Augmented two-step estimating equations with nuisance functionals and complex survey data

Puying Zhao
Department of Statistics, Yunnan University

(Joint work with Changbao Wu)

December 13, 2023

Design-based Inference for Surveys

- Survey population: $\mathbf{U} = \{1, 2, \ldots, N\}$; $\mathbf{U}$ is treated as fixed
- Associated with each unit $i \in \mathbf{U}$, survey variables Z_i are non-random;
- $\mathcal{S}$: the set of sampled units selected from the finite population $\mathbf{U}$ by a probability sampling design.
- n: the realized sample size which could be a random number under certain sampling designs.
- $\pi_i = \Pr(i \in \mathcal{S})$ and $\pi_{ij} = \Pr(i, j \in \mathcal{S})$: the first and second order inclusion probabilities
- Design-based inference: Frequentist interpretation with respect to the probability sampling design for the given finite population.

Census Estimating Equations

- Let $\theta \in \Theta \subset \mathcal{R}^p$ be a parameter of interest and $\varphi \in \Psi$ be some nuisance function
- The true parameter $\theta_N \in \Theta$ is uniquely determined by the system of equations

$$U_N(\theta_N, \varphi_N) = \frac{1}{N} \sum_{i=1}^{N} g(Z_i, \theta_N, \varphi_N) = 0,$$

where $\varphi_N = \varphi_N(\cdot, \theta_N) \in \Psi$,

- $g = (g_1, \cdots, g_r)^\top$ is a vector of q functions with $r \geq p$;
- $g(Z, \theta, \varphi)$ is non-smooth in both θ and φ.

Literature Reivew

- Smooth estimating equation inference without nuisance parameter; i.e., Binder (1983); Binder and Patak (1994) and Chen and Kim (2014);
- Smooth estimating equation inference with nuisance parameter; i.e., Oguz-Alper and Berger (2016);
- Nonsmooth estimating equation inference with/without nuisance parameter, in which the nuisance parameter should be estimated at a $\sqrt{n}$ rate; i.e., Wang and Opsomer (2011).
- Design-based two-step semiparametric GEE inference, in which the convergence rates for the "plug-in" estimators are slower than $n^{1/2}$, that is, the nuisance parameters are infinite-dimensional and are estimated nonparametrically; i.e., Zhao et al., (2020).

Design-Based Two-Step EL Inference

- Assume that we have at hand a suitable estimator $\hat{\varphi}$ for φ_N.
- Let $(p_1, \cdots, p_n)$ be the discrete probability measure assigned to the n sampled units. For any $\theta \in \Theta$, we define

$$L_n(\theta, \hat{\varphi}) = \sup \left\{ \prod_{i \in \mathcal{S}} (np_i) \,\middle|\, p_i \geq 0, \sum_{i \in \mathcal{S}} p_i = 1, \sum_{i \in \mathcal{S}} p_i \left[\pi_i^{-1} g(Z_i, \theta_N, \hat{\varphi}) \right] = 0 \right\}.$$

- This leads to the following two-step log EL ratio function

$$l_n(\theta, \hat{\varphi}) = -\log\{L_n(\theta, \hat{\varphi})\} = \sum_{i \in \mathcal{S}} \log\{1 + \lambda^\top \pi_i^{-1} g(Z_i, \theta, \hat{\varphi})\}$$

- The two-step maximum EL estimator $\hat{\theta}_{EL}$ for θ_N is the maximum point of $L_n(\theta, \hat{\varphi})$, or equivalently, the minimum point of $l_n(\theta, \hat{\varphi})$.

Design-Based Two-Step EL Inference

Problems with the design-based two-step EL approach:

- For complex survey data, the Wilks' theorem breaks down with the two-step EL approach even under simple random sampling.
- A bootstrap calibration procedure could be employed to compute the limiting distribution of $-2\log L_n(\theta, \hat{\varphi})$, but the method is computationally intensive and theoretical justifications are not available for general survey designs.
- Inferences based on the two-step EL approach do not use information on the main parameters and on the nuisance functionals simultaneously and therefore are not efficient.

Augmented Survey Weighted GEE

- We propose an augmented survey weighted generalized estimating equations (GEE) approach to restore Wilks' phenomenon in design-based two-step EL inferences.
- Assume that

$$\frac{1}{N}\sum_{i\in\mathcal{S}}\pi_i^{-1}g(Z_i,\theta_N,\hat{\varphi}) = \frac{1}{N}\sum_{i\in\mathcal{S}}\pi_i^{-1}\Big\{g(Z_i,\theta_N,\varphi_N) + \Xi(Z,\theta_N,\varphi_N)\Big\} + o_p(n_B^{-1/2}), \tag{1}$$

where $\Xi(Z,\theta_N,\varphi_N)$ has finite fourth population moments and $\sum_{i\in\mathcal{S}}\pi_i^{-1}\Xi(Z_i,\theta_N,\varphi_N)$ is asymptotically normally distributed with mean zero and variance-covariance matrix at the order $O(n_B^{-1}N^2)$.

Augmented Survey Weighted GEE

- The proposed augmented estimating functions are given by

$$\psi(Z, \theta, \varphi) = g(Z, \theta, \varphi) + \Xi(Z, \theta, \varphi), \tag{2}$$

  where the augmentation term $\Xi(Z, \theta, \varphi)$ is specified in (1).
- With the given finite population $\mathcal{F}_N = (Z_1, \cdots, Z_N)$, we define the following augmented population (census) estimating functions

$$\mathbb{U}_N(\theta, \varphi) = \frac{1}{N} \sum_{i=1}^{N} \psi(Z_i, \theta, \varphi). \tag{3}$$

  Note that $\mathbb{U}_N(\theta, \varphi) = 0$ has a unique root at $(\theta, \varphi) = (\theta_N, \varphi_N)$.

Augmented Survey Weighted GEE

- Given the set of sampled units $\mathcal{S}$ and the set of survey weights $\{\pi_i^{-1}, i \in \mathcal{S}\}$, the augmented survey weighted estimating functions are defined as

$$\hat{\mathbb{U}}_N(\theta, \varphi) = \frac{1}{N} \sum_{i \in \mathcal{S}} \pi_i^{-1} \psi(Z_i, \theta, \varphi). \tag{4}$$

- For scenarios where $r = p$, a design-based estimator of θ_N may be obtained by solving $\hat{\mathbb{U}}_N(\theta, \hat{\varphi}) = 0$. The resulting estimator for θ_N is bias-corrected in the sense that

$$\hat{\mathbb{U}}_N(\theta_N, \hat{\varphi}) = \hat{\mathbb{U}}_N(\theta_N, \varphi_N) + o_p(n_B^{-1/2}).$$

In other words, the estimation of the nuisance functional has no impact asymptotically on the estimation of the main parameters of interest.

Bias-Corrected GEL

- Let $\rho(v)$ be a concave function of the scalar $v \in \mathcal{V}$ (an open interval containing zero); let $\rho_j(v) = \partial^j \rho(v)/\partial v^j$ and $\rho_j = \rho_j(0)$ for $j = 0, 1, 2, \ldots$.
- Define the re-centred generalized empirical likelihood (GEL) objective function as

$$\hat{P}_N(\theta, \eta, \varphi) = \sum_{i \in \mathcal{S}} \left\{ \rho\left(\eta^\top \pi_i^{-1} \psi(Z_i, \theta, \varphi)\right) - \rho_0 \right\},$$

where η is an r-vector of "pseudo parameters" related to the Lagrange multipliers.

Bias-Corrected GEL

Given $\hat{\varphi}$, a class of augmented design-based two-step generalized empirical likelihood (GEL) estimators for θ_N can be defined as the solution to the following saddle-point problem

$$\hat{\theta}_{GEL} = \arg \inf_{\theta \in \Theta} \sup_{\eta \in \hat{\Lambda}_{N,\psi}(\theta,\hat{\varphi})} \hat{P}_N(\theta, \eta, \hat{\varphi}), \tag{5}$$

where $\hat{\Lambda}_{N,\psi}(\theta, \varphi) = \{\eta : \eta^\top \pi_i^{-1} \psi(Z_i, \theta, \varphi) \in \mathcal{V}, i \in \mathcal{S}\}$.

Consistency and Efficiency

Besides certain regularity assumptions on the probability sampling design and estimating equations, we further require the following high-level conditions for establishing consistency and efficiency of the proposed estimators.

B1. There exists real-valued functions $\mathbb{U}(\theta, \varphi)$ such that

$$\sup_{(\theta,\varphi)\in\Theta\times\Psi(\delta_N)} \|\mathbb{U}_N(\theta, \varphi) - \mathbb{U}(\theta, \varphi)\| = o(1)$$

for all sequences of positive numbers $\{\delta_N\}$ with $\delta_N = o(1)$.

B2. The ordinary derivative $\Gamma_2(\theta, \varphi)$ of $\mathbb{U}(\theta, \varphi)$ with respect to θ exists for $\theta \in \Theta(\delta)$, and is continuous at $\theta = \theta_N$; the matrix $\Gamma_2(\theta, \varphi)$ has full column rank p;

B3. For all $(\theta, \varphi), (\theta', \varphi') \in \Theta(\delta_N)\times\Psi(\delta_N)$ with $\delta_N = o(1)$, $\|\mathbb{U}(\theta, \varphi) - \mathbb{U}(\theta, \varphi')\| \leq c\|\varphi - \varphi'\|_\Psi^2$ for some constant $c \geq 0$.

Consistency and Efficiency

Theorem 1

Suppose that $\hat{\varphi} = \varphi_N + o_p(1)$, *and that conditions B1–B3 hold. Then, as* $N \to \infty$,

(i) $\lim_{N\to\infty} \Pr\{\|\hat{\theta}_{GEL} - \theta_N\| > \epsilon \mid \mathcal{F}_N\} = 0$ *for any* $\epsilon > 0$.

(ii) *the proposed two-step GEL estimators* $\hat{\theta}_{GEL}$ *are asymptotically normally distributed with mean* θ_N *and variance-covariance matrix* $V_2 = \Sigma_2 \Gamma_2^\top W_2^{-1} \Omega W_2^{-1} \Gamma_2 \Sigma_2$,, *where* $\Sigma_2 = (\Gamma_2^\top W_2^{-1} \Gamma_2)^{-1}$, $\Gamma_2 = \Gamma_2(\theta_N, \varphi_N)$, $W_2 = (n_B/N^2) \sum_{i=1}^N \pi_i^{-1} \psi(Z_i, \theta_N, \varphi_N)^{\otimes 2}$, $\Omega = (n_B/N^2) \mathrm{Var}\{\sum_{i\in\mathcal{S}} \pi_i^{-1} [g(Z_i, \theta_N, \varphi_N) + \Xi(Z_i, \theta_N, \varphi_N)] \mid \mathcal{F}_N\}$.

Corollary 2

Under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, the asymptotic variance-covariance matrix $V_2 = \Sigma_2$.

Hypothesis Testing

The GEL ratio statistic for testing $H_0 : \theta = \theta_N$ is

$$\mathrm{T}_N(\theta) = -2\{[\hat{P}_N(\hat{\theta}_{GEL}, \hat{\eta}_{GEL}, \hat{\varphi}) - \hat{P}_N(\theta, \eta_\theta, \hat{\varphi})\},$$

where $\hat{\eta}_{GEL} = \eta(\hat{\theta}_{GEL}, \hat{\varphi})$ and $\eta_\theta = \eta(\theta, \hat{\varphi})$.

Theorem 3

Suppose that the assumptions for Theorem 1 hold. Then, as $N \to \infty$,

$$\mathrm{T}_N(\theta_N) \xrightarrow{\mathcal{L}} Q^\top \Delta Q,$$

where $Q \sim N(0, I_r)$, $\Delta = \Omega^{1/2} W_2^{-1} \Gamma_2 (\Gamma_2^\top W_2^{-1} \Gamma_2)^{-1} \Gamma_2^\top W_2^{-1} \Omega^{1/2}$, and I_r is the $r \times r$ identity matrix.

Hypothesis Testing

Corollary 4

Under single-stage PPS sampling with replacement or single-stage PPS sampling without replacement with negligible sampling fractions, we have $\mathrm{T}_N(\theta_N) \xrightarrow{\mathcal{L}} \chi^2_p$ *as* $N \to \infty$.

Simulation Studies

- Consider the finite population quantile share $\theta_N(\tau_1, \tau_2)$ with some fixed quantile levels $\tau_1, \tau_2 \in (0,1)$, $\tau_1 \leq \tau_2$.
- Fixed finite population $(Z_1, \cdots, Z_N)$ with size $N = 20,000$ is generated from $Z_i = X_i + \varepsilon_i, \quad i = 1, \cdots, N$, where $X_i \sim 0.25 + \text{Weibull}(2,2)$ and $\varepsilon_i \sim \chi^2_3$.
- Repeated simulation samples of size $n = 300$ are selected from the finite population by the following four sampling methods:
  - (A) Single-stage PPS sampling without replacement with negligible sampling fractions;
  - (B) Single-stage PPS sampling without replacement with non-negligible sampling fractions;
  - (C) Stratified PPS sampling;
  - (D) Two-stage cluster sampling with self-weighting designs.

Simulation Studies

- We consider four scenarios for the quantile levels (τ_1, τ_2), i.e., S1: $(0, 0.25)$; S2: $(0.25, 0.5)$; S3: $(0.5, 0.75)$; S4: $(0.75, 1)$.
- For each selected sample, we use six different methods to construct the 95% confidence intervals for the quantile share $\theta_N(\tau_1, \tau_2)$ at each quantile levels (τ_1, τ_2):
  (1) The GEL ratio confidence intervals using the standard chi-square limiting distributions for each of EL, ET, CU and GMM;
  (2) The normal approximation confidence interval using the estimating equation based point estimator and a bootstrap estimate of the standard error (BC_n);
  (3) The bootstrap percentile interval with the estimating equation based point estimator (BC_p).

Simulation Studies

In all simulations, the proposed method (Augmented SWEE) is compared with the method of Zhao et al. (2020) (Conventional SWEE) under an assumed standard chi-square limiting distribution for the GEL ratio statistic.

95% Confidence Intervals Under the Survey Design A

		Augmented SWEE				Conventional SWEE			
Methods	Levels	LE	CP	UE	AL	LE	CP	UE	AL
EL	S1	0.037	0.940	0.023	0.024	0.000	1.000	0.000	0.068
	S2	0.031	0.947	0.022	0.019	0.000	1.000	0.000	0.091
	S3	0.032	0.940	0.028	0.020	0.000	1.000	0.000	0.111
	S4	0.017	0.944	0.039	0.040	0.000	1.000	0.000	0.137
ET	S1	0.051	0.932	0.017	0.023	0.000	1.000	0.000	0.068
	S2	0.033	0.944	0.023	0.019	0.000	1.000	0.000	0.091
	S3	0.031	0.940	0.029	0.020	0.000	1.000	0.000	0.111
	S4	0.013	0.943	0.044	0.039	0.000	1.000	0.000	0.138
CU	S1	0.057	0.932	0.011	0.023	0.000	1.000	0.000	0.068
	S2	0.031	0.948	0.021	0.019	0.000	1.000	0.000	0.091
	S3	0.031	0.942	0.027	0.020	0.000	1.000	0.000	0.112
	S4	0.009	0.944	0.047	0.040	0.000	1.000	0.000	0.139

95% Confidence Intervals Under the Survey Design A

		Augmented SWEE				Conventional SWEE			
Methods	Levels	LE	CP	UE	AL	LE	CP	UE	AL
GMM	S1	0.059	0.929	0.012	0.023	0.000	1.000	0.000	0.068
	S2	0.033	0.944	0.023	0.019	0.000	1.000	0.000	0.091
	S3	0.031	0.940	0.029	0.020	0.000	1.000	0.000	0.111
	S4	0.010	0.942	0.048	0.039	0.000	1.000	0.000	0.138
BC_n	S1	0.059	0.928	0.013	0.023	0.057	0.930	0.013	0.023
	S2	0.014	0.969	0.017	0.021	0.021	0.966	0.013	0.021
	S3	0.008	0.966	0.026	0.023	0.022	0.958	0.020	0.023
	S4	0.008	0.945	0.047	0.041	0.011	0.951	0.038	0.041
BC_p	S1	0.060	0.926	0.014	0.022	0.045	0.945	0.010	0.024
	S2	0.010	0.972	0.018	0.021	0.012	0.981	0.007	0.024
	S3	0.007	0.965	0.028	0.023	0.004	0.992	0.004	0.026
	S4	0.007	0.936	0.057	0.041	0.009	0.949	0.042	0.041

Thank you for your attention!

Puying Zhao
Department of Statistics, Yunnan University
pyzhao@live.cn