A Bayesian Mixed Multinomial Logit Model for Partially Microsimulated Data on Labor Supply

Un Modello Bayesiano Logit Multinomiale Misto per Dati Parzialmente Microsimulati sull'Offerta di Lavoro

Cinzia Carota and Consuelo R. Nava

Abstract We focus on the determinants of labor choices in the presence of partially microsimulated data and discrete choice sets not identical for all agents under examination. The independence of irrelevant alternative assumption is thus discussed and the variability of the available choice set is taken into account. By comparing a Bayesian mixed multinomial logit model to a model without random effects, we show how the above described scenario affects labor choices made by single females and females within couples when the same discrete choice set is assigned to both individuals in each couple and the partner's choice is known.

Abstract *Si studiano le determinanti delle scelte di lavoro in presenza di dati parzialmente microsimulati e di insiemi discreti di scelte non identici per tutti gli agenti in esame. Viene pertanto discussa l'assunzione di indipendenza dalle alternative irrilevanti e viene tenuta in conto la variabilità dell'insieme delle scelte disponibili. Attraverso un modello bayesiano logit multinomiale a effetti misti comparato a uno privo di effetti aleatori, si mostra come questo scenario impatta sull'analisi delle scelte lavorative delle donne single oppure inserite in una coppia in cui ambedue gli individui sono esposti allo stesso insieme di scelte ed è nota la scelta del partner.*

Key words: Bayesian mixed multinomial logit model, independence of the irrelevant alternatives assumption, labour supply, random discrete choice set.

Cinzia Carota
Università degli Studi di Torino, Via Verdi 8, Torino, e-mail: cinzia.carota@unito.it

Consuelo R. Nava
Università della Valle d'Aosta, Strada Cappuccini 2, Aosta, e-mail: c.nava@univda.it

1 Introduction

Investigating the determinants of individual choices via random utility models (RUMs) [13, 15] is nowadays a common practice. RUMs describe the agent preference scheme in terms of the utility assigned to each discrete choice option, in a set of mutually exclusive alternatives (choice set). RUMs define a mapping from observed individual and/or choice characteristics into preferences with challenging theoretical and empirical statistical implications. The latter are investigated in various research fields, among which psychology and economics are by far the most important.

Recently, also policy evaluations take advantage of RUMs [3]. Instead of mere comparisons between events before and after a policy implementation, suitable RUMs are combined with microsimulation methods to anticipate, simulate and estimate the effects of socio-economic interventions. Such methods can simulate changes caused not only by hypothetical policies, but also by individual behaviours. In [4, 7] these tools are jointly used to conduct a "controlled experiment" in order to predict effects of tax and benefit reform interventions by using micro-data from national household surveys.

In this context, as a result of the microsimulation of certain fiscal variables and/or a sampling procedure applied to the available alternative options [1], it is quite common that the choice set does not exhibit the required homogeneity across decision makers (households). Even if the latter face the same number of job types (defined on the basis of a discretization of the weekly working hours in intervals, hereafter referred to as classes), microsimulation needs for each decision maker a random selection of a specific amount of working hours within each class, in order to simulate net household incomes, taxes and benefits. This implies that the i^{th} choice option refers to the i^{th} class of weekly working hours, but each household makes his decision by comparing his punctual amounts of working hours and other characteristics of jobs included in his own choice set. In this study, gross and net wage rates, given the amount of working hours, are computed via EUROMOD, a static microsimulation model [10] for tax and benefits, while the remaining variables are based on the Italian Survey on Household Income and Wealth (SHIW)[1]. The resulting partially microsimulated database contains information on households (e.g. singles or couples), while the required sampling procedure for microsimulation produces eight distinct choice sets $\{\mathscr{C}_h\}_{h=1}^{8}$, each one defined by 10+1 jobs with a specific amount of working hours[2]. Formally, household j, with $j = 1,\ldots,J$, is assigned the h^{th} choice set when $\mathscr{C}_j = \mathscr{C}_h$. In what follows, we distinguish the variables available in such database in two groups: variables directly introduced in the analysis as explanatory variables, and variables used to create groups of individuals as homogeneous as possible by means of a preliminary cluster analysis. We focus only on

[1] Also tax and benefits are, therefore, simulated according to the current fiscal system. See [7] for further details.

[2] Each choice set is composed by jobs with a different and increasing amount of working hours. For instance, the first choice set, $\mathscr{C}_1$, proposes jobs with 0, 1, 9, 17, 25, 33, 41, 49, 57, 65, 73 working hours, while the last one, $\mathscr{C}_8$, proposes with 0, 8, 16, 24, 32, 40, 48, 56, 64, 72, 80 working hours.

labour choices made by single and non-single females, labelled female-single and female-couple.

Due to the particular structure of the available data, this article discusses the validity of the independence of irrelevant alternatives (IIA) assumption[3] associated with widely used RUMs, such as logit, conditional logit and multinomial logit models, and tries to overcome it by incorporating the decision maker heterogeneity. The main contribution is a statistical model, specifically a Bayesian mixed multinomial logit model (MMLM) [5], able to address both the violation of the IIA assumption and the just described choice set variability. In recent labour supply studies (see [6, 7] and references therein) the bias induced by the discretization of weekly working hours, and the random selection of the choice set are unaddressed problems yet[4].

2 Methods, data and results

RUMs assume a utility maximization process in order to select one of the available alternative options. For each agent j, with $j = 1,\dots,n$, with choice set $\mathscr{C}_j = \mathscr{C}_h$, we define a random variable describing his utility U_{ij} for each alternative c_i^h in $\mathscr{C}_h = \{c_1^h, c_2^h, \dots, c_I^h\}$. We assume, for every $j, i = 1,\dots,I$ and $h = 1,\dots,H$, the conditional distribution $U_{ij}|\mathscr{C}_h \sim f_{ih}(\cdot|V_{ij})$, where V_{ij} is the ground truth utility or the score assigned to each c_i^h in $\mathscr{C}_h$ [2]. In particular, $U_{ij}|\mathscr{C}_h : c_i^h \to \mathbb{R}$ and we assume $\mathbb{E}[U_{ij}] = V_{ij}$.

In our case, agents are provided with choice sets with the same cardinality, I, but made up of different alternatives across decision makers. The j^{th} agent decision process defines a permutation τ^j of $\{c_1^h, c_2^h, \dots, c_I^h\}$ such that a linear order can be defined $[c_{\tau^j(1)} \succ c_{\tau^j(2)} \succ \dots \succ c_{\tau^j(I)}]$. The latter manifests individual preferences to which correspond an equivalent order of the random utilities $\boldsymbol{U}_j = \{U_{1j}, U_{2j}, \dots, U_{Ij}\}$ such that

$$\mathsf{Pr}(c_{\tau^j(1)} \succ \dots \succ c_{\tau^j(I)} | \boldsymbol{V}_j = \{V_{1j}, \dots, V_{Ij}\}) = \mathsf{Pr}(U_{\tau^j(1)} > \dots > U_{\tau^j(I)}). \quad (1)$$

Under RUMs, conditionally on h, every U_{ij} is given by the sum of V_{ij} and a stochastic (unobserved) component, ε_{ij}, i.e. $U_{ij} = V_{ij} + \varepsilon_{ij}\ \forall\ i = 1,\dots,I;\ j = 1,\dots,n$. Therefore, the j^{th} agent selects the alternative c_i^h in the choice set $\mathscr{C}_j$, if and only if $U_{ij} > U_{kj} \forall\, k = 1,\dots,I; k \neq i$ where $\varepsilon_{ij} = U_{ij} - V_{ij}$ is a random variable (r.v.) whose mean is 0. In turn, $V_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta}$ where $\mathbf{x}_{ij}$ is a $r \times 1$ vector of observed explanatory variables (for individual j and choice i), and $\boldsymbol{\beta}$ denotes a $r \times 1$ vector of fixed effects.

[3] The IIA assumption has been first used in the Luce's Axiom of Choice [11] and postulates that, when estimating the probability to select a job across a particular slate of alternatives in $\mathscr{C}$, the likelihood of choosing job a over job b will not change based on whether a third job $c \notin \mathscr{C}$ is present.

[4] Only few contributions on voters, as for instance [8], address these issues with $|\mathscr{C}_j| \neq |\mathscr{C}_i|$ for some $j \neq i$, i.e. when the cardinality of choice sets is different across decision makers.

The stochastic component ε_{ij}, instead, represents subjective noises and accommodates different sources of uncertainty like unobservable characteristics, unobservable variations in individual utilities, measurement errors and functional misspecification [12].

Assuming i.i.d. standard Gumbel or Extreme Value Type I errors, we obtain the well known multinomial logit model (MLM) [14]. To avoid the IIA irrealistic assumption and to model household heterogeneity we introduce in the expected utility V_{ij} a random component $\mathbf{z}'_{ij}\boldsymbol{\gamma}_{ij}$, where $\boldsymbol{z}_{ij}$ is a $s \times 1$ design vector (assumed to be known) and $\boldsymbol{\gamma}_{ij}$ is a vector of s individual-specific and/or choice-specific random effects. In particular, here we exploit the individual-specific information stored in the variables not represented in $\mathbf{x}'_{ij}$ by performing a suitable hierarchical cluster analysis to assign each household j to a cluster k_j. Then, a random component $\alpha_j k_j$ is included in V_{ij}, with α_j denoting the random effect associated to cluster id k_j. Similarly, we introduce a second individual-specific random effect, δ_j, for female-couple, to make the choice made by individual j dependent on the choice made by her partner in the couple, labelled p_j. Notice that both individuals in the couple are assigned the same choice set. Finally, we take into account the above described heterogeneity of choice sets by including a random effect, η_j, of the choice set $\mathscr{C}_j$, to which decision maker j is assigned. Hence, in the more general case, the probability π_{ij} to select alternative i by agent j, for $\mathbf{z}_{ij} = \{k_j, p_j, \mathscr{C}_j\}$ and $\boldsymbol{\gamma}_{ij} = \{\alpha_j, \delta_j, \eta_j\}$, can be rewritten as

$$\pi_{ij} = \mathsf{Pr}(Y_j = i | \mathscr{C}_j = \mathscr{C}_h) = \frac{\exp\{\mathbf{x}'_{ij}\boldsymbol{\beta} + \alpha_j k_j + \delta_j p_j + \eta_j \mathscr{C}_j\}}{\sum_{i=1}^{I} \exp\{\mathbf{x}'_{ij}\boldsymbol{\beta} + \alpha_j k_j + \delta_j p_j + \eta_j \mathscr{C}_j\}} \tag{2}$$

where Y_j is a random variable that takes values between 1 and I, the cardinality of the choice set $\mathscr{C}_h$, $h = 1, \ldots, 8$. The probability in eq. (2) can be embedded in the following hierarchy:

$$Y_j \sim \mathsf{Multinom}(1, \boldsymbol{\pi}_j) \;\; \forall\, j = 1, \ldots, n \tag{3}$$

$$\mathsf{Logit}(\pi_{ij}) = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\boldsymbol{\gamma}_{ij} \tag{4}$$

$$\boldsymbol{\beta} \sim \mathsf{N}(\mu_{\boldsymbol{\beta}}, V_{\boldsymbol{\beta}}), \quad \boldsymbol{\gamma}_{ij} \sim \mathsf{N}(\mathbf{0}, V_{\boldsymbol{\gamma}}), \quad V_{\boldsymbol{\gamma}} \sim \mathsf{IW}(\Psi, \nu), \tag{5}$$

where $\boldsymbol{\pi}_j = \{\pi_{1j}, \ldots, \pi_{Ij}\}$ is a vector including the "success" probabilities for each alternative in the choice set $\mathscr{C}_j$. We assign an Inverse Wishart (IW) prior distribution to the random effect (co)variance matrix $V_{\boldsymbol{\gamma}}$ while fixed effects are assigned a normal distribution. Such a hierarchical model is now applied to study whether or not the probability of choosing job i, given its features and household characteristics, is the same in the two groups of female-single and female-couple. Estimates are based on simulated (tax, benefit, gross income) and real data at a micro level, respectively from EUROMOD and SHIW[5].

[5] Here variables microsimulated by EUROMOD are based on real data provided by the Bank of Italy from the SHIW-1998 and on the 1998 Italian fiscal policy. The microsimulation model enables to calculate, in a comparable way, the effects of taxes and benefits on household incomes and work incentives.

Available data consist of 291 observed female-single and 2955 female-couple. Agents are aged between 20 and 55, neither retired nor students. They can choose among 10 different types of jobs (i.e. $i = 1, \ldots, 10$) and the non labour-market participation (indexed as job 0)[6]. Type of jobs are defined on the basis of a discretization of the weekly working hours 0, 1–8, 9–16, 17–24, 25–32, 33–40, 41–48, 49–56, 57–64, 65–72, 73–80. Eight distinct choice sets $\{\mathscr{C}_h\}_{h=1}^{8}$, given the sampled weakly working hours, group singles and couples, and a new variable stores the choice set id h when $\mathscr{C}_j = \mathscr{C}_h$[7]. A hierarchical cluster analysis for each sub-population defines the cluster id for female-single and female-couple[8]. This implies that we ignore uncertainty about clusters resulting from this analysis, but we are currently working to implement more sophisticated grouping techniques, as fuzzy clustering. Variables used as predictors are the ones typically used in the literature: weekly hours of work, gross wages, age, son, choice set, taxes and benefits. Only the last two variables and gross wages were simulated with EUROMOD[9].

A Markov Chain Monte Carlo method with a block Gibbs sampling algorithm is implemented to estimate model coefficients using the `R` package `MCMCglmm` [9]. Algorithms are run for 30000 iterations, with a burn-in phase of 6000 and a thinning interval equal to 10. Hyperprior parameters were set to be: $\mu_{\boldsymbol{\beta}} = 0$, $V_{\boldsymbol{\beta}} = \mathbf{I}_4 \cdot 10^{10}$, with $\mathbf{I}_4$ denoting a 4×4 identity matrix. The residual covariance matrix was $\frac{1}{I \cdot 10^2} \cdot (\mathbf{I}_{I-1} + \mathbf{U})$, where $\mathbf{I}_{I-1}$ is a $(I-1) \times (I-1)$ identity matrix and $\mathbf{U}$ is a $(I-1) \times (I-1)$ unit matrix [9], given I as the number of possible choices. For the inverse-Wishart prior, Ψ was set to be equal to $\mathbf{I}_{I-1}$. Standard diagnostic tools confirmed the convergence of runs. Models did not include a global intercept, hence the first 10 estimated coefficients represented actual job type specific intercepts compared to the non-working alternative. Point estimates, under the proposed Bayesian MMLM compared to a MLM, are set out in Table 1.

The main improvements in the results under the MMLM can be appreciated both in the sign of the 10 choice coefficients, counterintuitive under the MLM, and in the large number of HPD intervals bounded away from zero.

References

1. Aaberge, R., Colombino, U., Strøm, S.: Labor Supply in Italy An Empirical Analysis of Joint Household Decisions, with Taxes and Quantity Constraints, Journal of Applied Econometrics, **14**, pp. 403–422 (1999)
2. Azari, H., Parks, D., and Xia, L., Random utility theory for social choice. In Advances in Neural Information Processing Systems pp. 126–134 (2012)

[6] In such a way, couples have 121 mixed alternatives among male and female.

[7] The number of agents from each sub-population for each choice set $h = 1, .., 8$ is 43, 43, 43, 42, 46, 52, 54, 43 for female-single and 388, 355, 356, 359, 353, 390, 382, 372 for female-couple.

[8] The hierarchical cluster analysis, based on the Ward's method and the Euclidean distance, identifies 8 groups both for female-single (with cardinality 44, 45, 69, 71, 20, 95, 19, 3) and female-couple (with cardinality 582, 557, 169, 646, 181, 367, 72, 381).

[9] Other details on data description or partial simulation can be found in [6, 7].

Table 1 Bayesian estimates for female-single and female-couple under MLM and MMLM; "***", "**", "*" and "." indicate respectively that the corresponding 99.9%, 99%, 95% and 90% HPD intervals are bounded away from zero

	Female-single		Female-couple	
	MLM	MMLM	MLM	MMLM
c_1	0,236	0,479 ***	-0,076	0,134 ***
c_2	0,153	0,427 **	-0,092	0,083 *
c_3	0,086	0,398 ***	-0,125	0,104 *
c_4	0,192	0,411 **	-0,044	0,166 ***
c_5	0,167	0,483 ***	-0,107	0,089 *
c_6	0,164	0,523 ***	-0,087	-0,001
c_7	0,134	0,482 ***	-0,106	0,034
c_8	0,292	0,383 **	-0,055	0,078 .
c_9	0,330	0,535 ***	-0,142	0,067
c_{10}	0,225	0,311 **	-0,122	0,141 **
hours	0,002	0,003 ***	0,010 ***	0,003 ***
wage	0,013 ***	0,009 ***	0,008 ***	-0,001 ***
tax	0,0001 .	-0,0001 ***	-0,00021 ***	-0,00001 *
benefit	0,0003 ***	0,0003 ***		
age	-0,011 ***	-0,020 ***	-0,007 ***	-0,002 ***
son	-0,031	0,163 ***	-0,029 ***	-0,027 ***

3. Blundell, R., MaCurdy, T.: Labour supply: a review of alternative approaches. In: Ascenfelter, O., Card, D. (eds.) Handbook of Labour Economics, pp. 1559–1695. North Holland, Amsterdam (1999)
4. Bourguignon, F., Spadaro, A.: Microsimulation as a tool for evaluating redistribution policies. Journal of Economic Inequality, **4**, pp. 77–106 (2006)
5. Cardell, N. and Dunbar, F.: Measuring the societal impacts of automobile downsizing. Transportation Research A, 14 423–434 (1980)
6. Colombino, U.: A new equilibrium simulation procedure with discrete choice models. International Journal of Microsimulation, **6**(3), pp. 25–49 (2013)
7. Colombino, U.: Five Crossroads on the Way to Basic Income. An Italian Tour. Italian Economic Journal, **1**, 353–389 (2015)
8. Gallego, M., Schofield, N., McAlister, K., Jeon, J. S.: The variable choice set logit model applied to the 2004 Canadian election. Public Choice, 158(3-4), 427-463 (2014)
9. Hadfield, J. D.: MCMC methods for Multi-Response Generalized Linear Mixed Models. The MCMCglmm R Package. Journal of Statistical Software, **33**(2), pp. 1–22 (2009)
10. Immervoll, H., O'donoghue, C., Sutherland, H.: An introduction to EUROMOD. Microsimulation Unit, Department of Applied Economics, University of Cambridge (1999)
11. Luce R. D.: On the possible psychophysical laws. Psychological Review, pp.66–81, (1959)
12. Manski, C. F.: The Structure of Random Utility Models, Theory and Decision, 8, 229–54 (1977)
13. Marschak, J.: Binary choice constraints on random utility indications. In K. Arrow, ed. Stanford Symposium on Mathematical Methods in the Social Sciences, pp. 312–329. Stanford University Press, Stanford (1960)
14. McFadden, D.: Conditional logit analysis of qualitative choice behavior. In P. Zarembka, ed., Frontiers in Econometrics, pp. 105–142. Academic Press, New York (1974)
15. Train, K. E.: Discrete Choice Methods with Simulation. Cambridge University Press (2003)