A LOGIT MODEL WITH ENDOGENOUS EXPLANATORY VARIABLES AND NETWORK EXTERNALITIES

Louis de Grange (corresponding author), Felipe González
Industrial Engineering Department, Diego Portales University, Santiago, Chile.
e-mail: louis.degrange@udp.cl, felipe.gonzalezr@udp.cl

Ignacio Vargas
Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, MA 02139, USA.
e-mail: ijv@mit.edu

Rodrigo Troncoso
Centro de Políticas Públicas (CPP), Facultad de Gobierno, Universidad del Desarrollo, Santiago, Chile.
e-mail: rtroncoso@udd.cl

ABSTRACT

A novel logit model is presented that explicitly includes endogeneity in explanatory variables whose values depend on individual choice decisions that involve network externalities or social interactions such as those impacting road congestion or public transport comfort and convenience. The proposed specification corrects for this particular type of endogeneity. The model is derived from a linearly constrained maximum entropy optimization problem that incorporates the network externalities or social interactions causing the endogeneity. It is validated through simulations.

Keywords: logit model; endogeneity; bias; network externalities; social interactions; fixed point; maximum entropy.

JEL Codes: C13, C5, C6, R4.

1. INTRODUCTION

Logit multinomial discrete choice models are frequently used in marketing to model consumer preferences, and are also employed in transport system planning to represent trip demand and ground use. They are traditionally built from random utility models and estimated using maximum likelihood (McFadden, 1974; Ortúzar, 1982, 1983; Train, 2003; Ortúzar and Willumsen, 2011). Alternatively, however, they have been derived as the solution to certain constrained entropy maximization problems in which the Lagrange multipliers of the constraints are the model's parameters (Anas, 1983; Donoso and De Grange, 2010; Donoso et al. 2011).

In discrete choice models, interaction between supply and demand (Berry et al., 1995; Petrin and Train, 2010) and the omission of variables unobservable to the researcher (Villas-Boas and Winer, 1999) may cause endogeneity. In both cases, the problem can be handled by introducing instrumental variables into the estimation. The existence of endogeneity has also been traced to the definition of the choice set facing the individual (Haab and Hicks, 1997; Louviere et al., 2005).

In this study, endogeneity is considered to be the result of network externalities or social interactions of the type peculiar to transport systems. More specifically, it emerges from the fact that the attractiveness of a particular alternative depends on the number of persons choosing it (Yamins et al., 2003; Blumenfeld-Lieberthal, 2009; Bogart, 2009). But whereas the attractiveness of a technology product, for example, often varies positively with user numbers, the choice of a transport mode can be negatively impacted by user numbers because it is influenced negatively by congestion. To deal with this phenomenon we propose a novel specification for logit models that corrects the endogeneity bias introduced by network externalities and social interactions.

The approach that will be taken is based on entropy maximization with explicit incorporation of the endogeneity caused by network externalities or social interactions into logit multinomial discrete choice models. An equivalent maximum entropy optimization problem is formulated in which the explanatory variables facing individuals depend on the decisions they make (e.g., with private transport, trip time depends on congestion; with public transport it depends on wait time or crowding at bus stops or metro stations). This interdependence produces endogeneity in the model's explanatory variables, whose values will therefore depend on the individuals' decisions.

The result of this approach is an alternative version of the logit multinomial model that has the functional form of a fixed-point equation in which the choice probabilities depend on themselves. Estimation is by maximum likelihood, and no additional variables (e.g., instruments) other than the original ones in the model are required, a major advantage of the proposed method.

The remainder of this article is organized into four sections. Section 2 introduces the theoretical framework for the proposed methodology and briefly surveys the literature on endogeneity in discrete choice models. Section 3 formulates the Logit model with endogenous explanatory variables. Section 4 presents a numerical example using simulations. Finally, Section 5 sums up the main conclusions and contributions of this study.

2. THEORETICAL FRAMEWORK AND LITERATURE SURVEY

Logit-type discrete choice models are developed using either of two approaches. One of them is based on random utility models (McFadden, 1974; Williams, 1977; Train, 2003; Ortúzar and Willumsen, 2011) while the other involves formulating a maximum entropy optimization problem (Anas, 1983; Boyce, 2007; Hasan and Dashti, 2007; De Cea et al., 2008; Donoso and De Grange, 2010; Donoso et al. 2011; De Grange et al., 2010, 2011, 2013; Kitthamkesorn et al., 2014). Our proposed model takes the latter approach, specifying a maximum entropy optimization problem with constraints that explicitly incorporates the endogeneity phenomenon in the model's explanatory variables.

In the random utility approach, an individual *i* faced with a set of alternatives chooses the one that produces the greatest utility. Thus, the individual will choose alternative *m* when $U_i^m > U_i^{m'} \; \forall m' \neq m$. The utility function U_i^m is typically decomposed additively as $U_i^m = V_i^m + \varepsilon_i^m$, where V_i^m is a deterministic component depending on observable variables and ε_i^m a random component that is not observable. The observer does not directly observe the individuals' actual utility functions U_i^m but rather the choices they make and the attributes of each alternative, these latter constituting the definition of V_i^m. The deterministic component is typically expressed as a linear function in the attributes. Thus, if the *k*th attribute or explanatory variable faced by individual *i* in alternative *m* is defined as $x_{ki}^m \; \forall i,m,k$, then $V_i^m = \sum_k \beta_k^m x_{ki}^m$, where β_k^m are the parameters to be estimated and represent the relative weights of each attribute.

The multinomial logit (MNL) model is obtained by assuming the random component of each utility function is independent and identically Gumbel-distributed (McFadden, 1974; Ben-Akiva and Lerman, 1985; Train, 1986, 2003; Ortúzar and Willumsen, 2011). The probability that individual *i* chooses alternative *m* is then given in general terms by

$$P_i^m = \frac{e^{V_i^m}}{\sum_{m'} e^{V_i^{m'}}} \tag{1}$$

Under the second approach, on the other hand, (1) is derived by solving the following maximum entropy optimization problem:

$$\begin{aligned} &\min_{\{p_i^m\}} \quad \sum_i \sum_m p_i^m \left(\ln p_i^m - 1 \right) \\ &s.t.: \\ &\sum_m p_i^m = 1 \qquad \left(\Phi_i \right) \\ &\sum_i p_i^m x_{ki}^m = c_k^m \qquad \left(\beta_k^m \right) \end{aligned} \tag{2}$$

where variables x_{ik}^m are the measurable exogenous attributes individual *i* perceives in alternative *m*; and c_k^m are the observed values for each attribute *k* in each alternative *m*. Also, $c_k^m = \sum_i \delta_i^m x_{ik}^m$, where δ_i^m is 1 if individual *i* chooses alternative *m* and 0 otherwise. The values of x_{ki}^m and δ_i^m are traditionally obtained from surveys, measurements and calibration samples.

Anas (1983) gives a formal statement of the equivalence between the multinomial model based on random utility theory (1) and the maximum entropy problem (2). Donoso and De Grange (2010) show that an analysis of (2) yields two useful interpretations of this equivalence: first, the maximum entropy problem is consistent with the rational decisions of welfare-maximizing individuals, and second, the likelihood function of the MNL model is equal to the problem's dual. However, the equivalence is valid only if the constraints in (2) are linear, which is the case only when the attributes or variables x_{ki}^m are exogenous.

In general terms, endogeneity in discrete choice models is present when part of the deterministic utility function V_i^m is correlated with the error term ε_i^m (Berry et al., 1995; Louviere et al., 2005; Guevara and Ben-Akiva, 2009; Walker et al., 2011). When this occurs, the estimates of the parameters β_k^m in $V_i^m = \sum_k \beta_k^m x_{ki}^m$ may be inconsistent.

Endogeneity may appear for a variety of reasons, such as a model specification error due to the omission of important variables or the ability of individuals to influence the formation of the choice sets (Louviere et al., 2005). In this study we consider it to be the product of social interactions, which may be positive or negative. More specifically, the attributes of the alternatives are assumed to depend on the level of choice aggregation. For example, as more individuals choose to travel by private car, trip times may increase due to congestion. Likewise, overcrowding in buses or the metro, or wait times at stops or stations, may increase if there are many travellers at peak hours.

In addition, individuals may make decisions based on the actions of others due to incomplete information, such as occurs in herd behaviour (Banerjee, 1992) or informational cascades (Bikhchandani et al., 1992). For example, diners may avoid a restaurant if there are few customers inside, taking it as a sign of high prices or poor quality. Similarly, a bus stop with no-one or many people waiting may indicate a disruption in service and thus discourage potential riders.

In all of these cases the explanatory or attribute variables x_{ki}^m become endogenous variables of the type $x_{ki}^m = x_{ki}^m\left(t^m\right)$, where $t^m = \sum_i p_i^m$ is the total demand for alternative *m* so that $x_{ki}^m = x_{ki}^m\left(\sum_i p_i^m\right)$.

The proposed method set out in the following section is similar to the control function approach but has the advantage of not requiring that a functional form be specified, nor does it require exogenous instruments.

3. FORMULATION AND ESTIMATION OF LOGIT MODEL WITH ENDOGENOUS EXPLANATORY VARIABLES (MNLE)

3.1 Mathematical Formulation

As was explained in Section 2, we represent the network externalities or social interactions of the explanatory variables by functions of the type $x_{ki}^m = x_{ki}^m\left(t^m\right)$, where $t^m = \sum_i p_i^m$ is the total demand for alternative *m* so that $x_{ki}^m = x_{ki}^m\left(\sum_i p_i^m\right)$. We can now solve the following equivalent optimization problem, incorporating explicitly the variables $x_{ki}^m = x_{ki}^m\left(t^m\right)$ with $\lambda = 1$:

$$
\begin{aligned}
&\min_{\left\{p_i^m\right\}} \quad \sum_i \sum_m p_i^m \left(\ln p_i^m - 1\right) \\
&s.t.: \\
&\sum_i p_i^m x_{ki}^m = \sum_i \delta_i^m x_{ki}^m \qquad \forall m \quad \left(\beta_k^m\right) \\
&\sum_m p_i^m = 1 \qquad \forall i \quad \left(\Phi_i\right)
\end{aligned}
\tag{3}
$$

The optimality conditions of problem (3) are

$$
p_i^m = \frac{\exp\left(\sum_k \beta_k^m \left(x_{ki}^m + \left(p_i^m - \delta_i^m\right)\frac{dx_{ki}^m}{dp_i^m}\right)\right)}{\sum_{m'} \exp\left(\sum_k \beta_k^{m'} \left(x_{ki}^{m'} + \left(p_i^{m'} - \delta_i^{m'}\right)\frac{dx_{ki}^{m'}}{dp_i^{m'}}\right)\right)}
\tag{4}
$$

This expression is similar to the traditional logit multinomial specification except that it explicitly incorporates the phenomenon of endogeneity of the variables x_{ki}^m due to network externalities or social interactions. We thus describe it as a logit model with endogenous explanatory variables (MNLE). It is a fixed-point function in p_i^m, whose estimation runs up against two difficulties: first, solving the fixed point; and second, estimating $\frac{dx_{ki}^m}{dp_i^m}$. MNLE can be considered a member of the family of discrete choice models with social interactions and heterogeneity (Brock and Durlauf, 2001, 2006; Soetevent and Kooreman, 2007; Dugundji and Gulyás, 2008; Amador et al., 2008). In these formulations, an individual's decision may be affected by the decisions of other groups of individuals in society. When $\frac{dx_{ki}^m}{dp_i^m} = 0$, (4) reduces to the traditional MNL model but when $\frac{dx_{ki}^m}{dp_i^m} \neq 0$, a traditional MNL such as (1) would be incorrectly specified because the term $\left(p_i^m - \delta_i^m\right)\frac{dx_{ki}^m}{dp_i^m}$ present in (4)is missing. This omission results in biased estimations of the parameters β_k^m.

The missing term corrects the explanatory variables x_{ki}^m by incorporating the effect of the network externality or social interaction type of endogeneity.

3.2 Parameter Estimation

Multinomial logit and other analogous models are typically estimated by maximum likelihood. For the fixed-point model (4) just described, however, a practical complication arises due to the presence of the p_i^m term on the right-hand side of the equation. A simple way of getting around this difficulty is to estimate the model in two steps. In the first step, the term $\frac{dx_{ki}^m}{dp_i^m}$ is estimated exogenously. For example, if the variable x_{ki}^m represents trip time by private transport, the parameter $\frac{dx_{ki}^m}{dp_i^m}$ can be estimated from the road network flow-delay functions,. If the model is being used for empirical work, either additional data on the parameter value or valid instruments will be needed to obtain a consistent estimate. Yet another way is to make a conjecture and then conduct a sensitivity analysis on it to get an idea of the order of magnitude of the possible bias. A starting point must also be specified for an estimate of $p_{i,0}^m$. This can be obtained using the multinomial logit model (Raveau et al, 2011; De Grange et al., 2013), which is the equivalent of setting $\frac{dx_{ki}^m}{dp_i^m} = 0$ in (4) and provides a warm start. Thus,

$$p_{i,0}^m = \frac{\exp\left(\sum_k \beta_{k,0}^m x_{ki}^m\right)}{\sum_{m'} \exp\left(\sum_k \beta_{k,0}^m x_{ki}^{m'}\right)} \tag{5}$$

where the values of $\beta_{k,0}^m$ are the maximum likelihood estimators of the parameters. Once $p_{i,0}^m$ and $\frac{dx_{ki}^m}{dp_i^m}$ have been estimated we proceed to the second step, which is to estimate directly via maximum likelihood the parameters β_k^m of the following model:

$$p_{i,1}^m = \frac{\exp\left(\sum_k \beta_{k,1}^m \left(x_{ki}^m + \left(p_{i,0}^m - \delta_i^m\right)\frac{dx_{ki}^m}{dp_i^m}\right)\right)}{\sum_{m'} \exp\left(\sum_k \beta_{k,1}^{m'} \left(x_{ki}^m + \left(p_{i,0}^m - \delta_i^m\right)\frac{dx_{ki}^m}{dp_i^m}\right)\right)} \tag{6}$$

This process is iterated until $\hat{\beta}_{k,n}^m \approx \hat{\beta}_{k,n-1}^m \ \forall k, m$, which implies $p_{i,n}^m \approx p_{i,n-1}^m$ at the fixed point defined in (6). For the case where $\frac{dx_{ki}^m}{dp_i^m}$ is constant $\forall i$, the sufficient conditions of existence and uniqueness for the equilibrium solution of this iterative process are set out in what follows.

Theorem: Let $\eta^m = \sum_k \left(\beta_k^m \frac{dx_{ki}^m}{dp_i^m} \right)$, where $\frac{dx_{ki}^m}{dp_i^m} = const, \forall i$. If either (i) $\max_m \left| \eta^m \right| < 2$, or (ii) $\sum_m \left| \eta^m \right| < 4$, then both the existence and the uniqueness of a fixed point are assured and the linear convergence of the iterative method just described is guaranteed.

Proof:

Let $f : p = (p_i^m) \in \mathbb{R}^{i \times n} \mapsto f(p) \in \mathbb{R}^{i \times n}$ be

$$f(p) = \left(\frac{\exp\left(\eta^m \left(p_i^m - \delta_i^m \right) + \sum_k \beta_k^m x_{ik}^m \right)}{\sum_{m'} \exp\left(\eta^{m'} \left(p_i^{m'} - \delta_i^{m'} \right) + \sum_k \beta_k^{m'} x_{ik}^{m'} \right)} \right)_i^m \tag{7}$$

where parameters β_k^m and η^m are the points that optimize the log-likelihood function of the MNL. This is a continuous function (De Grange et al., 2013) and the recursive estimation is such that $f\left(p^{(n)} \right) = p^{(n+1)}$.

Let p, q be any two points. For any norm,

$$\left\| f(p) - f(q) \right\| \le \left\| J(r)(p - q) \right\| \le \left\| J(r) \right\| \left\| p - q \right\| \tag{8}$$

where J is the Jacobian of $f(\cdot)$ and r is a point such that $r = \theta p + (1 - \theta) q, \quad \theta \in [0,1]$.

$$J_{ij}^{mk} = \frac{\partial f_i^m}{\partial p_j^k} = \begin{cases} \eta^m f_i^m \left(1 - f_i^m \right), & i = j; k = m \\ -\eta^k f_i^m f_i^k, & i = j; k \ne m \\ 0, & i \ne j \end{cases} \tag{9}$$

Matrix norm $\left\| \cdot \right\|_1$ and matrix norm $\left\| \ast \right\|_{¥}$ are applied to Jacobian (9) to get (10) and (11), respectively. Thus,

$$\left\| J(r) \right\|_1 = \max_i \max_m \left| \eta^m \right| f_i^m \left(1 - f_i^m \right) + \left| \eta^m \right| f_i^m \sum_{k \ne m} f_i^k \le 2 \max_m \left| \eta^m \right| \max_i f_i^m \left(1 - f_i^m \right) \le \frac{1}{2} \max_m \left| \eta^m \right|, \forall r$$

$$\left\| J(r) \right\|_\infty = \max_i \max_m \left| \eta^m \right| f_i^m \left(1 - f_i^m \right) + f_i^m \sum_{k \ne m} \left| \eta^k \right| f_i^k \le \max_i \max_m f_i^m \left(1 - f_i^m \right) \sum_k \left| \eta^k \right| \le \frac{1}{4} \sum_k \left| \eta^k \right|, \forall r$$

Since $f_i^m f_i^k \le f_i^m \left(1 - f_i^m \right) \le \frac{1}{4}, \qquad \forall i, k, m$, we obtain

$$\max_m \left|\eta^m\right| < 2 \Rightarrow \left\|f(p) - f(q)\right\|_1 < L\left\|p - q\right\|_1, \qquad L \in [0,1) \tag{10}$$

$$\sum_{m'} \left|\eta^{m'}\right| < 4 \Rightarrow \left\|f(p) - f(q)\right\|_\infty < L\left\|p - q\right\|_\infty, \qquad L \in [0,1) \tag{11}$$

If condition (i) is satisfied, $f(\cdot)$ is proven to be a contraction for norm $\|\cdot\|_1$, and if condition (ii) is satisfied, $f(\cdot)$ is a contraction for norm $\|\cdot\|_\infty$. In both cases, by the Banach fixed point theorem, a unique fixed point exists. Furthermore, since all norms are equivalent in a finite-dimensional space, the linear convergence of the iterative method is demonstrated for any norm.

4. NUMERICAL EXAMPLES

In this section we review the performance of the model set out above in the presence of endogeneity using data generated by a Monte Carlo simulation. Consider a simple case in which there are only two transport alternatives (private car and bus) and one explanatory variable (trip time) with a common parameter β_{time}. The utility functions for the two modes are

$$V_{it}^{car} = \beta_0 + \beta_{time} T_{it}^{car} \tag{12}$$

$$V_{it}^{bus} = \beta_{time} T_{it}^{bus} \tag{13}$$

where the private car trip time in minutes observed by individual i is a linear function of the total number of car trips in a given period t $\left(F_t^{car}\right)$ plus a random term ε_{it}^{car}, and is thus expressed as

$$T_{it}^{car} = \alpha + \gamma F_t^{car} + u_{it}^{car} \tag{14}$$

where $\alpha = 3$, $\gamma = 1/15$, and $F_t^{car} \sim U(100;500)$, so that the total flow of cars during period t is 100 to 500 vehicles uniformly distributed. Different trip time intervals can be chosen as a function of aggregate demand. Since the number of trips will be different in each simulation, the number of observations (i.e., the sample size) in each case will also differ (each traveller or user is an observation). A total of 130 simulations were conducted with sample sizes ranging from 150 to 800 users or observations. The variable u_{it}^{car} distributes uniformly between 0 and $\left(\alpha + \gamma F_t^{car}\right)/2$. The bus trip time is independent of demand and is defined as $T_{it}^{bus} \sim U(8;60)$. Finally, the population parameters are set at $\beta_{time} = -0.25$ and $\beta_0 = -0.15$.

Since by construction $\sum_i p_i^m = \sum_i \delta_i^m$, it is easily demonstrated for this simulation exercise that $\sum_i p_{it}^{car} = \sum_i \delta_{it}^{car} = F_t^{car}$, and therefore $\frac{dF_t^{car}}{dp_{it}^{car}} = 1$ and $\frac{dT_{it}^{car}}{dp_{it}^{car}} = \frac{dT_{it}^{car}}{dF_t^{car}} \frac{dF_t^{car}}{dp_{it}^{car}} = \frac{dT_{it}^{car}}{dF_t^{car}} = \gamma = \frac{1}{15}$.

The values estimated for β_{time} using the traditional MNL model (1) above and MNLE model (4), the latter explicitly reflecting the endogeneity in the private car trip time variable, are compared by the histograms in Figure 1. The histograms for the parameter β_0 are shown in Figure 2. The mean of the β_{time} estimator for both models in the presence of trip time congestion ($\gamma = 1/15$) is shown in Table 1 along with the result of a simulation using the same population parameters ($\beta_{time} = -0.25$ and $\beta_0 = -0.15$) but no congestion (i.e., $\gamma = 0$) and thus no endogeneity. The bias in the simulation estimate of the mean of β_{time} using MNL with endogeneity in trip time (e.g., congestion) is evident in the figure of -0.125. At this value the null hypothesis H_0: $\beta_{time} = -0.25$, the known population parameter, is rejected with a high level of significance. The mean estimate produced by MNLE, on the other hand, is not significantly different from the known parameter value and the null hypothesis cannot be rejected. In the bottom row of the table, it can be seen that when there is no endogeneity ($\gamma = 0$) the classic MNL's estimate is consistent. The results for parameter β_0 are given in Table 2. As with β_{time}, the MNL estimate in the presence of endogeneity is biased whereas the MNLE estimate is statistically unbiased (as is MNL without endogeneity in the bottom row of the table).

Figure 1
Distribution of the β_{time} Parameter Estimator for the MNL and MNLE Models ($\gamma = 1/15$)

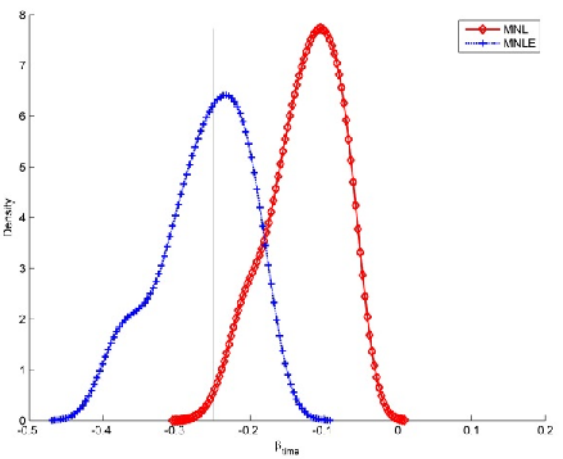


Figure 2
Distribution of the β_0 Parameter Estimator for the MNL and MNLE Models ($\gamma = 1/15$)

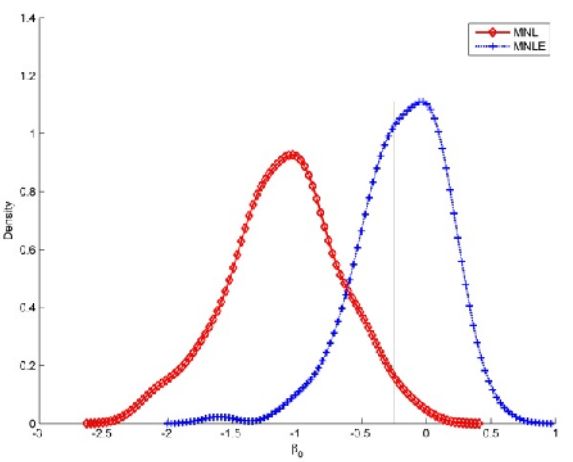


Table 1
β_{time} Parameter Estimators

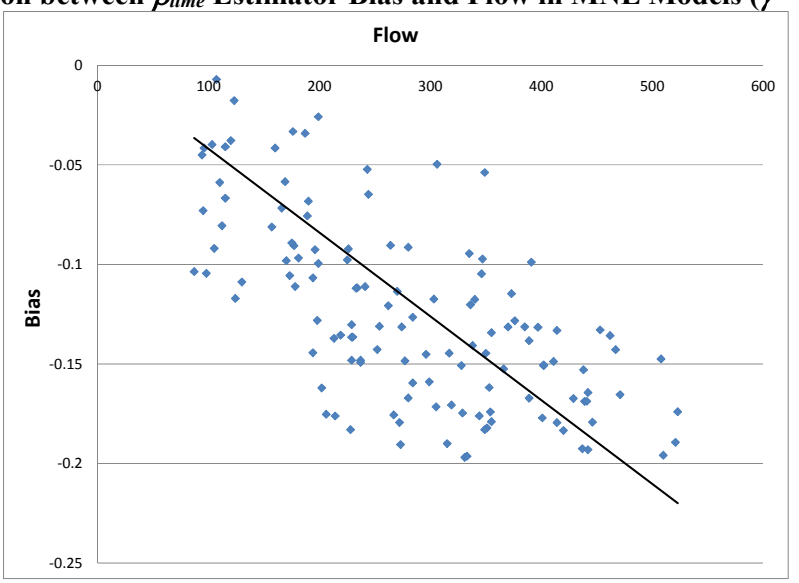
MODEL	$\hat{\beta}_{time}$	Standard Error	*t* Test (β_{time} = -0.25)
MNL (γ= 1/15)	-0.125	0.046	2.704
MNLE (γ= 1/15)	-0.262	0.058	-0.211
MNL (γ= 0)	-0.255	0.056	-0.084

Table 2

MODEL	$\hat{\beta}_0$	Standard Error	*t* Test (β_0 = -0.15)
MNL (γ= 1/15)	-1.102	0.431	-2.209
MNLE (γ= 1/15)	-0.194	0.342	-0.128
MNL (γ= 0)	-0.131	0.185	0.105

A dispersion graph of the bias estimated in each simulation for the β_{time} parameter and the level of demand or flow F_t^{car} impacting the level of congestion is shown in Figure 3. As is apparent, increasing congestion induces a downward bias in the MNL model estimates. Thus, the greater is the congestion the greater will be the MNL estimate bias unless the endogeneity is corrected.

Figure 3
Relation between β_{time} Estimator Bias and Flow in MNL Models (γ= 1/15)



As regards the models' goodness-of-fit, four indicators are reported in Table 3: log-likelihood evaluated at the parameter estimate values (L^*), log-likelihood evaluated at zero (L^0), and rho-square $\left(\rho^2\right)$ and adjusted rho-square $\left(\bar{\rho}^2\right)$ where

$$\rho^2 = 1 - \frac{L^*}{L^0}, \quad \bar{\rho}^2 = 1 - \frac{L^* - K}{L^0} \tag{15}$$

and K is the number of estimated parameters. The values shown are the averages of the simulation results for each statistic. They clearly show that MNLE has better goodness-of-fit on various indicators.

Table 3
Goodness-of-Fit for MNL and MNLE Models (γ= 1/15)

STATISTIC	MNL	MNLE
L^*	-274.80	-267.06
L^0	-308.991	-308.991
ρ^2	0.110	0.136
$\overline{\rho}^2$	0.104	0.129

The Horowitz test (1983) for comparing non-nested discrete choice models can also be used to compare the two models. The null hypothesis is the model with more parameters does not have a better fit. The value of this test statistic is given by

$$\left[\Phi\left\{-\left(-2\left(\overline{\rho}_h^2-\overline{\rho}_l^2\right)\cdot L^0+\left(K_h-K_l\right)\right)^{\frac{1}{2}}\right\}\right]^{-1}\sim N(0,1) \tag{16}$$

where:

$\overline{\rho}_l^2$ is the adjusted likelihood ratio index for the model with the lowest (l) value;
$\overline{\rho}_h^2$ is the adjusted likelihood ratio index for the model with the highest (h) value (in our case, it was the MNLE);
K_h, K_l are the numbers of parameters in models h and l, respectively;
Φ is the standard normal cumulative distribution function.

Using a 95% level of confidence the criterion for rejecting the null hypothesis if $\left|\Phi^{-1}\right|>1.96$. The average value of Φ^{-1} for various simulations was -3.934 indicating that the MNLE was had a better fit than MNL.

These findings can be complemented with the Hausman specification test statistic (Hausman, 1978; Marquez-Ramos et al, 2011), which in the present case is expressed as

$$H=\left[\hat{\ }_{\mathrm{MNL}}\quad\hat{\ }_{\mathrm{MNLE}}\right]^{\mathrm{T}}\left[\quad\left(\hat{\ }_{\mathrm{MNL}}\right)\quad\left(\hat{\ }_{\mathrm{MNLE}}\right)\right]^{-1}\left[\hat{\ }_{\mathrm{MNL}}\quad\hat{\ }_{\mathrm{MNLE}}\right]\sim\chi_q^2 \tag{17}$$

where $\left(\hat{\boldsymbol{\beta}}_{\mathbf{MNL}}\right)$ is the parameter vector estimated by the MNL model and $\left(\hat{\boldsymbol{\beta}}_{\mathbf{MNLE}}\right)$ the corresponding vector estimated by MNLE. Analogously, $\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\mathbf{MNL}}\right)$ is the variance and covariance matrix of the MNL parameter estimates and $\mathrm{var}\left(\hat{\boldsymbol{\beta}}_{\mathbf{MNLE}}\right)$ the corresponding matrix for the MNLE.

The null hypothesis of the test is that the selected parameters in both models are statistically equal. H is chi-squared distributed with q degrees of freedom (in this case 2, the number of parameters, and the critical value at the 5% level of significance is 5.99). The value of the statistic is $H = 18.58 > 5.99$ supporting that the MNL estimator is more biased than that of the MNLE. Thus, for the simulations that were conducted, the endogeneity in the trip time explanatory variable caused by congestion causes in the MNL model estimate whose order of magnitude is significant compared to the parameter values. By contrast, MNLE produces consistent estimators that fit the data better.

A sensitivity analysis was carried out on the $\frac{dx_{ki}^{m}}{dp_{i}^{m}}$ term. As noted above, the relationship between trip time and demand defined for the simulations was such that $\frac{dT_{it}^{car}}{dp_{it}^{car}} = \gamma = \frac{1}{15} \approx 0.07$. By modifying this value slightly in the MNLE model estimation, variations are produced in the β_{time} parameter bias and the log-likelihood value. The results are graphed for the former in Figure 4 and for the latter in Figure 5. Figure 4 shows the values taken by the likelihood function for different values of $\frac{dx_{ki}^{m}}{dp_{i}^{m}}$ (by trial and error). Figure 5 show the bias for different values of $\frac{dx_{ki}^{m}}{dp_{i}^{m}}$ ((by trial and error). As may be observed, the highest log-likelihood values are obtained when $\frac{dx_{ki}^{m}}{dp_{i}^{m}}$ is relatively close to $1/15 \approx 0.07$ while the lowest bias levels in β_{time} are also found close to that point. These relationships will be useful for estimating the MNLE model in real-world situations where the functional form of $\frac{dx_{ki}^{m}}{dp_{i}^{m}}$ is not known.

Figure 4
Sensitivity Analysis: dX/dP and Log-Likelihood

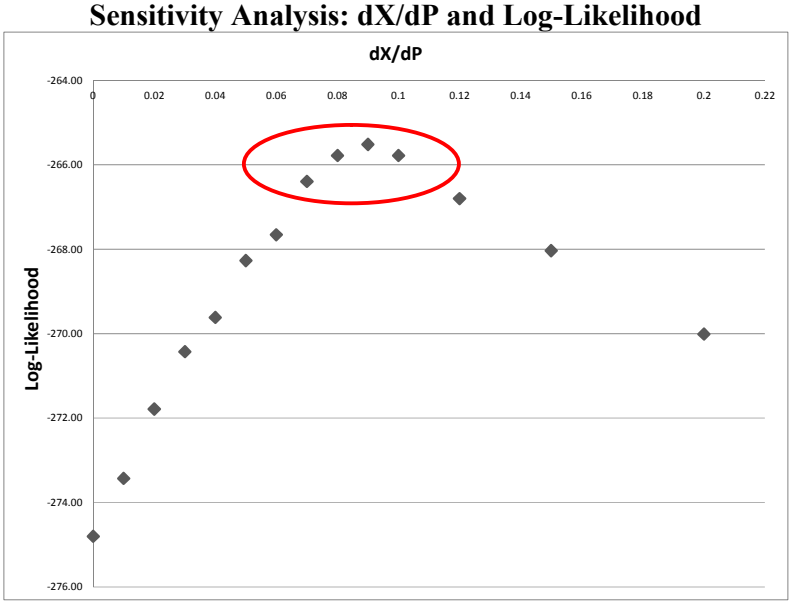
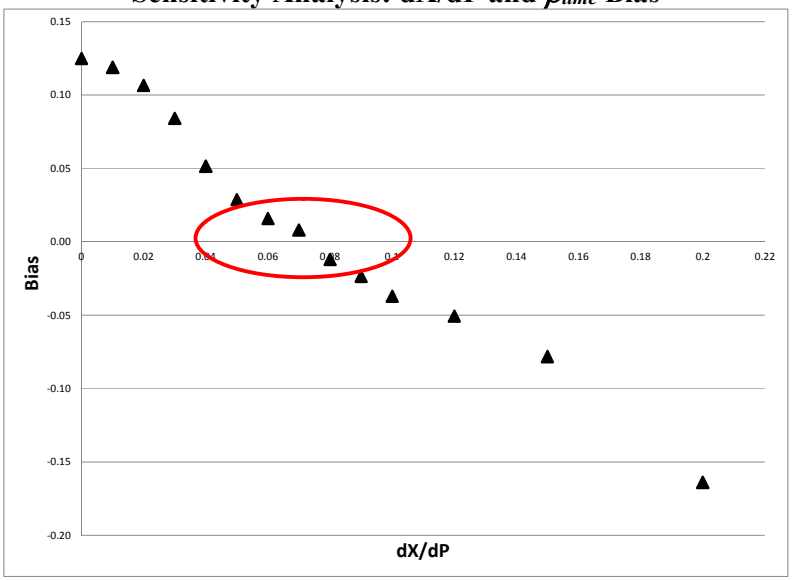


Figure 5
Sensitivity Analysis: dX/dP and β_{time} Bias



5. CONCLUSIONS

A new approach was presented for dealing with endogeneity in choice models where the attributes of the alternatives or the explanatory variables and their values are endogenous because they depend on individual choice decisions. This type of endogeneity typically arises in the context of social interactions or networks subject to network externalities. The article addressed a transport network setting in which public transport trip times depend on wait times or crowding at bus stops or train stations and private car trip times are subject to road congestion.

The proposed methodology explicitly incorporates this type of endogeneity into logit multinomial discrete choice models via the formulation of an equivalent maximum entropy optimization problem. The solution of the problem is a logit model with a fixed-point functional form that is calibrated through maximum likelihood estimation. This model is extendible to hierarchical multinomial formulations. The approach was tested by comparing the new model's performance with that of a traditional logit model in two different applications, one using simulated data. The simulated data were generated with endogeneity in the explanatory variable. The new model corrected the bias in the parameter estimates produced by the traditional formulation. The greater was the degree of endogeneity (i.e., level of congestion), the greater was the estimated bias. The proposed model also achieved a tighter fit to the data according to several goodness-of-fit indicators and statistical tests.

REFERENCES

Amador, F. J., González, R. M. and Ortúzar, J. de D. (2008). On Confounding Preference Heterogeneity and Income Effect in Discrete Choice Models. Networks and Spatial Economics, 8, 97–108.

Anas, A, (1983). Discrete Choice Theory, Information Theory and the Multinomial Logit and Gravity Models. Transportation Research, 17B, 13-23.

Banerjee, A. (1992). A Simple Model of Herd Behavior. Quarterly Journal of Economics, 107, 797-817.

Ben-Akiva, M. and Lerman, S. (1985). Discrete Choice Analysis: Theory and Application to Travel Demand. Cambridge, MA: The MIT Press.

Berry, S., Levinsohn, J. and A. Pakes. (1995). Automobile Prices in Market Equilibrium. Econometrica 63, 841-889.

Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of political Economy, 992-1026.

Blumenfeld-Lieberthal, E. (2009). The topology of transportation networks: a comparison between diferent economies. Networks and Spatial Economics, 9, 427-458.

Bogart, D. (2009). Inter-modal network externalities and transport development: Evidence from roads, canals, and ports during the english industrial revolution. Networks and Spatial Economics, 9, 309-338.

Boyce, D.E. (2007). Forecasting travel on congested urban transportation networks: Review and prospects for network equilibrium models. Networks and Spatial Economics, 7, 99–128.

Brock, W. and Durlauf, S. (2001). Discrete choice with social interactions. Review of Economic Studies, 68, 235–260.

Brock, W. and Durlauf, S. (2006). Multinomial Choice with Social Interactions. In: Blume, L., Durlauf, S. (Eds.), The economy as an evolving complex system. 3, Oxford University Press, Oxford.

De Cea, J.; Fernandez, J.E. and De Grange, L. (2008). Combined models with hierarchical demand choices: A multi-objective entropy optimization approach. Transport Reviews, 28, 415-438.

De Grange, L., Fernández, J. E., and De Cea, J. (2010). Combined Model Calibration and Spatial Aggregation, Networks and Spatial Economics, 10, 551-578.

De Grange, L., Ibeas, A. and González, F. (2011). A Hierarchical Gravity Model with Spatial Correlation: Mathematical Formulation and Parameter Estimation. Networks and Spatial Economics, 11, 439-463.

De Grange, L., González, F., Vargas, I. and Muñoz, J. C. (2013). A Polarized Logit Model, Transportation Research, 53A, 1-9.

Donoso, P. and De Grange, L. (2010). A Microeconomic Interpretation of the Maximum Entropy Estimator of Multinomial Logit Models and Its Equivalence to the Maximum Likelihood Estimator. Entropy, 12, 2077-2084.

Donoso, P.; De Grange, L. and González, F. (2011). A Maximum Entropy Estimator for the Aggregate Hierarchical Logit Model. Entropy, 13, 1425-1445.

Dugundji, E.R. and Gulyás, L. (2008). Socio-Dynamic Discrete Choice on Networks: Impacts of Agent Heterogeneity on Emergent Outcomes. Environment and Planning, 35B, 1028-1054.

Guevara, C.A. and Ben-Akiva, M. (2009). Addressing Endogeneity in Discrete Choice Models: Assessing Control-Function and Latent-Variable Methods. Working Paper Series, MIT Portugal, TSI-SOTUR-09-03.

Haab, T. and Hicks, R. (1997). Accounting for Choice Set Endogeneity in Random Utility Models of Recreation Demand. Journal of Environmental Economics and Management, 34, 127-147.

Hasan, M.K. and Dashti, H.M. (2007). A multiclass simultaneous transportation equilibrium model. Networks and Spatial Economics, 7, 197–211.

Hausman, J. (1978). Specification Tests in Econometrics. Econometrica, 46, 1251-1272.

Horowitz, J. H. (1983). Statistical Comparison of Non-Nested Probabilistic Discrete Choice Models. Transportation Science, 17, 319-350.

Kitthamkesorn, S., Chen, A., Xu, X. and Ryu, S. (2014). Modelling Mode and Route Similarities in Network Equilibrium Problem with Go-Green Modes. Online first, DOI 10.1007/s11067-013-9201-y.

Louviere, J.; Train, K.; Ben-Akiva, M.; Bhat, C.; Brownstone, D.; Cameron, T.; Carson, C.; Deshazo, J.; Fiebig, D.; Greene, W.; Hensher, D. and Waldman, D. (2005). Recent Progress on Endogeneity in Choice Modeling. Marketing Letters, 16, 255-265.

Marquez-Ramos, L., Martínez-Zarzoso, I., Pérez-García, E. and Wilmsmeier, G. (2011). "Special Issue on Latin-American Research" Maritime Networks, Services Structure and Maritime Trade. Networks and Spatial Economicas, 11, 555-576.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. Frontiers in Econometrics, Zarembka P., Ed.; Academic Press: New York, NY, USA.

Ortúzar, J. de D. (1982) Fundamentals of discrete multimodal choice modelling. Transport Reviews, 2, 47-78.

Ortúzar, J. de D. (1983) On the equivalence of modified logit models: some comments. Transportation, 11, 383-385.

Ortúzar, J. de D. and Willumsen, L.G. (2011). Modeling Transport. John Wiley & Sons: Chichester, UK.

Petrin, A. and Train, K. (2010). A control function approach to endogeneity in consumer choice models. Journal of Marketing Research, 47, 370-379.

Raveau, S., Muñoz, J.C. and De Grange, L. (2011). A Topological Route Choice Model for Metro. Transportation Research, 45A, 138 - 147.

Soetevent, A. and Kooreman, P. (2007). A Disecrete-Choice Model with Social Interaction: with an Application to Hight School Teen Behaviour. Journal of Applied Econometric, 22, 599–624.

Train, K. (1986). Qualitative Choice Analysis: Theory Econometrics, and an Application to Automobile Demand. Cambridge: The MIT Press.

Train, K. (2003). Discrete Choice Methods with Simulation. Cambridge University Press.

Villas-Boas, J. and R. Winer. (1999). Endogeneity in Brand Choice Models. Management Science, 45, 1324–1338.

Walker, J.; Ehlers, E.; Banerjee, I. and Dugundji, R. (2011). Correcting for endogeneity in behavioral choice models with social influence variables. Transportation Research, 45A, 362–374.

Williams, H.C.W.L. (1977) On the formation of travel demand models and economic evaluation measures of user benefit. Environment and Planning, 9A, 285-344.

Yamins, D., Rasmussen, S. and Fogel, D. (2003). Growing Urban Roads. Networks and Spatial Economics, 3, 69-85.