Econométrie 2

Première partie : variables dépendantes limitées

Xavier d'Haultfœuille

1 Introduction

Dans le cours d'économétrie 1, on a considéré des modèles de la forme :

$$Y = X'\beta_0 + \varepsilon, \tag{1}$$

où Y est une variable continue. On va considérer dans ce cours plusieurs extensions de ce cadre général :

- La variable dépendante est *discrète* : $Y \in \{1, ..., K\}$. Exemples : activité ou non, mode de transport, vote...
- La variable dépendante est *censurée* : on observe seulement $\min(Y, C)$ (ou $\max(Y, C)$) et $\mathbb{1}\{Y \geq C\}$. Exemple : durées de chômage ou de survie à une maladie, consommation d'un bien, etc.
- Modèles de *sélection* : on a bien (1) mais on observe Y seulement si $D = 1$ ($D \in \{0, 1\}$). Exemple : équations de salaire, non-réponse, participation à un programme (formation, emplois aidés...).
- On observe un même individu plusieurs fois dans le temps grâce à des *données de panel*. Comment utiliser cette dimension longitudinale pour résoudre l'endogénéité, analyser la dynamique du modèle ?

2 Plan

Première partie : **variables dépendantes limitées**.

- Modèles dichotomiques (2 séances)
  - Modèles logit et probit : identification, estimation, qualité du modèle, différence entre les deux.
  - Remise en cause des hypothèses du modèle : homoscédasticité, exogénéité.
- Extensions du modèle dichotomique (1 séance 1/2)
  - Modèles polytomiques ordonnés : seuils connus et inconnus.
  - Modèles polytomiques non ordonnés. Logit multinomial, Modèles alternatifs : probit multinomial, choix séquentiels, choix simultanés.
- Censure et sélection (1 séance 1/2)
  - Modèle de censure : tobit simple.
  - Modèles de sélection : sélection exogène, sélection généralisée.

Deuxième partie : **GMM et données de panels**.

- Méthode des moments généralisés (2 séances)
  - Définition, convergence, optimalité.
  - Tests de spécifications.
  - Applications aux variables instrumentales dans le cas hétéroscédastique.
- Introduction à l'économétrie des panels (3 séances)
  - Modèle à effets aléatoires, modèle à effet fixe, test d'effet aléatoire.
  - Estimation avec exogénéité faible : estimation par GMM, tests de suridentification.
  - Application aux panels autorégresssifs.

3 Bibliographie indicative

Ouvrages de références :

1. Amemiya, T. (1985), Advanced Econometrics, Basil Blackwell, Oxford.
2. Gouriéroux, C. (1989), Econométrie des variables qualitatives, Economica.
3. Greene, W.H. (1995), Econometric Analysis (chap. 13, 18, 21 et 22), Prentice Hall.
4. Maddala, G.S. (1983), Limited Dependent and Qualitative variables in econometrics, Cambridge University Press.
5. Thomas, A. (2000), Econométrie des variables qualitatives, Dunod.
6. **Wooldridge, J.M. (2002), Econometric Analysis of Cross Section and Panel Data (chap. 10, 11, 14 à 17), MIT Press.**

Chapitre 1

Modèles binaires

1 Introduction

On cherche à expliquer Y dichotomique par $X = (X_1, ..., X_K) \in \mathbb{R}^K$. Les deux valeurs possibles de Y étant arbitraires, on posera toujours $Y \in \{0, 1\}$. Les variables dichotomiques sont très largement répandues :

- En microéconomie : activité vs inactivité, emploi vs chômage, consommation ou non d'un bien durable etc.
- En risques de crédit : défaut ou non d'un emprunteur.
- En assurance : sinistre ou non.
- En biostatistique : individu malade ou non, traitement efficace ou non.
- En sciences sociales : obtention d'un diplôme ou non, couple vs célibat, vote vs abstention etc.

Les modèles linéaires sont mal adaptés pour étudier ce genre de variables. On a alors recours à des modèles binaires, les plus courants étant le logit et le probit.

- Ces modèles ont d'abord été introduits en biostatistique (Gaddum, 1933, Bliss, 1935, Berkson, 1944).
- Ils n'ont fait leur apparition en économie que dans les années 70 (cf. McFadden, 1974), avec le développement des bases de données individuelles permettant l'estimation de modèles microéconomiques.

2 Présentation des modèles binaires

Dans le modèle linéaire sous hypothèse d'exogénéité $E(\varepsilon|X) = 0$, on a

$$E(Y|X) = X'\beta_0.$$

Si $Y \in \{0, 1\}$, cette modélisation est mal adaptée. En effet, dans ce cas,

$$E(Y|X) = P(Y = 1|X) \in [0,\, 1].$$

Or rien n'assure que $X'\beta_0 \in [0,\, 1]$.

Pour que cette dernière condition soit satisfaite, on va supposer que

$$E(Y|X) = F(X'\beta_0) \tag{2}$$

où $F(.)$ est une fonction (connue) strictement croissante bijective de $\mathbb{R}$ dans $]0,\, 1[$, donc une fonction de répartition. Notons que l'équation (2) est celle d'un modèle linéaire généralisé (GLM en anglais), c'est-à-dire un modèle de la forme :

$$h(E(Y|X)) = X'\beta_0$$

où h est une fonction connue (dite fonction de lien).

Le modèle (2) peut s'interpréter en termes de *variables latentes*. Supposons qu'il existe une variable continue $Y^* \in \mathbb{R}$ telle que,

$$Y = \mathbb{1}\{Y^* \geq 0\}.$$

Supposons par ailleurs que Y^* suive un modèle linéaire :

$$Y^* = X'\beta_0 + \varepsilon \tag{3}$$

où $-\varepsilon$ est indépendante de X et a pour fonction de répartition F. Alors

$$P(Y = 1|X) = P(X'\beta_0 + \varepsilon \geq 0|X) = P(-\varepsilon \leq X'\beta_0|X) = F(X'\beta_0).$$

On retrouve donc l'équation (2).

L'interprétation en termes de variables latentes est, très souvent, naturelle.

Exemple 1 (microéconomie) : supposons que Y corresponde à un choix binaire de la part d'un agent. Soit U_1 l'utilité (espérée) de l'agent s'il décide $Y = 1$, U_0 son utilité s'il décide $Y = 0$. Posons également $Y^* = U_1 - U_0$ la différence d'utilité entre les deux choix. Si l'agent est rationnel, il décide en maximisant son utilité espérée :

$$Y = \mathbb{1}\{U_1 \geq U_0\} = \mathbb{1}\{Y^* \geq 0\}.$$

Exemple 2 (finance d'entreprise) : le défaut ($Y = 1$) d'une entreprise survient lorsque la dette de l'entreprise D dépasse un certain seuil S (éventuellement aléatoire). On a alors $Y^* = D - S$.

Exemple 3 (biostatistique) : un individu sera considéré comme guéri ($Y = 0$) lorsque le nombre de bactéries N (s'il s'agit d'une pathologie bactériologique) est descendu en dessous d'un certain seuil S (dépendant éventuellement de l'individu). On a alors $Y^* = N - S$.

Exemple 4 (éducation) : un individu réussira à obtenir son diplôme ($Y = 1$) si sa moyenne M est supérieure à un seuil fixé s. On a alors $Y^* = M - s$.

Deux cas particuliers importants : le probit et le logit

Plusieurs choix sont possibles pour F. Les plus courants sont les deux suivants :

- $F = \Phi$, fonction de répartition d'une loi normale standard. On parle alors de modèle probit. Ceci est équivalent à $\varepsilon \sim \mathcal{N}(0, 1)$ dans l'équation (3).

- $F(x) = \Lambda(x) = 1/(1 + \exp(-x))$, fonction de répartition d'une loi *logistique*, on parle de modèle logit.

Notons que ces deux lois sont proches en pratique. L'intérêt du modèle logit est sa simplicité.

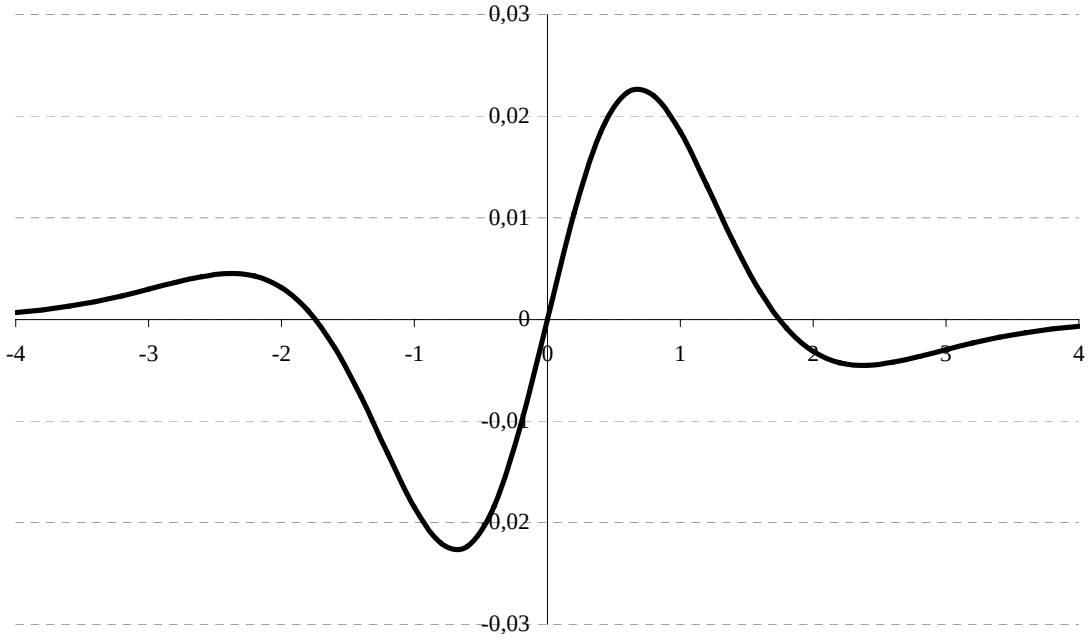

FIG. 1 – $x \mapsto \Lambda(x\sqrt{3}/\pi) - \Phi(x)$.

Lorsque $|x| \to +\infty$,

$$\varphi(x) = \Phi'(x) \propto e^{-x^2/2}, \quad \Lambda'(x) = \Lambda(x)(1 - \Lambda(x)) = O(e^{-|x|}).$$

En d'autres termes, la loi logistique possède des queues de distribution plus épaisses que la loi normale.

3 Interprétation des paramètres

- *Au niveau qualitatif*, la $k-$ème composante de X_i aura un effet positif sur $P(Y = 1|X)$ ssi $\beta_{0k} > 0$.

- *Au niveau quantitatif*, l'interprétation de β_{0k} est plus délicate.

Dans le modèle linéaire standard $E(Y|X) = X'\beta_0$, l'effet β_{0k} de la $k-$ème composante peut s'interpréter comme l'effet marginal d'une modification de X_k :

$$\frac{\partial E(Y|X_1 = x_1, ..., X_K = x_K)}{\partial x_k} = \beta_{0k}.$$

Cette valeur est indépendante de $x_{-k} = (x_1, ..., x_{k-1}, x_{k+1}, ..., x_K)$.

En revanche, dans les modèles binaires (et non-linéaires plus généralement), l'effet marginal d'une variable dépend de la valeur des autres variables :

$$\frac{\partial E(Y|X = x)}{\partial x_k} = f(x'\beta_0)\beta_{0k}$$

où $f = F'$. L'effet marginal dépend donc de x_{-k}. Pour des f usuels, l'effet d'une variable sur $P(Y = 1|X)$ est d'autant plus fort que $x'\beta_0$ est proche de 0 ou que $P(Y = 1|X) \simeq 0.5$.

Notons qu'on a toujours

$$\frac{\beta_{0k}}{\beta_{0j}} = \frac{\partial E(Y|X=x)/\partial x_k}{\partial E(Y|X=x)/\partial x_j}$$

Donc la comparaison des différents paramètres est licite.

Outre l'estimation de β_{0k}, il est intéressant d'estimer l'effet marginal à la moyenne des observations $f(E(X)'\beta_0)\beta_{0k}$ ou l'effet marginal moyen $E\left[f(X'\beta_0)\right]\beta_{0k}$.

N.B. : pour les variables discrètes (dichotomiques) l'effet marginal est remplacé par

$$F(x'_{-k}\beta_{0-k} + \beta_{0k}) - F(x'_{-k}\beta_{0-k}).$$

Une spécificité du modèle logit : les odds-ratios.

On définit le risque (ou odd) comme égal à $P(Y = 1|X)/P(Y = 0|X)$. Le risque est à peu près égal à $P(Y = 1|X)$ lorsque cette probabilité est faible (cf. cas de maladies).

Dans le cas du logit :

$$\frac{P(Y = 1|X = x)}{P(Y = 0|X = x)} = \frac{1/(1 + e^{-x'\beta_0})}{e^{-x'\beta_0}/(1 + e^{-x'\beta_0})} = e^{x'\beta_0}$$

Considérons une variable explicative $X_k \in \{0, 1\}$. On a alors :

$$e^{\beta_{0k}} = \frac{P(Y = 1|X_{-k} = x_{-k}, X_k = 1)/P(Y = 0|X_{-k} = x_{-k}, X_k = 1)}{P(Y = 1|X_{-k} = x_{-k}, X_k = 0)/P(Y = 0|X_{-k} = x_{-k}, X_k = 0)}.$$

Donc $e^{\beta_{0k}}$ est égal au rapport des risques (odd-ratio) correspondant à $X_k = 1$ et $X_k = 0$. Il est indépendant de la valeur de X_{-k}.

Exemple : probabilité d'occurrence de cancer en fonction de caractéristiques X_{-k} et du fait de fumer $X_k = 1$ ou non ($X_k = 0$). Si $\beta_{0k} = 1.1$, cela signifie que toutes choses égales par ailleurs, on multiplie son risque de contracter un cancer par trois en fumant.

4 Identification du modèle

Revenons à l'équation :

$$Y = \mathbb{1}\{X'\beta_0 + \varepsilon \geq 0\}.$$

Deux questions :

- pourquoi fixer le seuil à 0 ?
- pourquoi fixer la variance de ε (à 1 pour le probit, à $\pi^2/3$ pour le logit) ?

(remarque : ceci est équivalent à imposer F connue)

Raison : le modèle n'est pas identifiable sinon. En effet, en posant $X_1 = 1$, on a

$$Y = \mathbb{1}\{\beta_{01} + X'_{-1}\beta_{0-1} + \varepsilon \geq s\} \Longleftrightarrow Y = \mathbb{1}\{\beta_{01} - s + X'_{-1}\beta_{0-1} + \varepsilon \geq 0\}$$

En d'autres termes, on ne peut pas identifier séparément la constante β_{01} et le seuil s. On fixe donc (arbitrairement) s à 0.

De même, on ne peut pas identifier de façon jointe β_0 et la variance σ^2 du résidu ε. En effet,

$$Y = \mathbb{1}\{X'\beta_0 + \sigma\varepsilon \geq 0\} \Longleftrightarrow Y = \mathbb{1}\{X'(\beta_0/\sigma) + \varepsilon \geq 0\}.$$

On fixe donc arbitrairement σ à 1.

Théorème 1 *Si s et σ sont fixés et $E(XX')$ est inversible, le modèle est identifié.*

Preuve : soit P_β la loi des observations lorsque le vrai paramètre est β. Il s'agit de montrer que la fonction

$$\beta \mapsto P_\beta$$

est injective. Dans notre modèle conditionnel, on peut montrer que l'identification est équivalente à

$$P_\beta(Y = 1|X) = P_{\beta'}(Y = 1|X) \text{ p.s.} \Rightarrow \beta = \beta' \quad \forall(\beta, \beta').$$

Or

$$(E(XX') \text{ est inversible }) \Longleftrightarrow (X'\lambda = 0 \Longrightarrow \lambda = 0)$$

Par conséquent,

$$\begin{aligned} P_\beta(Y = 1|X) = P_{\beta'}(Y = 1|X) \text{ p.s.} &\iff F(X'\beta) = F(X'\beta') \text{ p.s.} \\ &\iff X'\beta = X'\beta' \text{ p.s.} \\ &\iff \beta = \beta'. \end{aligned}$$

Donc le modèle est identifié.

5 Estimation du modèle

On s'intéresse maintenant à l'estimation de β_0 à partir d'un échantillon i.i.d. $((Y_1, X_1), ..., (Y_n, X_n))$. Comme le modèle est paramétrique, on peut l'estimer par maximum de vraisemblance.

Rappel : si la loi de probabilité P_β de (Y, X) est absolument continue p/r à une mesure μ, on appelle vraisemblance la fonction L vérifiant

$$\frac{dP_\beta}{d\mu}(y, x) = L(y, x; \beta).$$

Dans un modèle conditionnel, la loi de X ne dépend pas de β donc on peut écrire

$$L(y, x; \beta) = L_1(y|x; \beta)L_x(x) \tag{4}$$

On s'intéresse uniquement à $L_1(.|.; .)$ puisque $L_x(.)$ ne dépend pas de β.

Ici, $Y \in \{0, 1\}$ donc on choisit $\mu =$ la mesure de comptage. On a alors

$$\begin{aligned} L_1(y|x; \beta) &= P(Y = y|X = x) \\ &= [P(Y = 1|X = x)]^y \ [P(Y = 0|X = x)]^{1-y} \\ &= F(x'\beta)^y \ (1 - F(x'\beta))^{1-y} \end{aligned}$$

La vraisemblance conditionnelle d'un échantillon i.i.d. $(\mathbf{Y}, \mathbf{X}) = ((Y_1, X_1), ..., (Y_n, X_n))$ s'écrit alors

$$L_n(\mathbf{Y}|\mathbf{X}; \beta) = \prod_{i=1}^{n} F(X_i'\beta)^{Y_i} \, (1 - F(X_i'\beta))^{1-Y_i}$$

Un estimateur du maximum de vraisemblance (EMV) est alors défini par :

$$\widehat{\beta} \in \arg\max_{\beta \in \mathbb{R}^k} L_n(\mathbf{Y}|\mathbf{X}; \beta).$$

Remarques :

i. en général, cet estimateur n'est pas unique, et il peut ne pas exister.

ii. On maximise plutôt la log-vraisemblance qui a une forme plus simple :

$$l_n(\mathbf{Y}|\mathbf{X}; \beta) = \sum_{i=1}^{n} Y_i \ln\left(F(X_i'\beta)\right) + (1 - Y_i) \ln\left(1 - F(X_i'\beta)\right)$$

Conditions du premier ordre.

On a $\partial F(X_i'\beta)/\partial\beta = f(X_i'\beta)X_i$. Donc :

$$\frac{\partial l_n}{\partial \beta}(\mathbf{Y}|\mathbf{X};\beta) = \sum_{i=1}^{n}\left[Y_i\frac{f(X_i'\beta)}{F(X_i'\beta)} + (1-Y_i)\frac{-f(X_i'\beta)}{1-F(X_i'\beta)}\right]X_i \tag{5}$$

Soit encore

$$\frac{\partial l_n}{\partial \beta}(\mathbf{Y}|\mathbf{X};\beta) = \sum_{i=1}^{n}\frac{f(X_i'\beta)}{F(X_i'\beta)(1-F(X_i'\beta))}\left[Y_i(1-F(X_i'\beta)) - (1-Y_i)F(X_i'\beta)\right]X_i.$$

Finalement

$$\frac{\partial l_n}{\partial \beta}(\mathbf{Y}|\mathbf{X};\beta) = \sum_{i=1}^{n}\frac{f(X_i'\beta)}{F(X_i'\beta)(1-F(X_i'\beta))}\left[Y_i - F(X_i'\beta)\right]X_i.$$

Les conditions du premier ordre s'écrivent donc :

$$\sum_{i=1}^{n}\frac{f(X_i'\widehat{\beta})}{F(X_i'\widehat{\beta})(1-F(X_i'\widehat{\beta}))}\left[Y_i - F(X_i'\widehat{\beta})\right]X_i = 0 \tag{6}$$

qui n'admet pas de solution analytique simple en général.

Conditions du second ordre.

- Dans le cas du logit, on a $\Lambda' = \Lambda(1-\Lambda)$, donc

$$\frac{\partial^2 l_n}{\partial\beta\partial\beta'}(\mathbf{Y}|\mathbf{X};\beta) = -\sum_{i=1}^{n} \Lambda'(X_i'\beta) X_i X_i' << 0.$$

La log-vraisemblance est bien strictement concave.

- Dans le cas du probit, en notant $Z_i = X_i'\beta$ et en utilisant (5), $1-\Phi(x) = \Phi(-x)$, la parité de φ et $\varphi'(x) = -x\varphi(x)$:

$$\frac{\partial^2 l_n}{\partial\beta\partial\beta'}(\mathbf{Y}|\mathbf{X};\beta) = -\sum_{i=1}^{n} \left[Y_i \frac{\varphi(Z_i)\left(Z_i\Phi(Z_i)+\varphi(Z_i)\right)}{\Phi^2(Z_i)} + (1-Y_i)\frac{\varphi(-Z_i)\left(-Z_i\Phi(-Z_i)+\varphi(-Z_i)\right)}{\Phi^2(-Z_i)} \right] X_i X_i'.$$

Puisque (exercice) pour tout x, $x\Phi(x)+\varphi(x) = \int_{-\infty}^{x} \Phi(t)dt > 0$, la log-vraisemblance est bien strictement concave.

- Dans le cas général, le programme n'est pas nécessairement concave et il peut y avoir plusieurs solutions.

Remarque 1 : l'estimateur peut être obtenu numériquement par un algorithme de Newton-Raphson. Partant de $\beta^{(0)}$ quelconque, on définit la suite $\left(\beta^{(m)}\right)_{m\in\mathbb{N}}$ par :

$$\beta^{(m+1)} = \beta^{(m)} - \left[H^{(m)}\right]^{-1} \frac{\partial l_n}{\partial \beta}(\beta^{(m)})$$

où $H^{(m)}$ est le gradient de $\partial l_n/\partial\beta$ en $\beta^{(m)}$ (donc une matrice ici) :

$$H^{(m)} = \sum_{i=1}^{n} \frac{\partial l_1^2}{\partial\beta\partial\beta'}(Y_i|X_i;\beta^{(m)}).$$

Par concavité de $\beta \mapsto l_n(\mathbf{Y}|\mathbf{X};\beta)$, la suite $\beta^{(m)}$, si elle converge, tend nécessairement vers l'EMV.

Cependant, si dans l'échantillon il existe une variable X_k dichotomique telle que $y_i = 1$ (ou $y_i = 0$) pour tout i tel que $x_{ik} = 1$, l'estimateur n'existe pas. En effet, la $k-$ème composante de $\partial l_n/\partial\beta$ s'écrit, si les $y_i = 1$,

$$\sum_{i=1}^{n} \frac{f(x_i'\beta)}{F(x_i'\beta)(1-F(x_i'\beta))} \left[y_i - F(x_i'\beta)\right] x_{ik} = \sum_{i=1}^{n} \frac{f(x_i'\beta)}{F(x_i'\beta)} x_{ik} > 0 \quad \forall\beta$$

Intuitivement, l'échantillon nous dit que l'effet de X_k est infini.

Remarque 2 : lien avec les moindres carrés non linéaires (MCNL).

Le modèle (2) peut se réécrire

$$Y = F(X'\beta_0) + \eta$$

où $E(\eta|X) = 0$. Plutôt que d'utiliser l'EMV, on peut donc, comme dans le modèle linéaire, penser à estimer β_0 par

$$\widetilde{\beta} = \arg\min_b \sum_{i=1}^{n} (Y_i - F(X_i'b))^2$$

Cette idée se justifie par la définition de l'espérance conditionnelle. En effet, en notant $m : x \mapsto E(Y|X = x)$,

$$m(.) = \arg\min_{g(.)} E\left[(Y - g(X))^2\right]$$

En fait, on peut montrer que pour toute fonction $h(.)$ positive,

$$m(.) = \arg\min_{g(.)} E\left[h(X)(Y - g(X))^2\right]$$

Ici,

$$\beta_0 = \arg\min_b E\left[h(X)(Y - F(X'b))^2\right]$$

On peut donc estimer β_0 par la contrepartie empirique de ce programme :

$$\widehat{\beta}_{MCNL} = \arg\min_b \sum_{i=1}^{n} h(X_i)(Y_i - F(X_i'b))^2$$

La CPO s'écrit :

$$\sum_{i=1}^{n} h(X_i) f(X_i'\widehat{\beta}_{MCNL})(Y_i - F(X_i'\widehat{\beta}_{MCNL}))X_i = 0. \tag{7}$$

L'estimateur $\widehat{\beta}_{MCNL}$ dépend de la fonction $h(.)$ choisie. On peut montrer que le h optimal s'écrit (comme dans le modèle linéaire !) :

$$h(X) = \frac{1}{V(Y|X)}.$$

Or ici, $Y|X \sim Be(F(X'\beta_0))$. Donc

$$V(Y|X) = F(X'\beta_0)\left(1 - F(X'\beta_0)\right).$$

Cependant, β_0 est inconnu. Si on le remplace par un estimateur convergent $\widehat{\beta}_1$ (obtenu par exemple avec $h = 1$), l'équation (7) se réécrit :

$$\sum_{i=1}^{n} \frac{f(X_i'\widehat{\beta}_{MCNL})}{F(X'\widehat{\beta}_1)\left(1 - F(X'\widehat{\beta}_1)\right)}(Y_i - F(X_i'\widehat{\beta}_{MCNL}))X_i = 0. \tag{8}$$

On retrouve pratiquement les conditions du premier ordre (6) satisfaites par l'EMV. On les retrouve exactement si on poursuit le procédé, en remplaçant $\widehat{\beta}_1$ par $\widehat{\beta}_2 = \widehat{\beta}_{MCNL}$ dans (8), et en recommençant.

Propriétés asymptotiques

Proposition 2 *si $|X| \leq M$ p.s. (pour simplifier), on a*

$$\widehat{\beta} \xrightarrow{P} \beta_0$$

$$\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, I_1^{-1}(\beta_0)\right).$$

où $I_1(\beta_0)$ est l'information de Fisher. De plus,

$$I_1(\beta_0) = E\left(\frac{f^2(X'\beta_0)}{F(X'\beta_0)(1 - F(X'\beta_0))} XX'\right).$$

On peut l'estimer de façon convergente par :

$$\widehat{I_1(\beta_0)} = \frac{1}{n}\sum_{i=1}^{n} \frac{f^2(X_i'\widehat{\beta})}{F(X_i'\widehat{\beta})(1 - F(X_i'\widehat{\beta}))} X_i X_i'.$$

Rappel : l'EMV présente l'intérêt d'être le meilleur estimateur "régulier" asymptotiquement. Ainsi, si l'on consière un autre estimateur $\widetilde{\beta}$ vérifiant :

$$\sqrt{n}\left(\widetilde{\beta} - \beta_0\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, V)$$

On aura nécessairement

$$V >> I_1^{-1}(\beta_0).$$

Preuve de la proposition : la 1ère partie se déduit des théorèmes généraux sur l'EMV (pour la preuve formelle du résultat dans le cas précis des modèles binaires, cf. par exemple Van der Vaart, Asymptotic Statistics, exemple 5.40).

Montrons la formule sur l'information de Fisher. On a

$$I_1(\beta_0) = V\left(\frac{\partial l_1}{\partial \beta}(Y|X;\beta_0)\right)$$

Par décomposition de la variance :

$$I_1(\beta) = E\left[V\left(\frac{\partial l_1}{\partial \beta}(Y|X;\beta_0)\middle|X\right)\right] + V\left[E\left(\frac{\partial l_1}{\partial \beta}(Y|X;\beta_0)\middle|X\right)\right]$$

Or (cf. équation (6))

$$\frac{\partial l_1}{\partial \beta}(Y|X;\beta_0) = \frac{f(X'\beta_0)}{F(X'\beta_0)(1-F(X'\beta_0))}\left[Y - F(X'\beta_0)\right]X$$

Donc

$$\begin{aligned} E\left(\frac{\partial l_1}{\partial \beta}(Y|X;\beta_0)\middle|X\right) &= 0. \\ V\left(\frac{\partial l_1}{\partial \beta}(Y|X;\beta_0)\middle|X\right) &= \frac{f^2(X'\beta_0)XX'}{F(X'\beta_0)(1-F(X'\beta_0))}. \end{aligned}$$

d'où le résultat.

Enfin, on peut prouver que $\widehat{I_1(\beta_0)}$ converge vers $I_1(\beta_0)$ en démontrant que

$$\sup_{\beta \in K} \left| \frac{1}{n} \sum_{i=1}^{n} \frac{f^2(X_i'\beta)}{F(X_i'\beta)(1-F(X_i'\beta))} X_i X_i' - E\left(\frac{f^2(X'\beta)}{F(X'\beta)(1-F(X'\beta))} XX' \right) \right| \xrightarrow{P} 0$$

où K est un compact incluant β_0 (cf. Van der Vaart, Asymptotic Statistics, chapitre 19) $\square$

Remarques :

i. On a montré que le score conditionnel est centré :

$$E\left(\frac{\partial l_1}{\partial \beta}(Y|X;\beta_0) \middle| X \right) = 0.$$

C'est un résultat général dans un modèle conditionnel (montrez-le !). On a donc toujours dans ce cas :

$$I_1(\beta_0) = E\left[V\left(\frac{\partial l_1}{\partial \beta}(Y|X;\beta_0) \middle| X \right) \right]$$

ii. On peut démontrer directement (exercice) que $\widehat{I_1(\beta_0)}$ converge vers $I_1(\beta_0)$ dans le modèle logit, en utilisant le fait que la dérivée de

$$g(x) = \frac{f^2(x)}{F(x)(1-F(x))} \ (x \in \mathbb{R})$$

est bornée.

6 Tests d'hypothèses.

On souhaite tester une hypothèse du type

$$H_0 \ : R\beta_0 = 0 \ \text{ contre } \ H_1 \ : R\beta_0 \neq 0 \quad (R \ \text{ matrice } p \times K, \ p \leq K).$$

Par exemple, $\beta_{0k} = 0$ ou $\beta_{02} = ... = \beta_{0K} = 0$.
On utilise l'un des trois tests liés au maximum de vraisemblance : le test de Wald, le test du score ou le test de rapport de vraisemblance. Les statistiques de test correspondantes s'écrivent :

$$\begin{aligned}
\xi_n^W &= n\widehat{\beta}' R' \left[R \, \widehat{I_1(\beta_0)}^{-1} \, R' \right]^{-1} R\widehat{\beta} \\
\xi_n^S &= \frac{1}{n} \frac{\partial l_n}{\partial \beta'}(\mathbf{Y}|\mathbf{X}; \widehat{\beta}_C) \, \widehat{I_1(\beta_0)}^{-1} \, \frac{\partial l_n}{\partial \beta}(\mathbf{Y}|\mathbf{X}; \widehat{\beta}_C) \\
\xi_n^R &= 2 \left[l_n(\mathbf{Y}|\mathbf{X}; \widehat{\beta}) - l_n(\mathbf{Y}|\mathbf{X}; \widehat{\beta}_C) \right]
\end{aligned}$$

où $\widehat{\beta}_C$ est l'estimateur du maximum de vraisemblance contraint, i.e. estimé sous H_0.

Sous H_0, ces trois statistiques tendent en loi vers un $\chi^2(p)$. Pour les trois tests, la région critique d'un test de niveau asymptotique α est donc de la forme $W = \{\xi_n > q_p(1-\alpha)\}$ où $q_p(y)$ est le quantile d'ordre y d'une $\chi^2(p)$.

7 Prévision.

Supposons qu'on connaisse X mais pas Y. On suppose pour l'instant β_0 connu. Il s'agit de décider si $Y = 0$ ou 1. On note :

- $\widehat{Y} = f(X) \in \{0, 1\}$ la décision prise ;

- C_1 le coût associé à la décision $\widehat{Y} = 0$ lorsque $Y = 1$;

- C_0 le coût associé à la décision $\widehat{Y} = 1$ lorsque $Y = 0$.

On cherche à minimiser le coût moyen sachant X. Ce coût $C(\widetilde{Y})$ s'écrit, pour une décision $\widetilde{Y}$ donnée,

$$\begin{aligned} C(\widetilde{Y}) &= \mathbb{1}\{\widetilde{Y} = 0\} C_1 E(\mathbb{1}\{Y = 1\}|X) + \mathbb{1}\{\widetilde{Y} = 1\} C_0 E(\mathbb{1}\{Y = 0\}|X) \\ &= \mathbb{1}\{\widetilde{Y} = 0\} C_1 F(X'\beta_0) + \mathbb{1}\{\widetilde{Y} = 1\} C_0 (1 - F(X'\beta_0)) \end{aligned}$$

La meilleure prévision $\widehat{Y}$ de Y vérifie :

$$\widehat{Y} = \arg\min_{\widetilde{Y}} C(\widetilde{Y})$$

On a alors, en notant $Z = F(X'\beta)$,

$$\widehat{Y} = \mathbb{1}\left\{ Z \geq \frac{C_0}{C_0 + C_1} \right\}.$$

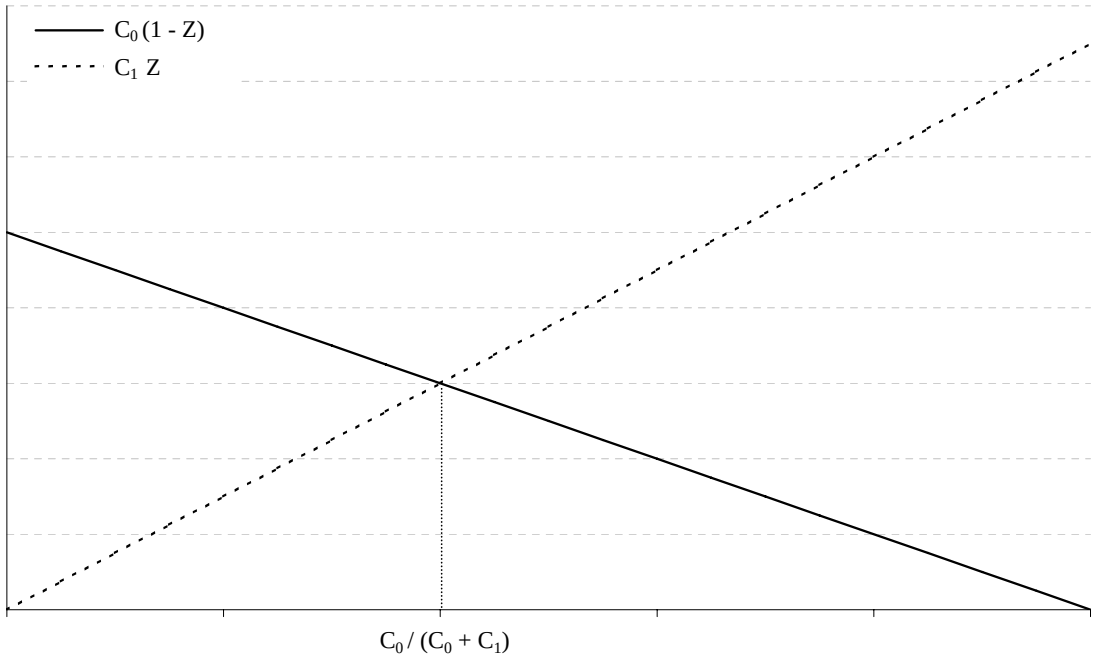

FIG. 2 – Comparaison des coûts associés à $\widetilde{Y} = 0$ et $\widetilde{Y} = 1$.

N.B. : comme β_0 est inconnu, on remplace en pratique Z par $F(X'\widehat{\beta})$.

Exemple d'application : acceptation (ou non) d'un prêt par une banque. Soit Y la variable de défaut de l'emprunteur ($Y = 1$ si défaut). Dans l'idéal on accepterait le prêt ssi $Y = 0$ mais Y est inobservé. Ici, C_0 et C_1 représentent respectivement (pour faire simple) le manque à gagner en termes d'intérêt et le capital prêté.

8 Qualité du modèle, sélection des variables explicatives.

Deux situations différentes :

1) le modèle théorique s'écrit $Y = \mathbb{1}\{X\beta_0 + \varepsilon \geq 0\}$ et l'on cherche à estimer l'effet causal de X, β_0. Juger la pertinence du modèle revient (principalement) à tester la nullité des paramètres ;

2) On cherche à estimer $P(Y = 1|X)$ (par exemple pour faire de la prévision). On utilise un logit ou un probit comme approximation commode. On souhaite alors savoir :

   a) si le modèle a un bon pouvoir explicatif (par rapport au modèle sans explicatives) ;

   b) quelles variables explicatives retenir pour prévoir au mieux cette probabilité.

Que l'on soit dans 1) ou 2), on peut également se demander s'il faut faire un logit ou un probit.

1) Pouvoir explicatif du modèle.

– On définit, de façon similaire au R^2, le pseudo-R^2 par :

$$\text{pseudo-}R^2 = 1 - \frac{l_n(\mathbf{Y}|\mathbf{X};\widehat{\beta})}{l_n(\mathbf{Y}|\mathbf{X};\widehat{\beta}_0)}$$

où $\widehat{\beta}_0$ est le paramètre estimé sous l'hypothèse nulle $\beta_{0(-1)} = 0$. Puisque $0 > l_n(\mathbf{Y}|\mathbf{X};\widehat{\beta}) > l_n(\mathbf{Y}|\mathbf{X};\widehat{\beta}_0)$, le pseudo-$R^2$ appartient à $[0,1]$. Il est proche de 1 lorsque

$$Y_i = 1 \text{ et } F(X_i'\widehat{\beta}) \simeq 1 \text{ ou } Y_i = 0 \text{ et } F(X_i'\widehat{\beta}) \simeq 0.$$

N.B. : comme le R^2, le pseudo-R^2 augmente mécaniquement avec le nombre de variables.

– On peut également s'appuyer sur la table de concordance. Si l'on note $\widehat{Y} = \mathbb{1}\{F(X'\widehat{\beta}) > s\}$, on peut calculer la table suivante :

	$\widehat{Y}=0$	$\widehat{Y}=1$	Total
$Y=0$	$n_{Y=0,\widehat{Y}=0}$	$n_{Y=0,\widehat{Y}=1}$	$n_{Y=0}$
$Y=1$	$n_{Y=1,\widehat{Y}=0}$	$n_{Y=1,\widehat{Y}=1}$	$n_{Y=1}$
Total	$n_{\widehat{Y}=0}$	$n_{\widehat{Y}=1}$	n

– On peut calculer, à partir de cette table, le taux de prédiction correct (ou score) :

$$S = \frac{1}{n}\sum_{i=1}^{n} Y_i\widehat{Y}_i + (1-Y_i)(1-\widehat{Y}_i).$$

Inconvénients de cet indicateur :

- si la proportion de $y = 1$ est faible, il pénalisera surtout les erreurs du type $(y = 0, \widehat{y} = 1)$, beaucoup moins les autres ;

- on peut avoir des X ayant un impact significatif et pourtant $S_0 > S$, où S_0 est le score dans un modèle avec la constante seule ! En effet, $\widehat{\beta}$ ne maximise pas le score.

– On peut également (cf. SAS) calculer le pourcentage de paires concordantes et discordantes. Soit (i, j) une paire d'individus telle que $Y_i \neq Y_j$ (disons $Y_i = 1, Y_j = 0$), on la considérera comme :

- concordante si $F(X_i'\widehat{\beta}) > F(X_j'\widehat{\beta})$;

- disconcordante si $F(X_i'\widehat{\beta}) < F(X_j'\widehat{\beta})$;

- nulle lorsque $F(X_i'\widehat{\beta}) = F(X_j'\widehat{\beta})$.

2) Choix des variables

Arbitrage entre :

- l'accroissement du pouvoir explicatif du modèle ;

- la perte de précision liée à l'estimation de nombreux paramètres.

1) on peut faire des tests de nullité des variables, éventuellement via des procédures séquentielles (forward, backward, stepwise...).
Inconvénient : lorsque n tend vers l'infini, on est conduit à accepter systématiquement toutes les variables explicatives.

2) on peut utiliser les critères d'information AIC (*Akaïke Information Criterion*, Akaïke, 1973) ou BIC (*Bayesian Information Criterion*, Schwarz, 1978).

Ces critères sont utilisés pour résoudre le problème du choix de modèles. Supposons que l'on ait K modèles paramétriques possibles :

$$\left\{ \left(P_{\beta^{(1)}}\right)_{\beta^{(1)} \in B^{(1)}}, ..., \left(P_{\beta^{(K)}}\right)_{\beta^{(K)} \in B^{(K)}} \right\}.$$

On souhaite sélectionner le vrai modèle.

- 1ère idée : faire des tests. Problème : du fait de l'asymétrie de l'hypothèse nulle et de l'hypothèse alternative, on peut être conduit à des incohérences suivant l'hypothèse nulle retenue.

- $2^{\text{ème}}$ idée : comparer les log-vraisemblances $l_n(\mathbf{Y}|\mathbf{X}; \widehat{\beta}^{(k)})$ obtenues aux paramètres estimés $\widehat{\beta}^{(k)}$. Problème : dans des modèles emboîtés, la log-vraisemblance augmente mécaniquement avec le nombre de paramètres estimés. De façon générale, les modèles ayant plus de paramètres ont très souvent une plus grande log-vraisemblance. Il faut donc les pénaliser.

Critère d'Akaïke pour le modèle k ayant p_k paramètres :

$$\text{AIC}(k) = l_n(\mathbf{Y}|\mathbf{X}; \widehat{\beta}^{(k)}) - p_k$$

On choisit alors le modèle $k_0 = \arg\max_k \text{AIC}(k)$.

Problème : ce critère n'est pas convergent. Il ne conduit pas au bon choix lorsque n tend vers l'infini. En effet, le critère ne pénalise pas assez le nombre de paramètres. Pour corriger cela, Schwarz (1978) propose le critère suivant :

$$\text{BIC}(k) = l_n(\mathbf{Y}|\mathbf{X}; \widehat{\beta}^{(k)}) - \frac{p_k}{2} \ln(n)$$

Ce critère est convergent (pour des observations i.i.d., lorsque les modèles sont exponentiels).

3) Différence entre le probit et le logit

Le choix d'une distribution logistique ou normale pour les résidus de la régression latente est arbitraire. Cependant, ces choix conduisent en général à des estimateurs différents des paramètres. Quel modèle croire ?

Pour les départager, on peut tester $H_0 \ : F = \Lambda$ contre $H_1 \ : F = \Phi$ (ou l'inverse). Problème de ce test : il est non emboîté (le modèle correspondant à l'hypothèse nulle n'est pas directement un sous-modèle du modèle initial). Pour faire un tel test, il faut supposer que :

$$P(Y = 1|X) = \frac{\Lambda(X'\beta_0)^{\alpha}\Phi(X'\beta_1)^{1-\alpha}}{\Lambda(X'\beta_0)^{\alpha}\Phi(X'\beta_1)^{1-\alpha} + (1 - \Lambda(X'\beta_0))^{\alpha}(1 - \Phi(X'\beta_1))^{1-\alpha}}.$$

Tester le logit contre le probit revient à tester $\alpha = 1$ contre $\alpha = 0$. Pour ce faire, on utilisera un test du score qui a l'avantage :
1) de ne pas nécessiter d'estimer le modèle général (on estime le modèle sous l'hypothèse nulle seulement) ;

2) d'avoir une distribution standard malgré le fait que le test se fasse à la frontière du domaine (puisque $0 \leq \alpha \leq 1$).

Pour plus de détails, cf. Silva (2001).

Cependant, en pratique, il y a très peu de différence entre les paramètres estimés sous les deux modèles. Une règle empirique (cf. Amemiya, 1981) donne $\widehat{\beta}_{logit} = 1.6\widehat{\beta}_{probit}$.

Ceci peut se comprendre à travers les effets marginaux. Sous le modèle logit,

$$\frac{\partial E(Y|X=x)}{\partial x_k} = \Lambda(x'\beta_0)\left(1-\Lambda(x'\beta_0)\right)\beta_{0k,logit}$$

Sous le modèle probit,

$$\frac{\partial E(Y|X=x)}{\partial x_k} = \varphi(x'\beta_0)\beta_{0k,probit}$$

Si $P(Y=1|X=x) \simeq 0.5$, $x'\beta_0 \simeq 0$, et donc

$$\Lambda(x'\beta_0)\left(1-\Lambda(x'\beta_0)\right) \simeq 0.25, \quad \varphi(x'\beta_0) \simeq 0.4$$

On obtient donc

$$\frac{\beta_{0k,logit}}{\beta_{0k,probit}} \simeq \frac{0.40}{0.25} = 1.6$$

Des différences peuvent apparaître lorsque l'une des occurrences de Y est très rare, car les queues de distribution des deux fonctions diffèrent.

9 Le modèle de probabilité linéaire

Parfois, pour des raisons de simplicité, on estime un modèle de probabilité linéaire plutôt qu'un modèle logit ou probit :

$$E(Y|X) = X'\beta_0$$

Exemple : données de panel. Supposons que

$$E(Y_{it}|X_{it}, u_i) = X'_{it}\beta_0 + u_i$$

où u_i est un effet individuel (a priori corrélé aux X_{it}). Dans un tel modèle on peut facilement éliminer l'effet fixe, par différence ou par within :

$$E(Y_{it} - Y_{it-1}|X_{it}, X_{it-1}) = (X_{it} - X_{it-1})'\beta_0$$

Dans les modèles non-linéaires, ce n'est pas aussi simple car

$$E(Y_{it} - Y_{it-1}|X_{it}, X_{it-1}, u_i) = F(X'_{it}\beta + u_i) - F(X'_{it-1}\beta + u_i).$$

Le modèle de probabilité linéaire peut se réécrire :

$$Y = X'\beta_0 + \varepsilon$$

avec

$$\varepsilon = \left| \begin{array}{lll} 1 - X'\beta_0 & \text{avec la probabilité (conditionnelle)} & X'\beta_0 \\ -X'\beta_0 & \text{avec une probabilité} & 1 - X'\beta_0 \end{array} \right.$$

On a donc

$$\begin{aligned} V(\varepsilon|X) &= E(\varepsilon^2|X) \\ &= X'\beta_0(1 - X'\beta_0)^2 + (1 - X'\beta_0)(X'\beta_0)^2 \\ &= X'\beta_0(1 - X'\beta_0) \end{aligned}$$

Le modèle est hétéroscédastique. On peut l'estimer par MCQG :

1) On estime β_0 par MCO : $\widehat{\beta}_{MCO}$.

2) On réestime β_0 par

$$\widehat{\beta}_{MCGQ} = \arg\min_{\beta} \sum_{i=1}^{n} \frac{1}{X_i'\widehat{\beta}_{MCO}(1 - X_i'\widehat{\beta}_{MCO})} \left[Y_i - X_i'\beta\right]^2$$

En pratique les résultats sont souvent proches de ceux du logit ou du probit.

10 Exemple : activité des femmes.

On cherche à expliquer l'activité ($Y = 1$, $Y = 0$ sinon) des femmes suivant leur diplôme, leur situation familiale, leur âge de fin d'étude (ADFE) et leur expérience (EXP). Modalités retenues :

DDIPL
1 Diplôme supérieur
3 Baccalauréat + 2 ans
4 Baccalauréat ou brevet professionnel ou autre diplôme de ce niveau
5 CAP, BEP ou autre diplôme de ce niveau
6 Brevet des collèges
7 Aucun diplôme ou CEP

TYPMEN5
1 Ménages d'une seule personne
2 Familles monoparentales
3 Couples sans enfant
4 Couples avec enfant(s)
5 Ménages complexes de plus d'une personne

```
clear
set mem 180m

use "X:\Cours 2A\Microéconométrie\Cours\2006-2007\emploi.dta", clear

* On garde l'année 2004 et le 4ème trimestre
keep if (annee == "2004" & trim == "4")

* on garde uniquement les femmes
keep if sexe == "2"

* On transforme les variables age et fordat
destring age, replace
destring fordat, replace

keep if age >=15 & age <= 65

if (act != ""){
     gen active = 1 - (act == "3")
}

if (fordat != . & age != .){
     * âge de fin d'études initiales
     gen adfe = fordat - (2004 - age)

     * âge de fin d'étude au carré
     gen adfe2 = adfe^2

     * niveau d'expérience
     gen exp = age - adfe

     * expérience au carré
     gen exp2 = exp^2
}

char typmen5[omit] 4
char ddipl[omit] 7

xi: logit active adfe adfe2 exp exp2 i.ddipl i.typmen5
```

```
i.ddipl           _Iddipl_1-6         (_Iddipl_6 for ddipl==7 omitted)
i.typmen5         _Itypmen5_1-5       (_Itypmen5_4 for typmen5==4 omitted)

Iteration 0:   log likelihood = -2704.1638
Iteration 1:   log likelihood = -2214.5833
Iteration 2:   log likelihood =  -2205.025
Iteration 3:   log likelihood = -2204.9767
Iteration 4:   log likelihood = -2204.9766

Logistic regression                               Number of obs   =       4569
                                                  LR chi2(13)     =     998.37
                                                  Prob > chi2     =     0.0000
Log likelihood = -2204.9766                       Pseudo R2       =     0.1846

------------------------------------------------------------------------------
      active |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        adfe |  -.0814008   .0597962    -1.36   0.173    -.1985991    .0357975
       adfe2 |   .0013144   .0014139     0.93   0.353    -.0014568    .0040856
         exp |   .1283979   .0105132    12.21   0.000     .1077925    .1490033
        exp2 |  -.0040017   .0002343   -17.08   0.000     -.004461   -.0035424
   _Iddipl_1 |   .9174023   .1898909     4.83   0.000      .545223    1.289582
   _Iddipl_2 |   .8556882   .1614762     5.30   0.000     .5392008    1.172176
   _Iddipl_3 |   .3728286   .1326774     2.81   0.005     .1127856    .6328716
   _Iddipl_4 |     .43674   .1100017     3.97   0.000     .2211407    .6523394
   _Iddipl_5 |   .3915286   .1532609     2.55   0.011     .0911428    .6919143
 _Itypmen5_1 |   .6923346    .133966     5.17   0.000     .4297661    .9549031
 _Itypmen5_2 |   .3866097   .1315943     2.94   0.003     .1286897    .6445297
 _Itypmen5_3 |   .5272825   .1043615     5.05   0.000     .3227377    .7318272
 _Itypmen5_5 |   .1350586    .205496     0.66   0.511    -.2677061    .5378233
       _cons |   1.430016   .6477024     2.21   0.027     .1605428     2.69949
------------------------------------------------------------------------------
```

CODE SAS

```
libname donnees "X:\Cours 2A\Microéconométrie\2006-2007\Cours";

data emploi;
     set donnees.emploi (keep=ddipl typmen5 age fordat act sexe trim annee ident);
     where annee="2004" and trim="4"
     and sexe='2'                                  /* on garde uniquement les femmes */
     and typmen5 ne '' and ddipl ne '';
     age_num = input(age,3.);                      /* âge en numérique */
     fordat_num = input(fordat,4.);                /* date fin d'études en numérique */

     if age_num >=15 and age_num<=65 ;

     if act ne '' then active = 1- (act = '3');

     if fordat_num ne . and age_num ne . then do;
          adfe = fordat_num - (2004 - age_num) ; /* âge de fin d'études initiales */
          adfe2 = adfe**2 ;                          /* âge de fin d'étude au carré */
          exp = age_num - adfe ;      /* niveau d'expérience, assimilé ici au nombre
                                         d'années depuis la fin des études */
          exp2 = exp**2;              /* expérience au carré */
     end;
     typmen_1 = (typmen5='1');
     typmen_2 = (typmen5='2');
     typmen_3 = (typmen5='3');
     typmen_5 = (typmen5='5');
run;

proc logistic data=emploi descending; /* descending pour préciser que l'on modélise
                                         Y=1 et non Y=0 */
     class ddipl / param= ref ; /* param = ref permet de fixer une modalité à 0 */
     model  active = ddipl typmen_1--typmen_5 adfe adfe2 exp exp2 /link=logit;
     /* pour faire un probit on indique link=probit */
     test1: test adfe = adfe2 = 0 ;
run;
```

Le Système SAS

The LOGISTIC Procedure

Informations sur le modèle	
Data Set	WORK.EMPLOI
Response Variable	active
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	5372
Number of Observations Used	4569

Profil de réponse		
Valeur ordonnée	**active**	**Fréquence totale**
1	1	3295
2	0	1274

Probability modeled is active=1.

Note: 803 observations were deleted due to missing values for the response or explanatory variables.

Informations sur le niveau de classe						
Classe	**Valeur**	**Variables de création**				
DDIPL	**1**	1	0	0	0	0
	3	0	1	0	0	0
	4	0	0	1	0	0
	5	0	0	0	1	0
	6	0	0	0	0	1
	7	0	0	0	0	0

État de convergence du modèle

Convergence criterion (GCONV=1E-8) satisfied.

Statistiques d'ajustement du modèle

Critère	Coordonnée à l'origine uniquement	Coordonnée à l'origine et covariables
AIC	5410.328	4437.953
SC	5416.755	4527.932
-2 Log L	5408.328	4409.953

Test de l'hypothèse nulle globale : BETA=0

Test	Khi 2	DF	Pr > Khi 2
Likelihood Ratio	998.3743	13	<.0001
Score	1024.5662	13	<.0001
Wald	738.6946	13	<.0001

Analyse des effets Type 3

Effet	DF	Khi 2 de Wald	Pr > Khi 2
DDIPL	5	37.4333	<.0001
typmen_1	1	26.7003	<.0001
typmen_2	1	8.6294	0.0033
typmen_3	1	25.5225	<.0001
typmen_5	1	0.4318	0.5111
adfe	1	1.9028	0.1678
adfe2	1	0.8636	0.3527
exp	1	154.9759	<.0001
exp2	1	291.5948	<.0001

Analyse des estimations de la vraisemblance maximum

Paramètre		DF	Estimation	Erreur std	Khi 2 de Wald	Pr > Khi 2
Intercept		1	1.2174	0.5933	4.2105	0.0402
DDIPL	**1**	1	0.9172	0.1899	23.3334	<.0001
DDIPL	**3**	1	0.8556	0.1615	28.0754	<.0001
DDIPL	**4**	1	0.3728	0.1327	7.8963	0.0050
DDIPL	**5**	1	0.4367	0.1100	15.7628	<.0001
DDIPL	**6**	1	0.3915	0.1533	6.5261	0.0106
typmen_1		1	0.6922	0.1340	26.7003	<.0001
typmen_2		1	0.3866	0.1316	8.6294	0.0033
typmen_3		1	0.5272	0.1044	25.5225	<.0001
typmen_5		1	0.1350	0.2055	0.4318	0.5111
adfe		1	-0.0787	0.0571	1.9028	0.1678
adfe2		1	0.00131	0.00141	0.8636	0.3527
exp		1	0.1364	0.0110	154.9759	<.0001
exp2		1	-0.00400	0.000234	291.5948	<.0001

Estimations des rapports de cotes			
Effet	**Point Estimate**	**95% Limites de confiance de Wald**	
DDIPL 1 vs 7	2.502	1.725	3.631
DDIPL 3 vs 7	2.353	1.714	3.229
DDIPL 4 vs 7	1.452	1.119	1.883
DDIPL 5 vs 7	1.548	1.247	1.920
DDIPL 6 vs 7	1.479	1.095	1.998
typmen_1	1.998	1.537	2.598
typmen_2	1.472	1.137	1.905
typmen_3	1.694	1.381	2.079
typmen_5	1.145	0.765	1.712
adfe	0.924	0.826	1.034
adfe2	1.001	0.999	1.004
exp	1.146	1.122	1.171
exp2	0.996	0.996	0.996

Association des probabilités prédites et des réponses observées			
Percent Concordant	76.3	**Somers' D**	0.530
Percent Discordant	23.3	**Gamma**	0.532
Percent Tied	0.4	**Tau-a**	0.213
Pairs	4197830	**c**	0.765

Résultats des tests des hypothèses linéaires			
Libellé	**Khi 2 de Wald**	**DF**	**Pr > Khi 2**
test1	3.5500	2	0.1695

Comparaison logit / probit.

	Logit		Probit		Ratio
Paramètre	Estimateur	p value	Estimateur	p value	logit / probit
Constante	1.22	0.04	0.59	0.07	2.07
DDIPL 1	0.92	$< 10^{-4}$	0.55	$< 10^{-4}$	1.65
DDIPL 3	0.86	$< 10^{-4}$	0.52	$< 10^{-4}$	1.63
DDIPL 4	0.37	0.01	0.24	0.002	1.57
DDIPL 5	0.44	$< 10^{-4}$	0.26	$< 10^{-4}$	1.66
DDIPL 6	0.39	0.01	0.24	0.01	1.63
typmen 1	0.69	$< 10^{-4}$	0.41	$< 10^{-4}$	1.68
typmen 2	0.39	0.003	0.22	0.003	1.73
typmen 3	0.53	$< 10^{-4}$	0.31	$< 10^{-4}$	1.68
typmen 5	0.14	0.51	0.08	0.50	1.68
adfe	-0.08	0.17	-0.03	0.27	2.35
adfe^2	0.0013	0.35	0.0005	0.52	2.84
exp	0.14	$< 10^{-4}$	0.08	$< 10^{-4}$	1.69
exp^2	-0.0040	$< 10^{-4}$	-0.0024	$< 10^{-4}$	1.69

Questions :

- Accepte-t-on l'hypothèse nulle du modèle sans explicative ?

- Quelles sont les variables significatives à 5% ?

- Calculer l'effet marginal de la variable expérience.

- Comment sont calculés les intervalles de confiance sur les odds-ratios ?

- Estime-t-on ici l'effet causal des variables ?

11 Remise en cause des hypothèses du modèles

On revient sur trois hypothèses du modèle :

– L'homoscédasticité : $V(\varepsilon|X) = \sigma^2$;

– L'exogénéité : $E(X'\varepsilon) = 0$;

– le modèle est paramétrique : la loi de ε est connue *(hors programme)*.

a) Hétéroscédasticité

Les paramètres estimés ne sont pas convergents si ε est hétéroscédastique. On peut cependant estimer le modèle sous une hypothèse paramétrique d'hétéroscédasticité :

$$\varepsilon | X, Z \sim \mathcal{N}\left(0, \exp(2Z'\gamma_0)\right)$$

où, pour des raisons d'identification, Z ne doit pas inclure la constante. On a

$$Y = \mathbb{1}\{X'\beta_0 + \varepsilon \geq 0\} = \mathbb{1}\{\exp(-Z'\gamma_0)X'\beta_0 + \exp(-Z'\gamma_0)\varepsilon \geq 0\}.$$

Donc

$$P(Y = 1|X, Z) = \Phi\left[\exp(-Z'\gamma_0)X'\beta_0\right]$$

La log-vraisemblance s'écrit

$$l_n(\mathbf{Y}|\mathbf{X}; \beta, \gamma) = \sum_{i=1}^{n} Y_i \ln\left\{\Phi\left[\exp(-Z'\gamma)X'\beta\right]\right\} + (1 - Y_i)\ln\left\{1 - \Phi\left[\exp(-Z'\gamma)X'\beta\right]\right\}$$

La log-vraisemblance peut être difficile à maximiser. Cependant, la forme des dérivées permet de mettre simplement en œuvre un test du score de $H_0 : \gamma_0 = 0$ (homoscédasticité) contre $H_1 : \gamma_0 \neq 0$ (hétéroscédasticité).

N.B. : ces modèles peuvent être estimés sous SAS à l'aide de la PROC QLIM.

b) Endogénéité

Supposons que l'on ait :

$$Y = \mathbb{1}\{X_1'\beta_{10} + X_2'\beta_{20} + \varepsilon \geq 0\} \tag{9}$$

avec X_1 indépendant de ε mais $\text{cov}(X_2, \varepsilon) \neq 0$ ($X_2 \in \mathbb{R}$ pour simplifier). Alors les estimateurs précédents ne sont pas convergents.

Supposons maintenant qu'il existe des instruments Z corrélés à X_2 mais pas à ε. Il est (très !) difficile d'appliquer les méthodes habituelles de régression instrumentale sans hypothèse supplémentaire. Supposons que

$$\left|\begin{array}{l} X_2 = X_1'\gamma_{10} + Z'\gamma_{20} + \eta \\ \varepsilon = \eta\delta_0 + \nu \\ (X_1, Z) \perp\!\!\!\perp (\eta, \varepsilon) \\ \nu \perp\!\!\!\perp \eta \text{ et } \nu \text{ est normale.} \end{array}\right.$$

La troisième condition est restrictive (rejetée si (X_1, X_2, Z) est discret par exemple).

La dernière condition est satisfaite par exemple si (ε, η) suit une loi normale bivariée.
(Pourquoi ?)

b) Endogénéité

Sous ces hypothèses, on applique le principe de la régression augmentée en remplaçant ε dans (9) :

$$Y = \mathbb{1}\{X_1'\beta_{10} + X_2'\beta_{20} + \eta\delta_0 + \nu \geq 0\}$$

- Supposons η connu. Les conditions précédentes assurent que ν est indépendante de (X_1, Z, η) (donc de (X_1, X_2, η)). Si $\gamma_{20} \neq 0$ (condition de rang), η n'est pas colinéaire à (X_1, X_2). Enfin, ν est normale. Donc on peut appliquer un probit classique pour estimer les paramètres (β_{10}, β_{20}).

- Ici, η est inconnu. Cependant, on peut l'estimer par le résidu de la régression de X_2 sur (X_1, Z). On procède donc en deux étapes :

i) Régresser X_2 sur (X_1, Z). Soit $\widehat{\eta}$ le résidu estimé de cette régression ;

ii) Probit "augmenté" de Y sur $(X_1, X_2, \widehat{\eta})$.

Cette procédure permet d'obtenir des estimateurs convergents et asymptotiquement normaux. Cependant, la variance asymptotique fournie par le logiciel est fausse car elle ne tient pas compte de la variance de première étape (η est estimé). Un moyen simple d'obtenir un estimateur convergent est de bootstrapper...

b) Endogénéité

Rappel : la méthode du bootstrap est souvent utilisée pour faire de l'inférence à partir d'estimateurs "complexes". On s'intéresse à $\widehat{\theta} = T(X_1, ..., X_n)$ où $(X_1, ..., X_n)$ sont i.i.d. de f.d.r. F. On approche alors la loi de $\widehat{\theta}$ par celle de $\widehat{\theta}^* = T(X_1^*, ..., X_n^*)$, où les X_i^* sont i.i.d. de loi F_n, la f.d.r. empirique.

Ici, pour $b = 1$ à B :

– on tire avec remise des quadruplets $(Y_i, X_{1i}, X_{2i}, Z_i)_{i\in\{1,...,n\}}$ à partir de l'échantillon initial. Soit $S^{(b)}$ l'échantillon ainsi constitué ;

– on régresse X_2 sur (X_1, Z) dans l'échantillon $S^{(b)}$. Soit $\widehat{\eta}^{(b)}$ le résidu estimé de cette régression ;

– on effectue un probit de Y sur $(X_1, X_2, \widehat{\eta}^{(b)})$ sur l'échantillon $S^{(b)}$. Soit $(\widehat{\beta}_1^{(b)}, \widehat{\beta}_2^{(b)}, \widehat{\delta}^{(b)})$ les paramètres estimés correspondant.

On peut alors estimer la variance de $\widehat{\beta}_k$ ($k \in \{1, 2\}$) par

$$\widehat{V}_k = \frac{1}{B} \sum_{b=1}^{B} \left(\widehat{\beta}_k^{(b)} - \widehat{\beta}_k \right)^2$$

Pour plus de détails sur le bootstrap, cf. :

- Davison et Hinkley (1997 ou 2003), Bootstrap Methods and Their Application, Cambridge University Press ;

- Le cours de 3A de méthodes simulées et rééchantillonnage.

c) Modèle semi-paramétrique *(hors programme)*

Le modèle hétéroscédastique précédent permet de lever l'hypothèse d'indépendance entre ε et X mais suppose une forme paramétrique ad hoc. Est-il possible d'estimer β_0 dans le modèle $Y = \mathbb{1}\{X'\beta_0 + \varepsilon \geq 0\}$ sans cette hypothèse ?

$1^{\text{ère}}$ idée : imposer $E(\varepsilon|X) = 0$, comme dans le modèle linéaire. Cependant, on peut montrer (cf. Manski, 1988) que le modèle n'est pas identifié dans ce cas.

$2^{\text{ème}}$ idée : imposer $\text{med}(\varepsilon|X) = 0$. Dans ce cas, le modèle est identifié. Plus précisément, on a

Proposition 3 *(Manski, 1988) :*
1) med$(\varepsilon|X) = 0$.
2) s'il existe une variable (disons X_K*) continue et dont la densité conditionnelle (à* X_{-K}*) est presque partout positive ;*
3) les $(X_k)_{1\leq k\leq K}$ *sont linéairement indépendants.*
Alors $\beta_0/||\beta_0||$ *est identifié.*

Remarque : β_0 n'est identifié qu'a un facteur d'échelle près. Pourquoi ?

c) Modèle semi-paramétrique *(hors programme)*

Preuve : Supposons $||\beta_0|| = 1$ et montrons que

$$\beta_0 = \arg \max_{\beta/||\beta||=1} E\left[(2Y-1)\mathbb{1}\{X'\beta \geq 0\}\right]. \tag{10}$$

Soit $p(x) = P(Y=1|X=x)$, $A = \{x/p(x) \geq 1/2\}$ et, pour tout β, $A_\beta = \{x/x'\beta \geq 0\}$. On a :

$$\begin{aligned} p(x) \geq 1/2 &\iff P(\varepsilon \geq -x'\beta_0 | X=x) \geq 1/2 \\ &\iff P(\varepsilon \geq -x'\beta_0 | X=x) \geq P(\varepsilon \geq 0 | X=x) \\ &\iff -x'\beta_0 \leq 0 \end{aligned}$$

Donc $A = A_{\beta_0}$. De plus, les hyp. 2 et 3 assurent que pour tout $\beta \neq \beta_0$, $P^X\left({}^cA \cap A_\beta\right) > 0$. Ainsi, pour tout $\beta \neq \beta_0$,

$$\begin{aligned} E\left((2Y-1)\mathbb{1}\{X'\beta \geq 0\}\right) &= E\left((2p(X)-1)\mathbb{1}\{X'\beta \geq 0\}\right) \\ &= \int_{A_\beta} (2p(x)-1) dP^X(x) \\ &= \int_{A \cap A_\beta} (2p(x)-1) dP^X(x) + \int_{{}^cA \cap A_\beta} (2p(x)-1) dP^X(x) \\ &< \int_{A} (2p(x)-1) dP^X(x) \\ &= \int_{A_{\beta_0}} (2p(x)-1) dP^X(z) = E\left[(2Y-1)\mathbb{1}\{X'\beta_0 \geq 0\}\right] \ \square \end{aligned}$$

c) Modèle semi-paramétrique *(hors programme)*

On a (exercice)

$$\begin{aligned}\beta_0 &= \arg\max_{\beta} E\left((2Y-1)\mathbb{1}\{X'\beta \geq 0\}\right) \\ &= \arg\max_{\beta} E\left(Y\mathbb{1}\{X'\beta \geq 0\} + (1-Y)\mathbb{1}\{X'\beta < 0\}\right).\end{aligned}$$

Ainsi, une méthode d'estimation naturelle est :

$$\widehat{\beta}_{MS} \in \arg\max_{\beta} \sum_{i=1}^{n} Y_i\mathbb{1}\{X_i'\beta \geq 0\} + (1-Y_i)\mathbb{1}\{X_i'\beta < 0\}.$$

Il s'agit d'un paramètre maximisant le score, d'où le nom d'*estimateur du maximum du score*.

N. B. : ce programme est difficile a résoudre car la fonction est non continue en β (utilisation du simplexe par exemple). Il peut admettre plusieurs solutions.

c) Modèle semi-paramétrique *(hors programme)*

Proposition 4 *(Manski, 1975, Kim et Pollard, 1990) On a*

$$\begin{aligned} \widehat{\beta}_{MS} &\xrightarrow{P} \beta_0 \\ n^{1/3}\left(\widehat{\beta}_{MS} - \beta_0\right) &\xrightarrow{\mathcal{L}} D \end{aligned}$$

cf. Kim et Pollard (1990) pour la définition exacte de la loi D.

Remarques :

i. sous des hypothèses faibles ($\text{med}(\varepsilon|X) = 0$) proches de celles du modèle linéaire, l'estimateur a une vitesse de convergence lente. L'information fournie par le modèle est beaucoup plus faible que dans le modèle linéaire.

ii. En pratique, l'estimateur est bien meilleur que le logit/probit lorsqu'il y a une forte hétéroscédasticité ou si la loi des résidus est très différente d'une logistique/normale (cf. Manski et Thomson, 1986).

iii. L'inférence sur β_0 est difficile car la loi est non standard (voir Delgado et al., 2001).

iv. Il n'existe pas de procédures calculant $\widehat{\beta}_{MS}$ sous SAS ou STATA. Cette méthode est très peu usitée en pratique.

v. Il existe d'autres méthodes semiparamétriques pour estimer β_0, sous des hypothèses alternatives : $X \perp\!\!\!\perp \varepsilon$ mais la loi de ε est inconnue (cf. Klein et Spady, 1993), $E(X'\varepsilon) = 0$ et existence d'un régresseur "spécial" (Lewbel, 2000), etc.