GSICS Working Paper Series

Resampling-Based Maximum Likelihood Estimation

Takahiro ITO

No. 40

Revised as of December 2023

Graduate School of International

Cooperation Studies

Kobe University

Resampling-Based Maximum Likelihood Estimation[*]

Takahiro Ito[†]

First version: March 20, 2023
This version: December 15, 2023

Abstract

This study develops a novel distribution-free maximum likelihood estimator and formulates it for linear and binary choice models. The estimator is consistent and asymptotically normally distributed (at the rate of $N^{-1/2}$). The Monte Carlo simulation results show that the estimator is strongly consistent and efficient. For the binary choice model, when the linear combination of regressors is leptokurtic, the efficiency loss of having no distribution assumption is virtually nonexistent, and the estimator is superior to the probit and other semiparametric estimators. The results further show that the estimator performs exceedingly well in the presence of a typical perfect prediction problem.

JEL codes: C14, C25

Keywords: semiparametric estimator, distribution-free maximum likelihood estimation, Monte Carlo resampling with replacement, binary choice model, perfect prediction problem,

[*] I am deeply grateful to Manabu Asai and Yuichiro Yoshida for their very insightful comments and suggestions on earlier versions of this paper. This study was supported by JSPS KAKENHI (Grant-in-Aid for Scientific Research) Grant Number 22K01425.
[†] Graduate School of International Cooperation Studies, Kobe University, 2–1, Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan. E-mail: takahiro.ito@lion.kobe-u.ac.jp, Phone: +81-78-803-7148.

1. Introduction

The fact that conventional maximum likelihood (ML) estimation depends on parametric assumptions about the data distribution has long been recognized as a fundamental drawback. In the literature, this issue has been addressed via two distinct approaches. In the first group of studies, where the likelihood is considered to be misspecified and a simplified form of the likelihood function is often used, the minimal assumptions for obtaining consistent estimators of the true parameters and their asymptotic variances have been discussed (Huber, 1967; Wedderburn, 1974; White, 1982; Ruud, 1983; Gourieroux et al., 1984).[1] This so-called quasi (pseudo) ML approach, however, still requires the assumption that the data follow a specific class of distributions.

In contrast, the second group of studies estimates the likelihood function nonparametrically. Distribution-free ML estimators for semiparametric regression models include Cosslett's (1983) infinite-dimensional ML estimator, the sieve ML estimator (Duncan, 1986; Fernandez, 1986; Gallant and Nychka, 1987), the local ML estimator (Tibshirani and Hastie, 1987; Fan et al., 1998), and the kernel ML estimator (Klein and Spady, 1993; Lee, 1995; Ai, 1997; Ichimura and Thompson, 1998). Most of these estimators exhibit consistency, and some are asymptotically normal and attain the semiparametric efficiency bound. However, none of these estimators is adaptive: Nonparametric estimation of an unknown likelihood function tends to lead to a nonnegligible loss of efficiency.

In this study, I propose a third avenue of ML estimation with the advantages of the quasi ML approach and the semiparametric ML approach. Specifically, the newly proposed method exploits a parametric likelihood function by leveraging the normality of the limit distribution of the data generated from the original data by Monte Carlo "in-sample" resampling with replacement. The proposed method, therefore, does not require assuming an ad hoc data distribution and computing the likelihood nonparametrically. This is a clear methodological advantage, and the estimator exhibits several desirable asymptotic properties of the conventional ML estimator: it is consistent and asymptotically normally distributed (at the rate of $N^{-1/2}$). In addition, for linear regression models, by achieving the Cramér–Rao lower bound, the new estimator is shown to be fully efficient. For binary choice models, where the parameters are identified up to a scale, a comparison of the asymptotic variance of the *scale-normalized* parameters shows that the proposed estimator can be as efficient as the probit estimator for a class of *regressor* distributions.

The Monte Carlo simulation results for linear regression and binary choice models indeed show that the proposed estimator is strongly consistent and efficient. For the binary choice model, when the linear combination of the regressors is leptokurtic, the estimator performs better than ML-based methods, including the probit and other semiparametric estimators, even in cases where the probit is the correct model. The superiority of the proposed method in the binary choice model holds even in the presence of conditionally

[1] Generalized linear models based on an assumed exponential family density have also been developed (see, for instance, Nelder and Wedderburn (1972) and McCullagh and Nelder (1983)).

heteroscedastic errors, consistent with the theoretical argument in a companion paper (Ito, 2024). Moreover, the simulation results also indicate that the new estimator is the most stable in the sense that the root mean square error (RMSE) of the estimated parameters varies least with the regressor and error distributions. Theoretically, this could be because the proposed method constructs the likelihood based on the normality of the generated data, regardless of the distribution of the original data.[2]

Furthermore, the new method is expected to be free from perfect prediction (or complete separation) problems in discrete choice models since it focuses on variations around the mean of (dependent and explanatory) variables, not on the one-to-one correspondence between them. The simulation results also support this theoretical implication, showing that in the presence of a typical perfect prediction problem, the new method is superior to other semiparametric methods in terms of the number of trials with convergence and the magnitude of the bias and RMSE.[3]

The remainder of this paper is organized as follows. In Section 2, I propose a new ML-based semiparametric estimator and establish the root-N consistency and asymptotic normality of the proposed estimator. Section 3 presents example applications of the proposed method and discusses the asymptotic efficiency for linear regression and binary choice models. The Monte Carlo simulation results are presented in Section 4, and the conclusions follow in Section 5.

2. Resampling-based maximum likelihood estimator

2.1. Definitions

Conventional (parametric) ML estimation assumes that the data are from a specific distribution, such as a normal distribution, a Poisson distribution, or another distribution from the exponential family, and the conditional mean of the dependent variable is correctly specified (Gourieroux et al., 1984). However, there is no guarantee that these distributional assumptions hold. To overcome this fundamental flaw in parametric ML estimation, a new ML estimator is proposed in this study based on the normality of the limit distribution of the data generated by Monte Carlo "in-sample" resampling.

Definition 1 (RBML data construction)

[2] The use of the parametric likelihood function provides another practical advantage. In general, semiparametric methods are likely to have difficulties in optimizing the objective function due to complex calculations of the unknown function (and probably its undulating shape). However, the proposed method is expected to be less problematic in maximizing the likelihood than are the conventional semiparametric methods. In the simulation analysis of the binary choice models in Section 4, the proposed method achieved convergence in all 23,000 trials, but the sieve ML, kernel ML, and semiparametric least squares (SLS) estimations fail to achieve convergence in 378 (1.64%), 2,327 (10.12%), 2,111 (9.18%) cases, respectively, despite estimating simple models.

[3] All simulation results presented in this study can be replicated using a software package for Stata that is available on the author's website. For details on the processes of obtaining and using the package, see Online Supplementary Material I.

Suppose there exist data with a sample size of N, that is, $\{z_i | \, i = 1, \cdots, N\}$. By employing Monte Carlo resampling, new data are obtained as follows.

(1) Draw M observations from $\{z_i | \, i = 1, \cdots, N\}$ by random resampling with replacement ($M > N$); multiply the difference between the mean of the M observations ($\bar{z} = M^{-1} \sum_{j=1}^{M} z_j$) and the original sample mean ($\mu_N = N^{-1} \sum_{i}^{N} z_i$) by $\sqrt{NM/(N-1)}$; and obtain an observation $\tilde{z} = \sqrt{NM/(N-1)} \cdot (\bar{z} - \mu_N)$.

(2) Repeat the first step T times and obtain a new sample with T observations:

$$\{\tilde{z}_t \mid t = 1, \cdots, T\}.$$

M and T are set to be sufficiently large. In particular, T increases in N (e.g., $T = T' + N$), where T' is thus sufficiently large. This procedure to obtain a new sample $\{\tilde{z}_t\}$ via Monte Carlo "in-sample" resampling is defined as resampling-based maximum likelihood (RBML) data construction.

Then, based on an approach in finite population theory (Cornfield, 1944; Raj and Khamis, 1958), it is shown that newly generated data have the following distribution property.

Proposition 1 (Distribution of a new sample obtained by RBML data construction)
Assume that an original sample $\{z_i | \, i = 1, \cdots, N\}$ is independent and identically distributed (i.i.d.) with finite mean μ_0 and finite variance σ_0^2 ($\sigma_0^2 \neq 0$). Define the sample mean and variance as $\mu_N = N^{-1} \sum_{i=1}^{N} z_i$ and $\sigma_N^2 = N^{-1} \sum_{i=1}^{N} (z_i - \mu_N)^2$, respectively. When M (the size in the resampling process in RBML data construction) is sufficiently large, new data $\{\tilde{z}_t \mid t = 1, \cdots, T\}$ obtained by RBML data construction are i.i.d. with $\mathrm{N}(0, \sigma_N^2)$ and converge in distribution to $\mathrm{N}(0, \sigma_0^2)$ as N goes to infinity:

$$\tilde{z}_t \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_N^2) \xrightarrow{d} \mathrm{N}(0, \sigma_0^2).$$

The proof is provided in Appendix A.1. Then, the RBML estimator is defined as follows.

Definition 2 (RBML estimator)
Let $\{\tilde{z}_t \mid t = 1, \cdots, T\}$ be an i.i.d. sample obtained by RBML data construction from the original sample $\{z_i | \, i = 1, \cdots, N\}$ and $p(\tilde{z}_t \mid \theta)$ be the probability of an event of interest based on the normal density of $\tilde{z}_t$. Then, the RBML estimator is defined as:

$$\hat{\theta}_{\mathrm{RB}} = \arg\max_{\theta \in \Theta} \ln L(\theta | \, \tilde{\mathbf{z}}) = \arg\max_{\theta \in \Theta} \, T^{-1} \sum_{t=1}^{T} \ln p(\tilde{z}_t \mid \theta), \tag{1}$$

where Θ is a compact parameter space.

Thus, exploiting the normality of the limit distribution of the generated data frees us from the nonparametric estimation of the likelihood function.

2.2. Asymptotic properties

Theorems 1 and 2 state the results for the consistency and limit distribution of the RBML estimator. The theorems are similar to those for the conventional ML estimator.[4]

Theorem 1 (Consistency)
If (i) $\theta_0 \in \Theta$, *which is compact, (ii)* $\ln L(\theta|\,\tilde{\mathbf{z}}) = T^{-1}\sum_t^T \ln p(\tilde{z}_t|\,\theta)$ *converges to* $\mathrm{E}[\ln p(\tilde{z}\mid\theta)]$ *in probability uniformly in* $\theta \in \Theta$ *as N goes to infinity, (iii)* $\mathrm{E}[\ln p(\tilde{z}\mid\theta)]$ *is continuous, and (iv)* $\mathrm{E}[\ln p(\tilde{z}\mid\theta)]$ *is uniquely maximized at* θ_0, *then* $\hat{\theta}_{\mathrm{RB}} \xrightarrow{p} \theta_0$.

See Appendix A.2 for the proof of Theorem 1. Note that conditions (ii) and (iv) of the theorem can be replaced with more primitive conditions. See Lemmas 2.2 and 2.4 in Newey and McFadden (1994).

Theorem 2 (Asymptotic normality)
Suppose that the conditions of Theorem 1 are satisfied. Furthermore, if (i) $\theta_0 \in \mathrm{interior}(\Theta)$, *(ii)* $p(\tilde{z}_t|\,\theta)$ *is twice continuously differentiable with respect to* θ *and* $p(\tilde{z}_t|\,\theta) > 0$ *in an open convex neighbourhood* $\mathcal{N}$ of θ_0 *contained in* Θ, *(iii)* $J = \mathrm{E}[\{\nabla_\theta \ln p(\tilde{z}\mid\theta_0)\}\{\nabla_\theta \ln p(\tilde{z}\mid\theta_0)\}']$ *exists and is nonsingular, where* ∇_θ *is the operator taking the first partial derivatives, (iv)* $T^{-1}\sum_t^T \nabla_{\theta\theta} \ln p(\tilde{z}_t|\,\theta)$ *converges uniformly in probability to* $H = \mathrm{E}[\nabla_{\theta\theta} \ln p(\tilde{z}\mid\theta)]$, *where* $\nabla_{\theta\theta}$ *is the operator taking the second partial derivatives and H is continuous at* θ_0, *and (v)* $\int \sup_{\theta\in\mathcal{N}} \|\nabla_\theta p(\tilde{z}\mid\theta)\|\,\mathrm{d}\tilde{z} < \infty$ *and* $\int \sup_{\theta\in\mathcal{N}} \|\nabla_{\theta\theta} p(\tilde{z}\mid\theta)\|\,\mathrm{d}\tilde{z} < \infty$, *then* $\sqrt{N}\big(\hat{\theta}_{\mathrm{RB}} - \theta_0\big) \xrightarrow{d} \mathrm{N}(0, J^{-1})$.

The proof is provided in Appendix A.3. Theorem 2 indicates that when N is large, the variance–covariance matrix of the RBML estimator is approximated by $(N \cdot J)^{-1}$, and it appears to be exactly the same as the Cramér–Rao lower bound for the estimator. However, the Cramér–Rao bound for the model expressed in Eq. (1) is $(T \cdot J)^{-1}$, and the efficiency of the RBML estimator must be discussed for each regression model based on the asymptotic variance of the conventional normal ML estimator. In the following section, I discuss the asymptotic efficiency of the RBML estimator for linear and binary choice models.

3. Applications

3.1. Linear regression model

Suppose there exists a sample $\{(y_i, \mathbf{x}_i)|\; i = 1, \cdots, N\}$, where $y_i \in \mathbb{R}$ and $\mathbf{x}_i' \in \mathbb{R}^K$ are independent random

[4] See, for example, Theorem 4.1.1 in Amemiya (1985) and Theorem 2.1 in Newey and McFadden (1994) for consistency. See Theorem 3.3 in Newey and McFadden (1994) for asymptotic normality.

variables with finite means and variances. Note that in the following, z_i, $\tilde{z}_t$, and $p(\tilde{z}_t|\theta)$ in the previous section correspond to $(y_i, \mathbf{x}_i)$, $(\tilde{y}_t, \tilde{\mathbf{x}}_t)$, and $p(\tilde{y}_t|\tilde{\mathbf{x}}_t, \boldsymbol{\theta})$, respectively. The first model considered here is a linear regression model expressed as:

$$y_i = \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0 + \varepsilon_i,$$

where $\varepsilon_i \in \mathbb{R}$ is an unobserved error and $\alpha_0 \in \mathbb{R}$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^K$ are unknown population parameters. It is also assumed that $\mathrm{Rank}\left[\sum_i^N \dot{\mathbf{x}}_i{}'\dot{\mathbf{x}}_i\right] = K+1$, where $\dot{\mathbf{x}}_i = (1, \mathbf{x}_i)$, and $\mathrm{E}[\varepsilon_i|\dot{\mathbf{x}}_i] = 0$, which implies that $\mathrm{E}[y_i|\dot{\mathbf{x}}_i] = \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0$.

Then, by means of RBML data construction, new data $\{(\tilde{y}_t, \tilde{\mathbf{x}}_t)|\ t = 1, \cdots, T\}$ are obtained, and we have

$$\tilde{y}_t = \tilde{\mathbf{x}}_t\boldsymbol{\beta}_0 + \tilde{\varepsilon}_t,$$

where $\tilde{\varepsilon}_t \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_N^2) \overset{d}{\to} \mathrm{N}(0, \sigma_0^2)$, $\sigma_N^2 = N^{-1}\sum_{i=1}^N \varepsilon_i^2$, and $\sigma_0^2 = \lim_{N\to\infty} \mathrm{E}[\sigma_N^2]$. The distribution property of the new error term $\tilde{\varepsilon}_t$ follows from Proposition 1 for the homoscedastic case (i.e., $\mathrm{E}[\varepsilon_i^2|\dot{\mathbf{x}}_i] = \sigma_0^2$ for all i) and Proposition A1 in Appendix A.4 for the heteroscedastic case (i.e., $\mathrm{E}[\varepsilon_i^2|\dot{\mathbf{x}}_i] = \sigma_i^2$). In particular, σ_i^2 is allowed to depend on the value of $\dot{\mathbf{x}}_i$, that is, $\sigma_i^2 = h_i(\dot{\mathbf{x}}_i)$. See Ito (2024) for the theoretical discussion.

Thus, the RBML estimator $\widehat{\boldsymbol{\theta}}'_{\mathrm{RB}} = \left(\widehat{\boldsymbol{\beta}}'_{\mathrm{RB}}, \hat{\sigma}_{\mathrm{RB}}\right)$ is defined as values that satisfy the following:

$$\widehat{\boldsymbol{\theta}}_{\mathrm{RB}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ln L\left(\boldsymbol{\theta}; \tilde{\mathbf{y}}, \tilde{\mathbf{X}}\right) = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{t=1}^{T}\left[\ln\phi\left\{\frac{(\tilde{y}_t - \tilde{\mathbf{x}}_t\boldsymbol{\beta})}{\sigma}\right\} - \ln\sigma\right], \tag{2}$$

where $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^{K+1}$, which contains the true value $\boldsymbol{\theta}_0$, and $\phi(\cdot)$ is the standard normal density.

The limit variance matrix of $\sqrt{N}\left(\widehat{\boldsymbol{\beta}}_{\mathrm{RB}} - \boldsymbol{\beta}_0\right)$ for the above linear model, according to Theorem 2, is $\sigma_0^2 \lim_{N\to\infty}(T^{-1}\sum_t^T \tilde{\mathbf{x}}_t'\tilde{\mathbf{x}}_t)^{-1} = \sigma_0^2 \lim_{N\to\infty}\left(\boldsymbol{\Sigma}_{N,\mathbf{x}}\right)^{-1}$, where $\boldsymbol{\Sigma}_{N,\mathbf{x}} = N^{-1}\sum_i^N\left(\mathbf{x}_i - \boldsymbol{\mu}_{N,\mathbf{x}}\right)'\left(\mathbf{x}_i - \boldsymbol{\mu}_{N,\mathbf{x}}\right)$. $\boldsymbol{\Sigma}_{N,\mathbf{x}}$ and $\sum_t^T \tilde{\mathbf{x}}_t'\tilde{\mathbf{x}}_t$ are, by assumption, nonsingular matrices. Note also that $T \to \infty$ when $N \to \infty$ because $T = T' + N$, and the equality comes from the fact that the variance of $\tilde{\mathbf{x}}$ is equal to the variance of $\mathbf{x}$ when N is large, that is, $\lim_{N\to\infty}(T^{-1}\sum_t^T \tilde{\mathbf{x}}_t'\tilde{\mathbf{x}}_t) = \lim_{N\to\infty}\boldsymbol{\Sigma}_{N,\mathbf{x}}$. Therefore, the variance of the RBML estimator, $\widehat{\boldsymbol{\beta}}_{\mathrm{RB}}$ in Eq. (2), achieves the Cramér–Rao lower bound for the linear regression model. Notably, even when N is small, $T^{-1}\sum_t^T \tilde{\mathbf{x}}_t'\tilde{\mathbf{x}}_t$ can equal $\boldsymbol{\Sigma}_{N,\mathbf{x}}$ if T' (and, hence, T) is sufficiently large. However, if T' is small, $T^{-1}\sum_t^T \tilde{\mathbf{x}}_t'\tilde{\mathbf{x}}_t = \boldsymbol{\Sigma}_{N,\mathbf{x}} + \mathrm{o}_p(1)$ and $T^{-1}\sum_t^T \hat{\tilde{\varepsilon}}_t{}^2 = \sigma_N^2 + o_p(1)$ for a given N. This indicates that a small T' leads to a loss of efficiency when the sample size N is small. Therefore, T' is assumed to be sufficiently large in Definition 1 (RBML data construction). The efficiency loss due to a small T' is verified in the Monte Carlo simulation in Section 4.

3.2. Binary choice model

The second example is a binary choice model, which is expressed as follows:

$$d_i = 1[y_i > 0] = 1[\alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0 + \varepsilon_i > 0], \tag{3}$$

where $d_i \in \{0,1\}$ and $\mathbf{x}_i' \in \mathbb{R}^K$ are observed, $y_i \in \mathbb{R}$ is an unobserved latent variable, $\varepsilon_i \in \mathbb{R}$ is an unobserved error, and $\alpha_0 \in \mathbb{R}$ and $\boldsymbol{\beta}_0 \in \mathbb{R}^K$ are unknown population parameters. It is assumed that y_i and $\mathbf{x}_i$ are independent random variables with finite means and variances. In addition, by letting $\grave{\mathbf{x}}_i = (1, \mathbf{x}_i)$, it is assumed that $\mathrm{Rank}[\sum_i^N \grave{\mathbf{x}}_i' \grave{\mathbf{x}}_i] = K+1$ and $\mathrm{E}[\varepsilon_i | \grave{\mathbf{x}}_i] = 0$. In particular, the second condition (zero conditional mean of errors) implies that $\mathrm{E}[y_i | \grave{\mathbf{x}}_i] = \alpha_0 + \mathbf{x}_i \boldsymbol{\beta}_0$. Note that both the linear and binary models here assume that y_i and $\mathbf{x}_i$ have a linear relationship. This linear assumption may seem stringent, but this is not always the case. Especially in typical experimental settings, where binary treatment status is often employed, the assumption can be regarded as less restrictive. These issues are discussed in detail by Ito (2024).

Then, by applying RBML data construction and arranging the expression of the latent outcome $\tilde{y}_t$, we obtain the following relationship (see Appendix A.5 for the derivation):

$$\begin{aligned} \tilde{d}_t > 0 \quad &\text{if } \tilde{y}_t = \tilde{\mathbf{x}}_t \boldsymbol{\beta}_0 + \tilde{\varepsilon}_t > \gamma_t \\ \tilde{d}_t \leq 0 \quad &\text{if } \tilde{y}_t = \tilde{\mathbf{x}}_t \boldsymbol{\beta}_0 + \tilde{\varepsilon}_t \leq \gamma_t. \end{aligned} \tag{4}$$

where γ_t is a threshold variable defined in Eq. (A-6) in the Appendix and is assumed to be independent of $\tilde{\mathbf{x}}_t$ and $\tilde{\varepsilon}_t$.[5] Note that $\gamma_t \overset{i.i.d.}{\sim} \mathrm{N}\left(0, n_{N,0}\sigma_{N,0}^2 + n_{N,1}\sigma_{N,1}^2\right)$, where by defining $\{a\} = \{y_i | \, i = 1, \cdots, N \;\&\; d_i = a\}$, $N_a = \#\{a\}$ (the number of observations for which $d_i = a$), and $\bar{y}_{N,a} = N_a^{-1} \sum_{y_i \in \{a\}} y_i$, $n_{N,a} = N_a / N$ and $\sigma_{N,a}^2 = N_a^{-1} \sum_{y_i \in \{a\}} \left(y_i - \bar{y}_{N,a}\right)^2$ for $a = 0,1$. Additionally, as shown in the linear regression model, $\tilde{\varepsilon}_t \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_N^2)$, where $\sigma_N^2 = N^{-1} \sum_{i=1}^N \varepsilon_i^2$. Letting $\delta_t = \gamma_t - \tilde{\varepsilon}_t$ and $\tau = \sqrt{n_{N,0}\sigma_{N,0}^2 + n_{N,1}\sigma_{N,1}^2 + \sigma_N^2}$, $\boldsymbol{\delta}/\tau = (\delta_1, \cdots, \delta_T)'/\tau$ are uncorrelated standard normal random variables and therefore are *i.i.d.* with $\mathrm{N}(0,1)$. Thus, $p\left(\tilde{d}_t | \, \tilde{\mathbf{x}}_t, \boldsymbol{\theta}\right) = [\Phi(\tilde{\mathbf{x}}_t \boldsymbol{\beta}/\tau)]^{1[\tilde{y}_t > 0]} \cdot [1 - \Phi(\tilde{\mathbf{x}}_t \boldsymbol{\beta}/\tau)]^{1[\tilde{y}_t \leq 0]}$, and the RBML estimator for the binary choice model is defined as values that satisfy

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{\mathrm{RB}} &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \ln L\left(\boldsymbol{\theta} \,\middle|\, \tilde{\mathbf{d}}, \tilde{\mathbf{X}}\right) \\ &= \arg\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{t=1}^{T} \left\{ 1\left[\tilde{d}_t > 0\right] \ln \Phi(\tilde{\mathbf{x}}_t \boldsymbol{\theta}) + 1\left[\tilde{d}_t \leq 0\right] \ln \Phi(-\tilde{\mathbf{x}}_t \boldsymbol{\theta}) \right\}, \end{aligned} \tag{5}$$

where $\boldsymbol{\theta} = \boldsymbol{\beta}/\tau$, $\boldsymbol{\Theta}$ is a compact subset of $\mathbb{R}^K$, which contains the true value $\boldsymbol{\theta}_0$, and $\Phi(\cdot)$ is the standard normal cumulative distribution function.

Notably, the proposed method is theoretically expected to be free from perfect prediction (or complete separation) problems in discrete choice models. This is because the new method focuses on variations around the mean of the (outcome and explanatory) variables and not the one-to-one correspondence between them. This notable property is also examined in the Monte Carlo simulation analysis in Section 4.

Also note that in the proposed estimation, the parameters in Eq. (5) are identified up to a scale, as in

[5] In the absence of the assumption that γ_t is independent of $\tilde{\mathbf{x}}_t$ and $\tilde{\varepsilon}_t$, it is impossible to estimate $\boldsymbol{\beta}_0$ unless an additional identification condition is assumed. Online Supplementary Material II examines the validity of this assumption using simulated data, and the results support the assumption.

the conventional probit estimation, but with a different scale from the probit. The estimand of the probit estimation is β_0/σ_0 and that of the RBML estimation is β_0/τ_0. Therefore, a direct comparison of the variance or efficiency between the two estimators is not meaningful. The small sample properties, including the efficiency of the proposed estimator for the binary choice model, are verified in the Monte Carlo simulations in the next section. See also Online Supplementary Material III, where a numerical simulation is conducted to compare the asymptotic variance of the *scale-normalized* parameters between the probit and RBML methods. The results in the Appendix show that the RBML estimator can be as efficient as the probit estimator when the linear combination of regressors is highly leptokurtic.

For applications other than the two examples above, the RBML method is applicable to, for example, sample selection models based on the two-step procedure and ordered response models.[6]

4. Monte Carlo analysis

4.1. Simulation design

In the simulation analysis, the RBML method is applied to linear regression and binary choice models, and the finite sample performance of the proposed estimator is examined. Specifically, I consider three estimation models, as explained below.

The first model (referred to as Model 1) is a linear regression model given by $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$. The regressors are generated by drawing randomly from a chi-square distribution with 16 degrees of freedom and adjusting the values to have a variance of 0.5 (i.e., $x_i, z_i \sim \chi^2(16)/8$). The parameters are set as $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$; thus, $\beta_1 x_i + \beta_2 z_i$ has unit variance. Regarding the error distribution, four designs are employed: (a) standard normal, $\mathrm{N}(0,1)$; (b) normal mixture (left skewed and leptokurtic), $0.6 \cdot \mathrm{N}(-0.3, 1.225) + 0.4 \cdot \mathrm{N}(0.45, 0.325)$;[7] (c) normal with heteroscedastic variance, $\mathrm{N}(0, (\mathbf{x}_i\boldsymbol{\beta})^2/\mathrm{E}[(\mathbf{x}_i\boldsymbol{\beta})^2])$, where $\mathbf{x}_i\boldsymbol{\beta} = \beta_1 x_i + \beta_2 z_i$, and (d) Student's t with two degrees of freedom, $\mathrm{T}(2)$.

The second model (Model 2) is a binary choice model given by $d_i = 1[\beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i > 0]$. The parameters are set as $\beta_1 = \beta_2 = 1$, and the nuisance parameter, β_0, is set to the negative population median of $\beta_1 x_i + \beta_2 z_i$ so that the probability of $d_i = 1$ in the population is 0.5. For the regressors, several distributions with different degrees of kurtosis are employed. This is because the relative efficiency of the RBML estimator for binary choice models is expected to depend on the kurtosis of the regressors, as indicated in Online Supplementary Material III. Thus, I employ the following cases where the (excess) kurtosis of x_i and z_i varies from −2 to 4: chi-square distributions with 3, 4, 6, 8, 12, and 24 degrees of freedom (kurtosis $= 12/\text{d.f.}$); normal distribution (kurtosis $= 0$); and beta distributions with parameter values of $(4.5, 4.5)$, $(1.5, 1.5)$, $(0.5, 0.5)$, and $(10^{-10}, 10^{-10})$ (kurtosis $= -6/(\text{2p.v.} + 3)$). Here, again, $\beta_1 x_i + \beta_2 z_i$ is designed to have the same unit variance in all cases (but different kurtosis values from −2 to 4). The error designs used

[6] For an application to ordered response models, see Ito (2024).

[7] Although not mentioned in the text, I also use a right skewed and platykurtic error distribution as another type of normal mixture distribution. See Online Supplementary Material IV for details.

for the binary model are the same as those for the linear model (Model 1).

Notably, when the expectation of each regressor (x_i and z_i), conditional on $\mathbf{x}_i\boldsymbol{\beta}$, is linear in $\mathbf{x}_i\boldsymbol{\beta}$, the probit estimator is known to be consistent (Ruud, 1983).[8] Therefore, for regressors with chi-square or normal distributions, the conditional expectations of x_i and z_i given $\mathbf{x}_i\boldsymbol{\beta}$ are linear, and the probit estimator is consistent (up to scale) across all error designs, although some efficiency may be lost when the errors are nonnormal. On the other hand, the RBML estimator is consistent for all error designs except one—error design (d)—because Student's t distribution with 2 degrees of freedom has no variance, the central limit theorem (CLT) cannot be applied in the RBML data construction, and the RBML estimator is theoretically inconsistent. Therefore, to determine the performance of the RBML estimator in an inconsistent case, while the errors in designs (a) to (c) are set to have unit variance in the population, the variance of the errors in design (d) is not standardized.

Finally, the third estimation model (Model 3) is a binary choice model with a perfect prediction (or complete separation) problem. Specifically, I consider a simple problem in which a regressor predicts the binary outcome perfectly. As mentioned in Section 3.2, the proposed method is theoretically expected to be able to estimate the impact of a variable that perfectly predicts the outcome in discrete choice models. Model 3 is given by $d_i = 1[\beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i > 0]$, where x_i is a binary variable that is unity with a probability of 50%, z_i and ε_i follow normal distributions defined as $z_i \sim \mathrm{N}(0, 0.2)$ and $\varepsilon_i \sim \mathrm{N}(0, 0.8)$, respectively, $\beta_0 = 0$, and $\beta_2 = 1$. Then, the following two cases with different degrees of perfect prediction are considered. In the first case, $\beta_1 = \Phi^{-1}(0.99) \approx 2.326$, and therefore, for 99% of the observations with $x_i = 1$, $\beta_1 x_i > -(\beta_0 + \beta_2 z_i + \varepsilon_i)$ and $y_i = 1$. This case is referred to as a *nearly* perfect prediction. Then, the second case is a *fully* perfect prediction, where $\beta_1 = |\mathrm{Min}(\beta_0 + \beta_2 z_i + \varepsilon_i)| + 0.01$, and therefore, for all observations with $x_i = 1$, $y_i = 1$. Note that it is generally not possible to estimate the impact of a variable causing a fully perfect prediction, as observed from the fact that doubling or tripling β_1 does not change the binary outcome (and hence, this may result in an inflated estimate of the coefficient). Therefore, in the presence of a perfect prediction problem, a reasonable goal is to estimate the impact at the threshold, and β_1 in this second case is set to a value close to the threshold such that $\beta_1 x_i > -(\beta_0 + \beta_2 z_i + \varepsilon_i)$. Table 1 summarizes the simulation designs for the three models. For all simulations, the sample size in a trial is set to 500 ($N = 500$), and each design consists of 500 independent trials.[9]

[Table 1 near here]

4.2. Results

4.2.1. Linear regression model (Model 1)

[8] For linear regression models, the normal ML estimator is equivalent to the ordinary least squares (OLS) estimator and is consistent, albeit with a loss of efficiency, even when the errors are nonnormal.

[9] Descriptive statistics of the variables used in a trial are presented in Table IV-1 in Online Supplementary Material IV.

Table 2 presents the simulation results for the linear regression model. To investigate the role of the size of T' (and thus, T), different values of T are employed: $T = 1{,}000$, $10{,}000$, and $100{,}000$. OLS results are also reported for comparison.

[Table 2 near here]

The results show that the RBML estimator is strongly consistent. As previously suggested, even if T is small, the bias is comparable to that of the OLS estimator across all four error designs. However, when T is relatively small ($T = 1{,}000$ or $10{,}000$), the RMSE is always greater for the RBML estimator than for the OLS estimator. Nonetheless, when $T = 100{,}000$, the RBML estimator performs as well as the OLS estimator in terms of the RMSE: the relative efficiency always exceeds 99%. As shown in Section 3.1, the asymptotic variance of the RBML and OLS estimators is identical for the linear regression model, and this is true when T is sufficiently large. The efficiency loss due to the small T' almost disappears when $T = 100{,}000$.

4.2.2. Binary choice model (Model 2)

Figure 1 presents the simulation results for the binary choice model. For comparison with the RBML method, the results obtained by probit (normal ML), Gallant and Nychka's (1987) Hermite polynomial sieve ML, Klein and Spady's (1993) Nadaraya–Watson kernel ML, and Ichimura's (1993) SLS estimations are also presented.[10] Note that the bias and RMSE reported in the figure are for the regression coefficient and not the marginal effect. For discrete choice models, researchers' interest is generally in the marginal effect on the choice probability, not the coefficient estimate. However, in this simulation, the ratio of coefficients of the two regressors is designed to equal the ratio of their marginal effects.[11]

[Figure 1 near here]

As shown in Figure 1 (lower lines), the RBML estimator is strongly consistent for all kurtosis values (horizontal axis) and error designs (Panels A to D). Notably, regardless of the error designs, the bias of the RBML estimator is smaller than those of the other estimators in most cases. The only exception is when the kurtosis is -2 (i.e., x_i and z_i are both binary variables). In this case, the RBML estimator tends to have a larger bias than those of the probit and sieve ML estimators.

Turning to the RMSE (upper lines), it is worth noting that when the linear combination $\mathbf{x}\boldsymbol{\beta}$ is leptokurtic (positive excess kurtosis), the RBML estimator always has the smallest RMSE compared to those of the probit and other semiparametric estimators for all error designs. This is also true for the case of normal errors (Panel A), where the probit specification is correct. Although the numerical simulation in Online

[10] The sieve ML, kernel ML, and SLS estimations are implemented using the –snp–, –sml–, and –sls– commands in Stata, respectively. See De Luca (2008) for the –snp– and –sml– commands and Barker (2014) for the –sls– command.

[11] Although this study focuses on coefficient estimates, it would be an interesting exercise to examine the various situations in which the choice probabilities are inconsistently estimated by probit estimation but consistently estimated by semiparametric estimators. Ito (2024) compared marginal effect estimates between the RBML and other parametric and semiparametric estimators for binary and ordered response models.

Supplementary Material III suggests that the RBML estimator is asymptotically less efficient than the probit estimator when the kurtosis of the regressors is positive but small, the results in this figure show that the RBML estimator performs exceedingly well for the small sample.

When $\mathbf{x}_i\boldsymbol{\beta}$ is platykurtic (negative excess kurtosis), on the other hand, the performance of the proposed estimator is worse than that of the probit and sieve estimators in most cases except for the case of a kurtosis of −2. When the regressors are both dummy variables (and the excess kurtosis is −2), semiparametric estimators tend to have difficulty estimating the likelihood nonparametrically. In fact, kernel-based semiparametric estimations require at least one variable to be continuous, and hence, kernel ML and SLS methods cannot be applied in this case. In addition, in the case of the sieve ML estimation, of 2,000 trials (500 trials × 4 error designs), there are 367 cases (or 18.4%) in which convergence cannot be achieved (while the proposed method achieves convergence in all 2,000 trials). Even when the estimation performed successfully, the RMSEs are sacrificed.[12]

Then, when $\mathbf{x}_i\boldsymbol{\beta}$ is normal (mesokurtic), the RBML, probit, and sieve ML estimators achieve comparable performance: the relative efficiency of the RBML and sieve ML estimators to the probit estimator always exceeds 97%. Moreover, we see that the probit estimator is clearly superior to both the RBML and sieve ML estimators only when the kurtosis of $\mathbf{x}_i\boldsymbol{\beta}$ is −2. Therefore, except in those special cases, the RBML and sieve ML estimations complement each other as alternatives to the probit estimation: the RBML estimator for the leptokurtic $\mathbf{x}_i\boldsymbol{\beta}$ and the sieve ML estimator for the platykurtic $\mathbf{x}_i\boldsymbol{\beta}$.

In addition, the RBML estimator performs better than the kernel ML and SLS estimators in terms of the RMSE in most cases. While kernel-based semiparametric estimators generally suffer a loss of efficiency when estimating an unknown function nonparametrically, the proposed method utilizing the parametric likelihood function does not. The results further indicate that the RBML estimator is considerably stable in the sense that the RMSE does not vary significantly with the degree of the regressors' kurtosis and error design. When the kurtosis of the regressors varies, the RMSEs of all the estimators also vary, but the RBML estimator fluctuates less than the other estimators. The proposed method provides robust estimates across the error and regressor distributions.

4.2.3. Binary choice model with a perfect prediction problem (Model 3)

Finally, Figure 2 presents the distribution of the estimation errors (differences between estimates and true parameter values) for the binary choice model with a perfect prediction problem (Model 3). Panel A represents the case where x_i predicts y_i nearly perfectly (for 99% of the observations with $x_i > 0, y_i = 1$), and Panel B is the case where x_i predicts y_i perfectly (for all observations with $x_i > 0, y_i = 1$). In both panels, in addition to the RBML estimation results, those of the sieve ML, kernel ML, and SLS estimations

[12] For sieve ML estimation, I started with third-order Hermite polynomials. If convergence was not achieved, I reestimated the model by increasing the order by one repeatedly until the eighth order.

are presented for comparison.

Panel A shows that the estimation error of the RBML method is almost symmetrically distributed around zero, and the method is superior to other semiparametric methods in terms of the number of trials with convergence and the magnitude of the bias and RMSE. The kernel-based semiparametric methods (i.e., kernel ML and SLS) tend to underestimate the parameter, and the sieve ML method has a relatively smaller bias but slightly larger RMSE.

[Figure 2 near here]

Turning to Panel B, where the model has a *fully* perfect prediction problem and is designed to estimate the lower bound of the impact, the results show that the RBML method has a relatively smaller bias but a tendency for underestimation. Nonetheless, it can be observed that the proposed method is superior to other methods in terms of the number of trials with convergence and the magnitude of the RMSE. The kernel ML method provides an error distribution roughly centred at zero and has the smallest bias, but it fails to converge in more than half of the 500 trials, and the RMSE is also relatively large because some errors are outside the range shown in the figure. Thus, these results indicate that the RBML method is the best to employ when estimating models with a perfect prediction problem.

5. Conclusions

In this study, an innovative ML estimation method that requires no distributional assumption was proposed and formulated for linear regression and binary choice models. Furthermore, this study showed that the proposed estimator is consistent and normally distributed for large samples. In particular, the estimator attains an asymptotic efficiency bound for the linear regression model. For the binary choice model, the proposed estimator can be as efficient as the probit estimator when comparing the asymptotic variance of the *scale-normalized* parameter when the excess kurtosis of the regressors is positive and high.

I also evaluated the estimator in a Monte Carlo analysis for linear regression and binary choice models under several error distributions. The results showed that the RBML estimator performed exceedingly well for the small sample case. For the linear regression model, the RBML estimator was almost equivalent to the OLS estimator in terms of bias and the RMSE when the sample size of the resampled data (T) was large. For the binary choice model, the RBML estimator outperformed the probit and other semiparametric estimators when the linear combination of regressors was leptokurtic, even if the probit model was, in theory, expected to be the best. Furthermore, the RBML method was shown to avoid a typical perfect prediction problem and to estimate (the lower bound of) the impact of the variable causing the problem.

The new semiparametric ML method proposed in this study has potential in the field of behavioural and experimental social sciences, where discrete choice models are extensively utilized. While the recent trend in empirical fields is to require robust estimation against model misspecification, conventional semiparametric estimators are seldom employed in practice, probably because of their practical inconvenience. The proposed estimator can potentially bridge the gap between the needs in empirical fields

and the sparsity of well-performing practicable semiparametric estimators. Although the regressor and error distributions in the simulation showed only a few possible examples, the RBML estimator could be the first choice for binary choice models, especially in the case of leptokurtic regressors or in the presence of a perfect prediction problem.

Appendix A: Proofs and other results

A.1. Proof of Proposition 1

Taking M observations from $\{z_i | \ i = 1, \cdots, N\}$ by resampling with replacement (in RBML data construction) is equivalent to taking one observation at random from $\{z_i | \ i = 1, \cdots, N\}$ and repeating M times. Therefore, $\tilde{z}_t$ can be expressed as follows:

$$\tilde{z}_t \equiv \sqrt{\frac{NM}{N-1}}\left(\frac{\sum_{j=1}^{M} z_{jt}}{M} - \mu_N\right) = \sqrt{\frac{NM}{N-1}}\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ijt} z_i}{M} - \mu_N\right),$$

where w_{ijt} is a random variable that is one if the i-th observation is drawn at the j-th iteration in the t-th resampling stage and zero otherwise.

Then, treating w_{ijt} as a variable following a Bernoulli distribution with probability of $1/N$ (Cornfield, 1944; Raj and Khamis, 1958), $M^{-1}\sum_j^M w_{ijt}$ has a mean of $1/N$ and a variance of $(N-1)/(N^2M)$. Therefore,

$$\left(\frac{N-1}{N^2M}\right)^{-\frac{1}{2}}\left(\frac{\sum_j^M w_{ijt}}{M} - \frac{1}{N}\right) \xrightarrow{d} \mathrm{N}(0,1)$$

as M goes to infinity by the Lindeberg–Levy CLT. Alternatively, we can apply the de Moivre–Laplace theorem by treating $\sum_j^M w_{ijt}$ as a binomial random variable. Thus, assuming that N is given (and hence the sample $\{z_i | \ i = 1, \cdots, N\}$ is given) and M is sufficiently large,

$$\acute{z}_i = \left(\frac{N-1}{N^2M}\right)^{-\frac{1}{2}}\left(\frac{\sum_{j=1}^{M} w_{ijt}}{M} - \frac{1}{N}\right)(z_i - \mu_N) \sim \mathrm{N}(0, (z_i - \mu_N)^2). \tag{A-1}$$

Therefore, $\tilde{z}_t$ $(= N^{-1/2}\sum_{i=1}^{N} \acute{z}_i)$ follows $\mathrm{N}(0, \sigma_N^2)$, where $\sigma_N^2 = N^{-1}\sum_i^N (z_i - \mu_N)^2$.

The *i.i.d.* property of the new data comes from the fact that $\{\tilde{z}_t | \ t = 1, \cdots, T\}$ are uncorrelated joint normal random variables. When N (and the sample $\{z_i | \ i = 1, \cdots, N\}$) is given, the covariance of any two observations, $\tilde{z}_t$ and $\tilde{z}_s$ $(t, s \in \{1, \cdots, T\}$ and $t \neq s)$, is expressed as follows:

$$\mathrm{Cov}_w(\tilde{z}_t, \tilde{z}_s) = \mathrm{E}_w\left[\sqrt{\frac{N^2M}{N-1}}\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ijt} z_i}{M} - \mu_N\right) \cdot \sqrt{\frac{N^2M}{N-1}}\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ijs} z_i}{M} - \mu_N\right)\right]$$

$$= \frac{N^2M}{N-1}\mathrm{E}_w\left[\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ijt} z_i}{M} - \frac{\sum_{i=1}^{N} z_i}{N}\right)\left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ijs} z_i}{M} - \frac{\sum_{i=1}^{N} z_i}{N}\right)\right]$$

$$= \frac{N^2M}{N-1}\mathrm{E}_w\left[\left(\frac{\sum_i^N\sum_j^N W_{it}W_{js} z_i z_j}{M^2} - \frac{\sum_i^N\sum_j^N W_{it} z_i z_j}{NM} - \frac{\sum_i^N\sum_j^N W_{js} z_i z_j}{NM} + \frac{\sum_i^N\sum_j^N z_i z_j}{N^2}\right)\right]$$

$$= \frac{N^2 M}{N-1}\left\{\frac{\mathrm{E}[W_{it}]\mathrm{E}[W_{js}]}{M^2} - \frac{\mathrm{E}[W_{it}]}{NM} - \frac{\mathrm{E}[W_{js}]}{NM} + \frac{1}{N^2}\right\}\sum_{i}^{N}\sum_{j}^{N} z_i z_j$$

$$= \frac{M}{N-1}(1-1-1+1)\sum_{i}^{N}\sum_{j}^{N} z_i z_j = 0,$$

where W_{ik} is defined as $\sum_{j=1}^{M} w_{ijk}$ and represents the number of times that z_i is drawn at the k-th resampling stage ($k = t, s$). $\mathrm{E}_w[\cdot]$ means that the expectation is taken solely with respect to the distribution of random weights w_{ijk}, given the other random variables in the expectation operator. Then, since w_{ijk} is a Bernoulli random variable, $\mathrm{E}[W_{ik}] = \mathrm{E}\left[\sum_{j=1}^{M} w_{ijk}\right] = M/N$. Therefore, $\tilde{\mathbf{z}} = (\tilde{z}_1, \cdots, \tilde{z}_T)$ follows $\mathrm{N}(\mathbf{0}, \sigma_N^2 \mathbf{I}_T)$, where $\mathbf{I}_T$ is a $T \times T$ identity matrix, and hence, the joint normal density can be expressed as the product of individual normal densities: $f(\tilde{\mathbf{z}}) = 1/\sqrt{(2\pi)^T |\sigma_N^2 \mathbf{I}_T|} \times \exp\{-(\tilde{\mathbf{z}})'(\sigma_N^2 \mathbf{I}_T)^{-1}(\tilde{\mathbf{z}})/2\} = \prod_{t=1}^{T} 1/\sqrt{2\pi\sigma_N^2} \times \exp\{-(\tilde{z}_t/\sigma_N)^2/2\} = \prod_{t=1}^{T} f(\tilde{z}_t)$, showing that $\{\tilde{z}_t | \, t = 1, \cdots, T\}$ are mutually independent.

The assumption that the original sample $\{z_i | \, i = 1, \cdots, N\}$ is *i.i.d.* with a finite mean and variance (μ_0, $\sigma_0^2 < \infty$) ensures that $\{\acute{z}_i | \, i = 1, \cdots, N\}$, where $\acute{z}_i$ is defined in Eq. (A-1), is *i.i.d.* with a mean of $\mathrm{E}[\acute{z}_i] = 0$ and a variance of $\mathrm{Var}[\acute{z}_i] = \sigma_0^2(N-1)/N$. Then, $\mathrm{E}[N^{-1}\sum_i^N \acute{z}_i] = 0$ and $\mathrm{Var}[N^{-1}\sum_i^N \acute{z}_i] = \sigma_0^2(N-1)/N^2$, and

$$\left\{\frac{\sigma_0^2(N-1)}{N^2}\right\}^{-\frac{1}{2}} \frac{\sum_i^N \acute{z}_i}{N} \xrightarrow{d} \mathrm{N}(0,1)$$

as N goes to infinity by the Lindeberg–Levy CLT. Therefore, applying the Slutzky theorem, $\tilde{z}_t$ ($= N^{-1/2}\sum_{i=1}^{N} \acute{z}_i$) converges in distribution to $\mathrm{N}(0, \sigma_0^2)$, and the conclusion is obtained.

A.2. Proof of Theorem 1

From condition (ii), $\operatorname{plim}_{N\to\infty} |T^{-1}\sum_{t=1}^{T} \ln p(\tilde{z}_t|\theta) - \mathrm{E}[\ln p(\tilde{z}|\theta)]| < \epsilon/3$. Note that $T = T' + N$ as in Definition 1 (RBML data construction). Therefore,

$$\mathrm{E}\left[\ln p\left(\tilde{z}|\hat{\theta}_{RB}\right)\right] > \frac{\sum_{t=1}^{T} \ln p\left(\tilde{z}_t|\hat{\theta}_{RB}\right)}{T} - \frac{\epsilon}{3}, \tag{A-2}$$

$$\frac{\sum_{t=1}^{T} \ln p(\tilde{z}_t|\theta_0)}{T} - \frac{2\epsilon}{3} > \mathrm{E}[\ln p(\tilde{z}|\theta_0)] - \epsilon, \tag{A-3}$$

with probability approaching one (w.p.a.1). In addition, by Definition 2 (RBML estimator), as expressed in Eq. (1),

$$\frac{\sum_{t=1}^{T} \ln p\left(\tilde{z}_t|\hat{\theta}_{RB}\right)}{T} > \frac{\sum_{t=1}^{T} \ln p(\tilde{z}_t|\theta_0)}{T} - \frac{\epsilon}{3}, \tag{A-4}$$

Hence, adding both sides of inequalities (A-2), (A-3), and (A-4) and arranging the expression yields

$$\mathrm{E}\left[\ln p\left(\tilde{z}|\hat{\theta}_{RB}\right)\right] > \mathrm{E}[\ln p(\tilde{z}|\theta_0)] - \epsilon,$$

for any $\epsilon > 0$, w.p.a.1. Then, let $\mathcal{N}$ be any open subset of Θ containing θ_0, and let $\mathcal{N}^c$ be the complement of $\mathcal{N}$. By $\Theta \cap \mathcal{N}^c$ compact (by condition (i) and the definition of $\mathcal{N}^c$) and conditions (iii) and (iv), $\sup_{\theta \in \Theta \cap \mathcal{N}^c} \mathrm{E}[\ln p(\tilde{z}|\theta)] = \mathrm{E}[\ln p(\tilde{z}|\theta^*)] < \mathrm{E}[\ln p(\tilde{z}|\theta_0)]$ for some $\theta^* \in \Theta \cap \mathcal{N}^c$. Therefore, by choosing $\epsilon = \mathrm{E}[\ln p(\tilde{z}|\theta_0)] - \sup_{\theta \in \Theta \cap \mathcal{N}^c} \mathrm{E}[\ln p(\tilde{z}|\theta)]$, it follows that

$$\mathrm{E}\left[\ln p\left(\tilde{z}|\hat{\theta}_{RB}\right)\right] > \sup_{\theta \in \Theta \cap \mathcal{N}^c} \mathrm{E}[\ln p(\tilde{z}|\theta)]$$

and thus, $\hat{\theta}_{RB} \in \mathcal{N}$ w.p.a.1 as N goes to infinity.

A.3. Proof of Theorem 2

Let $\hat{1}$ be the {0,1}-valued indicator function for the event that $\hat{\theta}_{RB} \in \mathcal{N}$. Note that $\hat{1} \xrightarrow{p} 1$ by $\hat{\theta}_{RB} \xrightarrow{p} \theta_0$ (as $N \to \infty$) from Theorem 1. From condition (ii) and the first-order conditions for a maximum, $0 = \hat{1} \cdot \nabla_\theta \ln L\left(\hat{\theta}_{RB} \middle| \tilde{\mathbf{z}}\right)$. The mean value theorem applied to each row (denoted by k) of the right-hand side yields

$$0 = \hat{1} \cdot \nabla_\theta \ln L\left(\hat{\theta}_{RB} \middle| \tilde{\mathbf{z}}\right)_k = \hat{1} \cdot \frac{\sum_{t=1}^{T} \nabla_\theta \ln p\left(\tilde{z}_t \middle| \hat{\theta}_{RB}\right)_k}{T}$$

$$= \hat{1} \cdot \frac{\left\{\sum_{t=1}^{T} \nabla_\theta \ln p(\tilde{z}_t|\, \theta_0)_k + \sum_{t=1}^{T} \nabla_{\theta\theta} \ln p(\tilde{z}_t \middle| \bar{\theta}_k)_k^{\mathrm{T}} \left(\hat{\theta}_{RB} - \theta_0\right)\right\}}{T},$$

where $\bar{\theta}_k$ denotes a random variable equal to a mean value located between θ_0 and $\hat{\theta}_{RB}$ and converges in probability to θ_0. Let g denote the vector with the k-th row $T^{-1} \sum_t^T \nabla_\theta \ln p(\tilde{z}_t|\, \theta_0)_k$, and let $\bar{H}$ denote the matrix with the k-th row $T^{-1} \sum_{t=1}^{T} \nabla_{\theta\theta} \ln p(\tilde{z}_t \middle| \bar{\theta}_k)_k^{\mathrm{T}}$. In addition, let $\bar{1}$ be an indicator variable that takes unity if $\hat{\theta}_{RB} \in \mathcal{N}$ and $\bar{H}$ is nonsingular and zero otherwise. Then, from $\bar{1} \xrightarrow{p} 1$ and $0 = \bar{1} \cdot g + \bar{1} \cdot \bar{H}\left(\hat{\theta}_{RB} - \theta_0\right)$, $\sqrt{N}\left(\hat{\theta}_{RB} - \theta_0\right) = \bar{1} \cdot \bar{H}^{-1} \cdot \sqrt{N} g + (1 - \bar{1})\sqrt{N}\left(\hat{\theta}_{RB} - \theta_0\right)$. By $\bar{\theta}_k \xrightarrow{p} \theta_0$ and condition (iv), $\bar{H} \xrightarrow{p} \mathrm{E}[\nabla_{\theta\theta} \ln p(\tilde{z}|\theta_0)] = H$. Furthermore, $H = -J$ by differentiating $\int p(\tilde{z} \mid \theta) dz$ twice and interchanging the order of differentiation and integration (by conditions (ii) and (v), applying Lemma 3.6 of Newey and McFadden (1994)). Then, $\sqrt{N} g \xrightarrow{d} \mathrm{N}(0, J)$ by $g \xrightarrow{p} \mathrm{E}[\nabla_\theta p(\tilde{z} \mid \theta_0)] = 0$, condition (iii), and the Lindeberg–Levy CLT. Additionally, $(1 - \bar{1})\sqrt{N}\left(\hat{\theta}_{RB} - \theta_0\right) \xrightarrow{p} 0$ by $\bar{1} \xrightarrow{p} 1$. Therefore, by the nonsingularity of J, the conclusion follows from the Slutzky theorem.

A.4. Proposition A-1 and its proof

Proposition A1 (Distribution properties of new data with heteroscedasticity)

Assume that the original data $\{(y_i, \mathbf{x}_i) | \, i = 1, \cdots, N\}$ *are independent with finite means and variances. The regression model is expressed by* $y_i = \alpha_0 + \mathbf{x}_i \boldsymbol{\beta}_0 + \varepsilon_i$*, as defined in the main text.* $\{\varepsilon_i | \, i = 1, \cdots, N\}$ *are independent with* $\mathrm{E}[\varepsilon_i] = 0$ *and* $\mathrm{Var}[\varepsilon_i] = \sigma_i^2$*. Further assume that* $\lim_{N \to \infty} \sum_{i=1}^{N} \mathrm{E}|\varepsilon_i|^{2+\delta} / \left(\sum_{i=1}^{N} \sigma_i^2\right)^{1+\delta/2} = 0$ *for some* $\delta > 0$*. Then, letting* $\sigma_N^2 = N^{-1} \sum_i^N \varepsilon_i^2$ *and* $\sigma_0^2 = \lim_{N \to \infty} N^{-1} \sum_i^N \sigma_i^2$*,* $\{\tilde{\varepsilon}_t | \, t = 1, \cdots, T\}$ *obtained in the*

first (data construction) step with sufficiently large M are i.i.d. with $\mathrm{N}(0,\sigma_N^2)$ *and converge in distribution to* $\mathrm{N}(0,\sigma_0^2)$ *as N goes infinity:*

$$\tilde{\varepsilon}_t \overset{i.i.d.}{\sim} \mathrm{N}(0,\sigma_N^2) \xrightarrow{d} \mathrm{N}(0,\sigma_0^2).$$

Note that in the above proposition, the variances can be dependent on $\mathbf{x}_i$, that is, $\sigma_i^2 = h_i(\mathbf{x}_i)$. In this case, the issue of conditional heteroscedasticity arises for the original error term (ε_i). However, the new error term ($\tilde{\varepsilon}_i$) obtained through RBML data construction is independent of $\tilde{\mathbf{x}}_i$ and is free from heteroscedasticity (see the discussion in Section 3.2 in Ito (2024)). The proof of Proposition A-1 is presented below.

Proof of Proposition A-1

The result that $\tilde{\varepsilon}_t \overset{i.i.d.}{\sim} \mathrm{N}(0,\sigma_N^2)$ follows from the same discussion in Appendix A.1 by replacing $\tilde{z}_t$ with $\tilde{\varepsilon}_t$. Then, similar to Eq. (A-1), $\acute{\varepsilon}_i$ is defined as follows:

$$\acute{\varepsilon}_i = \left(\frac{N-1}{N^2M}\right)^{-\frac{1}{2}} \left(\frac{\sum_{j=1}^M w_{ijt}}{M} - \frac{1}{N}\right)(\varepsilon_i - \mu_N), \tag{A-5}$$

with $\mathrm{E}[\acute{\varepsilon}_i] = 0$ and $\mathrm{Var}[\acute{\varepsilon}_i] = \sigma_i^2$. Then, $\{\acute{\varepsilon}_i |\ i = 1,\cdots,N\}$ satisfies the Lindeberg condition as shown below.

$$\begin{aligned}
\lim_{N\to\infty} \frac{\sum_{i=1}^N \mathrm{E}|\acute{\varepsilon}_i|^{2+\delta}}{(\sum_{i=1}^N \sigma_i^2)^{1+\delta/2}} &= \lim_{N\to\infty} \frac{\sum_{i=1}^N \mathrm{E}\left|\sqrt{\frac{N^2}{(N-1)M}}\left(\sum_{j=1}^M w_{ijt} - \frac{M}{N}\right)(\varepsilon_i - \mu_N)\right|^{2+\delta}}{(\sum_{i=1}^N \sigma_i^2)^{1+\delta/2}} \\
&= \lim_{N\to\infty} \frac{\left\{\frac{N^2}{(N-1)M}\right\}^{1+\frac{\delta}{2}} \sum_{i=1}^N \mathrm{E}\left|W_{it} - \frac{M}{N}\right|^{2+\delta} \mathrm{E}|\varepsilon_i - \mu_N|^{2+\delta}}{(\sum_{i=1}^N \sigma_i^2)^{1+\delta/2}} \\
&= \lim_{N\to\infty} \left[\left\{\frac{N^2}{(N-1)M}\right\}^{1+\frac{\delta}{2}} \mathrm{E}\left|W_{it} - \frac{M}{N}\right|^{2+\delta} \frac{\sum_{i=1}^N \mathrm{E}|\varepsilon_i - \mu_N|^{2+\delta}}{(\sum_{i=1}^N \sigma_i^2)^{1+\delta/2}}\right] \\
&= \lim_{N\to\infty} \left[\left\{\frac{N}{(N-1)}\right\}^{1+\frac{\delta}{2}} \left(\frac{N}{M}\right)^{1+\frac{\delta}{2}} \mathrm{E}\left|W_{it} - \frac{M}{N}\right|^{2+\delta} \frac{\sum_{i=1}^N \mathrm{E}|\varepsilon_i - \mu_N|^{2+\delta}}{(\sum_{i=1}^N \sigma_i^2)^{1+\delta/2}}\right] \\
&= \lim_{N\to\infty} \left[\left\{\frac{N}{(N-1)}\right\}^{1+\delta/2} \mathrm{E}\left|W_{it}\left(\frac{N}{M}\right)^{\frac{1}{2}} - \left(\frac{M}{N}\right)^{\frac{1}{2}}\right|^{2+\delta} \frac{\sum_{i=1}^N \mathrm{E}|\varepsilon_i - \mu_N|^{2+\delta}}{(\sum_{i=1}^N \sigma_i^2)^{1+\delta/2}}\right],
\end{aligned}$$

where $W_{it} = \sum_{j=1}^M w_{ijt}$, which can be viewed as a binomial random variable with $\mathrm{E}[W_{it}] = M/N$, $\mathrm{E}[W_{it} - M/N]^2 = M(N-1)/N^2$, and $\mathrm{E}[W_{it} - M/N]^4 = 3M(M-2)(N-1)^2/N^4$. Thus, we have

$$\mathrm{E}\left|W_{it}\left(\frac{N}{M}\right)^{\frac{1}{2}} - \left(\frac{M}{N}\right)^{\frac{1}{2}}\right|^{2+\delta}$$

$$= \int_{\left|W-\frac{M}{N}\right|<\left(\frac{M}{N}\right)^{\frac{1}{2}}} \left| W\left(\frac{N}{M}\right)^{\frac{1}{2}} - \left(\frac{M}{N}\right)^{\frac{1}{2}} \right|^{2+\delta} \mathrm{d}F(W) + \int_{\left|W-\frac{M}{N}\right|\geq\left(\frac{M}{N}\right)^{\frac{1}{2}}} \left| W\left(\frac{N}{M}\right)^{\frac{1}{2}} - \left(\frac{M}{N}\right)^{\frac{1}{2}} \right|^{2+\delta} \mathrm{d}F(W)$$

$$\leq \int_{\left|W-\frac{M}{N}\right|<\left(\frac{M}{N}\right)^{\frac{1}{2}}} \left\{ W\left(\frac{N}{M}\right)^{\frac{1}{2}} - \left(\frac{M}{N}\right)^{\frac{1}{2}} \right\}^{2} \mathrm{d}F(W) + \int_{\left|W-\frac{M}{N}\right|\geq\left(\frac{M}{N}\right)^{\frac{1}{2}}} \left\{ W\left(\frac{N}{M}\right)^{\frac{1}{2}} - \left(\frac{M}{N}\right)^{\frac{1}{2}} \right\}^{4} \mathrm{d}F(W)$$

$$\leq \mathrm{E}\left[W_{it}\left(\frac{N}{M}\right)^{\frac{1}{2}} - \left(\frac{M}{N}\right)^{\frac{1}{2}} \right]^{2} + \mathrm{E}\left[W_{it}\left(\frac{N}{M}\right)^{\frac{1}{2}} - \left(\frac{M}{N}\right)^{\frac{1}{2}} \right]^{4}$$

$$= \frac{N}{M} \cdot \frac{M(N-1)}{N^2} + \left(\frac{N}{M}\right)^{2} \cdot \frac{3M(M-2)(N-1)^2}{N^4}$$

$$= \frac{N-1}{N} + \frac{3(M-2)}{M} \cdot \left(\frac{N-1}{N}\right)^{2},$$

and therefore, $\lim_{N\to\infty} \{N/(N-1)\}^{1+\delta/2} \mathrm{E}\left| W_{it}(N/M)^{1/2} - (M/N)^{1/2} \right|^{2+\delta} < 4$. Note that in the above derivation, we implicitly assume that $\delta < 1$ (Rao (1973, pp. 127–128) presents the special case where $\delta = 1$, but here δ can be arbitrarily small to reduce the moment conditions). Then, since $\lim_{N\to\infty} \sum_i^N \mathrm{E}|\varepsilon_i - \mu_N|^{2+\delta} / \left(\sum_i^N \sigma_i^2\right)^{1+\delta/2} = \lim_{N\to\infty} \sum_i^N \mathrm{E}|\varepsilon_i|^{2+\delta} / \left(\sum_i^N \sigma_i^2\right)^{1+\delta/2} = 0$, we also have $\lim_{N\to\infty} \sum_{i=1}^N \mathrm{E}|\acute{\varepsilon}_\iota|^{2+\delta} / \left(\sum_{i=1}^N \sigma_i^2\right)^{1+\delta/2} = 0$, indicating that the Lindeberg condition is satisfied (see White, 2001, p. 119).

The mean and variance of $N^{-1}\sum_{i=1}^N \acute{\varepsilon}_i$ are $\mathrm{E}[N^{-1}\sum_{i=1}^N \acute{\varepsilon}_i] = 0$ and $\mathrm{Var}[N^{-1}\sum_{i=1}^N \acute{\varepsilon}_i] = N^{-2}\sum_{i=1}^N \sigma_i^2$, respectively, and thus by the Lyapunov CLT,

$$\left\{ \frac{\sum_{i=1}^N \sigma_i^2}{N^2} \right\}^{-\frac{1}{2}} \frac{\sum_{i=1}^N \acute{\varepsilon}_i}{N} \xrightarrow{d} \mathrm{N}(0,1)$$

as N goes to infinity. Then, applying the Slutzky theorem, $\tilde{\varepsilon}_t$ $(= N^{-1/2}\sum_i^N \acute{\varepsilon}_i)$ converges in distribution to $\mathrm{N}(0, \sigma_0^2)$, where $\sigma_0^2 = \lim_{N\to\infty} N^{-1}\sum_{i=1}^N \sigma_i^2$.

A.5. Derivation of Eq. (4)

Let $\tilde{y}_t$ be an outcome at the t-th resampling stage in the RBML data construction, which can be expressed as

$$\tilde{y}_t = \sqrt{\frac{NM}{N-1}} \left(\frac{1}{M}\sum_{j=1}^M y_{jt} - \bar{y}_N^* \right) = \sqrt{\frac{NM}{N-1}} \left(\frac{1}{M}\sum_{i=1}^N \sum_{j=1}^M w_{ijt} y_i - \frac{1}{N} \left(\sum_{y_i \in \mathcal{Y}_0} y_i + \sum_{y_i \in \mathcal{Y}_1} y_i \right) \right)$$

$$= \sqrt{\frac{NM}{N-1}} \left\{ \frac{1}{M} \left(\sum_{y_i \in \mathcal{Y}_0} \sum_{j=1}^M w_{ijt} y_i + \sum_{y_i \in \mathcal{Y}_1} \sum_{j=1}^M w_{ijt} y_i \right) - \left(n_{N,0}\bar{y}_{N,0} + n_{N,1}\bar{y}_{N,1} \right) \right\},$$

where for $a \in \{0,1\}$, $\mathcal{Y}_a = \{y_i |\, i = 1, \cdots, N \;\&\; d_i = a\}$, $\bar{y}_{N,a} = N_a^{-1}\sum_{y_i \in \mathcal{Y}_a} y_i$, $n_{N,a} = N_a/N$, and $N_a =$

$\#[\mathcal{Y}_a]$ (the number of observations in the original sample for which $d_i = a$). Then, defining γ_t as

$$\gamma_t = N^{-\frac{1}{2}} \sum_{y_i \in \{0\}} \left\{ \left(\frac{N-1}{N^2 M} \right)^{-\frac{1}{2}} \left(\frac{\sum_{j=1}^{M} w_{ijt}}{M} - \frac{1}{N} \right) \left(y_i - \bar{y}_{N,0} \right) \right\}$$

$$+ N^{-\frac{1}{2}} \sum_{y_i \in \{1\}} \left\{ \left(\frac{N-1}{N^2 M} \right)^{-\frac{1}{2}} \left(\frac{\sum_{j=1}^{M} w_{ijt}}{M} - \frac{1}{N} \right) \left(y_i - \bar{y}_{N,1} \right) \right\},$$

the above equation for $\tilde{y}_t$ is rewritten as

$$\tilde{y}_t = \gamma_t + \sqrt{\frac{NM}{N-1}} \left\{ m_{0,t} \bar{y}_{N,0} + m_{1,t} \bar{y}_{N,1} - \left(n_0 \bar{y}_{N,0} + n_1 \bar{y}_{N,1} \right) \right\}$$

$$= \gamma_t + \sqrt{\frac{NM}{N-1}} \left\{ \left(1 - m_{1,t} \right) \bar{y}_{N,0} + m_{1,t} \bar{y}_{N,1} - \left(\left(1 - n_{N,1} \right) \bar{y}_{N,0} + n_{N,1} \bar{y}_{N,1} \right) \right\}$$

$$= \gamma_t + \sqrt{\frac{NM}{N-1}} \left(m_{1,t} - n_{N,1} \right) \left(\bar{y}_{N,1} - \bar{y}_{N,0} \right), \tag{A-6}$$

where for $a \in \{0,1\}$, $m_{a,t} = M_{a,t}/M$ ($M_{a,t}$ is the number of draws from $\mathcal{Y}_a$ at the t-th resampling stage). Note that $\gamma_t \overset{i.i.d.}{\sim} \mathrm{N}\left(0, n_{N,0}\sigma_{N,0}^2 + n_{N,1}\sigma_{N,1}^2\right)$, where $\sigma_{N,a}^2 = N_a^{-1} \sum_{y_i \in \mathcal{Y}_a} \left(y_i - \bar{y}_{N,a} \right)^2$ for $a \in \{0,1\}$. If $\{y_i | \, i = 1, \cdots, N\}$ are *i.i.d.*, applying Proposition 1 with the finite variance assumption ($\mathrm{Var}[y_i] < \infty$), we obtain the result that $\gamma_t \xrightarrow{d} \mathrm{N}(0, n_0\sigma_0^2 + n_1\sigma_1^2)$ as $N \to \infty$, where $\sigma_a^2 = \mathrm{Var}(y_i | \, d_i = a)$, and n_a is the proportion of observations in the population for which $d_i = a$. In addition, when $\{y_i | \, i = 1, \cdots, N\}$ are independent and not identically distributed (*i.n.i.d*), by applying Proposition A1 with the additional assumption that $\lim_{N \to \infty} \sum_{i=1}^{N} \mathrm{E}|y_i|^{2+\delta} / \left(\sum_{i=1}^{N} \mathrm{Var}(y_i) \right)^{1+\delta/2} = 0$, we have $\gamma_t \xrightarrow{d} \mathrm{N}(0, n_0\sigma_0^2 + n_1\sigma_1^2)$ as $N \to \infty$, where $\sigma_a^2 = \lim_{N \to \infty} N_a^{-1} \sum_{y_i \in \mathcal{Y}_a} \mathrm{Var}(y_i)$ for $a \in \{0,1\}$.

The last expression in Eq. (A-6) implies that when $\tilde{y}_t > \gamma_t$, since $\bar{y}_{N,1} > 0 > \bar{y}_{N,0}$, we have $\left(m_{1,t} - n_{N,1} \right) > 0$, which means that more observations are taken from $\mathcal{Y}_1 = \{ y_i \mid i = 1, \cdots, N \ \& \ d_i = 1 \}$ at the t-th resampling stage than those in the sample, and therefore $\tilde{d}_t > 0$. On the other hand, when $\tilde{y}_t \leq \gamma_t$, we have $\left(m_{1,t} - n_{N,1} \right) \leq 0$ and $\tilde{d}_t \leq 0$. Therefore, the introduction of a threshold variable γ_t yields Eq. (4).

References

Ai, C. (1997). A semiparametric maximum likelihood estimator. *Econometrica*, 65(4), 933–963. https://doi.org/10.2307/2171945

Amemiya, T. (1985). *Advanced econometrics*. Harvard University Press.

Barker, M. (2014). *SLS: Stata module to perform semiparametric least squares. Statistical software*

components, S457927. Boston College Department of Economics.

Cornfield, J. (1944). On samples from finite populations. *Journal of the American Statistical Association*, *39*(226), 236–239. https://doi.org/10.1080/01621459.1944.10500680

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica*, *51*(3), 765–782. https://doi.org/10.2307/1912157

De Luca, G. (2008). SNP and SML estimation of univariate and bivariate binary-choice models. *The Stata Journal*, *8*(2), 190–220. https://doi.org/10.1177/1536867x0800800203

Duncan, G. M. (1986). A semi-parametric censored regression estimator. *Journal of Econometrics*, *32*(1), 5–34. https://doi.org/10.1016/0304-4076(86)90010-2

Fan, J., Farmen, M., & Gijbels, I. (1998). Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *60*(3), 591–608. https://doi.org/10.1111/1467-9868.00142

Fernandez, L. (1986). Non-parametric maximum likelihood estimation of censored regression models. *Journal of Econometrics*, *32*(1), 35–57. https://doi.org/10.1016/0304-4076(86)90011-4

Gallant, A. R., & Nychka, D. W. (1987). Semi-nonparametric maximum likelihood estimation. *Econometrica*, *55*(2), 363–390. https://doi.org/10.2307/1913241

Gourieroux, C., Monfort, A., & Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, *52*(3), 681–700. https://doi.org/10.2307/1913471

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In J. Neyman (Ed.), *Proceedings of the fifth Berkeley symposium* (pp. 221–233). University of California Press.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, *58*(1-2), 71–120. https://doi.org/10.1016/0304-4076(93)90114-k

Ichimura, H. & Thompson, T. S. (1998). Maximum likelihood estimation of a binary choice model with random coefficients of unknown distribution. *Journal of Econometrics*, 86(2), 269–295. https://doi.org/10.1016/S0304-4076(97)00117-6.

Ito, T. (2024). *Binary and ordered response models in randomized experiments: Applications of the resampling-based maximum likelihood method*. Unpublished Manuscript, Kobe University. (available at: http://dx.doi.org/10.2139/ssrn.3976010)

Klein, R. W., & Spady, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica*, *61*(2), 387–421. https://doi.org/10.2307/2951556

Lee, L.-F. (1995). Semiparametric maximum likelihood estimation of polychotomous and sequential choice models. *Journal of Econometrics*, 65(2), 381–428. https://doi.org/10.1016/0304-4076(93)01591-9

McCullagh, P., & Nelder, J. A. (1983). *Generalized linear models*. Chapman & Hall.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370–384. https://doi.org/10.2307/2344614

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In R. F. Engle & D. McFadden (Eds.), *Handbook of econometrics* (pp. 2111–2245). Elsevier.

Raj, D., & Khamis, S. H. (1958). Some remarks on sampling with replacement. *The Annals of Mathematical Statistics*, *29*(2), 550–557. https://doi.org/10.1214/aoms/1177706630

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, Second Edition. John Wikey & Sons, Inc.

Ruud, P. A. (1983). Sufficient conditions for the consistency of maximum likelihood estimation despite misspecification of distribution in multinomial discrete choice models. *Econometrica*, *51*(1), 225–228. https://doi.org/10.2307/1912257

Tibshirani, R., & Hastie, T. (1987). Local likelihood estimation. *Journal of the American Statistical Association*, *82*(398), 559–567. https://doi.org/10.1080/01621459.1987.10478466

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method. *Biometrika*, *61*(3), 439–447. https://doi.org/10.1093/biomet/61.3.439

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, *50*(1), 1–25. https://doi.org/10.2307/1912526

White, H. (2001) *Asymptotic Theory for Econometricians*, Revised Edition. Emerald Group Publishing Ltd.

Tables & Figures

Table 1. Model description

A) Model 1: Linear regression model

Dependent variable: $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$

Regressors: $x_i, z_i \sim \chi^2(16)/8$

Parameters: $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$

Error:

(a) Standard normal: $\varepsilon_i \sim \mathrm{N}(0,1)$

(b) Normal mixture (left skewed, leptokurtic):
$\varepsilon_i \sim 0.6\mathrm{N}(-0.3, 1.225) + 0.4\mathrm{N}(0.45, 0.325)$

(c) Normal with heteroscedasticity:
$\varepsilon_i \sim \mathrm{N}\left(0, \frac{(\beta_1 x_i + \beta_2 z_i)^2}{\mathrm{E}[(\beta_1 x_i + \beta_2 z_i)^2]}\right)$

(d) Student's t: $\varepsilon_i \sim \mathrm{T}(2)$

B) Model 2: Binary choice model

Dependent variable: $d_i = 1[y_i > 0] = 1[\beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i > 0]$

Regressors:

$$x_i, z_i \sim \begin{cases} \text{(i)-(vi)} & \chi^2(a)/2\sqrt{a} \\ \text{(vii)} & \mathrm{N}(3, 0.5) \\ \text{(vii)-(xi)} & \mathrm{Beta}(b,b) \times \sqrt{4b+2} \end{cases},$$

where $a \in \{3,4,6,8,12,24\}$ and $b \in \{4.5, 1.5, 0.5, 10^{-10}\}$

Parameters: $\beta_0 = -\mathrm{Med}(\beta_1 x_i + \beta_2 z_i)$ and $\beta_1 = \beta_2 = 1$

Error: Same as Model 1 (See Panel A above)

C) Model 3: Binary choice model with perfect prediction problem

Dependent variable: $d_i = 1[y_i > 0] = 1[\beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i > 0]$

Regressors:

$x_i = 1[a_i > 1/2]$ and $z_i \sim \mathrm{N}(0, 0.2)$,
where $a_i \sim \mathrm{Uniform}(0,1)$

Parameters:

$\beta_0 = 0$, $\beta_2 = 1$, and

$$\beta_1 = \begin{cases} \text{(i) nearly perfect prediction: } \Phi^{-1}(0.99) \approx 2.326 \\ \text{(ii) fully perfect prediction: } |\mathrm{Min}(\beta_0 + \beta_2 z_i + \varepsilon_i)| + 0.01 \end{cases}$$

Error: Normal: $\varepsilon_i \sim \mathrm{N}(0, 0.8)$

Table 2. Simulation results for Model 1 (linear regression model)

		(1)		(2)		(3)		(4)	
Error design:		(a) Standard normal		(b) Normal mixture (left skewed, leptokurtic)		(c) Normal with heteroscedastic variance		(d) Student's t	
Estimator		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
RBML ($M = 100{,}000$, $T = 1{,}000$)	β_1	−0.0035	0.0770	−0.0033	0.0794	−0.0008	0.0838	−0.0214	0.2967
	β_2	−0.0019	0.0780	−0.0045	0.0825	−0.0059	0.0849	−0.0150	0.3144
RBML ($M = 100{,}000$, $T = 10{,}000$)	β_1	−0.0026	0.0657	−0.0030	0.0664	−0.0035	0.0720	−0.0317	0.2310
	β_2	−0.0021	0.0678	−0.0044	0.0704	−0.0035	0.0764	−0.0037	0.2408
RBML ($M = T = 100{,}000$)	β_1	−0.0026	0.0642	−0.0033	0.0643	−0.0033	0.0707	−0.0295	0.2219
	β_2	−0.0017	0.0658	−0.0034	0.0684	−0.0015	0.0756	−0.0018	0.2349
OLS	β_1	−0.0025	0.0642	−0.0031	0.0639	−0.0030	0.0705	−0.0288	0.2218
	β_2	−0.0018	0.0659	−0.0035	0.0682	−0.0016	0.0754	−0.0019	0.2344

Notes: "RMSE" stands for the root mean square error and "RBML" is the resampling-based ML proposed in this paper.

Source: Author's estimates using simulated data.

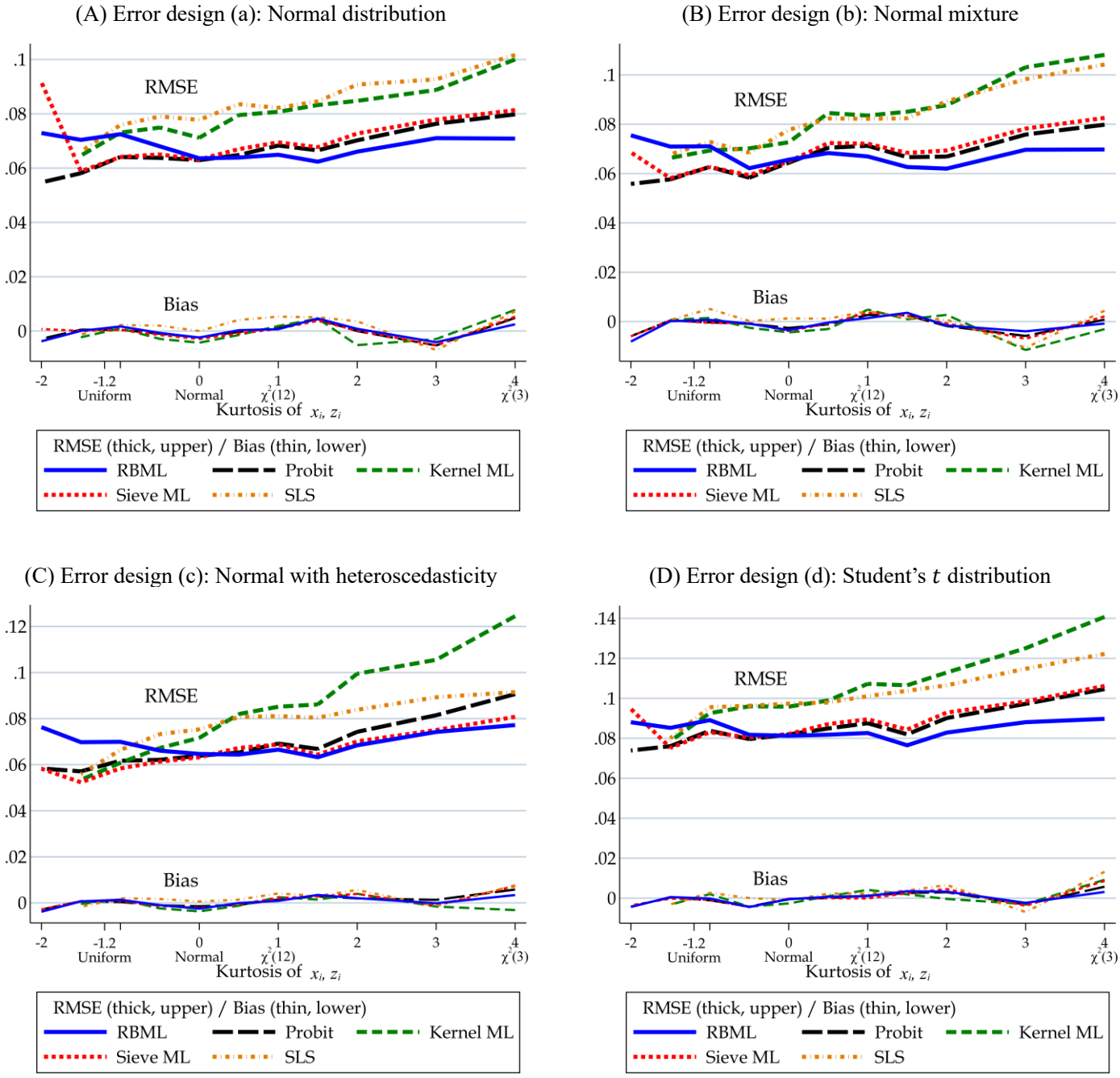
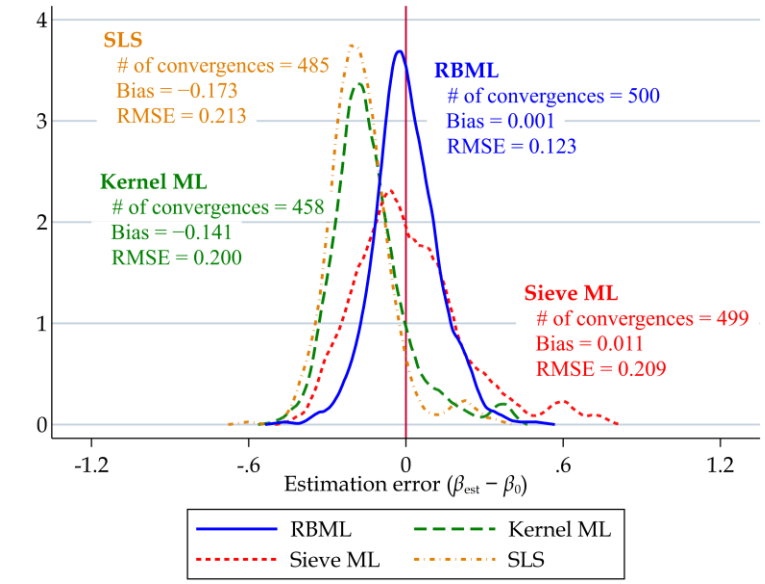


Figure 1. Simulation results for Model 2 (binary choice model)

Notes: The thick lines represent the root mean square errors (RMSEs), and the thin lines represent the biases. "RBML" is the resampling-based ML proposed in this study, "Kernel ML" is Klein and Spady's (1993) Nadaraya–Watson kernel ML, "sieve ML" is Gallant and Nychka's (1987) Hermite polynomial sieve ML, and "SLS" is Ichimura's (1993) semiparametric least squares.

Source: Author's estimates using simulated data.

(A) Nearly perfect prediction



(B) Fully perfect prediction

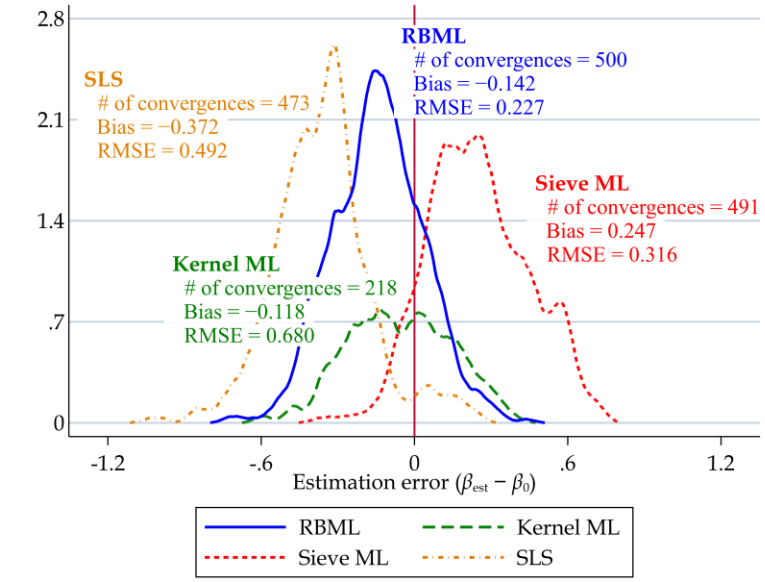


Figure 2. Simulation results for Model 3 (binary model with perfect prediction problem)

Notes: The figures show the distributions of the estimation errors. "RBML" is the resampling-based ML proposed in this paper, "sieve ML" is Gallant and Nychka's (1987) Hermite polynomial sieve ML, "Kernel ML" is Klein and Spady's (1993) Nadaraya–Watson kernel ML, and "SLS" is Ichimura's (1993) semiparametric least squares.

Source: Author's estimates using simulated data.

Online Supplementary Material for

"Resampling-Based Maximum Likelihood Estimation"

Takahiro Ito*

I. Stata program for the RBML estimators

This supplement describes how to obtain and use the estimation programs for the resampling-based maximum likelihood (RBML) estimators. The programs were written for Stata by the author and are available at http://www2.kobe-u.ac.jp/~takahiro/stata, using the –net– command, as follows.

First, in the Stata command line, specify the URL as:

```
. net from http://www2.kobe-u.ac.jp/~takahiro/stata
```

```
http://www2.kobe-u.ac.jp/~takahiro/stata/
Takahiro Ito, Kobe University

Here are some useful programs I have written

PACKAGES you could -net describe-:
    rbml              A program to implement the resampling-based maximum
                        likelihood (RBML) estimation.
```

Then, the description of the package is obtained:

* Graduate School of International Cooperation Studies, Kobe University, 2–1, Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan. E-mail: takahiro.ito@lion.kobe-u.ac.jp, Phone: +81-78-803-7148

```
. net describe rbml

------------------------------------------------------------------------------
package rbml from http://www2.kobe-u.ac.jp/~takahiro/stata
------------------------------------------------------------------------------

TITLE
      rbml.  Package to analyze data.

DESCRIPTION/AUTHOR(S)
      Program by Takahiro Ito.
      Other lines describing the package could appear here.

INSTALLATION FILES                                  (type net install rbml)
      rbml.ado
      rbml.hlp
      rbml_linear_lf2.ado
      rbml_binary_lf2.ado

ANCILLARY FILES                                     (type net get rbml)
      simulated_data.dta
------------------------------------------------------------------------------
```

Finally, the program package can be installed as follows:

```
. net install rbml
checking rbml consistency and verifying not already installed...
all files already exist and are up to date.
```

After installing the program, please see the help file for the syntax. The examples in the help file describe how to estimate the linear regression, binary choice and ordered response models via the RBML method.

```
. help rbml
```

II. The validity of the random threshold (γ_t) assumption

In this appendix, the validity of the assumption that γ_t is independent of $\tilde{\mathbf{x}}_t$ and $\tilde{\varepsilon}_t$ is discussed. In the absence of this assumption, we must consider the correlation among them. Specifically, we must calculate the conditional distribution of the new error component $\delta_\tau = (\gamma_t - \tilde{\varepsilon}_t)$ given $\tilde{\mathbf{x}}_t$ when estimating the log-likelihood in Eq. (5):

$$\delta_\tau | \tilde{\mathbf{x}}_t \sim \mathrm{N}(\tilde{\mathbf{x}}_t \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1} \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta}, \upsilon^2),$$

where $\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} = \mathrm{E}_w[\tilde{\mathbf{x}}_t' \tilde{\mathbf{x}}]$, $\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta} = \mathrm{E}_w[\tilde{\mathbf{x}}_t' \delta_t]$, and $\upsilon^2 = \sigma_\delta^2 - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta}' \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1} \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta}$. Since $\tilde{\mathbf{x}}_t$ and $\tilde{\varepsilon}_t$ can be decomposed into

two parts, similar to Eq. (B-1), we have

$$\tilde{\mathbf{x}}_t = \boldsymbol{\gamma}_{\mathbf{x},t} + \sqrt{M'}\left(m_{1,t} - n_1\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0),$$
$$\tilde{\varepsilon}_t = \gamma_{\varepsilon,t} + \sqrt{M'}\left(m_{1,t} - n_1\right)(\bar{\varepsilon}_1 - \bar{\varepsilon}_0),$$

where $\boldsymbol{\gamma}_{\mathbf{x},t} \sim \mathrm{N}\left(\mathbf{0}, n_0\boldsymbol{\Sigma}_{N,0} + n_1\boldsymbol{\Sigma}_{N,1}\right)$, $\boldsymbol{\Sigma}_{N,d} = N_d^{-1}\sum_{i\in\{d\}}\left(\mathbf{x}_{i,t} - \bar{\mathbf{x}}_d\right)^2$, $M' = NM/(N-1)$, $\bar{\mathbf{x}}_d = N_d^{-1}\sum_{i\in\{d\}}\mathbf{x}_i$, $\gamma_{\varepsilon,t} \sim \mathrm{N}\left(\mathbf{0}, n_0\sigma_{N,\varepsilon 1}^2 + n_1\sigma_{N,\varepsilon 0}^2\right)$, $\sigma_{N,\varepsilon d}^2 = N_d^{-1}\sum_{\{d\}}\left(\varepsilon_{i,t} - \bar{\varepsilon}_d\right)^2$, and $\bar{\varepsilon}_d = N_d^{-1}\sum_{i\in\{d\}}\varepsilon_i$ (for $d \in \{0,1\}$). Thus, δ_t is expressed as:

$$\begin{aligned}\delta_t &= \gamma_t - \tilde{\varepsilon}_t = \left(\boldsymbol{\gamma}_{\mathbf{x},t}\boldsymbol{\beta} + \gamma_{\varepsilon,t}\right) - \tilde{\varepsilon}_t \\ &= \boldsymbol{\gamma}_{\mathbf{x},t}\boldsymbol{\beta} + \left\{\tilde{\varepsilon}_t - \sqrt{M'}\left(m_{1,t} - n_1\right)(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)\right\} - \tilde{\varepsilon}_t \\ &= \boldsymbol{\gamma}_{\mathbf{x},t}\boldsymbol{\beta} - \sqrt{M'}\left(m_{1,t} - n_1\right)(\bar{\varepsilon}_1 - \bar{\varepsilon}_0).\end{aligned}$$

Note that $Mm_{1,t}$ can be regarded as a random number drawn from a binomial distribution $\mathrm{Bin}(M, n_1)$, which is approximated by $\mathrm{N}(n_1M, n_0n_1M)$ when M (the number of draws in each resampling stage) is sufficiently large. Therefore,

$$\begin{aligned}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta} &= \mathrm{E}_w[\tilde{\mathbf{x}}_t'\delta_t] = \mathrm{E}_w\left[\left\{\boldsymbol{\gamma}_{\mathbf{x},t} + \sqrt{M'}\left(m_{1,t} - n_1\right)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)\right\}'\left\{\boldsymbol{\gamma}_{\mathbf{x},t}\boldsymbol{\beta} - \sqrt{M'}\left(m_{1,t} - n_1\right)(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)\right\}\right] \\ &= \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{\beta} - \xi(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'(\bar{\varepsilon}_1 - \bar{\varepsilon}_0),\end{aligned}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} = \mathrm{E}_w\left[\boldsymbol{\gamma}_{\mathbf{x},t}'\boldsymbol{\gamma}_{\mathbf{x},t}\right] = n_0\boldsymbol{\Sigma}_{N,0} + n_1\boldsymbol{\Sigma}_{N,1}$, and $\xi = N(N-1)^{-1}n_0n_1$. The second equality comes from $\mathrm{E}\left[\sqrt{M'}\left(m_{1,t} - n_1\right)\right]^2 = M'\,\mathrm{E}\left[m_{1,t} - n_1\right]^2 = M'\,\mathrm{Var}\left[m_{1,t}\right] = N(N-1)^{-1}n_0n_1 = \xi$ and $\mathrm{E}_w\left[\boldsymbol{\gamma}_{\mathbf{x},t}{}^{\mathrm{T}}\sqrt{M'}\left(m_{1,t} - n_1\right)\right] = 0$. Therefore, we have $\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta} = \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{\beta} - \xi\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)$, $\sigma_\delta^2 = \mathrm{E}_w[\delta_t^2] = \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{\beta} + \xi(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)^2$, and

$$\begin{aligned}\upsilon^2 &= \sigma_\delta^2 - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta}'\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta} \\ &= \boldsymbol{\beta}'\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\left(\mathbf{I}_K - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\right)\boldsymbol{\beta} - 2\xi\boldsymbol{\beta}'\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'(\bar{\varepsilon}_1 - \bar{\varepsilon}_0) \\ &\quad + \xi(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)^2\{1 + \xi(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'\}\end{aligned} \tag{II-1}$$

Then, we eventually derive the following log-likelihood function:

$$\begin{aligned}&\ln L\left(\boldsymbol{\theta}^\upsilon \middle| \tilde{\mathbf{d}}, \tilde{\mathbf{X}}\right) \\ &= \sum_{t=1}^{T}\left[1\left[\tilde{d}_t \le 0\right]\ln\Phi\left(-\frac{\tilde{\mathbf{x}}_t(\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta})}{\upsilon}\right) + 1\left[\tilde{d}_t > 0\right]\ln\Phi\left(\frac{\tilde{\mathbf{x}}_t(\boldsymbol{\beta} - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta})}{\upsilon}\right)\right] \\ &= \sum_{t=1}^{T}\left[1\left[\tilde{d}_t \le 0\right]\ln\Phi\left(-\tilde{\mathbf{x}}_t\left\{\left(\mathbf{I}_K - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\right)\boldsymbol{\theta}^\upsilon + \xi\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'\frac{(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)}{\upsilon}\right\}\right)\right. \\ &\qquad \left. + 1\left[\tilde{d}_t > 0\right]\ln\Phi\left(\tilde{\mathbf{x}}_t\left\{\left(\mathbf{I}_K - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\right)\boldsymbol{\theta}^\upsilon + \xi\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'\frac{(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)}{\upsilon}\right\}\right)\right] \\ &= \sum_{t=1}^{T}\left[1\left[\tilde{d}_t \le 0\right]\ln\Phi\left(-\xi\tilde{\mathbf{x}}_t\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'\left\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)\boldsymbol{\theta}^\upsilon + \frac{(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)}{\upsilon}\right\}\right)\right. \\ &\qquad \left. + 1\left[\tilde{d}_t > 0\right]\ln\Phi\left(\xi\tilde{\mathbf{x}}_t\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'\left\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)\boldsymbol{\theta}^\upsilon + \frac{(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)}{\upsilon}\right\}\right)\right],\end{aligned}$$

where $\boldsymbol{\theta}^\upsilon = \boldsymbol{\beta}/\upsilon$. The third equality comes from $\left(\mathbf{I}_K - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\right) = \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\left(\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}} - \boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\right) = \xi\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)$. Therefore, when estimating the log-likelihood in Eq. (5):

$$\ln L\left(\boldsymbol{\eta} \mid \tilde{\mathbf{d}}, \tilde{\mathbf{X}}\right) = \sum_{t=1}^{T}\left[1\left[\tilde{d}_t \le 0\right]\ln\Phi(-\tilde{\mathbf{x}}_t\boldsymbol{\eta}) + 1\left[\tilde{d}_t > 0\right]\ln\Phi(\tilde{\mathbf{x}}_t\boldsymbol{\eta})\right],$$

we have the following relationship between $\boldsymbol{\eta}$ and $\boldsymbol{\theta}^{\upsilon}$:

$$\boldsymbol{\eta} = \xi\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)\boldsymbol{\theta}^{\upsilon} + (\bar{\varepsilon}_1 - \bar{\varepsilon}_0)/\upsilon\}. \tag{II-2}$$

Note that $(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)/\upsilon$ in the above equation can be calculated from Eq. (II-1) as:

$$e(\boldsymbol{\theta}) = \frac{(\bar{\varepsilon}_1 - \bar{\varepsilon}_0)}{\upsilon} = \frac{b \pm \sqrt{b^2 - ac}}{a}, \tag{II-3}$$

where $a = 1 + \xi(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'$, $b = \boldsymbol{\theta}'\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'$, and $c = \left\{\boldsymbol{\theta}'\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\left(\mathbf{I}_K - \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\right)\boldsymbol{\theta} - \mathbf{1}\right\}/\xi$.

In summary, if the random threshold assumption is true, the parameter to be estimated is $\boldsymbol{\theta}^{\tau}$ $(= \boldsymbol{\beta}/\tau)$, and it can be determined based on the above log-likelihood with $\boldsymbol{\eta} = \boldsymbol{\theta}^{\tau}$. On the other hand, if the assumption does not hold, the parameter to be estimated is $\boldsymbol{\theta}^{\upsilon}$ $(= \boldsymbol{\beta}/\upsilon)$. Therefore, we need to estimate the parameter based on Eq. (II-2). Unfortunately, however, $\boldsymbol{\theta}^{\upsilon}$ cannot be identified without an additional assumption because $\text{Rank}[(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)] = 1$ $(\neq K)$ and $\xi\,\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\,(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)$ is not invertible. Thus, it is impossible to obtain K solutions from one independent equation. A natural example of such an assumption to identify $\boldsymbol{\theta}^{\upsilon}$ is Assumption II-1 below.

Assumption II-1: Proportionality between $\boldsymbol{\eta}$ and $\boldsymbol{\theta}^{\upsilon}$

θ_k is proportional to η_k in the same ratio for all $k = 1, \cdots K$:

$$\rho\boldsymbol{\eta} = \boldsymbol{\theta}^{\upsilon}, \exists\ \rho \in \mathbb{R}.$$

Then, based on the above assumption, we can solve for $\boldsymbol{\theta}^{\upsilon}$ by, for example, the Newton–Raphson method as follows.

Iteration with the Newton–Raphson method

Step 1. Let ρ^s be the s-th value such that $\rho^s\boldsymbol{\eta} = \boldsymbol{\theta}^{\upsilon,s}$, where $\boldsymbol{\theta}^{\upsilon,s}$ is the s-th guess of $\boldsymbol{\theta}^{\upsilon}$. Compute the next guess (ρ^{s+1}) based on:

$$\rho^{s+1} = \rho^s - K^{-1}\sum_{k=1}^{K}\frac{g_k(\rho^s\boldsymbol{\eta})}{J_k(\rho^s\boldsymbol{\eta})},$$

where $g_k(\rho^s\boldsymbol{\eta})$ is the k-th row of $\boldsymbol{g}(\rho^s\boldsymbol{\eta}) = \boldsymbol{\eta} - \xi\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)'\{(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_0)\rho^s\boldsymbol{\eta} - e(\rho^s\boldsymbol{\eta})\}$ and $J_k(\rho^s\boldsymbol{\eta})$ is the k-th row of $\boldsymbol{J}(\rho^s\boldsymbol{\eta}) = \partial\boldsymbol{g}(\rho^s\boldsymbol{\eta})/\partial\rho^s$. Note that $e(\rho^s\boldsymbol{\eta})$ is calculated based on Eq. (I-3).

Step 2. Repeat Step 1 until $|\rho^{s+1} - \rho^s| \le \epsilon$, where ϵ is a stopping criterion $(\epsilon > 0)$.

Finally, I examine the validity of the random threshold assumption using simulated data from one of the designs presented in Section 4.1. Table II-1 reports τ and υ (and corresponding theoretical values of the parameters, i.e., $\boldsymbol{\beta}/\tau$ and $\boldsymbol{\beta}/\upsilon$) in a large sample $(N = 5{,}000{,}000)$ when the regressors are mesokurtic and the error is normal (specifically, $d_t = 1[\beta_0 + \beta_1 x_t + \beta_2 z_t + \varepsilon_t > 0]$, where x_t, $z_t \sim \text{N}(3, 0.5)$, $\varepsilon_t \sim \text{N}(0,1)$, $\beta_0 = -\text{Med}(\beta_1 x_t + \beta_2 z_t)$, and $\beta_1 = \beta_2 = 1$). Recall that τ is calculated as $\sqrt{n_0\sigma_{N,0}^2 + n_1\sigma_{N,1}^2 + \sigma_{N,\varepsilon}^2}$, which is the

standard deviation of the error component ($\delta_t = \gamma_t - \varepsilon_t$) when the random threshold assumption is true (i.e., γ_t is independent of $\tilde{\mathbf{x}}_t\boldsymbol{\beta}$ and $\tilde{\varepsilon}_t$), and $\upsilon = \sqrt{\sigma_\delta^2 - \boldsymbol{\Sigma}'_{\tilde{\mathbf{x}}\delta}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\tilde{\mathbf{x}}}^{-1}\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}\delta}}$ when the assumption is not true (i.e., γ_t is dependent on $\tilde{\mathbf{x}}_t\boldsymbol{\beta}$ and $\tilde{\varepsilon}_t$).

Columns 1 and 4 in the table report the theoretical values of τ and υ in this simulation, showing that τ is, as expected, larger than υ: 1.318 for the former and 0.477 for the latter. Consequently, θ_k^τ $(= \beta_k/\tau)$ should be larger than θ_k^υ $(= \beta_k/\upsilon)$. However, Columns 3 and 6, which report the average of the estimates with and without the random threshold assumption (500 trials with a sample size of 500), respectively, show that the average of $\hat{\theta}_k^\upsilon$ estimated by the above procedure is smaller than that of $\hat{\theta}_k^\tau$ estimated in the Monte Carlo simulation in Section 4. Additionally, $\hat{\rho}$ is, on average, 0.595 (the stopping criterion ϵ was set to 10^{-5}, resulting in an average Euclidean norm of $\boldsymbol{g}(\rho\boldsymbol{\eta})$ of 0.0048). Importantly, while $\hat{\theta}_k^\upsilon$ differs significantly from its theoretical value, $\hat{\theta}_k^\tau$ is close to its theoretical value, implying that the random threshold assumption holds. Also note that $\hat{\theta}_k^\tau$ is 9.4% smaller than θ_k^τ, a small but nonnegligible difference. This may be due to the presence of positive or negative errors present in each trial due to the small sample size ($N = 500$). Since τ is the standard deviation (of δ_t) calculated as $\tau = \sqrt{\tau^2}$, both types of errors may result in an increase in τ, while their effects on β are averaged out.

Table II-1. Simulation results for τ and υ

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	With the random threshold assumption			Without the random threshold assumption			
	Theoretical value ($N = 10{,}000{,}000$)		Ave. of estimates (500 trials with $N = 500$)	Theoretical value ($N = 10{,}000{,}000$)		Ave. of estimates (500 trials with $N = 500$)	
Model: $d_t = 1[-6 + x_t + z_t + \varepsilon_t > 0]$, where $x_t, z_t \sim \mathrm{N}(3, 0.5)$, and $\varepsilon_t \sim \mathrm{N}(0,1)$	τ	θ_k^τ $(= 1/\tau)$	$\hat{\theta}_k^\tau$	υ	θ_k^υ $(= 1/\tau)$	$\hat{\theta}_k^\upsilon$ $(\hat{\rho})$	
	1.318	0.759	0.688	0.447	2.236	0.410 (0.595)	

Source: Author's estimation using simulated data.

III. The asymptotic efficiency for the binary choice model

For the binary choice model in Section 3.2, the limit variance matrix of $\sqrt{N}\left(\widehat{\boldsymbol{\theta}}_{\mathrm{RB}} - \boldsymbol{\theta}_0\right)$, where $\boldsymbol{\theta} = \boldsymbol{\beta}/\tau$, is

$$-\mathrm{E}[\nabla_{\theta\theta} \ln f(\tilde{z}_t|\, \theta_0)]^{-1} = \lim_{N\to\infty}\left(\frac{\sum_t^T G(\tilde{\mathbf{x}}_t\boldsymbol{\theta}_0)\tilde{\mathbf{x}}_t'\tilde{\mathbf{x}}_t}{T}\right)^{-1}$$

$$= \lim_{N\to\infty}\left(\frac{\sum_t^T G(\tilde{\mathbf{x}}_t\boldsymbol{\theta}_0)}{T}\frac{\sum_t^T G(\tilde{\mathbf{x}}_t\boldsymbol{\theta}_0)\tilde{\mathbf{x}}_t'\tilde{\mathbf{x}}_t}{\sum_t^T G(\tilde{\mathbf{x}}_t\boldsymbol{\theta}_0)}\right)^{-1} = \lim_{N\to\infty}(\mu_{\tilde{G}}\cdot\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}}^{w})^{-1},$$

where $G(\tilde{\mathbf{x}}\boldsymbol{\theta}_0) = \phi^2(\tilde{\mathbf{x}}\boldsymbol{\theta}_0)[\Phi(\tilde{\mathbf{x}}\boldsymbol{\theta}_0)\{1 - \Phi(\tilde{\mathbf{x}}\boldsymbol{\theta}_0)\}]^{-1}$, $\mu_{\tilde{G}} = T^{-1}\sum_t^T G(\tilde{\mathbf{x}}_t\boldsymbol{\theta}_0)$, and $\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}}^{w} = \{\sum_t^T G(\tilde{\mathbf{x}}_t\boldsymbol{\theta}_0)\}^{-1}\sum_t^T G(\tilde{\mathbf{x}}_t\boldsymbol{\theta}_0)\tilde{\mathbf{x}}_t'\tilde{\mathbf{x}}_t$. Again, note that $T = T' + N$ and $T \to \infty$ when $N \to \infty$. Therefore, when N is large, the variance of the RBML estimator is approximated by $(N \cdot \mu_{\tilde{G}} \cdot \boldsymbol{\Sigma}_{\tilde{\mathbf{x}}}^{w})^{-1}$.

Meanwhile, the Cramér–Rao lower bound for the binary choice model in Eq. (3) with the normality assumption is

$$\left\{\sum_i^N G(\mathbf{w}_i)(1 \quad \mathbf{x}_i)'(1 \quad \mathbf{x}_i)\right\}^{-1},$$

where $G(\mathbf{w}_i) = \phi^2(\mathbf{w}_i)[\Phi(\mathbf{w}_i)\{1-\Phi(\mathbf{w}_i)\}]^{-1}$ and $\mathbf{w}_i = \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0$. Then, when N is large, the variance of the probit ML estimator ($\widehat{\boldsymbol{\beta}}_{\mathrm{P}}$) is approximated by

$$\left[\sum_i^N G(\mathbf{w}_i)\mathbf{x}_i'\mathbf{x}_i - \frac{\sum_i^N G(\mathbf{w}_i)\mathbf{x}_i' \sum_i^N G(\mathbf{w}_i)\mathbf{x}_i}{\sum_i^N G(\mathbf{w}_i)}\right]^{-1}$$

$$= \left[\sum_i^N G(\mathbf{w}_i) \cdot \left\{\frac{\sum_i^N G(\mathbf{w}_i)\mathbf{x}_i'\mathbf{x}_i}{\sum_i^N G(\mathbf{w}_i)} - \frac{\sum_i^N G(\mathbf{w}_i)\mathbf{x}_i'}{\sum_i^N G(\mathbf{w}_i)} \cdot \frac{\sum_i^N G(\mathbf{w}_i)\mathbf{x}_i}{\sum_i^N G(\mathbf{w}_i)}\right\}\right]^{-1}$$

$$= \left[\sum_i^N G(\mathbf{w}_i) \cdot \frac{\sum_i^N [G(\mathbf{w}_i)\{\mathbf{x}_i'\mathbf{x}_i - (\boldsymbol{\mu}_{\mathbf{x}}^w)'\boldsymbol{\mu}_{\mathbf{x}}^w\}]}{\sum_i^N G(\mathbf{w}_i)}\right]^{-1} = (N \cdot \mu_G \cdot \boldsymbol{\Sigma}_{\mathbf{x}}^w)^{-1},$$

where $\boldsymbol{\mu}_{\mathbf{x}}^w = \{\sum_i^N G(\mathbf{w}_i)\}^{-1}\sum_i^N G(\mathbf{w}_i)\mathbf{x}_i$, $\mu_G = N^{-1}\sum_i^N G(\mathbf{w}_i)$, and $\boldsymbol{\Sigma}_{\mathbf{x}}^w = \{\sum_i^N G(\mathbf{w}_i)\}^{-1}\ \sum_i^N [G(\mathbf{w}_i)\{\mathbf{x}_i'\mathbf{x}_i - (\boldsymbol{\mu}_{\mathbf{x}}^w)'\boldsymbol{\mu}_{\mathbf{x}}^w\}]$. Therefore, when the "average weight" (μ) and "weighted variance" ($\boldsymbol{\Sigma}^w$) components are equivalent between the RBML and probit estimators (i.e., $\mu_{\tilde{G}} = \mu_G$ and $\boldsymbol{\Sigma}_{\tilde{\mathbf{x}}}^w = \boldsymbol{\Sigma}_{\mathbf{x}}^w$), the RBML estimator appears to attain the Cramér–Rao lower bound for the binary choice model. However, this is not true because the parameters to be estimated in the RBML estimation are different from those in the probit estimation (i.e., $\theta_0 = \beta_0/\tau_0 \neq \beta_0/\sigma_0$). Therefore, it is meaningless to compare their variance–covariance estimators directly. Then, focusing on the *scale-normalized* parameter, I conduct numerical simulations to compare the asymptotic variance and discuss the asymptotic efficiency of the RBML estimator for the binary choice model.

The model considered here is expressed as follows:

$$d_i = 1[y_i > 0] = 1[\alpha_0 + \beta_0 x_i + \gamma_0 z_i + \varepsilon_i > 0] = 1[\alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_\mathbf{0} + \varepsilon_i > 0],$$

where $\alpha_0 = -\mathrm{Median}(\mathbf{x}_i\boldsymbol{\beta}_\mathbf{0})$, $\beta_0 = \gamma_0 = 1$, $\varepsilon_i|\mathbf{x}_i \sim \mathrm{N}(0, \sigma_0^2)$, and $\sigma_0^2 = 1$. For the regressors, because the variance estimator depends on the data distribution (i.e., the mean, variance, skewness, kurtosis, and other characteristics of $\mathbf{x}_i\boldsymbol{\beta}_\mathbf{0}$) and the error distribution, I consider eight different cases: $x_i, z_i \overset{i.i.d.}{\sim}$ (a) $\chi^2(k)/2\sqrt{k}$ ($k \in \{1, 2, 4, 8, 16\}$, (b) N(0,0.5), and (c) $\mathrm{Beta}(k,k) \times \sqrt{4k+2}$ ($k \in \{2.5, 0.5, 10^{-10}\}$). Note that $\mathbf{x}\boldsymbol{\beta}_\mathbf{0}$ is designed to have unit variance and zero skewness in all cases for the sake of simplicity. Since the probit estimate is $\widehat{\boldsymbol{\beta}}_{\mathrm{P}} = \widehat{\boldsymbol{\beta}_0/\sigma_0}$ and the RBML estimate is $\widehat{\boldsymbol{\beta}}_{\mathrm{RB}} = \widehat{\boldsymbol{\beta}_0/\tau_0}$, I compare the asymptotic variance of the *scale-normalized* estimate for β_m, that is, $\hat{\zeta}_m = \hat{\beta}_m \times 2 \times \left|\hat{\beta}_m + \hat{\gamma}_m\right|^{-1}$ ($m = \mathrm{PR}, \mathrm{RB}$). Then, the delta method implies that the variance of the normalized parameter estimate is $\mathrm{Var}(\hat{\zeta}_m) = \frac{\eta_m^2}{4}\{\mathrm{Var}(\hat{\beta}_m) + \mathrm{Var}(\hat{\gamma}_m)\}$, where $\eta_{\mathrm{P}}^2 = \sigma_0^2 = 1$ and $\eta_{\mathrm{RB}}^2 = \tau_0^2$.

Table III-1 reports the simulation results for the variance of the probit and RBML estimators in large samples. The results indicate that the variance of the probit estimator varies according to the data distribution through the average weight (μ_G) and weighted variance (Σ^w). As shown in the table, as the kurtosis of $\mathbf{x}_i\boldsymbol{\beta}_0$ increases, μ_G increases and Σ^w decreases. Regarding μ_G, $G(\mathbf{w}_i) = \phi^2(\mathbf{w}_i)[\Phi(\mathbf{w}_i)\{1-\Phi(\mathbf{w}_i)\}]^{-1}$ becomes large as $\mathbf{w}_i$ $(= \alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0)$ approaches zero. In this simulation, when the kurtosis is large, the distribution of $\mathbf{w}_i$ is concentrated around zero; therefore, μ_G increases as the kurtosis of $\mathbf{x}_i\boldsymbol{\beta}_0$ increases. Conversely, Σ^w decreases when the distribution becomes sharper. The asymptotic variance of the probit estimator increases primarily due to the latter effect. (It is also worth noting that the effect of skewness is very limited, although not examined in this appendix. As described above, $\mathbf{x}_i\boldsymbol{\beta}_0$ is adjusted to have zero skewness, but even if I allow the skewness of $\mathbf{x}_i\boldsymbol{\beta}_0$ to vary, the results remain unchanged.)

In contrast, the asymptotic variance of the RBML estimator is stable across the regressor distributions. Surprisingly, when the kurtosis of $\mathbf{x}_i\boldsymbol{\beta}_0$ is six, the RBML estimator outperforms the probit estimator in terms of the variance of the scale-normalized parameter. Although a kurtosis of six seems unlikely in reality, it is important that when comparing the scale-normalized estimates, a case exists in which a semiparametric estimator with no distributional assumption can be comparable to the probit estimator in terms of efficiency when the probit is the correct model. The Monte Carlo simulations in Section 4 compare the small sample performance of the probit and RBML estimators for a binary choice model when the error is normal or nonnormal.

Table III-1. Comparison of asymptotic variance for the scale-normalized parameter: RBML versus probit

Model: $d_i = 1(\alpha_0 + \mathbf{x}_i\boldsymbol{\beta}_0 + \varepsilon_i > 0)$			Probit			RBML				Variance for the normalized parameter, $N \cdot \mathrm{Var}(\hat{\zeta})$ ($N = 10{,}000{,}000$)		
Distribution of x_i, z_i		Kurtosis of $\mathbf{x}_i\boldsymbol{\beta}_0$	μ_G	Σ_x^w	$-\mathrm{H}_x$	τ_0	$\mu_{\tilde{G}}$	$\Sigma_{\tilde{x}}^w$	$-\mathrm{H}_{\tilde{x}}$	Probit	RBML	Relative efficiency
$\chi^2(1)/2$	(kurtosis = 12.00)	6.000	0.522	0.228	0.119	1.354	0.537	0.426	0.229	4.203	4.005	1.024
$\chi^2(2)/2\sqrt{2}$	(kurtosis = 6.000)	3.000	0.506	0.285	0.144	1.338	0.535	0.425	0.227	3.462	3.937	0.938
$\chi^2(4)/4$	(kurtosis = 3.000)	1.500	0.495	0.329	0.163	1.327	0.533	0.424	0.226	3.071	3.893	0.888
$\chi^2(8)/4\sqrt{2}$	(kurtosis = 1.500)	0.750	0.488	0.357	0.174	1.321	0.533	0.424	0.226	2.869	3.867	0.861
$\chi^2(16)/8$	(kurtosis = 0.750)	0.375	0.485	0.373	0.181	1.318	0.532	0.423	0.225	2.767	3.853	0.847
$\mathrm{N}(0, 0.5)$	(kurtosis = 0,000)	0.000	0.481	0.390	0.188	1.318	0.532	0.423	0.225	2.665	3.852	0.832
$\mathrm{Beta}(2.5, 2.5) \cdot 2\sqrt{3}$ (kurtosis = −0.750)		−0.375	0.476	0.424	0.202	1.310	0.531	0.423	0.225	2.479	3.827	0.805
$\mathrm{Beta}(0.5, 0.5) \cdot 2$ (kurtosis = −1.500)		−0.750	0.471	0.468	0.220	1.306	0.531	0.423	0.224	2.272	3.806	0.773
$\mathrm{Beta}(10^{-10}, 10^{-10}) \cdot \sqrt{2}$ (kurtosis = −2.000)		−1.000	0.467	0.500	0.233	1.303	0.530	0.422	0.224	2.141	3.770	0.754

Notes: α_0 is set to the negative median value of $\mathbf{x}_i\boldsymbol{\beta}_0$ ($= \beta_0 x_i + \gamma_0 z_i$) so that $\Pr(d_i = 1) = 0.5$, x_i and z_i are adjusted to have variance of 0.5, $\beta_0 = \gamma_0 = 1$, and $\varepsilon_i \sim \mathrm{N}(0,1)$. In addition, when x_i and z_i are from the chi-square distribution, z_i is multiplied by -1 (negative of chi-square random variable) so that $\mathbf{x}_i\boldsymbol{\beta}_0$ has zero skewness. Therefore, for all regressor distributions, the skewness of $\mathbf{x}_i\boldsymbol{\beta}_0$ is zero. The scale-normalized parameter is defined as $\hat{\zeta}_m = \hat{\beta}_m \times (\beta_0 + \gamma_0)/|\hat{\beta}_m + \hat{\gamma}_m|$ ($m =$ PR, RB), whose variance is calculated based on the delta method. The relative efficiency is calculated as the ratio of the standard errors, $\sqrt{\mathrm{Var}(\hat{\zeta}_{\mathrm{PR}})/\,\mathrm{Var}(\hat{\zeta}_{\mathrm{RB}})}$.

Source: Author's estimates using simulated data.

IV. Data and estimation results in the simulation analysis

This section provides supplementary information regarding the Monte Carlo simulation study (Section 4). Table IV-1 shows summary statistics of the simulated data used in a trial in the simulations for the linear regression model (Panel A), binary choice model (Panel B), and binary choice model with a perfect prediction problem (Panel C). Table IV-2 presents selected simulation results for the binary choice model with different regressors' kurtoses and error designs. These tables also include summary statistics and results for the case where the error design is a right-skewed platykurtic normal mixture (referred to as error design (b')). In addition, the estimation results for all regressor designs when the error distribution is this design (b') are shown in Figure IV-1.

Table IV-1. Summary statistics of simulated data in a trial

Variable	Obs.	Mean	Std. Dev.	Min	Max
A) Variables for linear regression model					
$x_i \sim \chi^2(16)/8$	500	2.010	0.696	0.535	4.244
$z_i \sim \chi^2(16)/8$	500	1.967	0.654	0.541	4.162
ε_i					
(a) Standard normal	500	0.015	0.954	−2.378	2.615
(b) Normal mixture (left skewed, leptokurtic)	500	−0.007	0.968	−2.932	2.595
(b') Normal mixture (right skewed, platykurtic)	500	−0.012	0.995	−2.104	2.729
(c) Normal with heteroscedastic variance	500	0.015	0.960	−3.306	3.283
(d) Student's t	500	0.054	1.865	−7.483	10.520
y_i ($= \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$, where $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$)					
(a) Standard normal	500	3.993	1.368	0.562	8.204
(b) Normal mixture (left skewed, leptokurtic)	500	3.970	1.373	0.008	8.071
(b') Normal mixture (right skewed, platykurtic)	500	3.965	1.380	0.786	8.349
(c) Normal with heteroscedastic variance	500	3.992	1.369	1.244	9.158
(d) Student's t	500	4.031	2.132	−4.543	14.616
B) Variables for binary choice model					
$x_i \sim \text{Beta}(4.5,4.5) \cdot 2\sqrt{5}$	500	2.209	0.681	0.448	4.006
$z_i \sim \text{Beta}(4.5,4.5) \cdot 2\sqrt{5}$	500	2.251	0.695	0.239	3.945
ε_i					
(a) Standard normal	500	−0.014	1.067	−3.378	2.615
(b) Normal mixture (left skewed, leptokurtic)	500	0.011	1.049	−3.932	2.595
(b') Normal mixture (right skewed, platykurtic)	500	0.035	1.038	−2.692	3.388
(c) Normal with heteroscedastic variance	500	0.001	1.075	−4.306	3.283
(d) Student's t	500	−0.015	3.028	−32.560	35.494
d_i ($= 1[\beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i > 0]$, where $\beta_0 = 0$ and $\beta_1 = \beta_2 = 1$))					
(a) Standard normal	500	0.494			
(b) Normal mixture (left skewed, leptokurtic)	500	0.506			
(b') Normal mixture (right skewed, platykurtic)	500	0.480			
(c) Normal with heteroscedastic variance	500	0.474			
(d) Student's t	500	0.498			
C) Variables for binary choice model with perfect prediction (PP) problem					
x_i (binary variable)	500	0.516			
z_i (normal random variable, $\text{N}(0, 0.2)$)	500	−0.020	0.465	−1.509	1.295
u_i (normal random variable, $\text{N}(0, 0.8)$)	500	0.042	0.944	−2.914	2.944
d_i ($= 1[\beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i > 0]$), where $\beta_0 = 0$, and $\beta_2 = 1$					
(i) Nearly PP, $\beta_1 = \Phi^{-1}(0.99)$	500	0.790			
(ii) Fully PP, $\beta_1 = \lvert\text{Min}(\beta_2 z_i + \varepsilon_i)\rvert + 0.01$	500	0.798			

Source: Author's calculations using simulated data.

Table IV-2. Selected simulation results for the binary choice model

	(1)		(2)		(3)		(4)		(5)	
Error design:	(a) Standard normal		(b) Normal mixture (left skewed, leptokurtic)		(b') Normal mixture (right skewed, platykurtic)		(c) Normal with heteroscedasticity		(d) Student's t	
Estimator	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
A) The kurtosis of x_i and z_i is 2 (that of $\mathbf{x}_i\boldsymbol{\beta}$ is 1)										
RBML ($M = T = 100{,}000$)	0.0010	0.0655	−0.0016	0.0628	−0.0038	0.0622	0.0021	0.0690	0.0036	0.0828
Probit	0.0000	0.0702	−0.0018	0.0669	−0.0044	0.0666	0.0020	0.0743	0.0030	0.0901
Sieve ML (Gallant & Nychka, 1987)	0.0002	0.0727	−0.0008	0.0693	−0.0044	0.0643	0.0039	0.0701	0.0046	0.0928
Kernel ML (Klein & Spady, 1993)	−0.0051	0.0848	0.0027	0.0877	−0.0040	0.0753	0.0040	0.0995	−0.0004	0.1129
SLS (Ichimura, 1993)	0.0035	0.0908	0.0008	0.0892	−0.0010	0.0828	0.0056	0.0838	0.0066	0.1065
B) The kurtosis of x_i and z_i is 1 (that of $\mathbf{x}_i\boldsymbol{\beta}$ is 0.5)										
RBML ($M = T = 100{,}000$)	0.0008	0.0650	0.0017	0.0670	−0.0022	0.0698	0.0007	0.0662	0.0011	0.0828
Probit	0.0012	0.0683	0.0027	0.0713	−0.0019	0.0725	0.0014	0.0692	0.0013	0.0876
Sieve ML (Gallant & Nychka, 1987)	0.0008	0.0695	0.0036	0.0721	−0.0027	0.0704	0.0021	0.0687	−0.0001	0.0895
Kernel ML (Klein & Spady, 1993)	0.0019	0.0807	0.0049	0.0835	−0.0046	0.0774	0.0025	0.0852	0.0042	0.1072
SLS (Ichimura, 1993)	0.0053	0.0822	0.0042	0.0822	−0.0025	0.0807	0.0041	0.0811	0.0023	0.1011
C) The kurtosis of x_i and z_i is 0 (that of $\mathbf{x}_i\boldsymbol{\beta}$ is also 0)										
RBML ($M = T = 100{,}000$)	−0.0021	0.0637	−0.0035	0.0656	−0.0009	0.0691	−0.0025	0.0647	−0.0005	0.0813
Probit	−0.0023	0.0629	−0.0026	0.0644	0.0004	0.0674	−0.0015	0.0639	−0.0003	0.0819
Sieve ML (Gallant & Nychka, 1987)	−0.0029	0.0632	−0.0038	0.0648	−0.0009	0.0673	−0.0022	0.0632	−0.0003	0.0821
Kernel ML (Klein & Spady, 1993)	−0.0043	0.0711	−0.0044	0.0727	0.0007	0.0772	−0.0037	0.0716	−0.0026	0.0959
SLS (Ichimura, 1993)	0.0000	0.0778	0.0013	0.0777	0.0013	0.0805	0.0007	0.0753	−0.0010	0.0973
D) the kurtosis of x_i and z_i is −1 (that of $\mathbf{x}_i\boldsymbol{\beta}$ is −0.5)										
RBML ($M = T = 100{,}000$)	0.0011	0.0727	0.0006	0.0704	−0.0018	0.0748	0.0012	0.0696	−0.0013	0.0894
Probit	0.0005	0.0641	−0.0003	0.0627	−0.0019	0.0669	0.0004	0.0617	0.0010	0.0838
Sieve ML (Gallant & Nychka, 1987)	0.0004	0.0641	−0.0005	0.0626	−0.0022	0.0656	0.0009	0.0584	−0.0008	0.0831
Kernel ML (Klein & Spady, 1993)	0.0010	0.0731	0.0015	0.0693	−0.0004	0.0755	0.0015	0.0608	0.0020	0.0926
SLS (Ichimura, 1993)	0.0021	0.0759	0.0051	0.0728	0.0017	0.0779	0.0020	0.0663	0.0027	0.0956

Source: Author's estimates using simulated data.

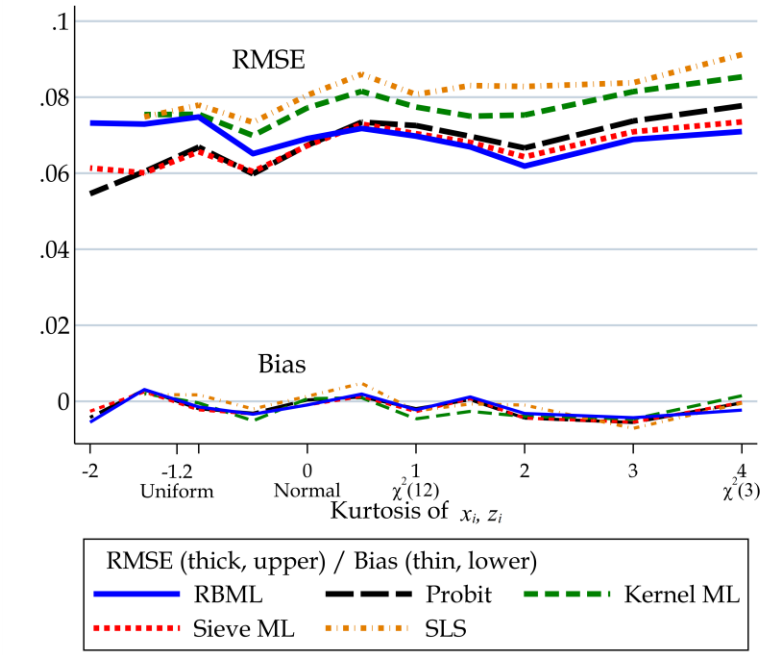


Figure IV-1: Simulation results for the binary model when the error distribution is a normal mixture (right skewed and platykurtic)

Notes: The thick lines represent the root mean square errors (RMSEs), and the thin lines represent the biases. "RBML" is the resampling-based ML proposed in this paper, "Kernel ML" is Klein and Spady's (1993) Nadaraya–Watson kernel ML, "sieve ML" is Gallant and Nychka's (1987) Hermite polynomial sieve ML, and "SLS" is Ichimura's (1993) SLS.

Source: Author's estimates using simulated data.