Monograph:
Murchland, J.D. (1980) What are the errors we should be seeking to minimise in the estimation process? Working Paper. Institute of Transport Studies, University of Leeds , Leeds, UK.

Working Paper 126

Published paper

Murchland, J.D. (1980) *What are the errors we should be seeking to minimise in the estimation process?.* Institute of Transport Studies, University of Leeds, Working Paper 126

CONTENTS

ABSTRACT

MURCHLAND, J.D. (1980) What are the errors we should be seeking to minimise in the estimation process? Leeds: Univ. Leeds, Inst. Transp. Stud., Work.Pap. 126

Three methods of fitting a gravity model - a triproportional model - to an observed trip matrix are compared. The first is the familiar practical method of choosing row, column and cost factors so that the model has the same row, column and cost sums as the grossed-up data. The second method, a true maximum likelihood estimation, chooses the factors so that the sums of the observed counts (not grossed-up) are matched. This differs from the first method only when the sampling probabilities vary from cell to cell. The third method applies the more modern approach of selecting a loss function which represents the practical effect of differences between the model values and the true values, and then chooses the model factors so that the expected loss, as far as it can be determined from the sample data and any prior information, is minimised. Squared error in flow times travel time is proposed as the loss function. It is noted that there is a loss function whose use is equivalent to maximum likelihood.

When the sample counts are large and the model fits well, each of the methods reduced to minimising the weighted squared, difference between the model and the saturated value. The variations in these weights show the differences between the three methods.

WHAT ARE THE ERRORS WE SHOULD BE SEEKING TO MINIMIZE IN THE ESTIMATION PROCESS?

J.D.Murchland
University of Leeds Institute for Transport Studies

1. Introduction

Gravity models of the distribution of trips from origins amongst destinations usually have their unknown parameters estimated by fitting to a sample of trips collected at one point in time. (This is already a state of sin since there is no guarantee that such 'cross-section' estimates are the right values in other circumstances - especially for parameters that are used as price elasticities.) Conventionally the most important parameters estimated from the base-year sample are the time or cost band parameters in a tabular attractance function, or the parameters in a functional attractance function (such as an exponential or gamma function). These are kept fixed for all forecasts.

Three different estimation methods are briefly described and compared below. The tabular attractance function (triproportional model) is mainly considered - the biproportional case is very similar.

2. Notation and the function e(x)

j	a row (origin)
k	a column (destination)
l	a 'file' (time or cost interval)
i	denotes jkl , for brevity (or jk in the biproportional case)
π_i	sampling probability for cell i (zero if unobserved)
n_i	sample count for cell i

* Version of 1979:6:17 with minor additions.

$D_i \equiv D_{jkl}$ prior value of flow for cell i , $\geq o$

D_i is zero if there can be no flow for i , and usually 1 otherwise. Commonly, there is only one l for which D_{jkl} is not zero.

$f_i \equiv f_{jkl}$ model flow for cell i , taken to be

$$f_{jkl} = a_j b_k c_l \; D_{jkl} \;=\; e^{u_j + v_k + w_l} \; D_{jkl} \qquad (1)$$

for a triproportional model

The function $e(x)$, defined as follows, is of great use here. The function is illustrated in Figure 1.

$$e(x) = x \ln x - x + 1 \qquad \text{for} \quad o < x$$

$$= 1 \qquad \text{for} \quad x = o \, . \qquad (2)$$

It is strictly convex over $[o, \infty)$. $e'(x)$ is $\ln x$. $e(o)$ is 1 , but $e'(o)$ is $-\infty$. For x close to 1

$$e(x) \simeq \frac{1}{2} (x-1)^2 \, . \qquad (3)$$

As a measure of the difference between x and 1 , $e(x)$ is approximately half the square error, but when included in an objective function, it won't allow x to become negative.

The functions

$$f_i \; e(D_i/f_i) \qquad \text{and} \qquad f_i \pi_i \; e(n_i/f_i \pi_i)$$

are convex functions of u_j, v_k, w_l (but not a_j, b_k, c_l) and for f_i close to D_i

$$f_i \, e\,(D_i/f_i) \simeq \tfrac{1}{2}(f_i - D_i)^2/f_i \qquad (4)$$

which is like the X^2 goodness of fit measure.

3. Cell sampling probabilities

The goal of the model is to forecast the number of trips from origin j to destination k in cost band l, averaged over relevant days (all annual weekdays, or all annual days, as the case may be). Hence the sample is drawn from the three-way array of origins by destinations by days of the survey period.

The sampling probability π_i is to be the probability of sampling any of the trips in cell i over the survey period, trips being taken in the sense of the model (as persons, vehicles or commodity volume, weight or value units). The count n_i for the sample is scaled up by $1/\pi_i$ to give the estimate for the cell.

For a roadside cordon point π_i is the product of the probability that the trip will pass through the point times the probability of being interviewed if it does. Each point gives a separate sample h, but these are easily combined : the counts n_{hi} are added, and so are the probabilities π_{hi}.

If household interviews are a uniform random sample of all zones over the survey period, and so of all trips, their sampling probabilities are the same for all cells.

Usually the sample counts n_i and π_i are not separately retained in survey processing, but they are there implicitly in the scaled-up values and it should be possible to recover them. For a cell in which n_i, and hence the scaled-up value, is zero, it will be necessary to use the sampling probability of some other, non-zero cell, judged to be similar.

Further work on sampling probabilities is undoubtedly needed, since there are various complications – such as home interview samples of non-home-based trips, and a distribution for the number of trip units per interview (giving a compound Poisson sampling distribution).

4. Multiproportional adjustment

A general method of adjusting a set of positive numbers so that particular subsets add up to specified totals, is to multiply each number by a factor for each sum in which it appears. As there are as many factors as sums this should be successful. An obvious algorithm is to adjust each factor in turn to make its sum correct.

Obviously this is not a statistical approach. It can be reformulated as a minimization problem: find numbers f_i which are as close as possible to given positive numbers D_i subject to the conditions that subsets of the f_i sum to the given totals, where closeness is measured by the function e,

$$\sum_i D_i \, e(f_i/D_i) \; . \tag{5}$$

As noted, if the fit is good this is approximately a minimization of the X^2 measure of disagreement. This minimization reformulation shows that if there is a solution, it is unique for the f_i. Also, the product form for f_i becomes a deduction, not an assumption.

Multiproportional adjustment is the original method of fitting the gravity model. In the biproportional case D_i would be the attractance function value if the cell was observed, and otherwise zero, and similarly unity and zero in the triproportional case. The row and column (and cost) sums of the scaled-up survey counts n_i/π_i would be taken as the required totals.

In its simple form this does not take account of different sample probabilities for different cells. This is easily remedied by using the $f_i \pi_i$ and $D_i \pi_i$ in place of f_i and D_i in the objective function (5) and changing to sums of un-scaled counts n_i in place of scaled ones. What is interesting about the adjustment process is that, once the prior values D_i have been chosen, it is only the observed sums which are relevant and not the individual observed cell values, even if known.

The measure of closeness may be changed to any similar function, but then the computational problem becomes more complex, and the property just mentioned is lost. It should be said that there are similar adjustment problems which are not multiproportional in the precise sense used here ('a factor for each sum in which it appears', only).

5. Maximum likelihood

A popular general method of parameter estimation from sample data is maximum likelihood: that is, choose those parameter values that maximize the probability (or probability density) of the sample. This is an intuitively plausible criterion, and as the sample size becomes indefinitely large the estimates tend to the true values, with the smallest variance that an estimator can have.

For a specified sampling method, either Poisson (at random according to the sampling probabilities) or multinomial (at random, according to the relative sampling probabilities, to obtain a chosen sample size), maximum likelihood estimation may be easily applied, assuming that the model is specified to be of product form

$$f_{jkl} = a_j b_k c_l D_{jkl} \qquad (6)$$

($D_{jkl} \geq 0$ given, usually zero or one according as the movement is possible or not). Indeed this gives a simple form of what is called a 'log-linear' model, about which there is a large literature.

If there are several independent Poisson samples they combine into one, as noted above. For this case maximum likelihood estimation becomes, computationally, a multiproportional adjustment problem, indeed the same one. Of course, the interpretation is different. Now it is the matching to observed sums of counts n_i which is deduced, instead of postulated.

Thw residuals are always of interest in judging the fit of the model for particular cells. In terms of what the minimization actually does in choosing parameters according to this criterion, the residual should be measured by

$$\check{f}_i\pi_i e(n_i/\check{f}_i\pi_i) \simeq \tfrac{1}{2}\pi_i(\check{f}_i - n_i/\pi_i)^2/\check{f}_i \,, \tag{7}$$

where $\check{f}_i$ denotes the minimizing value. (The appropriately signed square root can be taken to get a signed measure of departure which, if n_i is large and the model fits, is approximately a standardized deviate.)

A single multinomial sample is computationaly the same as a Poisson sample, but several overlapping multinomial samples, or one or more combined with a Poisson sample, give a little extra complication.

Maximum likelihood estimation naturally suggests a goodness of fit test. It is useful to compare the fit with the assumed product form model with that from the 'complete' or 'saturated' model which just takes the sample estimate n_i/π_i for each cell.

Now for the objections to maximum likelihood. The advantages listed are rather weak - any reasonable method will have the same large sample properties. For small samples there is no guarantee that the estimates are unbiassed. The method seems to depend on the truth of the specified model. However the abiding objection is that the maximized likelihood has no practical interpretation. It is one particular criterion, which cannot be changed to meet the needs of a specific problem. It will not indicate the value of collecting further data, except in the single sense of a reduction in the estimated asymptotic variance of the estimated parameters, and seems to make no sense at all if sampling error is negligible relative to the approximation inherent in using the model specified.

6. Optimal statistical decisions

The essence of an optimal decision approach is that the best choice of model parameters are those which minimize the expected loss. The loss function for the particular problem represents the cost incurred if the chosen model value differs from the actual value. The expectation is taken over the distribution of possible values for the actual value, conditional on any available prior information and on the available sample data. Choices of parameter values are constrained by the form of model adopted.

Consider one cell i in isolation. The indication i is omitted for the moment. Let $m > 0$ denote the actual parameter, $\phi_1(m)$ its probability distribution conditional on the sample, and f the chosen estimate.* If

$$L(f,m) \tag{8}$$

denotes the loss when f is chosen but m is the actual value, then the expected loss from choosing f is

$$\mathcal{E}L(f,m) = \int_0^\infty L(f,m)\phi_1(m)dm \ . \tag{9}$$

If a Poisson sample of n_i for the cell, obtained with a sampling probability of π_i, is available, it is convenient to assume that the prior distribution, $\phi_0(m)$, has a gamma form

*The notion that the true value of the parameter <u>has</u> a probability distribution is a new viewpoint.

$$\frac{\beta^{\alpha}}{\Gamma(\alpha)} m^{\alpha-1} e^{-\beta m} \quad , \quad o<\alpha,\beta \tag{10}$$

where α_o and β_o express the prior information. This has mean α_o/β_o and variance α_o/β_o^2, and the prior information may be interpreted as equivalent to a sample count of α_o obtained with a sampling probability β_o. The convenience of this distribution lies in the fact that the posterior distribution after the sample n,π has been obtained will also be of gamma form, with

$$\alpha_1 = \alpha_o + n \quad , \qquad \beta_1 = \beta_o + \pi \quad , \tag{11}$$

and so a mean of $(\alpha_o+n)/(\beta_o+\pi)$ and variance $(\alpha_o+n)/(\beta_o+\pi)^2$. (As a bonus, the integrals for the expectation can be looked up in a table of Laplace transforms.)

If no prior information is available, the limiting form $\alpha_o = o$, $\beta_o = o$ should be used (a so-called 'diffuse prior'). Note the importance of prior knowledge when n_i happens to be zero.

Two simple forms of loss function are the deviation

$$L(f,m) = |f-m| \tag{12}$$

and half the squared error (quadratic function)

$$L(f,m) = \tfrac{1}{2}(f-m)^2 \quad . \tag{13}$$

If f is not constrained (the saturated model), the expected loss is minimized in the deviation case by choosing a median, and in the quadratic case by the mean. When f is constrained the expected loss splits into a part which would be incurred even by the saturated model, and a part arising from the difference between the constrained value and the saturation value. Figure 2 shows several loss functions.

What is the most appropriate loss function for the fitting of a model of trip distribution? Amongst absolute or relative error in either flow or flow times evaluated travel time there is a strong argument for absolute error in flow times evaluated travel time, namely: this is the quantity which enters into economic and other evaluation. (Evaluated time, that is evaluated perceived generalized cost expressed in time units, will vary markedly between different purposes, because of the much higher values given to commercial or business purpose trips, but usually each purpose is fitted by a separate model.) Accident exposure depends on time on the road. Assignment is concerned to get the correct total time on the network, or possibly kilometerage, and also if a cell is misestimated, it will contribute incorrectly to every cordon which its routes cross. Again this argues for flow times time, rather than flow alone.

Let T_i denote the travel time, or other weight, for cell i. The weighted deviation loss function

$$T_i|f_i - m_i| \tag{14}$$

is probably the form that would be generally preferred. Although this is not intractable, it is harder to handle than the weighted quadratic loss function

$$\tfrac{1}{2}T_i^2(f_i - m_i)^2 \quad , \tag{15}$$

whose expected value over the gamma distribution is

$$\tfrac{1}{2}T_i^2\frac{\alpha_{1i}}{\beta_{1i}^2} + \tfrac{1}{2}T_i^2(f_1 - \alpha_{1i}/\beta_{1i})^2 , \tag{16}$$

$$\text{where} \quad \alpha_{1i} = \alpha_{oi} + n_i , \qquad \beta_{1i} = \beta_{oi} + \pi_i . \tag{17}$$

The optimal decision approach for the triproportional model (1) thus chooses a_j, b_k, c_l to minimize

$$\tfrac{1}{2}\sum_i T_i^2(f_i - \alpha_{1i}/\beta_{1i})^2 . \tag{18}$$

This objective could have been proposed directly, but the optimal decision approach has related it to a chosen determinate loss function, to the Poisson sampling used, and to any prior information available.

The minimization may be carried out by minimization with respect to each variable in turn, just like multiproportional adjustment. For a_j , for instance, this gives

$$a_j \leftarrow \sum_{k,l} T_{jkl}^2 (b_k c_l D_{jkl})^2 \Big/ \sum_{k,l} T_{jkl}^2 (\alpha_{1jkl}/\beta_{1jkl}) b_k c_l D_{jkl}, \tag{19}$$

and similarly for b_k and c_l. Unfortunately the objective function is not convex in a_j, b_k, c_l (or u_j, v_k, w_l) , but provided all the variables are bounded the algorithm will at least converge to a local minimum (Luenberger). Unlike the multiproportional or maximum likelihood problems, the observed row, column and cost sums are not used, and the cell estimates α_{1i}/β_{1i} (essentially the grossed-up observations) are constantly employed.

There lies the apparent disadvantage of this estimation method: the row and column sums play no special role, and the forecasting problem has to be rethought afresh. (One may comment on the asymmetry of the usual approach: why retain the base year cost factors and forecast future row and column sums? Why not retain base year row and column factors and forecast future cost sums? Why is it easier to forecast numbers of trips than total travel times?) It would be possible to retain the customary biproportional forecast, assuming the forecast for the base year was close enough to the base-year fitted model.

There is a curiosity, which perhaps has a deeper explanation. If the loss function is chosen to be the former error measure

$$\beta_1 fe(m/f) \tag{20}$$

the expected loss turns out to be

$$\mathcal{E}\beta_1 fe(m/f) = \mathcal{E}\alpha_1 e(m\beta_1/\alpha_1) + \beta_1 fe(\alpha_1/\beta_1 f) \,. \tag{21}$$

The first term on the right band side is the unavoidable expected loss from the saturated model value α_1/β_1. It evaluates to

$$\alpha_1(\psi(\alpha_1+1) - \ln \alpha_1) \approx \frac{\alpha_1}{.3652+2\alpha_1} \tag{22}$$

where ψ is the psi or digamma function (the derivative of the logarithm of the gamma function), and the approximation is accurate for large α_1 and is otherwise within .o2 of the correct value.

Hence with this loss function for a Poisson sample the decision approach is identical to maximum likelihood. Maximum likelihood has been justified!

If the loss function for the decision approach is changed from the quadratic form (15) to the weighted form of (2o) which has approximately the same value for small errors

$$L(f,m) = T^2 f^2 e(m/f) \simeq \tfrac{1}{2}T^2(f-m)^2 \tag{23}$$

the minimization problem (18) is now convex in the parameters u_j, v_k, w_l. This is desirable computationally. For large errors the true quadratic (15) gives a much greater penalty than (23) - for a poorly fitting model (15) will sacrifice everything else to a minimization of the very largest errors. Choosing a loss function requires not only a choice of what quantity to measure the loss in, but also how large errors should be weighted relative to small ones.

7. Conclusion

Three approaches to estimation of a base-year trip distribution model have been presented. They can be succintly compared: for Poisson sampling, subject to the model form imposed

$$F_i = f_{jkl} = a_j b_k c_l D_{jkl} , \tag{24}$$

the original form of multiproportional adjustment to sums of scaled-up counts minimized

$$\sum_{\substack{i \\ 0<\pi_i}} f_i e(n_i/f_i\pi_i) \simeq \tfrac{1}{2}\sum_{\substack{i \\ 0<\pi_i}} \frac{1}{f_i}(f_i - n_i/\pi_i)^2 \tag{25}$$

while the later form with sums of unscaled counts, and maximum likelihood, both minimize

$$\sum_{\substack{i \\ 0<\pi_i}} f_i \pi_i e(n_i/f_i\pi_i) \simeq \tfrac{1}{2}\sum_{\substack{i \\ 0<\pi_i}} \frac{\pi_i}{f_i}(f_i - n_i/\pi_i)^2 , \tag{26}$$

while the decision approach minimizes the chosen loss (15)

$$\frac{1}{2}\sum_{\substack{i\\0<\pi_i}} T_i^2(f_i-n_i/\pi_i)^2 \tag{27}$$

if there is no prior information, where T_i is the evaluated travel time. While the weighting factors π_i/f_i and T_i^2 have some general resemblance, because large times are associated with few trips, they will not be very close, and the fitted parameters are likely to be substantially different.

The optimal decision method has the advantage that the objective is meaningful practically, and the minimized value expresses the expected error in the fitted model. It can also utilize prior information. So far it has not been tried numerically. However, its adoption would seem to require rethinking the method by which forecasts are made from the base-year fitted model.

8. References

Hays W.L. and Winkler R.L.
Statistics: probability, inference and decision (1971)

Luenberger D.
Introduction to linear and nonlinear programming, Addison-Wesley (1973) 158-9

Murchland J.D.
The multiproportional problem, Transport Studies Group University College London, Gower Street, London WC1E 6BT, JDM-263 (1977)

Shenton L.R. and Bowman K.O.
Maximum likelihood estimation in small samples, Griffin (1977)