CHAPITRE 1 : LES MODELES DE CHOIX BINAIRE

Des exemples :

1- Après le bac, aller à la fac ou pas,
2- Acheter ou louer une maison,
3- Prendre le bus ou la voiture,
4- Faire une campagne de publicité à la télé ou à la radio,
5- Voter pour x ou y…

Ces choix sont représentés par une variable indicatrice (dummy) qui prend la valeur 1 quand un item est choisi et la valeur 0 pour l'autre.

Dans chaque cas, on choisit entre deux résultats mutuellement exclusifs.

La variable endogène est ici une variable dichotomique.

<u>Exemple historique</u> : choix individuel de transport (pour aller au travail)

La variable endogène se présente comme suit :

$$y = \begin{cases} \mathbf{1} \\ \mathbf{0} \end{cases} \tag{1.1}$$

Avec 1 si l'individu prend sa voiture pour aller travailler
0 si l'individu prend le bus pour aller travailler

y est une variable aléatoire (toujours le cas quelque-soit les cours). Donc en termes de probabilité nous avons :

p = probabilité que l'individu prenne sa voiture

$$\boldsymbol{Prob[y = 1] = p}$$

On en déduit naturellement (probabilité que l'individu prenne le bus) :

$$\boldsymbol{Prob[y = 0] = 1 - p}$$

La fonction de probabilité est :

$$\boldsymbol{f(y) = p^{y}(1 - p)^{1-y}}$$

Avec y = 0,1

Cette variable aléatoire discrète se caractérise par :

$$\boldsymbol{E(y) = p \quad \text{et} \quad Var(y) = p(1 - p)}$$

Quels sont les facteurs qui affectent la probabilité qu'un individu choisisse un mode de transport plutôt qu'un autre ? On va donc rechercher les variables exogènes.

On peut avoir :

1) Le temps de trajet,
2) Le coût,
3) La pollution,
4) …

Supposons une seule variable exogène. Le modèle s'écrit alors :

$$\boldsymbol{y = cste + bx}$$

Comment estimer cette équation mais surtout comment estimer la probabilité p ? Par OLS, par GLS ?...

SECTION 1 : LE MODELE DE PROBABILITE LINEAIRE (OLS classique)

Dans la régression classique on décompose y en deux parties : une part fixe et une part aléatoire. Dans notre étude cela donne :

$$\boldsymbol{y = E(y) = p + \varepsilon}$$

Si la relation est linéaire alors la partie fixe, avec une probabilité p pour y=1, s'écrit :

$$\boldsymbol{E(y) = p = cste + bx}$$

Le modèle complet (appelé aussi <u>modèle de probabilité linéaire</u>) s'écrit :

$$\boldsymbol{y = E(y) + \varepsilon = cste + bx + \varepsilon}$$

Or trois problèmes apparaissent ici :

1) Les résidus sont hétéroscédastiques,
2) Les prévisions des probabilités peuvent sortir de l'intervalle [0,1],
3) Les effets marginaux des changements dans les variables explicatives sont <u>constants</u>.

<u>Démonstrations :</u>

1 Dans le modèle (1.1) on constate que la matrice des variances-covariances des résidus varie entre les individus en fonction de leurs caractéristiques associées aux variables exogènes x. Puisque y ne peut prendre que 2 valeurs, il en va de même pour les résidus :

Si y = 1 alors ε = 1-(cste+bx) avec p=cste+bx

Si y = 0 alors ε = -(cste+bx) avec 1-p=1-(cste+bx)

Calculons la variance des résidus :

$$var(\varepsilon) = \big(1-(cste+bx)\big)^2 prob(y=1) + (-cste-bx)^2 prob(y=0)$$

On remplace :

$$var(\varepsilon) = \big(1-(cste+bx)\big)^2 (cste+bx) + (-cste-bx)^2 (1-(cste+bx))$$

On développe :

$$var(\varepsilon) = (1 - cste - bx)\,(cste + bx)[(1 - cste - bx) + (cste + bx)]$$

On obtient :

$$\boldsymbol{var(\varepsilon) = (cste + bx)(1 - cste - bx)}$$

Cette dernière expression n'est pas constante pour toutes les observations.

2 Si les résidus sont de moyenne nulle, la probabilité p associée à l'événement y=1 est déterminée de façon unique :

$$E(\varepsilon) = (1 - cste - bx)p + (-cste - bx)(1 - p) = 0$$

$$E(\varepsilon) = p - p.cste - bx.p - cste + p.cste - bx + bx.p = 0$$

$$E(\varepsilon) = p - cste - bx = 0$$

$\Leftrightarrow$ $$\boldsymbol{p = cste + bx}$$

Ainsi la quantité [cste+bx] correspond à une probabilité et doit par conséquent satisfaire la propriété d'appartenir à l'intervalle [0,1]. Or rien n'assure que de telles conditions soient satisfaites ici.

3 $\frac{dp}{dx} = b$. La dérivée est constante or p $\in$[0,1]. Ce qui est impossible

SECTION 2 : LE MODELE PROBIT

Pour avoir une probabilité p comprise entre [0, 1], une relation non linéaire entre x et p doit être utilisée. Une première forme naturelle qui nous vient à l'esprit est la fonction PROBIT. Cette fonction est en relation avec la distribution de probabilité de la loi normale (0,1). La fonction de densité de probabilité est :

(1.2)

$$\boldsymbol{\emptyset(z) = \frac{1}{\sqrt{2\pi}} e^{-0.5z^2}}$$

La fonction Probit :

(1.3)

$$\boldsymbol{\varphi(z) = P[Z \leq z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-0.5u^2}\, du}$$

NB : c’est la loi normale standard

La probabilité p s’écrit de fait quand y prend la valeur 1 :

$$\boxed{(1.4)} \qquad \boldsymbol{p = P[Z \leq cste + bx] = \varphi(cste + bx)}$$

Le modèle probit est dit non linéaire car l’expression (1.4) est une fonction non linéaire des paramètres : cste et b.

2.1 Interprétation du modèle PROBIT

L’effet marginal (d’une variation d’une unité de x sur la probabilité que y=1) se calcule comme suit :

$$\boldsymbol{\frac{dp}{dx} = \frac{d\varphi(t)}{dt} . \frac{dt}{dx} = \emptyset(cste + bx)b}$$

Avec t=cste+bx

L'effet marginal dépend de la pente de la fonction probit qui est donnée par $\emptyset(cste + bx)$ et de la valeur du paramètre b.

L’effet marginal a les caractéristiques suivantes :

1- Puisque $\emptyset(cste + bx)$ est une densité de probabilité, sa valeur est toujours positive. Le signe de l’effet marginal ne dépend que du signe du paramètre b,

2- Quand x change, la valeur de $\emptyset(cste + bx)$ change de fait. La densité de la loi normale atteint son maximum quand z=0 i.e. quand cste+bx = 0. Dans ce cas p = φ(0) = 0.5,

3- Quand on fera des prévisions, on pourra s’appuyer sur les 2 résultats précédents qui impliquent :

$$\hat{y} = \begin{cases} 1 & \hat{p} \geq .5 \\ 0 & \hat{p} < .5 \end{cases}$$

En ayant au préalable estimé :

$$\hat{p} = \varphi(\widehat{cste} + \hat{b}x)$$

2.2 L'estimation du modèle PROBIT : le maximum de vraisemblance

Cf. chapitre 7 et chapitre 9 du cours de L3.

NB : dans le chapitre 7 il faut remplacer [y-Xβ] par (1.3) qui est non linéaire (d'où le chapitre 9).

2.3 Application : un modèle dichotomique de transport (Cf. TD1)

Base de données : TRANSPORT.XLS

N=21 individus

Variable endogène :

$y = \begin{cases} 1 & on\ choisit\ la\ voiture \\ 0 & on\ choisit\ le\ bus \end{cases}$ pour aller travailler

Variable exogène : $X = (bustime - autotime)/10$

(i.e. différentiel de temps de trajet par tranche de 10 mn)

Avec : bustime : temps de trajet en bus en mn

autotime : temps de trajet en voiture en mn

Le modèle probit que nous voulons estimer est :

$$\boldsymbol{P(auto = 1) = \varphi(cste + bX)}$$

Les estimations par ML avec SAS 9.4 sont :

Model Fit Summary	
Number of Endogenous Variables	1
Endogenous Variable	Y
Number of Observations	21
Log Likelihood	-6.16516
Maximum Absolute Gradient	9.85946E-7
Number of Iterations	9
Optimization Method	Quasi-Newton
AIC	16.33032
Schwarz Criterion	18.41936

Mesures du critère qualificatif de lissage		
Mesure	Valeur	Formule
Likelihood Ratio (R)	16.734	2 * (LogL - LogL0)
Upper Bound of R (U)	29.065	- 2 * LogL0
Aldrich-Nelson	0.4435	R / (R+N)
Cragg-Uhler 1	0.5493	1 - exp(-R/N)
Cragg-Uhler 2	0.7329	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.6948	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.5497	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.5758	R / U
Veall-Zimmermann	0.7639	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.7352	
N = # d'observations, K = # de régresseurs		

Algorithm converged.

Résultats estimés des paramètres					
Paramètre	DDL	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|
Intercept	1	-0.064434	0.399244	-0.16	0.8718
X	1	0.299990	0.102869	2.92	0.0035

<u>Commentaires</u> :

cste non significative b significatif

Signe – pour cste i.e. quand les temps de trajet sont identiques (x=0) alors les individus ont un biais en défaveur de la voiture, par rapport aux transports en commun.

Signe + pour b i.e. une augmentation du temps de trajet du bus augmentera la probabilité qu'un individu choisisse la voiture.

Calculons <u>les effets marginaux</u> d'une augmentation des temps de transport en bus/au transport en voiture. Nous devons calculer l'expression suivante :

$$\frac{\boldsymbol{dp}}{\boldsymbol{dx}} = \emptyset(-.\mathbf{0644}+.\mathbf{3} * \boldsymbol{aug}).\mathbf{3}$$

Calcul effets marginaux

Aug	$\emptyset(-.0644 + .3 * aug)$ [Cf. (1.2)]	$\frac{dp}{dx}$
5 mn (.5)	0,397483358	0,119245007
10 mn (1)	0,388022398	0,11640672
15 mn (1.5)	0,370359096	0,111107729
20 mn (2)	0,345634917	0,103690475
25 mn (2.5)	0,315384669	0,094615401
30 mn (3)	0,281379155	0,084413746
35 mn (3.5)	0,245454861	0,073636458

Interprétation pour 20 mn : une augmentation graduelle du temps de déplacement par bus entraîne une augmentation de la probabilité de déplacement par auto d'environ 10.36%, (étant donné que prendre le bus nécessite déjà 20 mn de temps de déplacement en plus que la voiture).

On peut aussi prévoir le comportement d'un individu qui doit choisir entre ces deux modes de transport. Nous utilisons ici la formule (1.4) :

$$\hat{\boldsymbol{p}} = \boldsymbol{\varphi}(\widehat{\boldsymbol{cste}} + \hat{\boldsymbol{b}} * \boldsymbol{aug})$$

Calcul probabilités estimées

Aug	$\varphi(-.0644 + .3 * aug)$ [Cf. (1.3)]	$\hat{p}$
5 mn (.5)	0,0856	.5341
10 mn (1)	0,2356	.5931
15 mn (1.5)	0,3856	.6501
20 mn (2)	0,5356	.7038
25 mn (2.5)	0,6856	.7535
30 mn (3)	0,8356	.7983
35 mn (3.5)	0,9856	.8372

NB : pour le calcul de φ voir le site : http://calculis.net/probabilite-loi-normale
~~(NB : pour 'valeur de a' et 'valeur de b', mettre le même nombre i.e. 0.0856)~~

Puisque la probabilité estimée qu'un individu choisisse la voiture est de 70% (.7038), qui est > .5, on peut prédire que lorsque le trajet du bus prend 20 mn de plus que le trajet de la voiture, l'individu choisira la voiture pour aller travailler. Idem pour les autres cas.

NB : on calcule souvent l'EMM (effet marginal moyen) qui est la moyenne des effets marginaux à chaque point de l'échantillonnage :

$$\boldsymbol{EMM = \frac{1}{N}\sum_{i=1}^{N} \emptyset\left(\widehat{cste} + \hat{b}x\right)\hat{b}}$$

Dans notre exemple EMM = 0,0484, c'est l'augmentation estimée moyenne de la probabilité, compte tenu de 10 mn d'augmentation du temps de trajet en bus par rapport au temps de déplacement de l'automobile, soit 4.84%.