Estimating willingness-to-pay from discrete choice models: setting the record straight

Andrew Daly[a], Stephane Hess[b] and Juan de Dios Ortúzar[c]

[a] Choice Modelling Centre & Institute for Transport Studies, University of Leeds, Leeds, UK; e-mail: andrew@alogit.com (ORCID 0000-0001-5319-2745)

[b] Choice Modelling Centre & Institute for Transport Studies, University of Leeds, Leeds, UK; e-mail: s.hess@leeds.ac.uk (ORCID: 0000-0002-3650-2518)

[c] Department of Transport Engineering and Logistics, Instituto Sistemas Complejos de Ingeniería (ISCI), BRT+ Centre of Excellence, Pontificia Universidad Católica de Chile, Santiago, Chile; e-mail: jos@ing.puc.cl (ORCID: 0000-0003-3452-3574)

Abstract

A recent paper by Carson and Czajkowski (2019) motivates an investigation of the issues surrounding the estimation of the accuracy of trade-off ratios such as willingness-to-pay derived from choice models. We show that the exponential transformation proposed by Carson and Czajkowski is motivated by an incorrect (though seemingly widespread) understanding of the asymptotic normality of maximum likelihood estimates and also raises a number of practical issues. We illustrate these findings empirically and seek to clarify the distinction between asymptotic normality and a population level distribution. We point to better approaches to solving this problem, such as the Delta method, the likelihood ratio method of Armstrong et al. (2001) or the Bootstrap approach. We also present a theorem confirming the applicability of the Delta method to robust error calculation.

Keywords: willingness-to-pay; discrete choice; random utility; standard errors

1. Introduction

The estimation of indicators of willingness-to-pay (WTP) and the computation of associated measures of uncertainty has attracted much interest in choice modelling, but has also been characterised by a substantial amount of confusion and misguided discussions. We have addressed these points in a number of previous papers (Armstrong et al., 2001; Daly et al., 2012a, 2012b; Hess et al., 2005; Sillano & Ortúzar, 2005; but also Daly and Zachary, 1975), yet we have found that much confusion persists, especially outside our main discipline of transport research.

Our core motivation here is to respond to a recent paper by Carson & Czajkowski (2019), published in the Journal of Environmental Economics and Management (hereafter referred to as C&C). In this paper, C&C state that they … "*show a substantive problem exists with the widely-used ratio of coefficients approach to calculating willingness to pay (WTP) from discrete choice models*". The work of C&C goes back to a conference paper presented at the *International Choice Modelling Conference* in 2013. It is presumably in relation to this occasion where C&C wish to thank two of us for … "*valuable discussions and comments*". However, it seems that these comments were not considered when writing their paper, as it contains some misconceptions that we now discuss. The present paper addresses not just the issues in the C&C paper, but provides additional detail and background with the hope of clarifying the ongoing confusion in many other papers and conference presentations.

In choice models with utility functions linear in attributes and parameters, we would have:

$$V = \beta_c . c + \beta_x . x + \cdots \quad (1)$$

where V is the systematic utility (we omit subscripts for decision makers, choice situations and alternatives), x is the attribute for which we wish to calculate WTP[1], and c is the cost attribute. Assuming that the marginal utilities are the same across alternatives and choice situations, we have that, for a given decision maker (Beesley, 1965; Gaudry et al., 1989):

$$WTP = \frac{\delta V}{\delta x} / \frac{\delta V}{\delta c} = \frac{\beta_x}{\beta_c}, \quad (2)$$

that is, the WTP is simply calculated as the ratio of two marginal utility parameters.

While some complications arise in the situation where the marginal utilities $\frac{\delta V}{\delta x}$ and $\frac{\delta V}{\delta c}$ depend on the level of the attribute in question and/or observable characteristics of the decision-maker, much greater issues arise when they vary randomly across individuals, such as in a Mixed Logit model. The analyst then has the ability to make a choice of a sample (population) level distribution, and much work has looked at behavioural realism, for example, would a normal distribution make sense for a cost coefficient as it implies accepting both positive and negative values (cf. Hess et al., 2005; Sillano & Ortúzar, 2005). An equally important issue relates to the fact that we now have the possibility of the denominator in (1) not being bounded away from zero, that is, having a continuous distribution that includes zero. Daly et al. (2012b) provide extensive discussions in this context as well as a theorem showing which assumptions for the population-level distribution of β_c ensure that the moments of the distribution of WTP are defined.

A completely separate issue relates to the computation of measures of uncertainty for the component in (2), in the form of standard errors and/or confidence intervals. When the WTP is given by a ratio of two parameters, say $WTP = \frac{\beta_x}{\beta_c}$, the question arises of how to calculate an error measure for this ratio. This is the topic of C&C, and hence the focus of this note.

It needs to be stated categorically that these two issues are quite separate. Unfortunately, however, notably also in C&C, findings in relation to the former (assumed population level distributions) have been used in discussions about the latter (computation of measures of uncertainty), often as a result of misunderstanding the asymptotic properties of maximum likelihood estimates.

The remainder of this paper is organised as follows. We first discuss the specification put forward by C&C. Section 3 explains the differences between normality and asymptotic normality. Section 4 looks at the calculation of measures of uncertainty for ratios of marginal utilities, and Section 5 presents a brief empirical comparison. Section 6 offers some conclusions.

[1] Our focus here is on WTP, but the points apply to marginal rates of substitution more generally.

2. The new “baseline model” of C&C

C&C apply a negative exponential to the parameter estimated for the marginal cost sensitivity, reformulating the model from (1) as follows:

$$V = -\exp(\gamma)\,.\,c + \beta_x.\,x + \cdots \qquad (3)$$

The WTP from the C&C specification would then be calculated as WTP $= \frac{\beta_x}{-\exp(\gamma)}$.

C&C first state that … “*[t]his alternative specification imposes a standard restriction from neoclassical economic theory*”, that is that the cost coefficient must be negative. While this is of course reasonable in terms of economic theory, it is questionable whether it is beneficial in model building. Imagine a situation where due to either data problems or an endogenous price/quality relationship (e.g. Palma et al., 2016), the value for the cost coefficient is positive at the optimum. Such a finding is beneficial to the analyst in terms of then improving the model or dealing with any data issues. Preventing the estimation from highlighting the problem is not in anyone’s interest.

Second, C&C suggest that their formulation … “*avoids problems associated with non-existent moments of the resulting WTP ratio distribution*”. This point, unfortunately, mistakes asymptotic distributions for population level distributions, a point we address in detail in the following Section, where we show that the standard specification has just as much status and validity as their “new” specification.

Third, C&C also seem imply that, without this negative exponential restriction, the point value of the WTP may be incorrect, stating that … “*the mean and standard deviation of the resulting WTP ratio distribution are undefined*”. This is factually incorrect in the case of parameters without a population-level distribution (as in C&C), and again misunderstands the nature of asymptotic distributions. We show, later in the paper, that the results are exactly equivalent across specifications, as are their statistical properties.

Fourth, a key step in choice modelling involves testing the significance of individual model parameters; in the case of a cost coefficient testing whether the estimate is different from zero. With the C&C specification, the fact that we obtain an estimate of γ complicates matters. A t-ratio on γ against zero only tells us whether the cost coefficient is different from -1, which is not helpful. To establish whether the cost coefficient has a significant impact on the model, would require either a likelihood ratio test against a (separately estimated) model excluding the cost coefficient, or the calculation of a standard error for $-\exp(\gamma)$, a point we return to in Section 3.

Finally, a point not considered by C&C is that the use of an exponential transform in estimation can itself create more harm than good. A search for maximum likelihood estimates involves making changes to the parameter values, and changes of the same step size can have a much larger impact for those parameters to which an exponential transform is applied.

3. Population-level *vs* asymptotic distributions

C&C start their paper by correctly stating that … “*numerical approximation of the standard error associated with WTP generally requires the assumption that the ML estimates of the utility function parameters have an asymptotically normal distribution*”. However, they then follow that by saying that … *"[c]alculating moments of a resulting ratio distribution (e.g., the empirical distribution of WTP) becomes problematic. Non-existence of moments of a ratio distribution resulting from dividing two normally distributed variables has long been known in statistics … As a result, the mean and standard deviation of the resulting WTP ratio distribution are undefined, and the resulting distribution is not normal, being typically skewed and potentially bimodal."* The apparent issue highlighted by C&C would relate to normally distributed parameters, but not estimates that have the lesser property of **asymptotic** normality. Given this apparent confusion, it unfortunately seems necessary to explain the distinction between an assumption relating to the distribution of a parameter in the population and uncertainty affecting the estimated value of a parameter.

Let us assume that our model contains a vector of unknown parameters β which are estimated using the maximum likelihood criterion. We take a frequentist view that there exists a single true value β^*, but that our data is incomplete and we thus have sampling error. Provided reasonable conditions are met and the model is correctly specified, the expected score (first derivative of the likelihood function with respect to the model parameters) is zero and the maximum likelihood estimates (MLE) $\hat{\beta}$ of the model parameters converge, as the sample size N increases, to a normal distribution around the true values β^*:

$$\sqrt{N}\left(\hat{\beta} - \beta^*\right) \rightarrow N(0, \Omega) \qquad (4)$$

where Ω is the negative of the inverse of the Hessian (matrix of second derivatives) of the likelihood function with respect to the model parameters and forms the Cramér-Rao lower bound, so these estimates are minimum variance. The standard errors, which give a measure of the uncertainty affecting an estimate of a parameter in a model, are given by the square root of the diagonal of Ω.

Two key observations need to be made.

First, given finite sample sizes, the key property of MLEs is one of asymptotic normality, rather than normality, and this important distinction seems to be poorly understood. To be specific, if the estimated parameters were distributed exactly normal, this would imply that the log likelihood function was quadratic in the parameters. A shift away from the optimum $\widehat{\beta_k}$ for parameter β_k by a value $\Delta\beta_k$ would reduce the log likelihood by $\Delta LL = \Delta\beta_k^2/2\,\sigma_{\beta_k}^2$, where σ_{β_k} is the standard error for the estimated optimum value[2]. The 'asymptotic' qualifier means that the property applies exactly only for very large data sets. For finite data sets, it is an approximation that applies only in the neighbourhood of the optimum. Indeed, as is shown by Armstrong et al. (2001), for example, the log likelihood function is not exactly quadratic, so that reductions in likelihood do not follow this formula exactly at all. What happens as we move further away from the optimum is not defined by maximum likelihood theory. The parameter estimates are not distributed exactly normal, and any theoretical insights/claims based on an assumption of normality are misguided.

Second, unlike population level distributions, where an analyst makes a decision on the assumed shape of a distribution for each marginal utility component, asymptotic normality is a property of MLEs of model parameters. This applies to fixed marginal utility components as

[2] This can be seen by approximating the log likelihood by a quadratic function around the optimum and noting that the estimation variance ($\sigma_{\beta_k}^2$) is equal to the inverse of the second derivative of the log likelihood function.

well as those varying randomly across individuals. For the latter, the estimates of the parameters of the assumed distribution will have this property. This is independent of the distributional assumptions, with for example the estimates of the offset and range of a uniform distribution still having the property of asymptotic normality despite the sample level distribution being uniform.

The above two points highlight that asymptotic normality is a fundamentally different property from a population level distribution. The similarity of form is beguiling but needs to be resisted. It is widely known that the inverse of a normal distribution does not have defined moments, and this knowledge is at the heart of any guidance to not assume normal distributions over the population for a cost coefficient used as the denominator in (2). The confusion that seems to persist in the literature, and which is at the heart of the C&C paper, is the mistaken belief that this same finding also applies to dividing by a parameter the estimate of which follows an asymptotic normal distribution. However, such theoretical insights on population level distributions of ratios based on distributed parameters must not be used to say anything about the statistical properties of these ratios, that is, their error measures, as we will now discuss, drawing on Daly et al. (2012a).

Let us consider a reparametrisation of the model by an invertible one-to-one vector function to obtain a vector η with the same dimension as β:

$$\eta = g(\beta) \text{ and } \beta = g^{-1}(\eta) \tag{5}$$

With these conditions, Cramer (1986) shows that the essential MLE properties are not affected by the transformation, so that $\hat{\eta} = g(\hat{\beta})$ is a maximum likelihood estimate of η. Greene (2008) similarly indicates that, if β is asymptotically normally distributed and $g()$ is continuous, then $g()$ is also asymptotically normally distributed.

This highlights the clear distinction between normality and asymptotic normality. If we have that a given parameter β_k follows a normal distribution in a population, then the distribution of its inverse does not have finite moments. However, if $\widehat{\beta_k}$ is a parameter estimate that is asymptotically normal, then the inverse of this parameter estimate (i.e. the estimate of its inverse) maintains the property of asymptotic normality. As stated by Daly et al. (2012a) … "*the estimate of the reciprocal has just as much status as the initial estimate"*.

In summary, because of the invertible relationship (5), we could just as easily have estimated a value for η and derived a value for β as vice versa. This point is fundamental to the understanding of the reparametrisation of models and the status of the maximum likelihood estimates. The transformation does not affect the properties of minimum variance, asymptotic consistency and asymptotic normality that apply to MLE.

Let us now consider the implications of this discussion for the work in C&C. Estimation of the specification in (3) will yield estimates for γ, β_x and any other parameters included in the model. Using the notation from (5), we then have that $\eta = -\exp(\gamma)$, that is, the function g is the negative exponential. We can then state a number of facts:

- With g being a differentiable function, and with $\hat{\gamma}$ being a MLE estimate, the property of asymptotic normality also applies to $\eta = -\exp(\gamma)$, such that $\hat{\eta} = -\exp(\hat{\gamma})$ is a MLE of η, with the associated MLE properties of minimum variance and asymptotic consistency.

- The estimate we obtain for γ (i.e. $\hat{\gamma}$) is exactly equal to $\log(-\widehat{\beta_c})$, with $\widehat{\beta_c}$ being the estimate obtained when using a model of the form in (1).
- The estimate and standard errors for any other untransformed parameters (e.g. β_x) remain the same as those obtained when using a model of the form in (1).
- The estimate for the WTP (i.e. $\frac{\widehat{\beta_x}}{-\exp(\hat{\gamma})}$) is the same as that obtained using a model of the form in (1), which gives $\frac{\widehat{\beta_x}}{\widehat{\beta_c}}$.

This discussion highlights misconceptions in C&C and already, in itself, removes the need for their proposed solution. The reasoning that the exponential avoids issues with … "*dividing two normally distributed variables*" is misguided – the distributions were asymptotically normal. The approach will yield exactly the same results with exactly the same properties, thus invalidating the claim that without the restriction, the WTP is somehow "*undefined*". More importantly, the exponential transform does not **in any way** liberate the analyst from the asymptotic normality property – if the estimate of γ is distributed asymptotically normal, then so is the estimate of $-\exp(\gamma)$.

4. Calculating measures of uncertainty for WTP

The present section looks at the key question of how to calculate measures of uncertainty for WTP. We start by looking at the computation of standard errors before discussing the computation of confidence intervals.

A simple method for calculating the standard deviation of a function of random parameters is the Delta method, which goes back at least to Bessel (1838). It is generally presented as an approximation, based on a Taylor series expansion. Statistical textbooks show that, given appropriate conditions, a first-order approximation to the error in $\eta = g(\beta)$ induced by the error in β is given by (e.g. Greene, 2008):

$$cov(\eta) = g'^T \Omega g' \tag{6}$$

where g' is the matrix of first derivatives of the function g with respect to β, and Ω is the covariance matrix of the estimates of β.

However, in the present case of MLE, these formulae can be given a different status. Cramer (1986) shows that, in the context of (5), the covariance of η as given in (6) is the Cramér-Rao lower bound, so that the estimate of η has the full MLE properties. Daly et al. (2012a) use this result to show that a likelihood function which has been maximised with respect to one set of parameters can also be considered to have been maximised with respect to parameters derived by transformations of the first set. The optimum values of the new parameters are the transformed values of the old parameters and their estimation errors are given exactly by the formulae of the Delta method, which, contrary to prevailing views in the literature (including C&C), is thus not an approximation. Therefore, a model can be estimated with a specification that is convenient and then transformed to a different specification as required, without losing the maximum likelihood properties of the estimates. A statement of this central result is given in Appendix 1.

In the context of WTP estimation, it may be convenient to estimate a model with a utility function of the form shown in (1). This would yield optimal estimates $\widehat{\beta_c}$ (which we can assume to be negative) and $\widehat{\beta_x}$, along with estimation asymptotic standard errors σ_c and σ_x and covariance $\sigma_{c,x}$ of the estimates. Equation (2) is used to calculate the WTP from this model. We then have that $g = \frac{\beta_x}{\beta_c}$, and applying (6), we can calculate the standard error of the WTP as:

$$\sigma_{WTP} = \left(\widehat{\beta_x}/\widehat{\beta_c}\right)\sqrt{\left(\frac{\sigma_x^2}{\widehat{\beta_x}^2} + \frac{\sigma_c^2}{\widehat{\beta_c}^2} - \frac{2\sigma_{c,x}}{\widehat{\beta_x}\widehat{\beta_c}}\right)} \quad (7)$$

The results of Daly et al. (2012a) show that this is an exact error for the WTP rather than an approximation. Readers can easily convince themselves of this using a simple empirical proof, by estimating a model in WTP space, as we do in Section 5. It is well known (Train & Weeks, 2005; AHCG, 1996) that the model in (1) can be rewritten as:

$$V = \beta_c.(c + v_x.x + \cdots) \quad (8)$$

in which case we would have obtained optimal estimates $\widehat{\beta_c}$ (as before) and the WTP $\widehat{v_x}$, along with their standard errors and covariance. Train & Weeks (2005) show that (8) is consistent with (1), leading to the same model fit (in fixed parameter models, and in random coefficients models with consistent distributions). Similarly, it can then easily be seen that $\widehat{v_x} = \widehat{\beta_x}/\widehat{\beta_c}$, that is, the directly estimated WTP from a model using (8) is in line with that based on the estimates of a model in preference space (2). What the results in Daly et al. (2012a) show, is that the standard error obtained for $\widehat{v_x}$ is identical to the standard error calculated using (7) for a model estimated in preference space, that is, the value that would be given by the Delta method. The model can be estimated using (1) or equivalently using (8) as is convenient.

The same reasoning applies to the C&C specification in (3). We have already discussed how the specification yields the same WTP as (1) and hence also (8). We can now also state that:

- The Delta method can be used to calculate the standard error for $\hat{\eta}$ as $\sigma_\eta = \exp(\hat{\gamma})\,\sigma_\gamma$, where $\hat{\gamma}$ is the MLE, with standard error σ_γ. The resulting standard error σ_η is the same as that obtained for $\widehat{\beta_c}$ using a model of the form in (1). The asymptotic distribution of the estimate for the parameter without a restriction ($\widehat{\beta_c}$) is thus the same as the asymptotic distribution of the restricted parameter (i.e. $-\exp(\hat{\gamma})$). How can it be that an estimate to which we apply a negative exponential, which is thus by definition negative, still has an asymptotic normal distribution? The key lies in the word **asymptotic** – the normality applies only in a small neighbourhood around the MLE.
- Applying the Delta method, we have that $\sigma_{WTP,C\&C} = \sqrt{\frac{\sigma_x^2 + \widehat{\beta_x}^2\sigma_c^2 - 2\widehat{\beta_x}\sigma_{x,\gamma}}{\exp(2\hat{\gamma})}}$, which is the same as the standard error for the WTP for models using (1), and the same as σ_{v_x} for models using (7).

It should also be noted that the use of the Delta method does not, in any way, imply that the cost coefficient needs to be fixed. If β_c and possibly also β_x are distributed across individuals, the analyst needs to first use the results in Daly et al. (2012b) to determine the existence of moments for the distribution of WTP. If these moments can be expressed as a function of estimated parameters, then an analyst can again use the Delta method. We present a simple illustration of this in Appendix 2.

In addition to the Delta method, the work of Krinsky and Robb is often cited in environmental economics, though it is far less prevalent in other areas with large choice modelling communities. It should first be noted that the findings of Krinsky and Robb (1986), are incorrect, as shown by Krinsky and Robb (1991). Their simulation procedure is lengthy but correct in principle for the cases they present and in 1991, where they find close agreement with the Delta method after correcting for a bug in their 1986 code. It is also fair to note that in neither paper do Krinsky and Robb deal with the ratio of parameters; so, they cannot be blamed for the misapplication of their approach to calculating errors in ratios such as WTP. Repeated sampling will indeed lead to problems in calculating WTP, as illustrated by C&C, and we can conclude that this approach is not satisfactory for estimating the uncertainty in WTP.

The above discussion has explained how standard errors can be calculated accurately for WTP measures, and that for this purpose, there is no need for the C&C reparametrisation. Our attention then turns to confidence intervals around the estimates. Care is again required given the asymptotic nature of the distribution of parameter estimates. The natural tendency to calculate a C% confidence interval using $\hat{\beta} \pm z^{*}\sigma_{\beta}$, where z^{*} is the upper $\frac{1-C}{2}$ critical value for a $N(0,1)$ distribution, relies on the assumption of normality, which as discussed, applies only in the close neighbourhood of $\hat{\beta}$. In the case of parameters (or functions thereof, such as WTP) with small standard errors relative to their estimates (i.e. high t-ratios), the above calculation can be acceptable. However, it is far from clear what level of statistical significance is required to make the $\hat{\beta} \pm z^{*}\sigma_{\beta}$ calculation acceptable. Other approaches for calculating confidence intervals thus deserve attention too, and, in the context of the present note, a discussion of the implications for the C&C work.

A commonly used method in environmental economics is that of Fieller (1954), which calculates the confidence limits as:

$$V_{min}, V_{max} = \frac{S_{cx} \pm \sqrt{S_{cx}^{2} - S_{cc}S_{xx}}}{S_{cc}} \tag{9}$$

where $S_{cc} = \hat{\beta}_c^2 - t^2\sigma_c^2$; $S_{xx} = \hat{\beta}_x^2 - t^2\sigma_x^2$; $S_{cx} = \widehat{\beta_c}\widehat{\beta_t} - t^2\sigma_{cx}$, and where t is the critical t value. What is often not recognised is that this assumes exact normality rather than asymptotic normality. This calculation is also offered as one approach by Armstrong et al. (2001), though their formula looks rather different.

Another approach is the 'likelihood ratio' (LR) method also proposed by Armstrong et al. (2001), who draw a parallel between the test that a statistic is outside the confidence limit and the χ^2 test that estimating a free parameter gives a significant improvement from fixing the parameter at the confidence limit. This approach is ingenious and avoids the issue caused by the denominator approaching zero, but relies heavily on knowledge of the likelihood function. For example, if the function contains local optima, it is quite possible that spurious results can be obtained. The LR approach is, as acknowledged by Armstrong et al. (2001), also computationally intensive. They find that the confidence limits given by the Fieller approach are similar to those of the LR approach, at least for the data they tested. The Fieller formula is based only on the curvature of the likelihood function at the optimum, while the LR approach looks at likelihood values away from the optimum, when the curvature of the function may change. Moreover, the incorrect assumption (by Fieller) of exact normality in the distribution of the denominator of the coefficient ratio is likely to affect the similarity more strongly in specific circumstances. Further, as shown in the Delta method discussions above, the representation (by Fieller) of a function of parameters (such as WTP) as having two dimensions of uncertainty, is arbitrary and artificial.

An approach that has received less attention than one would expect in this context is bootstrapping[3]. The Bootstrap operates by sampling N observations from the original sample, *with replacement*, repeated a number of times as chosen by the analyst, say leading to S samples $D_1, \dots, D_S$. Individual models are then estimated yielding S sets of parameter values (e.g. the vector $\widehat{\beta^{(s)}}$ in run s). 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. For that reason, they give the sampling variance that may be expected. In the context of data with multiple observations per individual, it makes most sense to sample at the level of individuals rather than observations. The covariance matrix of the parameter estimates $cov(\widehat{\beta^{(1)}}, \dots, \widehat{\beta^{(S)}})$ is calculated as the covariance of the parameters over the Bootstrap samples and an empirical confidence interval for WTP can be obtained from the distribution of $WTP^{(s)} = \frac{\widehat{\beta_x^{(S)}}}{\widehat{\beta_c^{(S)}}}, \forall s$. This process is, of course, computationally expensive with a large value for S but does not rely on any assumptions about normality, asymptotic or otherwise.

5. Empirical example

We now present a brief empirical example using a stated preference dataset collected by Axhausen et al. (2008). A set of 388 people faced 9 choices each between two public transport routes, both using train. The alternatives are described by travel time (tt), travel cost (tc), headway (hw, time between successive trains) and the number of interchanges (ch). We estimate models in preference space , WTP space, and the C&C specification, thus using the utility specifications in (1), (8) and (3), respectively. All models were estimated using Apollo (Hess & Palma, 2019).

The results of the model estimation on the full sample are shown in Table 1. For the preference space and C&C models, we also show the calculated WTP values, along with the standard errors computed using the Delta method, where, for the C&C specification we also do this for the cost coefficient. We show both 'Classical' errors, based on the inverse Hessian matrix, and 'Robust' errors, based on the 'sandwich' matrix (cf., Train, 2009; Huber, 1967).

As expected, all three models produce exactly the same log-likelihood (LL), but the specifications where the utility is not linear in parameters (WTP space and C&C) require more iterations to reach convergence. Any untransformed parameters are the same across models, such as the non-cost parameters for the first and third model, and the cost parameter for the first and second model. In addition, the implied WTP are the same in the first and third model, and equal to the estimated WTP measures in the second model (up to numerical precision). More importantly, the standard errors calculated using the Delta method for the WTP measures in the first model are equal within the accuracy of the computation to the standard errors for the directly estimated WTP in the second model. Finally, the same applies for the calculated standard errors for WTP in the C&C formulation.

[3] An alternative resampling technique that could also be used is the Jackknife (Shao & Tu, 1995).

Table 1: Full sample estimation results on Swiss data

	Preference space (Eqn 1)			WTP space (Eqn 8)			C&C specification (Equation 9)		
Estimated parameters	4			4			4		
Iterations	15			21			18		
LL(final)	-1665.688			-1665.688			-1665.688		
Adj. $\rho^2(0)$	0.3102			0.3102			0.3102		
	estimate	**classical s.e.**	**robust s.e.**	**estimate**	**classical s.e.**	**robust s.e.**	**estimate**	**classical s.e.**	**robust s.e.**
β_{tt}	-0.0598	0.0043	0.0067	-		-	-0.0598	0.0043	0.0067
β_{tc}	-0.1318	0.0135	0.0236	-0.1318	0.0135	0.0236	-		-
β_{hw}	-0.0375	0.0018	0.0023	-		-	-0.0375	0.0018	0.0023
β_{ch}	-1.1521	0.0434	0.0613	-		-	-1.1521	0.0434	0.0613
v_{tt}	-		-	0.4534	0.0285	0.0555	-		-
v_{hw}	-		-	0.2841	0.0301	0.0518	-		-
v_{ch}	-		-	8.7393	0.8993	1.5323	-		-
γ	-		-	-		-	-2.0264	0.1025	0.1791
Implied values									
β_{tc}	-		-	-		-	-0.1318	0.0135	0.0236
v_{tt}	0.4534	0.0283	0.0555	-		-	0.4534	0.0285	0.0555
v_{hw}	0.2841	0.0302	0.0518	-		-	0.2841	0.0302	0.0518
v_{ch}	8.7400	0.8996	1.5329	-		-	8.7400	0.8996	1.5329

We have already acknowledged that asymptotic normality is a large sample property, a point also noted by C&C. We therefore also investigate small sample implications. We estimate 96 models for each specification, going from the full sample to retaining only the first 20 respondents, removing around 1% of respondents from the end of the data at each step.

Given the wealth of results, we present an overview in graphical format in Figure 1, where we rely on four key metrics, showing the evolution of the log-likelihood per observation, the estimate of one of the three WTP measures, namely the value of travel time (VTT), and finally the classical and robust standard error of the VTT, multiplied by the square root of the sample size used.

For the LL per observation, we see that the results are identical for the preference space and WTP space model, with a maximum difference of less than 0.001% for the WTP space model compared to the preference space model. For the model using the C&C specification, two larger differences arise, in the models removing 88% and 95% of the sample, respectively. For these two segments, the C&C model converges to an inferior solution than the other two models, dropping 3.29% and 6.07% in LL, respectively, in comparison with the preference space model. However, far from this being a result of some benefits of the C&C specification, it seems an illustration of the numerical issues that can arise in estimation due to the exponential transform.

Figure 1 approximately here

The findings for the estimates of the WTP are in line with theory and the findings from the LL plot. The WTP space results never differ by more than 0.57% compared to preference space (with an average of 0.047% deviation). The same is true for the C&C specification, with the exception of very large deviations in the models removing 88% and 95% of the sample, respectively, where we see vast overestimations of the VTT, by 1,409% and 11,042%, respectively. Removing these two outliers gives us differences of up to 0.51% compared to the preference space model, with an average of 0.049%.

Finally, for the standard errors for WTP, we see maximum differences of 2.37% (classical) and 2.85% (robust) when comparing WTP space with preference space, with average differences of 0.19% (classical) and 0.22% (robust). For the C&C specification, we again see the impact of the two runs with poor convergence, with infinite classical standard errors and very large robust standard errors in those two cases. Removing these, we see maximum differences of 1.96% (classical) and 2.07% (robust) compared to the preference space model, with averages of 0.18% for both classical and robust.

This analysis of progressively smaller samples has highlighted that the three model specifications, except for two outliers, remain entirely consistent with each other, in line with our theoretical points. In addition, not only are the differences negligible, but there is no indication that they are anything other than white noise, and do not become larger at smaller sample sizes

We finally look at the use of bootstrapping for confidence intervals, contrasting the findings for the base model with the C&C specification (omitting the WTP space model given its theoretical equivalence to the base model). We look at the full data (388 individuals), and reduced samples containing the first 100 individuals, and the first 50 individuals, respectively. This latter is a reasonable lower bound in terms of sample size for any robust study. We use 200 Bootstrap samples in each case. The results of this process are summarised in Table 2.

For the full data case, we see that the Bootstrap means are close to the MLEs, and the Bootstrap standard deviations are very similar to the robust standard errors. In addition, there is only a little skewness in the distribution, and no excess kurtosis (with 3 corresponding to normal). For the full sample, the distribution thus remains largely symmetrical, allowing an analyst to compute confidence intervals on the basis of the standard errors (i.e. the asymptotic results). We also see no difference between the base and C&C specifications in model fit.

Once we move to smaller samples (N=100 and N=50), we see evidence of asymmetry in the distribution of Bootstrap results for the base model, along with standard deviations that quickly outstrip the asymptotic standard errors, indicating the limits of the asymptotic assumption. We also see substantial excess kurtosis and greater skewness. The negative value for skewness for the base model with N=50 deserves attention. This is caused by a small number of the samples (4 out of 200) yielding a positive cost coefficient, which, by being close to zero, gives large negative WTP. The C&C model of course prohibits such positive cost coefficients – this leads to a loss in fit for these four samples (by up to 8.41%) given that the optimum is at a value not allowed by C&C. The further undesirable side effect of this, is that for these samples the C&C cost coefficient becomes arbitrarily close to zero, leading to an explosion in WTP, and making the standard deviation and confidence intervals unusable (exactly the problem that C&C are trying to avoid). Problems for the C&C specification in fact also arise with N=100. Even though none of the Bootstrap samples lead to positive cost coefficients with the base specification, C&C yields some outlying values, most likely due to slightly inferior solutions, a problem we attribute to the use of the exponential transform and the resulting numerical problems.

These findings suggest that, in those cases where the C&C specification is in line with the standard model, it offers no benefits. But those cases where the C&C results differ from the standard specification show that the model fit in repeated sampling is worse than for the base model, and the implied confidence intervals are not useful.

Table 2: Bootstrap results

		N=388		N=100		N=50	
		Base	C&C	Base	C&C	Base	C&C
MLE	$\widehat{WTP}$	27.21	27.21	23.43	23.43	34.24	34.25
	σ_{WTP} (robust s.e.)	3.33	3.33	5.32	5.32	13.88	13.90
Bootstrap results	Mean	27.84	27.85	24.79	43.20	34.65	636.33
	Median	27.57	27.57	23.85	23.85	32.97	33.44
	Standard deviation	3.26	3.26	7.08	117.30	38.67	6,655.84
	2.5^{th} percentile	22.40	22.39	15.54	15.54	17.50	19.76
	97.5^{th} percentile	34.75	34.75	40.47	441.14	95.00	3,176.13
	Skewness	0.37	0.37	2.16	6.87	-5.52	13.83
	Kurtosis	2.99	2.99	12.70	52.87	55.17	194.12
	$mean\left(\frac{LL_{C\&C} - LL_{base}}{LL_{base}}\right)$	-	0.00%	-	0.02%	-	0.11%
	$max\left(\frac{LL_{C\&C} - LL_{base}}{LL_{base}}\right)$	-	0.00%	-	1.25%	-	8.41%

6. Conclusions

In this note, we have addressed a number of misconceptions in the C&C paper and set out in general how modellers can deal with the issues that arise in estimating WTP and its accuracy.

The fundamental issue is that C&C (though our experience at conferences is that they are by no means alone in this) seemingly fail to appreciate the difference between asymptotic normality and normality. The lesser property of maximum likelihood estimates, that they are distributed asymptotically normal, cannot justify imposing a full normal distribution. It is the assumption of a full normal distribution that causes the problem that the denominator of the WTP ratio seems to be close to zero.

In contrast, the result in Daly et al. (2012a) shows that the same model could have been estimated with a different specification, in which the WTP is estimated directly and therefore itself has an asymptotic normal distribution. This apparent paradox is caused by the fact that the estimates are not distributed normal, but only asymptotically normal. Similarly, the model specification proposed by C&C also leads to the same model. Once again, the errors can be calculated exactly using the Delta method, giving the same standard errors as for preference and WTP space. Despite imposing a negative exponential to ensure a strictly signed cost coefficient, this cost coefficient also again has the property of asymptotic normality, as does the WTP.

Given these theoretical points, and our empirical results in Section 5, we therefore conclude that the model specification proposed by C&C is not technically incorrect, but it is both unnecessary and misleading. It is unnecessary because simpler formulations, less prone to practical difficulties, can give the same result and both the estimates and the estimation errors can be transformed as required. It is misleading because it suggests that there is an issue with the ratio calculation. Moreover, the negative exponential transformation of C&C, while indeed imposing a constraint implied by economic theory, causes difficulties by potentially obscuring data or specification issues and possibly introducing numerical problems into the estimation process.

In determining the most appropriate way to estimate the accuracy of WTP measures, it is useful to consider the reasons for requiring to know the accuracy.

- To test whether WTP is significantly different from zero is better and more easily done by testing the coefficient of the specific attribute, i.e. β_x in the equations above.
- The standard error of WTP may be needed to give a general indication of its accuracy, for comparison with other models and other studies, for example; in this case the Delta method gives a simple and suitable calculation.
- It may be required to indicate confidence limits for WTP, for example to give ranges for which a specific policy might be appropriate; in this case the Delta method can also be used, but it must be recognised that this gives only asymptotic results. When the estimation accuracy is low, alternative methods may be needed.

It is clear, and generally agreed, that when the relevant parameters are estimated with reasonable accuracy, any of the methods proposed will give reasonable results in good agreement with the other methods. Errors implying t ratios around 6 or 8 might be considered acceptable in this context[4]. In such cases, the Delta method can generally be used, as it is the simplest of the methods.

To enhance the applicability of the Delta method, we present a theorem (see Appendix 1), which shows that Delta calculations can be applied to ‘robust’ error measures derived from the ‘sandwich’ matrix, as well as to classical errors derived from the inverted Hessian matrix. These robust error measures have a wider applicability than classical errors, in terms of the assumptions needed.

When the parameters are less accurately estimated it is more difficult to choose an appropriate approach. Asymptotic approaches will be less appropriate, as the confidence limits will lie further from the optimum and the assumption of a quadratic log likelihood function will be less satisfactory. The respecification suggested by C&C does not help with this.

In these circumstances it may be useful to use the likelihood ratio approach of Armstrong et al. (2001) or a resampling approach such as bootstrapping, as in Section 5. While these approaches require considerable calculation, they are not asymptotic. The Armstrong et al. (2001) approach depends only on the correctness of the likelihood function, while bootstrapping is even more general.

[4] It should be noted that when the coefficients are positively correlated, then the WTP can be more accurately estimated than either of the components in the ratio. The opposite could apply in the case of negative correlation.

Finally, we should note that in some cases the accuracy of estimation of WTP may be inadequate for the purpose. Even a t ratio of 4, which may well be thought to show that the WTP is significantly different from zero, implies confidence limits of roughly ±50% and may well be inadequate as a basis for important investment or policy decisions.

Acknowledgements

Stephane Hess was supported by the European Research Council through the consolidator grant 615596-DECISIONS. Juan de Dios Ortúzar wishes to acknowledge the support of Instituto Sistemas Complejos de Ingeniería (ANID PIA/BASAL AFB180003) and the BRT+ Centre of Excellence funded by the Volvo Research and Educational Foundations (www.brt.cl).

References

AHCG (1996). The Value of Travel Time on UK Roads. Report to UK Department of Transport, Accent and Hague Consulting Group, London.

Armstrong, P.M., Garrido, R.A. & Ortúzar, J. de D. (2001) Confidence intervals to bound the value of time. *Transportation Research Part E: Logistics and Transportation Review* **37**, 143-161.

Axhausen, K.W., Hess, S., König, A., Abay, G., Bates, J.J. & Bierlaire, M. (2008). State of the art estimates of the Swiss value of travel time savings. *Transport Policy* **15**, 173-185.

Beesley, M. E. (1965). The value of time spent in travelling: some new evidence. *Economica* **32**, 174–185.

Berndt, E., Hall, B., Hall, R. & Hausman, J. (1974). Estimation and inference in nonlinear structural models, *Annals of Economic and Social Measurement* **3**, 653–665.

Bessel, F.W. (1838). Untersuchungen uber die Wahrscheinlichkeit der Beobachtungsfehler. *Astronomische Nachrichten* **15**, 369-404 (in German).

Carson, R.T. & Czajkowski, M. (2019). A new baseline model for estimating willingness-to-pay from discrete choice models. *Journal of Environmental Economics and Management* **95**, 57-61.

Cramer, J.S. (1986). *Econometric Applications of Maximum Likelihood Methods*. Cambridge University Press, Cambridge.

Daly, A. & Zachary, S. (1975). Commuters' values of time. *LGORU Report T55*, Local Government Operational Research Unit, Reading. Available at: (http://www.alogit.com/papers/CommutersValues.pdf.)

Daly, A., Hess, S. & de Jong, G. (2012a). Calculating errors for measures derived from choice modelling estimates. *Transportation Research Part B: Methodological* **46**, 333-341.

Daly, A., Hess, S. & Train, K. (2012b). Assuring finite moments for willingness to pay in random coefficient models. *Transportation* **39**, 19-31.

Fieller, E.C. (1954). Some problems in interval estimation. *Journal of the Royal Statistical Society*, *Series B* **16**, 175-185.

Gaudry, M.J.I., Jara-Díaz, S.R. & Ortúzar, J. de D. (1989). Value of time sensitivity to model specification. *Transportation Research Part B: Methodological* **23**, 151-158.

Greene, W.H. (2008). *Econometric Analysis* (Sixth Edition). Pearson Education Inc, Upper Saddle River, NJ.

Hess, S., Bierlaire, M. & Polak, J.W. (2005). Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice* **39**, 221-236.

Hess, S. & Palma, D. (2019). Apollo: a flexible, powerful and customisable freeware package for choice model estimation and application. *Journal of Choice Modelling* **32**, 100170.

Huber, P.J. (1967). The behaviour of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. I, 221–233.

Palma, D., Ortúzar, J. de D., Rizzi, L.I., Guevara, C.A., Casaubon, G. & Ma, H. (2016). Modelling choice when price is a cue for quality: a case study with Chinese consumers. *Journal of Choice Modelling* **19**, 24-39.

Shao, J. & Tu, D. (1995). *The Jackknife and Bootstrap.* Springer-Verlag, New York.

Sillano, M. & Ortúzar, J. de D. (2005). Willingness-to-pay estimation with mixed logit models: some new evidence. *Environment and Planning PartA: Economy and Space* **37**, 525-550.

Train, K. & Weeks, M. (2005). Discrete choice models in preference space and willingness-to-pay space. In R. Scarpa & A. Alberini (eds), *Application of Simulation Methods in Environmental and Resource Economics*, 1-16. Springer Publisher, Dordrecht.

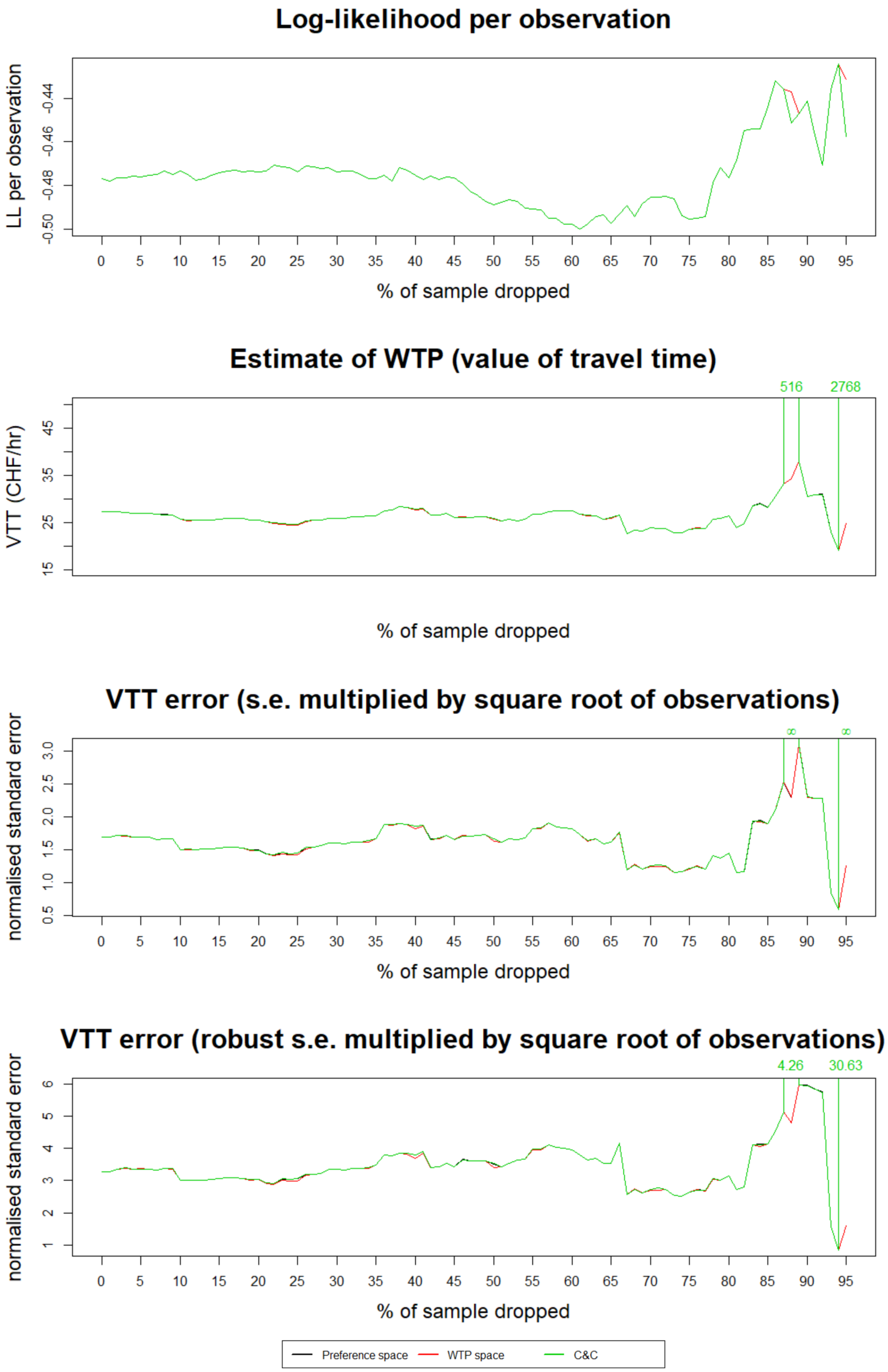

Figure 1: Small sample properties

Appendix 1

The theorems given in this Appendix justify the use of the Delta method with both classical and robust errors.

Classical Theorem

Let $\hat{\beta}$ be a correctly-specified maximum likelihood estimator of a vector with true value β^* of dimension L and let Ω be the covariance matrix of $\hat{\beta}$ around β^*, given by the inverse of the negative of the Hessian. Let $\Phi: R_L \rightarrow R_L$ be a differentiable and invertible function. Then:

1. $\Phi^* = \Phi(\beta^*)$ is the true value of $\Phi(\beta^*)$;
2. $\hat{\Phi} = \Phi(\hat{\beta})$ is a maximum likelihood estimator of Φ^*; and
3. the covariance matrix of $\hat{\Phi}$ around Φ^* attains the Cramér-Rao lower bound and is given by $cov(\Phi) = \Psi = \Phi'^T \Omega \Phi'^*$, where Φ' is the first derivative matrix of Φ.

Proof: outlined in Daly et al. (2012a) and given more completely in Cramer (1986).

The theorem above applies only when the model is correctly specified. However, provided the likelihood scores are zero at the true values of the parameters, we can use an alternative error estimate, usually termed the 'robust' error matrix (for a discussion see, e.g., Train, 2009). Using the same notation as in the Classical Theorem, the robust matrix is given by the 'sandwich' formula

$$cov^{(S)}(\hat{\beta}) = \Omega^{(S)} = \Omega\, B\, \Omega, \qquad \text{[A1]}$$

where B is the 'BHHH' matrix (Berndt et al., 1974), given by $b_{jk} = \sum_n b_{jk}^n = \sum_n \frac{\partial L^n}{\partial \beta_j}\frac{\partial L^n}{\partial \beta_k}$, that is, the sample covariance of the scores. We use the notation of lower-case letters to indicate elements of vectors or matrices indicated by the corresponding upper-case symbols.

The following theorem allows us to apply the Delta method to the robust matrix.

Robust Theorem

In the context defined above for the Classical Theorem, $\Phi^{'T} \Omega^{(S)}\, \Phi^{'}$ is the robust error matrix for $cov^{(S)}(\hat{\Phi})$ for maximum likelihood estimation over Φ.

Proof: The robust error matrix with respect to Φ is given by

$$\Psi^{(S)} = XMX \qquad \text{[A2]}$$

where X and M are respectively the classical error matrix and the BHHH matrix, in each case defined with respect to Φ. From the classical theorem we know that the transformation of the classical matrix from β to Φ gives $X = \Psi$ as defined above. The required BHHH matrix M is given by the sample covariance of the scores with respect to Φ

$$M = [m_{jk}] = [\sum_n m_{jk}^n] = \left[\sum_n \frac{\partial L^n}{\partial \phi_j}\frac{\partial L^n}{\partial \phi_k}\right], \qquad \text{[A3]}$$

using the notation that $[x_{jk}]$ is the matrix whose jk^{th} element is x_{jk} and n to index the sample. Now, by the chain rule,

$$\frac{\partial L^n}{\partial \beta_r} = \sum_i \frac{\partial L^n}{\partial \phi_i} . \phi'_{ri} \qquad \text{[A4]}$$

Hence, the β-based BHHH matrix can be expressed in terms of ϕ derivatives:

$$b^n_{rs} = \frac{\partial L^n}{\partial \beta_r}\frac{\partial L^n}{\partial \beta_s} = \left(\sum_i \frac{\partial L^n}{\partial \phi_i} . \phi'_{ri}\right)\left(\sum_i \frac{\partial L^n}{\partial \phi_i} . \phi'_{si}\right) = \sum_j \sum_k \phi'_{rj}\left(\frac{\partial L^n}{\partial \phi_j}\frac{\partial L^n}{\partial \phi_k}\right)\phi'_{sk} \qquad \text{[A5]}$$

Noting that Φ does not vary with n, terms can be summed and rearranged

$$b_{rs} = \sum_n b^n_{rs} = \sum_n \sum_j \sum_k \phi'_{rj}\left(\frac{\partial L^n}{\partial \phi_j}\frac{\partial L^n}{\partial \phi_k}\right)\phi'_{sk} = \sum_j \sum_k \phi'_{rj} m_{jk}\, \phi'_{sk} \qquad \text{[A6]}$$

B can therefore be expressed as a matrix multiplication[5]

$$B = \Phi'^T M\, \Phi', \text{ i.e. } M = (\Phi'^T)^{-1} B (\Phi')^{-1} \qquad \text{[A7]}$$

So, we can now write the required matrix

$$\Psi^{(S)} = XMX = \Psi\left(\Phi'^T\right)^{-1} B \left(\Phi'\right)^{-1} \Psi = \Phi'^T \Omega \Phi' \left(\Phi'\right)^{-1} B \left(\Phi'^T\right)^{-1} \Phi'^T \Omega \Phi'$$

$$= \Phi'^T \Omega B \Omega \Phi' = \Phi'^T \Omega^{(S)} \Phi' \qquad \text{[A8]}$$

which is the Φ transformation of the original robust error matrix as required.

Appendix 2

To illustrate the use of the Delta method in the case of random coefficients, let us look at the simple case of using two independent negative Lognormal distributions, with:

$$\log(-\beta_x) \sim N\left(\mu_{log-\beta_x}, \sigma_{log-\beta_x}\right) \ \& \ \log(-\beta_c) \sim N\left(\mu_{log-\beta_c}, \sigma_{log-\beta_c}\right), \qquad \text{[A9]}$$

i.e. the logarithms of the negatives of the two coefficients follow a normal distribution.

The ratio of two Lognormal distributions is itself a Lognormal distribution, such that:

$$\log(WTP) \sim N\left(\mu_{logWTP}, \sigma_{logWTP}\right), \qquad \text{[A10]}$$

with $\mu_{logWTP} = \mu_{log-\beta_x} - \mu_{log-\beta_c}$ and $\sigma_{logWTP=}\sqrt{\sigma^2_{log-\beta_x} + \sigma^2_{log-\beta_c}}$. With maximum likelihood estimates $\widehat{\mu_{log-\beta_x}}$, $\widehat{\sigma_{log-\beta_x}}$, $\widehat{\mu_{log-\beta_c}}$ and $\widehat{\sigma_{log-\beta_x}}$, we then now that $\widehat{\mu_{logWTP}}$ and $\widehat{\sigma_{logWTP}}$ have these same properties, and using the Delta method, we have that:

[5] Note that in the matrix multiplication $A = B^T CD$, elements of A are given by $a_{ij} = \sum_k \sum_l b_{ki} c_{kl} d_{lj}$.

$$s.e.\left(\widehat{\mu_{logWTP}}\right) = \sqrt{var(\widehat{\mu_{log-\beta_x}}) + var(\widehat{\mu_{log-\beta_c}}) - 2cov(\widehat{\mu_{log-\beta_x}}, \widehat{\mu_{log-\beta_c}})} \qquad [A11]$$

and

$$s.e.\left(\widehat{\sigma_{logWTP}}\right) = \sqrt{\frac{\widehat{\sigma_{log-\beta_x}}^2}{\widehat{\sigma_{log-\beta_x}}^2 + \widehat{\sigma_{log-\beta_c}}^2} var(\widehat{\sigma_{log-\beta_x}}) + \frac{\widehat{\sigma_{log-\beta_c}}^2}{\widehat{\sigma_{log-\beta_x}}^2 + \widehat{\sigma_{log-\beta_c}}^2} var(\widehat{\sigma_{log-\beta_c}}) + 2\frac{\widehat{\sigma_{log-\beta_x}}\widehat{\sigma_{log-\beta_c}}}{\widehat{\sigma_{log-\beta_x}}^2 + \widehat{\sigma_{log-\beta_c}}^2} cov(\widehat{\sigma_{log-\beta_x}}, \widehat{\sigma_{log-\beta_c}})} \qquad [A12]$$

where var and cov relate to the variance (i.e. square of the standard error) and covariance of MLEs, rather than the moments of the population level distribution.