MASTER ECONOMETRIE ET STATISTIQUE APPLIQUEE (ESA)

Université d'Orléans

Econométrie des Variables Qualitatives

Chapitre 1

Modèles Dichotomiques Univariés

Modèles Probit, Logit et Semi-Paramétriques

Christophe Hurlin

Polycopié de Cours

Master Econométrie et Statistique Appliquée (ESA)
Université d'Orléans
Faculté de Droit, d'Economie et de Gestion
Bureau A 224
Rue de Blois – BP 6739
45067 Orléans Cedex 2
www.univ-orleans.fr/deg/masters/ESA/

January 21, 2003

Contents

Introduction

Un des développements majeurs de l'économétrie dans les années 60 et 70, fut sans conteste lié à l'utilisation croissante des *données microéconomiques* relatives à des caractéristiques économiques d'agents individuels (firmes, consommateurs, centres de profits...). A cette époque, les bases de données microéconomiques ont en effet pu être constituées, puis exploitées principalement du fait de l'extension des capacités informatiques et de la réduction de leur coût. Bien souvent, les données statistiques disponibles dans ces bases sont relatives à des **caractères qualitatifs** comme par exemple la catégorie socio-professionnelle, le type d'études suivies, le fait de travailler ou au contraire d'être au chômage, d'acheter ou de ne pas acheter un certain produit etc.. Or, comme nous allons le voir dans ce chapitre, les méthodes d'inférence traditionnelles ne permettent pas de modéliser et d'étudier des caractères quantitatifs : des méthodes spécifiques doivent être utilisées tenant compte par exemple de l'absence de continuité des variables traitées ou de l'absence d'ordre naturel entre les modalités que peut prendre le caractère qualitatif. Ce sont ces méthodes spécifiques les plus usuelles qui seront l'objet de ce cours d'économétrie des variables qualitatives.

Historiquement l'étude des modèles décrivant les modalités prises par une ou plusieurs variables qualitatives date des années 1940-1950. Les travaux les plus marquants de cette époque sont sans conteste ceux de **Berkson (1944, 1951)** consacrés notamment aux **modèles dichotomiques simples (modèles logit et probit)**. Les premières applications ont alors essentiellement été menées dans le domaine de la biologie, de la sociologie et de la psychologie. Ainsi, ce n'est finalement que récemment, que ces modèles ont été utilisés pour décrire des données économiques avec notamment les travaux[1] de **Daniel L. MacFadden (1974)** et de **James J. Heckman (1976)**. Or, l'application des techniques économétriques propres aux variables qualitatives à des problématiques économiques a d'une part largement contribué à améliorer l'interprétation des modèles simples (comme par exemple le modèle logit avec les travaux de MacFadden), et d'autre part à identifier des problèmes économiques dont la structure, si elle n'est pas qualitative au sens propre du terme, en mathématiquement très proche (c'est par exemple le cas de la consommation de bien durable avec le modèle de Tobin de 1958). Ces développements ont ainsi conduit à introduire un modèle intermédiaire entre les modèles qualitatifs et le modèle linéaire habituel : le **modèle tobit**.

Dans la suite du cours, nous supposerons l'existence d'un caractère qualitatif qui peut prendre K modalités disjointes. Si $K = 2$, on dit que **la variable est dichotomique**. Exemple : être au chômage ou ne pas être au chômage. Dans le cas général $K \in \mathbb{N}^*$, on dit que la **variable est polytomique**. A ce niveau de l'exposé, la question qui se pose est de savoir comment représenter un caractère qualitatif dans le cadre d'un modèle économétrique ? Si l'on considère

[1] Il convient ici de rappeler que ces deux économètres ont obtenu conjointement le prix nobel d'économie en 2000, cf. document en annexe.

par exemple le type d'études suivies par un étudiant (université, école d'ingénieur etc..), la catégorie socio-professionnelle (ouvrier, employé, cadre..), ou le fait d'être au chômage, comment doit on représenter ces différents caractères qualitatifs ? **La réponse naturelle à ces questions consiste à associer une variable quantitative (ou codage) au caractère qualitatif**.

Considérons l'exemple de la variable qualitative $y =$ "*niveau d'étude*" pouvant prendre 3 modalités : *"licence", "master", "doctorat"*. Plusieurs choix sont possible pour coder cette variable qualitative. La première consiste tout simplement à associer à y une variable quantitative x pouvant prendre trois valeurs réelles distinctes $(a, b, c) \in \mathbb{R}^3$ suivant les modalités de y. La connaissance de la valeur prise par la variable x permet alors de connaître la modalité de la variable y et inversement. Le choix du triplet de valeurs (a, b, c) est alors à priori non contraint : on peut par exemple prendre $(1, 2, 3)$ ou $(3, 5, 8)$ en référence au nombre d'années d'étude suivies. Ainsi, on définit par exemple la variable x de la façon suivante :

$$x = \begin{cases} 3 & \text{si } y = "licence" \\ 5 & \text{si } y = "master" \\ 8 & \text{si } y = "doctorat" \end{cases}$$

Mais d'autres formes de codage auraient pu être envisagées dans ce cas. On peut par exemple représenter la variable qualitative par le vecteur $z = (z_1, z_2, z_3)$ où les variables z_i, $i = 1, 2, 3$ sont de type dichotomique avec :

$$z_1 = \begin{cases} 1 & \text{si } y = "licence" \\ 0 & \text{sinon} \end{cases}$$

$$z_2 = \begin{cases} 1 & \text{si } y = "master" \\ 0 & \text{sinon} \end{cases}$$

$$z_3 = \begin{cases} 1 & \text{si } y = "doctorat" \\ 0 & \text{sinon} \end{cases}$$

Les variables z_i sont appelées **variables dummy** ou **variables muettes**. Il s'agit ici d'une autre représentation quantitative de y à valeur cette fois dans $(0, 1)^3$. *Ainsi, de façon générale toutes les représentations quantitatives de y s'écrivent sous la forme d'une application injective de* $\{$*"licence","master","doctorat"*$\}$ *dans un espace* $\mathbb{R}^p$, $p \in \mathbb{N}^*$.

L'intérêt principal du codage (ou de la représentation quantitative des variables qualitatives) est de pouvoir se ramener à des lois discrètes sur $\mathbb{R}^p$. Ainsi, si l'on considère l'exemple précédent la loi de z est une loi multinomiale $\mathcal{M}(1; p_1, ., p_i, .., p_K)$ où p_i désigne la probabilité que la $i^{ème}$ modalité de la variable y se réalise. De la même façon, la variable z_1 suit une loi de Bernouilli $\mathcal{B}(1, p_1)$. *Il faut toutefois utiliser avec prudence la loi d'une telle représentation : elle est en effet, par nature, conditionnelle au codage choisi.* Les seules caractéristiques véritablement liées à la variable qualitative sont celles qui ne dépendent pas de la représentation choisie, et ne sont autres que les probabilités p_1,...,p_K. *Ainsi, les moments (moyenne, variance etc..) de la variable codée ont en général peu de sens.* Dans l'exemple précédent, l'espérance de la variable codée x n'a pas de signification particulière. En revanche, l'espérance des variables dummies z_i permet de retrouver les probabilités p_i. *De plus, le calcul d'un coefficient de corrélation entre deux variables codées x et z dépend naturellement des codages retenus, et ne peut donc être interprété économiquement.* En revanche, la notion d'indépendance entre deux variables codée reste indépendante du codage retenu.

Dans le cadre de ce premier chapitre, nous allons nous intéresser au modèle le plus simple, à savoir **le modèle dichotomique**, dans lequel la variable expliquée du modèle ne peut prendre que deux modalités. Le plan de ce chapitre est le suivant. Nous commencerons par présenter les principaux modèles dichotomiques, et en particulier les modèles logit et probit. Puis, dans une seconde section, nous intéresserons au problème de l'estimation des paramètres de ces modèles, notamment par la méthode du maximum de vraisemblance. Dans une troisième partie, nous étudierons la convergence des estimateurs du maximum de vraisemblance. Enfin, dans une dernière section nous aborderons les tests de spécification de ces modèles ainsi que les différents problèmes d'inférence.

1. Modèles Dichotomiques Univariés

Par modèle dichotomique, on entend un modèle statistique dans lequel la variable expliquée ne peut prendre que deux modalités (variable dichotomique). Il s'agit alors généralement d'expliquer la survenue ou la non survenue d'un événement.

Hypothèse *On considère un échantillon de N individus indicés $i = 1, .., N$. Pour chaque individu, on observe si un certain évenément s'est réalisé et l'on note y_i la variable codée associée à évenement. On pose, $\forall i \in [1, N]$:*

$$y_i = \begin{cases} 1 & \textit{si l'événement s'est réalisé pour l'individu } i \\ 0 & \textit{si l'événement ne s'est pas réalisé pour l'individu } i \end{cases} \tag{1.1}$$

On remarque ici le choix du codage $(0, 1)$ qui est traditionnellement retenu pour les modèles dichotomique. En effet, celui-ci permet définir la probabilité de survenue de l'événement comme l'espérance de la variable codée y_i, puisque :

$$E(y_i) = Prob(y_i = 1) \times 1 + Prob(y_i = 0) \times 0 = Prob(y_i = 1) = p_i$$

L'objectif des modèles dichotomiques consiste alors à expliquer la survenue de l'événement considéré en fonction d'un certain nombre de caractéristiques observées pour les individus de l'échantillon. Comme nous le verrons par la suite, on cherche dans ces modèles, à spécifier la probabilité d'apparition de cet événement.

Quels sont alors les principaux champs d'application des modèle dichotomiques ? Nous pouvons ici évoquer quelques pistes, sur lesquelles nous reviendrons par la suite. Un des domaines d'application traditionnel consiste en **l'étude des choix d'éducation**. Ainsi, parmi les premiers travaux utilisant les modèles à réponses qualitatives, plusieurs s'intéressaient aux comportements des étudiants que ce soit en terme de choix de filières, ou en termes de choix d'établissements. Il s'agissait alors de modéliser ces comportements en fonction d'un certain nombres de caractéristiques propres aux universités (présence de campus, débouchés professionnels etc..) ou aux étudiants (CSP des parents, études antérieures etc..). Typiquement, il s'agit par exemple, de modéliser le choix des étudiants entre une université en ville ou un campus, ce choix étant représenté par une variable dichotomique que l'on va cherche à modéliser en fonction de plusieurs facteurs comme le revenu, le sexe de l'étudiant, la distance domicile-université etc.. Du fait de l'organisation privée des études aux Etats-Unis, de telles modélisations ont connu un grand intérêt, que ce soit dans une perspective purement académique ou dans une perspective appliquée. On peut citer ici par exemple l'étude de Radner et Miller (1970).

Un autre domaine d'application consiste en *la modélisation des risques de défaillance dans une relation de prêt, ou dans tout autre forme de contrat d'engagement* (contrat d'abonnement téléphonique, contrat d'assistance etc...). On considère par exemple une variable dichotomique prenant deux modalités : "rupture du contrat" et "poursuite du contrat", et l'on cherche à expliquer variables par différents facteurs socio-économiques. Il s'agit ici des techniques de

bases des méthodes de scoring largement utilisées dans le secteur bancaire et dans le secteur des télécommunications.

Cette liste d'application n'est bien entendu pas exhaustive. Nous allons à présent montrer que la modélisation des variables dichotomiques ne peut se faire à l'aide d'une spécification linéaire standard.

1.1. Spécification linéaire des variables endogènes dichotomiques

En effet, la question que l'on peut naturellement se poser à ce stade de l'exposé, est de savoir en quoi les modèles dichotomiques, et plus généralement les modèles à variables endogènes qualitatives, se distinguent du modèle linéaire classique étudié en cours de licence. En effet, il s'agit de comprendre pourquoi l'utilisation de méthodes d'estimation particulières s'avère indispensable pour ce type de modèles. Pour ce faire, appliquons naïvement une modélisation linéaire simple au cas d'une variable endogène dichotomique.

Supposons que l'on dispose de N observations y_i, $\forall i = 1,..,N$ d'une variable endogène dichotomique codée $y_i = 1$ ou $y_i = 0$ par convention, lorsque parallèlement les observations de K variables exogènes sont $x_i = \left(x_i^1..x_i^K\right)$, $\forall i = 1,..,N$. Dans ce cas, le modèle linéaire simple s'écrit :

$$\underset{(1,1)}{y_i} = \underset{(1,K)}{x_i}\ \underset{(K,1)}{\beta} + \underset{(1,1)}{\varepsilon_i} \qquad \forall i = 1,..,N \tag{1.2}$$

où $\beta = (\beta_1...\beta_K)' \in \mathbb{R}^K$ désigne un vecteur de K paramètres inconnus et où les perturbations ε_i sont supposées être indépendamment distribuées. On peut alors mettre en évidence plusieurs problèmes liés à l'utilisation de cette spécification linéaire simple pour modéliser notre variable dichotomique.

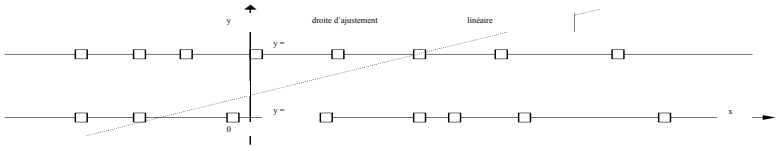
Premièrement, les termes de gauche et de droite de l'équation (1.1) sont de nature différentes. La variable y_i est de type qualitative tandis que la somme $x_i\beta+\varepsilon_i$ est une variable quantitative. On peut répondre à ceci que le membre de gauche correspond en fait au codage (ici 0 ou 1) associé à la variable qualitative; dès lors, il n'y aurait plus de problème. Mais il est évident que ce codage est lui même par nature arbitraire, et que les valeurs de β obtenues pour ce codage sont nécessairement différentes de celles obtenues pour tout autre codage. Elles seraient par exemple de $\alpha\beta$ si le codage était de type $(0, \alpha)$. **Ainsi, le premier problème de l'application du modèle linéaire simple à une variable dichotomique, est que le paramètre β du modèle (1.1) n'est pas interprétable.**

Deuxièmement, une étude graphique montre que l'approximation linéaire est peu adaptée au problème posé. Considérons pour cela le modèle linéaire avec une seule variable explicative ($K = 1$), notée x_i^1, et une constante. On pose $\beta = (\beta_0\ \beta_1)'$ et l'on considère le modèle linéaire suivant :

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i \qquad \forall i = 1,..,N \tag{1.3}$$

Pour constater l'inadéquation de ce modèle à reproduire correctement la variable endogène dichotomique y_i, il suffit de se placer dans un repère $\left(x^1, y\right)$ et de reproduire les N différents couples $\left(x_i^1, y_i\right)$, $\forall i = 1,..,N$. Naturellement, du fait du statut dichotomique de la variable endogène, le nuage de points ainsi obtenu se situe soit sur la droite $y = 0$, soit sur la parallèle $y = 1$. Ainsi, comme on l'observe sur la figure (**??**), **il est impossible d'ajuster de**

Figure 1.1: Ajustement Linéaire d'une Variable Endogène Dichotomique



façon satisfaisante, par une seule droite, le nuage de points, associé à une variable dichotomique qui, par nature, est réparti sur deux droites parallèles.

Troisièmement, la spécification linéaire standard ne convient pas aux variables dichotomiques, et plus généralement aux variables qualitatives, car elle pose un certain nombre de problèmes mathématiques.

1. Sachant que dans la cas d'une variable endogène y_i dichotomique, celle-ci ne peut prendre que les valeurs 0 ou 1, la spécification linéaire (1.1) implique que la perturbation ε_i ne peut prendre, elle aussi, que 2 valeurs, conditionnellement au vecteur x_i :

$$\varepsilon_i = 1 - x_i\beta \text{ avec une probabilité de } p_i = Prob\left(y_i = 1\right)$$

$$\varepsilon_i = -x_i\beta \text{ avec une probabilité de } 1 - p_i$$

Ainsi, la perturbation ε_i du modèle (1.1) admet nécessairement une loi discrète, ce qui exclut en particulier l'hypothèse de normalité des résidus.

2. Lorsque l'on suppose que les résidus ε_i sont de moyenne nulle, la probabilité p_i associée à l'événement $y_i = 1$ est alors déterminée de façon unique. En effet, écrivons l'espérance des résidus :

$$E\left(\varepsilon_i\right) = p_i\left(1 - x_i\beta\right) - \left(1 - p_i\right)x_i\beta = p_i - x_i\beta = 0$$

On en déduit immédiatement que :

$$p_i = x_i\beta = Prob\left(y_i = 1\right) \tag{1.4}$$

Ainsi la quantité $x_i\beta$ correspond à une probabilité et doit par conséquent satisfaire un certain nombre de propriétés et en particulier appartenir à l'intervalle fermé $[0, 1]$.

$$0 \leq x_i\beta \leq 1 \quad \forall i = 1, .., N \tag{1.5}$$

Or rien n'assure que de telles conditions soient satisfaites par l'estimateur des Moindres Carrés utilisé dans le modèle linéaire (1.1). Si de tels contraintes ne sont pas assurées, le modèle

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i \quad E\left(\varepsilon_i\right) = 0 \quad \forall i = 1, .., N$$

n'a pas de sens.

3. Enfin, même si l'on parvenait à assurer le fait que les contraintes (1.5) soient satisfaites par l'estimateur des Moindres Carrés des paramètres du modèle linéaire, il n'en demeurerait pas moins une difficulté liée à **la présence d'hétéroscedasticité**. En effet, on constate immédiatement que, dans le modèle (1.1), la matrice de variance covariance des résidus varie entre les individus en fonction de leur caractéristiques associées aux exogènes x_i puisque :

$$V(\varepsilon_i) = x_i\beta(1 - x_i\beta) \quad \forall i = 1, .., N \tag{1.6}$$

Pour démontrer ce résultat il suffit de considérer la loi discrète des résidus et de calculer la variance :

$$\begin{aligned} V(\varepsilon_i) &= E\left(\varepsilon_i^2\right) = (1 - x_i\beta)^2 \, Prob(y_i = 1) + (-x_i\beta)^2 \, Prob(y_i = 0) \\ &= (1 - x_i\beta)^2 p_i + (-x_i\beta)^2 (1 - p_i) \end{aligned}$$

Sachant que d'après la relation (1.4) on a $p_i = x_i\beta$, on en déduit que :

$$\begin{aligned} V(\varepsilon_i) &= (1 - x_i\beta)^2 x_i\beta + (-x_i\beta)^2 (1 - x_i\beta) \\ &= (1 - x_i\beta) x_i\beta [(1 - x_i\beta) + x_i\beta] \\ &= (1 - x_i\beta) x_i \end{aligned}$$

Or, de plus ce problème d'hétéroscédascticité ne peut pas être résolu par une méthode d'estimation des Moindres Carrés Généralisés tenant compte de la contrainte d'inégalité (1.5), puisque la matrice de variance covariance des perturbations (1.6) dépend du vecteur β des paramètres à estimer dans la spécification linéaire, qui est par nature supposé inconnu.

Pour toutes ces différentes raisons, la spécification linéaire des variables endogènes qualitatives, et plus spécialement dichotomiques, n'est jamais utilisée et l'on recourt à des modèles logit ou probit, que nous allons à présent étudier, pour représenter ces variables.

1.2. Modèles Logit et Probit

Les modèles dichotomiques probit et logit admettent pour variable expliquée, non pas un codage quantitatif associé à la réalisation d'un évenement (comme dans le cas de la spécification linéaire), mais la probabilité d'apparition de cet évenement, conditionnellement aux variables exogènes. Ainsi, on considère le modèle suivant :

$$p_i = Prob(y_i = 1 | x_i) = F(x_i\beta) \tag{1.7}$$

où la fonction $F(.)$ désigne une fonction de répartition. La choix de la fonction de répartition $F(.)$ est a priori non contraint. Toutefois, on utilise généralement deux types de fonction : la fonction de répartition de la loi logistique et la fonction de répartition de la loi normale centrée réduite. A chacune de ces fonctions correspond un nom attribué au modèle ainsi obtenu : modèle logit et modèle probit[2].

Definition 1.1. ***On considère le modèle dichotomique suivant :***

$$p_i = Prob(y_i = 1 | x_i) = F(x_i\beta) \quad \forall i = 1, .., N \tag{1.8}$$

[2] Qui selon toute logique aurait du être nommé modèle *nomit* et non modèle *probit*.

Dans le cas du modèle logit, la fonction de répartition $F(.)$ ***correspond à la fonction logistique*** $\forall w \in \mathbb{R}$ ***:***

$$F(w) = \frac{e^w}{1+e^w} = \frac{1}{1+e^{-w}} = \Lambda(w) \tag{1.9}$$

Dans le cas du modèle probit, la fonction de répartition $F(.)$ ***correspond à la fonction de répartition de la loi normale centrée réduite*** $\forall w \in \mathbb{R}$ ***:***

$$F(w) = \int_{-\infty}^{w} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \Phi(w) \tag{1.10}$$

Ainsi, pour une valeur donnée du vecteur des exogènes et du vecteur des paramètres β, on peut définir les deux modèles d'une façon équivalente :

Definition 1.2. ***Le modèle logit définit la probabilité***[3] ***associé à l'événement*** $y_i = 1$, ***comme la valeur de la fonction de répartition de la loi logistique considérée au point*** $x_i\beta$ *:*

$$\textbf{\textit{Modèle logit}} : \; p_i = \Lambda(x_i\beta) = \frac{1}{1+e^{-x_i\beta}} \quad \forall i = 1, .., N \tag{1.11}$$

Dans le cas du modèle probit, cette probabilité est définie comme la valeur de la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0,1)$ ***considérée au point*** $x_i\beta$ *:*

$$\textbf{\textit{Modèle probit}} : \; p_i = \Phi(x_i\beta) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \quad \forall i = 1, .., N \tag{1.12}$$

A ce stade de l'exposé, la question que l'on se pose immédiatement est de savoir *quelles sont les différences fondamentales entre les modèles probit et logit ?* Quand doit on utiliser l'un plutôt que l'autre ? Quelles sont les propriétés particulières de ces deux modèles ? Bien entendu, ces deux modèles ne diffèrent que par la forme de la fonction de répartition $F(.)$. Ainsi, il faut donc se rappeler quelles sont les propriétés respectives des lois logistiques et normales, pour comprendre quelles peuvent être les différences et les similitudes entre les modèle logit et probit.

1.3. Comparaison des modèles probit et logit

Historiquement, les modèles logit ont été introduits comme des approximations de modèles probit permettant des calculs plus simples. *Dès lors, il n'existe que peu de différences entre ces deux modèles dichotomiques. Ceci s'explique par la proximité des familles de lois logistiques et normales.* Les deux fonctions de répartition $\Lambda(w)$ et $\Phi(w)$ sont en effet sensiblement proches, comme on peut le constater à partir du tableau (1.1) où sont reportées les valeurs de ces fonctions pour différentes valeurs de w. Mais cette similitude est encore grande si l'on considère une *loi logistique transformée* de sorte à ce que la variance soit identique à celle de la loi normale réduite. En effet, nous avons vu que la loi logistique usuelle admet pour fonction de répartition

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

[3] La variable y_i étant dichotomique, la probabilité d'apparition de l'événement complémentaire $y_i = 0$ est définie par $1 - p_i$ avec :

$$1 - p_i = \frac{e^{-x_i\beta}}{1+e^{-x_i\beta}}$$

Cette loi a une espérance nulle et une variance égale à $\pi^2/3$. C'est pourquoi, il convient de normaliser la loi logistique de sorte à obtenir une distribution de variance unitaire, comparable à celle de la loi normale réduite. On définit pour cela une loi logistique transformée.

Definition 1.3. ***La loi logistique transformée de paramètre*** λ ***admet pour fonction de répartition***[4]***, notée*** $\Lambda_\lambda(w)$, $\forall w \in \mathbb{R}$

$$\Lambda_\lambda(w) = \frac{e^{\lambda w}}{1+e^{\lambda w}} = \frac{1}{1+e^{-\lambda w}} \tag{1.13}$$

A cette fonction de répartition correspond une variance de $\pi^2/\left(3\lambda^2\right)$. Ainsi, il convient de comparer la loi normale centrée réduite à la loi logistique transformée, de paramètre $\lambda = \pi/\sqrt{3}$, dont la fonction de répartition est définie comme suit :

$$\widetilde{\Lambda}(w) = \Lambda_{\pi/\sqrt{3}}(w) = \frac{1}{1+e^{-\frac{\pi w}{\sqrt{3}}}} \tag{1.14}$$

Cette loi admet par construction une variance unitaire. On observe ainsi à partir du tableau (1.1), que les réalisations de cette fonction $\Lambda_{\pi/\sqrt{3}}(.)$ sont très proches de celles de la fonction $\Phi(.)$ associée à la loi normale réduite et ce notamment pour des valeurs de w proche de 0, c'est à dire des valeurs dites centrales, car proches de la moyenne de la distribution.

Certains auteurs proposent d'utiliser d'autres paramètres λ afin de mieux reproduire encore la fonction de répartition de la loi normale pour des valeurs centrales. En particulier Amemiya (1981) propose d'utiliser un paramètre[5] $\lambda = 1.6$ et donc de retenir la loi logistique transformée $\Lambda_{1.6}(.)$. Comme on peut l'observer sur le tableau (1.1), la fonction de paramètre 1.6 est encore plus proche de $\Phi(.)$ que la fonction de paramètre $\pi/\sqrt{3}$. pour les valeurs centrales proches de 0 ($w < 1$ en l'occurrence dans le tableau).

Tableau 1.1: Comparaison des Fonctions de Répartition $\Lambda_\lambda(w)$ et $\Phi(w)$

w	0	0.1	0.2	0.3	0.4	0.5	1	2	3
$\Phi(w)$	0.5	0.5398	0.5793	0.6179	0.6554	0.6915	0.8413	0.9772	0.9987
$\Lambda(w)$	0.5	0.5250	0.5498	0.5744	0.5987	0.6225	0.7311	0.8808	0.9526
$\Lambda_{\pi/\sqrt{3}}(w)$	0.5	0.5452	0.5897	0.6328	0.6738	0.7124	0.8598	0.9741	0.9957
$\Lambda_{1.6}(w)$	0.5	0.5399	0.5793	0.6177	0.6548	0.6900	0.8320	0.9608	0.9918

Sources : Anemiya (1981), table 1, page 1487 et calculs de l'auteur.

Quoiqu'il en soit, il apparaît ainsi que les fonctions de répartition des lois normales centrées réduites et des lois logistiques simples ou transformées sont extrêmement proches. Par conséquent, *les modèles probit et logit donnent généralement des résultats relativement similaires.* De nombreuses études ont d'ailleurs été consacrées à ce sujet comme par exemple celle de Morimune (1979)[6] ou de Davidson et MacKinnon (1984). Ainsi a priori, la question du choix entre les deux modèle ne présente que peu d'importance. Toutefois, il convient d'être prudent quand à la comparaison directe des deux modèles.

[4] Par convention, la fonction de répartition de la loi logistique simple correspondant au cas $\lambda = 1$ sera noté $\Lambda(.)$ afin d'alléger les notations.

[5] Cette valeur 1.6 est dérivée du rapport des fonctions de densité $\phi(w)/\lambda(w)$ évalué au point $w = 0$.

[6] Morimune K. (1979), "Comparisons of Normal and Logistic Models in the Bivariate Dichitomous Analysis", Econometrica 47, 957-975.

En effet, les valeurs estimées des paramètres dans les modèles probit et logit ne sont pas directement comparables puisque les variances des lois logistiques et normale réduite ne sont pas identiques. Cette différence de variance implique que la normalisation des coefficients β n'est pas identique et que par conséquent les estimateurs de ces paramètres obtenus dans les deux modèles ne fournissent pas des réalisations identiques.

Proposition 1.4. ***Supposons que l'on note respectivement*** $\widehat{\beta}_P$ ***et*** $\widehat{\beta}_L$ ***les estimateurs des paramètres*** β ***obtenus dans les modèles probit et logit. Amemiya (1981) propose en première approximation d'utiliser la relation suivante entres les estimations probit et logit***[7] **:**

$$\widehat{\beta}_L \simeq 1.6\widehat{\beta}_P \tag{1.15}$$

Toutefois, si ces approximations sont relativement précises sur certains échantillons comportant peu de valeurs "extrêmes" (c'est à dire lorsque la moyenne des valeurs $x_i\beta$ est proche de zéro), elles seront moins précises en présence de nombreuses valeurs $x_i\beta$ éloignées de zéro. Une façon équivalente[8] de vérifier l'adéquation de cette approximation consiste à observer si la valeur moyenne des probabilités p_i est proche de 0.5 (Davidson et MacKinnon 1984). Si tel est le cas, les estimateurs des coefficients du modèle logit seront environ 1.6 fois supérieurs à ceux du modèle probit.

Considérons l'exemple des données de l'article de Spector et Mazzeo (1980), paru dans *Journal of Economic Education*, et intitulé "*Probit Analysis and Economic Education*". Il s'agit ici d'évaluer la probabilité pour un étudiant d'obtenir le passage en post-graduate (variable dichotomique *graduate*), l'équivalent du master. Cette probabilité est modélisée comme une fonction d'une constante (*cons*), du score obtenu au *tuce* (*test of understanding of college economics*) et de la moyenne obtenue au niveau du graduate (*grad*). Sur la figure (1.2) sont reportés les résultats d'estimation du modèle logit tandis que sur la figure (1.3) sont reportés les résultats d'estimation du même modèle probit. Considérons par exemple le coefficient de la variable *tuce*. Le modèle logit nous donne une estimation de 0.0855 pour ce paramètre alors que le modèle probit donne une estimation de 0.05266. On vérifie alors que, pour cet échantillon, les approximations (1.15) sont satisfaisantes puisque selon cette formule, on devrait obtenir une estimation logit de paramètre de l'ordre de $0.05266 * 1.6 = 0.0843$ ou 0.0955 si l'on considère l'approximation $0.05266 * \pi/\sqrt{3}$. Ces approximations sont en effet très proches de la vraie estimation du paramètre dans le modèle logit.

De la même façon, Amemiya (1981) propose différentes approximations permettant d'approcher les estimations des modèles logit et probit à partir des estimations obtenues dans le modèle linéaire simple, présenté précédemment.

Proposition 1.5. ***On note*** $\widehat{\beta}_P$ ***l'estimateur obtenu dans le modèle probit,*** $\widehat{\beta}_L$ ***l'estimateur obtenu dans le modèle logit et*** $\widehat{\beta}_{LP}$ ***l'estimateur obtenu dans le modèle linéaire. Amemiya (1981) propose les approximations suivantes pour les modèles***

[7] En utilisant la normalisation de la variance, on peut aussi retenir comme approximation un facteur $\pi/\sqrt{3} \simeq 1.81$, en posant $\widehat{\beta}_L \simeq \pi\widehat{\beta}_P/\sqrt{3}$.

[8] Sachant que $\Phi(0) = \Lambda(0) = 0.5$, il équivalent de vérifier si la moyenne des valeurs $x_i\beta$ est proche de 0 ou si la moyenne des probabilités $p_i = F(x_i\beta)$ est proche de 0.5, avec $F(x) = \Lambda(x)$ dans le cas du modèle logit et $F(x) = \Phi(x)$ dans le cas du probit.

Figure 1.2: Estimation d'un Modèle Logit

Dependent Variable: GRADE
Method: ML - Binary Logit
Date: 09/06/02 Time: 18:40
Sample: 1 32
Included observations: 32
Convergence achieved after 4 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-10.65600	4.057117	-2.626497	0.0086
TUCE	0.085551	0.133185	0.642352	0.5206
GPA	2.538281	1.181851	2.147716	0.0317

Mean dependent var	0.343750	S.D. dependent var	0.482559
S.E. of regression	0.419006	Akaike info criterion	1.186968
Sum squared resid	5.091415	Schwarz criterion	1.324380
Log likelihood	-15.99148	Hannan-Quinn criter.	1.232516
Restr. log likelihood	-20.59173	Avg. log likelihood	-0.499734
LR statistic (2 df)	9.200493	McFadden R-squared	0.223403
Probability(LR stat)	0.010049		

Obs with Dep=0	21	Total obs	32
Obs with Dep=1	11		

probit et linéaire :

$$\widehat{\beta}_{LP} \simeq 0.4\widehat{\beta}_P \quad \textbf{\textit{pour tous les paramètres à l'exception de la constante}} \tag{1.16}$$

$$\widehat{\beta}_{LP} \simeq 0.4\widehat{\beta}_P + 0.5 \quad \textbf{\textit{pour la constante}} \tag{1.17}$$

et les approximations suivantes pour les modèles logit et linéaire :

$$\widehat{\beta}_{LP} \simeq 0.25\widehat{\beta}_L \quad \textbf{\textit{pour tous les paramètres à l'exception de la constante}} \tag{1.18}$$

$$\widehat{\beta}_{LP} \simeq 0.25\widehat{\beta}_L + 0.5 \quad \textbf{\textit{pour la constante}} \tag{1.19}$$

Ainsi si l'on considère l'exemple des données de l'article de Spector et Mazzeo (1980), les estimations de la constante et des paramètres des variables *tuce* et *grad* obtenues dans le modèle linéaires sont respectivement égales à -1.4493, 0.0160 et 0.4619. Or, si l'on compare ces résultats à ceux obtenus à partir des modèles logit et probit (figures 1.2 et 1.3), on obtient les résultats relativement proches. Ainsi, dans le cas du modèle logit pour la variable *tuce* l'approximation donnerait $0.25 * 0.08555 = 0.0214$ et $0.25 * 2.53828 = 0.6346$ pour la variable *grad*. Pour la constante l'approximation donne une valeur approchée égale à $-0.25 * 10.656 + 0.5 = -2.164$. Cers approximations seront d'autant plus proches des valeurs estimées qu'il y a aura un grand nombre d'observations $x_i\beta$ proches de 0, car en effet les fonctions de répartition des lois logistiques et normales ne se démarquent pas d'une droite dans cette zone.

Figure 1.3: Estimation d'un Modèle Probit

Dependent Variable: GRADE
Method: ML - Binary Probit
Date: 09/06/02 Time: 18:45
Sample: 1 32
Included observations: 32
Convergence achieved after 4 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
C	-6.034326	2.121031	-2.844997	0.0044
TUCE	0.052667	0.075553	0.697094	0.4857
GPA	1.409575	0.635467	2.218172	0.0265

Mean dependent var	0.343750	S.D. dependent var	0.482559
S.E. of regression	0.420296	Akaike info criterion	1.197010
Sum squared resid	5.122808	Schwarz criterion	1.334423
Log likelihood	-16.15216	Hannan-Quinn criter.	1.242558
Restr. log likelihood	-20.59173	Avg. log likelihood	-0.504755
LR statistic (2 df)	8.879145	McFadden R-squared	0.215600
Probability(LR stat)	0.011801		

Obs with Dep=0	21	Total obs	32
Obs with Dep=1	11		

En conclusion, il apparaît que les résultats des modèles probit et logit sont généralement similaires que ce soit en termes de probabilité ou en termes d'estimation des coefficients β si l'on tient compte des problèmes de normalisation. C'est le sens de cette conclusion d'Amemiya.

> "*Because of the close similarity of the two distributions, it is difficult to distinguish between them statistically unless one has an extremely large number of observations. Thus, in the univariate dichotomous model, it does not matter much whether one uses a probit model or a logit model, except in cases where data are heavily concentrated in the tails due to the characteristics of the problem being studied.*", Amemiya T. (1981), page 1487.

Toutefois, comme le note Amemiya (1981), il convient d'être prudent dans l'utilisation des approximations pour comparer les modèles probit et logit. Il est toujours préférable de raisonner en termes de probabilités $p_i = F(x_i\beta)$ et non en termes d'estimation des paramètres β pour comparer ces résultats.

> "*The reader should keep in mind that this equality [equation (1.15)] constitutes only a rough approximation and that a different set of formulae may work better over a different domain. When one wants to compare models with different probability functions, it is generally better to compare probabilities directly rather than comparing the estimates of the coefficients even after an appropriate conversion*", Amemiya T. (1981), page 1488.

Si les deux modèle sont sensiblement identiques, il existe cependant certaines différences entre les modèles probit et logit, comme le souligne d'ailleurs Amemiya. Nous évoquerons ici deux principales différences :

1. **La loi logistique tend à attribuer aux événements "extrêmes" une probabilité plus forte que la distribution normale.**
2. **Le modèle logit facilite l'interprétation des paramètres β associées au variables explicatives x_i**

Nous allons à présent étudier successivement ces deux propriétés. Premièrement, la fonction de densité associée à la loi logistique possède en effet des queues de distribution plus épaisses que celles de la fonction de densité de la loi normale (distribution à queues "plates"). La loi logistique présente donc un excès de Kurtosis[9] : il s'agit d'une *distribution leptokurtique*. En d'autres termes, nous avons vu que les lois logistique et normale appartiennent à la même famille des lois exponentielles et sont par nature très proches, notamment pour les valeurs proches de la moyenne de la distribution. Toutefois, le profil de ces deux distributions diffère aux extrémités du support : pour la loi normale, les valeurs extrêmes sont moins pondérées, la fonction de répartition tendant plus vite vers 0 à gauche du support et vers 1 à droite.

Economiquement, **cela implique que le choix d'une fonction logistique (modèle logit) suppose une plus grande probabilité[10] attribuée aux évenements "extrêmes", comparativement au choix d'une loi normale (modèle probit)**, que ce soit à droite ou gauche de la moyenne de la distribution, les lois normales et logistiques étant symétriques. Pour visualiser ce phénomène, il convient de comparer la fonction de répartition associée à la loi normale centrée réduite avec la fonction de répartition associée à la loi logistique possédant les deux premiers moments identiques à la loi $\mathcal{N}(0,1)$.

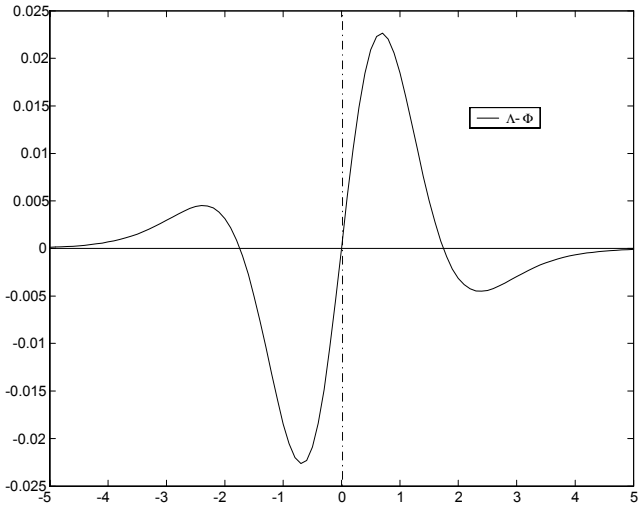
Sur le graphique (1.4) est reportée la différence $\widetilde{\Lambda}(w) - \Phi(w)$ en fonction de w :

On constate qu'à droite du support, pour des valeurs élevées de w ($w > 1.5$ environ), on a $\Phi(w) > \widetilde{\Lambda}(w)$. La fonction de répartition de la loi normale est au dessus de celle de la loi logistique. Etant donnée la définition de la fonction de répartition, $F(w) = Prob(W \leq w)$, cela signifie que la probabilité que la réalisation de la variable W soit inférieure au seuil w est plus grande dans le cas de la loi normale que dans le cas de la loi logistique. Inversement, pour un seuil w donnée, la probabilité d'obtenir des valeurs supérieures à ce seuil (des valeurs "extrêmes") est plus grande dans le cas de la loi logistique que dans le cas de la loi normale. On vérifie ainsi la propriété de la loi logistique qui sur-pondère les valeurs extrêmes en comparaison de la loi normale. Naturellement, puisque les distributions sont symétriques, on obtient le même résultat à gauche du support pour des valeurs très faibles de w ($w < -1.5$ environ).

[9] L'excès de Kurtosis est défini en référence au moment d'ordre d'une loi normale centrée réduite. Si X suit une loi normale $N(\mu, \sigma^2)$, la Kurtosis est égale à $\mu_4 = 3\sigma^4$. Par convention, le degré d'excès de Kurtosis, défini par $\mu_4/\sigma^4 - 3$, est nul.

[10] Bien entendu, la différence entre les résultats des modèles probit et logit ne pourra être observée que si l'on dispose de suffisament d'observations des exogènes se situant dans ces zones "extrêmes".

Figure 1.4: Différence des Fonctions de Répartition $\widetilde{\Lambda}(w) - \Phi(w)$



Deuxièmement, il existe une propriété particulièrement intéressante propre au modèle logit, qui facilite en particulier l'interprétation des paramètres β associées au variables explicatives x_i. Attention, comme nous le verrons par la suite, les valeurs numérique des estimations n'ont pas d'interprétation économique directe, en raison du problème de la normalisation de la variance résiduelle. Ainsi, il faut retenir que la seule information directe réellement utilisable est le signe des paramètres, indiquant si la variable associée influence à la hausse ou la baisse la probabilité de l'événement considéré. Toutefois, on peut en outre calculer les effets marginaux : *les effets marginaux mesurent la sensibilité de la probabilité de l'événement* $y_i = 1$ *par rapport à des variations dans les variables explicatives* x_i. Et c'est précisément dans ce contexte, que l'utilisation d'un modèle logit peut faciliter l'analyse de ces effets marginaux.

Au delà, de ces différences entre les lois logistiques et normales, il existe en effet certaines propriétés du modèle logit qui sont particulièrement utiles pour simplifier les calculs ainsi que l'interprétation économique des résultats d'estimation des paramètres β associées au variables explicatives. Tout d'abord, si l'on note $p_i = Prob(y_i = 1) = \Lambda(x_i\beta)$, étant donnée la définition de la loi logistique on remarque que plusieurs égalités, permettant de simplifier les calculs, peuvent être établies comme suit :

$$e^{x_i\beta} = p_i\left(1 + e^{x_i\beta}\right)$$

$$\log\left(\frac{p_i}{1-p_i}\right) = x_i\beta$$

$$1 - p_i = \frac{1}{1 + e^{x_i\beta}}$$

En plus de ces différentes relations, il existe une égalité qui est en outre particulièrement intéressante en ce qui concerne l'analyse économique des résultats d'estimation. Il s'agit de la relation suivante :

$$e^{x_i\beta} = \frac{p_i}{1-p_i}$$

En effet, on sait que la probabilité p_i désigne la probabilité associée à l'événement $y_i = 1$, et que la quantité $1 - p_i$ désigne par conséquent la probabilité associée à l'événement comlpémentaire $p_i = 0$.

Proposition 1.6. ***De façon générale, la quantité*** $c_i = p_i/(1 - p_i)$ ***représente le rapport de la probabilité associée à l'événement*** $y_i = 1$ ***à la probabilité de non survenue de cet événment : il s'agit de la cote ("odds"). Dans un modèle logit, cette cote correspond simplement à la quantité*** $e^{x_i\beta}$ ***:***

$$c_i = \frac{p_i}{1 - p_i} = e^{x_i\beta} \quad \textit{modèle logit} \tag{1.20}$$

Si ce rapport est égal à c_i pour l'individu i, cela signifie qu'il y a c_i fois plus de chance que l'événement associé au code $y_i = 1$ se réalise, qu'il ne se réalise pas (*" c_i contre 1" dans le langage usuel*).

Exemple : Considérons les 32 observations tirées de l'échantillon de Spector et Mazzeo (1980). Les données correspondant aux variables exogènes *tuce* et *grad*, ainsi que la variable endogène dichotomique *graduate* sont reportés sur les trois premiers quadrants de la figure (1.5). A partir des estimations obtenues dans le modèle logit (cf. figure 1.2), on a calculé la cote associée à l'événement "être en post graduate". Sans surprise on constante que par exemple l'individu 10, qui a obtenu la meilleure note de l'échantillon au *tuce* (29) et qui a obtenu une moyenne de 3.92/4 aux examens de graduate a une cote de 5.9. C'est à dire qu'il a 6 plus de chances d'obtenir le passage en post graduate que de ne pas l'obtenir alors que la moyenne des cotes pour l'échantillon est de 0.97. De la même façon, l'individu 5 qui obtenu la note maximale (4) aux examens de graduate à une cote de 3.64. Ces deux individus figurent parmi les étudiants qui ont effectivement obtenu le passage en post graduate ($graduate = 1$).

Au delà du simple calcul de la cote, on peut en outre chercher à mesure *les effets marginaux sur la cote*. Il s'agit alors de mesurer l'impact, pour le $i^{ème}$ individu d'une variation de la $j^{ème}$ variable explicative, notée $x_i^{[j]}$, sur la cote. Supposons que l'on considère une variation d'une unité de cette variable, et calculons alors la variation induite de la cote. En effet, étant donné la propriété (**??**) du modèle logit, on peut alors facilement mesurer l'impact d'une variation d'une unité d'une des variables explicatives sur cette cote. En effet, si l'on note c la cote de l'événement $y_i = 1$, $x_i = \left(x_i^{[1]}...x_i^{[K]}\right)$ le vecteur des variables explicatives et $\beta = (\beta_1...\beta_K)'$ le vecteur des paramètres associés, on a :

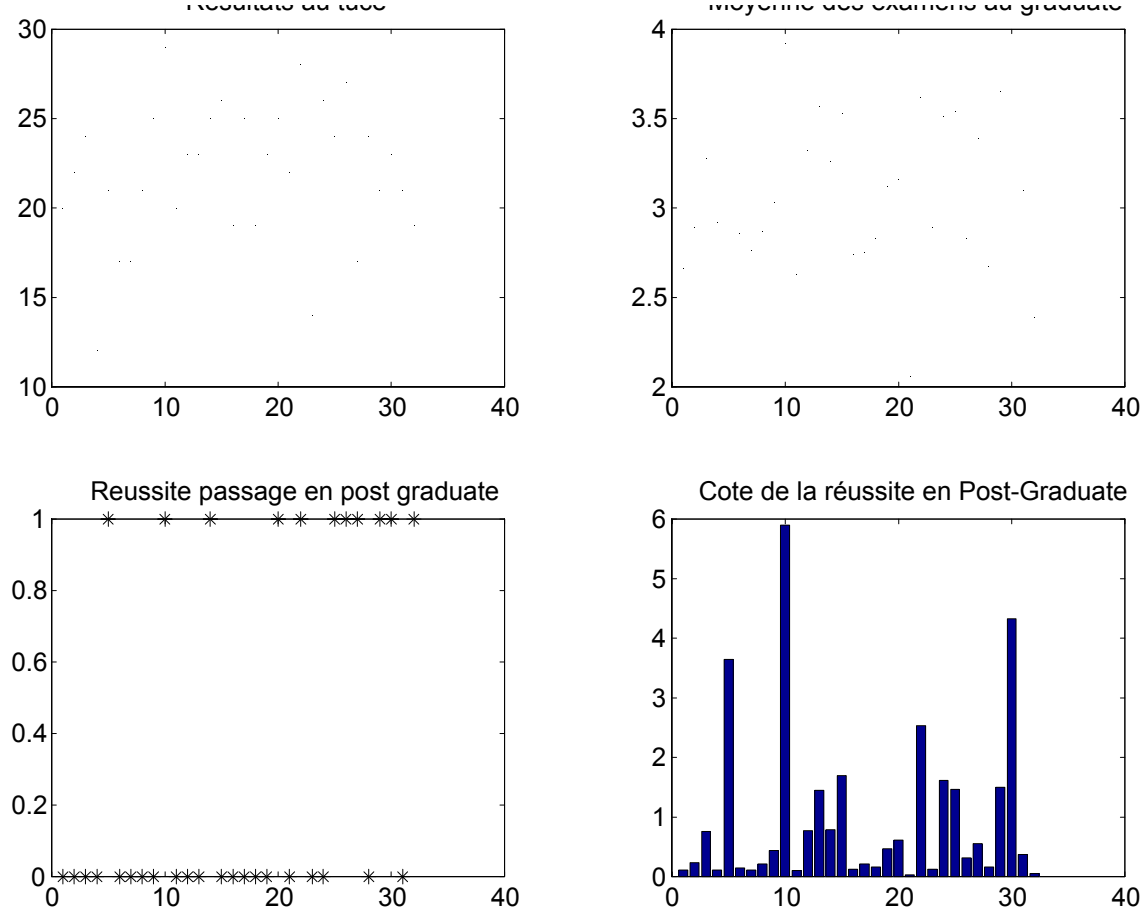
$$c_i = \frac{p_i}{1 - p_i} = \exp\left(\sum_{k=1}^{K} x_i^{[k]}\beta_k\right) = \prod_{k=1}^{K} \exp\left(x_i^{[k]}\beta_k\right)$$

On peut alors isoler la part de la cote imputable à une variable $x_i^{[j]}$ quelconque de la façon suivante. Supposons que la variable $x_i^{[j]}$ augmente de une unité, nouvelle cote notée $\overline{c}_i$ est égale à :

$$\overline{c}_i = \exp\left[\left(x_i^{[j]} + 1\right)\beta_j\right] \prod_{\substack{k=1 \\ k\neq j}}^{K} \exp\left(x_i^{[k]}\beta_k\right) = \exp\left(\beta_j\right) \prod_{k=1}^{K} \exp\left(x_i^{[k]}\beta_k\right)$$

Proposition 1.7. ***Dans un modèle logit, un accroissement d'une unité de la variable exogène*** $x_i^{[j]}$***,toutes choses égales par ailleurs, multiplie la valeur de la cote par***

Figure 1.5: Données et Calcul de la Cote à partir du Modèle Logit : Spector et Mazzeo (1980)



$\exp(\beta_j)$. ***Si l'on note*** c_i ***la cote initiale et*** $\overline{c}_i$ ***la cote obtenue après variation de la*** $j^{ème}$ ***variable explicative, on a :***

$$\overline{c}_i = \exp(\beta_j)\, c_i \tag{1.21}$$

Exemple : Considérons l'échantillon de Spector et Mazzeo.. Nous avons vu que le $10^{ème}$ individu de l'échantillon avait obtenu une note de 29 au *tuce*. Calculons la variation de sa cote s'il avait obtenu 30 au lieu de 29. Les estimations obtenues dans le modèle logit (cf. figure 1.2) nous donne une estimation du paramètre associé à *tuce* égale à 0.0855. Dès lors, le coefficient multiplicatif à appliquer à la cote est de $\exp(0.0855) = 1.0893$. La cote initiale du $10^{ème}$ individu était de 5.9. Donc après modification de la note au *tuce* sa cote doit passer à $5.9 * 1.0893 = 6.4269$. On vérifie en estimant à nouveau (non reproduit) le modèle logit avec la valeur modifiée (30) de l'exogène *tuce* pour le $10^{ème}$ individu que le cote estimée est égale à 6.43.

Toutefois, de façon plus générale, *on calcule les effets marginaux non pas à partir de la cote mais directement à partir des probabilité associé à l'événement de référence.* On cherche ainsi à

établir quelle est la variation de la probabilité de l'événement $y_i = 1$ en cas de variation d'une des variables exogène. *On considérera ici uniquement le cas de variables explicatives continues.* Dans ce cas, pour de petites variations de la j^{ème} variable explicative, on peut approximer la variation de probabilité p_i par la dérivée de celle-ci par rapport à la variable $x_i^{[j]}$:

$$\frac{\partial p_i}{\partial x_i^{[j]}} = \frac{\partial F(x_i\beta)}{\partial x_i^{[j]}} = \frac{\partial F(x_i\beta)}{\partial (x_i\beta)} \frac{\partial (x_i\beta)}{\partial x_i^{[j]}} = \frac{\partial F(x_i\beta)}{\partial (x_i\beta)} \beta_j$$

puisque $x_i\beta = \sum_{k=1}^{K} x_i^{[k]}\beta_k$.

Proposition 1.8. ***Dès lors, si l'on note*** $f(.)$ ***la fonction de densité des résidus du modèle dichotomique, l'effet marginal associé à la*** $j^{ème}$ ***variable explicative*** $x_i^{[j]}$ ***est défini par :***

$$\frac{\partial p_i}{\partial x_i^{[j]}} = f(x_i\beta).\beta_j \tag{1.22}$$

Suivant que l'on considère un modèle probit ou un modèle logit, cette dérivée s'écrit comme suit :

$$\frac{\partial p_i}{\partial x_i^{[j]}} = \frac{e^{x_i\beta}}{(1+e^{x_i\beta})^2}\beta_j \quad \textit{modèle logit} \tag{1.23}$$

$$\frac{\partial p_i}{\partial x_i^{[j]}} = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x_i\beta)^2\right].\beta_j \quad \textit{modèle probit} \tag{1.24}$$

Puisque par définition $f(.) > 0$, le signe de cette dérivée est donc identique à celui de β_j. Dès lors, l'augmentation d'une variable associée à un coefficient positif induit une hausse de la probabilité de réalisation de l'événement $y_i = 1$. Inversement, la hausse d'une variable associé à un coefficient négatif induit une baisse de la probabilité de réalisation de l'événement $y_i = 1$. Par exemple, si l'on considère les données de Spector et Mazzeo (190) et les résultats d'estimation des probit et logit (figures 1.2 et 1.3), les deux variables *tuce* et *grad* sont affectées d'un coefficient dont l'estimateur a une réalisation positive. Ainsi, une augmentation de la note au *tuce* ou une augmentation de la moyenne aux examens du graduate conduit à une amélioration de la probabilité de passage en *postgraduate*.

Enfin, plutôt que d'exprimer l'effet marginal sous la forme de la dérivée $\partial p_i/\partial x_i^{[j]}$, on préfère généralement calculer une élasticité, cette dernière ayant l'avantage d'être indépendante des unités de mesure.

Definition 1.9. ***Ainsi, on définit l'élasticité*** $\varepsilon_{p_i/x_i^{[j]}}$ ***comme la variation en pourcentage de la probabilité de survenue*** p_i ***de l'événement codé*** $y_i = 1$, ***suite à une variation de*** 1% ***de la*** $j^{ème}$ ***variable explicative*** $x_i^{[j]}$ ***:***

$$\varepsilon_{p_i/x_i^{[j]}} = \frac{\partial p_i}{\partial x_i^{[j]}} \frac{x_i^{[j]}}{p_i} = f(x_i\beta) \frac{x_i^{[j]}\beta_j}{F(x_i\beta)} \tag{1.25}$$

Cette expression peut se simplifier dans le cas du modèle logit sachant que $F(x) = e^x/(1+e^x)$ et que $f(x) = e^x/(1+e^x)^2$. Pour un logit, l'élasticité prend la valeur suivante :

$$\forall i \in [1, N] \quad \varepsilon_{p_i/x_i^{[j]}} = \frac{x_i^{[j]}\beta_j}{1+\exp(x_i\beta)} \quad \text{modèle logit} \tag{1.26}$$

Plusieurs remarques doivent être faites à ce niveau. Tout d'abord, pour les deux modèles, l'élasticité est une fonction non linéaire des autres composantes du vecteur x_i. On peut ainsi calculer l'influence des variables explicatives annexes sur la sensibilité du modèle à l'évolution d'une variable j particulière. On peut par exemple calculer :

$$\frac{\partial\left(\varepsilon_{p_i/x_i^{[j]}}\right)}{\partial x_i^{[k]}} \quad \forall k \neq j, \forall i \in [1,N] \tag{1.27}$$

Deuxième remarque, les fonctions de densité $f(.)$ des modèles logit et probit étant symétriques et unimodales,elles atteignent donc leur maximum en zéro. *Ainsi, l'impact d'une variable explicative est d'autant plus important pour les individus donc le scalaire $x_i\beta$ est proche de zéro. En d'autres termes, pour les individus pour lesquels on est pratiquement sûr de la survenue d'un évenement ($p_i = F(x_i\beta)$ proche de 1 ou $x_i\beta$, positif et très élevé), l'élasticité sera faible : seule une variation très importante des variables explicatives pourra modifier sensiblement la probabilité.* De la même façon, les individus pour lesquels on est pratiquement sûr de la non survenue d'un évenement ($p_i = F(x_i\beta)$ proche de 0 ou $x_i\beta$, négatif et très élevé en valeur absolue), l'élasticité sera faible.

Enfin, troisième et dernière remarque les formules ci-dessus fournissent des mesures individuelles des effets marginaux, et généralement il est utile de calculer l'élasticité au point moyen de l'échantillon afin de répondre à la question : quel est l'impact moyen (dans l'échantillon) de la variation de 1% de la j^{ème} variable explicatives ? Deux possibilité peuvent être retenues : soit on calcule l'élasticité en remplaçant les valeurs individuelles x_i par les moyennes empiriques de ces composantes sur l'échantillon, $\forall j \in [1,K]$:

$$\overline{\varepsilon}_{p/x^j} = \frac{f(\overline{x}\beta)}{F(\overline{x}\beta)}\overline{x}^{[j]}\beta_j \tag{1.28}$$

où le vecteur $\overline{x}$ est défini par $\overline{x} = (1/N)\, x_i$ et le scalaire $\overline{x}^{[j]}$ vaut $\overline{x}^{[j]} = (1/N)\, x^{[j]}$. La deuxième solution consiste à calculer la moyenne des élasticités individuelles sur l'ensemble de l'échantillon, $\forall j \in [1,K]$:

$$\overline{\varepsilon}_{p/x^j} = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_{p_i/x_i^{[j]}} \tag{1.29}$$

1.4. Présentation des modèles dichotomiques en termes de variable latente

Généralement, bien que cela ne soit pas nécessaire on présente les modèles dichotomiques en termes de variables latentes ou inobservée y_i^*, la variable observée y_i étant alors un indicateur des valeurs prises par y_i^*. Cette référence à une variable latente permet de mieux comprendre l'émergence des modèles dichotomiques à partir de certains problèmes ou de biologie.

L'exemple le plus célébre (repris dans Amemiya 1981) est tiré de la bio-économétrie (n'oublions que c'est dans ce domaine que furent proposées les premières applications) celui de l'insecticide : on diffuse dans un espace clos un insecticide et l'on cherche à dé terminer la dose minimale permettant de tuer les insectes. Pour cela, on observe au terme d'une période fixé les insectes i morts pour lesquels on adopte le code $y_i = 0$ et ceux encore vivants codés $y_i = 1$. On suppose alors que chaque insecte dispose d'une capacité de résistance propre qui se traduit par un seuil inobservable de produit, noté y_i^*, telle que si la dose de produit est supérieure à ce seuil l'insecte est mort ($y_i = 0$), et qu'il reste vivant (mais malade peut être) pour une dose

inférieure ($y_i = 1$). Il s'agit alors de modéliser la probabilité de survie de l'insecte i en fonction de la dose d'insecticide et des observations faites sur y_i. On suppose pour cela qu'un certain dosage γ est diffusé sur l'ensemble des insectes. On voit immédiatement que ce problème peut s'écrire de la façon suivante :

$$y_i = \begin{cases} 1 & \text{si } y_i^* > \gamma \\ 0 & \text{sinon} \end{cases} \tag{1.30}$$

où la variable latente y_i^* peut s'écrire comme la somme d'une combinaison linéaire de caractéristiques propres à chaque insecte et d'une terme aléatoire.

$$y_i^* = x_i\beta + \varepsilon_i \tag{1.31}$$

Si le terme aléatoire ε_i est distribué selon une loi normale, on retrouve un modèle probit, si ce terme est distribué selon une loi logistique on retrouve le modèle logit.

Un autre exemple, toujours tiré d'une étude biologique de Ashford et Sowden (1970), concerne la probabilité pour un mineur de contracter une maladie des poumons (événement codé $y_i = 1$) lorsque sa tolérance inobservable, notée y_i^*, aux conditions de travail et en particulier aux poussières de charbon est inférieure à certain seuil γ inconnue. On suppose que la tolérance est liée à l'âge du mineur noté x_i. De la même façon, ce modèle peut s'écrire sous la forme :

$$y_i = \begin{cases} 1 & \text{si } y_i^* = \beta_1 + \beta_2 x_i + \varepsilon_i < \gamma \\ 0 & \text{sinon} \end{cases} \tag{1.32}$$

où ε_i a une distribution normale ou logistique suivant les modèles. Ici l'événement $y_i = 1$ (maladie) apparaît quand la variable latente y_i^* est inférieure à un seuil γ. Mais il aurait parfaitement été possible de considérer une variable latente égale à $-y_i^*$ et un seuil $-\theta$ pour retomber sur une relation semblable à celle de l'exemple précédent où $y_i^* > \gamma$. Une autre manière aurait consister à coder l'événement "maladie" en 0. Par la suite, nous considérerons un modèle où l'on a $y_i = 1$ lorsque $y_i^* > \gamma$, ce qui permet d'écrire que $p_i = F(x_i\beta - \gamma)$. En effet, on a bien[11] :

$$p_i = Prob(y_i = 1) = Prob(y_i^* > \gamma)$$

$$\Longleftrightarrow p_i = Prob(\varepsilon_i > \gamma - x_i\beta) = 1 - Prob(\varepsilon_i < \gamma - x_i\beta)$$

$$\Longleftrightarrow p_i = F(x_i\beta - \gamma) \tag{1.33}$$

Dans le cas où $\gamma = 0$, on retrouve l'écriture des modèles dichotomiques proposée jusqu'à présent : $p_i = F(x_i\beta)$.

Proposition 1.10. ***Tout modèle dichotomique univarié peut s'écrire sous la forme d'une équation de mesure de la forme :***

$$y_i = \begin{cases} 1 & \textit{si } y_i^* > \gamma \\ 0 & \textit{sinon} \end{cases} \tag{1.34}$$

où $\gamma \in \mathbb{R}$ ***et où la variable latente*** y_i^* ***inobservable est définie en fonction de caractéristiques observables*** x_i ***et d'une perturbation*** ε_i *i.i.d.* $\left(0, \sigma_\varepsilon^2\right)$ ***:***

$$y_i^* = x_i\beta + \varepsilon_i \tag{1.35}$$

[11] On suppose que la loi des perturbations est symétrique $f(x) = f(-x)$, dès lors on a $F(x) = 1 - F(-x)$.

Ce modèle peut également s'exprimer sous la forme :

$$p_i = Prob\left(y_i = 1\right) = F\left(x_i\beta - \gamma\right) \tag{1.36}$$

où la fonction $F\left(.\right)$ ***désigne la fonction de répartition associée à la loi des perturbations*** ε_i.

Ainsi, si $F\left(.\right) = \Phi\left(.\right)$ on retrouve le modèle probit et si $F\left(.\right) = \Lambda\left(.\right)$ on retrouve le cas du modèle logit. De façon générale, l'équation (1.33) correspond en effet aux définitions des modèles logit et probit posées dans la section précédente.

A ce stade deux aspects doivent être discutés (Colletaz 2001). Le premier aspect concerne **la normalisation du seuil** γ qui évidement ne peut être identifié que si la combinaison linéaire $x_i\beta$ ne comporte pas de terme constant. Si la combinaison linéaire inclut un terme constant et s'écrit sous la forme $x_i\beta = \beta_1 + \sum_{j=2}^{K} x_{i,j}\beta_j$, alors il est seulement possible d'estimer la constante c telle que :

$$p_i = F\left(x_i\beta - \gamma\right) = F\left(\beta_1 + \sum_{j=2}^{K} x_{i,j}\beta_j - \gamma\right) = F\left(\overline{\beta}_1 + \sum_{j=2}^{K} x_{i,j}\beta_j\right)$$

Il y a alors indetermination du couple (β_1, γ) puisqu'il existe une infinité de couples tels que $\overline{\beta}_1 = \beta_1 - \gamma$. Deux choses l'une : ou l'on possède une information a priori sur le seuil γ qui permet alors de lever l'indétermination et d'identifier β_1, soit l'on impose a priori une contrainte sur l'une ou l'autre des paramètres pour identifier l'autre. Dans ce dernier cas, généralement on suppose $\gamma = 0$ ce qui permet d'obtenir l'égalité $\beta_1 = \overline{\beta}_1$. Sans perte de généralité, on considère donc une écriture de la forme :

$$p_i = F\left(x_i\beta\right) \tag{1.37}$$

Le second aspect du modèle à variable latente concerne **la normalisation de la variance des perturbations** ε_i. Partant de la relation (1.37) pour $\gamma = 0$, on a $p_i = F\left(x_i\beta\right) = Prob\left(\varepsilon_i < x_i\beta\right)$ et donc $\forall\lambda \in \mathbb{R}^+$, on obtient :

$$p_i = Prob\left(\frac{\varepsilon_i}{\lambda} < \frac{x_i\beta}{\lambda}\right) = Prob\left(\widetilde{\varepsilon}_i < x_i\widetilde{\beta}\right) \quad \forall\lambda > 0 \tag{1.38}$$

avec $\widetilde{\beta} = \beta/\lambda$ et $\widetilde{\varepsilon}_i = \varepsilon_i/\lambda$, $\forall i \in (1, N)$. En d'autres termes, la détermination de la probabilité p_i n'est pas unique par rapport au terme aléatoire ε_i et au vecteur de paramètres β : à caractéristiques (y_i, x_i) données, une infinité de couples $\left\{\widetilde{\varepsilon}_i, \widetilde{\beta}\right\}$ conduit à une même probabilité p_i de survenue de l'événement codé $y_i = 1$. Cette infinité de couples est définie par la proportionnalité :

$$\left\{\widetilde{\varepsilon}_i, \widetilde{\beta}\right\} = \frac{1}{\lambda}\left\{\varepsilon_i, \beta\right\} \quad \forall\lambda \in \mathbb{R}^+ \tag{1.39}$$

Le choix d'une solution unique s'effectue encore une fois en imposant une contrainte soit sur le vecteur des paramètres $\widetilde{\beta}$*, soit sur la loi des perturbations* $\widetilde{\varepsilon}_i$*, et plus précisément sur leur variance, la loi étant fixée par le choix du modèle logit ou probit. C'est cette dernière solution qui est généralement privilégiée. On sait en effet que la variance des résidus* ε_i *est égale à* $\pi^2/3$ *dans le cadre du modèle logit et que cette variance est égale à l'unité dans le modèle probit.* Les variances des perturbations étant fixée par le choix de la loi $F\left(.\right)$, c'est donc sur le vecteur de

paramètres β que porte l'incertitude puisque les composantes de ce vecteur sont définis à un facteur λ positif près. Naturellement, cette incertitude est sans conséquence pratique puisque toute composante non nulle dans le "vrai" vecteur β a une image dans le β contraint et que par ailleurs les deux valeurs étant de même signe cela n'affecte pas la mesure des effets marginaux.

Proposition 1.11. ***Dans les modèles logit et probit, la variance de l'erreur du modèle n'est pas identifiable : elle est normalisée à l'unité dans le cas du probit et est égale à*** $\pi^2/3$ ***dans le cas du logit. Par conséquent, la valeur numérique des paramètres estimés n'a pas d'intérêt en soi dans la mesure où il ne correspondent aux paramètres*** β ***de l'équation de la variable latente qu'à une constante multiplicative près. De plus, le seuil*** γ ***n'est pas identifiable car il se confond au terme constant du vecteur des explicatives*** x_i.

Ainsi, la seule information réellement utilisable est le signe des paramètres, indiquant si la variable associée influence à la hausse ou la baisse la probabilité de l'événement considéré. *Le signe des coefficients et le calcul des effets marginaux restent les deux seules informations directement exploitables en ce qui concerne les variables explicatives.*

Exemple : afin de mieux comprendre reprenons l'exemple du modèle de Ashford et Sowden (1970), où l'on considère la probabilité pour un mineur de contracter une maladie des poumons (événement codé $y_i = 1$) lorsque sa tolérance inobservable, notée y_i^*, aux conditions de travail et en particulier aux poussières de charbon est inférieure à certain seuil γ inconnue. On suppose que la tolérance est liée à l'âge du mineur noté x_i par une relation affine.

$$y_i = \begin{cases} 1 & \text{si } y_i^* = \beta_1 + x_i\beta_2 + \varepsilon_i > \gamma \\ 0 & \text{sinon} \end{cases}$$

On suppose que la variance des perturbations *i.i.d.* ε_i est égale à $\sigma_i^2 = \sigma^2$, $\forall i \in (1, N)$. Dès lors, pour un individu i la probabilité de décès s'écrit sous la forme :

$$\begin{aligned} p_i &= Prob\left(y_i = 1\right) \\ &= Prob\left(\varepsilon_i > \gamma - \beta_1 - x_i\beta_2\right) \\ &= F\left(\beta_1 - \gamma + x_i\beta_2\right) \end{aligned} \tag{1.40}$$

Si l'on considère un modèle probit, les perturbations du modèle doivent suivre une loi normale centrée réduite. La contrainte sur la variance égale à l'unité, impose d'écrire le modèle sous la forme suivante :

$$p_i = Prob\left(\frac{\varepsilon_i}{\sigma} > \frac{\gamma - \beta_1 - x_i\beta_2}{\sigma}\right) \tag{1.41}$$

$$= \Phi\left(\frac{\beta_1 - \gamma}{\sigma} + \frac{x_i\beta_2}{\sigma}\right) \tag{1.42}$$

$$= \Phi\left(\widetilde{\beta}_1 + x_i\widetilde{\beta}_2\right) \tag{1.43}$$

avec $\widetilde{\beta}_1 = \left(\beta_1 - \gamma\right)/\sigma$ et $\widetilde{\beta}_2 = \beta_2/\sigma$. Seuls deux paramètres $\widetilde{\beta}_1$ et $\widetilde{\beta}_2$ seront estimés, alors qu'il y a 4 paramètres structurels $(\beta_1, \beta_2, \gamma, \sigma)$. L'adoption d'une normalisation du type $\gamma = 0$ et $\sigma = 1$ permet alors d'identifier les paramètres β_1 et β_2.

Si l'on considère un modèle logit, on sait que la variance résiduelle doit être égale à $\pi^2/3$ dès lors que l'on impose le choix d'une loi logistique simple pour les perturbations du modèle.

Ainsi, la contrainte sur la variance résiduelle égale à $\pi^2/3$, impose d'écrire le modèle sous la forme suivante :

$$\begin{aligned} p_i &= Prob\left(\frac{\pi}{\sqrt{3}}\frac{\varepsilon_i}{\sigma} > \frac{\pi}{\sqrt{3}}\frac{\gamma-\beta_1-x_i\beta_2}{\sigma}\right) \\ &= \Lambda\left(\frac{\pi}{\sqrt{3}}\frac{\beta_1-\gamma}{\sigma}+\frac{\pi}{\sqrt{3}}\frac{x_i\beta_2}{\sigma}\right) \\ &= \Lambda\left(\widetilde{\beta}_1+x_i\widetilde{\beta}_2\right) \end{aligned} \tag{1.44}$$

avec $\widetilde{\beta}_1 = \pi\left(\beta_1-\gamma\right)/\sqrt{3}\sigma$ et $\widetilde{\beta}_2 = \pi\beta_2/\sqrt{3}\sigma$. En effet, dans ce cas les perturbations normalisées $\widetilde{\varepsilon}_i = \pi\varepsilon_i/\sigma\sqrt{3}$ vérifient la contrainte sur la variance puisque :

$$E\left(\widetilde{\varepsilon}_i^2\right) = \frac{\pi^2}{3\sigma^2}E\left(\varepsilon_i\right) = \frac{\pi^2}{3}$$

Encore une fois, seuls les paramètres $\widetilde{\beta}_1$ et $\widetilde{\beta}_2$ seront estimés, alors qu'il y a 4 paramètres structurels $(\beta_1, \beta_2, \gamma, \sigma)$ dans le modèle initial. L'adoption d'une normalisation du type $\gamma = 0$ et $\sigma = 1$ permet dans ce cas d'identifier les paramètres β_1 et β_2.

2. Estimation des Paramètres par la Méthode du Maximum de Vraisemblance

Considérons le modèle suivant :

Hypothèse *On considère un échantillon de N individus indicés $i = 1, .., N$. Pour chaque individu, on observe si un certain évenément s'est réalisé et l'on note y_i la variable codée associée à évenement. On pose $\forall i \in [1, N]$:*

$$y_i = \begin{cases} 1 & p_i = F(x_i\beta) \\ 0 & 1 - p_i = 1 - F(x_i\beta) \end{cases} \tag{2.1}$$

où $x_i = \left(x_i^1 ..x_i^K\right)$, $\forall i = 1, .., N$ désigne un vecteur de caractéristiques observables et où $\beta = (\beta_1 ...\beta_K)' \in \mathbb{R}^K$ est un vecteur de paramètres inconnus.

On cherche naturellement à estimer les composantes du vecteur β. Dans le cas des modèles dichotomiques univariés, plusieurs méthodes d'estimation sont envisageables (GMM par exemple). Toutefois la méthode la plus usitée lorsque la loi des perturbations est connue consiste en la méthode du maximum de vraisemblance. Nous ne considérerons pas ici le cas des observations répétées[12].

2.1. Estimation par maximum de vraisemblance

Dans le cas du modèle dichotomique univarié, la construction de la vraisemblance est extrêmement simple. En effet, à l'événement $y_i = 1$ est associée la probabilité $p_i = F(x_i\beta)$ et à l'événement $y_i = 0$ correspond la probabilité $1 - p_i = 1 - F(x_i\beta)$. Ceci permet de considérer les valeurs observées y_i comme les réalisations d'un processus binomial avec une probabilité de $F(x_i\beta)$. La vraisemblance des échantillons associés aux modèles dichotomiques s'écrit donc comme la vraisemblance d'échantillons associés à des modèles binomiaux. La seule particularité étant que les probabilités p_i varient avec l'individu puisqu'elles dépendent des caractéristiques x_i. Ainsi la vraisemblance associée à l'observation y_i s'écrit sous la forme :

$$L(y_i, \beta) = p_i^{y_i} (1 - p_i)^{1-y_i}$$

Dès lors, la vraisemblance associée à l'échantillon de taille N, noté $y = (y_1, .., y_N)$ s'écrit de la façon suivante.

Definition 2.1. ***Pour un modèle dichotomique univarié simple, la vraisemblance associée à l'échantillon de taille*** N, ***noté*** $y = (y_1, .., y_N)$, ***s'écrit sous la forme :***

$$L(y, \beta) = \prod_{i=1}^{N} p_i^{y_i} (1 - p_i)^{1-y_i} = \prod_{i=1}^{N} [F(x_i\beta)]^{y_i} [1 - F(x_i\beta)]^{1-y_i} \tag{2.2}$$

Il ne reste plus alors qu'à spécifier la fonction de distribution $F(.)$ pour obtenir la forme fonctionnelle de la vraisemblance. Ainsi, $\forall x_i\beta \in \mathbb{R}$ dans le cas du modèle logit, on a:

[12] Cas où à chaque valeur des caractéristiques exogènes correspondent plusieurs observations du caractère qualitatif. Ceci traduit la possibilité de répéter plusieurs fois l'expérience sous les mêmes conditions. Comme le note Anemiya (1980) ce cas est plus fréquent en biologie qu'en économie.

$$F(x_i\beta) = \frac{e^{x_i\beta}}{1+e^{x_i\beta}} = \Lambda(x_i\beta)$$

alors que pour le probit, on a :

$$F(x_i\beta) = \int_{-\infty}^{x_i\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz = \Phi(x_i\beta)$$

De cette définition, on déduit alors la log-vraisemblance comme suit :

$$\log L(y,\beta) = \sum_{i=1}^{N} y_i \log\left[F(x_i\beta)\right] + (1-y_i)\log\left[1-F(x_i\beta)\right] \tag{2.3}$$

En distinguant les observations $y_i = 1$ et celles pour lesquelles on a $y_i = 0$, la log-vraisemblance peut s'écrire sous la forme :

$$\log L(y,\beta) = \sum_{i\,:\,y_i=1} \log F(x_i\beta) + \sum_{i\,:\,y_i=0} \log\left[1-F(x_i\beta)\right] \tag{2.4}$$

L'estimateur du maximum de vraisemblance des paramètres β est obtenu en maximisant soit la fonction de vraisemblance $L(y,\beta)$ soit la fonction de log-vraisemblance $\log L(y,\beta)$. En dérivant la log vraisemblance (équation 2.3) par rapport aux éléments du vecteur β, de dimension $(K,1)$, on obtient un vecteur de dérivées, noté $G(\beta)$, appelé vecteur du gradient.

$$G(\beta) = \frac{\partial \log L(y,\beta)}{\partial\beta} = \sum_{i=1}^{N} y_i \frac{f(x_i\beta)}{F(x_i\beta)} x_i' + (y_i - 1)\frac{f(x_i\beta)}{1-F(x_i\beta)} x_i'$$

où $f(.)$ est la fonction de densité associée à $F(.)$ et où x_i'désigne la transposée du vecteur x_i de dimension $(1,K)$. En simplifiant, l'expression du gradient, on obtient alors :

$$G(\beta) = \sum_{i=1}^{N} \frac{\left[y_i - F(x_i\beta)\right] f(x_i\beta)}{F(x_i\beta)\left[1-F(x_i\beta)\right]} x_i' \tag{2.5}$$

On peut en outre exprimer le gradient en distinguant les observations $y_i = 1$ et celles pour lesquelles on a $y_i = 0$:

$$G(\beta) = \sum_{i\,:\,y_i=1} \frac{f(x_i\beta)}{F(x_i\beta)} x_i' - \sum_{i\,:\,y_i=0} \frac{f(x_i\beta)}{\left[1-F(x_i\beta)\right]} x_i' \tag{2.6}$$

Definition 2.2. ***L'estimateur $\widehat{\beta}$ du maximum de vraisemblance du vecteur de paramètre $\beta \in \mathbb{R}^K$ dans un modèle dichotomique est défini par la résolution du système de K équations non linéaires en β :***

$$\widehat{\beta} = \arg\max_{\{\beta\}} \left[\log L(y,\beta)\right] \tag{2.7}$$

$$\Longleftrightarrow \frac{\partial \log L\left(y,\widehat{\beta}\right)}{\partial\widehat{\beta}} = \sum_{i=1}^{N} \frac{\left[y_i - F\left(x_i\widehat{\beta}\right)\right] f\left(x_i\widehat{\beta}\right)}{F\left(x_i\widehat{\beta}\right)\left[1-F\left(x_i\widehat{\beta}\right)\right]} x_i' = G\left(\widehat{\beta}\right) = 0 \tag{2.8}$$

où $G(\beta)$ ***désigne le gradient associé à la log-vraisemblance*** $\partial \log L(y,\beta)$***, évalué au point*** $\widehat{\beta}$***.Dans le cas du modèle logit, ce système se ramène à :***

$$G_L\left(\widehat{\beta}\right) = \sum_{i=1}^{N}\left[y_i - \Lambda\left(x_i\widehat{\beta}\right)\right]x_i' = 0 \tag{2.9}$$

Dans le cas du modèle probit, on a :

$$G_P\left(\widehat{\beta}\right) = \sum_{i=1}^{N}\frac{\left[y_i - \Phi\left(x_i\widehat{\beta}\right)\right]\phi\left(x_i\widehat{\beta}\right)}{\Phi\left(x_i\widehat{\beta}\right)\left[1-\Phi\left(x_i\widehat{\beta}\right)\right]}x_i' = 0 \tag{2.10}$$

En effet, l'écriture du gradient dans le cas du modèle logit se simplifie en tenant compte de la propriété de la loi logistique selon laquelle, si l'on note $\lambda(x)$ la densité associée à $\Lambda(x)$, on a la relation suivante : $\forall x,\ \lambda(x) = \Lambda(x)[1-\Lambda(x)]$. Dès lors, l'expression (2.5) se simplifie puisque :

$$G_L(\beta) = \sum_{i=1}^{N}\frac{[y_i - \Lambda(x_i\beta)]\,\lambda(x_i\beta)}{\Lambda(x_i\beta)[1-\Lambda(x_i\beta)]}x_i' = \sum_{i=1}^{N}[y_i - \Lambda(x_i\beta)]\,x_i'$$

Première remarque : comme de façon générale avec la méthode d'estimation du maximum de vraisemblance, l'équation de définition (2.8) peut s'interpréter comme une condition d'orthogonalité imposée sur les variables explicatives et les résidus généralisés. Cette égalité est en effet l'équivalent empirique d'une condition de la forme $E[(x_i'w_i)\,\varepsilon_i]$ où ε_i est le résidu dans le modèle non linéaire $y_i = F(x_{i\beta}) + \varepsilon_i$ et où w_i est une variable de pondération. En effet, si l'on pose :

$$w_i = \frac{f(x_i\beta)}{F(x_i\beta)[1-F(x_i\beta)]} \qquad \varepsilon_i = y_i - F(x_i\beta)$$

alors l'équation (2.8) se réécrit sous la forme :

$$G(\beta) = \sum_{i=1}^{N}(x_i'w_i)[y_i - F(x_i\beta)] = 0 \Longleftrightarrow \frac{1}{N}\sum_{i=1}^{N}(x_i'w_i)\,\varepsilon_i = 0 \tag{2.11}$$

Cette propriété est particulièrement facile à visualiser dans le cas du modèle logit. De façon générale, *les estimateurs du maximum de vraisemblance constituent un cas particulier des estimateurs des moments.*

Deuxième remarque : le système défini par l'équation (2.8) est non linéaire. *L'estimateur* $\widehat{\beta}$ *ne peut être obtenu directement. Un algorithme d'optimisation numérique de la vraisemblance est donc nécessaire.* Comme nous le verrons dans la section suivante, ces algorithmes se fondent à la fois sur le gradient mais aussi sur la matrice hessienne des dérivées secondes. C'est pourquoi, nous allons donné l'expression des gradients et des matrice hessiennes, notées $H(\beta)$, dans le cas particulier des modèles logit et probit.

2.1.1. Matrices Hessiennes et Matrices d'information de Fischer

Commençons par définir les matrices hessiennes associée à la log vraisemblance des modèles dichotomiques univariés.

Definition 2.3. ***Pour un modèle dichotomique univarié, la matrice hessienne associée à la log vraisemblance d'un échantillon de taille*** N, ***noté*** $y = (y_1, .., y_N)$, ***s'écrit sous la forme :***

$$\begin{aligned} \underset{(K,K)}{H\ (\beta)} &= \frac{\partial^2 \log L(y,\beta)}{\partial\beta\partial\beta'} = -\sum_{i=1}^{N}\left[\frac{y_i}{F(x_i\beta)^2} + \frac{1-y_i}{[1-F(x_i\beta)]^2}\right] f(x_i\beta)^2 x_i' x_i \\ &\quad + \sum_{i=1}^{N}\left[\frac{y_i - F(x_i\beta)}{F(x_i\beta)[1-F(x_i\beta)]}\right] f'(x_i\beta) x_i' x_i \end{aligned} \tag{2.12}$$

où $f'(.)$ ***désigne la dérivée de la fonction de densité*** $f(.)$ ***associée à*** $F(.)$.

En effet, en omettant les arguments des fonctions et les indices il vient :

$$\begin{aligned} H\ (\beta) &= \frac{\partial}{\partial\beta}\left(\frac{\partial \log L(y,\beta)}{\partial\beta}\right)' = \frac{\partial}{\partial\beta} G(\beta)' \\ &= \frac{\partial}{\partial\beta}\left[\sum_{i=1}^{N}\frac{(y_i - F)\ f}{F(1-F)}\right] x_i \\ &= \sum \frac{F(1-F)}{F^2(1-F)^2}\frac{\partial[(y-F)f]}{\partial\beta}x - \sum\frac{(y-F)f}{F^2(1-F)^2}\frac{\partial[F(1-F)]}{\partial\beta}x \end{aligned}$$

En simplifiant, il vient :

$$\begin{aligned} H\ (\beta) &= \sum \frac{-f^2 + (y-F)f'}{F(1-F)}x'x - \sum\frac{(y-F)f}{F^2(1-F)^2}[f(1-F) - Ff]\ x'x \\ &= -\sum\frac{f^2}{F(1-F)}x'x + \sum\frac{(y-F)f'}{F(1-F)}x'x - \sum\frac{f^2(y-F)}{F^2(1-F)}\ x'x + \sum\frac{f^2(y-F)}{F(1-F)^2}x'x \end{aligned}$$

En regroupant les termes en f^2 et en f' on obtient alors :

$$\begin{aligned} H\ (\beta) &= \sum\frac{f^2 x'x}{F^2(1-F)^2}[F(1-F) + (y-F)F - (y-F)(1-F)] + \sum\frac{(y-F)f'}{F(1-F)}x'x \\ &= \sum\frac{f^2 x'x}{F^2(1-F)^2}\left[2yF - F^2 - y\right] + \sum\frac{(y-F)f'}{F(1-F)}x'x \\ &= -\sum\frac{f^2}{F^2(1-F)^2}\left[y(1-F)^2 + (1-y)F^2\right]\ x'x + \sum\frac{(y-F)f'}{F(1-F)}x'x \\ &= -\sum\frac{yf^2}{F^2}x'x - \sum\frac{(1-y)\ f^2}{(1-F)^2}x'x + \sum\frac{(y-F)f'}{F(1-F)}x'x \end{aligned}$$

En intégrant les indices et les arguments des fonctions $F(.)$, $f(.)$ et $f'(.)$ on retrouve alors l'expression de la matrice hessienne $H(\beta)$ donnée dans l'équation (2.12). Attention, *il n'existe pas d'expression simplifiée dans le cas des modèles logit et probit de la matrice hessienne.* En revanche, l'espérance de la matrice hessienne, qui intervient dans le calcul de la matrice de variance covariance asymptotique de l'estimateur de maximum de vraisemblance, a une écriture plus simple.

En effet, en partant de l'expression (2.12) de la matrice hessienne de la fonction de log vraisemblance et en considérant que dans le modèle dichotomique on a :

$$E(y_i) = F(x_i\beta) \tag{2.13}$$

on peut alors établir que :

$$\begin{aligned} E\left[H\left(\beta\right)\right] &= E\left[\frac{\partial^2 \log L\left(y,\beta\right)}{\partial\beta\partial\beta'}\right] = -\sum_{i=1}^{N}\left[\frac{E\left(y_i\right)}{F\left(x_i\beta\right)^2} + \frac{E\left(1-y_i\right)}{\left[1-F\left(x_i\beta\right)\right]^2}\right] f\left(x_i\beta\right)^2 x_i'x_i \\ &= -\sum_{i=1}^{N}\left[\frac{1}{F\left(x_i\beta\right)} + \frac{1}{1-F\left(x_i\beta\right)}\right] f\left(x_i\beta\right)^2 x_i'x_i \end{aligned}$$

En effet, le second terme de l'expression (2.12) s'annule lorsque l'on applique l'opérateur espérance. Cette expression peut alors se simplifier comme suit :

$$E\left[H\left(\beta\right)\right] = -\sum_{i=1}^{N}\frac{f\left(x_i\beta\right)^2}{F\left(x_i\beta\right)\left[1-F\left(x_i\beta\right)\right]} x_i'x_i$$

On reconnaît ici bien sûr, l'expression de l'opposé de la matrice d'information de Fischer.

Definition 2.4. ***Pour un modèle dichotomique univarié, la matrice d'information de Fischer*** $I\left(\beta\right)$ ***s'écrit sous la forme :***

$$I\left(\beta\right) = -E\left[\frac{\partial^2 \log L\left(y,\beta\right)}{\partial\beta\partial\beta'}\right] = \sum_{i=1}^{N}\frac{f^2\left(x_i\beta\right)}{F\left(x_i\beta\right)\left[1-F\left(x_i\beta\right)\right]}x_i'x_i \tag{2.14}$$

Dans le cas du modèle logit, cette matrice est définie par :

$$I\left(\beta\right) = \sum_{i=1}^{N}\lambda\left(x_i\beta\right)x_i'x_i = \sum_{i=1}^{N}\frac{\exp\left(x_i\beta\right)}{\left[1+\exp\left(x_i\beta\right)\right]^2}x_i'x_i \tag{2.15}$$

Dans le cas du modèle probit, cette matrice est définie par :

$$I\left(\beta\right) = \sum_{i=1}^{N}\frac{\phi^2\left(x_i\beta\right)}{\Phi\left(x_i\beta\right)\left[1-\Phi\left(x_i\beta\right)\right]}x_i'x_i \tag{2.16}$$

En effet, dans le cas du modèle logit on a $\Lambda\left(x\right)\left[1-\Lambda\left(x\right)\right] = \lambda\left(x\right)$, dès lors l'expression de la matrice d'information de Fischer se simplifie comme suit :

$$I\left(\beta\right) = \sum_{i=1}^{N}\frac{\lambda^2\left(x_i\beta\right)}{\Lambda\left(x_i\beta\right)\left[1-\Lambda\left(x_i\beta\right)\right]}x_i'x_i = \sum_{i=1}^{N}\lambda\left(x_i\beta\right)x_i'x_i \tag{2.17}$$

Il nous reste à présent à montrer que si la fonction de log vraisemblance admet un maximum global, ce dernier est unique.

2.1.2. Unicité du maximum global de la fonction de log-vraisemblance

Si l'on admet que le maximum global de $\log L\left(y,\beta\right)$ existe, la condition suffisante pour que ce maximum soit unique consiste à montrer que la fonction $\log L\left(y,\beta\right)$ est concave. Etant donnée l'écriture (2.4) de la log-vraisemblance, il suffit alors de montrer que les fonctions $\log\left[F\left(x\right)\right]$ et $\log\left[1-F\left(x\right)\right]$ sont concaves.

Dans le cas du modèle logit, les dérivées première et seconde de la fonction $\log[F(x)] = \log[\Lambda(x)]$ sont les suivantes :

$$\frac{\partial \log[\Lambda(x)]}{\partial x} = \frac{1}{\Lambda(x)}\frac{\partial \Lambda(x)}{\partial x} = \frac{(1+e^x)}{e^x}\frac{e^x}{(1+e^x)^2} = \frac{1}{1+e^x}$$

$$\frac{\partial^2 \log[\Lambda(x)]}{\partial x^2} = \frac{\partial}{\partial x}\left(\frac{1}{1+e^x}\right) = \frac{-e^x}{(1+e^x)^2} < 0$$

Les dérivées première et seconde de la fonction $\log[1-\Lambda(x)]$ sont les suivantes :

$$\frac{\partial \log[1-\Lambda(x)]}{\partial x} = -\frac{1}{1-\Lambda(x)}\frac{\partial \Lambda(x)}{\partial x} = -\frac{(1+e^x)}{1}\frac{e^x}{(1+e^x)^2} = -\frac{e^x}{1+e^x} = -\Lambda(x)$$

$$\frac{\partial^2 \log[1-\Lambda(x)]}{\partial x^2} = -\frac{\partial \Lambda(x)}{\partial x} = \frac{-e^x}{(1+e^x)^2} < 0$$

Dans le cas du logit, les fonctions $\log[F(x)]$ et $\log[1-F(x)]$ sont donc strictement concaves, donc la log-vraisemblance $\log L(y,\beta)$ est elle même strictement concave. S'il existe un maximum à cette fonction en β, ce maximum est global. Le même résultat peut être mis en évidence dans le cas du modèle probit.

Proposition 2.5. ***Dans un modèle dichotomique univarié, la fonction de log-vraisemblance*** $\log L(y,\beta)$ ***est strictement concave, ce qui garantit l'unicité du maximum de cette fonction. Dans la pratique, ce résultat garantit la convergence des estimateurs du maximum de vraisemblance vers la vraie valeur*** β_0 ***des paramètres, quel que soit le choix des conditions initiales et de l'algorithme d'optimisation utilisé.***

Comme le note Colletaz (2001), il peut toutefois arriver que l'on observe des difficultés dans la progression de l'algorithme vers la solution. Généralement ces difficultés conduisent à l'affichage de valeurs anormalement grandes, en valeur absolue, pour un ou plusieurs des paramètres du modèle. Ceci correspond au **cas de la classification parfaite** dans lequel une ou plusieurs combinaisons de variables explicatives permet de prévoir parfaitement la survenue ou la non survenue de l'événement considéré. Par exemple, considérons le cas où $K > 1$, et si pour une variable explicative notée $z_i = 1$ lorsque $y_i = 1$, alors que $y_i = 1$ ou $y_i = 0$ lorsque $z_i = 0$. Dans ce cas, $Prob(y_i = 1/z_i = 1) = 1$ quelles que soit les valeurs prises par les autres variables explicatives x_i. Cela contraint l'algorithme à donner une valeur extrêmement forte à la combinaison linéaire $\widetilde{\beta} z_i + \beta x_i$, c'est à dire à donner une valeur théoriquement infinie au vecteur $\widetilde{\beta}$, de sorte que l'on rencontre alors des problèmes numériques. Le plus souvent, on observera une valeur estimée de $\widetilde{\beta}$ particulièrement élévée en valeur absolue avec un écart type associé tendant vers la nullité. Pour résoudre ce problème, il suffit la ou les variables concernées ainsi que la totalité des observations parfaitement classées, soit celles associées aux observations telles que $z_i = 1$ et plus généralement aux variables ou aux combinaisons de variables autorisant cette classification parfaite.

2.2. Algorithmes de maximisation de la vraisemblance

Comme nous l'avons vu l'obtention de l'estimateur de maximum de vraisemblance $\widehat{\beta}$ du vecteur de paramètres $\beta \in \mathbb{R}^K$ implique de résoudre un système de K équations non linéaires de la forme :

$$G\left(\widehat{\beta}\right) = \frac{\partial \log L\left(y, \widehat{\beta}\right)}{\partial \widehat{\beta}} = \sum_{i=1}^{N} \frac{\left[y_i - F\left(x_i\widehat{\beta}\right)\right] f\left(x_i\widehat{\beta}\right)}{F\left(x_i\widehat{\beta}\right)\left[1 - F\left(x_i\widehat{\beta}\right)\right]} x_i' = 0 \tag{2.18}$$

avec $F(.) = \Lambda(.)$ dans le cas du logit et $F(.) = \Phi(.)$ dans le cas du probit. **Un tel problème n'admet pas de solution analytique**. *La résolution d'un tel système ne peut se faire qu'en utilisant une procédure d'optimisation numérique. Les algorithmes utilisées dans les principaux logiciels d'économétrie sont généralement*[13] *construit selon l'une ou l'autre de ces deux méthodes : la méthode de Newton Raphson et la méthode du score.* Nous n'évoquerons ici que la méthode de Newton Raphson.

Les méthodes d'optimisation numérique sont utilisées pour maximiser une fonction $f(\theta)$ lorsque la condition du premier ordre $\partial f(\theta)/\partial\theta = 0$ n'admet pas de solution analytique ; le θ optimal doit être déduit par tatônnement ou par un algorithme itératif. Dès lors, un algorithme itératif utilise trois principaux éléments :

1. Des valeurs initiales θ_0 pour amorcer le processus itératif
2. Une règle de passage d'un vecteur θ au suivant
3. Une règle d'arrêt si il y a convergence

****
**** INSERER GRAPHIQUE SUR LA PROCEDURE ****
****

En ce qui concerne le choix des conditions initiales, ce choix est d'autant plus important que le critère à maximiser $f(\theta)$ est complexe. Dans le cas des modèles dichotomiques, on sait que la fonction $f(\theta)$ à maximiser (la vraisemblance ou la log vraisemblance suivant les cas) est globalement concave : dès lors, on est assuré que l'algorithme converge vers la vraie valeur des paramètre, c'est à dire vers la solution[14] unique qui maximise $f(\theta)$, et cela quelles que soient les conditions initiales. Mais même dans ce cas particulièrement favorable, la convergence peut être extrêmement longue si les valeurs de départ sont trop éloignées de l'optimum. Pour les modèles logit et probit, les logiciels usuels considèrent des valeurs initiales pour l'algorithme de maximisation de la vraisemblance égales aux réalisations des estimateurs obtenus dans le modèle linéaire :

$$y_i = x_i\widehat{\beta}_{LP} + \varepsilon_i \quad \beta_0 = \widehat{\beta}_{LP} \tag{2.19}$$

La règle d'arrêt est généralement du type : arrêter le processus itératif si la variation de θ ou du critère $f(\theta)$ entre l'itération actuelle et la précédente est inférieure à une valeur seuil (souvent appelée tolérance).

[13] Sous Eviews et LimDep, la méthode utilisée est celle de Newton-Raphson.
[14] Si cette dernière existe. On admettra l'existence d'un maximum.

Reste à définir la règle de passage d'un vecteur θ au suivant. Une règle de passage consiste à partir des valeurs initiales θ_0, à trouver le prochain vecteur des paramètres θ_1 tel que :

$$f(\theta_1) \geq f(\theta_0)$$

et ainsi de suite à la i^{ème} étape :

$$f(\theta_i) \geq f(\theta_{i-1}) \tag{2.20}$$

Ainsi, on obtient une règle du type :

$$\theta_i = \theta_{i-1} + \lambda_{i-1} D_{i-1} \tag{2.21}$$

où λ_{i-1} désigne le pas à l'itération $i-1$ et D_{i-1} est la direction. D_{i-1} indique la direction que doivent prendre les composantes du nouveau vecteur θ_i et $_{\iota-1}$ indique l'amplitude du saut dans cette orientation. Dans une méthode du gradient, la direction est déterminée par le gradient de la fonction $f(\theta)$. dans le cas $K=1$, si le gradient est positif cela signifie que l'on se situe à gauche de l'optimum : donc on se déplace en augmentant $\theta_i > \theta_{i-1}$. En ce qui concerne le pas, on cherche alors λ_i tel que $\partial f(\theta_i + \lambda_i D_i)/\partial \lambda_i \approx 0$.

La méthode d'optimisation de Newton Raphson est une méthode du gradient[15] *qui est notamment recommandée lorsque le critère à maximiser est globalement concave, ce qui est le cas de la fonction de log vraisemblance dans un modèle dichotomique univarié.* Dans cette méthode, la direction est déterminée par le gradient de la fonction $f(\theta)$, noté $G(\theta)$, tandis que le pas est déterminé par le hessien, noté $H(\theta)$. En effet, cette méthode considère un développement limité de la condition du premier ordre du programme de maximisation de la fonction $f(\theta)$. Soit un point solution θ_i, satisfaisant la condition du premier ordre.

$$\forall i \quad \frac{\partial f(\theta_i)}{\partial \theta} = G(\theta_i) = 0$$

On peut alors donner l'expression d'un développement limité autour de ce point θ_i. Ainsi, pour tout point θ_{i+1}, on obtient la relation suivante au voisinage de θ_i :

$$G(\theta_{i+1}) = G(\theta_i) + \frac{\partial G(\theta_i)}{\partial \theta}(\theta_{i+1} - \theta_i) = 0$$

ou encore :

$$G(\theta_{i+1}) = G(\theta_i) + H(\theta_i)(\theta_{i+1} - \theta_i) = 0$$

On en déduit la relation suivante :

$$\forall i, \quad \theta_{i+1} = \theta_i - H(\theta_i)^{-1} G(\theta_i) \tag{2.22}$$

La méthode de d'optimisation de Newton Raphson ainsi fondé sur cette règle de passage, nécessite le calcul à chaque étape du hessien $H(\theta_i)$.

Proposition 2.6. ***Appliqué au problème de maximisation de la vraisemblance d'un modèle dichotomique, la règle de passage de l'algorithme d'optimisation de Newton Raphson, entre le vecteur d'estimation $\widehat{\beta}_{i-1}$ de la $i-1^{ème}$ itération et vecteur d'estimation $\widehat{\beta}_i$ de la $i^{ème}$ itération est alors définie par la relation :***

$$\widehat{\beta}_i = \widehat{\beta}_{i-1} - \left[\frac{\partial^2 \log L(y,\beta)}{\partial \beta \partial \beta'}\bigg|_{\beta=\widehat{\beta}_{i-1}} \right]^{-1} \left(\frac{\partial \log L(y,\beta)}{\partial \beta}\bigg|_{\beta=\widehat{\beta}_{i-1}} \right) \tag{2.23}$$

[15] Pour un exposé des méthodes du gradient en général voir Alban 2000, pages 49 et suivantes.

ou encore

$$\widehat{\beta}_i = \widehat{\beta}_{i-1} - H\left(\widehat{\beta}_{i-1}\right)^{-1} G\left(\widehat{\beta}_{i-1}\right) \tag{2.24}$$

L'itération est alors arrêté si la variation $\widehat{\beta}_i - \widehat{\beta}_{i-1}$ ou la variation du critère $\log L\left(y, \widehat{\beta}_i\right) - \log L\left(y, \widehat{\beta}_{i-1}\right)$ est inférieure à un certain seuil fixé dans le programme. Le dernier estimateur obtenu $\widehat{\beta}_i = \widehat{\beta}$ correspond alors à l'estimateur optimal du maximum de vraisemblance. Pour être plus précis, il convient de montrer que la suite des $\widehat{\beta}_i$ converge vers l'estimateur du maximum de vraisemblance.

On vérifie immédiatement que si la suite $\widehat{\beta}_i$ converge vers une limite $\widetilde{\beta}$, cette limite est forcement solution des équations de vraisemblance. En effet, si l'on pose $\widetilde{\beta} = \lim_{i \to \infty} \widehat{\beta}_i$, et en considérant la limite des membres de l'égalité (2.24) on a :

$$\widetilde{\beta} = \widetilde{\beta} - H\left(\widetilde{\beta}\right)^{-1} G\left(\widetilde{\beta}\right) \Longleftrightarrow H\left(\widetilde{\beta}\right)^{-1} G\left(\widetilde{\beta}\right) = 0$$

La matrice hessienne étant définie positive strictement, on a bien $G\left(\widetilde{\beta}\right) = \partial \log L\left(y, \widetilde{\beta}\right) / \partial \beta = 0$. Par conséquent, si la suite $\widehat{\beta}_i$ des estimateurs obtenus par l'algorithme de Newton Raphson, convergent vers une quantité $\widetilde{\beta}$, cette quantité est solution des équations du premier ordre du programme de maximisation de la vraisemblance. *Autrement dit, si la suite $\widehat{\beta}_i$ converge, elle converge alors nécessairement vers l'estimateur du maximum de vraisemblance $\widehat{\beta}$ défini par la condition* :

$$\frac{\partial \log L\left(y, \widehat{\beta}\right)}{\partial \beta} = G\left(\widehat{\beta}\right) = 0 \tag{2.25}$$

Reste maintenant à démontrer que l'estimateur du maximum de vraisemblance $\widehat{\beta}$, quel que soit l'algorithme d'optimisation utilisé, converge vers la vraie valeur β des paramètres des modèles logit et probit. Etudions pour cela les propriétés asymptotiques du maximum de vraisemblance.

3. Propriétés Asymptotiques des Estimateurs du Maximum de Vraisemblance

Lorsque l'on cherche à établir les propriétés asymptotiques des estimateurs du maximum de vraisemblance dans le cadre de modèles dichotomiques, et plus généralement dans le cadre de modèle à variables qualitatives, *toute la difficulté réside dans le fait que l'on dispose pas d'expression analytique pour ces estimateurs.* En effet, nous avons vu que les équations de vraisemblance associées au probit et au logit sont non linéaires dans les paramètres. Dès lors, il n'est pas possible alors d'exprimer les estimateurs, solutions de ces équations, comme des fonctions simples des observations. Nous avons vu qu'il était alors nécessaire de recourir à des algorithmes d'optimisation numériques. Mais devant l'impossibilité d'écrire les estimateurs du maximum de vraisemblance comme des fonctions simples des observations, il est alors difficile d'étudier la convergence de ces estimateurs comme nous avions pu le faire dans le cas des modèles linéaires standard. *Il convient ainsi d'adopter une démarche particulière où l'on va chercher à étudier la convergence du critère de maximum de vraisemblance, afin de démontrer la convergence des estimateurs du MV, solutions du programme de maximisation de ce critère.*

Un certain nombre de rappels sur les différentes notions de convergence sont proposés dans l'annexe (A.1). Toutefois, la lecture de ces rappels doit nécessairement s'accompagner d'une étude plus systématique des fondements probabilistes de ces notions[16].

3.1. Convergence du Critères de MV

On considère un modèle dichotomique univarié simple :

$$y_i = \begin{cases} 1 & \text{si } y_i^* \geq 0 \\ 0 & \text{sinon} \end{cases} \tag{3.1}$$

$$y_i^* = x_i\beta_0 + \varepsilon_i \tag{3.2}$$

avec $Prob\left(y_i = 1\right) = F\left(x_i\beta\right)$ où $F\left(.\right)$ désigne la fonction de répartition de ε_i, où $x_i = \left(x_i^1..x_i^K\right)$, $\forall i = 1,..,n$ désigne un vecteur de caractéristiques observables et où $\beta_0 \in \mathbb{R}^K$ est un vecteur de paramètres inconnus. On suppose que l'on dispose d'un échantillon de n individus indicés $i = 1,..,n$.

Nous avons vu précédemment que l'estimateur $\widehat{\beta}$ du maximum de vraisemblance du vecteur de paramètre β_0 dans ce modèle dichotomique est défini par la résolution du système de K équations non linéaires en β. En effet, si l'on pose :

$$\widehat{\beta} = \underset{\{\beta\}}{\arg\max}\left[\log L\left(y,\beta\right)\right] \tag{3.3}$$

où la fonction $\log L\left(y,\beta\right)$ est définie par l'équation (2.3) :

$$\log L\left(y,\beta\right) = \sum_{i=1}^{n} y_i \log\left[F\left(x_i\beta\right)\right] + \left(1-y_i\right)\log\left[1-F\left(x_i\beta\right)\right] \tag{3.4}$$

[16] Voir par exemple, "Méthodes Statistiques", Philippe Tassi, *Economica* 1989

on vérifie que la condition nécessaire de ce programme s'écrit :

$$\frac{\partial \log L\left(y,\widehat{\beta}\right)}{\partial \widehat{\beta}} = \sum_{i=1}^{n} \frac{\left[y_i - F\left(x_i\widehat{\beta}\right)\right] f\left(x_i\widehat{\beta}\right)}{F\left(x_i\widehat{\beta}\right)\left[1 - F\left(x_i\widehat{\beta}\right)\right]} x_i' = G\left(\widehat{\beta}\right) = 0 \tag{3.5}$$

où $G(\beta)$ désigne le gradient associé à la log-vraisemblance $\partial \log L(y,\beta)$, évalué au point $\widehat{\beta}$. On trouve alors un système de K équations non linéaires.

Ainsi, nous ne pouvons pas obtenir d'expression analytique de l'estimateur $\widehat{\beta}$ du maximum de vraisemblance. Dès lors, la question qui se pose est de savoir comment montrer que l'estimateur $\widehat{\beta}$ est convergent. Autrement dit, il s'agit de savoir comment établir le résultat suivant ?

$$\widehat{\beta} \xrightarrow[n\to\infty]{p} \beta_0 \tag{3.6}$$

où β_0 désigne la "varie" valeur des paramètres β. En effet, tout le problème consiste à établir une propriété de convergence de l'estimateur sans disposer d'une expression analytique de celui-ci. Tout ce que l'on sait pour l'instant, c'est que si la fonction de log-vraisemblance dans les modèles logit et probit admet un maximum, ce maximum est unique, puisque nous avons montré que la fonction $\log L(y,\beta)$ est dans ces deux cas concave.

3.1.1. Convergence d'estimateurs dans les modèles non linéaires

Pour résoudre ce problème, nous allons tout d'abord exposer une méthode générale permettant d'établir la convergence d'estimateur dans des modèles non linéaires. Considérons le problème suivant. On cherche à minimiser en θ un critère $C_n(y,\theta)$:

$$\min_{\{\theta\}} C_n(y,\theta) \tag{3.7}$$

$$\text{sous } \theta \in \Theta \tag{3.8}$$

Ce critère $C_n(y,\theta)$ peut être soit celui somme des carrés des résidus (critère des MCO), soit celui de la somme des carrés pondérés (critère des MCG), etc.. De façon générale, ce critère correspond à la classe des **M-estimateurs**. Soit θ_0 le vrai vecteur de paramètres permettant de minimiser le critère et soit y un vecteur de variables endogènes observables. On considère un M-estimateur quelconque noté $\widehat{\theta}_n$ défini par :

$$\widehat{\theta}_n = \arg\min_{\{\theta\}} \left[C_n(y,\theta)\right] \tag{3.9}$$

On cherche alors à établir que cet estimateur est convergent et cela sans spécifier le critère $C_n(y,\theta)$. La convergence de $\widehat{\theta}_n$ se traduit par la relation :

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{p.s.} \theta_0 \tag{3.10}$$

Pour établir ce résultat on a besoin de faire trois hypothèses :

Hypothèse 1 $\theta \in \Theta$, $\Theta \in \mathbb{R}^K$ *compact.*

Hypothèse 2 *Le critère $C_n(y,\theta)$ converge presque sûrement et uniformément par rapport à θ vers une fonction $C_\infty(\theta,\theta_0)$*

$$C_n(y,\theta) \underset{n\to\infty}{\overset{p.s.}{\longrightarrow}} C_\infty(\theta,\theta_0) \tag{3.11}$$

Hypothèse 3 *La fonction $C_\infty(\theta,\theta_0)$ admet un minimum unique en $\theta=\theta_0$:*

$$\forall\theta\in\Theta,\ \ C_\infty(\theta_0,\theta_0) \leq C_\infty(\theta,\theta_0)$$

L'idée de la démonstration du résultat (3.10) est alors la suivante. *On considère la suite des estimateurs $\left\{\widehat{\theta}_n\right\}$ définie sur un ensemble compact. On sait que toute suite définie sur un ensemble compact admet au moins une valeur limite. Soit θ^L une des valeurs d'adhérence de la suite $\left\{\widehat{\theta}_n\right\}$. Il suffit alors de montrer que cette valeur d'adhérence est unique et correspond à la vraie valeur θ_0 des paramètres du modèle.*

Soit θ^L une des valeurs d'adhérence particulière de la suite $\left\{\widehat{\theta}_n\right\}$. Il existe alors une sous suite $\left\{\widehat{\theta}_n^L\right\}$ qui converge vers θ^L.

$$\widehat{\theta}_n^L \underset{n\to\infty}{\overset{p.s.}{\longrightarrow}} \theta^L$$

Sachant que le M-estimateur $\widehat{\theta}_n$ minimise le critère $C_n(y,\theta)$, on a par construction $C_n\left(y,\widehat{\theta}_n\right) \leq C_n(y,\theta)$, $\forall\theta\in\Theta$. Ce résultat vaut aussi pour la sous suite $\widehat{\theta}_n^L$. Par conséquent :

$$\forall\theta\in\Theta \quad C_n\left(y,\widehat{\theta}_n^L\right) \leq C_n(y,\theta)$$

Cette inégalité est en particulier valable pour la valeur $\theta_0\in\Theta$:

$$C_n\left(y,\widehat{\theta}_n^L\right) \leq C_n(y,\theta_0) \tag{3.12}$$

Considérons à présent la limite en probabilité des termes de droite et de gauche de cette inégalité. Pour cela, on utilise le résultat de convergence suivant :

$$\begin{array}{c} f_n(.) \underset{n\to\infty}{\overset{p.s.}{\longrightarrow}} f(.) \\ x \underset{n\to\infty}{\overset{p.s.}{\longrightarrow}} x_0 \end{array} \Longrightarrow f_n(x_n) \underset{n\to\infty}{\overset{p.s.}{\longrightarrow}} f(x_0)$$

Sachant que $\left\{\widehat{\theta}_n^L\right\}$ converge vers θ^L, et que sous l'hypothèse 2 le critère $C_n(y,\theta)$ converge vers $C_\infty(\theta,\theta_0)$, on montre que la limite en probabilité du terme de gauche de l'inégalité (3.12) peut s'écrire sous la forme suivante :

$$C_n\left(y,\widehat{\theta}_n^L\right) \underset{n\to\infty}{\overset{p.s.}{\longrightarrow}} C_\infty\left(\widehat{\theta}^L,\theta_0\right) \tag{3.13}$$

De la même façon, on montre que le terme de droite de l'inégalité (3.12) converge en probabilité vers la quantité suivante :

$$C_n(y,\theta_0) \underset{n\to\infty}{\overset{p.s.}{\longrightarrow}} C_\infty(\theta_0,\theta_0) \tag{3.14}$$

Dès lors on obtient l'inégalité suivante définie sur les limites des critères :

$$C_\infty\left(\widehat{\theta}^L,\theta_0\right) \leq C_\infty(\theta_0,\theta_0) \tag{3.15}$$

Sachant que sous l'hypothèse 3, θ_0 *est la seule valeur qui assure le minimum global de la fonction* $C_\infty(\theta, \theta_0)$, c'est à dire que $\forall \theta \in \Theta$ on a $C_\infty(\theta_0, \theta_0) \leq C_\infty(\theta, \theta_0)$, on en conclut que θ^L correspond nécessairement à θ_0 :

$$\theta^L = \theta_0 \tag{3.16}$$

En d'autres termes, la sous suite $\left\{\widehat{\theta}_n^L\right\}$ converge vers la vraie valeur θ_0 des paramètres. Donc par conséquent, la suite $\left\{\widehat{\theta}_n\right\}$ converge elle aussi vers la vraie valeur θ_0 des paramètres.

$$\widehat{\theta}_n \xrightarrow[n\to\infty]{p.s.} \theta_0 \tag{3.17}$$

On ainsi réussi à démontrer la convergence de notre M-estimateur $\widehat{\theta}_n$ vers la vraie valeur des paramètres θ_0. Appliquons à présent cette méthode dans le cas de l'estimateur du maximum de vraisemblance dans le cadre des modèles dichotomiques univariés.

3.1.2. Application aux modèles Logit et Probit

Dans le cas d'un modèle dichotomique simple (logit ou probit), l'estimateur $\widehat{\beta}_n$ (noté aussi $\widehat{\beta}$) du maximum de vraisemblance du vecteur de paramètre β est défini par la maximisation d'un critère $C_n(y, \beta_0)$ qui correspond, bien évidemment à la log vraisemblance du modèle (équation 2.3) :

$$\widehat{\beta}_n = \arg\max_{\{\beta\}} C_n(y, \beta) \tag{3.18}$$

où l'on pose[17]

$$C_n(y, \beta) = \frac{1}{n} \log L(y, \beta) = \frac{1}{n} \sum_{i=1}^{n} y_i \log\left[F(x_i\beta)\right] + (1 - y_i) \log\left[1 - F(x_i\beta)\right] \tag{3.19}$$

où $F(.)$ désigne une fonction de répartition. On note β_0 la vraie valeur des paramètres. On suppose que l'hypothèse 1 est vérifiée, c'est à dire que $\beta \in \Theta,\ \Theta \in \mathbb{R}^K$ compact. Reste à établir que les hypothèses 2 et 3 sont valides.

Montrons que tout d'abord que le critère $C_N(y, \theta)$ converge presque sûrement et uniformément par rapport à θ vers une fonction $C_\infty(\theta, \theta_0)$, c'est à dire que :

$$C_n(y, \theta) \xrightarrow[n\to\infty]{p.s.} C_\infty(\theta, \theta_0)$$

Dans notre cas, on sait que

$$\begin{aligned} C_n(y, \beta) &= \frac{1}{n} \sum_{i=1}^{n} y_i \log F(x_i\beta) + (1 - y_i) \log\left[1 - F(x_i\beta)\right] \\ &= \frac{1}{n} \sum_{i=1}^{n} y_i \log F(x_i\beta) + \frac{1}{n} \sum_{i=1}^{n} (1 - y_i) \log\left[1 - F(x_i\beta)\right] \end{aligned} \tag{3.20}$$

Etudions la convergence des différents éléments de cette somme. On suppose que les variables x_i sont aléatoires. Sous certaines hypothèse de régularités, on sait que :

$$\frac{1}{n} \sum_{i=1}^{n} y_i \log F(x_i\beta) \xrightarrow[n\to\infty]{p} E\left\{y_i \log F(x_i\beta)\right\}$$

[17] Afin de simplifier les calculs, on pose que $C_n(y, \beta) = (1/N) \log L(y, \beta)$. On aurait pu assimiler le critère directement à la log vraisemblance. Quoiqu'il en soit ces deux définitions du critère laissent inchangée la définition de l'estimateur du maximum de vraisemblance $\widehat{\beta}$.

$$\frac{1}{n}\sum_{i=1}^{n}(1-y_i)\log\left[1-F\left(x_i\beta\right)\right]\xrightarrow[n\to\infty]{p}E\left\{(1-y_i)\log\left[1-F\left(x_i\beta\right)\right]\right\}$$

Or si l'on note E_x l'espérance conditionnelle à x_i, on a :

$$\begin{aligned}E\left[y_i\log F\left(x_i\beta\right)\right] &= E_x\left\{E\left[y_i\log F\left(x_i\beta\right)\right]\ /\ x_i\right\}\\ &= E_x\left[E\left(y_i/x_i\right).\log F\left(x_i\beta\right)\right]\end{aligned}$$

en appliquant la loi de Bayes, on sait que :

$$h\left(y,\theta\right)=f\left(y/\theta\right)g\left(\theta\right)=g\left(\theta/y\right)f\left(y\right) \tag{3.21}$$

où $h\left(.\right)$ désigne la densité jointe de y et de θ, et où $f\left(.\right)$ et $g\left(.\right)$ désignent suivant les cas les densités marginales et conditionnelles des v.a.r. y et θ. On en déduit le *théorème de Bayes* :

$$g\left(\theta/y\right)=\frac{f\left(y/\theta\right)g\left(\theta\right)}{f\left(y\right)} \tag{3.22}$$

****
**** Finir Demonstration *********
****

Donc finalement, on a :

$$\frac{1}{n}\log\ L\left(y,\beta\right)\xrightarrow[n\to\infty]{p.s.}\sum_{i=1}^{n}F\left(-x_i\beta_0\right)\log\left[F\left(x_i\beta_0\right)\right]+\left[1-F\left(-x_i\beta_0\right)\right]\log\left[1-F\left(x_i\beta_0\right)\right]=L_\infty\left(y,\beta\right)$$

3.2. Lois et variance asymptotiques de l'estimateur de MV

Nous avons vu précédemment que la fonction de vraisemblance des échantillons associés aux modèles logit et probit était concave. Par conséquent, si la solution des équations de vraisembalance existe, cette solution est unique et correspond bien au maximum de la focntion de log vraisemblance. Nous avons vu en outre, dans la section précédente, que sous certaines conditions, l'estimateur du maximum de vraisemblance ainsi obtenu est convergent. Dès lors, nous allons à présent nous intéresser à la loi asymptotique de ce estimateur ainsi qu'à sa variance asymptotique.

Pour garantir à la fois la convergence et la normalité asymptotique des estimateurs du maximum de vraisemblance dans les modèles logit et probit, un certain nombre de conditions doivent être validées (cf. Amemiya 1985, Greene 1997). Deux approches sont retenues suivant que l'on suppose que les variables explicatives sont des variables aléatoires continues ou des variables déterministes. Dans le cas de variables explicatives aléatoires continues, les conditions se ramènent à imposer l'indépendance des x_i, la même distribution pour tous les x_i $i=1,..N$, en admettant l'existence de moments d'ordre suffisant (Amameyia 1976). Dans le cas de variables explicatives déterministes, les conditions imposent alors aux valeurs x_i d'être bornées : $\exists m>0$ et $\exists M<\infty$, tels que $m<\left|x_i^k\right|<M$, $\forall k\in\mathbb{R}$, $\forall i=1,..,N$, et cela de sorte à assurer que la matrice de variance covariance asymptotique existe (Gourieroux et Monfort 1981). Nous supposerons ici que nous avons des variables explicatives aléatoires et que les conditions correspondantes sont satisfaites.

Proposition 3.1. ***Sous certaines conditions, l'estimateur du maximum de vraisemblance*** $\widehat{\beta}$ ***est convergent et suit asymptotiquement une loi normale de moyenne égale à la vraie valeur*** β_0 ***des paramètres et de matrice de variance covariance égale à l'inverse de la matrice d'information de Fischer*** $I(\beta_0)$ ***évaluée au point*** β_0 ***:***

$$\sqrt{N}\left(\widehat{\beta}-\beta_0\right) \underset{N\to\infty}{\xrightarrow{L}} N\left[0, I\left(\beta_0\right)^{-1}\right] \tag{3.23}$$

avec

$$I(\beta_0) = -E\left[\frac{\partial^2 \log L(y,\beta)}{\partial\beta\partial\beta'}\right]_{\beta=\beta_0} = \sum_{i=1}^{N} \frac{f^2(x_i\beta_0)}{F(x_i\beta_0)\left[1-F(x_i\beta_0)\right]} x_i' x_i \tag{3.24}$$

Nous avons vu précédemment que la matrice d'information de Fischer peut se simplifier notamment dans le cas du modèle logit. En effet, dans le cas où $F(.) = \Lambda(.)$, on a :

$$I(\beta) = \sum_{i=1}^{N} \lambda(x_i\beta)\, x_i' x_i = \sum_{i=1}^{N} \frac{\exp(x_i\beta)}{\left[1+\exp(x_i\beta)\right]^2} x_i' x_i$$

Dans le cas du modèle probit, il n'y a pas de simplification particulière.

$$I(\beta) = \sum_{i=1}^{N} \frac{\phi^2(x_i\beta)}{\Phi(x_i\beta)\left[1-\Phi(x_i\beta)\right]} x_i' x_i$$

L'idée de la démonstration[18] de cette propososition est la suivante. Si l'on note $G(\beta) = \partial \log L(.)/\partial\beta$ le vecteur de gradient et $H(\beta) = \partial^2 \log L(.)/\partial\beta\partial\beta'$ la matrice hessienne, on sait que l'estimateur du maximum de vraisemblance satisfait la condition du premier ordre $G\left(\widehat{\beta}\right) = 0$. Considérons un developpement limité à l'ordre 1 autour de cette condition autour de la vraie valeur des paramètres β_0. En ometant les termes de degré supérieurs à 2, il vient :

$$G\left(\widehat{\beta}\right) = G(\beta_0) + H(\beta_0)\left(\widehat{\beta}-\beta_0\right) = 0$$

En prémultipliant cette égalité par $H(\beta_0)^{-1}$, on obtient $\left(\widehat{\beta}-\beta_0\right) = -H(\beta_0)^{-1} G(\beta_0)$, ce qui peut se réecrire sous la forme :

$$\sqrt{N}\left(\widehat{\beta}-\beta_0\right) = -\left[\frac{1}{N}H(\beta_0)\right]^{-1}\left[\sqrt{N}\overline{g}(\beta_0)\right]$$

où le vecteur $\overline{g}(\beta_0)$ de dimension $(K,1)$ est défini par :

$$\underset{(K,1)}{\overline{g}(\beta_0)} = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^{N} \partial \log L(y_i,\beta)/\partial\beta_1 \\ .. \\ \sum_{i=1}^{N} \partial \log L(y_i,\beta)/\partial\beta_{K-1} \\ \sum_{i=1}^{N} \partial \log L(y_i,\beta)/\partial\beta_K \end{pmatrix} \tag{3.25}$$

En supposant que chaque composante $(1/N)\sum_{i=1}^{N} \partial \log L(y_i,\beta)/\partial\beta_1$ est *i.i.d*, on alors appliquer le théorème central limite à $\overline{g}(\beta_0)$. Parallèlement, si l'on applique une loi des grands nombres à $H(\beta_0)/N$, on montre finalement que la quantité $\sqrt{N}\left(\widehat{\beta}-\beta_0\right)$ a une distribution normale de moyenne 0 et de matrice de variance covariance $-E\left[H(\beta_0)\right]$.

[18] Pour une dsitribution rigorueuse voir le cours de A. Holly (1999).

Une remarque doit être faite ici concernant la matrice de variance covariance asympotique de $\widehat{\beta}$, notée $V_{as}\left(\widehat{\beta}\right) = I\left(\beta_0\right)^{-1}$. Naturellement, cette matrice de variance covariance dépend de la vraie valeur du paramètre β_0 qui est par définition inconnue. Dès lors, on retient généralement comme estimateur de la matrice de variance covariance asympotique la matrices $I\left(\widehat{\beta}\right)^{-1}$ dans laquelle la vraie valeur des paramètres β_0 a été remplacée par son estimateur $\widehat{\beta}$.

$$\widehat{V}_{as}\left(\widehat{\beta}\right) = I\left(\widehat{\beta}\right)^{-1} = \left[-E\left(\frac{\partial^2 \log L\left(y,\beta\right)}{\partial\beta\partial\beta'}\right)_{\beta=\widehat{\beta}}\right]^{-1} \tag{3.26}$$

4. Méthodes d'Estimation non Paramétriques

Un des problèmes qui peut se poser lors de la phase d'estimation des paramètres des modèles dichotomiques[19] par maximum de vraisemblance provient de l'hypothèse que l'on fait sur la distribution des résidus du modèle. Considérons le modèle dichotomique suivant :

$$y_i = \begin{cases} 1 & \text{si } y_i^* = x_i\beta_0 + \varepsilon_i \geq 0 \\ 0 & \text{sinon} \end{cases}$$

où ε_i est une perturbation *i.i.d.* $\left(0, \sigma_\varepsilon^2\right)$. Lorsque l'on cherche à estimer les paramètres β_0 par maximum de vraisemblance, on postule une certaine distribution pour les termes ε_i. On considère par exemple une distribution logistique dans le cas d'un modèle logit et une distribution normale dans le cas d'un probit. *Or, rien ne garantit a priori que cette distribution que l'on utilise pour construire la vraisemblance de l'échantillon corresponde réellement à la "vraie" distribution des perturbations* ε_i. Naturellement, une erreur sur la distribution des termes ε_i conduit alors nécessairement à une estimation du maximum de vraisemblance non efficace des paramètres β_0.

Une des solutions pour se prémunir contre ce risque de mauvaise spécification de la loi des perturbations du modèle, consiste tout à s'afranchir de toute de hypothèse sur la distribution paramétrique des résidus dans la phase d'estimation des paramètres β_0. On parle alors **de méthodes d'estimation non paramètriques**. Nous ne présenterons ici que les méthodes du score maximum et une méthode semi-paramétrique (Alban 2000).

4.1. La méthode du score maximum

Commençons par définir l'estimateur du score maximum.

Définition L'estimateur du score maximum est obtenu par la maximisation, par rapport au vecteur $\beta \in \mathbb{R}^K$, d'un critère constitué du nombre de fois où $x_i\beta > 0$ lorsque $y_i = 1$ et du nombre de fois où $x_i\beta < 0$ lorsque $y_i = 0$:

$$\widehat{\beta}_s = \underset{\{\beta\}}{\arg\max} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{y_i=1}\mathbb{I}_{x_i\beta>0} + \mathbb{I}_{y_i=0}\mathbb{I}_{x_i\beta<0} \tag{4.1}$$

où $\mathbb{I}_x$ désigne la fonction indicatrice.

L'idée générale de cette méthode est la suivante. On sait que la probabilité associée à l'événment $y_i = 1$ est définie par $p_i = Prob\left(\varepsilon_i < x_i\beta\right) = F\left(x_i\beta\right)$. En d'autres termes, on a $y_i = 1$ quand l'inégalité $\varepsilon_i < x_i\beta$ est vérifiée. Si l'on considère à présent des valeurs de ε_i suffisament faibles relativement à $x_i\beta$, cette relation peut être approximée de la façon suivante $x_i\beta - \varepsilon_i \simeq x_{i\beta} > 0$. Ainsi, on doit observer $y_i = 1$ quand $x_i\beta$ est positif, si tant ait que l'on dispose de la vraie valeur β_0 du vecteur β. Parralèment, on doit observer $y_i = 0$ quand $x_i\beta$ est négatif. En termes de probabilités on obtient les approximations suivantes :

$$Prob\left(y_i = 1\right) \simeq Prob\left(x_i\beta > 0\right)$$

[19] Problème qui n'est pas spécifique aux modèles à variable explicative dichotomique.

$$Prob\left(y_i = 0\right) \simeq Prob\left(x_i\beta \leq 0\right)$$

Le critère du score maximum consiste alors à maximiser en β la fréquence empirique (le score) des évements $(y_i = 1)$ et $(x_i\beta > 0)$.

Une autre interprétation de la méthode du score est qu'elle compare le signe de la prédiction, c'est à dire le signe de $x_i\beta$, avec celui de la variable transformée $\delta_i = 2y_i - 1$ qui prend la valeur -1 quand $y_i = 0$ et la valeur 1 quand $y_i = 1$. On compare donc une valeur observée δ_i qui est positive quand l'événement $y_i = 1$ se réalise avec la quantité $x_i\beta$, qui pour la vraie valeur β_0 du vecteur β, doit elle aussi être positive quand l'événement $y_i = 1$ se réalise. Ainsi, le critère du score maximum peut s'écrire sous la forme :

$$\widehat{\beta}_s = \arg\max_{\{\beta\}} \frac{1}{N}\sum_{i=1}^{N} \delta_i \text{sgn}\left(x_i\beta\right) \tag{4.2}$$

où la fonction $\text{sgn}(z)$ est définie de la façon suivante :

$$\text{sgn}\left(z\right) = \begin{cases} 1 & \text{si } z > 0 \\ 0 & \text{si } z = 0 \\ -1 & \text{si } z < 0 \end{cases}$$

Le principal avantage de cette méthode du score maximum est qu'elle ne nécessite aucune hypothèse sur la distribution des résidus ε_i. Mais cet avantage constitue en outre sa principale limite. *En effet, puisque l'on ne construit aucune vraisemblance pour obtenir l'estimateur $\widehat{\beta}_s$ et puisque le critère à maximiser n'est pas continument différentiable, le calcul des principales statistiques de tests sur cet estimateur ne peut pas se faire avec les techniques usuelles.* Par exemple, les écarts types associés au vecteur $\widehat{\beta}_s$ ne peuvent pas être calculés à partir des formules usuelles, fondées par exemple sur la dérivée seconde d'une fonction critère continue (fonction de log-vraisemblance dans le cas de l'estimateur du MV). Une possibilité consiste à calculer les estimateurs des variances des estimateurs $\widehat{\beta}_s$ par des méthodes de *bootstrap* (Greene 1997).

Ainsi, l'information fournie par la méthode du score minimum est limitée, et de plus l'estimateur $\widehat{\beta}_s$ est généralement inefficace par rapport à l'estimateur du maximum de vraisemblance. De plus, son exploitation est elle aussi très limitée : il n'est par exemple pas possible de calculer les effets marginaux associées aux variables explicatives sans postuler une hypothèse sur la distribution $F\left(.\right)$. De plus, le fait de ne pas imposer de dsitribution a priori n'assure aucunement que l'estimation sera plus précise ou que les prévisions seront plus satisfaisantes. C'est pour ces raisons que se sont développées des méthodes intermédiaires : *les méthodes d'estimation semi-paramétrique.*

4.2. Estimation semi-paramétrique

L'idée des méthodes semi-paramétrique dans ce contexte (Klein et Spady 1993) consiste tout simplement *à séparer le modèle en deux : une partie paramétrique correspondant au scalaire $x_i\beta$ et une partie non paramétrique correspondant à la fonction de répartition $F\left(.\right)$.*

Dans un modèle dichotomique simple, nous avons vu que l'on l'égalité $p_i = E\left(y_i\right)$ dès lors que le modèle s'écrit sous la forme $p_i = Prob\left(y_i = 1\right)$. De façon plus précise, on obtient donc

l'égalité suivante :

$$p_i = E\left(y_i | x_i\right) = F\left(x_i\beta\right) \tag{4.3}$$

Ainsi, décrire l'espérance conditionnelle de y_i sachant x_i revient en fait à décrire la fonction de répartition $F(.)$, que l'on cherche à maximiser en β. On définit $r\left(x_i\right)$, appelée **fonction de lien**, cette espérance conditionnelle :

$$r\left(x_i\right) = E\left(y_i | x_i\right) = \int_{-\infty}^{\infty} y_i \frac{f\left(x_i, x_i\right)}{f\left(x_i\right)} dy_i \tag{4.4}$$

La démarche est alors la suivante : on cherche dans un premier temps à estimer la fonction de lien $r(z)$, qui n'est autre que la fonction de répartition $F(z)$. Une fois que l'on dispose d'un estimateur de $F(z)$, noté $\widehat{F}(z)$, en tout point z, il suffit d'écrire la log-vraisemblance de l'échantillon en fonction de la loi estimée $\widehat{F}\left(x_i\beta\right)$, et de maximiser cette quantité par rapport à β pour obtenir un estimateur $\widehat{\beta}_{sp}$.

Comment estimer cette fonction de lien, qui correspond en fait la fonction de répartition $F(z)$? On utilise ici une méthode non paramétrique fondateur sur un estimateur à noyau. Sans le démontrer, on admettra le résultat suivant :

Proposition 4.1. ***La probabilité associée à l'observation y_i en tout point x_i^0 peut être estimée par la moyenne pondérée :***

$$\widehat{r}\left(x_i^0\right) = \frac{\left[\sum_{i=1}^{N} w_i\left(x_i^0\right) y_i\right]}{\left[\sum_{i=1}^{N} w_i\left(x_i^0\right)\right]} \tag{4.5}$$

où la pondération $w_i\left(x_i^0\right)$ est définie par la relation :

$$w_i\left(x_i^0\right) = K\left(\frac{x_i - x_i^0}{h}\right) \tag{4.6}$$

où $K(.)$ désigne un opérateur noyau et h une fenêtre.

Ainsi, cette proposition nous permet de reconstruire toute la fonction de répartition $F\left(x_i\right)$ en appliquant la formule (4.5) pour chaque observations x_i, $i = 1, .., N$. On dispose alors d'une suite de N réalisations d'un estimateur $\widehat{F}\left(x_i\beta\right)$ pour une valeur donnée du vecteur β.

L'opérateur noyau $K(.)$, ou *kernel*, fournit une mesure de la distance entre le point considéré x_{i0} et n'importe quel autre point x_i de l'échantillon. Plus la distance est importante, plus l'on attribue une faible valeur à la pondération, donc plus la valeur du kernel est faible. C'est une fonction continue, symétrique autour de zéro, intégrant à 1, et nulle pour de grandes valeurs de son argument. Les fonctions kernel les plus souvent utilisées sont les suivants :

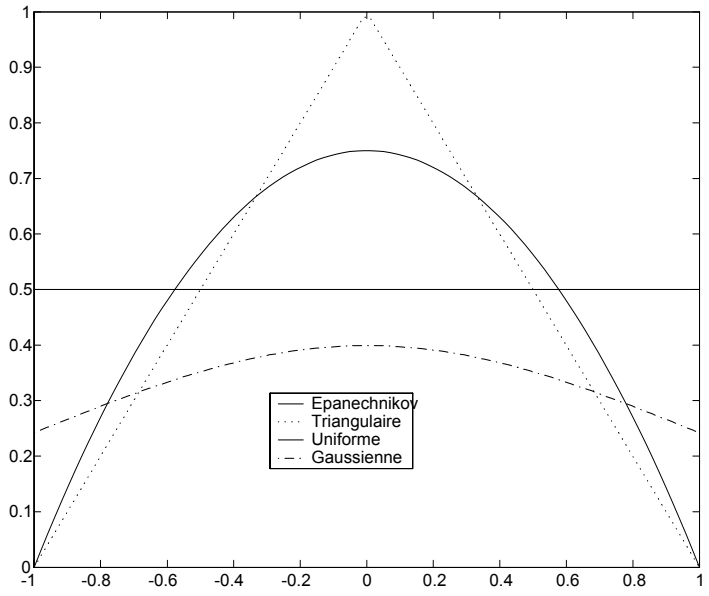
Sur la figure (4.1) ont été reportées les valeurs de ces différentes fonctions, ce qui permet de visualer la décroissance du poids accordé aux observations éloignées du point central x_i^0.

Le paramètre h de la pondération (4.6) est appelé fenêtre (ou *bandwidth parameter*) sert à calibrer la distance entre x_i et x_{i0}, en pénalisant plus ou moins les poids éloignés de x_{i0}. Plus h est petit, plus l'opérateur $w_i\left(x_{i0}\right)$ privilégie les points proches de x_{i0}. Un exemple de valeur de la fenêtre correspond à $h = 0.15\left(x_v - x_u\right)$ où $x_v - x_u$ désigne l'écart maximal entre les observations (*upper* moins *lower*). Naturellement, il convient d'évaluer l'impact de ce choix sur l'estimateur de β en faisant varier h.

Tableau 4.1: Définition des Principales Fonctions Kernel

Noyau	Définition
*G*aussien	$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$
Epanechnikov	$K(x) = \frac{3}{4}\left(1 - x^2\right).\mathbb{I}_{\lvert x\rvert \leq 1}$
Triangulaire	$K(x) = (1 - \lvert x\rvert).\mathbb{I}_{\lvert x\rvert \leq 1}$
Uniforme	$K(x) = \frac{1}{2}.\mathbb{I}_{\lvert x\rvert \leq 1}$

Figure 4.1: Fonctions Kernel $K(x)$ Usuelles



Remarque Dans le cas des estimateurs semi-paramétriques, le choix de la fenêtre h permet d'arbitrer entre le biais de l'estimateur non paramétrique et sa variance. Une fenêtre petite fournira un biais plus faible mais un estimateur moins efficace (de plus grande variance), alors qu'une fenêtre plus large s'approchera de l'estimation par les moindres carrés linéaires dans lesquels tous les points sont pris en compte avec la même pondération.

En résumé, l'approche semi-paramétrique consiste à construire un estimateur à noyau de la vraisemblance évalué pour une valeur quelconque de β, et à maximiser cette fonction pour obtenir l'estimateur semi-paramétrique noté $\widehat{\beta}_s$. La construction de l'estimateur à noyau de fonction de log-vraisemblance se réalise de la façon suivante. Pour une valeur quelconque $\beta \in \mathbb{R}^K$, les étapes de la construction sont les suivantes :

1. **Première étape :** On estime pour le premier individu ($i = 1$), la fonction de lien au

voisinage du point $z_1^0 = x_1\beta$ pour la valeur retenue de β.

$$\widehat{r}\left(z_1^0\right) = \frac{\left[\sum_{i=1}^{N} w_i\left(z_1^0\right) y_i\right]}{\left[\sum_{i=1}^{N} w_i\left(z_1^0\right)\right]} \quad \text{avec } w_i\left(z_1^0\right) = K\left(\frac{z_i - z_1^0}{h}\right)$$

Dans le cas d'une fonction kernel gaussienne, on a par exemple $\forall i = 1, .., N$:

$$\begin{aligned} K\left(\frac{z_i - z_1^0}{h}\right) &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{z_i - z_1^0}{h}\right)^2\right] \\ &= \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_i\beta - x_1\beta}{h}\right)^2\right] \end{aligned}$$

On obtient ainsi une estimation de la fonction de répartition $\widehat{F}\left(z_1^0\right) = \widehat{r}\left(z_1^0\right)$ au point $z_1^0 = x_1\beta$. On répète alors l'opération pour les N individus, $j = 1, .., N$, et ce faisant on obtient N réalisations d'un estimateur à noyau $\widehat{F}\left(x_j\beta\right)$ de la fonction de répartition $F\left(.\right)$ évaluée aux N points $z_j^0 = x_j\beta$ conditionnellement à la valeur β.

$$\forall j = 1, .., N \quad \widehat{r}\left(z_j^0\right) = \frac{\left[\sum_{i=1}^{N} w_i\left(z_j^0\right) y_i\right]}{\left[\sum_{i=1}^{N} w_i\left(z_j^0\right)\right]} \quad avec\ w_i\left(z_j^0\right) = K\left(\frac{z_i - z_j^0}{h}\right)$$

2. **Deuxième étape :** A partir des N réalisations de l'estimateur à noyau $\widehat{F}\left(x_j\beta\right)$ on construit un estimateur de la fonction de log vraisemblance du modèle associée à l'échantillon $y = (y_1, y_2, ..., y_N)$:

$$\log \widehat{L}\left(y, \beta\right) = \sum_{i=j}^{N} y_j \log\left[\widehat{F}\left(x_j\beta^0\right)\right] + \left(1 - y_j\right) \log\left[1 - \widehat{F}\left(x_j\beta^0\right)\right] \tag{4.7}$$

On peut ainsi finallement obtenir une valeur estimée de la log-vraisemblance $\log \widehat{L}\left(y, \beta\right)$ pour toute valeur du vecteur $\beta \in \mathbb{R}^K$.

Il ne reste plus alors qu'à maximiser la fonction $\log \widehat{L}\left(y, \beta\right)$ en β. Pour cela on utilisera une procédure numérique d'optimisation (par exemple une méthode du gradient Newton Raphson) qui à partir d'une condition initiale sur β permettra d'obtenir l'estimateur semi-paramétrique $\widehat{\beta}_s$:

$$\widehat{\beta}_s = \underset{\{\beta\}}{\arg\max} \left\{\log \widehat{L}\left(y, \beta\right)\right\} \tag{4.8}$$

Généralement, la condition initiale choisie dans les algorithmes d'optimisation, notée β^0, correspond à un estimateur simple comme par exemple l'estimateur du score maximum ou un estimateur des MCO :

$$\beta^0 = \widehat{\beta}_s = \underset{\{\beta\}}{\arg\max} \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{y_i=1}\mathbb{I}_{x_i\beta>0} + \mathbb{I}_{y_i=0}\mathbb{I}_{x_i\beta<0}$$

4.3. Comparaison des estimateurs paramétriques, non paramétriques et semi paramétriques

****
**** INSERER Programme Matlab et Résultats *********
****

5. Tests de Spécification et Inférence

Comment tester le modèle dichotomique ? Comment tester les paramètres de ce modèle ? Autant de questions auxquelles nous allons à présent tacher de répondre. Nous commencerons par évoquer les tests d'hypothèse sur les coefficients, puis dans une seconde sous section nous envisagerons les principaux tests de spécification sur les modèles dichotomiques.

5.1. Tests d'hypothèse sur les paramètres

Les différentes méthodes d'estimation présentées précédemment conduisent à des estimateurs asymptotiquement normaux lorsque le nombre d'observations tend vers l'infini. Il est donc facile d'utiliser ces divers estimateurs pour construire des procédures de tests dont certaines seront asymptotiquement équivalentes. Nous présentarons ici les principales procédures de test à partir de la méthode d'estimation du maximum de vraisemblance qui est la plus souvent utilisée. On retouve alors la trilogie :

1. Test de Wald
2. Test du score ou multiplicateur de Lagrange : LM (*Lagrange Mulitplier*)
3. Test du rapport des maxima de vraisemblance : LRT (*Likelihood Ratio Test*)

On rappelle que ces trois tests sont asymptotiquement équivalents, ce qui implique qu'ils peuvent notamment se contredire sur poetits échantillons. De plus, leur distribution n'étant valide qu'asympotitquement, il convient d'être prudent dans leur utilisation sur de petits échantillons. On sait en outre que le test LRT est localement le plus puissant et que donc il devrait être a priori préféré. Nous n'envisagerons ici que le cas d'un test bidirectionnel[20] sur un coefficient ou sur un ensemble de coefficients.

5.1.1. Test de Wald

On considère le test $H_0 : \beta_j = a$ contre $H_1 : \beta_j \neq a$ où β_j désigne la $j^{ème}$ composante du vecteur de paramètres $\beta = (\beta_1, .., \beta_K)' \in \mathbb{R}^K$ d'un modèle dichotomique. L'idée du test de Wald est d'accepter l'hypothèse nulle si l'estimateur non contraint $\widehat{\beta}_j$ de β_j est proche de a. La stratistique de test est une mesure bien choisie de la proximité de $\beta_j - a$ à zéro.

On sait que dans la formulation générale d'un test de contraintes de type $H_0 : g(\beta) = r$, où r est un vecteur de dimension $(c, 1)$, on a le résultat suivant :

$$\left[g\left(\widehat{\beta}\right) - r\right]' \left[G\widehat{V}\left(\widehat{\beta}\right)G'\right] \left[g\left(\widehat{\beta}\right) - r\right] \xrightarrow[N\to\infty]{L} \chi(c)$$

[20] Le passage à un test unidirectionnel tel que $H_0 : \beta = a$ contre $H_1 : \beta > a$ peut se faire simplment en considérant les statistiques des test bidirectionnels et en adaptant la valeur critique.Pour un test de Wald sur un seul coefficient, l'intervalle d'acceptation à 5% est $[-1.96, 1.96]$ pour un test $H_1 : \beta \neq a$ alors qu'il devient $]-\infty, 1.96]$ pour le test $H_1 : \beta > a$.

où $\widehat{\beta}$ désigne l'estimateur du maximum de vraisemblance non contraint, avec $G = \partial g(.)/\partial \beta'$, et $\widehat{V}\left(\widehat{\beta}\right)$ l'estimateur de la matriuce de variance covariance des coefficients. Dans le cas qui nous intéresse, on a $g(\beta) = \beta_j$ et $r = a$. Le vecteur G, de dimension $(K, 1)$, comporte $K-1$ zéros et 1 à la $j^{ème}$ position. Ainbis, on obtient le résultat suivant :

Definition 5.1. ***La statistique du test de Wald associée au test unidirectionnel*** $H_0 : \beta_j = a$ ***contre*** $H_1 : \beta_j \neq a$ ***admet la loi suivante sous*** H_0 ***:***

$$\left[\widehat{\beta}_j - a\right]' (\widehat{v}_{jj})^{-1} \left[\widehat{\beta}_j - a\right] = \frac{\left(\widehat{\beta}_j - a\right)^2}{\widehat{v}_{jj}} \xrightarrow[N\to\infty]{L} \chi^2(1) \tag{5.1}$$

où $\widehat{v}_{jj}$ ***désigne l'estimateur de la variance de l'estimateur du*** $j^{ème}$ ***coefficient*** β_j***.***

Ainsi, si l'on note $\chi^2_{95\%}(1)$ le quantile à 95% de la loi $\chi^2(1)$, le test de Wald au seuil de 5% de l'hypothèse H_0 consiste à accepter H_0 si $\left(\widehat{\beta}_j - a\right)^2 / \widehat{v}_{jj}$ est inférieur à $\chi^2_{95\%}(1)$, et à refuser H_0 si cette quantité est supérieure à $\chi^2_{95\%}(1)$.

La plupart des logiciels (sauf SAS) ne propose pas cette statistique de Wald, mais une statistique z_j définie comme la racine carré de la précédente. Compte tenu du lien entre la loi normale centrée réduite et la loi du Chi2 à un degré de liberté, on a immédiatement sous H_0 :

$$z_j = \frac{\widehat{\beta}_j - a}{\sqrt{\widehat{v}_{jj}}} \xrightarrow[N\to\infty]{L} N(0, 1) \tag{5.2}$$

et en particulier pour un test de nullité $H_0 : \beta_j = 0$, on retrouve :

$$z_j = \frac{\widehat{\beta}_j}{\sqrt{\widehat{v}_{jj}}} \xrightarrow[N\to\infty]{L} N(0, 1) \tag{5.3}$$

5.1.2. Tests du rapport des maxima de vraisemblance

Dans le cas des modèles dichotomiques, on peut appliquer sans difficulté particulière la logique du test du rapport des maxima de vraisemblance. Ainsi, on estime le modèle non contraint et d'autre part le modèle contraint : soient $\widehat{\beta}_j$ et $\widehat{\beta}^c_j$ les deux estimations ainsi obtenues. La statistique LRT correspond alors tout simplemnt à l'écart des log-vraisemblance:

Definition 5.2. ***La statistique*** LRT_j ***du test du rapport des maxima de vraisemblance associée au test unidirectionnel*** $H_0 : \beta_j = a$ ***contre*** $H_1 : \beta_j \neq a$ ***admet la loi suivante sous*** H_0 ***:***

$$LRT_j = -2\left[\log L\left(y, \widehat{\beta}_j\right) - \log L\left(y, \widehat{\beta}^c_j\right)\right] \xrightarrow[N\to\infty]{L} \chi^2(1) \tag{5.4}$$

où $\widehat{\beta}_j$ ***et*** $\widehat{\beta}^c_j$***désignent respectivement les estimateurs non contraint et contraint de*** β_j***.***

Naturellement si l'on note $\chi^2_{95\%}(1)$ le quantile à 95% de la loi $\chi^2(1)$, le test du rapport des maxima de vraisemblance au seuil de 5% de l'hypothèse H_0 consiste à accepter H_0 si $LRT_j < \chi^2_{95\%}(1)$, et à refuser H_0 si $LRT_j > \chi^2_{95\%}(1)$. Cette porcédure est asymptotiquement équivalente à celle d'un test de Wald.

Dans le cas d'un test portant sur plus d'un paramètre, on utilise la statistique suivante

$$LRT = -2\left[\log L\left(y,\widehat{\beta}\right) - \log L\left(y,\widehat{\beta}^{c}\right)\right] \underset{N\to\infty}{\overset{L}{\longrightarrow}} \chi^{2}(r) \tag{5.5}$$

où r désigne le nombre de restrictions imposées sur les paramètres, et où $\widehat{\beta}$ et $\widehat{\beta}^{c}$ désigne les estimateurs respectivement non contraint et contraint du vecteur complet β.

5.1.3. Test du score ou du multiplicateur de Lagrange

Le principe de ce test est le suivant. On sait que si l'hypothèse nulle est satisfaite, les deux estimateurs non contraint $\widehat{\beta}_j$ et contraint $\widehat{\beta}_j^{c}$ doivent relativement proches l'un de l'autre, et que donc la même propriété doit être vérifiée pour le vecteur des des conditions du premier ordre de la maximisation de la log varisemblance.

Definition 5.3. ***La statistique*** LM_j ***du test du multiplicateur de Lagrange associée au test unidirectionnel*** $H_0 : \beta_j = a$ ***contre*** $H_1 : \beta_j \neq a$ ***admet la loi suivante sous*** H_0 ***:***

$$LM_j = \left(\left.\frac{\partial \log L(y,\beta)}{\partial \beta'}\right|_{\beta=\widehat{\beta}^{c}}\right)' \widehat{I}^{-1} \left(\left.\frac{\partial \log L(y,\beta)}{\partial \beta'}\right|_{\beta=\widehat{\beta}^{c}}\right) \underset{N\to\infty}{\overset{L}{\longrightarrow}} \chi^{2}(1) \tag{5.6}$$

où $\widehat{\beta}_j$ ***et*** $\widehat{\beta}_j^{c}$ ***désignent respectivement les estimateurs non contraint et contraint de*** β_j***.***

L'estimateur $\widehat{I}$ de la matrice d'iinformation de Fischer peut être obtenu par :

$$\widehat{I} = \sum_{i=1}^{N} \left(\left.\frac{\partial \log L(y_i,\beta)}{\partial \beta'}\right|_{\beta=\widehat{\beta}^{c}}\right) \left(\left.\frac{\partial \log L(y_i,\beta)}{\partial \beta'}\right|_{\beta=\widehat{\beta}^{c}}\right)'$$

et où

$$\left.\frac{\partial \log L(y,\beta)}{\partial \beta'}\right|_{\beta=\widehat{\beta}^{c}} = \sum_{i=1}^{N} \left.\frac{\partial \log L(y_i,\beta)}{\partial \beta'}\right|_{\beta=\widehat{\beta}^{c}}$$

5.2. Tests de spécification des modèles dichotomiques

Reste à présent à étudier les tests de spécifications qui permettent d'évaluer la qualité de l'ajustement par les modèles dichotomiques. Plusieurs solutions peuvent être adoptées à ce niveau pour comparer les différents modèles : comparaison tant au niveau du choix de la fonction $F(.)$ qu'au niveau du choix des variables explicatuves x_i^k. Par la suite, on notera $\widehat{F}(x_i\beta)$ la quantité $F\left(x_i\widehat{\beta}\right)$. Les différents critères présentés ici sont comme des fonctions de perte et il ne faut pas croire trouver un critère optimal pour chaque situation.

Nombre de prédictions fausses : le critère s'écrit sous la forme

$$\text{Nombre de fausses prédictions } \sum_{i=1}^{N} (y_i - \widehat{y}_i)^2 \tag{5.7}$$

où $\widehat{y}_i = 1$ si $\widehat{F}(x_i\beta) \geq 1/2$ et $\widehat{y}_i = 0$ si $\widehat{F}(x_i\beta) < 1/2$. Cette quantité donne le nombre de fausses prédictions puisque $(y_i - \widehat{y}_i)^2$ si seullement $y_i \neq \widehat{y}_i$: c'est à dire dans le cas où $\widehat{y}_i = 1$

alors que $y_i = 0$, ou dans le cas où $\widehat{y}_i = 0$ alors que $y_i = 1$. Ce critère est souvent utilisé en analyse discriminante. Le problème avec ce critère est que l'on considère de la même façon un individu ayant une probabilité $p_i = \widehat{F}(x_i\beta) = 0.49$ et un individu ayant une probabilité $p_i = \widehat{F}(x_i\beta) = 0$: on pénalise ces deux individus de la même façon dans le cas d'un échec du modèle (c'est à dire lorsque our les deux individus on a $y_i = 1$) et on les valorise de la même façon en cas de réussite. En, particulier, lorsque l'on considère des événements avec une forte probabilité (par exemple de sortir du chômage) ou au contraire une très faible probabilité (par exemple de tomber malade), la plupart des modèles obtiendront de bons résultats selon ce critère.

Somme des Carrés des Résidus (SCR) : ce critère traditionnel s'écrit sous la forme

$$\text{Somme des carrés des résidus } \sum_{i=1}^{N}\left[y_i - \widehat{F}(x_i\beta)\right]^2 \tag{5.8}$$

Rappelons que dans les modèles dichotomiques, on modélise la probabilité $p_i = E(y_i) = F(x_i\beta)$. Ce critère ne souffre pas de la critique précédente concernant le critère du nombre de fausses prédictions. C'est un crtère naturel puisquu'il correspond à la somme des carrés des résidus dans un modèle de régression linéaire standard à partir de laquell le R^2 est construit. Toutefois, l'utilisation de ce critère ne peut pas être défendue de la même façon dans le modèle linéaire simple et dans les modèles dichotomiques. En effet, nous avons vu que les modèles dichotomiques étaient des modèles hétéroscédastiques. C'est pourquoi Efron (1978) propose une mesure analogue au R^2 :

$$R^2 \text{ de Efron (1978) } = 1 - \frac{\sum_{i=1}^{N}\left[y_i - \widehat{F}(x_i\beta)\right]^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{5.9}$$

où $\overline{y} = N^{-1}\sum_{i=1}^{N} y_i$. Cette mesure alternative peut être défendue par une approche axiomatique (cf. Efron 1978)L.

SCR pondérée par les probabilités estimées : ce critère s'écrit sous la forme

$$\text{SCR pondérée } \sum_{i=1}^{N}\frac{\left[y_i - \widehat{F}(x_i\beta)\right]^2}{\widehat{F}(x_i\beta)\left[1 - \widehat{F}(x_i\beta)\right]} \tag{5.10}$$

La principale raison de préférer ce critère à la somme non pondérée est la suivante. Il paraît raisonnable d'attacher une plus grande perte aux erreurs faites en prévoyant des variables de faible variance, étant donné qu'il est plus facile de pérvoir des variables de faible variance que des variables de plus forte variance. Dès lors, il paraît raisonnable de pondérer la somme des carrés des résidus par un poids qui est inversement proprtionnel à la variance.

Coefficient de Corrélation des Carrés : ce critère s'écrit sous la forme

$$\text{Coefficient de corrélation des carrés } \frac{\left[\sum_{i=1}^{N}(y_i - \overline{y})\widehat{F}(x_i\beta)\right]^2}{\left\{\sum_{i=1}^{N}(y_i - \overline{y})^2\right\}\left\{\sum_{i=1}^{N}\left[\widehat{F}(x_i\beta) - \overline{F}\right]^2\right\}} \tag{5.11}$$

Cette mesure est liée à la SCR non pondérée. Dans un modèle de régression standard, cette mesure serait identique au R^2 de Effron. Bien que cette égalité ne soit pas vraie dans

les modèles dichotomiques, les mêmes critiques s'appliquent au coefficient de corrélation des carrésqu'à la SCR.

Log - Vraisemblance : ce critère s'écrit sous la forme

$$\text{Log-Vraisemblance } \log L\left(y,\widehat{\beta}\right) = \sum_{i=1}^{N} y_i \log\left[F\left(x_i\widehat{\beta}\right)\right] + (1-y_i)\log\left[1-F\left(x_i\widehat{\beta}\right)\right] \quad (5.12)$$

Ce critère est particulièrement bien adapté pour comparre des modèles qui ne possèdent pas les mêmes dimensions. En effet, on sait que si l'on désire tester r contraintes linéaires sur les paramètres la $-2\left[\log L\left(y,\widehat{\beta}_j\right) - \log L\left(y,\widehat{\beta}_j^c\right)\right]$ suit asympotiquement un $\chi^2(r)$. Une normlisation de la quantité $\log L\left(y,\widehat{\beta}\right)$ a été proposée par McFadden pour se ramener à une quantité similaire à un R^2 :

$$R^2 \text{ de McFadden (1974) } = 1 - \frac{\log L\left(y,\widehat{\beta}\right)}{\log L(y,0)} \quad (5.13)$$

où $\log L(y,0)$ désigne le maximum de la fonction de log vraisemblance obtenu lorsque tous les coefficients de la regrssion β sont nuls à l'exception du terme constant.

6. Application

Proposer une application avec :

1. Problème économique et spécification en variable latente
2. Estimation Logit Probit
3. Comparaison avec estimation non paramétrique (score maximum et semi paramétrique)
4. Calcul des cotes et des probabilités individuelles
5. Calcul des effets marginaux : calcul des elasticités moyennes selon les deux formules et des elasticités individuelles
6. Vérification des calculs de l'estimateur de la matrice de variance covariance asymptotique
7. Calcul des principaux critères d'évaluation (R^2 de McFadden etc..)
8. Tests d'hypothèse sur les paramètres : Wald, LRT et LM

****
*** A FINIR ****
****

A. Annexes

A.1. Rappels sur les notions de convergence

Les rappels proposés dans le cadre de cette section portent sur les différentes notions de convergence. Toutefois, la lecture de ces rappels doit nécessairement s'accompagner d'une étude plus systématique des fondements probabilistes de ces notions[21].

Considérons une séquence de T v.a.r. $\{X_1, X_2, ..., X_i, ..., X_T\}$, indicées par i. Supposons que l'on souhaite étudier le comportement de la moyenne empirique de ces v.a.r. lorsque T augmente. On cherche ainsi à déterminer *le comportement asymptotique* de la v.a.r. transformée, $\overline{X}_T$, telle que :

$$\overline{X}_T = \frac{1}{T}\sum_{i=1}^{T} X_i \tag{A.1}$$

Pour cela, il convient d'utiliser *la notion de convergences.*

A.1.1. Convergence en probabilité

La notion de convergence en probabilité est définie de la façon suivante :

Definition A.1. *(Convergence en Probabilité)* ***Soit*** $\{X_T\}_{T=1}^{\infty}$ ***une séquence de variables aléatoires scalaires. Cette séquence converge en probabilité vers*** $c, \forall c \in \mathbb{C}$***, si pour toute valeurs arbitraires*** $\varepsilon > 0$ ***et*** $\delta > 0$***, il existe une valeur*** N***, telle que*** $\forall T \geq N$ ***:***

$$P\left[|X_T - c| > \delta\right] < \varepsilon \tag{A.2}$$

Alors, on note :

$$X_T \xrightarrow{p} c \Longleftrightarrow plim\, X_T = c \tag{A.3}$$

Exprimée autrement, cette définition signifie que pour un échantillon de taille infinie, la probabilité que la réalisation de la variable X_T diffère de la valeur c de plus ou moins δ (δ étant aussi petit que l'on veut) est inférieure à toute valeur ε aussi petite soit-elle. En d'autres termes, les réalisations de la variable X_T sont concentrées au voisinage de la valeur c.

Propriété *Une suite de matrices de v.a.r.* $\{X_T\}_{T=1}^{\infty}$*, de dimension* (m, n) *,converge en probabilité vers une matrice* C*, de dimension* (m, n)*, si chaque élément de* X_t *converge en probabilité vers l'élément correspondant de* C*. De façon plus générale, si l'on considère deux séquences de v.a.r.* $\{X_T\}_{T=1}^{\infty}$ *et* $\{Y_T\}_{T=1}^{\infty}$*, de dimension* (m, n)*, alors :*

$$X_T \xrightarrow{p} Y_T \tag{A.4}$$

si et seulement si, la différence entre les deux suites converge en probabilité vers zero :

$$X_T - Y_T \xrightarrow{p} 0 \tag{A.5}$$

Enfin, il convient de rappeler deux propriétés qui nous serons utiles dans la caractérisation des distributions asymptotiques des estimateurs usuels.

[21] Voir par exemple, "Méthodes Statistiques", Philippe Tassi, *Economica* 1989

Theorem A.2. *(Théorème de Slutsky)* ***Soit*** $\{X_T\}_{T=1}^{\infty}$ ***une suite de*** $(n,1)$ ***vecteurs admettant une limite en probabilité définie par*** c, ***et soit*** $g(.)$ ***une fonction continue en*** c, ***satisfaisant*** $g:\mathbb{R}^n \longrightarrow \mathbb{R}^n$, ***et ne dépendant pas de*** T, ***alors :***

$$g(X_T) \xrightarrow[T\to\infty]{p} g(c) \tag{A.6}$$

L'idée est la suivante. Si la fonction $g(.)$ est continue, la quantité $g(X_T)$ se situera au voisinage de $g(c)$, dès lors que X_T se situe au voisinage de c. En choisissant une valeur de T suffisamment grande, la probabilité que la réalisation de X_T se situe au voisinage de c peut être définie aussi proche de l'unité que l'on le désire. Un exemple simple est le suivant. Considérons deux séquences de v.a.r. telles que $plim\ X_{1,T} = c_1$ et $plim\ X_{2,T} = c_2$, alors $plim\,(X_{1,T} + X_{2,T}) = c_1 + c_2$. La démonstration de ce résultat est immédiate dès lors que l'on montre que la fonction $g(X_{1,T}, X_{1,T}) = X_{1,T} + X_{2,T}$ est une fonction continue en (c_1, c_2).

Propriété 1 *Une condition suffisante pour qu'une suite de v.a.r.* $\{X_T\}_{T=1}^{\infty}$ *converge en probabilité vers une constante réelle* c *est :*

$$\lim_{T\to\infty} E(X_T) = c \tag{A.7}$$

$$\lim_{T\to\infty} V(X_T) = 0 \tag{A.8}$$

L'intuition de cette propriété est simple. Si pour un ordre T suffisamment grand, la variable X_T admet c pour espérance et a une variance qui tend vers 0, alors la fonction de distribution de X_T sera infiniment concentrée autour de la valeur c.

A.1.2. Convergence en moyenne quadratique

Une forme de convergence plus restrictive que la convergence en probabilité est la convergence *en moyenne quadratique* (*m.s. pour mean square convergence*).

Definition A.3. ***Une suite de suite de v.a.r.*** $\{X_T\}_{T=1}^{\infty}$ ***converge en moyenne quadratique vers*** c, ***si pour tout*** $\varepsilon > 0$, ***il existe une valeur*** N, ***telle*** $\forall T \geq N$ ***:***

$$E(X_T - c)^2 < \varepsilon \tag{A.9}$$

Alors, on note :

$$X_T \xrightarrow{m.s.} c \tag{A.10}$$

Naturellement, étant donné cette définition, *la convergence en moyenne quadratique implique la convergence en probabilité, mais la réciproque n'est pas vraie* :

$$X_T \xrightarrow{m.s.} c \Longrightarrow X_T \xrightarrow{p} c$$

La notion de convergence en m.q. nous permet alors d'introduire l'inégalité de Chebyshev.

Proposition A.4. *(Inégalité de Chebyshev)* ***Soit*** X ***une v.a.r. telle que la quantité*** $E(|X|^r)$ ***existe et soit finie pour*** $r > 0$. ***Pour tout*** $\delta > 0$, ***et toute valeur de*** c, ***on montre que :***

$$P\{|X - c| > \delta\} \leq \frac{E(|X - c|^r)}{\delta^r} \tag{A.11}$$

Le résultat selon lequel la convergence en moyenne quadratique implique la convergence en probabilité peut être démontré à partir de l'inégalité de Chebyshev. Pour cela, il suffit de remarquer que si $X_T \xrightarrow{m.s.} c$, alors il existe un couple de valeurs positives (δ, ε) et une valeur N, tel que $E\left(X_T - c\right)^2 < \delta^2\varepsilon$, pour tout $T \geq N$**.** Il s'ensuit que :

$$\frac{E\left(X-c\right)^2}{\delta^2} = \frac{E\left(\left|X-c\right|^2\right)}{\delta^2} < \varepsilon \quad \forall T \geq N$$

L'inégalité de Chebyshev implique alors que :

$$P\left\{\left|X-c\right| > \delta\right\} < \varepsilon \quad \forall T \geq N$$

Donc, on montre ainsi que $X_T \xrightarrow{p.} c$.

A.1.3. Convergence en loi

Le troisième type de convergence que nous utiliserons cette année est la *convergence en loi* ou *convergence en distribution.*

Theorem A.5. *(Théorème de Paul Levy)* ***Soit*** $\{X_T\}_{T=1}^{\infty}$ ***une suite de v.a.r. et soit*** $F_{X_T}(x)$ ***la fonction de distribution cumulative de*** X_T**.** ***Si*** X_T ***converge en loi vers une v.a.r.*** X ***admettant*** $F_X(x)$ ***pour fonction caractéristique, alors :***

$$\lim_{T\to\infty} F_{X_T}(x) = F_X(x) \quad \forall x \in \mathbb{R} \tag{A.12}$$

On note alors :

$$X_T \xrightarrow[T\to\infty]{loi} X \quad ou \quad X_T \xrightarrow[T\to\infty]{\mathcal{L}} X \tag{A.13}$$

Un certain nombre de propriétés nous serons particulièrement utiles par la suite :

Propriété 1 *La convergence en probabilité implique la convergence en loi :*

$$X_T - X \xrightarrow[T\to\infty]{p} 0 \Longrightarrow X_T \xrightarrow[T\to\infty]{\mathcal{L}} X \tag{A.14}$$

Propriété 2 *La convergence en loi vers une constante réelle implique la convergence en probabilité :*

$$\forall c \in \mathbb{R} \quad X_T \xrightarrow[T\to\infty]{\mathcal{L}} c \Longrightarrow X_T \xrightarrow[T\to\infty]{p} c \tag{A.15}$$

Propriétés 3 *Soient deux suites de v.a.r.* $\{X_T\}_{T=1}^{\infty}$ *et* $\{Y_T\}_{T=1}^{\infty}$ *telle que* $X_T \xrightarrow{\mathcal{L}} X$ *et* $Y_T \xrightarrow{p} c$*, alors :*
(i) $X_T + Y_T \xrightarrow{\mathcal{L}} X + c$
(ii) $X_T Y_T \xrightarrow{\mathcal{L}} c\,X$
(iii) $\frac{X_T}{Y_T} \xrightarrow{\mathcal{L}} \frac{X}{c}$ avec $c \neq 0$

Propriété 4 *Soient* X_T *et* X *des vecteurs aléatoires de* $\mathbb{R}^p$*, tels que* $X_T \xrightarrow[T\to\infty]{\mathcal{L}} X$ *,et soit* $g(.)$ *une fonction continue définie de* $\mathbb{R}^p$ *and* $\mathbb{R}^n$*, alors :*

$$g(X_T) \xrightarrow[T\to\infty]{\mathcal{L}} g(X) \tag{A.16}$$

Bibliographie

Amemiya T. (1976), "The ML, the Minimum Chi-*Square and the Non Linear Weighted Least Squares Estimator in the General Qualtitative Response Model", *Journal of the Ameraican Statistical Association,* 71, 347-351

Amemiya T. (1981), "Qualitative Response Models : A Survey", *Journal of Economic Litterature,* 19(4), 481-536

Amemiya T. (1985), "Advanced Econometrics", *Cambridge, Harvard University Press.*

Alban T. (2000), "Econométrie des Variables Qualitatives", *Dunod.*

Berkson J. (1944), "Application of the Logistique Function to Bio-Assay", *JASA*, 39, 357-365.

Berkson J. (1951), "Why I prefer Logit to Probit", *Biometrics*, 7, 327-339.

Colletaz G. (2001), "Modèles à Variables Expliquées Qualitatives", *Miméo Université Orléans*

Davidson R. et MacKinnon J.G. (1984), "Convenient Tests for Logit and Probit Models", *Journal of Econometrics*, 25, 241-262.

Gourieroux C. (1989), "Econométrie des Variables Qualitatives", *Economica.*

Gourieroux C. et Montfort A. (1981), "Asymptotic Properties of the Maximuùm Likelihood Estimator in Dichotomous Logit Models", *Journal of Econometrics,* 17, 83-97.

Greene W.H. (1997), "Econometric Analysis", *Londres, Prentice Hall.*

Judge G.G., Miller D.J. et Mittelhammer R.C. (2000), "Econometric Foundations", *Cambridge University Press.*

Klein R.W. et Spady R.H. (1993), "An Efficient Semi Parametric Estimator for Binary Response Models", *Econometrica*, 61, 387-421

Maddala. G.S. (1983), "Limited-dependent and Qualitative Variables in Econometrics", *Econometric Society Monographs,* 3, Cambrige University Press.

Morimune K. (1979), "Comparisons of Normal and Logistic Models in the Bivariate Dichitomous Analysis", *Econometrica,* 47, 957-975.

Radner R. et Miller L. (1970), "Demand and Supply in U.S. Higher Education : A Progress Report", *American Economic Review,* 60.

Spector L.C. et Mazzeo M. (1980), "Probit Analysis and Economic Education", *Journal of Economic Education*, 11(2), 37-44

Tobin J. (1958), "Estimation of Relationships for Limited Dependent Variables", *Econometrica*, 26, 24-36.

Figure A.1: L'économie a travers les prix nobel, Problèmes Economiques 2001

2000

Les économètres du quotidien

James J. Heckman
Daniel L. McFadden

Le prix 2000 a été attribué conjointement à James Heckman « pour avoir développé des théories et des méthodes d'analyse des échantillons sélectifs » et à Daniel McFadden « pour avoir développé des théories et des méthodes d'analyse des choix discrets. »

James Heckman

Biais d'échantillonnage...

Toute recherche économétrique utilise des statistiques. La microéconométrie se sert de données qui décrivent une population d'individus. Comme la population n'est généralement pas observable dans son intégralité, les économètres ont recours à un échantillon de celle-ci. Cet échantillon est construit à partir d'un tirage aléatoire dans la population étudiée. Sa représentativité n'est cependant pas toujours garantie car certaines données ne sont pas directement observables ; on parle alors de biais d'échantillonnage ou de biais de sélection. Par exemple, le nombre d'heures travaillées et les salaires ne sont disponibles que pour ceux qui ont effectivement choisi de travailler.

... et auto-sélection

Les statistiques sont donc parfois trompeuses car les échantillons sélectifs ne rendent pas compte avec exactitude des populations étudiées. Ces biais peuvent également être le résultat des choix des agents : on parle alors d'auto-sélection. Par exemple, des agents peuvent choisir volontairement de ne pas demander à bénéficier d'un plan d'aide social. L'évaluation de la politique publique devra donc tenir compte de ces auto-sélections.

James Heckman a mis au point une méthode économétrique qui permet de résoudre, relativement simplement, ces problèmes d'auto-sélection. Elle porte son nom : la correction de Heckman. Ses travaux ont eu de nombreuses implications empiriques, aussi bien en économie que dans les autres sciences sociales.

De l'économétrie utile !

Deux domaines ont particulièrement profité des avancées de Heckman : l'évaluation des politiques actives de l'emploi et les modèles de durée.

Les travaux de l'Américain ont notamment permis de répondre à la difficile question : que serait-il arrivé à un agent s'il n'avait pas bénéficié de la politique publique ? Y répondre pose un problème de sélection. Heckman a montré que leurs effets sont souvent plus faibles qu'on ne l'imaginait auparavant, quand ils ne sont pas négatifs.

Les modèles de durée, appliqués par exemple au chômage, s'intéressent au lien entre la durée du chômage et la probabilité de retrouver un emploi. Ici aussi, un problème de sélection se pose : les individus susceptibles d'être au chômage, de par l'inadaptation de leur qualifications, sont surreprésentés dans les chômeurs de longue durée. Heckman a développé des méthodes de résolution de ce type de biais.

Daniel McFadden

Traditionnellement dans l'analyse économique de la demande, le choix individuel est représenté par une variable continue. Or, dans la réalité, l'individu est souvent confronté à un nombre restreint de choix possibles, autrement dit des choix discrets. C'est le cas, par exemple, pour la demande de transport. L'individu choisit entre la voiture ou les transports en commun. Les travaux de McFadden ont consisté à élaborer une méthodologie économétrique afin d'analyser ce type de choix.

L'analyse logit conditionnelle

Sa contribution fondamentale date de 1974 (a,b) et porte sur l'analyse logit conditionnelle. Dans ce modèle, chaque individu d'une population fait face à un nombre fini de choix possibles et cherche à maximiser son utilité. L'analyse consiste à lier les caractéristiques de chaque choix possible à celles des individus. Toutefois, toutes les données n'étant pas observables, les différences entre les caractéristiques des choix et des individus seront représentées par un vecteur de termes d'erreurs aléatoires. Ces derniers ont une distribution statistique particulière appelée distribution des valeurs extrêmes. Sous certaines conditions, on peut calculer la probabilité qu'un individu choisisse tel ou tel moyen de transport, tel ou tel type de logement.

Le succès de ces modèles, appelés aussi logit multivariés, vient de la combinaison de fondements microéconomiques solides et de la simplicité du calcul. Cette simplicité a pour origine l'hypothèse d'indépendance statistique des termes d'utilité aléatoires, hypothèse qui implique la propriété d'indépendance des choix non pertinents (*independance of irrelevant alternatives* ou IIA). Autrement dit, le ratio des probabilités de choisir entre deux moyens de transport est indépendant des propriétés de tous les autres modes de transport possibles.

Les développements

Cette hypothèse est toutefois restrictive. Il est invraisemblable qu'un des deux choix soit indépendant de l'introduction d'un nouveau choix.

McFadden a donc montré l'intérêt de relâcher cette hypothèse au travers des modèles logit emboîté (*nested logit*) et de valeurs extrêmes généralisées (*generalized extreme value*). Un des prolongements dans l'analyse des choix individuels a été de chercher à développer des modèles et méthodes qui expliquent à la fois les choix discrets et continus.

McFadden a contribué à d'autres domaines importants de la recherche. C'est le cas dans le domaine des impôts et de la nouvelle économie publique. C'est également le cas dans le domaine des problèmes de l'économie de l'environnement notamment en matière de méthodes de valorisation contingente pour l'estimation de la valeur des ressources naturelles.

■ **La rédaction de Problèmes économiques**

Bibliographie indicative

Heckman J.J. (1987), « Selection Bias and Self-Selection », in P. Newman, M. Milgate and J. Eatwell (sous la dir.), *The New Palgrave - A Dictionary of Economics*, Macmillan.
Heckman J. J., LaLonde R. et Smith J. (1999), « The Economics and Econometrics of Active Labor Market Programs », *Handbook of Labor Economics*, 3, North-Holland.
Heckman J. J. et Singer B. (1984), « A Method of Minimizing the Impact of Distributional Assumptions for Duration Data », *Econometrica*, 52, p. 271 à 320.
Heckman J. J. and J. Smith (1995), « Assessing the Case for Social Experiments », *Journal of Economic Perspectives*, 9, p. 85 à 110.
McFadden D. (1974a), « Conditionnal logit analysis of qualitative choice behavior », in P. Zarembka (sous la dir.), *Frontiers of econometrics*, Academic Press.
McFadden D. (1974b), « The measurement of urban travel demand », Journal of Public Economics, 3, p. 303 à 328.
McFadden D. (1987), « Regression based specification tests for the multinomial logit model », *Journal of Econometrics*, 34, p. 63 à 82.
McFadden (1989), « A method of simulated moments for estimation of discrete response models without numerical integration », *Econometrica*, 57, p. 995 à 1026.
McFadden D. (1994), « Contigent valuationand social choice », *American Journal of Agricultural Economics*, 74, p. 689 à 708.