DOI: 10.15439/2021F113 ISSN 2300-5963 ACSIS, Vol. 26

Maximum Simulated Likelihood: Don't Stop Believin'?

Christopher Schrey
Lipsiusstrasse 44, 04317 Leipzig, Germany
Email: christopher.schrey@outlook.de

***Abstract*—Unobserved heterogeneity may complicate model estimation in econometrics. To integrate out the effect of unobserved heterogeneity via maximum simulated likelihood (MSL) estimation, assumptions regarding the underlying distribution need to be made. Researchers seldomly discuss these assumptions. This raises the question, to what extent estimation results in the MSL-context are robust to potential distributional mismatch. This work-in-progress derives the research question from the literature. A simulation study is conducted that underpins the relevance of this matter, where results imply that mismatch may introduce significant bias. Intended future work to properly address and answer this question is defined and discussed.**

I. Introduction

UNOBSERVED heterogeneity may complicate model estimation in (health) econometrics. When modelling discrete choice, such as patients decisions regarding health insurance plans, unobserved heterogeneity may come in the form of private information regarding awareness of and attitudes towards an individuals health risks, resulting in self-selection into healthcare plans [1], [2], [3]. Similarly, unobserved heterogeneity may occur in every aspect of commerce, such as when consumers choose among alternatively-fuelled vehicles [4], among energy efficient refrigerators [5] or among modes of transportation [6], while their preferences (i.e., coefficients) are allowed to vary randomly among their choices. Generally speaking, unobserved heterogeneity may be considered whenever researchers cannot measure patient or consumer characteristics that determine preferences or equivalently, whenever features of the alternatives that are chosen from remain unrecorded [4].

Econometricians need to address unobserved heterogeneity, that materialises either though self-selection or varying preferences among alternatives. When researchers make an assumption regarding the distribution of these unobservable factors, their effect can be integrated out. This can be achieved, among others, by conducting *maximum simulated likelihood* (MSL) estimation. Simulation refers to the fact that integration over a density is but a form of averaging [7]. By averaging the likelihood function over a sufficiently large number of draws from the assumed distribution, MSL-estimation becomes feasible. Put differently, researchers need to make an as assumption, which distribution to choose, herein after referred to as *assumed distribution*, to approximate the *true distribution* which is unknown. While several distributional forms may be assumed, researchers most frequently assume that their unobserved heterogeneity follows a normal distribution [8], [9].

Accordingly, the researchers' assumption regarding the assumed distribution seems to be a critical one. MSL-estimation may be sensitive to poor approximations of the simulated probabilities [10] and even the wrong amount (i.e., too little) or quality of random draws may jeopardise the reliability of the results [11]. But what if researchers choose the assumed distribution incorrectly, resulting in *distributional mismatch*? The consequences of such distributional mismatch do not seem to be adequately addressed within the relevant literature. Many [12], [13], [3], [14], [9], [15], [6], state they assume unobserved heterogWas folgt eneity to follow a normal distribution without any justification or further elaboration. Some [1], [2], [16] provide little context regarding their choice.

[1] state that they obtained similar results with the uniform and beta as assumed distribution as with choosing the standard normal distribution. [2] justify their assumption regarding the standard normal distribution to handle location invariance. The readers are informed by [16] that distributional mismatch within their model *" (...) would potentially lead to biased parameter estimates"*.

As such, the research question of this piece of work-in-progress is to investigate bias in parameter estimates due to distributional mismatch between assumed and true distribution. Specifically, the mismatch will be limited to mismatch within the normal distribution, i.e., mean and standard deviation. Addressing this research problem will be beneficial to both econometricians conducting analysis with MSL-estimation as well as the research community interpreting the respective results. Further tools and methods to detect such biases and to potentially correct them may follow.

To this end, the MSL-method and its features will be introduced and a simulation study conducted, which aims at identifying bias due to distributional mismatch. The results of the simulation study will be discussed and interpreted. The bias introduced by the mismatch, i.e., through mismatch in mean and standard deviation, will be approximated by two equations, that will serve as basis for further discussion. Intended future work to properly address and answer this question is defined and discussed.

[0] This work was not supported by any organisation

II. Unobserved Heterogeneity

An example of unobserved heterogeneity that can be addressed with MSL-estimation is provided by [8]. Their example will serve as basis and will be enhanced to serve as a simulation study subsequently.

Let y_i be individual i's (with $i = 1 \dots N$) outcome of a sample with size N. Here, y_i depends on the observable variable x_i times its coefficient α, which is additionally be influenced by unobservable heterogeneity u_i with coefficient β and a standard normally distributed error term ε, such that [8]

$$y_i = \alpha x_i + \beta u_i + \varepsilon_i. \tag{1}$$

While the standard normally distributed error terms ε might similarly be viewed as a source of unobserved heterogeneity, their effect could simply be taken into account by OLS-regression or regular maximum likelihood estimation.

The density of *y* conditional on *u* is given by [8]

$$f(y_i|x_i, u_i) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(y_i - \alpha x_i - \beta u_i)^2}{2}\}. \tag{2}$$

Inference on *x* is based on the marginal density $f(y|u)$, which requires to integrate out the effect of *u* [8]. In the original case study by [8], the u's (true) distribution is the extreme value type 1 distribution. Here, for simplicity u's true distribution will be the normal distribution in different settings (regarding mean and standard deviation, as will be explained later). By drawing a number of *S* random draws from the distribution of *u*, their effect can be integrated out via simulation, hence the name maximum *simulated* likelihood. Given that the number of simulation draws S and sample size N both $S, N \to \infty$ while S increases faster than $\sqrt{N}$, such that $\sqrt{N}/S \to 0$, MSL is asymptotically normal, efficient and equivalent to maximum likelihood estimation [17], [7].[1] Here, MSL-estimation is achieved by drawing *S* random draws from the assumed distribution $\hat{u}$ of the unobserved heterogeneity *u* for each individual and averaging over each individual, such that [8]:

$$\ln \mathrm{L}_N = \frac{1}{N} \sum_{i=1}^{N} \ln \left(\frac{1}{S} \sum_{s=1}^{S} \frac{1}{\sqrt{2\pi}} \exp\{-\frac{(y_i - \alpha x_i - \beta \hat{u}_i^s)^2}{2}\} \right) \tag{3}$$

Put differently, an assumption regarding the true distribution of the unobserved heterogeneity needs to be made, so that it can be approximated by this assumed distribution. In this case study, the true distribution of the unobserved heterogeneity is known, such that the assumed distribution can be chosen correctly. This is a crucial point and the main focus of the study at hand: What if the assumption by the researcher does not match the true distribution, i.e., distributional mismatch occurs? Only a few of the before mentioned pieces of research offer a theoretical or practical justification for choosing the (standard) normal as the assumed distribution to match unobserved heterogeneity. Similarly, only few make the reader aware that their assumption may have consequences on the estimation results.

To this end, the here introduced unobserved heterogeneity example will be employed and modified to gain insights on the consequences of mismatching true and assumed distribution in the MSL-context. Although many distributional forms of unobservable heterogeneity seem plausible, e.g., extreme value or the uniform distribution, within this example the mismatch will be achieved by mismatching mean, i.e., μ vs. $\hat{\mu}$ (0 vs. 1), and standard deviation, i.e., $\hat{\sigma}$ vs. σ (1 vs. 2) across the normal distribution, as summarised in Table I. The underlying parameter choice is purely for experimental purposes and is not justified by any other reference. Each of the four constellations will serve as the true data-generating (i.e., unobserved heterogeneity) distribution and will be benchmarked against each of the other four as an assumed distribution which will be employed in MSL-estimation. This will result in sixteen cases, of which four times true and assumed distribution match, whereas in twelve scenarios a mismatch will occur. Table II provides an overview.

Within the simulation study, the α and β coefficients (cf. Equation 1) are to be estimated. Each time the assumed and true distribution match one another, the estimates for α and β, i.e., $\hat{\alpha}$ and $\hat{\beta}$, are hypothesised to be fairly close to their true values, i.e., $\alpha = \frac{1}{2}$ and $\beta = 1$. Yet, interest lies in the situation when a mismatch between assumed and true distribution occurs. It is unclear beforehand whether or not results will be biased and if so how much. This is the central question of this piece of research.

Due to the study design, mismatches will occur along two dimensions: Firstly, there will be four mismatches only among the mean of the assumed and true distribution. Secondly, there will be four mismatches only among the standard deviation of the assumed and true distribution. Also, there will be four mismatches along both dimensions. These twelve mismatches will be exploited for further analysis. Interest lies in the bias of the estimated $\hat{\alpha}$ vs. the true α, as well as the estimated $\hat{\beta}$ vs. the true β. If possible, the bias will be explained by the deviation in μ vs. $\hat{\mu}$ and $\hat{\sigma}$ vs. σ.

III. Preliminary findings

Each of the sixteen scenarios, as described in Table II was estimated 500 times, using [20], [21], [22]. Results are summarised in Figure 1, where the upper part displays the results for the estimates $\hat{\alpha}$, whereas the bottom presents results for $\hat{\beta}$. For each of the two coefficients the diagonal from the top left to the bottom right displays the four scenarios, in which the distributional parameters of u, i.e., $\sim \mathcal{N}(\mu, \sigma)$ and $\hat{u}$, i.e., $\sim \mathcal{N}(\hat{\mu}, \hat{\sigma})$ match one another. As was expected, the observed values are fairly close to their respective true values, i.e, $\alpha = \frac{1}{2}$ and $\beta = 1$, which are represented by a grey vertical line in Figure 1.

Surprisingly, $\hat{\alpha}$ seems to respond differently to mismatches in mean and standard deviation of u than $\hat{\beta}$ does, which was

[1] How to know whether or not one has employed a sufficient amount of simulation draws is subject to another discussion [18], [19].

TABLE I
PARAMETER SUMMARY

Variable	Value	Description
α	.5	true coefficient of x
β	1	true coefficient of u
$\hat{\alpha}$		estimated coefficient of x
$\hat{\beta}$		estimated coefficient of u
x	1	observable characteristics
u	$\sim \mathcal{N}(\mu,\, \sigma)$	true unobservable heterogeneity
μ	$\{0, 1\}$	true mean
σ	$\{1, 2\}$	true standard deviation
$\hat{u}$	$\sim \mathcal{N}(\hat{\mu},\, \hat{\sigma})$	assumed unobservable heterogeneity
$\hat{\mu}$	$\{0, 1\}$	assumed mean
$\hat{\sigma}$	$\{1, 2\}$	assumed standard deviation
ε	$\sim \mathcal{N}(0,\, 1)$	error term
S	1,000	Number of simulation draws
N	1,000	Sample size
R	500	Number of repetitions

not anticipated. Yet, in hindsight, it makes sense, as $\hat{\beta}$ belongs to the unobservable u variable that is incorrectly approximated, whereas $\hat{\alpha}$ belongs to the x variable which can be observed. $\hat{\alpha}$ seems to be shifted away from the true value of α by the difference in true mean and assumed mean, amplified by the relation in mismatch of the standard deviation. The reaction of $\hat{\alpha}$ seems to be described by:

$$\hat{\alpha} = \alpha(1 + \mu - \hat{\mu}\frac{\sigma}{\hat{\sigma}}). \tag{4}$$

For each of the sixteen scenarios in the upper part of Figure 1, this Equation 4 is represented by a blue vertical line.

The reaction of $\hat{\beta}$ on the other hand does not seem to be influenced by any difference in true mean and assumed mean. Nevertheless, it seems to be shifted away from the true value of β by the relation in mismatch of the standard deviation. The reaction of $\hat{\beta}$ can be approximated by:

$$\hat{\beta} = \beta\frac{\sigma}{\hat{\sigma}}. \tag{5}$$

For each of the sixteen scenarios in the bottom part of Figure 1, this Equation 5 is represented by a red vertical line. One notable exception for the latter Equation 5 is the behaviour of $\hat{\beta}$ where the true $u \sim \mathcal{N}(\mu = 1,\, \sigma = 1)$ and the assumed $\hat{u} \sim \mathcal{N}(\hat{\mu} = 0,\, \hat{\sigma} = 2)$ (second row from the top, third column from the left, bottom part of Figure 1). In this case, $\hat{\beta}$ seems to be represented both as implied by Equation 5 as well as its negative, even though the former occurred more often than the latter.

IV. Discussion and Outlook

The lack in guidance regarding potential bias due to mismatch in true and assumed distribution in MSL-estimation motivated this simulation study. It seemed unclear, to what extent the estimation coefficients may be biased from distributional mismatch of mean and standard deviation within the normal distribution. This lead to an back-of-the-envelope calculation, resulting in Equation 4 and Equation 5. These two equations were deduced from the underlying results and seem to approximate the bias in $\hat{\alpha}$ vs. α and $\hat{\beta}$ vs. β fairly well, except for one notable exception, as mentioned in section III. Nevertheless, they are only trial-and-error approximations of the observed results.

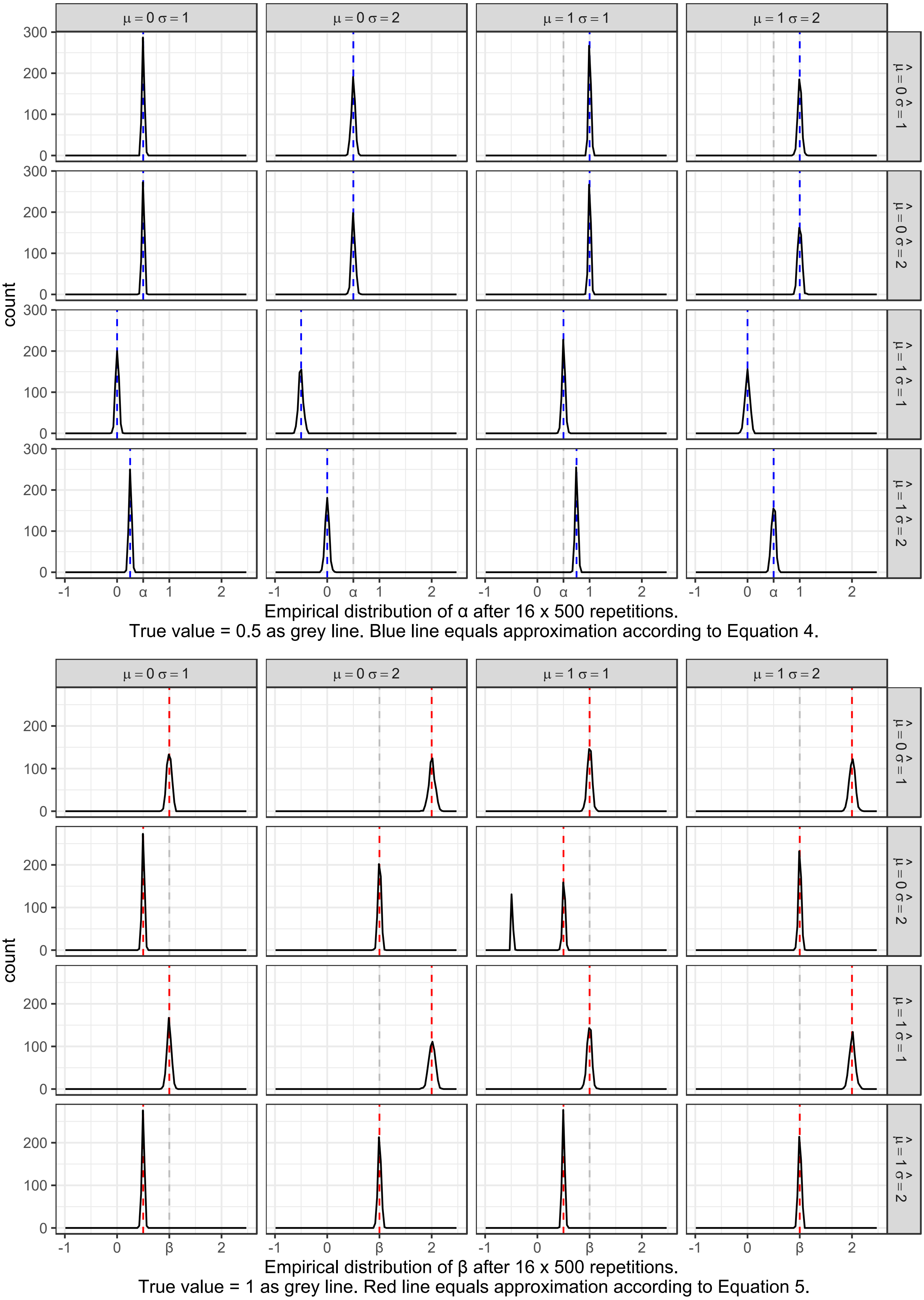
While the lack of guidance, such as provided by Equation 4 and Equation 5, was the motivation to looking for it in the first place, it needs to be assumed that such relation were found and discussed earlier. Yet, this would similarly raise the question why, if it was already common knowledge, none of the found pieces of research that apply MSL-estimation pointed out to this direction when discussing limitations of their models and findings?

Future intended work is motivated by this question: A more quantitatively comprehensive and qualitatively structured literature research will be conducted in the realm of what is described by [23] as *Maximum Approximated Likelihood*, i.e., MSL-estimation, Gaussian-quadrature and integration on sparse grids. The main focus will be placed on the *distributional assumption* regarding the assumed distributions, its theoretical materialisation, i.e., whether it is applied to varying preferences or endogeneity. Variation in the latter findings will then be structured among the dimensions:

- **scope**: theoretical vs. applied papers,
- **estimation method**: e.g., MSL-estimation, Gaussian-quadrature, integration on sparse grids,
- **models**: e.g., mixed multinomial, multinomial treatment regression [24] and
- **field of research**: e.g., healthcare, commerce, transportation.

Additional interest lies in finding pieces of applied research that already had similar findings as given by Equation 4 and Equation 5, as it is assumed that these findings were made already earlier.

Additionally, and equivalently important, remains the further exploration of the bias induced by distributional mismatch between assumed and true distribution in the simulation-context. Depending on the findings of the literature review, Equation 4 and Equation 5 may be further explored, as especially Equation 5 could not approximate all of the sixteen scenarios. As of now it remains unclear, whether or not the findings of Equation 4 and Equation 5 may be applicable to any other situation than the underlying (toy) example. To potentially detect distributional mismatch, consequences regarding the log-Likelihood seems promising with respect to diagnostic tests, such as the Likelihood-ratio test. Similarly, consequences of variance reduction techniques will be discussed.



Empirical distribution of α after 16 x 500 repetitions.
True value = 0.5 as grey line. Blue line equals approximation according to Equation 4.

Empirical distribution of β after 16 x 500 repetitions.
True value = 1 as grey line. Red line equals approximation according to Equation 5.

Fig. 1: RESULTS OF THE SIMULATION STUDY: EMPIRICAL DISTRIBUTION OF THE $\hat{\alpha}$ (TOP) AND $\hat{\beta}$ (BOTTOM) COEFFICIENTS.

TABLE II
SETUP OF SIMULATION STUDY: DISTRIBUTIONAL MISMATCH

		True			
		$\mu = 0, \sigma = 1$	$\mu = 0, \sigma = 2$	$\mu = 1, \sigma = 1$	$\mu = 1, \sigma = 2$
Assumed	$\hat{\mu} = 0, \hat{\sigma} = 1$	match	mismatch $(\hat{\sigma})$	mismatch $(\hat{\mu})$	mismatch $(\hat{\mu}, \hat{\sigma})$
	$\hat{\mu} = 0, \hat{\sigma} = 2$	mismatch $(\hat{\sigma})$	match	mismatch $(\hat{\mu}, \hat{\sigma})$	mismatch$(\hat{\mu})$
	$\hat{\mu} = 1, \hat{\sigma} = 1$	mismatch $(\hat{\mu})$	mismatch $(\hat{\mu}, \hat{\sigma})$	match	mismatch $(\hat{\sigma})$
	$\hat{\mu} = 1, \hat{\sigma} = 2$	mismatch $(\hat{\mu}, \hat{\sigma})$	mismatch $(\hat{\mu})$	mismatch $(\hat{\sigma})$	match

REFERENCES

[1] P. Deb, C. Li, P. K. Trivedi, and D. M. Zimmer. "The effect of managed care on use of health care services: results from two contemporaneous household surveys". In: *Health economics* 15.7 2006, pp. 743–760. DOI: http://dx.doi.org/10.1002/hec.1096.

[2] P. Deb and P. K. Trivedi. "Specification and simulated likelihood estimation of a non–normal treatment–outcome model with selection: Application to health care utilization". In: *The Econometrics Journal* 9.2 2006, pp. 307–331. DOI: http://dx.doi.org/10.1111/j.1368-423X.2006.00187.x.

[3] D. Shane and P. K. Trivedi. "What drives differences in health care eemand? The role of health insurance and selection bias". In: *Health, Econometrics and Data Group (HEDG) Working Papers* 2012. URL: https://www.york.ac.uk/media/economics/documents/herc/wp/12_09.pdf.

[4] D. McFadden and K. Train. "Mixed MNL models for discrete response". In: *Journal of Applied Econometrics* 15.5 2000, pp. 447–470.

[5] D. Revelt and K. Train. "Mixed logit with repeated choices: households' choices of appliance efficiency level". In: *The Review of Economics and Statistics* 80.4 1998, pp. 647–657. DOI: http://dx.doi.org/10.1162/003465398557735. URL: https://direct.mit.edu/rest/article/80/4/647/57083/Mixed-Logit-with-Repeated-Choices-Households.

[6] D. Munger, P. L'Ecuyer, F. Bastin, C. Cirillo, and B. Tuffin. "Estimation of the mixed logit likelihood function by randomized quasi-Monte Carlo". In: *Transportation Research Part B: Methodological* 46.2 2012, pp. 305–320. DOI: http://dx.doi.org/10.1016/j.trb.2011.10.005.

[7] K. E. Train. *Discrete choice methods with simulation*. Cambridge: Cambridge University Press, 2009. DOI: http://dx.doi.org/10.1017/CBO9780511805271.

[8] A. Cameron and P. K. Trivedi. *Microeconometrics: Methods and applications*. New York, NY: Cambridge University Press, 2005.

[9] C. R. Bhat. "Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model". In: *Transportation Research Part B: Methodological* 35.7 2001, pp. 677–693. DOI: http://dx.doi.org/10.1016/S0191-2615(00)00014-X.

[10] L. Chiou and J. L. Walker. "Masking identification of discrete choice models under simulation methods". In: *Journal of Econometrics* 141.2 2007, pp. 683–703. DOI: http://dx.doi.org/10.1016/j.jeconom.2006.10.012.

[11] L. M. Andersen. "Obtaining reliable likelihood ratio tests from simulated likelihood functions". In: *PloS one* 9.10 2014, e106136. DOI: http://dx.doi.org/10.1371/journal.pone.0106136.

[12] M. Bratti and A. Miranda. "Endogenous treatment effects for count data models with endogenous participation or sample selection". In: *Health economics* 20.9 2011, pp. 1090–1109. DOI: http://dx.doi.org/10.1002/hec.1764.

[13] M. B. Buntin, C. H. Colla, P. Deb, N. Sood, and J. J. Escarce. "Medicare spending and outcomes after postacute care for stroke and hip fracture". In: *Medical care* 48.9 2010, pp. 776–784. DOI: http://dx.doi.org/10.1097/MLR.0b013e3181e359df.

[14] M. M. Garrido, P. Deb, J. F. Burgess, and J. D. Penrod. "Choosing models for health care cost analyses: issues of nonlinearity and endogeneity". In: *Health services research* 47.6 2012, pp. 2377–2397. DOI: http://dx.doi.org/10.1111/j.1475-6773.2012.01414.x.

[15] Z. Sándor and K. Train. "Quasi-random simulation of discrete choice models". In: *Transportation Research Part B: Methodological* 38.4 2004, pp. 313–327. DOI: http://dx.doi.org/10.1016/S0191-2615(03)00014-6.

[16] D. Brunner, F. Heiss, A. Romahn, and C. Weiser. *Reliable estimation of random coefficient logit demand models: DICE Discussion Papers*. 2017. URL: https://EconPapers.repec.org/RePEc:zbw:dicedp:267.

[17] C. Gouriéroux and A. Monfort. *Simulation-based econometric methods*. Oxford University Press, 1997. DOI: http://dx.doi.org/10.1093/0198774753.001.0001.

[18] M. Czajkowski and W. Budziski. "Simulation error in maximum likelihood estimation of discrete choice models". In: *Journal of Choice Modelling* 31 2019, pp. 73–85. DOI: http://dx.doi.org/10.1016/j.jocm.2019.04.003.

[19] C. Schrey, T. Schäffer, C. Militzer-Horstmann, and N. Kossack. "Maximum Simulated Likelihood: Don't stop 'til you get enough?" In: *Position Papers of the 2019 Federated Conference on Computer Science and Information Systems*. Annals of Computer Science and Information Systems. PTI, 2019, pp. 79–82. DOI: http://dx.doi.org/10.15439/2019F354.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2020. URL: https://www.R-project.org/.

[21] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer-Verlag New York, 2016. URL: https://ggplot2.tidyverse.org.

[22] Lionel Henry and Hadley Wickham. *purrr: functional programming tools*. 2020. URL: https://CRAN.R-project.org/package=purrr.

[23] M. Griebel, F. Heiss, J. Oettershagen, and C. Weiser. "Maximum approximated likelihood estimation". In: INS Preprint No. 1905 2019. URL: https://ins.uni-bonn.de/media/public/publication-media/INSPreprint1905.pdf?pk=1424.

[24] P. Deb and P. K. Trivedi. "Maximum simulated likelihood estimation of a negative binomial regression model with multinomial endogenous treatment". In: *Stata Journal* 6.2 2006, pp. 246–255.