ECONOMETRIC ISSUES IN
STATED PREFERENCE ANALYSIS

By John Bates*

INTRODUCTION
It is a commonplace in transport modelling, as in other forms of science, that the
theoretical justification of a method appears only after the method itself has been
in use for a number of years. Well-known examples include the "gravity" model
of trip distribution and the (not entirely unrelated) logit model of mode choice:
the popularity of both these models was initially due to their relative simplicity
and reasonable success in reproducing observed patterns, rather than to the
elegant rationales produced for them by Wilson (1969) and McFadden (1973).
Likewise, the basic ideas of Stated Preference, that the apparent weights
placed on different "product" attributes can be derived on the basis of responses
to hypothetical questions relating to preference, have been around for a long
time, and were widely used within the field of market research long before the
current interest in their applications to transport. However, as yet the theoretical
basis is still relatively weak in one crucial point – the appropriate error structure
for models based on stated preference data.

As is pointed out elsewhere in this issue (see the article by Hensher et al.), it
seems generally appropriate to view models based on SP data as falling within
the overall scope of random utility models – that is, models which attempt to
construct a suitable formula for utility but which allow for random effects. But
there has been little discussion in the literature of the source and nature of these
random effects. The object of this paper is to extend the discussion in an earlier
paper (Bates, 1984), and to pose a number of questions which need to be
answered to increase confidence in the technique of stated preference.

The models that have been reported in SP work are in general fairly straight-
forward. For a start, most SP designs contain a small number of variables
(typically between three and six, though much larger designs have been reported),
and these variables are typically limited to two or three different values. This
precludes extensive model development: the basic candidate variables are clearly
defined, and the small number of possible values makes it unlikely that anything

* Consultant in Transport Economics, Gillingham, Dorset. I am most grateful to other contri-
bu tors in this issue, all of whom made useful comments on an earlier draft of this paper. I
hope that the resulting paper is substantially improved.
departing far from linear-in-parameter formulations will be practicable. Such complexity as does exist is normally introduced through:

(a) limited transformations of individual design variables (quadratic effects, etc.);
(b) interactions between design variables;
(c) the interaction of design variables with socio-economic factors.

Given the relatively unambitious model formulation found in most SP studies, it is perhaps not surprising that the ratio of parameters has been shown not to be very sensitive to the assumptions about error structure. For example, Bates and Roberts (1983) report that "rating" data analysed on the basis of a number of different assumptions (for example, converting to different probability scales, or moving between normal and binomial error structures) gave very similar results for coefficient ratios such as the value of time; and Kroes and Sheldon (in this issue) claim that ranked data treated by MONANOVA give similar results to more complicated treatments (like the exploded logit model of Chapman and Staelin, 1982). When the aim of an SP study is primarily to evaluate certain attributes in terms of others (equivalent to estimating marginal rates of substitution), explicit specification of the errors may well be fairly unimportant.

However, as will be explained in the main part of this paper, this is emphatically not true if the aim is to make forecasts of demand (and, as Louviere and Woodworth (1983) note, this is a primary requirement in many studies). In order to forecast demand we need to predict what choices will actually be made in the market, and this requires knowledge of the relative scales of the fixed and random components of utility. To understand this point more clearly, it is helpful to return to the basic tenets of random utility theory.

OUTLINE OF RANDOM UTILITY THEORY

The fundamental axiom of utility theory is that alternative \( i \) will be preferred to alternative \( j \) if \( U_i > U_j \). To convert preference into choice requires certain enabling factors, such as income, to be taken into account. It is now widely understood and accepted that the kind of utility that is dealt with in discrete choice theory is indirect utility (see, for example, Deaton and Muellbauer, 1980) — that is, the measure of utility internalises the constraints arising from income and other sources. Some further discussion is provided in the paper by Hensher et al. in this issue.

The discrete choice paradigm can therefore postulate that, if alternative \( i \) is chosen from a set of available alternatives \( S \), then it necessarily follows that

\[
U_i > U_j, \quad j \neq i \quad \& \quad j \in S.
\]

But, the argument goes, the modeller can only calculate a part of what goes into making up the utility \( U_i \): there is a random element which cannot be determined. Thus the basic random utility formulation is that

\[
U_i = U(V_i, \epsilon_i)
\]

which is commonly simplified to the additive formulation

60
ECONOMETRIC ISSUES IN STATED PREFERENCE ANALYSIS

\[ U_i = V_i + \epsilon_i. \] (1a)

and, as a further simplification, \( V \) is specified as a linear-in-parameters function of a coefficient vector \( b \) and a vector of explanatory variables.

In relation to a typical Revealed Preference problem, the modeller's interest is in defining suitable forms for \( V_i \) (the "deterministic" part of the utility function) in terms of coefficients which, when estimated by an appropriate statistical technique, explain as many of the observed choices as possible.

Let us first consider the estimation of coefficients, given that we know the choices made: thus, for a given individual, we know, say, that \( U_i \) is greater than all other \( U_j \). However, this does not imply that \( V_i \) must be greater than all other \( V_j \), because we need to take account of the size of the random terms \( \epsilon_i \), etc. We do not know the size of these terms as they vary across alternatives and individuals. The best we can do is to assume a distribution for them; then, by integrating over the distribution, we obtain, for any given formula for \( V_i \), the probability that \( U_i \) is greater than all \( U_j \). The standard estimation technique for this kind of problem, that of Maximum Likelihood, then estimates that set of coefficients which, when inserted into the formula \( V \), maximises the joint probability across all the observations of the choices actually made. On the assumption of a distribution for \( \epsilon_i \), we allow the relative scale of \( V_i \) and \( \epsilon_i \) to be determined.

The forecasting problem is effectively the reverse: now we assume that we know \( V_i \) and we need to predict \( U_i \). Again, we cannot do this without allowing for the random terms. Once more, however, we can make assumptions about the distribution of \( \epsilon_i \), and hence calculate the probability that a particular alternative \( i \) will in fact be chosen; with suitable qualifications, this can be considered as the demand function for alternative \( i \). In the case of Revealed Preference data, it appears generally reasonable to assume that the distribution of the random terms used for the estimation is identical to that which should be used for forecasting, though it is certainly possible to imagine situations where this would not be valid.

Clearly, the form of the demand function depends critically on the assumptions about the random terms. By far the best known form of discrete choice model, the multinomial logit (MNL), depends on a particular, rather restrictive, error structure that has been widely discussed (for example, by Ben-Akiva and Lerman, 1985). Theoretically there is no need to restrict ourselves to the MNL error structure, but in practice the estimation problems quickly become severe as we depart from it.

Two assumptions which are central to most standard forms of estimation are that the variance of the error term is constant (homoscedasticity) and that the error terms are independent. In the discrete choice model, these assumptions typically apply both within an individual respondent across the set of alternatives, and between respondents.

Though serious computational problems may be encountered, it is in principle possible, given any error structure, to use the Maximum Likelihood approach to obtain the optimum coefficient vector \( b^* \) according to standard statistical procedures. The advantage of this approach is that useful statistical indicators are available to assess the overall fit of the model for \( V \) and the significance of individual coefficients, etc.
This approach has been discussed at great length in the literature. However, the source of the random terms has received very much less discussion, and in general, while the classic econometric textbooks pay considerable attention to the random elements which provide the basis of all estimation, the average practitioner is content to take a more cavalier approach. There appear to be at least three main sources of “error” in the standard discrete choice approach:

(a) *The “unobservable” factors which affect choice.* These may be specific to the individual, representing prejudices in favour of certain alternatives.

(b) *Measurement error in the explanatory variables entering the function for* $V$. In the transport field common examples are the calculations of journey times based on coarse zoning systems, and the rounding of reported times to the nearest five minutes.

(c) *Model specification errors.* These arise not only in the decision on which variables should enter the formula for $V$, but also in the way in which they are entered.

All these sources are conveniently assumed to be subsumed in the single additive element $\varepsilon_i$, with assumed known distribution.

Of course, in practice the distinctions between these sources of error are not so clear-cut as is being suggested here. We define the unobservable factors as those elements which either cannot be identified, or, where they can be identified, cannot be measured. Even after we have identified all the measurable factors which could reasonably be expected to influence choice between a given set of alternatives, there will be additional, idiosyncratic, factors which enter the utility calculation $U$. But to make such a statement implies that the analyst has succeeded in identifying all relevant measurable effects, and has measured them without error. Since such a state of omniscience is unlikely, there are reasonable practical grounds for conflating the three sources of error discussed above.

The usual approach in this kind of modelling is to assume either that the same utility function applies to all respondents in the sample, or that variations in “tastes” can be dealt with by means of dummy variables relating to identifiable market segments, possibly interacting with other explanatory variables in the utility function. In the MNL model no allowance is made for “random taste variation” — that is, the stochastic variation in the coefficients of the utility function. The error term thus effectively reflects the variation between individuals, as far as it can be captured by an additive effect, with respect to the model formulation $V$ and the measurement of the variables which compose $V$.

A recent paper by McFadden (1986) attempts to make the choice process more explicit, by dealing with “latent variables” which can take account of underlying perceptions and attitudes, and by proposing a series of equations which can be formulated as a LISREL model (Joreskog, 1973). This is an important step towards identifying the sources of error, and potentially allowing them to be treated in different ways. Needless to say, it is in practice usually necessary to make quite severe simplifications in the error structure to obtain coefficient estimates.
ECONOMETRIC ISSUES IN STATED PREFERENCE ANALYSIS

J. Bates

Application to SP data

One of the main features of the SP approach is the existence of repeated measurements, or "treatments": each respondent contributes a number of observations, typically between 8 and 16. As is explained elsewhere in this issue (see, for example, the paper by Louviere), there are three main types of responses: "rating", "ranking" and "choice". If we assume that in any case there are $N$ treatments, the response data can be described as follows:

(a) Rating
For each of $N$ treatments, a response is given on a numeric or semantic scale, which can be transformed to a utility scale by making further assumptions; this gives a measure $U_i$ which can be related to the values of the variables for treatment $i$. In some examples of this approach, the rating consists of a "relative preference" for one of two alternatives; in that case the response variable is more correctly interpreted as $\Delta U_i$ (see Bates (1984) for further discussion).

(b) Ranking
In the ranking approach, a treatment corresponds to an alternative. The $N$ alternatives are ranked in order of preference. If $r_1$ denotes the alternative which is ranked highest, and $r_2$ the alternative ranked second, and so on, the response implies that

$$U_{r_1} > U_{r_2} > \ldots > U_{r_N}$$

(c) Choice
Here, a treatment refers to a choice set, and the respondent is merely asked to select his preferred option from the alternatives in the set. If $S_i$ is the choice set for the $i$th treatment, and $c_i$ denotes the alternative chosen, then the response implies that

$$U_{c_i} > U_j, j \neq c_i, j \in S_i$$

This latter condition corresponds with the usual discrete choice (Revealed Preference) approach, except that both the alternatives and the responses are hypothetical.

The analysis of such data is in principle the same as the analysis of Revealed Preference data — the observed responses, as reflected in the measures $U_i$, are assumed to be explained by a model for $V$ together with a random effect. The data can either be analysed separately for each individual (subject to adequate degrees of freedom) or pooled over a number of respondents. We will begin with the case of individual analysis.

Analysis of individual data

Of the three sources of error mentioned in the RP case, neither (a), unobservables, nor (b), measurement, is in general appropriate to SP. There is no measurement error, since the values are presented to the respondent (there may, nevertheless, be some problems of perception). Further, since the alternatives are in general
fully represented by their attributes (in "abstract"), there is no reason to make allowance for unobservables. An exception to this occurs when respondents are asked to express a preference between alternatives with identifiable properties (for example, between bus and rail), where choice may depend on variables not included in the design. Even here, however, it is generally possible, with a number of treatments, to deal with this by means of, say, a modal constant, which does not vary across treatments.

Source (c), specification, on the other hand, clearly does still apply, though for a single individual there is no reason to expect random taste variation. Although in theory designs can be drawn up which allow a useful number of alternative model forms to be tested, the practical limitations of administering the design tend to counteract this, and it is arguable whether the scope for model specification is in practice any greater with SP data than with RP.

With SP, there would seem to be a further serious source of error, and that relates to the response variable itself. As is made clear by the contributors to this issue, the practical results of carrying out SP surveys and analysis are generally encouraging, and suggest that the great majority of respondents do understand what is expected of them. This is a far cry, however, from suggesting that they are able to carry out the task with complete accuracy. As McFadden (1986, p. 289) writes:

Another issue is the stability of elicited preferences over the sequences of task performed by each subject. Factors such as learning, boredom, or anchoring to earlier tasks may distort the measurement of preferences, and cast doubt on the cognitive congruence of the time frames in which experimental versus market decisions are made.

"Respondent fatigue" has been confirmed in some studies. This possibility is typically dealt with by randomising the order of the treatments between respondents; but that of course has no useful effect unless the responses from several people are pooled.

It is also possible, of course, that respondents may deliberately give biased responses, in the hope of affecting the outcome of the analysis (policy-response bias) or of casting their existing behaviour in a better light (justification bias). These and other forms of bias response have been classified by Bonsall (1985). SP researchers are usually careful to avoid presenting the most obvious invitations for such biases, but they may not always be successful.

The upshot is that we have measurement error in the dependent variable: we are not getting a true estimate of $U$, but rather some pseudo-utility which we may call $\hat{U}$, where the general linking framework is

$$U_i = V_i + \varepsilon_i = \hat{U}_i + \eta_i$$

(2)

It is quite possible that the variance of $\eta_i$ may vary across treatments — for example, it may increase (response fatigue) or it may decrease, as the respondent "learns" how to perform the task. In this discussion, however, for simplicity's sake we will assume homoscedasticity, as is typically assumed for the other error term $\varepsilon_i$.

Now the revised formulation for measurement error presents no estimation
problems in terms of the coefficients in $V$, since the equation is easily rewritten as

$$\hat{U}_i = V_i + (e_i - \eta_i)$$

and, assuming that the error term can be conflated in this way, the normal methods, as explained for the Revealed Preference case, apply. The problem comes in forecasting, where we are interested in making estimates of $U$. If we now assume that the distribution of error appropriate to estimation applies to forecasting, we will be making estimates of the pseudo-utility $\hat{U}$ rather than of the “true” utility $U$; in other words, we are making estimates of relative preferences as expressed in a Stated Preference experiment rather than of what would occur in the market. The only way to get round this is to apportion the error between $e_i$ and $\eta_i$.

It must be stressed that this is so, whether we use “choice” data or “preference” data (see the discussion elsewhere in this issue). It may be true (as has been argued for example, by Louviere in this issue) that the error is less in the case of choice data, because the task is easier, but it is still likely that some error will be present. We should therefore amend the formulations given above for the three kinds of SP response data by substituting $\hat{U}$ for $U$ throughout. For example, in the case of choice data we are in reality observing

$$\hat{U}_{c_i} > \hat{U}_j, j \neq c_i, j \in S_i$$

or

$$U_{c_i} - \eta_{c_i} > U_j - \eta_i, j \neq c_i, j \in S_i$$

which does not in general imply $U_{c_i} > U_j$.

The simplest case is when $e_i$, which we are using to relate to the kind of error which is compatible with models fitted to RP data, and $\eta_i$, which relates to the inability of the respondent to reply to the SP exercise in a way which corresponds with his actual behaviour, are independently distributed with the same type of distribution, and differ only in their variances. Suppose the variances are $\sigma_e^2$ and $\sigma_\eta^2$ respectively. Then the variance of the random component used in estimation ($\sigma_e^2 + \sigma_\eta^2$) is greater than that which would be used for forecasting. The practical result is that the scale of the coefficients in $V$ relative to the random terms has altered, and this will affect the demand function when we integrate out across the error distribution.

An understanding of the magnitude of $\eta$ is thus of crucial importance to the use of SP in forecasting. Only if $\eta$ is insignificant relative to $e$ can the estimated model be used directly to give forecasts: in all other cases, some kind of scaling of the coefficients of $V$ relative to the random terms is required, and in general the knowledge of how to do this is lacking. In practice, users of the technique have been content either to assume that the alternative with highest ‘deterministic utility’ ($V$) will be chosen, implying that the size of the error term is small relative to the modelled difference in $V$ between alternatives, or that the size of the random term for predicting behaviour is identical with that relating to the estim-

---

1 It is not, of course, implied that this is a general result.
ation. As the above discussion shows, these assumptions cannot be justified theoretically.

Before concluding this discussion, let us return to the questions of independence and constant variance for the error terms. We have already referred in passing to possible sources for heteroscedasticity, but a well-designed experiment should be able to avoid the grosser effects. Are there, however, reasons for expecting non-independence between the error terms? Except for the case of biased response (a respondent may deliberately downgrade alternatives involving a tolled road, for instance, in order to affect policy), it is difficult to see why this should be so. In general, it seems safe to conclude that the analysis of the responses of a given individual to a well-conceived experiment does not present serious statistical problems (apart from those associated with low degrees of freedom), even though the forecasting issues remain severe.

Analysis of pooled data

Thus far, the discussion has related to models calibrated on the treatments for a single individual. We must now expand it to the more general case where the same model is deemed to apply to a group of individuals.

Returning to our general statement of theory that we have

\[
\hat{U}_t = V_I + (e_t - \eta_t)
\]  

(3)

The question turns on how the magnitudes of \(e_t\) and \(\eta_t\) are likely to be affected. It seems reasonable to expect that the variance of \(e_t\), which we have suggested in the case of SP is largely related to model specification, will increase, since we are now imposing the further constraint that the same utility function must apply to all individuals in the group (though it should be noted that the ability of SP data to assist with the analysis of "taste variation" to some extent mitigates this problem). It also seems likely that the variance of \(\eta_t\) will differ between individuals, as some individuals find the task easier than others. At the least, then, we should expect some kind of "structured heteroscedasticity", with the variance of part of the error term varying from one individual to the next, in a largely random way. Thus the estimation itself becomes more complicated, quite apart from the question of forecasting.

Hitherto, the nature of this complication does not seem to have been taken into account in analysis, and standard error assumptions have been used. Louviere and Woodworth (1983) have drawn attention to the problem of "repeated measurements", and suggested that the result of ignoring it is likely to be an underestimation of the standard errors of the parameters. They suggest that, on a conservative approach, the estimated variance matrix should be multiplied by the number of treatments.

This is a reasonable approach as far as it goes; it is almost certainly an over-correction. Informally, the reasoning appears to be that the treatments do not constitute separate pieces of information, and should not therefore contribute to the calculation of degrees of freedom. However, a more rigorous justification would seem to imply that the error terms pertaining to the treatments for any given individual are completely correlated, and this does not seem at all likely.
ECONOMETRIC ISSUES IN STATED PREFERENCE ANALYSIS

J. Bates

It seems much more likely that the variance of the error terms differs between individuals, as already explained. The general principles of estimation suggest that the consequence will be a decrease in the efficiency of the estimation of the coefficients, but without leading to bias — this provides some consolation in relation to estimation.

However, the only reliable method of dealing with this problem is to adopt an estimation method that explicitly allows for the hypothesised error structure. Similar conclusions are reached by Davies and Pickles (1985) in the context of longitudinal data analysis: they conclude that, though the approach raises quite complicated technical questions, these are unlikely to remain an impediment for long, and that appropriate macros for packages like GLIM are already becoming available. It would seem important that SP researchers keep abreast of such developments. Nevertheless, though these methods would give more realistic estimates of the standard errors, they cannot provide information on the partitioning between $\varepsilon$ and $\eta$: for that, additional data are needed.

DISCUSSION

The preceding arguments suggest that the statistical treatment of SP data remains rather simplistic, and that there are particular problems for forecasting. On the assumption that the remedies for these deficiencies are not going to be immediately available, we present in this section some interim solutions which may do much to alleviate the problems.

In the first place, any potential biases involved in current methods of analysis could be fairly readily quantified by some careful simulation work. This would involve specifying random errors from a number of sources along the lines discussed, and generating response data which would then be analysed by current methods. This should provide an understanding of the importance or otherwise of various sources of error, along the lines of the work done by, for instance, Horowitz (1982) for discrete choice models. From this, it might be concluded that relatively straightforward variants on current methods, while not being "exact", were sufficiently accurate for most purposes.

The second obvious course of action is to attempt to remove the sources of variation in the error term, as far as possible, by judicious model specification. We will deal with two possible cases. In the first case, there is a considerable literature on taste variation — that is, interpersonal variations in the utility function. Where this taste variation is random, serious problems may be raised by the assumption that the random element is purely additive. However, much of the variation may in fact be related to measurable characteristics (such as income, age and sex). Introducing these characteristics into the utility function in an appropriate way may alleviate much of the problem. An example given by Fischer and Nagin (1981) is reassuring in this respect, and relates specifically to SP data.

In the second case, it may be possible to circumvent in a similar way the hypothesised heteroscedastic effects due to variations in the competence of individuals to perform SP tasks. At present, this is an unresearched area; but,
for the sake of example, if it could be demonstrated that, say, level of education played an important part, this variable could be incorporated explicitly into the analysis, or some kind of weighting scheme could be devised.

Even after the estimation problems have been dealt with, the forecasting problems remain. By careful design of the SP tasks, much can be done to reduce the scale of the response error term $\eta$, and indeed the other contributors to this issue make important proposals for this. But it still seems unlikely that a utility function as derived from current SP analysis will be correctly scaled relative to the random effects which we hypothesise to be active in real choices. There thus remains a need for external data. Two methods are widely used, though in essence they are the same — the rescaling of the utility function to Revealed Preference data.

The first is the method of elasticities (described by Kroes and Sheldon in this issue). This requires that external evidence on elasticity of demand is available for at least one of the variables in the SP experiment. The rescaling necessary to reproduce the ‘known’ elasticity is applied to all coefficients in the utility function.

The second method requires actual choice data comparable with the alternatives presented in the SP experiment. Using the formula for deterministic utility $V$ as estimated from SP data, we calculate $V$ for each alternative available in an actual choice context. These values are then entered as the independent variables in a discrete choice estimation, and a coefficient estimated for the variable. In certain circumstances, it may be necessary to include alternative-specific constants as well. This is essentially the method proposed by Kocur et al. (1982), though it should be noted that here the motivation is rather different: it is not a question of validation, and there is no a priori expectation about the size of the ‘scaling’ coefficient (though it should of course be greater than zero).\(^2\)

Of course, in carrying out this rescaling it is implicitly assumed that the error terms in both the estimation model and the forecasting model are homoscedastic, and that all that needs to be done in moving between the two models is to adjust the scale of the variance relative to the coefficients. For the reasons already given, this is likely to be a considerable simplification.

**CONCLUSIONS**

The error structure of models based on SP data is a somewhat neglected topic. However, the discipline of random utility theory presents a framework for discussing it. Because the error pertaining to the actual SP response is unlikely to be negligible, the utility functions obtained from SP analysis will not usually have the scale properties to allow them to be used directly in demand forecasting. At the least, it will continue to be necessary to scale them by means of independent observed data.

Moreover, when data are pooled across individuals the error structure becomes

---

\(^2\) See, in this respect, a recent paper by Ben-Akiva and Bocca (1987), proposing the simultaneous estimation of $V$ using SP and Revealed Preference data.
ECONOMETRIC ISSUES IN STATED PREFERENCE ANALYSIS

J. Bates

even more complex. Techniques are being developed to deal with ad hoc error structure, but these are not yet in widespread circulation. In the absence of such techniques, the use of market segment techniques to deal with likely sources of variation is strongly advised.

Finally, the observations made in this paper derive almost entirely from theoretical speculation: empirical evidence is clearly needed about the magnitudes of the differing effects. Since there is considerable scope for investigating the problems discussed here by means of simulation techniques, it is recommended that such research be carried out as a matter of some urgency.

REFERENCES