

A COMPARISON OF APPROACHES TO ESTIMATING CONFIDENCE INTERVALS FOR WILLINGNESS TO PAY MEASURES

ARNE RISA HOLE*

National Primary Care Research and Development Centre, Centre for Health Economics, University of York, UK

SUMMARY

This paper describes four approaches to estimating confidence intervals for willingness to pay measures: the delta, Fieller, Krinsky Robb and bootstrap methods. The accuracy of the various methods is compared using a number of simulated datasets. In the majority of the scenarios considered all four methods are found to be reasonably accurate as well as yielding similar results. The delta method is the most accurate when the data is well-conditioned, while the bootstrap is more robust to noisy data and misspecification of the model. These conclusions are illustrated by empirical data from a study of willingness to pay for a reduction in waiting time for a general practitioner appointment in which all the methods produce fairly similar confidence intervals. Copyright © 2007 John Wiley & Sons, Ltd.

Received 11 September 2005; Revised 10 August 2006; Accepted 27 October 2006

KEY WORDS: willingness to pay; confidence interval; delta method; bootstrap

INTRODUCTION

It is well-known that the marginal rate of substitution between two attributes in a discrete choice model is given by the ratio of the attribute coefficients when the model is linear in the attributes. This result is frequently used in the discrete choice experiment (DCE) literature to derive estimates of willingness to pay (WTP) for an improvement in a given attribute. For a recent review of studies applying DCEs in health care see Ryan and Gerard (2003).

Most analysts are aware that since WTP is derived as the ratio of two random variables, WTP is itself a random variable. In spite of this, however, standard errors and confidence intervals for WTP estimates are rarely reported in applied work (for an exception see McIntosh and Ryan, 2002). This paper describes four approaches to estimating confidence intervals for willingness to pay measures: the delta, Fieller, Krinsky Robb and bootstrap methods, and compares their accuracy using simulated data. Although the issues surrounding confidence intervals for WTP measures are similar to those discussed in the large literature on confidence intervals for cost effectiveness ratios (see e.g. Briggs *et al.*, 1999; Chaudhary and Stearns, 1996), the data generating processes are fundamentally different. It is therefore not given *a priori* that the small sample properties of the various methods will be the same. Consequently, I argue that it is necessary to investigate the methods' suitability for estimating confidence intervals for WTP measures by explicitly simulating the data generating process relevant for DCE studies.

*Correspondence to: National Primary Care Research and Development Centre, Centre for Health Economics, Alcuin 'A' Block, University of York, York YO10 5DD, UK. E-mail: ah522@york.ac.uk

The accuracy of the various methods is compared using a number of simulated datasets with varying characteristics. In the majority of the cases considered all four methods are found to be reasonably accurate as well as yielding similar results. While the bootstrap is found to be the least accurate method when the model is correctly specified and the data well-conditioned, it has the advantage of being the only method which is robust to ignoring unobserved heterogeneity when present. The findings of an empirical application support the conclusions drawn from the simulation study in that all the methods produce fairly similar confidence intervals.

A recent discussion in the literature (Lancsar and Savage, 2004; Ryan, 2004; Santos Silva, 2004) distinguishes between ‘state-of-the-world’ and ‘multiple alternative’ models and show that, in general, marginal rates of substitution are appropriate welfare measures for ‘state-of-the-world’ models only.¹ Since the present paper focuses on willingness to pay estimates derived from marginal rates of substitution the findings are relevant for ‘state-of-the-world’ rather than ‘multiple-alternative’ models, since the marginal rates of substitution do not have a valid welfare interpretation in the latter case.

The paper is organised as follows: the next section provides an outline of random utility maximisation and the logit model, the following section describes the various methods for estimating confidence intervals, the next following section describes the simulated data, the succeeding sections present the simulation results and the empirical application, while the penultimate section provides a discussion. The final section offers some concluding remarks.

RANDOM UTILITY MAXIMISATION AND THE LOGIT MODEL

I assume a sample of N consumers with the choice of J discrete alternatives in T choice scenarios. Let U_{njt} be the utility individual n derives from choosing alternative j in choice scenario t . It is assumed that the utility can be partitioned into a systematic component or ‘representative utility’, V_{njt} , and a random component, ε_{njt} , such that:

$$U_{njt} = V_{njt} + \varepsilon_{njt} \quad (1)$$

The systematic component, V_{njt} , is a function of the attributes of alternative j while ε_{njt} represents characteristics and attributes unknown to the researcher, measurement error and/or heterogeneity of tastes in the sample. Since the unknown variable, ε_{njt} , is treated as random by the researcher, this class of utility models is called random utility models. The probability that individual n chooses alternative i rather than alternative j is the probability that the utility of choosing i is higher than the utility of choosing j :

$$P_{nit} = P(U_{nit} > U_{njt}) = P(V_{nit} + \varepsilon_{nit} > V_{njt} + \varepsilon_{njt}) = P(\varepsilon_{njt} - \varepsilon_{nit} < V_{nit} - V_{njt}) \quad (2)$$

Assuming that the difference of the random terms, $\varepsilon_{nt} = \varepsilon_{njt} - \varepsilon_{nit}$, is logistically distributed and the number of alternatives, $J = 2$, we get the binomial logit model (see e.g. Ben-Akiva and Lerman, 1985) in which the probability that alternative i is chosen in scenario t is given by

$$P_{nit} = \frac{1}{1 + e^{-\mu(V_{nit} - V_{njt})}} \quad (3)$$

where μ is a positive scale parameter which can be shown to be inversely proportional to the error variance, σ_ε^2 :

$$\mu = \frac{\pi}{\sqrt{3}\sigma_\varepsilon^2} \quad (4)$$

¹In ‘state-of-the-world’ models the respondents are assumed to choose the same alternative following a change in an alternative attribute, while in ‘multiple alternative’ models they may switch to another alternative. See Lancsar and Savage (2004), Ryan (2004) and Santos Silva (2004) for a discussion of the implications of these assumptions for constructing welfare measures.

The representative utility, V_{njt} , is usually specified to be linear in the alternative attributes:

$$V_{njt} = \beta_{0i} + \beta_1 X_{1njt} + \cdots + \beta_K X_{Knjt} + \beta_C C_{njt} \quad (5)$$

where β_{0i} is a constant which reflects the mean impact of the unobservable components on the utility of alternative i . $\beta_1, \dots, \beta_K$ are the coefficients for attributes $X_1, \dots, X_K$ and β_C is the coefficient for the cost of the alternatives.

The total derivative of U_{njt} with respect to changes in attribute X_k and cost is given by $dU_{njt} = \beta_k dX_k + \beta_C dC$. Setting this expression equal to zero and solving for dC/dX_k yields the change in cost that keeps utility unchanged given a change in X_k :

$$\frac{dC}{dX_k} = \text{WTP}_k = -\frac{\beta_k}{\beta_C} \quad (6)$$

which equals the willingness to pay for an improvement in X_k , WTP_k . It can be seen from Equation (6) that WTP_k is given by the negative of the ratio of the coefficients for X_k and C .²

Since the logit model is typically estimated using maximum likelihood, which implies that the coefficients in the model are asymptotically normally distributed, it is reasonable to assume that WTP is given by the ratio of two normally distributed variables when the model is estimated using a large sample. The distribution of the ratio of two normally distributed variables has been derived by Fieller (1932) and Hinkley (1969), who show that the distribution is approximately normal when the coefficient of variation of the denominator variate (in this case β_C), is negligible. In other words, if the ratio of the standard error of β_C to its mean is low, the distribution of WTP is likely to be approximately normal (note that the t -statistic, which may be a more intuitive way of representing the precision of the coefficient estimate, is simply the inverse of the coefficient of variation). As will become clear in the following section, this result is of importance when comparing the various approaches to estimating confidence intervals for WTP .

WTP CONFIDENCE INTERVALS

The delta method

The delta method estimate of the variance of a non-linear function of two (or more) random variables is given by taking a first-order Taylor expansion around the mean value of the variables and calculating the variance for this expression (see e.g. Greene, 2003). In the case of WTP the delta estimate of the variance is given by:

$$\begin{aligned} \text{var}(\hat{\text{WTP}}_k) &= [(\hat{\text{WTP}}_{\beta_k})^2 \text{var}(\hat{\beta}_k) + (\hat{\text{WTP}}_{\beta_C})^2 \text{var}(\hat{\beta}_C) \\ &\quad + 2\hat{\text{WTP}}_{\beta_k} \hat{\text{WTP}}_{\beta_C} \text{covar}(\hat{\beta}_k, \hat{\beta}_C)] \\ &= [(-1/\hat{\beta}_C)^2 \text{var}(\hat{\beta}_k) + (\hat{\beta}_k/\hat{\beta}_C^2)^2 \text{var}(\hat{\beta}_C) \\ &\quad + 2(-1/\hat{\beta}_C)(\hat{\beta}_k/\hat{\beta}_C^2) \text{covar}(\hat{\beta}_k, \hat{\beta}_C)] \end{aligned} \quad (7)$$

where $\hat{\text{WTP}}_{\beta_k}$ and $\hat{\text{WTP}}_{\beta_C}$ are the partial derivatives of WTP_k w.r.t. β_k and β_C , respectively, evaluated at the estimates. The confidence interval can then be created in the standard fashion:

$$\hat{\text{WTP}}_k \pm z_{\alpha/2} \sqrt{\text{var}(\hat{\text{WTP}}_k)} \quad (8)$$

²This is a consequence of the representative utility being linear in the alternative attributes. Evaluating the properties of the various methods for constructing confidence intervals for WTP measures derived from non-linear utility functions is beyond the scope of the present paper.

where $z_{\alpha/2} = \Phi^{-1}[1 - \alpha/2]$, Φ^{-1} is the inverse of the cumulative standard normal distribution and the confidence level is $100(1 - \alpha)\%$. This assumes that WTP is normally distributed and thus symmetrical around its mean. As discussed in the previous section it is likely that WTP is approximately normally distributed when the model is estimated using a large sample *and* the estimate of the coefficient for the cost attribute is sufficiently precise. The assumption of normality is clearly strong, however, as there is no guarantee that WTP will be normally distributed if these conditions do not hold. There is little theory to inform us as to what distribution WTP will have if the coefficients are not normally distributed, which may be expected if the model is estimated using smaller samples. Shanmugalingham (1982) has conducted some Monte Carlo experiments to investigate how the shape of the distribution of the ratio of two normal variables is affected by the relative magnitude of the mean and standard deviation of the variables as well as the correlation between them, and finds that in many cases the distribution is far from normal. In particular, when the standard deviation of the denominator variable is large relative to its mean the distribution will be skewed. This suggests that when the cost coefficient is not precisely estimated the delta method confidence interval may be inaccurate, since it will not reflect the skewness of the distribution of WTP.

The Fieller method

Since WTP can be assumed to be a ratio of two normally distributed variables the method suggested by Fieller (1954) provides an alternative to the delta method. The Fieller confidence interval for $\hat{WTP}_k$ is given by

$$\frac{\hat{\beta}_k - z_{\alpha/2}^2 \text{covar}(\hat{\beta}_k, \hat{\beta}_C) / \hat{\beta}_C \pm \Omega^{1/2}}{-\hat{\beta}_C + z_{\alpha/2}^2 \text{var}(\hat{\beta}_C) / \hat{\beta}_C} \quad (9)$$

where $\Omega = (\hat{\beta}_k - z_{\alpha/2}^2 \text{covar}(\hat{\beta}_k, \hat{\beta}_C) / \hat{\beta}_C)^2 - (1 / \hat{\beta}_C^2)(\hat{\beta}_k^2 - z_{\alpha/2}^2 \text{var}(\hat{\beta}_k))(\hat{\beta}_C^2 - z_{\alpha/2}^2 \text{var}(\hat{\beta}_C))$. The advantage of Fieller's method relative to the delta method is that it does not assume that $\hat{WTP}_k$ is normally distributed. The only assumption required is that the coefficients are joint normally distributed, which may not be unrealistic when the sample is relatively large. It can be seen from (9) that the confidence interval will not be symmetric around $\hat{WTP}_k$ in general, but will tend towards symmetry when $\text{var}(\hat{\beta}_C) / \hat{\beta}_C$ and $\text{covar}(\hat{\beta}_k, \hat{\beta}_C) / \hat{\beta}_C$ approach zero. This suggests that the Fieller method will yield more accurate confidence intervals than the delta method when WTP is not symmetrically distributed. A drawback of the Fieller method is that the confidence limits are imaginary when Ω is negative. While this is unlikely when the coefficients of variation for $\hat{\beta}_k$ and $\hat{\beta}_C$ are lower than 0.3 (Cochran, 1977), it cannot be ruled out in practice. Moreover, when the denominator in (9) is negative the confidence interval is given by the union of the intervals $(-\infty, L^+)$ and $(L^-, +\infty)$, where L^- and L^+ denote (9) with $\pm$ set to $-$ and $+$, respectively. This is arguably not a major issue since the denominator will only be negative if $\hat{\beta}_C$ is insignificant, which is a relatively infrequent occurrence in DCE applications.

The Krinsky Robb method

Krinsky and Robb (1986, 1990) suggest an alternative method for estimating confidence intervals which is based on taking a large number of draws from a multivariate normal distribution with means given by the estimated coefficients and covariance given by the estimated covariance matrix of the coefficients. This method is also referred to as the parametric bootstrap (Efron and Tibshirani, 1993). Based on R draws taken from the joint distribution of the coefficients, R simulated values of WTP are calculated. These R values can then be used to calculate the percentiles of the simulated distribution reflecting the desired level of confidence. For instance, if 1000 simulated values of WTP are estimated, the lower and upper limits of a 95% confidence interval are given by the 26th and 975th sorted estimates of WTP, respectively. Confidence intervals derived in this fashion are usually referred to as *percentile intervals*.

(Efron and Tibshirani, 1993; Mooney and Duval, 1993). The confidence interval could also be derived by using the draws to calculate the variance of WTP and plugging the estimated variance into Equation (8), but this approach, like the delta method confidence interval, hinges on the assumption that WTP is symmetrically distributed. The percentile interval, on the other hand, does not assume that WTP is symmetrically distributed. Like the Fieller method the only assumption required is that the coefficients are joint normally distributed. The advantage of the Krinsky Robb method relative to the Fieller method is that it produces confidence intervals which are defined in all samples. The disadvantage is that it is more computationally demanding, since it requires a large number of draws being taken from the joint distribution of the coefficients. Considering the speed of modern PCs, however, this is not a major obstacle.

The bootstrap

The bootstrap (Efron, 1979; Efron and Tibshirani, 1993; Mooney and Duval, 1993) has been used extensively to estimate standard errors and confidence intervals in economics in recent years (see e.g. Horowitz, 2001). The bootstrap is similar to the Krinsky Robb method in that a simulated distribution for the variable of interest is generated. In contrast to the Krinsky Robb method, however, the bootstrap makes no assumptions about the distribution of the coefficients in the model. The simulated distribution of WTP is generated by drawing a large number of samples of size N (with replacement) from the estimation sample. Each of these samples are used to derive an estimate of WTP by estimating the model and calculating WTP using Equation (6). The confidence interval can then be derived in an analogous fashion to the Krinsky Robb percentile interval. Again, an alternative would be to calculate the variance of the simulated distribution and plug this into Equation (8), but as discussed above this approach is likely to be less accurate since it imposes the additional assumption of symmetry.

The bootstrap, therefore, has the same advantage as the Fieller and Krinsky Robb methods in that it does not rely on the assumption that WTP is symmetrically distributed, but unlike these methods it does not require that the coefficients themselves are joint normally distributed. It is therefore possible that the bootstrap will perform better than the Fieller and Krinsky Robb methods when the sample size is small. The bootstrap is by far the most computationally demanding method, however, since it requires that the model is re-estimated for each bootstrap sample. The gains of the bootstrap must therefore be weighed against the additional computational cost it imposes on the analyst.

THE SIMULATED DISCRETE CHOICE EXPERIMENT

To compare the various approaches to constructing confidence intervals for WTP described in the previous section numerous artificial datasets are constructed. Common to all the datasets is the assumption that a number of hypothetical individuals are presented with a set of scenarios in which they must choose between two alternatives which differ only in three attributes: X_1 , X_2 and cost. X_1 and X_2 are two-level attributes while cost has four levels. A summary of the attributes and their respective levels is given in Table I.

The full factorial design is given by $4^2 \times 2^4 = 256$, which was reduced to a design with 16 choice scenarios using the MKTEX SAS macro (Kuhfeld, 2005) in order to keep the data as similar as possible to an actual choice experiment. Suppressing the individual and scenario subscripts for simplicity, the difference in representative utility between choosing alternative 1 and 2, respectively, is given by

$$V_1 - V_2 = \beta_0 + \beta_1(X_{11} - X_{12}) + \beta_2(X_{21} - X_{22}) + \beta_C(C_1 - C_2) \quad (10)$$

where X_{1j} , X_{2j} and C_j are the values of attributes X_1 , X_2 and cost for alternative j , respectively. The values of the coefficients are set to $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$ and $\beta_C = -1$. It follows that the

Table I. The attributes of the simulated discrete choice experiment

Attribute	Levels	Coding
X_1	Low, High	1,2
X_2	Low, High	1,2
Cost	£1, £ 2, £ 3, £ 4	1,2,3,4

willingness to pay for an improvement in attribute X_1 (WTP_1) is £ 1 and the willingness to pay for an improvement in attribute X_2 (WTP_2) is £ 0.5. The final step necessary in order to create the simulated data is to take a number of draws from the logistical distribution where each draw represents the error difference for a hypothetical individual in a given choice scenario. If the error difference is less than the difference in representative utility, $V_1 - V_2$, alternative 1 is chosen. Otherwise, alternative 2 is chosen.

SIMULATION RESULTS

As discussed in the previous sections there are a number of factors which are expected to influence the accuracy of the confidence intervals. The delta, Fieller and Krinsky Robb methods assume that the coefficients in the model are normally distributed, which is influenced by the sample size. In addition the delta method assumes that WTP itself is normal which also requires that the precision of the cost coefficient estimate is sufficiently high. The precision of the coefficients depends on the sample size as well as the amount of ‘noise’ in the data, or in other words, the magnitude of the error variance. These two factors are considered in turn in the next two subsections and the next following subsection considers the impact of neglected unobserved heterogeneity in the model. Since neglecting unobserved heterogeneity will lead to biased estimates of the coefficient standard errors, confidence intervals based on these standard errors will also be biased. It is therefore expected that the bootstrap will be the superior method in this case, since it is the only method that does not rely on the estimated covariance matrix.

The impact of changes in the sample size

Four different sample sizes are considered, $N = 10, 25, 50$ and 100 . Since each hypothetical respondent ‘completes’ 16 choice scenarios, however, the total number of observations are 160, 400, 800 and 1600, respectively. For now it is assumed that the logit model is the correct specification. The scale parameter is set equal to unity, which is equivalent to an error variance of $\pi^2/3$. The results for WTP_1 and WTP_2 are given in Tables II and III, respectively.

The first two columns in the tables give the sample size and the method used for calculating the confidence interval. Both the Krinsky Robb and bootstrap confidence intervals are derived using the percentile method based on 1000 resamples in each trial.³ Columns 3 and 4 give the lower and upper limits of the estimated 95% confidence intervals averaged over 10 000 trials. The Monte Carlo estimates of the confidence limits are derived by calculating the 2.5 and 97.5 percentiles of the 10 000 WTP estimates. These estimates serve as a benchmark for the accuracy of the other methods. The figures in brackets are the mean squared errors of the confidence limits based on the Monte Carlo estimates.

³ It was found that increasing the number of resamples to 10 000 did not have a marked impact on the accuracy of the Krinsky Robb method. This is in line with the findings documented by Krinsky and Robb. Increasing the number of bootstrap resamples beyond 1000 was considered impractical since this would lead to a considerable increase in computation time (it should be noted that the bootstrap already involves estimating $10\,000 \text{ (runs)} \times 1000 \text{ (resamples)} = 10\,000\,000$ logit models, which for $N = 100$ takes about 4 days to run on a PC with a 3.06 GHz Xeon processor).

Table II. The impact of changes in the sample size – results for $WTP_1 = £1$ ($\mu = 1$)

<i>N</i>	Method	Lower (0.025) ^a	Upper (0.975) ^a	Proportion of type I error ^b
10	Monte Carlo	0.418	1.616	—
10	Delta	0.419 (0.0838)	1.587 (0.1102)	0.0508 (0.0465–0.0551)
10	Fieller	0.412 (0.0906)	1.654 (0.1311)	0.0476 (0.0434–0.0518)
10	KR	0.399 (0.0906)	1.654 (0.1368)	0.0459 (0.0418–0.0500)
10	Bootstrap	—	—	—
25	Monte Carlo	0.637	1.379	—
25	Delta	0.637 (0.0340)	1.375 (0.0397)	0.0501 (0.0458–0.0543)
25	Fieller	0.638 (0.0351)	1.393 (0.0419)	0.0495 (0.0452–0.0538)
25	KR	0.629 (0.0342)	1.388 (0.0420)	0.0464 (0.0423–0.0505)
25	Bootstrap	0.644 (0.0373)	1.375 (0.0452)	0.0668 (0.0619–0.0717)
50	Monte Carlo	0.741	1.268	—
50	Delta	0.744 (0.0175)	1.264 (0.0192)	0.0518 (0.0475–0.0561)
50	Fieller	0.746 (0.0178)	1.272 (0.0197)	0.0518 (0.0475–0.0561)
50	KR	0.735 (0.0175)	1.268 (0.0198)	0.0482 (0.0440–0.0524)
50	Bootstrap	0.747 (0.0184)	1.264 (0.0205)	0.0611 (0.0564–0.0658)
100	Monte Carlo	0.815	1.189	—
100	Delta	0.821 (0.0087)	1.188 (0.0093)	0.0538 (0.0494–0.0582)
100	Fieller	0.822 (0.0088)	1.191 (0.0094)	0.0545 (0.0501–0.0589)
100	KR	0.814 (0.0088)	1.189 (0.0092)	0.0496 (0.0453–0.0539)
100	Bootstrap	0.822 (0.0089)	1.188 (0.0096)	0.0574 (0.0528–0.0620)

^aMean squared errors of the confidence limits based on the Monte Carlo estimates are reported in parenthesis.^b95% confidence intervals for the estimates are reported in parenthesis.Table III. The impact of changes in the sample size – results for $WTP_2 = £0.50$ ($\mu = 1$)

<i>N</i>	Method	Lower (0.025) ^a	Upper (0.975) ^a	Proportion of type I error ^b
10	Monte Carlo	−0.113	1.126	—
10	Delta	−0.114 (0.0838)	1.103 (0.1102)	0.0509 (0.0466–0.0552)
10	Fieller	−0.113 (0.0961)	1.182 (0.1397)	0.0420 (0.0381–0.0459)
10	KR	−0.159 (0.1029)	1.148 (0.1318)	0.0424 (0.0385–0.0463)
10	BS	—	—	—
25	Monte Carlo	0.110	0.872	—
25	Delta	0.106 (0.0356)	0.876 (0.0423)	0.0478 (0.0436–0.0520)
25	Fieller	0.111 (0.0365)	0.899 (0.0448)	0.0444 (0.0404–0.0484)
25	KR	0.092 (0.0382)	0.886 (0.0427)	0.0430 (0.0390–0.0470)
25	BS	0.118 (0.0398)	0.879 (0.0483)	0.0616 (0.0569–0.0663)
50	Monte Carlo	0.214	0.763	—
50	Delta	0.217 (0.0184)	0.760 (0.0208)	0.0520 (0.0476–0.0564)
50	Fieller	0.221 (0.0187)	0.770 (0.0212)	0.0506 (0.0463–0.0549)
50	KR	0.211 (0.0187)	0.762 (0.0208)	0.0496 (0.0453–0.0539)
50	BS	0.223 (0.0196)	0.761 (0.0225)	0.0612 (0.0565–0.0659)
100	Monte Carlo	0.293	0.680	—
100	Delta	0.297 (0.0094)	0.680 (0.0101)	0.0552 (0.0507–0.0597)
100	Fieller	0.299 (0.0095)	0.684 (0.0102)	0.0542 (0.0498–0.0586)
100	KR	0.293 (0.0094)	0.681 (0.0102)	0.0535 (0.0491–0.0579)
100	BS	0.299 (0.0099)	0.680 (0.0106)	0.0603 (0.0556–0.0650)

^aMean squared errors of the confidence limits based on the Monte Carlo estimates are reported in parenthesis.^b95% confidence intervals for the estimates are reported in parenthesis.

Finally, column 5 gives the proportion of times the estimated confidence intervals do not include the true WTP, which is probably the most important indicator of the precision of the confidence interval (Mooney and Duval, 1993). If the estimates are accurate this proportion should not be significantly different from the nominal $\alpha = 0.05$. Since the proportion is binomially distributed it is possible to derive confidence intervals for the estimates of alpha (see e.g. Conover, 1999). The 95% confidence intervals are given in brackets in column 5.

It can be seen from the tables that all the methods yield fairly similar results, which is reassuring. Somewhat surprisingly the delta method confidence intervals are found to be the most accurate overall, and only fail to include the true WTP the correct number of times for WTP_2 when $N = 100$ (at the 5% significance level). The Fieller and Krinsky Robb methods perform well when $N > 25$, but for lower sample sizes the confidence intervals for WTP_2 have significantly lower than nominal alphas. This may suggest that the Fieller and Krinsky Robb methods are more sensitive to departures from the assumed normality of the coefficients than the delta method. The bootstrap method did not work well when $N = 10$, since some of the resamples caused problems for the convergence of the model (this happened when one or two of the 'respondents' was resampled a large number of times), and because of this the bootstrap results for $N = 10$ are not reported. For $N > 10$ the bootstrap confidence intervals have a slightly higher than nominal alpha in all the cases. This suggest that some form of bias adjustment might be appropriate. Briggs *et al.* (1999) found that the bias adjusted and accelerated bootstrap (Efron, 1987) performed better than the percentile bootstrap in their application. The bias adjusted and accelerated bootstrap is substantially more computationally demanding than the percentile bootstrap, however, since it requires an additional round of Jackknife replications for each bootstrap replication. In the cases considered here the bias adjusted and accelerated bootstrap was not found to be significantly more accurate than the percentile bootstrap (the results are available from the author upon request).

The impact of changes in the error variance

The amount of noise in the data is expected to have an influence on the precision of the estimated confidence intervals. Recall from the section 'WTP confidence intervals' that the scale parameter, μ , is inversely related to the error variance in the model (Equation (4)). Three values of μ are considered: $\mu = 1, 0.5$ and 0.25 , which implies an error variance of $\pi^2/3, 4\pi^2/3$ and $16\pi^2/3$, respectively. The sample size is held constant at $N = 50$. The results for WTP_1 and WTP_2 are given in Tables IV and V, respectively.

The Monte Carlo estimates of the confidence intervals become wider when the error variance increases, reflecting the lower precision of the parameter estimates. Moreover, the confidence intervals derived using the delta method become less precise when the error variance increases. The delta method confidence intervals include the true WTP the correct number of times when $\mu > 0.25$, but have lower

Table IV. The impact of changes in the error variance – results for $WTP_1 = £1$ ($N = 50$)

μ	Method	Lower (0.025) ^a	Upper (0.975) ^a	Proportion of type I error ^b
1	Monte Carlo	0.741	1.268	—
1	Delta	0.744 (0.0175)	1.264 (0.0192)	0.0518 (0.0475–0.0561)
1	Fieller	0.746 (0.0178)	1.272 (0.0197)	0.0518 (0.0475–0.0561)
1	KR	0.735 (0.0175)	1.268 (0.0198)	0.0482 (0.0440–0.0524)
1	Bootstrap	0.747 (0.0184)	1.264 (0.0205)	0.0611 (0.0564–0.0658)
0.5	Monte Carlo	0.565	1.489	—
0.5	Delta	0.546 (0.0452)	1.471 (0.0723)	0.0467 (0.0426–0.0508)
0.5	Fieller	0.565 (0.0484)	1.514 (0.0812)	0.0484 (0.0442–0.0526)
0.5	KR	0.547 (0.0488)	1.517 (0.0844)	0.0466 (0.0425–0.0507)
0.5	Bootstrap	0.573 (0.0506)	1.500 (0.0842)	0.0596 (0.0550–0.0642)
0.25	Monte Carlo	0.192	2.082	—
0.25	Delta	0.108 (0.1442)	1.974 (0.4833)	0.0387 (0.0349–0.0425)
0.25	Fieller ^c	0.185 (0.1872)	2.299 (1.6874)	0.0508 (0.0465–0.0551)
0.25	KR	0.152 (0.2979)	2.307 (1.3659)	0.0498 (0.0455–0.0541)
0.25	Bootstrap	0.186 (0.5572)	2.273 (1.4710)	0.0599 (0.0552–0.0646)

^a Mean squared errors of the confidence limits based on the Monte Carlo estimates are reported in parenthesis.

^b 95% confidence intervals for the estimates are reported in parenthesis.

^c Based on 9997 observations.

Table V. The impact of changes in the error variance – results for $WTP_2 = £0.50$ ($N = 50$)

μ	Method	Lower (0.025) ^a	Upper (0.975) ^a	Proportion of type I error ^b
1	Monte Carlo	0.214	0.763	—
1	Delta	0.217 (0.0184)	0.760 (0.0208)	0.0520 (0.0476–0.0564)
1	Fieller	0.221 (0.0187)	0.770 (0.0212)	0.0506 (0.0463–0.0549)
1	KR	0.211 (0.0187)	0.762 (0.0208)	0.0496 (0.0453–0.0539)
1	Bootstrap	0.223 (0.0196)	0.761 (0.0225)	0.0612 (0.0565–0.0659)
0.5	Monte Carlo	0.067	0.987	—
0.5	Delta	0.020 (0.0514)	0.961 (0.0708)	0.0495 (0.0452–0.0538)
0.5	Fieller	0.030 (0.0548)	0.994 (0.0768)	0.0512 (0.0469–0.0555)
0.5	KR	0.007 (0.0571)	0.970 (0.0739)	0.0503 (0.0460–0.0546)
0.5	Bootstrap	0.043 (0.0545)	0.981 (0.0803)	0.0632 (0.0584–0.0680)
0.25	Monte Carlo	−0.365	1.508	—
0.25	Delta	−0.426 (0.1922)	1.433 (0.3670)	0.0362 (0.0325–0.0399)
0.25	Fieller ^c	−0.427 (0.2871)	1.639 (0.7353)	0.0485 (0.0443–0.0527)
0.25	KR	−0.492 (0.4364)	1.570 (0.7197)	0.0480 (0.0438–0.0522)
0.25	Bootstrap	−0.397 (0.5211)	1.613 (0.8650)	0.0623 (0.0576–0.0670)

^a Mean squared errors of the confidence limits based on the Monte Carlo estimates are reported in parenthesis.

^b 95% confidence intervals for the estimates are reported in parenthesis.

^c Based on 9997 observations.

Table VI. Coefficients of variation for β_1 , β_2 and β_C by sample size and scale

N	μ	$s.e.(\hat{\beta}_1)/\hat{\beta}_1$	$s.e.(\hat{\beta}_2)/\hat{\beta}_2$	$s.e.(\hat{\beta}_C)/\hat{\beta}_C$
10	1	0.332	0.638	−0.176
25	1	0.207	0.395	−0.109
50	1	0.149	0.282	−0.077
100	1	0.130	0.200	−0.054
10	0.5	0.505	1.074	−0.241
25	0.5	0.320	0.674	−0.152
50	0.5	0.228	0.480	−0.108
100	0.5	0.159	0.341	−0.077
10	0.25	0.917	2.011	−0.424
25	0.25	0.579	1.250	−0.267
50	0.25	0.414	0.897	−0.190
100	0.25	0.286	0.626	−0.134

than nominal alphas when $\mu = 0.25$. The Fieller method confidence interval could not be derived in 2 out of the 10 000 trials when $\mu = 0.25$ due to Ω in Equation (9) being negative. A further observation was dropped due to the denominator in Equation (9) being negative. Based on the remaining 9997 trials the Fieller method performs well, with alphas insignificantly different from the nominal value in all the cases. Like the Fieller method the accuracy of the Krinsky Robb and bootstrap methods seem largely unaffected by the increase in error variance. As before the bootstrap confidence intervals have a somewhat larger than nominal alpha in all the cases.

To investigate possible interaction effects between the sample size and the error variance the above experiments were re-run with N set to 100 and 25. It was found that increasing the sample size to 100 did not change the results qualitatively for either value of μ . In the case of $N = 25$, only the Fieller method alpha was found to be insignificantly different from 0.05. However, for $\mu = 0.25$, the Fieller method did not lead to satisfactory results in 270 out of the 10 000 trials, either due to the confidence limits being imaginary or the denominator of Equation (9) being negative. This finding is not too discouraging, however, as the $N = 25$, $\mu = 0.25$ combination is quite extreme in terms of the accuracy of the parameter estimates. Table VI presents Monte Carlo estimates of the coefficients of variation for β_1 , β_2 and β_C by sample size and scale (recall that the t -statistic is simply the inverse of the coefficient of variation). It can be seen from the table that neither β_1 or β_2 will be significant in general when $N = 25$

and $\mu = 0.25$, with average t -statistics of 1.73 and 0.8, respectively. When an attribute coefficient is insignificant it could be argued that deriving the willingness to pay for the attribute is not very meaningful in the first place, since its marginal utility is not significantly different from zero.

Returning to the $N = 50$ case it can be seen from the table that the Monte Carlo estimates of the coefficients of variation for β_C are -0.077 , -0.108 and -0.190 for $\mu = 1$, 0.5 and 0.25 , respectively, which corresponds to t -statistics of -13.02 , -9.25 and -5.27 . Although one should be careful not to draw strong conclusions on the basis of one study, this seems to imply that for moderately sized samples a cost coefficient with a t -statistic of around 10 or higher in absolute value (this is high, but not uncommon for models estimated using data from DCEs) suggest that the delta method will produce accurate confidence intervals. In cases where the cost coefficient is less precisely estimated it is likely that the Fieller, Krinsky Robb and bootstrap methods will produce more accurate results.

The impact of neglected unobserved heterogeneity

Until this point it has been assumed that the logit model is the correct specification. It is often argued, however, that it is appropriate to correct for unobserved individual heterogeneity when estimating discrete choice models with DCE data by using either the random effects logit or probit estimator (see e.g. Ryan and Gerard, 2003). If the random effects logit is the true specification the difference in utility between the alternatives is given by

$$U_{nit} - U_{njt} = (V_{nit} - V_{njt}) + z_n - (\varepsilon_{njt} - \varepsilon_{nit}) \quad (11)$$

where z_n is an individual-specific and time invariant unobserved effect which is normally distributed with mean zero and standard deviation σ_z (see Greene, 2003, for a detailed description of this model). In previous simulation studies the coefficients have been found to be unbiased if the random effect is ignored and a standard logit is used for the analysis, but the estimated standard errors of the coefficients will be biased in this case (Bradley and Daly, 1999; Cirillo *et al.*, 1999) (see Guilkey and Murphy, 1993, for a similar result for the probit model).⁴ Since the estimated standard errors are biased it follows that a confidence interval based on these standard errors will also be biased, and it is therefore expected that the delta, Fieller and Krinsky Robb methods will produce less accurate estimates of the confidence intervals in this case. In order to investigate the influence of neglected unobserved heterogeneity on the precision of the confidence intervals, draws from the normal distribution with mean zero and standard deviation 2 was added to Equation (10) in the data generation process. Tables VII and VIII present the results for $N = 50$.

Similar to the previous studies conducted it is found that the logit estimates of WTP are unbiased in spite of ignoring the random effect. It can be seen from the tables that the Monte Carlo estimates of the confidence intervals are slightly wider than in the case of no unobserved heterogeneity, but not much so, implying that the misspecification does not lead to a substantial loss in efficiency. The delta, Fieller and Krinsky Robb confidence intervals are less accurate in this case, which is to be expected since they are both based on the biased estimate of the covariance matrix. This is not redeemed by increasing the sample size. The bootstrap method is the most accurate, and although the bootstrap confidence intervals have a higher than nominal alpha they are about as accurate as in the previous cases without unobserved heterogeneity. The accuracy of the bootstrap was not found to be affected by reducing the

⁴ It should be noted that the coefficients will be also be biased if z_n is correlated with the alternative attributes, in which case the fixed effects logit model (Chamberlain, 1980) is the appropriate specification. Unless the attributes are interacted with the socio-demographic characteristics of the respondents, however, this possibility can be ruled out due to the experimental nature of the data.

Table VII. The impact of neglected unobserved heterogeneity – results for $WTP_1 = £1$ ($N = 50$, $\mu = 1$)

Method	Lower (0.025) ^a	Upper (0.975) ^a	Proportion of type I error ^b
Monte Carlo	0.730	1.332	—
Delta	0.643 (0.0282)	1.398 (0.0334)	0.0132 (0.0110–0.0154)
Fieller	0.654 (0.0274)	1.423 (0.0395)	0.0139 (0.0116–0.0162)
KR	0.639 (0.0300)	1.421 (0.0407)	0.0124 (0.0102–0.0146)
Bootstrap	0.734 (0.0222)	1.327 (0.0289)	0.0590 (0.0544–0.0636)

^a Mean squared errors of the confidence limits based on the Monte Carlo estimates are reported in parenthesis.

^b 95% confidence intervals for the estimates are reported in parenthesis.

Table VIII. The impact of neglected unobserved heterogeneity – results for $WTP_2 = £0.50$ ($N = 50$, $\mu = 1$)

Method	Lower (0.025) ^a	Upper (0.975) ^a	Proportion of type I error ^b
Monte Carlo	0.180	0.804	—
Delta	0.100 (0.0293)	0.873 (0.0341)	0.0145 (0.0122–0.0168)
Fieller	0.108 (0.0292)	0.894 (0.0392)	0.0142 (0.0119–0.0165)
KR	0.091 (0.0321)	0.877 (0.0353)	0.0134 (0.0111–0.0157)
Bootstrap	0.192 (0.0239)	0.801 (0.0295)	0.0612 (0.0565–0.0659)

^a Mean squared errors of the confidence limits based on the Monte Carlo estimates are reported in parenthesis.

^b 95% confidence intervals for the estimates are reported in parenthesis.

sample size and increasing the error variance. This suggests that the bootstrap is the appropriate method to employ if one suspects that there may be unobserved heterogeneity present in the data. It would also be possible, of course, to estimate a random effects logit model and use any of the four methods to construct the confidence interval, but investigating the properties of confidence intervals derived in this fashion is beyond the scope of the present paper.⁵

AN EMPIRICAL APPLICATION

To illustrate how the various methods compare with empirical data I use data from the pilot study of the National Primary Care Research and Development Centre's PAPRICA project. The aims of the PAPRICA project include examining priorities among key attributes of primary care for a representative sample of UK patients. The attributes of the pilot experiment include waiting time for the appointment, the cost of seeing the GP, whether the patient is offered a choice of appointment times, the doctor's manner, whether the doctor knows the patient's medical history and the thoroughness of the examination. The data consists of 30 respondents who completed 16 choices each, yielding a total of 480 choices. A simple logit model is estimated on this data with the aim of estimating the willingness to pay for a reduction in waiting time. The estimate of the coefficient for waiting time is -0.193 , while the cost coefficient is -0.054 , implying that the willingness to pay for a one day decrease in waiting time equals £ 3.57 (the standard errors of the waiting time and cost coefficients are 0.042 and 0.008, respectively, implying t -statistics of -4.63 and -6.77). Table IX presents confidence intervals for the willingness to pay for a reduction in waiting time derived by the various methods.

As before, both the Krinsky Robb and bootstrap intervals are derived using the percentile method based on 1000 resamples. It can be seen that all the methods yield fairly similar results, which is consistent with the findings in the simulation study. The confidence limits of the Fieller, Krinsky Robb

⁵ The reason for this is the considerable increase in computing time needed to estimate the random effects logit model which would make such an experiment impractical.

Table IX. Confidence intervals for the willingness to pay for a 1-day reduction in waiting time

Method	Lower (0.025)	Upper (0.975)
Delta	2.09	5.05
Fieller	2.16	5.28
KR	2.22	5.27
Bootstrap	2.14	5.19

and bootstrap intervals imply that the distribution of WTP is somewhat skewed, which is not reflected in the delta method estimate. This finding, together with the findings from the simulation study suggesting that the delta method is less accurate than the other methods when the *t*-statistic for the cost coefficient is less than 10 in absolute value, suggests that the Fieller, Krinsky Robb and bootstrap confidence intervals are the most accurate in this case. The estimates are so similar, however, that the choice of method is unlikely to make a difference from a policy point of view.

DISCUSSION

The general finding of the simulation study and the empirical application is that the methods yield fairly similar results in a broad range of situations, with the notable exception of neglected unobserved heterogeneity. This is in line with the findings in some (Kling, 1991; Krinsky and Robb, 1991), but not all (Briggs *et al.*, 1999; Chaudhary and Stearns, 1996), of the other contexts in which these methods have been compared. The only other study known to the author which compares various methods of estimating confidence intervals for willingness to pay estimates is the study by Armstrong *et al.* (2001), who compare a method devised by the authors with the Jackknife (Efron and Tibshirani, 1993), Fieller and Krinsky Robb methods (they call the latter 'simulation of multivariate normal variates') using real data on commuters' mode choice in Chile. Since Armstrong *et al.* do not employ the delta and bootstrap methods the results in their paper cannot be directly compared with those reported here, but it should be noted that they find that the Jackknife method yields different results from the Fieller and Krinsky Robb methods. Since the bootstrap is likely to be more accurate than the Jackknife in this context, however, this finding is not at odds with the bootstrap, Fieller and Krinsky Robb confidence intervals being similar. Also, since Armstrong *et al.* use real rather than simulated data they are unable to evaluate which of the methods produce the more accurate results.

The main difference between the results in the present study and the cost-effectiveness literature is the good performance of the delta method, as the Fieller, Krinsky Robb and bootstrap methods are generally found to outperform the delta method in cost-effectiveness applications (e.g. Briggs *et al.*, 1999; Chaudhary and Stearns, 1996). (The Krinsky Robb method is usually referred to as the parametric bootstrap in this literature.) One reason for this difference is the fact that the coefficient of variation for the denominator variate tends to be relatively high in cost-effectiveness applications, which as we have seen will worsen the performance of the delta method. Another factor is the difference in the data generation process. While both cost-effectiveness ratios and WTP measures are ratios of two variables which tend asymptotically toward normality, the variables in cost-effectiveness studies are estimated differences in the means of cost and effectiveness between two groups receiving different treatments. The distributions of these variables may have very different properties from the parameters in a discrete choice model estimated using data from a DCE in small samples.

CONCLUDING REMARKS

This paper compares four approaches to estimating confidence intervals for willingness to pay measures: the delta, Fieller, Krinsky Robb and the bootstrap methods. It is found that all of the methods produce reasonably accurate confidence intervals in the majority of the cases considered. The delta method is, somewhat surprisingly, found to be the most accurate when the data is well conditioned, while the bootstrap is more robust to noisy data and misspecification of the model. None of the approaches produces wildly misleading results in any of the cases considered, which suggests that estimating confidence intervals using any of the methods considered here is far superior to not estimating confidence intervals at all. The conclusions drawn from the simulation study are supported by the findings of the empirical application in that all the methods produce fairly similar confidence intervals. The simulation results are all based on the condition that the logit (and in some cases the random effects logit) is the correct specification. Although it is expected that similar results will apply to the probit model, this should be properly investigated in future research.

ACKNOWLEDGEMENTS

NPCRDC receives funding from the Department of Health. The views expressed are not necessarily those of the funders. The author is grateful to two anonymous referees, Hugh Gravelle, Andrea Manca, participants at the 3rd 'Advancing the methodology of discrete choice experiments in health economics' workshop at the University of Las Palmas and seminar participants at the University of York and the University of Bergen for helpful comments.

REFERENCES

- Armstrong P, Garrido R, Ortúzar JdeD. 2001. Confidence intervals to bound the value of time. *Transportation Research Part E* **37**: 143–161.
- Ben-Akiva M, Lerman S. 1985. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press: Cambridge.
- Bradley M, Daly AJ. 1999. New analysis issues in stated preference research. In *Stated Preference Modelling Techniques*, Ortúzar JdeD (ed.). PTRC Education and Research Services Limited: London.
- Briggs AH, Mooney CZ, Wonderling DE. 1999. Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Statistics in Medicine* **18**: 3245–3262.
- Chamberlain G. 1980. Analysis of covariance with qualitative data. *Review of Economic Studies* **47**: 225–238.
- Chaudhary MA, Stearns SC. 1996. Estimating confidence intervals for cost-effectiveness ratios: an example from a randomised trial. *Statistics in Medicine* **15**: 1447–1458.
- Cirillo C, Daly A, Lindveld K. 1999. Eliminating bias due to the repeated measurements problem in SP data. In *Stated Preference Modelling Techniques*, Ortúzar JdeD (ed.). PTRC Education and Research Services Limited: London.
- Cochran WG. 1977. *Sampling Techniques* (3rd edn). Wiley: New York.
- Conover WJ. 1999. *Practical Nonparametric Statistics*. Wiley: New York.
- Efron B. 1979. Bootstrap methods: another look at the Jackknife. *Annals of Statistics* **7**: 1–26.
- Efron B. 1987. Better bootstrap confidence intervals. *Journal of the American Statistical Association* **82**: 171–200.
- Efron B, Tibshirani R. 1993. *An Introduction to the Bootstrap*. Chapman & Hall: New York.
- Fieller EC. 1932. The distribution of the index in a normal bivariate population. *Biometrika* **24**: 428–440.
- Fieller EC. 1954. Some problems in interval estimation. *Journal of the Royal Statistical Society, Series B* **16**: 175–185.
- Greene WH. 2003. *Econometric Analysis* (5th edn). Prentice-Hall: Englewood Cliffs.
- Guilkey DK, Murphy JL. 1993. Estimation and testing in the random effects probit model. *Journal of Econometrics* **59**: 301–317.
- Hinkley DV. 1969. On the ratio of two correlated normal variables. *Biometrika* **56**: 635–639.

- Horowitz JL. 2001. The bootstrap in econometrics. In *Handbook of Econometrics*, Leamer E, Heckman JJ (eds), vol. 5. North Holland: Amsterdam.
- Kling CL. 1991. Estimating the precision of welfare measures. *Journal of Environmental Economics and Management* **21**: 244–259.
- Krinsky I, Robb AL. 1986. On approximating the statistical properties of elasticities. *Review of Economics and Statistics* **68**: 715–719.
- Krinsky I, Robb AL. 1990. On approximating the statistical properties of elasticities: a correction. *Review of Economics and Statistics* **72**: 189–190.
- Krinsky I, Robb AL. 1991. Three methods for calculating the statistical properties of elasticities: a comparison. *Empirical Economics* **16**: 199–209.
- Kuhfeld WF. 2005. *Marketing Research Methods in SAS*. SAS Institute Inc.: Cary.
- Lancsar E, Savage E. 2004. Deriving welfare measures from discrete choice experiments: inconsistency between current methods and random utility and welfare theory. *Health Economics Letters* **13**: 901–907.
- McIntosh E, Ryan M. 2002. Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: implications of discontinuous preferences. *Journal of Economic Psychology* **23**: 367–382.
- Mooney CZ, Duval RD. 1993. *Bootstrapping: A Nonparametric Approach to Statistical Inference*. Sage University Paper Series on Quantitative Applications in the Social Sciences, Series No. 07-095. Sage: Newbury Park.
- Ryan M, Gerard K. 2003. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied Health Economics and Health Policy* **2**: 1–10.
- Ryan M. 2004. Deriving welfare measures in discrete choice experiments: a comment to Lancsar and Savage. *Health Economics* **13**: 909–912.
- Santos Silva JMC. 2004. Deriving welfare measures in discrete choice experiments: a comment to Lancsar and Savage. *Health Economics* **13**: 913–918.
- Shanmugalingham S. 1982. On the analysis of the ratio of two correlated normal variables. *Statistician* **31**: 251–258.