Experiment design and logit errors

Maria Cleci Martins de Carvalho

Senior Research Lecturer Business School
Universidade Luterana do Brasil

Preference surveys report individual choices regarding an alternative choice set. This type of data is paramount to forecast the product/service demand as it makes possible to determine the importance the consumers/users pay to the different product/service characteristics. The data modelling is usually performed through Logit. As a probabilistic model, Logit is based on assumptions about the consumer's behaviour, which might not be real (for instance, the model assumes there is no taste variation among individuals). Besides, researchers have been using a linear utility function in order to avoid model complications, even though it is known that the individual's behaviour in a choice process is not linear. This paper reports literature main research findings on the subject, as well as, the influence of designs in the calibration results. Experiments are conducted using simulation tools. Results showed that, for data holding Logit assumptions, the bigger the distance between the attribute vectors, the better the calibration results from Logit modelling as it reduces the possibility of the optimisation algorithm to get stuck in a flat region. On the other hand, it is likely that the data could break the homoscedasticity Logit assumption on the error term.

Keywords: Stated Preference Design, Logit Optimisation, Errors

1 Introduction

Our main objective is to avoid erroneous market share forecasting from the modelling of discrete choices made by travellers when faced with travel options with different characteristics.

This information is vital to transport planners who wish to forecast demand and maximise resource utilisation. Choice processes can be explored using the technique of stated preference, in which consumers are presented with set of hypothetical scenarios and are asked to choose among them. This yields discrete choice data, the modelling of which has to be approached in a different way to continuous variables. Logit modelling is the most popular technique to cope with discrete choice data. It provides a framework to explore the trade-off between the attributes of the various alternatives, each of which is associated with a *utility*.

The maximum of a Logit likelihood function can be obtained by the Newton-Raphson method (Ben-Akiva & Lerman, 1985), which uses first and second derivatives to reach the optimum. In case of a linear utility function, a unique maximum is guaranteed, if it exists. However, if a linear utility function is used in a process that was truly performed on a non-linear function, the modelling will have its source of errors increased considerably.

Unfortunately, if the utility function is non-linear in parameters, there is no guarantee that a single maximum exists, as the error surface can be non-convex (Ortuzar & Willumsen, 1994).

An additional problem is encountered when data characteristics are not according to Logit modelling assumptions (Horowitz, 1981;1981b) (Ben-Akiva,1985). A main cause of trouble occurs when the tastes are not homogeneous through the population. Therefore, besides a generalised random term (to account for factors that influences the choice but are not considered in the modelling for some reason), there is a noise from the individuals taste variations expressed in their choices (Fowkes & Wardman,1988).

Furthermore, there is no guarantee that the total noise is Weibul distributed as assumed by Logit.

Probit models are a possible solution to it (Daganzo, 1979). However, they are complex and very difficult to solve, especially if the number of alternatives is greater than three. Other models have also being investigated, however, Logit is still the most employed by far.

Horowitz (1981, 1981a) developed experiments based on simulation tools to show situations where the Logit modelling is inappropriate. His experiments are based on an only data composition and design, although considering different error situations (taste variation, correlation, etc.).

Experiments from De Carvalho (1998) showed that, for the data being analysed, Logit modelling only failed in extreme situations that hardly happen in real life. Both experiments have used simulation approach and held same data design (variables and attributes).

Apparently, there is a contradiction: on the one hand, researchers state that the model fails due to the assumptions it is based on; on the other hand, experiments with synthetic data showed that Logit is a quite robust modelling approach. This research investigates whether the trade-offs considered in the design of the SP experiment are the reason for those contradictions. The paper is divided in four main topics: a theoretical review of the subject, explanation of the experiments procedure, results and conclusions.

2 Theoretical review

Economic Consumer Theory states that individuals make choices among alternatives by evaluating their attributes combined in some way. The attributes being considered refer to a features vector that best represents a choice alternative.

The total value of an alternative is known as utility and the alternative with the highest utility is chosen. The Random utility theory (Ben-Akiva,1985) considers that the analyst does not know the whole range of factors influencing the choice and introduces a random error in the model to account for it. A random utility (Ui) has two components: the observed element (Vi), and the stochastic term (ε_i):

$$U_i = V_i + \varepsilon_i \tag{1}$$

In order to predict if an alternative will be chosen, its utility must be contrasted with those of all other alternatives and transformed into a choice probability value between 0 and 1.

Logit models hold the assumption that the unobserved components of utility are independently and identically Weibul distributed. Thus, for the Logit model, ε_{ni} and ε_{nj}, $i \neq j$, are assumed to have the same distribution, with the same mean and variance, and also to be unrelated. These random variables being unrelated means that any factor that the researcher does not observe and that affects the utility of an alternative does not affect the utility of the others. These assumptions do not hold where exist correlation between alternatives.

For the Logit model the probability of an observation 'n' choosing 'i'th alternative is a function of the utilities of the alternatives involved in the choice set, as follows:

$$P_{in} = \frac{e^{\mu V_{in}}}{\sum_i e^{\mu V_{in}}} \tag{2}$$

Where

V_i = Systematic part of the alternative 'i' utility ($V_i = f(\beta_k, X_{ik})$)

β_k = Unknown parameter related to attribute 'k', to be found by calibration

X_{ik} = Known value of attribute 'k' for the 'i' alternative

μ = Scaling factor

The choice data is composed of N observations on the occurrence of an event. A vector can represent each observation as follows:

$$[x_{nc1k} \ldots x_{nc1K}\ x_{nc2k} \ldots x_{nc2k}\ x_{nc3k} \ldots x_{nc3K}]\ [y_{nc}]$$

Where:

X_{nc1k} - is the known of the 'k' attribute of the alternative '1', of the 'c' choice set, for individual 'n'

y is the choice

Alternatives 1,2,and 3

The stochastic part of the Utility is scaled by a factor 'μ', that is inversely related to the standard deviation of the random error through the following expression:

$$\mu = \frac{\pi}{\sqrt{6}\sigma_{\varepsilon i}} \tag{3}$$

In real situations, this scaling factor is not found separately from the coefficients (β).

The true utility function of an alternative (U) can be linear or non-linear. In fact, there are enough reasons to believe that such decisions are made using non-linear functions in the individual's mind. However, researchers have preferred to model it (V) using a linear function due to its convenience (Ben-Akiva, 1985)(Ortuzar,1994). Some of the reasons for that are reported as follows:

i. Calibration results can be analysed straight way, without the need of calculations (variable transformations, etc.).
ii. It might reduce errors from modelling due to taste variation. (Horowitz, 1981, 1981a).
iii. The likelihood function is guaranteed to be well behaved and therefore an only optimum point is reached (if exists). (Ben-Akiva,1985)

2.1 Main Logit Model Assumption

The variance-covariance matrix (Σ) for the random term in a choice set is stated as:

$$\Sigma = \begin{pmatrix} \sigma_1^2 & cov_{12} & cov_{1j} \\ cov_{21} & \sigma_2^2 & cov_{2j} \\ cov_{i1} & cov_{i2} & \sigma_j^2 \end{pmatrix}$$

Where

σ_j^2 = variance of the alternative 'j' random term

cov_{ij} = covariance of the random term between alternatives 'i' and 'j' ($cov_{ij} = cov_{ji}$)

The Multinomial Logit assumes that all covariances (off diagonal of Σ) are equal to zero ($cov_{ij}=0$, $i \neq j$), and variances hold same value ($\sigma_1 = \sigma_2 = \ldots = \sigma_j$). This means that the noise inside the data must be independent and identically distributed. The Weibul (Extreme Value) distribution guarantees that the utilities difference is logistic distributed (Ben-Akiva, 1985). This is a convenient property as it allows the introduction/elimination of new alternatives into the model without the need of re-estimating it. However, it yields the property of irrelevant and independent alternatives (IIA) where the odds of one alternative against others remain unchanged in a same choice set. Therefore, the market share of the new introduced alternative is proportional to the original market share. For instance, a choice set with the attributes showed in the next table is analysed.

	Alternative 1	Alternative 2	Alternative 3
Cost	15	20	25
Time	10	7	5
Headway	45	90	105

Table 1 - Three alternative choice set

According to the IIA property, a new option (alternative 4) in the choice set would reduce the other alternatives (1, 2, 3) probability of choice in a proportional way. Supposing the following attributes for the new alternative: cost=10, time=15, headway=35. Applying the coefficients previously found (_) given alternatives 1 to 3, and data set that follows the model assumptions, the new probabilities of each alternatives are found (see TABLE 2). It can be noticed that the rate between the new and the current probabilities remain unchanged (0.687).

Situation	p(1)	p(2)	P(3)	p(4)
Current	0.529	0.258	0.213	-
New	0.364	0.178	0.146	0.313
rate new/current	0.687	0.687	0.687	-

Table 2 - IIA Property

These situation yields only if the utilities cross elasticity are constant (Ortuzar & Willumsen, 1994). For instance, a new mode of public transport would have to equally affect both individuals using public or private transport. This is a situation that is hard to happen in a real situation as the owner of a car would not switch from private to public transport only because a new service has been introduced.

Several authors have studied discrete choice modelling errors: Daly and Ortuzar (1990); Horowitz (1981, 1981a); Koppelman (1979); Ben-Akiva and Lerman (1985); Ortuzar and Willumsen (1994). A summary of their research is showed in Errore. L'origine riferimento non è stata trovata.. The types of errors are classified according to whether they cause the development of an incorrect model or not. The table main objective is to define the scope of the research being developed regarding the errors that can be reduced by improving the design of the choice experiment.

Logit Modelling errors come from the analyst lack of information on the decision-making process.

Horowitz(1981a) has reported the following specification errors in the Multinomial Logit modelling:

I. Including an independent variable that is irrelevant to explain the behaviour of the dependent variable
II. The random part inside the data is not independent and identically distributed (for instance, alternatives are related somehow)
III. Individual taste variations. It is not difficult to find situations where a particular characteristic of an individual influences his taste.
IV. Omitting explanatory variables that were relevant to the decision process.
V. Existence of correlation between the random and the deterministic part of the utility.

According to Horowitz(1981a), specification errors can develop the problems showed in Errore. L'origine riferimento non è stata trovata.. Consistency is a parameter quality that indicates whether the estimator tends to its true value as the sample size increases or not.

CORRECT MODEL	INCORRECT MODEL
-Transfer errors: a model developed in one context is incorrectly applied to a different context.	- Measurement error: misinterpretation of the questions by respondents, etc.
- Data aggregation: originated from using disaggregate data to find aggregate forecasts. It can also leads to errors due to incorrect modes as the distribution of data randomness does not fulfil the model's assumptions.	- Sampling error: derived from applying results based on a sample, to the whole population.
	- Computational error: to find solutions to difficult problems that might be costly.
	- Specification errors: omission of a relevant variable, inclusion of an irrelevant variable, exclusion of taste variation, simplification of the form of variables, other specification errors.

Table 3 - Modelling errors from correct and incorrect models

Error source	Consequences to Logit modelling
1. Inclusion of an irrelevant explanatory variable	
Linear-in-parameters utility function	It can increase sampling errors, but it does not cause inconsistency or asymptotic bias in the model or its forecasts
non-linear utility function:	
-if there is a coincident parameter in the true and erroneous model	It increases the sampling errors, but the model is consistent.
-if a coincident parameter does not exist	It can cause the model to be inconsistent
2. Not independent and identically distributed random utility components	
linear-in-parameters utility function	It produces consistent parameter estimates, but inconsistent estimates of choice probabilities
non-linear utility function	It produces inconsistent parameter and choice probabilities estimates
3. Random taste variations	
linear-in-parameters utility function	It produces consistent estimates of the mean values of the taste parameters, but inconsistent estimates of choice probabilities.(Horowitz reported experiments with normal distributed taste)
non-linear utility function	It produces inconsistent estimates of utility function parameters, and choice probabilities
4. Omission of a relevant explanatory variable	It can cause inconsistent estimates of choice probabilities
5. Correlated random utility component and explanatory variable	The estimates of models' parameters and of choice probabilities are inconsistent

Table 4 - Error Specification - Consequences to Logit Modelling

Ortuzar (1994), Ben-Akiva and Lerman (1985), Fowkes and Wardman (1988) have stated that an important source of Logit modelling errors is related to the different individual's taste. Logit does not allow such characteristic inside the data and assume a mean value for taste. This might affect the calibration results. In spite of that, De Carvalho's experiments showed that individual taste variation can be a source of errors only in extreme situations (that report ρ^2 higher than 0.4). Fowkes e Wardman (1988) also did experiments using simulation tools to investigate the problem. Their experiments report ρ^2 of 0.63, 0.45, 0.54, e 0.23. Stated Preference experiments in real situations do not show such high figures for ρ^2 (usually around 0.2). For synthetic data showing a ρ^2 measure lower than 0.2, the model is adequate, no matter how much taste is included. This is so because the random term (which is Weibul distributed) is high, meaning that the attributes used to characterise the choice experiment alternatives are not relevant to the decision making process. These findings are apparently controversial.

3 Experiments

Experiments main objective is to investigate the influence of the trade-offs between alternatives to the calibration results. Two data sets with different attribute values difference for cost were framed using Logit modelling assumptions. FIGURE 1 shows the mean of the differences between the two data sets. The random error (Weibul) accounts for approximately 30% of the total alternative utility. A linear utility function was employed to find the utility value of an alternative. No constants (ASC's) were allowed. The true coefficients used to create the data were cost attribute= -1; time attribute= -1; headway attribute= -1. Data was used to calibrate the Logit model and results from both data sets are compared to each other. As the data set follows the model assumptions, the error coming from the absolute difference between both true (obtained from using the Limit Central Theorem) and forecast probabilities should be very small for both data sets.

The design of the experiments was developed following the fractional factorial design approach (Kokur et. al, 1982). Therefore orthogonality is considered, but some interactions between attributes are disregarded (this is normal procedure in real situations). The aim is to investigate the influence of the distance between alternative attribute values in the same choice set to calibration.

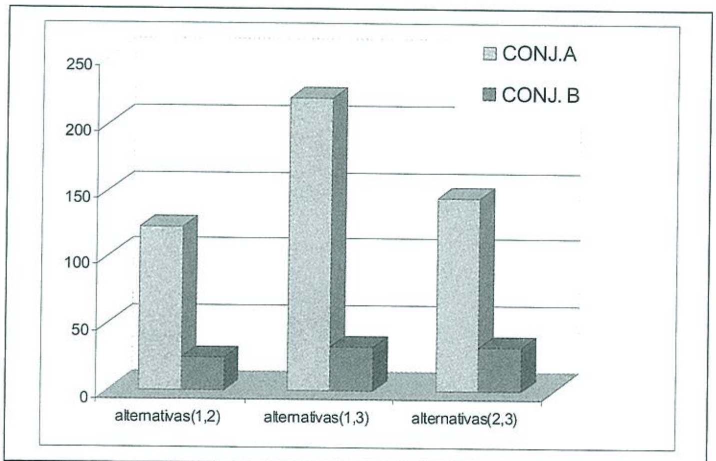

Figure 1 - Difference between the Deterministic Utilities in the choice sets

The mean of the distances between alternatives for both choice sets can be seen in FIGURE 1. It is emphasized that besides considering a trade-off between attributes, it was necessary to them to show different values for the difference between alternatives. In such way, the 'A' data set shows alternatives that are very much far from each other, whilst alternatives in the data set 'B' they are very close. Furthermore, both data sets have approximately the same amount of random inside as it can be seen in see Errore. L'origine riferimento non è stata trovata..

4 Results

Calibration results from both data sets (Errore. L'origine riferimento non è stata trovata.) show that, even though they have approximately the same amount of randomness, the data set with the higher difference between alternatives produced the best performance. It is worth to emphasise that both data were created and calibrated using the same parameters (true coefficients, linear utility function, and optimisation algorithm stopping criteria).

	DATA A (bigger difference)		DATA B (small difference)	
	Logit Coefficient (t)	Recovered Coefficient	Logit Coefficient (t)	Recovered Coefficient
Cost	-0.008728 (-54.24)	-1.00	-0.01848(-10.2)	-1.198
Time	-0.008688 (-7.91)	-0.99	-0.01718(-18.0)	-1.114
Headway	-0.008598 (-11.47)	-0.98	-0.01607(-25.7)	-1.042
Rho squared	0.2493	-	0.0479	-
Mean error*	0.0043	-	0.0111	-

* Refers to the absolute difference between the forecast and the true choice probabilities (obtained from simulation)

Table 5 - Logit Calibration Results

Analysis was performed using the tests: signs of coefficients, rho squared (ρ^2), mean of absolute error and the true coefficient recovering (through Weibul scaling factor). The analysis of the 't' test showed that the coefficients from Logit calibration are highly significant as expected because the data was created according to Logit assumptions. Yet, the cost attribute for data set A is far superior than the one obtained for data set B. This is line with the rho squared test that measures the calibration goodness-of-fit (data set A = 0.2493, data set B= 0.0479). The recovered coefficients indicate that the calibration performance of data set B was poorer. Finally, it was observed that forecasting error from calibration results of data A is smaller than the one obtained from calibration of data B. This analysis was developed according to simulation procedures (Convolution Theorem for 10,000 observations of each choice set). These testing results are in line with those obtained from calibration.

4.1 Discussion

Calibration results herein showed report experiments developed with synthetic data created according to Logit modelling assumptions using a linear utility function. In such a way, the only problem to be faced would be the random term inside the data (even though, Logit modelling would manage to scale it). Data set 'A' considered a bigger difference between alternatives, and had a random error higher than the one inside data 'B'. In such a way, it would be expected that results from data set 'A' were worse than those found for data set 'B'. Yet it did not happen. A closer analysis on the calibration procedure brought some light to this point. The function being optimised to get those calibration results is the log likelihood function as specified in (Ben-Akiva, 1985):

$$\log L(Y / X, \beta)=\sum_{n=1}^{N} \sum_{i \in C_n} Y_{ni} \left(\Sigma_k (\beta_k X_{nik}) - 1n \sum_{j \in C_n} e^{(\sum_k \beta_k X_{njk})}\right)$$

Where

Y= 'dummy' dependent variable

X= attribute value

ß = parameter to be found in the optimisation procedure

Index:

n= choice set
i, j = alternatives
k = parameter

This analysis indicated that the best point is dependent of the difference between alternative attribute values within the choice set. Therefore, the bigger that difference, the higher the value of the function at its best point. Small differences between the maximum and minimum points indicate that the function is flatter than the one with more dramatic differences. A flat function, even in case of linear shape of the utility function, increases the possibility of the algorithm getting stuck in a flat region. To overcome this problem it is necessary to reduce the algorithm convergence criteria what creates an additional problem: the necessary time for reaching convergence. FIGURE 2 shows the functions to be optimised in both situations investigated. It can be seen that function that models data 'B' (small difference) is a flatter that the one modelling data 'A'.

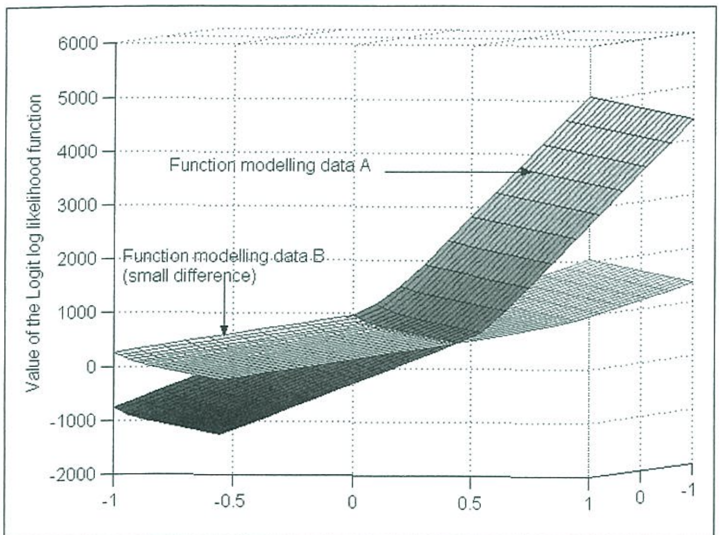

Figure 2 - Graph of the Logit log likelihood functions that model data sets 'A' and 'B'

4.2 *Homoscedasticity assumption*

So far experiments were conducted based on homoscedasticity (i.e. the Weibul error spread is the same in all alternatives). This assumption comes from the modelling approach about the data. The random error behind the data accounts for many different sources of choice influences that might not be significant (or known) to the process (i.e. income, fashion, etc.). However, there is no mention to whether these sources of error are in fact similar to all alternatives. For instance, if the source of the error were independent of the deterministic utility (the value known by the analyst), in order to hold homoscedasticity, the choice sets with large difference between alternatives would favour the alternative with highest attribute values. The following table shows results of experiments to investigate this fact. The data set A created ten times using homoscedasticity assumption and another ten times considering heteroscedasticity.

	DATA A				DATA B			
Variable	Mean	StdDev	Min.	Max.	Mean	StdDev	Min.	Max.
UTIL1	-264.83	182.93	-730.82	931.29	-102.41	86.44	-303.24	543.61
WEIBUL1	-161.31	166.71	-552.15	1073.04	-85.72	84.69	-285.34	485.53
UTIL2	-43.42	151.11	-394.16	1018.75	-72.42	84.30	-267.59	600.91
WEIBUL2	65.17	146.85	-260.82	1151.29	36.78	82.06	-150.19	630.53
UTIL3	66.19	147.41	-265.48	1198.04	37.58	82.81	-142.44	730.07
WEIBUL3	66.58	148.67	-254.16	1133.75	37.59	82.96	-148.76	693.61
Weibul/util	0.124				0.0827			

Table 6 - Utility and Random Error Statistics

	Homoscedasticity*				Hetereoscedasticity**			
Variable	Mean	StdDev	Min.	Max.	Mean	StdDev	Min.	Max.
UTIL1	-40.74	150.64	-389.18	1064.74	-77.97	77.64	-309.12	459.34
WEIBUL1	69.26	147.93	-234.18	1204.74	32.03	75.70	-154.12	614.34
UTIL2	-161.37	165.91	-557.44	1051.55	-159.92	168.67	-665.18	1112.9
WEIBUL2	66.13	146.69	-267.44	1191.55	67.58	161.59	-335.18	1442.9
UTIL3	-266.29	180.74	-711.26	770.01	-229.51	241.27	-995.03	2258.51
WEIBUL3	63.71	144.36	-267.88	990.01	100.49	231.82	-525.03	2698.51
LOGIT CALIBRATION								
Likelihood	-9939				-11872			
Rho	0.2460				0.0995			
Cost	-0.008445				-0.004390			
Time	-0.009886				-0.002073			
Head	-0.009225				-0.006708			
Ratio Cost	-53.7				-36.0			
Ratio Time	-9.0				-2.1			
Ratio Head	-12.2				-10.1			
Iterations	5				4			

Table 7 - Heteroscedastic effects on calibration

Both experiments pictured data containing Weibul error of 42% the size of the deterministic utility. The only difference is that in the first column, the Weibul random error was considered homoscedastic, while in the second column, it was heteroscedastic. Coefficients were all significant, although the heteroscedastic model was poorer than the homoscedastic one. There is no evidence whether or not real data random error is homoscedastic, and it needs further investigation. Even though using small trade off's in the experiment design can reduce the heteroscedasticity there is also a concern of two additional problems:

i. Decision-makers might not perceive the real difference between the alternatives, resulting that the choice process becomes purely random;
ii. The optimization algorithm might get stuck in a local minimum, resulting in poorer calibration.

Furthermore, other error sources such as taste variation, correlation, etc. also can happen and make the modelling task very difficult. There is also the question of which errors are worse for the forecastings (that might be reduced by an experiment design with alternatives very distinct to each other).

Although completely controlled, this is an empirical investigation. Next step is to investigate index representing the differences between alternatives that could be considered as guideline to reach quality calibration results and forecasting.

It is expected that these results be even worse in real situations as the difference between alternatives might not be a trigger to make the individual changes his/her choice. For instance, that would increase the error behind the data and the choice would be at random. The calibration could show worse results, as the alternatives in the choice set are closer.

5 Conclusion

Errors from Logit modelling calibration of discrete choice data have been analysed. This paper aims to investigate the influence of the differences between alternatives attribute values within a choice set to data calibration (optimisation). Experiments were based on simulation tools as a way to control the random inside the decision making process.

To the extent of the experiments developed, it was found that, Logit assumptions held the bigger the distance between alternative attribute values within the choice set the best the calibration results. This conclusion comes from both results: calibration, and the model ability to produce accurate forecasts. It was found that a small distance between alternatives attribute values creates an optimisation function that is quite flat. Then the optimisation algorithm (Newton-Raphson) might get trapped in a flat region and therefore a better point cannot be pursued. Results are expected to worsen in real situations, as the decision-maker might not reach a threshold value to which it would be worth to exchange the choice.

A closer analysis of the effects of random heteroscedasticity showed that it could deeply increase the calibration errors as it violates an IID Logit assumption. A question that needs further investigation is whether these source of errors represented by the Weibul distribution are in fact homoscedastic or are related to the size of the alternative deterministic and known utility.

References

(A) Books and Books chapters

Ben-Akiva, M. & Lerman R. (1985). *Discrete choice analysis.* Massachusetts Institute of Technology. USA.

Daganzo, C. F. (1979). *Multinomial probit: the theory and its applications to demand forecasting.* Academic Press, New York.

Horowitz, J. L. (1981a). *Sampling, Specification and Data Errors in Probabilistic Discrete Choice Models.* In Applied discrete choice modelling, Hensher, D. A., Appendix C. John Wiley & Sons, Inc. USA.

Ortuzar, J. D. & Willumsen, L. G. (1994). *Modelling Transport.* John Willey & Sons. West Sussex, UK.

(B) Journal papers

Daly, A. & Ortuzar J. D. (1990). *Forecasting and data aggregation: theory and practice.* Traffic Engineering + Control (11), December, pp. 632-643.

Fowkes, A. S. & Wardman, M. (1988). *Design of SP travel choice experiments, with special reference to taste variations.* Journal of Transport Economics and Policy XXII (1), pp.27-44.

Horowitz, J. L. (1981). *Identification and Diagnosis of Specification Errors in the Multinomial Logit Model.* Transp. Research-B, vol. 15, no.5, pp. 345-360.

Kokur, G.T.; Adler, T.; Hyman, W., & Aunnet, B. (1982). *Grade to forecasting travel demand with direct direct utility assessment.* United States Department of Transportation, Urban Mass Transportation Administration.

Koppelman, F.S. (1979). *Methodology for Analysing Errors in Prediction with Disaggregate Choice Models.* Transportation Research Record, n. 592, pp.17-23.

(C) Unpublished references

De Carvalho, M. C. M. (1998). *A comparison of Neural Networks and Econometric Discrete Choice Models in Transport.* Unpublished PhD thesis, University of Leeds, UK.

Acknowledgements

The author thanks CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazilian Ministry of Science and Technology- for financially supporting this research project.