Empirical Likelihood Methods for Survey Data

J.N.K. Rao (Carleton University)
and
Changbao Wu (University of Waterloo)

Invite paper presented at the DASF Conference, Toronto
April 30, 2010

Traditional Design-Based Approach

- Strategies (design and estimation) that appeared reasonable were entertained (accounting for costs). Relative properties carefully studied by analytical and/or empirical methods, mainly through comparison of MSE and anticipated MSE under plausible models
- Design unbiasedness not insisted upon because it "often results in much larger MSE than necessary". Instead, design consistency is deemed necessary for large samples
- Working models used to obtain efficient design-consistent estimators: model-assisted, GREG estimators

Unified Theory

- Finite population: $U = \{1, 2, \cdots, N\}$
- Sampling design: s, $p(s)$
- Sample data: $\{(i, y_i), i \in s\}$
- Godambe class: Estimator of population total $Y = \sum_{i=1}^{N} y_i$ uses weights $d_i(s)$ that may depend on both i and s

$$\hat{Y} = \sum_{i \in s} d_i(s) y_i$$

- **Theorem**: BLUE of Y does not exist in the Godambe class even for simple random sampling.

Non-Parametric Likelihood

- Parameter vector $\tilde{\mathbf{y}} = (\tilde{y}_1, \cdots, \tilde{y}_N)'$; labels i
  Sample data: $\{(i, y_i),\ i \in s\}$ minimal sufficient
- Godambe (1966) likelihood function:

$$L(\tilde{\mathbf{y}}) = \begin{cases} p(s), & \text{if sample data consistent with } \tilde{\mathbf{y}} \\ 0, & \text{otherwise} \end{cases}$$

- Godambe likelihood is **uninformative**: all possible non-observed $y_i, i \notin s$ lead to the same (flat) likelihood.

Non-Parametric Likelihood (Cont'd)

- Resolution I: Bayesian route (Ericson, 1969)
  Specify an informative (exchangeable) prior distribution:
  Given a joint N-dimensional prior on $\tilde{\boldsymbol{y}}$ with pdf $g(\tilde{\boldsymbol{y}})$ and assume the sampling design is independent of $\tilde{\boldsymbol{y}}$, the posterior density is given by

$$h(\tilde{\boldsymbol{y}}|y_i, i \in s) = \begin{cases} g(\tilde{\boldsymbol{y}})/g(\tilde{\boldsymbol{y}}_s) & \text{if } \; y_i = \tilde{y}_i \; \text{ for } \; i \in s\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

- Problems:
  - How to specify $g(\tilde{\boldsymbol{y}})$?
  - Posterior inferences are independent of the sampling design, usually invalid under the design-based frameowrk

Non-Parametric Likelihood (cont'd)

- Resolution II: Likelihood route (Hartley and Rao, 1968)
  Ignore certain aspects of data. For example, for SRS suppress labels i and use $(y_i,\ i \in s)$. Likelihood now becomes informative and inference depends on the sample design.

- C. R. Rao (1970): "*In situations where the full likelihood does not satisfy our purpose, we may have to depend on a statistic T which for every observed value supplies information (however poor it may be) on parameters of interest. Unfortunately, no unique choice of T may be possible.*"

Scale-Load Approach (Hartley and Rao, 1968): SRSWOR

- Finite set of known scale points $y_1^*, \cdots, y_D^*$ with scale loads $N_1, \cdots, N_D$
- Population mean $\bar{Y} = \sum_{j=1}^{D} p_j y_j^*$ where $p_j = N_j/N$
  Sample scale loads $n_1, \cdots, n_D$
- Likelihood function $L(N_1, \cdots, N_D)$ is hypergeometric likelihood with support on $n_j > 0$ $(j = 1, \cdots, d)$: $\prod_{j=1}^{d} \binom{N_j}{n_j}$
- For SRS with replacement, $L(p_1, \cdots, p_d)$ reduces to multinomial likelihood, now popularly known as empirical likelihood (Owen, 1988)

Empirical Likelihood (EL) for Independent Observations

- $y_1, \cdots, y_n$ IID with CDF $F(t)$; Empirical likelihood function

$$L(\boldsymbol{p}) = \prod_{i=1}^{n} p_i$$

Maximizing $L(\boldsymbol{p})$ subject to $p_i > 0$ and $\sum_{i=1}^{n} p_i = 1$ gives
$\hat{p}_i = 1/n$
$\hat{F}(t) = \sum_{i=1}^{n} \hat{p}_i I(y_i \leq t) = F_n(t)$

Empirical Likelihood (Cont'd)

- Owen (1988): Empirical likelihood ratio statistic for $\mu = \int y dF(y)$

$$R(\mu) = \max\left\{ \prod_{i=1}^{n} (np_i) \;\middle|\; \sum_{i=1}^{n} p_i y_i = \mu \,,\; \sum_{i=1}^{n} p_i = 1 \right\}$$

  $-2\log R(\mu)$ is asymptotically distributed as χ_1^2
- EL ratio confidence intervals: Shape and orientation of CI determined entirely by the data; CI are range preserving and transformation respecting
- Qin and Lawless (1994): Estimating equations and EL; side information; additional constraints

Empirical Likelihood (Cont'd)

- Chen and Qin (1993): EL for survey data under simple random sampling
  Maximum EL estimator of $\bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ is given by $\hat{\bar{Y}}_{EL} = \sum_{i\in s}\hat{p}_i y_i$, where $\hat{p}_i$ maximize

$$l(\boldsymbol{p}) = \sum_{i\in s}\log(p_i)$$

  subject to

$$\sum_{i\in s} p_i = 1 \quad \text{and} \quad \sum_{i\in s} p_i \boldsymbol{x}_i = \bar{\boldsymbol{X}}$$

- $\hat{\bar{Y}}_{EL}$ is asymptotically equivalent to the regression estimator under SRS

Empirical Likelihood (Cont'd)

- Zhong and Rao (2000): EL for stratified simple random sampling

$$l(\boldsymbol{p}) = \sum_{h=1}^{L} \sum_{i \in s_h} \log(p_{hi})$$

Constraints:

$$\sum_{i \in s_h} p_{hi} = 1 \;(h = 1, \cdots, L) \;\text{ and }\; \sum_{h=1}^{L} W_h \sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \bar{X}$$

- Maximum EL estimator of $\bar{Y}$ is asymptotically equivalent to optimal regression estimator

Empirical Likelihood (Cont'd)

- An application: Population containing many zero values (Chen, Chen and Rao, 2003)
  Accounting practice: Amount of money owed to government
  Audit sampling: Estimate μ, average amount of excessive claim
- Parametric mixture models (Kvanli, Shen and Deng, 1998): Normal mixture, Exponential mixture

Empirical Likelihood (Cont'd)

p: Population error rate (% Non-zeros)
LNR: Lower non-coverage rate (nominal value: 2.5%)
LB: Average lower bound
True model: Normal mixture

p	Normal Approxi.		Exponential Mixture		Normal Mixture		EL	
	LNR	LB	LNR	LB	LNR	LB	LNR	LB
0.10	0.58	0.19	0.87	0.25	2.08	0.28	2.21	0.28
0.20	1.17	0.63	0.74	0.65	2.13	0.71	2.20	0.71

1. Preliminaries
2. Likelihood-based Approaches
3. Empirical Likelihood Approach: SRS and STSRS
4. Pseudo Empirical Likelihood Approach for Complex Surveys
5. Additional Remarks

Pseudo Empirical Likelihood

- Design-based inference using complex survey data $\{(y_i, \boldsymbol{x}_i), i \in s\}$; $\pi_i = P(i \in s)$, $\pi_{ij} = P(i,j \in s)$; $d_i = 1/\pi_i$
- Pseudo empirical log-likelihood (PEL) function (Chen and Sitter, 1999)
$$l(\boldsymbol{p}) = \sum_{i \in s} d_i \log(p_i)$$
- $l(\boldsymbol{p})$ is the Horvitz-Thompson estimator of the census empirical log-likelihood function $l_N(\boldsymbol{p}) = \sum_{i=1}^{N} \log(p_i)$
- Maximum PEL estimator $\hat{\bar{Y}}_{PEL} = \sum_{i \in s} \hat{p}_i y_i$, where $\hat{p}_i$ maximize $l(\boldsymbol{p})$ subject to
$$\sum_{i \in s} p_i = 1 \quad \text{and} \quad \sum_{i \in s} p_i \boldsymbol{x}_i = \bar{\boldsymbol{X}}\,,$$
is asymptotically equivalent to the generalized regression (GREG) estimator of $\bar{Y}$

Pseudo Empirical Likelihood (Cont'd)

- The PEL function of Chen and Sitter does not involve π_{ij}
- PEL function adjusted by the design effect (Wu and Rao, 2006)

$$l(\boldsymbol{p}) = n^* \sum_{i \in s} \tilde{d}_i(s) \log(p_i) ,$$

  where $n^* = n/\text{deff}$ (effective sample size), "deff" is the design effect and $\tilde{d}_i(s) = d_i / \sum_{i \in s} d_i$
- Under simple random sampling with replacement, $l(\boldsymbol{p}) = \sum_{i \in s} \log(p_i)$

Pseudo Empirical Likelihood (Cont'd)

- PEL ratio function of $\theta = \bar{Y}$

$$r(\theta) = n^* \sum_{i \in s} \tilde{d}_i(s) \log(\hat{p}_i(\theta)) - n^* \sum_{i \in s} \tilde{d}_i(s) \log(\hat{p}_i)$$

$\hat{p}_i(\theta)$ subject to the additional parameter constraint

$$\sum_{i \in s} p_i y_i = \theta$$

- $-2r(\theta)$ converges in distribution to the χ_1^2 random variable (Wu and Rao, 2006)
- $(1-\alpha)$-level PEL ratio confidence interval on $\bar{Y}$

$$\mathcal{C} = \left\{ \theta \mid -2r(\theta) \leq \chi_1^2(\alpha) \right\}$$

Pseudo Empirical Likelihood (Cont'd)

- Confidence intervals on $F(t) = N^{-1}\sum_{i=1}^{N} I(y_i \le t)$ at $t = t_q$
- NA: $\hat{\theta} \pm Z_{\alpha/2}\{v(\hat{\theta})\}^{1/2}$; EL: EL confidence intervals $\mathcal{C}$
- PPS sampling
- CP: Coverage probability; L, U: Lower and Upper tail error rates; AL: Average length

n	q	CI	CP	L	U	AL
80	0.10	NA	90.7	0.2	9.1	0.134
		EL	94.1	1.7	4.2	0.134
	0.50	NA	95.3	2.4	2.3	0.212
		EL	95.5	2.4	2.1	0.208
	0.90	NA	93.9	5.0	1.1	0.116
		EL	95.2	2.7	2.1	0.115

PEL: Multiple Surveys

- Two independent surveys

$$\{(y_i, \boldsymbol{x}_i),\ i \in s_1\} \quad \text{and} \quad \{(y_i, \boldsymbol{x}_i),\ i \in s_2\}$$

- Joint PEL function (Rao and Wu, 2005)

$$l(\boldsymbol{p}_1, \boldsymbol{p}_2) = n_1^* \sum_{i \in s_1} \tilde{d}_{i1}(s_1) \log(p_{i1}) + n_2^* \sum_{i \in s_2} \tilde{d}_{i2}(s_2) \log(p_{i2})$$

- Maximum PEL estimator $\hat{\bar{Y}}_{PEL} = \sum_{i \in s_1} \hat{p}_{i1} y_i = \sum_{i \in s_2} \hat{p}_{i2} y_i$ is asymptotically optimal
- PEL ratio confidence intervals available
- Very flexible in using auxiliary information through added constraints

PEL: Multiple Frame Surveys

- Multiple sampling frames: each of them can be incomplete; together they cover the entire finite population
- Dualframe A and B: $U = a \cup ab \cup b$ (three domains)
- Q-frame survey samples: $\{(y_i, \boldsymbol{x}_i), i \in s_q\}$, $q = 1, \cdots, Q$
- Multiplicity-based PEL function (Rao and Wu, 2009)

$$l_M(\boldsymbol{p}_1, \cdots, \boldsymbol{p}_Q) = \frac{n_M}{\hat{N}_M} \sum_{q=1}^{Q} \sum_{i \in s_q} \frac{d_{qi}}{m_{qi}} \log(p_{qi})$$

– $n_M = \sum_{q=1}^{Q} n_q$; n_q: qth frame sample size
– $\hat{N}_M = \sum_{q=1}^{Q} \sum_{i \in s_q} d_{qi}/m_{qi}$
– d_{qi}: qth frame sampling weights
– m_{qi}: number of frames to which unit i on frame q belongs

PEL: Multiple Frame Surveys

- Pooling together the Q samples into a single one without removing duplicated units
- Auxiliary information can be used through added constraints
- PEL ratio function for $\bar{Y}$ is asymptotically χ_1^2
- Excellent performance in estimating population proportions of rare items

Bayesian Pseudo Empirical Likelihood

- Three issues:
  – (i) likelihood function for complex survey data
  – (ii) prior distribution
  – (iii) posterior distribution providing valid inference under design-based set-up
- Bayesian PEL formulation I: Profile PEL function on $\theta = \bar{Y}$ and flat prior on θ
- Bayesian PEL formulation II: PEL function $l(p_1, \cdots, p_n)$ and Dirichlet-Haldane prior on $(p_1, \cdots, p_n)$
- Both formulations provide posterior inferences which are valid under the design-based set-up (Rao and Wu, 2010)

Bayesian PEL for $\theta = \bar{Y}$

- Profile PEL for θ

$$l_{PEL}(\theta) = n^* \sum_{i \in s} \tilde{d}_i(s) \log \hat{p}_i(\theta) \, ,$$

where the $\hat{p}_i(\theta)$ maximize $\sum_{i \in s} \tilde{d}_i(s) \log p_i$ subject to

$$\begin{aligned} \sum_{i \in s} p_i &= 1 \\ \sum_{i \in s} p_i y_i &= \theta \\ \sum_{i \in s} p_i \boldsymbol{x}_i &= \bar{\boldsymbol{X}} \, . \end{aligned}$$

Bayesian PEL for $\theta = \bar{Y}$

- Posterior distribution of θ under noninformative prior $p(\theta) \propto 1$

$$\pi(\theta|\boldsymbol{y},\boldsymbol{x}) = c(\boldsymbol{y},\boldsymbol{x}) \exp\left\{ -n^* \sum_{i\in s} \tilde{d}_i(s) \log(1 + \boldsymbol{\lambda}'\boldsymbol{u}_i) \right\}$$

where $\boldsymbol{\lambda}$ solves

$$g(\boldsymbol{\lambda}) = \sum_{i\in s} \frac{\tilde{d}_i(s)\boldsymbol{u}_i}{1 + \boldsymbol{\lambda}'\boldsymbol{u}_i} = \boldsymbol{0}$$

with $\boldsymbol{u}_i = (y_i - \theta, \boldsymbol{x}_i' - \bar{\boldsymbol{X}}')'$

Bayesian PEL for $\theta = \bar{Y}$

- The posterior distribution $\pi(\theta | \boldsymbol{y}, \boldsymbol{x})$ is asymptotically normal
- The posterior mean is asymptotically equivalent to the GREG estimator of $\bar{Y}$ and hence is design consistent
- The posterior variance matches the design-based variance of the GREG estimator
- Posterior inferences are valid under the design-based framework

Bayesian PEL based on $(p_1, \cdots, p_n)$

- Treat $p_1, \cdots, p_n$ as parameters
- The pseudo empirical log-likelihood

$$l_{PEL}(\boldsymbol{p}) = n^* \sum_{i \in s} \tilde{d}_i(s) \log p_i$$

- The pseudo empirical likelihood

$$L_{PEL}(\boldsymbol{p}) = \exp\{l_{PEL}(\boldsymbol{p})\} = \prod_{i \in s} p_i^{\gamma_i}$$

$\gamma_i = n^* \tilde{d}_i(s)$

- With the Dirichlet-Haldane prior $\pi(\boldsymbol{p}) \propto \prod p_i^{-1}$, the posterior distribution of $(p_1, \cdots, p_n)$ is also Dirichlet:

$$\pi(p_1, \cdots, p_n | s) \propto \prod_{i=1}^{n} p_i^{\gamma_i - 1}$$

Bayesian PEL based on $(p_1, \cdots, p_n)$

- This is a generalization of Hartley-Rao scale-load method to an arbitrary sampling design
- The posterior distribution of $\theta = \bar{Y}$ is the distribution of $\theta = \sum_{i \in s} p_i y_i$ based on the Dirichlet distribution for $(p_1, \cdots, p_n)$
- Posterior mean and variance of θ match the design-based GREG estimator and its variance
- Valid posterior inference for the mean under the design
- May have an advantage in handling other type of parameters such as quantiles

Other topics

- Adaptive sampling
- Gini and income inequality measures
- Bootstrap procedures
- Missing data
- Analytic use of survey data
- Longitudinal surveys

Empirical Likelihood: Canadian Connections

- Art Owen: BMath, University of Waterloo
- Jing Qin: PhD in Statistics, University of Waterloo
- Jerry Lawless: University of Waterloo
- J.N.K. Rao: Carleton University
- Jiahua Chen: Waterloo and UBC
- Randy Sitter: PhD Waterloo; Carleton; Simon Fraser
- Changbao Wu: PhD Simon Fraser; University of Waterloo