Fabio Rossera

The use of log-linear models in transport economics : the problem of commuters'choice of mode

Quaderno N. 00-12

Decanato della Facoltà di Scienze economiche
Via Ospedale, 13 CH-6900 Lugano. Tel. 091 912 46 09 / 08. Fax 091 912 46 29

The use of log-linear models
in transport economics :
the problem of commuters' choice of mode

Fabio Rossera
Istituto di ricerche economiche
Università di Lugano

September 2000

Abstract

Before implementing a model of mode choice in urban travel, one needs to solve several problems relating to the framing of the analysis. The present study argues that loglinear models may be usefully applied to the analysis of categorical variables. Two problems are considered. Firstly, the methods proposed by loglinear models for identifying the interrelationships existing between a set of variables are exploited in order to evaluate the possibilities of reducing the dimensions of the table of reference. The case under examination relates to the possible extension of results obtained at national level as estimates of the characteristics of individual regions. The alternative choices open to the user differ according to several circumstances, the length of the journey being among the most important. In the second part of the study choice sets are defined to narrow down the analysis of the essential alternatives, without however sacrificing the representativity of the results.

1. Introduction

Estimating a model on the choice of mode of travel in an urban context requires some preliminary clarifications, which are usually carried out too hastily and cursorily. In my opinion, some problems could be investigated with satisfying results with the aid of methods devised for the analysis of categorical variables. Nowadays, surveys of users' behaviour generally provide information that can be ordered into multilevel contingency tables. In such cases, log-linear models have proved to be particularly suitable for identifying the interrelationships existing among variables of interest. In the present study I will consider two such problems.

The first concerns the possibility of extending the results obtained at a broad spatial context - national, say - to a lower (regional) level. The advantages are first of all the greater robustness of the former adjustments, owing to the more substantial quantity of data available at this level. This is of particular relevance when interrelationships among variables are considered, and an important number of cells may be empty at the lower level.

The second problem concerns the definition of the choice sets of the individual categories of users, when the heterogeneity present in the sample precludes a uniform representation. Determining which means of travel is available for which kind of user also implies evaluating the consequences of eliminating some alternatives which, at first sight, seem little relevant. The log-linear methodology has the right potential for testing hypotheses in such a context.

Section 2 of this study contains a brief reminder of the main characteristics of log-linear models, in particular the interpretation of the parameters and the classes of models used. Attention is drawn to the kinds of three-dimensional models used in the present study. A brief presentation of the frame from which I derived the most important issues follows in section 3. For the sake of exemplification, I have considered the commuters' journeys to work. The data are derived from a survey carried out by the Swiss Federal Statistical Office for the whole of Switzerland. Sections 4 and 5 present the results obtained. In section 4, some methodological aspects are examined in more detail in order to specify the hypotheses submitted to a formal test. Some final remarks conclude this paper.

2. Hierarchical models for three-dimensional tables[1]

In the case of three interdependent variables the standard model reads

$$\log m_{ijk} = m + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{ik}^{XZ} + l_{jk}^{YZ} + l_{ijk}^{XYZ}$$

The capital letters in the indices refer to the individual variables – mode, distance, region, in the example presented below – the small letters indicate the categories considered: car, bike, etc. Parameters with a single index refer to main effects; those with two symbols to interaction between two effects; while the last parameter measures the joint interaction of the variables. The exact interpretation of these parameters will be given below.

In this model – called the ‘saturated model’ as all possible effects are taken into account – *1+I+J+K+IJ+IK+JK+IJK* parameters stand against *IJK* observations. Imposing identification constraints - setting to zero one of the parameters in each of the dimensions, e.g. those codified by *I,J,K* - their number is reduced to *IJK* and the model is completely determined. More formally, in this case the parameters satisfy the constraints:

$$l_I^X = l_J^Y = l_K^Z = l_{Ij}^{XY} = l_{iJ}^{XY} = l_{IJ}^{XY} = ... = l_{IJK}^{XYZ} = 0 \quad \forall i, j, k$$

Writing out the system in full and solving for the individual parameters, the following interpretations result for each of them (Table 1)[2]. For the interpretation of parameters I use the probability of an event occurring in a particular cell (the p_{ijk}’s) instead of the corresponding expected count (the m_{ijk}’s). Apart from parameter μ, which has to be re-scaled, this has no effect on the other coefficients.

μ is a scale parameter referring to the common reference cell. In conformity with the point just mentioned, the asterisk indicates that it differs from that given in the previous model equation. The three following parameters λ measure the deviation of a category of an individual variable from its reference category, the other variables staying put at their respective reference value. To anticipate on the adjustment following below: one may consider, say, the log of the odds of cycling to work rather than using the car, the confrontation being

[1] This presentation is mainly based on: Agresti (1990, ch. 5 to 7), Bishop et al. (1975, ch. 2) and Wrigley and Brouwer (1986, ch. 11). For a more econometric stance, see Maddala (1983, ch.5).

[2] See, in particular, Wrigley, cit., 451, where a system with these kinds of constraints is termed “corner-effect” system, as distinguished from a “centred-effect” system, where the sums of the various categories of parameters, and not the parameters themselves, are set to zero.

carried out at the level of reference of the two other variables (in the example presented below: over a distance of 5 to 10 km and concerning the whole Swiss territory excepting the Canton Ticino).

Table 1 Interpretation of coefficients in I xJ xK models

$$m^{*} = \log p_{IJK}$$

$$l_{i}^{X} = \log\left(p_{iJK} \big/ p_{IJK} \right)$$

$$l_{j}^{Y} = \log\left(p_{IjK} \big/ p_{IJK} \right)$$

$$l_{k}^{Z} = \log\left(p_{IJk} \big/ p_{IJK} \right)$$

$$l_{ij}^{XY} = \log\left\{ \left(p_{ijK} \big/ p_{iJK} \right) \Big/ \left(p_{IjK} \big/ p_{IJK} \right) \right\}$$

$$l_{ik}^{XZ} = \log\left\{ \left(p_{iJk} \big/ p_{iJK} \right) \Big/ \left(p_{IJk} \big/ p_{IJK} \right) \right\}$$

$$l_{jk}^{YZ} = \log\left\{ \left(p_{Ijk} \big/ p_{IjK} \right) \Big/ \left(p_{IJk} \big/ p_{IJK} \right) \right\}$$

$$l_{ijk}^{XYZ} = \log\left\{ \frac{p_{ijk}\, p_{IJk}}{p_{iJk}\, p_{Ijk}} \Big/ \frac{p_{ijK}\, p_{IJK}}{p_{iJK}\, p_{IjK}} \right\}$$

In the three following parameters, the logs of the cross odds are considered, the comparisons between two variables being carried out at a fixed level for the third. To resume the example above: the ratio bike / car is considered, say, over a distance under 1 km compared to one between 1 and 5 km.

In the case of the last parameter, for the three-way interaction, the comparison between pairs of two variables is made at the different levels of the third. I want to deal exclusively with the two-way interaction parameters.

The saturated model is of limited empirical interest, as it reconstructs the data set exactly. Usually, it acts as a reference for other more parsimonious models, in terms of number of parameters. In the latter cases, supplementary restrictions have to be imposed on the coefficients. The most interesting are those allowing to build classes of models, each of them taking into account a particular form of dependence among the variables. Hereunder we list the most relevant to the problems at hand[3].

Three classes can be distinguished in this table. A first model presents only main effects and no interaction effects between the variables, the three following specifications with two-way interactions – respectively: one (three

[3] See Agresti, cit.,144

possible cases), two (also three possible cases), and all possible combinations -, the last – the 'saturated' model – with also the three-way interaction effect. With the exception of this last model, only models with some kind of interaction are of practical interest.

Table 2 Typology of models

Type	Symbol
$\log m_{ijk} = m + l_i^X + l_j^Y + l_k^Z$	*(X,Y,Z)*
$\log m_{ijk} = m + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY}$	*(XY,Z)*
$\log m_{ijk} = m + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{ik}^{XZ}$	*(XY,XZ)*
$\log m_{ijk} = m + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{ik}^{XZ} + l_{jk}^{YZ}$	*(XY,XZ,YZ)*
$\log m_{ijk} = m + l_i^X + l_j^Y + l_k^Z + l_{ij}^{XY} + l_{ik}^{XZ} + l_{jk}^{YZ} + l_{ijk}^{XYZ}$	*(XYZ)*

3. The frame of reference supporting this study

Commuters' choice of a means of transportation is strongly related to the distance to be covered. This may be easily seen in Table 3, where data referring to the whole of Switzerland are presented. Only movements from home to work have been considered. The data are taken from a survey on the mobility characteristics of the Swiss population in 1994. About 18'000 households were surveyed in a two-day interview lasting 45 minutes on average.

On very short distances, journeys on foot clearly prevail. The predominance is, however, quickly transferred to the car. This is what happens on distances longer than one kilometre, although bicycles and public transport (like buses and trams) make their appearance, too. The use of public transport drops considerably beyond 10 km, where the railway plays its major role. For these trips, too, however, the private vehicle occupies centre stage.

The two problems on which I wish to dwell briefly are clearly in evidence in this table.

At first, a large number of cells show very small figures, even zeroes (even looking at the aggregate national level!). In these circumstances, two cases must be distinguished. On the one hand, it may be impossible to have observations in the cell (structural zeroes). On the other hand, the event may have a very small probability that can only be estimated by means of a very large sample (sampling zeroes). The fact that one person in the table reported a trip to work longer than 100 km made me consider all zeroes as sampling zeroes. Anyway, sparseness in a table may have disturbing consequences for the analysis, even generating indeterminacy in the estimation[*], and should therefore be avoided. One way to avoid it would be to aggregate the categories of the variables or even the variables themselves. Another, more interesting, possibility would be to increase the sample size. In the present case this can be done when, by analysing some strata in the sample, one is able to prove that the variable of stratification has no impact on the variables of interest; in other words, if one proves that a table analogous to that presented above, for a sub-region, will look exactly like the one shown. In such a case, the analysis will resort directly to the aggregated table. Our evaluation will be based on the case of a small region, the Canton Ticino. The test results are presented in the following section.

Table 3 Mode by distance distribution in the sample

	➙ ***Distance in km***	< 1	1-5	5-10	10-20	20-50	50-100	> 100	Total
				counts					
↘ ***Means***	On foot	852	302	7	1	1		1	1164
	By bike	132	550	80	28	1			791
	By moped	5	88	47	13	2			155
	By motorcycle	2	41	26	29	6	2	1	107
	By car (as driver)	168	875	984	1169	725	90	30	4041
	By car (as passenger)	1	54	37	30	23	1		146
	By train	7	31	70	203	198	51	16	576
	By public transport	17	448	329	154	28	1	1	978
	Total	1184	2389	1580	1627	984	145	49	7958
				Percentage (referred to the sample total)					
↘ ***Means***	On foot	*0,11*	*0,04*	*0,00*	*0,00*	*0,00*	*0,00*	*0,00*	*0,15*
	By bike	*0,02*	*0,07*	*0,01*	*0,00*	*0,00*	*0,00*	*0,00*	*0,10*
	By moped	*0,00*	*0,01*	*0,01*	*0,00*	*0,00*	*0,00*	*0,00*	*0,02*
	By motorcycle	*0,00*	*0,01*	*0,00*	*0,00*	*0,00*	*0,00*	*0,00*	*0,01*
	By car (as driver)	*0,02*	*0,11*	*0,12*	*0,15*	*0,09*	*0,01*	*0,00*	*0,51*
	By car (as passenger)	*0,00*	*0,01*	*0,00*	*0,00*	*0,00*	*0,00*	*0,00*	*0,02*
	By train	*0,00*	*0,00*	*0,01*	*0,03*	*0,02*	*0,01*	*0,00*	*0,07*
	By public transport	*0,00*	*0,06*	*0,04*	*0,02*	*0,00*	*0,00*	*0,00*	*0,12*
	Total	*0,15*	*0,30*	*0,20*	*0,20*	*0,12*	*0,02*	*0,01*	*1,00*

Source: Microcensus on mobility 1994, Federal statistical office, Berne

A second problem is connected to the first. The table shows that, over different distances, only a small subset of alternatives is actually open to the user. These sets (the choice sets) diverge substantially from one category to the other in terms of length of the journeys involved. A question comes to mind

[*] The SPSS software I used for estimation puts a figure of 0.5 in the cells where a zero is encountered.

about the possibility of restricting the number of alternatives at the individual levels, by emptying the greatest possible number of cells. Obviously, one can go as far as to say that the conclusions derived from the analysis may not be affected. In statistical terms, this is the problem of imposing a set of supplementary restrictions on the model, beyond those usually taken into account. The findings are presented in section 5.

Both problems lie one step ahead of the evaluation of the modal choice as considered in traditional urban transport economics[4]. They might provide a solid basis from which to conduct that kind of analysis.

4. Testing for independence

a. Alternative hypotheses

In order to test for independence among the variables, in view of exploiting the consequent possibility of collapsing the table to smaller dimensions, several forms of independence may be considered. In this case, the taxonomy of models presented above needs further clarification.

The practical question may be formulated as follows: when can one confine the analysis to aggregated tables? And, for which variables and when, on the contrary, does one need a more detailed examination? This problem can be solved by considering the cross-odds deriving in the different models for the individual pairs of variables. When these statistics for partial and for marginal tables coincide, one may simplify the analysis by considering only the latter. Formally this means that one is looking for situations where the following is true[5]:

$$q_{ij}^{XY} = q_{ij(1)} = q_{ij(2)} = ... = q_{ij(K)} = q_{ij}, \qquad 1 \leq i \leq I-1, 1 \leq j \leq J-1$$

with

$$q_{ij}^{XY} = \frac{p_{ij+}p_{i+1,j+1,+}}{p_{i+1,j,+}p_{i,j+1,+}}, \qquad 1 \leq i \leq I-1, 1 \leq j \leq J-1$$

$$q_{ij(k)} = \frac{p_{ijk}p_{i+1,j+1,k}}{p_{i+1,j,k}p_{i,j+1,k}}, \qquad 1 \leq i \leq I-1, 1 \leq j \leq J-1$$

These are the cross-odds in the marginal and partial tables. By partial is meant the series of tables each representing the relationships between variable *X* and variable *Y* at each level of the third variable *Z*. By marginal is meant the table where all values of the relationships between *X-Y* are added up over the different levels of variable *Z*.

[4] Among others: Domencich and McFadden (1975) and Ben-Akiva and Lerman (1985).

[5] Agresti, cit., 145-6.

Three forms of independence may be distinguished[6]. The conditions in terms of parameters θ relating to each of them are indicated in Table 4, below. In its central part are arranged the parameters (cross-odds) relating to the individual configurations. For each model its sub-table is subdivided into two parts. The three symbols/figures under the diagonal line report the values calculated in the partial tables, those in the upper triangle the values for the marginal table. An example will clarify what is meant. In the second model - *(XY,Z)*, where *X* and *Y* are related, *Z* is jointly independent of both – the cross-odds between *X* and *Y* estimated at different levels of *Z* are all equal and also the same as the odds estimated in the table obtained by summing up the values at the different levels of *Z*. To arrive at this conclusion one only needs to confront the figures in cells (1,2) and (2,1) of the sub-table.

Table 4 Partial and marginal associations in *IxJxK* models

	?	*marginal*			*symbols*
?		*X*	*Y*	*Z*	
	X	-	$1.00, \forall i, j$	$1.00, \forall i, k$	*(X,Y,Z)*
	Y	$1.00, \forall i, j$	-	$1.00, \forall j, k$	
	Z	$1.00, \forall i, k$	$1.00, \forall j, k$	-	
partial	*X*	-	$q_{ij}, \forall i, j$	$1.00, \forall i, k$	*(XY,Z)*
	Y	$q_{ij}, \forall i, j$	-	$1.00, \forall j, k$	
	Z	$1.00, \forall i, k$	$1.00, \forall j, k$	-	
	X	-	$q_{ij}, \forall i, j$	$q_{ik}, \forall i, k$	*(XY,XZ)*
	Y	$q_{ij}, \forall i, j$	-	$q_{jk}^{YZ}, \forall j, k$	
	Z	$q_{ik}, \forall i, k$	$1.00, \forall j, k$	-	
	X	-	$q_{ij}^{XY}, \forall i, j$	$q_{ik}^{XZ}, \forall i, k$	*(XY,XZ,YZ)*
	Y	$q_{ij(k)}, \forall i, j$	-	$q_{jk}^{YZ}, \forall j, k$	
	Z	$q_{ik(j)}, \forall i, k$	$q_{jk(i)}, \forall j, k$	-	

a) *Mutual independence* between the variables. (Model *(X,Y,Z)*). In this case, all variables are independent, considered two at a time. The characteristics of this situation are shown in the following formula

$$p_{ijk} = p_{i++} p_{+j+} p_{++k}$$

The probability associated with any individual cell may be fully reconstructed with the marginal probabilities referring to the individual effects. One recognises the first model in Table 2. In view of the fact that all variables are mutually independent, the parameters all equal 1, in the

[6] The exposition follows Agresti, cit., 146-9, and Haberman (1978, 197-207). For the evaluation of marginal and partial coefficients see also Bishop et al., 37-42.

partial as well as in the marginal tables. Once again, with all interactions missing, this case is of no particular interest in the present context.

b) *Joint independence*. (Model *(XY,Z)*). One variable is independent of the other two. In the case indicated in the table, *Z* is independent of *X*; it is also independent of *Y*. The probability may be represented as a two-way independence for *Z* and a new variable with the combination of levels of *X* and *Y*, as in the following formula

$$p_{ijk} = p_{ij+}\, p_{++k}$$

Only the parameters measuring the interaction of *X* with *Y* are present in this model. In our context, this represents clearly the most interesting configuration. Considering that only one relationship matters -- that between distance and means of transport chosen -- I could concentrate on it without fear of neglecting a probable cross-effect produced by the third variable.

In view of the independence existing between Z and X and Z and Y, the related parameters, as shown in Table 4, reveal these properties at both levels. A relationship exists between variables *X* and *Y*, but the measure of its intensity is the same in the partial as in the marginal tables.

c) *Conditional independence.* (Model *(XY,XZ)*). In this case one finds independence between two variables but only when the relationship is considered at the different levels of the third variable. In this case the formula reads

$$p_{ik/j} = p_{i+/j}\, p_{+k/j}$$

On the contrary, this independence is no longer present when one considers the relationship in the marginal distribution. In the present case this could be interpreted in the following way. Considering the tables combining mode used by region at different levels of distance covered, one may find no interrelationships between the two variables. Summing up these different tables, i.e. collapsing the whole table on the distance dimension, a dependence between mode and region might be found. This would be the case when one of the variables (e.g. region) is related to the third. This would give: in principle, the choice of mode doesn't differ in the different regions but as the characteristics of distance differ in the individual regions, and region and distance are interrelated, a relationship between mode and region is introduced in the global table. In this case given *Y*, *X* and *Z* are independent. This shows in the coefficient in the partial tables but not in the collapsed table, as the relation between *Y* and *Z* is interfering in this case. The other relationships are the same at the different levels. This model is also viable in the present case, provided that the desirable relationships obtain.

To the three models of independence a fourth could be associated. It is of interest when no independence in either form has been discovered.

d) *No three-factor interaction* (Model *(XY,XZ,YZ)*). It may be of interest to notice that, at least in principle, this model may not be completely devoid of interest either. Although all parameters differ on two levels of analysis, they are the same at each level of the third variable considered. In this context, the third variable, *Z* say, is represented by the regional subdivisions. For all regions the parameters θ_{ij} are the same. For instance:

$$q_{ij(1)} = q_{ij(2)} = ... = q_{ij(K)} = q_{ij}, \qquad 1 \leq i \leq I-1, 1 \leq j \leq J-1$$

So, when the various regions are above a certain size, it may be useful to take into account the data of a large region in order to estimate the parameters of a smaller one. However, a three-way interaction is likely to exist in such cases.

b. Tests results[7]

In Table 5 I have recapped the statistics supporting to the evaluation of degree of accuracy of the models. G^2 is the likelihood ratio of the model considered, compared to the saturated model of exact fit; DG^2 is the increment of this statistics derived by comparing two models contiguous in this hierarchy. The latter may be interpreted in the sense that the simpler model could be used according to the principle of parsimony, if the two models are not found to differ in terms of statistical fit. The degrees of freedom related to these tests as well as *P*-values in the $c2$ distribution are reported in the last two columns.

Considering that these tests are not independent of one another, the significance should be subdivided between them. In view of the fact that four comparisons are carried out one should perform each of these test at a level of 1 - (0.90)**0.25 = 0.026 if one wants to ensure an overall Type-I error probability of, say, α=0,10.

All but the first model match the data well, so that no significant departure from the saturated model is observed. Comparing the three, we see that the fit of the model without three-way interaction doesn't differ significantly – in this case obviously $G^2 = DG^2$ – from that of the saturated model; the fit of the model with conditional independence lies at the limit compared with that of the former model. The model with joint independence is at the same level with that showing conditional independence. Following the principle of parsimonious estimation one would accept the model of joint independence as the most convenient. It shows a good general fit when compared with the saturated model, and the difference between it and the model with conditional

[7] Agresti, cit., 174-6 and 211-3.

independence is immaterial (cf. DG^2 at a significance level of 0.5593). From this, one may conclude that the mode choice at different levels of distance is not conditioned by the regional context considered.

Table 5 Goodness-of-fit statistics

Model	G^2	DG^2	*Df*	*c2*
(X,Y,Z)	5253.99	5208.47	97 42	0.0000 0.0000
(XY,Z)	45.52	4.88	55 6	0.8152 0.5593
(XY,XZ)	40.64	14.43	49 7	0.7966 0.0440
(XY,XZ,YZ)	26.21	26.21	42 42	0.9731 0.9731
(XYZ)	0.0		0	

4. Defining the choice sets

After verifying the independence of the combination of the variables relating to mode used and distance covered from the variable related to the spatial subdivisions, let us look at the most interesting characteristics evidenced by the parameters estimated. Using the estimates of the expected counts (the m_{ijk}'s), Table 6 presents the odds of the different choice of mode at different levels of spatial separation. These indicators can be interpreted most easily.

Table 6 Odds of binary choices of means, at different distance levels

	→ ***Distance in km***	< 1	1-5	5-10	10-20	20-50	50-100	> 100
↘	On foot / by car	3,40	0,12	0,00	0.00	0,00	0,00	0,04
M	By bike / by car	1,00	0,49	0,05	0.05	0,00	0,00	0,00
e	By moped by car	0,06	0,09	0,03	0.03	0,00	0,00	0,00
a	By motorcycle / by car	0,03	0,04	0,03	0.03	0,01	0,02	0,00
n	By car (as passenger) / by car	0,03	0,05	0,04	0.04	0,03	0,02	0,00
s	By train / by car	0,03	0,04	0,12	0.12	0,29	0,68	0,39
	By public transport / by car	0,20	0,50	0,26	0.26	0,02	0,00	0,04

Remarks: The figures refer to the odds of different modes, at different levels of distance, derived from the table of counts adjusted to the *(XY,Z)* model, collapsed over variable “Regions”.

One can see from the figures in this table how, for individual distances, the choice is restricted to a limited number of alternatives. For instance, on distances under one kilometre, users prefer walking rather then using the car in more than three cases to one. On this distance, the bike is the only other competitor after the car. On distances between one and five kilometres, bike and one of the public means of transports break even, but scoring less than he car by half. From this category of distance upwards, only one mode competes with the car: public transport on shorter, the railway on longer distances.

Analogous conclusions can be drawn from the direct examination of the parameters of the model relating to the interaction between mode and distance. Table 7 reports a transformation of these parameters which may be understood most easily, their exponentiation. This is tantamount to transforming a λ-parameter, measuring the interaction between two variables at a defined combination of levels, into a cross-odds ratio. In fact, as explained in section 2, these figures may then be interpreted as the ratios of two odds: that of the two means indicated by the wording of the line of the table, taken at one level of distance, confronted with that related to the same odds taken at another distance, both distances being indicated in the respective column titles.

Table 7 Cross-odds of means by distance interaction

↘	→ ***Distance in km***	< 1 / 5-10	1-5/ 5-10	10-20/ 5-10	20-50/ 5-10	50-100/ 5-10	> 100/ 5-10
M	On foot / by car	857,3	31,5	(0,2)	(0,5)	(0,0)	11,0
e	By bike / by car	20,6	8,1	0,3	(0,0)	(0,0)	(0,0)
a	By moped by car	2,2	3,3	0,5	0,1	(0,0)	(0,0)
n	By motorcycle / by car	(1,0)	(1,4)	(0,6)	0,3	(0,6)	(0,0)
s	By car (as passenger) / by car	(0,8)	(1,3)	0,6	(0,7)	(0,4)	(0,0)
	By train / by car	0,3	0,3	1,6	2,4	5,7	3,2
	By public transport / by car	(0,8)	2,0	0,4	0,1	(0,0)	(0,2)

Remarks: The figures refer to the $q_{ij} = exp(l_{ij})$, related to means by distance combinations and are derived from the *(XY,Z)* model. The entries are identical – up to approximation errors – to the rates of those in the individual columns of the preceding table to those in the third column of data in the same table.

The table reads: for example, travelling by bike instead of by car is 20,6 times more probable on distances shorter than 1 km than on those lying between 5 and 10 km.

Figures in parentheses are not significant at 95% two-tailed test (carried out on the original parameters λ)

These figures are complementary to those in the previous table. However, they are far from self-evident[8]. They seem to be easier to read along the individual rows, where the ordinality present in the categories of the distance variable helps. In any case, one should be careful about what to make of the extreme values in this table. For instance, the figure of 857,3 only signals that the use of this means is highly concentrated on the first distance category, when it is practically absent from other segments of the market. Also, the fact that the use of a public means of transport occupies a perfectly dignified position at one to five on distances below one kilometre (Table 6, col 1 of data, last row) as on those between one and five kilometres (third column) is totally obscured by a 0.8, and what's more, insignificant, coefficient.

Let us now turn to the question: Which alternatives are to be considered for each of these strata, and which can be eliminated?

A first solution would consist in eliminating from the table all the counts with an insignificant probability. A more elegant and systematic procedure would envisage considering the estimated coefficients and imposing a set of restrictions on the model; hence one would be able to test directly for the statistical validity of the frame imposed. This is the way I will now follow.

Capitalising on the results obtained so far, I will move on to the last step of my analysis by using the cases classified in the following reduced table of data. Only a two-way table will be considered.

Table 8 Mode choice by distance, reduced table

↘	→ ***Distance in km***	< 1	1-5	5-10	10-20	20-50
M	On foot	852	302	7	1	1
e	By bike	132	550	80	28	1
a	By car	168	875	984	1169	725
n	By train	7	31	70	203	198
s	By public transport	17	448	329	154	28

One can see that four categories have been eliminated: moped, motorcycle, motorcycle (passenger) and automobile (passenger). As to the other variable, distances over 50 km have been dropped.

At this point, it becomes interesting to subject to a test a set of restrictions that have usually been considered as valid in a great number of studies, especially those carried out in a practical planning context. The design that has undergone a test is presented in Table 9.

[8] An expert goes as far as stating that "it is tempting to avoid contact with parameter estimates"(Fingleton, 1990, 6).

Table 9 Constraints design

↘	→ ***Distance in km***	< 1	1-5	5-10	10-20	20-50
M	On foot			= 0	= 0	= 0
e	By bike			= 0	= 0	= 0
a	By car					
n	By train	= 0	= 0			
s	By public transport	= 0				= 0

In this way, there would be three alternatives available under 1 km. – on foot, by bike, and by car –, a further alternative between 1 and 5 km., public means, three alternatives – car, train, public means from 5 to 20 km – and in the last category the choice would be restricted to car and railway.

To evaluate this design a set of restrictions has to be tested:

$$R\,l = q$$

R is a q by K design matrix for specifying the restrictions, where q is the number of restrictions tested and K is the total number of coefficients in the model. l is a K by 1 vector of the coefficients, and ***q*** is the q by 1 vector of values to be satisfied by the restrictions. In the present case ***q*** is a zero vector, containing the values of the cells to be removed. To clarify matters let's take an example, the restriction to be imposed on a journey on foot at a distance of 5 to 10 km. In a two-way saturated model the corresponding equation would read, still considering the last category of every variable as the identifying restriction:

$$\log m_{on\ foot,5-10km} = m + l^{\text{mod}\,e}_{on\ foot} + l^{dis\,\text{tan}\,ce}_{5-10km} + l^{\text{mod}\,e\ and\ dis\,\text{tan}\,ce}_{on\ foot,\,5-10km}$$

Now, we would have:

$$r_1^{'} = [1111] \quad x \quad transpose \left[m \quad l^{\text{mod}\,e}_{on\ foot} \quad l^{distan\ ce}_{5-10km} \quad l^{\text{mod}\,e and\ dis\,\text{tan}\,ce}_{on\ foot,5-10km}\right] = 0$$

The test statistics for the full set converges in distribution to a c 2 variable.

These figures show very clearly how the number of restrictions has been overestimated, as the hypothesis is clearly rejected. Some supplementary tests have been carried out, which have demonstrated that all data but only two are significant in the table. In fact, only the categories referring to the journeys on foot, at distances, respectively, covering 10 – 20 km and 20 - 50 km, may be omitted. In a purely statistical sense, these results imply that models as the first presented in Table 10 are not exempt from mis-specification flaws. However, caution is advisable if we consider that the consequences of the considerable sparseness of the data in these tables are still to be assessed.

Table 10 Tests of hypotheses

Model restrictions	*DG^2*	*Df*	*c2*
as presented in Table 9	1702.97	10	0.0000
reintroduced *(4)*	949.81	9	0.0000
reintroduced *(4), (5), (8), (9), (10)*	69.03	5	0.0000
reintroduced *(1), (4), (5), (7), (8), (9), (10)*	10.66	3	0.0137
reintroduced *(1), (4), (5), (6), (7), (8), (9), (10)*	0.49	2	0.7824

4. Conclusions

Up to now, there has been little consensus or even no great liking for log-linear models in transport economics. This is understandable, as what planners mostly need are fine tuned calibrations of flows and loadings of the transport infrastructure. However, some use may be found for these methods, exploiting in particular their aptitude at identifying intricate relationships existing among sets of variables. In the present study I have shown how this peculiarity can be used for solving some problems before a behavioural analysis or a planning exercise is carried out.

Two problems have been dealt with, both relating to a difficulty inherent in the application of survey data to mobility behaviour, namely the substantial imbalance present in such samples. To begin with, we considered the possibility of enlarging the information set by eliminating all superfluous subdivisions in the sample. Secondly, we addressed the question of the largest possible number of entries to be eliminated from a table in order to focus on the most relevant elements.

Both investigations have shown the utility of the methodology considered, although the second question may require and deserve a more detailed and in-depth analysis.

References

Agresti, A. (1990) Categorical Data Analysis. Wiley, New York.

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

Bishop, Y., Fienberg, S.E., and Holland, P.W. (1975) Discrete Multivariate Analysis. MIT Press, Cambridge, Mass.

Domencich, T. and McFadden, D. (1975) Urban Travel Demand - A Behavioral Analysis. North Holland, Amsterdam.

Fingleton, B. Loglinear Models:Theory and Applications. Paper presented at the workshop on Discrete Choice Models. CSI Piemonte. May 30th, 1990. Turin.

Haberman, S.J. (1978) The analysis of qualitative data. University of Chicago Press, Chicago.

Maddala, G.S. (1983) Limited-dependent and qualitative variables in econometrics. Cambridge University Press, Cambridge.

Wrigley, N. and Brouwer, F. (1986) Qualitative Statistical Models for Regional Economic Analysis. In: Nijkamp, P. (ed), pp. 443-492, North Holland, Amsterdam.

QUADERNI DELLA FACOLTÀ

I quaderni sono richiedibili (nell'edizione a stampa) alla Biblioteca universitaria di Lugano
via Ospedale 13 CH 6900 Lugano
tel. +41 91 9124675 ; fax +41 91 9124647 ; e-mail: biblioteca@lu.unisi.ch
La versione elettronica (file PDF) è disponibile all'URL:
http://www.lu.unisi.ch/biblioteca/Pubblicazioni/f_pubblicazioni.htm

The working papers (printed version) may be obtained by contacting the Biblioteca universitaria di Lugano
via Ospedale 13 CH 6900 Lugano
tel. +41 91 9124675 ; fax +41 91 9124647 ; e-mail: biblioteca@lu.unisi.ch
The electronic version (PDF files) is available at URL:
http://www.lu.unisi.ch/biblioteca/Pubblicazioni/f_pubblicazioni.htm

Quaderno n. 98-01
P. Balestra, *Efficient (and parsimonious) estimation of structural dynamic error component models*

Quaderno n. 99-01
M. Filippini, *Cost and scale efficiency in the nursing home sector : evidence from Switzerland*

Quaderno n. 99-02
L. Bernardi, *I sistemi tributari di oggi : da dove vengono e dove vanno*

Quaderno n. 99-03
L.L. Pasinetti, *Economic theory and technical progress*

Quaderno n. 99-04
G. Barone-Adesi, K. Giannopoulos, L. Vosper, *VaR without correlations for portfolios of derivative securities*

Quaderno n. 99-05
G. Barone-Adesi, Y. Kim, *Incomplete information and the closed-end fund discount*

Quaderno n. 99-06
G. Barone-Adesi, W. Allegretto, E. Dinenis, G. Sorwar, *Valuation of derivatives based on CKLS interest rate models*

Quaderno n. 99-07

M. Filippini, R. Maggi, J. Mägerle, *Skalenerträge und optimale Betriebsgrösse bei den schweizerische Privatbahnen*

Quaderno n. 99-08

E. Ronchetti, F. Trojani, *Robust inference with GMM estimators*

Quaderno n. 99-09

G.P. Torricelli, *I cambiamenti strutturali dello sviluppo urbano e regionale in Svizzera e nel Ticino sulla base dei dati dei censimenti federali delle aziende 1985, 1991 e 1995*

Quaderno n. 00-01

E. Barone, G. Barone-Adesi, R. Masera, *Requisiti patrimoniali, adeguatezza del capitale e gestione del rischio*

Quaderno n. 00-02

G. Barone-Adesi, *Does volatility pay?*

Quaderno n. 00-03

G. Barone-Adesi, Y. Kim, *Incomplete information and the closed-end fund discount*

Quaderno n. 00-04

R. Ineichen, *Dadi, astragali e gli inizi del calcolo delle probabilità*

Quaderno n. 00-05

W. Allegretto, G. Barone-Adesi, E. Dinenis, Y. Lin, G. Sorwar, *A new approach to check the free boundary of single factor interest rate put option*

Quaderno n. 00-06

G.D.Marangoni, *The Leontief model and economic theory*

Quaderno n. 00-07

B. Antonioli, R, Fazioli, M. Filippini, *Il servizio di igiene urbana italiano tra concorrenza e monopolio*

Quaderno n. 00-08

L. Crivelli, M. Filippini, D. Lunati. *Dimensione ottima degli ospedali in uno Stato federale*

Quaderno n. 00-09

L. Buchli, M. Filippini, *Estimating the benefits of low flow alleviation in rivers: the case of the Ticino River*

Quaderno n. 00-10

L. Bernardi, *Fiscalità pubblica centralizzata e federale: aspetti generali e il caso italiano attuale*

Quaderno n. 00-11

M. Alderighi, R. Maggi, *Adoption and use of new information technology*

Quaderno n. 00-12

F. Rossera, *The use of log-linear models in transport economics: the problem of commuters' choice of mode*