ÉCONOMÉTRIE DES VARIABLES QUALITATIVES

Examen 2nd session L3 — Corrigé détaillé

UNIVERSITÉ DU MANS

Mardi 16 juin 2026

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

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

On rappelle que pour $|x| < 1$, $\sum_{k=0}^{\infty} x^k = (1-x)^{-1}$, d'où, en dérivant, $\sum_{k=1}^{\infty} k\, x^{k-1} = (1-x)^{-2}$ et $\sum_{k=2}^{\infty} k(k-1)\, x^{k-2} = 2(1-x)^{-3}$. En posant $q = 1-p$, l'espérance vaut

$$\mathbb{E}[y] = \sum_{k=1}^{\infty} k\,(1-p)^{k-1} p = p \sum_{k=1}^{\infty} k\, q^{k-1} = p\,\frac{1}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}.$$

Pour la variance, on calcule d'abord (le terme $k=1$ étant nul)

$$\mathbb{E}[y(y-1)] = \sum_{k=1}^{\infty} k(k-1)\, q^{k-1} p = p\,q \sum_{k=2}^{\infty} k(k-1)\, q^{k-2} = p\,q\,\frac{2}{(1-q)^3} = \frac{2q}{p^2},$$

d'où $\mathbb{E}[y^2] = \dfrac{2q}{p^2} + \dfrac{1}{p}$ et donc

$$\mathbb{V}[y] = \mathbb{E}[y^2] - \mathbb{E}[y]^2 = \frac{2q}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{2q - 1 + p}{p^2} = \frac{q}{p^2} = \frac{1-p}{p^2}.$$

Comparaison : on a $\mathbb{E}[y] = 1/p$ et $\mathbb{V}[y] = (1-p)/p^2$, soit $\mathbb{V}[y] - \mathbb{E}[y] = (1-2p)/p^2$. La loi géométrique est donc *sur-dispersée* ($\mathbb{V}[y] > \mathbb{E}[y]$) lorsque $p < \frac{1}{2}$, *équi-dispersée* en $p = \frac{1}{2}$ et *sous-dispersée* lorsque $p > \frac{1}{2}$. C'est notamment lorsque p est faible (espérance élevée, variance encore plus élevée) qu'elle constitue une alternative à la loi de Poisson (équidispersée) pour des données de comptage présentant une variance supérieure à la moyenne.

(2) Log-vraisemblance.

Les observations étant i.i.d.,

$$L(p; y) = \prod_{i=1}^{N} (1-p)^{y_i - 1}\, p = p^N\, (1-p)^{\sum_i y_i - N},$$

et donc

$$\log L(p; y) = \Big(\sum_{i=1}^{N} y_i - N\Big) \log(1-p) + N \log p.$$

(3) Estimateur du maximum de vraisemblance.

La condition du premier ordre s'écrit

$$\frac{\partial \log L}{\partial p} = -\frac{1}{1-p}\Big(\sum_{i=1}^{N} y_i - N\Big) + \frac{N}{p} = 0 \iff N(1-p) = p\Big(\sum_{i=1}^{N} y_i - N\Big),$$

soit $N = p\sum_i y_i$ et donc

$$\hat{p} = \frac{N}{\sum_{i=1}^{N} y_i} = \frac{1}{\bar{y}}.$$

La condition du second ordre, $\partial^2 \log L/\partial p^2 = -\left(\sum_i y_i - N\right)(1-p)^{-2} - Np^{-2} < 0$ (puisque $\sum_i y_i \geq N$), confirme qu'il s'agit bien d'un maximum.

(4) Variance asymptotique par la méthode delta.

L'estimateur $\hat{p} = g(\bar{y})$ avec $g(m) = 1/m$ est une fonction lisse de la moyenne empirique. Le théorème central limite appliqué à $\bar{y}$ (moyenne de variables i.i.d. d'espérance $\mu = \mathbb{E}[y] = 1/p$ et de variance $\mathbb{V}[y] = (1-p)/p^2$) donne

$$\sqrt{N}\,(\bar{y} - \mu) \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0,\ \tfrac{1-p}{p^2}\Big).$$

La méthode delta fournit alors

$$\sqrt{N}\,(\hat{p} - p) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0,\ g'(\mu)^2\,\mathbb{V}[y]\big).$$

On a $g'(m) = -1/m^2$. Or $\mu = 1/p$, donc $g'(\mu) = -p^2$ (on retrouve au passage $g(\mu) = p$). D'où

$$g'(\mu)^2\,\mathbb{V}[y] = p^4 \cdot \frac{1-p}{p^2} = p^2(1-p),$$

et finalement

$$\sqrt{N}\,(\hat{p} - p) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0,\ p^2(1-p)\big), \qquad \mathbb{V}[\hat{p}] \approx \frac{p^2(1-p)}{N}.$$

On a donc $f(N) = \sqrt{N}$, et la loi limite est centrée de variance $p^2(1-p)$. Comme $\mathbb{V}[\hat{p}] \to 0$, $\hat{p}$ est convergent.

Vérification par l'information de Fisher. La log-vraisemblance d'une observation est $\ell(p) = (y-1)\log(1-p) + \log p$, de dérivée seconde $\ell''(p) = -(y-1)(1-p)^{-2} - p^{-2}$. En prenant l'espérance et en utilisant $\mathbb{E}[y] = 1/p$ (donc $\mathbb{E}[y] - 1 = (1-p)/p$),

$$\mathcal{I}_1(p) = -\mathbb{E}[\ell''(p)] = \frac{\mathbb{E}[y]-1}{(1-p)^2} + \frac{1}{p^2} = \frac{1}{p(1-p)} + \frac{1}{p^2} = \frac{1}{p^2(1-p)}.$$

La variance asymptotique de l'EMV est $\mathcal{I}_1(p)^{-1}/N = p^2(1-p)/N$, ce qui confirme le résultat obtenu par la méthode delta (l'EMV est asymptotiquement efficace).

(5) Régression géométrique.

Le paramètre et l'espérance sont liés par $\mu_i = 1/p_i$, soit, en inversant,

$$p_i = \frac{1}{\mu_i} = \exp(-X_i\beta), \qquad 1 - p_i = \frac{\mu_i - 1}{\mu_i} = 1 - \exp(-X_i\beta).$$

La log-vraisemblance d'une observation devient

$$\ell_i(\beta) = (y_i - 1)\log(1-p_i) + \log p_i = (y_i - 1)\log\big(e^{X_i\beta} - 1\big) - y_i\,X_i\beta,$$

d'où

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

Le score s'obtient en dérivant. Comme $\partial_\beta \log\big(e^{X_i\beta}-1\big) = \frac{e^{X_i\beta}}{e^{X_i\beta}-1}\,X_i' = \frac{\mu_i}{\mu_i-1}\,X_i'$,

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{N}\Big[\,(y_i-1)\frac{\mu_i}{\mu_i-1} - y_i\,\Big]X_i'.$$

Le crochet se simplifie :

$$(y_i-1)\frac{\mu_i}{\mu_i-1} - y_i = \frac{(y_i-1)\mu_i - y_i(\mu_i-1)}{\mu_i-1} = \frac{y_i-\mu_i}{\mu_i-1},$$

de sorte que

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

On retrouve la structure « résidu × régresseur » du modèle de Poisson, ici pondérée par $1/(\mu_i-1)$; à l'optimum, l'erreur de prévision $y_i-\mu_i$ est orthogonale aux régresseurs (au poids près).

Pourquoi la forme exponentielle ? L'espérance d'une loi géométrique sur $\mathbb{N}^*$ vaut $\mu_i = 1/p_i > 1$, donc en particulier strictement positive. La forme $\mu_i = \exp(X_i\beta)$ garantit $\mu_i > 0$ quel que soit $\beta \in \mathbb{R}^K$, donc $p_i = 1/\mu_i > 0$. Une forme linéaire $\mu_i = X_i\beta$ pourrait au contraire produire une espérance négative (donc un p_i négatif), ce qui n'a pas de sens et rendrait la log-vraisemblance non définie sur une partie de l'espace des paramètres. (On notera que le lien exponentiel n'impose pas mécaniquement la borne supérieure $p_i < 1$, c'est-à-dire $\mu_i > 1$; il reste néanmoins le choix canonique — lien log — pour modéliser une espérance positive.)

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 syndiqués et non-syndiqués 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 (type de contrat, ancienneté, secteur), il ne peut être interprété comme un effet « causal » de l'appartenance syndicale.

(2) La somme $C_{1,i} + C_{2,i} + C_{3,i} = 1$ pour tout i est exactement la colonne de constante. Les quatre régresseurs constante, C_1, C_2, C_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 C_3), et l'on estime $y_i = \alpha + \gamma_1 C_{1,i} + \gamma_2 C_{2,i} + u_i$. Alors $\hat{\alpha} = \bar{y}_{C_3=1}$ 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 non-syndiqué ($D = 0$), il vaut γ ; pour un syndiqué ($D = 1$), 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'ancienneté 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 syndiqués et non-syndiqués *pour un individu d'ancienneté nulle*. Si $X = 0$ est hors du support empirique (par exemple si tous les individus ont au moins quelques années d'ancienneté), 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 syndiqués et non-syndiqués *pour un individu d'ancienneté 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 Logit et interprétation

(1) On a :

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

en utilisant le fait que u_i/s suit une loi logistique standard (de fonction de répartition Λ) et la symétrie de cette loi ($1 - \Lambda(-t) = \Lambda(t)$, qui découle de $\Lambda(-t) = \frac{1}{1+e^t} = 1 - \Lambda(t)$).

(2) Non-identification.

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

Normalisation conventionnelle : on impose $s = 1$, c'est-à-dire que u_i suit la loi logistique standard, de variance $\pi^2/3$. Ce choix est sans perte de généralité puisque seul β/s est identifié ; les coefficients estimés doivent alors s'interpréter comme des estimations de β/s et non de β en niveau. Cette normalisation est faite implicitement par tous les logiciels qui estiment un Logit.

(3) Densité, log-vraisemblance et score.

La densité de la loi logistique standard est la dérivée de Λ :

$$\Lambda'(t) = \frac{d}{dt}\,\frac{1}{1+e^{-t}} = \frac{e^{-t}}{(1+e^{-t})^2} = \frac{1}{1+e^{-t}} \cdot \frac{e^{-t}}{1+e^{-t}} = \Lambda(t)\big(1 - \Lambda(t)\big),$$

puisque $\dfrac{e^{-t}}{1+e^{-t}} = 1 - \Lambda(t)$.

Avec $s = 1$, $\Lambda_i = \Lambda(X_i\beta)$ et la vraisemblance d'une Bernoulli i.i.d. :

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

soit

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

On a $\partial\Lambda_i/\partial\beta = \Lambda_i(1 - \Lambda_i)\,X_i'$, d'où le score :

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

Le score prend exactement la forme « résidu × régresseur », $\sum_i (y_i - \Lambda_i)X_i'$: c'est une simplification spécifique au Logit (le facteur de pondération $\Lambda_i(1 - \Lambda_i)$ disparaît, ce qui n'est pas le cas pour le Probit). La condition du premier ordre n'a néanmoins pas de solution analytique en β (puisque Λ_i dépend de β de façon non linéaire) : l'estimation passe par un algorithme numérique (Newton-Raphson, BHHH, etc.).

(4) Information de Fisher.

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

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

Chaque terme est positif puisque
— $\Lambda_i \in (0,1)$ donc $\Lambda_i(1-\Lambda_i) > 0$;
— $(X_i a)^2 \geq 0$.
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 de la log-vraisemblance vaut exactement $\partial^2 \log L / \partial\beta\partial\beta' = -\sum_i \Lambda_i(1-\Lambda_i) X_i' X_i = -I(\beta)$ (elle ne dépend pas de y). Elle est définie négative, donc la log-vraisemblance du Logit est *globalement* strictement concave. L'éventuel maximum est donc 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) Cote, rapport de cotes et lien avec le Probit.

Sous la normalisation $s = 1$, $\mathbb{P}(y_i = 1 \mid X_i) = \Lambda(X_i\beta) = \frac{1}{1+e^{-X_i\beta}}$, donc $\mathbb{P}(y_i = 0 \mid X_i) = 1 - \Lambda(X_i\beta) = \frac{e^{-X_i\beta}}{1+e^{-X_i\beta}}$. La cote (*odds*) est le rapport de ces deux probabilités :

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

d'où, en prenant le logarithme,

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

Le Logit est donc linéaire *dans le log de la cote* (d'où son nom). En conséquence, augmenter x_k d'une unité (les autres variables étant fixées) multiplie la cote par e^{β_k} : la quantité e^{β_k} est le **rapport de cotes** (*odds ratio*) associé à x_k. Si $\beta_k > 0$ ($e^{\beta_k} > 1$) la cote augmente avec x_k ; si $\beta_k < 0$ elle diminue. C'est l'avantage interprétatif du Logit sur le Probit.

Lien avec le Probit. Le modèle latent n'identifie que β/σ_u. Le Probit fixe $\sigma_u = 1$ tandis que le Logit fixe $\sigma_u = \pi/\sqrt{3}$; en raisonnant comme à la question (2) sur un même modèle générateur, on obtient la règle empirique

$$\hat\beta_{\text{Logit}} \approx \frac{\pi}{\sqrt{3}}\,\hat\beta_{\text{Probit}} \approx 1{,}81\,\hat\beta_{\text{Probit}}, \quad \text{soit} \quad \hat\beta_{\text{Probit}} \approx \frac{\sqrt{3}}{\pi}\,\hat\beta_{\text{Logit}}.$$

Cette correspondance n'est qu'approximative, car les deux lois n'ont pas la même forme (la logistique a des queues plus épaisses que la normale). En revanche, les *probabilités* prédites $\hat F(X_i\hat\beta)$ sont presque identiques d'un modèle à l'autre : le choix entre Logit et Probit est, en pratique, surtout une affaire de convention et de commodité (le Logit fournissant des rapports de cotes faciles à interpréter via e^{β}).