PUBLIC TRANSPORTATION IN CARTAGENA, COLOMBIA: UNDERSTANDING PREFERENCES USING DISCRETE CHOICE MODELS

DANIEL TORO GONZÁLEZ*

ABSTRACT

This paper revisits the data set collected by Toro, Alvis and Arellano (2005) in order to evaluate three main models in transportation: probit, conditional logit, and the nested logit. The results show that the advantages of other transportation modes over buses are their lower travel and waiting times. Additionally, it was found that a traveler facing equal times and costs would prefer bus rather than other alternatives.

JEL **Classifications:** C25, D12, R41.

Key words: Cartagena, Colombia, transportation, discrete choice models, multinomials, conditional, logit, probit.

* The author is professor of Economics, Universidad Tecnológica de Bolívar. He thanks Jia Yan, Haroldo Calvo, Roberto Fortich and two anonymous reviewers for their comments and guidance. Submitted: October 9, 2012; accepted: April 28, 2013.

RESUMEN

El transporte público en Cartagena de Indias: Un estudio de preferencias con modelos de elección discreta

Este trabajo emplea los datos recolectados por Toro, Alvis and Arellano (2005) con el fin de evaluar tres de los modelos más conocidos aplicados al análisis de preferencias por transporte: el probit, el logit condicional y el logit anidado. Los resultados revelan que la ventaja de otros modos de transporte diferentes al bus radica en sus menores tiempos de viaje y de espera. Adicionalmente, se encontró que, en un escenario de costos y tiempos de viaje similares, los consumidores preferirían bus ante cualquier otra alternativa.

Clasificaciones JEL: C25, D12, R41.

Palabras Clave: Cartagena, Colombia, Transporte, modelos de elección discreta, multinomiales, condicional, logit, probit.

I. INTRODUCTION

Cartagena, Colombia's main port on the Caribbean, has about one million inhabitants. One of the main problems people face daily is urban mobility and commuting. Cartagena's public transportation network is a mixture of formal transportation modes and routes, such as buses and taxies, and informal transportation modes, such as collective transportation vehicles and motorcycles, also known as «mototaxis». The formal transportation system of Cartagena is well known for its low quality and efficiency.

The purpose of this paper is to study some of the factors that influence consumers' transportation mode choice in Cartagena, using the most common empirical strategies in the literature. The data was collected in a survey undertaken by Universidad Tecnológica de Bolívar in September, 2004 (Toro, Alvis and Arellano, 2005).[1] In that study, individual binary choice probit models were used to identify the individual factors leading to the choice of each transportation

[1] For a descriptive analysis of the dataset see Toro, Alvis and Arellano (2005).

mode. In the present study, different choice models are used, where consumers confront all the transportation options in order to make decisions. By design, the exercise undertaken here shows very different but complementary results to Toro *et. al.* (2005) and, hence, has different policy implications.

The next section describes the random utility models (RUM), which is the main theoretical framework used in the study. Then a binary probit model evaluating the choice of public transportation is estimated, followed by a conditional logit model of the choice of urban transportation. Finally, a nested logit model and the conclusions are presented.

II. RANDOM UTILITY MODELS

Following the Random Utility Model framework (Small and Winston, 1999; Cameron and Trivedi, 2005, and Greene, 2003), suppose the decision maker chooses among j alternative transportation modes. Then, the chosen mode is assumed to maximize the decision maker's utility, which for mode j may be represented as in equation 1.

$$U_j = V(X_j, S; \beta) + \varepsilon_j \qquad (1)$$

Where,

X_j is a set of modal attributes such as cost, travel and waiting time,

S denotes characteristics of the decision maker, such as income and education,

β is a set of unknown parameters representing the user's preferences, and

ε_j is an unobserved («random») utility component capturing other decision makers' individual influences, including idiosyncratic preferences for mode j.

The modal X_j attributes may contain a dummy variable for mode j; in that case its coefficient represents an average preference for mode j, while ε_j represents the deviation from that average preference. The function $V(\cdot)$ is called «systematic» utility because the same functional form applies to all decision makers, unlike the random component that varies across decision makers.

Because utility is partly random, choices can be predicted only as probabilities. The probability that the decision maker will choose mode i is as seen in equation 2.

$$P_i = \text{Prob}\,[U_i > U_j \text{ for all } i \neq j] \qquad (2)$$
$$= \text{Prob}\,[V_i + \varepsilon_i > V_j + \varepsilon_j \text{ for all } i \neq j]$$
$$= \text{Prob}\,[V_i - V_j > \varepsilon_j - \varepsilon_i \text{ for all } i \neq j]$$

Where,
V_i is shorthand for $V(X_j, S; \beta)$.

Thus the choice probability depends not only on the systematic utility differences $(V_i - V_j)$, but also on how the random utility differences $(\varepsilon_j - \varepsilon_i)$ are distributed across the population.

III. BINARY PROBIT AND THE VALUE OF TRAVEL TIME

The simplest choice model involves two alternatives. Consider the transportation mode choice for an urban work trip. Let's call the public mode *Alternative 1* and the private mode *Alternative 2*. Then there is only one difference, namely $\varepsilon_2 - \varepsilon_1$, in equation 2. A reasonable assumption is that this utility difference is normally distributed across the population. This assumption leads to the probit choice probability in equation 3.

$$P_1 = \phi\,(V_1 - V_2) \qquad (3)$$

Where,
P_1 is the probability of choosing public, and
ϕ is the cumulative standard normal distribution function.

This choice model is simple to estimate and was used earlier by Lave (1970) to measure the value of time in urban commuting. Lave analyzed a sample, taken in the mid-1960's, of 280 urban commuters in the Chicago area who chose between automobile and public transportation. Here, as mentioned before, I use data collected from commuters of Cartagena, Colombia, in 2004 (Toro, Alvis and Arellano, 2005) to obtain estimates for the same parameters, replicating Lave's exercise.[2] The results of the estimated systematic utility function obtained are:

[2] The values of the Lave estimation are presented in the Appendix.

$$V = 2{,}56^{*} \cdot D^{T} - 0{,}00233^{*} \cdot w \cdot t - 0{,}9235^{*} \cdot c - 0{,}0022 \cdot (Inc \cdot Dist \cdot D^{T}) - 0{,}0221^{*} \cdot (Age \cdot D^{T}) + 0{,}459^{*} \cdot (Female \cdot D^{T})$$

$$R^{2} = 0{,}248$$

$$N = 449$$

Where D^T is an alternative-specific dummy variable equal to 1 for public transit and 0 for private. It enters the model independently (in the first term) and also interacts with the traveler's income (*Inc*), trip distance (*Dist*), age, and a dummy variable indicating whether the traveler is female. The traveler's wage rate is denoted by *w*, travel time by *t*, and travel cost by *c*. Note that D^T, *t* and *c* all vary from one mode to the other, whereas *w*, *Inc*, *Dist*, *Age*, and *Female* do not.

For Cartagena the estimated parameters that are statistically significant are indicated in the equation by an asterisk. In this case, the model suggests that travelers are less likely to take public transportation as they become older; on the contrary, they are likelier to do so if they are female.

The utility function is linear in travel time and cost. The value of travel time (VOT), defined as the marginal rate of substitution between time and cost, is just the ratio of the time and cost coefficients of that linear relation:

$$\text{VOT} = \frac{-0{,}00233397}{-0{,}92355887} w = 0{,}0025\, w$$

In other words, time of travel is valuated at 0,25 percent of the average wage rate. Note that the variables in this model were specified so that VOT is proportional to the wage rate.

Even though this approach is consistent with models of time allocation, which suggest that a person's trade-off between travel time and money is strongly related to his or her possibilities of earning money in the labor market, the results of the estimation for Cartagena are very low to be interpreted in such manner. However, the VOT discriminating the estimation by students and workers yields that the VOT for workers is almost one third higher than the VOT for students. This result is economically intuitive since the workers are paid by their time; thus the opportunity cost in the short run is higher. In this context, short run means that the parameter is not capturing all the effects of the returns of education; otherwise, there will be no incentives to study. The omission of these effects may be corrected by an exercise that allows us to discriminate by type of work: if it is

high or low skilled labor according to their type of education. However it is not possible to address this problem using the current survey.

The estimated VOT might also be low due to problems with the measurement of the variables. For example, travel costs (*c*) actually represent the daily consumption of transportation. Even though travel costs were adjusted for public transportation users by its frequency of use and for the private transportation users by adding the cost of gas, taxes and general maintenance; it still may contain measurement errors.

In the case of income (*w*), many of the students surveyed did not report any value. Because of missing information in the survey this variable generates some non randomness in the error term. Thus we may have a sample selection problem, causing sample selection bias. If the sample selection problem is present (for example, because the students tend to not report any income) then we have an identification problem.

In order to empirically test for a sample selection problem a Heckman two-step approach was used. This was done by introducing the variables household size, education, and age, to predict if the commuter is a student or not. The results show that there is not statistical evidence to reject the null hypothesis of rho equal to zero. There is, thus, no evidence of a sample selection problem (Sweeney, n.d.). Because of the non significance of the test parameter (rho=0,5485892, with a p-value of 0,6395) we cannot reject the null hypothesis of no sample selection bias in the coefficients. Similar results are obtained when using a work dummy variable instead of school (rho=0,642, with a p-value of 0,26).

Although this modeling strategy yields useful information about commuter's choices between public and private types of transportation, such as the differences in VOT between workers and students, it does not predict which specific transportation mode, such as bus, taxi or any other, people are choosing.

IV. CONDITIONAL LOGIT AND THE CHOICE OF URBAN TRANSPORTATION MODE

The data base collected with the survey not only provides information about commuters' choices with respect to public or private transportation; it also provides information about what type of public transportation mode commuters choose. For this model $J = 4$ observable categories of public transportation which we define as mototaxi (1), bus (2), taxi (3), and automobile (4).

According to Small and Winston (1999), the key to obtaining a computationally convenient choice model is to have an easily calculated expression for the choice probability. McFadden derives such a model by assuming that each of the random utilities follows the extreme value distribution, which is almost indistinguishable from the normal distribution in practice. Thus the resulting choice probability given in equation 4.

$$P_i = \frac{e^{V_i}}{\sum_{j=1}^{J} e^{V_j}} \qquad (4)$$

This model is known as a conditional logit. One of its advantages is that its complexity does not increase with the number of alternatives *J*. However, though the conditional logit model seems more appropriate to understand commuters' behavior than the simple probit, the problem is that there is not enough information in the data base to identify the model. According to Schmidheiny (2007), in the conditional logit model individuals only care about utility differences across alternatives. Therefore factors that influence the level of utility for all alternatives in the same way cannot explain the individual's decision. Then, if the utility function in equation 1 is specified as V_n (X_{nj}, X_j, S_n; β), where S_n are the individual *n* specific independent variables, X_j are the variables related only on the alternative *j*, and X_{nj} are the variables related with both, then individual specific independent variables will be canceled in the choice probability, and the correspondent parameter (β_3) is not identified, and we get equation 5.

$$P_{nj} = \frac{e^{x_{nj'}\beta_1 + x_{j'}\beta_2 + S_{n'}\beta_3}}{\sum_{i=1}^{j} e^{x_{ni'}\beta_1 + x_{i'}\beta_2 + S_{n'}\beta_3}} = \frac{e^{x_{nj'}\beta_1} e^{x_{j'}\beta_2} e^{S_{n'}\beta_3}}{e^{S_{n'}\beta_3} \sum_{i=1}^{j} e^{x_{ni'}\beta_1} e^{x_{i'}\beta_2}} = \frac{e^{x_{nj'}\beta_1} e^{x_{j'}\beta_2}}{\sum_{i=1}^{j} e^{x_{ni'}\beta_1} e^{x_{i'}\beta_2}} \qquad (5)$$

A constant that does not vary with individuals or with alternatives is, of course, not identified by the same argument. Thus, individual characteristics start playing a role when they interact with alternative characteristics, like dummy variables by alternative. It is often useful to include alternative specific constants α_j. These alternative fixed effects capture all observed and unobserved characteristics that describe the alternative but are identical across individuals. According to this, since it is not possible to build a conditional logit model based only on individual characteristics, because of the identification problem, then the identi-

TABLE 1
Conditional Logit Regression

Variable	Mode	Coefficient	P-Val
Cost/Income	Mototaxi	2,162	0,003
Cost/Income	Taxi	4,029	0,000
Cost/Income	Auto	3,368	0,000
Travel Time	Mototaxi	-0,074	0,000
Travel Time	Taxi	-0,085	0,001
Travel Time	Auto	-0,033	0,014
Waiting Time	Mototaxi	-0,003	0,819
Waiting Time	Taxi	-0,055	0,081
Waiting Time	Auto	-0,021	0,214
Dummy	Mototaxi	-0,918	0,032
Dummy	Taxi	-1,471	0,015
Dummy	Auto	-2,012	0,000
Observations			1308
R2			0,48

Sources: Author's calculations based on survey data in Toro, Alvis and Arellano (2005).

fication strategy was to use dummy variables by alternative. The empirical results of the estimated model are presented in Table 1.

Following a similar specification structure as that presented by McFadden and Talvitie (1977) and Small and Winston (1999), mode choice is explained by four observed attributes: costs (*c*), in this case as a proportion of the wage rate (*w*), in vehicle travel time (*tin*) and out of vehicle travel time or waiting time (*tout*). The unobserved attributes for mode *j* are captured by the mode specific dummy variable D^j, which like D^T in the previous model, is defined as one for the mode *j* and zero for the other modes. In this case mode 2 (bus transportation) is dropped off the specification and thus is defined as the *base mode* with which the parameters estimated for other modes are compared.

According to the results, higher travel costs will increase the probability of taking any alternative (mototaxi, taxi, and automobile) with respect to buses. In

the case of the variables time-in-vehicle and time-out-vehicle, in both cases greater values of the variables decreases the probability that commuters will select each mode over bus. This basically means that the competitive advantage of other transportation modes over bus is their lower times in and out vehicle. Specifically, with respect to the waiting time for the cases of mototaxi and auto the coefficients are not statistically different from zero.

Finally, the negative mode-specific constants of D^j show that a traveler facing equal times and costs would prefer bus (alternative 4), rather than the other alternatives. This last finding is very important to support some policies that seek to increase the use of public transportation (such as buses) in Cartagena.

One of the main problems of the conditional logit model is that it imposes the restriction that the choice between any two pairs of alternatives is simply a binary logit model. This means that it implicitly assumes that all transportation modes are perfect substitutes for each other. In this sense, in the absence of buses as a transportation option, the probability of choosing mototaxi is increased in the same amount that all the other options — taxi and auto — which is clearly unrealistic. This is the well known problem of Independence of Irrelevant Alternatives (IIA).

V. NESTED LOGIT AND THE CHOICE OF URBAN TRANSPORTATION MODE

The problem for the conditional logit model is that we rely on the IIA assumption. For this reason, following Cameron and Trivedi (2005), we use a Nested Logit Model, which is a generalization of the multinomial model, to approach the IIA assumption with a model in which the substitution patterns among alternatives are more flexible. For example, we can estimate a structure like the one presented in Graph 1.

According to this structure, the unavailability of one of the public transportation modes will not affect the consumer's probability of choosing private transportation, but will increase the probability of choosing taxi or mototaxi. The equivalent output for the Nested Logit Model is shown in Table 2.

GRAPH 1
Nested Logit Structure

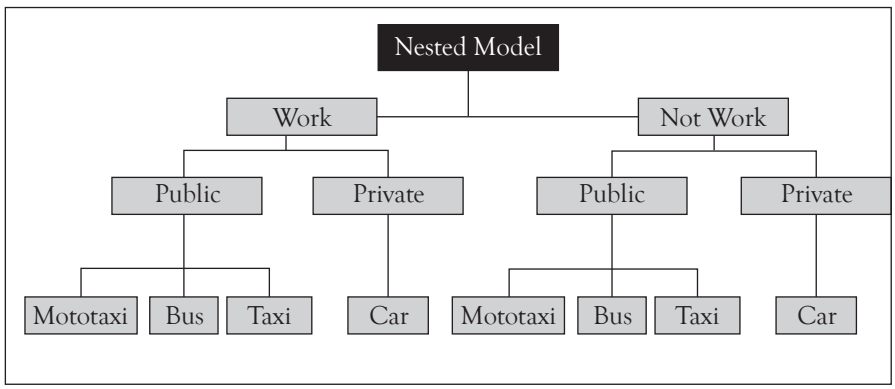

TABLE 2
Nested Logit Regression

Variable	Level 1	Level 2	Level 3	Level 4	P-Val
Cost/Income	Work	Public	Mototaxi	89,045	0,147
Cost/Income	Work	Public	Bus	84,314	0,166
Cost/Income	Work	Public	Taxi	87,131	0,153
Cost/Income	Work	Private	Car	66,905	0,239
Cost/Income	Not Work	Public	Mototaxi	85,299	0,164
Cost/Income	Not Work	Public	Bus	80,619	0,191
Cost/Income	Not Work	Public	Taxi	86,341	0,158
Travel Time	Work	Public	Mototaxi	-0,095	0,326
Travel Time	Work	Public	Bus	0,121	0,286
Travel Time	Work	Public	Taxi	-0,066	0,503
Travel Time	Work	Private	Car	0,078	0,452
Travel Time	Not Work	Public	Mototaxi	0,055	0,628
Travel Time	Not Work	Public	Bus	0,106	0,357
Travel Time	Not Work	Public	Taxi	0,068	0,546
Waiting Time	Work	Public	Mototaxi	0,231	0,191
Waiting Time	Work	Public	Bus	0,254	0,156
Waiting Time	Work	Public	Taxi	0,111	0,521
Waiting Time	Work	Private	Car	0,234	0,182
Waiting Time	Not Work	Public	Mototaxi	0,225	0,220
Waiting Time	Not Work	Public	Bus	0,243	0,185
Waiting Time	Not Work	Public	Taxi	0,174	0,360
Observations					3368
LR Test for IIA (tau=1):	Chi2(5)=7.90		Prob>Chi2 =		0,1616

Sources: Author's calculations based on survey data in Toro, Alvis and Arellano (2005).

The outputs of this model are weaker than in the previous case. First, the flexibility of the model seems to affect the statistical significance of the variables. The confidence level of most variables drops to 80%. Second, the bottom of the output includes a Likelihood Ratio test statistic of the IIA restriction that leads to the rejection of nested logit in favor of conditional logit.

VI. CONCLUSIONS

Different discrete choice models were evaluated to identify some of the factors that affect consumers' choice of transportation in Cartagena. The most reliable model for this specific case seems to be the conditional logit model, for which the results show that the competitive advantage of other transportation modes over buses is their lower travel time and the time of waiting. This points to a systematic inefficiency of public transportation in the city, which drives the consumers to other alternatives. However, market structure can be changed with relative ease by means of public policy. Since travelers facing equal times and costs prefer buses over other alternatives, the trend can be corrected in order to re-orient consumers to use the public transportation system. According to the results shown here, this new system just has to offer competitive travel and waiting times with respect to other transportation modes.

REFERENCES

Cameron, A. Colin, and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*, New York: Cambridge University Press.

Greene, William H. (2003), *Econometric Analysis*, Fifth Edition, New York: Pearson Education, Inc.

Lave, Charles A. (1970), «The Demand for Urban Mass Transportation», *The Review of Economics and Statistics*, Vol. 52, No. 3, August

McFadden, Daniel, and Antti Talvitie (1977), «Demand Model Estimation and Validation» http://elsa.berkeley.edu/~mcfadden/utdfp5.html (accessed 12/12/2009).

Schmidheiny, Kurt (2007), «'UPF' Teaching», Universitat Pompeu Fabra http://kurt.schmidheiny.name/teaching/multinomialchoice2up.pdf (accessed 11/24/2009).

Small, Kenneth A., and Clifford Winston (1999), «The Demand for Transportation: Models and Applications», in José Gómez-Ibáñez, William B. Tye and Clifford Winston, editors, *Essays in Transportation Economics and Policy*, Washington, D.C.: Brookings Institution Press.

Sweeney, Kevin (n. d.), «Implementing and Interpreting Sample Selection Models», Department of Political Science, Ohio State University, http://psweb.sbs.ohio-state.edu/prl/Selection%20Models.pdf (accessed 12/12/2009).

Toro, Daniel, Jorge Luis Alvis, and William Arellano (2005), «Transportation Systems in Cartagena: The Behavior of Public Urban Transportation System Users» http://ssrn.com/abstract=962709.

APPENDIX

Probit Model for Students (public vs. private)

$$V = 2{,}02D^T - 0{,}0031^* \, w \cdot t - 1{,}7243 \, c - 0{,}00254 \, (Inc \cdot Dist \cdot D^T) - 0{,}0489 \, (Age \cdot D^T) - 0{,}3964 \, (Female \cdot D^T)$$

$$R^2 = 0{,}398 \text{ and } N = 77$$

$$\text{VOT} = \frac{-0{,}00331309}{-1{,}7243448} \, w = 0{,}019w$$

Probit Model for Workers (public vs. private)

$$V = 2{,}23^* D^T - 0{,}00222^* \, w \cdot t - 0{,}9012^* \, c - 0{,}00275 \, (Inc \cdot Dist \cdot D^T) - 0{,}0132 \, (Age \cdot D^T) - 0{,}504 \, (Female \cdot D^T)$$

$$R^2 = 0{,}251 \text{ and } N = 351$$

$$\text{VOT} = \frac{-0{,}000222}{-0{,}9012} \, w = 0{,}025w$$

Lave's results for the same estimation (public vs. private)

The results of the Lave's (Lave 1970) estimated systematic utility function obtained are:

$$V = -2{,}08D^T - 0{,}00759^* \, w \cdot t - 0{,}0186^* \, c - 0{,}0254 \, (Inc \cdot Dist \cdot D^T) + 0{,}0255^* \, (Age \cdot D^T) - 0{,}057 \, (Female \cdot D^T)$$

$$R^2 = 0{,}279 \text{ and } N = 280$$

In this case all estimated parameters are statistically significant except the last. The model shows that travelers are less likely to take public transportation as their income or trip distance increases, but more likely to take it as they become older.

This utility function is linear in travel time and cost. The value of travel time (VOT), defined as the marginal rate of substitution between time and cost, is just the ratio of the time and cost coefficients of that linear relation:

$$\text{VOT} = \frac{-0{,}00759}{-0{,}0186} w = 0{,}41w$$

This result means that travel time is valuated at 41 percent of the average wage rate. Note that the variables in this model were specified so that VOT is proportional to the wage rate. This approach is consistent with models of time allocation, which suggest that a person's trade-off between travel time and money is strongly related to his or her possibilities for earning money in the labor market.