THE UNIVERSITY of York

A comparison of approaches to estimating confidence intervals for willingness to pay measures

CHE Research Paper 8

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, York, YO10 5DD
E-mail: ah522@york.ac.uk

January 2006

Acknowledgements

The National Primary Care Research and Development Centre receives funding from the Department of Health. The views expressed are not necessarily those of the funders. The author is grateful to 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.

Background

CHE Discussion Papers (DPs) began publication in 1983 as a means of making current research material more widely available to health economists and other potential users. So as to speed up the dissemination process, papers were originally published by CHE and distributed by post to a worldwide readership.

The new CHE Research Paper series takes over that function and provides access to current research output via web-based publication, although hard copy will continue to be available (but subject to charge).

Disclaimer

Papers published in the CHE Research Paper (RP) series are intended as a contribution to current research. Work and ideas reported in RPs may not always represent the final position and as such may sometimes need to be treated as work in progress. The material and views expressed in RPs are solely those of the authors and should not be interpreted as representing the collective views of CHE research staff or their research funders.

Further copies

Copies of this paper are freely available to download from the CHE website www.york.ac.uk/inst/che/pubs. Access to downloaded material is provided on the understanding that it is intended for personal use. Copies of downloaded papers may be distributed to third-parties subject to the proviso that the CHE publication source is properly acknowledged and that such distribution is not subject to any payment.

Printed copies are available on request at a charge of £5.00 per copy. Please contact the CHE Publications Office, email che-pub@york.ac.uk, telephone 01904 321458 for further details.

Centre for Health Economics
Alcuin College
University of York
York, UK
www.york.ac.uk/inst/che

Abstract

This paper describes three approaches to estimating confidence intervals for willingness to pay measures, the delta, Krinsky and 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 three 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.

Keywords: willingness to pay, confidence interval, delta method, boot-strap

1 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. 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 derived in applied work (for an exception see [1]). For a recent review of studies applying DCEs in health care see [2].

This paper describes three approaches to estimating confidence intervals for willingness to pay measures, the delta, Krinsky and Robb and bootstrap methods. 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 three 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.

The paper is organised as follows: section 2 provides an outline of random utility maximisation and the logit model, section 3 describes the various methods for estimating confidence intervals, section 4 describes the simulated data, while sections 5 and 6 present the simulation results and the empirical application, respectively. Finally section 7 offers some concluding remarks.

2 Random utility maximisation and the logit model

We 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:

$$U_{njt} = V_{njt} + \varepsilon_{njt} \tag{1}$$

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:

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

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. [3]) in which the probability that alternative i is chosen in scenario t is given by:

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

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

$$\mu = \frac{\pi}{\sqrt{3\sigma_\varepsilon^2}} \tag{4}$$

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

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

where β_{0i} is a constant which reflects the mean impact of the unobservable components on the utility of alternative i. β_1,...,β_K are vectors of coefficients for attributes X_1,...,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:

$$\frac{dC}{dX_k} = WTP_k = -\frac{\beta_k}{\beta_C} \tag{6}$$

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 respectively. 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 [4] and Hinkley [5], 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 deviation of β_C to its mean is low, the distribution of WTP is likely to be approximately normal. As will become clear in the following section, this result is of importance when comparing the various approaches to estimating confidence intervals for WTP.

3 WTP confidence intervals

3.1 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. [6]). In the case of WTP the variance is given by:

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

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

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

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 [7] 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.

3.2 The Krinsky and Robb method

Krinsky and Robb [8, 9] suggest an alternative to the delta method 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 [10]. 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* [10, 11]. 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

above, hinges on the assumption that WTP is symmetrically distributed. The percentile interval, on the other hand, does not assume that WTP is symmetrically distributed. The only assumption required is that the coefficients are joint normally distributed, which may not be unrealistic when the sample is relatively large. This suggests that the (percentile) Krinsky and Robb method will yield more accurate confidence intervals than the delta method when WTP is not symmetrically distributed. The downside of the Krinsky and Robb method relative to the delta method 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.

3.3 The bootstrap

The bootstrap [10, 11, 12] has been used extensively to estimate standard errors and confidence intervals in economics in recent years (see e.g. [13]). The bootstrap is similar to the Krinsky and Robb method in that a simulated distribution for the variable of interest is generated. In contrast to the Krinsky and 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 and 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 Krinsky and Robb method in that it does not rely on the assumption that WTP is symmetrically distributed, but unlike the Krinsky and Robb method it does not require that the coefficients themselves are joint normally distributed. It is therefore possible that the bootstrap will perform better than the Krinsky and Robb method 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.

4 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 1.

[Table 1 about here]

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 [14] 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:

$$V_1 - V_2 = \beta_0 + \beta_1(X_{11} - X_{12}) + \beta_2(X_{21} - X_{22}) + \beta_C(C_1 - C_2) \qquad (9)$$

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 willingness to pay for an improvement in attribute X_1 (WTP_1) is £1 and the willingness to pay for an 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 indirect utility, $V_1 - V_2$, alternative 1 is chosen. Otherwise, alternative 2 is chosen.

5 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. Both the delta and Krinsky and 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 sections 5.1 and 5.2. Finally section 5.3 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.

5.1 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 2 and 3 respectively.

[Tables 2 and 3 about here]

The first two columns in the tables give the sample size and the method used for calculating the confidence interval. Both the Krinsky and Robb and bootstrap confidence intervals are derived using the percentile method based on 1000 resamples in each trial.[1] Columns 3 and 4 gives the lower and upper

[1] It was found that increasing the number of resamples to 10,000 did not have a marked impact on the accuracy of the Krinsky and Robb method. This is in line with the findings documented by Krinsky and Robb. Increasing the number of boostrap resamples beyond 1,000 was considered impractical since this would lead to a considerable increase in computation

limits of the estimated 95% confidence intervals averaged over 10000 trials. The Monte Carlo estimates of the confidence limits are derived by calculating the 2.5 and 97.5 percentiles of the 10000 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. 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 [11]. 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. [15]). The 95% confidence intervals are given in brackets in column 5.

A number of interesting findings can be derived from the tables. Firstly, it can be seen that the Monte Carlo estimates of the confidence intervals become narrower as the sample size increases. The marginal improvement in precision decreases with the sample size, however: the biggest gain comes from increasing the sample size from 10 to 25. Moreover it can be seen that all the methods yield fairly similar results. This is reassuring and in line with the findings in some [16, 17], but not all [18] 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. [19], who compares two methods devised by the authors with the Jackknife [10] and Krinsky and 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 and Krinsky and Robb methods yield different results. 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 and Krinsky and 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.

time (it should be noted that the bootstrap already involves estimating 10,000 (runs) $\times$ 1,000 (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).

It can be seen from the tables that, somewhat surprisingly, the delta method is the most accurate overall, and only fails to include the true WTP the correct number of times for WTP_2 when N=100 (at the 5% significance level). The Krinsky and Robb method performs well when N>25, but for lower sample sizes the confidence intervals for WTP_2 have significantly lower than nominal alphas while the null hypothesis of $\alpha = 0.05$ is only marginally 'accepted' for WTP_1 when N=10. This may suggest that the Krinsky and Robb method is more sensitive to departures from normality 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. found that the bias adjusted and accelerated bootstrap [20] 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).

5.2 The impact of changes in the error variance

In addition to the sample size, the amount of noise in the data is expected to have an influence on the precision of the estimated confidence intervals. It should be recalled from section 3 that the scale parameter, μ, is inversely related to the error variance in the model (eq. 4). Three values of μ are considered: μ =1, μ =0.5 and μ =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 4 and 5 respectively.

[Tables 4 and 5 about here]

It can be seen that 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 than nominal alphas when $\mu = 0.25$. The Krinsky and Robb method is somewhat more accurate, with alphas insignificantly different from the nominal value in all the cases. Like the Krinsky and Robb method the accuracy of the bootstrap 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.

As discussed in section 3 the delta method relies on the assumption that WTP is normally distributed, which is likely to hold when the sample size is large and the coefficient of variation for the cost coefficient is sufficiently small. In the cases considered here 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. This corresponds to t-statistics of -13.02, -9.25 and -5.27, which may be a more intuitive representation of the precision of the estimates. Although one should be careful to draw strong conclusions on the basis of one study, this seems to imply that when the cost coefficient has a t-statistic of around 10 or higher in absolute value (this is high, but not uncommon for models estimated using data from DCEs) the delta method will produce accurate confidence intervals, while in cases where the cost coefficient is less precisely estimated it is likely that the Krinsky and Robb and bootstrap methods will produce more accurate results.

5.3 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. [2]). If the random effects logit is the true specification the difference in utility between the alternatives is given by:

$$U_{nit} - U_{njt} = (V_{nit} - V_{njt}) + z_n - (\varepsilon_{njt} - \varepsilon_{nit}) \tag{10}$$

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 [6] for a detailed description of this model). In previous simulation studies the coef-

ficients 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 [21, 22] (see [23] for a similar result for the probit model).[2] 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 and Krinsky and 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 eq. 9 in the data generation process. Tables 6 and 7 present the results for N=50.

[Tables 6 and 7 about here]

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. Both the delta and Krinsky and 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. 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. 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 three 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.

[2] It should be noted that the coefficicents will be also be biased if z_n is correlated with the alternative attributes, in which case the fixed effects logit model [24] 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.

6 An empirical application

To illustrate how the various methods compare with empirical data we use data from the pilot study of the National Primary Care Research and Development Centres'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 8 presents confidence intervals for the willingness to pay for a reduction in waiting time derived by the various methods. As before, both the Krinsky and 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 Krinsky and Robb 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 Krinsky and 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.

7 Concluding Remarks

This paper compares three approaches to estimating confidence intervals for willingness to pay measures, the delta, Krinsky and 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. Although one should be careful to draw strong conclusions on the basis of a single study, it is interesting to note that none of the approaches produces wildly misleading results in any of the cases, 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. Finally, it should be noted that the simulation results were 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.

References

[1] McIntosh E, Ryan M. Using discrete choice experiments to derive welfare estimates for the provision of elective surgery: Implications of discontinuous preferences. *J Econ Psychol* 2002; **23**: 367-382.

[2] Ryan M, Gerard K. Using discrete choice experiments to value health care programmes: Current practice and future research reflections. *Appl Health Econ Health Policy* 2003; **2**: 1-10.

[3] Ben-Akiva M, Lerman S. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press: Cambridge, 1985.

[4] Fieller EC. The distribution of the index in a normal bivariate population. *Biometrika* 1932; **24**: 428-440.

[5] Hinkley DV. On the ratio of two correlated normal variables. *Biometrika* 1969; **56**: 635-639.

[6] Greene WH. *Econometric Analysis, 5th ed*. Prentice Hall: Englewood Cliffs, 2003.

[7] Shanmugalingham S. On the analysis of the ratio of two correlated normal variables. *Statistician* 1982; **31**: 251-258.

[8] Krinsky I, Robb AL. On approximating the statistical properties of elasticities. *Rev Econ Stat* 1986; **68**: 715-719.

[9] Krinsky I, Robb AL. On approximating the statistical properties of elasticities: A correction. *Rev Econ Stat* 1990; **72**: 189-190.

[10] Efron B, Tibshirani R. *An Introduction to the Bootstrap*. Chapman and Hall: New York, 1993.

[11] Mooney CZ, Duval RD. *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, 1993.

[12] Efron B. Bootstrap methods: Another look at the Jackknife. *Annals of Statistics* 1979; **7**: 1-26.

[13] Horowitz JL. The bootstrap in econometrics. In *Handbook of Econometrics, vol. 5*, Leamer E, Heckman JJ (eds). North Holland: Amsterdam, 2001, 3159-3228.

[14] Kuhfeld WF. *Marketing Research Methods in SAS*. SAS Institute Inc.: Cary, 2005..

[15] Conover WJ. *Practical Nonparametric Statistics*. John Wiley and Sons, Inc.: New York.

[16] Kling CL. Estimating the Precision of Welfare Measures. *J Environ Econ Manage* 1991; **21**: 244-259.

[17] Krinsky I, Robb AL. Three methods for calculating the statistical properties of elasticities: A comparison. *Empir Econ* 1991; **16**: 199-209.

[18] Briggs AH, Mooney CZ, Wonderling DE. Constructing confidence intervals for cost-effectiveness ratios: An evaluation of parametric and nonparametric techniques using Monte Carlo simulation. *Stat Med* 1999; **18**: 3245-3262.

[19] Armstrong P, Garrido R, Ortúzar JdeD. Confidence Intervals to Bound the Value of Time. *Transportation Research Part E* 2001; **37**:143-161.

[20] Efron B. Better bootstrap confidence intervals, *J Am Stat Assoc* 1987; **82**: 171-200.

[21] Bradley M, Daly AJ. New analysis issues in stated preference research. In *Stated Preference Modelling Techniques*, Ortúzar JdeD (ed.). PTRC Education and Research Services Limited: London, 1999, 121-136.

[22] Cirillo C, Daly A, Lindveld K. 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, 1999, 137-154.

[23] Guilkey DK, Murphy JL. Estimation and testing in the random effects probit model. *J Econom* 1993; **59**: 301-317.

[24] Chamberlain G. Analysis of covariance with qualitative data. *Rev Econ Stud* 1980; **47**: 225-238

Table 1. 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

Table 2. The impact of changes in the sample size - results for WTP_1 ($\mu = 1$).

n	Method	Lower (0.025)*	Upper (0.975)*	Proportion of type I error†
10	Monte Carlo	0.418	1.616	-
10	Delta	0.419 (0.0838)	1.587 (0.1102)	0.0508 (0.0465 - 0.0551)
10	K&R	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	K&R	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	K&R	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	K&R	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)

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

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

Table 3. The impact of changes in the sample size - results for WTP_2 ($\mu = 1$)

n	Method	Lower (0.025)*	Upper (0.975)*	Proportion of type I error†
10	Monte Carlo	-0.113	1.126	-
10	Delta	-0.114 (0.0838)	1.103 (0.1102)	0.0509 (0.0466 - 0.0552)
10	K&R	-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	K&R	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	K&R	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	K&R	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)

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

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

Table 4. The impact of changes in the error variance - results for WTP_1
($n = 50$)

μ	Method	Lower (0.025)*	Upper (0.975)*	Proportion of type I error†
1	Monte Carlo	0.741	1.268	-
1	Delta	0.744 (0.0175)	1.264 (0.0192)	0.0518 (0.0475 - 0.0561)
1	K&R	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	K&R	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	K&R	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)

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

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

Table 5. The impact of changes in the error variance - results for WTP_2
($n = 50$)

μ	Method	Lower (0.025)*	Upper (0.975)*	Proportion of type I error†
1	Monte Carlo	0.214	0.763	-
1	Delta	0.217 (0.0184)	0.760 (0.0208)	0.0520 (0.0476 - 0.0564)
1	K&R	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	K&R	0.007 (0.0057)	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	K&R	-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)

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

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

Table 6. The impact of neglected unobserved heterogeneity - results for WTP_1 ($n = 50$, $\mu = 1$)

Method	Lower (0.025)*	Upper (0.975)*	Proportion of type I error†
Monte Carlo	0.730	1.332	-
Delta	0.643 (0.0282)	1.398 (0.0334)	0.0132 (0.0110 - 0.0154)
K&R	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)

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

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

Table 7. The impact of neglected unobserved heterogeneity - results for WTP_2 ($n = 50$, $\mu = 1$)

Method	Lower (0.025)*	Upper (0.975)*	Proportion of type I error†
Monte Carlo	0.180	0.804	-
Delta	0.100 (0.0293)	0.873 (0.0341)	0.0145 (0.0122 - 0.0168)
K&R	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)

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

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

Table 8. Confidence intervals for the willingness to pay for a one-day reduction in waiting time

Method	Lower (0.025)	Upper (0.975)
Delta	2.09	5.05
K&R	2.22	5.27
Bootstrap	2.14	5.19