Willingness to pay confidence interval estimation methods: a comparison

Edoardo Marcucci, Valerio Gatta, Luisa Scaccia

Abstract:
The derivation of reliable willingness to pay (WTP) measures is fundamental in transportation economics and its importance cannot be underestimated. WTP is the amount of money an agent would pay to receive a desired good or service and it can be calculated as the ratio between the coefficient of a given attribute and the cost coefficient. When estimating coefficients via maximum likelihood (ML) WTP is, itself, an estimate with a given but unknown distribution. The analyst cannot therefore simply rely on point estimates and should calculate robust confidence intervals (CI).

Different methods have been proposed in the literature: i) the Delta method (DM), assumes WTP is normally distributed and is the most frequently used; ii) Krinsky and Robb (1986, 1990) propose a sort of parametric bootstrap; iii) the non-parametric bootstrap constitutes an alternative; iv) Bolduc et al. (2010) use the Fieller method (FM) inverting a Wald-type test associated with a conveniently specified null hypothesis; v) analogously, Armstrong et al. (2001) suggest the likelihood ratio inversion method (LRIM) inverting the likelihood ratio test. All methods have advantages and drawbacks. Various comparisons have been performed (e.g. Krinsky and Robb, 1986 and 1990; Armstrong et al., 2001; Hole, 2007; Bolduc et al., 2010) but no consensus has emerged. Several studies concluded that DM generally performs well. However, Hirschberg and Lye (2010) underline that most Monte Carlo tests (e.g., Hole 2007; Dorfman et al. 1990) consider situations congenial to Dm thus biasing the comparison. In particular, they show that, when estimates of attribute and cost coefficients are correlated and with the same sign of WTP, DM- and FM-based CI may diverge even if the estimate of the cost parameter is highly significant. Such situations often arise in practice (e.g. efficient designs).

The paper provides motivated suggestions concerning which method to use when constructing robust CI for WTP measures in different contexts. It contributes in three ways to the extant literature: i) provides a comprehensive illustration and a systematic comparison; ii) proposes new methods (e.g. bootstrap bias-corrected accelerated, Student-t test bootstrap inversion); iii) suggests new performance indicators (e.g. left and right rejection rate).

The Monte Carlo study reported uses, as in Hole (2007), different data sets. For each model specification and various sample sizes, M=1000 different data sets are generated. A multinomial logit model is fitted to each data set, its parameters estimated via ML and WTP measures and CI are calculated. These M sample values of the CI are used to evaluate coverage rates, left and right rejection probabilities, average interval length and interval shape attained by different methods. Furthermore, the different approaches are tested using real data coming from two surveys with specific characteristics and goals: i) stated preferences for measuring service quality in local public transport; ii) revealed preferences for studying airport choice. The contexts considered when assessing the performance of the various methods are: i) correct model specification, ii) cost parameter approaching zero, iii) non-orthogonal experimental designs characterized by correlation among estimates having the same sign of WTP, iv) neglected unobserved heterogeneity, v) neglected unobserved heteroschedasticity.

The main findings are:

1) All the scenarios revealed a degree of skewness in WTP distribution translating into asymmetric CI while DM produces symmetric intervals around WTP point estimates. This problem has been neglected so far since it usually does not affect CI coverage rates even if it impacts left and right rejection probabilities. In fact, the number of times the real WTP value lies below (above) the inferior (superior) limit of the CI significantly differs from its correct value, being too large on one side and too small on the other, even if the global coverage rate is not altered. Since in practice the mean values of the intervals are generally greater than point estimates (e.g. Armstrong et al., 2001), symmetric intervals would undervalue WTP estimates.
2) WTP skewed distributions are more relevant in case of model misspecification and cost parameter approaching zero. Skewness tends to decrease as sample size rises thus using symmetric CI becomes less problematic with larger samples. Bolduc et al. (2010) find, however, very large samples are needed to compensate for small cost parameter values.
3) WTP distribution can be very skewed even for well-specified models and reasonable cost parameter values in presence of non-orthogonal designs and correlation between estimates having the same sign of WTP. A weak correlation too can highly deteriorate DM coverage rate. When using real data DM-based CI are quite different from those calculated employing FM, LRIM, and other bootstrap methods. DM-based CI are less informative and often include zero at the (1-α)□level even with a statistically significant coefficient.
4) Test-inversion-based methods are not affected by small cost parameter values and are simple/fast to calculate when compared to bootstrap ones. Monte Carlo simulations confirm the intuition by Armstrong et al. (2001) that LRIM-based CI are usually contained in FM-based ones, making the first preferable to the second.

In summary, the paper shows that DM-based CI can deliver misleading results in situations that frequently arise in practice. We suggest using LRIM since they produce CI: i) not necessarily symmetric; ii) not affected by small cost parameter values; iii) having the correct coverage rate under all the scenarios considered and producing, on average, shorter CI than FM. Provided the cost parameter is not too small, one could also use bootstrap methods of the percentile family. Despite requiring a greater computational time, these methods deliver, as a by-product, the entire simulated WTP distribution, which might be of interest for policy evaluation.

References

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

Bolduc, D., Khalaf, L. and Yélou, C. (2010) Identification robust confidence set methods for inference on parameter ratios with application to discrete choice models. Journal of Econometrics, 157, 317-327.

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

Fieller, E.C. (1932) The distribution of the index in a normal bivariate population. Biometrika, 24, 428-440.

Fieller, E.C. (1940) The biological standardization of insulin. Supplement to the Journal of the Royal Statistical Society, 7, 1-64.

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

Hinkley, D.V. (1969) On the ratio of two correlated normal variables. Biometrika, 56, 635-639.

Hirschberg, J. and Lye, J. (2010) A Geometric Comparison of the Delta and Fieller Confidence Intervals. The American Statistician, 64, 234-241.

Hole, A.R. (2007) A comparison of approaches to estimating confidence intervals for willingness to pay measures. Health Economics, 16, 827-840.

Krinsky, I. and Robb, A.L. (1986) On approximating the statistical properties of elasticities. The Review of Economics and Statistics, 68, 715-719.

Krinsky, I. and Robb, A.L. (1990) On approximating the statistical properties of elasticities: A correction. The Review of Economics and Statistics, 72, 189-190.