Statistics and Model Estimation

General Advice

All model estimations should be based on statistical procedures. As no such procedures are available through traditional transport packages, statistical packages should be used. The two widely used techniques are multiple regression (least squares) and maximum likelihood.

Statistics are the means of testing the fit to the data of a model proposed by the analyst and of comparing the fit of one model with another. If the analyst's ideas of a model specification are misplaced, then so will be the resulting statistically-estimated models.

Analysts should understand the capability of the observed data to support the estimation of models. If the data are insufficient to support the estimation of complex models, then attempts to estimate such models will inevitably give unsatisfactory outcomes.

The principle of parsimony applies to transport models - that is (quoting from an early GLIM manual): "Systematic effects should only be included in a model only if there is convincing evidence of the need for them. Time and effort should not be wasted interpreting effects which could just as well be random variation and our aim should always be to simplify the model as much as possible, consistent with the observed data."

Transport modelling of the type covered by this website started in the late 50s/early 60s. Since then, extremely clever researchers and modellers have explored how these models work, how best to estimate them and what are the most appropriate specifications. It follows that a very important first step in any modelling exercise is ensuring that this knowledge is available to the modelling team.

Car Ownership Models

There are so many different car ownership model specifications that I can offer no general rules. However, the use of a nonlinear maximum likelihood approach (with LIMDEP) in a specific example is described in Task 6.4 of the templates.

Trip Generation Models

For home-based trips, the standard procedure is to classify trip ends into productions (the home end of trips) and attractions (the non-home end of a trip).

Trip Production Models

Good practice is to estimate trip productions using individual household or person data from a household travel survey in a disaggregate estimation process.

Regression of household or person trip rates against household and person characteristics and location parameters is good practice. There is much evidence on the most appropriate specifications.

For multiple regression to achieve the best unbiased estimators of the model coefficients and to correspond to the maximum likelihood values, the data are required to meet a number of conditions, that the residual errors of the model are:

- normally distributed,
- with mean zero,
- constant variance across all observations (no heterocedasticity),
- and zero covariance across all observations (ie no correlations between the residuals, or multicollinearity).

Unfortunately these conditions are not usually met precisely. In theory, the sample data are counts for which a Poisson distribution might be more appropriate (although this turns out not to be the case) and because many trips are in pairs, the covariance condition is not met.

In practice, the data are assumed not to break these rules by much and regression is used. But it is probably worthwhile checking these conditions for any models developed and particularly ensuring that there are no outliers (as might be expected from a Poisson error distribution) which may upset the estimation process.

In determining the specifications to be tested and finalising the models be aware:

- that there is questionable value in including variables to which the forecasts would be largely insensitive, and
- that the more complex and extensive specifications will demand much more in terms of family structure modelling.

Trip Attraction Models

Trip attraction models are usually estimated on aggregate data with multiple regression techniques. The process fails most statistical requirements:

- the residuals are correlated with the independent variables, being related to zone size and to survey sample size, that is heterocedasticity is an issue;
- they do not accord with the normal distribution assumption but more closely relate to a Poisson distribution;
- there is multicollinearity in the independent aggregate variables, due to the relationships with the size of the zone and between employment types;
- the range of aggregate observations is the such that the estimation is highly affected by data outliers and high leverage points.

A further problem is that the observed estimates of trip attractions are subject to significant measurement errors - for example, it is usual to find that address coding inaccuracies are focused on the non-home end of trips, for which survey respondents typically are able to give less precise address information than for their home residence. Consequently, the allocation of trip ends from travel surveys to attraction zones is often subject to many approximations.

It follows that the estimation of trip attraction models should be approached with considerable caution. For suggestions, see Task 6.4 of the templates.

Parsimony and T-tests

In regression modelling, t-tests are usually used to determine the significance of a model coefficient, a t-value of 1.96 or more signifying 95% significance. This test confirms that we have 95% confidence that the coefficient is different from zero.

But for parsimony, we also generally want model coefficients to be significantly different from one another. That is we want trip rates for different household or person types to be distinguished only if they are significantly different, coefficients sub-categories of employment in an attraction model to be distinguished only if they are significantly different from each other.

In both cases, the unnecessary inclusion of additional explanatory variables further complicates the modelling, requiring a more complex family structure model in the first case or the forward projection of extra sub-categories of employment in the second.

The t-test offers a convenient way of testing the significance of the difference between coefficients of a regression model.

Trip Distribution and Mode Choice Model

This is the most complex estimation exercise requiring specialised software, and as a result there appear to be many poorly estimated models of this sort developed on inadequate estimation packages.

These models should be estimated (together if possible) as hierarchical (or nested) logit models with a logsum interface, that is the distribution model is treated as a destination choice decision in a hierarchy with the mode decision.

The hierarchy must meet nested model requirements to avoid irrational forecasts.

For aggregate models, and aggregate estimation, if expanded data is used then the statistics should be corrected (t-statistics should be divided by the square root of the average expansion factor).

In such estimations, there is rarely enough data to estimate the parameters of generalised cost (which combine the various components of cost and time into an overall generalised cost) - it is therefore far safer to draw on local and international evidence to pre-specify the generalised cost weights. The estimation can then focus on establishing the sensitivity of mode and destination choices to overall generalised cost.

Aggregate mode choice models are difficult to estimate on household travel survey data in cities for which the overall public transport shares are low - there are too few observed public transport trips. In such circumstances, it is therefore advisable to extend the survey programme to increase the sample of public transport trips, perhaps through a public transport passenger intercept survey.

This may be illustrated by the data from Auckland. The trip samples used for the distribution and mode choice calibration were 51,000 car trips and 24,000 public

transport trips for 6 segments implying on average 8,500 car trips and 4,000 public transport trips per segment[1].

Had we not collected additional public transport data through an intercept survey, we would have had to base the calibrations on the data derived from the household travel survey, around 2,400 trips or 400 trips per segment on average. For an aggregate calibration exercise, such a small sample of public transport trips would not permit any precision in the estimation of the mode choice model (ie the trade-off between car and public transport) or the distribution model (ie the variation of public transport trips by trip length and location).

The public transport intercept survey increased the public transport trip sample tenfold and this facilitated the successful model calibration.

[1] Of course, the trips are not distributed uniformly across the segments.