Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Bootstrap procedures for the pseudo empirical likelihood method in sample surveys

Changbao Wu [a,*], J.N.K. Rao [b]

[a] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1 Canada
[b] School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6 Canada

ARTICLE INFO

Article history:
Received 28 May 2009
Received in revised form 15 May 2010
Accepted 23 May 2010
Available online 4 June 2010

Keywords:
Auxiliary information
Confidence interval
Design effect
Profile likelihood
Stratified sampling
Unequal probability sampling

ABSTRACT

Pseudo empirical likelihood ratio confidence intervals for finite population parameters are based on asymptotic χ^2 approximation to an adjusted pseudo empirical likelihood ratio statistic, with the adjustment factor related to the design effect. The calculation of the design effect involves variance estimation and hence requires second order inclusion probabilities. It also depends on how auxiliary information is used, and needs to be derived one-at-a-time for different scenarios. This paper presents bootstrap procedures for constructing pseudo empirical likelihood ratio confidence intervals. The proposed method bypasses the need for design effects and is valid under general single-stage unequal probability sampling designs with small sampling fractions. Different scenarios in using auxiliary information are handled by simply including the same type of benchmark constraints with the bootstrap procedures. Simulation results show that the bootstrap calibrated intervals perform very well and have much improved coverage probabilities over the χ^2-based intervals when the sample sizes are small or moderate.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

The first major result in the development of the empirical likelihood (EL) method was the establishment of asymptotic χ^2 distribution of the EL ratio statistic for the population mean with independent and identically distributed (*iid*) sample data (Owen, 1988). The profile EL ratio confidence intervals have several advantages over the traditional normal theory intervals, such as range-respecting, transformation invariant and data-determined shapes. The EL intervals on the population mean based on the χ^2 approximation, however, tend to have coverage probabilities lower than nominal values when sample sizes are not large. Significant improvement can be made when the χ^2 approximation is replaced by a bootstrap calibration (Owen, 2001, Section 3.3). Bootstrap procedures for the EL method and related theoretical justifications are often straightforward for simple cases involving *iid* samples.

The EL method has been extended to complex survey data through a pseudo EL approach. The pseudo EL function, formulated by Chen and Sitter (1999), is the Narain–Horvitz–Thompson (NHT) estimator (Narain, 1951; Horvitz and Thompson, 1952) of the so-called "census" EL function and hence involves first order inclusion probabilities. It is shown by Wu and Rao (2006) that the pseudo EL ratio confidence intervals based on the χ^2 approximation require an adjustment factor which involves variance estimation and hence evaluation of second order inclusion probabilities. In addition, the adjustment factor also depends on how auxiliary information is used and needs to be derived one-at-a-time for different scenarios. Rao and Wu (2009) presented a detailed account of the EL methods for finite populations.

[*] Corresponding author. Tel.: +1 519 888 4567x35537; fax: +1 519 746 1875.
E-mail address: cbwu@uwaterloo.ca (C. Wu).

0167-7152/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2010.05.015

In this paper we present bootstrap procedures which bypass the adjustment factor previously required for the construction of pseudo EL ratio confidence intervals. The use of auxiliary information is handled in a unified manner by simply including the same type of benchmark constraints or calibration equations with the bootstrap procedure. The proposed methods are theoretically justified for single-stage with-replacement unequal probability sampling designs. They are also valid for without-replacement sampling designs when sampling fractions are small. Simulation results show that the bootstrap calibrated pseudo EL confidence intervals perform much better than those based on the χ^2 approximation using the adjustment factor when sample sizes are small. For simple random sampling without replacement with non-negligible sampling fractions, a simple correction can be made to the proposed bootstrap method and the resulting confidence intervals have correct asymptotic coverage probabilities. Simulation results show that the simple correction can also provide improved performance when it is applied to without-replacement unequal probability sampling designs with large sampling fractions.

A brief review of the pseudo EL ratio confidence intervals for complex surveys is given in Section 2. The proposed bootstrap procedures are presented in Section 3. Results from an extensive simulation study are reported in Section 4. We conclude with a few additional remarks in Section 5.

2. The pseudo empirical likelihood method

Consider a finite population $\mathcal{U}$ consisting of N units. Let $\{(y_i, \boldsymbol{x}_i),\ i \in s\}$ be a non-stratified probability sample with fixed sample size n, where y_i and $\boldsymbol{x}_i$ are the values of the response variable y and the vector of auxiliary variables $\boldsymbol{x}$ associated with the ith unit, and s is the set of sample units selected using a probability sampling design. Let $\pi_i = P\ (i \in s)$ be the inclusion probabilities and $d_i = 1/\pi_i$ be the basic design weights. The pseudo EL function, first proposed by Chen and Sitter (1999), is given by $l(\boldsymbol{p}) = \sum_{i\in s} d_i \log(p_i)$ which is the NHT estimator of the so-called "census" empirical likelihood $\sum_{i=1}^N \log(p_i)$. This definition works fine for point estimation of population parameters but is not convenient for interval estimation or hypothesis testing.

The pseudo empirical likelihood (PEL) function defined in Wu and Rao (2006) is given by

$$l_{ns}(\boldsymbol{p}) = n \sum_{i\in s} \tilde{d}_i(s) \log(p_i), \tag{2.1}$$

where $\tilde{d}_i(s) = d_i / \sum_{i\in s} d_i$ are the normalized design weights and $\boldsymbol{p} = (p_1, \ldots, p_n)'$ is the discrete probability measure imposed over the sampled units. Maximizing $l_{ns}(\boldsymbol{p})$ subject to $p_i > 0$ and $\sum_{i\in s} p_i = 1$ gives $\hat{p}_i = \tilde{d}_i(s)$. The maximum PEL estimator for the population mean $\bar{Y} = N^{-1} \sum_{i=1}^N y_i$ is given by $\hat{\bar{Y}}_{\text{PEL}} = \sum_{i\in s} \hat{p}_i y_i = \sum_{i\in s} \tilde{d}_i(s) y_i$, and $\hat{\bar{Y}}_{\text{PEL}}$ is identical to the well-known Hájek estimator of $\bar{Y}$.

For PEL ratio confidence intervals on $\bar{Y}$, we consider the general case where the vector of population means, $\bar{\boldsymbol{X}} = N^{-1} \sum_{i=1}^N \boldsymbol{x}_i$, is known and needs to be incorporated into inferences. Let $\hat{p}_i$, $i \in s$ be the maximizer of $l_{ns}(\boldsymbol{p})$ subject to

$$\sum_{i\in s} p_i = 1, \tag{2.2}$$

$$\sum_{i\in s} p_i \boldsymbol{x}_i = \bar{\boldsymbol{X}}. \tag{2.3}$$

The maximum PEL estimator of $\bar{Y}$ in this case is again defined as $\hat{\bar{Y}}_{\text{PEL}} = \sum_{i\in s} \hat{p}_i y_i$. Let $\tilde{p}_i(\theta)$ be the maximizer of $l_{ns}(\boldsymbol{p})$ subject to (2.2), (2.3) and an additional constraint induced by the parameter of interest, $\bar{Y}$,

$$\sum_{i\in s} p_i y_i = \theta \tag{2.4}$$

for a fixed θ. Let $r_{ns}(\theta) = -2\{l_{ns}(\tilde{\boldsymbol{p}}(\theta)) - l_{ns}(\hat{\boldsymbol{p}})\}$ be the PEL ratio function. It is shown by Wu and Rao (2006) that, under suitable regularity conditions, the adjusted PEL ratio function $r_{ns}^{[a]}(\theta) = r_{ns}(\theta)/\text{deff}_{GR}$ converges in distribution to a χ^2 random variable with one degree of freedom when $\theta = \bar{Y}$, where the design effect deff_{GR} is calculated based on a generalized regression (GR) estimator and requires second order inclusion probabilities $\pi_{ij} = P\ (i, j \in s)$.

A stratified probability sample is given by $\{(y_{hi}, \boldsymbol{x}_{hi}),\ i \in s_h,\ h = 1, \ldots, H\}$ where $(y_{hi}, \boldsymbol{x}_{hi})$ is the value of $(y, \boldsymbol{x})$ associated with the ith unit in stratum h, s_h is the set of sample units selected from stratum h with fixed stratum sample size n_h, and H is the total number of strata in the population. Let $n = \sum_{h=1}^H n_h$ be the overall sample size. Let d_{hi} be the stratum design weights and $\tilde{d}_{hi}(s_h) = d_{hi} / \sum_{i\in s_h} d_{hi}$ be the normalized stratum design weights. The PEL function under stratified sampling is defined as

$$l_{st}(\boldsymbol{p}_1, \ldots, \boldsymbol{p}_H) = n \sum_{h=1}^H W_h \sum_{i\in s_h} \tilde{d}_{hi}(s_h) \log(p_{hi}), \tag{2.5}$$

Download English Version:

https://daneshyari.com/en/article/1152849

Download Persian Version:

https://daneshyari.com/article/1152849

Daneshyari.com