Nouveaux modèles de choix qualitatifs prenant en compte des caractéristiques individuelles et des caractéristiques de choix

Jean Peyhardi [1,2]

[1] *Université de Montpellier, Faculté de Pharmacie, 15 avenue Charles Flahault, 34093 Montpellier, France - jean.peyhardi@gmail.com*

[2] *Virtual Plants, CIRAD, AGAP and Inria, 860 rue de St Priest, 34095 Montpellier, France*

Résumé. En économétrie, les modèles logit multinomial et logit conditionnel sont des modèles de choix qualitatifs très utilisés qui prennent en compte respectivement des caractéristiques individuelles et des caractéristiques de choix. Ils se différencient par leur paramétrisation bien qu'ils partagent la fonction de lien canonique. Cette fonction de lien se décompose en le ratio de probabilités référence et la fonction de répartition logistique. Nous proposons alors de conserver le ratio référence, approprié pour des modalités de choix qualitatives, mais de sélectionner la fonction de répartition parmi une plus grande famille, contenant par exemple celle associée à la loi de Student. Ces nouveaux modèles donnent bien souvent de meilleurs résultats que les modèles classiques et restent pour autant facilement estimables et interprétables. Ceci est vérifié sur un jeu de données classique sur les modes de transport entre Sydney et Melbourne.

Mots-clés. Choix qualitatifs, logit conditionnel, fonction de lien, matrice de design.

Abstract. In the econometric framework, multinomial and conditional logit models are very usual regression models for qualitative choices incorporating respectively individual caracteristics and choice caracteristics. They differ by their parametrization while sharing the canonical link function. This link function can be decomposed into the reference ratio of probabilities and the logistic cumulative distribution function (cdf). We propose to conserve the reference ratio, appropriate for qualitative choices, but to select the cdf among an enlarged family containing the Student cdf for instance. These new qualitative choice models often outperform logit models in terms of likelihood and error rate of classification and stay easily interpretable. This is illustrated with a benchmark dataset of travel demand between Sydney and Melbourn.

Keywords. Qualitative choices, conditional logit, link function, design matrix.

1 Les modèles logit multinomial et logit conditionnel

La variable réponse Y_i est le choix de l'individu i (avec les alternatives $j = 1, \dots, J$), x_i est le vecteur des caractéristiques individuelles (sexe, age par exemple). Dans le cas

de modèle de choix de transport, des caractéristiques dépendant du choix $\omega_{i,j}$ sont aussi utilisées, comme le coût de l'alternative j pour l'individu i par exemple. Dans la suite nous n'utiliseront plus la notation i de l'individu sans perte de généralité.

L'axiomatique de choix de Luce (1959) et le principe de maximisation de l'utilité aléatoire mènent au modèle logit définit par

$$\pi_j = \frac{\exp(\eta_j)}{1 + \sum_{k=1}^{J-1} \exp(\eta_k)}$$

pour $j = 1, \ldots, J-1$, où π_j est la probabilité $P(Y = j)$. Selon la forme du prédicteur linéaire η_j, différents modèles logit sont obtenus :

- $\eta_j = \alpha_j + x^t \delta_j$. Les caractéristique individuelles x sont utilisées avec $J-1$ différentes pentes δ_j ; c'est le modèle classique logit multinomial.
- $\eta_j = \alpha_j + \tilde{\omega}_j^t \gamma$ où $\tilde{\omega}_j = \omega_j - \omega_J$. Les caractéristiques de choix ω_j sont utilisées avec une pente commune δ ; c'est le modèle logit conditionnel introduit par McFadden (1974).
- $\eta_j = \alpha_j + x^t \delta_j + \tilde{\omega}_j^t \gamma$. Les caractéristique individuelles et celles dépendant du choix sont utilisées avec respectivement des pentes différentes et une pente commune ; c'est une combinaison des deux paramétrisations précédentes.

2 Généralisation des modèles logit multinomial et logit conditionnel

Tous les modèles classiques de régression multinomiale, décrits dans la littérature (Tutz, 2011), partagent le même type d'équation (Peyhardi et al., 2014)

$$r(\pi) = F(Z\beta)$$

spécifiée par trois composantes

- le ratio de probabilités r qui est un C^1-difféomorphisme entre le simplexe $\Delta = \{\pi \in]0,1[^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ (coin de l'hypercube) et un ouvert de l'hypercube $]0,1[^{J-1}$,
- la fonction de répartition F strictement croissante et continue,
- la matrice de design Z qui dépends des variables explicatives,

où π est le vecteur de probabilités $(\pi_1, \ldots, \pi_{J-1})^t$ et β le vecteur de paramètres.

Les trois modèles logit (définit dans la section 1) utilisent la fonction de lien canonique qui se décompose en le ratio référence

$$r_j(\pi) = \frac{\pi_j}{\pi_j + \pi_J}$$

et la fonction de répartition logistique $F(\eta) = \exp(\eta)/\{1+\exp(\eta)\}$. Ces trois modèles sont ainsi complètement spécifiés par les triplets (référence, logistique, Z) avec respectivement les trois matrices de design suivantes (correspondant aux trois paramétrisations décrites dans la section 1)

$$Z_1 = \begin{pmatrix} 1 & & & x^t & & \\ & \ddots & & & \ddots & \\ & & 1 & & & x^t \end{pmatrix},$$

$$Z_2 = \begin{pmatrix} 1 & & & \tilde{\omega}_1^t \\ & \ddots & & \vdots \\ & & 1 & \tilde{\omega}_{J-1}^t \end{pmatrix},$$

$$Z_3 = \begin{pmatrix} 1 & & & x^t & & & \tilde{\omega}_1^t \\ & \ddots & & & \ddots & & \vdots \\ & & 1 & & & x^t & \tilde{\omega}_{J-1}^t \end{pmatrix}.$$

Il s'avère que dans la fonction de lien canonique, seul le ratio référence est nécessaire pour des modalités de choix non-ordonnées tandis que la fonction de répartition logistique n'est pas indispensable (Peyhardi et al., 2014). Nous proposons alors une nouvelle classe de modèles de régression appropriée pour des choix qualitatifs contenant les triplets (référence, F, Z_i) pour $i = 1, 2, 3$ avec F pouvant être sélectionnée parmi les fonction de répartitions suivantes par exemple : logistique, Gauss, Laplace, Gumbel, Gompertz et Student (avec différent degré de liberté $\nu \in \mathbb{R}_+^*$).

La lourdeur des queues des distributions de Student peut amélioré considérablement l'ajustement du modèle et réduire l'erreur de classification (Peyhardi et al., 2014). Pour autant les paramètres de ces modèles de régressions restent facilement interprétables puisque $\pi_j/\pi_J = F(\eta_j)/\{1 - F(\eta_j)\}$ est une fonction strictement croissante de η_j (égale à la fonction exponentielle dans le cas de la fonction de répartition logistique). Enfin ces paramètres sont aussi facilement estimables en utilisant l'algorithme des scores de Fisher.

2.1 Algorithme des scores de Fisher

La procédure est décrite de manière générale par Peyhardi et al. (2014) pour les quatre ratios : référence, adjacent, cumulatif et séquentiel. Nous décrivons ici uniquement la procédure pour le ratio référence, dont les calculs se simplifient puisque le ce ratio est une partie de la fonction de lien canonique. Pour cela il suffit simplement de décrire, pour une seule observation, le score

$$\frac{\partial l}{\partial \beta} = Z^T \, D \, (y - \pi), \tag{1}$$

et la matrice d'information de Fisher

$$\mathrm{E}\left(\frac{\partial^2 l}{\partial\beta^T\partial\beta}\right) = -Z^T\ D\ \mathrm{Cov}(Y)\ D\ Z, \tag{2}$$

avec

$$D = \mathrm{diag}_{1\leq j\leq J-1}\left[\frac{f(\eta_j)}{F(\eta_j)\{1-F(\eta_j)\}}\right],$$

et f la fonction de densité associée à F.

3 Application au choix du mode de transport

Le jeu de données, utilisé par Greene (2003), contient les informations de 210 ménages sur leur choix de mode de transport entre Sydney et Melbourne (Australie) pris parmi les $J = 4$ modalités suivantes : avion (1), bus (2), train (3) et voiture (4). Deux caractéristiques individuelles sont utilisées : le revenu du ménage x^1 et le nombre de personnes voyageant dans le ménage x^2. Trois caractéristiques de choix sont utilisées : le temps de transfert ω_j^1 ($\omega_4^1 = 0$ pour la voiture), le temps de trajet ω_j^2 et le coût ω_j^3. L'échantillon est stratifié afin de rééquilibrer les proportions de chaque mode de transport puisque la vraie population est principalement composée d'automobilistes.

Table 1: Inférence des six modèles (référence, logistique, Z_i) et (référence, Student, Z_i).

	$F = \text{logistic}$			$F = \text{Student}_{\nu=1}$		
	Z_1	Z_2	Z_3	Z_1	Z_2	Z_3
α_1	0.9435	4.7399	6.0351	0.6438	13.7305	15.2387
α_2	1.978	3.3062	4.5045	1.9446	7.955	7.3668
α_3	2.4938	3.9532	5.5735	2.354	8.6827	9.5013
δ_1^1	0.003544		0.007481	0.00496		0.02897
δ_2^1	-0.03033		-0.0209	-0.02581		-0.004521
δ_3^1	-0.05731		-0.05923	-0.06026		-0.0571
δ_1^2	-0.6006		-0.9224	-0.4946		-1.0745
δ_2^2	-0.9404		-0.1478	-1.0836		0.7765
δ_3^2	-0.3098		0.2163	-0.2489		0.7777
γ^1		-0.09689	-0.1012		-0.2548	-0.2597
γ^2		-0.003995	-0.004131		-0.00426	-0.003878
γ^3		-0.01391	-0.008667		-0.01849	-0.01746
$\mathcal{L}$	-253.34	-192.89	-172.47	-253.36	-169.79	-159.02
BIC	554.8	417.86	409.1	554.84	**371.66**	382.21
Error	53.33%	26.19 %	27.14 %	54.29 %	22.38%	21.9%

Nous avons estimés les trois modèles logit classiques ainsi que d'autres modèles références (c.a.d. avec F non logistique). Les meilleurs résultats ont été obtenus avec les modèles (référence, Student$_{\nu=1}$, Z_i) qui se démarquent nettement des modèles logit avec pourtant les mêmes nombres de paramètres (par exemple la log-vraisemblance vaut $\mathcal{L} = -192.89$ pour logistique contre $\mathcal{L} = -169.79$ pour Student avec la même matrice de design Z_2). En particulier le modèle (référence, Student$_{\nu=1}$, Z_2) est le meilleur au sens du BIC ; voir Table 1.

Les proportions entre les paramètres de ce modèle (lorsque ces paramètres sont significatifs) sont relativement bien conservées comparées au cas logistique ($\alpha_1/\alpha_2 \simeq 1.43$ pour logistique et $\alpha_1/\alpha_2 \simeq 1.72$ pour Student par exemple). Une différence intéressante concerne l'estimation de la pente γ^1 puisque $\gamma^1/\gamma^2 \simeq 24$ pour logistique tandis que $\gamma^1/\gamma^2 \simeq 60$ pour Student. Cela signifie que, selon le modèle (référence, Student$_{\nu=1}$, Z_2), le temps de transfert a un effet beaucoup plus important sur le mode de transport choisit contrairement à ce préconise le modèle logit conditionnel classique (référence, logistique, Z_2). Dans cette étude les gens privilégient la voiture en grande partie pour sa praticité.

En conclusion, le choix de F ne change pas à priori les signes des paramètres significatifs mais peut changer les proportions entre eux. Une caractéristique peut alors s'avérer avoir plus d'effet que prévu par le modèle logit. De plus une meilleur fonction de répartition F peut réduire l'erreur de classification et donc améliorer les prédictions.

Bibliographie

[1] Luce, R. D. (1959), *Individual choice behavior: a theoretical analysis*, John Wiley & Sons.

[2] Greene, W. H. (2003), *Econometric analysis*, New Jersey, 729–735.

[3] McFadden, D. (1974), Conditional logit analysis of qualitative choice analysis, *Frontiers in Econometrics*, 105–142.

[4] Peyhardi, J., Trottier, C. and Guédon, Y. (2014), A new specification of generalized linear models, *arXiv preprint arXiv:1404.7331*.

[5] Tutz, G. (2012), *Regression for categorical data*, Cambridge University Press.