The Problem Of Infrequent Trips And Excess Zeros In Modelling Trip Generation[1]

Dr Helen Johnson[1], Dr Mary Kynn[2],
[1] Queensland University of Technology, Brisbane, QLD, Australia and QLD Main Roads, Brisbane.
[2] University of Lancaster, UK

1 Introduction

The purpose of this paper is to analyse the effectiveness of the current method of modelling trip generation used in the Brisbane Strategic Transport Model (BSTM), and to propose an alternative methodology.

The current method of predicting trip productions is to use ordinary linear regression with dummy variables. Ortuzar and Willumsen (1994) provide an introduction into this methodology for transport modelling. In brief, trips are split by purpose and for each purpose a separate regression equation is calculated relating the number of trips (for each purpose) to household attribute (such as the number of workers, dependents or vehicles). This is a simple method of allowing zone totals to be used as inputs for the model equations. However, some of these models have a very low explanatory power and violate some basic and extremely important statistical assumptions of linear regression. These assumptions are that the range of dependent variable is unrestricted, the dependent and independent variables are linearly related, and that errors are independent, normally distributed and uncorrelated with a constant variance. The dummy formulation is used to avoid issues of non-linear relationships, yet does not address the problems created by the restricted range of the dependent variable (that is, the number of trips can only be positive) nor the non-normal, non-constant errors. The use of simple regression is further complicated by the presence of a large number of zero trips for all trip purposes, which will be a large focus of this paper.

Therefore, the main objection to the current models is that the distribution of the variables is not appropriately considered. Many published trip generation models use a log transform to account for the restricted range of the dependent variable and the non-constant errors (for example, Washington (2000)). However, count data with values close to zero may be more effectively modelled using a Generalised Linear Model (GLM) framework with a Poisson or a negative Binomial distribution; Said and Young (1990); Hellerstein (1991). In addition, where there is a large proportion of zeros, a more appropriate model may be a two-stage model or a zero-inflated regression model. A two-stage model was applied to trip generation data by Monzon, Goulias and Kitamura (1989), where the authors concluded that the extra time involved in calculating a two-stage model was not justified by the relative improvement in model performance. However, this was not explicitly used on zero-inflated data and so the conclusions are not clear. A variation on the two-stage model is a zero-inflated model developed by Lambert (1992), in which the parameters are estimated simultaneously and for which software is freely available. In terms of computing time there is little additional effort involved in calculating a zero-inflated model over a linear model. This has yet to be trialled in the trip generation literature. As such, this paper compares the following models:

1. Linear model: stratified dummy linear regression.
2. Log-linear model: log transform of dependent variable.
3. Zero-Inflated Poisson/Negative Binomial (ZIP/ZINB) regression.

[1] This work was carried out by Dr Mary Kynn whilst working for the Planning, Design and Environment Division, Department of Mainroads, Brisbane.

2 Dataset description

The Brisbane Household Travel Survey collected data in the form of travel diaries from approximately 4000 households. Each household filled in the diary for a representative working day, during October-November 2003 or February-March 2004 (note that this did include some school and university holidays). The previous 1992 Household Travel Survey had drawn a random sample from residential Energex billing addresses and used a mail out–mail back survey. This sampling methodology was changed slightly for the 2003/04 SEQTS to improve a dropping response rate observed in other cities. It was decided to change to a hand delivery–hand collection system with follow-up telephone calls to encourage the completion of questionnaires. To facilitate this more labour intensive method a two-step process was used when sampling. First, census collection districts were sampled (176 out of 3000), and then households were randomly sampled within each district (at the rate of approximately 6%). Using these techniques the target response rate was 60%, giving a final sample size of 4057. This information is explained in more detail in The Urban Transport Institute (2005).

The variables pertinent to this analysis are: the number of trips (by purpose); household structure (blue collar, white collar and undefined workers), dependents by age groups A[0-17], B[18-64],C[≥65]), and household vehicles where those used in the subsequent analyses are given in Appendix 1.

3 Model Formulations

As described in the introduction, three models will be discussed in this paper: a Linear model; a Log-linear model; and a Zero-Inflated Poisson/Negative Binomial (ZIP/ZINB) model.

3.1 A Linear Model

The current method of modelling trip production in the Brisbane Strategic Transport Model (BSTM) is to use simple linear regression with no intercept with stratified household attributes included as dummy 0-1 variables. While this is effective for calculating average trip rates for various household attributes, the nature of the data is such that the assumptions of linear regression are seriously violated. This means that the usefulness of these averages applied at various levels is questionable and measures of model accuracy are not to be trusted. However, the results from this particular technique will be compared with the models fitted in Section 3.2.

3.2 A Log-linear model

A common technique in the transport modelling literature is to transform the dependent variable (the number of trips) onto the log scale. This has the effect of 'normalizing' the distribution (under the hypothesis that the data is actually from a Poisson distribution) and tempering the effects of increasing variance. The histograms in Figure 1 show the trip rates transformed by first adding 1 (since it is not possible to take the log of zero) and then by taking the log to base 10. In this case, it makes the spike at zero even more prominent, and so it is unlikely it will improve the model at all. Other transformations are possible but there is no transformation that can 'spread out' the zeros, and none would significantly improve the model overall.

The model coefficients are not presented here as there are some considerations when transforming the parameters back from the log scale and the parameters should not be directly interpreted in terms of trip rates. Rather, by looking at the histograms, it is clear that

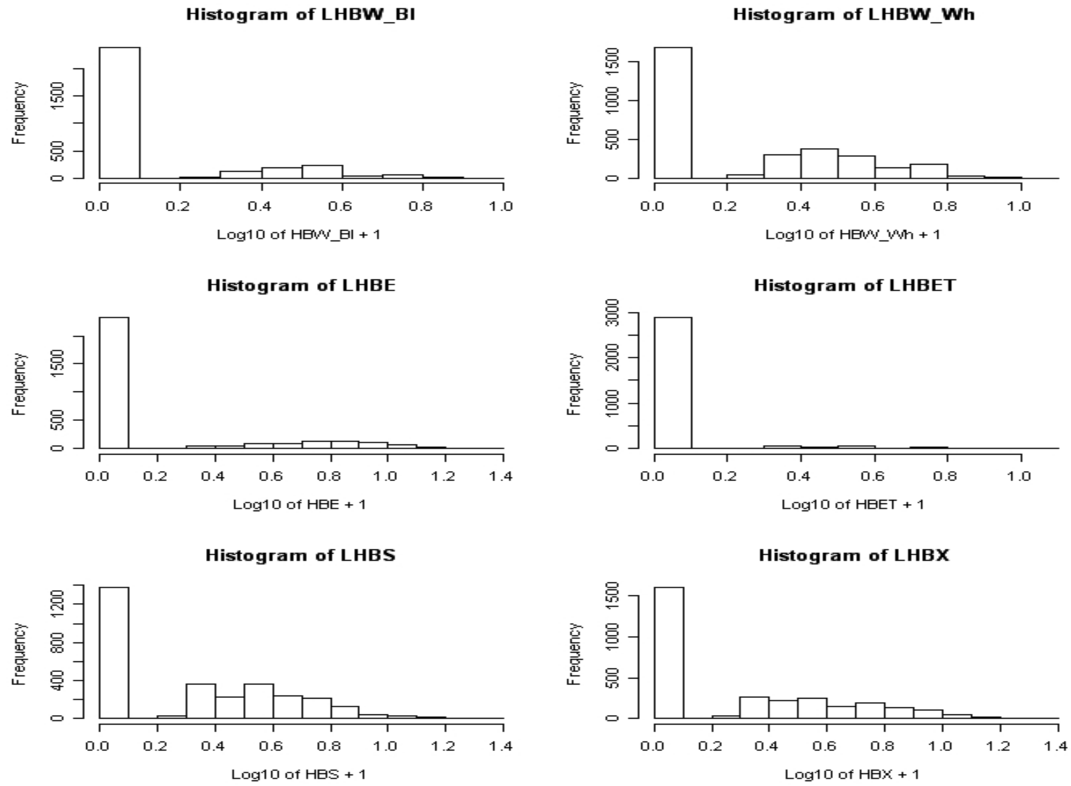

Figure 1: Log-linear model: histograms of the transformed trip rates by each trip purpose.

a log transform is not appropriate in this case. Observation of the residual plots and prediction intervals do confirm this (although these are not shown here) where rather than accommodating the assumptions of normally distributed, constant errors, the non-normality has in fact been exacerbated. This was indicated in the initial histograms that showed a spike at zero for each trip purpose. It is clearly evident that when the data is heavy in zeros taking a log transform of the trip rates is not an appropriate way to 'fix' the problems of linear regression.

3.3 Zero Inflated Poisson regression

3.3.1 Assumptions

Count data with a large number of zeros have been analysed in a number of fields and there exists methodology and supporting software to conduct such analyses. The original zero inflated Poisson model (Lambert, 1992) describes of a factory line which fluctuates between perfect and imperfect states. The perfect state producing no faults (that is, all zeros), and the imperfect state producing some faults according to a Poisson process, where this may also include some zeros. The overall percentage of zeros in this example was 81%, which is clearly higher than can be explained by the Poisson distribution alone (for example, a Poisson distribution with a mean of 1 predicts only 37% zeros, a mean of 2 predicts 14% zeros and a mean of 3 predicts 5% zeros). For the trip production data the percentage of zeros by trip purpose is given in Table 1.

Other count distributions have been tested in accounting for cases with a high proportion of zeros such as the binomial and negative binomial. However these cannot adequately model data with a high proportion of zeros either, unless also mixed with a distribution giving a point mass at zero. For this analysis, the zero-inflated Poisson model of Lambert (1992) was used in three cases, and a modified form using the negative binomial distribution (Heilbron (1994); Lewsey and Thomson (2004)) was used to model the count part in the

Table 1: Percentage of zeros for each trip purpose in the whole dataset.

Purpose	% zero
Blue	78
White	56
Edu	76
Edu Tert	95
Shopping	44
Rec/Other	53

remaining three cases. This is because some of the trip purposes showed over-dispersion in the trip rates which is better accounted for by the negative binomial distribution. This relaxes the strict mean-variance relationship of the Poisson distribution and allows for greater dispersion. The effect of the high proportion of zeros on the overall distribution can be seen in Figure 2, which shows boxplots of trip rates by purpose including and excluding the zeros.

As with the study of defects in a factory line, there are also two types of zeros in trip generation. Zeros can result when a trip purpose is not applicable to any member of the household, and also when no member of the household chooses to make a trip for a particular purpose. For example, a house with no blue collar workers will not make any blue collar work trips. However, when a household does have blue collar worker(s) there may be still be zero trips produced on that day due to illness or personal holidays, for example. In the case of shopping and recreation trips there is no category which precludes these trips, so zeros are solely due to ‘choice’.

As a result, the model has two components; the probability that no trips are made which may include both the ‘not applicable’ groups and the ‘free choice’ groups, and the distribution of trips modelled by a Poisson or negative binomial distribution (which may also contain some zeros). For a zero-inflated Poisson regression, Lambert (1992) describes the model as follows:

The responses $\boldsymbol{Y} = \left(Y_1, ..., Y_n\right)$ are assumed to be independent, and

$$Y_i \begin{cases} \sim 0 \text{ with probability } p_i \\ \sim Poisson\left(\lambda_i\right) \text{ with probability } 1 - p_i \end{cases}$$

so that

$$P\left(Y_i = 0\right) = p_i + \left(1 - p_i\right)\exp\left(-\lambda_i\right)$$
$$P\left(Y_i = k\right) = \left(1 - p_i\right)\exp\left(-\lambda_i\right) \times \lambda_i^k / k!$$

where the expected value is given by $E\left(Y_i\right) = \left(1 - p_i\right)\lambda_i$, since the mean of the Poisson distribution is λ itself.

The case is similar for zero-inflated negative Binomial regression, although the equation for the count part of the model is more complicated (see Cheung (2002) for details). Therefore, in words, the expected number of trips is the probability that a trip is made multiplied by the mean trip rate as defined by the Poisson distribution. These particular models were developed specifically for count data, whereas the trip data had been weighted and was no longer discrete. It was decided to round the trip data back to the closest count to facilitate

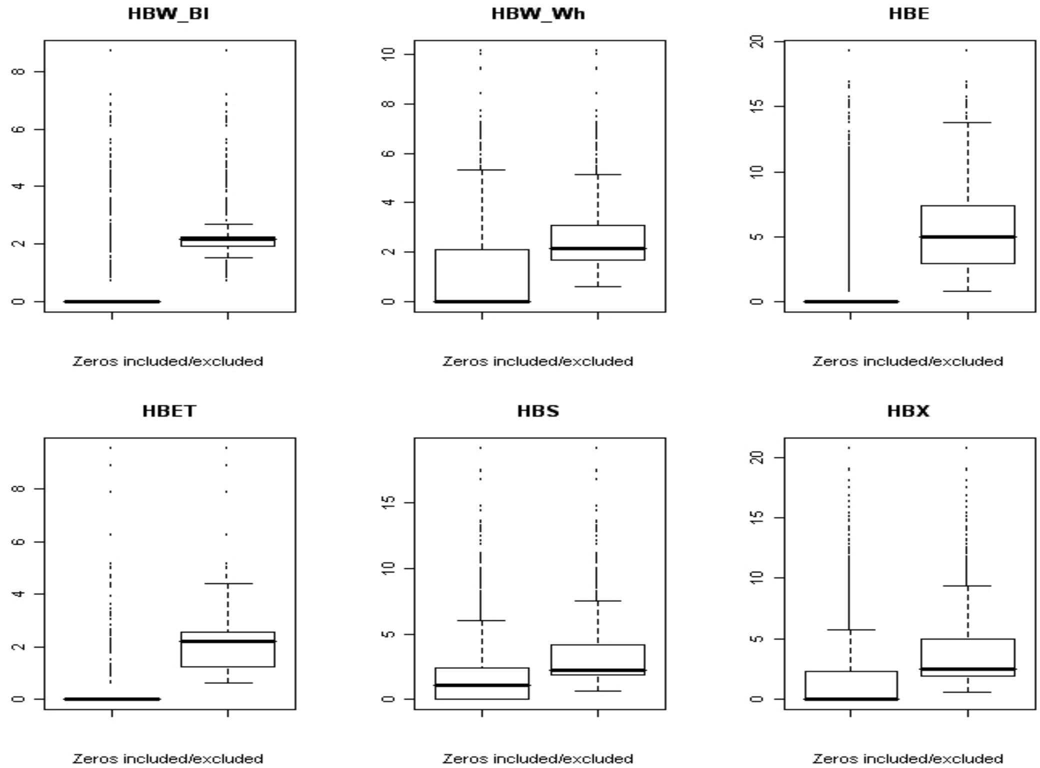

Figure 2: Boxplots of trip rates for each trip purpose including and excluding the zeros.

this investigation, and to make sure that all models were compared on equal grounds the linear models were rerun on the rounded data. This had very little impact on the actual model coefficients or measures of fit.

3.3.2 Formulation and fit

The parameter values of the zero-inflated models are given in Appendix 2. It is interesting to compare the variables that were significant in the zero-inflated models with those in the linear models, and to make comparisons of model fit. Table 2 shows the variables significant in the linear regressions and those significant in the zero-inflated regressions. In most cases, the same variable that was significant in the linear regression was significant for predicting the zero-inflated part of the model (represented by a Z), when these zeros were accounted for, other variables then became significant in predicting the counts (represented by an X). So for blue collar workers, the number of blue collar workers in the household was the best predictor of whether or not any blue collar trips were made. Having accounted for the mass of zeros, the household structure then became significant in determining the number of trips made. Note that the number of vehicles owned by the household was tried for each model but since this was highly correlated with the number of workers (particularly the number of white collar workers) it was not sensible to include it. The final choice of variables relied on variable significance, or practical importance, and lowering the Akaike's Information Criterion (AIC) value where the AIC measure the 'distance' from the true model (that is, the data) and the approximating model. The larger the distance, the worse the model, and this will increase with sample size irrespective of the type of model.

When comparing the performance of zero-inflated regression with linear regression it was also of interest to see what improvement could be made with no additional variables. So for each trip purpose a simple zero-inflated model was created using only the variables significant in the linear regression for both the zero and the count parts of the model. This comparison makes it clear how much of the improvement in the model performance is

Table 2: Variables significant in the (L) linear and (ZI) zero-inflated models. Z represents a variable included in the zero-inflated part of the model, and X represents a variable included in the count part of the model.

Stratified Household Attribute Variable	Blue		White		Edu		Edu T		Shop		Rec/ Oth	
	Variables included											
	L	ZI	L	ZI	L	ZI	L	ZI	L	ZI	L	ZI
Households (intercept)						Z		Z				
Blue Collar Workers	X	ZX		X		X		Z		X		
White Collar Workers			X	ZX			X	Z		ZX		
Dependents A (0-17)		X		X	X	ZX				X		
Dependents B (18-64)		X		X		X	X	ZX		ZX		
Dependents C (>=65)		X		X						ZX		
Total Persons									X		X	ZX

achieved by more accurately describing the underlying structure of the data, without relying on any additional predictive variables. Table 3 gives a comparison of AIC values for the linear regression, simple zero inflated regression and ‘free’ zero-inflated regression. It is clear that most of the improvement in the model has come from the different representation of the data, and not from the additional variables. This is not to imply that accounting for household structure is not important, but rather it is to demonstrate the impact of the underlying structure of the data on the final model.

3.3.3 Diagnostics

Tools for diagnosing non-linear models are given in Huet, Bouvier, Poursat and Joliet (2004). Briefly however, although there are no assumptions made about the normality of errors as there are with simple linear regression, it is important to examine residual plots to see if errors are random or if any ‘patterns’ remain. There are several types of residual plots that can be examined. Two of these are presented here. These are: absolute residuals versus the fitted values, and the standardised residual versus the fitted values.

The plots of the absolute residuals versus the fitted values (given in Figure 3) show random scatter about zero, although there may still be a slight tendency for increasing variance which could perhaps be investigated further. This is a marked improvement on the linear models although the plots are not shown here for brevity.

The plots of the standardised residuals versus the fitted values, given in Figure 4, show that predicting zeros was vastly more accurate for the blue and white-collar workers than for the other purposes. This is due to the type of zeros. For work trips it is easy to predict that a household with no workers of the specified trip will not make any of those work trips, and there will be little error involved in those predictions. However, for the other trips purposes there is no defining category that precludes a certain trip. Instead, different households have a different probability of making a trip, and the predicted or expected value is this probability multiplied by the predicted number of trips from the count part of the model. So if the household has a 0.3 chance of making an education trip, and if the expected number of (person) trips is 10, given that they do go out, the overall expected value is 3.33. Thus giving a large residual if the observed value is 0 or 10. Given the nature of trip making for less clearly defined purposes, predictions at the household level are very difficult. However, lack of fit for the education trips may also be affected by the data, which was partly collected during the holidays. In either case, in terms of predictions it is more instructive to look at the overall probability of each of the counts and these are presented for each of the models in section 4.

Table 3: comparison of AIC values for the linear regression, simple zero inflated regression and 'free' zero-inflated regression

	3.3.3.1.1.1 Blue	White	Edu	Edu T	Shop	Rec/ Oth
Model	**AIC**					
linear	6162	8972	12574	5358	13524	14320
simple ZI	2773	5853	5047	1522	10400	9932
free ZI	2768	5789	5014	1518	10360	9932
Model	**3.3.3.1.1.2 Difference from best AIC**					
linear	3394	3183	7560	3840	3164	4388
simple ZI	5	64	33	4	40	0
free ZI	0	0	0	0	0	0
type of ZI model	*ZIP*	*ZIP*	*ZINB*	*ZIP*	*ZINB*	*ZINB*

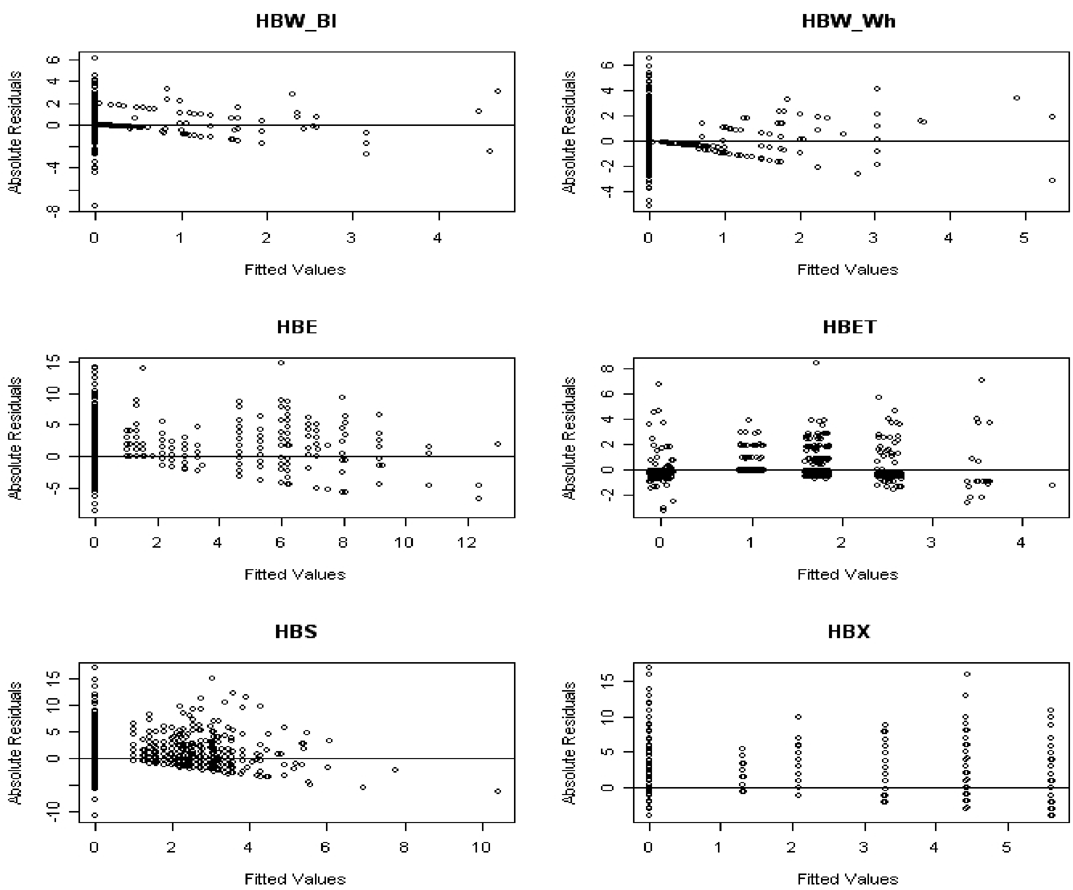

Figure 3: Absolute values versus the fitted values for the zero-inflated models.

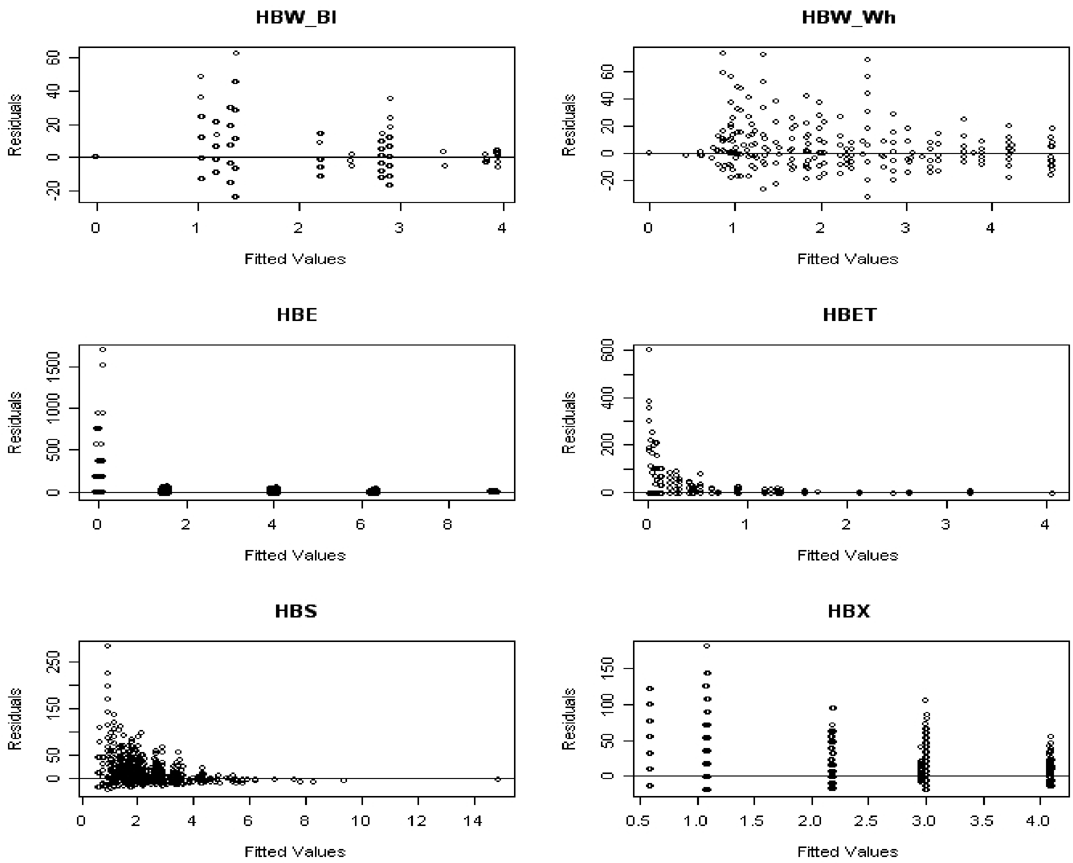

Figure 4: Standardised residuals versus the fitted values.

3.3.4 Performance

Figure 5 shows the predicted versus observed values for each trip purpose, with confidence intervals about the expected values. Unfortunately it was not possible to calculate the prediction intervals from the model output, although with extra time a routine could be written to simulate these[2]. The confidence intervals for the expected values for the zero-inflated models are actually slightly wider than those for the linear models. This is a more realistic picture of the error, and although the confidence intervals are slightly wider, due to structure the prediction intervals would be more realistic. They would not include the negative values of the linear models or the extremely large values of the log-linear model.

4 Results

In order to compare how effectively each regression technique can model the household data, we can compare the predicted distributions. That is, the number of zeros, ones, twos, etc that each model predicts. Given the inability of the log-linear model to 'fix' the problems with the linear regression, it is not discussed further in this section and only the linear and

[2] Prediction intervals can be estimated by drawing MC samples from a multivariate normal distribution centred on the MLE's of the parameters and variance-covariance matrix equal to the MLE's variance-covariance matrix. From the sampled set of parameters values predictions can be generated, and the prediction interval can be taken from the percentiles of a large number of these simulated observations. A personal communication with Simon Jackman, the author of the Pscl library, indicates that this may written into the existing functionality of the prediction routine in the near future.

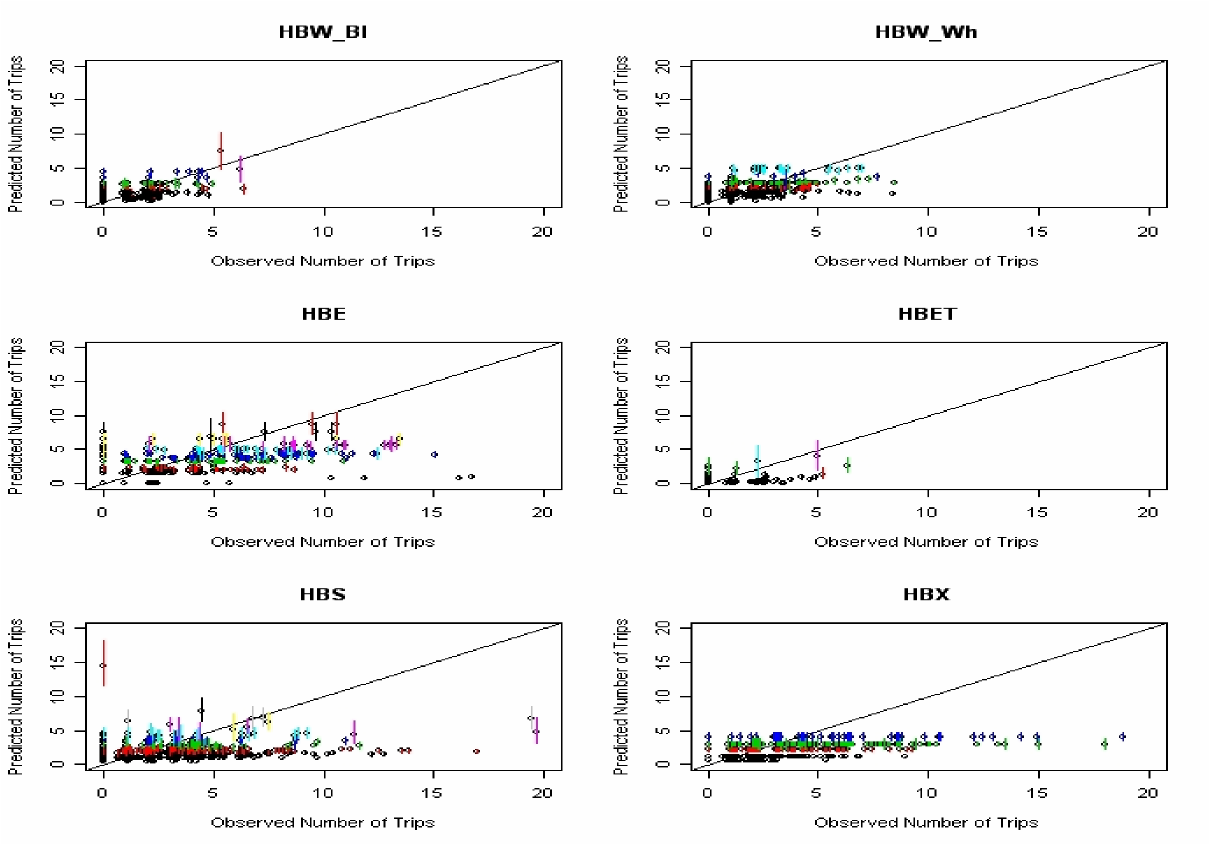

Figure 5: Predicted versus observed values with 95% confidence intervals for the expected values

zero-inflated models are compared. The values in table 4 were calculated using a hold-out/validation data set that was not used in the model calibration and shows the percentage of zeros predicted by the linear and zero-inflated models, as well as the observed percentage of zeros, and the percentage that are zero though have no applicable (NA) members of the household. Thus the percentage of NA's for blue collar trips is equal to the percentage of the households with no blue collar workers. The percentage of NA's for education are those household with no dependants 0-17, and the percentage of NA's for tertiary education are those households with no dependant 18-65 AND no white collar workers. However, although the data show that blue collar workers and dependants aged over 65 or under 18 are less likely to be involved in tertiary education, they are certainly not excluded. These results demonstrate that the linear model only captures absolute zeros resulting from the trip purpose not being applicable to the household, whereas the percentage of zeros in the zero-inflated model is a very close match to the observed values.

The following figure plots the percentage for each count, from each model, for each trip purpose. The percentage of the observed counts are marked with an 'O', the percentage of linear counts are marked with an 'L' and the percentage of zero-inflated counts are marked with a 'Z'. In all cases the zero-inflated model is a much closer fit to the observed counts than the linear model is. In some cases (such as the zero prediction for shopping and rec/other trips) there is a vast difference in the predictions.

5 Conclusions

This analysis has investigated stratified linear dummy regression for predicting trip rates for various purposes by household characteristics, and proposed zero-inflated regression techniques an alternative. Although stratified dummy linear regression can be used to

Table 4: Percentage of zeros predicted by the linear and zero-inflated models, to compare with the observed and 'trip not applicable to the household' percentage of zeros.

Purpose	Probability of a zero trip			
	NA	Linear	Observed	Zero Inflated
Blue	*0.70*	0.70	0.78	0.79
White	*0.42*	0.42	0.58	0.56
Edu	*0.65*	0.66	0.74	0.76
Edu Tert	*0.21?*	0.23	0.95	0.95
Shopping	*0.0*	0.0	0.41	0.46
Rec/Other	*0.0*	0.0	0.55	0.52

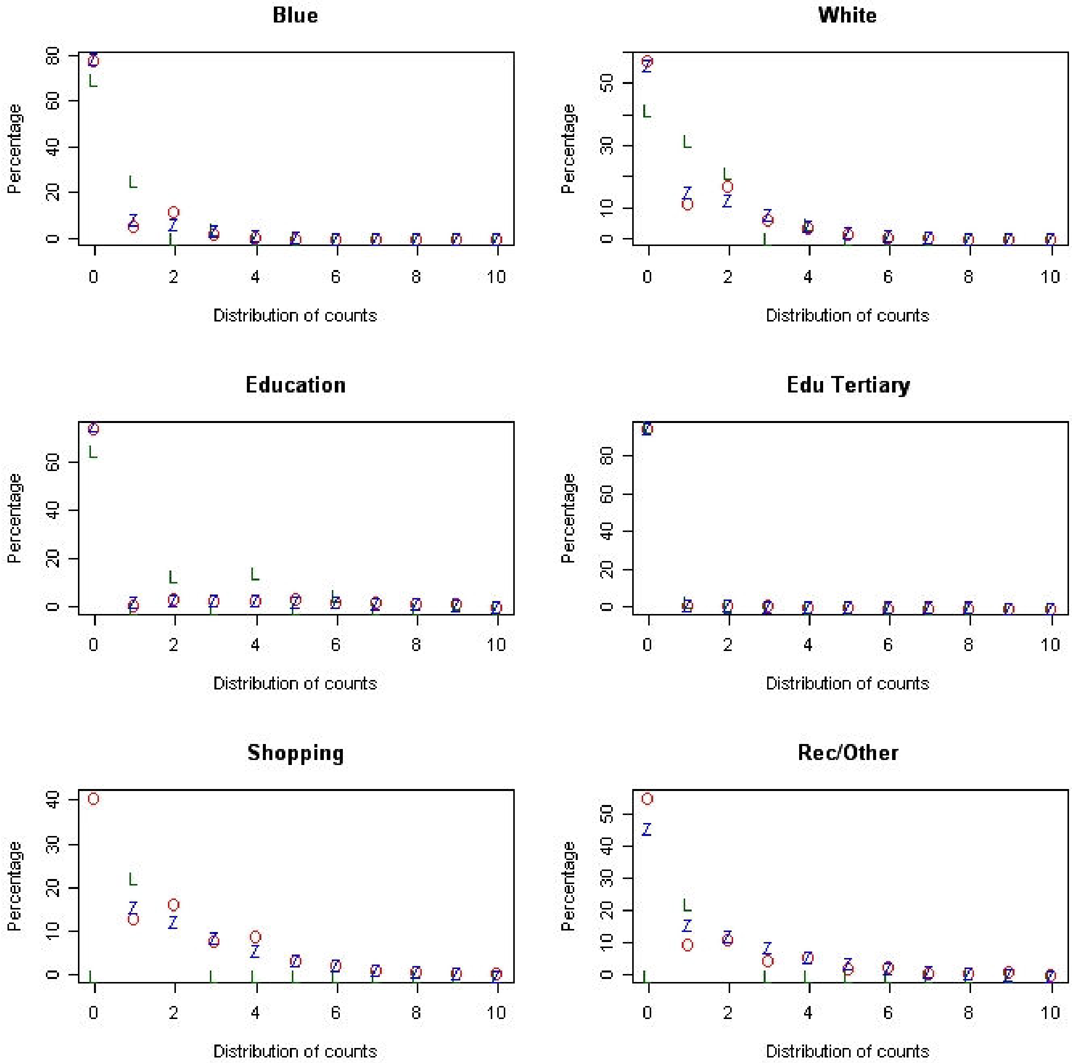

Figure 6: The distribution of counts predicted by the linear model (L), the zero-inflated model (Z) and the percentages observed in the data (O).

calculate averages, the measures of fit and performance given are not reliable and the model should not be interpreted as a description of the relationship between trip making and household characteristics. In contrast, the zero-inflated Poisson and negative binomial regression models relax the assumptions of linear regression and provide a much closer fit to the data.

The current formulation of the BSTM model uses aggregated census data in the linear regression equations to predict total trips for each zone (see the following equations). This is very simple and effective, but unfortunately not possible with the non-linear models presented here.

Linear Model Expected Value:

$$E\left(Y_i\right) = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots$$

Aggregated Expected Value:

$$\sum_{i=1}^{n} E\left(Y_i\right) = \sum_{i=1}^{n}\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots\right) = \beta_1 \sum_{i=1}^{n} x_{1i} + \beta_2 \sum_{i=1}^{n} x_{2i} + \beta_3 \sum_{i=1}^{n} x_{3i} + \ldots$$ Σ E(yi) = Σ (β1x1i +

ZIP Model Expected Value:

$$\begin{aligned} E\left(\boldsymbol{Y}\right) &= \left(1-\boldsymbol{p}\right)\boldsymbol{\lambda} \\ &= \frac{\exp\left(\beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \ldots\right)}{1+\exp\left(\gamma_1 \boldsymbol{z}_1 + \gamma_2 \boldsymbol{z}_2 + \gamma_3 \boldsymbol{z}_3 + \ldots\right)} \end{aligned}$$

Aggregated Expected Value:

$$\sum_{i=1}^{n} E\left(Y_i\right) = \sum_{i=1}^{n} \frac{\exp\left(\beta_1 \boldsymbol{x}_1 + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \ldots\right)}{1+\exp\left(\gamma_1 \boldsymbol{z}_1 + \gamma_2 \boldsymbol{z}_2 + \gamma_3 \boldsymbol{z}_3 + \ldots\right)}.$$

An alternative to using aggregated data is to simulate household data and calculate all predictions at the disaggregated level. Simulated data has been used in many other cities, for example, see studies in Sydney and Adelaide (Pointer, Stopher and Bullock (2004), Stopher, Rose and Bullock (2002)). There are three (possibly more) reasons why a move to a more comprehensive model would be justified:

1. The linear model provides mean estimates that are based on a large, random, heterogeneous sample of Brisbane.

    - Sampling theory predicts that the averages for small homogeneous zones will likely deviate from this overall mean.
    - With smaller zone sizes greater deviations from the means can be expected.
    - With more homogeneous zones greater deviations from the means can be expected.

2. In terms of traffic planning, an upper percentile of the predictive distribution may be more useful than the mean value. That is, it may be more desirable to plan traffic models around a ‘heavy’ traffic day than an ‘average’ day.

    - The linear model cannot produce useful upper or lower predictions, as it does not capture the underlying distribution of the data.

3. In terms of future planning, linear models may not be reliable in predicting future trip generation of future households if household structures change.

- For example, household structure was significant in predicting work trips when the zeros were accounted for, but the linear model cannot reliably detect this. Thus, if household structures change, the linear model will still predict the same number of trips.

Further comparison of these models could be conducted before moving to simulate household data. This analysis was constructed around a random split of the data into calibration and validation datasets. A better alternative would be to bootstrap the model using the collection districts as groups, so that each district could be independently predicted based on the other districts. There are 176 districts and approximately 20 households in each one, so the bootstrap simulation would need to run each model 176 times each time leaving out a different district. An alternative could be to use the 11 zones described in the survey documentation (The Urban Transport Institute, 2005). This would also facilitate the investigation of transferability of the trip generation equations to areas outside of the Brisbane Statistical Division. If the temporal or spatial transferability of models were a priority, then the use of the zero-inflated models within a Bayesian framework would also allow for the input of expert knowledge or past data.

Appendix

Appendix 1: Description of Dependent Variables

Dependent Variables	Code
Home Based Work_White	HBW_Wh
Home Based Work_Blue	HBW_Bl
Home Based Education_Total	HBE
Home Based Education_Tertiary	HBET
Home Based Shopping	HBS
Home Based Rec_Other	HBX

Appendix 2: Model output for zero-inflated models.

There a few cases where a variable is not significant at the '5%' level, but is included in the model nonetheless for practical reasons. That is, it is sensible to include all levels of a categorical variable where there is no intercept to account for a 'missing' level. Otherwise the model is forced to choose between an incomplete listing of categories. The alternative is to group levels within a category, but in some cases this is not advisable as it may affect the AIC values and comparison with the linear models. In most cases all variables are highly significant.

```
HBW-Bl – Blue Collar work trips

Total Log-likelihood: -1373.99911927419

 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error  z value  Pr(>|z|)
Wrkrs_Bl_0    20.566   570.7725  0.03603 9.713e-01
Wrkrs_Bl_1    -2.155     0.3035 -7.10200 1.230e-12
Wrkrs_Bl_2    -3.146     0.6916 -4.54879 5.396e-06
Wrkrs_Bl_3p   -1.606     0.6726 -2.38725 1.697e-02
-------------------------------------------------------------------

Count Model (Poisson)
Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
Wrkrs_Bl_1   0.42211    0.04895  8.6237 6.485e-18
```

```
Wrkrs_Bl_2   1.10526    0.05706 19.3687 1.418e-83
Wrkrs_Bl_3p  1.55744    0.12830 12.1394 6.533e-34
DepsA_1     -0.02817    0.07421 -0.3796 7.042e-01
DepsA_2     -0.27165    0.08517 -3.1893 1.426e-03
DepsA_3p    -0.14183    0.11344 -1.2503 2.112e-01
```

HBW_Wh – White Collar work trips

```
Total Log-likelihood: -2879.47718914046

 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error  z value  Pr(>|z|)
Wrkrs_Wh_0    19.566   474.9114   0.0412 9.671e-01
Wrkrs_Wh_1    -2.568     0.4021  -6.3871 1.690e-10
Wrkrs_Wh_2    -2.521     0.2412 -10.4503 1.461e-25
Wrkrs_Wh_3p   -3.052     0.5360  -5.6945 1.237e-08
---------------------------------------------------------------------

Count Model (Poisson)
Coefficients:
            Estimate Std. Error z value   Pr(>|z|)
Wrkrs_Wh_1   0.36271    0.04403  8.2375  1.758e-16
Wrkrs_Wh_2   1.01464    0.03249 31.2309 4.056e-214
Wrkrs_Wh_3p  1.48178    0.05406 27.4107 2.047e-165
DepsA_1     -0.24714    0.05340 -4.6281  3.691e-06
DepsA_2     -0.32941    0.05533 -5.9532  2.629e-09
DepsA_3p    -0.42608    0.07724 -5.5161  3.467e-08
DepsB_1      0.11155    0.04410  2.5297  1.142e-02
DepsB_2     -0.07714    0.08700 -0.8866  3.753e-01
DepsB_3p    -0.27366    0.23658 -1.1567  2.474e-01
DepsC_1     -0.10208    0.10727 -0.9516  3.413e-01
DepsC_2p    -0.80529    0.41695 -1.9314  5.344e-02
```

HBE – Eduction (preschool, primary & secondary trips)

```
Total Log-likelihood: -2521.04771254125

 Zero-Inflated Model was fit with a logit link
Coefficients:
         Estimate Std. Error z value  Pr(>|z|)
DepsA_0   3.42059     0.1638 20.8766 8.732e-97
DepsA_1  -0.06275     0.1115 -0.5628 5.736e-01
DepsA_2  -0.96883     0.1123 -8.6284 6.222e-18
DepsA_3  -2.00448     0.2428 -8.2542 1.530e-16
DepsA_4p -2.15187     0.5342 -4.0279 5.628e-05
---------------------------------------------------------------------

Count Model (Negative Binomial)
Coefficients:
           Estimate Std. Error z value   Pr(>|z|)
DepsA_1       1.092    0.05244   20.82  2.678e-96
DepsA_2       1.713    0.03193   53.64  0.000e+00
DepsA_3       1.965    0.04171   47.11  0.000e+00
DepsA_4p      2.310    0.07951   29.05 1.374e-185
log(theta)    2.165    0.17067   12.69  6.992e-37
```

HBET – Education Tertiary trips.

```
Total Log-likelihood: -743.288044231178

 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
DepsB_0       4.7861     0.2788  17.165 4.827e-66
DepsB_1       2.9304     0.2101  13.947 3.268e-44
```

```
DepsB_2        2.0550     0.2332   8.811 1.245e-18
DepsB_3        1.0459     0.4401   2.377 1.747e-02
DepsB_4p      -1.2114     1.2019  -1.008 3.135e-01
Wrkrs_Wh_1    -0.5543     0.2347  -2.362 1.816e-02
Wrkrs_Wh_2    -1.4740     0.2488  -5.925 3.129e-09
Wrkrs_Wh_3p   -2.1269     0.3145  -6.762 1.364e-11
Wrkrs_Bl_1p   -0.4349     0.1958  -2.221 2.635e-02
---------------------------------------------------------------------

Count Model (Poisson)
Coefficients:
         Estimate Std. Error z value  Pr(>|z|)
DepsB_0    0.2895     0.1815   1.595 1.107e-01
DepsB_1    0.5490     0.1014   5.416 6.088e-08
DepsB_2    0.9272     0.1102   8.413 3.985e-17
DepsB_3    1.2569     0.1686   7.455 8.956e-14
DepsB_4p   1.4350     0.2489   5.765 8.154e-09
```

HBS – Shopping trips

```
Total Log-likelihood: -5169.18773743085

 Zero-Inflated Model was fit with a logit link
Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
Wrkrs_Wh_1p  -0.6711    0.09013  -7.446 9.610e-14
DepsB_1p     -0.7382    0.10677  -6.914 4.707e-12
DepsC_1p     -0.6773    0.13841  -4.894 9.902e-07
---------------------------------------------------------------------

Count Model (Negative Binomial)
Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
Wrkrs_Bl_1    0.2290    0.04951   4.625 3.749e-06
Wrkrs_Bl_2p   0.4366    0.09133   4.781 1.744e-06
Wrkrs_Wh_1p   0.3396    0.03983   8.526 1.517e-17
DepsA_1       0.4157    0.06543   6.353 2.111e-10
DepsA_2       0.4819    0.06286   7.666 1.775e-14
DepsA_3p      0.6622    0.07676   8.626 6.355e-18
DepsB_1       0.4518    0.04577   9.871 5.548e-23
DepsB_2       0.9235    0.07088  13.029 8.324e-39
DepsB_3p      0.8808    0.17582   5.010 5.449e-07
DepsC_1       0.4455    0.06860   6.494 8.368e-11
DepsC_2p      1.1145    0.07408  15.045 3.717e-51
log(theta)    0.9823    0.10145   9.683 3.555e-22
```

HBX – Recreation and Other trips

```
Total Log-likelihood: -4952.83872812529

 Zero-Inflated Model was fit with a logit link
Coefficients:
        Estimate Std. Error z value  Pr(>|z|)
Pers_1   0.21638     0.1389   1.558 1.193e-01
Pers_2  -0.08907     0.0880  -1.012 3.114e-01
Pers_3  -0.67713     0.1224  -5.530 3.204e-08
Pers_4  -0.75368     0.1117  -6.749 1.491e-11
Pers_5  -0.98855     0.1699  -5.817 5.986e-09
Pers_6p -0.69392     0.2742  -2.530 1.139e-02
---------------------------------------------------------------------

Count Model (Negative Binomial)
Coefficients:
           Estimate Std. Error z value   Pr(>|z|)
Pers_1       0.2714    0.08999   3.016  2.564e-03
Pers_2       0.7336    0.04950  14.822  1.062e-49
```

```
Pers_3        1.1917    0.05307  22.454 1.178e-111
Pers_4        1.4869    0.04626  32.140 1.216e-226
Pers_5        1.7262    0.06335  27.249 1.704e-163
Pers_6p       1.4909    0.11784  12.651  1.101e-36
log(theta)    0.9259    0.10009   9.251  2.222e-20
```

6 References

Cheung, Y. B. (2002). "Zero-inflated models for regression analysis of count data: a study of growth and development." *Statistics in Medicine,* 1461-1469.

Heilbron, D. C. (1994). "Zero-altered and other regression models for count data with added zeros." *Biometrical Journal,* 36 531-547.

Hellerstein, D. M. (1991). "Using count data models in travel cost analysis with aggregate data." *American Journal of Agricultural Economics,* 73(3), 860-866.

Huet, S., Bouvier, A., Poursat, M., and Joliet, E. (2004). *Statistical tools for nonlinear regression: A practical guide with S-PLUS and R examples.* Springer-Verlag, New York.

Jackman, S. (2005). "PSCL: R in the Political Science Computational Laboratory." 0.51.

Lambert, D. (1992). "Zero-inflated Poisson regression, with an application to defects in manufacturing." *Technometrics,* 34(1), 1-14.

Lewsey, J. D., and Thomson, W. M. (2004). "The utility of the zero-inflated Poisson and zero-inflated negative binomial models: a case study of cross-sectional and longitudinal DMF data examining the effect of socio-economic status." *Community Dentistry and Oral Epidemiology,* 32 183-189.

Monzon, J., Goulias, K., and Kitamura, R. (1989). "Trip generation models for infrequent trips." *Transportation Research Record,* 1220 40-46.

Ortuzar, J. de D., and Willumsen, L. G. (1994). *Modelling Transport.* John Wiley and Sons, Inc, New York, USA.

Pointer, G., Stopher, P. R., and Bullock, P. (2004). "Monte Carlo simulation of Sydney Household Travel Survey data with Bayesian updating using different local sample sizes." *Rep. No. Working Paper ITS-WP-04-05,* Institute of Transport Studies (Sydney & Monash), Sydney, Australia.

R Core Development Team. (2005). "The R project for statistical computing." 2.1.0.

Said, G. M., and Young, D. H. (1990). "A general linear model framework for estimating work trip rates for households in Kuwait." *Transportation Research A,* 24A(3), 187-200.

Stopher, P., Rose, J., and Bullock, P. (2002). "Simulated Household Travel Survey Data: Synthetic Data in Australia." Working Paper ITS-WP-02-12.

The Urban Transport Institute. (2005). "Survey Procedures and Documentation - Brisbane Survey." I-view Pty Ltd, Data Analysis Australia, Queensland Government, Queensland Transport and Main Roads.

Washington, S. (2000). "Iteratively specified tree-based regression: theory and trip generation example." *Journal of Transportation Engineering,* 126(6), 482-491.