

ÉCONOMÉTRIE DES VARIABLES QUALITATIVES

UNIVERSITÉ DU MANS (EXAMEN 2ND SESSION, L3)

Mardi 16 juin 2026

EXERCICE I

(1) Soit un échantillon de variables aléatoires $(y_1, \dots, y_N)$ identiquement et indépendamment distribuées selon une loi géométrique de paramètre $p \in]0, 1[$, c'est-à-dire telles que :

$$\mathbb{P}(y_i = k) = (1 - p)^{k-1} p, \quad k \in \mathbb{N}^*$$

(y_i s'interprète comme le nombre d'épreuves nécessaires pour obtenir le premier succès dans une suite d'épreuves de Bernoulli indépendantes de probabilité de succès p .) Calculer l'espérance et la variance de y_i . Comparer les deux quantités.

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

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

(4) L'estimateur $\hat{p}$ étant une fonction non linéaire de la moyenne empirique $\bar{y}$, on utilisera la méthode delta. Calculer la variance asymptotique de $\hat{p}$. En déduire le comportement asymptotique de $\hat{p}$. Déterminer une fonction $f(N)$ telle que $f(N)(\hat{p} - p)$ 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 note $\mu_i = \mathbb{E}[y_i \mid X_i]$ l'espérance conditionnelle et l'on postule la spécification :

$$\mu_i = \exp(X_i \beta).$$

Exprimer p_i en fonction de μ_i , puis écrire la log-vraisemblance et le score (gradient de la log-vraisemblance) du modèle. Justifier pourquoi

le choix d'une forme exponentielle pour μ_i est préférable à une forme linéaire $\mu_i = X_i \beta$.

EXERCICE II

On dispose d'un échantillon de N individus, indexés par $i = 1, \dots, N$, et l'on observe pour chacun une variable quantitative y_i (le salaire horaire, en euros), une variable indicatrice $D_i \in \{0, 1\}$ identifiant le statut syndical ($D_i = 1$ si l'individu est syndiqué, $D_i = 0$ sinon) et une variable quantitative X_i (l'ancienneté dans l'entreprise, 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 de contrat, repérées par les indicatrices C_1, C_2 et C_3 ($C_{j,i} = 1$ si l'individu i appartient à la catégorie j , 0 sinon ; un seul $C_{j,i}$ vaut 1 pour chaque i). Pourquoi ne peut-on pas estimer le modèle

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

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'ancienneté 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 non-syndiqué, puis pour un syndiqué. 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'ancienneté : $\tilde{X}_i = X_i - \bar{X}$. 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 syndiqués et les non-syndiqués ont des variances de salaires différentes (à ancienneté 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 $i = 1, \dots, N$, on définit :

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

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 logistique. On note

$$\Lambda(t) = \frac{1}{1 + e^{-t}}$$

la fonction de répartition de la loi logistique standard (d'espérance nulle et de variance $\pi^2/3$).

(1) On suppose que u_i admet la fonction de répartition $F(t) = \Lambda(t/s)$, où $s > 0$ est un paramètre d'échelle (la variance de u_i vaut alors $s^2 \pi^2/3$). Montrer que

$$\mathbb{P}(y_i = 1 \mid X_i) = \Lambda(X_i \beta / s).$$

(2) En déduire que β et s 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, montrer d'abord que la densité de la loi logistique standard vérifie $\Lambda'(t) = \Lambda(t)(1 - \Lambda(t))$. Écrire ensuite la log-vraisemblance du modèle Logit, puis calculer le score $\partial \log L / \partial \beta$. On pourra introduire la notation $\Lambda_i = \Lambda(X_i \beta)$.

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

$$I(\beta) = \sum_{i=1}^N \Lambda_i (1 - \Lambda_i) X_i' X_i.$$

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) Montrer que, dans le modèle Logit, le logarithme de la cote (*odds*) est linéaire :

$$\log \frac{\mathbb{P}(y_i = 1 \mid X_i)}{\mathbb{P}(y_i = 0 \mid X_i)} = X_i \beta,$$

et en déduire l'interprétation de e^{β_k} comme rapport de cotes (*odds ratio*) associé à la variable x_k . Enfin, si l'on avait estimé à la place un modèle Probit (u_i normal), quelle relation approximative lierait $\hat{\beta}_{\text{Logit}}$ et $\hat{\beta}_{\text{Probit}}$? Que peut-on attendre des probabilités prédites par les deux modèles ?