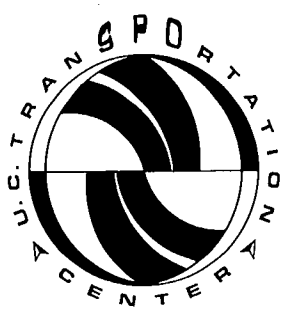
Title
Analysis of Binary Choice Frequencies With Limit Cases

Authors
Goulias, Konstadinos G.
Kitamura, Ryuichi

Publication Date
1992-04-01



Analysis of Binary Choice Frequencies With Limit Cases

Konstadinos G. Goulias
Ryuichi Kitamura

April 1992
Reprint, No. 96

The University of California
Transportation Center

University of California
Berkeley, CA 94720

The University of California
Transportation Center

The University of California Transportation Center (UCTC) is one of ten regional units mandated by Congress and established in Fall 1988 to support research, education, and training in surface transportation. The UC Center serves federal Region IX and is supported by matching grants from the U.S. Department of Transportation, the California State Department of Transportation (Caltrans), and the University.

Based on the Berkeley Campus, UCTC draws upon existing capabilities and resources of the Institutes of Transportation Studies at Berkeley, Davis, and Irvine; the Institute of Urban and Regional Development at Berkeley; the Graduate School of Architecture and Urban Planning at Los Angeles; and several academic departments at the Berkeley, Davis, Irvine, and Los Angeles campuses. Faculty and students on other University of California campuses may participate in Center activities. Researchers at other universities within the region also have opportunities to collaborate on selected studies. Currently faculty at California State University, Long Beach, and at Arizona State University, Tempe, are active participants.

UCTC's educational and research programs are focused on strategic planning for improving metropolitan accessibility, with emphasis on the special conditions in Region IX. Particular attention is directed to strategies for using transportation as an instrument of economic development, while also accommodating to the region's persistent expansion and while maintaining and enhancing the quality of life there.

The Center distributes reports on its research in working papers, monographs, and in reprints of published articles. For a list of publications in print, write to the address below.

University of California
Transportation Center

108 Naval Architecture Building
Berkeley, California 94720
Tel: 415/643-7378
FAX: 415/643-5456

Analysis of Binary Choice Frequencies With Limit Cases: Comparison of Alternative Estimation Methods and Application to Weekly Household Mode Choice

Konstadinos G. Goulias
Department of Civil Engineering
The Pennsylvania Transportation Institute
The Pennsylvania State University

Ryuichi Kitamura
Institute of Transportation Studies
Department of Civil Engineering
University of California at Davis

Reprint, No. 96

To be published in
Transportation Research, Part B: Methodological

The University of California Transportation Center
University of California at Berkeley

ANALYSIS OF BINARY CHOICE FREQUENCIES WITH LIMIT CASES: COMPARISON OF ALTERNATIVE ESTIMATION METHODS AND APPLICATION TO WEEKLY HOUSEHOLD MODE CHOICE

by

Konstadinos G. Goulias*

and

Ryuichi Kitamura**

Abstract

An extensive evaluation of alternative estimation methods for logistic binary choice probabilities when applied to binary frequency data with limit cases is presented in this paper. The methods examined are: binomial-logistic (BL) model, Berkson's (BK) method, and Haldane's (HL) method. These models are applied to weekly household mode choice data that contained a substantial number of limit cases in which one of the alternatives was never chosen. The results obtained indicate that the BL model is a practical tool that outperforms all the other methods examined in this study. The BL models accommodate limit cases without requiring any additional assumptions or approximations. The BK and HL methods have been shown to offer coefficient estimates similar to, and fits that are somewhat worse than, those obtained by the BL models. These methods remain to be useful tools for the analysis of binary frequency data, especially in initial phases of analysis. In this paper it is also shown that coefficient estimates of HL method are sensitive to the value of the adjustment constant, δ, used to incorporate limit cases and its optimal value may depend on the data at hand.

*Department of Civil Engineering
The Pennsylvania State University
University Park, PA 16802
U.S.A.

**Institute of Transportation Studies and
Department of Civil Engineering
University of California, Davis
Davis, California 95616
U.S.A.

September 1991

1. INTRODUCTION

Suppose data containing the outcome of repeated binary choices are available, e.g., the choice of travel mode for commuting (automobile versus public transit) observed over a one-week period. Then suppose a constant probability exists for each individual that governs his repeated binary choices, and this probability is a function of person attributes and other variables which are invariant during the observation period. This probability can then be estimated using the observed relative frequency of choices (see, e.g., Cox, 1970).

One method used for this estimation that has received considerable attention is the minimum logit chi-square estimation. Berkson (1944) proposed it as an operationally simple alternative to the maximum likelihood method. The method assumes that the binary choice probability can be written in the form of a logistic function and uses as the dependent variable the "logit," i.e., the logarithm of the ratio of choice frequencies for the binary alternatives. Estimates obtained using this method have been proven to possess all the desirable properties. Another important advantage of this method is the convenience; a weighted least squares procedure can be used for parameter estimation.

The problem arises, however, when choices made by an individual are all identical, e.g., an individual always chose to travel by automobile and made no public transit trips (such cases shall be termed "limit cases"). This leads to a frequency ratio of either zero or infinity, for which logarithm is undefined. Berkson's method cannot be applied to these cases without modification.

The limit case is not a common phenomenon in many disciplines and thus has tended to be neglected in previous studies. Berkson (1944) presented justifications for excluding limit cases from the sample when the data are generated through an experiment in bio-assay and re-sampling can be performed. On the other hand, limit cases are frequent in the analysis of travel behavior. A good example is travel mode choice where, unless a choice based sample is used, the frequency of public transit trips tends to be small (especially in the United States). When individuals' travel behavior is observed,

experimentation and re-sampling are often infeasible while limit cases can not be legitimately discarded. In fact, they contain as much information on the individual's decisions and choices as do non-limiting cases.

This study is concerned with the estimation of binary choice models using choice frequency data which contain a non-negligible number of limit cases. It focuses on the treatment of limit cases to which limited effort appears to have been directed in the past. Two groups of alternative methods are examined in this study[1]. They all share the same formulation of the binary choice probability using a logistic function.

One group of methods stems directly from the studies in bio-assay by Berkson (1944) and Haldane (1955). In this approach limit-case frequencies are modified by an approximation (Berkson's method, or BK method) or by adding a constant (say, 0.5, as in Haldane's method, or HL method) to obtain estimates of frequency ratios. In this study, a sensitivity analysis is performed to determine the effect of the value of a small constant (δ) added to observed frequencies, on estimated coefficients (β) and their standard errors.

As the second approach of this study, we propose the use of binomial density functions together with binary choice probabilities formulated as a logistic function of explanatory variables (we will term the model the "binomial-logistic model," or BL model). The first approach (BK and HL) is compared with the more exact, maximum likelihood estimation of BL models which does not require any assumption or approximation to incorporate limit cases (the BL model contains as its special case the ordinary binary logit model for unrepeated choice). The BL models are estimated using codes developed by the authors in this study.

The objective of this study is to evaluate the relative usefulness of these alternative estimation methods when applied to binary frequency data with limit cases. The focus of the analysis is on: 1) the practicality of the binomial-logistic (BL) method, and 2) the accuracy of the Berkson (BK), and Haldane (HL) methods relative to that of the BL method. Underlying the study is the uncertain accuracy of BK and HL models in

accounting for limit cases. These models, however, have an advantage in that they can be estimated using commonly available statistical or econometric software packages. As they require no assumption or approximation to accommodate limit cases, BL models are expected to offer the most accurate results. But the effectiveness of commercially available estimation software when applied to BL models is not evident. It is thus desired to determine whether BK or HL methods are adequate and practical substitutes for the maximum likelihood estimation of BL models.

The secondary objective of the study is to determine an optimal value of the added constant (δ) used in the HL method and to evaluate the effect of δ on coefficient and standard error estimates. For reasons discussed in Section 2, it is conceivable that the analysis of mode choice cannot be performed for each individual trip; it may not be uncommon that a modal split model must be formulated for some aggregate of trips, e.g., the set of all trips by an individual or by a household on a day, or work trips produced in a zone (trip-end modal split models fall in this category). BK or HL methods will be effective estimation tools in such instances. The study intends to offer guidelines as to the selection of a constant term to be used in BK and HL methods when limit cases are prevalent in the data.

Following the discussions in the next section on the background of this study, the framework of binary choice analysis is summarized in Section 3 together with a description of the BL model. Following this, the methods by Berkson and Haldane are illustrated in Section 4. Estimation results with the BL models are presented in Section 5. Results with BK and HL methods are presented in Section 6. Section 7 offers a brief summary of this study.

2. USE OF FREQUENCY DATA

Analysis of travel survey data quite often encounters the problem of limited supply-side information. While the interview survey data offer detailed measurements of household

and person characteristics, measurements of urban land development and transportation system characteristics are available only in terms of aggregate zonal averages, or often not available at all. This is almost inevitably the case when a sample is taken from many geographical areas to represent a nation-wide population. Collecting land use information to cover the entire sample area and network data by travel mode for all trip records in the data set, would be too costly, if possible at all. Consequently forecasting models need to be developed using data with limited information, together with whatever supplementary information is available.

This applies to the development of modal choice models using the Dutch National Mobility Panel data set (see Golob, et al, 1986; van Wissen & Meurs, 1989). The data contain information from weekly travel diaries prepared by household members of 12 years and older. Thus an average of approximately 50 trips are available per household in each survey wave. Trip attributes that are typically collected in person trip surveys are all available from the travel diary. Also available are demographic and socioeconomic attributes of the household and its members. Land use and transportation network data for the 20 municipalities from which the panel sample was drawn, are yet to be compiled; the only measures available in this category are a rough indicator of transit service level by municipality, and accessibility measures by mode based on destination choice models developed in an earlier study (Geinzer & Daly, 1981).

Mode choice models that can be developed using such data do not focus on modal competition at the trip level. This, although more desirable, is not possible because information on the attributes of alternative modes is not available. However, because the data set contains weekly travel information, it presents many travel mode choices repeated by the same household members. These repeated choices may be collectively explained by accessibility and other macroscopic measures.

Furthermore, mode choice may be made considering not each individual trip but a series of linked trips to be made by the individual as a whole. Then the attributes of trips

by alternative modes between a given origin and destination pair may not be as influential as might be thought. To the contrary, household car ownership, the number of drivers in the household, overall level of transit development, and other socio-demographic attributes may be the major determinant of weekly household modal split. From this viewpoint, the appropriate measure of mode choice is the relative frequency of trips made by a particular mode rather than the mode chosen for each trip. These considerations motivate the modeling effort reported here.

3. BINOMIAL-LOGISTIC MODEL

Consider the binary choice of one alternative out of two (A or B). For observation i, let $R_i = 1$ if the choice is alternative A and $R_i = 0$ otherwise. Suppose all choices made by i have the same probability of success (i.e., A is chosen) and let $Pr[R_i = 1] = P$, where P is the probability that alternative A is chosen. Then, $E[R_i] = P$ and $Var[R_i] = P(1 - P)$. The probability density of R_i is

$$f_{R_i}(x) = Pr[R_i = x] = \begin{cases} P^x(1-P)^{1-x}, & x = 0,1. \\ 0, & \text{otherwise} \end{cases} \qquad [1]$$

It is often of interest to explain the variation of P from individual to individual. The underlying choice probability, P, may then be parametrized and expressed as a function of the person's attributes (thus denoted as P_i). One possible parametrization of P_i is:

$$P_i = Pr[R_i = 1] = F(X_i'\beta), \quad i = 1, 2, ..., N \qquad [2]$$

where X_i is a vector of explanatory variables for individual i, β is a vector of coefficients, and N is the sample size. X_i is assumed to be exogenous to the process under study.

Expression [2] is a general expression for $E[R_i] = P_i$ (Amemiya, 1985) in which the independent variables can be transformed to reflect a rich number of non-linear relations. For a given individual i, R_i is assumed to be an independently and identically distributed random variable with mean, $E[R_i|X_i] = P_i = F(X_i'\beta)$. The function, F, can be specified to assume any suitable form.

Now consider an individual, i, on whom we observe a vector of attribute variables, X_i, and choices between two alternatives repeated T_i times. The total frequency of choices, T_i, is assumed to be constant. The observations at hand can be treated as repeated T_i Bernoulli trials with probability P_i. Then, the probability that alternative A is chosen k times and alternative B (t - k) times, conditional on that $T_i = t$, is given by the following binomial density[2]:

$$Pr[K_i = k \mid T_i = t] = \binom{t}{k} P_i^k (1-P_i)^{t-k} \qquad [3]$$

where K_i is the frequency with which alternative A is chosen by i. Now suppose a logistic function can be adopted for F:

$$P_i = F(X_i'\beta) = \exp(X_i'\beta)/[1 + \exp(X_i'\beta)] = 1/[1 + \exp(-X_i'\beta)] \qquad [4]$$

Then,

$$Pr[K_i = k|T_i = t] = \binom{t}{k} \left(\frac{1}{1 + \exp(-X_i'\beta)}\right)^k \left(\frac{\exp(-X_i'\beta)}{1 + \exp(-X_i'\beta)}\right)^{t-k} \qquad [5]$$

This model shall be called a binomial-logistic (BL) model of repeated binary choices.

The parameter vector, β, can be estimated by maximizing the log-likelihood function:

$$L = \sum_i \left\{ ln\binom{t_i}{k_i} + k_i \, ln\left(\frac{1}{1+\exp(-X_i'\beta)}\right) + (t_i - k_i)\, ln\left(\frac{\exp(-X_i'\beta)}{1+\exp(-X_i'\beta)}\right)\right\}$$

$$= \sum_i \left\{ \ln\binom{t_i}{k_i} - (t_i - k_i)\beta' X_i + t_i \, ln\left(\frac{1}{1+\exp(-X_i'\beta)}\right)\right\} \qquad [6]$$

where subscript i is added to t and k.

It can be easily shown that the function, $ln(1 + \exp(-Z))$, is convex in Z. It follows that $ln(1 + \exp(-X_i'\beta))$ is convex in β (see Avriel, 1976, Theorem 6.9, p. 154). Therefore the log-likelihood function in eq. [6] is concave everywhere. This allows the use of the Newton-Raphson algorithm which guarantees quick and unique convergence.

From eq. [5] it is clear that the model reduces to the ordinary binary logit model when t = 1. In fact the formulation of a log-likelihood function while treating each of the repeated choices by an individual as a separate observation, will lead to the same expression as eq. [6], except that the first term with the binomial coefficient will be absent. This, however, does not affect the first-order conditions for a maximum. An important benefit of using the BL formulation is computational; the log-likelihood associated with choices repeated by an individual can be evaluated more efficiently by using choice frequencies. The statements here can be immediately extended to multinomial cases.

The first-order conditions for a maximum can be expressed as

$$\partial L/\partial\beta = \sum [t_i/(1 + \exp(-\beta' X_i)) - k_i]X_i = 0. \qquad [7]$$

Therefore if the model contains a constant term, i.e., one element of X_i is set to unity, then, the first-order conditions contain

$$\sum [t_i/(1 + \exp(-\hat{\beta}'X_i)) - k_i] = 0, \qquad [8]$$

or

$$\sum [t_i/(1 + \exp(-\hat{\beta}'X_i))] = \sum t_i \hat{P}_i = \sum k_i \qquad [8']$$

Namely, an average predicted frequency of choice A ($= \sum t_i \hat{P}_i$) equals the observed frequency ($= \sum k_i$).

4. MINIMUM LOGIT CHI-SQUARE ANALYSIS OF BINARY CHOICE

As before, it is assumed that individual i has made t_i repeated binary choices with a constant probability, P_i, of choosing alternative A. Let k_i be the frequency with which alternative A has been chosen, and $(t_i - k_i)$ be that of alternative B. The empirical probability for individual i is

$$\hat{P}_i = k_i/t_i \qquad [9]$$

We retain the assumption that

$$P_i = 1/[1 + \exp(-X_i'\beta)] \qquad [4']$$

The discussions on this section are concerned with the linearization of this expression such that weighted least squares estimators can be used to obtain estimates of parameter vector β.

Multiple Observations and Minimum Logit Chi-Square Estimation[3]

Equation [4'] can be written as the logarithm of the odds-ratio (log-odds):

$$ln[P_i/(1 - P_i)] = X_i'\beta \qquad [10]$$

This equation can be rewritten in terms of the empirical probability (Berkson, 1953; Amemiya, 1985; and Maddala, 1983) as

$$ln[P_i/(1 - P_i)] = X_i'\beta + u_i \qquad [11]$$

The error term, u_i, in [11] can be shown to have $E[u_i] = 0$, and its variance can be approximated as (Berkson, 1953; and Maddala, 1983)[4]

$$Var[u_i] = 1/(t_iP_i(1 - P_i)) \qquad [12]$$

Note that the empirical log-odds, $\ln[P_i/(1 - P_i)]$, can be expressed as $\ln[k_i/m_i]$, where $m_i = t_i - k_i$. This is called the "logit."

Equations [11] and [12] can be used for estimation of the parameter vector, β. Berkson's logit chi-square method (Berkson, 1953) is a generalized least squares approach. First, an estimate of the variance of the heteroskedastic error term is obtained by replacing the theoretical probability in Eq. [12] with the empirical probability in Eq. [9]. Second, weighted least squares is applied to equation [11]. The weight can be defined as the reciprocal of the square root of the estimated error variance (Maddala, 1983):

$$W_i = (t_i\hat{P}_i(1 - \hat{P}_i))^{1/2} = (k_im_i/t_i)^{1/2} \qquad [13]$$

where, as before, $m_i = t_i - k_i$.

Amemiya (1985, pp. 275-278) has proven the consistency and asymptotic normality of these estimators. Both Berkson (1980) and Amemiya (1985) agree that the minimum chi-square estimator is furnishing efficient estimates. However, there appears to exist no general consensus among researchers. For a comparison on the efficiency of the maximum likelihood estimator versus the minimum logit chi-square, see Berkson (1980) and the discussions that follow on the paper by other authors.

Berkson's 2n Rule

One practical problem often occurs in Berkson's procedure, i.e., empirical probabilities may assume values for which the log-odds are undefined. This happens with limit cases, or what Berkson calls "0 or 100 percent observations." This problem was recognized by Berkson in his first exposition of the estimation procedure he devised (Berkson, 1944). For situations where limit cases must be included in the analysis, Berkson proposed several means to circumvent this problem.

One of the methods proposed is called the "2n rule" (Berkson, 1953). For limit cases with a null frequency, i.e. $k_i = 0$, the frequency is replaced with $k_i = 1/2$. Likewise for limit cases with $k_i = t_i$, the frequency is modified as $k_i = t_i - 1/2$. The logit, the dependent variable of the analysis, can then be rewritten as

$$Y_i = \begin{cases} ln[1/(2t_i - 1)], & \text{if } k_i = 0 \\ ln[k_i/m_i], & \text{if } 0 < k_i < t_i \\ ln[2t_i - 1], & \text{if } k_i = t_i \end{cases} \qquad [14]$$

The logit, Y_i, is now defined for all cases and can be used as the dependent variable in the weighted least squares estimation of Eq. [11], with the weight of Eq. [13] redefined using

the 2n rule. Apparently this rule is strictly empirical and is not supported by statistical theory.

Haldane's Method

Gart & Zweifel (1967) present the relative advantages and disadvantages of Berkson's and other similar estimators that account for the presence of limit cases by adding a "small" constant to each frequency to allow the log-odds to assume values that are neither zero nor infinite. The comparison by Gart & Zweifel (1967) includes a method devised by Haldane (1955) in which the value 1/2 is added to all the frequencies. The dependent variable of these methods can be written as

$$Y_i = ln[(k_i + \delta)/(m_i + \delta)] \qquad [15]$$

where δ is a positive constant ($\delta = 1/2$ in Haldane, 1955).

The difference between this method and Berkson's method lies in the treatment of non-limit observations. Haldane suggests modifying all the observations by adding 1/2, whereas Berkson suggests the modification of limit cases only. Haldane's estimator is an unbiased estimator of $ln[P_i/(1 - P_i)]$ up to terms of order superior to (n^{-2}) (Gart & Zweifel, 1967). The same procedure is suggested in BMDP (1985 & 1988) when an empty cell is present in multi-way frequency tables.

The most rigorous comparison of a variety of estimators was performed by Gart, et al (1985). In their study the bias introduced when using Taylor series expansions in the empirical logit formulation is derived, and a variety of estimation methods are compared in terms of this bias. In addition, the variance, skewness and kurtosis of the empirical logit are analyzed both through asymptotic expansions and exact computations. Their results indicate that the correlation between the estimated weights and the empirical transformation (adding a constant) may result in biased estimates. However, as in Gart & Zweifel (1967),

the results do not indicate any estimator to be superior in all cases. As the sample size and the distribution of empirical probabilities change, the performance of one estimator may prove better or worse than the performance of another.

It is of interest to identify the magnitude of change in the parameter estimates (β) as the value of the added "small" constant (δ) changes. In addition it is useful to identify which estimator replicates the observed proportions better. For the HL method, a sensitivity analysis over a range of added constant values is performed to identify any systematic differences in the value of the parameter estimates and predicted choice probabilities. This is the subject of Section 6.

5. ESTIMATION RESULTS: BINOMIAL-LOGISTIC (BL) MODELS

The dependent variable in the empirical analysis of this study is the relative frequency of transit trips among the total motorized trips, or equivalently, the log-odds ratio of the number of transit trips over the number of car trips (including passenger trips), made by each household over a one-week period[5]. The data set used is the Dutch National Mobility Panel Data set. This data set is derived from a large scale panel survey and contains a set of observations on the same households over a period of time (for reviews of the Dutch panel, see Golob et. al. 1986; and van Wissen and Meurs, 1989).

Observations from four panel survey waves (contacts), conducted in the spring of 1984, 1985, 1986, and 1987, are used to estimate the models discussed in this paper. Each data set contains over 1,600 households. A fifth data set was created by pooling these four data sets. These multiple data sets, although not mutually independent, enable more rigorous comparison of the estimation methods. The model presented in this study has been validated using 1988 and 1989 data sets that were not used in model estimation. The results are reported in Kitamura and Goulias (1991). The number of limit cases in the estimation sample is summarized in Appendix Table.

The set of variables included in the models reflects those in previous empirical studies of mode choice. An initial screening of variables was made using various multivariate analysis methods to identify significant factors affecting household modal split in the data set. As discussed in Section 2, the modeling effort here evolves within the limit that no information on the attributes of competing modes is available in the data set.

The binary choice probability is formulated using a logistic function of Eq. [4] with the explanatory variables shown in Table 1. The same formulation is used throughout the study in BL models, and BK and HL methods. The explanatory variables comprise those depicting household demographics, household socioeconomic status, car ownership level and number of drivers, household lifecycle category, and the type of area where the household is located.

Binomial-Logistic (BL) Models

As noted earlier, the BL model handles limit observations most adequately because their probabilities are well defined in the binomial density functions without requiring any approximation. Estimation results are summarized in Table 2. Altogether five models are estimated; four models estimated respectively on 1984, 1985, 1986, and 1987 data sets, and another model on the pooled data set. The convergence was rapid with Netwon-Raphson algorithm in all estimation cases. All models are highly significant with likelihood-ratio chi-squares exceeding 6,800 with 17 degrees of freedom.

The number of diary-keepers in the household, number of cars available, number of drivers and level of public transit availability are the major variables that most significantly influence mode choice (a positive coefficient implies a positive effect of the variable on transit use). In particular, the results indicate that households without a car available (ZEROCAR) and households in a large urban area with a regional transit district (BOV-Large) tend to have higher fractions of public transit trips.

Table 1
The variables used in model formulations

Variable	Definition
Household Demographics	
NRECORDS	Number of respondents in the household
NWOMEN	Number of female members in the household
Household Socio-economics	
NWORKERS	Number of employed persons in the household
INCOME1*	1 if annual household income is less than dfl 17,000
INCOME2	1 if annual household income is between dfl 17,000 and dfl 24,000
INCOME3	1 if annual household income is between dfl 24,000 and dfl 36,000
INCOME4	1 if annual household income is more than dfl 36,000
Household Carownership and drivers license holdings	
ZEROCAR	1 if the household owns no cars
ONECAR	1 if the household owns one car
MULTI-CAR*	1 if the household owns more than one car
NDRIVERS	Number of persons with drivers license in the household
Household Type	
SINGLE	1 if the household is a single person
COUPLE	1 if the household is composed of two adults of different gender
FAMILY	1 if the household is composed of two adults of different gender and there is at least one child
SGLPARENT	1 if the household is composed of an adult and at least one child
OTHER*	1 if the household is not part of any of the above categories
Residence Area Type (City Class)	
BOV-Large	1 if the household resides in a large metropolitan area with highly developed multi-mode transit systems
BOV-Small	1 if the household resides in a small metropolitan area with highly developed multi-mode transit systems
RAIL	1 if the household resides in a small community that is served by rail
NORAIL	1 if the household resides in a small community that is not served by rail

The variables used in the model specification are all indicator (dummy) variables. The results reported here are the same when part or all of the independent variables used are continuous.
* Omitted dummy variable, unless otherwise indicated

Table 2
Binomial-Logistic Models

	Pooled		1984		1985		1986		1987	
Variable	β	t	β	t	β	t	β	t	β	t
Constant	-3.91	-57.4	-4.19	-30.7	-3.59	-26.3	-3.44	-26.3	-4.60	-31.0
NRECORDS	0.46	39.1	0.54	24.1	0.51	20.6	0.38	15.8	0.42	16.7
NWOMEN	0.15	9.6	0.13	3.9	0.06	1.9	0.19	5.9	0.24	7.7
NWORKERS	0.09	7.4	-0.08	-3.2	0.14	5.4	0.21	8.1	0.17	6.4
INCOME2	-0.09	-2.5	0.19	2.9	-0.17	-2.4	-0.15	-2.3	-0.19	-2.6
INCOME3	0.16	5.0	0.22	3.6	0.03	0.5	0.26	4.1	0.15	2.1
INCOME4	0.41	11.4	0.57	8.7	0.31	4.2	0.29	3.8	0.41	5.2
ZEROCAR	3.02	77.2	3.25	39.6	3.11	38.6	2.76	36.4	3.10	39.7
ONECAR	0.70	23.1	0.96	15.1	0.77	12.4	0.44	7.5	0.68	10.9
NDRIVERS	-0.40	-28.6	-0.40	-14.7	-0.37	-13.1	-0.46	-16.2	-0.34	-12.4
SINGLE	-0.08	-1.6	0.03	0.3	-0.48	-4.5	-0.26	-2.5	0.36	3.3
COUPLE	-0.60	-12.9	-0.62	-6.8	-1.00	-10.8	-0.72	-7.4	-0.06	-0.6
FAMILY	-0.74	-16.5	-0.68	-7.9	-1.18	-13.4	-0.76	-8.0	-0.40	-4.2
SGLPARENT	-0.32	-6.0	-0.18	-1.8	-0.77	-7.1	-0.58	-5.2	0.24	2.1
BOV-Large	1.20	51.5	1.26	26.3	1.18	25.8	1.07	21.8	1.29	28.5
BOV-Small	0.30	11.6	0.12	2.4	0.26	4.5	0.38	7.6	0.44	8.9
RAIL	0.41	16.6	0.28	5.6	0.46	8.7	0.48	10.2	0.43	8.8
NORAIL	-0.45	-10.3	-0.41	-4.7	-0.29	-3.5	-0.50	-5.7	-0.58	-6.4
L(C)	-37209		-9227		-8867		-9317		-9794	
L(β)	-22878		-5647		-5454		-5870		-5702	
χ^2	28661		7159		6827		6895		8185	
N	6787		1652		1611		1726		1798	
P										
Observed	0.165		0.168		0.162		0.167		0.163	
Predicted	0.164		0.168		0.158		0.167		0.162	
% Error	-0.7%		0.1%		-2.7%		0.3%		-0.9%	
R^2	0.666		0.671		0.674		0.653		0.675	
MAE	0.139		0.139		0.134		0.144		0.135	
MSE	0.046		0.046		0.044		0.049		0.045	
NT										
Observed	3.05		3.16		3.03		3.02		3.00	
Predicted	3.05		3.16		3.03		3.02		3.00	
% Error	0.0%		0.0%		0.0%		0.0%		0.0%	
R^2	0.661		0.666		0.671		0.654		0.687	
MAE	2.70		2.76		2.71		2.69		2.55	
MSE	19.13		20.32		19.88		18.55		16.41	

Stationarity $\chi^2 = 809$(df = 54)

L(C) = Value of Log-likelihood function with constant only
L(β) = Value of Log-likelihood function at convergence
P = Proportion of transit trips
NT = Number of transit trips
MAE = Mean absolute error, average of the absolute difference between observed and estimated value
MSE = Mean square error, average of the squared difference between observed and estimated value

The models' replication capability is excellent, with correlation coefficients between observed and predicted relative frequencies (or, probabilities) of transit trips over 0.65 for all cases. Relative errors in mean choice probability are within 1% for most cases (the 1985 model is the only exception with a relative error of -2.7%).

One advantage of the BL model with a constant term is its ability to replicate observed frequency of choices exactly, in this case the number of transit trips. Correlation coefficients between observed and predicted numbers of transit trips are again high over 0.65 in all cases. The model coefficients are reasonably stable across the four periods, although a log-likelihood ratio chi-square of 809 implies that the hypothesis of stability must be rejected.

7. ESTIMATION RESULTS: BERKSON'S (BK) AND HALDANE'S (HL) METHODS

One important advantage of the BK and HL methods is the computational convenience offered by the fact that both methods are based on weighted least squares procedures[6]. Whether these methods lead to adequate models is the main focus of this section. Also of concern is the selection of the value of the added constant, δ, in the HL method. The constant may affect the estimates β in two ways. The first is directly, through the value of the dependent variable (see Eq. [15]), while the second is indirectly through the value of the weights computed to perform weighted least squares. The sensitivity of coefficient estimates and the model's goodness-of-fit to the value of δ is examined in this section.

When the value of δ is small, so are the weights applied to limit cases because, with δ added to both k_i and m_i, Eq. [13] becomes

$$\begin{aligned} W_i &= [(k_i + \delta)(m_i + \delta)/(k_i + m_i + 2\delta)]^{1/2} \\ &= [\delta(t_i + \delta)/(t_i + 2\delta)]^{1/2} \\ &\sim \delta^{1/2}, \qquad \text{if } k_i = 0 \text{ or } m_i = 0. \end{aligned} \qquad [19]$$

Therefore the contribution of limit cases diminishes as the value of δ approaches 0. When δ is large, on the other hand, large errors arise in the value of Y_i for limit cases, especially when t_i is small. It is then anticipated that there exists a δ that maximizes the model's fit by balancing these two sources of error.

The estimation results obtained using the BK method and the HL method with five values of δ (0.01, 0.05, 0.1, 0.5 and 1.0) are summarized in Table 3 for the pooled data. The estimation results confirm previous comparisons (Gart & Zweifel, 1967; and Gart, et al., 1985) where the BK and HL methods produced very similar results. The coefficient estimates are similar to those obtained by the BL model, especially by HL method with δ = 0.05 or 0.1, according to the square error measures presented in Table 3.

The coefficient estimates vary greatly depending on the value of δ. In general, their absolute values are smallest when δ = 1.0, while they tend to take on the largest values with δ = 0.05 or 0.1. The extent of variation is different from variable to variable. For example, the coefficients of NORAIL range from -0.23 to -0.47, those of RAIL from 0.26 to 0.45, NWORKERS from 0.05 to 0.08, and INCOME4 from 0.19 to 0.39. For the five most significant variables, i.e., NRECORDS, ZEROCAR, ONECAR, NDRIVERS, and BOV-Large, the ratios of the largest to smallest coefficients are 1.49, 1.24, 1.36, 1.13, and 1.28, respectively. The influence of δ on coefficient estimates is evident.

The goodness of fit of the HL model, measured in terms of percent error or mean absolute error, are best with δ = 0.1, for both the probability of a transit trip and the number of transit trips. The mean square error is minimal when δ = 0.5, while the correlation coefficient between predicted choice probability and observed relative choice frequency is maximum at δ = 1.0 for the probability and δ = 0.5 for the number of transit trips. From the viewpoint that predicting the average number of transit trips or their relative frequency is the most frequent application of such models, an optimum choice for δ appears to lie between 0.1 and 0.5. The results of this analysis thus lend support to the

Table 3
Models obtained with Haldane's and Berkson's Methods

			Haldane's Method									
	Berkson		δ=0.01		δ=0.05		δ=0.1		δ=0.5		δ=1.0	
Variable	β	t	β	t	β	t	β	t	β	t	β	t
Constant	-3.54	-26.6	-3.54	-18.9	-3.99	-21.8	-4.04	-23.3	-3.52	-26.4	-3.03	-27.2
NRECORDS	0.43	15.8	0.44	12.0	0.52	14.4	0.52	15.2	0.43	15.7	0.35	14.9
NWOMEN	0.14	3.7	0.16	3.3	0.17	3.6	0.17	3.7	0.14	3.7	0.11	3.5
NWORKERS	0.06	2.3	0.07	1.8	0.08	2.3	0.08	2.5	0.07	2.4	0.05	2.1
INCOME2	-0.11	-1.7	-0.15	-1.6	-0.18	-2.0	-0.17	-2.0	-0.11	-1.7	-0.08	-1.4
INCOME3	0.06	0.9	0.11	1.3	0.10	1.1	0.09	1.1	0.06	0.9	0.04	0.8
INCOME4	0.26	3.6	0.38	3.8	0.39	4.0	0.37	4.0	0.26	3.6	0.19	3.2
ZEROCAR	2.78	35.0	2.73	24.9	2.99	27.6	3.02	29.3	2.73	34.2	2.44	36.5
ONECAR	0.65	10.9	0.73	8.7	0.76	9.3	0.75	9.7	0.64	10.7	0.56	11.2
NDRIVERS	-0.34	-11.3	-0.32	-8.2	-0.35	-8.9	-0.35	-9.4	-0.33	-11.1	-0.31	-12.1
SINGLE	-0.09	-0.8	-0.09	-0.6	-0.10	-0.7	-0.10	-0.7	-0.09	-0.9	-0.09	-1.0
COUPLE	-0.57	-6.0	-0.51	-4.0	-0.58	-4.6	-0.60	-4.9	-0.56	-5.8	-0.50	-6.2
FAMILY	-0.70	-7.4	-0.66	-5.3	-0.75	-6.0	-0.76	-6.4	-0.69	-7.2	-0.61	-7.5
SGLPARENT	-0.25	-2.3	-0.20	-1.4	-0.24	-1.7	-0.26	-1.9	-0.26	-2.3	-0.23	-2.5
BOV-Large	1.09	20.6	1.00	14.5	1.18	17.2	1.22	18.5	1.10	20.7	0.95	21.1
BOV-Small	0.33	6.1	0.32	4.4	0.41	5.7	0.42	6.1	0.35	6.6	0.29	6.5
RAIL	0.33	6.2	0.40	5.4	0.45	6.2	0.44	6.4	0.33	6.3	0.26	5.9
NORAIL	-0.32	-4.5	-0.47	-4.3	-0.46	-4.4	-0.43	-4.4	-0.30	-4.2	-0.23	-4.0
S.E.	1.51		1.70		1.77		1.74		1.54		1.40	
F*	936.4		342.3		462.8		554.0		897.7		1104.3	
P												
Observed	0.165		0.165		0.165		0.165		0.165		0.165	
Predicted	0.175		0.190		0.174		0.170		0.177		0.189	
% Error	6.3%		14.9%		5.2%		3.1%		7.0%		14.4%	
R^2	0.665		0.662		0.661		0.662		0.665		0.666	
MAE	0.144		0.151		0.143		0.142		0.145		0.153	
MSE	0.046		0.047		0.047		0.047		0.046		0.047	
NT												
Observed	3.051		3.051		3.051		3.051		3.051		3.051	
Predicted	3.353		3.845		3.419		3.314		3.416		3.720	
% Error	9.9%		26.0%		12.1%		8.6%		12.0%		21.9%	
R^2	0.658		0.647		0.650		0.651		0.656		0.653	
MAE	2.837		3.058		2.836		2.791		2.868		3.059	
MSE	19.372		20.748		20.183		19.976		19.499		19.898	
$\Sigma (\beta_j^{BL} - \beta_j)^2$	0.281		0.302		0.050		0.059		0.328		1.384	
$\Sigma (\beta_j^{BL} - \beta_j)^2/\beta_j^{BL}$	0.294		0.253		0.215		0.216		0.322		0.892	

*df = (18, 6769), N = 6787

S.E. = Standard error of the white noise
P = Proportion of transit trips
NT = Number of transit trips
MAE = Mean absolute error, average of the absolute difference between observed and estimated value
MSE = Mean square error, average of the squared difference between observed and estimated value
Pred(1) = Obtained by "naive aggregation"

conventional value of 0.5; however, the results also indicate that a value smaller than 0.5 may offer better replication of sample averages. The BK model shows similar performance in replication as the HL models with $\delta = 0.1$ or 0.5. Note that all models over-predict the probability of transit trips and the number of transit trips.

In summary, this exercise has also shown that small values of δ are not advisable. Contrary to the expectation that a smaller δ would lead to more accurate estimates of logits and therefore better coefficient estimates, models estimated with small δ have exhibited poorer predictive performances, presumably due to extremely small weights computed for limit cases. An optimal value of δ appears to lie between 0.1 and 0.5. The BK method performed nearly as well as the HL method with δ in this range. It has also been shown that HL estimates are close to BL estimates with these δ values. None of these models replicate observation better than the BL model. The most important finding of this analysis is that the best value of δ appears to be smaller than the conventionally used value of 0.5, but not by very much. The results have also shown that coefficient and standard error estimates are sensitive to the value of δ. It is desirable, however, that the generality of these results be determined in future analysis.

7. CONCLUSION

The objective of this study has been to evaluate and compare alternative estimation methods for logistic binary choice probabilities when applied to binary frequency data with limit cases. The methods examined are: binomial-logistic (BL) model, Berkson's (BK) method, and Haldane's (HL) method. These models were applied to weekly household mode choice data that contained a substantial number of limit cases in which one of the alternatives was never chosen.

It has been shown that the BL model is a practical tool whose performance surpasses those of the other methods examined in the study. BL models accommodate limit cases without requiring any additional assumptions or approximations. The use of

binomial probability mass functions is ideally suited for binary frequency data. With the everywhere concave log-likelihood functions of BL models, the Newton-Raphson algorithm converged very quickly at the maximum likelihood. The resulting models best replicated observed choice probabilities and number of trips by mode.

The BK and HL methods have been shown to offer coefficient estimates similar to, and fits that are somewhat worse than, those obtained by the BL models. An advantage of these methods is that they can be applied with weighted least squares procedures for which software packages are ubiquitously available. They remain to be useful tools for the analysis of binary frequency data, especially in initial phases of analysis. A summary on the comparison of the two groups of models is reported in Table 4.

The study has also shown that coefficient estimates of the HL method are sensitive to the value of the adjustment constant, δ, used to incorporate limit cases. An important result is that the HL method performs best between 0.1 and 0.5. It also produces coefficient estimates that are closest to those produced by the BL model when δ is in this range. The BK method performs slightly worse than the best HL models. The results support the BK method and the convention of using $\delta = 0.5$ in HL method as a reasonable substitute for the BL model. Estimating a BL model, however, is not at all a computationally onerous task[7]. This study has extended the scope of binary frequency data analysis by showing the accuracy and practicality of the BL method.

Acknowledgements--The authors wish to acknowledge the support given by the U.S. Department of Transportation through the Region Nine Transportation Center at University of California, Berkeley, and the Projectbureau for Integrated Traffic and Transportation Studies, The Netherlands. The data set used in this study was obtained from the Dutch National Mobility Panel survey, sponsored by the Projectbureau for Integrated Traffic and Transportation Studies, and the Directorate General of Transport of the Netherlands Ministry of Transport and Public Works. Two anonymous referees offered extremely useful comments on an earlier version of the paper.

Table 4
Comparison of the BL with the BK-HL group of models

Method	Advantages	Disadvantages
BK-HL (Estimated via Minimum Chi-Square)	• Easy to Estimate • Requires No Programming • Relatively Accurate	• Approximation Required for the Limit Cases
BL (Estimated via Maximum Likelihood)	• No Approximation for Limit Cases Required • Very Accurate • Low Computational Cost • Extendable to Multinomial Choices	• Requires Programming

REFERENCES

Amemiya, T. (1985) *Advanced Econometrics*. Harvard University Press, Cambridge, MA.

Avriel, M. (1976) *Nonlinear Programming: Analysis and Methods*. Prentice-Hall, Englewood Cliffs, NJ.

Berkson, J. (1944) Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, **39**, 357-365.

Berkson, J. (1953) A statistically precise and relatively simple method of estimating the bio-assay with quantal response, based on the logistic function. *Journal of the American Statistical Association*, **48**, 565-599.

Berkson, J. (1980) Minimum chi-square, not maximum likelihood! *Annals of Statistics*, **8**, 457-487.

BMDP (1985 & 1988) *BMDP Statistical Software Manual*. University of California Press, Berkeley, CA.

Cox, D.R. (1970) *Analysis of Binary Data*. Methuen, London.

Gart, J.J. and Zweifel J.R. (1967) On the bias of various estimators of the logit and its variance with application to quantal bio-assay. *Biometrika*, **54**, 181-187.

Gart, J.J., Pettigrew H.M. and Thomas D.G.(1985) The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika*, **72**,179-190.

Geinzer, J. and Daly A (1981) Zuidvleugel Study - Report 9, Models of Car Ownership and License Holding. Cambridge Systematics Europe, The Hague, The Netherlands.

Golob, J.M.,Schreurs L.J.M. and Smit J.G. (1986) The design and policy applications of a panel for studying changes in mobility over time. In *Behavioural Research for Transport Policy*, VNU Press, Utrecht, The Netherlands, pp. 81-95.

Goulias, K.G. and Kitamura R. (1991) Analysis of Binary Choice Frequencies with Limit Cases. Research Report No. UCD-ITS-RR-91-01. Institute of Transportation Studies, University of California, Davis.

Haldane, J. B. S. (1955) The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics*, **20**, 309-311.

Kitamura, R. (1988) A Dynamic model system of household car ownership, trip generation, and modal split: Model development and simulation experiments. In *Proceedings of the 14th Australian Road Research Board Conference, Part 3,* Australian Road Research Board, Vermont South, Victoria, Australia, 1988, pp. 96-111.

Kitamura, R. and Goulias K.G.(1991) MIDAS: A Travel Demand Forecasting Tool Based on a Dynamic Model System of Household Demographics and Mobility. Final report for Projectbureau Integrale Verkeer- en Vervoerstudies, Ministerie van Verkeer en Waterstaat, The Netherlands.

Maddala, G. S. (1983) *Limited Dependent and Qualitative Variables in Econometrics.* Cambridge University Press, Cambridge.MA.

van Wissen, L.J.G. and Meurs H.J. (1989) The Dutch mobility panel: Experiences and evaluation. *Transportation*, **16**, 99-119.

Appendix Table
Presence of Limit Cases in the Sample used

	No Transit Trips	No Car Trips	Non-Limit Cases	Total
Wave 1 (1984)	952	89	611	1652
Wave 3 (1985)	939	75	597	1611
Wave 5 (1986)	996	98	632	1726
Wave 7 (1987)	1048	86	664	1798
Pooled	3935	348	2504	6787

FOOTNOTES

[1] A third group which is based on a heuristic formulation and the Tobit model can be found in Goulias and Kitamura (1991).

[2] In this formulation the effect of possible unobserved factors is neglected, assuming that the vector X_i is sufficient for determining the choice. The task of including possible individual specific effects is left as future extension of the study here. However, an alternative interpretation, offered by one of the referees, would be that on average for a set of individuals t with the same values of vector attributes (X_i), alternative A is chosen k times and alternative B (t-k) times, while the inclusion of unobserved factors may yield choices of A or B exclusively. This may also be the source of limit cases which remains to be proven in future extensions.

[3]Maddala (1983) classifies this method as minimum chi-square method applied to multiple observations. In the analysis of contingency tables literature the same method is referred to as minimum chi-square method applied to grouped data.

[4]Maddala (1983) offers a simplified proof of the formula of the variance. Amemiya (1985) reports the same results, however, his proofs are more rigorous and more general.

[5]The weekly modal split model is a component of a long-range forecasting model system for car ownership and utilization (Kitamura & Goulias, 1991). The model structure presented here follows an earlier specification proposed by Kitamura (1988). That specification has been enriched by the inclusion of a set of variables representing the household type. Kitamura (1988) found that the use of Haldane's estimator led to over - prediction of transit use.

[6]Note that if the total frequency is not observed, then we cannot develop weights.

[7]The binomial-logistic estimation code used in the analysis is available from the authors.