Issues in the Analysis of Stated Preference Data

John Polak

Head of Transport Studies

Imperial College, London

<j.polak@ic.ac.uk>

Outline

- General approach to the analysis of SP data
- SP data and discrete choice modelling
- Scaling and data enrichment strategies
- Repeated measures issues
- Dealing with heterogeneity
- Future challenges
- Conclusions

General approach/1

- Relate the attribute values of hypothetical travel alternatives to the responses obtained

$$R_i = F(\mathbf{X}_i, \boldsymbol{\beta})$$

- The objective is to determine an appropriate functional form for F and appropriate values of $\boldsymbol{\beta}$
- Typically F is assumed to be linear in $\boldsymbol{\beta}$ but this is not necessary
- The response may be a rating, a ranking or a choice
- The details of the analysis depend on the type of response

General approach/2

- For rating responses, simple linear regression models are often used

$$R_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_n X_{ni} + \varepsilon_i$$

- For ranking responses, we can use multivariate analysis of variance techniques to find that values of $\boldsymbol{\beta}$ that best reproduces the reported ranks
- For choice responses, we can use logit or probit models to model the discrete outcomes
- The relative values of $\boldsymbol{\beta}$ tell us about how people trade off between different attributes

SP data and discrete choice models/1

- Although there is no *necessary* link between SP data and discrete choice/random utility models, the discrete choice framework of analysis is especially important because:
  - choice oriented tasks have a number of advantages over rating and ranking tasks and are now the favoured data collection approach
  - it is very flexible (e.g., does not require strict orthogonality in the SP design)
  - it enables links to be established between SP data and other aspects of travel demand modelling
- However, it is important to understand what assumptions are being made, and their limitations

SP data and discrete choice models/2

- The utility (U_i) of a hypothetical travel alternative is composed of an observable component (V_i) and an error (ε_l)

$$U_i = V_i + \varepsilon_i = \quad {}_j \beta_j X_{ij} + \varepsilon_i$$

- Alternative *i* is chosen in comparison to alternative *k* if

$$U_i > U_k$$

$$V_i + \varepsilon_i > V_k + \varepsilon_k$$

$$\varepsilon_i - \varepsilon_k > V_k - V_i$$

- Hence choice behaviour depends the magnitude of the errors as well as the magnitude of the *V*s.

SP data and discrete choice models/3

- Under standard assumptions this leads to the familiar logit discrete choice model

$$\Pr(i) = \frac{e^{\lambda V_i}}{\sum_k e^{\lambda V_k}}$$

- The parameter λ is inversely related to Var(ε)
- With RP data it is conventional to assume that $\lambda = 1$ (this fixes the scale of the errors), but what about SP?
- The value of λ does not affect the relative values of the βs (i.e., values of time) but it does affect elasticities and hence demand forecasts

Scaling and enrichment/1

- This ambiguity of scale is a serious issue in the use of SP data for forecasting
- One approach to this problem is to combine SP and RP data in model development. This is called scaling or data enrichment
- The idea is to use RP data to help fix the scale of the SP errors
- It requires RP and SP data that ‘overlap’ in at least some choice alternatives and some choice attributes

Scaling and enrichment/2

- A typical formulation would have overlapping attributes ***X*** and corresponding parameters β and RP and SP specific errors:

$$U_{SP} = V_{SP} + \varepsilon_{SP} = \quad {}_j \beta_j X_j + \quad {}_j \gamma_j Y_j + \varepsilon_{SP}$$

$$U_{RP} = V_{RP} + \varepsilon_{RP} = \quad {}_j \beta_j X_j + \quad {}_j \phi_j Z_j + \varepsilon_{RP}$$

- We assume that $\mathrm{Var}(\varepsilon_{RP}) = \mu^2 \mathrm{Var}(\varepsilon_{SP})$ and estimate μ as well as β, γ and ϕ
- Although more complex, this type of estimation can be easily accomplished with standard software

Scaling and enrichment/3

- Empirical experience suggests that often $\mu > 1$ (i.e., $Var(\varepsilon_{RP}) > Var(\varepsilon_{SP})$) hence that in these cases unadjusted SP models may over estimate the sensitivity of the market
- Essentially the same approach can be used to:
  - combine different SP data sets
  - combine SP and attitudinal data
  - investigate the effect of respondent fatigue and other methodological issues

Repeated measurement/1

- It is conventional to assume that the observations collected in SP exercises are all independent
- In fact however, we typically make a series repeated observations on the same individual
- The data therefore are hierarchical, with observations nested within individuals
- This effectively reduces the amount of information in the data
- We must take this into account in model estimation

Repeated measurement/2

- The principal effect of the repeated measurements taken on an individual will be to induce a structure in the error term:

$$U_{ik} = V_{ik} + \nu_i + \varepsilon_{ik}$$

where ν_I is a component of error associated with person *i* and ε_{ik} is the component of error associated with person *i* in observation *k*

- These models are more complicated to estimate but the capability is beginning to appear in standard software

Heterogeneity/1

- Individuals typically vary in their experiences, preferences and responses
- We can accommodate these inter-personal variations in a number of ways in SP exercises
- At the design stage, by customising the design to reflect the experiences of particular respondents
- At the modelling stage, by allowing parameters to vary amongst individuals
  - deterministically via segmentation
  - stochastically via random coefficients models

Heterogeneity/2

- In segmentation approaches we allow the value of a model parameter to vary across different categories of one or more classification variables

$$\beta \rightarrow \beta_0 + \beta_i \delta_i$$

where β_0 is the value of the parameter for observations in the reference category and β_i is the marginal effect associated with the *i*th category (δ_i=1 if observation is in the *i*th category, 0 otherwise)

- Segmentation is very easy to implement, but depends upon the apriori determination of suitable segmentation categories

Heterogeneity/3

- In random coefficients models, we replace a single deterministic value of β with a probability distribution of values across the population

$$\beta \rightarrow f(\beta, \theta)$$

- The objective is then to estimate the functional form of *f* and the unknown parameters θ of the distribution
- For reasons of convenience *f* is usually assumed to be one of the standard distributions such as gamma or log normal
- The estimation of random coefficients models is more complex, but suitable software is available

Future challenges for analysis

- Accommodating non-compensatory decision mechanisms
- Diagnosing and treating β non-response and non-informative response
- Accommodating uncertainty
- Extending SP to deal with long term decisions

Conclusions

- SP techniques are now a key component of the transport analysts toolbox β
- However, to get the best out of them requires careful and rigorous analysis
- Significant advances in analysis methods and tools have been made during the past decade
- These tools are becoming more widely available
- Further advances, especially in econometric estimation can be expected in the future