MODELLING TRIP FREQUENCY AS A POISSON VARIABLE

By Tim Barmby and Jurgen Doornik*

Data on the spatial behaviour of households are typically not continuous. Destination and mode choice are qualitative variables, and the datum of interest here, trip frequency, is a discrete variable which can only take non-negative values. It can be considered inappropriate to use the classical linear regression model (CLRM) for the analysis of data of this nature, as in Robinson and Vickerman (1976) and Vickerman and Barmby (1984). There are three distinct reasons for this, which are discussed by Gourieroux, Monfort and Trognon (1984). Firstly, the observation set is not that of the CLRM; secondly, the assumption of normality for the error term cannot be made; and, thirdly, the predictions from the CLRM would allow for impossible values. Use of the CLRM would be advocated on the grounds that it is empirically tractable, and that it will give estimates which are in some sense robust. The first point is undeniable; but as we hope to demonstrate here, it would not necessarily require a prohibitive amount of extra investment to construct and estimate a statistical model which would describe the data generating process more coherently. On the question of robustness, we are essentially trying here to reduce the inherent specification error; this must intuitively result in estimates which are more informative about the underlying processes. To emphasise this point we will draw comparisons with previous regression-based analyses.

ECONOMIC MODELS

The modelling of shopping trip behaviour is of practical importance. For instance, important implications for the planning of new shopping facilities may be drawn from the pattern of this behaviour and the extent to which it is influenced by the nature of the spatial market and the characteristics of the population of households.

The conventional approach to the analysis of trip data is to use a trip generation relationship of the form

* The authors are respectively in the Department of Economics at Loughborough University and in the School of Management and Economic Studies at the University of Sheffield. They thank Roger Vickerman, the editors of this journal, and an anonymous referee for valuable comments, and Ray Robinson for making the data available to them. All remaining errors are of course theirs.
\[ T_i = f(X_i, Z; \beta) \] (1)

where

- \( T_i \) is the number of trips for a specific purpose undertaken during a given period of time,
- \( X_i \) is a vector of characteristics of household \( i \),
- \( Z \) is a vector of characteristics of the spatial environment,
- \( \beta \) is a vector of unknown parameters characterising this relationship.

(See Vickerman, 1980, and Vickerman and Barmby, 1984.)

We will maintain in our analysis this distinction between household and environmental characteristics. We use car ownership, size of household and expenditure to measure household characteristics. The spatial environment is characterised by cost and attraction indices of shopping centres.

As stated in the introduction, one empirical approach would be to approximate the trip generation function as a linear regression function. Our approach, which we think has both empirical and conceptual advantages, is to find a more data-coherent way of representing (1), without the drawbacks outlined in the introduction.

POISSON MODELS

One possible course which could be fruitful is to model the number of observed trips, \( T_i \), as a Poisson variable. This would have two distinct advantages. Firstly, the model could not predict a negative number of trips for certain values of the regressor variables (this is possible for the CLRM, as explained in the introduction). Secondly, the estimates of the model show underlying probabilities for actual numbers of trips, whereas in the regression situation we are given only the expectation and variance of the number of trips, as implicitly \( T \) would be a continuous variable. A Poisson model could be described as:

\[ T_i \sim f(t_i) = \frac{e^{-\lambda_i} \lambda_i^{t_i}}{t_i!} ; t_i = 0, 1, 2, \ldots \] (2)

\[ E(T_i) = \lambda_i = \exp(X_i' \beta); \ i = 1, \ldots, n \]

\( X_i \) is a vector of characteristics of the household which defines the mean of the distribution. The log likelihood function can be formed as the sum over the sample observations of the log of the above densities. This can be maximised with respect to the parameters \( \beta \).

Of course the choice of the Poisson as a probability function describing the realised distribution of trip frequencies is not without its implications. We might have allowed for a more coherent description of the data generating process, but we need to examine more closely the basis of the Poisson distribution to make a clearer statement about the assumptions now underlying our model.

A Poisson variable is obtained as the number of events of interest occurring in a given interval of time. To generate the Poisson form for the probability function, these events must have occurred independently through time. This condition is
MODELLING TRIP FREQUENCY AS A POISSON VARIABLE  
T. Barmby and J. Doornik

clearly restrictive, as the occurrence of events can be dependent in several ways. The occurrence of an event can condition the probability of the occurrence of a subsequent event, either increasing this probability or reducing it (see Cameron and Trivedi, 1986).

A second problem inherent in the use of the Poisson distribution is that the variance is constrained to be equal to the mean. If this same relationship does not characterise the data, then the data are said to be under or over dispersed, according to whether the variance is less than or greater than the mean. Cox (1983) outlines both the consequences of this situation and a method of detecting its existence. The structure of the model can be made less restrictive, and in particular the over (under) dispersion can be circumvented, by modelling \( \lambda \), the Poisson parameter, as a Gamma distribution, \( h(\lambda) \). The new distribution of the observed number of trips can be obtained by mixing the distribution as

\[
g(t) = \int_0^\infty f(t; \lambda)h(\lambda)d\lambda
\]

The resulting form will be a Negative Binomial distribution

\[
T_i \sim g(t_i) = \frac{\Gamma(t_i + \gamma)}{\Gamma(t_i + 1)\Gamma(\gamma)} \left( \frac{\gamma}{\gamma + \mu_i} \right)^{t_i} \left( \frac{\mu_i}{\gamma + \mu_i} \right)^{\gamma_i} t_i = 0, 1, \ldots
\]

Following Cameron and Trivedi (1986), we parameterise the above model as

\[
\begin{align*}
\mu_i &= \exp(X_i' \beta) \\
\gamma &= \frac{1}{\alpha} \left[ \exp(X_i' \beta) \right]^k \quad \alpha > 0 \\
E(T_i) &= \exp(X_i' \beta) \\
VAR(T_i) &= E(T_i) + \alpha[E(T_i)]^2 - k
\end{align*}
\]

It can be seen now that the variance and mean are no longer constrained to be equal, and the parameters \( \alpha, k \) will determine the form of the relationship between \( E(T) \) and \( VAR(T) \). However, whatever the value of \( k \) is, the hypothesis \( H_0: \alpha = 0 \) is the hypothesis of no over (under) dispersion, which corresponds to the Poisson model. This is of central interest here because rejecting or accepting this hypothesis will decide whether the simpler Poisson framework is appropriate for our data. Cameron and Trivedi (1986) use two different Negative Binomial models, obtained by setting \( k = 1 \) and \( k = 0 \) respectively. However, this is unnecessarily restrictive; so, to achieve maximum flexibility in modelling the relationship between the mean and variance, we established \( k \) by a grid search.

Rickard (1988) compares the use of the Poisson distribution and the Negative Binomial to model long-distance rail trips as a generalised linear model (GLM) (McCullagh and Nelder, 1983). She finds the Negative Binomial distribution to be the more appropriate, and postulates that this is because the overall distribution is the sum of those of a number of sub-groups, each following its own Poisson distribution. Although both the context and the methods used are different, in essence her conclusions are similar to those derived in this paper.
DATA CONSIDERATIONS

The data are taken from the Sussex Household Shopping Survey, and are the same as are used in Robinson and Vickerman (1976) and Vickerman and Barmby (1984, 1985). The purpose of this survey was to measure accurately the trip making behaviour, for a period of one week, of a sample of households and to relate this to a set of household characteristics. The questionnaire used did not record more than seven trips, so there is an upper truncation point at seven. We take account of this truncation in estimating the Negative Binomial model; in general, if $T \sim f(t)$, truncation at $T^*$ will result in a truncated density of the following form (see also Cohen, 1961):

\[
g(t) = \frac{f(t)}{T^*}; t=0, \ldots, T^*
\]

\[
\sum_{j=0}^{T^*} f(j)
\]

(5)

We have omitted the specific functional form for the truncated density corresponding to (3), as the derivation is straightforward but tedious.

SOME ESTIMATION RESULTS

The truncated Negative Binomial model was estimated for a simple specification, and the results are reported in Table 1. The variables used are: number of cars owned or available for use, $CAR$; number of individuals in the household, $NUM$; total expenditure in the survey week, $EXP$; and two indexes for attraction, $SA$, and cost, $SC$, to the nearest $n$ centres (where $n$ is varied so as to include 95 per cent of all trips). These variables are defined in exactly the same way as in Vickerman and Barmby (1984, 1985). These results are compared against regression results similar to those reported in Vickerman and Barmby (1984, Table 2, p. 117); the only difference is that income is not included in the same specification as observed expenditure, as income is likely to be an important determinant of expenditure and should be modelled separately. The highly significant coefficient for $\alpha$ rejects the simpler Poisson model, so these estimates are not reported.

CONCLUSION

We conclude that a generalisation of the Poisson distribution, the Negative Binomial distribution, can be usefully employed in constructing a statistical model of trip frequency. It appears that the data are characterised by overdispersion, so the simple Poisson distribution would be too restrictive. The comparison with the regression results is interesting. The two models have similar implications for the effects of car ownership and of the attraction and cost indexes. Car ownership is again shown to have no clear effect, and this result must now be regarded as being established for this sample of data. The attraction and cost indices also
MODELLING TRIP FREQUENCY AS A POISSON VARIABLE  T. Barmby and J. Doornik

TABLE 1

Estimation Results

<table>
<thead>
<tr>
<th></th>
<th>Truncated Negative Binomial model (k grid searched)</th>
<th>Regression model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>CONST</strong></td>
<td>1.0280 (7.2915)</td>
<td>3.0980 (9.8506)</td>
</tr>
<tr>
<td><strong>CAR</strong></td>
<td>-0.0472 (1.0555)</td>
<td>0.07912 (0.9052)</td>
</tr>
<tr>
<td><strong>NUM</strong></td>
<td>0.0108 (0.5174)</td>
<td>0.1486 (3.4558)</td>
</tr>
<tr>
<td><strong>EXP</strong></td>
<td>0.0404 (12.0763)</td>
<td>0.0067 (3.4167)</td>
</tr>
<tr>
<td><strong>SA</strong></td>
<td>0.0062 (2.0334)</td>
<td>0.1802 (2.7591)</td>
</tr>
<tr>
<td><strong>SC</strong></td>
<td>-0.0330 (5.3454)</td>
<td>-0.8746 (5.5145)</td>
</tr>
<tr>
<td><strong>α</strong></td>
<td>0.0463 (4.7792)</td>
<td></td>
</tr>
<tr>
<td><strong>k</strong></td>
<td>-0.25</td>
<td></td>
</tr>
<tr>
<td><strong>Log likelihood</strong></td>
<td>-1585.8846</td>
<td></td>
</tr>
</tbody>
</table>

\( n = 850 \)

asymptotic \( t \) values in parentheses

maintain their effect in terms of sign and significance.

Figure 1 shows the advantage of the Negative Binomial in fitting the observed data; the implied relative frequencies are computed as the mean of the implied individual probabilities. Though both the normal frequency curve, implied by the regression results, and the Negative Binomial lack the flexibility to pick up the bimodality in the observed data at trip level one, the Negative Binomial clearly tracks the relative frequencies of the observed data better than does the normal curve. We also computed the relative frequencies for the Negative Binomial by evaluating the probabilities at the sample mean values of the regressors; this substantially underestimated the proportion making trip level one. Our alternative method of giving an aggregate representation of our results rests on the use in our estimation of individual level and not aggregate data. Though comparison of the predicted and observed proportions at the aggregate level is a useful
informal diagnostic device, we feel that an equally important aspect for this paper is the identification of significant effects at the individual level, and this would not be achieved by an aggregate study. Modelling the bimodality directly could be achieved by considering a mix of discrete densities. We leave this for future research.

The contrast between the two sets of results is seen in the way in which the Negative Binomial model highlights the expenditure variable, EXP, at the expense, in terms of significance, of NUM, the number of individuals in the household. As we remember that the observed trips are for the specific purpose of shopping, the data seem to be indicating that expenditure levels are more informative than a crude head count of individuals in the household. One possible interpretation of this result is that simply counting the number of individuals in a specific household will not give an accurate picture of the household's composition, which can

314
MODELLING TRIP FREQUENCY AS A POISSON VARIABLE T. Barmby and J. Doornik

be constructively thought of as determining its shopping trip behaviour. However, information on this composition will be present in realised expenditure in a given week. This effect is more likely to be picked up when the explanatory variables are related to the dependent variable in a more coherent structure, such as that explored in this paper.

Date of receipt of final typescript: August 1988

REFERENCES