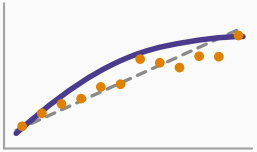
Économétrie avancée

Chapitre 17 : Variables dépendantes limitées et sélection d'échantillon



Jaouad Madkour

Faculté d'Économie et de Gestion de Tanger · 27 août 2026

Plan du chapitre

Avant de commencer

Pourquoi ce chapitre ?

Le modèle de probabilité linéaire (chapitre 7) traite y binaire comme n'importe quelle variable continue — avec des probabilités prédites parfois hors de $[0, 1]$. Beaucoup d'autres variables économiques sont **limitées** de façon similaire : nulles pour une fraction de la population (dépenses caritatives), dénombrées (nombre d'arrestations), censurées (durées suivies jusqu'à une date fixe) ou observées seulement pour un sous-échantillon non aléatoire (salaire offert, non observé pour les inactifs). Ce chapitre présente les modèles non linéaires — logit, probit, Tobit, Poisson, régression censurée/tronquée, correction de Heckman — conçus pour ces situations.

Modèles logit et probit pour réponse binaire

Spécification générale

Pour $y \in \{0, 1\}$, on remplace $P(y = 1 \mid x) = \beta_0 + x\beta$ (MPL) par

$$P(y = 1 \mid x) = G(\beta_0 + x\beta), \qquad 0 < G(z) < 1 \ \ \forall z,$$

ce qui garantit des probabilités toujours comprises entre 0 et 1. Le **modèle logit** prend $G(z) = \Lambda(z) = e^z/(1 + e^z)$ (cdf logistique) ; le **modèle probit** prend $G(z) = \Phi(z)$, la cdf de la loi normale centrée réduite. Les deux fonctions sont croissantes, en forme de S, et quasi indiscernables graphiquement.

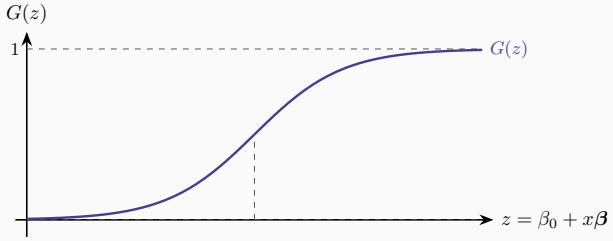
La variable latente sous-jacente

Logit et probit se déduisent d'un modèle à **variable latente** $y^* = \beta_0 + x\beta + e$, $y = \mathbb{1}[y^* > 0]$, où e suit une loi logistique ou normale standard indépendante de x. Le signe de β_j indique le sens de l'effet sur y^* — et donc sur $P(y = 1 \mid x)$ — mais y^* n'a pas d'unité de mesure interprétable ; seuls les effets marginaux sur la probabilité sont économiquement significatifs.

Maximum de vraisemblance

La densité conditionnelle est $f(y \mid x; \beta) = [G(x\beta)]^y [1 - G(x\beta)]^{1-y}$, d'où la log-vraisemblance individuelle $\ell_i(\beta) = y_i \log G(x_i\beta) + (1 - y_i) \log[1 - G(x_i\beta)]$. L'estimateur du maximum de vraisemblance (EMV) qui maximise $\sum_i \ell_i$ est consistant, asymptotiquement normal et asymptotiquement efficace sous conditions de régularité usuelles ; les tests t, de Wald et du rapport de vraisemblance ($LR = 2(\ell_{ur} - \ell_r) \sim \chi^2_q$) s'appliquent comme au chapitre 5.

Estimation et effets marginaux (suite)



Effet marginal

Pour x_j continue, $\dfrac{\partial p(x)}{\partial x_j} = g(\beta_0 + x\beta)\beta_j$, où $g = G'$ est une densité. Le signe de l'effet est toujours celui de β_j, mais son **ampleur dépend de** x — contrairement au MPL. En pratique on évalue $g(\cdot)$ aux valeurs moyennes des régresseurs ; pour x_j binaire, l'effet exact est la différence $G(\dots, x_j{=}1, \dots) - G(\dots, x_j{=}0, \dots)$.

Participation des femmes mariées au marché du travail

Sur les données MROZ, MCO (MPL), logit et probit s'accordent sur les signes et la significativité : *kidslt6* réduit fortement la probabilité de travailler. À la moyenne de l'échantillon, un enfant de moins de six ans fait chuter la probabilité estimée de participation d'environ 0{,}33 (calcul probit exact), contre un coefficient MPL brut de $-0{,}262$ — l'écart illustre les rendements décroissants qu'implique la forme en S, absents du modèle linéaire.

Le modèle Tobit

Variable latente censurée à zéro

Le **modèle Tobit** s'applique à y continue sur les valeurs strictement positives mais nulle avec probabilité non négligeable (dépenses caritatives, heures travaillées, montant investi en actions). Formellement, $y^* = \beta_0 + x\beta + u$, $u \mid x \sim \mathcal{N}(0, \sigma^2)$, et $y = \max(0, y^*)$. La vraisemblance combine une masse ponctuelle en $y = 0$, $P(y = 0 \mid x) = 1 - \Phi(x\beta/\sigma)$, et une densité continue pour $y > 0$.

Pourquoi pas MCO sur le sous-échantillon $y > 0$?

Estimer par MCO uniquement sur les observations positives revient à conditionner sur $y^* > 0$, ce qui corrèle le terme d'erreur retenu à x (troncature) : les MCO sont incohérents. Estimer par MCO sur l'échantillon complet, en traitant les zéros comme des observations ordinaires, est également incohérent car $E(y \mid x)$ n'est pas linéaire en $x\beta$.

Interprétation : deux espérances distinctes

Espérance conditionnelle et non conditionnelle

$$E(y \mid y > 0, x) = x\beta + \sigma\lambda(x\beta/\sigma), \qquad E(y \mid x) = \Phi(x\beta/\sigma)\, E(y \mid y > 0, x),$$

où $\lambda(c) = \phi(c)/\Phi(c)$ est le **ratio inverse de Mills**. Les effets partiels de x_j sur ces deux espérances ne sont **pas** simplement β_j : chacun est multiplié par un facteur d'ajustement strictement compris entre 0 et 1, qui dépend de $x\beta/\sigma$. Remarquablement, $\partial E(y \mid x)/\partial x_j = \beta_j\, \Phi(x\beta/\sigma)$ – le facteur le plus simple des trois.

Heures travaillées annuelles (MROZ)

Sur 753 femmes mariées (325 avec zéro heure), le coefficient Tobit sur *kidslt6* (-894) est environ le double du coefficient MCO (-442) – comparaison trompeuse en valeur brute. Une fois multiplié par le facteur d'ajustement pertinent ($\approx 0{,}451$ pour l'espérance conditionnelle à $y > 0$), l'effet Tobit redevient comparable en ordre de grandeur à l'estimation MCO, confirmant que le Tobit "moyenne" implicitement la décision de participer et la décision du volume d'heures.

Le modèle de régression de Poisson

Espérance exponentielle

Pour $y \in \{0, 1, 2, \dots\}$ (nombre d'arrestations, de brevets déposés, d'enfants), on spécifie $E(y \mid x) = \exp(\beta_0 + x\beta)$, garantissant des prédictions positives. Comme $\log E(y \mid x) = \beta_0 + x\beta$, les β_j s'interprètent en semi-élasticité — exactement comme dans un modèle linéaire pour $\log(y)$: $100\beta_j$ est approximativement le pourcentage de variation de $E(y \mid x)$ pour une hausse d'une unité de x_j ; l'effet exact discret est $\exp(\beta_j) - 1$.

Vraisemblance de Poisson et QMLE

Sous la loi de Poisson, $P(y = h \mid x) = \dfrac{e^{-\exp(x\beta)}[\exp(x\beta)]^h}{h!}$, d'où une log-vraisemblance simple à maximiser. La loi de Poisson impose $\mathrm{Var}(y \mid x) = E(y \mid x)$; en pratique la variance excède souvent la moyenne (**surdispersion**). L'EMV de Poisson reste néanmoins consistant même si cette égalité échoue (**quasi-maximum de vraisemblance**, QMLE) ; il suffit de corriger les erreurs-types par un facteur $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ estimé à partir des résidus standardisés.

Variables de comptage (suite)

Nombre d'arrestations en 1986

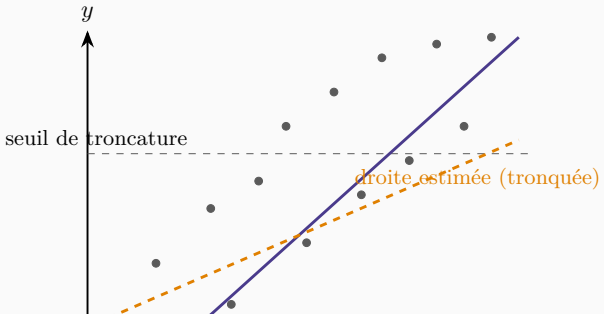
Sur 2 725 jeunes hommes (1 970 avec zéro arrestation), le coefficient de Poisson sur *pcnv* (proportion de condamnations antérieures) implique qu'une hausse de 0{,}10 de ce taux réduit le nombre attendu d'arrestations d'environ 4 %. L'estimation de $\hat{\sigma} \approx 1{,}23$ signale une surdispersion modérée : les erreurs-types Poisson brutes doivent être gonflées d'environ 23 % pour rester valides.

Régression censurée et régression tronquée

Deux problèmes de données distincts

Censure vs troncature

La **régression censurée** conserve toutes les observations, y compris celles pour lesquelles y n'est connu qu'au-delà (ou en-deçà) d'un seuil c_i : $w_i = \min(y_i, c_i)$ — typiquement un plafonnement (*top coding*) ou une durée suivie jusqu'à une date de fin d'étude. La **régression tronquée** exclut entièrement de l'échantillon les unités situées au-delà du seuil : on n'observe ni y_i ni x_i pour ces unités. Les deux problèmes se traitent par maximum de vraisemblance sous normalité, mais la troncature perd strictement plus d'information.



Corrections de sélection d'échantillon

Sélection exogène versus endogène

Quand les MCO sur l'échantillon sélectionné restent valides

Soit $s_i = 1$ si l'observation i est retenue, 0 sinon. Régresser y sur x sur le sous-échantillon $\{s_i = 1\}$ équivaut à régresser $s_i y_i$ sur $s_i x_i$ sur l'échantillon complet. Les MCO restent consistants si $E(s\,u \mid s\,x_1, \ldots, s\,x_k) = 0$ — condition satisfaite dès lors que la sélection s ne dépend que des régresseurs exogènes (**sélection exogène**) ou d'un tirage indépendant de u. La sélection est **endogène**, et les MCO incohérents, dès que s dépend directement de u — c'est le cas typique de la troncature sur y.

Troncature incidente et équation de sélection

Pour le modèle d'intérêt $y = x\beta + u$ observé seulement si $s = 1$, avec $s = \mathbb{1}[z\gamma + v \geq 0]$ et (u, v) bivariée normale, on montre que

$$E(y \mid z, s = 1) = x\beta + \rho\,\lambda(z\gamma).$$

La **méthode de Heckman en deux étapes** : (i) estimer γ par probit de s sur z (échantillon complet), calculer $\widehat{\lambda}_i = \lambda(z_i\widehat{\gamma})$; (ii) régresser y sur x et $\widehat{\lambda}$ sur le sous-échantillon sélectionné. Le test t sur $\widehat{\lambda}$ teste $H_0 : \rho = 0$ (absence de biais de sélection). La procédure exige idéalement au moins une variable de z absente de x — sinon $\widehat{\lambda}$ est presque colinéaire avec x.

Équation de salaire offert pour femmes mariées (Heckit)

Sur MROZ, l'équation de $\log(\text{salaire})$ estimée sur les 428 femmes actives donne un coefficient d'éducation de 0{,}108 (MCO) contre 0{,}109 (Heckit) — pratiquement identique. Le coefficient sur $\widehat{\lambda}$ est non significatif ($t = 0{,}239$) : rien n'indique de biais de sélection important pour cette relation, une fois contrôlé par éducation et expérience.

Synthèse

Ce qu'il faut retenir

L'essentiel du chapitre 16

- Logit et probit corrigent les défauts du MPL (probabilités hors $[0, 1]$, effets constants) au prix d'une interprétation plus lourde ; estimation par maximum de vraisemblance.
- Le Tobit modélise des solutions en coin ($y \geq 0$ avec masse en zéro) ; les effets partiels sur $E(y \mid x)$ et $E(y \mid y > 0, x)$ diffèrent des coefficients bruts par des facteurs d'ajustement non linéaires.
- La régression de Poisson (et son extension QMLE) traite les variables de comptage ; interprétation en semi-élasticités, correction de surdispersion.
- Censure et troncature sont deux problèmes de données distincts ; la censure conserve plus d'information et produit des estimations plus fiables.
- La sélection d'échantillon ne biaise les MCO que si elle est endogène ; la méthode de Heckman (probit puis régression augmentée du ratio inverse de Mills) permet de tester et corriger ce biais.

Vers le chapitre 17

Après les modèles pour variables limitées en coupe transversale, le chapitre suivant revient aux séries temporelles pour des sujets avancés : dynamique de court terme via les modèles à correction d'erreur, tests de racine unitaire, et estimation de systèmes avec variables prédéterminées.