CHAPITRE 3 : LE MODELE LOGIT (PROBIT) MULTINOMIAL

SECTION 1 : Exemple : le choix d'éducation post-secondaire

Base de données : EDUCATION.XLS

N=1000 individus

Variable endogène :

$$y = \begin{cases} 1 & \textbf{pas au collège} \\ 2 & \textbf{2 ans au collège} \\ 3 & \textbf{4 ans au collège} \end{cases}$$

Variables exogènes :

ETCATHO = 1 si études secondaires catholiques

NIVEAU = index moyen en mathématiques, anglais et études sociales sur échelle de 13 points avec 1 le plus élevé et 13 le plus faible

REVENU = revenu familial brut en $

PERSON = Nombre de personnes dans la famille

DIPARENT = 1 si la plupart des parents formés étaient diplômés du collège ou avaient un diplôme plus élevé

FEMME = 1 si femme, 0 si homme

NOIR = 1 si noir, 0 si autres

SECTION 2 : ECRITURE DU MODELE LOGIT

On doit exprimer la probabilité que la i^{ième} personne choisira la possibilité (alternative) j :

P_{ij}=Proba[personne i choisisse l'événement j]

Si j=3, on a les 3 probabilités suivantes avec la distribution logistique et une seule variable exogène :

$$P_{i1} = \frac{\exp(cste_1 + b_1 x)}{\exp(cste_1 + b_1 x) + \exp(cste_2 + b_2 x) + exp(cste_3 + b_3 x)}$$

$$P_{i2} = \frac{\exp(cste_2 + b_2 x)}{\exp(cste_1 + b_1 x) + \exp(cste_2 + b_2 x) + exp(cste_3 + b_3 x)}$$

$$P_{i3} = \frac{\exp(cste_3 + b_3 x)}{\exp(cste_1 + b_1 x) + \exp(cste_2 + b_2 x) + exp(cste_3 + b_3 x)}$$

Avec : $cste_1$ et b_1 représentent le premier choix.
$cste_2$ et b_2 représentent le deuxième choix.
$cste_3$ et b_3 représentent le troisième choix.

Or il n'est pas possible de calculer toutes ces probabilités. Pour résoudre ce problème d'identification on impose que $cste_1$ = b_1 =0 et on obtient les probabilités suivantes :

$$P_{i1} = \frac{1}{1 + \exp(cste_2 + b_2 x) + exp(cste_3 + b_3 x)}$$

$$P_{i2} = \frac{\exp(cste_2 + b_2 x)}{1 + \exp(cste_2 + b_2 x) + exp(cste_3 + b_3 x)}$$

$$P_{i3} = \frac{\exp(cste_3 + b_3 x)}{1 + \exp(cste_2 + b_2 x) + exp(cste_3 + b_3 x)}$$

Avec : $P_{i1} + P_{i2} + P_{i3} = 1$

On peut écrire cela plus schématiquement :

$$P_j = \frac{\exp(cste_j + b_j x)}{1 + \sum_{k=2}^{m} \exp(cste_k + b_k x)} \quad \forall i$$

Avec j=1,…,m et le vecteur de paramètres ($cste_1$ b_1) est normalisé à zéro. m représente le nombre de choix.

On estime toutes ces probabilités par la méthode du maximum de vraisemblance. On en déduit normalement les effets marginaux :

$$\frac{dp_j}{dx} = \exp(cste_j + b_j x)\, b_j$$

Nb : on peut faire un lien entre les effets marginaux et les probabilités de la manière suivante (Cf. J. Scott Long et P. H. Franses) : (Avec $b_1 = 0$)

$$\frac{dp_j}{dx} = p_j \left[b_j - \sum_{i=1}^{3} b_i * p_i \right]$$

La nouveauté dans le cas multinomial est que nous pouvons calculer des ratios de probabilité appelés ratios de risques ou *odds ratios* :

$$\frac{p_j}{p_1} = \exp(cste_j + b_j x)$$

On remarque que ces *odds ratios* sont indépendants des autres alternatives.

SECTION 3 : APPLICATION ECONOMETRIQUE

On reprend l'exemple de la section 1 en ne gardant qu'une seule variable exogène : **NIVEAU**. On a donc le modèle suivant :

$$P(choix = 1) = \varphi(cste_1 + b_1 niveau)$$

$$P(choix = 2) = \varphi(cste_2 + b_2 niveau)$$

$$P(choix = 3) = \varphi(cste_3 + b_3 niveau)$$

NB : le choix 1 n'est pas estimé. Il sert de référence. Les paramètres sont égaux à zéro par définition.

Profil réponse discrète			
Index	CHOICE	Fréquence	Pourcentage
0	1	222	22.20
1	2	251	25.10
2	3	527	52.70

Mesures du critère qualificatif de lissage		
Mesure	Valeur	Formule
Likelihood Ratio (R)	446.6	2 * (LogL - LogL0)
Upper Bound of R (U)	2197.2	- 2 * LogL0
Aldrich-Nelson	0.3087	R / (R+N)
Cragg-Uhler 1	0.3602	1 - exp(-R/N)
Cragg-Uhler 2	0.4052	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.393	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.3869	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.2033	R / U
Veall-Zimmermann	0.4492	(R * (U+N)) / (U * (R+N))
N = # d'observations, K = # de régresseurs		

multinomial logit education choice

The MDC Procedure

Résultats estimés des paramètres					
Paramètre	DDL	Valeur estimée	Erreur type	Valeur du test t	Approx. de Pr > |t|
cst2	1	2.5064	0.4184	5.99	<.0001
b2	1	-0.3088	0.0523	-5.91	<.0001
cst3	1	5.7699	0.4043	14.27	<.0001
b3	1	-0.7062	0.0529	-13.34	<.0001

On en déduit les probabilités (p_i) et les effets marginaux (em_i) pour chaque choix au point médian (i.e. niveau=6.64) :

Effets marginaux (au point median)

Obs.	p1	p2	p3	em1	em2	em3
1	0.18101	0.28558	0.53341	0.084148	0.044574	-0.12872

Avec un niveau médian de 6.64 points (sur 13), une personne à 18% de chance de ne pas aller au collège ; 28% de chance de rester que 2 ans au collège et 53% de chance de faire les 4 ans (et ensuite lycée…..).

Interprétation des EM_i :

Une augmentation de 1 point du NIVEAU (i.e. d'être plus mauvais) augmentent les probabilités de ne pas aller au collège/de ne faire que 2 ans et réduit la probabilité de faire 4 ans au collège.

A comparer avec les 5% les plus forts i.e. le décile 5 avec niveau=2.635 :

Effets marginaux (au point 5ième décile)

Obs.	p1	p2	p3	em1	em2	em3
1	0.017766	0.096545	0.88569	0.011642	0.033452	-0.045094

La probabilité passe de 18% à 1.77% pour p_1.

Le EM1 passe de 8.4% à 1.16%. C'est-à-dire pour les meilleurs élèves une augmentation de 1 point du NIVEAU n'augmente la probabilité de ne pas aller au collège que de 1.16 point.

Remarque : la somme des probabilités est bien égale à 1 et la somme des effets marginaux à zéro.

Calcul des Odds Ratio

	P_2/P_1	P_3/P_1	P_3/P_2	P_2/P_3
ODDS ratio 6.64	1.57	2.94	1.86	.53
ODDS ratio 2.635	5.43	49.85	9.17	.10

Si Odds ratio = 1 indépendance
Si Odds ratio > 1 effet plus fréquent dans le groupe A que dans le groupe B
Si Odds ratio largement > 1 effet beaucoup plus fréquent dans le groupe A que dans le groupe B

Si Odds ratio < 1 effet moins fréquent dans le groupe A que dans le groupe B
Si Odds ratio ≈ 0, effet beaucoup moins fréquent dans le groupe A que dans le groupe B

Avec Odds Ratio ≥ 0

On peut aussi l'interpréter comme suit :

P2 = 1.57 P1

Cf. EDUCATION.SAS en TDs pour faire d'autres scénarii. Et ajouter d'autres variables exogènes.