SIMPLE APPROACHES FOR RANDOM UTILITY MODELLING WITH PANEL DATA

Andrew Daly
RAND Europe and Institute for Transport Studies, University of Leeds

Stephane Hess
Institute for Transport Studies, University of Leeds

ABSTRACT

In the last few years a large share of the choice modelling work in transport and other fields has been based on the use of data in which surveyed individuals give multiple responses. While the availability and use of multiple choices per individual has a number of significant advantages, we cannot avoid the problem that responses from a given individual cannot be treated as independent. Advanced modelling frameworks, especially those accommodating random variations in preferences, allow for an explicit treatment of the repeated choice nature of the data, but important questions need to be addressed in terms of choice of random parameters, the joint distributions used for these parameters, and the sequence of calculation of their integrals in the log-likelihood function. We discuss three different approaches that avoid the use of advanced modelling approaches while ensuring that the estimated errors are not unduly affected by the repeated choice nature of the data. We illustrate some of the properties of these approaches with real and simulated data. In this context, we show how a well used error correction method, the sandwich estimator, can be easily extended to deal with panel effects in a seemingly very reliable manner. We also explain how to do this with well known software. As an aside, we illustrate the well known theorem that if any correlation across choice tasks if independent of the attributes of the alternatives, then a naive model such as MNL will produce consistent estimates. If on the other hand, the correlation is due for example to unobserved taste heterogeneity, then the MNL estimates will no longer be consistent.

Keywords: *panel data; repeated choice; sandwich estimator; Jack-knife; Bootstrap; random utility*

1. INTRODUCTION

In the last few years there have been numerous choice modelling publications in journals relating to transport, environmental economics, marketing, health economics and other fields based on the use of data, typically from stated choice surveys, in which surveyed individuals each give multiple responses. This is what we mean by 'panel data'.

The advantages of panel data relative to data in which single responses are available from each individual can be substantial, by providing evidence of the preferences of each respondent in different circumstances. However, for both stated choice and repeated actual choice data we are presented with the analytical complication that responses from the same individual are not 'independent', in the way that responses from different individuals would be. It is this latter complication that is the main subject of the current paper. The apparatus that was developed from the early 1970's onwards to model discrete choice behaviour was based on the assumption of the independence of the observations, so that the kernel of the log likelihood function L could be formulated in the simplest possible way, i.e. as if they were independent:

$$L = \sum_n \log \prod_t p_{cnt} = \sum_{nt} \log p_{cnt} \qquad (1)$$

where p_{cnt} is the probability given by the model for the observed choice c made by individual n at choice occasion t.

The assumption of independence initially remained thoroughly defensible, since most of the data used in practice contained a single response per respondent, until stated choice data began to be analysed more regularly, starting in the early 1980's (see e.g. Louviere & Hensher, 1982; Hensher & Louviere, 1983; Kroes *et al*., 1986).

For several years analysts continued to use the methods developed for independent observations to apply to data in which the observations are in fact correlated. Model estimation which is based on this incorrect assumption, i.e. equation (1), is what we mean by 'naïve' estimation.

Naïve estimation is by no means to be discarded from the analyst's tool-kit. It is relatively simple to show that naïve estimation of a correct model yields consistent estimates of the true parameters (e.g. Liang and Zeger, 1986)[1]. However, it was soon realised by marketing analysts that the information content of N respondents each giving T responses is usually considerably less than that of NT respondents each giving a single response (e.g. Louviere and Woodworth, 1983). More fundamentally, perhaps, equation (1) simply no longer applies, because the probability of a sequence of choices is simply not equal to the product of the individual probabilities if those choices are not independent.

For these reasons the estimates of accuracy given by naïve modelling are incorrect. Soon after this was realised, correction procedures began to be applied, and a number of these methods are discussed in Section 2 of this paper. Some of these methods have not been fully understood and, perhaps for that reason, have not received the attention they merit. Transportation analysts were rather late in realising that corrections were required but by the mid-1990s the Jack-knife procedure began to be used in practical transport studies (Accent and Hague Consulting Group, 1996).

More recently however, there has been a trend to move away from correction approaches, attempting instead to account for aspects of the repeated choice nature of the data during model specification and estimation. The emphasis has been on extending the model to include a description of the precise correlation between

[1] This is not the case if the unobserved correlation across choice tasks is not independent of the explanatory variables, e.g. in the presence of unobserved taste heterogeneity, as illustrated in Appendix 1. Appendix 2 shows the application of Liang and Zeger (1986) to binary logit.

observations that undermines the basis of naïve estimation. Crucially however, it is not clear what the correct modelling approach for such data should be. Rather than addressing this question, analysts have made use of models including random factors, and have imposed assumptions on how these random terms vary to incorporate the correlation across choices made by the same individual. While very flexible when used in the right context, these methods also introduce many complexities, which are sometimes unwelcome, and a number of potentially misleading assumptions have been made in past work. Worryingly, the main motivation for advanced specification in at least some studies is seemingly simply to safeguard against the effects of the repeated choice nature on the error structure, but the use of advanced model structures in fact leads to a different set of results that may not in fact be relevant to the main issues of interest to the analyst, as discussed in Section 3.

The aim of this paper is to revisit the issue of how to account for the repeated choice nature of the data when the main emphasis is on obtaining unbiased error measures for the model estimates. We do this by reassessing, in the following section, the two most commonly used correction approaches, the Jack-knife and the Bootstrap, and highlighting a number of issues with their use in this context. We then show how another method commonly used to deal with general model misspecification, namely the sandwich (or ‘robust’) estimator, can be adapted to correct for panel effects. In Section 3, we show the possible pitfalls that can arise when relying on advanced specifications, and highlight a number of past mistakes. Finally, conclusions are drawn in Section 4 for appropriate practice.

2. SIMPLE CORRECTION APPROACHES

The basic concept of the methods discussed in this section is that they retain the naïve estimation model (1) but correct the results *a posteriori*. The main emphasis is therefore on correcting error measures, which are clearly biased, rather than on improving parameter estimates. The simplest correction method is due to Louviere and Woodworth (1983), who appear to have been the first to recognise that a problem existed. They implemented a very simple correction of increasing the standard errors estimated by the naïve model by a factor of $\sqrt{T}$. The basis for this correction was that the apparent amount of information in the data was increased by a factor of T, relative to a data set with one observation per individual, when T responses were collected per individual. Multiplying the error by $\sqrt{T}$ would then give a conservative estimate of the true strength of their findings.

More recent work along these lines has focussed on two more sophisticated techniques, namely the two standard statistical re-sampling techniques known as the Jack-knife and the Bootstrap procedures. Section 2.1 revisits these two techniques and discusses their application to the specific case of panel data. Both methods have a number of shortcomings in the present context that are highlighted in this discussion. Section 2.2 then puts forward an alternative but seemingly rarely used method, namely a simple extension of the well known ‘sandwich’ estimator to the case of panel data. Finally, Section 2.3 presents the findings of an empirical analysis comparing the three different correction approaches.

With any of the three approaches, it is recognised that the true log likelihood function is not of the form of equation (1) but rather

$$L = \sum_n \log p_{\{c\}n} \qquad (2)$$

where $p_{\{c\}n}$ is the probability given by the model for the observed *sequence* of choices $\{c\}$ made by individual n.

The issue is that the two equations (1) and (2) are equivalent only when we can write

$$p_{\{c\}n} = \prod_t p_{cnt} \qquad (3)$$

and this is true in general only when the successive choices t are independent for the individuals. If there is an issue that there may be dependence between the choices made by an individual, correct analysis would require the use of equation (2) rather than equation (1).

In the context of the correction methods discussed in this section, estimates are made using the theoretically incorrect equation (1), which can be acceptable because we know the maximum likelihood coefficient estimators for that model to be consistent for the true coefficients if and only if model (2) is the true model (e.g. Liang and Zeger, 1986, see Appendix 2; see also Hess and Train, 2010). Thus any unexplained variance across choice tasks must not be a function of the explanatory variables. However, the error estimators from the naïve model (1) are by no means consistent for the true coefficients and the main objective of the correction procedures is to get better estimates of these errors.

2.1 'Re-sampling' methods

Errors arise in estimating the coefficients in a choice model because the data that is used is a sample from the total population that the model is to represent. If the estimation is consistent, i.e. its limit as the data set increases is the correct answer, the estimation error is then exactly how much the coefficient estimates vary around the true value when the sample changes. The idea of re-sampling methods is to measure the variation of the estimates as we change the data.

We need to generate data sets 'like' the data set we have, but of course we do not have any further data. The re-sampling procedures offer varying methods of 'generating' new data sets, closely resembling the existing data, but varying enough to let us measure the variation in the estimates. The basic idea is to use the existing data to generate a 'new' data set, estimate a model on the new data and repeat this process sufficiently to make reasonable estimates of the variation of the coefficients.

The two most frequently used methods of re-sampling are the 'Bootstrap' and the 'Jack-knife'. Both are well-known statistical techniques, but in the present context there are specific issues. In particular, because the true model is (2), not (1) and to ensure that the new data is 'like' the original data, it is necessary to formulate the re-sampling in terms of the N respondents, not the NT total responses. The logic is thus that (2) is the true model: because of the fortunate consistency result quoted above we can estimate the parameters using model (1) but for estimation of the errors we must conduct the re-sampling on the basis of respondents in order to pick up the specific misspecification that (1) omits the panel effect.

The two methods are discussed in detail for this context of repeated stated choice by Cirillo *et al*. (2000), who indicate a slight preference for the Jack-knife, but other experience (e.g. Ortúzar *et al.*, 2000; Daly *et al*., 2008) calls this into question.

2.1.1 The Jack-knife

The Jack-knife is a method for eliminating bias and estimating the accuracy of parameters estimated in a model that may not be correct. An overview is given, for example, by Miller (1974). The Jack-knife corrects bias by eliminating the bias component proportional to $1/N$. Whether such bias exists in naïve estimates from panel data is not entirely clear, so that the main function of the Jack-knife in this context is to give an improved estimate of the accuracy of estimation.

In practice, the Jack-knife operates by repeatedly removing a small number of *different* observations from the data. Runs are made until every observation has been removed once. In some cases the number removed on each run is 1, but in applications with substantial numbers of observations, this is usually considered too time-consuming and N/K observations are removed, with K of the order of 20 or 30. This implies that K runs are required.

For data with repeated observations it is necessary to consider an 'observation' to be all the choices made by an individual, i.e. as in the correct model of equation (2), not treating these choices as independent as in equation (1). That is, we need to remove entire respondents, rather than removing just some of their responses.

Once the runs are made the parameters can be estimated by

$$\hat{\beta} = K\beta - \frac{K-1}{K}\sum_k \beta_{(k)} \tag{4}$$

where β is the naïve estimate and

$\beta_{(k)}$ is the estimate made on the k^{th} Jack-knife run.

Component ij of the error covariance matrix can be estimated as

$$\sigma_{ij}^{JK} = \frac{K-1}{K}\sum_k \left(\beta_{i,(k)} - \overline{\beta_i}\right)\left(\beta_{j,(k)} - \overline{\beta_j}\right) \tag{5}$$

where $\overline{\beta}_j = \sum_k \beta_{j,(k)} \big/ K$ and

$\beta_{j,(k)}$ is the j^{th} component of β estimated on the k^{th} Jack-knife run.

The difficulty that has been encountered in practice is that substantially different results can be obtained for different values of K and there may also be variation when different sampling procedures are adopted, as described below. To get round this point, some analysts have proposed working with $K = N$ (see Miller 1974), but, as noted above, this is frequently computationally infeasible when N is large and does not entirely escape from the suggestion of an arbitrary choice.

2.1.2 The Bootstrap

The Bootstrap is another re-sampling technique, related to but distinct from the Jack-knife (Efron 1981), which can also be used to obtain estimates of the true accuracy of estimation of the parameters. Unlike the Jack-knife, it is not often used to obtain improved parameter estimates, although it can be used in that way (Efron and Tibshirani, 1993).

The Bootstrap operates by sampling N observations from the original sample, *with replacement*. In consequence, some observations will be sampled multiple times, while others will be omitted from the Bootstrap sample. The process is repeated a number of times and the covariance matrix of the parameter estimates is calculated as the covariance of the parameters over the Bootstrap samples.

The number of Bootstrap samples may be chosen by the analyst. In distinction to the Jack-knife, Bootstrap runs that have already been made may be used to contribute to further estimates[2]. For this reason, the results tend to converge to a final average value, so that an impression may be obtained of the stability of the estimates.

The concept on which the Bootstrap is based is that, if the original data is a representative sample from the population being studied, then the Bootstrap samples also resemble samples that might be drawn if the sampling were done again. Intuitively, for that reason they give the sampling variance that may be expected.

2.2 Sandwich estimator

An alternative approach to dealing with the issue of bias in the estimation errors is to calculate the errors using the robust or 'sandwich' estimator. The sandwich estimator, sometimes ascribed to Godambe (1960), Huber (1967), or White (1982), is defined by

$$S = (-H)^{-1} B (-H)^{-1} \tag{6}$$

where H is the Hessian matrix, i.e. the matrix of second derivatives of the log likelihood function with respect to the model parameters to be estimated; and

[2] In other words, if an analyst had results for 20 Jack-knife and Bootstrap samples and wanted to instead make use of 30 samples, he or she would have to do a Jack-knife run with 30 new samples, but could simply complement the 20 Bootstrap samples with 10 additional samples.

B is the Berndt-Hall-Hall-Hausman (BHHH, 1974) matrix, defined as the matrix which has in cell jk the value

$$B_{jk} = \sum_n L_{jn} L_{kn} \qquad (7)$$

where L_{jn} is the derivative with respect to model parameter j of the contribution to the log likelihood function from observation n.

At the optimum value of the parameters, the average value of the first derivatives is zero and B becomes the covariance matrix of the scores (first derivatives for each observation), calculated over the observations on which the likelihood is calculated. The matrix B is identifiable as the information matrix and, when the model is correctly defined the values of B and $(-H)$ are equal at the optimum, so that S reduces to $(-H)^{-1}$. For a correctly specified model the maximum likelihood estimate of the model parameters can reasonably be taken to be distributed around the correct value with covariance matrix $(-H)^{-1}/N$, where N is the number of observations and H is calculated from the estimation sample. However, in the present context we need to consider the impact of model misspecifications and in such cases we can no longer consider $(-H)$ to be an approximation of B.

The sandwich matrix can be used more generally to give the estimation errors in a model when the model is not correctly defined in some respects. The reasoning on which the use of the sandwich matrix is based (e.g. Froot, 1989; Greene, 2008) applies when the estimate of a parameter vector is based on the maximisation of the expectation over N observations of some criterion. Then, under quite general conditions, the estimate of that parameter is distributed around the true value with covariance matrix S/N. This means that, even if the criterion on which the estimation is based is not a true maximum likelihood function, or is the maximum likelihood function for a related problem, the sandwich matrix gives a good estimate of the error distribution of the estimate around the true maximiser of the criterion. The error assessed by S will be larger than the error assessed by the information matrix, because the latter forms the Cramér-Rao lower bound. Liang and Zeger (1986), for example, show explicitly that the sandwich matrix gives the error in estimates made by the naïve method for a model of panel data.

To be clear, when there is no modelled retention of information between observations, the contribution of an observation to the first and second derivatives of the likelihood function will be the same whether the data is considered as a panel or not. That is, H will not be different. However, although the first derivative components used in calculating B are the same, the calculation procedure for S is different for the panel case and the non-panel cases, because the specification of an 'observation' is different. Indeed, the errors *should* be different, as the information content in the panel case is less than if we make the naïve assumption that the observations are independent.

Specifically, if we treat the data as being panel observations, we make the calculation

$$B_{jk} = \sum_n L_{jn} L_{kn} = \sum_n \left(\sum_t L_{jnt}\right)\left(\sum_t L_{knt}\right) \qquad (8)$$

whereas, if we treat panel data as being independent, the calculation is

$$B_{jk} = \sum_n \sum_t L_{jnt} L_{knt} \qquad (9)$$

which is clearly different and will typically give a matrix (B and hence S) with larger components on the diagonal, so that inverting the matrix will (incorrectly) indicate smaller estimation errors. Intuitively, the correct calculation (8) gives the between-individual covariance matrix of the scores, whereas the incorrect calculation (9) also includes the within-individual covariance.

It may be noted that the calculations above do not imply any specific form of model. In particular, it is not necessary for the calculation that there should actually be any *modelled* correlation between the choices made by an individual respondent. The BHHH matrix differs between panel and non-panel calculation independently of the form of the model.

This is in contrast to the calculation of the derivatives, because if the *model* represents no correlation between the choices made by an individual, the calculation of the choice probability satisfies equation (1) and we can write the overall log-likelihood as

$$L = \sum_n \sum_t \log(p_{nt}) = \sum_n \sum_t L_{nt} \qquad (10)$$

This simple summation survives differentiation, so that, unless correlation is modelled, the first and second derivatives of the model are independent of whether or not a panel effect actually exists. Thus the BHHH matrix and the Hessian differ as a function of the panel nature of the data, quite apart from any other specification issues.

The sandwich estimator is used in several discrete choice software packages with a view to correcting for general model misspecification. In the context of the analysis of panel data, the primary interest in the sandwich matrix is to calculate the potential downwards bias in the standard errors caused by not accounting for the repeated choice nature of the data. In order to apply this calculation correctly, it is necessary to use the proper formulation (2) of the likelihood function in which the probability for each individual of the observed *sequence of choices* is used. When a model is used that represents between-individual variation explicitly, e.g. a mixed logit model using integration at the respondent rather than observation level (cf. Revelt & Train, 1997), this would be the natural approach. However, if the analyst wishes to use the sandwich estimator to correct for panel effects, the sequence of choices must similarly be considered even for models such as MNL where there is *no* modelled retention of information between the choices for an individual.

Although not explicitly stated as a *feature*, it is possible to undertake such an approach with two of the most widely used packages, namely Biogeme (Bierlaire, 2008) and NLogit (Econometric Software, 2007). In Biogeme, it is sufficient for the analyst to specify a [PanelData] section in the model file, containing only the variable identifying individual respondents. The software will respond with an *error* message stating "*Warning: No random parameter capturing panel effect have been selected*", but the correct specification will be in fact be used. In NLogit, the analyst needs to use the *pds* command but needs to additionally specify a *fake* random term to *trick* the software into using the panel specification – here, an error component with fixed zero mean and standard deviations will suffice, and only a single draw needs to be specified for estimation.

Before moving on, it is important in this discussion to take note of the point made by Freedman (2006): "On the other hand, if the model is seriously in error, the sandwich may help on the variance side, but the parameters estimated by the MLE are likely to be meaningless..". In other words, we have to convince ourselves that the estimator for which the sandwich is defined is in fact a good estimator for the problem to which it is being applied. Specifically, it is required that the expected scores for the erroneous model are zero when calculated at the true values of the parameters. If that condition is satisfied, then the estimate derived from the erroneous model is distributed around the true value with covariance matrix S/N .

As we have noted, an estimator of the naïve model (1) is a consistent estimator of the model (2). This implies, of course, that Freedman's point does not apply. Here, a brief empirical analysis in Appendix 1 illustrates this concept, i.e. that a mis-specification involving only correlation of the errors does not affect the naïve estimation. On the other hand, if there is, for example, heteroskedasticity (e.g. in the form of unmeasured taste heterogeneity across respondents) in the data then issue is not whether equation (2) can be replaced by equation (1), but whether equation (2) is itself correct. This point is also illustrated in Appendix 1, where it is shown that an error caused by not modelling random taste heterogeneity leads to erroneous estimates, irrespective of the panel effect.

In the present context we also need to take account of the fact that simulation methods will often be used to estimate models for panel data. Train (2003) discusses the impact of the use of maximum simulated likelihood on error estimates and concludes that the noise and bias introduced by simulation reduce towards zero as the number of observations and the simulation accuracy (number of draws) increase. For practical purposes we do not consider that these issues affect the use of the sandwich estimator.

In summary, the sandwich estimator is a very useful means of more accurately assessing the error introduced by using an imperfect model. However, it assesses sampling error, not bias, and this needs to be recognised in its use. However, the main interest in the context of the present paper is in correcting the standard errors in models that do not accommodate the panel nature of the data, where we believe it to be very useful, when it has so far been primarily used with a view to correcting for more general misspecification. In this context, the application of the sandwich estimator with and without accounting for the panel nature of the data may give some insight into the relative importance of the panel and other mis-specifications.

2.3 Tests of the correction procedures

To illustrate the impact of these correction procedures, tests were made using two convenient data sets, both based on Stated Choice exercises.

2.3.1 First data set

The first data set used for these analyses was collected in Denmark to support the estimation of national values of travel time. The data contained 13,408 responses from 1,677 individuals travelling for private purposes, with up to 8 responses from any one individual. Each response gave a preference for either the *fast* alternative or the *cheap* alternative.

Jack-knife and Bootstrap estimators on full sample

As a first step, we made use of the Jack-knife (JK) and Bootstrap estimators on the full sample. Two series of JK runs were made, making 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 re-samples from the respondents. These runs analysed a simple model with three estimated parameters (time, cost, and a constant for the *fast* alternative), all of which were well estimated, with t ratios in the naïve model of –22.2 for the travel cost coefficient, –22.5 for the travel time coefficient, and –21.6 for the constant for the fast alternative. consistently over 20. The JK runs indicated that the bias in the parameters was in the range 0.03% to 0.3% and that the bias was to overstate the importance of the parameters. While the bias was small, it was typically at least as large as the variation of the parameter estimates over the JK runs.

The results for error estimates are shown in Figure 1, which gives the t ratios for the three coefficients over the series of runs. Additionally the geometric mean (GM) of the t ratios is presented, giving a summary statistic for the variability between runs. Compared with the naïve run, which had a GM for the t ratios of –22.1, for the same data, the GM indicates an average decrease in absolute t ratio of around 35%.

Both series showed considerable variation, but in the second series variation was reduced as the number of samples increased. For example, the coefficient of variation in the GM statistic was 7.3% in the first series and 10.7% in the second series; however, if runs 5, 10 and 15 were omitted, the variation was 8.2% and 8.6%. Thus a reduction in variation as the number of runs increases appears to be a random effect.

A further two series of runs (omitting the samples of 5, 10 and 15) confirmed that the reduction in t ratio averaged around 33%, while the variation across the runs was slightly less, giving an overall average of 7.4%.

Figure 1: Jack-knife results on full sample of Danish data

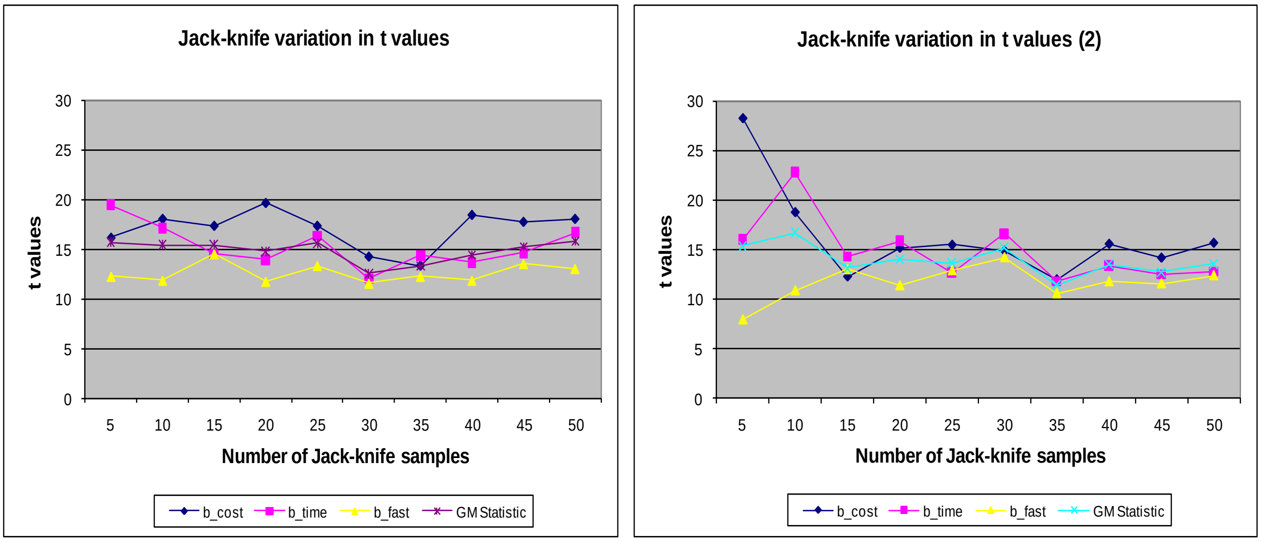
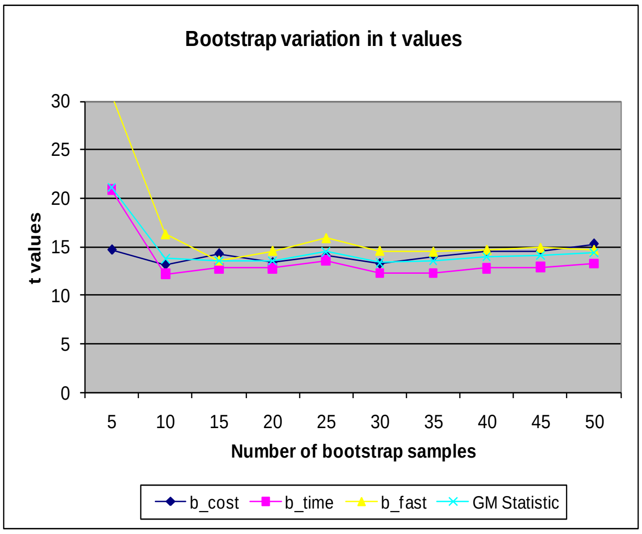


Figure 2 shows corresponding Bootstrap results. Here we see that after initial high estimates of the t ratio there is apparent convergence. The reduction in t ratio (omitting runs 5, 10 and 15) is 37.0% and the variation between runs is 3.2%. A further series of BS runs (omitting runs 5, 10 and 15; this series is not shown) gave average BS estimates of 35.8% reduction in t ratio and 4.8% variation between runs.

Compared with the JK it seems that the correction to errors is similar, but that the BS gives an estimate that is more stable over the different numbers of runs and therefore more reliable. Moreover, the JK does not seem to be stable even with 50 samples, which is a considerably higher number than seems to be used in practice: Daly *et al*. (2008) use 30 samples, for example.

Figure 2: Bootstrap results on full sample of Danish data



Sandwich matrix for full data set

The impact of the sandwich matrix is illustrated in Table 1. The runs to be made here are of a naïve model and the optimum value of the coefficients is always the same. However, different error measures are obtained by considering:

- whether the data is treated as panel data or as independent observations;
- whether we use an exact calculation of the second derivatives or the BHHH approximation;
- whether we use the classical estimation of error or the 'robust' sandwich estimator discussed in this section.

The BHHH calculations do not vary between classic and robust procedures. That is, columns 3 and 4 are the same, as are 7 and 8. This follows from the form of the robust (sandwich) estimator, which is given by equation (6) as $S = (-H)^{-1} B (-H)^{-1}$, with H being the Hessian and B being the BHHH matrix. When the Hessian is replaced by BHHH we just get H^{-1}, which is just the same as in the classic case. Columns 1 and 5 are the same, i.e. the Hessian is the same whether or not we take account of the panel effect. The reason for this is given in equation (10) above, showing that the simple summation required when correlation is not modelled is not affected by whether or not the data comes from a panel. There are therefore only 5 different error measures.

Table 1: Results of Sandwich Estimations, full sample of Danish data

	no panel				panel			
	exact		BHHH		exact		BHHH	
	classic	robust	classic	robust	classic	robust	classic	robust
t ratios	1	2	3	4	5	6	7	8
cost	-22.21	-15.64	-31.08	-31.08	-22.21	-14.70	-33.52	-33.52
fast	-21.58	-20.74	-22.21	-22.21	-21.58	-13.29	-35.00	-35.00
time	-22.54	-18.70	-26.57	-26.57	-22.54	-13.86	-36.23	-36.23
GM	-22.11	-18.24	-26.37	-26.37	-22.11	-13.94	-34.90	-34.90

When the panel effect is not included, BHHH is a better approximation of the Hessian, i.e. column 1 is more like 3-4. than when the panel effect is included, i.e. 5 is not like 7-8. The BHHH error calculations are not reliable approximations in the panel context.

The robust estimators in columns 2 and 6 are less than the classical estimators in columns 1 and 5, as expected. However, this is much more true when the panel effect is included in column 6. This follows of course from the results for BHHH, as BHHH contributes to the robust estimation and, as we saw, the BHHH differs more strongly from the classical estimator when the panel effect is considered. Thus, if the naïve specification were correct, the Hessian matrix calculations of columns 1-5 would be correct. But we may conclude already from column 2 that there are specification errors other than the panel that cause 1-5 to be incorrect, while column 6 shows that if we properly account for the nature of the data we obtain still lower estimates of the accuracy of the model.

Comparison of results on full data

Comparing the results obtained from the Jack-knife and Bootstrap procedures illustrated in Figures 1 and 2 with those from the sandwich matrix presented in Table 1 (column 6), we see that the overall t ratios indicated are similar and that the sandwich estimate appears to lie within the fluctuations observed in the re-sampling estimates. However, the variation between runs for the re-sampling procedures is estimated by comparing samples of different size, which is determined by the analyst's choice of the sample, rather than random variation which is also part of these procedures and arises from the way in which samples of a given size are drawn. For this reason, a more systematic examination of this random variation was made and this is described in the next section.

Reduced data

In practice, analysts often have smaller data sets, so that the results above may not be typical of common practice. For this reason, the remainder of the analysis is based on a 25% subset of the sample. To give a better indication of the true inherent variation of the JK and BS methods across runs, we made use of ten different seeds for the random sampling, each time using 30 re-samples.

In these runs, the JK and BS indicated similar reductions in the t ratios, as can be seen in Table 2, which focuses on the GM of the t-ratios. The variability between runs (coefficient of variation) of the reduction in GM t value was however larger in the JK than the BS. The JK runs showed a reduction in bias for three of the coefficients that was significantly higher than the variation between runs. In summary, it seems that for estimating the increased error that results from recognising the panel nature of this data, the BS gives more stable results than the JK.

Table 2 also shows the reductions obtained with the sandwich estimator, in the cross-sectional as well as panel specifications, analogous to the results shown in Table 1. A comparison of the results of the sandwich estimator with and without account of the panel effect indicates that the cross-sectional correction lies between zero and the full panel correction. We conclude that some of the correction is due to allowing for general misspecification, some due to the specific panel misspecification. While the Jack-knife and Bootstrap approaches give varying reductions, we note that their coefficients of variation are large enough that the sandwich estimate (as panel) lies within one standard deviation of each. We conclude that the three approaches are giving consistent results.

The Jack-knife indicates that the coefficient values should be increased, i.e., since they are negative, that the naïve run indicates too strong an effect, repeating the finding with the full data set. While the effect is less than 2% and unlikely to be important, the coefficient of variation shows that it is significantly different from zero.

Table 2: Comparisons of correction methods, reduced Danish data (geometric mean of the t ratios)

	Reduction	Coefficient of variation of reductions (across 10 runs)
Jack-knife t ratios	37.87%	0.15
Jack-knife coefficients	-1.78%	0.25
Bootstrap t ratios	31.87%	0.21
Sandwich t ratios (as cross-section)	20.37%	N/A
Sandwich t ratios (as panel)	35.38%	N/A

2.3.2 Second data set

For this case study, we make use of data from a survey looking at rail travel behaviour, collected through an online panel in the United Kingdom in early 2010. In particular, we rely on a sample of 7,936 observations collected from 992 respondents, each faced with 8 exercises involving a choice among three alternatives, where the attributes were pivoted around those of a reported trip (but without a reference alternative being included). The alternatives were described on the basis of travel time, fare, the guarantee of a reserved seat, the provision of free wifi, and whether the ticket offered flexibility (in terms of rescheduling).

We make use of the same approach as for the analysis on the reduced Danish data, i.e. using 30 samples for Jack-knife and Bootstrap. For the presentation of the results, we focus on the geometric mean, with results reported in Table 3.

Table 3: Comparisons of correction methods, rail SP data (geometric mean of the t ratios)

	Reduction	Coefficient of variation of reductions (across 10 runs)
Jack-knife t ratios	22.90%	0.22
Jack-knife coefficients	-0.08%	0.78
Bootstrap t ratios	21.92%	0.27
Sandwich t ratios (as x-section)	0.26%	N/A
Sandwich t ratios (as panel)	23.46%	N/A

The Jack-knife and Bootstrap approaches give very similar reductions, although in this case the Jack-knife is more stable, and again we note that their coefficients of variation are sufficient that the sandwich estimate (as panel) lies well within one standard deviation of each. We conclude again that the three approaches are giving consistent results. The Jack-knife again indicates that the coefficient values should be increased, i.e., since they are negative, that the naïve run indicates too strong an effect, repeating the finding with the Danish data; however, in this case the effect is very small so that the bias is not significantly different from zero. Finally, comparison of the results of the sandwich estimator with and without account of the panel effect indicates in this case that the cross-sectional correction lies close to zero and far from that of the full panel correction. We conclude that almost all of the correction is due to panel effects.

3. MODEL EXTENSIONS

Recently, there has been a shift away from correction approaches, with an increasing emphasis on a direct treatment of the repeated choice nature of panel data. To a large extent, this has come about as a result of the growing popularity of random parameter models such as Mixed Multinomial Logit (MMNL). In this section, we review two main approaches in this context, namely random coefficients approaches, and error components approaches. If used for the right reasons, these advanced models can lead to very significant improvements in understanding behaviour and in modelling flexibility and performance. However, there is evidence of authors deploying these methods (occasionally at a reviewer's request) solely with a view to accommodating the panel nature of the data. Inappropriate specifications are regularly used in this context, and the present section discusses why this is ill informed and how the use of the methods discussed in Section 2 may in such cases be more appropriate.

3.1 Random coefficients

An increasing number of discrete choice applications make use of the random coefficients specification of the MMNL model. This model allows for random variations in sensitivities in the data, leading to often very significant increases in explanatory power and insights into behavioural patterns. In the most basic approach, the variation in sensitivities is across all observations in the data, but this was extended by Revelt & Train (1997) to a specification in which the variation is across respondents, with intra-respondent homogeneity, with the exception of a residual 'white noise' error. Through the specific assumptions made in this model, there is an accommodation of the panel nature of the data by allowing for correlation across observations for the same respondent. Analysts have seemingly increasingly come to see this as an ideal solution to the issue of appropriately dealing with the repeated choice nature of panel data.

While it is undeniable that the Revelt & Train (1997) approach offers a treatment of the intra-respondent correlation across choices, and while this is generally observed to lead to substantial gains in fit and increased ability to recover taste heterogeneity, it is of course impossible to know whether this is in fact the *true* nature of the correlation in place in the data and there is significant scope for confounding. Indeed, it is very easy for inter-personal variation in one parameter to be confounded with similar variation in another, while adding a coefficient that is correlated between choices made by an individual may apparently explain *any* correlation between those choices, not just correlation relating to that coefficient. Moreover, if only inter-personal preference variation is investigated, this runs the risk of being a proxy for variation that occurs between the responses of an individual.

On the whole, analysts seem to be happy with the assumptions made in MMNL models in relation to the interpersonal variation in values. What is more of an issue however is that the use of Mixed Logit may be motivated solely by its ability to incorporate a treatment of the repeated choice nature of the data. Indeed, using the MMNL model brings many aspects of opening a can of worms. Numerous assumptions need to be made in the specification of this model, relating not just to the treatment of panel effects (see e.g. recent discussions in Hess & Rose, 2009), but also the decision on the choice of which parameters should follow a random distribution, what distribution this should be (cf. Hensher & Greene, 2003; Hess et al., 2005), and whether a (correlated) multivariate distribution should be used. Aside from this, there are significant computational and identification issues, and estimating the numerous parameters of a fully specified MMNL model for inter and intra-personal variation may be beyond the budget or scope of a study, the capabilities of available software (e.g. in drawing correlated samples from suitable distributions[3]) and/or the information content of a data set. Especially when the only motivation for using MMNL is to accommodate the panel nature of the data, this may lead an analyst to limit the heterogeneity to an arbitrary subset of parameters (or even just a single one) or make other misguided assumptions, potentially leading to biased results. Even if this can be avoided, and a fully specified model can be estimated, the issue of interpretation remains, most notably in relation to the computation of WTP indicators (see e.g. Daly et al., 2009). Furthermore, especially in applied work, analysts may in fact only be interested in producing point estimates of average valuations, and obtaining unbiased point estimates of WTP measures from models allowing for taste heterogeneity is by no means straightforward.

3.2 Error components

Some analysts have attempted to avoid the pitfalls of introducing a representation of random heterogeneity by instead turning to error components with a view to capturing correlation across choices for the same respondents. Error components are a flexible way to allow for heteroskedasticity or various correlation patterns (cf. Walker, 2001). These methods are attractive, but require care in specifying the way in which additional variation enters the model.

The effect that analysts have striven to capture is an individual-specific effect that creates correlation across choices for the same respondent but is not related to a specific alternative. This would involve the use of error components that are distributed at the level of individual respondents, much as in a Revelt & Train (1997) specification for random heterogeneity. However, it is clearly impossible for identification reasons to add the same error components to all J of the alternatives. As for example reported by Yáñez *et al*. (2010), it has become common practice to instead include the same error component in J-1 of the utility functions (see also example in Bierlaire, 2008). Unfortunately, this ignores the fact that the models now introduce correlation across those J-1 alternatives, as well as heteroskedasticity, potentially once again unduly influencing those results that are of interest to the analyst who simply wanted to obtain a correction of the error estimates. Adding an error component to just a single alternative is no better, while randomly varying the

[3] Here, the procedure for drawing correlated Gumbel variables used in the Appendix is of interest for other non-Normal multivariate distributions.

omitted error component across respondents (cf. Yáñez *et al*. 2010) is similarly misguided, as it creates random variations in the correlation/heteroskedasticity structure across individuals. The only scenario in which the addition of an error component to J-1 alternatives makes sense is in the case of binary data, where issues with correlation and heteroskedasticity (only differences matter) are avoided. More generally, however, it seems that these methods are simply mistaken.

Seemingly the only reasonable approach based on error components is an approach put forward by Hess *et al*. (2008) which consists of adding independently but identically distributed error components to all of the alternatives, with the integration carried out at the level of individuals rather than choices. Such a model is identified in panel data given the independent nature of the error components, and avoids issues with inter-alternative correlation through the independence assumption, while the use of identically distributed error components maintains homoskedasticity. The approach has been found to work well, for example in the work cited above, but in theory it seems not to be suitable for truly unlabelled data (if such data exists).

4. CONCLUSIONS

With a growing reliance on stated choice data and other data sources containing multiple responses for each individual, numerous studies need to address the question on how to best accommodate the repeated choice nature of the data with a view to avoiding biased standard errors. In studies making use of methods that explicitly model the relationship between choices, such as Mixed Logit, the risk of biased standard errors is reduced, though the analyst is faced with the difficulty of needing to make multiple assumptions in relation to the specification of the random terms in the models, in terms of which parameters they apply to, what distributions are used, and how they are jointly distributed (e.g. across respondents only, or with additional intra-respondent variation; also the issue of correlation between the parameters needs to be considered, together with the sampling problems that this introduces).

As stressed by numerous authors in a more general context (e.g. Hensher and Greene, 2003), the pitfalls with Mixed Logit are already quite severe, and analysts should in our view therefore be wary of deploying this structure solely for the aim of capturing correlation across individual choices made by the same respondent. In many ways, the situation can be described as opening a can of worms. The specification issues listed above still arise, and numerous mistakes have been made in the literature with a view to bypassing some of the complexities, arguably with a view to seeking a quick fix of the panel effects issue. While not wishing to discredit Mixed Logit, which currently offers the best approach to understanding inter and intra-personal heterogeneity in taste, this paper thus seeks to highlight the risks and remind authors that other methods exist to correct the standard errors if an explicit treatment of the panel nature of the data is impossible or undesirable (e.g. for interpretation reasons).

We present three correction approaches to dealing with the issue of error bias in panel data, the Jack-knife, the Bootstrap, and the (in our view) thus far under-used panel specification of the Sandwich estimator. An outline is given of the theory and how these three approaches seem to be applicable to the correction of error bias in naïve model estimation from panel data. We are not aware of a comparative treatment of these three approaches for this important context.

The approaches are illustrated by application to two data sets based on Stated Choice experiments; one data set is tested in larger and smaller variants. Without suggesting that these tests are exhaustive, it is notable that the results show that the three approaches seem to be consistent.

In comparison with the more widely used Jack-Knife and Bootstrap methods, the computation for the sandwich approach is much less onerous, there is no issue of variation with sampling and the issue of choosing a sample size does not occur. Moreover, it would be possible to program the sandwich approach so that the panel and cross-sectional variants were calculated at the same time, a saving that would not be possible with the re-sampling approaches, where new samples would have to be drawn and a whole new series of runs undertaken to obtain these results. Giving both panel and cross-sectional calculations would have the advantage of giving an indication as to what share of the correction was due to general misspecification, and what share was due to the panel nature of the data.

The only possible shortcoming we see is that the sandwich method is not able to correct any bias in the actual estimates. However, in our own work and in the literature it does not appear that such biases are large and for large data sets we have the reassurance that, when the correlation is independent of the attributes, the consistency results mean that we can rely on the naïve estimates.

It seems that the sandwich approach is thus very attractive among these correction approaches.

Acknowledgements

This work described in this paper was supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the project "Estimation of travel demand models from panel data", grant EP/G033609/1. The second author also acknowledges the support of the Leverhulme Trust in the form of a Leverhulme Early Career Fellowship.

References

Accent and Hague Consulting Group (1996) The value of travel time on UK roads, The Hague and London.

Berndt, E. U., Hall, B. H., Hall, R. E. and Hausman, J. A. (1974.) Estimation and inference in non-linear structural models, Annals of Economic and Social Measurement 3/4, pp. 653-665.

Bierlaire, M. (2008). An introduction to BIOGEME Version 1.7, biogeme.epfl.ch

Cirillo, C., Daly, A. and Lindveld, K. (2000) Eliminating bias due to the repeated measurement problem in SP data, in Ortúzar, J. de D. (ed.) Stated Preference Modelling Techniques: PTRC Perspectives, PTRC.

Daly, A., Hess, S. & Train, K. (2009), Assuring finite moments for willingness to pay in random coefficient models, presented at the European Transport Conference, Noordwijkerhout, Netherlands.

Daly, A., Choudhury, C., Burge, P. and Sivakumar, A. (2008) Dealing with repeated choices in stated preference data, presented to European Transport Conference, Noordwijkerhout, Netherlands.

Econometric Software (2007), Nlogit 4.0, Econometric Software, New York and Sydney.

Efron, B. (1981) Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods, *Biometrika*, **68**, pp. 589-99.

Efron, B. and Tibshirani, R. (1993) An introduction to the bootstrap, Chapman and Hall/CRC.

Froot, K. A. (1989) Consistent covariance estimation with cross-sectional dependence and heteroskedasticity in financial data, J. of Financial and Quantitative Analysis, 24 (3).

Godambe, V. P. (1960) An optimum property of regular maximum likelihood estimation, *Annals of Math. Stats.*, **31** (4), pp. 1208-1211.

Greene, W. H. (2008) Econometric Analysis (Sixth Edition), Pearson.

Hensher, D.A. and Louviere, J.J. (1983) Identifying individual preferences for international air fares, Journal of Transport Economics and Policy, XVII (2), 225-245.

Hensher, D. A., Greene, W. H., 2003. The Mixed Logit Model: The State of Practice. Transportation 30 (2), pp. 133-176.

Hess, S., Bierlaire, M. & Polak, J.W. (2005), Estimation of value of travel-time savings using Mixed Logit models, Transportation Research Part A, 39(2-3), pp. 221-236.

Hess, S., Rose, J.M. & Hensher, D.A. (2008), Asymmetrical Preference Formation in Willingness to Pay Estimates in Discrete Choice Models, *Transportation Research Part E*, 44(5), pp. 847-863.

Hess, S. & Rose, J.M. (2009), Allowing for intra-respondent variations in coefficients estimated on stated preference data, Transportation Research B, 43(6), pp. 708-719.

Hess, S. & Train, K.E. (2010), Recovery of inter- and intra-personal heterogeneity using mixed logit models, ITS working paper, Institute for Transport Studies, University of Leeds.

Kroes, E., Sheldon, R. and Beswick, M. (1986): Stated Preference Microsimulation Models from Qualitative Inputs to Estimate Market Shares in InterCity Travel, proceedings of the 1986 Esomar Conference, Monte Carlo.

Liang, K.-Y. and Zeger, S. L. (1986) Longitudinal data analysis using generalized linear models, Biometrika, Vol. 73, pp. 13-22.

Louviere, J.J. and Hensher, D.A. (1983) Using discrete choice models with experimental design data to forecast consumer demand for a unique cultural event, Journal of Consumer Research, 10 (3), 348-361

Louviere, J. and Woodworth, G. (1983) Design and analysis of simulated consumer choice or allocation experiments: an approach based on aggregate data, Journal of Marketing Research, **20** (4), pp. 350-367.

Miller, R. G. (1974) The jackknife – a review, *Biometrika*, **61**, pp. 1-15.

Ortúzar, J. de D., D.A. Roncagliolo and U.C. Velarde (2000) Interactions and Independence in Stated Preference Modelling. In: J. de D. Ortúzar (ed.) Stated Preference Modelling Techniques: PTRC Perspectives, Vol. 4, PTRC Education and Research Services Ltd, London.

Revelt, D. and Train, K. (1998) Mixed Logit with repeated choices: households choices of appliance efficiency level, Review of Economics and Statistics, **80** (4), pp. 647-657.

Walker, J.L. (2001), Extended Discrete Choice Models: Integrated Framework, Flexible Error Structures, and Latent Variables, PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica, 50, 1-25.

Yáñez, M.F., Cherchi, E., Heydecker, B.G. and Ortúzar JdD. (2010), On the treatment of repeated observations in panel data: efficiency of mixed logit parameter estimates, Networks & Spatial Economics, Special Issue on Latin-American Research, forthcoming.

Appendix 1: Simple runs to illustrate mis-specification effects

In order to illustrate the consistency of naïve models under specific circumstances two series of very simple simulation runs were made. In the first series, data was generated from the binary choice model with utility difference between the alternatives given by

$$\Delta U = \beta_0 + \beta_1 x + \Lambda^{-1}\left(\Phi\left(\frac{\xi_1+\lambda\xi_2}{\sqrt{1+\lambda^2}}\right)\right)$$

where x, ξ_1 and ξ_2 are distributed independently with standard normal distributions;
β are coefficients fixed to the values $\beta_0 = 0.5, \beta_1 = 1$;
Λ^{-1} is the inverse logistic distribution, $\Lambda^{-1}(t) = log(t/(1-t))$;[4]
Φ is the cumulative standard normal distribution.

The values x are treated as data and drawn independently for each of one million 'observations'; the value ξ_1 is treated as noise and is also drawn independently for each 'observation', while the value ξ_2 is treated as individual-specific preference and is drawn randomly for each of 100,000 'individuals', each of whom is treated as contributing 10 'observations'.

The effect of the final term in the equation is that we obtain:

- first, normally distributed variables with variance 1, independent between individuals but with a correlation within individuals of $\lambda^2/(1+\lambda^2)$;
- second, applying the transformation Φ, uniformly distributed variables in the range (0, 1), with approximate within-individual correlation as before;
- finally, applying the transformation Λ^{-1}, we get standard logistically distributed variables, with the approximate correlations as before.

Setting $\lambda = 1$ we get approximate within-individual logistic correlation of 0.5.

Having made the calculations of ΔU, 'choices' are assigned: choice is 1 if $\Delta U < 0$, otherwise choice is 2. From this 'data', the model is then estimated naïvely, assuming that the error is simply independent logistic variables. More correct error measures are then calculated using the sandwich matrix method. The results are shown in the table.

Variables	True values	Naïve estimates	Robust estimates (sandwich matrix)
β_0	0.5	0.4971	
t statistic		218	116
β_1	1.0	1.002	
t statistic		375	343

It is clear from these results that, for this data, the naïve estimates are very close to the true values. Comparing those error estimates with robust estimates using the sandwich matrix we find a large difference for β_0, though very much less for β_1.

[4] Note that the standard logistic distribution, with cumulative form $\Lambda(t) = \exp(t)/(1+\exp(t))$, is the distribution of the difference of two independent standard Gumbel-distributed variables, and has variance $\pi^2/3$, i.e. twice that of the standard Gumbel distribution.

Because the specification error caused by correlation is not itself correlated with x, this result (in terms of unbiased estimate) is not unexpected. Of course a single data set cannot prove a general result (although it could *disprove* one if opposite results had been found), it is certain, given the size of this data, that the result is not a random effect and the finding is therefore strongly suggestive that the theoretical result of consistency operates as expected.

In the second series of runs, the impact of taste variation was tested by generating data with the model:

$$\Delta U = \beta_0 + (\beta_1 + \xi_1)x + \Lambda^{-1}(\xi_2)$$

where ξ_2 is distributed independently of ξ_1 with a uniform distribution in (0, 1); other variables are defined as in the previous model.

Two variants of the data were generated. In the first case ξ_1 is drawn independently for each ‘observation’ (i.e. cross-sectional data), while in the second case ξ_1 is treated as individual-specific preference and is drawn for each of the 100,000 ‘individuals’ (i.e. panel data).

The two models were then estimated naïvely, assuming that there was no taste variation. The results are shown in the table.

	Cross-sectional variation		Panel variation	
Variables	True values	Naïve estimates	True values	Naïve estimates
β_0	0.5	0.4439	0.5	0.4439
t statistic		204		204
β_1	1.0	0.7468	1.0	0.7414
t statistic		310		308

In considering these results, it is necessary to consider that the scale of error in the model is greater than given by the standard logistic alone. Specifically, the term $\xi_1 x$ adds variance: since ξ_1 and x are independent, have mean zero and variance 1, their product has variance 1, adding to the variance $\pi^2/3$ of the logistic distribution. The error scale is therefore increased by a factor $\sqrt{\frac{1+\pi^2/3}{\pi^2/3}} \approx 1.142$. Thus the 11.2% reduction in the value of β_0 is not unexpected, but the reduction of more than 25% in the value of β_1 is an indication of bias caused by unexplained taste heterogeneity

We conclude that the naïve model is not capable of giving a consistent estimate of the parameters in the presence of taste heterogeneity. We also see that bias affects one coefficient much more than the other, so that taste heterogeneity can bias the *relative* values of parameters, not just the overall scale. Finally we see that it scarcely affects the results, either parameters or errors, whether we consider ‘respondent’-specific or observation-specific variation, suggesting that the biasing effect of taste heterogeneity is not specifically a panel effect, although of course this data set may be too simple to show the full ramifications of panel data.

While not unexpected, this appendix confirms two things. If the correlation in errors between choice tasks is unrelated to interactions with explanatory variables, then a model not accounting for this within individual correlation will still provide consistent estimates, but biased standard errors. If on the other hand, the data contains unexplained taste heterogeneity then a model assuming taste homogeneity is prone to producing biased estimates with differential bias across coefficients, hence also affecting the relative values.

Appendix 2: Application of Liang and Zeger Theorem to Logit Models

The widely quoted paper of Liang and Zeger (1986) presents a theorem on 'generalised linear models'. By this they mean models formulated in terms of a parameter θ which is a function of a linear combination of the unknown parameters, i.e. $\theta = h(\eta)$ and $\eta = x\beta$, where β are the parameters to be estimated and x is data.

Theorem 1 of Liang and Zeger then states (in the notation of the current paper, but the equation numbers are theirs):

> Let X_n be the T*K matrix of covariates and let Y_n be the T-vector of outcomes, both for individual n.
>
> Assume that the marginal density of y_{nt} is given by
>
> $$f(y_{nt}) = exp[\{y_{nt}\theta_{nt} - a(\theta_{nt}) + b(y_{nt})\}\phi] \quad (1)$$
>
> where $\theta_{nt} = h(\eta_{nt})$ and $\eta_{nt} = x_{nt}\beta$; and let $\hat{\beta}$ be the solution of the equation
>
> $$\sum_n X_n^T \Delta_n L_n^{'} = 0 \quad (3)$$
>
> where $\Delta_n = diag(\partial\theta_{nt}/\partial\eta_{nt})$ is T*T and $L_n^{'} = Y_n - a_n^{'}(\theta)$ is T*1.
>
> Then, under mild regularity conditions, $\hat{\beta}$ is consistent for β and $\sqrt{N}(\hat{\beta} - \beta)$ is asymptotically normally distributed as $N \rightarrow \infty$ with zero mean and covariance matrix given by
>
> $$S = \lim_{N\rightarrow\infty} N\{H_1(\beta)\}^{-1} H_2(\beta) \{H_1(\beta)\}^{-1} \quad (4)$$
>
> where $H_1(\beta) = \sum_n X_n^T \Delta_n A_n \Delta_n X_n$ and $A_n = diag\left(a^{''}(\theta_{nt})\right)$
> $H_2(\beta) = \sum_n X_n^T \Delta_n cov(Y_n) \Delta_n X_n$.

For example, for linear binary logit we can set ϕ = 1, b = 0, $h(z) = z$ and $a(\theta) = log\big(1 + exp(\theta)\big)$, obtaining

$$f(y_{nt}) = exp[\{y_{nt}\theta_{nt} - a(\theta_{nt}) + b(y_{nt})\}\phi] = \frac{exp[y_{nt}x_{nt}\beta]}{1+exp[x_{nt}\beta]}$$

i.e. the usual likelihood contribution as required.

In this case, because of the simple form of h, we get $\Delta = I$, the identity matrix, so there is a considerable simplification. In particular

$$H_1(\beta) = \sum_n X_n^T A_n X_n \text{ and } H_2(\beta) = \sum_n X_n^T cov(Y_n) X_n$$

It is clear that H_1 is the analogue of the Hessian, while H_2 is analogous to the BHHH matrix, calculated taking account of the panel nature of the data, so that S (equation 4 above) represents the sandwich matrix.

Thus the theorem assures us that the naïve estimates for binary logit models are consistent if the true model is as specified. Note that Liang and Zeger do not specify exactly how observations in the real data generating process might be correlated. That is, all they assume is that the *marginal* equation for the density of *y* is as given by equation (1) above and that $\hat{\beta}$ is the solution of equations that assume independence of observations.

The way in which the theorem is formulated means that it applies only to binary logit models.