Universiteit Antwerpen
CSB | Centrum voor Sociaal Beleid Herman Deleeck

Lukas Leitner

Imprecision in the Estimation of Willingness to Pay Using Subjective Well-Being Data

WORKING PAPER

No. 23/10

September 2023

Universiteit Antwerpen

University of Antwerp
Herman Deleeck Centre for Social Policy
https://www.uantwerpen.be/en/research-groups/csb/

Imprecision in the Estimation of Willingness to Pay Using Subjective Well-Being Data

Lukas Leitner [1]

[1] Centre for Social Policy, University of Antwerp (Belgium)

Working Paper No. 23/10

September 2023

Abstract

The subjective well-being (SWB) method has become a popular tool to estimate the willingness to pay for non-market goods. In this method, the willingness to pay measure contains the ratio of two coefficients (of the nonmarket good and consumption), which are both estimated in a regression on subjective well-being. Computing confidence intervals for such ratios turns out to be error-prone, in particular when the consumption coefficient is imprecisely estimated. In this paper, five different ways of computing the confidence intervals are compared: the delta, Fieller, parametric bootstrapping, and bootstrapping method, and a numerical integration of Hinkley's formula. Using a large number of simulated SWB data sets, confidence intervals and their coverage rates are computed for each method. The findings suggest that the delta method is accurate only if the consumption coefficient is estimated with very high precision. All other methods turn out to be more robust, with minor differences in accuracy.

Keywords: Willingness to pay; preference estimation; subjective well-being; normal ratio distribution; confidence intervals.

JEL-classification: C46, C15, I31.

Acknowledgments: The author would like to thank FWO/FNRS for their support through the EOS project 30544469. This paper has benefited substantially from comments by Koen Decancq, Santiago Burone, Sebastiaan Maes, as well as discussions at the lunch seminar series at the Centre for Social Policy (University of Antwerp) and at the WAP2023 conference in Bordeaux.

Corresponding author: Lukas Leitner, Lukas.Leitner@uantwerpen.be

1 Introduction

In many fields of economics and other social sciences, researchers are interested in the willingness to pay (WTP) of individuals for a variety of goods. This can help to predict the benefit, expected outcome, or political acceptance of certain policies, or to find the optimal price for a product. One of the advantages of the WTP measure is that the non-monetary good can be any good for which individuals have preferences. Some examples are applications on air quality (Luechinger 2009), travel time (Amador, González, and Dios Ortúzar 2005), crime reduction (Brenig and Proeger 2018), mental distress from bereavement (Oswald and Powdthavee 2008) or from other life events (Clark and Oswald 2002). To obtain WTP estimates, a method to elicit these preferences is needed. Different approaches to do so are the revealed preference approach, contingent valuation, discrete choice experiments, or regressions on reported subjective well-being (SWB). The last two methods, discrete choice experiments and the SWB method, are prone to a common source of error; for all differentiable and additively separable utility functions, the estimated WTP measure contains the ratio of two coefficient estimators.

In case an ordinary least squares or a maximum likelihood estimation is used, the estimators are asymptotically normally distributed. Accordingly, their ratio asymptotically follows a normal ratio distribution. It is known for a long time that the statistical properties of the normal ratio distribution differ from those of the normal distribution in many ways (Geary 1930; Fieller 1940; Fieller 1954; Marsaglia 1965; Hinkley 1969), and that one should be cautious when using the normal distribution to approximate it. Most notably, the moments of the normal ratio distribution are not defined, such that the mean and standard deviation of the final estimate are generally meaningless (Daly, Hess, and Train 2011). Theoretically, these are infinite, even though empirical applications always generate finite (and often even seemingly reasonable) moments, luring researchers into believing their results were correct.

While a growing body of literature compares different methods to construct confidence intervals for the normal ratio distribution in the context of discrete choice experiments (see, for instance, Hole 2007; Bolduc, Khalaf, and Yélou 2010; Gatta, Marcucci, and Scaccia 2015; Wang et al. 2020), I am not aware of any study published to this date doing the same in the context of SWB data. The aim of this paper is to find out which of the most commonly used methods (Fieller, bootstrap, parametric bootstrap, and delta) yield sufficiently accurate confidence intervals for WTP when using the SWB method. I further propose a new method where the formula given by Hinkley (1969) is numerically integrated. While this list of methods is not exhaustive, it covers a spectrum of different assumptions and approaches. Moreover, not all methods use the same value around which the confidence interval is located. It is not evident that the median of the distribution of estimated WTP should also be the median of the confidence interval for WTP. I discuss why it can be advantageous to choose intervals which are not located around the median. Finally, the accuracy and robustness of these different methods are compared by means of a simulation, using a large number of data sets based on the same underlying preferences. By varying the cut-off levels on the t-values of the coefficient estimators, I investigate whether the popular cut-off level of 1.96 is sufficient in the given context, or whether the t-value of the denominator should be at least 3, as proposed by Geary (1930).

The findings in this study underline that special attention needs to be given to the monetary coefficient estimator. As it enters the formula for WTP in the denominator, larger imprecision in its estimate can lead to an exorbitant level of imprecision in the WTP estimate. When applying the conventional cut-off level of 1.96 on the t-value of the monetary coefficient estimator, the confidence intervals for the final estimate may be substantially less accurate than for a cut-off level of 3. Further, the accuracy depends not only on the chosen cut-off values, but also on whether the expected t-values are sufficiently large. Accordingly, the statistical power of a SWB survey aimed at estimating WTP should be larger than that of a SWB survey investigating the determinants of well-being. Another finding is that all methods presented here except for the delta method perform reasonably well when a cut-off on the monetary coefficient estimator is applied. The delta method fails to incorporate the skewness of the normal ratio distribution, such that it is particularly inaccurate for certain significance levels while being fairly accurate for others. It is also the least robust method to changes in correlation between the estimators, smaller sample sizes, and lower statistical power. Although the focus of this paper is on the SWB method, most of the findings also apply to applications with discrete choice experiments.

In section 2, I briefly outline how the willingness to pay measure is estimated using the SWB method, and show why the normal ratio distribution appears in the final estimate. Section 3 gives an overview of the five methods to construct confidence intervals and discusses the choice of the interval mid-point. I describe the set up of the simulation in section 4 and the results in section 5. In section 6, I discuss the results and draw inference for practitioners before concluding in the final section.

2 Estimation

2.1 Subjective well-being method

The SWB method requires a data set which includes reported life satisfaction, individual outcomes in the monetary good and the non-monetary dimension for which WTP is estimated, and ideally some additional variables which are correlated with SWB and individual outcomes (e.g. socio-demographics, personality traits, or answers to locus of control questions). It is further required to assume a parametric model of preferences. Reported life satisfaction LS_i is then regressed on all the other variables to measure their influence on life satisfaction. For the example below, and without loss of generality, I use a simple log-linear utility function over two life dimensions, consumption c and the non-monetary good k.[1] The log-linear specification is obtained by transforming one life dimension, in this case consumption, with the logarithm. Given some additional variables z_i, and an unobserved, random component ε_i, the utility function looks as follows:

$$\mathrm{LS}_i = \beta_0 + \beta_c \cdot \log(c_i) + \beta_k \cdot k_i + \beta_z \cdot z_i + \varepsilon_i. \tag{1}$$

The idea behind WTP is to keep the level of utility constant while changing the outcome in the non-monetary good of interest. It is given by the level of con-

[1] As shown in the course of this paper, the findings hold for a much wider set of utility functions and for more than two dimensions.

sumption which individual i would forego to attain k^*, the hypothetical outcome level of good k.[2] Figure 1 illustrates this concept:

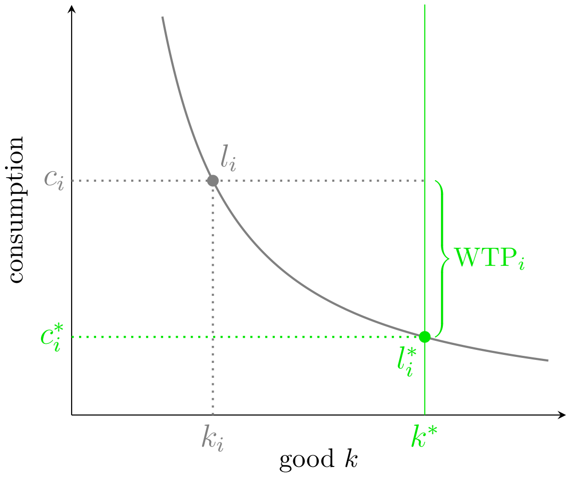


Figure 1: Illustration of willingness to pay measure.

Individual i is endowed with the outcome $l_i = \begin{pmatrix} c_i & k_i \end{pmatrix}$ in both life dimensions and has complete, continuous, and transitive preferences over the space of outcomes $\mathbb{R}_+^2$, here depicted by the indifference curve through l_i. The intersection of the indifference curve through l_i with the vertical line at k^* shows the level of consumption c_i^* at which i is equally well off as in her observed situation, but with the hypothetical attainment k^* in the non-monetary dimension.[3] This level of consumption c_i^* is also called "equivalent consumption".[4] The vertical distance between c_i^* and the observed level of consumption c_i then represents WTP$_i$. To derive the formula for WTP, both l_i and $l_i^* = \begin{pmatrix} c_i^* & k^* \end{pmatrix}$ are inserted in the right-hand

[2] It is implicitly assumed here that all individuals attain the same reference outcome k^*, which is obviously not the case in every real-world application. One could alternatively use an individual-specific value k_i^*.

[3] Note that good k may as well be a discrete or binary variable. In that case, the indifference curve would be an indifference set. It is however required that consumption is continuous, otherwise there may be no consumption level which yields the same predicted life satisfaction as the observed outcome.

[4] Equivalent consumption itself can also be used as a measure to compare individual well-being in a multi-dimensional setting. It combines multiple life dimensions into one by deducting the WTP for each non-monetary dimension from the consumption level, hence respecting individual preferences (Decancq, Fleurbaey, and Schokkaert 2015).

side of equation 1, and the two expressions equated. After some transformations, one obtains:

$$\mathrm{WTP}_i = c_i \cdot \left(1 - \exp\left[\frac{\beta_k}{\beta_c} \cdot \big(k_i - k^*\big)\right]\right). \tag{2}$$

Two features of equation 2 are important for the concepts presented in this paper. First, only the outcome bundle of individual i, the new attainment in the non-monetary dimension, and the coefficients β_c and β_k are needed to determine i's WTP. This means that preferences are not individual, but assumed to be equal for the whole population.[5] Hence, when all individual outcomes l_i, and the coefficients β_c and β_k are known, the WTP of every individual is determined, independent of their reported life satisfaction or other individual characteristics. Second, the coefficients β_c and β_k appear as a ratio where β_c is in the denominator, which is always the case for any additively separable utility function.[6] In this example, these two coefficients are also the only coefficients which need to be estimated. For additively separable functions with more parameters, equation 2 may contain more coefficients to be estimated, but β_c and β_k still appear as a ratio. The regression is only a means to an end; its purpose is to yield an unbiased and precise estimate of the coefficients determining the shape of the indifference curves.

2.2 Normal ratio distribution

Imprecision is the natural by-product of any estimation using regressions. When interpreting the results, it is imperative to answer the question how this imprecision is captured in the final estimate. In this paper, imprecision is defined from a frequentist perspective. It is assumed that the coefficients which determine the utility function have a true value, which is unknown to the observer. By means of a regression, the observer can form an expectation in which range these values may be located, or in other words, she can compute confidence intervals. The width of

[5] It is possible to allow for interaction effects between the life dimensions and, typically, socio-demographic characteristics. In that case, preferences are equal for everyone belonging to the same socio-demographic group.

[6] The proof is fairly simple: Following the notation of this paper, any additively separable utility function can be written as $\mathrm{LS}_i = \beta_0 + \beta_c \cdot f_c(c_i) + \beta_k \cdot f_k(k_i) + \cdots + \varepsilon_i$, where f_c and f_k denote the transformation functions of c_i and k_i, respectively. Setting $\mathrm{LS}_i(c_i, k_i) = \mathrm{LS}_i(c_i^*, k^*)$ and rewriting the equation yields: $\mathrm{WTP}_i = c_i - f_c^{-1}\big[f_c(c_i) + \big(\frac{\beta_k}{\beta_c} \cdot [f_k(k_i) - f_k(k^*)]\big)\big]$.

a confidence interval is determined by the significance level α, its mid-point, and the level of imprecision. The significance level denotes the likelihood α with which the confidence interval of a random sample from the population contains the true parameter value, and the mid-point divides the probability mass of the confidence interval in half. Hence, if the significance level and mid-point are given, the size of the confidence interval only depends on the level of imprecision. In contrast to a Bayesian perspective, imprecision is *not* interpreted as a distribution of coefficients in the population, but as a data-driven limitation which hinders the observer to narrow down confidence intervals further.

In the context of the SWB method, the parameters β_c and β_k are typically estimated using an ordinary least squares or a maximum likelihood regression.[7] If so, their estimators are asymptotically normally distributed. Under the assumption that the estimators are jointly distributed, it follows that their ratio asymptotically follows a normal ratio distribution. For the remainder of the paper, the estimators of β_c and β_k are denoted by $\hat{\beta}_c$ and $\hat{\beta}_k$ respectively, and the estimator of their ratio by $\hat{\pi} := \hat{\beta}_k / \hat{\beta}_c$. Using the assumption of joint distribution, the multivariate distribution is given by:

$$\begin{gathered}\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_c \\ \hat{\beta}_k \end{pmatrix} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \\ \text{where } \boldsymbol{\mu} = \begin{pmatrix} \mu_c \\ \mu_k \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_c^2 & \sigma_{ck} \\ \sigma_{ck} & \sigma_k^2 \end{pmatrix}. \end{gathered} \tag{3}$$

Some characteristics of the normal ratio distribution severely complicate the analysis. Four of these have immediate consequences for the construction of confidence intervals: First, the normal ratio distribution is "heavy tailed", meaning that its moments (mean, variance, etc.) are infinite, and hence not defined. Empirically, one would always obtain finite moments when taking some random samples, but

[7] Life satisfaction is usually reported on discrete scales, e.g. on the scale of integers from 0 to 10. Since a one point difference in reported SWB at the top of the scale may not represent an equally large difference in real subjective well-being as a one point difference at the bottom of the scale, it seems more sensible to use an ordered logit model for the regression. However, as shown by Ferrer-i-Carbonell and Frijters (2004), the coefficient estimators obtained when using ordered logit models are very similar to those obtained with ordinary least squares.

since the law of large numbers does not apply to moments which are not defined, the sample moments do not converge. Daly, Hess, and Train (2011) provide a proof and an intuition for this finding: Since $\hat{\beta}_c$ appears in the denominator and has a "relatively high" likelihood for values arbitrarily close to 0, the resulting ratio can be arbitrarily large with "relatively high" likelihood. The theorem in Daly, Hess, and Train (2011) states the exact condition that determines which moments are undefined for which type of ratio distribution (e.g. for the inverse gamma, Weibull, and uniform distribution, etc.). Truncating $\hat{\beta}_c$ such that values around 0 were excluded would guarantee that $\hat{\pi}$ had finite moments, but it might be difficult to find defensible *a priori* arguments for this choice.[8] Since the moments of the normal distribution are well-defined and finite, any approximation using the normal distribution always underestimates the likelihood of extreme outcomes. Practitioners who estimate confidence intervals for WTP using such an approximation may strongly underestimate the width of these intervals.

Second, the normal ratio distribution can be bimodal in some cases. Marsaglia (1965) shows for which combination of values the distribution is uni- or bimodal. Bimodality occurs when the probability mass of the denominator is located closely around 0 and the probability mass of the nominator is not, i.e. when $\hat{\beta}_k$ is relatively far away from 0 and $\hat{\beta}_c$ is relatively close to 0.[9] As described above, the mean of the normal ratio distribution is not defined, and its mode is not necessarily unique. Additionally, the median of the normal ratio distribution is generally not equal to the ratio μ_k/μ_c. Hence, it is unclear which central measure describes the location of the normal ratio distribution.

Third, there exists no closed-form solution to compute the quantiles of $\hat{\pi}$. Moreover, the normal ratio distribution depends on five different continuous parameters (the means μ_c and μ_k, the variances σ_c^2 and σ_k^2, and the covariance σ_{ck}), such that tables with pre-calculated values, as could be found for, for instance, Student's

[8] Note that a truncation of the values around 0 is different from assuming strict monotonicity. As long as $\hat{\beta}_c$ can take values arbitrarily close to 0, the moments of $\hat{\pi}$ are infinite. Thus, one would need to specify a value $\epsilon \neq 0$ for which $\hat{\beta}_c \geq \epsilon$, or $\hat{\beta}_c \leq -\epsilon$, or $\hat{\beta}_c \notin (-\epsilon, \epsilon)$ is assumed.

[9] These conditions provide that values of $\hat{\pi}$ close to 0 are relatively unlikely, since $\hat{\pi}$ can only be close to 0 when $\hat{\beta}_k$ is close to 0 and/or when $\hat{\beta}_c$ is very large. On the other hand, $\hat{\beta}_c$ can take both negative and positive values with a high likelihood, such that $\hat{\pi}$ can peak on both sides of the vertical axis. If there are peaks on both sides, the distribution is bimodal.

t-distribution, are not available.

And fourth, as noted in Dufour (1997), the normal ratio distribution is "locally almost unidentified". This means that every valid method to construct its confidence intervals must allow for unbounded intervals when the parameters β_c and β_k cannot be identified with the given data.

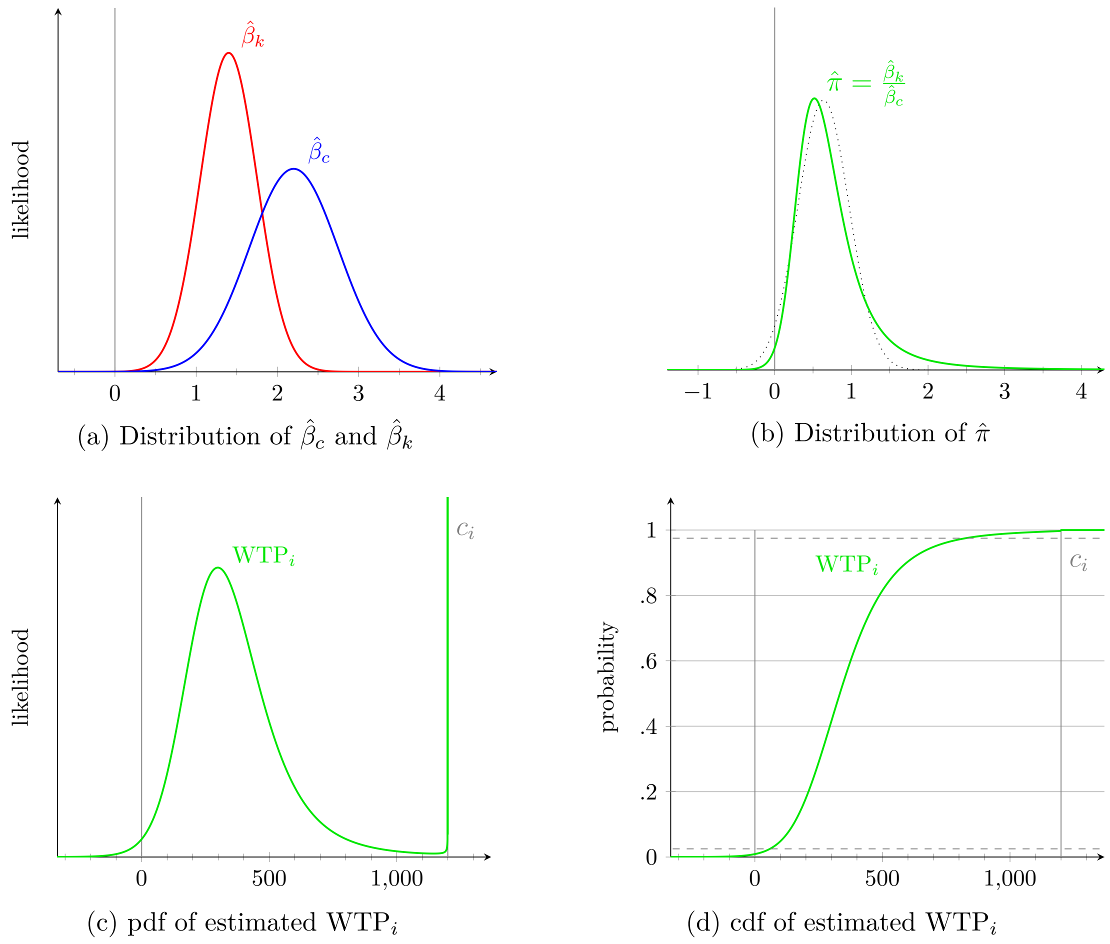
When plugging $\hat{\pi}$ into the formula for WTP, some of these characteristics might potentiate even further.[10] Incorrectly specified confidence intervals may be strongly biased and may fail to reflect the skewness as well as the heavy tails of the normal ratio distribution. Figure 2 illustrates the intermediate steps when estimating WTP for some hypothetical case, where $\hat{\beta}_c$ and $\hat{\beta}_k$ are jointly and normally distributed.[11] The t-values of $\hat{\beta}_c$ and $\hat{\beta}_k$ are equal to 4, meaning that both estimators are estimated with high precision. Yet, we observe that the distribution of $\hat{\pi}$ is considerably skewed to the right. The function visibly converges to 0 more slowly than a normal distribution, as indicated by the dotted line. Inserting $\hat{\pi}$ into equation 2 yields the distribution of estimated WTP, as shown in figure 2c. We observe that the probability mass of estimated WTP is distributed over a wide range of values. The dashed lines in figure 2d indicate the lower and upper bounds of a confidence interval centred around the median of $\hat{\pi}$. Even though the estimators are highly significant compared to conventional significance levels, the interval spreads from a value close to 0 to a value which is about two thirds of the total consumption.

3 Construction of Confidence Intervals

As shown in the previous section, it is futile to capture the imprecision of $\hat{\pi}$ using its standard deviation, since the standard deviation of a normal ratio distribution is not defined. In this section, I present five methods to obtain confidence intervals for a normal ratio distribution. Afterwards, I discuss why the central points around

[10] Consider the formula for WTP with a log-linear utility function shown in equation 2. Due to the exponentiation of β_k/β_c, small differences in the preference parameter can lead to large differences in the WTP estimate.

[11] For illustrative purposes, a Cobb-Douglas utility function is used here, where consumption is transformed by a logarithm. This implies that WTP cannot be higher than the consumption value, which in this hypothetical case is equal to 1200.



(a) Distribution of $\hat{\beta}_c$ and $\hat{\beta}_k$

(b) Distribution of $\hat{\pi}$

(c) pdf of estimated WTP$_i$

(d) cdf of estimated WTP$_i$

Figure 2: Illustration of WTP estimation step-by-step.

which these confidence intervals are located differ between methods and why this may or may not be desirable in the given context.

3.1 Five methods

The five methods presented here are the Fieller method, the Hinkley method,[12] the naïve bootstrap, the parametric bootstrap, and the delta method. Neither is this list of methods exhaustive, nor are these methods considered optimal in terms

[12] This method has, to the best of my knowledge, not been described before. It utilises the formula for the normal ratio distribution given by Hinkley (1969), wherefore I use his name.

of accuracy.[13] It is not the aim of this paper to find the most accurate method; instead, I test whether the most commonly used ones are suitable, and whether the different assumptions and approaches employed for each method are robust.

Fieller method

The Fieller method was introduced by Fieller (1940) and discussed in more detail in Fieller (1954). It gives confidence bounds for the ratio of two jointly normally distributed variables.[14] Hence, when applying it to the SWB method, one needs to assume that $\hat{\beta}_c$ and $\hat{\beta}_k$ follow a multivariate normal distribution. The accuracy of the confidence set thus partly depends on whether the sample size is large enough to elicit the asymptotic properties of the two estimators.

Figure 3 illustrates how the Fieller confidence set is constructed, given some hypothetical values.[15] In this example, both t-values are equal to 4, and the point (μ_c, μ_k) representing the ratio of means is surrounded by the 95% confidence ellipse. The construction starts by projecting the ratio of means, i.e. point (μ_c, μ_k), onto the vertical line at $\beta_c = 1$. This projection is equivalent to dividing β_k by β_c, as all points on the ray through the origin have the same ratio π. The first step already yields the mid-point π_M of the confidence set. Next, the rays through the origin which are tangent to the confidence ellipse are drawn. The intersections of these lines with the vertical line at $\beta_c = 1$ yield the lower and upper bound of the Fieller confidence set.[16] Again, this projection of the outmost points of the confidence ellipse onto $\beta_c = 1$ is equivalent to calculating the ratio of their coordinates. Observe that the Fieller method allows for skewness, as π_L and π_U are not equidistant from π_M. On that account, the mid-point of the Fieller

[13] See for instance Armstrong, Garrido, and Ortúzar (2001), Gatta, Marcucci, and Scaccia (2015), Puth, Neuhäuser, and Ruxton (2015), or Wang et al. (2020) for further examples and variations of the methods described here, some of which are considered to be more accurate or to have better small-sample properties. Carson and Czajkowski (2019) develop an entirely different approach where the cost coefficient is re-parametrised, such that the ratio is estimated directly.

[14] Interested readers can find the formula in appendix B.2.

[15] The illustration is adapted from von Luxburg and Franz (2009), who advocate for using the Fieller method when the nominator and denominator are normally, or even approximately normally distributed.

[16] In case the point where the ray through the origin is tangent to the confidence ellipse lies left of $\beta_c = 0$, the ray needs to extended through the origin to find the intersection with $\beta_c = 1$ (see figure 4).

confidence set should not be understood as its geometric centre, but as the point which splits the probability mass contained between π_L and π_U in half. Note that the area between the dashed lines through π_L and π_U covers the entire confidence ellipse, and the dashed line through π_M splits this area equally.[17]

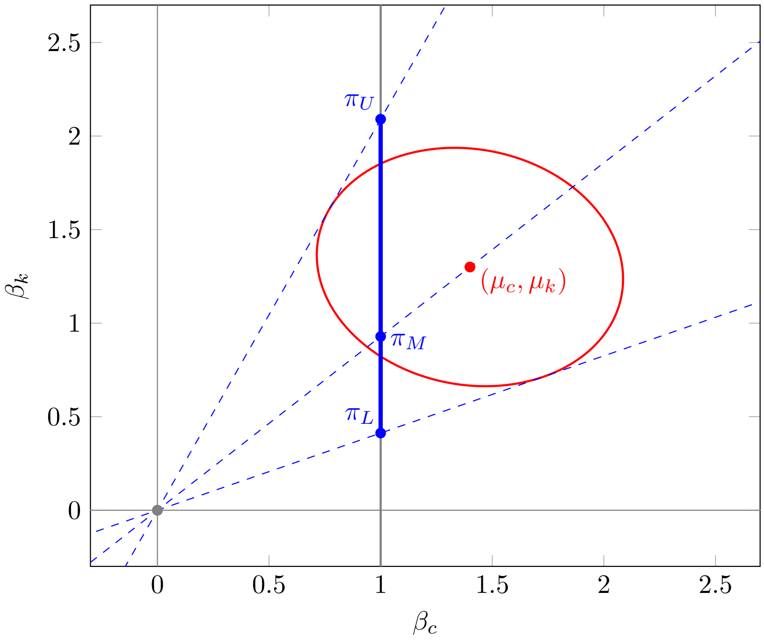


Figure 3: Construction of Fieller confidence set.

Careful readers might have noticed that I use the terms "confidence set" and "confidence bounds" instead of "confidence interval". The reason for this is that the Fieller method does not necessarily yield an (inclusive) interval. Consider the case illustrated in figure 4, where $\hat{\beta}_c$ is not significant at the significance level of the confidence ellipse. Here, the t-values are set to $t_c = -0.3$ and $t_k = 2.2$. Again, all values are hypothetical and only chosen to illustrate the concept. Using the same method to construct the Fieller confidence set as shown above yields a lower bound which is larger than the upper bound. Moreover, the mid-point does not lie

[17] As not only the confidence ellipse, but also a large area outside of it are captured by the confidence set, one might be tempted to think that the estimated confidence set is far too conservative. However, recall that the confidence ellipse is not equivalent to the quantile ellipse. The first describes the confidence set of the mean value, while the latter contains a given percentage of some sample data. For jointly normally distributed variables, the quantile ellipse is slightly larger than the confidence ellipse.

between the two. In this case, the confidence interval is exclusive, meaning that all values between the lower and upper bound are *not* part of the confidence set, while all other values of the set of real numbers are. In mathematical notation, the confidence set is given by $(-\infty, \pi_U) \cup (\pi_L, \infty)$. Such an interval can be understood as follows: If $\hat{\beta}_c$ is not significantly different from 0, but $\hat{\beta}_k$ is, the resulting ratio is expected to be different from 0, and potentially very far away from it.

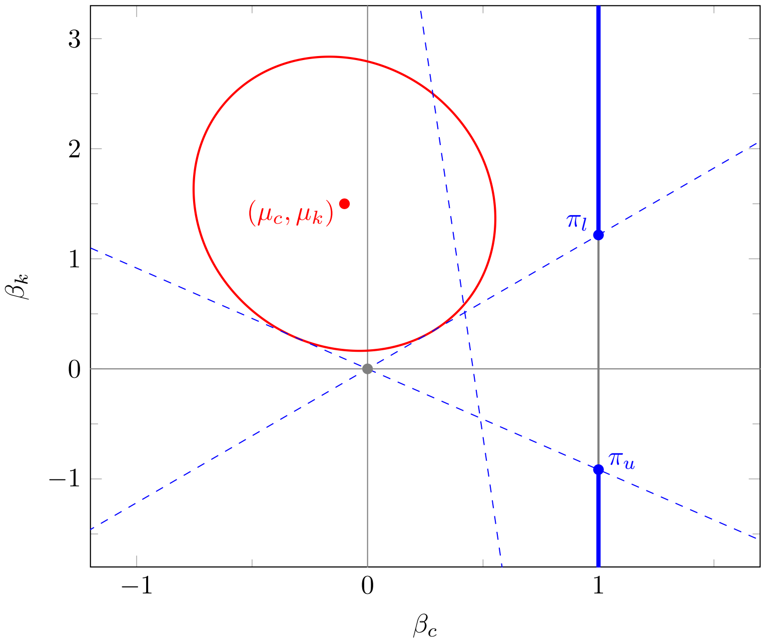


Figure 4: Construction of Fieller confidence set when $\hat{\beta}_c$ is not significant.

A third type of confidence set occurs when the confidence ellipse surrounds the origin. In that case, it is impossible to find a ray through the origin which is tangent to the confidence ellipse. Thus, the Fieller bounds do not exist and the confidence set contains the entire set of real numbers. Note that it is a necessary, but not a sufficient condition that both estimators are individually insignificant at the significance level α for the confidence set to be unbounded. If the bounds exist, but $\hat{\beta}_c$ is not significant, the confidence set is an exclusive interval. And finally, if $\hat{\beta}_c$ is significant, the confidence set is an inclusive interval. In previous simulation studies (Hole 2007; Gatta, Marcucci, and Scaccia 2015), iterations with exclusive or unbounded intervals are simply discarded. Here, in order not to insert an artificial bias, exclusive or unbounded intervals are not discarded.[18]

Hinkley method

Hinkley (1969) provides the exact formula for the probability density function of a normal ratio distribution (see appendix B.1). Ideally, one would integrate this function to obtain the cumulative distribution, find its inverse, and read off the quantiles of π to construct a confidence set. Unfortunately, there exists no closed-form solution for its cumulative distribution function, and due to the large number of parameters, tables with pre-calculated values are not available. However, it should be a fairly simple exercise to find a sufficiently precise numerical integration method, which is usually part of any statistical software package, and to apply it on the formula given by Hinkley. In the simulation, the `uniroot` function, which is part of the R stats package (R Core Team 2023), is used to find the inverse of the cumulative density function by means of numerical integration. To the best of the author's knowledge, there exists no publication to this date in which this method has been applied or tested.

The quantiles which limit the confidence set can be chosen freely, as long as the probability mass between them equals the significance level. If μ_k/μ_c is chosen as its mid-point, the Hinkley method is equivalent to computing the Fieller confidence set. One important difference is that the Hinkley method always finds confidence bounds, even when the Fieller method does not. However, taking into account the implications of Dufour (1997), this may not be desirable. Since the normal ratio distribution is locally almost unidentified, a valid method to construct confidence intervals should yield unbounded intervals with non-zero probability. Analogue to the Fieller method, the accuracy of the Hinkley method depends on whether the assumption of joint normality of $\hat{\beta}_c$ and $\hat{\beta}_k$ is satisfied. When the sample size is too small, the asymptotic properties may not be given, rendering this method unsuitable. Additionally, the heavy tails of the normal ratio distribution may pose a challenge for the numerical integration method.

[18] Since Fieller confidence intervals are always bounded when the denominator is significant, this decision makes no difference when sufficiently large cut-off values for the t-value of $\hat{\beta}_c$ are applied.

Bootstrap

The widely applied bootstrapping technique has been introduced by Efron (1979). It is often used to obtain robust results, or any results if no other method is feasible. In this paper, I test the so-called “naïve” bootstrap. It works as follows: First, resample B times from the original sample (in this case from each simulated sample), and perform the regression with each of the B artificial samples. The regressions yield B simulated values $\tilde{\mu}_c$ and $\tilde{\mu}_k$. Then, take their ratios to obtain the corresponding quantiles of $\tilde{\pi}$, which can then be used to construct confidence intervals.

One advantage of the bootstrap is that it avoids any assumptions on the joint distribution of $\hat{\beta}_c$ and $\hat{\beta}_k$. Hence, it is robust in case the estimators have not converged to their normal distributions, or in case they are not jointly distributed. However, a known drawback of the bootstrap is that it is biased in small samples. One of the robustness checks focuses on whether this poses a problem for WTP estimation. While there exist several versions of the bootstrap which correct for the small sample bias, I use the naïve bootstrap as a baseline.

Parametric bootstrap

In Krinsky and Robb (1986) and Krinsky and Robb (1990), the authors propose another variant of the bootstrap.[19] Assuming that $\hat{\beta}_c$ and $\hat{\beta}_k$ are jointly and normally distributed, their multivariate distribution can be used to simulate random values. As in the naïve bootstrap, all pairs of $\tilde{\beta}_c$ and $\tilde{\beta}_k$ are divided to obtain the quantiles of $\tilde{\pi}$, which yield the confidence interval for $\hat{\pi}$.

In contrast to the naïve bootstrap, it requires only one regression, which is computationally far less demanding. However, given the speed of computers nowadays and the possibility of cloud computing, this should not be a deciding factor any longer. More importantly, it requires an additional assumption, which may be violated in some cases. Asymptotically, the two bootstrapping methods should be equivalent, since the naïve bootstrap is asymptotically unbiased and the distribu-

[19] This method is sometimes called “Krinsky-Robb method”. To distinguish it from the naïve bootstrap, I use the more expressive term “parametric bootstrap” instead.

tions of $\hat{\beta}_c$ and $\hat{\beta}_k$ are asymptotically normal. Ultimately, the speed at which both converge determines which bootstrapping method is more accurate.

Delta method

The delta method is a flexible approach to approximate any statistical estimator by a normal distribution, and it is known at least since the 19th century (Portnoy and Ver Hoef 2013). Hole (2007) uses a simulation of discrete choice experiments to test the delta method's accuracy in providing confidence intervals, and he finds that it is more accurate than the other methods when "the data is well-conditioned". It uses a first-order Taylor approximation around one value of the distribution function, in this case around the ratio of the means of $\hat{\beta}_c$ and $\hat{\beta}_k$. The crucial assumption is that the estimator of interest is asymptotically normally distributed, which is not fulfilled in the case of a normal ratio distribution. Hence, the objective here is to test whether the most common method of calculating standard errors of partial effects (Dowd, Greene, and Norton 2013) and willingness to pay estimates (Mott, Chami, and Tervonen 2020) is suitable in the given context. The variance of $\hat{\pi}$ as computed in the delta method is given by:

$$\mathrm{Var}[\hat{\pi}] = \begin{pmatrix} {}^{\delta\hat{\pi}}/_{\delta\hat{\beta}_c} \\ {}^{\delta\hat{\pi}}/_{\delta\hat{\beta}_k} \end{pmatrix}' \boldsymbol{\Sigma}^{-1} \begin{pmatrix} {}^{\delta\hat{\pi}}/_{\delta\hat{\beta}_c} \\ {}^{\delta\hat{\pi}}/_{\delta\hat{\beta}_k} \end{pmatrix}. \tag{4}$$

Note that, as stated in the previous section, the moments of $\hat{\pi}$ are not defined. Hence, the ratio of the means of $\hat{\beta}_c$ and $\hat{\beta}_k$ cannot be equal to the (theoretically infinite) mean of $\hat{\pi}$. It is also not equal to the median of $\hat{\pi}$, such that one needs to be careful when interpreting the results of the delta method. The mean and standard deviation stemming from the delta method can be used to find approximate confidence intervals, but aside from that the calculated moments have no deeper meaning.

Figure 5, which is adapted from Hirschberg and Lye (2010), illustrates how delta confidence intervals are constructed. The hypothetical values are the same as in figure 3, where the t-values of both estimators are equal to 4. Again, the construction starts by projecting the point (μ_c, μ_k) onto the vertical line at $\beta_c = 1$ using the ray through the origin, and thus computing the ratio of means. This determines π_M, the mid-point of the interval for $\hat{\pi}$. Next, the tangents of the confidence ellipse which are parallel to the ray through the origin are drawn, here depicted by the dotted lines. The intersections of these tangents with the vertical line at $\beta_c = \mu_c$ are marked with a dot. Finally, these points are projected onto the vertical line at $\beta_c = 1$ to find the lower and upper bound π_L and π_U of the interval.

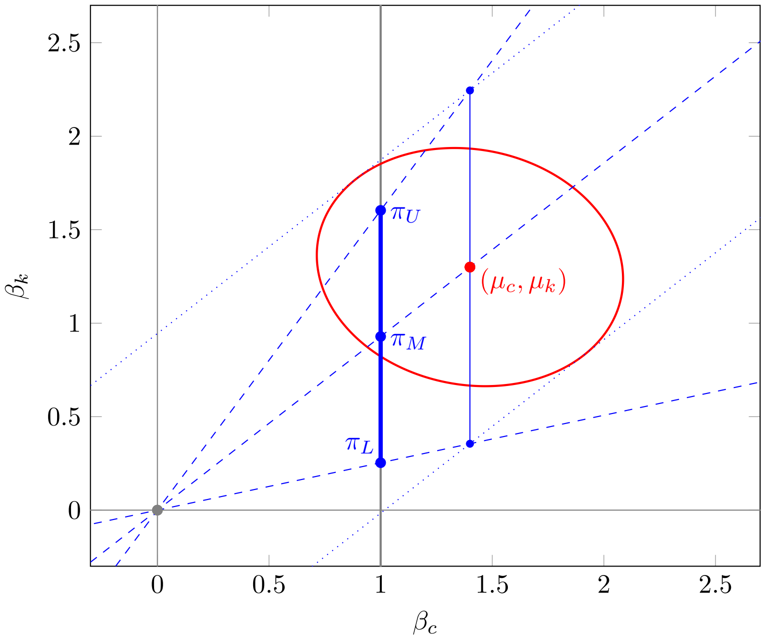


Figure 5: Construction of delta confidence interval.

One weakness of the delta method can be seen right away. Since the tangents are equidistant from the point (μ_c, μ_k), the lower and upper bound must be equidistant from π_M by construction. Accordingly, the delta method does not capture the skewness of the normal ratio distribution. Thus, the intervals generally do not cover very large values (in absolute terms), even though a substantial part of the probability mass is located in the heavy tails of the normal ratio distribution (see figure 2). Additionally, the probability mass between π_L and π_M is generally not equal to that between π_M and π_U. The area between the dashed lines through π_L and π_M represents the set of parameter combinations (β_c, β_k) which fall into the interval between π_L and π_M (analogously, the same is true for π_M and π_U). Observe how the dashed line through (μ_c, μ_k) divides the confidence ellipse in two halves. In this example it can be seen clearly that the entire lower right half of the confidence ellipse is contained in the area between the dashed lines through π_L and π_M, while only a part of the upper left half is contained in the area between the dashed lines through π_M and π_U. Consequently, the hypothesis to be tested is that the confidence interval for $\hat{\pi}$ given by the delta method generally overestimates the likelihood of values close to 0, and underestimates the likelihood of extreme values.

Hirschberg and Lye (2010) provide an illustration of how the delta method and Fieller method relate to each other. They find that when $\mu_c, \mu_k > 0$ or $\mu_c, \mu_k < 0$, the confidence intervals given by the two methods are more similar for a positive correlation between $\hat{\beta}_c$ and $\hat{\beta}_k$ than for a negative correlation. The opposite is true when the signs of μ_c and μ_k differ. The empirical relevance of this finding is tested in the simulation by increasing the expected correlation between $\hat{\beta}_c$ and $\hat{\beta}_k$ (in absolute terms).

3.2 Interval mid-point

It is *a priori* not clear where the confidence interval for $\hat{\pi}$ should be located. Considering the cumulative distribution of an estimator, it would seem like the most natural choice to centre the interval around the median, such that the probability masses between 0 and the lower limit, and between the upper limit and 1 are equal. However, given the properties of the normal ratio distribution described in section 2.2, it is unclear what its central measure is. Its mean is not defined and its mode is not necessarily unique. Further, the median of $\hat{\pi}$ is generally not equal to the ratio of the means of $\hat{\beta}_c$ and $\hat{\beta}_k$, and a closed-form solution to compute the median of $\hat{\pi}$ does not exist.

Another aspect to consider when choosing the central point is whether one wants to respect the confidence intervals for the underlying distributions $\hat{\beta}_c$ and $\hat{\beta}_k$. To see why this may be desirable, consider an example: Given is a multivariate distribution $\hat{\boldsymbol{\beta}}$ where $\hat{\beta}_c$ is not significantly different from 0 at a given significance level α, while $\hat{\beta}_k$ is significantly different from 0. After taking a random draw from the multivariate distribution and calculating the ratio $\tilde{\beta}_k/\tilde{\beta}_c$ of the simulated values, we observe an extremely large result for $\tilde{\pi}$ (in absolute terms). There are three ways this could have occurred; either the value of $\tilde{\beta}_k$ is extremely large, or

$\tilde{\beta}_c$ is very close to 0, or both. The first and the third case represent outliers of the multivariate distribution $\hat{\boldsymbol{\beta}}$, which should not be contained in the confidence interval for $\hat{\pi}$. In the second case, $\tilde{\beta}_c$ and $\tilde{\beta}_k$ may lie within the confidence set of $\hat{\boldsymbol{\beta}}$, hence $\tilde{\pi}$ should be contained in the confidence interval for $\hat{\pi}$. In order to decide whether or not to include $\tilde{\pi}$ in the confidence interval, one may want to consider the likelihood of each case occurring. As the second case is relatively likely (values close to 0 are part of the confidence interval for $\hat{\beta}_c$), extreme values should be included in the confidence interval. A confidence interval centred around the median of $\hat{\pi}$, however, never includes extreme values, such that it disregards a large set of points $(\tilde{\beta}_c, \tilde{\beta}_k)$ which lie within the confidence ellipse for $\hat{\boldsymbol{\beta}}$.

Two of the methods presented here, the delta and Fieller method, are not centred around the median, but around the ratio of means. The other three methods are flexible in that regard; they may be centred around the median, but one can also choose to centre them around any other value. For all methods except the bootstrap, it is assumed that the estimators $\hat{\beta}_c$ and $\hat{\beta}_k$ are normally distributed. As the Fieller method gives the exact confidence intervals for $\hat{\pi}$ when $\hat{\beta}_c$ and $\hat{\beta}_k$ are normally distributed, it is needless to check whether other methods requiring this assumption perform better. Instead, the delta and Fieller method are compared with the other methods centred around the median. To conclude this section, table 1 gives an overview of the five methods and their characteristics.

4 Simulation

The goal of the simulation is to find out which methods yield accurate confidence intervals for π under different circumstances. Since WTP is a monotonic transformation of π, the confidence interval for π finally determines the confidence interval for the WTP measure. To resemble a real-world application, a large number of data sets following equation 1 is generated, using some hypothetical parameter values. The five methods described in section 3 are then applied to obtain confidence intervals for π and to calculate the corresponding coverage rates, i.e. the shares of confidence intervals which contain the true value of π. Following a frequentist interpretation, the coverage rate of all confidence intervals should converge to their respective significance levels as the number of simulated data sets increases.

Method	Approach	Assumptions	Mid-points
Fieller	Compute interval bounds	$\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	μ_k/μ_c
Hinkley	Numerically integrate density function	$\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	median
Bootstrap	Resample from observations	—	median
Parametric bootstrap	Resample from estimator distribution	$\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	median
Delta	Approximate by normal distribution	$\hat{\pi} \xrightarrow{d} \mathcal{N}$	μ_k/μ_c

Table 1: Overview over the five methods to construct confidence intervals.

Continuing the example in section 2.1, the consumption variable c follows a log-normal distribution, while good k, the additional variable z, and the error term ε are normally distributed. Equation 5 shows how the true variance-covariance matrix $\boldsymbol{\Delta}_{ck}$ of the variables c and k relates to the expected variance-covariance matrix $\boldsymbol{\Sigma}_{ck}$ of the estimators $\hat{\beta}_c$ and $\hat{\beta}_k$. By setting the values of $\boldsymbol{\Delta}_{ck}$ accordingly, $\boldsymbol{\Sigma}_{ck}$ can be determined by the researcher:[20]

$$\boldsymbol{\Delta}_{ck} = \begin{pmatrix} \delta_c^2 & \delta_{ck} \\ \delta_{ck} & \delta_k^2 \end{pmatrix} \approx \begin{pmatrix} \dfrac{\sigma_e^2}{N\sigma_c^2\,(1-\rho_{ck}^2)} & \dfrac{-\rho_{ck}\,\sigma_e^2}{N\sigma_c^2\,\sigma_k^2\,(1-\rho_{ck}^2)} \\ \dfrac{-\rho_{ck}\,\sigma_e^2}{N\sigma_c^2\,\sigma_k^2\,(1-\rho_{ck}^2)} & \dfrac{\sigma_e^2}{N\sigma_k^2\,(1-\rho_{ck}^2)} \end{pmatrix}. \tag{5}$$

where σ_e^2 denotes the variance of the error term, N denotes the sample size in each data set, and ρ_{ck} denotes the correlation between $\hat{\beta}_c$ and $\hat{\beta}_k$.

This also determines the expected t-values τ_c and τ_k of the estimators; since $t_c = \mu_c/\sigma_c$ and $\mathrm{E}[\mu_c] = \beta_c$, it follows that $\mathrm{E}[t_c] = \tau_c = \beta_c/\sigma_c$ as long as the estimator $\hat{\beta}_c$ is unbiased (the same holds for τ_k, respectively). In the baseline case, τ_c and τ_k are set to 3. With an expected t-value of 3, about half the simulated data

[20] Appendix B.3 shows the derivation of this equation. Note that it only holds approximately.

sets are expected to satisfy the threshold of $t_c \geq 3$ proposed by Geary (1930). A table describing the remaining variables and their values can be found in appendix A.1. By choosing different significance levels $\alpha = \{0.01, 0.05, 0.1, 0.2, 0.4\}$ for the confidence intervals for $\hat{\pi}$, the extent to which each method captures the shape and skewness of the normal ratio distribution can be checked. Besides the baseline configuration, the robustness of each method is tested in three scenarios: (i) when the statistical power to reject $\hat{\beta}_c = 0$ and $\hat{\beta}_k = 0$ is low, (ii) when the sample size is small, and (iii) when the estimators for $\hat{\beta}_c$ and $\hat{\beta}_k$ are strongly negatively correlated.

Another test in the simulation deals with the behaviour of researchers. When researchers apply the SWB method on a real-world example to estimate WTP, they usually discard insignificant results. It is tested whether this induces a bias on the accuracy of confidence intervals given by each method. Further, it is unclear which cut-off value should be used to obtain sufficiently accurate confidence intervals. To find a guideline, different cut-off values are applied on the t-value of $\hat{\beta}_c$. First, the performance of each method is examined without a cut-off on t_c. Second, the most commonly used 5% significance level is applied, which corresponds to a cut-off value of about 1.96. And third, a cut-off value of 3 is used, as proposed by Geary (1930). An ideal method to find confidence intervals should always be accurate for any given cut-off value, and a second-best method should always be accurate for a known set of cut-off values.

The simulation mechanism works as follows: First, draw N random values from the distributions of c, k, and ε. Depending on the underlying utility function, transform the values for c and k (in this case by taking the logarithm of c to obtain a log-linear utility function). Second, multiply the transformed values for c and k with β_c and β_k, add the error term and obtain the life satisfaction LS.[21] Third, using c, k, and LS, perform the different methods to calculate the confidence intervals for π. Perform the regression once for each data set and then apply the Fieller, Hinkley, parametric bootstrapping, and delta method. For the naïve bootstrap, resample many times from the simulated samples, perform the

[21] Typically, life satisfaction is reported on a scale of integers, e.g. from 0 to 10. However, I choose not to transform or restrict the life satisfaction variable, as this may bias the expected t-values τ_c and τ_k. Hence, LS can take any value of the set of real numbers in the simulation.

regression for each resampled sample, and store the distribution of values to construct confidence intervals. Fourth, apply the cut-off values and check how many confidence intervals contain the true value β_k/β_k for each method, significance level, and cut-off value.

5 Results

5.1 Baseline values

Figure 6 shows the distribution of t-values over all 100 000 simulated data sets in the baseline case. We see that both t_c and t_k are concentrated between 2 and 4; this means that most of the times, the estimators $\hat{\beta}_c$ and $\hat{\beta}_k$ are significant at the 5% significance level, but not highly significant. Moreover, no simulated data set yields t-values which are significantly smaller than 0 at the 5% significance level.[22] The dot in the middle of the data cloud shows the means of both t-values, and as expected, it is located close to the point $(3, 3)$. Around the means, we find the quantile ellipse at the 95% level, here represented by the dotted line. We observe that it is slightly stretched in the north-west to south-east direction, indicating a small negative correlation between the estimators. Note, however, that this graph does not resemble figures 3 and 5. The quantile ellipse depicted here contains 95% of the values, while the confidence ellipse depicted in the other graphs represents the distribution of the sample means. This graph does not allow for the construction of confidence intervals for π either, as the values shown are the t-values of the estimators $\hat{\beta}_c$ and $\hat{\beta}_k$, and not the distributions of the estimators themselves.

Table 2 shows the frequency with which the confidence intervals include the true value of $\pi = 0.06$ for a given method, significance level, and cut-off value in the first simulation, using the baseline values. This frequency is also known as the coverage rate. The table can be understood as follows: The closer a value in the table is to the targeted coverage rate given in the top row, the more accurate a method is. When the value is too large, the confidence intervals are too conservative, and

[22] While this cannot be deduced from the graph, since the leftmost hexagons may contain the critical values for a one-sided test of $\beta_c < 0$, it can be seen in the data.

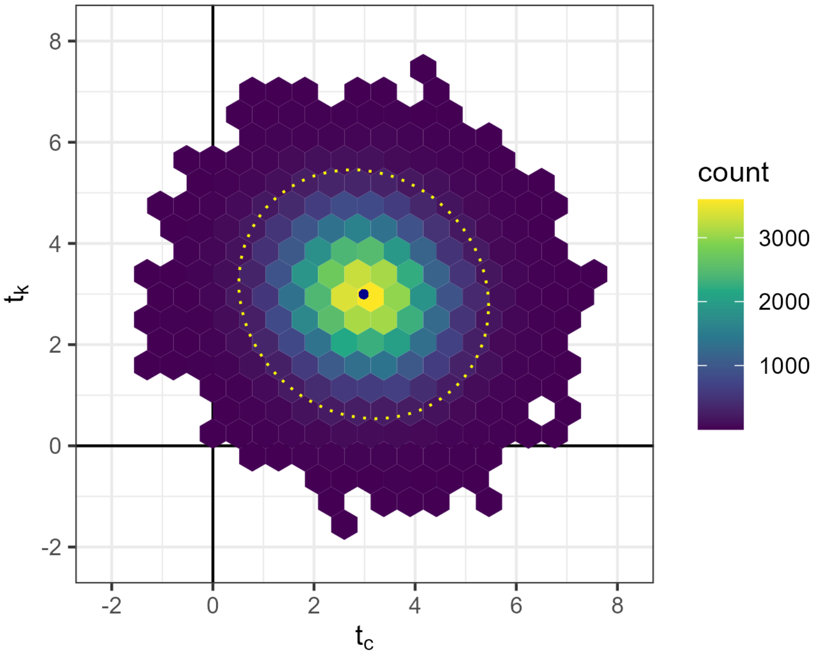


Figure 6: Distribution of simulated t-values in the baseline case.

when it is too small, the confidence intervals are too narrow. Note that, contrary to other contributions, it is not required here that all methods yield confidence intervals which lie above or below the true value equally often for a method to be accurate. This test would only be meaningful if all confidence interval were centred around the median.[23]

We observe that the confidence intervals calculated using the Fieller, Hinkley, bootstrapping, and parametric bootstrapping method are too conservative for every significance level when no cut-off on the t-value is applied. Out of these four, the Fieller method stands out as a slightly more accurate method, specifically for lower significance levels. After applying the "conventional" cut-off value of 1.96, the four methods become slightly more accurate for higher significance levels, but slightly less accurate for the 60% confidence interval. Notably, the Fieller method also becomes more inaccurate for the 80% confidence interval and does not stand

[23] To see why such a test would be meaningless, imagine applying it on the Fieller method. Every time the interval was exclusive, the confidence set would either lie both above and below the true value at the same time, or contain it.

	$\mathbf{1-\alpha}$	**0.99**	**0.95**	**0.9**	**0.8**	**0.6**
no cut-off	Fieller	0.9945	0.9708	0.9296	0.8296	0.6125
	Hinkley	0.9950	0.9733	0.9396	0.8535	0.6410
	Bootstrap	0.9943	0.9726	0.9379	0.8509	0.6394
	Param. bootstrap	0.9944	0.9733	0.9389	0.8521	0.6404
	Delta	0.9710	0.9330	0.9013	0.8489	0.6909
$t_c \geq 1.96$	Fieller	0.9941	0.9655	0.9261	0.8399	0.6430
	Hinkley	0.9941	0.9684	0.9302	0.8453	0.6481
	Bootstrap	0.9933	0.9675	0.9287	0.8432	0.6467
	Param. bootstrap	0.9934	0.9683	0.9295	0.8438	0.6473
	Delta	0.9656	0.9206	0.8832	0.8214	0.6630
$t_c \geq 3$	Fieller	0.9900	0.9488	0.8976	0.7984	0.5912
	Hinkley	0.9899	0.9488	0.8975	0.7984	0.5912
	Bootstrap	0.9886	0.9478	0.8960	0.7961	0.5906
	Param. bootstrap	0.9888	0.9489	0.8967	0.7967	0.5905
	Delta	0.9423	0.8697	0.8112	0.7180	0.5568

Table 2: Coverage rates of confidence intervals in the baseline case.

out any longer. Out of 100 000 simulated data sets, 84 341 data sets remain after applying the cut-off value of 1.96. After applying the cut-off value of 3, only 49 403 simulated data sets remain. Here, the four methods become more accurate and yield very similar coverage rates. We even observe two subgroups: The Fieller and the Hinkley method yield almost identical coverage rates, and so do the two bootstrapping methods. However, the intervals given by the four methods are now slightly too narrow, especially for lower significance levels.

Interestingly, the coverage rates obtained by the delta method differ strongly from those of the other methods. When no cut-off is applied, the delta method appears to yield highly accurate 90% confidence intervals. For higher significance levels, though, the intervals are too narrow, and for lower significance levels too conservative. This pattern is not found in any of the other methods. It seems that the delta method fails to incorporate the shape of the normal ratio distribution. When applying the cut-off values, the coverage rates drop everywhere, such that the intervals are too narrow at every significance level. While the accuracy increases for all other methods when the cut-off value of 3 is applied, it sharply decreases for the delta method.

Some readers may wonder why no cut-off is applied on the t-values of $\hat{\beta}_k$, even though it is a wide-spread practice that results only get published when the coefficients of interest are significantly different from 0. First, it should be noted that there is no theoretically compelling reason to do so. Confidence intervals for the normal ratio distribution are intended to capture the possibility that the nominator and denominator are equal or close to 0. Hence, any truncation of the values for t_c and t_k induces a bias on the size and location of the confidence interval for π. Yet, this bias is accepted when truncating the values for t_c, as allowing for values of t_c close to 0 biases the confidence intervals even more strongly. Table 4 shows the impact on the coverage rates when a cut-off of 1.96 is applied on t_k. The estimated confidence intervals are far too conservative across all methods, significance levels, and cut-off values for t_c.[24]

5.2 Robustness Checks

In the first robustness check, the expected t-values τ_c and τ_k are set altered. Decreasing both to a value of 2 is expected to reduce the share of simulated data sets with significant estimators, mimicking a situation where a survey or experiment is conducted with less statistical power. Figure 7 shows the distribution of the t-values, which are concentrated much closer to the origin than in the baseline case. Insignificant results and significant results in the wrong direction are much more likely under this configuration. Table 5 contains the coverage rates of this scenario. Analogously to the baseline case, all except the delta method show very similar levels of accuracy. Only the Fieller method is slightly more accurate when no cut-off is applied. In contrast, however, the highest level of accuracy is reached when the cut-off value of 1.96 is applied, and the intervals become too narrow for $t_c \geq 3$. As in the baseline case, the delta method underperforms compared to all

[24] Since this is also true for all robustness checks, further coverage rates after applying a cut-off value on t_k are not shown here. Only the treatment where $\tau_c = \tau_k = 5$ yields similar coverage rates before and after applying the additional cut-off. However, since t_k is rarely smaller than the chosen cut-off values here, this is to be expected.

other methods. Only the 90% confidence interval appears to be fairly accurate when no cut-off is applied. With $t_c \geq 1.96$, the intervals are too narrow across all significance levels, and this even worsens with $t_c \geq 3$. Note that under this configuration, 51 206 simulated data sets pass the cut-off level of 1.96, and only 15 782 simulated data sets pass the one of 3.

When τ_c and τ_k are set to 5, the level of imprecision is expected to be much smaller, such that a large share of observation is expected to be significant at the chosen set of significance levels. In figure 8 we observe that no simulated data set yields t-values below 0, implying that the simulated point estimates for β_c and β_k are always positive. Only a small share of simulated data sets is cut off at the level 1.96, such that 99 873 data sets remain, and 97 440 data sets remain for a cut-off level of 3. Indeed, table 6 confirms that applying the cut-offs barely changes the results. The coverage rates of the Fieller, Hinkley, bootstrapping and parametric bootstrapping method are almost identical and highly accurate when applying no cut-off or a cut-off of 1.96, and only slightly too conservative when applying the stricter cut-off level of 3. Even though the delta method is more accurate compared to the baseline case across all significance and cut-off levels, it still lags behind the other methods. For higher significance levels, its intervals are too narrow, and for lower significance levels too conservative.

Second, the correlation between the estimators is changed to check the impact on the accuracy of each method, specifically the delta method. As described in Hirschberg and Lye (2010), the signs of μ_c and μ_k play a crucial role here. In the baseline case, both are positive, such that a stronger negative correlation between the estimators would lead to the intervals given by the delta and Fieller method being less similar. Accordingly, setting $\rho_{ck} = -0.4$ is expected to increase the difference between the two methods. Figure 9 illustrates how the t-values are correlated more strongly than in the baseline case, as the diagonal stretch of the confidence ellipse is more pronounced. The coverage rates for all methods, which can be found in table 7, resonate with the baseline case: All except the delta method yield very similar results, with the Fieller method being slightly more accurate when no cut-off is applied. The coverage rates of each method but the delta method are very similar to those in the baseline case. As expected, the delta method becomes less accurate under this scenario, with the exception of the 80%

and 60% confidence intervals when no cut-off, or a cut-off of 1.96 is applied.

In the final robustness check, the sample size is decreased to $N = 200$ in order to test whether the asymptotic properties assumed for each method hold. While a sample size of 200 might seem large for such a test, it is very small compared to the sample sizes one typically encounters in SWB data sets. Figure 10 indicates no major differences compared to the baseline case, and the coverage rates in table 8 are very similar. When $t_c \geq 3$ is applied, all methods yield slightly narrower intervals. Interestingly, the bootstrap seems to be slightly more inaccurate than the other methods (excluding the delta method), which may give an indication that a large sample size is more relevant for the asymptotic properties of the bootstrap than it is for the assumed normal distribution of the estimators.

6 Discussion

The results of the simulation indicate that all methods to construct confidence interval for a normal ratio distribution presented here, except for the delta method, are fairly accurate and robust. When no cut-off value is applied, the intervals given by these methods are typically too conservative, which is reassuring. The delta method, however, is very inconsistent in terms of accuracy. For different significance and cut-off levels, it can be too conservative or too narrow, and only sometimes fairly accurate. Even when imprecision is relatively small, it is less accurate than the other methods. Moreover, it is neither robust against low statistical power, nor against changes in the correlation between the estimators $\hat{\beta}_c$ and $\hat{\beta}_k$, nor is it more robust against small sample size bias than the other methods. Since all other options presented here are easily implementable and perform better than the delta method under almost every configuration, it seems logical that the delta method should not be used in this context.

Yet, it is not clear which of the four remaining methods is the best. Their coverage rates are very similar for most significance and cut-off levels, and under almost every configuration. Among the four methods, we can identify two subgroups: the normal ratio distribution-specific methods (Fieller and Hinkley), and the bootstrapping methods. In the first sub-groups, the Fieller method is slightly more accurate when no cut-off is applied, but otherwise both yield almost identical results. However, the Hinkley method uses the median as the interval mid-point and its intervals are never exclusive or unbounded, which some practitioners may prefer. Among the bootstrapping techniques, the parametric bootstrap seems to be more robust against small sample size. Yet, there are also reasons in favour of a non-parametric bootstrap: First, the sample size is typically known *a priori* in a real-world application, such that it can easily be controlled for. Second, there exist bias-corrected bootstrapping techniques, which may mitigate this potential issue.[25] And third, the coverage rates of both methods are almost identical under every other configuration. Hence, there is no clear winner in this simulation study, and all methods except for the delta method turn out to be viable options.

Another finding from this study is that the statistical power and the cut-off values play a role for the accuracy of a method. While the latter is chosen by the practitioner, the former is typically unobserved in real-world applications. Thus, if the practitioner has prior knowledge or a belief about the statistical power of a study, this should be taken into account. When the statistical power of a study is expected to be low, and the estimators of interest turn out to be significant, this may represent an outlier which should be treated carefully. Ideally, studies to estimate WTP should be designed with a higher level of statistical power than studies which are only aimed at finding significant partial effects. For instance, when it is expected that $\tau_c = \tau_k = 5$, all methods but the delta method are accurate, even without cut-off values. This would make the discussion about cut-off levels superfluous.

Finally, it should be noted that when reporting results of a WTP estimation, covariances between the estimators should be provided to allow for replication of the data. In Mott, Chami, and Tervonen (2020), the authors assume zero covariance between the estimators to measure the impact of failing to report the entire variance-covariance matrix, and they find a considerable effect. The results of the simulation presented here support their recommendation to report such information: Specifically when applying the delta method, the effect on the accuracy of confidence intervals may be sizeable.

[25] See Gatta, Marcucci, and Scaccia (2015) for a comparison of different bootstrapping techniques, many of which are presented in more detail in DiCiccio and Efron (1996).

7 Conclusion

In this paper, five methods to capture the imprecision when estimating willingness to pay using the subjective well-being method are discussed and compared in a simulation. For any additively separable utility function, the estimate of willingness to pay contains the ratio of two estimated coefficients. As these are asymptotically normally distributed, their ratio asymptotically follows a normal ratio distribution. This normal ratio distribution has several characteristics which impede the construction of accurate confidence intervals: It is heavy-tailed, its moments are infinite, it can be bimodal, there exists no closed-form solution, and it is locally almost unidentified. Five methods to construct confidence intervals are compared in this paper: the Fieller and delta method, the naïve and parametric bootstrap, and a proposed numerical integration of Hinkley's formula for the likelihood function of the normal ratio distribution. The selection of the first four methods is based on their popularity in applied work, and covers a spectrum of assumptions and approaches. The performance of each method in constructing confidence intervals for the ratio of estimators is compared in terms of accuracy in their coverage rates, and robustness against low statistical power, strong correlation between the estimators, and small sample size.

It is found that all but the delta method perform reasonably well under each configuration. With only minor differences in accuracy, each of them can be recommended for use. The delta method is not robust and fails to reflect the skewness of the normal ratio distribution. It is not recommended to use the delta method, unless the monetary coefficient is estimated with very high precision. Further, the statistical power of a survey aiming at estimating WTP should be increased compared to a survey aiming at identifying partial effects.

References

Amador, Francisco Javier, Rosa Marina González, and Juan de Dios Ortúzar (2005). "Preference Heterogeneity and Willingness to Pay for Travel Time Savings". In: *Transportation* 32.6, pp. 627–647.

Armstrong, Paula, Rodrigo Garrido, and Juan de Dios Ortúzar (2001). “Confidence intervals to bound the value of time”. In: *Transportation Research Part E: Logistics and Transportation Review* 37.2-3, pp. 143–161.

Bolduc, Denis, Lynda Khalaf, and Clément Yélou (2010). “Identification robust confidence set methods for inference on parameter ratios with application to discrete choice models”. In: *Journal of Econometrics* 157.2, pp. 317–327.

Brenig, Mattheus and Till Proeger (2018). “Putting a Price Tag on Security: Subjective Well-Being and Willingness-to-Pay for Crime Reduction in Europe”. In: *Journal of Happiness Studies* 19.1, pp. 145–166.

Carson, Richard T. and Mikołaj Czajkowski (2019). “A new baseline model for estimating willingness to pay from discrete choice models”. In: *Journal of Environmental Economics and Management* 95, pp. 57–61.

Clark, Andrew E. and Andrew J. Oswald (2002). “A simple statistical method for measuring how life events affect happiness”. In: *International Journal of Epidemiology* 31.6, pp. 1139–1144.

Daly, Andrew, Stephane Hess, and Kenneth Train (2011). “Assuring finite moments for willingness to pay in random coefficient models”. In: *Transportation* 39.1, pp. 19–31.

Decancq, Koen, Marc Fleurbaey, and Erik Schokkaert (2015). “Happiness, Equivalent Incomes and Respect for Individual Preferences”. In: *Economica* 82.5, pp. 1082–1106.

DiCiccio, Thomas J. and Bradley Efron (1996). “Bootstrap Confidence Intervals”. In: *Statistical Science* 11.3, pp. 189–228.

Dowd, Bryan E., William H. Greene, and Edward C. Norton (2013). “Computation of Standard Errors”. In: *Health Services Research* 49.2, pp. 731–750.

Dufour, Jean-Marie (1997). “Some Impossibility Theorems in Econometrics With Applications to Structural and Dynamic Models”. In: *Econometrica* 65.6, pp. 1365–1387.

Efron, Bradley (1979). “Bootstrap Methods: Another Look at the Jackknife”. In: *The Annals of Statistics* 7.1, pp. 1–26.

Ferrer-i-Carbonell, Ada and Paul Frijters (2004). “How Important is Methodology for the estimates of the determinants of Happiness?” In: *The Economic Journal* 114, pp. 641–659.

Fieller, Edgar C. (1940). “The Biological Standardization of Insulin”. In: *Journal of the Royal Statistical Society* 7.1, pp. 1–64.

— (1954). “Some Problems in Interval Estimation”. In: *Journal of the Royal Statistical Society* 16.2, pp. 175–185.

Gatta, Valerio, Edoardo Marcucci, and Luisa Scaccia (2015). “On finite sample performance of confidence intervals methods for willingness to pay measures”. In: *Transportation Research Part A: Policy and Practice* 82, pp. 169–192.

Geary, Robert Charles (1930). “The Frequency Distribution of the Quotient of Two Normal Variates”. In: *Journal of the Royal Statistical Society* 93.3, pp. 442–446.

Greene, William H. (2012). *Econometric Analysis*. Ed. by Sally Yagan and Donna Battista. 7th ed. Pearson Education Limited, p. 1188.

Hinkley, David Victor (1969). “On the Ratio of Two Correlated Normal Random Variables”. In: *Biometrika* 56.3, pp. 635–639.

Hirschberg, Joseph Gerald and Jenny Ngaire Lye (2010). “A Geometric Comparison of the Delta and Fieller Confidence Intervals”. In: *The American Statistician* 64.3, pp. 234–241.

Hole, Arne Risa (2007). “A Comparison of Approaches to Estimating Confidence Intervals for Willingness to Pay Measures”. In: *Health Economics* 16.8, pp. 827–840.

Krinsky, Itzhak and A. Leslie Robb (1986). “On Approximating the Statistical Properties of Elasticities”. In: *The Review of Economics and Statistics* 68.4, pp. 715–719.

— (1990). “On Approximating the Statistical Properties of Elasticities: A Correction”. In: *The Review of Economics and Statistics* 72.1, pp. 189–190.

Luechinger, Simon (2009). “Valuing Air Quality Using the Life Satisfaction Approach”. In: *The Economic Journal* 119, pp. 482–515.

Marsaglia, George (1965). “Ratios of normal variables and ratios of sums of uniform variables”. In: *Journal of the American Statistical Association* 60.309, pp. 193–204.

Mott, David J., Nour Chami, and Tommi Tervonen (2020). “Reporting Quality of Marginal Rates of Substitution in Discrete Choice Experiments That Elicit Patient Preferences”. In: *Value in Health* 23.8, pp. 979–984.

Oswald, Andrew J. and Nattavudh Powdthavee (2008). “Death, Happiness, and the Calculation of Compensatory Damages”. In: *The Journal of Legal Studies* 37.2, pp. 217–251.

Portnoy, Stephen and Jay M. Ver Hoef (2013). “Letter to the Editor”. In: *The American Statistician* 67.3, pp. 190–190.

Puth, Marie-Therese, Markus Neuhäuser, and Graeme D. Ruxton (2015). “On the variety of methods for calculating confidence intervals by bootstrapping”. In: *Journal of Animal Ecology* 84.4. Ed. by Dylan Childs, pp. 892–897.

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing. Vienna, Austria.

von Luxburg, Ulrike and Volker H. Franz (2009). “A Geometric Approach to Confidence Sets for Ratios: Fieller’s Theorem, Generalizations and Bootstrap”. In: *Statistica Sinica* 19, pp. 1095–1117.

Wang, Peng et al. (2020). “Penalized Fieller’s confidence interval for the ratio of bivariate normal means”. In: *Biometrics* 77.4, pp. 1355–1368.

A Tables and Figures

A.1 Default values

Description	**Symbol**	**Value**
Number of observations per data set	N	1000
Number of simulated data sets	–	100 000
Number of bootstrap samples	B	2000
True consumption coefficient	β_c	0.5
True coefficient for good k	β_k	0.03
True coefficient for variable z	β_z	0.03
Standard deviation of error term	σ_e	3
Expected t-value of $\hat{\beta}_c$	τ_c	3
Expected t-value of $\hat{\beta}_k$	τ_k	3
Expected correlation between $\hat{\beta}_c$ and $\hat{\beta}_k$	ρ_{ck}	-0.1
Expected correlation between c and z	ρ_{cz}	0.2
Mean value of c	γ_c	$e^{6.5}(\approx 665)$
Mean value of k	γ_k	70
Mean value of z	γ_z	10

Table 3: Overview over the variables and default values used in the simulation.

A.2 Robustness checks

	$1-\alpha$	0.99	0.95	0.9	0.8	0.6
no cut-off	Fieller	0.9991	0.9906	0.9614	0.8741	0.6597
	Hinkley	0.9997	0.9936	0.9733	0.9027	0.6947
	Bootstrap	0.9996	0.9928	0.9718	0.8999	0.6932
	Param. bootstrap	0.9996	0.9935	0.9724	0.9014	0.6941
	Delta	0.9954	0.9797	0.9614	0.9238	0.7649
$t_c \geq 1.96$	Fieller	0.9996	0.9890	0.9652	0.8993	0.7144
	Hinkley	0.9996	0.9924	0.9703	0.9062	0.7217
	Bootstrap	0.9995	0.9914	0.9690	0.9040	0.7203
	Param. bootstrap	0.9995	0.9922	0.9694	0.9049	0.7208
	Delta	0.9945	0.9758	0.9540	0.9090	0.7524
$t_c \geq 3$	Fieller	0.9993	0.9897	0.9664	0.9023	0.7070
	Hinkley	0.9993	0.9897	0.9665	0.9024	0.7072
	Bootstrap	0.9992	0.9886	0.9653	0.9000	0.7063
	Param. bootstrap	0.9992	0.9896	0.9655	0.9011	0.7062
	Delta	0.9905	0.9581	0.9204	0.8438	0.6695

Table 4: Coverage rates of confidence intervals with $t_k \geq 1.96$.

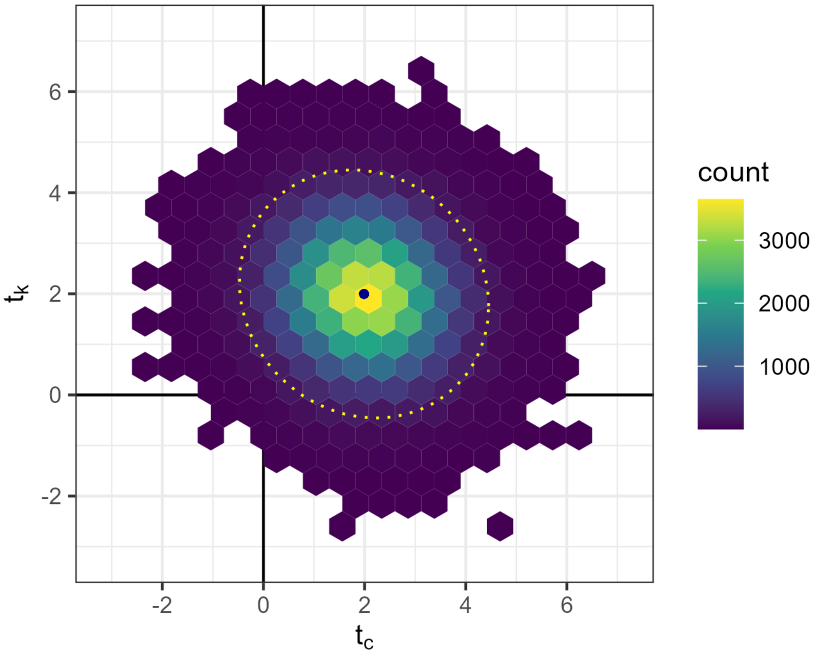


Figure 7: Distribution of simulated t-values with $\tau_c = \tau_k = 2$.

	$\mathbf{1-\alpha}$	**0.99**	**0.95**	**0.9**	**0.8**	**0.6**
no cut-off	Fieller	0.9892	0.9692	0.9430	0.8777	0.6849
	Hinkley	0.9951	0.9747	0.9478	0.8896	0.7277
	Bootstrap	0.9945	0.9740	0.9469	0.8879	0.7254
	Param. bootstrap	0.9946	0.9747	0.9474	0.8888	0.7271
	Delta	0.9646	0.9215	0.8871	0.8336	0.7354
$t_c \geq 1.96$	Fieller	0.9904	0.9505	0.9009	0.8041	0.5987
	Hinkley	0.9905	0.9506	0.8996	0.8009	0.5956
	Bootstrap	0.9892	0.9492	0.8979	0.7989	0.5947
	Param. bootstrap	0.9895	0.9507	0.8988	0.7999	0.5955
	Delta	0.9314	0.8528	0.7928	0.7027	0.5584
$t_c \geq 3$	Fieller	0.9704	0.8643	0.7544	0.5845	0.3410
	Hinkley	0.9703	0.8637	0.7532	0.5834	0.3404
	Bootstrap	0.9660	0.8609	0.7508	0.5805	0.3410
	Param. bootstrap	0.9672	0.8638	0.7505	0.5815	0.3410
	Delta	0.8268	0.6768	0.5761	0.4477	0.2877

Table 5: Coverage rates of confidence intervals with $\tau_c = \tau_k = 2$.

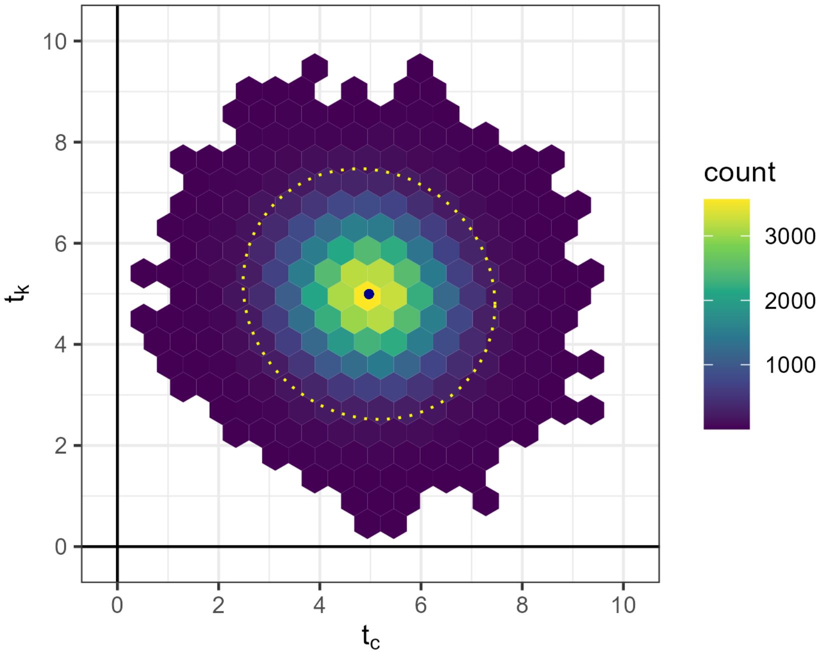


Figure 8: Distribution of simulated t-values with $\tau_c = \tau_k = 5$.

	$\mathbf{1-\alpha}$	**0.99**	**0.95**	**0.9**	**0.8**	**0.6**
no cut-off	Fieller	0.9918	0.9505	0.8994	0.8001	0.5972
	Hinkley	0.9928	0.9522	0.9004	0.8006	0.5974
	Bootstrap	0.9915	0.9504	0.8982	0.7986	0.5963
	Param. bootstrap	0.9914	0.9509	0.8996	0.7998	0.5973
	Delta	0.9795	0.9464	0.9149	0.8330	0.6167
$t_c \geq 1.96$	Fieller	0.9917	0.9504	0.9000	0.8009	0.5980
	Hinkley	0.9927	0.9521	0.9008	0.8013	0.5981
	Bootstrap	0.9914	0.9504	0.8986	0.7993	0.5971
	Param. bootstrap	0.9914	0.9509	0.9000	0.8005	0.5980
	Delta	0.9795	0.9464	0.9148	0.8328	0.6174
$t_c \geq 3$	Fieller	0.9928	0.9578	0.9102	0.8137	0.6098
	Hinkley	0.9931	0.9580	0.9103	0.8138	0.6099
	Bootstrap	0.9920	0.9564	0.9082	0.8117	0.6088
	Param. bootstrap	0.9919	0.9568	0.9097	0.8130	0.6097
	Delta	0.9790	0.9450	0.9128	0.8343	0.6272

Table 6: Coverage rates of confidence intervals with $\tau_c = \tau_k = 5$.

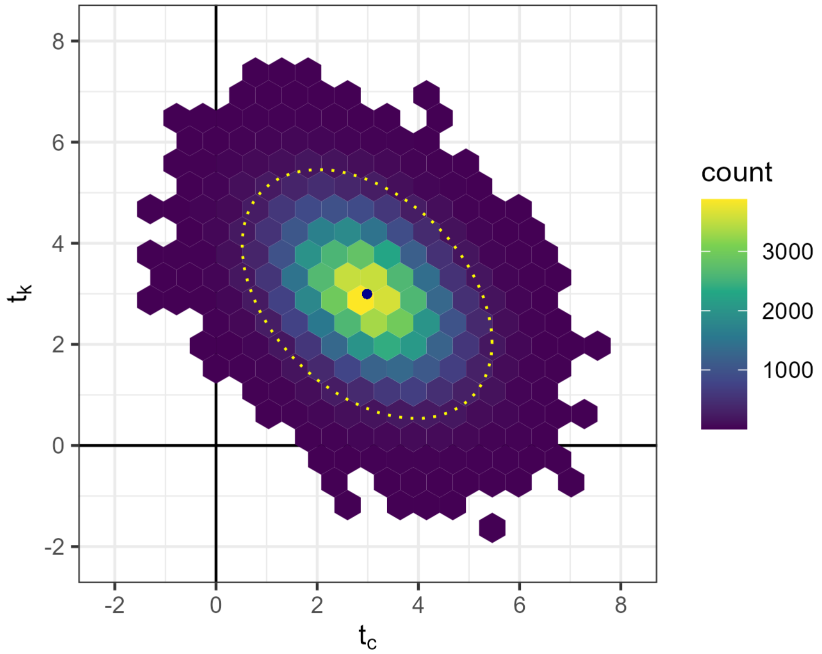


Figure 9: Distribution of simulated t-values with $\rho_{ck} = -0.4$.

	$\mathbf{1-\alpha}$	**0.99**	**0.95**	**0.9**	**0.8**	**0.6**
no cut-off	Fieller	0.9945	0.9734	0.9356	0.8349	0.6131
	Hinkley	0.9950	0.9748	0.9451	0.8639	0.6456
	Bootstrap	0.9944	0.9735	0.9434	0.8618	0.6445
	Param. bootstrap	0.9946	0.9741	0.9446	0.8624	0.6450
	Delta	0.9632	0.9244	0.8931	0.8424	0.7091
$t_c \geq 1.96$	Fieller	0.9940	0.9686	0.9325	0.8521	0.6598
	Hinkley	0.9940	0.9701	0.9361	0.8574	0.6663
	Bootstrap	0.9933	0.9686	0.9343	0.8560	0.6651
	Param. bootstrap	0.9935	0.9693	0.9356	0.8561	0.6654
	Delta	0.9564	0.9105	0.8734	0.8133	0.6797
$t_c \geq 3$	Fieller	0.9898	0.9498	0.8982	0.7971	0.5929
	Hinkley	0.9898	0.9498	0.8981	0.7969	0.5928
	Bootstrap	0.9886	0.9476	0.8958	0.7954	0.5919
	Param. bootstrap	0.9890	0.9484	0.8978	0.7960	0.5915
	Delta	0.9261	0.8491	0.7880	0.6916	0.5351

Table 7: Coverage rates of confidence intervals with $\rho_{ck} = -0.4$.

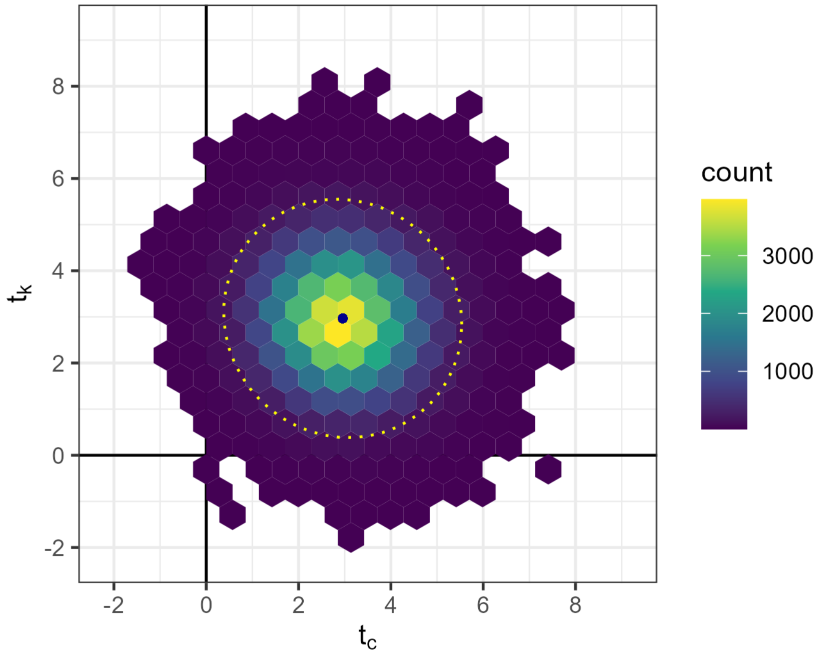


Figure 10: Distribution of simulated t-values with $N = 200$.

	$\mathbf{1-\alpha}$	**0.99**	**0.95**	**0.9**	**0.8**	**0.6**
no cut-off	Fieller	0.9934	0.9682	0.9268	0.8285	0.6156
	Hinkley	0.9943	0.9712	0.9362	0.8516	0.6463
	Bootstrap	0.9927	0.9674	0.9293	0.8433	0.6377
	Param. bootstrap	0.9936	0.9705	0.9349	0.8508	0.6450
	Delta	0.9695	0.9309	0.8993	0.8459	0.6906
$t_c \geq 1.96$	Fieller	0.9930	0.9620	0.9210	0.8340	0.6413
	Hinkley	0.9931	0.9652	0.9248	0.8386	0.6455
	Bootstrap	0.9912	0.9609	0.9185	0.8315	0.6381
	Param. bootstrap	0.9923	0.9644	0.9233	0.8380	0.6444
	Delta	0.9631	0.9168	0.8788	0.8149	0.6547
$t_c \geq 3$	Fieller	0.9881	0.9442	0.8899	0.7856	0.5812
	Hinkley	0.9881	0.9442	0.8898	0.7853	0.5810
	Bootstrap	0.9851	0.9389	0.8825	0.7782	0.5734
	Param. bootstrap	0.9868	0.9434	0.8884	0.7847	0.5802
	Delta	0.9371	0.8620	0.8027	0.7093	0.5461

Table 8: Coverage rates of confidence intervals with $N = 200$.

B Technical Appendix

B.1 Hinkley's Formula

Using the same notation as in the rest of this paper, Hinkley's formula for the density function f_π of the estimator for preference parameter π is the following:

$$
\begin{aligned}
f_\pi(p) = {} & \frac{b(p)\, d(p)}{\sqrt{2\pi}\, \sigma_c\, \sigma_k\, a^3(p)} \left[\Phi\left(\frac{b(p)}{a(p)\, r} \right) - \Phi\left(-\frac{b(p)}{a(p)\, r} \right) \right] \\
& + \frac{r}{\pi\, \sigma_c\, \sigma_k\, a^2(p)} \exp\left(-\frac{c}{2r^2} \right),
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
\text{with} \quad a(p) &= \left(\frac{p^2}{\sigma_k^2} - \frac{2\, \rho_{ck}\, p}{\sigma_c\, \sigma_k} + \frac{1}{\sigma_c^2} \right)^{\frac{1}{2}}, \\
b(p) &= \frac{\mu_k\, p}{\sigma_k^2} - \frac{\rho_{ck}\, (\mu_c\, p + \mu_k)}{\sigma_c\, \sigma_k} + \frac{\mu_c}{\sigma_c^2}, \\
c &= \frac{\mu_k^2}{\sigma_k^2} - \frac{2\, \rho_{ck}\, \mu_c\, \mu_k}{\sigma_c\, \sigma_k} + \frac{\mu_c^2}{\sigma_c^2}, \\
d(p) &= \exp\left(\frac{b^2(p) - c\, a^2(p)}{2\, r^2\, a^2(p)} \right), \\
r &= \sqrt{1 - \rho_{ck}^2},
\end{aligned}
$$

where Φ denotes the cumulative distribution function of the standard normal distribution, and ρ_{ck} denotes the correlation between $\hat{\beta}_c$ and $\hat{\beta}_k$. Note that p is used as the argument of the function instead of π since the mathematical constant π appears in the equation.

B.2 Fieller Confidence Intervals

Sticking to the notation in the rest of this paper, the formula to compute the Fieller bounds for the preference parameter π is given by:

$$
\begin{aligned}
(\pi_L, \pi_U) = \frac{1}{1-g} \cdot \Bigg[& \frac{\mu_k}{\mu_c} - \frac{g\,\sigma_{ck}}{\sigma_c^2} \\
& \mp \frac{t_{r,\alpha}}{\mu_c} \cdot \left(\sigma_k^2 - 2\,\sigma_{ck} \cdot \frac{\mu_k}{\mu_c} + \sigma_c^2 \cdot \frac{\mu_k^2}{\mu_c^2} - g \cdot \left(\sigma_k^2 - \frac{\sigma_{ck}^2}{\sigma_c^2} \right) \right)^{\frac{1}{2}} \Bigg],
\end{aligned}
\tag{7}
$$

$$
\text{where } \; g = \left(t_{r,\alpha} \cdot \frac{\sigma_c}{\mu_c} \right)^2,
$$

and $t_{r,\alpha}$ denotes the value of Student's t-distribution at confidence level α with r degrees of freedom. Note that despite their names, the upper bound π_U need not be larger than the lower bound π_L.

B.3 Variance-Covariance Matrix

In this paper, the statistical properties of the regression estimators of an equation with additively separable coefficients are examined. To do so, a large number of independent data sets is simulated and the regression is performed on each of them. In order to obtain t-values and correlations of the estimators which are roughly equal for each generated data set, it is necessary to determine the expected variance-covariance matrix of the estimators. To achieve this, the values of the independent variables must be simulated using a specific variance-covariance matrix, as will be shown below.

Consider a regression of the dependent variable Y on the independent variables c and k where the independent variables are additively separable. For each simulated data set, an independent sample of N individuals is drawn from the whole population. Let $\boldsymbol{X}$ be the design matrix $\left(\mathbf{1}\;\boldsymbol{c}\;\boldsymbol{k}\right)$, where $\boldsymbol{c}$ and $\boldsymbol{k}$ denote the vectors of variables c and k after being transformed according to the underlying utility function.[26] The regression equation is given by:

$$
\boldsymbol{Y} = \beta_0 + \beta_c \cdot \boldsymbol{c} + \beta_k \cdot \boldsymbol{k} + \boldsymbol{\epsilon} = \boldsymbol{\beta}'\boldsymbol{X} + \boldsymbol{\epsilon}, \tag{8}
$$

[26] For instance, if the underlying utility was of the Cobb-Douglas type, both vectors would be transformed with the logarithm.

where the error term $\boldsymbol{\epsilon}$ is distributed with $\mathrm{E}[\boldsymbol{\epsilon} \,|\, \boldsymbol{X}] = 0$ and $\mathrm{Var}[\boldsymbol{\epsilon} \,|\, \boldsymbol{X}] = \sigma_e^2$. Note that the last equation entails homoscedastic error terms.

As proved in, for instance, Greene (2012), the variance of the least squares estimator is given by:

$$\mathrm{Var}\big[\hat{\boldsymbol{\beta}}\big] = \begin{pmatrix} \mathrm{Var}\big[\hat{\beta}_0\big] & \mathrm{Cov}\big[\hat{\beta}_0, \hat{\beta}_c\big] & \mathrm{Cov}\big[\hat{\beta}_0, \hat{\beta}_k\big] \\ \mathrm{Cov}\big[\hat{\beta}_0, \hat{\beta}_c\big] & \mathrm{Var}\big[\hat{\beta}_c\big] & \mathrm{Cov}\big[\hat{\beta}_c, \hat{\beta}_k\big] \\ \mathrm{Cov}\big[\hat{\beta}_0, \hat{\beta}_k\big] & \mathrm{Cov}\big[\hat{\beta}_c, \hat{\beta}_k\big] & \mathrm{Var}\big[\hat{\beta}_k\big] \end{pmatrix} = \hat{\sigma}_e^2\,(\boldsymbol{X}'\boldsymbol{X})_{.}^{-1} \tag{9}$$

Defining $\boldsymbol{\Sigma}$ as the expected variance-covariance matrix of $\hat{\boldsymbol{\beta}}$, for which $\mathrm{Var}\big[\hat{\boldsymbol{\beta}}\big]$ is a consistent estimator, let us denote:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & \sigma_{0c} & \sigma_{0k} \\ \sigma_{0c} & \sigma_c^2 & \sigma_{ck} \\ \sigma_{0k} & \sigma_{ck} & \sigma_k^2 \end{pmatrix} = \sigma_e^2\,\mathrm{E}\big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\big], \tag{10}$$

where $\mathrm{E}\big[\hat{\sigma}_e^2\,(\boldsymbol{X}'\boldsymbol{X})^{-1}\big] = \mathrm{E}\big[\hat{\sigma}_e^2\big]\,\mathrm{E}\big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\big] = \sigma_e^2\,\mathrm{E}\big[(\boldsymbol{X}'\boldsymbol{X})^{-1}\big]$ holds only if $\mathrm{Var}[\boldsymbol{\epsilon} \,|\, \boldsymbol{X}] = \sigma_e^2$, hence the assumption of homoscedastic error terms. To see why this holds, recall that $x \perp y \Rightarrow \mathrm{E}\big[g(x)\,h(y)\big] = \mathrm{E}\big[g(x)\big]\,\mathrm{E}\big[h(y)\big]$.

Since the estimator $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$, it follows a distribution with means $\boldsymbol{\beta} = \big(\beta_0 \;\; \beta_c \;\; \beta_k\big)'$ and the variance-covariance matrix $\boldsymbol{\Sigma}$. The plan for the following steps is to reformulate $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ in equation 10, such that $\boldsymbol{\Sigma}$ is given by the moments of $\boldsymbol{X}$. Let us first define the corresponding vector of means $\boldsymbol{\gamma}$ and the variance-covariance matrix $\boldsymbol{\Delta}$ of $\boldsymbol{X}$:

$$\boldsymbol{\gamma} = \begin{pmatrix} 1 \\ \gamma_c \\ \gamma_k \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Delta} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & \delta_c^2 & \delta_{ck} \\ 0 & \delta_{ck} & \delta_k^2 \end{pmatrix}. \tag{11}$$

After writing out and multiplying $\boldsymbol{X}'\boldsymbol{X}$ in equation 10, we obtain:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_0^2 & \sigma_{0c} & \sigma_{0k} \\ \sigma_{0c} & \sigma_c^2 & \sigma_{ck} \\ \sigma_{0k} & \sigma_{ck} & \sigma_k^2 \end{pmatrix} = \sigma_e^2\,\mathrm{E}\left[\begin{pmatrix} N & \sum_i c_i & \sum_i k_i \\ \sum_i c_i & \sum_i c_i^2 & \sum_i c_i k_i \\ \sum_i k_i & \sum_i c_i k_i & \sum_i k_i^2 \end{pmatrix}^{-1}\right], \tag{12}$$

where the index i in $\sum_i$ runs from 1 to N.

Since the objective is to write $\boldsymbol{\Sigma}$ as a function of $\boldsymbol{\gamma}$ and $\boldsymbol{\Delta}$, we form the expectations over the expressions contained in $\boldsymbol{X}'\boldsymbol{X}$, such that we can replace the sums in equation 12 by the moments of $\boldsymbol{X}$. Note that Bessel's correction for the degrees of freedom of the sample variance and covariance is not applied here. Hence, the final result will only be asymptotically unbiased.[27] After rearranging the general formulas for the sample mean, variance, and covariance, and taking the expectations, we obtain:

$$\begin{gathered} \textstyle\sum_i c_i = N\gamma_c\,, \qquad\qquad \textstyle\sum_i k_i = N\gamma_k\,, \\ \textstyle\sum_i c_i^2 \approx N\big(\delta_c^2 + \gamma_c^2\big), \qquad\qquad \textstyle\sum_i k_i^2 \approx N\big(\delta_k^2 + \gamma_k^2\big), \\ \textstyle\sum_i c_i k_i \approx N(\delta_{ck} + \gamma_k\gamma_c). \end{gathered} \tag{13}$$

Inserting these moments into the right-hand side of equation 12 and factoring out N^{-1} yields:

$$\begin{pmatrix} \sigma_0^2 & \sigma_{0c} & \sigma_{0k} \\ \sigma_{0c} & \sigma_c^2 & \sigma_{ck} \\ \sigma_{0k} & \sigma_{ck} & \sigma_k^2 \end{pmatrix} \approx \frac{\sigma_e^2}{N} \begin{pmatrix} 1 & \gamma_c & \gamma_k \\ \gamma_c & \delta_c^2 + \gamma_c^2 & \delta_{ck} + \gamma_c\gamma_k \\ \gamma_k & \delta_{ck} + \gamma_c\gamma_k & \delta_k^2 + \gamma_k^2 \end{pmatrix}^{-1}. \tag{14}$$

Next, the matrix on the right-hand side needs to be inverted. We will refer to it as matrix $\boldsymbol{A}$, and since $\boldsymbol{A}^{-1} = \frac{1}{\det(\boldsymbol{A})} \cdot \operatorname{adj}(\boldsymbol{A})$, the next step will be to find its determinant and adjugate:

$$\begin{aligned} \det(\boldsymbol{A}) &= (\delta_c^2 + \gamma_c^2)(\delta_k^2 + \gamma_k^2) + 2\,\gamma_c\,\gamma_k\,(\delta_{ck} + \gamma_c\,\gamma_k) \\ &\quad - (\delta_{ck} + \gamma_c\,\gamma_k)^2 - \gamma_c^2\,(\delta_k^2 + \gamma_k^2) - \gamma_k^2\,(\delta_c^2 + \gamma_c^2) \\ &= \delta_c^2\,\gamma_k^2 + \delta_k^2\,\gamma_c^2 + \gamma_c^2\,\gamma_k^2 + \delta_c^2\,\delta_k^2 + 2\,\delta_{ck}\,\gamma_c\,\gamma_k + 2\,\gamma_c^2\,\gamma_k^2 \\ &\quad - \delta_{ck}^2 - 2\,\delta_{ck}\,\gamma_c\,\gamma_k - \gamma_c^2\,\gamma_k^2 - \delta_k^2\gamma_c^2 - \gamma_c^2\,\gamma_k^2 - \delta_c^2\gamma_k^2 - \gamma_c^2\,\gamma_k^2 \\ &= \delta_c^2\,\delta_k^2 - \delta_{ck}^2, \end{aligned} \tag{15}$$

[27] The exact method is certainly feasible, but results in much longer expressions of $\boldsymbol{\Delta}$. For the application presented here asymptotically unbiased results are sufficient.

$$
\begin{aligned}
\operatorname{adj}(\boldsymbol{A}) &= \begin{pmatrix} (\delta_c^2+\gamma_c^2)(\delta_k^2+\gamma_k^2)-(\delta_{ck}+\gamma_c\,\gamma_k)^2 & \gamma_k\,(\delta_{ck}+\gamma_c\,\gamma_k)-\gamma_c\,(\delta_k^2+\gamma_k^2) & \gamma_c\,(\delta_{ck}+\gamma_c\,\gamma_k)-\gamma_k\,(\delta_c^2+\gamma_c^2) \\ \gamma_k\,(\delta_{ck}+\gamma_c\,\gamma_k)-\gamma_c\,(\delta_k^2+\gamma_k^2) & (\delta_k^2+\gamma_k^2)-\gamma_k^2 & \gamma_c\,\gamma_k-(\delta_{ck}+\gamma_c\,\gamma_k) \\ \gamma_c\,(\delta_{ck}+\gamma_c\,\gamma_k)-\gamma_k\,(\delta_c^2+\gamma_c^2) & \gamma_c\,\gamma_k-(\delta_{ck}+\gamma_c\,\gamma_k) & (\delta_c^2+\gamma_c^2)-\gamma_c^2 \end{pmatrix} \\
&= \begin{pmatrix} \delta_c^2\delta_k^2+\delta_c^2\gamma_k^2+\delta_k^2\gamma_c^2-\delta_{ck}^2-2\gamma_c\gamma_k & \delta_{ck}\gamma_k-\delta_k^2\gamma_c & \delta_{ck}\gamma_c-\delta_c^2\gamma_k \\ \delta_{ck}\gamma_k-\delta_k^2\gamma_c & \delta_k^2 & -\delta_{ck} \\ \delta_{ck}\gamma_c-\delta_c^2\gamma_k & -\delta_{ck} & \delta_c^2 \end{pmatrix}.
\end{aligned} \tag{16}
$$

For clearness, let us define $\boldsymbol{A}^* := \operatorname{adj}(\boldsymbol{A})$. Replacing the inverse of $\boldsymbol{A}$ in the right-hand side of equation 14 by its determinant and adjugate yields:

$$
\begin{pmatrix} \sigma_0^2 & \sigma_{0c} & \sigma_{0k} \\ \sigma_{0c} & \sigma_c^2 & \sigma_{ck} \\ \sigma_{0k} & \sigma_{ck} & \sigma_k^2 \end{pmatrix} \approx \frac{\sigma_e^2}{N(\delta_c^2\,\delta_k^2-\delta_{ck}^2)} \begin{pmatrix} \boldsymbol{A}_{11}^* & \boldsymbol{A}_{12}^* & \boldsymbol{A}_{13}^* \\ \boldsymbol{A}_{21}^* & \delta_k^2 & -\delta_{ck} \\ \boldsymbol{A}_{31}^* & -\delta_{ck} & \delta_c^2 \end{pmatrix}. \tag{17}
$$

The equation above relates the expected variance-covariance matrix of the estimator $\hat{\boldsymbol{\beta}}$ to the variance-covariance matrix of the process which generates $\boldsymbol{X}$. Note that it is only defined when $\det(\boldsymbol{A}) \neq 0$, which is equivalent to the assumption of non-multicollinearity.

The parameters σ_c^2, σ_k^2, and σ_{ck}, which comprise the relevant part of the variance-covariance matrix $\boldsymbol{\Delta}$, should be freely selectable for the simulation. This can be achieved by choosing the parameters δ_c^2, δ_k^2, and δ_{ck} accordingly, while σ_e^2 and N are taken as fixed values. To compute the required values, we extract the corresponding matrix elements from the left- and right-hand side of equation 17:

$$
\begin{gathered}
\sigma_c^2 \approx \delta_k^2 \cdot \frac{\sigma_e^2}{N(\delta_c^2\,\delta_k^2-\delta_{ck}^2)}, \quad \sigma_k^2 \approx \delta_c^2 \cdot \frac{\sigma_e^2}{N(\delta_c^2\,\delta_k^2-\delta_{ck}^2)}, \\
\text{and} \quad \sigma_{ck} \approx -\delta_{ck} \cdot \frac{\sigma_e^2}{N(\delta_c^2\,\delta_k^2-\delta_{ck}^2)}.
\end{gathered} \tag{18}
$$

To get the formulas for δ_c^2, δ_k^2, and δ_{ck}, the equations above need to be rearranged. For this step, let us denote the expected correlation between $\hat{\beta}_c$ and $\hat{\beta}_k$ by

ρ_{ck}. After simplifying, we obtain:

$$\begin{aligned} \delta_c^2 \approx \frac{1}{\sigma_c^2} \cdot \frac{\sigma_e^2}{N(1-\rho_{ck}^2)}, \quad \delta_k^2 \approx \frac{1}{\sigma_k^2} \cdot \frac{\sigma_e^2}{N(1-\rho_{ck}^2)}, \\ \text{and} \quad \delta_{ck}^2 \approx -\frac{\rho_{ck}}{\sigma_c\,\sigma_k} \cdot \frac{\sigma_e^2}{N(1-\rho_{ck}^2)}. \end{aligned} \tag{19}$$

Plugging in these expressions into $\mathbf{\Delta}$ in equation 11 yields the desired variance-covariance matrix. Since the first row and column of $\mathbf{\Delta}$ only consist of zeros, the restricted matrix $\mathbf{\Delta}_{ck}$ (without zeros) is shown here:

$$\mathbf{\Delta}_{ck} = \begin{pmatrix} \delta_c^2 & \delta_{ck} \\ \delta_{ck} & \delta_k^2 \end{pmatrix} \approx \begin{pmatrix} \dfrac{\sigma_e^2}{N\sigma_c^2\,(1-\rho_{ck}^2)} & \dfrac{-\rho_{ck}\,\sigma_e^2}{N\sigma_c\,\sigma_k\,(1-\rho_{ck}^2)} \\ \dfrac{-\rho_{ck}\,\sigma_e^2}{N\sigma_c\,\sigma_k\,(1-\rho_{ck}^2)} & \dfrac{\sigma_e^2}{N\sigma_k^2\,(1-\rho_{ck}^2)} \end{pmatrix}. \tag{20}$$

The following assumptions are required to obtain the result given above:

1. The distribution of the error term is homoscedastic, i.e. $\text{Var}[\boldsymbol{\epsilon}|\boldsymbol{X}] = \sigma_e^2$.
2. The sample size N is large, such that the uncorrected sample moments are approximately unbiased in expectation.
3. There is no multicollinearity in the data, such that the inverse of $\boldsymbol{X}'\boldsymbol{X}$ exists.