Exploratory analysis of endogeneity in discrete choice models

Anna Fernández Antolín

Amanda Stathopoulos

Michel Bierlaire

Transportation and Mobility Laboratory, ENAC, EPFL

April 2014

STRC

14th Swiss Transport Research Conference

Monte Verità / Ascona, May 14-16, 2014

Transportation and Mobility Laboratory, ENAC, EPFL

Exploratory analysis of endogeneity in discrete choice models

Anna Fernández Antolín
TRANSP-OR
Ecole Polytechnique Fédérale de Lausanne

phone: +41-21-693 93 29
fax: +41-21-693 80 60
anna.fernandezantolin@epfl.ch

Amanda Stathopoulos
TRANSP-OR
Ecole Polytechnique Fédérale de Lausanne

phone: +41-21-693 24 35
fax: +41-21-693 80 60
amanda.stathopoulos@epfl.ch

Michel Bierlaire
TRANSP-OR
Ecole Polytechnique Fédérale de Lausanne

phone: +41-21-693 25 37
fax: +41-21-693 80 60
michel.bierlaire@epfl.ch

April 2014

Abstract

A model is said to be affected by endogeneity when its deterministic part is correlated with the error term. This is an issue that affects both linear models such as regression and non-linear models like discrete choice models. It is a classical and well studied problem in regression models, in particular in econometric literature. However, how to deal with endogeneity in discrete choice models is not as developed, and in particular still less has been done regarding endogeneity and transportation.

The aim of this paper is to review what has already been developed in literature in order to identify potential contributions in this field. The paper contains a literature review and a description of three of the most common developed methods to deal with endogeneity.

Keywords

Discrete choice models, endogeneity, instrumental variables

1 Introduction

Endogeneity is an important problem that arises when modeling that is often not taken into account. Standard parameter estimation in presence of endogeneity leads to inconsistent estimates. A typical example in transportation mode choice in presence of endogeneity when not corrected for, is to obtain positive values for the parameters associated with cost. This is due to the fact that a common unobserved factor is comfort which is usually related to price. The positive effect in the utility of the unobserved comfort can offset the negative impact of cost. If this is the case, then it is easy to identify, but it is not always as trivial to do so. This can lead to not correcting for it due to the fact that its presence is not identified. If it is present and ignored, it has, of course, fatal consequences when it comes to forecasting. Guevara-Cue (2010) presents a test for it, but this test needs of valid and relevant instruments which are not always easy to identify.

The paper is structured as follows: section 2 describes the problem of endogeneity, section 3 is an overview of the main publications on the topic from the first proposed solutions to what is being done nowadays, section 4 presents the methodology introduced by Berry *et al.* (1995) which can only be used when endogeneity happens at a market level. Endogeneity is removed from the model by including a constant for each market group, with the help of instrumental variables. Section 5 describes the theoretical derivation of the method introduced by Hausman (1978) and Heckman (1978) which can be used even if endogeneity happens at an individual level. It also relies on finding valid and relevant instruments for the endogenous variable. Section 6 presents a similar solution to the control function approach, but the estimation of the parameters is simultaneous instead of sequential. It is a more general but less efficient method. Section 7 presents a comparison between the three described methods. Finally section 8 gives some final conclusions and future research directions.

2 The problem

In the derivation of the most common discrete choice models (such as probit, logit, nested logit, cross nested logit...) one of the assumptions is that the explanatory variables are independent from the error term, this is, from the unobserved factors. This may not always be the case. When the explanatory variables are not independent of the unobserved factors, endogeneity arises. If this is not taken into account, the usual estimation techniques lead to inconsistent parameters (Train (2003)). A very illustrative example of this phenomenon in the context of transportation is if comfort (which is positively correlated with cost) is not included in the model. Then we have an explanatory variable of the model (travel cost) which is correlated with an unobserved

factor (comfort). In this case, the parameter associated with travel cost will account both for the price and for the level of comfort and therefore the estimated parameter will be inconsistent.

The direction of this bias can usually be determined beforehand. In the above example where comfort (or any other desirable and unobserved attribute for that matter) is positively correlated with price, the parameter associated with price will capture both the negative impact of price and the positive impact of comfort resulting in a less negative estimated parameter compared to its real value.

Guevara-Cue (2010) points out two other causes of endogeneity: errors in variables and simultaneous determination. If there is an error in the measurement of a variable, this error will be propagated to the error term, which will then be correlated with the (wrongly) measured variable. To correct for this source of endogeneity all that can be done is to have all measurements as precise as possible or to use measurement equations to explicitly model the measurement errors. . Regarding simultaneous determination, it is the type of endogeneity that can be observed in the joint determination of residential location and mode choice. Those favoring public transportation will more probably live closer to locations with better accessibility and will therefore have smaller travel times compared to the rest of people living in the city. He points out that this source of endogeneity might not be significant if the demand and supply are treated at a microscopic scale, as the price (of each dwelling in his example, as he is dealing with residential location choice, but can be generalized to transportation modes in our case) is not likely to be determined by the choice made by any particular individual.

For the reasons stated in the paragraph above the next sections will focus on several techniques that have been developed to correct for endogeneity when it is caused by the explanatory variables being correlated with the unobserved attributes. The most popular ones are the BLP approach introduced by Berry *et al.* (1995), which is described in section 4, and the control function approach introduced by Heckman (1978) and Hausman (1978), described in section 5. The BLP is useful when endogeneity occurs at a market level and it consists of estimating an Alternative Specific Constant (ASC) for each market. It is clear that if there is a large number of markets this will add a lot of parameters to the model so its estimation can become an issue, but Berry *et al.* (1995) also proposed an algorithm within the iterative process of the other parameters to estimate the ASCs in a quick way. This algorithm is also described in section 4.

The control function approach allows to correct for endogeneity in more general scenarios, where it does not necessarily happen at a market level. As already stated, it is a method to correct for endogeneity when the observed variables are correlated with the unobserved factors. This means that the error term conditional on the observed variables don't have a zero mean, as is usually required. A control function is a variable that captures this conditional mean, therefore

"controlling" for this correlation.

3 Literature review

Louviere *et al.* (2005) present the recent progress that has been done in the field of endogeneity in discrete choice models. However, they give a very broad definition of endogeneity and focus also on choice set formation, interactions among decision makers and models of multiple discrete/continuous choice amongst other topics. In this review we are going to focus only on how to correct for endogenous explanatory variables.

A very used methodology is the BLP (Berry *et al.* (1995), Berry *et al.* (2004)) which recieves its name from the names of the authors. This approach consists on removing the endogeneity from the non-linear choice model and dealing with it in linear regressions. This is done by adding an ASC for each product and each market. By doing this the instrumental variables method can be used in the linear regression. A description of the instrumental variable methodology can be found in most of the basic econometric textbooks such as Baum (2006) or Lancaster (2004). Guevara-Cue (2010) describes in his thesis why it is more complex to deal with endogeneity in discrete choice models compared to linear models. The problem encountered when trying to correct for endogeneity in non-linear models is that these corrections lead to changes in the error term which imply a change of scale in the discrete choice models.

There are many studies that use the BLP approach to deal with endogeneity in discrete choice models. To name some examples, Walker *et al.* (2011) introduce a social influence variable in a behavioral model which is endogenous, as the factors that will impact the peer group will also influence the decision maker and this will cause correlation between the field effect variable and the error. Train and Winston (2007) use the BLP approach to correct for price endogeneity in automobile ownership choice. Crawford (2000) uses it for consumers' choice among TV options and Nevo (2001) uses it for a study of the cereal industry. It is also the approach chosen by Goolsbee and Petrin (2004) where they examine the direct broadcast satellites as a competitor to cable TV.

A second very used approach in literature is the control function methodology. The concept dates back to Hausman (1978) and Heckman (1978), although the term *control function* was introduced by Heckman and Robb Jr. (1985). Petrin and Train (2009) describe a control function approach to handle endogeneity in choice models. They apply both the control function and the BLP methodologies in a case study and find similar and more realistic demand elasticities than without correcting for endogeneity. They describe the control function methodology in

detail. Guevara-Cue (2010) also uses this method to study the choice of residential location. He also shows that there is a link between the control-function methods and a latent-variable approach.

The third frquently used approach is the one that Guevara-Cue (2010) calls *the control-function method in a maximum-likelihood framework* and Train (2003) calls *maximum-likelihood method*. Here, we will follow Train's terminology. It is the same formulation used by Villas-Boas and Winer (1999) in brand choice models and Park and Gupta (2012). In particular, Park and Gupta (2012) propose what they describe as a "new statistical instrument-free method to tackle the endogeneity problem". They model the joint distribution of the endogenous regressor and the structural error term by a Gaussian copula and use nonparametric density estimation to construct the marginal distribution of the endogenous regressor. Also, Bayesian methods to handle endogeneity have been introduced by Yang *et al.* (2003) and Jiang *et al.* (2009).

There are also other methods, but are less used because they are outperformed by the methods reviewed above. For example, the analogous to the standard 2-stage instrumental variable approach used in regression, described by Newey (1985) does not provide correct estimates of the aggregate elasticities of the models. Guevara-Cue (2010) shows it with a case study. Another method, developed by Amemiya (1978), is as efficient as the control function approach, as shown by Newey (1987), and is globally efficient under some circumstances, but is much more complex to calculate because it involves the estimation of auxiliary models.

An excelent review of the main methods presented in this paper can be found in Train (2003). Also Guevara-Cue (2010) describes in detail the control function and the maximum likelihood approaches.

4 BLP

When endogeneity occurs at a market or group level the approach proposed by Berry *et al.* (1995), Berry *et al.* (2004) can be used. It consists of the estimation of an Alternative Specific Constant (ASC) for each market in order to account for the endogeneity problem. Berry *et al.* (1995) apply the method to the choice of automobile markets, where price is supposed to be endogeneous by market. Markets are defined by geographical areas. In a second stage, the ASCs are regressed as a linear function of model variables. The number of ASCs to be estimated can be very large, so they also propose a method called Contraction to make the estimation process faster.

4.1 Specification

Let:

M be the number of markets,
J_m the number of options available to each consumer in market m,
p_{jm} the price of product j in market m,
x_{jm} the observed attributes different from price of product j in market m, and
ξ_{jm} the unobserved attributes.

Then the utility of consumer n in market m and product j can be expressed as:

$$U_{njm} = V(p_{jm}, x_{jm}, s_n, \beta_n) + \xi_{jm} + \varepsilon_{njm} \tag{1}$$

where s_n is a vector of socioeconomic characteristics of consumer n, β_n captures the tastes of consumer n and ε_{njm} are distributed as independent and identically distributed (iid) extreme value. $V(\cdot)$ is a function of the observed variables and the tastes of the consumer. Finally, ξ_{jm} represents the average utility that consumers obtain from the unobserved attributes of product j in market m.

Endogeneity arrises due to the fact that price depends on ξ_{jm}. The idea is to decompose the error term in the endogenous-causing part (in this case ξ_{jm}) and the random part (in this case ε_{njm}), and then to move ξ_{jm} into the observed part of the utility. This is achieved by introducing a constant for each product in each market.

$V(\cdot)$ can be decomposed as follows:

$$V(p_{jm}, x_{jm}, s_n, \beta_n) = \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)$$

where $\bar{V}(\cdot)$ varies over products and markets but not over consumers, and $\tilde{V}(\cdot)$ varies over consumers, markets and products. It is natural to think of $\bar{V}(\cdot)$ as representing the average of $V(\cdot)$ in the population.

Then equation 1 can be rewritten as:

$$\begin{aligned} U_{njm} &= \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \xi_{jm} + \varepsilon_{njm} \\ &= [\bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}] + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \varepsilon_{njm}, \end{aligned} \tag{2}$$

and defining $\delta_{jm} := \bar{V}(p_{jm}, x_{jm}, \bar{\beta}) + \xi_{jm}$, which does not vary among consumers (so is constant for each product in each market) and substituting it in (2) we obtain:

$$U_{njm} = \delta_{jm} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n) + \varepsilon_{njm}. \tag{3}$$

A choice model based on this utility specification does not entail any endogeneity as the constant for each product in each market absorbs ξ_{jm}.

Let $f(\tilde{\beta}_n|\theta)$ be the density of $\tilde{\beta}_n$, where θ are the parameter of the distribution, then the choice probability can be written as:

$$P_{nim} = \int \left[\frac{e^{\delta_{im} + \tilde{V}(p_{im}, x_{im}, s_n, \tilde{\beta}_n)}}{\sum_j e^{\delta_{jm} + \tilde{V}(p_{jm}, x_{jm}, s_n, \tilde{\beta}_n)}} \right] f(\tilde{\beta}_n|\theta)\, d\tilde{\beta}_n \tag{4}$$

The estimation of the choice model 4 provides estimates for δ_{im} and $\tilde{\beta}_n$, but it does not provide estimates of $\bar{\beta}_n$ in $\bar{V}(\cdot)$. To obtain these estimates what is usually done is to express $\bar{V}$ as linear-in-parameters, such tha δ_{jn} can be expressed as:

$$\delta_{jm} = \bar{\beta}\bar{V}(p_{jm}, x_{jm}) + \xi_{jm}, \tag{5}$$

which is a regression model that can be used to estimate $\bar{\beta}$. It is important to notice that this regression is affected by endogeneity, as $\text{corr}(p_{jm}, \xi_{jm}) \neq 0$, but endogeneity in linear regression is more streight-forward to deal with. Procedures for handling it are well established and can be found in econometric literature. This regression can be estimated by instrumental variables. An instrumental variable is an exogenous variable that can be used instead of price.

There are different ways to estimate the parameters. I will present estimation by maximum simulated likelihood and instrumental variables, but Train (2003) also describes in detail the estimation by the generalized method of moments (GMM). Berry *et al.* (1995),Berry *et al.* (2004), Nevo (2001) and Petrin (2002) use this generalized method of moments estimator in their work.

4.2 Estimation: Maximum Simulated Likelihood and Instrumental variables

The first step is to estimate $\bar{\beta}$ and θ by maximum simulated likelihood (MSL) and the constants δ_{jm} by the contraction method, described in the following section. Once the choice model is estimated, the estimated parameters are used in the linear regression (5) in order to obtain the estimates of $\bar{\beta}$. Since price is endogenous in this regression, instrumental variables can be used for the estimation.

The instrumental variables estimator is defined as the value $\hat{\bar{\beta}}$ that satisfies:

$$\sum_j \sum_m \{\hat{\delta}_{jm} - \bar{\beta}\bar{V}(p_{jm}, x_{jm})\} z_{jm} = 0$$

where $\hat{\delta}_{jm}$ are the estimated constants from the choice model and z_{jm} is the vector of all instruments used. Rearranging this expression we obtain

$$\hat{\bar{\beta}} = (\sum_j \sum_m z_{jm}\bar{V}(p_{jm}x_{jm})')^{-1}(\sum_j \sum_m z_{jm}\hat{\delta}_{jm}) \tag{6}$$

The main issue is to identify which variables should be used as instruments. They are usually the observed nonprice attributes, under the assumption that they are not affected by endogeneity. Berry *et al.* (1995) suggest two instruments: the average nonprice attributes of other products by the same manufacturer and the average nonprice attributes of other firms' products. However, the appropriate instruments depend on the context, and may be difficult to find as they have to be relevant (correlated with price) and valid (not correlated with the unobserved factors). Guevara-Cue (2010) proposes in Chapter 4 a method to validate instruments, but to do so you need to have some candidates, which could not be trivial.

4.3 The contraction

As mentioned before, the estimation of δ_{jm} can be numerically difficult, as there are as many δ as the number of markets multiplied by the number of alternatives. To deal with this, Berry *et al.* (1995) provide an algorithm which is very well explained by Train (2003) and that is

summarized here. The basic idea behind this methodology is that the constants can be set such that the predicted shares by the model equal the real shares.

Let

S_{jm} be the share of consumers in market m who choose product j,
$\delta =< \delta_{jm}, \forall j, m >$,
N_m the number of sampled consumers in market m, and
$\hat{S}_{jm}(\delta) = \sum_n \frac{P_{njm}}{N_m}$ the predicted shares.

Then δ is estimated following the iterative process:

$$\delta_{jm}^{t+1} = \delta_{jm}^{t} + \ln\left(\frac{S_{jn}}{\hat{S}_{jm}(\delta^t)}\right) \tag{7}$$

starting with any given values of the constants δ_{jm}^t.

For more details about this method refer to Chapter 13 of Train (2003)

5 Control function

The BLP approach presented in section 4 is not always applicable. For instance, if the observed share for a product in a market is zero its corresponding constant δ is not identifiable. Also, it may happen that endogeneity arises over decision markers instead of over markets, in which case the BLP approach will not solve the problem.

In this cases an alternative to the above proposed method is the control function method. Its derivation follows closely the specification of simultaneous equation regression models but is more complex due to the non-linearity.

5.1 Specification

The utility that consumer n obtains for product j is written as:

$$U_{nj} = V(y_{nj}, x_{nj}, \beta_n) + \varepsilon_{nj}, \tag{8}$$

where:

y_{nj} is the endogeneous explanatory variable for decision maker n and product j,

x_{nj} are the observed exogenous variables related to decision maker n and product j

ε_{jn} is the unobserved term, which is not independent from y_{nj} as required for standard estimation.

Let the endogeneous variable be expressed as a function of observed instruments and unobserved factors:

$$y_{nj} = W(z_{jn}, \gamma) + \mu_{nj}, \tag{9}$$

where ε_{nj} and μ_{nj} are independent of z_{nj}, but ε_{nj} and μ_{nj} are correlated, which implies that y_{nj} and ε_{nj} are also correlated.

The control function method consists of defining an auxiliary variable that will make endogeneity disappear when added to the systematic part of the utility. To define this variable, note that ε_{nj} can be decomposed in the following way:

$$\varepsilon_{nj} = E(\varepsilon_{nj}|\mu_{nj}) + \tilde{\varepsilon}_{nj}.$$

The error term $\tilde{\varepsilon}_{nj}$ is ortogonal of μ_{nj}. The expectation of ε_{nj} conditional on μ_{nj} is called the control function, and is denoted $CF(\mu_{nj}, \lambda)$, where λ are the parameters of the function. Then (8) can be rewritten as:

$$\begin{aligned} U_{nj} &= V(y_{nj}, x_{nj}, \beta_n) + E(\varepsilon_{nj}|\mu_{nj}) + \tilde{\varepsilon}_{nj} \\ &= V(y_{nj}, x_{nj}, \beta_n) + CF(\mu_{nj}, \lambda) + \tilde{\varepsilon}_{nj}, \end{aligned} \tag{10}$$

which leads to the following choice probabilities:

$$\begin{aligned} P_{nj} &= \text{Prob}(U_{nj} > U_{nk}, \forall k \neq j) \\ &= \iint I(V_{nj} + CF_{nj} + \tilde{\varepsilon}_n j > V_{nk} + CF_{nk} + \tilde{\varepsilon}_n k, \forall k \neq j)\, g(\tilde{\varepsilon}_{nj}|\mu_n)\, f(\beta_n|\theta)\, \mathrm{d}\tilde{\varepsilon}_n\, \mathrm{d}\beta_n \end{aligned} \tag{11}$$

where the following notation is used:

$\tilde{\varepsilon}_n =< \tilde{\varepsilon}_{nj}, \forall j >$,

$\tilde{\mu}_n =< \tilde{\mu}_{nj}, \forall j >$,

$V_{nj} = V(y_{nj}, x_{nj}, \beta_n)$,
$CF_{nj} = CF(\mu_{nj}, \lambda)$,
$g(\tilde{\varepsilon}_{nj}|\mu_n)$ is the conditional distribution of $\tilde{\varepsilon}_n$, and
$f(\beta_n|\theta)$ is the distribution of β_n.

This is a usual choice model, with the control function entering as an extra explanatory variable.

5.2 Estimation

Model (11) is estimated in two steps. Guevara-Cue (2010) calls this precedure the two-stage control-function (2SCF).

1. Equation (9) is estimated, and the residuals of this regression provide estimates for μ_{jn}: $\hat{\mu}_{jn} = y_{nj} - W(z_{jn}, \hat{\gamma})$, where $\hat{\gamma}$ are the estimated parameters from the regression.
2. The choice model in (11) is estimated by maximum likelihood, and $\hat{\mu}_{nj}$ enters it as an explanatory variable.

Depending on the specification of the control function, different choice models can be obtained. Train (2003) presents three different examples.

6 Maximum likelihood approach

It is a similar approach to the control function described in section 5 but the parameters of the model are estimated simultaneously rather than in two stages. We consider again equations (8) and (9), and instead of specifying the distribution of ε_{nj} conditional on μ_{nj}, their joint distribution is specified. It will be noted as $g(\varepsilon_{jn}, \mu_{jn})$ which can be rewritten as $g(\varepsilon_{jn}, y_{nj} - W(z_{nj}, \gamma))$ using equation (9).

Then the probability of choosing alternative i conditional on β_n is:

$$P_n(\beta_n) = \int I(U_{ni} > U_{nj}, \forall j \neq i)\, g(\varepsilon_n, y_n - W(z_n, \gamma))\, \mathrm{d}\varepsilon_n.$$

If β_n is random then the choice probability is also mixed over its distribution.

The log likelihood function can be then expressed as:

$$LL = \sum_n \ln(P_n),$$

which is then maximized over the parameters of the model. Guevara-Cue (2010) calls this model the full information maximum likelihood.

7 Comparison of different approaches

We have presented three different approaches to deal with endogeneity. The natural question now is which one is more appropriate to use.

The simplest and most straight forward method to use is the BLP approach, but as stated before it is not always possible due to some limitations. Between the maximum likelihood and the control function approach there is a trade-off between efficiency and generality.

As summarized in Table 1, the maximum likelihood approach requires the specification of the joint distribution of ε_n and μ_n. For every possible defined joint distribution there is a particular conditional distribution, but the reciprocal is not true. This is, two or more joint distributions can have the same conditional distribution. This implies that the control function approach is more general, but the maximum likelihood approach is more efficient as it is the maximum likelihood for all the parameters. A disadvantage of the control function method is that the standard errors of the parameters can not be calculated from the inverse of the Fisher-matrix information, as Guevara-Cue (2010) points out in his thesis.

	Max. Likelihood	Control funtion
Requires	Specification of joint distribution of ε_n and μ_n	Specification of conditional distribution of ε_n given μ_n
Advantages	More efficient (if joint distribution can be correctly specified); more robust to misspecifications in the error term	More general

Table 1: Comparison between maximum likelihood and control function approaches

8 Conclusion

After reviewing what has been done on endogeneity so far, it seems that theoretical methodologies to tackle it exist, but they often depend on instrumental variables that are not easy to find. To the best of our knowledge, there is no methodology to identify what a good instrument can be in different contexts. A possible future research direction would be to identify some valid and relevant instruments in the transportation context. However, an ideal procedure to handle endogeneity would not be dependent on finding valid and relevant instruments, as is the case of the methodology presented by Park and Gupta (2012). However, then other assumptions have to be made that may not be valid. This is topic for future work.

Also, it would be very interesting to develop statistical tests that could determine if there is the presence of endogeneity in a given model. The tests, summarized and generalized by Guevara-Cue (2010), depend always on the instruments used. This is, if an invalid or irrelevant instrument is used, it will appear as if endogeneity was not present, although it might happen that with a different instrument the result of the test was also different.

9 References

Amemiya, T. (1978) The estimation of a simultaneous equation generalized probit model, *Econometrica*, **46** (5) 1193–1205, September 1978.

Baum, C. F. (2006) *An Introduction to Modern Econometrics Using Stata*, Stata Press, August 2006, ISBN 9781597180139.

Berry, S., J. Levinsohn and A. Pakes (1995) Automobile prices in market equilibrium, *Econometrica*, **63** (4) 841–890, July 1995.

Berry, S., O. B. Linton and A. Pakes (2004) Limit theorems for estimating the parameters of differentiated product demand systems, *The Review of Economic Studies*, **71** (3) 613–654, July 2004, ISSN 0034-6527, 1467-937X.

Crawford, G. S. (2000) The impact of the 1992 cable act on household demand and welfare, *SSRN Scholarly Paper*, **ID 235290**, Social Science Research Network, Rochester, NY, July 2000.

Goolsbee, A. and A. Petrin (2004) The consumer gains from direct broadcast satellites and the competition with cable TV, *Econometrica*, **72** (2) 351–381, March 2004, ISSN 1468-0262.

Guevara-Cue, C. A. (2010) Endogeneity and sampling of alternatives in spatial choice models, Thesis, Massachusetts Institute of Technology. Thesis (Ph. D.)–Massachusetts Institute of Technology, Dept. of Civil and Environmental Engineering, 2010.

Hausman, J. A. (1978) Specification tests in econometrics, *Econometrica: Journal of the Econometric Society*, 1251â1271.

Heckman, J. J. (1978) Dummy endogenous variables in a simultaneous equation system, *Working Paper*, **177**, National Bureau of Economic Research, May 1978.

Heckman, J. J. and R. Robb Jr. (1985) Alternative methods for evaluating the impact of interventions: An overview, *Journal of Econometrics*, **30** (1â2) 239–267, October 1985, ISSN 0304-4076.

Jiang, R., P. Manchanda and P. E. Rossi (2009) Bayesian analysis of random coefficient logit models using aggregate data, *Journal of Econometrics*, **149** (2) 136–148, April 2009, ISSN 0304-4076.

Lancaster, T. (2004) *Introduction to Modern Bayesian Econometrics*, Wiley, June 2004, ISBN 9781405117203.

Louviere, J., K. Train, M. Ben-Akiva, C. Bhat, D. Brownstone, T. A. Cameron, R. T. Carson, J. R. Deshazo, D. Fiebig and W. Greene (2005) Recent progress on endogeneity in choice modeling, *Marketing Letters*, **16** (3-4) 255â265.

Nevo, A. (2001) Measuring market power in the ready-to-eat cereal industry, *Econometrica*, **69** (2) 307–342, March 2001, ISSN 1468-0262.

Newey, W. K. (1985) Semiparametric estimation of limited dependent variable models with endogenous explanatory variables, *Annales de l'inséé*, (59/60) 219–237, July 1985, ISSN 00190209.

Newey, W. K. (1987) Efficient estimation of limited dependent variable models with endogenous explanatory variables, *Journal of Econometrics*, **36** (3) 231–250, November 1987, ISSN 0304-4076.

Park, S. and S. Gupta (2012) Handling endogenous regressors by joint estimation using copulas, *Marketing Science*, **31** (4) 567â586.

Petrin, A. (2002) Quantifying the benefits of new products: The case of the minivan, *Journal of Political Economy*, **110** (4) 705–729, August 2002.

Petrin, A. and K. Train (2009) A control function approach to endogeneity in consumer choice models, *Journal of Marketing Research*, **47** (1) 3–13, February 2009, ISSN 0022-2437.

Train, K. (2003) *Discrete Choice Methods with Simulation*, Cambridge University Press, ISBN 9780521017152.

Train, K. E. and C. Winston (2007) VEHICLE CHOICE BEHAVIOR AND THE DECLINING MARKET SHARE OF US AUTOMAKERS*, *International Economic Review*, **48** (4) 1469â1496.

Villas-Boas, J. M. and R. S. Winer (1999) Endogeneity in brand choice models, *Management Science*, **45** (10) 1324–1338, October 1999, ISSN 0025-1909.

Walker, J. L., E. Ehlers, I. Banerjee and E. R. Dugundji (2011) Correcting for endogeneity in behavioral choice models with social influence variables, *Transportation Research Part A: Policy and Practice*, **45** (4) 362–374, May 2011, ISSN 09658564.

Yang, S., Y. Chen and G. M. Allenby (2003) Bayesian analysis of simultaneous demand and supply, *Quantitative Marketing and Economics*, **1** (3) 251–275, September 2003, ISSN 1570-7156, 1573-711X.