CANADIAN TRANSPORTATION RESEARCH FORUM
LE GROUPE DE RECHERCHES SUR LES TRANSPORTS AU CANADA

21st ANNUAL MEETING

PROCEEDINGS

VANCOUVER, B.C.
MAY 1986

EMPIRICAL BAYES ESTIMATES OF DEMAND FOR TRANSPORTATION MODE: A PROPOSED METHODOLOGY AND IMPLICATIONS FOR NETWORK SCHEDULING

by

T. J. Tomberlin, K. Buyukkurt, and M. Buyukkurt

Concordia University

1.0 Introduction

Theoretically, routing and scheduling of urban area transportation networks should be based on consumer preferences as well as a consideration of operating costs and availability of resources. Although there has been considerable research on transportation mode choice, attempts to relate the estimated probabilistic choice models to network scheduling issues have been scarce. (Daganzo 1979, Manski and McFadden 1981, McFadden 1976). Estimates for different transportation modes in different subareas of an urban community can be compared with the current scheduling of public transportation networks to identify the subareas where consumer preferences are poorly matched. Such information in conjunction with the data regarding operating costs and available resources can be used to the improve quality and scheduling of existing public transportation services. In order to make these subarea comparisons, a methodology for predicting consumer preferences for a large number of subareas needs to be developed.

The various choice models suggested in the transportation literature have dealt with estimation of parameters and prediction of probability of choice for different modes of transportation at the aggregate level. These models include logit, probit multinomial logit, multinomial probit and dogit, among others (Currim 1982, Daganzo 1979, Gaundry and Dagenais 1979, Manski and McFadden 1981, and McFadden 1976). Analysts who have recognised the importance of segments, for example subareas in a geographically deliniated market or "benefit" segments as defined in marketing, have partitioned the total sample into subgroups, and then have carried out estimation and prediction at the subgroup level (Currim 1981, Hauser and Urban 1977, Westin and Watson 1975).

This method of prediction at the subarea level can be improved by considering techniques which have recently become available in the area of multiple parameter estimation. (Rubin 1981 and Morris 1983). In the words of Morris (1983), these methods "borrow strength from the ensemble." That is, prediction in a subarea is improved by borrowing strength from data set as a whole.

Within this framework, we explore the application of empirical Bayes techniques via logit and multinomial logit models to the task of improving multiple predictions. In section 2, we consider a set of potential models which could be employed for this purpose. In section 3, we consider the problem of obtaining data for purposes of prediction, and map out directions for further research.

2.0 Models for Prediction

We propose to estimate consumer preference among a set of transportation mode alternatives, by subarea, using a combination of special purpose sample survey data and publicly available census data. Typically, the sample design for such a special purpose survey would be multi-stage, with a nested structure of primary sampling units (PSU's), secondary sampling units (SSU's) within PSU's, and finally, individuals within SSU's. Obtaining global estimates of choice probabilities using such samples is a problem which is well understood by survey practicioners. See, for example, Cochran (1963). However, only recently have researchers begun to consider ways of using thinly spread sample survey data for the purpose of drawing inference about small subgroups of the population.

Currently available estimation techniques for small domains have been reviewed by Purcell and Kish (1979). Although often applied to estimating counts such as unemployment or mortality statistics, most of the available techniques were designed primarily for continuous variables. Examples include the regression models of Ericksen (1974), the synthetic estimation techniques of Gonzalez and Hoza (1975), the prediction models of Holt, Smith and Tomberlin (1979) and Laake (1979), and the James-Stein methods of Fay and Herriot (1979).

Here, we adopt a model-based prediction approach similar to that used by Holt, Smith and Tomberlin (1979) and Laake (1979), but we concentrate entirely on models built for discrete response variables. By including random effects corresponding to small geographic subareas of the population, we propose to obtain empirical Bayes estimates. These methods, like those of Fay and Harriot (1979), are sometimes referred to as "shrinkage estimates".

2.1 Fixed effects logistic models

In order to facilitate an introduction to the proposed methods, let us limit consideration to the case where there are two transportation mode choices, for example, bus vs. subway. Let the symbol p with appropriate subscripts represent the probability of an individual choosing to take the subway, and q = 1 - p the complementary probability of preferring the bus. Let the vector **X** represent a set of characteristics associated with individuals, such as age, sex, income and distance from home to nearest subway station. Let the symbol v denote i,j,k, where i represents PSU, j represents SSU within PSU, and k represents an individual within SSU. A typical logistic model might assume the following form:

(1) $\text{logit}(p_v) = \boldsymbol{\beta}\, \mathbf{X}_v$,

where, the logit function is defined by

(2) $\text{logit}(p) = \ln(p/q)$.

For a more detailed discussion of logistic models, see Haberman (1978). In model (1), the logit is allowed to vary depending on individual characteristics, but note that there is no local area effect. Further, the absence of the subscript v on the vector of regression coefficients, **β**, indicates that the relationship between the logits and the individual characteristics is the same for the whole population. For example, this would imply that the effect of distance from the nearest subway station would have the same effect on the probability of choosing the subway over the bus for residents of the downtown core as it would for those living in outlying suburban areas. Model (1) allows for only that variation in transportation mode preference from area to area which can be explained by variation in the makeup of the individual characteristics $\mathbf{X}_v$.

By introducing a local area effect, the model could be improved to allow for variation beyond that due to differences in the $\mathbf{X}_v$. One such model would be

(3) $\text{logit}(p_v) = \boldsymbol{\beta}\, \mathbf{X}_v + \gamma_i$

Here, the geographic parameter, γ_i depends only on the ith PSU. The additive form means that there is no interaction is between the effect of the vector of individual characteristics, $\mathbf{X}_v$, and the PSU effect γ_i.

A more realistic model might permit the effect of any particular factor, such as the distance from the nearest subway station, to vary from PSU to PSU. A general framework for such models would be obtained by allowing the vector of regression coefficients **β**, to vary among PSU's, as with the following model:

(4) $\mathrm{logit}(p_\nu) = \beta_i X_\nu + \gamma_i$

Models (3) and (4) have the advantage of increasing realism, but the disadvantage that only data within the local area which constitutes the ith PSU can be used to estimate the local area effect, γ_i in either model, and the vector of coefficients β_i in model (4). Furthermore, for either of models (3) or (4), the possibility of estimation for local areas corresponding to PSU's not included in the sample is not immediately obvious. A major reason for introducing random effects into these models, as we suggest in Section 2.2 below, is to allow for a compromise between models such as model (1) and models (3) and (4).

Once a model such as (1), (3), or (4) has been fit to the data, there remains the task of applying this model to obtain estimates of numbers of consumers preferring the various modes of transportation for each of the subareas. This can be accomplished in the following manner. Information regarding the vector X_ν must be available in an aggregate form for each of the subareas under consideration. In particular, if $\bar{X}_i$ is the mean of the X_ν vectors for the ith subarea, then the proportion of consumers preferring the subway over the bus for subarea i would be obtained by simply applying the estimated model to this mean vector. For example, if model (1) were being used, the logit of the predicted proportion would be given by

(5) $\mathrm{logit}(\hat{p}_i) = \hat{\beta}\,\bar{X}_i$.

The predicted proportion itself would be obtained by inverting the logit transformation as follows,

(6) $\hat{p}_i = \exp(\hat{\beta}\,\bar{X}_i) / [1 + \exp(\hat{\beta}\,\bar{X}_i)]$

Finally, the total number of consumers preferring the subway, m_i, could be obtained by merely applying this estimated proportion to an estimate of the population of the subarea, n_i ,

(7) $\hat{m}_i = n_i\,\hat{p}_i$

Notice that this requires up-to-date, detailed information regarding the individual characteristics for small areas. For urban areas, information on demographic characteristics is generally available down to the block level from the preceeding census. In Canada, the census is taken every five years, so that reasonably up-to-date, detailed information is readily available. Any other information which can be obtained in aggregate form for relevant geographic subdivisions, can be included in the model. For example, average distance from the nearest subway station could be easily obtained by inspection of maps.

2.2 Random effects logistic models

As mentioned earlier, models (3) and (4) suffer a serious drawback. Any such model which attempts to allow for variation among subareas more than that which could be ascribed to variation in the individual characteristic variables X_ν, requires that sufficient sample survey data be available from each of the subareas under consideration. Estimation of the local area specific parameters associated with such fixed effects models is based on data obtained from the corresponding local areas.

In addition, standard estimation techniques for these models are based on the assumption that individual choices can be considered as independent binomial observations. This assumption fails to incorporate the intraclasss correlation among the p_ν from the same SSU or PSU. Even though this independence assumption is generally false, as long as the sample size is sufficiently large, the estimates themselves will be consistent. However, estimates of the sampling variability associated with these estimates will be overly optimistic.

One way of addressing both of these problems in a way which allows data from subareas to be used to an extent which is appropriate and which yields a valid assessment of the reliability of estimates obtained, is to incorporate a series of nested random effects into these logistic models with components of variability corresponding to each stage of the multi-stage sample design. Such a model might look like,

$$(8) \qquad \text{logit}(p_\nu) = \beta X_\nu + \gamma_i + \gamma_{j(i)},$$

where the γ_i are regarded as being drawn from a Normal $(0,\sigma_1^2)$ population and the $\gamma_{j(i)}$ from a Normal $(0,\sigma_2^2)$ population. These random effects imply that individuals from the same PSU have a common element entering into their p_ν, and the same occurs for the nested class of individuals in a common SSU.

Estimates of consumer preference based on random effects models such as (8) are obtained using an empirical Bayes framework. Once values of σ_1^2 and σ_2^2 have been obtained, it is possible to produce approximate normal posterior distributions for the p_ν. These distributions contain factors which automatically, and correctly weight sample consumer preferences observed at the various levels of the multi-stage design. Furthermore, a posterior distribution is obtained for the logit p_ν's corresponding to individuals in PSU's not included in the sample. For such individuals, this posterior depends entirely on the model-based prediction obtained as a function of the individual characteristic variables in X_ν.

The basic mathematical development facilitating approximate computation of the posterior means and variances appears in Laird (1975, 1978). Initial experience with variance estimation for a non-nested model is found in Miao (1977) in a study of regional variation in leukemia incidence rates. In a study predicting automobile accident rates, Tomberlin (1982, 1986) used similar nested random effects models for Poisson data. In that study, estimates of variance components were obtained using maximum likelihood methods, and subsequent empirical Bayes predictions of individual accident rates were demonstrated to be superior to the more commonly used fixed effects model alternatives. The methodology for analysing these nested random effects logistic models has now been firmly established.

2.3 Multinomial logistic models

As noted earlier, there are usually more than two transportation modes available to consumers. For example, in addition to public bus and subway, one might choose to travel by commercial taxi, or private automobile, or by any combination of these alternatives. In order to obtain estimates for a range of choices, we must consider expanding the logistic models described in (1), (3), (4), (5), and (8) to models which are appropriate for multinomial response variables.

One formulation for multinomial logit models is presented by Amemiya (1981). In this formulation, one category, for example the subway, is chosen as a base category. The log odds ratio of favoring other categories as opposed to the base category are then modeled as linear functions of the vector of individual characteristics, for an analog of a multivariate multiple regression model. For example, let $p_{m\nu}$ be the probability of individual ν opting for the subway. Then if $p_{k\nu}$ is the corresponding probability for some other mode of transportation, we have,

$$(9) \qquad \log(p_{k\nu}/p_{m\nu}) = \text{mlogit}(p_{k\nu}) = \beta_k X_\nu .$$

Just as with the ordinary logit model (1), the maximum likelihood solution for the parameters β_k requires a non-linear optimization algorithm. Computer programs for estimation of the parameters in (9) are available in standard statistical software packages such as SAS, though the convergence properties of these procedures can be somewhat more problematical.

Analogous to the nested random effects models described in (8), we have the following multivariate model:

(10) $\text{mlogit}(p_{k\nu}) = \beta_k X_\nu + \gamma_{ki} + \gamma_{kj(i)}, k \neq m,$

where the vector of γ_{ki}, $\boldsymbol{\gamma}_i$ is regarded as drawn from a multivariate Normal $(\mathbf{0},\Sigma_1)$ population, and the vectors of $\gamma_{kj(i)}$, $\boldsymbol{\gamma}_{j(i)}$ are considered as being drawn independently from a multivariate Normal $(\mathbf{0},\Sigma_2)$ population.

Without further research, it remains unclear just how accurately the covariance matrices Σ_1 and Σ_2 can be estimated from a typical special purpose consumer survey, nor is it clear how sensitive the final empirical Bayes estimates of disaggregate consumer preference would be on these estimated covariance components. However, such models do capture the levels of variation which a priori judgement strongly suggest must underlie such survey data.

3.0 Implications for Further Research

Several models have been suggested for predicting probabilities of transportation mode choice for subareas of a larger market. These models are based on empirical Bayes techniques, and are expected to improve prediction accuracy by using information available from other subareas. Just how much improvement and under what conditions such an improvement is possible can be studied by empirical comparisons (Carter and Rolph 1974) or by Monte Carlo experimentation.

Improved models for choice probabilities and census data associated with subareas can be used together to estimate demand for various modes of transportation from each subarea to specified destinations, for example, the central business district. Such estimates for all subareas would present the geographic distribution of demand. An overall comparison of subarea demand levels with currently allocated public transportation alternatives may be helpful in signaling scheduling problems. It should be noted, however, that a proper analysis of scheduling and networking issues should consider frequency and timing of trips to multiple destinations. The suggested models, in their current forms, do not address such crucial issues.

Further research in random effects models of choice probabilities seems warranted. First, empirical Bayes parameter estimation methods should be developed for each of the models discussed. Next, an empirical study based on data collected via a multi-stage sample design should be conducted to estimate the parameters of the suggested models. Finally, using disaggregate census data, these models can be employed to estimate demand for the various transportation modes. Creative ways of incorporating information on the frequency and timing of trips to specific destinations is likely to lead to useful application of these models for transportation network design and scheduling.

References

Amemiya, Takeshi (1981), "Qualitative Response Models: A Survey," The Journal of Economic Literature, 19, 1483-1536.

Carter, Grace M. and Rolph, John E. (1974), "Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities," Journal of the American Statistical Association, 69, 880-885.

Cochran, W. G. (1963), Sampling Techniques, New York: Wiley.

Currim, I. S. (1981), "Using Segmentation Approaches for Better Prediction and Understanding from Consumer Mode Choice Models," Journal of Marketing Research, 18, 301-309.

Currim, I. S. (1982), "Predictive Testing of Consumer Choice Models Not Subject to Independence of Irrelevant Alternatives," Journal of Marketing Research, 19, 208-222.

Daganzo, C. (1979), "Multinomial Probit: The Theory and Its Applications to Demand Forecasting," New York: Academic Press.

Ericksen, E. P. (1974), "A Regression Method for Estimating Population Change in Local Areas," Journal of the American Statistical Association, 69, 867-875.

Fay, R. E., and Herriot, R. A. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data," Journal of the American Statistical Association, 74, 269-277.

Gaudry, M. J. I. and Dagenais, M. G. (1979), "The Dogit Model," Transportation Research, 13B, 105-111.

Gonzalez, M. and Hoza, C. (1975), "Small Area Estimation of Unemployment," Proceedings of the American Statistical Association, Social Statistics Section, 437-443.

Haberman, S. J. (1978), Analysis of Qualitative Data, New York: Academic Press.

Hauser, J. and Urban, G. L. (1977), "A Normative Methodology for Modelling Consumer Response to Innovation," Operations Research, 25, 579-619.

Holt, D., Smith, T. M. F., and Tomberlin, T. J. (1979), "A Model-Based Approach to Estimation for Small Subgroups of a Population," Journal of the American Statistical Association, 74, 405-510.

Laake, P. (1979), "A Predictive Approach to Subdomain Estimation in Finite Populations," Journal of the American Statistical Association, 74, 355-358.

Laird, N. (1975), "Log-Linear Models with Random Parameters: An Empirical Bayes Approach," unpublished Ph.D. dissertation, Harvard University, Department of Statistics.

Laird, N. (1978), "Empirical Bayes for Two-Way Tables," Biometrika, 65, 805-811.

Manski, C. and McFadden, D. (1981), Structural Analysis of Discrete Data, Cambridge, MA: MIT Press.

McFadden, D. (1976), "Quantal Choice Analysis: A Survey,: Annals of Economic and Social Measurement, 5, 363-390.

Miao, L. L. (1977), "An Empirical Bayes Approach to Analysis of Inter-Area Variation," unpublished Ph.D. disseration, Harvard University, Department of Statistics.

Morris, C. N. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," Journal of the American Statistical Association, 78, 47-55, with discussion.

Purcell, N. J. and Kish L. (1979), "Estimation for Small Domains," Biometrics, 35, 365-384.

Rubin, D. (1981), "Using Empirical Bayes Techniques in the Law School Validity Studies," Journal of the American Statistical Association, 75, 801-827, with discussion.

Tomberlin, T. J. (1982), "A Statistical Perspective on Predicting Losses in Automobile Insurance," unpublished Ph.D. dissertation, Harvard University, Department of Statistics.

Tomberlin, T. J. (1986), "Predicting Accident Frequencies for Drivers in a Two-Way Classification", submitted for publication.

Westin, R. and Watson, P. (1975), "Reported and Revealed Preferences as Determinants of Mode Choice Behavior," Journal of Marketing Research, 12, 282-289.