Bernoulli **24**(2), 2018, 929–955
DOI: 10.3150/16-BEJ809

Asymptotics for the maximum sample likelihood estimator under informative selection from a finite population

DANIEL BONNÉRY[1], F. JAY BREIDT[2] and FRANÇOIS COQUET[3]

[1]*JPSM, 1218 LeFrak, College Park, MD 20742, USA.*
E-mail: dbonnery@umd.edu, url: jpsm.umd.edu/facultyprofile/Bonnéry/Daniel
[2]*Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877.*
E-mail: jbreidt@stat.colostate.edu, url: www.stat.colostate.edu/statpeople/people/jbreidt.html
[3]*Ensai, Campus de Ker-Lann, Rue Blaise Pascal – BP 37203, 35172 Bruz - Cedex, France.*
E-mail: fcoquet@ensai.fr, url: www.ensai.com/francois-coquet-rub,30.html

Inference for the parametric distribution of a response given covariates is considered under informative selection of a sample from a finite population. Under this selection, the conditional distribution of a response in the sample, given the covariates and given that it was selected for observation, is not the same as the conditional distribution of the response in the finite population, given only the covariates. It is instead a weighted version of the conditional distribution of interest. Inference must be modified to account for this informative selection. An established approach in this context is maximum "sample likelihood", developing a weight function that reflects the informative sampling design, then treating the observations as if they were independently distributed according to the weighted distribution. While the sample likelihood methodology has been widely applied, its theoretical foundation has been less developed. A precise asymptotic description of a wide range of informative selection mechanisms is proposed. Under this framework, consistency and asymptotic normality of the maximum sample likelihood estimators are established. The theory allows for the possibility of nuisance parameters that describe the selection mechanism. The proposed regularity conditions are verifiable for various sample schemes, motivated by real problems in surveys. Simulation results for these examples illustrate the quality of the asymptotic approximations, and demonstrate a practical approach to variance estimation that combines aspects of model-based information theory and design-based variance estimation.

Keywords: complex survey; pseudo-likelihood; stratified sampling; weighted distribution

1. Introduction

Consider a finite population $U = \{1, \ldots, N\}$ and a joint probability density function (p.d.f.) $\mathrm{f}(x, y, z)$, called a superpopulation model, generating independent and identically distributed (i.i.d.) random vectors (X_k, Y_k, Z_k) for each $k \in U$. Here, X_k is a vector of covariates, Y_k is a response (possibly vector-valued), and Z_k is a vector of design variables. Suppose that $\mathrm{f}(x, y, z) = \mathrm{f}_\xi(z|x, y)\mathrm{f}_\theta(y|x)g(x)$ where ξ and θ are vectors of unknown parameters. The factor f_ξ describes the selection mechanism, and contains only nuisance parameters ξ. The factor g describes the marginal distribution of X and is not of interest. It is of interest to conduct inference for θ based on observations of (X_k, Y_k, Z_k) for a sample selected from U. In the absence of

sampling, the conditional distribution of the Y_k's given the X_k's for the entire finite population is $\prod_{k\in U} \mathrm{f}_\theta(y_k|x_k)$, and Z_k's are irrelevant for inference on θ.

In general, however, the sample selection depends on $(Z_1, \ldots, Z_N)$. For positive Z_k's, a common design would be to include element k in the sample of size n with probability $nZ_k/\sum_{k\in U} Z_k$. As another example, the population could be sorted on an index Z_k and then grouped into disjoint strata of elements with similar Z_k values, with independent sample selections from each such stratum.

Since Y_k and Z_k are dependent in general, the conditional distribution of the Y_k's given the X_k's *in the sample* is typically not equal to the product of the conditionals: the distribution of Y_k given X_k for selected k is not $\mathrm{f}_\theta(y_k|x_k)$ and the informative selection mechanism may induce dependence among the selected observations. Ignoring the informative selection results in inconsistent estimation of θ.

One approach to estimation in this context of informative selection is *pseudo-likelihood*, consisting of maximizing the Horvitz and Thompson [13] estimator of $\sum_{k\in U} \ln \mathrm{f}_\theta(y_k|x_k)$, the population level log-likelihood. This method is straightforward and is a standard feature of software designed for the analysis of survey data. The resulting estimators are consistent for θ under mild conditions, but may be inefficient compared to other likelihood-based methods.

Another approach is based on the *sample likelihood*, consisting of ignoring the dependence among the observations and treating them as if they were independently distributed according to the *sample p.d.f.*, $\rho(x, y; \theta, \xi)\mathrm{f}_\theta(y|x)$, which is a weighted version of the population p.d.f. The weight function $\rho(\cdot)$ requires the specification of the model f_ξ for the selection mechanism. The sample p.d.f. is the basis for an estimation approach under length-biased sampling (Patil and Rao [17]). It has been developed for more general informative selection schemes in the complex survey context (Krieger and Pfeffermann [15], Section 3, Pfeffermann, Krieger and Rinott [19], Pfeffermann and Sverchkov [24]). The weight $\rho(x, y; \theta, \xi)$ is the expected inclusion probability of an element given its covariate is x and response is y, divided by the expected inclusion probability given its covariate is x. (This definition can be extended for random sample sizes and with-replacement sampling, replacing expected inclusion probability by expected number of selections (Bonnéry, Breidt and Coquet [2]). As shown by Landsman and Graubard [16], the Breslow and Cain [3] approach is a special case of sample likelihood estimation.

The sample likelihood methodology has been extended in a number of directions (Pfeffermann and Sverchkov [22,25], Pfeffermann and Sikov [21], Pfeffermann [18]), including longitudinal surveys (Eideh and Nathan [7–9]), small area estimation (Ghosh and Maiti [11], Pfeffermann and Sverchkov [23], Eideh and Nathan [6]), and multi-level modeling (Pfeffermann, Moura and Silva [20], Cai [4]). A review of these and other approaches to inference under informative selection is given by Pfeffermann and Sverchkov [24]. See also Chambers *et al.* [5], Section 3.3, for an overview and simulation-based comparisons of pseudo-likelihood, sample likelihood and full maximum likelihood.

Under a strong set of assumptions, including that sample size remains fixed as population size goes to infinity, Pfeffermann, Krieger and Rinott [19] have established the pointwise convergence of the joint distribution of the responses to the product of the sample p.d.f.'s. This is taken as partial justification of the sample likelihood approach; see, for example, Landsman and Graubard [16], Section 3. In this paper, we develop a more complete theoretical foundation for the sample likelihood methodology: a central limit theorem for model parameters that also accounts for estimation of nuisance parameters in the selection mechanism.

In Bonnéry, Breidt and Coquet [2], a precise asymptotic framework for informative selection is given and weak conditions on the informative selection mechanism are stated under which the (unweighted) empirical cumulative distribution function (c.d.f.) converges uniformly to the c.d.f. associated with the limiting sample p.d.f. That is, the classical Glivenko–Cantelli theorem holds in the case of a weak dependence among the draws, perhaps suggestive of good behavior for the sample likelihood approach.

In this paper, we use the same asymptotic framework and further conditions on the selection mechanism and on the regularity of the sample p.d.f. to show consistency and asymptotic normality of sample likelihood estimators of θ. Our theory assumes the existence of a $\sqrt{n}$-consistent and asymptotically normal estimator $\widehat{\xi}$ of the selection mechanism parameters ξ, which can often be obtained via Horvitz–Thompson estimation. These parameters are not of independent interest but do appear in $\rho(x, y; \theta, \xi)$ and hence in the sample likelihood. Our criterion function is then the log sample likelihood with $\widehat{\xi}$ replacing ξ. We study the properties of the estimator of θ that maximizes this criterion. Adapting Gong and Samaniego [12], we establish existence, consistency and central limit theory for such an estimator of θ.

Pfeffermann and Sverchkov [25], Section 12.4 and Eideh and Nathan [8], Section 4.3, have considered the use of inverse information for estimating the variance of the maximum sample likelihood estimators, but they treat the informativeness parameter estimates $\widehat{\xi}$ as fixed. Landsman and Graubard [16], Section 5.1, observe that under independence, Yuan and Jennrich [28] could be used to correct the variance estimator for the variation in $\widehat{\xi}$, but these results do not allow for the dependence due to selection that we consider here. Our results lead immediately to appropriate variance estimators.

As in Bonnéry, Breidt and Coquet [2], the conditions we propose are verifiable for various sample schemes, commonly encountered in real problems in surveys, and involve computing conditional versions of first and second-order inclusion probabilities. We derive the asymptotic distribution in detail for the aforementioned scheme in which the population is stratified on ordered values of Z. Given sufficient information on the sample design and the covariates, the maximum sample likelihood estimators are feasible and are more efficient than the maximum pseudo-likelihood estimators. We illustrate this final result by simulations applied in some basic sample schemes. We also show how to obtain useful variance estimators, combining information computations as in standard likelihood theory with design-based covariance matrix estimation.

2. Notation and definitions

2.1. General framework

Let $\{N_\gamma\}_{\gamma\in\mathbb{N}}$ denote an increasing sequence of positive integers and consider a sequence of finite populations $\{U_\gamma\}_{\gamma\in\mathbb{N}}$ with $U_\gamma = \{1, \ldots, N_\gamma\}$.

All random variables are defined on a common measured space $(\Omega, \mathscr{A}, \mathrm{P})$. We assume that P belongs to a parametric family $(\mathrm{P}_{\theta,\xi})_{(\theta,\xi)\in\Theta\times\Xi}$ where Θ and Ξ are subsets of real vector spaces; that P is dominated by some measure; and that there exists a unique $(\theta_0, \xi_0) \in \Theta \times \Xi$ such that $\mathrm{P} = \mathrm{P}_{\theta_0,\xi_0}$.

Densities are generically denoted by f and are given without specification of the measure. Equalities of conditional distributions or densities are almost sure equalities. Writing a density implies that it is defined:, $\mathrm{f}_{Y|X=x}$ means there exists a conditional distribution $\mathrm{P}^{Y|X=x}$ satisfying standard conditions, and $\mathrm{f}_{Y|X=x}$ is the density of $\mathrm{P}^{Y|X=x}$ with respect to some σ-finite measure. We write $\mathrm{f}_{Y|X=x;\theta,\xi}$, for example, when the explicit parameterization is important.

We define a sequence of real random vectors $\{(X_k, Y_k, Z_k)\}_{k\in\mathbb{N}}$, independently and identically distributed, where X_k is a vector of finite dimension of real covariates, Y_k corresponds to the response (possibly vector-valued) of interest and Z_k are the design characteristics of the element k.

The finite population matrices of covariate, response and design vectors are denoted respectively by $\mathcal{X}_\gamma = (X_k)_{k\in U_\gamma}$, $\mathcal{Y}_\gamma = (Y_k)_{k\in U_\gamma}$ and $\mathcal{Z}_\gamma = (Z_k)_{k\in U_\gamma}$. We assume that for all x and y, $\mathrm{f}_{Z|X=x,Y=y;\theta,\xi} = \mathrm{f}_{Z|X=x,Y=y;\xi}$, and that for all x, $\mathrm{f}_{Y|X=x;\theta,\xi} = \mathrm{f}_{Y|X=x,;\theta}$. We shall sometimes write $\mathrm{f}_{Y|X=x,;\theta} = f_\theta(\cdot|x)$. We are interested in inference for the distribution of Y given X, parameterized by θ. The parameter ξ is a nuisance parameter controlling the distribution of Z given X and Y. Let $j = (j_k)_{k\in U_\gamma} \in \{0,1\}^{N_\gamma}$ denote a realized sample from U_γ where $j_k = 1$ if element k is selected and $j_k = 0$ otherwise. Let p denote a probability measure on $\{0,1\}^{N_\gamma}$, which we will refer to as a design measure. Assume that there exists a sequence of functions $\{D_\gamma\}_{\gamma\in\mathbb{N}}$, which we will call design measure functions, such that $D_\gamma(z)$ is a design measure for all $z \in \mathcal{Z}_\gamma(\Omega)$. Then $\Pi_\gamma = D_\gamma(\mathcal{Z}_\gamma)$ is a random design measure on U_γ. Assume that the index of the element k of the population plays no role in the way elements are selected. Specifically, for all $z \in \mathcal{Z}_\gamma(\Omega)$, r a permutation of U_γ, and $A \subset \{0,1\}^{N_\gamma}$,

$$\big(D_\gamma(z)\big)(A) = \big(D_\gamma(r.z)\big)(r.A),$$

where $r.z = (z_{r(1)}\cdots z_{r(N)})$ and $r.A = \{(a_{r(1)}\cdots a_{r(N)}) | a \in A\}$. A sample from U_γ selected according to the random design measure Π_γ is a random vector $\mathcal{J}_\gamma = (J_{\gamma,k})_{k\in U_\gamma}$ that takes values in $\{0,1\}^{N_\gamma}$ and that satisfies

$$\begin{cases} \mathrm{f}_{\mathcal{J}_\gamma|\Pi_\gamma,\mathcal{X}_\gamma,\mathcal{Y}_\gamma,\mathcal{Z}_\gamma;\theta,\xi} = \mathrm{f}_{\mathcal{J}_\gamma|\Pi_\gamma}, & \text{(1a)} \\ \mathrm{P}^{\Pi_\gamma}_{\theta,\xi}\text{-a.s. (p)}, \qquad \mathrm{P}^{\mathcal{J}_\gamma|\Pi_\gamma=\mathrm{p}}_{\theta,\xi} = \mathrm{p} & \text{(1b)} \end{cases}$$

for all $(\theta,\xi) \in \Theta \times \Xi$.

For $\gamma \in \mathbb{N}$, define the sample size as the random variable $n_\gamma = \sum_{k=1}^{N_\gamma} J_{\gamma,k}$. Define the inclusion probability of element $k \in U_\gamma$ as the random variable $\pi_{\gamma,k} = \Pi_\gamma(\{j \in \{0,1\}^{N_\gamma} | j_k = 1\})$, and the second order inclusion probability of elements k and ℓ as the random variable $\pi_{\gamma,k,\ell} = \Pi_\gamma(\{j \in \{0,1\}^{N_\gamma} | j_k = 1, j_\ell = 1\})$.

To illustrate, consider stratified simple random sampling without replacement where H strata are formed by sorting the values of a scalar index Z_k. Designs such as this are common in establishment surveys and in retrospective studies of existing records, such as the National Maternal and Infant Health Survey described in Section 4.4. Let $(N_{\gamma h})_{\gamma\in\mathbb{N},h\in\{1,\ldots,H\}}$ be an array of strictly positive integers such that for all $\gamma \in \mathbb{N}$, $N_{\gamma h}$ denotes the size of the hth stratum of the γth population and $N_\gamma = \sum_{h=1}^{H} N_{\gamma h}$. Define $(\nu_\gamma(1),\ldots,\nu_\gamma(N_\gamma))$ as the permutation of $(1,\ldots,N_\gamma)$ such that $Z_{\nu_\gamma(1)} < \cdots < Z_{\nu_\gamma(N_\gamma)}$. The permutation is a random vector which is a function of $\mathcal{Z}_\gamma$.

The hth stratum of the γth population is $U_{\gamma h} = (\nu_\gamma(T_{\gamma h-1}+1), \ldots, \nu_\gamma(T_{\gamma h}))$, with $T_{\gamma 0} = 0$, $T_{\gamma h} = \sum_{1 \le h' \le h} N_{\gamma h'}$.

Let $n_{\gamma h} \in \{1, \ldots, N_{\gamma h}\}$ denote the non-random number of elements selected from the hth stratum of the γth population via simple random sampling without replacement. The probability of the atomic event $j \in \{0,1\}^{N_\gamma}$ is then

$$\Pi_\gamma\big(\{j\}\big) = \begin{cases} \displaystyle\prod_{h=1}^{H} \binom{N_{\gamma h}}{n_{\gamma h}}^{-1}, & \text{if } \forall h \in \{1, \ldots, H\}, \displaystyle\sum_{k \in U_{\gamma h}} j_k = n_{\gamma h}, \\ 0, & \text{otherwise,} \end{cases}$$

and the probability of any event $A \subset \{0,1\}^{N_\gamma}$ is obtained by summing the above probabilities over distinct vectors $j \in A$. The first-order inclusion probabilities are $\pi_{\gamma,k} = n_{\gamma h} N_{\gamma h}^{-1}$ for $k \in U_{\gamma h}$ and the second-order inclusion probabilities are $\pi_{\gamma,k,\ell} = n_{\gamma h}(n_{\gamma h}-1)\{N_{\gamma h}(N_{\gamma h}-1)\}^{-1}$ for $k, \ell \in U_{\gamma h}$, $k \neq \ell$; and $\pi_{\gamma,k,\ell} = n_{\gamma h} N_{\gamma h}^{-1} n_{\gamma h'} N_{\gamma h'}^{-1}$ for $k \in U_{\gamma h}$, $\ell \in U_{\gamma h'}$, $h \neq h'$.

In the following, by convention $0/0 = 0$. For $\gamma \in \mathbb{N}$, let $\mathbf{x}_\gamma = (x_1, \ldots, x_{N_\gamma}) \in X(\Omega)^{N_\gamma}$ denote a matrix of fixed covariate vectors for the entire finite population.

Remark 1 (Bounded in probability and convergence in probability). Let W_γ be a sequence of random variables in a real vector space of finite dimension and let $\|\cdot\|$ denote the Euclidean norm. Write $W_\gamma = o_{\mathrm{P}_{\theta,\xi}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1)$ if $\forall \varepsilon > 0$, $\lim_\gamma \mathrm{P}_{\theta,\xi}(\|W_\gamma\| > \varepsilon | \mathcal{X}_\gamma = \mathbf{x}_\gamma) = 0$, and $W_\gamma = O_{\mathrm{P}_{\theta,\xi}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1)$ if $\forall \varepsilon > 0$, $\exists M > 0$ such that $\sup_{\gamma \in \mathbb{N}} \mathrm{P}_{\theta_0,\xi_0}(\|W_\gamma\| > M | \mathcal{X}_\gamma = \mathbf{x}_\gamma) < \varepsilon$.

2.2. The limit sample p.d.f.

We now define the limit conditional sample p.d.f. of Y_k given $X_k = x_k$. In general, the conditional sample p.d.f. depends not only on X_k but also on numerical summaries $h_\gamma(\mathcal{X}_\gamma)$ of all the finite population X-values, and the limit conditional sample p.d.f. will depend on a limit h_∞ of those summaries.

Assumption A0. *There exists $d \in \mathbb{N}$ such that $\forall(\theta, \xi) \in \Theta \times \Xi$, there exists a sequence of functions $h_\gamma : \mathcal{X}_\gamma(\Omega) \to \mathbb{R}^d$, a vector $h_\infty \in \mathbb{R}^d$, a sequence of functions $m_{\gamma,\theta,\xi} : X(\Omega) \times \mathbb{R}^d \times Y(\Omega) \to \mathbb{R}$ and a function $m_{\infty,\theta,\xi} : X(\Omega) \times Y(\Omega) \to \mathbb{R}$, with*

$$\lim_{\gamma \to \infty} h_\gamma(\mathbf{x}_\gamma) = h_\infty, \tag{A0.a}$$

$$\begin{aligned} &\forall \gamma \in \mathbb{N}, \mathbf{x}_\gamma \in \mathcal{X}_\gamma(\Omega), y \in Y(\Omega), \\ &\quad \mathrm{E}_{\theta,\xi}[J_{\gamma,1} | \mathcal{X}_\gamma = \mathbf{x}_\gamma, Y_1 = y] = m_\gamma\big(x_1, h_\gamma(\mathbf{x}_\gamma), y; \theta, \xi\big), \end{aligned} \tag{A0.b}$$

$$\begin{aligned} &\forall x, y \in X(\Omega) \times Y(\Omega), \\ &\quad \lim_{\gamma \to \infty} m_\gamma\big(x, h_\gamma(\mathbf{x}_\gamma), y; \theta, \xi\big) = m_\infty(x, h_\infty, y; \theta, \xi). \end{aligned} \tag{A0.c}$$

Definition 1 (The limit conditional sample p.d.f.). *Under* A0, *the limit sample p.d.f. of response* Y *given covariate* $X = x$ *is*

$$\rho_\infty(x, \cdot; \theta, \xi) f_\theta(\cdot|x),$$

where $\rho_\infty(x, y; \theta, \xi) = (\int m_\infty(x, h_\infty, y; \theta, \xi) \,\mathrm{d}\,\mathrm{P}^{Y|X=x})^{-1} m_\infty(x, h_\infty, y; \theta, \xi)$.

We also define the weight function for the variable X: assume that $N_\gamma^{-1} \sum_{k=1}^{N_\gamma} \delta_{x_k}$ converges weakly to P^X, then for $x \in X(\Omega)$, $\bar{\rho}_\infty(x; \theta, \xi)$ is defined as

$$\bar{\rho}_\infty(x; \theta, \xi) = \left(\int m_\infty(x, h_\infty, y; \theta, \xi) \,\mathrm{d}\,\mathrm{P}_{\theta,\xi}^{X,Y}(x, y) \right)^{-1} \int m_\infty(x, h_\infty, y; \theta, \xi) \,\mathrm{d}\,\mathrm{P}_{\theta,\xi}^{Y|X=x}(y),$$

and under the assumption that $\lim_{\gamma\to\infty} N_\gamma^{-1} \mathrm{E}_{\theta,\xi}[n_\gamma] = \tau \in (0, 1)$,

$$\bar{\rho}_\infty(x; \theta, \xi) = \tau^{-1} \int m_\infty(x, h_\infty, y; \theta, \xi) \,\mathrm{d}\,\mathrm{P}_{\theta,\xi}^{Y|X=x}(y).$$

2.3. Results for stratified sampling

In this section, we prove a result of independent interest, showing that suitably normalized sums over the sample under this informative selection scheme are asymptotically normal. This result will be used in deriving the asymptotic distribution of the maximum sample likelihood estimator in an example of Section 4.

For $h \in \{0, \ldots, H\}$, define $t_{\gamma h} = T_{\gamma h} N_\gamma^{-1}$ and assume that $t_{\infty,h} = \lim_{\gamma\to\infty} t_{\gamma h}$ is defined. Further, for $h \in \{1, \ldots, H\}$, assume that $\tau_h = \lim_{\gamma\to\infty} n_{\gamma h} N_{\gamma h}^{-1}$ is well-defined and strictly positive. Then $\tau = \lim_{\gamma\to\infty} n_\gamma N_\gamma^{-1} = \sum_{h=1}^H \tau_h (t_{\infty,h} - t_{\infty,h-1})$, which we assume to be strictly positive.

Theorem 1. *Let* g *be a measurable function from* $(Y(\Omega) \times Z(\Omega) \times [0, 1])$ *to some finite-dimensional real vector space. Assume that there exists* $G : Y(\Omega) \times Z(\Omega) \to [0, \infty)$ *such that* $\mathrm{E}[G(Y, Z)^2] < \infty$, *and* $\|g(y, z, \pi)\| \le G(y, z)$ *for all* $y, z, \pi \in Y(\Omega) \times Z(\Omega) \times [\min\{\tau_{\infty,h} | h \in \{1, \ldots, H\}\}, 1]$. *Assume that* $\forall y \in Y(\Omega)$, $g(y, \cdot, \cdot) : Z(\Omega) \times (0, 1] \to \mathbb{R}$ *is continuous. Define* $S_{\gamma h} = \sum_{k \in U_{\gamma h}} g(Y_k, Z_k, \pi_{\gamma,k}) J_{\gamma,k}$ *and* $S_\gamma = \sum_{h=1}^H S_{\gamma h}$.

Let $\zeta(t) = \inf\{x | P(Z \le x) \ge t\}$ *be the quantile function of* Z (*see Serfling* [26], *page* 74), *which implicitly depends on* θ *and* ξ. *Assume that for all* θ *and* ξ, $\zeta(\cdot)$ *is a continuous function. For* $h \in \{1, \ldots, H\}$, *define* $\zeta_h = \zeta(t_{\infty,h})$ *and*

$$E_{\infty,h} = \mathrm{E}\big[g(Y_k, Z_k, \tau_{\infty,h}) | Z_k \in (\zeta_{h-1}, \zeta_h]\big], \qquad E_\infty = \sum_{h=1}^H \tau_{\infty,h} E_{\infty,h},$$

$$V_{\infty,h} = \mathrm{Var}\big[g(Y_k, Z_k, \tau_{\infty,h}) | Z_k \in (\zeta_{h-1}, \zeta_h]\big], \qquad V_\infty = \tau^{-1} \sum_{h=1}^H (t_{\infty,h} - t_{\infty,h-1}) \tau_{\infty,h} V_{\infty,h}.$$

Assume that $\forall z_0 \in \mathbb{R}$, $\exists O$ *an open subset of* $\mathbb{R}$, $\exists M : \mathbb{R} \to \mathbb{R}$ *a positive and measurable function such that* $\int M(y)\,\mathrm{d}\lambda < \infty$, $\int \sup_{z\in O}\{G(y,z)\}M(y)\,\mathrm{d}\lambda(y) < \infty$, $\int \sup_{z\in O}\{G^2(y,z)\} \times M(y)\,\mathrm{d}\lambda(y) < \infty$ *and* $\forall z \in K$, $y \in Y(\Omega)$, $\mathrm{f}_{Y|Z=z} < M(y)$. *Then*

$$\sqrt{n_\gamma}\big(n_\gamma^{-1}S_\gamma - E_\infty\big) \xrightarrow[\gamma\to\infty]{\mathscr{L}} \mathscr{N}(0, V_\infty).$$

Proof. See Appendix A. □

3. Maximum sample likelihood estimation

The following series of results follows the development in Gong and Samaniego [12], adapted to the context of informative selection which we have described above.

3.1. Plug-in maximum sample likelihood estimation

Definition 2. *Under* A0, *define*

$$\Delta(x, y; \theta, \xi) = \ln\big(\rho_\infty(x, y; \theta, \xi)\, f_\theta(y|x)\big)$$

and for $\gamma \in \mathbb{N}$, *define the mean log sample likelihood as*

$$\bar{\mathscr{L}}_\gamma(\theta, \xi) = n_\gamma^{-1} \sum_{k=1}^{N_\gamma} J_{\gamma,k}\Delta(x_k, Y_k, \theta, \xi).$$

Assume that $\widehat{\xi}_\gamma$ *is a sequence of estimators of* ξ *and define the maximum sample likelihood estimator of* θ *adapted to* $\widehat{\xi}_\gamma$ *as*

$$\widehat{\theta}_\gamma = \arg\max_{\theta\in\Theta}\big\{\bar{\mathscr{L}}_\gamma(\theta, \widehat{\xi}_\gamma)\big\}.$$

Assume that the following information matrices are defined:

$$\mathscr{I}_{\gamma,1,1} = -N_\gamma^{-1}\sum_{k=1}^{N_\gamma}\int\left(\frac{\partial^2\Delta}{\partial\theta\,\partial\theta^{\mathrm{T}}}(x_k, y, \theta_0, \xi_0)\right)\rho_\infty(x_k, y; \theta_0, \xi_0)\bar{\rho}_\infty(x_k; \theta, \xi)\,\mathrm{d}\,\mathrm{P}_{\theta_0,\xi_0}^{Y|X=x_k}(y)$$

and

$$\mathscr{I}_{\gamma,1,2} = -N_\gamma^{-1}\sum_{k=1}^{N_\gamma}\int\left(\frac{\partial^2\Delta}{\partial\xi\,\partial\theta^{\mathrm{T}}}(x_k, y, \theta_0, \xi_0)\right)\rho_\infty(x_k, y; \theta_0, \xi_0)\bar{\rho}_\infty(x_k; \theta, \xi)\,\mathrm{d}\,\mathrm{P}_{\theta_0,\xi_0}^{Y|X=x_k}(y).$$

Assumption A1 (Asymptotic normality conditions). *Assume* A0 *and assume that*:
for $u \in \left\{ \left(\frac{\partial^2 \Delta}{\partial\theta\,\partial\theta^{\mathrm{T}}}(\cdot,\theta_0,\xi_0) \right), \left(\frac{\partial^2 \Delta}{\partial\theta\,\partial\xi^{\mathrm{T}}}(\cdot,\theta_0,\xi_0) \right) \right\}$

$$N_\gamma^{-1} \sum_{k=1}^{N_\gamma} \mathrm{E}_{\theta,\xi}\big[u(x_k,Y_k)\rho_\infty(x_k,Y_k;\theta,\xi)|X_k = x_k\big]\bar{\rho}_\infty(x_k;\theta,\xi)$$
$$- n_\gamma^{-1} \sum_{k=1}^{N_\gamma} u(x_k,Y_k) J_{\gamma,k} = o_{\mathrm{P}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}_{\theta_0,\xi_0}}(1), \tag{A1.a}$$

$$\sqrt{n_\gamma}(\widehat{\xi}_\gamma - \xi_0) = O_{\mathrm{P}_{\theta_0,\xi_0}{}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1), \tag{A1.b}$$

$$\mathscr{I}_{\gamma,1,1} \text{ and } \mathscr{I}_{\gamma,1,2} \text{ are finite and } \mathscr{I}_{\gamma,1,1} \text{ is positive-definite}, \tag{A1.c}$$

$$\mathscr{I}_{\gamma,1,1} \text{ converges to a finite and positive-definite matrix}, \tag{A1.d}$$

$$-\sqrt{n_\gamma}\frac{\partial\bar{\mathscr{L}}}{\partial\theta}(\theta_0,\widehat{\xi}_\gamma) = \sqrt{n_\gamma}\left(\frac{\partial^2\bar{\mathscr{L}}}{\partial\theta\,\partial\theta^{\mathrm{T}}}(\theta_0,\widehat{\xi}_\gamma)\right)(\widehat{\theta}_\gamma - \theta_0) + o_{\mathrm{P}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}_{\theta_0,\xi_0}}(1), \tag{A1.e}$$

$$\begin{aligned}\sqrt{n_\gamma}\frac{\partial\bar{\mathscr{L}}}{\partial\theta}(\theta_0,\widehat{\xi}_\gamma) &= \sqrt{n_\gamma}\frac{\partial\bar{\mathscr{L}}}{\partial\theta}(\theta_0,\xi_0)\\ &\quad + \sqrt{n_\gamma}\left(\frac{\partial}{\partial\xi}\left(\frac{\partial}{\partial\theta}\bar{\mathscr{L}}\right)(\theta_0,\xi_0)\right)(\widehat{\xi}_\gamma - \xi_0) + o_{\mathrm{P}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}_{\theta_0,\xi_0}}(1).\end{aligned} \tag{A1.f}$$

There exists a sequence of positive-definite matrices $\left\{\Sigma_\gamma = \left[\begin{smallmatrix}\Sigma_{\gamma,1,1} & \Sigma_{\gamma,1,2}^{\mathrm{T}}\\ \Sigma_{\gamma,1,2} & \Sigma_{\gamma,2,2}\end{smallmatrix}\right]\right\}_{\gamma\in\mathbb{N}}$ *such that* $\Sigma_{\gamma,1,1}$ *is a* $\dim(\theta)\times\dim(\theta)$ *matrix and given* $\mathcal{X}_\gamma = \mathbf{x}_\gamma$,

$$\sqrt{n_\gamma}\Sigma_\gamma^{-1/2}\begin{bmatrix}\left(\frac{\partial}{\partial\theta}\bar{\mathscr{L}}\right)(\theta_0,\xi_0)\\ \widehat{\xi}_\gamma - \xi_0\end{bmatrix} \xrightarrow[\gamma\to\infty]{\mathcal{L}} \mathscr{N}(0,I). \tag{A1.g}$$

Theorem 2. *Under* A1, *the maximum sample likelihood estimator adapted to* $\widehat{\xi}_\gamma$ *is asymptotically normal, that is*

$$\sqrt{n_\gamma}V_\gamma^{-1/2}(\widehat{\theta}_\gamma - \theta_0) \xrightarrow[\gamma\to\infty]{\mathcal{L}} \mathscr{N}(0,I),$$

where the square root of a positive-definite matrix is its only positive square root by convention and $V_\gamma = \mathscr{I}_{\gamma,1,1}^{-1}(\Sigma_{\gamma,1,1} + \mathscr{I}_{\gamma,1,2}\Sigma_{\gamma,2,2}\mathscr{I}_{\gamma,1,2}^{\mathrm{T}} - \mathscr{I}_{\gamma,1,2}\Sigma_{\gamma,1,2}^{\mathrm{T}} - \Sigma_{\gamma,1,2}\mathscr{I}_{\gamma,1,2}^{\mathrm{T}})\mathscr{I}_{\gamma,1,1}^{-1}$.

Proof. See Appendix B. □

3.2. Variance estimation

Depending on what is known about the covariates, various estimators of the information matrix and of Σ can be proposed. If P^X is known, then $\mathscr{I}_{\gamma,1,1}$ can be estimated by

$$\int \left(\int \left(\frac{\partial^2 \Delta}{\partial\theta\, \partial\theta^{\mathrm{T}}}(x, y, \widehat{\theta}, \widehat{\xi}) \right) \rho_\infty(x, y; \widehat{\theta}, \widehat{\xi}) \bar{\rho}_\infty(x; \widehat{\theta}, \widehat{\xi}) \, \mathrm{d}\, \mathrm{P}_{\widehat{\theta}}^{Y|X=x}(y) \right) \mathrm{d}\, \mathrm{P}_{\widehat{\theta}}^{X}(x).$$

If the population vector $\mathbf{x}_\gamma$ is known, then $\mathscr{I}_{\gamma,1,1}$ can be estimated by

$$-{N_\gamma}^{-1} \sum_{k=1}^{N_\gamma} \int \left(\frac{\partial^2 \Delta}{\partial\theta\, \partial\theta^{\mathrm{T}}}(x_k, Y, \widehat{\theta}, \widehat{\xi}) \right) \rho_\infty(x_k, Y; \widehat{\theta}, \widehat{\xi}) \bar{\rho}_\infty(x_k; \widehat{\theta}, \widehat{\xi}) \, \mathrm{d}\, \mathrm{P}_{\widehat{\theta}}^{Y|X=x_k}.$$

If $\mathbf{x}$ is known on the sample, then $\mathscr{I}_{\gamma,1,1}$ can be estimated by its Horvitz–Thompson estimator:

$$\frac{-1}{N_\gamma} \sum_{k=1}^{N_\gamma} \frac{J_{\gamma,k}}{\pi_{\gamma,k}} \int \left(\frac{\partial^2 \Delta}{\partial\theta\, \partial\theta^{\mathrm{T}}}(x_k, Y, \widehat{\theta}, \widehat{\xi}) \right) \rho_\infty(x_k, Y; \widehat{\theta}, \widehat{\xi}) \bar{\rho}_\infty(x_k; \widehat{\theta}, \widehat{\xi}) \, \mathrm{d}\, \mathrm{P}_{\widehat{\theta}}^{Y|X=x_k}.$$

The matrix $\mathscr{I}_{\gamma,1,2}$ can also be estimated using the same methods.

Since $\widehat{\xi}$ will often be obtained via pseudo-likelihood methods, it is convenient to estimate Σ_γ in (A1.g) using design-based methods from standard software. To do so requires expressing the score vector $(\frac{\partial}{\partial\theta}\bar{\mathscr{L}})(\theta_0, \xi_0)$ as a design-weighted sum over the sample, then plugging in $(\widehat{\theta}, \widehat{\xi})$ for (θ_0, ξ_0). Design-based variance estimation might also require linearization of $\widehat{\xi}$. We illustrate these ideas in Section 4.4. Alternatively, the matrix Σ_γ could be estimated by analytically computing $\mathrm{Var}_{\theta,\xi}\big[\big[\begin{smallmatrix} (\frac{\partial}{\partial\theta}\bar{\mathscr{L}})(\theta_0,\xi_0) \\ \widehat{\xi}_\gamma - \xi_0 \end{smallmatrix}\big] | \mathcal{X}_\gamma = \mathbf{x}_\gamma\big]$ and plugging in the estimates of θ and ξ, or it could be computed by Monte Carlo methods.

4. Examples and simulations

4.1. Pareto distribution and Bernoulli sampling

We begin with a simple example with no covariate. Let $Y_1, \ldots, Y_{N_\gamma}$ be i.i.d. Pareto with p.d.f. $f_\theta(y) = \theta y^{-(\theta+1)} \mathbb{1}_{[1,\infty)}(y)$.

Assume that P^{Y_k}-a.s. (y), $\mathrm{P}^{Z_k|Y_k=y} = \mathscr{B}(y^{-\xi})$, the Bernoulli distribution with success probability $y^{-\xi}$, and $\Xi = (0, \infty)$. The sample scheme is stratified Bernoulli sampling with two strata determined by the realized Z_k's, and with sampling rates $\tau_0 = 0.02$ in stratum 0 and $\tau_1 = 0.1$ in stratum 1. Then $\forall\, j \in \{0,1\}^{N_\gamma}, z \in \{0,1\}^{N_\gamma}$,

$$D_\gamma(z)\big(\{j\}\big) = \prod_{k=1}^{N_\gamma} \big(\tau_1^{j_k}(1-\tau_1)^{1-j_k}\big)^{z_{\gamma,k}} \big(\tau_0^{j_k}(1-\tau_0)^{1-j_k}\big)^{1-z_{\gamma,k}}.$$

Ignoring the informative selection mechanism and maximizing the log-likelihood leads to the naive estimator

$$\bar{\theta}_\gamma = \left(\sum_{k=1}^{N_\gamma} \ln(Y_k) J_{\gamma,k}\right)^{-1} \left(\sum_{k=1}^{N_\gamma} J_{\gamma,k}\right),$$

which is biased and inconsistent. The maximum pseudo-likelihood estimator is obtained by maximizing the weighted log-likelihood, yielding

$$\widetilde{\theta}_\gamma = \left(\sum_{k=1}^{N_\gamma} \ln(Y_k) J_{\gamma,k} \Big/ \pi_{\gamma,k}\right)^{-1} \left(\sum_{k=1}^{N_\gamma} J_{\gamma,k} \Big/ \pi_{\gamma,k}\right);$$

this estimator is asymptotically unbiased and consistent. To obtain the maximum sample likelihood estimator, first compute

$$\rho_\gamma(y;\theta,\xi) = \rho_\infty(y;\theta,\xi) = \big(\tau_0 + (\tau_1 - \tau_0)\theta(\xi+\theta)^{-1}\big)^{-1}\big(\tau_0 + (\tau_1 - \tau_0) y^{-\xi}\big)$$

for all $\gamma \in \mathbb{N}$ and $y \in [1, \infty)$. Let

$$\theta^*_\gamma = \left(\sum_{k=1}^{N_\gamma} (Y_k - 1) J_{\gamma,k} \Big/ \pi_{\gamma,k}\right)^{-1} \left(\sum_{k=1}^{N_\gamma} Y_k J_{\gamma,k} \Big/ \pi_{\gamma,k}\right)$$

denote the Horvitz–Thompson plug-in estimator of $\theta = (\mathrm{E}_\theta[Y] - 1)^{-1}(\mathrm{E}_\theta[Y])$. A Horvitz–Thompson plug-in estimator of ξ is then obtained as

$$\widehat{\xi}_\gamma = 1 + \theta^*_\gamma \left(\sum_{k=1}^{N_\gamma} (Y_k Z_k) J_{\gamma,k} \Big/ \pi_{\gamma,k}\right)^{-1} \left(\sum_{k=1}^{N_\gamma} (1 - Y_k Z_k) J_{\gamma,k} \Big/ \pi_{\gamma,k}\right).$$

Finally, the maximum sample likelihood estimator of θ is obtained by numerical maximization: $\widehat{\theta}_\gamma = \arg\max_{\theta\in\Theta}\{\bar{\mathscr{Z}}_\gamma(\theta, \widehat{\xi}_\gamma)\}$. Straightforward calculations then yield the asymptotic variance of Theorem 2. For comparison, we also consider the full likelihood with N_γ considered as unknown. (The naive, pseudo-likelihood and sample likelihood estimators do not require N_γ known.) We maximize

$$\ln\binom{N_\gamma}{n_\gamma} + (N_\gamma - n_\gamma)\ln\left[1 - \left\{\tau_0 + (\tau_1 - \tau_0)\frac{\theta}{\theta+\xi}\right\}\right]$$
$$+ \sum_{k=1}^{N_\gamma} J_{\gamma,k} \ln\big(f_\theta(Y_k)\big(Z_k\, Y_k^{-\xi} + (1 - Z_k)\big(1 - Y_k^{-\xi}\big)\big)\big(\tau_0 + (\tau_1 - \tau_0) Z_k\big)\big)$$

with respect to $(N_\gamma \in \mathbb{N}, \theta \in \Theta, \xi \in \Xi)$ to obtain the "Full" estimators in the following tables.

Our first simulation results are presented in Table 1. We generated 1000 independent replicates of $(\mathcal{Y}_\gamma, \mathcal{Z}_\gamma, \mathcal{J}_\gamma)$ under the Pareto model with $N_\gamma = 10\,000$ elements each. For each replicate, we computed the naive, pseudo-likelihood, sample likelihood, and full likelihood estimators. Even in this simple example, the full likelihood estimator is much more complicated

Table 1. Simulation results based on 1000 replications for Pareto example of Section 4.1 with $\theta = 4$, $\tau_0 = 0.01$, $\tau_1 = 0.1$, and $N_\gamma = 10\,000$

	Estimator	Mean	% Relative bias	RMSE ratio	Empirical variance	Asymptotic variance
$\xi = 0.1$	Naive	4.09	2.22	1.17	0.017	
	Pseudo	4.01	0.24	1.04	0.020	
	Sample	4.01	0.26	1.00	0.018	0.018
	Full	4.01	0.74	0.98	0.018	
$\xi = 1$	Naive	4.73	18.19	4.08	0.026	
	Pseudo	4.02	0.46	1.16	0.045	
	Sample	4.01	0.31	1.00	0.033	0.033
	Full	4.02	0.40	0.86	0.030	
$\xi = 2$	Naive	5.30	32.38	6.12	0.043	
	Pseudo	4.02	0.47	1.11	0.056	
	Sample	4.02	0.38	1.00	0.046	0.044
	Full	4.01	0.27	0.76	0.033	

to implement than the sample likelihood estimator. Empirical means, percent relative biases $((\text{mean} - \theta)/\theta) \times 100\%$, root mean squared error (RMSE) ratios relative to the maximum sample likelihood estimator, and empirical variances across these 1000 replicates are summarized in Table 1 for $\theta = 4$ and different values of ξ, with larger ξ corresponding to greater informativeness. Also tabled is the asymptotic variance computed from Theorem 2 of the maximum sample likelihood estimator.

As expected, the naive estimator $\bar{\theta}_\gamma$ that ignores informative selection is badly biased except in the case closest to noninformative selection. (All four estimators are identical when $\xi = 0$.) The maximum pseudo-likelihood estimator $\widetilde{\theta}_\gamma$ is essentially unbiased, but is somewhat less efficient than the maximum sample likelihood estimator $\widehat{\theta}_\gamma$. In all but the most informative case, the maximum sample likelihood estimator has performance comparable to that of the estimator that maximizes the full likelihood. The asymptotic variance of $\widehat{\theta}_\gamma$ computed from Theorem 2 is an excellent approximation to the empirical variance in all cases.

We extended the comparisons of Table 1 to other population sizes and sampling rates within strata. Additional results for $N_\gamma \in \{100, 1000, 10\,000\}$ and $\tau \in \{(0.01, 0.1), (0.04, 0.5), (0.09, 0.95)\}$ are not shown but are qualitatively similar. In each case, the full likelihood estimator performs best with respect to mean squared error. Of the remaining estimators, the maximum sample likelihood estimator outperforms the maximum pseudo-likelihood estimator, and often approaches the efficiency of the maximum likelihood estimator. This same ordering of efficiency, with sample likelihood dominating pseudo-likelihood, is found in our other simulation studies below, and elsewhere in the literature (e.g., Krieger and Pfeffermann [15]; Chambers *et al.* [5], Section 3.3).

4.2. Normal distribution and stratified sampling: Without covariate

We next consider the stratified design described in Section 2.3, with normal distributions for the stratification variable Z_k and the response variable Y_k. First, we derive analytic results in detail for the case of no covariates. Similar results hold for the case with covariates, but the notation is considerably more complicated. We therefore omit such derivations and instead report on simulation results for the case with a covariate in Section 4.3.

Assume that $Y_k \sim \mathcal{N}(\theta, 1)$ and $\mathrm{P}^{Z_k|Y_k} = \mathcal{N}(\xi Y_k, \sigma_\eta^2)$, where σ_η is known; that is, $Z_k = \xi Y_k + \eta_k$ and $\eta_k \sim \mathcal{N}(0, \sigma_\eta^2)$. Further, assume that η_k are mutually independent and independent of $\mathcal{Y}_\gamma$.

Result 1. *Under this asymptotic framework,* A0 *holds and*

$$\rho_\infty(y;\theta,\xi) = \left(\tau_H + \sum_{h=1}^{H-1} t_{\infty,h}(\tau_h - \tau_{h+1})\right)^{-1} \left(\tau_H + \sum_{h=1}^{H-1} (\tau_h - \tau_{h+1})\Phi\left(\frac{\zeta_h - \xi y}{\sigma_\eta}\right)\right),$$

where $\zeta_h = \sqrt{\xi^2 + \sigma_\eta^2}\,\Phi^{-1}(t_{\infty,h}) + \xi\theta$ *and* Φ *is the c.d.f. of* $\mathcal{N}(0,1)$.

Proof. See Appendix C.1. □

The sample likelihood is then defined. Consider $\widehat{\xi}_\gamma = (\sum_{k=1}^{N_\gamma} Y_k^2 J_{\gamma,k}/\pi_{\gamma,k})^{-1}(\sum_{k=1}^{N_\gamma} Z_k \times Y_k J_{\gamma,k}/\pi_{\gamma,k})$; this is a standard Horvitz–Thompson plug-in estimator, but its asymptotic behavior is not immediate due to the dependence of $\pi_{\gamma,k}$ on the ordering of the Z_k's. Under the asymptotic framework described above, we can establish the following result:

Result 2. *The statistic* $\widehat{\xi}_\gamma$ *is a consistent estimator of* ξ.

Proof. See Appendix C.2. □

We then define $\widehat{\theta}_\gamma$ to be the maximum sample likelihood estimator of θ adapted to $\widehat{\xi}_\gamma$, that is,

$$\widehat{\theta}_\gamma = \arg\max_{\theta\in\Theta} \{\bar{\mathscr{L}}(\theta, \widehat{\xi}_\gamma)\}.$$

Result 3. *Under the above asymptotic framework, assumption* A1 *(the conditions of Theorem 2) is satisfied, and*

$$\Sigma = \sum_{h=1}^{H} (t_{\infty,h} - t_{\infty,h-1}) \frac{\tau_{\infty,h}}{\tau} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \dfrac{1}{\theta_0^2+1} & \dfrac{-\xi_0}{\theta_0^2+1} \end{bmatrix}$$
$$\times \mathrm{Var}_{\theta_0,\xi_0} \left[\begin{array}{c} \dfrac{\partial \Delta(Y,\theta_0,\xi_0)}{\partial\theta} \\ YZ/\tau_{\infty,h} \\ Y^2/\tau_{\infty,h} \end{array} \middle| Z \in (\zeta_{h-1}, \zeta_h] \right] \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{1}{\theta_0^2+1} \\ 0 & \dfrac{-\xi_0}{\theta_0^2+1} \end{bmatrix},$$

$$\left(\frac{\partial}{\partial\theta}\Delta\right)(y,\theta_0,\xi_0) = (y-\theta) + \frac{\sum_{h=1}^{H-1}(\tau_h-\tau_{h+1})\frac{\xi}{\sigma_\eta}f_0(\frac{\zeta_h-\xi y}{\sigma_\eta})}{\tau_H+\sum_{h=1}^{H-1}(\tau_h-\tau_{h+1})\Phi(\frac{\zeta_h-\xi y}{\sigma_\eta})}$$

and

$$\left(\frac{\partial}{\partial\xi}\Delta\right)(y,\theta_0,\xi_0) = \frac{\sum_{h=1}^{H-1}(\tau_h-\tau_{h+1})(\frac{\frac{\xi}{\sqrt{\xi^2+\sigma_\eta^2}}\Phi^{-1}(t_{\infty,h})+(\theta-y)}{\sigma_\eta})f_0(\frac{\zeta_h-\xi y}{\sigma_\eta})}{\tau_H+\sum_{h=1}^{H-1}(\tau_h-\tau_{h+1})\Phi(\frac{\zeta_h-\xi y}{\sigma_\eta})}.$$

Proof. See Appendix C.3. □

4.3. Normal distribution and stratified sampling: Covariate case

We now extend the previous model with a continuous covariate, denoted X. Assume $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0,\sigma_\varepsilon^2)$, and β_0, β_1, σ_ε are unknown real numbers. We are interested in the estimation of $\theta = (\beta_0,\beta_1,\sigma_\varepsilon)^{\mathrm{T}}$. Assume that the empirical population c.d.f. of $\mathbf{x}_\gamma$ converges uniformly to a normal c.d.f. with known parameters μ_X and σ_X^2.

Result 4. *Under this setting, the assumptions of Theorem 2 hold, and*

$$\rho_\infty(x,y;\theta,\xi) = \left(\tau_H + \sum_{h=1}^{H-1}(\tau_h-\tau_{h+1})\Phi\left(\frac{\zeta_h-\xi\beta_0-\xi\beta_1 x_k}{\sqrt{\xi^2\sigma_\varepsilon^2+\sigma_\eta^2}}\right)\right)^{-1} \times\left(\tau_H+\sum_{h=1}^{H-1}(\tau_h-\tau_{h+1})\Phi\left(\frac{\zeta_h-\xi y}{\sqrt{\sigma_\eta^2}}\right)\right),$$

where $\zeta_h = (\sqrt{\xi^2\beta_1^2\sigma_X^2+\xi^2\sigma_\varepsilon^2+\sigma_\eta^2})\Phi^{-1}(t_{\infty,h}) + \xi(\beta_0+\beta_1\mu_X)$, *and* $\Phi(\cdot)$ *is the c.d.f. of* $\mathcal{N}(0,1)$.

Proof. The proof is similar to that of Result 3 and is omitted. □

As in the previous example, the Horvitz–Thompson plug-in estimator of ξ is

$$\widehat{\xi}_\gamma = \left(\sum_{k=1}^{N_\gamma} Y_k^2 J_{\gamma,k}\Big/\pi_{\gamma,k}\right)^{-1}\left(\sum_{k=1}^{N_\gamma} Z_k Y_k J_{\gamma,k}\Big/\pi_{\gamma,k}\right).$$

The maximum sample likelihood estimators are then obtained as $\widehat{\theta}_\gamma = \arg\max\{\sum_{k=1}^{N_\gamma}\ln(\rho_\infty(X_k, Y_k;\theta,\widehat{\xi}_\gamma)f_{\theta,\xi}(Y_k|X_k)J_{\gamma,k})\}$.

Ignoring the informative selection mechanism and maximizing the log-likelihood leads to the ordinary least squares estimators $\bar{\theta}_\gamma$ of the parameters θ. These are biased and inconsistent under

informative selection. The maximum pseudo-likelihood estimators are obtained by maximizing the weighted log-likelihood, yielding weighted least squares estimators $\widetilde{\theta}_\gamma$, with weights given by inverse inclusion probabilities. These estimators are asymptotically unbiased and consistent.

We simulated a realization of $N_\gamma = 5000$ i.i.d. values $\mathbf{x}_\gamma$ from a normal distribution with parameters $\mu_X = 1$ and $\sigma_X^2 = 1$. Keeping $\mathbf{x}_\gamma$ fixed and choosing $\sigma_\eta \in \{0.1, 1, 10\}$, we then generated 1000 independent replicates of $(\mathcal{Y}_\gamma, \mathcal{Z}_\gamma, \mathcal{J}_\gamma)$, using $H = 2$ strata with $N_{\gamma 1} = 3500$, $N_{\gamma 2} = 1500$, $n_{\gamma 1} = 50$, $n_{\gamma 2} = 200$, $\beta_0 = 1/2$, $\beta_1 = 1$, $\sigma_\varepsilon = 2$ and $\xi = 2$. For each replicate, we computed $\bar{\theta}_\gamma$, $\widetilde{\theta}_\gamma$ and $\widehat{\theta}_\gamma$. For the latter, we computed ρ_∞ by approximating the limits τ_h and $t_{\infty,h}$: setting $\tau_1 = N_{\gamma 1}^{-1} n_{\gamma 1} = 1/70$, $\tau_2 = N_{\gamma 2}^{-1} n_{\gamma 2} = 2/15$, $t_{\infty,1} = N_\gamma^{-1} N_{\gamma 1} = 7/10$ and $t_{\infty,2} = N_\gamma^{-1} N_{\gamma 2} = 3/10$. Empirical means, percent relative biases, root mean squared error (RMSE) ratios relative to the maximum sample likelihood estimator, and empirical variances across these 1000 replicates are summarized in Table 2, with larger σ_η corresponding to greater informativeness. Also tabled is the asymptotic variance of the maximum sample likelihood estimators.

The ordinary least squares estimators $\bar{\theta}_\gamma$, which ignore informative selection, are badly biased except for the slope and noise variance under the least informative selection. The weighted least squares estimators $\widetilde{\theta}_\gamma$ obtained by maximum pseudo-likelihood estimation are essentially unbiased in every case, but are substantially less efficient than the maximum sample likelihood estimators $\widehat{\theta}_\gamma$. The asymptotic variance of $\widehat{\theta}_\gamma$ computed from Theorem 2 approximates the corresponding empirical variances very well in all cases.

4.4. Gestational age example

Our final example illustrates the use of the sample likelihood estimation and our asymptotic theory, along with variance estimation, when ρ_∞ is obtained empirically by the analyst, via regression of the design weights on the response variable as in Krieger and Pfeffermann [15]. Our simulation is motivated by a textbook example of at-risk infants from the 1988 National Maternal and Infant Health Survey (NMIHS), described in Korn and Graubard [14], Example 4.3-1. The design used birth certificates to oversample low birthweight infants. Fuller [10], Example 6.3.1, simulated a five-per-stratum design with 18 strata (90 observations) to mimic properties of NMIHS, and we used those data to generate parameter values for our simulation. Assume that only stratum and weight information is available to the analyst, and the goal is to model gestational age $Y_k \sim \mathcal{N}(\mu, \sigma^2)$, with no covariates. Since gestational age is highly correlated with birthweight, the design is informative. For this problem, an empirical model with considerable predictive power is $\mathrm{P}^{\ln W_k | Y_k} = \mathcal{N}(\xi_0 + \xi Y_k, \tau^2)$; that is, the design weights $W_k = \pi_k^{-1}$ are log-normally distributed. It is then easy to show that $\rho_\infty(y; \theta, \xi) f_\theta(y) = \mathcal{N}(\mu - \xi\sigma^2, \sigma^2)$, independent of ξ_0 and τ^2. Estimating ξ via pseudo-likelihood is equivalent to regressing $\ln W_k$ on Y_k using the design weights W_k, yielding

$$\widehat{\xi} = \frac{\sum_{k=1}^{N_\gamma} W_k \ln(W_k) Y_k J_{\gamma,k} - (\sum_{k=1}^{N_\gamma} W_k \ln(W_k) J_{\gamma,k})(\sum_{k=1}^{N_\gamma} W_k Y_k J_{\gamma,k}) / \sum_{k=1}^{N_\gamma} W_k J_{\gamma,k}}{\sum_{k=1}^{N_\gamma} W_k Y_k^2 J_{\gamma,k} - (\sum_{k=1}^{N_\gamma} W_k Y_k J_{\gamma,k})^2 / \sum_{k=1}^{N_\gamma} W_k J_{\gamma,k}}.$$

Table 2. Simulation results for normal example of Section 4.3, with $\theta^{\mathrm{T}} = (\beta_0, \beta_1, \sigma_\varepsilon) = (1/2, 1, 2)$

	Estimator	Mean	% Relative bias	RMSE ratio	Empirical variance	Asymptotic variance
$\sigma_\eta = 10$	Naive	$\begin{bmatrix} 1.14 \\ 0.966 \\ 1.96 \end{bmatrix}$	$\begin{bmatrix} 128 \\ -3.42 \\ -1.93 \end{bmatrix}$	$\begin{bmatrix} 2.95 \\ 0.999 \\ 1.03 \end{bmatrix}$	$\begin{bmatrix} 0.037 \\ 0.0161 \\ 7.47\ 10^{-3} \end{bmatrix}$	
	Pseudo	$\begin{bmatrix} 0.499 \\ 1.00 \\ 1.97 \end{bmatrix}$	$\begin{bmatrix} -0.21 \\ 0.43 \\ -1.4 \end{bmatrix}$	$\begin{bmatrix} 1.27 \\ 1.50 \\ 1.52 \end{bmatrix}$	$\begin{bmatrix} 0.0829 \\ 0.0388 \\ 0.0186 \end{bmatrix}$	
	Sample	$\begin{bmatrix} 0.496 \\ 1.00 \\ 1.99 \end{bmatrix}$	$\begin{bmatrix} -0.85 \\ 0.02 \\ -0.6 \end{bmatrix}$	1	$\begin{bmatrix} 0.0511 \\ 0.0173 \\ 8.24\ 10^{-3} \end{bmatrix}$	$\begin{bmatrix} 0.0507 \\ 0.0167 \\ 9.50\ 10^{-3} \end{bmatrix}$
$\sigma_\eta = 1$	Naive	$\begin{bmatrix} 2.21 \\ 0.804 \\ 1.79 \end{bmatrix}$	$\begin{bmatrix} 342 \\ -19.6 \\ -10.7 \end{bmatrix}$	$\begin{bmatrix} 8.07 \\ 1.73 \\ 2.23 \end{bmatrix}$	$\begin{bmatrix} 0.0359 \\ 0.0141 \\ 7.75\ 10^{-3} \end{bmatrix}$	
	Pseudo	$\begin{bmatrix} 0.499 \\ 1.01 \\ 1.98 \end{bmatrix}$	$\begin{bmatrix} -0.27 \\ 0.68 \\ -1.17 \end{bmatrix}$	$\begin{bmatrix} 1.29 \\ 1.43 \\ 1.25 \end{bmatrix}$	$\begin{bmatrix} 0.0758 \\ 0.036 \\ 0.0163 \end{bmatrix}$	
	Sample	$\begin{bmatrix} 0.506 \\ 1.00 \\ 1.99 \end{bmatrix}$	$\begin{bmatrix} 1.14 \\ 0.3 \\ -0.53 \end{bmatrix}$	1	$\begin{bmatrix} 0.0454 \\ 0.0176 \\ 0.0107 \end{bmatrix}$	$\begin{bmatrix} 0.0413 \\ 0.0182 \\ 0.0102 \end{bmatrix}$
$\sigma_\eta = 0.1$	Naive	$\begin{bmatrix} 2.28 \\ 0.782 \\ 1.78 \end{bmatrix}$	$\begin{bmatrix} 356 \\ -21.8 \\ -11.1 \end{bmatrix}$	$\begin{bmatrix} 9.88 \\ 1.85 \\ 2.5 \end{bmatrix}$	$\begin{bmatrix} 0.0348 \\ 0.0145 \\ 7.91\ 10^{-3} \end{bmatrix}$	
	Pseudo	$\begin{bmatrix} 0.506 \\ 0.998 \\ 1.98 \end{bmatrix}$	$\begin{bmatrix} 1.17 \\ -0.16 \\ -0.84 \end{bmatrix}$	$\begin{bmatrix} 1.47 \\ 1.41 \\ 1.34 \end{bmatrix}$	$\begin{bmatrix} 0.0712 \\ 0.0362 \\ 0.0162 \end{bmatrix}$	
	Sample	$\begin{bmatrix} 0.516 \\ 0.988 \\ 2.00 \end{bmatrix}$	$\begin{bmatrix} 3.16 \\ -1.16 \\ -0.08 \end{bmatrix}$	1	$\begin{bmatrix} 0.0326 \\ 0.0180 \\ 9.17\ 10^{-3} \end{bmatrix}$	$\begin{bmatrix} 0.0350 \\ 0.0188 \\ 8.48\ 10^{-3} \end{bmatrix}$

With $n_\gamma = \sum_{k=1}^{N_\gamma} J_{\gamma,k}$, the plug-in sample maximum likelihood estimators are then

$$\widehat{\sigma}^2 = \frac{\sum_{k=1}^{N_\gamma} Y_k^2 J_{\gamma,k} - (\sum_{k=1}^{N_\gamma} Y_k J_{\gamma,k})^2 / n_\gamma}{n_\gamma}, \qquad \widehat{\mu} = \frac{\sum_{k=1}^{N_\gamma} Y_k J_{\gamma,k}}{n_\gamma} + \widehat{\xi}\widehat{\sigma}^2.$$

These estimates can then be plugged into the information matrices, given by

$$\mathscr{I}_{\gamma,1,1} = \begin{bmatrix} \sigma^{-2} & -\xi\sigma^{-2} \\ -\xi\sigma^{-2} & (1/2)\sigma^{-4} + \xi^2\sigma^{-2} \end{bmatrix}, \qquad \mathscr{I}_{\gamma,1,2} = \begin{bmatrix} -1 \\ \xi \end{bmatrix}.$$

The score vector in (A1.g) has elements

$$\left(\frac{\partial}{\partial\mu}\bar{\mathscr{L}}\right)=\sum_{k=1}^{N_\gamma}\frac{1}{W_k n_\gamma}\left\{\frac{1}{\sigma^2}\left(Y_k-\mu+\xi\sigma^2\right)\right\}W_k J_{\gamma,k},$$

$$\left(\frac{\partial}{\partial\sigma^2}\bar{\mathscr{L}}\right)=\sum_{k=1}^{N_\gamma}\frac{1}{W_k n_\gamma}\left\{\frac{-1}{2\sigma^2}-\frac{\xi}{\sigma^2}\left(Y_k-\mu+\xi\sigma^2\right)+\frac{1}{2\sigma^4}\left(Y_k-\mu+\xi\sigma^2\right)^2\right\}W_k J_{\gamma,k}.$$

Plugging in $(\mu,\sigma^2,\xi)=(\widehat{\mu},\widehat{\sigma}^2,\widehat{\xi})$ yields two estimated totals, $\sum_{k=1}^{N_\gamma}S_{1k}W_k J_{\gamma,k}$ and $\sum_{k=1}^{N_\gamma}S_{2k}W_k J_{\gamma,k}$. The linearization of $\widehat{\xi}$ used in design-based variance estimation is also an estimated total, $\sum_{k=1}^{N_\gamma}L_k W_k J_{\gamma,k}$. Thus, by creating a data set with design information and the variables S_{1k}, S_{2k}, L_k, we can use standard survey software to obtain $\widehat{\Sigma}_\gamma$, the design-based covariance matrix estimate for the three estimated totals.

We chose $N_\gamma=15\,000$ and generated 1000 independent replicates of $\{Y_k\}_{k=1}^{N_\gamma}$ i.i.d. $\mathscr{N}(\mu,\sigma^2)$ and $\ln W_k=\exp(\xi_0+\xi Y_k+\varepsilon_k)$ with $\{\varepsilon_k\}_{k=1}^{N_\gamma}$ i.i.d. $\mathscr{N}(0,\tau^2)$ independent of $\{Y_k\}_{k=1}^{N_\gamma}$. We set $\mu=39.853$, $\sigma^2=16.723$, $\xi=0.175$, $\tau^2=0.087$, $n=90$ and

$$\xi_0=-\ln\left(\frac{n}{N_\gamma}\right)+\frac{\tau^2}{2}-\xi\mu+\frac{\xi^2\sigma^2}{2}\qquad\text{so that } N_\gamma^{-1}\sum_{k=1}^{N_\gamma}\pi_k\simeq N_\gamma^{-1}n.$$

For each simulated replicate, we used the $\{\pi_k\}=\{W_k^{-1}\}$ to draw two different without-replacement samples: an unstratified sample of size $n=90$ and a stratified five-per-stratum sample with 18 strata. The strata were formed by sorting the population on Y_k, cumulating the π_k's, and forming a new stratum each time the cumulative π_k's exceeded an integer multiple of five. For each sample, we computed the naive estimator $\bar{\mu}={n_\gamma}^{-1}\sum_{k=1}^{N_\gamma}Y_k J_{\gamma,k}$ that ignores informative selection, the maximum pseudo-likelihood estimator $\widetilde{\mu}=(\sum_{k=1}^{N_\gamma}W_k J_{\gamma,k})^{-1}\sum_{k=1}^{N_\gamma}W_k Y_k\times J_{\gamma,k}$, and the maximum sample likelihood estimator $\widehat{\mu}$. We also computed the standard design-based variance estimator for $\widetilde{\mu}$, and the new variance estimator as described above for $\widehat{\mu}$.

Empirical means, percent relative biases, root mean squared error (RMSE) ratios relative to the maximum sample likelihood estimator, and empirical variances across the 1000 replicates and both designs are summarized in Table 3. Also tabled are the average estimated variances and the ratio of average estimated variance to empirical variance. The naive estimators that ignore informative selection are badly biased under both designs. The weighted sample mean (Hájek estimator) obtained by maximum pseudo-likelihood estimation is essentially unbiased under each design, but is also less efficient under each design than the maximum sample likelihood estimator. The estimated variance for $\widehat{\mu}$ shows some tendency to underestimate, but its downward bias is comparable to that seen with standard variance estimation for $\widetilde{\mu}$.

Table 3. Simulation results based on 1000 replications for gestational age example of Section 4.4 with $\mu = 39.853$, $\sigma^2 = 16.723$, $\xi = 0.175$, $\tau^2 = 0.087$, $n = 90$ and $N_\gamma = 15\,000$

	Estimator	Mean	% Relative bias	RMSE ratio	Empirical variance	Average estimated variance	Variance ratio
Unstratified	Naive	36.949	−7.286	20.995			
	Pseudo	39.805	−0.122	1.106	0.452	0.419	0.927
	Sample	39.827	−0.065	1.000	0.410	0.388	0.946
Stratified	Naive	36.932	−7.328	113.911			
	Pseudo	39.858	0.013	2.448	0.184	0.169	0.918
	Sample	39.848	−0.012	1.000	0.075	0.066	0.880

5. Conclusion

In this paper, we have developed a precise asymptotic description of the behavior of the maximum sample likelihood estimator under informative selection from a finite population. We have shown that maximizing the sample likelihood, which treats the observations as if they were independently distributed according to a weighted distribution induced by the sample selection mechanism, is valid in the sense that the resulting estimators are consistent and asymptotically normal. This continues to hold even if nuisance parameters, which describe the selection mechanism but are not of scientific interest otherwise, must be estimated. We verified the conditions of our theory for the important special case of stratified sampling on an ordered index; many real designs incorporate this or similar methods. The asymptotic theory leads to excellent variance approximations in our simulations. The variance estimation method suggested by our theory combines analytic information computations, familiar from standard likelihood estimation, with design covariance matrix estimation, readily obtained from survey software.

Appendix A: Proof of Theorem 1

The continuity of ζ ensures the strong consistency of the sample quantiles (see Serfling [26], page 75), and further: $\mathrm{P}(\bigcap_{h=1}^{H-1}\{\omega \in \Omega | \lim_{\gamma\to\infty} Z_{\nu_\gamma(T_{\gamma,h})}(\omega) = \zeta(t_{\infty,h})\}) = 1$. Let $t^*_{\gamma,h}$ be a sequence defined for all $\gamma \in \mathbb{N}$, $h \in \{1,\dots,H\}$, such that $h < h' \Rightarrow 0 < t^*_{\gamma,h} \le t^*_{\gamma,h'} \le 1$, and such that $\forall h \in \{1,\dots,H\}$, $\lim_{\gamma\to\infty} t^*_{\gamma,h} = t_{\infty,h}$. Then conditionally on $Z_{\nu_\gamma(T_{\gamma,h-1})} = \zeta(t^*_{\gamma,h-1})$, $Z_{\nu_\gamma(T_{\gamma,h})} = \zeta(t^*_{\gamma,h})$, we have independence within the same stratum: $\forall h \in \{1,\dots,H-1\}$,

$$\begin{aligned}&\mathrm{P}^{(Z_k,Y_k)_{k\in r_{\gamma h}\circ\nu_\gamma(U_{\gamma,h})}|Z_{\nu_\gamma(T_{\gamma,h-1})}=\zeta(t^*_{\gamma,h-1}),Z_{\nu_\gamma(T_{\gamma,h})}=\zeta(t^*_{\gamma,h})}\\&\quad=\left(\mathrm{P}^{(Z,Y)|Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}\right)^{\otimes N_{\gamma,h}-1}\otimes\mathrm{P}^{(Z,Y)|Z=\zeta(t^*_{\gamma,h})},\end{aligned}\tag{1}$$

where $\otimes$ denotes product measure and $P^{\otimes k}$ denotes the product measure of k independent and identically distributed random variables, and

$$\mathrm{P}^{(Z_k,Y_k)_{k\in r_{\gamma H}\circ v_\gamma(U_{\gamma,H})}|Z_{v_\gamma(T_{\gamma,H-1})}=\zeta(t^*_{\gamma,H-1})} = \big(\mathrm{P}^{(Z,Y)|Z\in(\zeta(t^*_{\gamma,H-1}),\zeta(t^*_{\gamma,h})]}\big)^{\otimes N_{\gamma,H}}, \tag{2}$$

where for $h \in \{1,\dots,H-1\}$, $r_{\gamma h}$ is a random permutation of the ordered set $v_\gamma(U_{\gamma,h})$ such that $r_{\gamma h}\circ v_\gamma(T_{\gamma,h}) = T_{\gamma,h}$, and $r_{\gamma H}$ is a random permutation of the ordered set $v_\gamma(U_{\gamma,H})$. If we consider the distribution of the sample responses on hth stratum, we have: $\forall h \in \{1,\dots,H-1\}$,

$$\begin{aligned}
&\mathrm{P}^{(Z_k,Y_k)_{k\in r_{\gamma h}\circ v_\gamma(U_{\gamma,h}),J_{\gamma,k}=1}|Z_{v_\gamma(T_{\gamma,h-1})}=\zeta(t^*_{\gamma,h-1}),Z_{v_\gamma(T_{\gamma,h})}=\zeta(t^*_{\gamma,h})}\\
&\quad=\big(\mathrm{P}^{(Z,Y)|Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}\big)^{\otimes n_{\gamma,h}-1}\\
&\qquad\otimes\big(\tau_{\gamma h}\mathrm{P}^{(Z,Y)|Z=\zeta(t^*_{\gamma,h})}+(1-\tau_{\gamma h})\big(\mathrm{P}^{(Z,Y)|Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}\big)\big)
\end{aligned} \tag{3}$$

and

$$\mathrm{P}^{(Z_k,Y_k)_{k\in r_{\gamma H}\circ v_\gamma(U_{\gamma,H}),J_{\gamma,k}=1}|Z_{v_\gamma(T_{\gamma,H-1})}=\zeta(t^*_{\gamma,H-1})} = \big(\mathrm{P}^{(Z,Y)|Z\in(\zeta(t^*_{\gamma,H-1}),\infty]}\big)^{\otimes n_{\gamma,H}}. \tag{4}$$

Equation (3) implies that $\forall h \in \{1,\dots,H-1\}$,

$$\begin{aligned}
&\mathrm{P}^{(g(Y,Z,\pi_{\gamma H}))_{k\in v_\gamma(U_{\gamma,h}),J_{\gamma,k}=1}|Z_{v_\gamma(T_{\gamma,h-1})}=\zeta(t^*_{\gamma,h-1}),Z_{v_\gamma(T_{\gamma,h})}=\zeta(t^*_{\gamma,h})}\\
&\quad=\tau_{\gamma h}\mathrm{P}^{g(Y,Z,\pi_{\gamma h})|Z_{v_\gamma(T_{\gamma,h})}=\zeta(t^*_{\gamma,h})}\otimes\big(\mathrm{P}^{g(Y,Z,\pi_{\gamma h})|Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}\big)^{\otimes n_{\gamma,h}-1}\\
&\qquad+(1-\tau_{\gamma h})\big(\mathrm{P}^{g(Y,Z,\tau_{\gamma h})|Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}\big)^{\otimes n_{\gamma,h}},
\end{aligned} \tag{5}$$

and equation (4) implies that

$$\mathrm{P}^{(g(Y,Z,\tau_{\gamma H}))_{k\in v_\gamma(U_{\gamma,H}),J_{\gamma,k}=1}|Z_{v_\gamma(T_{\gamma,H-1})}=\zeta(t^*_{\gamma,H-1})} = \big(\mathrm{P}^{g(Y,Z,\tau_{\gamma H})|Z\in(\zeta(t^*_{\gamma,h-1}),\infty]}\big)^{\otimes n_{\gamma,H}}. \tag{6}$$

We will show that $\forall h \in \{0,\dots,H\}$,

$$\mathrm{P}^{\sqrt{n_{\gamma h}}(n_{\gamma h}^{-1}S_{\gamma,h}-E_{\gamma h})|Z_{v_\gamma(T_{\gamma,h-1})}=\zeta(t^*_{\gamma,h-1}),Z_{v_\gamma(T_{\gamma,h})}=\zeta(t^*_{\gamma,h})} \xrightarrow[\gamma\to\infty]{\mathcal{L}} \mathcal{N}(0,V_{\infty,h}). \tag{7}$$

Using (5) we calculate, for $h \in \{0,\dots,H-1\}$:

$$\begin{aligned}
&\mathrm{Var}\big[S_{\gamma h}|Z_{v_\gamma(T_{\gamma,h-1})}=\zeta\big(t^*_{\gamma,h-1}\big),Z_{v_\gamma(T_{\gamma,h})}=\zeta\big(t^*_{\gamma,h}\big)\big]\\
&\quad=\big(n^*_\gamma-1\big)\,\mathrm{Var}\big[g(Y,Z,\tau_{\gamma h})|Z\in\big(\zeta\big(t^*_{\gamma,h-1}\big),\zeta\big(t^*_{\gamma,h}\big)\big]\big]\\
&\qquad+\tau_{\gamma h}\,\mathrm{E}\big[\big(gg^{\mathrm{T}}\big)(Y,Z,\tau_{\gamma h})|Z=\zeta\big(t^*_{\gamma,h}\big)\big]\\
&\qquad+(1-\tau_{\gamma h})\,\mathrm{E}\big[\big(gg^{\mathrm{T}}\big)(Y,Z,\tau_{\gamma h})|Z\in\big(\zeta\big(t^*_{\gamma,h-1}\big),\zeta\big(t^*_{\gamma,h}\big)\big]\big]\\
&\qquad-\big(\tau_{\gamma h}\,\mathrm{E}\big[g(Y,Z,\tau_{\gamma h})|Z=\zeta\big(t^*_{\gamma,h}\big)\big]\\
&\qquad+(1-\tau_{\gamma h})\,\mathrm{E}\big[g(Y,Z,\tau_{\gamma h})|Z\in\big(\zeta\big(t^*_{\gamma,h-1}\big),\zeta\big(t^*_{\gamma,h}\big)\big]\big]\big)
\end{aligned} \tag{8}$$

$$\times \left(\tau_{\gamma h}\,\mathrm{E}\big[g(Y,Z,\tau_{\gamma h})|Z=\zeta\left(t^*_{\gamma,h}\right)\big]\right.$$
$$\left.+\,(1-\tau_{\gamma h})\,\mathrm{E}\big[g(Y,Z,\tau_{\gamma h})|Z\in\left(\zeta\left(t^*_{\gamma,h-1}\right),\zeta\left(t^*_{\gamma,h}\right)\right]\big]\right)^{\mathrm{T}}$$

and

$$\mathrm{Var}\big[S_{\gamma H}|Z_{v_\gamma(T_{\gamma,H-1})}=\zeta\left(t^*_{\gamma,H-1}\right)\big]=n^*_\gamma\,\mathrm{Var}\big[g(Y,Z,\tau_{\gamma h})|Z\in\left(\zeta\left(t^*_{\gamma,h-1}\right),\infty\right]\big]. \tag{9}$$

In addition,

$$\mathrm{E}\big[g(Y,Z,\tau_{\gamma h})|Z\in\left(\zeta\left(t^*_{\gamma,h-1}\right),\zeta\left(t^*_{\gamma,h}\right)\right]\big]=\frac{\int g(y,z,\tau_{\gamma h})\mathbb{1}_{(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}(z)\,\mathrm{d}\,\mathrm{P}^{Y,Z}(y,z)}{\mathrm{P}(Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})])}$$

and

$$\mathrm{E}\big[\left(gg^{\mathrm{T}}\right)(Y,Z,\tau_{\gamma h})|Z\in\left(\zeta\left(t^*_{\gamma,h-1}\right),\zeta\left(t^*_{\gamma,h}\right)\right]\big]$$
$$=\frac{\int(gg^{\mathrm{T}})(y,z,\tau_{\gamma h})\mathbb{1}_{(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}(z)\,\mathrm{d}\,\mathrm{P}^{Y,Z}(y,z)}{\mathrm{P}(Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})])}.$$

Because $\forall(y,z)\in Y(\Omega)\times Z(\Omega)$,

$$\lim_{\gamma\to\infty} g(y,z,\tau_{\gamma h})\mathbb{1}_{(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}(z)=g(y,z,\tau_{\infty,h})\mathbb{1}_{(\zeta(t_{\infty,h-1}),\zeta(t_{\infty,h})]}(z),$$

and $\|g(y,z,\tau_{\gamma h})\mathbb{1}_{(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}(z)\|\le G(y,z)$, we conclude by the Lebesgue dominated convergence theorem that

$$\lim_{\gamma\to\infty}\mathrm{E}\big[g(Y,Z,\tau_{\gamma h})|Z\in\left(\zeta\left(t^*_{\gamma,h-1}\right),\zeta\left(t^*_{\gamma,h}\right)\right]\big]$$
$$=\mathrm{E}\big[g(Y,Z,\tau_{\infty,h})|Z\in\left(\zeta(t_{\infty,h-1}),\zeta(t_{\infty,h})\right]\big],$$
$$\lim_{\gamma\to\infty}\mathrm{Var}\big[g(Y,Z,\tau_{\gamma h})|Z\in\left(\zeta\left(t^*_{\gamma,h-1}\right),\zeta\left(t^*_{\gamma,h}\right)\right]\big]$$
$$=\mathrm{Var}\big[g(Y,Z,\tau_{\infty,h})|Z\in\left(\zeta(t_{\infty,h-1}),\zeta(t_{\infty,h})\right]\big].$$

Also, as $\forall y\in Y(\Omega)\times Z(\Omega)$,

$$\lim_{\gamma\to\infty} g\left(y,\zeta\left(t^*_{\gamma,h}\right),\tau_{\gamma h}\right)\frac{\mathrm{d}\,\mathrm{P}^{Y|Z=\zeta(t^*_{\gamma,h})}}{\mathrm{d}\lambda}(y)=g\left(y,\zeta\left(t^*_{\infty,h}\right),\tau_{\gamma h}\right)\frac{\mathrm{d}\,\mathrm{P}^{Y|Z=\zeta(t^*_{\infty,h})}}{\mathrm{d}\lambda}(y)$$

and for γ large enough, $\|g(y,\zeta(t^*_{\gamma,h}),\tau_{\gamma h})\frac{\mathrm{d}\,\mathrm{P}^{Y|Z=\zeta(t^*_{\gamma,h})}}{\mathrm{d}\lambda}(y)\|\le G(y,z)M(y)$, we conclude by the Lebesgue dominated convergence theorem that:

$$\lim_{\gamma\to\infty}\mathrm{E}\big[g(Y,Z,\tau_{\gamma h})|Z=\zeta\left(t^*_{\gamma,h}\right)\big]=\mathrm{E}\big[g(Y,Z,\tau_{\infty,h})|Z=\zeta(t_{\infty,h})\big]<\infty, \tag{10}$$

$$\lim_{\gamma\to\infty}\mathrm{Var}\big[g(Y,Z,\tau_{\gamma h})|Z=\zeta\left(t^*_{\gamma,h}\right)\big]=\mathrm{Var}\big[g(Y,Z,\tau_{\infty,h})|Z=\zeta(t_{\infty,h})\big]<\infty. \tag{11}$$

Equations (8), (9) and the preceding imply that

$$\lim_{\gamma\to\infty}(n_{\gamma h})^{-1}\operatorname{Var}\big[S_{\gamma h}|Z_{v_\gamma(T_{\gamma,h-1})}=\zeta\big(t^*_{\gamma,h-1}\big),\,Z_{v_\gamma(T_{\gamma,h})}=\zeta\big(t^*_{\gamma,h}\big)\big]$$
$$=\operatorname{Var}\big[g(Y,Z,\tau_{\infty,h})|Z\in\big(\zeta(t_{\infty,h-1}),\zeta(t_{\infty,h})\big)\big]\big]$$

and

$$\lim_{\gamma\to\infty}(n_{\gamma H})^{-1}\operatorname{Var}\big[S_{\gamma H}|Z_{v_\gamma(T_{\gamma,H-1})}=\zeta\big(t^*_{\gamma,h-1}\big)\big]$$
$$=\operatorname{Var}\big[g(Y,Z,\tau_{\infty,h})|Z\in\big(\zeta(t_{\infty,H-1}),\infty\big)\big]\big].$$

For $\gamma\in\mathbb{N}$, $h\in\{1,\dots,H-1\}$ introduce the random variables $X^*_{\gamma,h,1}\cdots X^*_{\gamma,h,n_{\gamma h}}$ that satisfy

$$\mathrm{P}^{X^*_{\gamma,h,1}\cdots X^*_{\gamma,h,n_{\gamma h}}}=\big(\mathrm{P}^{g(Z,Y,\tau_{\gamma h})|Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}\big)^{\otimes n_{\gamma,h}-1}$$
$$\otimes\big(\tau_{\gamma h}\mathrm{P}^{g(Z,Y,\tau_{\gamma h})|Z=\zeta(t^*_{\gamma,h})}$$
$$+(1-\tau_{\gamma h})\big(\mathrm{P}^{g(Z,Y,\tau_{\gamma h})|Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h})]}\big)\big),$$

and the random variables $X^*_{\gamma,H,1}\cdots X^*_{\gamma,H,n_{\gamma H}}$ that satisfy

$$\mathrm{P}^{X^*_{\gamma,H,1}\cdots X^*_{\gamma,H,n_{\gamma H}}}=\big(\mathrm{P}^{g(Z,Y,\tau_{\gamma H})|Z>\zeta(t^*_{\gamma,H-1})}\big)^{\otimes n_{\gamma H}}.$$

For $\alpha\in Y(\Omega)$, $h\in\{1,\dots,H\}$, $\mathrm{P}^{\alpha^{\mathrm{T}}S_{\gamma h}}=\mathrm{P}^{\sum_{k=1}^{n_{\gamma h}}\alpha^{\mathrm{T}}X^*_{\gamma,h,k}}$. For $\alpha\in Y(\Omega)\setminus\{0\}$, $\gamma\in\mathbb{N}$, $\varepsilon\in(0,\infty]$, we define

$$A_{\gamma,h,\varepsilon,\alpha}=\big(\alpha^{\mathrm{T}}\operatorname{Var}[S_{\gamma h}]\alpha\big)^{-1}\sum_{k=1}^{n_{\gamma h}}\mathrm{E}\big[\big|\alpha^{\mathrm{T}}\big(X^*_{\gamma,h,k}-\mathrm{E}\big[X^*_{\gamma,h,k}\big]\big)\big|^2$$
$$\times\mathbb{1}_{(\varepsilon\sqrt{\alpha^{\mathrm{T}}\operatorname{Var}[S_\gamma]\alpha},\infty]}\big(\big|\alpha^{\mathrm{T}}\big(X^*_{\gamma,h,k}-\mathrm{E}\big[X^*_{\gamma,h,k}\big]\big)\big|\big)\big].$$

Let $\alpha\in Y(\Omega)\setminus\{0\}$. To prove the asymptotic normality of $\alpha^{\mathrm{T}}S_{\gamma h}$, we will show that the Lindeberg condition

$$\forall\varepsilon\in(0,\infty),\qquad\lim_{\gamma\to\infty}A_{\gamma,h,\varepsilon,\alpha}=0 \tag{12}$$

is satisfied. Let $\varepsilon\in(0,\infty)$, $h\in\{1,\dots,H\}$. Then as $\gamma\to\infty$,

$$A_{\gamma,h,\varepsilon,\alpha}\sim\frac{(n_{\gamma h}-1)}{n_{\gamma h}\alpha^{\mathrm{T}}V_{\infty,h}\alpha}\mathrm{E}\big[\big|\alpha^{\mathrm{T}}\big(g(Y,Z,\tau_{\gamma h})-\mathrm{E}\big[X^*_{\gamma,h,k}\big]\big)\big|^2$$
$$\times\mathbb{1}_{(\varepsilon\sqrt{\alpha^{\mathrm{T}}\operatorname{Var}[S_\gamma]\alpha},\infty]}\big(\alpha^{\mathrm{T}}\big(g(Y,Z,\tau_{\gamma h})-\mathrm{E}\big[X^*_{\gamma hk}\big]\big)\big)\big]. \tag{13}$$

In addition,

$$\begin{aligned}
&\int \left|\alpha^{\mathrm{T}}\left(g(Y,Z,\tau_{\gamma h})-\mathrm{E}\left[X^*_{\gamma h1}\right]\right)\right|^2 \mathbb{1}_{(\varepsilon\sqrt{\alpha^{\mathrm{T}}\operatorname{Var}[S_\gamma]\alpha},\infty]}\left(\alpha^{\mathrm{T}}\left(g(Y,Z,\tau_{\gamma h})-\mathrm{E}\left[X^*_{\gamma h1}\right]\right)\right)\mathrm{d}\,\mathrm{P}^{Y,Z}\\
&\quad\le \int \|\alpha\|^2\left(\left\|G(Y,Z)\right\|+\left\|\mathrm{E}\left[X^*_{\gamma h1}\right]\right\|\right)^2\\
&\qquad\times \mathbb{1}_{(\varepsilon\sqrt{\alpha^{\mathrm{T}}\operatorname{Var}[S_\gamma]\alpha},\infty]}\left(\|\alpha\|^2\left(\left\|G(Y,Z)\right\|+\left\|\mathrm{E}\left[X^*_{\gamma h1}\right]\right\|\right)\right)\mathrm{d}\,\mathrm{P}^{Y,Z}\\
&\quad\le \int \|\alpha\|^2\left(\left\|G(Y,Z)\right\|+\frac{\mathrm{E}[G(Y,Z)]}{\min\{t^*_{\gamma h}-t^*_{\gamma,h-1}|\gamma\in\mathbb{N}\}}\right)^2\\
&\qquad\times \mathbb{1}_{(\varepsilon\sqrt{\alpha^{\mathrm{T}}\operatorname{Var}[S_\gamma]\alpha},\infty]}\left(\|\alpha\|\left(\left\|G(Y,Z)\right\|+\frac{\mathrm{E}[G(Y,Z)]}{\min\{t^*_{\gamma h}-t^*_{\gamma,h-1}|\gamma\in\mathbb{N}\}}\right)\right)\mathrm{d}\,\mathrm{P}^{Y,Z}
\end{aligned}$$

because, for $h\in\{1,\dots,H\}$, with the convention $t^*_{\gamma H}=1$,

$$\begin{aligned}
\left\|\mathrm{E}\left[X^*_{\gamma h1}\right]\right\| &\le \frac{\int G(Y,Z)\,\mathrm{d}\,\mathrm{P}^{Y,Z}}{\mathrm{P}(Z\in(\zeta(t^*_{\gamma,h-1}),\zeta(t^*_{\gamma,h}))])}\le\frac{\int G(Y,Z)\,\mathrm{d}\,\mathrm{P}^{Y,Z}}{t^*_{\gamma h}-t^*_{\gamma,h-1}}\\
&\le\frac{\int G(Y,Z)\,\mathrm{d}\,\mathrm{P}^{Y,Z}}{\min\{t^*_{\gamma h}-t^*_{\gamma,h-1}|\gamma\in\mathbb{N}\}},
\end{aligned}$$

because $\mathrm{P}^{(T^*_{\gamma h},T^*_{\gamma,h-1})_{\gamma\in\mathbb{N}}}$-a.s. $((t^*_{\gamma h}-t^*_{\gamma,h-1})_{\gamma\in\mathbb{N}})$, $\lim_{\gamma\to\infty}t^*_{\gamma h}-t^*_{\gamma,h-1}=t^*_{\infty,h}-t^*_{\infty,h-1}$ and because $\min\{t^*_{\gamma h}-t^*_{\gamma,h-1}|\gamma\in\mathbb{N}\}>0$. As $\lim_{\gamma\to\infty}\varepsilon\sqrt{\alpha^{\mathrm{T}}\operatorname{Var}[S_\gamma]\alpha}=\infty$, and $\mathrm{E}[G(Y,Z)^2]<\infty$, we conclude that

$$\begin{aligned}
&\lim_{\gamma\to\infty}\int \left|\alpha^{\mathrm{T}}\left(g(Y,Z,\tau_{\gamma h})-\mathrm{E}\left[X^*_{\gamma,h,1}\right]\right)\right|^2\mathbb{1}_{(\varepsilon\sqrt{\alpha^{\mathrm{T}}\operatorname{Var}[S_\gamma]\alpha},\infty]}\left(\alpha^{\mathrm{T}}\left(g(Y,Z,\tau_{\gamma h})-\mathrm{E}\left[X^*_{\gamma,h,1}\right]\right)\right)\\
&\qquad=0,
\end{aligned}$$

which implies via (13) that the Lindeberg condition (12) is satisfied. We apply the Lindeberg–Feller theorem (see Serfling [26], Theorem page 31), and conclude by the asymptotic normality of $\alpha^{\mathrm{T}}S_{\gamma h}$ conditionally on $(T_{\gamma h})$ $\forall\alpha\in Y(\Omega)$ (which terminates the proof of (7)). Then, we remark that conditionally on $Z_{\nu_\gamma(T_{\gamma,h-1})}=\zeta(t^*_{\gamma,h-1})$, $Z_{\nu_\gamma(T_{\gamma,h})}=\zeta(t^*_{\gamma,h})$, we have independence between strata:

$$\begin{aligned}
h\neq h'\Rightarrow \mathrm{P}^{\binom{S_{\gamma,h}}{S_{\gamma h'}}\left|\begin{cases} Z_{\nu_\gamma(T_{\gamma,h-1})}=\zeta(t^*_{\gamma,h-1}),\, Z_{\nu_\gamma(T_{\gamma,h})}=\zeta(t^*_{\gamma,h}),\\ Z_{\nu_\gamma(T_{\gamma h'-1})}=\zeta(t^*_{\gamma h'-1}),\, Z_{\nu_\gamma(T_{\gamma h'})}=\zeta(t^*_{\gamma h'})\end{cases}\right.}&\\
=\mathrm{P}^{(S_{\gamma,h})|Z_{\nu_\gamma(T_{\gamma,h-1})}=\zeta(t^*_{\gamma,h-1}),Z_{\nu_\gamma(T_{\gamma,h})}=\zeta(t^*_{\gamma,h})}\otimes\mathrm{P}^{(S_{\gamma h'})|Z_{\nu_\gamma(T_{\gamma h'-1})}=\zeta(t^*_{\gamma h'-1}),Z_{\nu_\gamma(T_{\gamma h'})}=\zeta(t^*_{\gamma h'})}&\\
\Rightarrow S_{\gamma,h}\text{ and }S_{\gamma h'}\text{ are independent conditionally on }T_{\gamma,h},T_{\gamma,h-1},T_{\gamma,h'}T_{\gamma,h'-1}.&
\end{aligned}$$

Together with equation (7), this implies that

$$\begin{aligned} &\mathrm{P}^{(T_{\gamma h})_{\gamma\in\mathbb{N},h\in\{1,\dots,H-1\}}}\text{-a.s. } \left(\left(t^*_{\gamma h}\right)_{\gamma\in\mathbb{N},h\in\{1,\dots,H-1\}}\right), \\ &\mathrm{P}^{\sqrt{n_\gamma}(n_\gamma^{-1}S_\gamma-E_\infty)|T_{\gamma,1}=t^*_{\gamma,1},\dots,T_{\gamma,H-1}=t^*_{\gamma,H-1}} \xrightarrow[\gamma\to\infty]{\mathcal{L}} \mathcal{N}(0,V_\infty). \end{aligned} \tag{14}$$

The almost sure asymptotic normality implies the global asymptotic normality. For $\gamma\in\mathbb{N}$, $x\in\mathbb{R}$, $\alpha\in Y(\Omega)$ we define:

$$h_{\gamma,\alpha,x}: t^* \mapsto \mathrm{E}\left[\exp\left(\mathrm{i}x\frac{\alpha^{\mathrm{T}}(S_\gamma-E_\infty)}{\sqrt{n_\gamma}\sqrt{\alpha^{\mathrm{T}}V_\infty\alpha}}\right)|T_{\gamma,1}=t^*_{\gamma,1},\dots,T_{\gamma,H-1}=t^*_{\gamma,H-1}\right].$$

Then equation (14) implies that $\mathrm{P}^{(T_{\gamma h})_{\gamma\in\mathbb{N},h\in\{1,\dots,H-1\}}}$-a.s. (t^*), $\lim_{\gamma\to\infty}h_{\gamma,\alpha,x,}(t^*)=\exp(\mathrm{i}x-t^2/2)$. Besides, $\mathrm{P}^{(T_{\gamma h})_{\gamma\in\mathbb{N},h\in\{1,\dots,H-1\}}}$-a.s (t^*), $\forall\gamma\in\mathbb{N}$ $|h_{\gamma,\alpha,x}(t^*)|\le 1$. We apply the Lebesgue dominated convergence theorem:

$$\begin{aligned} &\lim_{\gamma\to\infty}\mathrm{E}\left[\exp\left(\mathrm{i}x\left(\sqrt{n_\gamma\alpha^{\mathrm{T}}V_\infty\alpha}\right)^{-1}\alpha^{\mathrm{T}}(S_\gamma-E_\infty)\right)\right] \\ &\quad=\lim_{\gamma\to\infty}\mathrm{E}\left[\mathrm{E}\left[\exp\left(\mathrm{i}x\left(\sqrt{n_\gamma\alpha^{\mathrm{T}}V_\infty\alpha}\right)^{-1}\alpha^{\mathrm{T}}(S_\gamma-E_\infty)\right)|T_{\gamma,1},\dots,T_{\gamma,H-1}\right]\right] \\ &\quad=\mathrm{E}\left[\lim_{\gamma\to\infty}\mathrm{E}\left[\exp\left(\mathrm{i}x\left(\sqrt{n_\gamma\alpha^{\mathrm{T}}V_\infty\alpha}\right)^{-1}\alpha^{\mathrm{T}}(S_\gamma-E_\infty)\right)|T_{\gamma,1},\dots,T_{\gamma,H-1}\right]\right] \\ &\quad=\int\exp\left(\mathrm{i}x-x^2/2\right)\mathrm{d}\,\mathrm{P}^{(T_{\gamma h})_{\gamma\in\mathbb{N},h\in\{1,\dots,H-1\}}}=\exp\left(\mathrm{i}x-x^2/2\right). \end{aligned}$$

The convergence of the characteristic function then implies the convergence to the normal distribution, which ends the demonstration of the theorem. □

Appendix B: Proof of Theorem 2

By assumption (A1.a),

$$\frac{\partial^2}{\partial\theta\,\partial\xi}\bar{\mathscr{L}}(\theta_0,\xi_0)+\mathscr{I}_{\gamma,1,2}=o_{\mathrm{P}_{\theta_0,\xi_0}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1),$$

and combining with (A1.b),

$$\sqrt{n_\gamma}\left(\frac{\partial^2}{\partial\theta\,\partial\xi}\bar{\mathscr{L}}(\theta_0,\xi_0)\right)(\widehat{\xi}_\gamma-\xi_0)=-\sqrt{n_\gamma}\mathscr{I}_{\gamma,1,2}(\widehat{\xi}_\gamma-\xi_0)+o_{\mathrm{P}_{\theta_0,\xi_0}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1). \tag{15}$$

By assumption (A1.a),

$$\frac{\partial^2}{\partial\theta^2}\bar{\mathscr{L}}(\theta_0,\widehat{\xi}_\gamma)+\mathscr{I}_{\gamma,1,1}=o_{\mathrm{P}_{\theta_0,\xi_0}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1). \tag{16}$$

Combining with (A1.e), we have

$$-\sqrt{n_\gamma}\frac{\partial}{\partial\theta}\bar{\mathscr{L}}(\theta_0,\widehat{\xi}_\gamma)=\sqrt{n_\gamma}\big(-\mathscr{I}_{\gamma,1,1}+o_{\mathrm{P}_{\theta_0,\xi_0}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1)\big)(\widehat{\theta}_\gamma-\theta_0)+o_{\mathrm{P}_{\theta_0,\xi_0}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1). \tag{17}$$

Combining (A1.f) and (15), we have

$$\sqrt{n_\gamma}\frac{\partial}{\partial\theta}\bar{\mathscr{L}}(\theta_0,\widehat{\xi}_\gamma)=\sqrt{n_\gamma}\frac{\partial}{\partial\theta}\bar{\mathscr{L}}(\theta_0,\xi_0)-\sqrt{n_\gamma}\mathscr{I}_{\gamma,1,2}(\widehat{\xi}_\gamma-\xi_0)+o_{\mathrm{P}_{\theta_0,\xi_0}^{\cdot|\mathcal{X}_\gamma=\mathbf{x}_\gamma}}(1). \tag{18}$$

Applying what precedes ((17) and (18)) and (A1.d), we see that $\sqrt{n_\gamma}(\widehat{\theta}_\gamma-\theta_0)$ is asymptotically equivalent to $\mathscr{I}_{\gamma,1,1}^{-1}(\sqrt{n_\gamma}\bar{\mathscr{L}}(\theta_0,\xi_0)-\sqrt{n_\gamma}\mathscr{I}_{\gamma,1,2}(\widehat{\xi}_\gamma-\xi_0))$. Thus $\sqrt{n_\gamma}V_\gamma^{-1/2}(\widehat{\theta}_\gamma-\theta_0)$ converges in distribution to $\mathscr{N}(0,I)$ conditionally on $\mathcal{X}_\gamma=\mathbf{x}_\gamma$, with $V_\gamma=\mathscr{I}_{\gamma,1,1}^{-1}(\Sigma_{\gamma,1,1}+\mathscr{I}_{\gamma,1,2}\Sigma_{\gamma,2,2}\mathscr{I}_{\gamma,1,2}-\mathscr{I}_{\gamma,1,2}\Sigma_{\gamma,1,2}^{\mathrm{T}}-\Sigma_{\gamma,1,2}\mathscr{I}_{\gamma,1,2}^{\mathrm{T}})\mathscr{I}_{\gamma,1,1}^{-1}$, establishing Theorem 2.

Appendix C: Proofs for stratified sampling

C.1. Proof of Result 1

Proof. We first show that A0 is satisfied. As there are no covariates, we can choose $d=1$, h_γ as any constant function, and $m_{\gamma,\theta,\xi}$ defined by $m_{\gamma,\theta,\xi}(y,x_1,h_\gamma(\mathbf{x}_\gamma))=\mathrm{f}_{J_{\gamma,a}|Y_1=y}(1)$, and assumption (A0.b) holds.

To show that A0, is satisfied, we compute:

$$\begin{aligned}&\lim_{\gamma\to\infty}\mathrm{P}_{\theta,\xi}\left(J_{\gamma,k}=1|Y_k=y,Z_k=z\right)\\&\quad=\lim_{\gamma\to\infty}\mathrm{P}_{\theta,\xi}\left(J_{\gamma,k}=1|Z_k=z\right)\\&\quad=\sum_{h=1}^{H}\tau_{\infty,h}\mathbb{1}_{[\Phi^{-1}(t_{\infty,h-1}),\Phi^{-1}(t_{\infty,h})]}\left(\frac{Z_k-\theta\xi}{\sqrt{\xi^2+\sigma_\eta^2}}\right).\end{aligned}$$

We deduce from the preceding that $m_\infty(y;\theta,\xi)$ is defined:

$$\begin{aligned}m_{\infty,\theta,\xi}(y)&=\sum_{h=1}^{H}\tau_{\infty,h}\,\mathrm{P}_{\theta,\xi}\left(\Phi^{-1}(t_{\infty,h})<\frac{\xi(y-\theta)+\varepsilon}{\sqrt{\xi^2+\sigma_\eta^2}}<\Phi^{-1}(t_{\infty,h-1})\right)\\&=\tau_H+\sum_{h=1}^{H-1}(\tau_{\infty,h}-\tau_{\infty,h+1})\Phi\left(\sqrt{\frac{\xi}{\sigma_\eta}^2+1}\big(\Phi^{-1}(t_{\infty,h})\big)+\frac{\xi}{\sigma_\eta}(\theta-y)\right).\end{aligned}$$

By assumption, $\int m_{\infty,\theta,\xi} f_\theta \, \mathrm{d}\lambda = \lim_{\gamma\to\infty} {N_\gamma}^{-1} n_\gamma > 0$, so $\rho_{\infty;\theta,\xi}$ is defined:

$$\rho_\infty(y;\theta,\xi) = \frac{\tau_H + \sum_{h=1}^{H-1}(\tau_{\infty,h} - \tau_{\infty,h+1})\Phi\left(\sqrt{\frac{\xi}{\sigma_\eta}^2 + 1}(\Phi^{-1}(t_{\infty,h})) + \frac{\xi}{\sigma_\eta}(\theta - y)\right)}{\tau_H + \sum_{h=1}^{H-1}(\tau_{\infty,h} - \tau_{\infty,h+1})t_{\infty,h}}.$$

□

C.2. Proof of Result 2

The quantile function $\zeta(t)$ of Z is continuous on $(0,1)$ and $\mathrm{f}_{Y|Z=z}(y)$ is continuous in z $\forall y \in \mathbb{R}$. The function $\zeta(\cdot)$ depends on σ_η, θ, and ξ via $\zeta(t) = \sqrt{\xi^2 + \sigma_\eta^2}\Phi^{-1}(t) + \xi\theta$. Applying Theorem 1 with $g(y,z,\tau) = [\, y \times z/\tau, \quad y^2/\tau \,]^{\mathrm{T}}$, we obtain the asymptotic normality of the vector $\sqrt{n_\gamma} S_\gamma$:

$$\sqrt{n_\gamma}\left(\begin{bmatrix} n_\gamma^{-1}\sum_{k=1}^{N_\gamma} Y_k Z_k/\tau_{\gamma,k} J_{\gamma,k} \\ n_\gamma^{-1}\sum_{k=1}^{N_\gamma} Y_k^2/\tau_{\gamma,k} J_{\gamma,k} \end{bmatrix} - E_\infty\right) \xrightarrow[\gamma\to\infty]{\mathcal{L}} \mathcal{N}(0, V_\infty),$$

with $E_\infty = [\, \tau^{-1}\xi_0(\theta_0^2+1) \quad \tau^{-1}(\theta_0^2+1) \,]^{\mathrm{T}}$, and

$$V_\infty = \sum_{h=1}^{H}(t_{\infty,h} - t_{\infty,h-1})\tau_{\infty,h} \times \mathrm{Var}_{\theta_0,\xi_0}\left[\begin{bmatrix} YZ/\tau_{\infty,h} \\ Y^2/\tau_{\infty,h} \end{bmatrix} \middle| Z \in \left(\zeta(t_{\infty,h-1}), \zeta(t_{\infty,h})\right]\right].$$

Applying the Delta method (see van der Vaart [27], Theorem 3.1 page 26),

$$\sqrt{n_\gamma}(\widehat{\xi}_\gamma - \xi_0) \xrightarrow[\gamma\to\infty]{\mathcal{L}} \mathcal{N}\left(0, \begin{bmatrix} \tau\left(\theta_0^2+1\right)^{-1} \\ -\tau\xi_0\left(\theta_0^2+1\right)^{-1} \end{bmatrix}^{\mathrm{T}} V_\infty \begin{bmatrix} \tau\left(\theta_0^2+1\right)^{-1} \\ -\tau\xi_0\left(\theta_0^2+1\right)^{-1} \end{bmatrix}\right).$$

C.3. Proof of Result 3

Assumption (A1.a), (A1.e), (A1.f) are satisfied (see Bonnéry [1], page 115 for details). Assumption (A1.b) is a consequence of Result 2. We now show (A1.c) and (A1.d). First,

$$\mathscr{I}_{\gamma,1,1} < \tau^{-1} \max_{h=1}^{H}\{\tau_{\infty,h}\} \int_{Y(\Omega)} \left(\frac{\partial\Delta}{\partial\theta}(y,\theta_0,\xi_0)\right)^2 \mathrm{d}\,\mathrm{P}^Y(y) < \infty,$$

because

$$\begin{aligned}
&\left(\frac{\partial}{\partial\theta}\Delta\right)(Y,\theta_0,\xi_0)\\
&\quad=(y-\theta)+\frac{\sum_{h=1}^{H-1}(\tau_{\infty,h}-\tau_{\infty,h+1})\frac{\xi}{\sigma_\eta}f_0(\sqrt{(\frac{\xi}{\sigma_\eta})^2+1}\Phi^{-1}(t_{\infty,h})+\frac{\xi}{\sigma_\eta}(\theta-y))}{\tau_H+\sum_{h=1}^{H-1}(\tau_{\infty,h}-\tau_{\infty,h+1})\Phi(\sqrt{(\frac{\xi}{\sigma_\eta})^2+1}\Phi^{-1}(t_{\infty,h})+\frac{\xi}{\sigma_\eta}(\theta-y))}\\
&\quad\le(y-\theta)+\frac{\sum_{h=1}^{H-1}(\tau_{\infty,h}-\tau_{\infty,h+1})\frac{\xi}{\sigma_\eta}\frac{1}{\sqrt{2\pi}}}{\min\{\tau_{\infty,h}\}}.
\end{aligned}$$

We have

$$\begin{aligned}
&\left(\frac{\partial}{\partial\xi}\Delta\right)(Y,\theta_0,\xi_0)\\
&\quad=\left(\tau_H+\sum_{h=1}^{H-1}(\tau_{\infty,h}-\tau_{\infty,h+1})\Phi\left(\sqrt{\left(\frac{\xi}{\sigma_\eta}\right)^2+1}\Phi^{-1}(t_{\infty,h})+\frac{\xi}{\sigma_\eta}(\theta-y)\right)\right)^{-1}\\
&\qquad\times\sum_{h=1}^{H-1}\left((\tau_{\infty,h}-\tau_{\infty,h+1})\left(\frac{\xi/\sigma_\eta^2}{\sqrt{\frac{\xi^2}{\sigma_\eta^2}+1}}\Phi^{-1}(t_{\infty,h})+\frac{(\theta-y)}{\sigma_\eta}\right)\right.\\
&\qquad\left.\times f_0\left(\sqrt{\left(\frac{\xi}{\sigma_\eta}\right)^2+1}\Phi^{-1}(t_{\infty,h})+\frac{\xi}{\sigma_\eta}(\theta-y)\right)\right)\\
&\quad\le\frac{\sum_{h=1}^{H-1}(\tau_{\infty,h}-\tau_{\infty,h+1})(\frac{\xi/\sigma_\eta^2}{\sqrt{\frac{\xi^2}{\sigma_\eta^2}+1}}+\frac{(\theta-y)}{\sigma_\eta})\frac{1}{\sqrt{2\pi}}}{\min\{\tau_{\infty,h}\}}.
\end{aligned}$$

Finally, $|\frac{\partial\Delta}{\partial\theta}\frac{\partial\Delta}{\partial\xi}(y,\theta_0,\xi_0)|\rho_{\infty,\theta_0,\xi_0}$ can be bounded above by a function of the form $|ay^2+by+c|$ so $\mathrm{E}_{\theta_0,\xi_0}[|(\frac{\partial}{\partial\theta}\Delta)(\frac{\partial}{\partial\psi}\Delta)|]<\infty$ and $\mathscr{I}_{\gamma,1,2}$ is defined.

To show that (A1.g) is satisfied, we apply Theorem 1 with

$$g(y,z,\pi)=\begin{bmatrix}(\partial\Delta/\partial\theta)(y,\theta_0,\xi_0) & y\times z/\pi & y^2/\pi\end{bmatrix}^{\mathrm{T}}.$$

Then we obtain the asymptotic normality of the vector $\sqrt{n_\gamma}(n_\gamma^{-1}S_\gamma-E_\infty)$:

$$\sqrt{n_\gamma}\begin{bmatrix}n_\gamma^{-1}\sum_{k=1}^{N_\gamma}(\partial\Delta/\partial\theta)(Y_k,\theta_0,\xi_0)J_{\gamma,k}\\ n_\gamma^{-1}\sum_{k=1}^{N_\gamma}Y_kZ_k/\pi_{\gamma,k}J_{\gamma,k}\\ n_\gamma^{-1}\sum_{k=1}^{N_\gamma}Y_k^2/\pi_{\gamma,k}J_{\gamma,k}\end{bmatrix}\xrightarrow[\gamma\to\infty]{\mathcal{L}}\mathcal{N}(E_\infty,V_\infty),$$

with $E_\infty = [\,0 \quad \pi^{-1}\xi(\theta^2+1) \quad \pi^{-1}(\theta^2+1)\,]^{\mathrm{T}}$. By applying the Delta method (see van der Vaart [27], Theorem 3.1 page 26), we obtain that

$$\sqrt{n_\gamma}\begin{bmatrix}\left(\dfrac{\partial}{\partial\theta}\bar{\mathscr{L}}\right)(\theta_0,\xi_0)\\ \widehat{\xi}_\gamma - \xi_0\end{bmatrix} \xrightarrow[\gamma\to\infty]{\mathcal{L}} \mathscr{N}\left(0, \begin{bmatrix}1 & 0\\ 0 & \pi\left(\theta_0^2+1\right)^{-1}\\ 0 & -\pi\xi_0\left(\theta_0^2+1\right)^{-1}\end{bmatrix}^{\mathrm{T}} V_\infty \begin{bmatrix}1 & 0\\ 0 & \pi\left(\theta_0^2+1\right)^{-1}\\ 0 & -\pi\xi_0\left(\theta_0^2+1\right)^{-1}\end{bmatrix}\right).$$

Hence, assumptions (A1.a)–(A1.g) are satisfied, and so Result 3 follows from Theorem 2.

Acknowledgements

This research was supported in part by the US National Science Foundation (SES–0922142). Daniel Bonnéry has worked at the Ensai, Toulouse School of Economics, and JPSM during this research.

References

[1] Bonnéry, D. (2011). Asymptotic properties of the sample distribution under informative selection. Ph.D. thesis, Université de Rennes 1. Available at http://tel.archives-ouvertes.fr/tel-00658990.
[2] Bonnéry, D., Breidt, F.J. and Coquet, F. (2012). Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. *Bernoulli* **18** 1361–1385. MR2995800
[3] Breslow, N.E. and Cain, K.C. (1988). Logistic regression for two-stage case–control data. *Biometrika* **75** 11–20. MR0932812
[4] Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. *Sociological Methodology* **43** 178–219.
[5] Chambers, R.L., Steel, D.G., Wang, S. and Welsh, A. (2012). *Maximum Likelihood Estimation for Sample Surveys*. *Monographs on Statistics and Applied Probability* **125**. Boca Raton, FL: CRC Press. MR2963765
[6] Eideh, A. and Nathan, G. (2009). Two-stage informative cluster sampling-estimation and prediction with applications for small-area models. *J. Statist. Plann. Inference* **139** 3088–3101. MR2535185
[7] Eideh, A.A.H. and Nathan, G. (2006). The analysis of data from sampling surveys under informative sampling. *Acta Comment. Univ. Tartu. Math.* **10** 41–51. MR2309744
[8] Eideh, A.A.H. and Nathan, G. (2006). Fitting time series models for longitudinal survey data under informative sampling. *J. Statist. Plann. Inference* **136** 3052–3069. MR2256216
[9] Eideh, A.A.H. and Nathan, G. (2006). Corrigendum to: "Fitting time series models for longitudinal survey data under informative sampling" [*J. Statist. Plann. Inference* **136** (2006) 3052–3069]. *J. Statist. Plann. Inference* **137** 628. MR2297656
[10] Fuller, W.A. (2009). *Sampling Statistics* **560**. New York: Wiley.
[11] Ghosh, M. and Maiti, T. (2004). Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika* **91** 95–112. MR2050462

[12] Gong, G. and Samaniego, F.J. (1981). Pseudomaximum likelihood estimation: Theory and applications. *Ann. Statist.* **9** 861–869. MR0619289
[13] Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.* **47** 663–685. MR0053460
[14] Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys. Wiley Series in Probability and Mathematical Statistics*. New York: Wiley.
[15] Krieger, A.M. and Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology* **18** 225–239.
[16] Landsman, V. and Graubard, B.I. (2013). Efficient analysis of case–control studies with sample weights. *Stat. Med.* **32** 347–360. MR3041872
[17] Patil, G.P. and Rao, C.R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34** 179–189. MR0507202
[18] Pfeffermann, D. (2011). Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology* **37** 115–136.
[19] Pfeffermann, D., Krieger, A.M. and Rinott, Y. (1998). Parametric distributions of complex survey data under informative probability sampling. *Statist. Sinica* **8** 1087–1114. MR1666233
[20] Pfeffermann, D., Moura, F.A.D.S. and Silva, P.L.d.N. (2006). Multi-level modelling under informative sampling. *Biometrika* **93** 943–959. MR2285081
[21] Pfeffermann, D. and Sikov, A. (2011). Imputation and estimation under nonignorable nonresponse for household surveys with missing covariate information. *Journal of Official Statistics* **27** 181–209.
[22] Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhyā Ser. B* **61** 166–186. MR1720710
[23] Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *J. Amer. Statist. Assoc.* **102** 1427–1439. MR2412558
[24] Pfeffermann, D. and Sverchkov, M. (2009). Inference under informative sampling. In *Sample Surveys: Inference and Analysis* (D. Pfefermann, ed.). *Handbook of Statistics* **29B** 455–487. Elsevier/North-Holland, Amsterdam.
[25] Pfeffermann, D. and Sverchkov, M.Y. (2003). Fitting generalized linear models under informative sampling. In *Analysis of Survey Data* (*Southampton*, 1999). *Wiley Ser. Surv. Methodol.* 175–195. Wiley, Chichester. MR1978851
[26] Serfling, R.J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley. MR0595165
[27] van der Vaart, A.W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge: Cambridge Univ. Press. MR1652247
[28] Yuan, K. and Jennrich, R.I. (2000). Estimating equations with nuisance parameters: Theory and applications. *Ann. Inst. Statist. Math.* **52** 343–350. MR1763567

Received August 2015 and revised November 2015