Transportation Research Part B 167 (2023) 54–78

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Prediction and confidence intervals of willingness-to-pay for mixed logit models

Luisa Scaccia [a,*], Edoardo Marcucci [b,c], Valerio Gatta [b,c]

[a] Dipartimento di Economia e Diritto, Università degli Studi di Macerata, Italy
[b] Dipartimento di Scienze Politiche, Università degli Studi Roma Tre, Italy
[c] Faculty of Logistics, Molde University College, Norway

ARTICLE INFO

Keywords:
Willingness to pay
Mixed logit
Confidence intervals
Prediction intervals
Delta method
Krinsky and Robb method

ABSTRACT

Heterogeneity in agents' preferences is generally analysed through mixed logit models, which assume taste parameters are distributed in the population according to a certain mixing distribution. As a result, if the utility function is linear in attributes, the willingness to pay is the ratio of two random parameters and is itself random. This paper proposes a technique built on the Delta method, partly analytical and partly based on simulations, to obtain the sampling distribution of the willingness to pay, accounting for both heterogeneity and sampling error. The paper contributes to the literature by: (i) redressing some imprecisions in Bliemer and Rose (2013) that produce biased results; (ii) proposing a faster estimation process, compared to the Krinsky and Robb (1986, 1990) method that, relying on simulation only, proves computationally more demanding; (iii) comparing the performance of different methods using both synthetic and real data sets. The paper shows, via a Monte Carlo study, that the method we develop and the Krinsky and Robb one produce similar results, while outperforming that proposed by Bliemer and Rose.

1. Introduction

Willingness to pay (WTP[1]) is the price an agent would pay to obtain an improvement in a specific attribute of a desired good or service. WTP plays a central role in many cases such as, for instance, when a private company has to decide on the selling price of a given good or when a policy maker has to set the price agents have to pay when using a public service (e.g., public transportation). Furthermore, it is often used for economic evaluations such as in cost–benefit analysis.

Discrete choice models, representing a widespread technique to derive WTP measures, rest upon random utility theory. Stated/revealed preferences data are used to estimate parameters of the utility function. Whenever the utility is linear in the attributes, one can calculate the WTP for a given attribute as the ratio of that attribute parameter and the estimate of the marginal utility of income. The latter is typically obtained as the negative of a price parameter (Train, 2009).

Utility parameters for all agents and WTP are constant when using the traditional multinomial logit (MNL) of McFadden (1974). However, the computed WTP is itself an estimate since model estimation yields an estimate of the true parameters. Informed policy-making needs the calculation of WTP standard errors and confidence intervals characterized by good statistical properties. Both the Delta method and the Krinsky and Robb (KR) method (Krinsky and Robb, 1986, 1990) represent the standard statistical tools for achieving this goal. Alternative methods used rest upon the inversion of Wald-type test (Bolduc et al., 2010), t-test or likelihood ratio

* Correspondence to: Dipartimento di Economia e Diritto, Università degli Studi di Macerata, Via Crescimbeni 14, 62100 Macerata, Italy.
E-mail address: luisa.scaccia@unimc.it (L. Scaccia).

[1] All acronyms used in this paper are listed in Appendix A.

https://doi.org/10.1016/j.trb.2022.11.007
Received 13 April 2022; Received in revised form 20 September 2022; Accepted 16 November 2022
Available online 29 November 2022

test (Armstrong et al., 2001). Gatta et al. (2015) provide, for all these methods, a comprehensive comparison of their performances when estimating WTP confidence intervals within a MNL model framework (see also Hole, 2007).

Utility parameters are random and vary according to some mixing distribution, accounting for agent's heterogeneity in preferences, when using mixed multinomial logit (MXL) model (McFadden and Train, 2000), also referred to as random parameters logit model. WTP is, therefore, random in this case. This is true also when MXL model is estimated in WTP space, instead of preference space (Train and Weeks, 2005; Scarpa et al., 2008).

When WTP is random, the policy maker will, for sure, be interested in many different quantities, such as the average WTP (i.e. the WTP of an agent with average attribute and price parameters), its *standard error* and/or its *confidence interval*. Additionally, one might be interested in the WTP *prediction interval*, accounting for the uncertainty due both to sampling error and WTP heterogeneity. This provides an estimate of the WTP lower and upper bound of a certain percentage of the population. Furthermore, the probability of WTP exceeding a given level might help estimating the proportion of agents willing to pay more than a pre-set threshold. Conversely, one might also want to know the quantiles of the WTP distribution, representing the maximum price a given proportion of agents is willing to pay.

The KR procedure is the only one that allows obtaining any possible information, since it delivers the whole simulated sampling distribution of the WTP. Within a MXL framework, this method, discussed by Hensher and Greene (2003) and used in both Sillano and Ortúzar (2005) and Michaud et al. (2013), proves particularly time consuming since it involves Monte Carlo simulations in a high number of dimensions.

Bliemer and Rose (2013) provide an alternative, based on an extension of the Delta method to the case of random parameters. This method, partly analytical and partly based on simulations, allows saving computing time. However, it does not deliver the whole WTP sampling distribution. It just provides mean and variance, thus enabling the computation of only prediction intervals and the standard errors of the predictions.

Other related works deal with standard errors and interval estimates at the individual-specific level. Among these, one should recall (Daziano and Achtnicht, 2014) considering the MXL in a Bayesian framework, and Sarrias and Daziano (2018) focusing on the latent class logit model within a frequentist setting. Fosgerau (2006), instead, proposes a different way of obtaining the WTP distribution. He employs non parametric techniques that do not require the preliminary specification of the mixing distribution for the taste parameters.

This paper focuses on parametric mixing distributions adopting a frequentist framework, since this is the most commonly used approach. It extends the Delta method to the random parameter case. It differs from the one proposed by Bliemer and Rose (2013) (denoted D_{BR} in the rest of the paper) since it treats the WTP sampling distribution as a mixture of normal distributions with the standard deviations of each component estimated through the Delta method. The paper contributes to the literature by: (i) improving with respect to D_{BR}, characterized by some technical imprecisions giving rise to biased results; (ii) improving with respect to KR that, representing a simulation-based approach to the calculation of confidence and prediction intervals, is a much more computationally demanding method compared to the one here proposed; (iii) comparing the performances of the three methods via synthetic data sets and the same empirical data (Bliemer and Rose, 2013) used in their study, so to clarify the relative advantages not only from an abstract perspective but also by showing their practical implications.

In more detail, the approach here proposed (denoted D_{SMG} in the rest of the paper) provides both the advantages of KR and those of D_{BR}. In fact, D_{SMG}, as it is also true for KR, produces the whole WTP simulated sampling distribution, thus allowing the estimation of any desired indicator characterizing the WTP. Additionally, as it is also true for D_{BR}, D_{SMG} being partly analytical allows reducing the computational burden characterizing KR.

The illustration and discussion of the different methods take place within a preference space context. However, they can seamlessly be transferred in WTP space without any adjustment. The paper compares D_{BR}, D_{SMG} and KR via a Monte Carlo study. Data are generated under different scenarios mimicking real situations and considering various taste parameter distributions. A real data set is also used to illustrate the practical relevance of the issues raised in the simulation study.

The rest of the paper is organized as follows. Section 2 briefly recalls random utility models and WTP measures. Section 3 illustrates the Delta method within a fixed parameters context, summarizes the D_{BR} extension to the random parameters case, discusses its shortcomings, and develops the alternative D_{SMG} approach. Section 4 considers the KR method. The Monte Carlo comparison between the three methods is proposed in Section 5, while Section 6 shows an application to real data. Section 7 summarizes the findings, concludes by suggesting general guidelines and provides readers with the link to a spreadsheet that implements the D_{SMG} methodology.

2. Logit models and WTP

Consider a sample of N decision makers facing J different alternatives in T different choice experiments. The choice of agent n, for $n = 1, \dots, N$, is typically described as follows:

$$y_{int} = \begin{cases} 1 & \text{if } U_{int} \geq U_{jnt} \text{ for } j = 1, \dots, J \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where

$$U_{int} = V_{int} + \epsilon_{int} \tag{2}$$

is the unobservable utility that n derives from alternative i (for $i = 1, \dots, J$), in choice experiment t (for $t = 1, \dots, T$), V_{int} is the observable utility and ϵ_{int} is an error term. Observable utility is a (linear or non-linear) function g_i of some known attribute levels for that alternative and a vector of unknown parameters:

$$V_{int} = g_i(X_{int}|\beta),$$

where X_{int} is a $(1 \times K)$ vector of attributes and β is a $(K \times 1)$ vector of parameters. Often, a linear formulation is assumed, so that

$$V_{int} = X_{int}\beta. \tag{3}$$

The choice probability can, therefore, be represented as follows:

$$P_{int} = P\left(U_{int} \geq U_{jnt}, \text{ for } j = 1, \dots, J\right).$$

The assumptions made on the error term in (2) determine different analytical models. In particular, when the error terms are independently and identically distributed (i.i.d.) as an extreme value type 1 distribution, one can calculate the probability of choosing a certain alternative following a logit type model (McFadden, 1974).

Dropping all the subscripts to lighten notation, one can write $U = V + \epsilon$. When V is specified as linear in the alternative attributes, as in (3), the total variation of U with respect to joint variations in the kth attribute X_k and the cost attribute X_c is $dU = \beta_k dX_k + \beta_c dX_c$. Setting this expression equal to zero and solving for dX_c/dX_k yields the change in cost that keeps utility unchanged given a change in X_k:

$$\frac{dX_c}{dX_k} = w_k = -\frac{\beta_k}{\beta_c},$$

which represents the WTP, w_k, for an improvement in kth attribute.[2] In the more general case of a non-linear utility function, the WTP of attribute k is $w_k = -\dfrac{dg_j/dX_k}{dg_j/dX_c}$. While this paper focuses on the case of linear utility functions, the theory exposed and the results obtained herein also apply to the general non-linear case.

Since the parameters β, describing agents' preferences, are unknown, one needs to estimate them. Assuming homogeneous agents implies fixed parameters. Alternatively, when unobserved heterogeneity[3] in agents' preferences is present, parameters are assumed random and distributed according to some distributions over agents. In this case, one can estimate the structural parameters of these distributions. For example, when assuming a normal distribution for β_k one can estimate the mean μ_k and standard deviation σ_k. When errors are i.i.d. extreme value type 1, and agents are homogeneous, this implies a MNL model as well as a fixed unknown value for WTP. Assuming heterogeneous agents, instead, implies a MXL model with random WTP distributed over the agents' population. In both cases, the WTP estimate or that of its distributional parameters suffer from sampling error. Therefore, calculating WTP confidence or prediction intervals as well as providing a measure of uncertainty in estimates or predictions become interesting issues that the next sections explore.

3. Delta method

3.1. The fixed parameter case

The Delta method is commonly used to determine the standard error for a function of the parameter estimates. If $\hat{\beta} \xrightarrow{D} N(\beta, \Omega_{\hat{\beta}})$ and h is a differentiable function, then

$$h(\hat{\beta}) \xrightarrow{D} N(h(\beta), \nabla_\beta h(\beta)^T \Omega_{\hat{\beta}} \nabla_\beta h(\beta)),$$

where $\nabla_\beta h(\beta)$ is the Jacobian of $h(\beta)$.

When considering the WTP of a MNL model, i.e. $w_k = h(\beta_k, \beta_c) = -\frac{\beta_k}{\beta_c}$, the Delta method yields

$$\hat{w}_k \xrightarrow{D} N\left(w_k, \begin{pmatrix} \nabla_{\beta_k} w_k \\ \nabla_{\beta_c} w_k \end{pmatrix}^T \Omega_{\hat{\beta}_k, \hat{\beta}_c} \begin{pmatrix} \nabla_{\beta_k} w_k \\ \nabla_{\beta_c} w_k \end{pmatrix}\right),$$

with $\Omega_{\hat{\beta}_k, \hat{\beta}_c}$ being the variance–covariance sub-matrix for parameter estimates $\hat{\beta}_k$ and $\hat{\beta}_c$, $\nabla_{\beta_k} w_k = -1/\beta_c$ and $\nabla_{\beta_c} w_k = \beta_k/\beta_c^2$. This leads to the well known asymptotic standard error formula for $\hat{w}_k$ found in Armstrong et al. (2001) and Daly et al. (2012a), among others, i.e.

$$\text{se}(\hat{w}_k) = \frac{1}{\beta_c}\sqrt{\text{Var}(\hat{\beta}_k) + 2w_k\text{Cov}(\hat{\beta}_k, \hat{\beta}_c) + w_k^2\text{Var}(\hat{\beta}_c)}. \tag{4}$$

[2] Note that the kth attribute can be either numerical or categorical. In this last case, dummy or effects coding are commonly used, but the definition of w_k is still valid. For example, X_k might be a dummy equal to 1 if class travel is "comfort" and 0 if it is "economy". Then, $dX_k = 1$ indicates the variation of the class travel from "economy" to "comfort", and $w_k = -\frac{\beta_k}{\beta_c}$ is the WTP for moving from "economy" to "comfort" class.

[3] Unobserved heterogeneity refers to taste variations that are purely random or due to unobserved agents' characteristics. Observed heterogeneity refers, instead, to tastes that vary systematically with respect to observed socio-economic or demographic characteristics. Fixed parameters logit models can capture observed heterogeneity by interacting attributes with agents' characteristics, while random parameters are used to deal with unobserved heterogeneity.

Substituting parameters with their estimates in (4) allows calculating $\widehat{se}(\hat{w}_k)$ that can be used in the construction of confidence intervals at the $(1-\alpha)$ level as:

$$\hat{w}_k \pm z_{\alpha/2}\widehat{se}(\hat{w}_k) \tag{5}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)$th percentile of the standard normal distribution.

3.2. The random parameter case

When assuming stochastic taste parameters, w_k is a random variable itself characterized by a certain probability density function $f_{w_k}(x)$, which is determined by the type of distributions chosen for the taste parameters β. For example, if β_k is assumed fixed and β_c is assumed log-normal, w_k will also have a log-normal distribution. If, instead, a normal distribution is chosen for β_c, w_k will have a distribution with infinite moments (Daly et al., 2012b).

When w_k is random, one might want to compute confidence intervals, prediction intervals, or both, as well as other quantities of interest. Let us suppose that the moments of w_k do exist, $\mathrm{E}(w_k)$ be its expected value, and $\bar{w}_k$ be an asymptotically unbiased and normally distributed sample estimator of $\mathrm{E}(w_k)$, with standard error $se(\bar{w}_k)$. Then the confidence interval for the expected WTP, $\mathrm{E}(w_k)$, is

$$\bar{w}_k \pm z_{\alpha/2}\widehat{se}(\bar{w}_k), \tag{6}$$

where $\widehat{se}(\bar{w}_k)$ is a sample estimate of $se(\bar{w}_k)$. If, instead, the aim is to predict an outcome of the variable w_k, rather than to estimate its expected value, one could still use $\bar{w}_k$ as a point estimate of this outcome. However, while the uncertainty in estimating $\mathrm{E}(w_k)$ is only due to the *sampling error*, the uncertainty in predicting an outcome of w_k, must also account for the natural variability of the distribution $f_{w_k}(x)$ along with the sampling error. The *prediction standard error* quantifies both uncertainty types. It might be tempting, then, to calculate prediction intervals for the random variable w_k similarly to (6), thus simply substituting the standard error with the prediction standard error. This is exactly what Bliemer and Rose (2013) do. However, unless β_k is assumed normal and β_c fixed, w_k is not normally distributed, thus prediction intervals calculated in such a way might suffer from a coverage probability being different from $(1-\alpha)$. One should, therefore, develop a different strategy such as that illustrated in Section 3.2.2.

3.2.1. The D_{BR} approach

Bliemer and Rose (2013) extend the Delta method to compute prediction intervals[4] for w_k in the mixed logit context. They start with the case of β_k and β_c independently distributed. They express the parameters as functions of structural parameters $\theta_k \in \mathbb{R}^{p_k}$ and $\theta_c \in \mathbb{R}^{p_c}$, of dimension p_k and p_c, respectively, using parameter-free distributions (standard normal or standard uniform distribution), i.e.

$$\beta_k = \beta_k(\theta_k, z_k) \qquad \text{and} \qquad \beta_c = \beta_c(\theta_c, z_c), \tag{7}$$

where $z_k \in \mathbb{R}^{s_k}$ and $z_c \in \mathbb{R}^{s_c}$ are vectors of draws (of dimension s_k and s_c, respectively) from standard distributions. Subsequently, they use these functions within the Delta method to map the variance–covariance matrix of the structural parameter estimates $\hat{\theta} = (\hat{\theta}_k, \hat{\theta}_c)$ to that of $\hat{\beta} = (\hat{\beta}_k, \hat{\beta}_c)$, so to obtain the standard error of $-\hat{\beta}_k/\hat{\beta}_c, \forall k$. For example, if $\beta_k \sim N(\mu_k, \sigma_k^2)$ one can write $\beta_k = \mu_k + \sigma_k z_k$, where $z_k \sim N(0,1)$, so that $\theta_k = (\mu_k, \sigma_k)$. Bliemer and Rose (2013) provide a useful list of distributions which can be obtained from a standard normal or uniform distribution.

Under this framework, one can write the variable w_k as follows:

$$w_k = -\frac{\beta_k(\theta_k, z_k)}{\beta_c(\theta_c, z_c)}, \tag{8}$$

i.e. as a function of the fixed parameters $\theta = (\theta_k, \theta_c) \in \mathbb{R}^{p_k+p_c}$ and the random variables $z = (z_k, z_c) \in \mathbb{R}^{s_k+s_c}$ accounting for heterogeneity. Considering a particular value z and conditioning on it, the model collapses to the fixed parameter case, where the WTP is no longer a random variable but a fixed value:

$$w_k(z) = -\frac{\beta_k(\theta_k|z_k)}{\beta_c(\theta_c|z_c)}. \tag{9}$$

Substituting θ with its estimate $\hat{\theta}$ leads to $\hat{w}_k(z) = -\frac{\beta_k(\hat{\theta}_k|z_k)}{\beta_c(\hat{\theta}_c|z_c)}$, which is an estimate of the fixed value $w_k(z)$ affected by a standard error. Bliemer and Rose (2013) state that, according to the Delta method:

$$\hat{w}_k(z) \xrightarrow{D} N\left(w_k(z), \begin{pmatrix} \nabla_{\theta_k} w_k(z) \\ \nabla_{\theta_c} w_k(z) \\ \nabla_{z_k} w_k(z) \\ \nabla_{z_c} w_k(z) \end{pmatrix}^T \begin{pmatrix} \Omega_{\hat{\theta}_k,\hat{\theta}_c} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{1} \end{pmatrix} \begin{pmatrix} \nabla_{\theta_k} w_k(z) \\ \nabla_{\theta_c} w_k(z) \\ \nabla_{z_k} w_k(z) \\ \nabla_{z_c} w_k(z) \end{pmatrix}\right), \tag{10}$$

[4] It is important to note that the econometric literature typically uses "confidence intervals" for both types of intervals. Bliemer and Rose (2013) use the term "confidence interval" while computing a "prediction interval". This paper, instead, keeps the two concepts separate and compute them both using an extension of the Delta method.

where $\mathbf{0} \in \mathbb{R}^{(p_k+p_c)\times(s_k+s_c)}$ is a matrix of zeros and $\mathbf{1} \in \mathbb{R}^{(s_k+s_c)\times(s_k+s_c)}$ is a diagonal matrix of ones due to the fact that z_k and z_c are independently distributed and they are also independent from the estimators $\hat{\theta}_k$ and $\hat{\theta}_c$. The *unconditional* mean WTP estimate, according to Bliemer and Rose (2013), is defined as[5]

$$\bar{w}_k = \int_z \hat{w}_k(z) d F_z(z) \tag{11}$$

where $F_z(z) = F_{z_k}(z_k)F_{z_c}(z_c)$ is the cumulative distribution function of the standard independently distributed z_k and z_c. Since integrals are linear operators, the resulting asymptotic distribution of $\bar{w}_k$ is also normally distributed, with expectation and variance given by the integrals over z of the conditional expectation and variance in (10).

Using Monte Carlo simulation, one can obtain the following approximation of the unconditional mean WTP estimate:

$$\bar{w}_k \approx \frac{1}{R}\sum_{r=1}^{R} \hat{w}_k\left(z^{(r)}\right), \tag{12}$$

whose prediction standard error[6] is approximated in Bliemer and Rose (2013) as the average, over z of the conditional standard errors, that is:

$$\text{pse}(\bar{w}_k) \approx \frac{1}{R}\sum_{r=1}^{R} \sqrt{\begin{pmatrix} \nabla_{\theta_k} w_k(z^{(r)}) \\ \nabla_{\theta_c} w_k(z^{(r)}) \\ \nabla_{z_k} w_k(z^{(r)}) \\ \nabla_{z_c} w_k(z^{(r)}) \end{pmatrix}^T \begin{pmatrix} \Omega_{\hat{\theta}_k,\hat{\theta}_c} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{1} \end{pmatrix} \begin{pmatrix} \nabla_{\theta_k} w_k(z^{(r)}) \\ \nabla_{\theta_c} w_k(z^{(r)}) \\ \nabla_{z_k} w_k(z^{(r)}) \\ \nabla_{z_c} w_k(z^{(r)}) \end{pmatrix}}, \tag{13}$$

where $z^{(r)} = \left(z_k^{(r)}, z_c^{(r)}\right)$ are draws from $F_{z_k}(z_k)$ and $F_{z_c}(z_c)$.

One can estimate the prediction standard error in (13) substituting parameter and variance–covariance matrix estimates in its equation and use it to compute prediction intervals as in (6). Bliemer and Rose (2013) clearly state that "... the [prediction] standard error of the unconditional WTP computed directly from the conditional WTP's, considering only simulated values of $\hat{w}_k\left(z_k^{(r)}, z_c^{(r)}\right)$ is incorrect as it ignores the uncertainty (expressed in the variance–covariance matrix) of the distributional parameter estimates $\hat{\theta}$, while we explicitly take this into account ...". They correctly stress that one should account for both the variability due to heterogeneity and for the variability linked to the uncertainty in the estimate of θ when constructing prediction intervals for random WTP. However, we do not completely agree with this procedure resting on Eqs. (10) and (13) (see Section 3.2.2). Moreover, computing prediction intervals as in (6) produces symmetric intervals. This might represent a shortcoming compared to the KR method and the one proposed here. In fact, in some cases, it might deliver questionable results, as discussed in Section 3.2.2 and illustrated in Sections 5 and 6.2.

3.2.2. The D_{SMG} approach

Let us go back to the conditional WTP in (9). When conditioning on z, z is no longer a random variable but a fixed known value instead. As a consequence, $w_k(z)$ is also a fixed and unknown value dependent on the unknown parameter vector θ. When estimating θ and $w_k(z)$, one can calculate the standard error of $\hat{w}_k(z)$ via the Delta method. We claim that the Delta method is to be applied in exactly the same way as it would be in the fixed parameter case. This is correct given we are conditioning on z, thus β becomes a vector of fixed parameters since θ is fixed. We suggest substituting Eq. (10) with:

$$\hat{w}_k(z) \xrightarrow{D} N\left(w_k(z), \begin{pmatrix} \nabla_{\theta_k} w_k(z) \\ \nabla_{\theta_c} w_k(z) \end{pmatrix}^T \Omega_{\hat{\theta}_k,\hat{\theta}_c} \begin{pmatrix} \nabla_{\theta_k} w_k(z) \\ \nabla_{\theta_c} w_k(z) \end{pmatrix}\right), \tag{14}$$

where, unlike in (10), the variability of $\hat{w}_k(z)$ only depends on the sample variability of the estimates $\hat{\theta}$, while z is fixed.

Eq. (11) reports the unconditional mean WTP estimate, $\bar{w}_k$, that can be evaluated through a Monte Carlo approximation as in (12). It is asymptotically normal with mean $\text{E}(w_k)$, given by the integral over z of the conditional expectations $\text{E}(\hat{w}_k(z)) = w_k(z)$. We propose an alternative method to compute the standard error of $\bar{w}_k$, with respect to what Bliemer and Rose (2013) suggest. In fact, we explicitly consider that the $\hat{w}_k(z)$ are not independent over z, thus we do not take the integral over z of the conditional standard errors.[7] In conclusion, to calculate the variance of $\bar{w}_k$ one should use the following formula:

$$\text{Var}(\bar{w}_k) = \text{Var}\int_z \hat{w}_k(z) d F_z(z) = \int_{z_1}\int_{z_2} \text{Cov}(\hat{w}_k(z_1), \hat{w}_k(z_2)) d F_z(z_1) d F_z(z_2),$$

[5] Note that Bliemer and Rose (2013) use the notation $\hat{w}_k$, while we prefer $\bar{w}_k$, since we use the former to denote the random variable obtained from (8) when substituting the parameters θ with their sample estimates.

[6] Bliemer and Rose (2013) refer to it simply as standard error and indicate it as se$(\cdot)$. We prefer distinguishing between prediction standard error, accounting for the uncertainty due to both sampling error and heterogeneity, and standard error, accounting for sampling variability alone, since we compute both quantities using the Delta method.

[7] In fact, even for independent variables, the standard error of their sum is not the sum of their standard errors, while it is true that the variance of their sums is the sum of their variances. We show in Appendix B that the prediction standard error in (13) correctly accounts for uncertainty due to heterogeneity but not for that due to sampling error, for the normal divided by fixed case. The sampling error, however, becomes progressively less important compared to heterogeneity when the sample size increases. In the large sample size case, the bias is negligible. Nevertheless, for different mixing distributions, even the variability due to heterogeneity is incorrectly estimated. The problem does not vanish even when the sample size increases. Thus, the estimate of the prediction standard error in (13) may be both biased and inconsistent, depending on the chosen mixing distributions.

where one could use the Delta method (Klein, 1953, p. 258) to approximate $\mathrm{Cov}(\hat{w}_k(z_1), \hat{w}_k(z_2))$. The standard error of $\bar{w}_k$ can be estimated as[8]:

$$\mathrm{se}(\bar{w}_k) \approx \frac{1}{R}\sqrt{\sum_{r=1}^{R}\sum_{s=1}^{R}\begin{pmatrix}\nabla_{\theta_k} w_k(z^{(r)})\\ \nabla_{\theta_c} w_k(z^{(r)})\end{pmatrix}^T \Omega_{\hat{\theta}_k,\hat{\theta}_c}\begin{pmatrix}\nabla_{\theta_k} w_k(z^{(s)})\\ \nabla_{\theta_c} w_k(z^{(s)})\end{pmatrix}}. \tag{15}$$

We stress that the standard error in (15) only accounts for the variability of $\bar{w}_k$ due to sampling error, but not for the variability due to heterogeneity.[9] As such, one can use it (once estimated on the data) to build a confidence interval for $E(w_k)$ as in (6), since $\bar{w}_k$ in (11) is an asymptotically unbiased and normally distributed estimator of $E(w_k)$.

Let us now discuss the estimation of a prediction interval for the random variable w_k and a prediction standard error. Consider the conditional estimates $\hat{w}_k(z)$ and their asymptotic normal distribution in (14). Integrating the normal distribution in (14) over z, one obtains the asymptotic distribution of the unconditional variable $\hat{w}_k$:

$$f_{\hat{w}_k}(x) = \int_z \phi\left(x; w_k(z), \mathrm{Var}\left(\hat{w}_k(z)\right)\right) dF_z(z), \tag{16}$$

where $\mathrm{Var}(\hat{w}_k(z))$ is specified in (14) and $\phi(x;\mu,\sigma^2)$ denotes the density of a normal distribution with mean μ and variance σ^2, evaluated in x. Notice the difference between (11) and (16). While (11) is the integral of the conditional $\hat{w}_k(z)$ over z, providing the expected value of these quantities with respect to z, (16) is the integral of the *densities* of the conditional $\hat{w}_k(z)$ and produces the density of the unconditional $\hat{w}_k$, which is a continuous mixture of normal distributions with weights given by the joint density of (z_k, z_c).

The moments of a mixture can be easily computed (Withers et al., 2015). In particular, the mean of a mixture is the average of the components' means, while the variance is the mean of the components' variances plus the variance for the components' means. In our framework, this implies both

$$\mathrm{E}(\hat{w}_k) = \int_z w_k(z) dF_z(z) = \mathrm{E}(w_k),$$

that one can estimate as in (11) and approximate as in (12), as well as

$$\mathrm{Var}(\hat{w}_k) = \int_z \mathrm{Var}(\hat{w}_k(z)) dF_z(z) + \int_z (w_k(z) - \mathrm{E}(w_k))^2 dF_z(z). \tag{17}$$

The first integral is the part of variability of $\hat{w}_k$ due to the sampling error. This can be estimated via a Monte Carlo approximation as $\frac{1}{R}\sum_{r=1}^{R}\widehat{\mathrm{Var}}(\hat{w}_k(z^{(r)}))$, where

$$\widehat{\mathrm{Var}}(\hat{w}_k(z^{(r)})) = \begin{pmatrix}\nabla_{\hat{\theta}_k} \hat{w}_k(z^{(r)})\\ \nabla_{\hat{\theta}_c} \hat{w}_k(z^{(r)})\end{pmatrix}^T \hat{\Omega}_{\hat{\theta}_k,\hat{\theta}_c}\begin{pmatrix}\nabla_{\hat{\theta}_k} \hat{w}_k(z^{(r)})\\ \nabla_{\hat{\theta}_c} \hat{w}_k(z^{(r)})\end{pmatrix} \tag{18}$$

represents the sample estimate of $\mathrm{Var}(\hat{w}_k(z))$ appearing in (14) and used in (17). The second integral in (17) accounts for the dispersion of the conditional $w_k(z)$ around their mean. This is $\mathrm{Var}(w_k)$ and constitutes the variability due to heterogeneity. One can estimate it as

$$\widehat{\mathrm{Var}}(w_k) = \frac{1}{R}\sum_{r=1}^{R}(\hat{w}_k(z^{(r)}) - \bar{w}_k)^2. \tag{19}$$

Therefore, the distribution of $\hat{w}_k$ in (16) accounts for both the variability due to sampling error and the variability due to heterogeneity. One can safely use it to build prediction intervals. However, $f_{\hat{w}_k}(x)$, being a mixture of normal distributions, might assume various shapes and be very far from symmetric. We, thus, consider wise to refrain from building a prediction interval using normal theory. Instead, we propose using the percentiles of $f_{\hat{w}_k}(x)$. After substituting $w_k(z)$ and $\mathrm{Var}(\hat{w}_k(z))$ with their sample estimates, one can use Monte Carlo approximation to get:

$$\hat{f}_{\hat{w}_k}(x) \approx \frac{1}{R}\sum_{r=1}^{R}\phi\left(x; \hat{w}_k(z^{(r)}), \widehat{\mathrm{Var}}(\hat{w}_k(z^{(r)}))\right). \tag{20}$$

A $(1-\alpha)$ level prediction interval for w_k is given by

$$\left[\hat{F}^{-1}_{\hat{w}_k}(\alpha/2) \quad ; \quad \hat{F}^{-1}_{\hat{w}_k}(1-\alpha/2)\right], \tag{21}$$

where $\hat{F}^{-1}_{\hat{w}_k}$ is the inverse of the estimated cumulative distribution of $\hat{w}_k$. Although there is no simple formula for the percentiles of a mixture of normal distributions, a line search can be performed to get the percentiles. In practice, $\hat{F}_{\hat{w}_k}(x)$ can be computed from any statistical package as

$$\hat{F}_{\hat{w}_k}(x) \approx \frac{1}{R}\sum_{r=1}^{R}\Phi\left(x; \hat{w}_k(z^{(r)}), \widehat{\mathrm{Var}}(\hat{w}_k(z^{(r)}))\right), \tag{22}$$

[8] Notice the difference with Eq. (13).

[9] This motivates why we refer to it simply as standard error and not as prediction standard error. We denote it by $\mathrm{se}(\cdot)$.

with Φ denoting the cumulative of ϕ. The percentile $\hat{F}^{-1}_{\hat{w}_k}(\alpha)$ can be obtained through a minimization routine as the value x minimizing $|\hat{F}_{\hat{w}_k}(x) - \alpha|$.

Notice that the approach we propose, as well as the D_{BR}, only requires simulations in a number of dimensions equal to the number of random parameters in the WTP ratio. Therefore, simulations in at most two dimensions are required, i.e. the draws for z_k and z_c. KR requires, instead, simulations in a number of dimensions equal to the number of random parameters in the WTP ratio plus the total number of structural parameters (see Section 4). Nevertheless, as for KR, our approach delivers an estimate of the whole $\hat{w}_k$ distribution, not only of its moments, which can be used to calculate whatever one needs to know about the WTP.

In conclusion, we stress that both D_{SMG} and D_{BR} assume a normal sampling distribution for the estimated structural parameters and for $\hat{w}_k(z)$. This assumption is fulfilled for large enough samples given asymptotic normality of simulated maximum likelihood estimates. Please also notice that, conditioning on z, $\hat{w}_k(z)$ becomes a continuous and differentiable function of the estimates of the structural parameters θ_k and θ_c. Since these estimates are asymptotically normal, the results of Daly et al. (2012a) apply and one can state that $\hat{w}_k(z)$ is also asymptotically normal.

To summarize, D_{SMG} prescribes the following steps:

1. Write the parameters β in their parameter-free distribution form as in (7).
2. Use standard software to get parameter vector estimates $(\hat{\theta}_1, \dots, \hat{\theta}_K, \hat{\theta}_c)$ and estimated variance–covariance matrix $\hat{\Omega}_{(\hat{\theta}_1,\dots,\hat{\theta}_K,\hat{\theta}_c)}$.
3. Obtain a large number R of pseudo-random or quasi-random draws (such as Halton, Sobol, Modified Latin Hypercube Sampling, etc.) for z from the appropriate standard distributions and let $(z_1^{(r)}, \dots, z_K^{(r)}, z_c^{(r)})$, for $r = 1, \dots, R$, denote the set of draws.
4. For each $r = 1, \dots, R$, compute $\beta_k(\hat{\theta}_k, z_k^{(r)})$, for $k = 1, \dots, K$, and $\beta_c(\hat{\theta}_c, z_c^{(r)})$ and, then, compute $\hat{w}_k(z^{(r)}) = -\beta_k(\hat{\theta}_k, z_k^{(r)})/\beta_c(\hat{\theta}_c, z_c^{(r)})$.
5. Use the $\hat{w}_k(z^{(r)})$ to get different quantities of interest, such as:

   - **estimated average WTP,** $\bar{w}_k$, obtained as in (12);

   - **estimated standard error** for $\bar{w}_k$, obtained as

$$\hat{\text{se}}(\bar{w}_k) \approx \frac{1}{R}\sqrt{\sum_{r=1}^{R}\sum_{s=1}^{R}\begin{pmatrix}\nabla_{\hat{\theta}_k}\hat{w}_k(z^{(r)})\\ \nabla_{\hat{\theta}_c}\hat{w}_k(z^{(r)})\end{pmatrix}^T \hat{\Omega}_{\hat{\theta}_k,\hat{\theta}_c}\begin{pmatrix}\nabla_{\hat{\theta}_k}\hat{w}_k(z^{(s)})\\ \nabla_{\hat{\theta}_c}\hat{w}_k(z^{(s)})\end{pmatrix}}; \tag{23}$$

   - **estimated confidence intervals** for the average WTP, $E(w_k)$, obtained as in (6), with $\bar{w}_k$ given in (12) and $\hat{\text{se}}(\bar{w}_k)$ given in (23);

   - **estimated prediction standard error,** $\widehat{\text{pse}}(\bar{w}_k)$, obtained as the square root of an estimate of (17), that is

$$\widehat{\text{pse}}(\bar{w}_k) = \sqrt{\frac{1}{R}\sum_{r=1}^{R}\widehat{\text{Var}}(\hat{w}_k(z^{(r)})) + \widehat{\text{Var}}(w_k)},$$

     with $\widehat{\text{Var}}(\hat{w}_k(z^{(r)}))$ given in (18) and $\widehat{\text{Var}}(w_k)$ given in (19);

   - **estimated prediction intervals** for w_k, obtained as in (21), where $\hat{F}^{-1}_{\hat{w}_k}(\alpha) = x : |\hat{F}_{\hat{w}_k}(x) - \alpha| = \min_x\left(|\hat{F}_{\hat{w}_k}(x) - \alpha|\right)$, with $\hat{F}_{\hat{w}_k}(x)$ given in (22).

4. Krinsky and Robb method

In this section, we briefly review the KR method. This has long been the only viable option to determine confidence and prediction intervals, as well as standard errors and prediction standard errors for the WTP in the random parameter framework. We use this as a benchmark for D_{SMG}. KR consists of the following steps:

1. Write the parameters β in their parameter-free distribution form as in (7).
2. Use standard software to get parameter vector estimates $(\hat{\theta}_1, \dots, \hat{\theta}_K, \hat{\theta}_c)$ and estimated variance–covariance matrix $\hat{\Omega}_{(\hat{\theta}_1,\dots,\hat{\theta}_K,\hat{\theta}_c)}$.
3. Obtain a large number B of first stage structural parameter draws from a multivariate normal distribution with mean $(\hat{\theta}_1, \dots, \hat{\theta}_K, \hat{\theta}_c)$ and variance–covariance matrix $\hat{\Omega}_{(\hat{\theta}_1,\dots,\hat{\theta}_K,\hat{\theta}_c)}$. Let $(\hat{\theta}_1^{(b)}, \dots, \hat{\theta}_K^{(b)}, \hat{\theta}_c^{(b)})$, for $b = 1, \dots, B$ denote these first stage draws.
4. Obtain a large number R of draws $\left(z^{(1)}, \dots, z^{(R)}\right)$ from the appropriate standard distributions of z and compute the second stage draws of the random parameters as $\hat{\beta}_k^{(b,r)} = \beta_k(\hat{\theta}_k^{(b)}, z_k^{(r)})$ and $\hat{\beta}_c^{(b,r)} = \beta_c(\hat{\theta}_c^{(b)}, z_c^{(r)})$.
5. Compute $\hat{w}_k^{(b,r)} = -\hat{\beta}_k^{(b,r)}/\hat{\beta}_c^{(b,r)}$, $\forall b$, $\forall r$ and $\forall k$.
6. Use the $\hat{w}_k^{(b,r)}$ to get different quantities of interest, such as:

   - **estimated average WTP,** $\bar{w}_k$, obtained as:

$$\bar{w}_k = \frac{1}{B}\frac{1}{R}\sum_{b=1}^{B}\sum_{r=1}^{R}\hat{w}_k^{(b,r)};$$

- **estimated standard error** for $\bar{w}_k$, obtained as:

$$\hat{\text{se}}(\bar{w}_k) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\bar{w}_k^{(b)} - \bar{w}_k\right)^2},$$

where $\bar{w}_k^{(b)} = \sum_{r=1}^{R} \hat{w}_k^{(b,r)}/R$;

- **estimated confidence intervals** for the average WTP, $\text{E}(w_t)$, obtained as:

$$\left[\bar{w}_k^{[L]}; \bar{w}_k^{[U]}\right],$$

where $\bar{w}_k^{[L]}$ and $\bar{w}_k^{[U]}$ are the $100\alpha/2$th and $100(1-\alpha/2)$th percentiles of the set $(\bar{w}_k^{(1)}, \ldots, \bar{w}_k^{(B)})$ of replicates for the sample estimate of $\text{E}(w_t)$, sorted in ascending order;

- **estimated prediction standard error,** $\widehat{\text{pse}}(\bar{w}_k)$, obtained as:

$$\widehat{\text{pse}}(\bar{w}_k) = \sqrt{\frac{1}{B}\frac{1}{R}\sum_{b=1}^{B}\sum_{r=1}^{R}\left(\hat{w}_k^{(b,r)} - \bar{w}_k\right)^2};$$

- **estimated prediction intervals** for w_k, obtained as:

$$\left[\hat{w}_k^{[L]}; \hat{w}_k^{[U]}\right],$$

where $\hat{w}_k^{[L]}$ and $\hat{w}_k^{[U]}$ are the $100\alpha/2$th and $100(1-\alpha/2)$th percentiles of the set $(\hat{w}_k^{(1,1)}, \ldots, \hat{w}_k^{(B,R)})$ sorted in ascending order.

KR requires simulations in a number of dimensions equal to the number of random parameters in the WTP ratio plus the total number of structural parameters. The steps above still apply when some of the parameters are fixed. Suppose that β_k is a non-random parameter. We can always think of it as $\beta_k \sim N(\theta_k, 0)$, i.e. as distributed according to a degenerate normal distribution, so that $P(\beta_k = \theta_k) = 1$. Therefore, in step 4 we would obtain $\hat{\beta}_k^{(b,r)} = \hat{\theta}_k^{(b)}$, $\forall b$ and $\forall r$. A small change in step 4 would also accommodate the case of dependently distributed random parameters by drawing the random parameters from their multivariate mixing distribution.

Notice that step 3 provides the simulated sampling distribution of each structural parameter estimator $\hat{\theta}_k$ and $\hat{\theta}_c$, accounting for sampling error, while step 4 provides, for a particular b, the simulated distribution of each random parameter, accounting for heterogeneity. As a consequence, the whole set (i.e. for $b = 1, \ldots, B$ and $r = 1, \ldots, R$) of second stage draws takes into account both the heterogeneity of the random parameters and the sampling error of structural parameter estimates. Similarly, the values $\hat{w}_k^{(b,r)}$, for $b = 1, \ldots, B$ and $r = 1, \ldots, R$, incorporate the variability due to both heterogeneity and sampling error thus providing the simulated sampling distribution of the random variable $\hat{w}_k$. This distribution is asymptotically equivalent to the mixture distribution in (20) obtained through the D_{SMG} (see Section 6). We remark that the normality of the sampling distribution of the structural parameters is the only assumption KR requires. Asymptotic normality of simulated maximum likelihood estimates assures that this assumption is fulfilled for large enough samples.

5. Simulation study

This section compares D_{SMG}, D_{BR}, and KR via a simulation study. We construct a number of data sets mimicking an actual choice experiment, in line with Gatta et al. (2015) and Hole (2007). A number N of hypothetical agents is faced with $T = 16$ different choice tasks, each one presenting $J = 2$ alternatives characterized by three different attributes, denoted X_1, X_2 and X_c, with X_c being the attribute "cost". As in Gatta et al. (2015), X_1 and X_2 have two levels, coded 1 and 2, while X_c has four levels, coded 1, 2, 3 and 4. Dropping all the subscripts for simplicity, the observed difference in utility between the two alternatives is:

$$V_1 - V_2 = \beta_0 + \beta_1(X_{11} - X_{12}) + \beta_2(X_{21} - X_{22}) + \beta_c(X_{c1} - X_{c2}). \tag{24}$$

A single data set can be drawn from a certain MXL model through the following steps: (1) draw values for β_1, β_2 and β_c, independently for each agent, from their mixing distribution and substitute them into (24); (2) draw a value for the error difference $\epsilon_1 - \epsilon_2$, independently for each agent and for each choice task, from a logistic distribution; (3) if this value is less than the difference $V_1 - V_2$, than the first alternative is chosen and the choice variable y takes the value 1 for the first alternative and 0 otherwise. If not, the opposite is true.

We consider different sample sizes and mixing distributions, and generate $M = 1000$ different data sets for each of them. We fit, for each data set, a correctly specified MXL model and estimate its parameters via simulated maximum likelihood. We use D_{SMG}, D_{BR}, and KR to estimate $\bar{w}_1$ and $\bar{w}_2$ and get confidence and prediction intervals, along with standard errors and prediction standard errors. The M Monte Carlo replicates allow evaluating the performance of the various estimators.

We illustrate the asymptotic behaviour of the different estimators by progressively increasing the sample size from $N = 150$ to $N = 300$ and finally using $N = 450$. We report simulations for $N = 300$ and $N = 450$ only when those for $N = 150$ are not satisfactory. The following sections discuss the results.

Table 1
Normal divided by fixed case: mean and standard deviation (in brackets) over Monte Carlo replicates for the estimated average WTP, its standard error and prediction standard error. The superscript (MC) denotes Monte Carlo estimates of benchmark values.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	$\bar{w}_1$	1.000	0.998 (0.074)	0.998 (0.074)	0.998 (0.074)
	$\widehat{se}(\bar{w}_1)$	$0.074^{(MC)}$	0.077 (0.004)	–	0.077 (0.004)
	$\widehat{pse}(\bar{w}_1)$	$0.505^{(MC)}$	0.501 (0.082)	0.500 (0.082)	0.501 (0.082)
$k = 2$	$\bar{w}_2$	0.500	0.503 (0.069)	0.503 (0.069)	0.503 (0.069)
	$\widehat{se}(\bar{w}_2)$	$0.069^{(MC)}$	0.065 (0.004)	–	0.065 (0.004)
	$\widehat{pse}(\bar{w}_2)$	$0.406^{(MC)}$	0.401 (0.084)	0.400 (0.086)	0.401 (0.084)

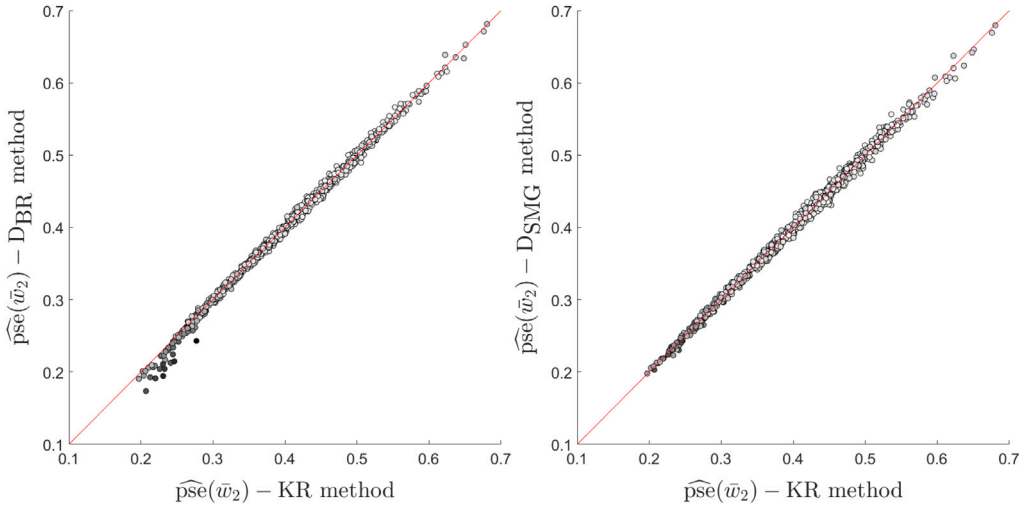

Fig. 1. Estimates of the predictive standard error for w_2. Left panel: estimates obtained through the D_{BR} method vs. those obtained through the KR method. Right panel: estimates obtained through the D_{SMG} method vs. those obtained through the KR method. Darker points correspond to Monte Carlo replicates with larger estimated sampling error. Red line represents the bisector of the first quadrant.

5.1. The normal divided by fixed case

When assuming the parameter at the numerator of WTP is distributed normal and the one at the denominator is fixed, the distribution of the WTP for X_k is known exactly: $w_k \sim N(\mu_{w_k}, \sigma^2_{w_k})$, with $\mu_{w_k} = -\mu_k/\beta_c$ and $\sigma^2_{w_k} = \sigma^2_k/\beta^2_c$. In particular, setting $\beta_0 = 0.5$, $\beta_1 \sim N(\mu_1, \sigma^2_1)$, with $\mu_1 = 1$ and $\sigma^2_1 = 0.25$, $\beta_2 \sim N(\mu_2, \sigma^2_2)$, with $\mu_2 = 0.5$ and $\sigma^2_2 = 0.16$, and $\beta_c = -1$, leads to $w_1 \sim N(1, 0.25)$ and $w_2 \sim N(0.5, 0.16)$.

Table 1 shows the mean of different estimated quantities over Monte Carlo replicates, separately for the D_{SMG}, D_{BR} and KR, in the case of $N = 150$. In brackets please find the standard deviation over replicates, providing a measure of the stability of these estimates. Note that the estimate of the average WTP is simply given by $-\hat{\mu}_k/\hat{\beta}_c$ in the normal divided by fixed case, thus, all the three methods provide the same result. However, only D_{SMG} and KR can estimate the standard error, $se(\bar{w}_k)$, while D_{BR} cannot. Both provide, on average, nearly the same result, which is almost equal to the benchmark value and stable across Monte Carlo replicates.

The three methods are equivalent and reliable, on average, when estimating the prediction standard error. One might, in principle, interpret this result as suggesting that the concerns expressed about the estimator of the prediction standard error in (13) are excessive. The Appendix B shows, for the normal divided by fixed case, that the prediction standard error in (13) correctly accounts for uncertainty deriving from heterogeneity while this is not true for the one arising from sampling error. However, in this simulation, both the choice of the parameters and the sample size are such that sampling error is much less important compared to the variability due to heterogeneity. This is the only reason why the prediction standard error in (13) yields reasonable results, on average. However, the single estimates in the left panel of Fig. 1 show that the prediction standard errors produced by D_{BR} are systematically smaller in comparison to those KR produces, whenever heterogeneity is not too large and, thus, sampling error variability becomes more relevant (darker points in the bottom left corner of left panel). On the other hand, the right panel of Fig. 1 shows the nearly perfect agreement between KR and D_{SMG}.

Table 2 reports and compares the performances of the three methods in computing WTP prediction intervals. For each replicate, one can calculate various measures of interest. The *length* of the interval is the difference between upper and lower bound. The

Table 2
Normal divided by fixed case: length, shape, LRP, RRP, and coverage of 95% level prediction intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. The benchmark for the length of the interval is in italic and represents a lower limit for the length, accounting only for variability due to heterogeneity.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	*1.960*	1.985	1.958	1.989
	Shape	1.000	1.013	1.000	1.015
	LRP	0.025	0.030	0.032	0.030
	RRP	0.025	0.030	0.033	0.030
	Coverage	0.950	0.940	0.935*	0.940
$k = 2$	Length	*1.568*	1.612	1.567	1.613
	Shape	1.000	1.007	1.000	1.009
	LRP	0.025	0.032	0.037*	0.032
	RRP	0.025	0.031	0.037*	0.031
	Coverage	0.950	0.937	0.926***	0.937
$N = 300$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	*1.960*	1.983	1.971	1.984
	Shape	1.000	1.006	1.000	1.009
	LRP	0.025	0.027	0.027	0.027
	RRP	0.025	0.027	0.028	0.027
	Coverage	0.950	0.946	0.944	0.946
$k = 2$	Length	*1.568*	1.587	1.567	1.588
	Shape	1.000	1.002	1.000	1.006
	LRP	0.025	0.029	0.031	0.029
	RRP	0.025	0.030	0.032	0.030
	Coverage	0.950	0.941	0.937	0.941

Table 3
Normal divided by fixed case: length, shape, LRP, RRP and coverage of 95% level confidence intervals for the average WTP. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. The superscript (MC) denotes Monte Carlo estimates of benchmark values.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	$0.298^{(MC)}$	0.300	–	0.301
	Shape	$1.173^{(MC)}$	1.000	–	1.097
	LRP	0.025	0.026	–	0.028
	RRP	0.025	0.021	–	0.018
	Coverage	0.950	0.953	–	0.954
$k = 2$	Length	$0.275^{(MC)}$	0.254	–	0.256
	Shape	$1.155^{(MC)}$	1.000	–	1.054
	LRP	0.025	0.026	–	0.033
	RRP	0.025	0.028	–	0.027
	Coverage	0.950	0.946	–	0.940

shape is the difference between the upper bound and the point estimate, divided by the difference between the point estimate and the lower bound. The *coverage* is the difference between the normal cumulative distribution function, evaluated at the upper bound, and that evaluated at the lower bound. The *left rejection probability* (LRP) is the normal cumulative distribution function evaluated at the lower bound. The *right rejection probability* (RRP) is one minus the normal cumulative distribution function evaluated at the upper bound. All these measures are averaged over the M Monte Carlo replicates. The difference between the $100(1-\alpha/2)$th and the $100\alpha/2$th percentiles of the $N(\mu_{w_k}, \sigma^2_{w_k})$, for $k = 1, 2$ is the benchmark for the length, representing a lower limit for the length of the prediction interval, including variability only due to heterogeneity. The known theoretical values for the shape, LRP, RRP and coverage represent the other benchmarks. Table 2 shows, for $N = 150$, that both D_{SMG} and KR provide prediction intervals with correct right/left rejection probabilities and global coverage. Prediction intervals are larger than the benchmark lower limit. D_{BR}, instead, produces too short prediction intervals with a coverage significantly smaller compared to the theoretical confidence level. This indicates that the WTP sampling distribution is characterized by heavier tails compared to the normal distribution D_{BR} uses to build prediction intervals. In particular, the WTP prediction interval for $k = 2$ performs worse than that for $k = 1$, due to the smaller heterogeneity of β_2. Thus, the sampling error is comparatively more relevant in determining the prediction standard error, making D_{BR} estimates questionable (Fig. 1, left panel). When N increases to 300, the coverage rate improves since the sampling error decreases.

D_{SMG} and KR can produce confidence intervals for the average WTP, while this is not possible for D_{BR} (Table 3). We use $\bar{w}_k$ Monte Carlo replicates to calculate benchmark values for WTP confidence intervals' length and shape. We compute LRP, RRP and

Table 4
Fixed divided by log-normal case: mean and standard deviation (in brackets) over Monte Carlo replicates for the estimated average WTP, its standard error and prediction standard error. The superscript (MC) denotes Monte Carlo estimates of benchmark values.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	$\bar{w}_1$	4.482	4.600 (1.214)	4.600 (1.214)	4.600 (1.214)
	$\hat{se}(\bar{w}_1)$	$1.214^{(MC)}$	1.141 (0.571)	–	1.240 (0.694)
	$\widehat{pse}(\bar{w}_1)$	$5.999^{(MC)}$	7.120 (3.999)	4.883 (2.014)	8.138 (6.055)
$k = 2$	$\bar{w}_2$	2.241	2.294 (0.655)	2.294 (0.655)	2.294 (0.655)
	$\hat{se}(\bar{w}_2)$	$0.655^{(MC)}$	0.606 (0.298)	–	0.657 (0.361)
	$\widehat{pse}(\bar{w}_2)$	$3.009^{(MC)}$	3.576 (2.086)	2.447 (1.054)	4.088 (3.109)
$N = 300$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	$\bar{w}_1$	4.482	4.627 (0.877)	4.627 (0.877)	4.627 (0.877)
	$\hat{se}(\bar{w}_1)$	$0.877^{(MC)}$	0.802 (0.269)	–	0.834 (0.287)
	$\hat{pse}(\bar{w}_1)$	$5.940^{(MC)}$	6.713 (2.432)	4.820 (1.353)	7.050 (2.711)
$k = 2$	$\bar{w}_2$	2.241	2.307 (0.458)	2.307 (0.458)	2.307 (0.458)
	$\hat{se}(\bar{w}_2)$	$0.458^{(MC)}$	0.425 (0.137)	–	0.441 (0.147)
	$\hat{pse}(\bar{w}_2)$	$2.973^{(MC)}$	3.355 (1.230)	2.407 (0.692)	3.524 (1.376)
$N = 450$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	$\bar{w}_1$	4.482	4.549 (0.669)	4.549 (0.669)	4.549 (0.669)
	$\hat{se}(\bar{w}_1)$	$0.669^{(MC)}$	0.634 (0.169)	–	0.647 (0.175)
	$\hat{pse}(\bar{w}_1)$	$5.913^{(MC)}$	6.414 (2.281)	4.683 (1.037)	6.583 (1.961)
$k = 2$	$\bar{w}_2$	2.241	2.279 (0.357)	2.279 (0.357)	2.279 (0.357)
	$\hat{se}(\bar{w}_2)$	$0.357^{(MC)}$	0.338 (0.087)	–	0.344 (0.090)
	$\hat{pse}(\bar{w}_2)$	$2.959^{(MC)}$	3.218 (1.131)	2.349 (0.538)	3.304 (1.004)

coverage for the intervals D_{SMG} and KR produce as the relative number of times the known μ_{w_k} falls below the lower bound, above the upper one, or within the two, respectively. Both methods yield correct coverage and left/right rejection probabilities. Notice that KR does not provide necessarily symmetric confidence intervals, while D_{SMG} always does by construction. This, in principle, undesirable characteristic only pertains to small samples or to particular choices of the mixing distribution not assuring a quick convergence (e.g. in the fixed divided by log-normal case). Symmetric confidence intervals are not a problem for large samples, where the central limit theorem ensures asymptotic normality.

As a final remark, we underline that the results obtained in this section, for the normal divided by fixed case, are also expected to hold for models estimated in WTP space, when assuming a normally distributed WTP parameter.

5.2. The fixed divided by log-normal case

When assuming the parameter at the numerator to be fixed and that at the denominator to be distributed log-normal, the distribution of the WTP for X_k is known exactly, as well as its moments: $w_k \sim \text{sgn}(\beta_k) LN(\mu_{w_k}, \sigma^2_{w_k})$, with $\mu_{w_k} = -\mu_c + \log(|\beta_k|)$ and $\sigma_{w_k} = \sigma_c$. We set $\beta_0 = 0.5$, $\beta_1 = 1$, $\beta_2 = 0.5$ and $-\beta_c \sim LN(\mu_c, \sigma^2_c)$ with $\mu_c = -1$ and $\sigma^2_c = 1$, so that $w_1 \sim LN(1, 1)$ and $w_2 \sim LN(0.307, 1)$. The expected WTP value is $E(w_k) = \exp(-\mu_k + \sigma^2_{w_k}/2)$, thus $E(w_1) = 4.482$ and $E(w_2) = 2.241$, and the variance is $\text{Var}(w_k) = (\exp(\sigma^2_c) - 1)\exp(2(-\mu_c + \log(|\beta_k|)) + \sigma^2_c)$, thus $\text{Var}(w_1) = 34.513$ and $\text{Var}(w_2) = 8.628$.

Table 4, analogously to Table 1, reports the mean of different estimated quantities over Monte Carlo replicates for the three methods.[10] The estimated WTP average is $\exp(-\hat{\mu}_k + \hat{\sigma}^2_c/2)$ for all three methods. Both D_{SMG} and KR provide standard error estimates, $\hat{se}(\bar{w}_k)$, very close to the benchmark value, but show a large variability across Monte Carlo replicates.[11]

Table 4 shows that D_{BR} systematically underestimates the prediction standard error. For $N = 150$, $\widehat{pse}(\hat{w}_k)$ averages are equal to 4.883 and 2.447, respectively for $k = 1$ and $k = 2$, which are much smaller than the benchmarks. Furthermore, these values are smaller than the known w_k standard deviations. These represent the variability component ascribable to heterogeneity alone (i.e. $\sqrt{\text{Var}(w_1)} = 5.874$ and $\sqrt{\text{Var}(w_2)} = 2.937$). To clarify, we compute, for each replicate, the estimated value of the standard

[10] For the fixed divided by log-normal case, Bliemer and Rose (2013) suggest, as a possible alternative, replacing the mean of both the conditional $\hat{w}_k(z)$ and the conditional standard errors with the median. They warn that dividing by the log-normal distribution might imply rather large WTPs, producing a large value for the mean of the conditional standard errors. The median of the conditional standard errors is expected to be less affected by extreme values. However, following this suggestion would introduce a larger bias in the prediction standard error, deteriorating also the prediction intervals' coverage. This section only reports results without the correction. They show a better performance. Complete results are available from the authors upon request.

[11] In fact, $\bar{w}_k$ is affected by large (absolute) WTP values arising from the log-normal distribution that can be more/less extreme depending on the parameter estimates obtained at each Monte Carlo replicate.

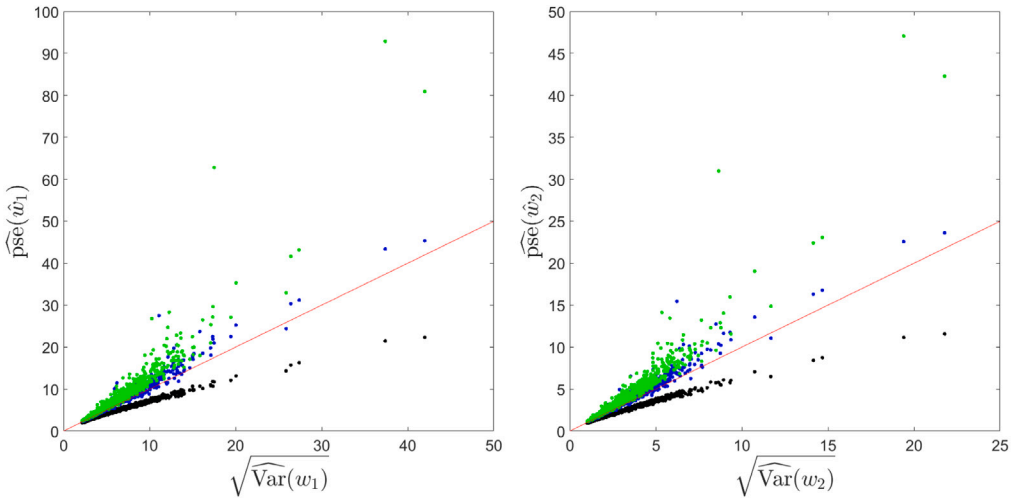

Fig. 2. Estimates of the predictive standard error $\widehat{\text{pse}}(\hat{w}_k)$ at each Monte Carlo replicate plotted against estimates of standard deviation $\sqrt{\widehat{\text{Var}}(w_k)}$ for the same replicate. The comparison is performed for D_{BR} (black points), D_{SMG} (blue points) and KR (green points). The $\sqrt{\widehat{\text{Var}}(w_k)}$ values are the same for the three methods. Estimates are computed for $N = 150$. The red line represents the bisector of the first quadrant. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

deviation as $\sqrt{\widehat{\text{Var}}(w_k)} = \sqrt{(\exp(\hat{\sigma}_c^2) - 1)\exp(2(-\hat{\mu}_c + \log(|\hat{\beta}_k|)) + \hat{\sigma}_c^2)}$ so to compare it with the estimated prediction standard error $\widehat{\text{pse}}(\hat{w}_k)$ for the same replicate. One would expect $\widehat{\text{pse}}(\hat{w}_k) > \sqrt{\widehat{\text{Var}}(w_k)}$, for each replicate, since $\sqrt{\widehat{\text{Var}}(w_k)}$ accounts for variability due to heterogeneity alone, while $\widehat{\text{pse}}(\hat{w}_k)$ incorporates the sampling error too. However, when using D_{BR} this is never true[12] for $k = 1, 2$. On the contrary, this is true in 99.6% ($k = 1, 2$) of times for KR, and in 93.1% ($k = 1$) and 93.9% ($k = 2$) of times for D_{SMG}. Fig. 2 depicts this comparison. D_{SMG} assumes $\hat{w}_k(z)$ asymptotic normality. This more stringent assumption explains its lower performance compared to KR. Table 4 reports a $\widehat{\text{pse}}(\hat{w}_k)$ mean over replicates sensibly larger than the benchmark for both D_{SMG} and KR. This is mainly due to some very large $\widehat{\text{pse}}(\hat{w}_k)$ values. Standard deviation, in brackets, and points on the top right corner of the scatter plots in Fig. 2 are evidence of this. The average $\widehat{\text{pse}}(\hat{w}_k)$ values D_{SMG} and KR produce get progressively closer to the benchmark when the sample size increases. The opposite is true for D_{BR}. In fact, the average shifts further away from the benchmark, and increasingly below the lower limits (i.e., $\sqrt{\text{Var}(w_1)} = 5.874$ and $\sqrt{\text{Var}(w_2)} = 2.937$).

Table 5 shows prediction intervals for the three methods. We calculate coverage, LRP and RRP using the known log-normal cumulative distribution function for w_1 and w_2. Analogously to the normal divided by fixed case, we calculate the benchmark for the interval length as the difference between the $100(1-\alpha/2)$th and the $100\alpha/2$th percentiles of the $\text{sgn}(\beta_k)LN(\hat{\mu}_{w_k}, \hat{\sigma}_{w_k})$ distribution for $k = 1, 2$. This represents a lower limit of the length of the prediction interval, accounting only for the variability due to heterogeneity. The computation for the interval shape benchmark uses the same percentiles. For $N = 150$, both D_{SMG} and KR produce prediction intervals with correct LRP, RRP, and global coverage. They are larger than the length lower limit. D_{BR}, instead, has a significantly smaller than 95% global coverage. While this problem seems to vanish as the sample size increases, LRP and RRP are still significantly far from 2.5% even when $N = 450$. Larger sample sizes will not improve the situation. In fact, the prediction interval is calculated on the basis of a normal distribution, while the $\hat{w}_k$ sampling distribution is a mixture of normals, which tends to the distribution of w_k (a highly skewed log-normal distribution) as the sample size increases and the sampling error becomes progressively less relevant. Therefore, even asymptotically, the correct LRP and RRP cannot be achieved by a symmetric prediction interval, as the one D_{BR} produces. In particular, the lower D_{BR} prediction interval bound always turns out negative in the M simulations,[13] for any sample size N. D_{SMG} and KR, instead, always correctly produce positive lower prediction intervals for all M replicates and sample sizes.

Table 6 summarizes D_{SMG} and KR performances when computing confidence intervals for $E(w_k)$. Both methods show unsatisfactory LRP and RRP for $N = 150$, with KR performing slightly better and achieving, at least, a reasonable global coverage. As N increases, KR improves faster than D_{SMG}, which still shows LRP and RRP significantly different from 2.5% even for $N = 450$. The symmetric confidence intervals produced by D_{SMG} are not realistic in the fixed divided by log-normal case, unless using very large sample sizes. We show in Table 7 the results of the skewness test for the composite hypothesis of normality (Shapiro et al., 1968), with respect to the estimators of the parameters and $\bar{w}_k$. When $N = 150$, the sampling distribution of the fixed parameters β_1 and β_2 estimator does not significantly diverge from normality at the 5% level. The estimators of the structural parameters μ_c and σ_c require, instead, a larger sample size to favour convergence. The $\bar{w}_k$ distribution is even more skewed. Skewness decreases as N

[12] This does not change for $N = 300, 450$.

[13] The lower bound of D_{BR} prediction intervals ranges from -29.332 up to -1.068 for $k = 1$ and from -15.250 up to -0.496 for $k = 2$, when $N = 150$.

Table 5
Fixed divided by log-normal case: length, shape, LRP, RRP and coverage of 95% level prediction intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. The benchmark for the length of the interval is in italic as it represents a lower limit for the length, only accounting for variability due to heterogeneity.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	*18.914*	20.459	19.141	21.440
	Shape	3.615	3.720	1.000	3.941
	LRP	0.025	0.025	0.000***	0.027
	RRP	0.025	0.031	0.064***	0.029
	Coverage	0.950	0.944	0.936*	0.944
$k = 2$	Length	*9.457*	10.258	9.590	10.740
	Shape	3.615	3.736	1.000	3.953
	LRP	0.025	0.024	0.000***	0.026
	RRP	0.025	0.032	0.066***	0.030
	Coverage	0.950	0.944	0.934*	0.944
$N = 300$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	*18.914*	20.223	18.894	20.786
	Shape	3.615	3.695	1.000	3.824
	LRP	0.025	0.025	0.000***	0.025
	RRP	0.025	0.027	0.058***	0.026
	Coverage	0.950	0.948	0.942	0.949
$k = 2$	Length	*9.457*	10.104	9.437	10.382
	Shape	3.615	3.703	1.000	3.831
	LRP	0.025	0.024	0.000***	0.025
	RRP	0.025	0.027	0.059***	0.026
	Coverage	0.950	0.949	0.941	0.949
$N = 450$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	*18.914*	19.678	18.356	19.995
	Shape	3.615	3.676	1.000	3.752
	LRP	0.025	0.024	0.000***	0.025
	RRP	0.025	0.026	0.058***	0.026
	Coverage	0.950	0.950	0.942	0.949
$k = 2$	Length	*9.457*	9.873	9.207	10.027
	Shape	3.615	3.681	1.000	3.756
	LRP	0.025	0.024	0.000***	0.024
	RRP	0.025	0.026	0.058***	0.026
	Coverage	0.950	0.950	0.942	0.950

increases, even if $N = 450$ is far too small to achieve symmetry. KR only assumes the normality of the estimators of the structural parameters, thus it is not surprising it outperforms D_{SMG} in this case.

As for the normal divided by fixed case, we stress that the conclusions reached for the fixed divided by log-normal case should also be valid for models estimated in WTP space, when assuming a log-normally distributed WTP parameter. In addition, also the log-normal divided by log-normal case, which is illustrated in Appendix C, produces results that resemble very closely those of the fixed divided by log-normal case, as both cases determine a log-normal distribution of the WTP.

5.3. The normal divided by normal case

For the normal divided by normal case, we set $\beta_0 = 0.5$ and $\beta_k \sim N(\mu_k, \sigma_k^2)$, for $k = 1, 2$, with $\mu_1 = 1$, $\mu_2 = 0.5$, $\sigma_1 = 0.5$ and $\sigma_2 = 0.4$, and assume $\beta_c \sim N(\mu_c, \sigma_c^2)$, with $\mu_c = -1$ and $\sigma_c = 0.5$ for the cost parameter. For simplicity, we also assume β_k to be independent from β_c. One can deal with the dependently distributed case by appropriately rewriting random parameters using standard distributions (see equation (23), p. 206, in Bliemer and Rose, 2013). When assuming both the numerator and denominator parameters are distributed normal, the WTP distribution is complex. This is also true when β_k and β_c are independent. In particular, Daly et al. (2012b) show that the moments of the WTP distribution do not exist, when assuming the cost parameter distribution is normal. It is, thus, meaningless to estimate $E(w_k)$ or its confidence interval, while perfectly reasonable to calculate a prediction interval for w_k. Bliemer and Rose (2013) suggest using the median to replace the mean of the conditional $\hat{w}_k(z)$ in (12), and the mean of their conditional standard errors in (13). We denote this adjusted method as $D_{BR^\star}$.

D_{SMG} does not require any adjustment to estimate prediction intervals. We also use the median as a measure of central tendency for WTP and estimate it as the median of $\hat{w}_k(z)$. The various conditional $\hat{w}_k(z)$, whose distribution is asymptotically normal, determine the $\hat{w}_k$ mixture distribution as in Eq. (20), by averaging $\hat{w}_k(z)$ distributions over z. Subsequently, we compute

Table 6
Fixed divided by log-normal case: length, shape, LRP, RRP and coverage of 95% level confidence intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. The superscript (MC) denotes Monte Carlo estimates of benchmark values.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	$4.278^{(MC)}$	4.473	–	4.788
	Shape	$1.854^{(MC)}$	1.000	–	1.769
	LRP	0.025	0.000^{***}	–	0.007^{***}
	RRP	0.025	0.087^{***}	–	0.051^{***}
	Coverage	0.950	0.913^{***}	–	0.942
$k = 2$	Length	$2.312^{(MC)}$	2.374	–	2.541
	Shape	$1.851^{(MC)}$	1.000	–	1.751
	LRP	0.025	0.000^{***}	–	0.008^{***}
	RRP	0.025	0.099^{***}	–	0.037^{*}
	Coverage	0.950	0.901^{***}	–	0.955
$N = 300$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	$3.271^{(MC)}$	3.145	–	3.246
	Shape	$1.643^{(MC)}$	1.000	–	1.497
	LRP	0.025	0.000^{***}	–	0.019
	RRP	0.025	0.074^{***}	–	0.045^{***}
	Coverage	0.950	0.926^{***}	–	0.936^{*}
$k = 2$	Length	$1.714^{(MC)}$	1.667	–	1.718
	Shape	$1.690^{(MC)}$	1.000	–	1.488
	LRP	0.025	0.001^{***}	–	0.022
	RRP	0.025	0.075^{***}	–	0.044^{***}
	Coverage	0.950	0.924^{***}	–	0.934^{*}
$N = 450$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	$2.518^{(MC)}$	2.486	–	2.524
	Shape	$1.768^{(MC)}$	1.000	–	1.385
	LRP	0.025	0.003^{***}	–	0.018
	RRP	0.025	0.046^{***}	–	0.025
	Coverage	0.950	0.951	–	0.957
$k = 2$	Length	$2.518^{(MC)}$	2.486	–	2.524
	Shape	$1.768^{(MC)}$	1.000	–	1.385
	LRP	0.025	0.003^{***}	–	0.018
	RRP	0.025	0.046^{***}	–	0.025
	Coverage	0.950	0.951	–	0.957

Table 7
Fixed divided by log-normal case: skewness test for the composite hypothesis of normality.

Parameter	$N = 150$		$N = 300$		$N = 450$	
estimator	Skewness	p-value	Skewness	p-value	Skewness	p-value
$\hat{\beta}_1$	−0.020	0.802	0.034	0.664	−0.001	0.992
$\hat{\beta}_2$	0.040	0.611	0.017	0.804	0.088	0.252
$\hat{\mu}_c$	−0.287	0.001	−0.364	0.000	−0.083	0.280
$\hat{\sigma}_c$	0.219	0.006	0.136	0.082	0.116	0.132
$\iota\tilde{w}_1$	1.825	0.000	1.007	0.000	0.814	0.000
$\iota\tilde{w}_2$	1.808	0.000	1.048	0.000	0.929	0.000

Table 8
Normal divided by normal case: mean and standard deviation (in brackets) over Monte Carlo replicates for the estimated median WTP. The superscript (MC) denotes Monte Carlo estimates of benchmark values.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR^*}	KR
$k = 1$	$\tilde{w}_1$	$0.961^{(MC)}$	0.962 (0.092)	0.962 (0.092)	0.961 (0.090)
$k = 2$	$\tilde{w}_2$	$0.478^{(MC)}$	0.476 (0.069)	0.476 (0.069)	0.475 (0.069)

the empirical percentiles of this mixture distribution to produce the prediction interval. We avoid estimating a prediction standard error that, in fact, does not exist.

Table 9
Normal divided by normal case: length, shape, LRP, RRP and coverage of 95% level prediction intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. The superscript (MC) denotes Monte Carlo estimates of benchmark values. The benchmark for the length of the interval is in italic as it represents a lower limit for the length, only accounting for variability due to heterogeneity.

$N = 150$					
		Benchmark	D_{SMG}	$D_{BR^\star}$	KR
$k = 1$	Length	*6.910*$^{(MC)}$	7.499	2.789	7.847
	Shape	3.813$^{(MC)}$	2.800	1.000	2.767
	LRP	0.025	0.028	0.027	0.028
	RRP	0.025	0.030	0.123***	0.029
	Coverage	0.950	0.942	0.850***	0.943
$k = 2$	Length	*4.333*$^{(MC)}$	4.556	1.944	4.681
	Shape	2.495$^{(MC)}$	2.098	1.000	2.067
	LRP	0.025	0.027	0.037*	0.027
	RRP	0.025	0.029	0.110***	0.028
	Coverage	0.950	0.944	0.853***	0.945
$N = 300$					
		Benchmark	D_{SMG}	$D_{BR^\star}$	KR
$k = 1$	Length	*6.910*$^{(MC)}$	7.220	2.777	7.731
	Shape	3.813$^{(MC)}$	2.960	1.000	3.030
	LRP	0.025	0.027	0.026	0.027
	RRP	0.025	0.026	0.119***	0.024
	Coverage	0.950	0.947	0.855***	0.949
$k = 2$	Length	*4.333*$^{(MC)}$	4.345	1.901	4.576
	Shape	2.495$^{(MC)}$	2.239	1.000	2.248
	LRP	0.025	0.026	0.034	0.025
	RRP	0.025	0.029	0.110***	0.027
	Coverage	0.950	0.945	0.856***	0.948
$N = 450$					
		Benchmark	D_{SMG}	$D_{BR^\star}$	KR
$k = 1$	Length	*6.910*$^{(MC)}$	6.983	2.770	7.504
	Shape	3.813$^{(MC)}$	3.067	1.000	3.185
	LRP	0.025	0.024	0.025	0.024
	RRP	0.025	0.030	0.127***	0.028
	Coverage	0.950	0.946	0.848***	0.948
$k = 2$	Length	*4.333*$^{(MC)}$	4.242	1.887	4.473
	Shape	2.495$^{(MC)}$	2.301	1.000	2.351
	LRP	0.025	0.026	0.035*	0.025
	RRP	0.025	0.029	0.111***	0.027
	Coverage	0.950	0.945	0.854***	0.948

Table 8 compares D_{SMG}, $D_{BR^\star}$ and KR performance in estimating the WTP median. We do not provide estimates of standard errors and prediction standard errors since they do not exist. We use the estimated median $\tilde{w}_k$ as a point predictor of WTP. All three methods produce unbiased estimates. Both D_{SMG} and $D_{BR^\star}$ calculate $\tilde{w}_k$ in exactly the same way. We draw a very large number of values for β_k from a $N(\mu_k, \sigma_k^2)$ and for β_c from a $N(\mu_c, \sigma_c^2)$ to obtain the benchmark for $\tilde{w}_k$ since the w_k distribution is non conventional. We calculate w_k for each couple of values and finally take the median of all the simulated w_k.

Table 9 compares the performance of the three methods in estimating prediction intervals. We use the simulated w_k and compute the percentage of these values falling within each replicate of the prediction interval, below the lower bound and above its upper one, to determine the coverage rate of the prediction intervals. Averaging these percentages over the M replicates of the prediction intervals allows calculating global coverage, LRP and RRP. Substituting the point estimate of $E(w_k)$ with the point estimate of the median of w_k allows calculating the shape of the prediction intervals. The empirical distribution of w_k permits obtaining the shape benchmark. A symmetric w_k distribution would have a shape index close to one. The difference between the $100(1-\alpha/2)$th and the $100\alpha/2$th percentiles of the empirical distribution of w_k is the benchmark for the prediction interval length, representing a lower limit since it accounts only for the variability due to heterogeneity.

Table 9 suggests that the adjustment needed so to apply the D_{BR} method produces too small and left shifted prediction intervals. One should note that, without this adjustment, the intervals would be unreasonable (median length of 1300.493 and 643.255, and coverage of 0.998 and 0.998, respectively for w_1 and w_2, for $N = 150$). Increasing sample size does not improve this. Both D_{SMG} and KR produce prediction intervals with the expected LRP, RRP and global coverage, also for $N = 150$.

6. Real data application

This section investigates the empirical relevance of the adoption of the three different methods when using the real data as in the example in Bliemer and Rose (2013, sec. 5). We report the results obtained for each method under the three different mixing

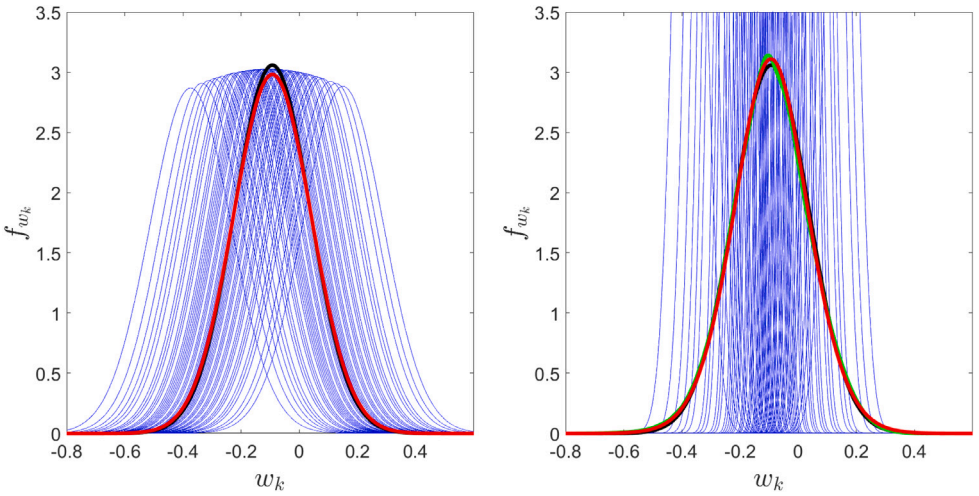

Fig. 3. Left panel: distributions of the conditional $\hat{w}_k(z)$ (blue lines) and the unconditional $\bar{w}_k$ based on the mean and mean standard error of $\hat{w}_k(z)$ (red line), according to the Bliemer and Rose (2013) approach to the Delta method. Right panel: distributions of the conditional $\hat{w}_k(z)$ (blue lines) and of the unconditional $\hat{w}_k$ (red line) according to our approach to the Delta method, and simulated sampling distribution of $\hat{w}_k$ according to the KR method (green line). Both panels: estimated distribution of w_k (black line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distributions in Section 5. The data set refers to a route choice experiment, where agents choose between the status quo route, and two hypothetical ones. The attributes describing routes characteristics are: time spent in free-flow/congested conditions on non-tolled road, and free-flow/congested conditions on toll road, toll/petrol costs, and number of traffic lights. The data set collected in 2011 includes 148 respondents, answering 12 choice tasks each, and generating 1776 observations.

6.1. The normal divided by fixed case

Recall that here (see Section 5.1) $\beta_k = \mu_k + \sigma_k z_k$, β_c is fixed and $w_k \sim N(-\mu_k/\beta_c, \sigma_k^2/\beta_c^2)$. Bliemer and Rose (2013) obtain the following estimates: $\hat{\mu}_k = -0.047$, $\hat{\sigma}_k = 0.066$, $\hat{\beta}_c = -0.506$ and

$$\hat{\Omega}_{\hat{\mu}_k,\hat{\sigma}_k,\hat{\beta}_c} = \begin{pmatrix} 0.00010 & 0.00000 & 0.00005 \\ 0.00000 & 0.00014 & -0.00011 \\ 0.00005 & -0.00011 & 0.00043 \end{pmatrix}.$$

Thus, the estimated distribution of w_k, without accounting for sampling error, is:

$$\hat{f}_{w_k}(x) = \phi(x; -0.0929, 0.0170). \tag{25}$$

The left panel in Fig. 3 reproduces Figure 2 in Bliemer and Rose (2013).[14] Blue lines represent the normal distributions of the conditional $\hat{w}_k(z^{(r)})$, calculated accordingly to Bliemer and Rose (2013) and given in (10), with means and variances estimated using $\hat{\mu}_k$, $\hat{\sigma}_k$ and $\hat{\beta}_c$, and corresponding to the first 50 Halton draws of z, i.e. for $r = 1, \ldots, 50$. The red line represents the distribution of the unconditional $\bar{w}_k$, a normal distribution with mean -0.0929 and standard deviation 0.1338, calculated as in (12) and (13). We underline that this red line does not provide the sampling distribution of $\hat{w}_k$, as Section 6.2 will show, but is instrumental to build a normal theory based prediction interval for w_k, which, in this case, is equal to $(-0.3571, 0.1741)$ (see Bliemer and Rose (2013, sec. 5.2)). The black line represents the estimated distribution of w_k given in (25), and accounts only for heterogeneity. The red distribution is, correctly, more dispersed than the black one and the prediction standard error is reasonably estimated. Section 5.1 discusses and Appendix B demonstrates that if the standard error is negligible compared to the variability due to heterogeneity, as in this case, D_{BR} provides plausible results.

The blue lines, in the right panel of Fig. 3, represent the normal distributions of the conditional $\hat{w}_k(z^{(r)})$. These are calculated, using the first 50 Halton draws of z, via D_{SMG} as in (14), with means and variances estimated using $\hat{\mu}_k$, $\hat{\sigma}_k$ and $\hat{\beta}_c$. The variances of these blue distributions only depend on sampling error since they are conditional on z. This is why the blue distributions in the right panel are much less dispersed than those in the left one, which encompass also variability due to heterogeneity (please note that the 50 distributions have the same mean in the two panels). The red line represents the estimate of the mixture distribution in (16), approximated as in (20) using $R = 10{,}000$ Halton draws for z. The green line represents, instead, the simulated sampling

[14] Apart from the fact of being reflected on the y-axes because of the different definition used for the WTP, and from the addition of the estimated distribution of w_k (black line).

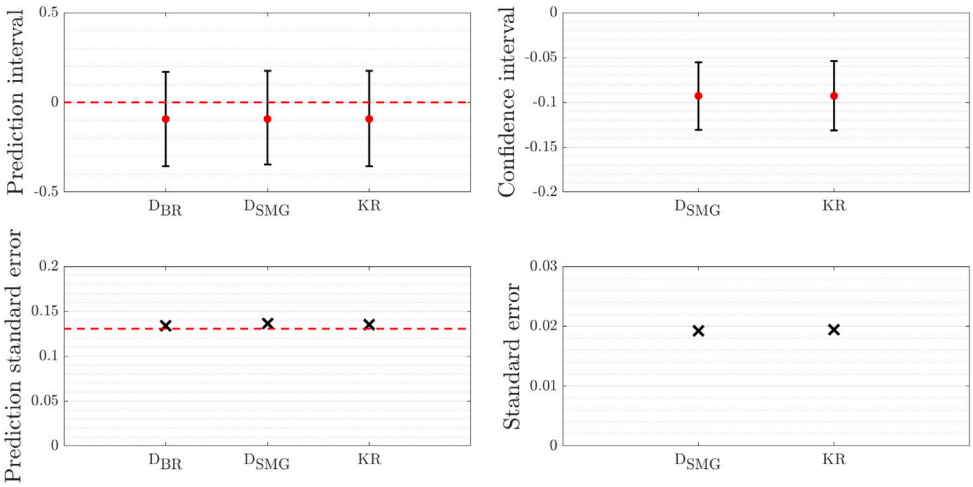

Fig. 4. Comparison between the performance of D_{BR}, D_{SMG} and KR in estimating prediction interval, confidence interval, prediction standard error and standard error on real data from Bliemer and Rose (2013), for the normal divided by fixed case.

distribution of $\hat{w}_k$ obtained via KR, and based on $B = 2,000$ first stage draws from the multivariate normal distribution of $(\hat{\mu}_k, \hat{\sigma}_k, \hat{\mu}_c)$, and $R = 10{,}000$ second stage draws for z. The black distribution is the same in the two panels and its variability is only due to taste heterogeneity. Note the similarity between red and green distributions as well as their close resemblance to the black one. They, in fact, are characterized by a slightly larger dispersion (i.e. slightly higher tails) due to sampling error. In practice, the variance of the black distribution arises from the second integral in (17), while that of the red line is the sum of the two integrals, so that the difference in the variances of the two distributions is equal to the first integral, i.e. is due to sampling error. The closeness of the red and green distributions to the black one depends on the occurrence of very precise parameter estimates for the data considered. Less informative data might well result in a larger difference between the black distribution and the other two. Finally, notice how the red and green distributions, in the right panel, even if not normal[15] are not too different from the normal red distribution in the left panel, testifying that the three methods perform very similarly for the data at hand and under the current specification of the mixing distributions.

Fig. 4, top-left panel, shows the similarity between the prediction intervals obtained through the three methods. All include the value zero, as underlined by the dashed red line, due to the large heterogeneity of β_k. The prediction standard errors estimated by the three methods (Fig. 4, bottom-left panel) are also very close to each other with D_{BR} providing only slightly smaller values compared to D_{SMG} and KR. All are correctly larger than the variability due to heterogeneity alone (the dashed red line corresponding to the value $\sqrt{\widehat{\mathrm{Var}}(w_k)} = \sqrt{\hat{\sigma}_k^2/\hat{\beta}_c^2} = 0.1304$). Finally, D_{SMG} and KR also perform very similarly in estimating confidence intervals (Fig. 4, top-right panel) and standard errors (Fig. 4, bottom-right panel) for $\hat{w}_k$.

6.2. The fixed divided by log-normal case

Consider the case of a fixed β_k and a log-normal distribution for the negative of β_c (i.e. $-\beta_c = \exp(\mu_c + \sigma_c z_c)$). The log-normal always produces negative values for the cost parameter, as discussed in Section 5.2. Thus, w_k can only take positive values if β_k is positive or negative values if β_k is negative. In particular,

$$w_k = \mathrm{sgn}(\beta_k)\exp(-\mu_c + \log(|\beta_k|) + \sigma_c z_c), \tag{26}$$

that is, w_k has a positive or negative (according to the sign of β_k) log-normal distribution with parameters $-\mu_c + \log(|\beta_k|)$ and σ_c^2. Bliemer and Rose (2013) obtain the estimates $\hat{\beta}_k = -0.035$, $\hat{\mu}_c = -0.994$ and $\hat{\sigma}_c = 1.223$, and the variance–covariance matrix

$$\hat{\Omega}_{\hat{\beta}_k,\hat{\mu}_c,\hat{\sigma}_c} = \begin{pmatrix} 0.00001 & -0.00002 & 0.00001 \\ -0.00002 & 0.01985 & -0.00652 \\ 0.00001 & -0.00652 & 0.00215 \end{pmatrix}.$$

The left panel in Fig. 5 reproduces Figure 3 in Bliemer and Rose (2013).[16] Blue lines represent the normal distributions of the conditional $\hat{w}_k(z^{(r)})$, calculated accordingly to Bliemer and Rose (2013) and given in (10), with means and variances estimated using $\hat{\beta}_k$, $\hat{\mu}_c$ and $\hat{\sigma}_c$, and corresponding to the first 50 Halton draws of z, i.e. for $r = 1, \ldots, 50$. Bliemer and Rose (2013) claim that

[15] They are leptokurtic with respect to the black normal distribution.

[16] See footnote 14.

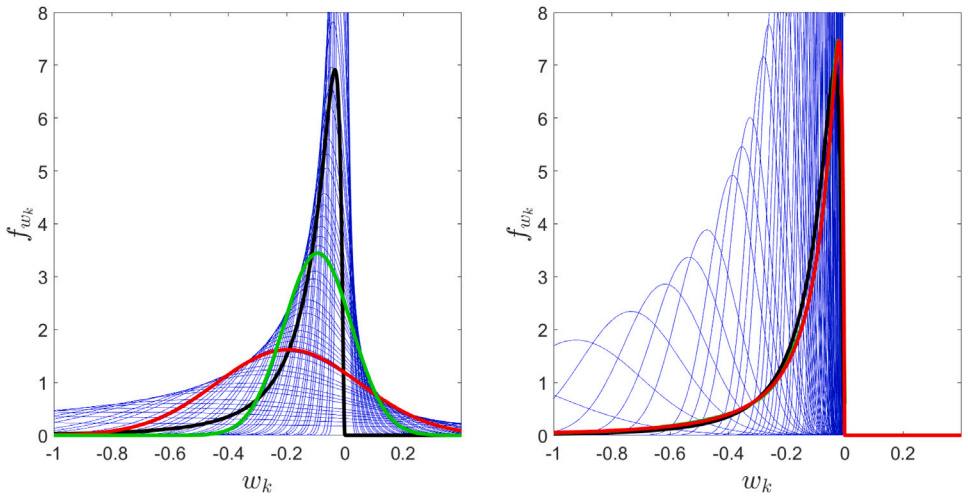

Fig. 5. Left panel: distributions of the conditional $\hat{w}_k(z)$ (blue lines), of the unconditional $\bar{w}_k$ based on the mean and mean standard error of $\hat{w}_k(z)$ (red line), of the unconditional $\bar{w}_k$ based on the median and median standard error of $\hat{w}_k(z)$ (green line), according to the Bliemer and Rose (2013) approach to the Delta method. Right panel: distributions of the conditional $\hat{w}_k(z)$ (blue lines) and of the unconditional $\hat{w}_k$ (red line) according to our approach to the Delta method, and simulated sampling distribution of $\hat{w}_k$ according to the KR method (green line, hardly distinguishable from red line). Both panels: estimated distribution of w_k (black line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the variance of these distributions encompasses both sampling error and heterogeneity. The red line represents the distribution of the unconditional $\bar{w}_k$, a normal distribution with mean -0.1959 and standard deviation 0.2463, calculated as in (12) and (13). The green line represents an alternative distribution obtained by using the median (equal to -0.0941) instead of the mean and the median of the conditional standard errors (equal to 0.1158). Bliemer and Rose (2013) suggest using this second distribution to compute prediction intervals. Since dividing by the log-normal distribution results in some cases in rather large WTP values, the mean standard error is large. Using the median reduces the impact of these extreme values. We then add the black line, representing the distribution of w_k in (26), with estimated parameters in place of real ones, that is the distribution of w_k assuming parameters were correctly estimated. As such, the black distribution accounts for variability due to heterogeneity, while the red and green distributions should, in some way, incorporate the variability also due to sampling error. The black distribution only takes negative values ($\hat{\beta}_k$ being negative), while the red and green also take positive ones due to sampling error in $\hat{\beta}_k$ estimates. However, the proportion of either red or green distribution taking positive values is definitively too large. We underline that positive $\hat{w}_k$ values might only arise due to positive $\hat{\beta}_k$ values ($\hat{\beta}_c$ always being negative). In our case, $\hat{\beta}_k$ is negative, with a t-value of -10.77 and a p-value substantially equal to 0. Thus, β_k is significantly different from 0 as well as from any positive value, implying that also w_k is significantly different from 0 as well as from any positive value, $\forall z$. As a consequence, the prediction interval for w_k should not include 0 nor any positive value. On the contrary, the percentage of positive values included in the prediction intervals based on both red and green distributions is approximately 30%. Fig. 6, top-right panel, depicts these intervals which are respectively equal to $(-0.6786, 0.2868)$ and $(-0.3210, 0.1328)$.

The distribution of w_k, estimated on the basis of the sample (i.e. the black line in left panel of Fig. 5), has mean equal to $-\exp(-\hat{\mu}_c + \log(|\hat{\beta}_k|) + \hat{\sigma}_c^2/2) = -0.1998$, which is very close to the mean of the unconditional $\bar{w}_k$, -0.1959, obtained by Bliemer and Rose (2013), and standard deviation equal to $\sqrt{(\exp(\hat{\sigma}_c^2) - 1)\exp(2(-\hat{\mu}_c + \log(|\hat{\beta}_k|)) + \hat{\sigma}_c^2)} = 0.3717$. This value, identified by the red dashed line in the bottom-left panel of Fig. 6, accounts for the (estimated) variability due to heterogeneity alone. The estimated prediction standard error of 0.2463, provided in Bliemer and Rose (2013), incorporating variability due to heterogeneity and sampling error, is, therefore, far too small. In summary, the formula in (13) fails to estimate the prediction standard error. Using the median of the conditional standard errors produces an even more biased estimate of 0.1158. Both estimates are represented in the bottom-left panel of Fig. 6 and fall well below the red-dashed line.

The blue lines in the right panel of Fig. 5 represent the normal distributions of the conditional $\hat{w}_k(z^{(r)})$, calculated according to $\mathrm{D}_{\mathrm{SMG}}$ and corresponding to the first 50 Halton draws of z. The variance of these blue distributions only depends on sampling error, thus they are much less dispersed in comparison to those in the left panel. Those in the right panel show a negligible probability of taking positive values. The red distribution is the estimate of the mixture distribution in (16), approximated as in (20). The black distribution is the same in the two panels. Notice how the red distribution closely resembles the black one, simply showing slightly higher tails due to the fact that, besides heterogeneity, it also accounts for sampling error. Finally, the right panel of Fig. 5 reports (green line) the sampling distribution obtained through KR, which is hard to distinguish from that $\mathrm{D}_{\mathrm{SMG}}$ produces. This similarity determines nearly identical prediction intervals, equal to $(-1.0552, -0.0085)$ and $(-1.0644, -0.0084)$, respectively for $\mathrm{D}_{\mathrm{SMG}}$ and KR (top-left panel of Fig. 6). Unlike those based on D_{BR} or $\mathrm{D}_{\mathrm{BR}^\star}$, they correctly exclude the value 0 and are highly non symmetric. The prediction standard errors, depicted in the bottom-left panel of Fig. 6, estimated via $\mathrm{D}_{\mathrm{SMG}}$ and KR are correctly above the dashed red line and fairly similar. Both methods also produce good standard errors estimates (bottom right panel) and comparable

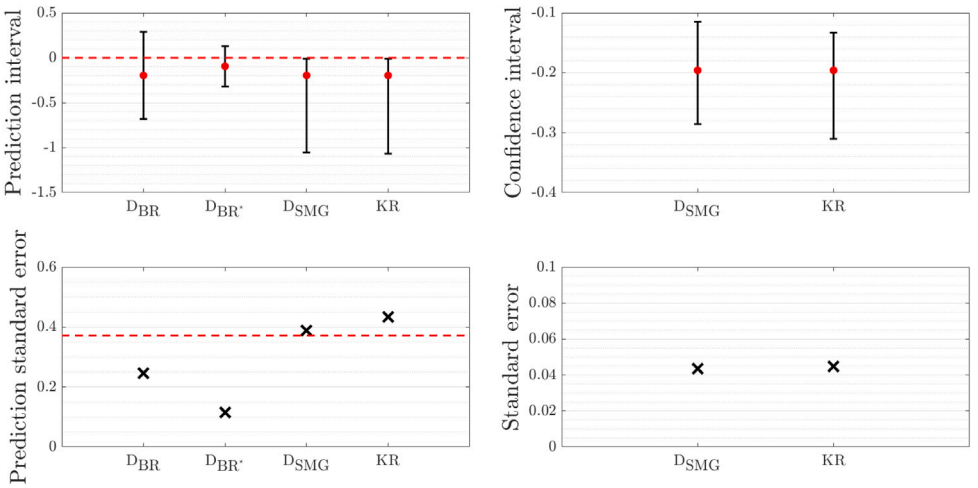

Fig. 6. Comparison between the performance of the D_{BR}, D_{BR^*}, D_{SMG} and KR methods in estimating the prediction interval, confidence interval, prediction standard error and standard error on real data from Bliemer and Rose (2013), for the fixed divided by log-normal case.

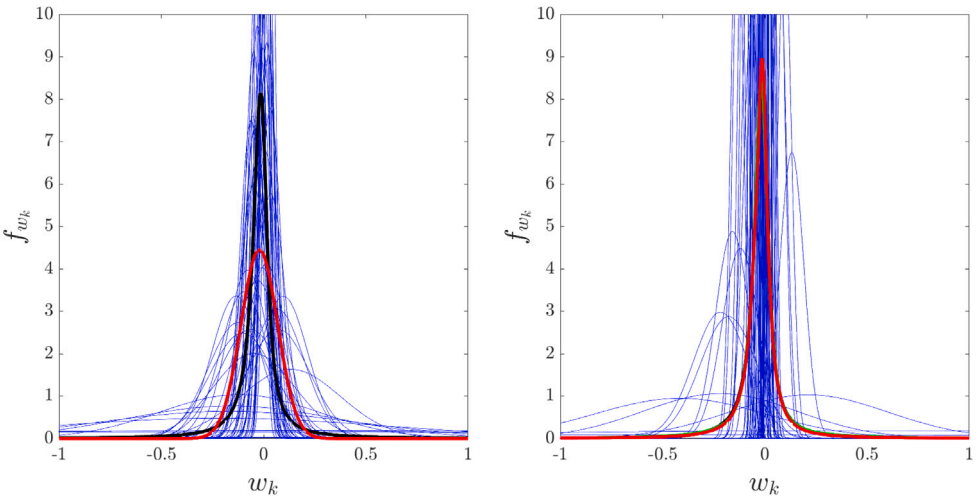

Fig. 7. Left panel: distributions of the conditional $\hat{w}_k(z)$ (blue lines) and the unconditional $\bar{w}_k$ based on the median and median standard error of $\hat{w}_k(z)$ (red line), according to the Bliemer and Rose (2013) approach to the Delta method. Right panel: distributions of the conditional $\hat{w}_k(z)$ (blue lines) and of the unconditional $\hat{w}_k$ (red line) according to our approach to the Delta method, and simulated sampling distribution of $\hat{w}_k$ according to the KR method (green line). Both panels: estimated distribution of w_k (black line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

confidence intervals (top right panel), with a slight difference between the two due to symmetry of the interval D_{SMG} necessarily produces.

6.3. The normal divided by normal case

In this case $\beta_k = \mu_k + \sigma_k z_k$, $\beta_c = \mu_c + \sigma_c z_c$ and w_k has an unknown distribution with infinite moments (see Section 5.3). Bliemer and Rose (2013) obtain the following estimates: $\hat{\mu}_k = -0.029$, $\hat{\sigma}_k = 0.051$, $\hat{\mu}_c = -0.951$, $\hat{\sigma}_c = 0.913$ and

$$\hat{\Omega}_{\hat{\mu}_k,\hat{\sigma}_k,\hat{\beta}_c,\hat{\beta}_k} = \begin{pmatrix} 0.00008 & -0.00002 & 0.00007 & 0.00001 \\ -0.00002 & 0.00014 & 0.00001 & -0.00007 \\ 0.00007 & 0.00001 & 0.00999 & 0.00463 \\ 0.00001 & -0.00007 & 0.00463 & 0.00762 \end{pmatrix}.$$

Fig. 7 reports in blue the conditional $\hat{w}_k(z)$, as obtained via D_{BR} (left panel) and D_{SMG} (right panel). We use random rather than Halton draws to avoid perfect collinearity between numerator and denominator of $\hat{w}_k$. The estimated distribution of w_k is, in this

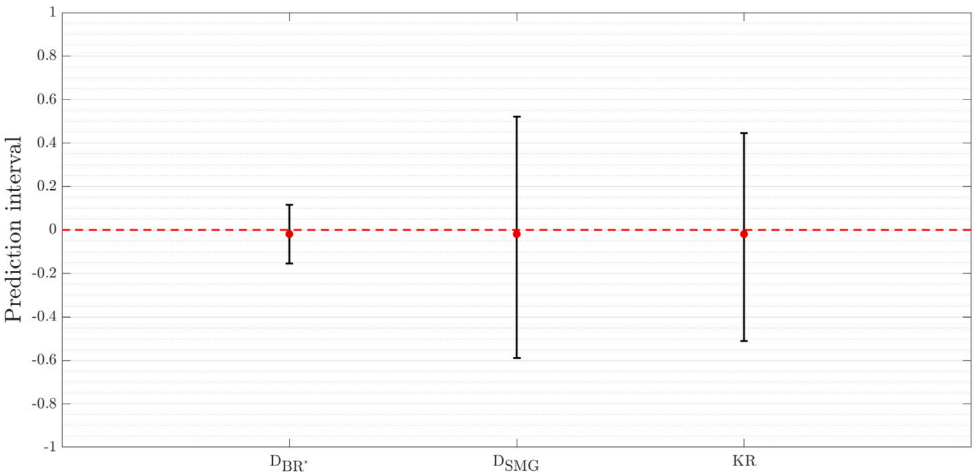

Fig. 8. Comparison between the performance of $D_{BR^\star}$, D_{SMG} and KR in estimating the prediction interval on real data from Bliemer and Rose (2013), for the normal divided by normal case.

case, unknown. We simulate it by drawing a large number of values from a $N(\hat{\mu}_k, \hat{\sigma}_k^2)$, and a $N(\hat{\mu}_c, \hat{\sigma}_c^2)$ so to calculate the negative of their ratio. In black, we report the resulting distribution in both panels. The red curve, in the left panel, represents the normal distribution with mean equal to the median of the conditional $\hat{w}_k(z)$ and standard deviation equal to the median of the conditional standard errors. This is equivalent to calculate moments adopting $D_{BR^\star}$. The red distribution has lower tails compared to the black one, determining a too small prediction interval, as Fig. 8 shows. This outcome is not surprising, given the simulation results in Section 5.3.

Considering the right panel of Fig. 7 one notices that the sampling distribution of $\hat{w}_k$, using D_{SMG} (red curve), closely resembles the one obtained via KR (green curve). Additionally, both are very similar to the black one. This determines larger and more reliable prediction intervals compared to the one $D_{BR^\star}$ produces. They are also reasonably close to each other (i.e. $(-0.5889, 0.5224)$ and $(-0.5119, 0.4454)$, respectively for D_{SMG} and KR).

7. Discussion and conclusion

This paper develops a method (D_{SMG}) to estimate prediction intervals, confidence intervals, standard errors, prediction standard errors, as well as other quantities of interest for WTP in a MXL framework.

D_{SMG}, similarly to D_{BR}, reformulates WTP as a function of the distributional parameters along with some parameter-free standard distributions so that one can use the Delta method. The uncertainty in the WTP measure depends on the variance–covariance matrix of the model parameter estimates. The paper develops a novel strategy to estimate this uncertainty and combine it with the variability of WTP due to heterogeneity. This strategy is not only theoretically sound but also delivers the whole sampling distribution of WTP, showing that this is, asymptotically, a mixture of normal distributions. The paper also shows how to exploit this sampling distribution to obtain many quantities of interest for WTP. This represents an important advantage D_{SMG} has over D_{BR}, that can only produce prediction standard errors and prediction intervals for the WTP. Additionally, we show that D_{BR} might yield, for some choices of the mixing distributions for the taste parameters, incongruous prediction standard errors and prediction intervals.

The paper evaluates D_{SMG} finite sample performances via a Monte Carlo study, under different choices of the mixing distributions for the MXL model parameters, comparing it with D_{BR} and KR. The latter is far more computationally intensive compared to D_{SMG} and D_{BR}, which are partly analytical and partly simulation-based. However, KR, as D_{SMG}, can provide the whole sampling distribution of WTP. Monte Carlo study results suggest that:

- D_{BR} produces reasonable results only when the mixing distribution of the taste parameter is normal with a fixed cost parameter, and, even in this case, the sample size needs to be sufficiently large and the sampling error of estimates much smaller compared to the variability due to taste heterogeneity;
- under different choices of the mixing distributions, i.e. different from the normal divided by fixed choice, D_{BR} produces questionable prediction intervals while the estimates of the prediction standard error can be seriously biased, independently of the sample size;
- D_{SMG} and KR provide asymptotically equivalent results and outperform D_{BR};
- D_{SMG} and KR produce valid prediction intervals, even for relatively small sample sizes, while they might require larger samples to compute confidence intervals for the mean WTP, depending on the choice of the mixing distribution of the parameters;
- for some choices of the mixing distributions (e.g. the fixed divided by log-normal case), D_{SMG} might require larger sample sizes compared to KR for the asymptotic theory to hold;

- even when a distribution with a positive probability mass in zero is chosen for the cost parameter, D_{SMG} and KR yield reasonable prediction intervals, while mean WTP, standard errors, and prediction standard errors do not exist (Daly et al., 2012b).

We also evaluate, in a comparative fashion, the three methods via the real data (Bliemer and Rose, 2013) use, to demonstrate, from a policy perspective, the relevance of our results. This confirms the risks linked to using D_{BR} when computing prediction standard errors or prediction intervals given specific choices of the mixing distribution for the taste parameters.

In summary, we recommend always using D_{SMG} to estimate prediction standard error and prediction intervals, since it is less computationally demanding compared to KR. One could use the latter, instead, for estimating confidence intervals for mean WTP whenever the sample size is not large enough to use D_{SMG}. To aid in using D_{SMG}, we provide an Excel spreadsheet implementing the methodology at the following link: http://www.trelab.it/utilities-wtpsmg-xlsx/. Routines used in this paper have been written in R (R Core Team, 2022) and a dedicated R package will be also shortly available.

To conclude, we underline that D_{SMG} can easily accommodate for the case of dependently distributed random parameters. This would imply rewriting the parameters as functions of parameter-free independent distributions multiplied by an appropriate Cholesky factorization matrix of the variance–covariance matrix of the parameters, as Bliemer and Rose (2013) suggest (see, for example, their equations (23) and (25)). In addition, one could also use D_{SMG} in the case of non linear utility functions, as well as when estimating the model in WTP space. In this case, not investigated in this paper, we expect the three methods to perform as in the normal divided by fixed case, when assuming normally distributed WTP parameters, and as in the fixed divided by log-normal case, when assuming log-normally distributed WTP parameters.

CRediT authorship contribution statement

Luisa Scaccia: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft. **Edoardo Marcucci:** Conceptualization, Validation, Writing – review & editing. **Valerio Gatta:** Conceptualization, Validation, Writing – review & editing.

Appendix A. List of abbreviations

Abbreviation	Definition
D_{BR}	Delta method computed according to Bliemer and Rose (2013) approach.
D_{BR^*}	Delta method computed according to Bliemer and Rose (2013) modified approach.
D_{SMG}	Delta method computed according to our approach.
KR	Krinsky and Robb (1986, 1990) method.
MC	Monte Carlo.
MNL	Multinomial logit model.
MXL	Mixed logit model.
LRP	Left rejection probability.
RRP	Right rejection probability.
WTP	Willingness to pay.

Appendix B. Bliemer and Rose formula for the normal divided by fixed case

Consider the variance of $\hat{w}_k(z)$ as given in Bliemer and Rose (2013) (see Eq. (10)):

$$\mathrm{Var}(\hat{w}_k(z)) = \begin{pmatrix} \nabla_{\theta_k} w_k(z) \\ \nabla_{\theta_c} w_k(z) \\ \nabla_{z_k} w_k(z) \\ \nabla_{z_c} w_k(z) \end{pmatrix}^T \begin{pmatrix} \Omega_{\hat{\theta}_k,\hat{\theta}_c} & \mathbf{0} \\ \mathbf{0}^T & \mathbf{1} \end{pmatrix} \begin{pmatrix} \nabla_{\theta_k} w_k(z) \\ \nabla_{\theta_c} w_k(z) \\ \nabla_{z_k} w_k(z) \\ \nabla_{z_c} w_k(z) \end{pmatrix}.$$

This can be rewritten as $\mathrm{Var}(\hat{w}_k(z)) = A(z) + B(z)$ where

$$A(z) = \begin{pmatrix} \nabla_{\theta_k} w_k(z) \\ \nabla_{\theta_c} w_k(z) \end{pmatrix}^T \Omega_{\hat{\theta}_k,\hat{\theta}_c} \begin{pmatrix} \nabla_{\theta_k} w_k(z) \\ \nabla_{\theta_c} w_k(z) \end{pmatrix}$$

and

$$B(z) = \begin{pmatrix} \nabla_{z_k} w_k(z) \\ \nabla_{z_c} w_k(z) \end{pmatrix}^T \mathbf{1} \begin{pmatrix} \nabla_{z_k} w_k(z) \\ \nabla_{z_c} w_k(z) \end{pmatrix}.$$

So, the prediction standard error considered in Bliemer and Rose (2013) and given in (13) can be rewritten as:

$$\mathrm{pse}(\bar{w}_k) \approx \frac{1}{R} \sum_{r=1}^{R} \sqrt{A(z^{(r)}) + B(z^{(r)})}. \tag{B.1}$$

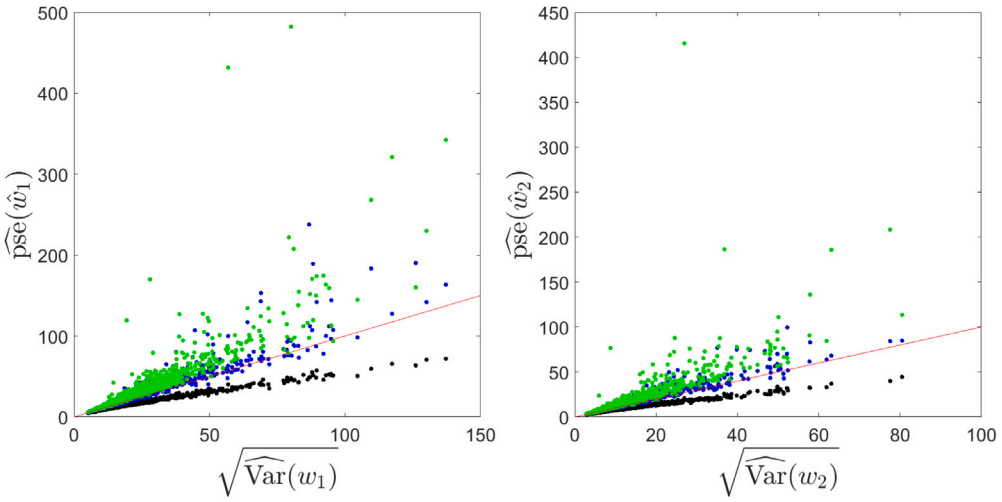

Fig. C.9. Estimates of the predictive standard error $\widehat{\text{pse}}(\hat{w}_k)$ at each Monte Carlo replicate plotted against estimates of standard deviation $\sqrt{\widehat{\text{Var}}(w_k)}$ for the same replicate. The comparison is performed for D_{BR} (black points), D_{SMG} (blue points) and KR (green points). The $\sqrt{\widehat{\text{Var}}(w_k)}$ values are the same for the three methods. Estimates are computed for $N = 150$. The red line represents the bisector of the first quadrant. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Consider, now, the normal divided by fixed case, where $w_k(z) = -(\mu_k + \sigma_k z_k)/\beta_c$. Then, $\nabla_{z_k} w_k(z) = -\sigma_k/\beta_c$ and $\nabla_{z_c} w_k(z) = 0$ so that $B(z) = \sigma_k^2/\beta_c^2$, which does not depend on z and is equal to the variance of w_k, that is the variance due to heterogeneity.[17] Notice, instead, that $A(z)$, in our approach, represents the variance of the conditional estimates $\hat{w}_k(z)$ (see Eq. (14)), which only accounts for sampling error.[18] Therefore, $A(z)$ is generally smaller than $B(z)$ and decreases as sample size increases, so that, for large samples, the prediction standard error in (B.1) is approximately equal to

$$\frac{1}{R}\sum_{r=1}^{R}\sqrt{B(z^{(r)})} = \frac{1}{R}\sum_{r=1}^{R}\sqrt{\sigma_k^2/\beta_c^2} = \sigma_k/\beta_c.$$

Thus, when the sample size is large enough for the sampling error to be negligible compared to heterogeneity, the formula in Bliemer and Rose (2013) gives a reasonable approximation of the prediction standard error for the normal divided by fixed case.[19]

We remark that this approximation is not valid in general. For example, for the fixed divided by the log-normal case, where $w_k(z) = \beta_k/\exp(\mu_c + \sigma_c z_c)$, we have $\nabla_{z_k} w_k(z) = 0$ and $\nabla_{z_c} w_k(z) = -w_k(z)\sigma_c$ so that $B(z) = w_k^2(z)\sigma_c^2$. Then, the Monte Carlo average $\frac{1}{R}\sum_{r=1}^{R}\sqrt{B(z^{(r)})} = \frac{1}{R}\sum_{r=1}^{R}\sqrt{w_k^2(z^{(r)})\sigma_c^2}$ is clearly different from the variance due to heterogeneity, given by $\text{Var}(w_k) = \left(\exp(\sigma_c^2) - 1\right)\exp\left(2(-\mu_c + \log|\beta_k|) + \sigma_c^2\right)$. Therefore, even for large sample sizes, the formula in Bliemer and Rose (2013) does not provide a reasonable approximation of the prediction standard error for the fixed divided by log-normal case.

Appendix C. The log-normal divided by log-normal case

When assuming both the parameters at the numerator and the denominator to be log-normal, the distribution of the WTP for X_k is known exactly, as well as its moments: $w_k \sim \text{sgn}(\beta_k)LN(\mu_{w_k}, \sigma_{w_k}^2)$, with $\mu_{w_k} = \mu_k - \mu_c$ and $\sigma_{w_k}^2 = \sigma_k^2 + \sigma_c^2$. We set $\beta_0 = 0.5$, $\beta_k \sim LN(\mu_k, \sigma_k^2)$, with $\mu_1 = 1$, $\mu_2 = 0.5$, $\sigma_1 = 0.5$, $\sigma_2 = 0.4$, and $-\beta_c \sim LN(\mu_c, \sigma_c^2)$, with $\mu_c = -1$ and $\sigma_c^2 = 1$, so that $w_1 \sim LN(2, 1.25)$ and $w_2 \sim LN(1.5, 1.16)$. The expected WTP value is $E(w_k) = \exp(-\mu_k + \sigma_{w_k}^2/2)$, thus $E(w_1) = 13.805$ and $E(w_2) = 8.004$, and the variance is $\text{Var}(w_k) = (\exp(\sigma_k^2 + \sigma_c^2) - 1)\exp(2(\mu_k - \mu_c) + \sigma_k^2 + \sigma_c^2)$, thus $\text{Var}(w_1) = 474.575$ and $\text{Var}(w_2) = 140.312$.

The log-normal divided by log-normal case, closely resembles the fixed divided by log-normal case. Thus, we just present simulation results here and summarize main findings, addressing the reader to the more in-depth discussion provided in Section 5.2.

Table C.10, clearly shows that both D_{SMG} and KR provide standard error estimates, $\hat{\text{se}}(\bar{w}_k)$, very close to the benchmark value but characterized by a large variability across Monte Carlo replicates. D_{BR} only provides prediction standard error estimates and these are systematically downward biased. For $N = 150$, the $\widehat{\text{pse}}(\hat{w}_k)$ averages are equal to 18.154 and 10.148, respectively for $k = 1$ and $k = 2$, which are much smaller than the benchmarks and smaller than the known w_k standard deviations (equal respectively to

[17] We are considering w_k and not its estimate, so its variance is only due to heterogeneity.

[18] More precisely, the expected value of $A(z)$ with respect to z represents the mean of the conditional variances and is the part of variability of $\hat{w}_k$ due to sampling error. See Eq. (17).

[19] Note that, in the normal divided by fixed case only, (B.1) would correctly account for both sampling error and heterogeneity if it were written as $\text{pse}(\bar{w}_k) \approx \sqrt{\frac{1}{R}\sum_{r=1}^{R} A(z^{(r)}) + B(z^{(r)})}$.

Table C.10
Log-normal divided by log-normal case: mean and standard deviation (in brackets) over Monte Carlo replicates for the estimated average WTP, its standard error and prediction standard error. The superscript (MC) denotes Monte Carlo estimates of benchmark values.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	$\bar{w}_1$	13.805	14.860 (5.542)	14.860 (5.542)	14.860 (5.542)
	$\hat{se}(\bar{w}_1)$	5.542$^{(MC)}$	4.712 (3.075)	–	5.476 (4.309)
	$\hat{pse}(\bar{w}_1)$	22.479$^{(MC)}$	30.253 (23.277)	18.154 (9.465)	37.256 (37.179)
$k = 2$	$\bar{w}_2$	8.004	8.591 (3.235)	8.591 (3.235)	8.591 (3.235)
	$\hat{se}(\bar{w}_2)$	3.235$^{(MC)}$	2.753 (1.778)	–	3.216 (2.528)
	$\hat{pse}(\bar{w}_2)$	12.279$^{(MC)}$	16.438 (12.034)	10.148 (5.414)	20.652 (22.412)
$N = 300$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	$\bar{w}_1$	13.805	14.446 (3.650)	14.446 (3.650)	14.446 (3.650)
	$\hat{se}(\bar{w}_1)$	3.650$^{(MC)}$	3.188 (1.432)	–	3.392 (1.634)
	$\hat{pse}(\bar{w}_1)$	22.088$^{(MC)}$	26.416 (13.481)	17.138 (6.073)	28.297 (15.480)
$k = 2$	$\bar{w}_2$	8.004	8.369 (2.132)	8.369 (2.132)	8.369 (2.132)
	$\hat{se}(\bar{w}_2)$	2.132$^{(MC)}$	1.869 (0.831)	–	1.995 (0.954)
	$\hat{pse}(\bar{w}_2)$	12.036$^{(MC)}$	14.527 (7.769)	9.592 (3.462)	15.677 (8.519)
$N = 450$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	$\bar{w}_1$	13.805	14.289 (2.872)	14.289 (2.872)	14.289 (2.872)
	$\hat{se}(\bar{w}_1)$	2.872$^{(MC)}$	2.540 (0.906)	–	2.641 (0.994)
	$\hat{pse}(\bar{w}_1)$	21.973$^{(MC)}$	25.143 (9.558)	16.729 (4.744)	26.430 (10.894)
$k = 2$	$\bar{w}_2$	8.004	8.290 (1.659)	8.290 (1.659)	8.290 (1.659)
	$\hat{se}(\bar{w}_2)$	1.659$^{(MC)}$	1.491 (0.523)	–	1.554 (0.581)
	$\hat{pse}(\bar{w}_2)$	11.961$^{(MC)}$	13.630 (5.026)	9.358 (2.678)	14.483 (6.003)

Table C.11
Log-normal divided by log-normal case: length, shape, LRP, RRP and coverage of 95% level prediction intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. The benchmark for the length of the interval is in italic as it represents a lower limit for the length, only accounting for variability due to heterogeneity.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	*65.283*	75.642	71.161	81.307
	Shape	4.030	4.187	1.000	4.558
	LRP	0.025	0.023	0.000***	0.026
	RRP	0.025	0.031	0.062***	0.028
	Coverage	0.950	0.945	0.938	0.946
$k = 2$	Length	*36.458*	42.411	39.781	45.817
	Shape	3.886	4.056	1.000	4.448
	LRP	0.025	0.023	0.000***	0.027
	RRP	0.025	0.032	0.064***	0.029
	Coverage	0.950	0.946	0.936*	0.944
$N = 300$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	*65.283*	71.764	67.180	74.134
	Shape	4.030	4.167	1.000	4.330
	LRP	0.025	0.023	0.000***	0.024
	RRP	0.025	0.026	0.057***	0.026
	Coverage	0.950	0.949	0.943	0.950
$k = 2$	Length	*36.458*	40.205	37.600	41.640
	Shape	3.886	4.026	1.000	4.202
	LRP	0.025	0.023	0.000***	0.024
	RRP	0.025	0.027	0.058***	0.026
	Coverage	0.950	0.950	0.942	0.950

(*continued on next page*)

$\sqrt{\text{Var}(w_1)} = 21.785$ and $\sqrt{\text{Var}(w_2)} = 140.312$). Fig. C.9 helps in depicting the underestimation problem of D_{BR}. The downward bias further increases with sample size. On the other hand, $\widehat{pse}(\hat{w}_k)$ mean over replicates is sensibly larger than the benchmark for both D_{SMG} and KR. This is mainly due to some very large $\widehat{pse}(\hat{w}_k)$ values and results improve rapidly as the sample size increases.

Table C.11 (*continued*).

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$N = 450$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	*65.283*	70.175	65.578	71.851
	Shape	4.030	4.138	1.000	4.257
	LRP	0.025	0.023	0.000	0.024
	RRP	0.025	0.026	0.056	0.025
	Coverage	0.950	0.951	0.944	0.951
$k = 2$	Length	*36.458*	39.277	36.684	40.357
	Shape	3.886	3.994	1.000	4.128
	LRP	0.025	0.024	0.000***	0.024
	RRP	0.025	0.026	0.056***	0.025
	Coverage	0.950	0.949	0.944	0.951

Table C.12
Log-normal divided by log-normal case: length, shape, LRP, RRP and coverage of 95% level confidence intervals. Significance codes: *** for p-value < 0.001; ** for p-value < 0.01; * for p-value < 0.05. The superscript (MC) denotes Monte Carlo estimates of benchmark values.

$N = 150$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	$22.386^{(MC)}$	18.471	–	20.820
	Shape	$2.860^{(MC)}$	1.000	–	2.038
	LRP	0.025	0.000***	–	0.020
	RRP	0.025	0.112***	–	0.072***
	Coverage	0.950	0.888***	–	0.908***
$k = 2$	Length	$12.552^{(MC)}$	10.793	–	12.227
	Shape	$2.742^{(MC)}$	1.000	–	2.068
	LRP	0.025	0.000***	–	0.020
	RRP	0.025	0.108***	–	0.066***
	Coverage	0.950	0.892***	–	0.914***
$N = 300$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	$13.625^{(MC)}$	12.496	–	13.153
	Shape	$2.040^{(MC)}$	1.000	–	1.662
	LRP	0.025	0.000***	–	0.024
	RRP	0.025	0.086***	–	0.060***
	Coverage	0.950	0.914***	–	0.916***
$k = 2$	Length	$8.089^{(MC)}$	7.326	–	7.733
	Shape	$2.072^{(MC)}$	1.000	–	1.676
	LRP	0.025	0.000***	–	0.024
	RRP	0.025	0.082***	–	0.050***
	Coverage	0.950	0.918***	–	0.926***
$N = 450$					
		Benchmark	D_{SMG}	D_{BR}	KR
$k = 1$	Length	$10.408^{(MC)}$	9.956	–	10.283
	Shape	$1.773^{(MC)}$	1.000	–	1.512
	LRP	0.025	0.003***	–	0.023
	RRP	0.025	0.071***	–	0.046***
	Coverage	0.950	0.926***	–	0.931**
$k = 2$	Length	$6.145^{(MC)}$	5.846	–	6.047
	Shape	$1.784^{(MC)}$	1.000	–	1.521
	LRP	0.025	0.002***	–	0.025
	RRP	0.025	0.072***	–	0.043***
	Coverage	0.950	0.926***	–	0.932**

Table C.11 shows prediction intervals for the three methods. For $N = 150$, both D_{SMG} and KR produce prediction intervals with correct LRP, RRP, and global coverage. D_{BR}, instead, has a significantly smaller than 95% global coverage for $k = 2$. While this problem seems to vanish as the sample size increases, LRP and RRP are still significantly far from 2.5% even when $N = 450$. As the sample size increases, the situation does not improve: the prediction interval is calculated on the basis of a normal distribution, while the $\hat{w}_k$ sampling distribution is a mixture of normals, which tends to the distribution of w_k (a highly skewed log-normal distribution) when N increases and the sampling error becomes progressively less relevant. In particular, the lower D_{BR} prediction

Table C.13
Log-normal divided by log-normal case: skewness test for the composite hypothesis of normality.

Parameter	$N = 150$		$N = 300$		$N = 450$	
estimator	Skewness	p-value	Skewness	p-value	Skewness	*p*-value
$\hat{\mu}_1$	0.075	0.350	−0.132	0.089	−0.048	0.532
$\hat{\mu}_2$	−0.155	0.043	−0.194	0.010	−0.062	0.420
$\hat{\mu}_c$	−0.434	0.000	−0.325	0.000	−0.147	0.053
$\hat{\beta}_1$	−0.046	0.535	0.179	0.021	0.072	0.364
$\hat{\beta}_2$	−0.224	0.003	−0.008	0.917	−0.068	0.361
$\hat{\beta}_c$	0.111	0.155	0.162	0.040	0.132	0.100
$\hat{\omega}_1$	1.583	0.000	1.238	0.000	1.149	0.000
$\hat{\omega}_2$	1.748	0.000	1.140	0.000	1.043	0.000

interval bound is always negative in the M simulations, for any sample size N. D_{SMG} and KR, instead, correctly produce positive lower prediction intervals for all M replicates and sample sizes.

Table C.12 summarizes D_{SMG} and KR performances when computing confidence intervals for $E(w_k)$. Both methods show unsatisfactory results even for $N = 450$, with KR performing slightly better and achieving, at least, a reasonable LRP. As N increases, both methods improve and KR maintains its superiority.

We show in Table C.13 the results of the skewness test for the composite hypothesis of normality (Shapiro et al., 1968), with respect to the estimators of the parameters and $\bar{w}_k$. The distribution of $\bar{w}_k$ is still very skewed and far from normality even for $N = 450$. KR only assumes the normality of the estimators of the structural parameters, which explains its better performance compared to D_{SMG}.

References

Train, K., 2009. Discrete Choice Methods with Simulation, second ed. Cambridge University Press.
McFadden, D., 1974. Conditional logit analysis of qualitative choice behaviour. In: Frontiers in Econometrics. Zarembka, P.
Krinsky, I., Robb, A.L., 1986. On approximating the statistical properties of elasticities. Rev. Econ. Stat. 68, 715–719.
Krinsky, I., Robb, A.L., 1990. On approximating the statistical properties of elasticities: A correction. Rev. Econ. Stat. 72, 189–190.
Bolduc, D., Khalaf, L., Yélou, C., 2010. Identification robust confidence set methods for inference on parameter ratios with application to discrete choice models. J. Econometrics 157, 317–327.
Armstrong, P., Garrido, R., Ortúzar, J.D., 2001. Confidence intervals to bound the value of time. Transp. Res. E 37, 143–161.
Gatta, V., Marcucci, E., Scaccia, L., 2015. On finite sample performance of confidence intervals methods for willingness to pay measures. Transp. Res. A 82, 169–192.
Hole, A.R., 2007. A comparison of approaches to estimating confidence intervals for willingness to pay measures. Health Econ. 16, 827–840.
McFadden, D., Train, K., 2000. Mixed MNL models for discrete response. J. Appl. Econometrics 15, 447–470.
Train, K., Weeks, M., 2005. Discrete choice models in preference space and willingness-to-pay space. In: Alberini, A., Scarpa, R. (Eds.), Applications of Simulations Methods in Environmental Resource Economics. Kluwer Academics Publisher, Dordrecht, The Netherlands.
Scarpa, R., Thiene, M., Train, K., 2008. Utility in willingness to pay space: a tool to address confounding random scale effects in destination choice to the alps. Am. J. Agric. Econ. 90, 994–1010.
Hensher, D.A., Greene, W.H., 2003. The mixed logit model: The state of practice. Transportation 30, 133–176.
Sillano, M., Ortúzar, J., 2005. Willingness-to-pay estimation with mixed logit models: some new evidence. Environ. Plan. A 37, 525–550.
Michaud, C., Llerena, D., Joly, I., 2013. Willingness to pay for environmental attributes of non-food products: a real choice experiment. Eur. Rev. Agric. Econ. 40, 313–329.
Bliemer, M.C., Rose, J.M., 2013. Confidence intervals of willingness-to-pay for random coefficient logit models. Transp. Res. B 58, 199–214.
Daziano, R., Achtnicht, M., 2014. Accounting for uncertainty in willingness to pay for environmental benefits. Energy Econ. 44, 166–177.
Sarrias, M., Daziano, R., 2018. Individual-specific point and interval conditional estimates of latent class logit parameters. J. Choice Model. 27, 50–61.
Fosgerau, M., 2006. Investigating the distribution of the value of travel time savings. Transp. Res. B 40, 688–707.
Daly, A., Hess, S., de Jong, G., 2012a. Calculating errors for measures derived from choice modelling estimates. Transp. Res. B 46, 333–341.
Daly, A., Hess, S., Train, K., 2012b. Assuring finite moments for willingness to pay in random coefficient models. Transportation 39, 19–31.
Klein, L., 1953. A Textbook of Econometrics. Prentice-Hall.
Withers, C.S., Nadarajah, S., Shih, S.H., 2015. Moments and cumulants of a mixture. Methodol. Comput. Appl. Probab. 17, 541–564.
Shapiro, S., Wilk, M., Chen, H., 1968. A comparative study of various tests for normality. J. Amer. Statist. Assoc. 63, 1343–1372.
R Core Team, 2022. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL: https://www.R-project.org/.