ÉCONOMÉTRIE DES VARIABLES QUALITATIVES

UNIVERSITÉ DU MANS (EXAMEN, L3)

EXERCICE I

(1) Soit un échantillon de variables aléatoires $(y_1, \ldots, y_N)$ identiquement et indépendamment distribuées selon une loi de Poisson de paramètre $\lambda > 0$, c'est-à-dire telles que :

$$\mathbb{P}(y_i = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \mathbb{N}$$

Calculer l'espérance et la variance de y_i. Que remarque-t-on ?

(2) Écrire la fonction de log-vraisemblance de l'échantillon.

(3) Calculer l'estimateur du maximum de vraisemblance $\hat{\lambda}$ de λ.

(4) Calculer la variance de $\hat{\lambda}$ à partir de son expression. En déduire le comportement asymptotique de $\hat{\lambda}$. Déterminer une fonction $f(N)$ telle que $f(N)(\hat{\lambda} - \lambda)$ converge vers une loi normale dont on précisera les paramètres.

(5) On considère maintenant que le paramètre varie d'un individu à l'autre, en fonction d'un vecteur de variables exogènes X_i (de dimension $1 \times K$) et d'un vecteur de paramètres β. On postule la spécification :

$$\lambda_i = \exp(X_i \beta)$$

Écrire la log-vraisemblance, puis le score (gradient de la log-vraisemblance) du modèle. Justifier pourquoi le choix d'une forme exponentielle pour λ_i est préférable à une forme linéaire $\lambda_i = X_i \beta$.

EXERCICE II

On dispose d'un échantillon de N individus, indexés par $i = 1, \ldots, N$, et l'on observe pour chacun une variable quantitative y_i (le salaire mensuel, en euros), une variable indicatrice $D_i \in \{0, 1\}$ identifiant le sexe ($D_i = 1$ si l'individu est une femme, $D_i = 0$ sinon) et une variable quantitative X_i (l'expérience professionnelle, en années).

(1) On estime par moindres carrés ordinaires le modèle :

$$y_i = \alpha + \delta D_i + u_i$$

Montrer que $\hat{\alpha} = \bar{y}_{D=0}$ et que $\hat{\alpha} + \hat{\delta} = \bar{y}_{D=1}$, où $\bar{y}_{D=d}$ désigne la moyenne empirique des salaires dans le sous-échantillon $\{i : D_i = d\}$. Interpréter $\hat{\delta}$.

(2) On souhaite désormais distinguer trois catégories d'éducation, repérées par les indicatrices E_1, E_2 et E_3 ($E_{j,i} = 1$ si l'individu i appartient à la catégorie j, 0 sinon ; un seul $E_{j,i}$ vaut 1 pour chaque i). Pourquoi ne peut-on pas estimer le modèle

$$y_i = \alpha + \gamma_1 E_{1,i} + \gamma_2 E_{2,i} + \gamma_3 E_{3,i} + u_i$$

par les MCO ? Indiquer une solution et donner l'interprétation des coefficients dans la spécification retenue.

(3) On revient au modèle à un seul groupe D, en y introduisant l'expérience X_i et son interaction avec D_i :

$$y_i = \alpha + \delta D_i + \gamma X_i + \theta D_i X_i + u_i$$

Calculer l'effet marginal de X sur y pour un homme, puis pour une femme. À quelle condition les effets marginaux sont-ils identiques dans les deux groupes ?

(4) Que représente exactement $\hat{\delta}$ dans cette spécification (autrement dit, à quelle sous-population le coefficient correspond-il) ? On centre désormais l'expérience : $\tilde{X}_i = X_i - \bar{X}$. Réécrire le modèle

avec $\tilde{X}$ à la place de X et expliquer en quoi l'interprétation de δ change.

(5) On suppose que les hommes et les femmes ont des variances de salaires différentes (à expérience donnée), de sorte que $\mathbb{V}(u_i \mid D_i = 0) = \sigma_0^2$ et $\mathbb{V}(u_i \mid D_i = 1) = \sigma_1^2$ avec $\sigma_0^2 \neq \sigma_1^2$. Quel problème cela pose-t-il pour l'inférence sur $\hat{\delta}$ et $\hat{\theta}$? Comment le résout-on en pratique ?

EXERCICE III

On reprend le cadre du modèle dichotomique à variable latente. Pour chaque individu $i = 1, \ldots, N$, on définit :

$$z_i = X_i\beta + u_i, \qquad y_i = \begin{cases} 1 & \text{si } z_i > 0, \\ 0 & \text{sinon,} \end{cases}$$

où X_i est un vecteur $1 \times K$ de variables exogènes, β un vecteur $K \times 1$ de paramètres et u_i une variable aléatoire identiquement et indépendamment distribuée, d'espérance nulle. On suppose dans tout l'exercice que u_i suit une loi normale $\mathcal{N}(0, \sigma^2)$. On note Φ la fonction de répartition de la loi normale centrée réduite et ϕ sa densité.

(1) Montrer que

$$\mathbb{P}(y_i = 1 \mid X_i) = \Phi(X_i\beta/\sigma)$$

(2) En déduire que β et σ ne sont pas identifiables séparément. Quelle normalisation impose-t-on traditionnellement pour résoudre ce problème ? Justifier.

(3) Sous la normalisation retenue à la question précédente, écrire la log-vraisemblance du modèle Probit. Calculer le score $\partial \log L / \partial \beta$. On pourra introduire les notations $\Phi_i = \Phi(X_i\beta)$ et $\phi_i = \phi(X_i\beta)$.

(4) La matrice d'information de Fisher du modèle Probit s'écrit :

$$I(\beta) = \sum_{i=1}^{N} \frac{\phi_i^2}{\Phi_i(1-\Phi_i)} X_i' X_i$$

Justifier qu'elle est définie positive (en supposant la matrice des régresseurs de rang plein). Quelle conséquence cela a-t-il sur l'estimation ?

(5) Si l'on suppose à la place que u_i suit une loi logistique standard (modèle Logit), de variance $\pi^2/3$, justifier la règle empirique selon laquelle

$$\hat{\beta}_{\text{Logit}} \approx \frac{\pi}{\sqrt{3}}\,\hat{\beta}_{\text{Probit}}$$

Pourquoi cette correspondance n'est-elle qu'approximative ? Que peut-on en revanche attendre des probabilités prédites par les deux modèles ?