Z_i\beta\right) y_i^{\alpha} \end{aligned}$$

Une fois l'estimation réalisée, on obtient l'estimateur du maximum de vraisemblance de h par :

$$\widehat{h} = \exp\left(\widehat{\theta}_1\right),$$

et on estime sa variance asymptotique comme dans le modèle exponentiel par :

$$\widehat{\text{Vas}}\left(\widehat{h}\right) = \widehat{\text{Vas}}\left(\widehat{\theta}_1\right) \exp\left(2\widehat{\theta}_1\right).$$

9.6.3 Modèle log-normal

Il ne s'agit pas d'un modèle à hasard proportionnel. Avec un modèle log normal, on fait directement une hypothèse sur la durée elle-même puisque l'on pose que :

$$\ln T = m + \sigma U,$$

où U suit une loi normale centrée et réduite. On peut donc voir ce modèle comme une simple extension du modèle linéaire standard normal. S'il n'y avait pas de censure des données, la méthode d'estimation adaptée serait simplement celle des moindres carrés ordinaires appliqués au logarithme de la durée. Toutefois, comme nous supposons la présence d'une censure droite, on ne peut pas appliquer les moindres carrés ordinaires. Il faut recourir à la méthode du maximum de vraisemblance. Une manière naturelle d'introduire des variables explicatives dans ce type de modèle consiste à poser simplement $m = Xb$.

La densité est alors donnée directement par :

$$f_i\left(y_i\right) = \frac{1}{\sigma y_i}\varphi\left(\frac{\ln y_i - X_i b}{\sigma}\right),$$

et la fonction de survie par :

$$S_i\left(y_i\right) = 1 - \Phi\left(\frac{\ln y_i - X_i b}{\sigma}\right),$$

ce qui donne la log-vraisemblance de l'échantillon :

$$\begin{aligned}
\ell\left(y|X,\theta\right) &= \sum_{i=1}^{N}\left(1-d_i\right)\ln f_i\left(y_i\right) + d_i \ln S_i\left(y_i\right) \\
&= \sum_{i=1}^{N}\left(1-d_i\right)\left[\ln\varphi\left(\frac{\ln y_i - X_i b}{\sigma}\right) - \ln\sigma - \ln y_i\right] \\
&+ d_i \ln\left[1 - \Phi\left(\frac{\ln y_i - X_i b}{\sigma}\right)\right],
\end{aligned}$$

on peut simplifier l'écriture du modèle en faisant le changement de paramètres suivant :

$$\beta = \frac{b}{\sigma},\ \gamma = \frac{1}{\sigma} \text{ et } \theta = \begin{pmatrix} \beta \\ \gamma \end{pmatrix}$$

ce qui donne :

$$\ell(y|X,\theta) = \sum_{i=1}^{N} (1-d_i)\left[\ln\varphi(\gamma \ln y_i - X_i\beta) + \ln\gamma - \ln y_i\right] + d_i \ln\left[1 - \Phi(\gamma \ln y_i - X_i\beta)\right].$$

Pour alléger les notations, on pose :

$$u_i = \gamma \ln y_i - X_i\beta.$$

La première partie du vecteur du score est égal à :

$$\frac{\partial \ell}{\partial \beta}(y|X,\theta) = \sum_{i=1}^{N} X_i' \left\{ (1-d_i)\left[-\frac{\varphi'(u_i)}{\varphi(u_i)}\right] - d_i \frac{\varphi(u_i)}{1-\Phi(u_i)} \right\},$$

à ce stade on utilise $\varphi'(u) = -u\varphi(u)$, ce qui permet de simplifier l'expression précédente :

$$\frac{\partial \ell}{\partial \beta}(y|X,\theta) = \sum_{i=1}^{N} X_i' \left\{ (1-d_i)\, u_i - d_i \frac{\varphi(u_i)}{1-\Phi(u_i)} \right\},$$

si $c_i = 0\ \forall i$, on retrouve les moindres carrés ordinaires, sinon on ajoute un terme pour corriger la censure droite. Pour l'autre paramètre, on trouve :

$$\frac{\partial \ell}{\partial \gamma}(y|X,\theta) = \sum_{i=1}^{N} \left\{ \ln y_i \left((1-d_i)\, u_i - d_i \frac{\varphi(u_i)}{1-\Phi(u_i)} \right) + \frac{1-d_i}{\gamma} \right\}.$$

9.6.4 Généralisation

Les variables de durée peuvent également être censurées à gauche. On remarque qu'une même durée peut être censurée à la fois à gauche et à droite. Le fait d'avoir des censures à gauche ne change toutefois rien à notre analyse. En effet, si l'on observe une durée censurée y_i, on sait juste que la vraie durée est supérieure ou égale à y_i, et ce qu'elle soit censurée à gauche, à droite ou des deux côtés. Dans ce cas la contribution à la vraisemblance reste égale à $S_i(y_i)$. Il faut juste penser à définir une indicatrice de censure égale au maximum des deux indicatrices de censure

gauche et droite. Soit d_{1i} une indicatrice de censure gauche et d_{2i} une indicatrice de censure droite, on doit prendre :

$$d_i = \max\left(d_{1i}, d_{2i}\right),$$

dans les expressions de la section précédente. Cette règle reste valable si la censure a lieu avec des "trous" différents d'une observation à l'autre, car dans tous les cas la seule information disponible est que la vraie durée est supérieure à la durée observée y_i.

CHAPITRE 10

Les variables tronquées

10.1 Le modèle tronqué

On dit qu'un modèle est tronqué lorsques les variables explicatives X_i ne sont pas observables lorsque la varible expliquée z_i^* passe en dessous d'un certain seuil C_i. Ce cas peut se produire soit lorsque l'on n'interroge que les individus pour lesquels $z_i^* > C_i$ soit lorsque les réponses aux variables explicatives X_i n'ont de sens que lorsque $z_i^* > C_i$. Pour simplifier l'écriture du modèle, on pose :

$$y_i^* = z_i^* - C_i,$$

quantité qui peut toujours être calculée lorsque les seuils C_i sont connus. Avec ce changement de variable, on observe y_i^* lorsque $y_i^* > 0$ ($\Leftrightarrow z_i^* > C_i$). La variable latente est décrite par le modèle linéaire suivant :

$$y_i^* = X_i b + \sigma\, u_i, \quad u_i \overset{\text{iid}}{\rightsquigarrow} \mathrm{N}(0,1), \quad i = 1, ..., N$$

avec $\sigma > 0$. La perturbation du modèle est donc égale à :

$$v_i = \sigma\, u_i,$$

ce qui implique que $v_i \overset{\text{iid}}{\rightsquigarrow} \mathrm{N}\left(0, \sigma^2\right)$. La fonction de répartition de la loi normale centrée et réduite est notée $\Phi(z)$ et sa densité $\varphi(z)$. La variable observable, notée y_i, est définie par :

$$y_i = \begin{cases} \text{manquant} & \text{si } y_i^* \leq 0 \\ y_i^* & \text{sinon} \end{cases}$$

Pour procéder à l'estimation il nous faut l'expression de la densité de y_i^* tronquée en 0. Elle est égale, par définition, à :

$$f(y_i) = 1_{(y_i^* > 0)} \frac{f(y_i)}{\Pr[y_i^* > 0]}.$$

La probabilité d'observer la variable endogène est donnée par :

$$\begin{aligned}\Pr\left[y_i^* > 0\right] &= \Pr\left[X_i b + \sigma\, u_i > 0\right] \\ &= \Pr\left[u_i > -\frac{X_i b}{\sigma}\right] \\ &= 1 - \Phi\left(-\frac{X_i b}{\sigma}\right) \\ &= \Phi\left(\frac{X_i b}{\sigma}\right),\end{aligned}$$

et la vraisemblance est donnée par :

$$\ell_i = \sum_{y_i>0} \ln\left[\frac{1}{\sigma}\varphi\left(\frac{y_i - X_i b}{\sigma}\right)\right] - \ln\Phi\left(\frac{X_i b}{\sigma}\right).$$

Il faut noter ici que seules les observations strictement positives de y_i^* sont utilisables, contrairement au modèle Tobit que nous verrons plus loin où toutes les observations sont utilisables. L'espérance mathématique de y_i est donnée par :

$$\mathrm{E}\left(y_i\right) = \int_0^{+\infty} y_i \frac{\frac{1}{\sigma}\varphi\left(\frac{y_i - X_i b}{\sigma}\right)}{\Phi\left(\frac{X_i b}{\sigma}\right)}\,\mathrm{d}y_i,$$

on effectue le changement de variable :

$$u = \frac{y_i - X_i b}{\sigma},$$

ce qui implique :

$$\lim_{y_i \to 0} u = -\frac{X_i b}{\sigma}, \quad \lim_{y_i \to +\infty} u = +\infty \quad \text{et} \quad \mathrm{d}y_i = \sigma\,\mathrm{d}u,$$

d'où :

$$\begin{aligned}\mathrm{E}\left[y_i \middle| y_i > 0\right] &= \frac{1}{\Phi\left(\frac{X_i b}{\sigma}\right)} \int_{\frac{-X_i b}{\sigma}}^{+\infty} \left(X_i b + \sigma\, u\right) \varphi\left(u\right)\,\mathrm{d}u \\ &= \frac{1}{\Phi\left(\frac{X_i b}{\sigma}\right)} X_i b \underbrace{\int_{\frac{-X_i b}{\sigma}}^{+\infty} \varphi\left(u\right)\,\mathrm{d}u}_{\Phi\left(\frac{X_i b}{\sigma}\right)} + \frac{\sigma}{\Phi\left(\frac{X_i b}{\sigma}\right)} \int_{\frac{-X_i b}{\sigma}}^{+\infty} \underbrace{u\varphi\left(u\right)}_{-\varphi'(u)}\mathrm{d}u \\ &= X_i b + \frac{\sigma}{\Phi\left(\frac{X_i b}{\sigma}\right)} \left[-\varphi\left(u\right)\right]_{\frac{-X_i b}{\sigma}}^{+\infty} \\ &= X_i b + \sigma\,\frac{\varphi\left(\frac{X_i b}{\sigma}\right)}{\Phi\left(\frac{X_i b}{\sigma}\right)},\end{aligned}$$

cette espérance est valable sur les observations strictement postives et pourra être utilisée lors de l'estimation du modèle Tobit. On effectue les changements de paramètres suivant :

$$\beta = \frac{b}{\sigma} \text{ et } \quad h = \frac{1}{\sigma},$$

ce qui permet d'écrire la log-vraisemblance sous la forme :

$$\ell\left(y_1, ..., y_N\right) = \sum_{i=1}^{N} \left\{ \ln h - \frac{1}{2}\ln\left(2\pi\right) - \frac{1}{2}\left(hy_i - X_i\beta\right)^2 - \ln\Phi\left(X_i\beta\right) \right\},$$

le score est donc égal à :

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} X_i' \left[h\, y_i - X_i\beta - \frac{\varphi\left(X_i\beta\right)}{\Phi\left(X_i\beta\right)} \right]$$

$$\frac{\partial \ell}{\partial h} = \sum_{i=1}^{N} \left[\frac{1}{h} - h\, y_i^2 + \, X_i\beta\, y_i \right]$$

Pour simplifier les notations, on pose :

$$m_i = X_i b \quad \text{et} \quad \lambda_i = \frac{\varphi\left(X_i\beta\right)}{\Phi\left(X_i\beta\right)},$$

où λ_i est l'inverse du ratio de Mills. La première partie du score peut donc se réécrire :

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} X_i' \left[h\left(y_i - m_i\right) - \lambda_i \right]$$

On voit que le score $\partial\ell/\partial\beta$ est d'espérance nulle, puisque :

$$\mathrm{E}\left(y_i | y_i > 0\right) = m_i + \frac{1}{h}\lambda_i.$$

C'est également le cas pour $\partial\ell/\partial h$ et l'on peut donc écrire :

$$\mathrm{E}\left[\frac{1}{h} - h\, y_i^2 + h\, m_i\, y_i \right] = 0$$

$$\Leftrightarrow \mathrm{E}\left[y_i^2 | y_i > 0\right] = \frac{1}{h}\left[\frac{1}{h} + h\, m_i\, \mathrm{E}\left(y_i | y_i > 0\right) \right]$$

$$\Leftrightarrow \mathrm{E}\left[y_i^2 | y_i > 0\right] = \frac{1}{h^2} + m_i\, \mathrm{E}\left(y_i | y_i > 0\right)$$

$$\Leftrightarrow \mathrm{E}\left[y_i^2 | y_i > 0\right] = \frac{1}{h^2} + m_i^2 + \frac{m_i \lambda_i}{h}.$$

Cette expression nous servira pour déterminer l'algorithme du score. Les dérivées secondes sont égales à :

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = -\sum_{i=1}^{N} X_i' X_i \left[1 - \lambda_i \left(m_i + \lambda_i\right)\right]$$
$$\frac{\partial^2 \ell}{\partial \beta \partial h} = \sum_{i=1}^{N} X_i' y_i$$
$$\frac{\partial^2 \ell}{\partial h^2} = -\sum_{i=1}^{N} \left[\frac{1}{h^2} + y_i^2\right]$$

d'où l'espérance de l'opposé des dérivées secondes :

$$\mathrm{E}\left[-\frac{\partial^2 \ell}{\partial \beta \partial \beta'}\right] = \sum_{i=1}^{N} X_i' X_i \left[1 - \lambda_i \left(m_i + \lambda_i\right)\right]$$
$$\mathrm{E}\left[-\frac{\partial^2 \ell}{\partial \beta \partial h}\right] = -\sum_{i=1}^{N} X_i' \left[m_i + \frac{1}{h}\lambda_i\right]$$
$$\mathrm{E}\left[-\frac{\partial^2 \ell}{\partial h^2}\right] = \sum_{i=1}^{N} \left[\frac{2}{h^2} + m_i^2 + \frac{m_i \lambda_i}{h}\right]$$

On peut alors utiliser un algorithme du score en prenant les moindres carrés ordinaires comme valeur initiale, cet estimateur n'étant pas convergent. L'algorithme de Newton-Raphson est ici plus simple que l'algorithme du score en raison de la forme particulière des espérances de la variable tronquée y_i.

10.2 Le modèle Tobit

Le modèle Tobit est un modèle censuré, ce qui signifie que l'on observe les variables explicatives X_i dans tous les cas. On peut donc utiliser cette information supplémentaire.

10.2.1 Estimation

La probabilité que la variable latente y_i^* soit négative est donnée par :

$$\Pr\left[y_i = 0\right] = \Pr\left[y_i^* < 0\right] = 1 - \Phi\left(\frac{X_i b}{\sigma}\right),$$

et la probabilité d'observer une valeur strictement positive est simplement égale à :

$$f\left(y_i\right) = \frac{1}{\sigma}\varphi\left(\frac{y_i - X_i b}{\sigma}\right), \quad y_i > 0,$$

la vraisemblance est donc égale à :

$$\ell(y_1, ..., y_N) = (1 - d_i) \ln [1 - \Phi(X_i\beta)] + d_i \ln \left\{ \frac{h}{\sqrt{2\pi}} - \frac{1}{2} (h\, y_i - X_i\beta)^2 \right\},$$

avec :

$$d_i = \begin{cases} 0 & \text{si } y_i^* \leq 0 \\ 1 & \text{sinon} \end{cases}$$

où la variable dichotomique d_i suit une loi de Bernoulli de paramètre $\Phi(X_i\beta)$. Le score est égal à :

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{N} X_i' \left[d_i (h\, y_i - X_i\beta) - (1 - d_i) \frac{\varphi(X_i\beta)}{1 - \Phi(X_i\beta)} \right]$$

$$\frac{\partial \ell}{\partial h} = \sum_{i=1}^{N} d_i \left[\frac{1}{h} - h\, y_i^2 + h\, m_i\, y_i \right]$$

et l'on vérifie que l'espérance du score est nulle en utilisant :

$$\begin{aligned} \mathrm{E}(d_i y_i) &= \mathrm{E}(y_i | y_i > 0) \times \Pr(y_i > 0) \\ &= \left(m_i + \frac{1}{h} \frac{\varphi_i}{\Phi_i} \right) \Phi_i \\ &= m_i \Phi_i + \frac{1}{h} \varphi_i, \end{aligned}$$

ainsi que :

$$\begin{aligned} \mathrm{E}\left(d_i y_i^2\right) &= \mathrm{E}\left(y_i^2 | y_i > 0\right) \times \Pr(y_i > 0) \\ &= \left[\frac{1}{h^2} + m_i^2 + \frac{m_i}{h} \frac{\varphi_i}{\Phi_i} \right] \Phi_i \\ &= \Phi_i \left(\frac{1}{h^2} + m_i^2 \right) + \frac{m_i \varphi_i}{h} \end{aligned}$$

Les dérivées secondes sont données par :

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = - \sum_{i=1}^{N} X_i' X_i \left\{ d_i + (1 - d_i) \frac{\varphi_i}{1 - \Phi_i} \left(\frac{\varphi_i}{1 - \Phi_i} - h\, m_i \right) \right\}$$

$$\frac{\partial^2 \ell}{\partial \beta \partial h} = \sum_{i=1}^{N} d_i X_i' y_i$$

$$\frac{\partial^2 \ell}{\partial h^2} = - \sum_{i=1}^{N} d_i \left(\frac{1}{h^2} + y_i^2 \right)$$

d'où les espérances mathématiques nécessaires à l'algorithme du score :

$$\mathrm{E}\left[-\frac{\partial^2 \ell}{\partial\beta\partial\beta'}\right] = \sum_{i=1}^{N} X_i' X_i \left[\Phi_i + \varphi_i\left(\frac{\varphi_i}{1-\Phi_i} - h\, m_i\right)\right]$$

$$\mathrm{E}\left[-\frac{\partial^2 \ell}{\partial\beta\partial h}\right] = -\sum_{i=1}^{N} X_i' \left[m_i\Phi_i + \frac{1}{h}\varphi_i\right]$$

$$\mathrm{E}\left[-\frac{\partial^2 \ell}{\partial h^2}\right] = \sum_{i=1}^{N} \left[\Phi_i\left(\frac{2}{h^2} + m_i^2\right) + \frac{m_i\varphi_i}{h}\right]$$

10.2.2 Valeur initiale

Le fait que l'on observe toujours les variables explicatives permet de calculer facilement des valeurs initiales. Ceci provient du fait que l'on peut estimer un modèle Probit en prenant d_i comme variable expliquée. Pour les observations strictement positives, on utilise :

$$\begin{aligned} \mathrm{E}\left(y_i | y_i > 0\right) &= X_i b + \sigma \frac{\varphi\left(X_i\beta\right)}{\Phi\left(X_i\beta\right)} \\ &= \sigma\left(X_i\beta + \frac{\varphi\left(X_i\beta\right)}{\Phi\left(X_i\beta\right)}\right) \\ &= \sigma\mu_i, \end{aligned} \tag{10.1}$$

avec $\mu_i = X_i\beta + \varphi\left(X_i\beta\right)/\Phi\left(X_i\beta\right)$. Cette quantité peut facilement être estimée à partir de l'estimateur $\widehat{\beta}$ de la partie Probit du modèle :

$$\widehat{\mu}_i = X_i\widehat{\beta} + \frac{\varphi\left(X_i\widehat{\beta}\right)}{\Phi\left(X_i\widehat{\beta}\right)}.$$

En utilisant directement (10.1) on peut obtenir un estimateur convergent de σ en régressant y_i sur un estimateur convergent de μ_i par les moindres carrés ordinaires sans terme constant; ce qui donne :

$$\widehat{\sigma} = \frac{\sum_{y_i>0} \widehat{\mu}_i y_i}{\sum_{y_i>0} \widehat{\mu}_i^2},$$

dont on déduit la valeur initiale convergente pour le paramètre h :

$$\hat{h} = \frac{1}{\hat{\sigma}} = \frac{\sum_{y_i>0} \widehat{\mu}_i^2}{\sum_{y_i>0} \widehat{\mu}_i y_i}.$$

10.2.3 Retour aux paramètres structurels

La méthode précédente permet d'obtenir des estimateurs convergents des paramètres β et h. Pour revenir aux paramètres de départ du modèle, on utilise la propriété d'invariance fonctionnelle et le théorème de Slutsky. L'invariance fonctionnelle implique que $\hat{b} = \hat{\beta}/\hat{h}$ est l'estimateur du maximum de vraisemblance de b et que $\hat{\sigma} = 1/\hat{h}$ est l'estimateur du maximum de vraisemblance de σ. Le théorème de Slutstky permet de trouver la matrice de covariance asymptotique de $\left(\hat{b}', \sigma\right)'$. Soit :

$$\theta = \begin{pmatrix} \beta \\ h \end{pmatrix},$$

la distribution asymptotique de l'estimateur du maximum de vraisemblance est normale :

$$\sqrt{N}\left(\hat{\theta} - \theta\right) \xrightarrow{L} \mathrm{N}\left(0, \Omega_{\hat{\theta}}\right),$$

ce qui implique :

$$\sqrt{N}\left(g\left(\hat{\theta}\right) - g\left(\theta\right)\right) \xrightarrow{L} \mathrm{N}\left(0, \frac{\partial g}{\partial \theta'}\Omega_{\hat{\theta}}\frac{\partial g}{\partial \theta'}\right),$$

avec

$$g\left(\theta\right) = \left(\beta/h, 1/h\right) = \left(b, \sigma\right) \quad \text{et} \quad \frac{\partial g}{\partial \theta}\left(\theta\right) = \begin{pmatrix} \mathrm{I}_k/h & 0 \\ -\beta/h^2 & -1/h^2 \end{pmatrix},$$

où k est le nombre de paramètres du vecteur β. On en déduit que :

$$\begin{pmatrix} \widehat{b} \\ \widehat{\sigma} \end{pmatrix} = \begin{pmatrix} \widehat{\beta}/\widehat{h} \\ 1/\widehat{h} \end{pmatrix}$$

$$\text{et } \widehat{\mathrm{V}}\begin{pmatrix} \widehat{b} \\ \widehat{\sigma} \end{pmatrix} = \begin{pmatrix} \mathrm{I}_k/\widehat{h} & 0 \\ -\widehat{\beta}/\widehat{h}^2 & -1/\widehat{h}^2 \end{pmatrix} \widehat{\mathrm{V}}\begin{pmatrix} \widehat{\beta} \\ \widehat{h} \end{pmatrix} \begin{pmatrix} \mathrm{I}_k/\widehat{h} & -\widehat{\beta}/\widehat{h}^2 \\ 0 & -1/\widehat{h}^2 \end{pmatrix}$$

10.3 Le modèle Tobit généralisé

On présente ici le modèle développé par Heckman dans ses articles de 1976 et 1979.

10.3.1 Définition

La forme latente comporte maintenant deux équations. Une première variable latente y_1^* détermine la décision et une seconde variable latente

y_2^* détermine le montant observé quand la décision est prise. On a :

$$\begin{aligned} y_{1i}^* &= m_{1i} + \sigma_1 u_{1i} \\ y_{2i}^* &= m_{2i} + \sigma_2 u_{2i}, \end{aligned}$$

avec $m_{ji} = X_i b_j$ et :

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \overset{\text{iid}}{\rightsquigarrow} \mathrm{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right].$$

On observe la décision :

$$y_{1i} = \begin{cases} 0 & \text{si } y_{1i}^* \leq 0 \\ 1 & \text{sinon} \end{cases},$$

ainsi que le montant lorsque $y_{1i} = 1$:

$$y_{2i} = \begin{cases} \text{manquant} & \text{si } y_{1i}^* \leq 0 \\ y_{2i}^* & \text{sinon} \end{cases}$$

10.3.2 Estimation

Pour écrire la log vraisemblance, on a besoin de la loi normale bivariée. Sa densité est égale à :

$$f\left(y_1^*, y_2^*\right) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2\left(1-\rho^2\right)} \left(u_1^2 - 2\rho u_1 u_2 + u_2^2\right) \right\} \quad (10.2)$$

Pour les observations nulles, la probabilité est simplement :

$$\Pr\left[y_{1i} = 0\right] = 1 - \Phi\left(\frac{m_{1i}}{\sigma_1}\right) = 1 - \Phi\left(X_{1i}\beta_1\right),$$

avec $\beta_1 = b_1/\sigma_1$. Pour les observations positives, il faut calculer :

$$\begin{aligned} f\left(y_2^* \cap \left(y_1^* > 0\right)\right) &= \int_{-X_{1i}\beta_1}^{+\infty} f\left(u_1, u_2\right) \mathrm{d}u_1 \\ &= \int_{-X_{1i}\beta_1}^{+\infty} f\left(u_1 | u_2\right) \varphi\left(u_2\right) \mathrm{d}u_1 \\ &= \varphi\left(u_2\right) \int_{-X_{1i}\beta_1}^{+\infty} f\left(u_1 | u_2\right) \mathrm{d}u_1. \end{aligned}$$

D'après la densité (10.2) on a :

$$\begin{aligned} f\left(u_1 | u_2\right) &= \frac{\frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{ -\frac{1}{2(1-\rho^2)} \left(u_1^2 - 2\rho u_1 u_2 + u_2^2\right) \right\}}{\frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{u_2^2}{2} \right\}} \\ &= \frac{1}{\sqrt{1-\rho^2}} \varphi\left(\frac{u_1 - \rho u_2}{\sqrt{1-\rho^2}} \right) \end{aligned}$$

ce qui permet d'écrire :

$$I = \int_{-X_{1i}\beta_1}^{+\infty} f\left(u_1|u_2\right) \mathrm{d}u_1 = \int_{-X_{1i}\beta_1}^{+\infty} \frac{1}{\sqrt{1-\rho^2}} \varphi\left(\frac{u_1 - \rho u_2}{\sqrt{1-\rho^2}}\right) \mathrm{d}u_1,$$

en effectuant le changement de variable :

$$z = \frac{u_1 - \rho u_2}{\sqrt{1-\rho^2}},$$

on a :

$$\mathrm{d}u_1 = \sqrt{1-\rho^2}\mathrm{d}z, \quad \lim_{u_1 \to -X_{1i}\beta_1} z = \frac{-X_{1i}\beta_1 - \rho u_2}{\sqrt{1-\rho^2}}, \quad \lim_{u_1 \to +\infty} z = +\infty,$$

ce qui implique :

$$\begin{aligned}
I &= \int_{\frac{-X_{1i}\beta_1 - \rho u_2}{\sqrt{1-\rho^2}}}^{+\infty} \varphi\left(z\right) \mathrm{d}z \\
&= 1 - \int_{-\infty}^{\frac{-X_{1i}\beta_1 - \rho u_2}{\sqrt{1-\rho^2}}} \varphi\left(z\right) \mathrm{d}z \\
&= 1 - \Phi\left(\frac{-X_{1i}\beta_1 - \rho u_2}{\sqrt{1-\rho^2}}\right) \\
&= \Phi\left(\frac{X_{1i}\beta_1 + \rho u_2}{\sqrt{1-\rho^2}}\right) \\
&= \Phi\left(\frac{X_{1i}\beta_1 + \rho\left(y_{2i} - X_{2i}b_2\right)/\sigma_2}{\sqrt{1-\rho^2}}\right),
\end{aligned}$$

en posant :

$$h = \frac{1}{\sigma_2} \quad \text{et} \quad \beta_2 = \frac{b_2}{\sigma_2},$$

on obtient finalement :

$$I = \Phi\left(\frac{X_{1i}\beta_1 + \rho\left(h\, y_{2i} - X_{2i}\beta_2\right)}{\sqrt{1-\rho^2}}\right).$$

La log vraisemblance du modèle Tobit généralisé est donc égale à :

$$\begin{aligned}
\ell\left(\beta_1, \beta_2, h\right) = &\sum_{y_{1i}=0} \ln\left(1 - \Phi\left(X_{1i}\beta_1\right)\right) \\
&+ \sum_{y_{1i}=1} \ln \Phi\left(\frac{X_{1i}\beta_1 + \rho\left(h\, y_{2i} - X_{2i}\beta_2\right)}{\sqrt{1-\rho^2}}\right) \\
&+ \sum_{y_{1i}=1} \ln\left\{\frac{h}{\sqrt{2\pi}} \exp -\frac{1}{2}\left(h\, y_{2i} - X_{2i}\beta_2\right)^2\right\}
\end{aligned}$$

10.3.3 Valeur initiale

Comme dans le modèle Tobit simple, il est possible de trouver une valeur initiale à partir d'une méthode en deux étapes. Pour cela, on utilise l'espérance conditionnelle suivante :

$$\begin{aligned}\mathrm{E}\left(y_{2i}^{*}|y_{1i}^{*}>0\right) &= X_{2i}b_2+\sigma_2\,\mathrm{E}\left(u_{2i}|u_{1i}>-X_{1i}\beta_1\right)\\ &= X_{2i}b_2+\rho\sigma_2\frac{\varphi\left(X_{1i}\beta_1\right)}{\Phi\left(X_{1i}\beta_1\right)}.\end{aligned}$$

Dans un premier temps, on estime donc un modèle Probit sur toutes les observations, ce qui permet d'obtenir un estimateur de β_1 noté $\hat{\beta}_1$. On estime ensuite l'inverse du ratio de Mills noté $\widehat{\lambda}_i$:

$$\widehat{\lambda}_i=\frac{\varphi\left(X_{1i}\hat{\beta}_1\right)}{\Phi\left(X_{1i}\hat{\beta}_1\right)},$$

en régressant les observations positives de y_2^* sur X_2 et $\hat{M}_i$ on obtient un estimateur convergent de b_2 et de $\rho\sigma_2$. On peut ensuite soit estimer le modèle en faisant un balayage sur ρ soit utiliser une expression similaire sur la variance conditionnelle de y_2^* pour estimer ρ. Le lecteur intéressé est invité à se reporter à l'ouvrage de C. Gouriéroux.

10.3.4 Amélioration de l'estimation

La plupart des logiciels n'ont besoin que de la log-vraisemblance pour déterminer le maximum de la fonction précédente, surtout si elle prend pour valeur initiale l'estimateur en deux étapes de Heckman, parce qu'il est convergent. Toutefois, pour pouvoir estimer un système d'équation incluant une variable modélisée par un Tobit généralisé, il faut disposer des dérivées premières analytiques. Elles permettent de calculer la matrice de covariance de la forme réduite du modèle. Ceci permet également d'accélérer les procédures d'optimisation numériques. On pose les notations suivantes :

$$\mu_{1i}=X_{1i}\beta_1,\ \mu_{2i}=X_{2i}\beta_{2i}$$

et

$$\lambda_{2i}=\varphi\left(\frac{\mu_{1i}+\rho\left(h\,y_{2i}-\mu_{2i}\right)}{\sqrt{1-\rho^2}}\right)\Phi\left(\frac{\mu_{1i}+\rho\left(h\,y_{2i}-\mu_{2i}\right)}{\sqrt{1-\rho^2}}\right)^{-1}.$$

La log vraisemblance pour une observation s'écrit :

$$\begin{aligned}\ell_i = (1-y_{1i})\ln\left(1-\Phi\left(\mu_{1i}\right)\right) \\ + y_{1i}\ln\Phi\left(\frac{\mu_{1i}+\rho\left(h\,y_{2i}-\mu_{2i}\right)}{\sqrt{1-\rho^2}}\right) \\ + y_{1i}\left(\ln h-\frac{1}{2}\ln\left(2\pi\right)-\frac{1}{2}\left(h\,y_{2i}-\mu_{2i}\right)^2\right).\end{aligned}$$

Les dérivées pour chaque observation s'écrivent donc :

$$\begin{aligned}\frac{\partial\ell_i}{\partial\mu_{1i}} &= \frac{y_{1i}\lambda_{2i}}{\sqrt{1-\rho^2}}-\frac{\left(1-y_{1i}\right)\varphi_{1i}}{1-\Phi_{1i}}, \\ \frac{\partial\ell_i}{\partial\mu_{2i}} &= y_{1i}\left\{hy_{2i}-\mu_2-\frac{\rho y_{1i}\lambda_{2i}}{\sqrt{1-\rho^2}}\right\}, \\ \frac{\partial\ell_i}{\partial h} &= y_{1i}\left\{\frac{\rho y_{2i}\lambda_{2i}}{\sqrt{1-\rho^2}}+\frac{1}{h}-y_{2i}\left(hy_{2i}-\mu_{2i}\right)\right\}, \\ \frac{\partial\ell_i}{\partial\rho} &= y_{1i}\lambda_{2i}\left(1-\rho^2\right)^{-3/2}\left\{\rho\mu_{1i}+hy_{2i}-\mu_{2i}\right\}.\end{aligned}$$

On en déduit les dérivées par rapport aux paramètres :

$$\frac{\partial\ell}{\partial\beta_1}=\sum_{i=1}^{N}X'_{1i}\frac{\partial\ell_i}{\partial\mu_{1i}},\ \frac{\partial\ell}{\partial\beta_2}=\sum_{i=1}^{N}X'_{2i}\frac{\partial\ell_i}{\partial\mu_{2i}},\ \frac{\partial\ell}{\partial h}=\sum_{i=1}^{N}\frac{\partial\ell_i}{\partial h}\text{ et }\frac{\partial\ell}{\partial\rho}=\sum_{i=1}^{N}\frac{\partial\ell_i}{\partial\rho}.$$

10.3.5 Programmation

Pour procéder à l'optimisation de la log vraisemblance, il est pratique de procéder aux changements de paramètres suivants :

$$h=\exp\left(\gamma_1\right)>0,\ \rho=\sin\left(\gamma_2\right)\in\left[-1,+1\right],$$

on obtient alors les dérivées de la manière suivante :

$$\frac{\partial\ell_i}{\partial\gamma_1}=\frac{\partial\ell_i}{\partial h}\frac{\partial h}{\partial\gamma_1}=\frac{\partial\ell_i}{\partial h}\exp\left(\gamma_1\right),$$

et

$$\frac{\partial\ell_i}{\partial\gamma_2}=\frac{\partial\ell_i}{\partial\rho}\frac{\partial\rho}{\partial\gamma_1}=\frac{\partial\ell_i}{\partial\rho}\cos\left(\gamma_2\right).$$

Pour les valeurs initiales, on utilise la régression sur les données quantitatives observables :

$$\left\{\forall i|y_1=1\right\},\ \ y_{2i}=X_{2i}b_2+c\times\widehat{\lambda}_{1i}+v_{2i},$$

avec :

$$\widehat{\lambda}_{1i} = \frac{\varphi\left(X_{1i}\widehat{\beta}_1\right)}{\Phi\left(X_{1i}\widehat{\beta}_1\right)},\ c = \rho \times \sigma_2.$$

A la suite de Gouriéroux (1989), on prend comme valeur initiale

$$\widehat{\sigma}_2^2 = \frac{1}{N_1} \sum_{\forall i | y_1 = 1} \left(\widehat{v}_{2i}^2 + \widehat{c}^2 \widehat{\lambda}_{1i} \left(X_{1i}\widehat{\beta}_1 + \widehat{\lambda}_{1i} \right) \right),$$

ce qui implique :

$$\widehat{h} = \frac{1}{\sqrt{\widehat{\sigma}_2^2}} \ \text{ donc } \ \widehat{\gamma}_1 = \ln \widehat{h},$$

ainsi que :

$$\widehat{\rho} = \frac{\widehat{c}}{\widehat{\sigma}_2} = \widehat{h} \times \widehat{c} \ \text{ donc } \ \widehat{\gamma}_2 = \sin^{-1}\left(\widehat{\rho}\right).$$

CHAPITRE 11

Estimation de modèles à plusieurs équations

11.1 Estimation de la forme réduite

Pour fixer les idées, on cherche à estimer le système à deux équations suivant :

$$\left\{ \begin{array}{rcl} y_1^* & = & a_{12}\, y_2^* + X_1 b_1 + u_1 \\ y_2^* & = & a_{21}\, y_1^* + X_2 b_2 + u_2 \end{array} \right. \tag{11.1}$$

en résolvant ce système par rapport aux variables expliquées (y_1^*, y_2^*) en fonction des variables explicatives et des perturbations, on obtient la forme réduite du modèle :

$$\left\{ \begin{array}{rcl} y_1^* & = & X\pi_1 + v_1 \\ y_2^* & = & X\pi_2 + v_2 \end{array} \right.$$

où X est la matrice de toutes les variables explicatives et :

$$v_1 = \frac{u_1 + a_{12} u_2}{1 - a_{12}\, a_{21}}, \quad v_2 = \frac{a_{21} u_1 + u_2}{1 - a_{12}\, a_{21}}. \tag{11.2}$$

On voit que ce système peut être estimé très simplement, équation par équation, puisqu'il n'y a plus de variable qualitative endogène dans les membres de droite des équations de la forme réduite. Le seul problème consiste à estimer la matrice de covariance globale des ces estimateurs obtenus séparément.

Il est facile de voir que les estimateurs de la forme réduite $(\widehat{\pi}_1, \widehat{\pi}_2)$ peuvent être obtenus par la maximisation d'un objectif de la forme suivante :

$$\widehat{\pi} = \arg \max_{(\pi_1, \pi_2)} \sum_{i=1}^{N} \Psi_1 \left(\pi_1; y_{1i}, X_i\right) + \sum_{i=1}^{N} \Psi_2 \left(\pi_2; y_{2i}, X_i\right) \tag{11.3}$$

En effet, la dérivée par rapport à π_1 ne fait intervenir que la première partie de l'objectif, qui réalise l'estimation par le (pseudo) maximum de vraisemblance, alors que la dérivée par rapport à π_2 ne fait intervenir que la dérivée par rapport à la seconde partie de l'objectif. En conséquence, les conditions du premier ordre de ce problème sont identiques à celles des estimations séparées ce qui implique que les estimateurs obtenus en maximisant l'objectif (11.3) sont numériquement identiques à ceux obtenus par les estimations séparées. Il nous reste à voir comment calculer la matrice de covariance de $\widehat{\pi}$ pour résoudre notre problème. Le problème d'optimisation se réécrit :

$$\widehat{\pi} = \arg\max_{\pi} \sum_{\iota=1}^{N} \Psi\left(\pi; y_i, X_i\right)$$

avec $y = (y_1, y_2)$ et $\Psi\left(\pi; y_i, X_i\right) = \Psi_1\left(\pi_1; y_{1i}, X_i\right) + \Psi_2\left(\pi_2; y_{2i}, X_i\right)$. L'estimateur de la forme réduite est défini par :

$$\sum_{i=1}^{N} \frac{\partial \Psi}{\partial \pi}\left(\widehat{\pi}; y_i, X_i\right) = 0.$$

La matrice de covariance est donnée par le résultat suivant, qui s'applique aux M-estimateurs en général (Gouriéroux et Monfort, 1989) :

$$\sqrt{N}\left(\widehat{\pi} - \pi\right) \xrightarrow{L} \mathrm{N}\left(0, \Sigma\right),$$

avec

$$\begin{aligned}
\Sigma &= J^{-1} I J^{-1}, \\
J &= \mathrm{E}\left[-\frac{\partial^2 \Psi}{\partial \pi \partial \pi'}\left(\pi; y, X\right)\right], \\
I &= \mathrm{E}\left[\frac{\partial \Psi}{\partial \pi}\left(\pi; y, X\right) \frac{\partial \Psi}{\partial \pi}\left(\pi; y, X\right)\right].
\end{aligned}$$

Dans la pratique, on estimera ces quantités par :

$$\widehat{J} = -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial^2 \Psi}{\partial \pi \partial \pi'}\left(\widehat{\pi}; y_i, X_i\right)$$

$$\text{et} \quad \widehat{I} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \Psi}{\partial \pi}\left(\widehat{\pi}; y_i, X_i\right) \frac{\partial \Psi}{\partial \pi'}\left(\widehat{\pi}; y_i, X_i\right).$$

On en déduit les remarques importantes suivantes :

1. Les dérivées secondes croisées entre équations sont toutes nulles puisque le paramètre d'une équation de la forme réduite n'apparaît que dans cette équation.

2. De la première remarque, on déduit le résultat suivant :

$$\Omega_{\hat{\pi}} = \begin{pmatrix} J_{11}^{-1} & 0 \\ 0 & J_{22}^{-1} \end{pmatrix} \begin{pmatrix} I_{11} & I_{12} \\ I_{12}' & I_{22} \end{pmatrix} \begin{pmatrix} J_{11}^{-1} & 0 \\ 0 & J_{22}^{-1} \end{pmatrix} = \begin{pmatrix} J_{11}^{-1} I_{11} J_{11}^{-1} & J_{11}^{-1} I_{12} J_{22}^{-1} \\ J_{22}^{-1} I_{12}' J_{11}^{-1} & J_{22}^{-1} I_{22} J_{22}^{-1} \end{pmatrix}.$$

3. Les estimateurs des matrices de covariance asymptotiques de chaque équation pris séparément, qui sont situées sur la diagonale, sont identiques à ceux du pseudo maximum de vraisemblance.

4. La covariance asymptotique entre les estimateurs des deux équations est donnée par :

$$\text{Covas}\left(\sqrt{N}\left(\widehat{\pi}_1 - \pi_1\right), \sqrt{N}\left(\widehat{\pi}_2 - \pi_2\right)\right) = J_{11}^{-1} I_{12} J_{22}^{-1}.$$

5. Il faut donc sauvegarder les dérivées premières individu par individu pour pouvoir estimer la matrice de covariance de l'estimateur de la forme réduite. La seule nouveauté est donc la matrice I_{12}, que l'on estimera par :

$$\widehat{I}_{12} = \frac{1}{N} \sum_{i=1}^{N} \frac{\partial \Psi_1}{\partial \pi_1}\left(\widehat{\pi}_1; y_{1i}, X_{1i}\right) \frac{\partial \Psi_2}{\partial \pi_2'}\left(\widehat{\pi}_2; y_{2i}, X_{2i}\right),$$

et que l'on obtient par un simple produit croisé des matrices des dérivées premières individuelles.

11.2 Estimation de la forme structurelle

En reportant les formes réduites de (y_1^*, y_2^*) dans le système (11.1) on obtient les identités suivantes :

$$\begin{aligned} X\pi_1 + v_1 &= a_{12}\left(X\pi_2 + v_2\right) + X_1 b_1 + u_1 \\ X\pi_2 + v_2 &= a_{21}\left(X\pi_1 + v_1\right) + X_2 b_2 + u_2 \end{aligned}$$

en prenant l'espérance mathématique du système précédent, on obtient les égalités :

$$\left\{ \begin{array}{lcl} X\pi_1 & = & X\pi_2\, a_{12} + X_1 b_1 \\ X\pi_2 & = & X\pi_1\, a_{21} + X_2 b_2 \end{array} \right. \tag{11.4}$$

Pour obtenir une relation entre les paramètres du modèle, on introduit les matrices d'exclusion E_1 et E_2, définies de la manière suivante :

$$X_1 = X\,E_1, \quad X = X\,E_2.$$

Ces matrices résument les contraintes qui permettent d'identifier le modèle, c'est-à-dire de remonter de la forme réduite du modèle à sa forme structurelle. On obtient la propriété suivante :

$$\left\{ \begin{array}{rcl} X\left(\pi_1 - \pi_2\, a_{12} - E_1 b_1\right) & = & 0 \\ X\left(\pi_2 - \pi_1\, a_{21} - E_2 b_2\right) & = & 0 \end{array} \right. \Rightarrow \left\{ \begin{array}{rcl} \pi_1 - \pi_2\, a_{12} - E_1 b_1 & = & 0 \\ \pi_2 - \pi_1\, a_{21} - E_2 b_2 & = & 0 \end{array} \right. \tag{11.5}$$

car X est de plein rang colonne. Les relations (11.5) s'appelent les contraintes identifiantes. La méthode des moindres carrés asymptotiques permet d'estimer la forme structurelle du modèle à partir d'un estimateur convergent et asymptotiquement normal (CAN) de la forme réduite. On note cet estimateur :

$$\sqrt{N}\left(\hat{\pi} - \pi\right) \xrightarrow{L} \mathrm{N}\left(0, \Omega_{\hat{\pi}}\right).$$

Les équations auxilliaires sont définies par :

$$\left\{ \begin{array}{rcl} \hat{\pi}_1 & = & \hat{\pi}_2\, a_{12} + E_1 b_1 + \omega_1 \\ \hat{\pi}_2 & = & \hat{\pi}_1\, a_{21} + E_2 b_2 + \omega_2 \end{array} \right. \tag{11.6}$$

où $\omega = (\omega_1, \omega_2)'$ est un terme d'erreur qui vérifie $\operatorname{Plim} \sqrt{N}\omega = 0$. Ce système peut être estimé en deux étapes. Une première étape set à estimer la matrice de covariance asymptotique de ω; la seconde étape sert à obtenir l'estimateur optimal. On estime d'abord la relation suivante par les moindres carrés ordinaires :

$$\underbrace{\begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \end{pmatrix}}_{\hat{\pi}} = \underbrace{\begin{pmatrix} \hat{\pi}_2 & E_1 & 0 & 0 \\ 0 & 0 & \hat{\pi}_1 & E_2 \end{pmatrix}}_{\hat{H}} \underbrace{\begin{pmatrix} a_{12} \\ b_1 \\ a_{21} \\ b_2 \end{pmatrix}}_{\beta} + \underbrace{\begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}}_{\omega},$$

on obtient l'estimateur suivant :

$$\hat{\beta} = \left(\hat{H}'\hat{H}\right)^{-1} \hat{H}'\hat{\pi}.$$

Cet estimateur de première étape permet de calculer la variance de ω de la manière suivante :

$$\omega = \underbrace{\left[\begin{pmatrix} 1 & -a_{12} \\ -a_{21} & 1 \end{pmatrix} \otimes \mathrm{I}_k\right]}_{A} \hat{\pi} + \begin{pmatrix} E_1 b_1 \\ E_2 b_2 \end{pmatrix},$$

ce qui implique :

$$\mathrm{V}(\omega) = A\,\mathrm{V}(\hat{\pi})\,A',$$

on obtient un estimateur de cette variance en remplaçant (a_{12}, a_{21}) par leurs estimations :

$$\hat{A} = \begin{pmatrix} 1 & -\hat{a}_{12} \\ -\hat{a}_{21} & 1 \end{pmatrix} \otimes \mathrm{I}_k \quad \text{et} \quad \widehat{\mathrm{V}}(\omega) = \hat{A}\,\mathrm{V}(\hat{\pi})\,\hat{A}'.$$

L'estimateur optimal β^* est obtenu en appliquant les moindres carrés généralisés à la relation (11.6) :

$$\beta^* = \left(\hat{H}'\widehat{\mathrm{V}}(\omega)^{-1}\hat{H}\right)^{-1}\hat{H}'\widehat{\mathrm{V}}(\omega)^{-1}\hat{\pi}, \tag{11.7}$$

et sa matrice de covariance peut être estimée par :

$$\mathrm{V}(\beta^*) = \left(\hat{H}'\,\mathrm{V}^*(\omega)^{-1}\hat{H}\right)^{-1}$$

avec

$$\mathrm{V}^*(\omega) = A^*\,\mathrm{V}(\hat{\pi})\,A^{*\prime} \quad \text{et} \quad A^* = \begin{pmatrix} 1 & -a_{12}^* \\ -a_{21}^* & 1 \end{pmatrix} \otimes \mathrm{I}_k\,.$$

On peut effectuer une troisième itération en remplaçant $\widehat{\mathrm{V}}(\omega)$ par $\mathrm{V}^*(\omega)$ dans la relation (11.7) .

Annexe A

Moments empiriques et moments théoriques

A.1 Moments empiriques des vecteurs

Le but de cette section est de se familiariser avec les notations de calcul matriciel, car c'est sous cette forme qu'apparaissent le plus souvent les moments empiriques. Il faut donc savoir les simplifier quand on les recontre dans une expression.

A.1.1 Moyenne arithmétique

La moyenne arithmétique d'un vecteur colonne $z = (z_1, z_2, ..., z_N)'$ peut se trouver sous les formes équivalentes suivantes :

$$\overline{z} = \frac{z'e}{e'e} = \frac{z'e}{N} = \frac{1}{N}\sum_{i=1}^{N} z_i,$$

car on a :

$$z'e = (z_1, z_2, ..., z_N)\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = z_1 + z_2 + ... + z_N = \sum_{i=1}^{N} z_i,$$

et :

$$e'e = (1, 1, ..., 1)\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \underbrace{1 + 1 + ... + 1}_{N \text{ fois}} = N.$$

A.1.2 Variance empirique

La variance empirique de la série z, notée $\mathrm{V}_e(z)$, peut se trouver sous les formes équivalentes :

$$\begin{aligned}\mathrm{V}_e(z) &= \frac{1}{N}\sum_{i=1}^{N}(z_i-\overline{z})^2 \\ &= \frac{1}{N}\sum_{i=1}^{N}z_i^2-(\overline{z})^2 \\ &= \frac{1}{N}(z-\overline{z}e)'(z-\overline{z}e), \\ &= \frac{z'z}{N}-(\overline{z})^2\end{aligned}$$

car

$$z-\overline{z}e=\begin{pmatrix}z_1\\z_2\\\vdots\\z_N\end{pmatrix}-\begin{pmatrix}\overline{z}\\\overline{z}\\\vdots\\\overline{z}\end{pmatrix}=\begin{pmatrix}z_1-\overline{z}\\z_2-\overline{z}\\\vdots\\z_N-\overline{z}\end{pmatrix},$$

ce qui implique :

$$\begin{aligned}(z-\overline{z}e)'(z-\overline{z}e) &= (z_1-\overline{z},z_2-\overline{z},...,z_N-\overline{z})\begin{pmatrix}z_1-\overline{z}\\z_2-\overline{z}\\\vdots\\z_N-\overline{z}\end{pmatrix} \\ &= (z_1-\overline{z})^2+(z_2-\overline{z})^2+...+(z_N-\overline{z})^2 \\ &= \sum_{i=1}^{N}(z_i-\overline{z})^2.\end{aligned}$$

En posant $\overline{z}=0$, on trouve :

$$z'z=\sum_{i=1}^{N}z_i^2.$$

A.1.3 Ecart-type empirique

Il s'agit simplement de la racine carrée de la variance empirique. On le note :

$$\sigma_e(x)=\sqrt{\mathrm{V}_e(x)}.$$

A.1.4 Covariance empirique

La covariance empirique entre le vecteur $z = (z_1, z_2, ..., z_N)'$ et le vecteur $x = (x_1, x_2, ..., x_N)'$, $\mathrm{Cov}_e(z, x)$, s'écrit :

$$\begin{aligned}
\mathrm{Cov}_e(x, z) &= \frac{1}{N}\sum_{i=1}^{N}(z_i - \overline{z})(x_i - \overline{x}) \\
&= \frac{1}{N}\sum_{i=1}^{N} z_i x_i - \overline{z}\,\overline{x} \\
&= \frac{1}{N}(z - \overline{z}e)'(x - \overline{x}e) \\
&= \frac{z'x}{N} - \overline{z}\,\overline{x}.
\end{aligned}$$

En effet :

$$\begin{aligned}
(z - \overline{z}e)'(x - \overline{x}e) &= (z_1 - \overline{z}, z_2 - \overline{z}, ..., z_N - \overline{z}) \begin{pmatrix} x_1 - \overline{x} \\ x_2 - \overline{x} \\ \vdots \\ x_N - \overline{x} \end{pmatrix} \\
&= (z_1 - \overline{z})(x_1 - \overline{x}) + ... + (z_N - \overline{z})(x_N - \overline{x}) \\
&= \sum_{i=1}^{N}(z_i - \overline{z})(x_i - \overline{x}).
\end{aligned}$$

En posant $\overline{z} = 0 = \overline{x}$ dans l'expression précédente, on a :

$$z'x = \sum_{i=1}^{N} z_i x_i.$$

On remarque de plus que lorsque $z = x$:

$$\begin{aligned}
\mathrm{Cov}_e(x, x) &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x}) \\
&= \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x})^2 \\
&= \mathrm{V}_e(x).
\end{aligned}$$

A.1.5 Corrélation empirique

Le coefficient de corrélation linéaire empirique entre les séries z et x, noté $\rho_e(x,z)$ est défini par :

$$\rho_e(x,z) = \frac{\mathrm{Cov}_e(x,z)}{\sqrt{\mathrm{V}_e(x)\,\mathrm{V}_e(z)}} = \frac{\mathrm{Cov}_e(x,z)}{\sigma_e(x)\,\sigma_e(z)}.$$

Il peut donc prendre différentes formes en fonction des expressions que nous avons vu plus haut. On peut faire apparaître son expression dans la définition des différents estimateurs.

A.2 Moments empiriques des matrices

A.2.1 Moyenne arithmétique

On considère maintenant une matrice X de dimension (N,p). Chaque ligne de X correspond à une observation et chaque colonne de X corrrespond à une variable. On note ces variables $X = \left(X^{(1)}|X^{(2)}|\ldots|X^{(p)}\right)$. On a :

$$\overline{X} = \underbrace{\frac{X'e}{N}}_{(p,1)} = \frac{1}{N}\begin{pmatrix} X^{(1)\prime} \\ X^{(2)\prime} \\ \vdots \\ X^{(p)\prime} \end{pmatrix} e = \frac{1}{N}\begin{pmatrix} X^{(1)\prime}e \\ X^{(2)\prime}e \\ \vdots \\ X^{(p)\prime}e \end{pmatrix} = \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{pmatrix}.$$

A.2.2 Matrice de covariance empirique

Contrairement au cas univarié, on définit une matrice qui contient à la fois les variances et les covariances des variables. Les variances sont sur la diagonale de la matrice de covariance. On a :

$$\mathrm{V}_e(X) = \frac{X'X}{N} - \overline{X}\,\overline{X}'$$

On peut définir la matrice des produits croisés des variables explicatives $X'X$ à partir du modèle écrit par observations ou par variables. Selon le contexte une expression peut s'avérer plus pratique que l'autre, et il faut pouvoir passer facilement entre les différentes expressions.

Par rapport aux variables, on a:

$$\underset{(N,p)(N,p)}{X'X} = \begin{pmatrix} X^{(1)\prime} \\ X^{(2)\prime} \\ \vdots \\ X^{(p)\prime} \end{pmatrix} \left(X^{(1)} | X^{(2)} | ... | X^{(p)} \right)$$

$$= \begin{pmatrix} X^{(1)\prime}X^{(1)} & X^{(1)\prime}X^{(2)} & ... & X^{(1)\prime}X^{(p)} \\ X^{(1)\prime}X^{(2)} & X^{(2)\prime}X^{(2)} & ... & X^{(2)\prime}X^{(p)} \\ \vdots & \vdots & \ddots & \vdots \\ X^{(p)\prime}X^{(1)} & X^{(p)\prime}X^{(2)} & ... & X^{(p)\prime}X^{(p)} \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{N} x_{i1}^2 & \sum_{i=1}^{N} x_{i1}x_{i2} & ... & \sum_{i=1}^{N} x_{i1}x_{ip} \\ \sum_{i=1}^{N} x_{i1}x_{i2} & \sum_{i=1}^{N} x_{i2}^2 & ... & \sum_{i=1}^{N} x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_{i1}x_{ip} & \sum_{i=1}^{N} x_{i2}x_{ip} & ... & \sum_{i=1}^{N} x_{ip}^2 \end{pmatrix}$$

La matrice des moments empiriques non centrés de X est définie par :

$$\frac{X'X}{N} = \begin{pmatrix} N^{-1}\sum_{i=1}^{N} x_{i1}^2 & ... & N^{-1}\sum_{i=1}^{N} x_{i1}x_{ip} \\ N^{-1}\sum_{i=1}^{N} x_{i1}x_{i2} & ... & N^{-1}\sum_{i=1}^{N} x_{i2}x_{ip} \\ \vdots & \ddots & \vdots \\ N^{-1}\sum_{i=1}^{N} x_{i1}x_{ip} & ... & N^{-1}\sum_{i=1}^{N} x_{ip}^2 \end{pmatrix}$$

On en déduit la matrice de covariance empirique :

$$\mathrm{V}_e(X) = \begin{pmatrix} N^{-1}\sum_{i=1}^{N} x_{i1}^2 & ... & N^{-1}\sum_{i=1}^{N} x_{i1}x_{ip} \\ N^{-1}\sum_{i=1}^{N} x_{i1}x_{i2} & ... & N^{-1}\sum_{i=1}^{N} x_{i2}x_{ip} \\ \vdots & \ddots & \vdots \\ N^{-1}\sum_{i=1}^{N} x_{i1}x_{ip} & ... & N^{-1}\sum_{i=1}^{N} x_{ip}^2 \end{pmatrix} - \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{pmatrix} \begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \cdots & \overline{x}_p \end{pmatrix}$$

$$= \begin{pmatrix} N^{-1}\sum_{i=1}^{N} x_{i1}^2 & ... & N^{-1}\sum_{i=1}^{N} x_{i1}x_{ip} \\ N^{-1}\sum_{i=1}^{N} x_{i1}x_{i2} & ... & N^{-1}\sum_{i=1}^{N} x_{i2}x_{ip} \\ \vdots & \ddots & \vdots \\ N^{-1}\sum_{i=1}^{N} x_{i1}x_{ip} & ... & N^{-1}\sum_{i=1}^{N} x_{ip}^2 \end{pmatrix} - \begin{pmatrix} \overline{x}_1^2 & ... & \overline{x}_1\overline{x}_p \\ \overline{x}_1\overline{x}_2 & ... & \overline{x}_2\overline{x}_p \\ \vdots & \ddots & \vdots \\ \overline{x}_1\overline{x}_p & ... & \overline{x}_p^2 \end{pmatrix}$$

$$= \begin{pmatrix} N^{-1}\sum_{i=1}^{N} x_{i1}^2 - \overline{x}_1^2 & ... & N^{-1}\sum_{i=1}^{N} x_{i1}x_{ip} - \overline{x}_1\overline{x}_p \\ N^{-1}\sum_{i=1}^{N} x_{i1}x_{i2} - \overline{x}_1\overline{x}_2 & ... & N^{-1}\sum_{i=1}^{N} x_{i2}x_{ip} - \overline{x}_2\overline{x}_p \\ \vdots & \ddots & \vdots \\ N^{-1}\sum_{i=1}^{N} x_{i1}x_{ip} - \overline{x}_1\overline{x}_p & ... & N^{-1}\sum_{i=1}^{N} x_{ip}^2 - \overline{x}_p^2 \end{pmatrix}$$

On obtient donc finalement :

$$\mathrm{V}_e(X) = \begin{pmatrix} \mathrm{V}_e(x_1) & \mathrm{Cov}_e(x_1, x_2) & ... & \mathrm{Cov}_e(x_1, x_p) \\ \mathrm{Cov}_e(x_1, x_2) & \mathrm{V}_e(x_2) & ... & \mathrm{Cov}_e(x_2, x_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}_e(x_1, x_p) & \mathrm{Cov}_e(x_2, x_p) & ... & \mathrm{V}_e(x_p) \end{pmatrix}$$

Par rapport aux observations. La matrice de covariance empirique peut s'écrire :

$$\mathrm{V}_e(X) = \frac{1}{N}\sum_{i=1}^{N} X_i' X_i - \overline{X}\,\overline{X}'$$

on a :

$$\sum_{i=1}^{N} X_i' X_i = \sum_{i=1}^{N} (x_{i1}, x_{i2}, ..., x_{ip}) \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ip} \end{pmatrix}$$

$$= \sum_{i=1}^{N} \begin{pmatrix} x_{i1}^2 & x_{i1}x_{i2} & ... & x_{i1}x_{ip} \\ x_{i1}x_{i2} & x_{i2}^2 & ... & x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i1}x_{ip} & x_{i2}x_{ip} & ... & x_{ip}^2 \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{N} x_{i1}^2 & \sum_{i=1}^{N} x_{i1}x_{i2} & ... & \sum_{i=1}^{N} x_{i1}x_{ip} \\ \sum_{i=1}^{N} x_{i1}x_{i2} & \sum_{i=1}^{N} x_{i1}^2 & ... & \sum_{i=1}^{N} x_{i2}x_{ip} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} x_{i1}x_{ip} & \sum_{i=1}^{N} x_{i2}x_{ip} & ... & \sum_{i=1}^{N} x_{ip}^2 \end{pmatrix}$$

$$= X'X$$

On retrouve donc le même résultat que précédemment. De même pour les produits croisés entre les variables explicatives et la variable expliquée, on a :

$$\underset{(N,p)}{X'} \underset{(N,1)}{y} = \begin{pmatrix} X^{(1)\prime} \\ X^{(2)\prime} \\ \vdots \\ X^{(p)\prime} \end{pmatrix} y = \begin{pmatrix} X^{(1)\prime} y \\ X^{(2)\prime} y \\ \vdots \\ X^{(p)\prime} y \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_{i1} y_i \\ \sum_{i=1}^{N} x_{i2} y_i \\ \vdots \\ \sum_{i=1}^{N} x_{ip} y_i \end{pmatrix} = \sum_{i=1}^{N} X_i' y_i.$$

Les moments centrés donnent donc :

$$\mathrm{Cov}_e\,(X, y) = \frac{1}{N} \sum_{i=1}^{N} X_i' y_i - \overline{X}\,\overline{y} = \frac{X'y}{N} - \overline{X}\,\overline{y}.$$

Le vecteur correspondant est égal à :

$$\mathrm{Cov}_e\,(X, y) = \begin{pmatrix} N^{-1} \sum_{i=1}^{N} x_{i1} y_i \\ N^{-1} \sum_{i=1}^{N} x_{i2} y_i \\ \vdots \\ N^{-1} \sum_{i=1}^{N} x_{ip} y_i \end{pmatrix} - \begin{pmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{pmatrix} \overline{y}$$

$$= \begin{pmatrix} N^{-1}\sum_{i=1}^{N} x_{i1}y_i \\ N^{-1}\sum_{i=1}^{N} x_{i2}y_i \\ \vdots \\ N^{-1}\sum_{i=1}^{N} x_{ip}y_i \end{pmatrix} - \begin{pmatrix} \overline{x}_1\overline{y} \\ \overline{x}_2\overline{y} \\ \vdots \\ \overline{x}_p\overline{y} \end{pmatrix}$$

$$= \begin{pmatrix} N^{-1}\sum_{i=1}^{N} x_{i1}y_i - \overline{x}_1\overline{y} \\ N^{-1}\sum_{i=1}^{N} x_{i2}y_i - \overline{x}_2\overline{y} \\ \vdots \\ N^{-1}\sum_{i=1}^{N} x_{ip}y_i - \overline{x}_p\overline{y} \end{pmatrix}$$

$$= \begin{pmatrix} \mathrm{Cov}_e\,(x_1, y) \\ \mathrm{Cov}_e\,(x_2, y) \\ \vdots \\ \mathrm{Cov}_e\,(x_p, y) \end{pmatrix}.$$

Sous certaines conditions, les moments empiriques que nous venons de voir convergent en probabilité vers les moments théoriques correspondants. Ce point est examiné dans la section suivante.

A.3 Convergence en probabilité

DÉFINITION A.1 *Soit $\widehat{b}_N$ une variable aléatoire dont la réalisation dépend du nombre d'observations disponibles dans un échantillon (noté N). On dit que cette suite de variables aléatoires $\widehat{b}_N$ converge en probabilité vers une valeur b lorsque le nombre d'observations N tend vers l'infini, si elle vérifie la propriété suivante :*

$$\forall \varepsilon > 0,\ \Pr\left[\left|\widehat{b}_N - b\right| > \varepsilon\right] \underset{N\to+\infty}{\longrightarrow} 0.$$

La convergence en probabilité de $\widehat{b}_N$ vers b est notée de manière abrégée par l'expression :

$$\mathrm{Plim}\,\widehat{b}_N = b,$$

où Plim *est l'abréviation de "probability limit" (i.e., limite en probabilité). Elle s'écrit également :*

$$\widehat{b}_N \underset{N\to+\infty}{\xrightarrow{P}} b.$$

Cette définition signifie que l'évènement "$\widehat{b}_N$ s'écarte de b d'une distance supérieure à ε" est de probabilité nulle (i.e., impossible) lorsque $N \to +\infty$. Cette propriété s'étend à certaines fonctions de $\widehat{b}_N$, comme le montre le théorème suivant.

THÉORÈME A.1 *[Slutsky]*
Soit $\widehat{b}_N$ une suite de variables aléatoires qui converge en probabilité vers b :

$$\operatorname{Plim}\ \widehat{b}_N = b,$$

et soit $g(.)$ une fonction continue définie au point b. On a :

$$\operatorname{Plim} g\left(\widehat{b}_N\right) = g\left(\operatorname{Plim} \widehat{b}_N\right) = g(b).$$

Les définitions précédentes et le théorème de Slutsky s'étendent au cas vectoriel en raisonnant composante par composante. En particulier le théorème de Slutsky permet de simplifier considérablement le calcul des limites en probabilités. Prenons deux estimateurs convergents, $\widehat{a}$ d'un paramètre a et $\widehat{b}$ d'un paramètre b. On a :

$$\operatorname{Plim} \widehat{a} + \widehat{b} = \operatorname{Plim} \widehat{a} + \operatorname{Plim} \widehat{b} = a + b,$$

car la fonction $g(a, b) = a + b$ est continue et les estimateurs convergent en probabilité. De même, en utilisant les fonctions $g(a, b) = ab$, $g(a, b) = a/b$ (pour $b \neq 0$) on obtient les propriétés :

$$\operatorname{Plim} \widehat{a}\,\widehat{b} = \operatorname{Plim} \widehat{a}\ \operatorname{Plim} \widehat{b} = a\,b,$$
$$\operatorname{Plim} \frac{\widehat{a}}{\widehat{b}} = \frac{\operatorname{Plim} \widehat{a}}{\operatorname{Plim} \widehat{b}} = \frac{a}{b}, \quad b \neq 0.$$

A.4 Inégalité de Bienaymé-Chebichev

Le théorème suivant est très important. Il nous permet notamment de démontrer la loi des grands nombres et le fait que la convergence en moyenne quadratique implique la convergence en probabilité...en une seule ligne.

THÉORÈME A.2 *[Inégalité de Bienaymé-Chebichev]*
Soit Z une variable de carré intégrable, on a :

$$\forall \delta > 0, \Pr[|Z| \geq \delta] \leq \frac{1}{\delta^2} \mathrm{E}\left(Z^2\right).$$

PREUVE :
Soit la variable de Bernoulli :

$$D = \begin{cases} 1 & \text{si } |Z| \geq \delta \\ 0 & \text{sinon} \end{cases}$$

son espérance mathématique est égale à :

$$\mathrm{E}(D) = 1 \times \Pr[|Z| \geq \delta] + 0 \times \Pr[|Z| < \delta] = \Pr[|Z| \geq \delta].$$

D'autre part :

1. Si $|Z| \geq \delta$ on a $D = 1$ donc :

$$\frac{|Z|}{\delta} \geq 1 \Rightarrow \frac{Z^2}{\delta^2} \geq D = 1.$$

2. Si $|Z| < \delta$ on a $D = 0$ donc :

$$\frac{|Z|}{\delta} \geq 0 \Rightarrow \frac{Z^2}{\delta^2} \geq D = 0.$$

donc dans tous les cas on a :

$$\begin{aligned} \frac{Z^2}{\delta^2} \geq D &\Rightarrow \mathrm{E}\left(\frac{Z^2}{\delta^2}\right) \geq \mathrm{E}(D) \\ &\Leftrightarrow \frac{1}{\delta^2}\mathrm{E}\left(Z^2\right) \geq \Pr[|Z| \geq \delta]. \end{aligned}$$

□

Remarque A.1 *En posant* $Z = X - \mathrm{E}(X)$, *on obtient l'expression :*

$$\forall \delta > 0, \Pr[|X - \mathrm{E}(X)| \geq \delta] \leq \frac{1}{\delta^2}\mathrm{V}(X),$$

car $\mathrm{V}(X) = \mathrm{E}\left[(X - \mathrm{E}(X))^2\right]$.

La convergence en probabilité est parfois difficile à vérifier directement, on utilise alors une conditions suffisante, qui correspond en fait à la convergence en moyenne quadratique.

DÉFINITION A.2 *Soit* $\widehat{b}_N$ *une variable aléatoire dont la réalisation dépend du nombre d'observations disponibles dans un échantillon (noté* N*). On dit que cette suite de variables aléatoires* $\widehat{b}_N$ *converge en moyenne quadratique vers une valeur* b *lorsque le nombre d'observations* N *tend vers l'infini, si elle vérifie une des deux propriétés équivalentes suivantes :*

1. $\mathrm{E}\left[\left(\widehat{b}_N - b\right)^2\right] \to 0 \quad$ *lorsque* $\quad N \to +\infty.$
2. $\mathrm{E}\left(\widehat{b}_N\right) \to b \quad$ *et* $\quad \mathrm{V}\left(\widehat{b}_N\right) \to 0 \quad$ *lorsque* $\quad N \to +\infty.$

On note ce résultat :

$$\widehat{b}_N \underset{N\to+\infty}{\overset{m.q.}{\longrightarrow}} b.$$

Cette définition porte directement sur la distance entre $\widehat{b}_N$ et b. Elle impose que cette distance s'annule quand le nombre d'observations devient suffisamment grand. L'équivalence entre les deux définitions vient du développement suivant :[1]

$$\begin{aligned}\mathrm{E}\left[\left(\widehat{b}_N - b\right)^2\right] &= \mathrm{V}\left[\widehat{b}_N - b\right] + \left[\mathrm{E}\left(\widehat{b}_N - b\right)\right]^2 \\ &= \mathrm{V}\left(\widehat{b}_N\right) + \left(\mathrm{E}\left(\widehat{b}_N\right) - b\right)^2 \geq 0.\end{aligned}$$

Les deux termes précédents sont positifs ou nuls donc pour que l'expression s'annule lorsque $N \to +\infty$, il faut que l'on ait simultanément $\mathrm{V}\left(\widehat{b}_N\right) \to 0$ et $\mathrm{E}\left(\widehat{b}_N\right) \to b$.

PROPRIÉTÉ A.1 *Soit $\widehat{b}_N$ une suite de variables aléatoires, on a :*

$$\widehat{b}_N \underset{N\to+\infty}{\overset{m.q.}{\longrightarrow}} b \quad \Rightarrow \quad \operatorname{Plim} \widehat{b}_N = b,$$

la convergence en moyenne quadratique implique la convergence en probabilité.

PREUVE :

C'est une conséquence de l'inégalité de Bienaymé-Chebichev. En posant $Z = \widehat{b}_N - b$ et $\delta = \varepsilon > 0$ dans le théorème $[A.2]$, on obtient :

$$\forall \varepsilon > 0,\ 0 \leq \Pr\left[\left|\widehat{b}_N - b\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2}\,\mathrm{E}\left[\left(\widehat{b}_N - b\right)^2\right] \underset{N\to+\infty}{\longrightarrow} 0.$$

□

A.5 La loi faible des grands nombres

Cette section permet de faire le lien entre les moments empiriques que nous avons vu plus haut et la convergence en probabilité que nous venons de voir. Elle signifie que sous certaines conditions, les moments empiriques convergent en probabilité vers les moments théoriques correspondants. On l'appelle loi faible des grands nombres, car la convergence en probabilité est également appelée convergence faible. La version de cette loi que nous utilisons est due à Markov (cf. Petrov 1995, p.134).

[1] On rappelle que : $\mathrm{V}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \Leftrightarrow \mathrm{E}(X^2) = \mathrm{V}(X) + \mathrm{E}(X)^2$. Ici on pose $X = \widehat{b}_n - b$.

THÉORÈME A.3 *[Markov]*

Soit $(X_1, .., X_N)$ *une suite de variables aléatoires qui admettent une espérance mathématique* $\mathrm{E}(X_k) = m_k$ *pour toute valeur de* $k \in \{1, ..., N\}$, *et qui vérifient la propriété suivante :*

$$\frac{1}{N^2}\,\mathrm{V}\left[\sum_{k=1}^{N} X_k\right] \to 0 \quad \textit{lorsque} \quad N \to +\infty,$$

alors

$$\mathrm{Plim}\left[\frac{1}{N}\sum_{k=1}^{N} X_k - \frac{1}{N}\sum_{k=1}^{N} m_k\right] = 0.$$

PREUVE :

Il suffit de poser $Z = N^{-1}\sum_{k=1}^{N}(X_k - m_k)$ dans l'inégalité de Bienaymé-Chebichev (théorème $[A.2]$) :

$$\forall \delta > 0, \Pr\left[\left|\frac{1}{N}\sum_{k=1}^{N} X_k - \frac{1}{N}\sum_{k=1}^{N} m_k\right| \geq \delta\right] \leq \frac{1}{\delta^2 N^2}\,\mathrm{V}\left[\sum_{k=1}^{N} X_k\right] \underset{N\to+\infty}{\longrightarrow} 0.$$

En effet, on a :

$$\mathrm{E}(Z) = \frac{1}{N}\sum_{k=1}^{N}\left[\mathrm{E}(X_k) - m_k\right] = 0$$

$$\mathrm{V}(Z) = \mathrm{V}\left[\frac{1}{N}\sum_{k=1}^{N} X_k - \frac{1}{N}\sum_{k=1}^{N} m_k\right] = \mathrm{V}\left[\frac{1}{N}\sum_{k=1}^{N} X_k\right]$$

car $N^{-1}\sum_{k=1}^{N} m_k$ est une quantité certaine et que l'on a :$\mathrm{V}\left[N^{-1}\sum_{k=1}^{N} X_k\right] = N^{-2}\,\mathrm{V}\left[\sum_{k=1}^{N} X_k\right]$.

□

Une moyenne arithmétique de variable aléatoires converge donc vers la moyenne des espérances mathématiques des variables aléatoires, à condition que la variance de leur moyenne $\mathrm{V}\left[N^{-1}\sum_{k=1}^{N} X_k\right]$ tende vers 0 lorsque $N \to +\infty$.

Exemple A.1 *On considère un échantillon de variables* $(X_1, ..., X_k)$ *indépendantes, d'espérance et de variance constantes :* $\forall k, m_k = m$ *et* $\mathrm{V}(X_k) = \sigma^2$. *Sous hypothèse d'indépendance, on obtient la condition suivante :*

$$\frac{1}{N^2}\,\mathrm{V}\left[\sum_{k=1}^{N} X_k\right] = \frac{1}{N^2}\sum_{k=1}^{N}\mathrm{V}(X_k) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N} \to 0 \quad \textit{quand} \quad N \to +\infty.$$

D'autre part $N^{-1}\sum_{k=1}^{N} m_k = N^{-1}(N \times m) = m$. *On a donc le résultat de convergence suivant :*

$$\operatorname{Plim} \overline{X} = m,$$

la moyenne empirique converge vers l'espérance mathématique commune des variables $(X_1, ..., X_k)$.

Exemple A.2 *On considère un échantillon de variables* $(X_1, ..., X_k)$ *indépendantes de variances différentes et finies :* $\mathrm{V}(X_k) = \sigma_k^2$. *La moyenne arithmétique de ces variances* $N^{-1}\sum_{k=1}^{N} \sigma_k^2 = \overline{\sigma}$ *est également finie. En effet :*

$$\overline{\sigma} \leq \max_{k=1,...,N} \sigma_k^2 \text{ qui est finie.}$$

ce qui implique :

$$\frac{1}{N^2}\mathrm{V}\left[\sum_{k=1}^{N} X_k\right] = \frac{1}{N^2}\sum_{k=1}^{N}\sigma_k^2 = \frac{\overline{\sigma}}{N} \to 0 \quad \textit{quand} \quad N \to +\infty.$$

On en déduit que :

$$\operatorname{Plim} \overline{X} = \operatorname{Plim} \frac{1}{N}\sum_{k=1}^{N} \mathrm{E}(X_k).$$

A.6 Théorème de la limite centrale

Le théorème suivant nous permet de déterminer la loi asymptotique de la plupart de nos estimateurs.

THÉORÈME A.4 (Liapunov) *Soit* $u_1, u_2, ..., u_N$ *une suite de variables aléatoires indépendantes d'espérances mathématiques* $\mathrm{E}(u_i) = \mu_i$ *et de variances respectives* $\mathrm{V}(u_i) = \mathrm{E}(u_i - \mu_i)^2 = \sigma_i^2 \neq 0$, $i = 1, ..., n$. *On suppose également que le moment absolu d'ordre trois existe* $\mathrm{E}|u_i - \mu_i|^3 = \beta_i \ \forall i$. *Soient :*

$$B_N = \left(\sum_{i=1}^{N}\beta_i\right)^{1/3}, D_N = \left(\sum_{i=1}^{N}\sigma_i^2\right)^{1/2},$$

alors, si $\lim B_N/D_N = 0$ *lorsque* $N \to +\infty$, *on a :*

$$\sum_{i=1}^{N}\frac{u_i - \mu_i}{D_N} \underset{N\to+\infty}{\longrightarrow} N(0,1).$$

Annexe B

Algèbre linéaire

B.1 Calcul matriciel

On considère une matrice $A = [A_{ij}]$ de format (m, n).

1. La transposée de A, notée A', est définie par $A' = [A_{ji}]$, on intervertit donc les lignes et les colonnes.

2. A est de plein rang colonne si ses colonnes sont linéairement indépendantes. C'est-à-dire si :

$$\forall \alpha \in \mathbb{R}^n, \quad A\alpha = 0 \Rightarrow \alpha = 0.$$

3. A est de plein rang ligne si ses lignes sont linéairement indépendantes (i.e., si A' est de plein rang colonne).

On considère maintenant deux matrices A de format (m, n) et B de format (r, p).

1. Le produit matriciel de A par B n'existe que si le nombre de colonnes de A est égal au nombre de lignes de B : $n = r$. Dans ce cas, on le note $F = AB$ et il est de format (m, p).

2. Les élements de la matrice produit $F = [F_{ij}] = AB$ sont définis comme les produits scalaires de la $i-$ème ligne de A et de la $j-$ième colonne de B.

3. AB n'est généralement pas égal à BA, le produit matriciel n'est pas commutatif.

4. $A(B + C) = AB + BC$.

5. $(A+B)\,C = AC + BC$.
6. $(AB)' = B'A'$.

On considère maintenant deux matrices *carrées* A de format (m,m) et B de format (r,r).

1. Une matrice est carrée si elle a autant de lignes que de colonnes.
2. Une matrice carrée A est symétrique si $A' = A$.
3. La trace d'une matrice carrée A est définie par la somme de ses élements diagonaux. On la note $\mathrm{tr}\,(A) = \sum_{i=1}^{m} A_{ii}$.
4. $\mathrm{tr}\,(A+B) = \mathrm{tr}\,(A) + \mathrm{tr}\,(B)$.
5. Si ABC est une matrice carrée et si les formats sont compatibles : $\mathrm{tr}\,(ABC) = \mathrm{tr}\,(CAB) = \mathrm{tr}\,(BCA)$. Il n'est pas nécessaire que chaque matrice soit carrée à l'intérieur des produits précédents.
6. Si A est une matrice carrée de plein rang (ligne ou colonne), elle admet une inverse notée A^{-1} telle que $AA^{-1} = A^{-1}A = \mathrm{I}$.
7. Si les matrices A et B sont inversibles : $(AB)^{-1} = B^{-1}A^{-1}$.
8. Une matrice carrée A est idempotente si $A^2 = A$.

B.2 Matrices définies positives

DÉFINITION B.1 *Une matrice A de format (m,m) est semi définie positive lorsque :*

$$\forall \alpha \in I\!R^m,\ s\,(\alpha, A) = \alpha' A \alpha \geq 0.$$

DÉFINITION B.2 *Une matrice A de format (m,m) est définie positive lorsque :*

$$\forall \alpha \in I\!R^m,\ \alpha \neq 0,\ s\,(\alpha, A) = \alpha' A \alpha > 0.$$

La propriété suivante est utile pour comparer les variances des différents estimateurs.

PROPRIÉTÉ B.1 *Soit $X_{(n,p)}$ une matrice quelconque, alors $X'X$ est semi définie positive.*

PREUVE :
En posant $A = X'X$, on obtient :

$$s\,(\alpha, X'X) = \alpha' X'X \alpha = \underbrace{(X\alpha)'}_{(1,n)} \underbrace{(X\alpha)}_{(n,1)} = \|X\alpha\|^2 \geq 0.$$

□

La propriété suivante est utile pour montrer l'existence de certains estimateurs.

PROPRIÉTÉ B.2 *Soit $X_{(n,p)}$ une matrice de plein rang colonne,* $\text{rang}\,(X) = p$*, alors $X'X$ est définie positive (donc de rang égal à p).*

PREUVE :
La matrice X est de plein rang colonne :

$$\forall \alpha \in \mathbb{R}^p, \quad X'\alpha = 0 \Rightarrow \alpha = 0$$

donc $\|X\alpha\|^2$ ne peut être nul que dans le cas $\alpha = 0$. En conséquence :

$$\forall \alpha \in \mathbb{R}^p,\ \alpha \neq 0,\ \|X\alpha\|^2 > 0.$$

□

B.3 Produits de Kronecker

Soient deux matrices $A = [A_{ij}]$ de format (a, b) et $B = [B_{ij}]$ de format (c, d). Le produit de Kronecker de la matrice A par la matrice B, noté $A \otimes B$, donne une matrice $F = [F_{ij}]$ de format (ac, bd). Cette matrice est définie par :

$$F = [A_{ij}B] = \begin{pmatrix} A_{1,1}B & A_{1,2}B & \cdots & A_{1,b}B \\ A_{2,1}B & A_{2,2}B & \cdots & A_{2,b}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{a,1}B & A_{a,2}B & \cdots & A_{a,b}B \end{pmatrix},$$

chaque élément originel de la matrice A se voit multiplié par la totalité de la matrice B. Chacun des éléments de la matrice ci-dessus est donc de dimensions égales à celles de B, et C est de format (ac, bd). Les propriétés suivantes sont valables sous réserve que les formats des matrices autorisent les multiplications matricielles indiquées.

1. Dans le cas général $(A \otimes B)$ n'est pas égal à $(B \otimes A)$, le produit de Kronecker n'est donc pas commutatif.
2. $0 \otimes A = 0$.
3. $A \otimes 0 = 0$, mais attention, le format de ce 0 n'est pas nécessairement le même que celui de la propriété précédente.
4. $A \otimes (B + C) = A \otimes B + A \otimes C$.

5. $(A+B)\otimes C = A\otimes B + B\otimes C$.
6. $\forall\,(x,y)\in\mathbb{R}^2,\ (xA)\otimes(yB) = xy\,(A\otimes B)$.
7. $(A\otimes B)\,(C\otimes D) = (AC\otimes BD)$.
8. $(A\otimes B)' = (A'\otimes B')$.
9. Si A et B sont inversibles : $(A\otimes B)^{-1} = A^{-1}\otimes B^{-1}$.
10. $\operatorname{tr}(A\otimes B) = \operatorname{tr} A.\operatorname{tr} B$.

Annexe C

La loi normale

La loi normale centrée réduite admet pour densité :

$$\phi(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\},$$

elle est d'espérance nulle et variance unitaire. Plus généralement, on peut définir une loi normale d'espérance m et de variance σ^2 en définissant la variable suivante :

$$Y = g(U) = m + \sigma U,\ U \rightsquigarrow \mathrm{N}(0,1),$$

la réciproque de la fonction est :

$$g^{-1}(y) = \frac{y-m}{\sigma},$$

et la densité de Y est donnée par :

$$\begin{aligned} f(y) &= \phi\left(g^{-1}(y)\right)\left|\frac{\mathrm{d}g^{-1}(y)}{\mathrm{d}y}\right| \\ &= \frac{1}{\sigma}\phi\left(\frac{y-m}{\sigma}\right). \end{aligned}$$

La fonction génératrice des moments de la loi normale centrée réduite est définie par :

$$\begin{aligned} M(s) &= \mathrm{E}\left(e^{sU}\right) \\ &= \int_{-\infty}^{+\infty} \exp(su)\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{u^2}{2}\right)\mathrm{d}u \\ &= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(u^2-2su\right)\right)\mathrm{d}u \end{aligned}$$

C.1 Loi normale univariée tronquée

On considère une loi normale de densité :

$$f(y) = \frac{1}{\sigma}\phi\left(\frac{y-m}{\sigma}\right),$$

et on cherche la densité de la loi tronquée en un seuil c. La densité de la loi tronquée est définie par :

$$f(y|y>c) = \frac{f(y)\,1_{(y>c)}}{\Pr(y>c)}.$$

Pour calculer l'espérance mathématique de la loi tronquée, il nous faut la quantité :

$$\begin{aligned} I &= \int_{-\infty}^{+\infty} y f(y)\,1_{(y>c)}\,\mathrm{d}y \\ &= \int_{c}^{+\infty} y\frac{1}{\sigma}\phi\left(\frac{y-m}{\sigma}\right)\mathrm{d}y, \end{aligned}$$

on fait le changement de variable :

$$z = \frac{y-m}{\sigma},$$

ce qui implique :

$$\lim_{y\to c} z = \frac{y-c}{\sigma},\ \lim_{y\to+\infty} z = +\infty,\ y = m+\sigma z \ \text{ et } \ \mathrm{d}y = \sigma\mathrm{d}z,$$

on obtient donc :

$$\begin{aligned} I &= \int_{(c-m)/\sigma}^{+\infty} (m+\sigma z)\,\phi(z)\,\mathrm{d}z \\ &= m\int_{(c-m)/\sigma}^{+\infty} \phi(z)\,\mathrm{d}z + \sigma\int_{(c-m)/\sigma}^{+\infty} z\phi(z)\,\mathrm{d}z \\ &= m\left(1-\Phi\left(c-m\right)/\sigma\right) + \sigma\int_{(c-m)/\sigma}^{+\infty} -\phi'(z)\,\mathrm{d}z \\ &= m\Phi\left(\left(m-c\right)/\sigma\right) + \sigma\phi\left(\left(m-c\right)/\sigma\right), \end{aligned}$$

d'autre part :

$$\begin{aligned}\Pr(y > c) &= \Pr\left(\frac{y-m}{\sigma} > \frac{c-m}{\sigma}\right) \\ &= 1 - \Pr\left(\frac{y-m}{\sigma} \leq \frac{c-m}{\sigma}\right) \\ &= 1 - \Phi\left(\frac{c-m}{\sigma}\right) \\ &= \Phi\left(\frac{m-c}{\sigma}\right),\end{aligned}$$

ce qui implique :

$$\mathrm{E}(y|y > m) = \frac{I}{\Pr(y > c)} = m + \sigma\frac{\phi\left(\frac{m-c}{\sigma}\right)}{\Phi\left(\frac{m-c}{\sigma}\right)}, \tag{C.1}$$

la quantité :

$$\lambda\left(\frac{m-c}{\sigma}\right) = \frac{\phi\left(\left(\frac{m-c}{\sigma}\right)\right)}{\Phi\left(\frac{m-c}{\sigma}\right)},$$

est égale à l'inverse du ratio de Mills.

C.2 Loi normale bivariée

Cette annexe présente la loi normale bivariée ainsi que les distributions conditionnelles qui y sont associées. On se limite ici à deux variables, mais l'extension à un nombre quelconque est possible. On considère deux variables (y_1, y_2) d'espérance (m_1, m_2) et de matrice de covariance :

$$\mathrm{V}\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

La loi normale bivariée est définie par la densité :

$$\phi_2(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\left[\frac{y_1-m_1}{\sigma_1}\right]^2 + \left[\frac{y_2-m_2}{\sigma_2}\right]^2 - 2\rho\frac{y_1-m_1}{\sigma_1}\frac{y_2-m_2}{\sigma_2}\right)\right\}$$

C.3 Loi normale conditionnelle

La densité de la loi conditionnelle de y_1 sachant y_2 est définie par :

$$\phi_c(y_1|y_2) = \frac{\phi_2(y_1, y_2)}{f_2(y_2)}, \tag{C.2}$$

où $f_2(y_2)$ est la densité marginale de y_2. On a :

$$f_2(y_2) = \frac{1}{\sigma_2}\phi\left(\frac{y_2 - m_2}{\sigma_2}\right),$$

où $\phi(.)$ est la densité de la loi normale centrée réduite. En prenant le ratio $(C.2)$ on obtient :

$$\begin{aligned}
\phi_c(y_1|y_2) &= \frac{1}{\sigma_1\sqrt{1-\rho^2}}\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{y_1 - m_1}{\sigma_1} - \rho\frac{y_2 - m_2}{\sigma_2}\right)^2\right\} \\
&= \frac{1}{\sigma_1\sqrt{1-\rho^2}}\phi\left(\frac{1}{\sigma_1\sqrt{1-\rho^2}}\left[y_1 - m_1 - \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2)\right]\right) \\
&= \frac{1}{\sigma_1\sqrt{1-\rho^2}}\phi\left(\frac{1}{\sigma_1\sqrt{1-\rho^2}}\left[y_1 - \left(m_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2)\right)\right]\right)
\end{aligned}$$

il s'agit de la densité d'une loi normale d'espérance :

$$\mathrm{E}(y_1|y_2) = m_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2).$$

et de variance :

$$\mathrm{V}(y_1|y_2) = \sigma_1^2\left(1-\rho^2\right).$$

Plus généralement, on peut montrer directement que :

$$y_1 = m_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2) + \varepsilon_1, \tag{C.3}$$

où ε_1 est une perturbation normale indépendante de y_2. Pour voir cela, il suffit de remarquer que ε_1 est une combinaison linéaire de lois normales et est donc une variable normale. Pour l'indépendance, il suffit de calculer la covariance entre ε_1 et y_2, puisque l'indépendance est équivalente à l'absence de corrélation pour cette loi. On a :

$$\begin{aligned}
\mathrm{Cov}(\varepsilon_1, y_2) &= \mathrm{Cov}\left(y_1 - m_1 - \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2), y_2\right) \\
&= \mathrm{Cov}(y_1, y_2) - \rho\frac{\sigma_1}{\sigma_2}\mathrm{V}(y_2) \\
&= \rho\sigma_1\sigma_2 - \rho\frac{\sigma_1}{\sigma_2}\sigma_2^2 \\
&= 0,
\end{aligned}$$

de sorte que ε_1 et y_2 sont indépendantes. L'espérance de ε_1 est nulle :

$$\mathrm{E}(\varepsilon_1) = \mathrm{E}\left(y_1 - m_1 - \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2)\right) = 0,$$

et sa variance est égale à :

$$\begin{aligned}
\mathrm{V}(\varepsilon_1) &= \mathrm{V}\left(y_1 - m_1 - \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2)\right) \\
&= \sigma_1^2\,\mathrm{V}\left(\frac{y_1 - m_1}{\sigma_1} - \rho\frac{y_2 - m_2}{\sigma_2}\right) \\
&= \sigma_1^2\left(1 + \rho^2 - 2\rho^2\right) \\
&= \sigma_1^2\left(1 - \rho^2\right).
\end{aligned}$$

La propriété ($C.3$) est très pratique lorsque l'on étudie la troncature d'une variable normale par une autre variable normale.

C.4 Loi normale bivariée tronquée

Ici, on recherche l'espérance conditionnelle d'une première variable tronquée par la valeur d'une seconde variable avec laquelle elle est corrélée. Ce cas se retrouve lorsque l'on estime une équation de salaire en tenant compte de la participation. On cherche donc la valeur de l'espérance conditionnelle suivante :

$$\begin{aligned}
\mathrm{E}(y_1|y_2 > c) &= \mathrm{E}\left(m_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2) + \varepsilon_1 \middle| y_2 > c\right) \\
&= m_1 + \rho\frac{\sigma_1}{\sigma_2}\left(\mathrm{E}(y_2|\, y_2 > c) - m_2\right) + \mathrm{E}(\varepsilon_1|\, y_2 > c) \\
&= m_1 + \rho\frac{\sigma_1}{\sigma_2}\left(m_2 + \sigma_2\frac{\phi\left(\frac{m_2 - c}{\sigma_2}\right)}{\Phi\left(\frac{m_2 - c}{\sigma_2}\right)} - m_2\right) + \mathrm{E}(\varepsilon_1),
\end{aligned}$$

en utilisant la propriété ($C.1$) et l'indépendance entre y_2 et ε_1. Après simplification, on obtient :

$$\mathrm{E}(y_1|y_2 > c) = m_1 + \rho\sigma_1\lambda\left(\frac{m_2 - c}{\sigma_2}\right). \tag{C.4}$$

Annexe D

Simplification du calcul des dérivées

La plupart des modèles font intervenir dans la log-vraisemblance des termes linéaires :

$$m(X, b) = Xb$$

où X est un vecteur ligne $1 \times p$ et b un vecteur colonne $1 \times p$. Cette propriété vient de la forme latente linéaire de la plupart des modèles à variables qualitatives. Pour estimer le modèle, on a besoin des dérivées de la fonction à maximiser par rapport à un vecteur b. En fait, nous allons voir un certain nombre de simplifications qui permettent de se limiter à des dérivées par rapport à une variable réelle non vectorielle.

Tout d'abord, la fonction à maximiser est la somme de N fonctions qui ne diffèrent que par les valeurs que prennent les variables expliquée et explicatives. La forme fonctionnelle reste la même quelle que soit l'observation. Dans le cas le plus simple :

$$\ell(y, X, b) = \sum_{i=1}^{N} g(y_i, X_i, b),$$

où la fonction g est identique pour tous les individus. C'est la forme que l'on obtient systématiquement sous hypothèse d'indépendance, où la fonction g est le logarithme de la densité de probabilité. Elle se simplifie souvent comme :

$$\ell(y, X, b) = \sum_{i=1}^{N} g(y_i, m_i, \xi) \text{ avec } m_i = X_i b,$$

où ξ est un paramètre, indépendant de b, généralement du second ordre (i.e., de variance ou de corrélation). Les observations des variables explicatives pour le $i-$ème individu sont rangées dans un vecteur ligne $X_i =$

$(X_{1i}, ..., X_{pi})$ et b est le vecteur colonne correspondant $b = (b_1, ..., b_p)'$. On traite les paramètres de ξ séparément comme des paramètres réels car ils sont en petit nombre dans les cas usuels. La dérivation ne pose donc pas de problème particulier par rapport à ξ. Nous sommes donc ramenés au calcul de la dérivée par rapport au *vecteur* b. Il est clair qu'il suffit de dériver :

$$g\left(y_i, m_i, \xi\right),$$

et de faire la somme des dérivées ensuite. Ceci est valable aussi bien pour les dérivées premières que pour les dérivées secondes. En dérivant en chaîne, on a :

$$\frac{\partial g}{\partial b}\left(y_i, m_i, \xi\right) = \frac{\partial g}{\partial m_i}\left(y_i, m_i, \xi\right) \frac{\partial m_i}{\partial b}.$$

La première dérivée est celle d'une fonction réelle et s'effectue comme d'habitude. La deuxième dérivée est obtenue en empilant les dérivées dans le même sens que le vecteur par rapport auquel on dérive (i.e., en ligne ou en colonne). En effet, par convention :

$$\frac{\partial m_i}{\partial b} = \frac{\partial m_i}{\partial \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix}} = \begin{pmatrix} \partial m_i / b_1 \\ \vdots \\ \partial m_i / b_p \end{pmatrix},$$

en remarquant que $m_i = X_i b = X_{1i} b_1 + ... + X_{pi} b_p$, on a $\partial m_i / b_j = X_{ji}$ pour $j = 1, ..., p$. Donc, en empilant ces dérivées :

$$\frac{\partial m_i}{\partial b} = \begin{pmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{pmatrix} = X_i'.$$

En conséquence le gradient, pour une observation i, est égal à :

$$\frac{\partial g}{\partial b}\left(y_i, m_i, \xi\right) = X_i' \underbrace{\frac{\partial g}{\partial m_i}\left(y_i, m_i, \xi\right)}_{\in R}.$$

Le calcul du hessien est simplifié du fait de la linéarité de m_i en b, qui implique que la dérivée seconde de m_i par rapport à b est nulle :

$$\frac{\partial^2 g}{\partial b \partial b'}(y_i, m_i, \xi) = \frac{\partial^2 g}{\partial m_i^2}(y_i, m_i, \xi) \frac{\partial m_i}{\partial b} \frac{\partial m_i}{\partial b'} + \frac{\partial g}{\partial m_i}(y_i, m_i, \xi) \underbrace{\frac{\partial^2 m_i}{\partial b \partial b'}}_{=0}$$

$$= \frac{\partial^2 g}{\partial m_i^2}(y_i, m_i, \xi) \frac{\partial m_i}{\partial b} \frac{\partial m_i}{\partial b'}$$

$$= X_i' X_i \underbrace{\frac{\partial^2 g}{\partial m_i^2}(y_i, m_i, \xi)}_{\in R}.$$

Le calcul se fait donc en trois étapes :

1. Calcul des dérivées première et seconde par rapport à une variable réelle $m = Xb$.
2. Multiplication par X_i' pour le gradient et par $X_i' X_i$ pour le hessien.
3. Addition des dérivées sur l'ensemble des observations.

Exemple D.1 *Nous verrons plus loin que la log-vraisemblance du modèle Logit pour une observation i peut s'écrire sous la forme :*

$$\ell_i = y_i \ln p(m_i) + (1 - y_i) \ln(1 - p(m_i)),$$

où $y_i \in \{0, 1\}$ est la réponse qualitative que l'on étudie (0 pour "non" et 1 pour "oui") et p la fonction de répartition de la loi logistique. Comme précédemment $m_i = X_i b$ résume l'influence des variables explicatives X_i sur le choix y_i qui a été effectué par l'individu i. On remarque dès maintenant que la fonction de répartition de la loi logistique est égale à :

$$p(m) = \frac{1}{1 + \exp(-m)}, \quad X \in R,$$

ce qui entraîne que

$$p'(m) = \frac{\exp(-m)}{1 + \exp(-m)} = p(m)(1 - p(m)).$$

Nous pouvons donc écrire la log-vraisemblance sous la forme :

$$\ell(b) = \sum_{i=1}^{N} g(y_i, m_i) \quad \textit{avec } g(y_i, m) = y_i \ln p(m) + (1 - y_i) \ln(1 - p(m)),$$

$$\forall m \in R.$$

La dérivée première de g par rapport à m est égale à :

$$\frac{\partial g}{\partial m}\left(y_i, m\right) = y_i \frac{p'\left(m\right)}{p\left(m\right)} - \left(1 - y_i\right) \frac{p'\left(m\right)}{1 - p\left(m\right)} = y_i - p\left(m\right),$$

après simplification. La dérivée seconde est égale à :

$$\frac{\partial^2 g}{\partial m^2}\left(y_i, X\right) = -p'\left(m\right) = -p\left(m\right)\left(1 - p\left(m\right)\right).$$

Le score s'écrit donc :

$$s\left(b\right) = \sum_{i=1}^{N} X_i' \frac{\partial g}{\partial m}\left(y_i, m_i\right) = \sum_{i=1}^{N} X_i' \left(y_i - p\left(m_i\right)\right),$$

et le hessien est égal à :

$$H\left(b\right) = \sum_{i=1}^{N} X_i' X_i \frac{\partial^2 g}{\partial m^2}\left(y_i, m_i\right) = -\sum_{i=1}^{N} X_i' X_i p\left(m_i\right)\left(1 - p\left(m_i\right)\right).$$