New qualitative choice models incorporating individual and choice characteristics

Jean Peyhardi

New qualitative choice models incorporating individual and choice characteristics

Jean Peyhardi[1]

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

E-mail for correspondence: jean.peyhardi@gmail.com

Abstract: In the econometric framework, multinomial and conditional logit models are the most usual regression models for qualitative choices. 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 model; Link function; Design matrix.

1 Multinomial and conditional logit models for qualitative choices

Let Y_i be the response variable corresponding to the choice of individual i (with alternatives $j = 1, \dots, J$) and x_i be the vector of individual characteristics (e.g. sex, age). In the context of travel demand, some choice characteristics $\omega_{i,j}$ are also used, such as the cost of alternative j for individual i. In the following we will suppress the individual subscript i without loss of generality.

Luce's choice axiom (Luce, 1959) and the principle of random utility maximisation lead to the logit model defined by

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

This paper was published as a part of the proceedings of the 30th International Workshop on Statistical Modelling, Johannes Kepler Universität Linz, 6–10 July 2015.

for $j = 1, \ldots, J-1$, where $\pi_j = P(Y = j)$. Depending on the form of the linear predictors η_j, we obtain different logit models:

- $\eta_j = \alpha_j + x^{\mathrm{T}}\delta_j$. Individual characteristics x are used with $J-1$ different slopes δ_j. This is the classical multinomial logit model.
- $\eta_j = \alpha_j + \tilde{\omega}_j^{\mathrm{T}}\gamma$ where $\tilde{\omega}_j = \omega_j - \omega_J$. Choice characteristics ω_j are used with common slope δ. This is the conditional logit model introduced by McFadden (1974).
- $\eta_j = \alpha_j + x^{\mathrm{T}}\delta_j + \tilde{\omega}_j^{\mathrm{T}}\gamma$. Individual and choice characteristics are used with respectively different slopes and common slope. This is a combination of the two previous parametrizations.

2 Generalisation of multinomial and conditional logit models

All the classical regression models for categorical data (Tutz, 2012) share the generic equations (Peyhardi et al., 2014)

$$r_j(\pi) = F(\eta_j)$$

for $j = 1, \ldots, J-1$, where r is a C^1-diffeomorphism from the simplex $\Delta = \{\pi \in (0,1)^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ (corner of hypercube) to an open subset of the hypercube $(0,1)^{J-1}$, π is the vector of probabilities $(\pi_1, \ldots, \pi_{J-1})^{\mathrm{T}}$ and F is a continuous and strictly increasing cdf.
Let us remark that the three logit models (defined in Section 1) share the canonical link function, specified by the reference ratio

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

and the logistic cdf

$$F(\eta) = \frac{\exp(\eta)}{1 + \exp(\eta)}.$$

Therefore, these three logit models are specified by the (reference, logistic, Z) triplet with different design matrices Z respectively equal to

$$Z_1 = \begin{pmatrix} 1 & & & x^{\mathrm{T}} & & \\ & \ddots & & & \ddots & \\ & & 1 & & & x^{\mathrm{T}} \end{pmatrix}, \quad Z_2 = \begin{pmatrix} 1 & & & \tilde{\omega}_1^{\mathrm{T}} \\ & \ddots & & \vdots \\ & & 1 & \tilde{\omega}_{J-1}^{\mathrm{T}} \end{pmatrix},$$

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

The reference ratio is mandatory for non-ordered choices whereas the logistic cdf is not (Peyhardi et al., 2014). We thus propose a new class of regression models appropriate for qualitative choices defined by (reference, F, Z_i) models ($i = 1, 2, 3$) where the cdf F can be selected among e.g. the logistic, Gaussian, Laplace, Gumbel, Gompertz, and Student cdfs (with different degrees of freedom $\nu \in \mathbb{R}_+^*$). The heavy tails of Student distributions may markedly improve the model fit and reduce the classification error rates (Peyhardi et al., 2014). Parameter estimates stay interpretable since

$$\frac{\pi_j}{\pi_J} = \frac{F}{1-F}(\eta_j)$$

is strictly increasing with η_j (we have $\pi_j/\pi_J = \exp(\eta_j)$ in the case of the logistic cdf). Finally, this family of reference models for qualitative choices is easily estimated using the standard Fisher's scoring algorithm.

2.1 Fisher's scoring algorithm for reference models

Let us remark that the Fisher's scoring algorithm is simplified in the particular case of the reference ratio (compared to the adjacent, cumulative and sequential ratios) which is a part of the canonical link function. Using the chain rule we obtain the score

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

and the Fisher's information matrix

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

where

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

and f is the density function. Remarking that $f = F(1 - F)$ for the logistic distribution, the Fisher's scoring algorithm turns out to be, in this particular case, the algorithm for the canonical link function.

3 Application to travel mode demand

The dataset, used by Greene (2003), contains 210 observations of choice among $J = 4$ travel modes between Sydney and Melbourn (Australia): *air* (1), *bus* (2), *train* (3), and *car* (4). The two individual characteristics are the household income x^1 and the number of people travelling x^2. The three choice characteristics are the terminal time ω_j^1 ($\omega_4^1 = 0$ for car), the amount of time spent traveling ω_j^2 and the in-vehicle cost ω_j^3. The sample is choices

based so as to balance it among the four choices knowing that the true population is dominated by drivers.
The three logit models and other reference models (i.e. $F \neq$ logistic) were estimated. Best results were obtained with (reference, Student$_{\nu=1}$, Z_i) models that markedly outperformed logit models; see Table 1 (we have for instance $l = -192.89$ for logistic versus $l = -169.79$ for Student with the same parametrization Z_2). The (reference, Student$_{\nu=1}$, Z_2) is the best model according to BIC. The proportions between parameters when significant are approximatively conserved comparing logistic and Student models ($\alpha_1/\alpha_2 \simeq 1.43$ for logistic cdf and $\alpha_1/\alpha_2 \simeq 1.72$ for Student cdf for instance). The interesting difference concerns estimate of the slope γ^1 since $\gamma^1/\gamma^2 \simeq 24$ for logistic cdf and $\gamma^1/\gamma^2 \simeq 60$ for Student cdf. In the Student case, the terminal time has a stronger impact on the travel mode choice.
Finally, the selection of F does not a priori change the sign of parameters but may change the proportion between them. Moreover it may reduce the classification error rate and thus increase the precision of predictions.

TABLE 1. Parameters estimates for six reference regression models.

	$F =$ logistic			$F =$ 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%

References

Luce, R. (1959). *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons.

Greene, W. (2003). *Econometric Analysis*. New Jersey: Prentice Hall. 729 – 735.

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

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

Tutz, G. (2012). *Regression for Categorical Data*. Cambridge University Press.