ÉCONOMÉTRIE DES VARIABLES QUALITATIVES

Examen L3 — Corrigé détaillé

UNIVERSITÉ DU MANS

EXERCICE I — Estimation par maximum de vraisemblance d'une loi de Poisson

(1) Calcul de l'espérance et de la variance.

L'espérance vaut

$$\mathbb{E}[y] = \sum_{k=0}^{\infty} k \, \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} = \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda.$$

Pour la variance, on calcule d'abord

$$\mathbb{E}[y(y-1)] = \sum_{k=0}^{\infty} k(k-1) \, \frac{\lambda^k e^{-\lambda}}{k!} = \lambda^2 e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} = \lambda^2,$$

d'où $\mathbb{E}[y^2] = \lambda^2 + \lambda$ et donc

$$\mathbb{V}[y] = \mathbb{E}[y^2] - \mathbb{E}[y]^2 = \lambda.$$

Remarque : l'espérance et la variance coïncident ($\mathbb{E}[y] = \mathbb{V}[y] = \lambda$), c'est la propriété d'*équidispersion* caractéristique de la loi de Poisson. Il s'agit d'une contrainte forte : en pratique on observe souvent une sur-dispersion ($\mathbb{V}[y] > \mathbb{E}[y]$) qui motive l'usage de modèles plus généraux (binomiale négative).

(2) Log-vraisemblance.

Les observations étant i.i.d.,

$$L(\lambda; y) = \prod_{i=1}^{N} \frac{\lambda^{y_i} e^{-\lambda}}{y_i!},$$

et donc

$$\log L(\lambda; y) = \log \lambda \sum_{i=1}^{N} y_i - N\lambda - \sum_{i=1}^{N} \log(y_i!).$$

(3) Estimateur du maximum de vraisemblance.

La condition du premier ordre s'écrit

$$\frac{\partial \log L}{\partial \lambda} = \frac{1}{\lambda} \sum_{i=1}^{N} y_i - N = 0 \iff \hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} y_i = \bar{y}.$$

La condition du second ordre, $\partial^2 \log L / \partial \lambda^2 = -\lambda^{-2} \sum_i y_i < 0$, confirme qu'il s'agit bien d'un maximum.

(4) Variance et comportement asymptotique.

Comme $\hat{\lambda} = \bar{y}$ est une moyenne empirique de variables i.i.d. de variance λ,

$$\mathbb{V}[\hat{\lambda}] = \frac{\mathbb{V}[y_i]}{N} = \frac{\lambda}{N}.$$

On a $\mathbb{E}[\hat{\lambda}] = \lambda$ (sans biais) et $\mathbb{V}[\hat{\lambda}] \to 0$ quand $N \to \infty$: $\hat{\lambda}$ converge en probabilité vers λ (loi des grands nombres).

Le théorème central limite appliqué à $\bar{y}$ donne

$$\sqrt{N}\big(\hat{\lambda} - \lambda\big) \;\xrightarrow{\mathcal{L}}\; \mathcal{N}(0, \lambda),$$

soit $f(N) = \sqrt{N}$; la loi limite est centrée de variance λ.

(5) Régression de Poisson.

Avec $\lambda_i = \exp(X_i\beta)$, la log-vraisemblance devient

$$\log L(\beta) = \sum_{i=1}^{N} \Big[\, y_i\, X_i\beta - \exp(X_i\beta) - \log(y_i!) \,\Big].$$

Le score s'obtient en dérivant terme à terme :

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{N} \big(y_i - \exp(X_i\beta)\big) X_i' = \sum_{i=1}^{N} (y_i - \lambda_i)\, X_i'.$$

Pourquoi la forme exponentielle ? Le paramètre d'une loi de Poisson est strictement positif. La forme $\lambda_i = \exp(X_i\beta)$ garantit cette positivité quel que soit $\beta \in \mathbb{R}^K$. Une forme linéaire $\lambda_i = X_i\beta$ pourrait produire des valeurs négatives, ce qui n'a pas de sens pour une loi de Poisson et rendrait la log-vraisemblance non définie sur une partie de l'espace des paramètres.

EXERCICE II — Variables explicatives qualitatives et interactions

(1) On note N_0 et N_1 le nombre d'individus tels que $D_i = 0$ et $D_i = 1$, avec $N_0 + N_1 = N$. Les conditions du premier ordre des MCO sont :

$$\sum_{i=1}^{N} \Big(y_i - \hat{\alpha} - \hat{\delta} D_i\Big) = 0, \qquad \sum_{i=1}^{N} D_i \Big(y_i - \hat{\alpha} - \hat{\delta} D_i\Big) = 0.$$

La seconde équation, en utilisant $D_i^2 = D_i$, donne

$$\sum_{i\,:\,D_i=1} y_i = (\hat{\alpha} + \hat{\delta})\, N_1 \iff \hat{\alpha} + \hat{\delta} = \bar{y}_{D=1}.$$

La première équation se réécrit $N\bar{y} = N\hat{\alpha} + \hat{\delta} N_1$, soit $\bar{y} = \hat{\alpha} + \hat{\delta}\,(N_1/N)$. En décomposant $\bar{y} = (N_0\bar{y}_{D=0} + N_1\bar{y}_{D=1})/N$ et en utilisant $\hat{\alpha} + \hat{\delta} = \bar{y}_{D=1}$, on obtient après simplification

$$\hat{\alpha} = \bar{y}_{D=0}, \qquad \hat{\delta} = \bar{y}_{D=1} - \bar{y}_{D=0}.$$

Interprétation : $\hat{\delta}$ est l'écart de salaire moyen entre femmes et hommes dans l'échantillon. Sous l'hypothèse d'exogénéité $\mathbb{E}[u_i \mid D_i] = 0$, c'est aussi l'écart de salaire *moyen* dans la population ; mais sans contrôle d'autres caractéristiques (éducation, expérience, secteur), il ne peut être interprété comme un effet « causal » du sexe.

(2) La somme $E_{1,i} + E_{2,i} + E_{3,i} = 1$ pour tout i est exactement la colonne de constante. Les quatre régresseurs constante, E_1, E_2, E_3 sont parfaitement colinéaires : la matrice $X'X$ n'est pas inversible et l'estimateur MCO n'est pas défini (*trappe à variables muettes*).

Deux solutions équivalentes :

— **Catégorie de référence :** on conserve la constante et on retire une indicatrice (par exemple E_3), et l'on estime $y_i = \alpha + \gamma_1 E_{1,i} + \gamma_2 E_{2,i} + u_i$. Alors $\hat{\alpha} = \bar{y}_{E_3=1}$ est le salaire moyen de la catégorie de référence et $\hat{\gamma}_j$ ($j = 1, 2$) est l'écart de salaire moyen entre la catégorie j et la catégorie 3.
— **Sans constante :** on garde les trois indicatrices mais on supprime la constante ; les coefficients sont alors directement les salaires moyens de chaque catégorie.

(3) Effets marginaux dans le modèle avec interaction.

L'effet marginal de X sur y vaut

$$\frac{\partial \mathbb{E}[y \mid D, X]}{\partial X} = \gamma + \theta D.$$

Pour un homme ($D = 0$), il vaut γ ; pour une femme ($D = 1$), il vaut $\gamma + \theta$. Les deux effets coïncident si et seulement si $\theta = 0$ (on retrouve alors un modèle « à pentes communes »). La significativité de $\hat{\theta}$ teste l'égalité du rendement de l'expérience entre les deux groupes.

(4) Interprétation de $\hat{\delta}$ et centrage.

Pour $X = 0$, le modèle donne $\mathbb{E}[y \mid D = 0, X = 0] = \alpha$ et $\mathbb{E}[y \mid D = 1, X = 0] = \alpha + \delta$. Donc $\hat{\delta}$ représente l'écart de salaire entre femmes et hommes *pour un individu d'expérience nulle*. Si $X = 0$ est hors du support empirique (par exemple si tous les individus ont au moins quelques années d'expérience), ce coefficient est extrapolé en dehors des données et n'est pas directement interprétable.

En remplaçant X_i par $\tilde{X}_i = X_i - \bar{X}$, le modèle devient

$$\begin{aligned} y_i &= \alpha + \delta D_i + \gamma(\tilde{X}_i + \bar{X}) + \theta D_i(\tilde{X}_i + \bar{X}) + u_i \\ &= \underbrace{(\alpha + \gamma\bar{X})}_{\alpha^*} + \underbrace{(\delta + \theta\bar{X})}_{\delta^*} D_i + \gamma\tilde{X}_i + \theta D_i\tilde{X}_i + u_i. \end{aligned}$$

Le nouveau coefficient $\delta^* = \delta + \theta\bar{X}$ s'interprète comme l'écart de salaire entre femmes et hommes *pour un individu d'expérience moyenne* $\bar{X}$. C'est en général un point du support empirique, donc une quantité économiquement pertinente. Les coefficients γ et θ sont, eux, inchangés : le centrage ne déplace que la lecture de la constante et de δ.

(5) Hétéroscédasticité par blocs.

L'estimateur MCO $\hat{\beta}_{\text{MCO}}$ reste sans biais et convergent (sa cohérence ne dépend pas de l'hypothèse d'homoscédasticité, mais seulement de $\mathbb{E}[u_i \mid D_i, X_i] = 0$). En revanche, la matrice de variance asymptotique usuelle $\sigma^2(X'X)^{-1}$ *n'est plus correcte* : les variances de $\hat{\delta}$ et $\hat{\theta}$ estimées sous homoscédasticité sont biaisées (généralement vers le bas), ce qui invalide les tests de Student et de Fisher classiques.

Deux solutions usuelles :
— **Variances robustes (White) :** on remplace l'estimateur naïf par $\widehat{\mathbb{V}}(\hat{\beta}) = (X'X)^{-1}\big(\sum_i \hat{u}_i^2 X_i' X_i\big)(X'X)^{-1}$, qui est convergent sous hétéroscédasticité de forme arbitraire.
— **Moindres carrés généralisés (Aitken) :** on estime σ_0^2 et σ_1^2 par sous-échantillon, puis on ré-estime par MCO pondérés en pondérant chaque observation par $1/\hat{\sigma}_{D_i}$. L'estimateur obtenu est plus efficace que les MCO sous hétéroscédasticité.

EXERCICE III — Modèle Probit et lien avec le Logit

(1) On a :

$$\begin{aligned}\mathbb{P}(y_i = 1 \mid X_i) &= \mathbb{P}(z_i > 0 \mid X_i) = \mathbb{P}(u_i > -X_i\beta \mid X_i) \\ &= \mathbb{P}\left(\frac{u_i}{\sigma} > -\frac{X_i\beta}{\sigma}\right) = 1 - \Phi\left(-\frac{X_i\beta}{\sigma}\right) \\ &= \Phi\left(\frac{X_i\beta}{\sigma}\right),\end{aligned}$$

en utilisant la symétrie de la loi normale ($1 - \Phi(-t) = \Phi(t)$) et le fait que $u_i/\sigma \sim \mathcal{N}(0,1)$.

(2) Non-identification.

Seul le rapport β/σ apparaît dans la probabilité : pour toute constante $c > 0$, les couples (β, σ) et $(c\beta, c\sigma)$ produisent la même probabilité conditionnelle. Le modèle n'identifie donc qu'un *rapport*, pas β et σ séparément.

Normalisation conventionnelle : on impose $\sigma = 1$. Ce choix est sans perte de généralité puisque seul β/σ est identifié ; les coefficients estimés doivent alors s'interpréter comme des estimations de β/σ et non de β en niveau. Cette normalisation est faite implicitement par tous les logiciels qui estiment un Probit.

(3) Log-vraisemblance et score.

Avec $\sigma = 1$, $\Phi_i = \Phi(X_i\beta)$ et la vraisemblance d'une Bernoulli i.id. :

$$L(\beta) = \prod_{i=1}^{N} \Phi_i^{y_i}(1 - \Phi_i)^{1-y_i},$$

soit

$$\log L(\beta) = \sum_{i=1}^{N} \big[y_i \log \Phi_i + (1 - y_i)\log(1 - \Phi_i)\big].$$

On a $\partial\Phi_i/\partial\beta = \phi_i\, X_i'$, d'où le score :

$$\begin{aligned}\frac{\partial \log L}{\partial \beta} &= \sum_{i=1}^{N} \left[\frac{y_i\,\phi_i}{\Phi_i} - \frac{(1 - y_i)\,\phi_i}{1 - \Phi_i}\right] X_i' \\ &= \sum_{i=1}^{N} \frac{y_i - \Phi_i}{\Phi_i(1 - \Phi_i)}\,\phi_i\, X_i'.\end{aligned}$$

Le score est de la forme « résidu généralisé × régresseur », analogue au score MCO mais avec une pondération $\phi_i/[\Phi_i(1-\Phi_i)]$ qui dépend de β : contrairement au modèle linéaire ou au Logit, il n'existe pas d'expression analytique de $\hat{\beta}$, et l'estimation passe nécessairement par un algorithme numérique (Newton-Raphson, BHHH, etc.).

(4) Information de Fisher.

Soit $a \in \mathbb{R}^K$ un vecteur non nul. Alors

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

Chaque terme est positif puisque
- $\phi_i^2 > 0$ (la densité normale est strictement positive sur $\mathbb{R}$) ;
- $\Phi_i \in (0,1)$ donc $\Phi_i(1 - \Phi_i) > 0$;

— $(X_i a)^2 \geq 0$.

La somme est donc positive ou nulle. Si la matrice $X = (X_1'; \ldots; X_N')$ est de rang plein (rang K), il existe au moins une observation i telle que $X_i a \neq 0$ (sinon a appartiendrait à $\ker X$ et serait nul, ce qui contredit $a \neq 0$). On a alors $a' I(\beta) a > 0$: $I(\beta)$ est **définie positive**.

Conséquences : la matrice hessienne moyenne $\mathbb{E}[\partial^2 \log L / \partial\beta \partial\beta'] = -I(\beta)$ est définie négative, donc la log-vraisemblance est strictement concave (au sens de l'espérance). L'éventuel maximum est unique et l'estimateur du maximum de vraisemblance $\hat{\beta}$ existe asymptotiquement, est convergent et asymptotiquement normal :

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0,\ \big[\tfrac{1}{N} I(\beta)\big]^{-1}\Big).$$

(5) Lien Logit/Probit.

Dans le modèle dichotomique à variable latente, on identifie β/σ_u, où σ_u^2 est la variance des chocs latents. Les deux modèles fixent cette variance par convention :

— Probit : $u_i \sim \mathcal{N}(0,1)$, donc $\sigma_u^{\text{Probit}} = 1$.

— Logit : u_i logistique standard, donc $\sigma_u^{\text{Logit}} = \pi/\sqrt{3}$.

Considérons un modèle latent générateur unique $z_i = X_i b + \varepsilon_i$, où b est le « vrai » paramètre et ε_i une variable centrée de variance σ_ε^2 (inconnue). Selon la question (1), la probabilité conditionnelle s'écrit $\mathbb{P}(y_i = 1 \mid X_i) = G(X_i b/\sigma_\varepsilon)$ pour une certaine fonction de répartition G. Quelle que soit la spécification postulée, on identifie le rapport b/σ_ε, jamais b isolément. Comme la spécification Probit fixe $\sigma = 1$ et la spécification Logit fixe $\sigma = \pi/\sqrt{3}$, on a :

$$\frac{\hat{\beta}_{\text{Probit}}}{1} \approx \frac{b}{\sigma_\varepsilon} \approx \frac{\hat{\beta}_{\text{Logit}}}{\pi/\sqrt{3}},$$

d'où

$$\hat{\beta}_{\text{Logit}} \approx \frac{\pi}{\sqrt{3}}\,\hat{\beta}_{\text{Probit}} \approx 1{,}81\,\hat{\beta}_{\text{Probit}}.$$

Pourquoi seulement approximative ? Les deux distributions n'ont pas la même *forme*. La loi logistique a des queues plus épaisses que la loi normale : pour des observations dans les queues ($|X_i\beta|$ grand), les probabilités prédites par les deux modèles diffèrent légèrement. La règle $\pi/\sqrt{3}$ équilibre les deux modèles « au centre » de la distribution, mais pas dans les queues.

Probabilités prédites : malgré la différence d'échelle entre $\hat{\beta}_{\text{Logit}}$ et $\hat{\beta}_{\text{Probit}}$, les *probabilités* prédites $\hat{F}(X_i\hat{\beta})$ sont presque identiques d'un modèle à l'autre (les divergences sont marginales sauf pour les valeurs extrêmes). C'est pour cela que le choix entre Logit et Probit est, en pratique, principalement une affaire de convention et de commodité (le Logit donne des odds-ratios faciles à interpréter via $\exp(\beta)$).