Statistica Sinica **9**(1999), 385-406

A PSEUDO EMPIRICAL LIKELIHOOD APPROACH TO THE EFFECTIVE USE OF AUXILIARY INFORMATION IN COMPLEX SURVEYS

Jiahua Chen and R. R. Sitter

University of Waterloo and Simon Fraser University

Abstract: In this paper, we develop a pseudo empirical likelihood approach to incorporating auxiliary information into estimates from complex surveys. In simple random sampling without replacement, the method reduces to the empirical likelihood approach of Chen and Qin (1993). We show that the method is asymptotically equivalent to a generalized regression estimator in the case of estimating a mean or population distribution function with known population means for a vector of auxiliary variables. We go on to investigate, in a simple case, the incorporation of more complex auxiliary information, and demonstrate the resulting increase in efficiency using the proposed approach both theoretically and through a limited simulation.

Key words and phrases: Calibration, generalized regression estimator, jackknife, optimal regression estimator.

1. Introduction

In sample surveys, auxiliary information on the finite population is regularly used to increase the precision of estimators, most commonly estimators of the population mean or total. Ratio and regression estimators incorporate known finite population means of auxiliary variables. Calibration estimators adjust basic survey weights so the sample sum of a weighted auxiliary variable equals its known population total. Deville and Sarndal (1992) propose a general calibration method which minimizes a chosen distance measure between the adjusted survey weights, called calibration weights, and the basic survey weights subject to consistency constraints, called calibration equations. They show that a "chi-square distance" leads to the generalized regression estimators (GREG) (Sarndal (1980), Bethlehem and Keller (1987)). For a good overview and recent related developments, in particular consideration of the population distribution function, see Rao (1994).

Chen and Qin (1993) propose an empirical likelihood approach to the use of auxiliary information in simple random sampling without replacement (srswor). Their theoretical and simulation results suggest the approach has desirable properties when estimating means, totals and population distribution functions, as

well as quantiles. This holds when the auxiliary information is in the form of a known population mean or total of some auxiliary variable, and when it is in the form of a known population quantile. Unfortunately, their formulation of the method does not extend to more complex survey designs.

In this paper, we develop a pseudo empirical likelihood approach for complex surveys which reduces to Chen and Qin's method in the case of simple random sampling. In Section 4, we show that, in some situations, the method is asymptotically equivalent to a GREG in the case of estimating a mean or population distribution function with known population means for a vector of auxiliary variables. We consider stratified sampling and incorporation of known strata size information, and demonstrate the resulting increase in efficiency both theoretically, in Section 5, and through a limited simulation, in Section 7.

2. Empirical Likelihood Estimation in Finite Populations

Suppose a finite population, S, consists of N distinct units with measurements z_i, $i = 1, \ldots, N$, which themselves are a random sample from a superpopulation. The simplest case is when the z_i are assumed to be independent and identically distributed with population distribution $F(z)$. If the entire finite population was available, the corresponding likelihood function would be $L(F) = \prod_{i=1}^{N} p_i$ with log–likelihood function

$$l(p) = \sum_{i=1}^{N} \log(p_i), \tag{1}$$

where $p_i = p(z_i)$ is the density at observation z_i.

This density function could be modelled parametrically as $p(z_i, \theta)$. In this case θ is an unknown superpopulation parameter. Let θ_N be an estimator of θ based upon $(z_1, \ldots, z_N)$. In this context, the purpose may be to estimate the superpopulation parameter, θ, or to estimate θ_N with θ_n based on a sample $s \subset S$ of size n. The argument for the latter, as described in Godambe and Thompson (1986), is that, since N is typically large, θ_N will be very close to θ, while if the true superpopulation departs from the model, θ_N may still be of interest as a finite population characteristic (see also Binder (1983), Godambe and Thompson (1996)).

Now consider $F_N(z) = N^{-1} \sum_{i=1}^{N} I_{[z_i \le z]}$, where $I_{[Z \le z]}$ is the componentwise indicator function. One could view $F_N(z)$ as the nonparametric maximum likelihood estimate of $F(z)$, based on $(z_1, \ldots, z_N)$. A nonparametric analogue of the above rationale is then available, with F or $\theta(F)$ and F_N or $\theta(F_N)$ analogous to θ and θ_N. If we proceed further and assume nothing is known about $p_i = p(z_i)$ and require $\sum p_i \le 1$, then (1) is the empirical log-likelihood (see Owen (1990), Chen

and Qin (1993)). If the entire finite population were available, we could then maximize (1), possibly subject to additional constraints based on some auxiliary information. In practice, we have only a sample, s, of size n of the entire finite population. For the purpose of illustration, let us first consider srswor and suppose we have obtained the sample $(z_i, i \in s)$ with $z_i = (y_i, x_i)^T$ a two dimensional vector. Let the superpopulation parameter of interest be $F(y)$ or $\theta(F)$ for some θ (eg., $\bar{Y} = \int y dF(y)$) with corresponding finite population parameter $F_N(y)$ or $\theta(F_N)$ (eg., $\bar{Y}_N = \int y dF_N(y)$). To overcome the difficulty of not knowing z_i for the entire finite population, we view the log–likelihood function in (1) as a finite population total. Then we have available a design unbiased estimate of $l(p)$, namely

$$\hat{l}(p) = \frac{N}{n} \sum_{i \in s} \log(p_i). \tag{2}$$

If we require $0 < \sum_{i \in s} p_i \leq 1$, this is the empirical log-likelihood function for srswor as defined in Chen and Qin (1993). When no auxiliary information is available, maximizing (2) yields $p_i = 1/n$ for all $i \in s$; the "maximum empirical likelihood estimator" of the finite population distribution function, $F_N(y)$, is the usual empirical distribution function, $F_n(y) = (1/n) \sum_{i \in s} I_{[y_i \leq y]}$, and the estimate of θ_N (eg., $\bar{Y}$) is $\theta_n = \theta(F_n)$ (eg., the sample mean, $\bar{y}_n = \int y dF_n(y)$). In the spirit of Godambe and Thompson (1986), note that if the entire finite population is known, $F_n(y)$ and θ_n become $F_N(y)$ and θ_N, respectively, and can be viewed as estimates of the corresponding superpopulation parameters, $F(y)$, and $\theta = \theta(F)$.

One can now incorporate auxiliary information by placing additional constraints on the maximization as in Chen and Qin (1993). For example, suppose $\bar{X}_N = \int x dF_N(x)$ is known. Then one could maximize (2) subject to $0 < \sum_{i \in s} p_i \leq 1$ and

$$\sum_{i \in s} p_i (x_i - \bar{X}_N) = 0. \tag{3}$$

If $\hat{p}_i$ results, $\sum_{i \in s} \hat{p}_i I_{[y_i \leq y]}$ is the empirical likelihood estimate of $F_N(y)$, and, for example, $\theta_n = \sum_{i \in s} \hat{p}_i y_i$ is the empirical likelihood estimate of $\theta_N = \bar{Y}_N$. Note that if the entire finite population is known, constraint (3) is satisfied.

Chen and Qin (1993) show that the empirical likelihood approach indeed has some desirable properties in this context. In simple situations, this approach coincides with commonly used approaches such as the poststratification method, and the raking method. Also, the maximum empirical likelihood estimate is asymptotically normal for smooth functions of the mean, and its asymptotic variance is the same as the regression estimator. A Bahadur-type representation for quantiles was also established. Simulations indicate that it has favorable finite sample properties as well.

Chen and Qin (1993), however, motivates the use of (2) in two entirely different ways: (i) by noting that the model-based likelihood for any non-informative sample s is $\prod p_i$ (see Jagers (1986) and Cassel, Sarndal and Wretman (1977), p. 109); and (ii) by noting that, if we assume the values y_i can take only a finite number of values, under simple random sampling the design based-likelihood is a multi-dimensional hypergeometric distribution. Hartley and Rao's (1968) use of this approach (what they termed the scale load approach) in estimating $\bar{Y}_N$ with $\bar{X}_N$ known is likely the first application of the concepts behind empirical likelihood. These motivations do not, unfortunately, extend well to more complex sampling designs.

3. Pseudo Empirical Likelihood Estimation in Finite Populations

Consider a finite population, S, of N distinct units with measurements z_i as in the previous section. But now suppose the sample, s, is drawn using some sampling design, $p(\cdot)$. That is, the sample $s \subset S$ is drawn with probability $p(s)$. It now becomes very difficult to extend to this more general setup either of the motivations for empirical likelihood of Chen and Qin (1993). The development of the previous section, however, extends quite naturally. We have available a design unbiased estimate of $l(p)$, namely

$$\hat{l}(p) = \sum_{i \in s} d_i(s) \log p_i, \tag{4}$$

where the $d_i(s)$ are the design weights, with $E(\sum_{i\in s} d_i(s) \log p_i) = \sum_{i=1}^{N} \log p_i$. Here, E refers to expectation under the sampling design. We term (4) the "pseudo-empirical likelihood". For auxiliary information of the form $E\{u(Z)\} = \sum_{i=1}^{N} u(z_i)/N = 0$, the problem then reduces to maximizing (4) subject to

$$\sum_{i \in s} p_i = 1, \quad \sum_{i \in s} p_i u(z_i) = 0 \quad (0 \le p_i \le 1). \tag{5}$$

For example in (3), $u(z_i) = (x_i - \bar{X}_N)$. Using the Lagrange multiplier method it is easily shown that, for any finite population parameters that can be written as $\theta_N = \theta(F_N)$, the resulting *pseudo empirical maximum likelihood estimator* (PEMLE) is $\hat{\theta}_n = \theta(\hat{F}_n)$, $\hat{F}_n = \sum_{i\in s} \hat{p}_i \delta_{z_i}$, where δ_{z_i} is the point measure at z_i, the $\hat{p}_i = w_i(s)/[1 + \lambda u(z_i)]$ for $i \in s$, and the Lagrange multiplier, λ, is the solution to

$$\sum_{i \in s} \frac{w_i(s) u(z_i)}{\{1 + \lambda u(z_i)\}} = 0, \tag{6}$$

where $w_i(s) = d_i(s)/\sum_{i\in s} d_i(s)$. If $u(\cdot)$ is vector valued, this extends naturally using a vector valued λ.

Note that if there is no auxiliary information, i.e. $u(z_i) = 0$, this approach yields $\hat{p}_i = d_i(s)/\sum_{i\in s} d_i(s)$. This is a very attractive property. For example, if the characteristic of interest is the population mean, $\bar{Y}_N$, and one uses the Horwitz-Thompson estimator $\hat{l}(p) = \sum_{i\in s}(1/\pi_i)\log p_i$ of $l(p)$, where π_i is the inclusion probability of the ith unit, then with no auxiliary information one gets $\hat{\bar{Y}}_N = \sum_{i\in s}(1/\pi_i)y_i/\sum_{i\in s}(1/\pi_i)$ and not $\sum_{i\in s}(1/\pi_i)y_i/N$. It was illustrated in Rao (1966), and later in the more well known Basu (1971) elephant example, that even though the first estimator estimates the the population size N and the second uses its known quantity, the first has better properties.

Also note, this is not equivalent to the suggested approach in Chen and Qin (1993) for unequal probability sampling. For example, in the case of no auxiliary information, $u(z_i) = 0$, their method yields $p_i = 1/n$ and the resulting estimator of F_N is not design unbiased.

4. PEMLE's and GREG's

For the remainder of the paper, we consider the situation where $z = (y, x)$, $u(z) = h(x) - \bar{H}_N$ and $\theta_N = \int g(y)dF_N(y)$ in the setting of Section 3. Note that x and $h(x)$ may be vector valued. Here $\bar{H}_N = \sum_{i=1}^N h(x_i)/N$ for some function h. The choice $g(y) = y$ and $h(x) = x$ corresponds to estimating the population mean $\bar{Y}_N$ when $\bar{X}_N$ is known, while $g(y) = \Delta(t-y)$ and $h(x) = x$, where $\Delta(a) = 1$ when $a \geq 0$ and $\Delta(a) = 0$ otherwise, corresponds to estimating the population distribution function at t when $\bar{X}_N$ is known.

Hartley and Rao (1968) consider the problem of estimating the population mean $\bar{Y}_N$ when $\bar{X}_N$ is a known scalar in the case of srswor, and in essence showed that maximizing the empirical likelihood is asymptotically equivalent to a regression estimator. In this more general setting a similar result holds. For simplicity, we state results for a scalar $h(x) = x$ and $g(y) = y$, though they hold generally.

Theorem 1. *Under conditions* 1 *and* 2 *(below), the PEMLE of* $\bar{Y}_N$*, when* $\bar{X}_N$ *is known, is asymptotically equivalent to a generalized regression estimator (GREG). That is,*

$$\lambda \doteq (\bar{x}_w - \bar{X}_N)/\sum_{i\in s} w_i(s)(x_i - \bar{x}_w)^2 + o_p(n^{-1/2})$$

and thus $\hat{\bar{Y}}_N = \bar{y}_{GREG} + o_p(n^{-1/2})$*, where* $\bar{y}_{GREG} = \sum_{i\in s}\tilde{w}_i(s)y_i$*,* $\tilde{w}_i(s) = w_i(s)[1 - (x_i - \bar{x}_w)(\bar{x}_w - \bar{X}_N)/\sum_{i\in s} w_i(s)(x_i - \bar{x}_w)^2]$*,* $\bar{y}_w = \sum_{i\in s} w_i(s)y_i$*,* $\bar{x}_w = \sum_{i\in s} w_i(s)x_i$ *and* $w_i(s) = d_i(s)/\sum_{j\in s} d_j(s)$*.*

The proof is given in Appendix 1. Defining $u_i = x_i - \bar{X}_N$, the conditions needed are:

1) $u^* = \max_{i\in s} |u_i| = o_p(n^{1/2})$;
2) $\sum_{i\in s} d_i(s)u_i / \sum_{i\in s} d_i(s)u_i^2 = O_p(n^{-1/2})$.

We have stated these necessary conditions in a general form which is compact but not very enlightening. Many commonly used sampling designs satisfy these conditions under some moderate assumptions. Appendix 2 gives such for three common designs, namely pps sampling with replacement, the Rao-Hartley-Cochran method, and cluster sampling.

Note that the above three designs do not involve stratification. It turns out that stratified designs offer an excellent opportunity to explore the pseudo empirical likelihood approach relative to possible competitors. This we do theoretically in Section 5, and through simulation in Section 7.

5. PEMLE, GREG and ORE in Stratified Sampling

In this section, we consider the PEMLE for stratified single-stage, and stratified multi-stage sampling, and demonstrate how the method is well-suited to incorporating different types of auxiliary information to advantage. We point out that the pseudo empirical likelihood approach provides both method and motivation for efficiently using information on the stratum population sizes, which is left out by the GREG and the optimal regression estimator (ORE) proposed by Rao (1994). Under stratified sampling, the ORE has been shown to be more efficient than the GREG. This is because the ORE explicitly makes use of the correlation structure between y and x. In the case of stratified srswor, the sampling weights are constant within each stratum and including the stratum size information is equivalent to including the correlation structure information. Thus the PEMLE and ORE are equivalent, and both are better than the GREG. When other sampling plans are used within each stratum, for example pps sampling, the stratum sizes contain important information that is not provided by the sampling weights nor the correlation structure. In this case, direct theoretical comparisons are difficult. However in Section 7 we demonstrate, via simulation, that the improvement of the pseudo empirical likelihood approach over the optimal regression approach can be substantial.

5.1. Stratified single-stage sampling

In stratified sampling the population of N units is first partitioned into non-overlapping sub-populations called strata, of size $N_1, \ldots, N_L$ units, respectively. Independent samples of size n_h are drawn from each stratum h.

If we assume that stratum h consists of N_h distinct Z_{hi}, themselves independently distributed from F_h, independent for $h = 1, \ldots, L$, then the log–likelihood function is $l(p) = \sum_{h=1}^{L} \sum_{j=1}^{N_h} \log(p_{hj})$. Viewing $l(p)$ as a population total, the

most commonly used design unbiased estimate of the population empirical log–likelihood would be

$$\hat{l}(p) = \sum_{h=1}^{L} \sum_{j \in s_h} d_{hi}(s) \log(p_{hj}), \tag{7}$$

as in Section 3. In the case of stratified srswor, $d_{hj}(s) = N_h/n_h$.

Suppose we naively apply the method in Section 3 (with hi replacing i throughout) to obtain the PEMLE of $\bar{Y}_N$ when $\bar{X}_N$ is known, without considering the fact that we have the additional information contained in the knowledge of $N_1, \ldots, N_L$. Then from Theorem 1,

$$\hat{\bar{Y}}_N = \bar{y}_w - \frac{\sum_h \sum_{i \in s_h} w_{hi}(s)(x_{hi} - \bar{x}_w) y_{hi}}{\sum_h \sum_{i \in s_h} w_{hi}(s)(x_{hi} - \bar{x}_w)^2} (\bar{x}_w - \bar{X}_N) + o_p(n^{-1/2}), \tag{8}$$

where $n = \sum_h n_h$, $\bar{y}_w = \sum_h \sum_i w_{hi}(s) y_{hi}$ and $\bar{x}_w = \sum_h \sum_i w_{hi}(s) x_{hi}$ and $w_{hi}(s) = d_{hi}(s) / \sum_h \sum_i d_{hi}$.

Consider stratified srswor, i.e. $d_{hi}(s) = N_h/n_h$. In this case, (8) reduces to

$$\begin{aligned} \hat{\bar{Y}}_N &= \bar{y}_{st} - \frac{\sum_h \sum_{i \in s_h} W_h (x_{hi} - \bar{x}_{st}) y_{hi}/n_h}{\sum_h \sum_{i \in s_h} W_h (x_{hi} - \bar{x}_{st})^2/n_h} (\bar{x}_{st} - \bar{X}_N) + o_p(n^{-1/2}) \\ &= \bar{y}_{GERG} + o_p(n^{-1/2}), \end{aligned}$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$. This, of course, cannot be the best possible approach. The optimal regression estimator, which is known to be more efficient than the above one, replaces $\bar{x}_{st}$ by $\bar{x}_h$ in the ratio of summations.

To see why we call this application of Theorem 1 naive, note that $F_N(z) = \sum_h W_h F_{N_h}(z)$, $\bar{Z}_N = \int z dF_N(z) = \sum_h W_h \int z dF_{N_h}(z)$ and $\int u(z) dF_N(z) = 0$ can be written $\sum_h W_h \int u(z) dF_{N_h}(z) = 0$. This knowledge of the form of F_N contained in W_h should be incorporated in constructing the PEMLE. The empirical likelihood approach is well-suited to incorporate auxiliary information and can accommodate this information contained in the population size for each stratum quite naturally. To see this, let $z_i = (y_i, U_i^T)^T$ for $i = 1, \ldots, N$, where $U_i = (x_i, v_{1i}, \ldots, v_{Li})^T$ and $v_{hi} = 1$ if $i \in h$ and 0 otherwise. Then $\bar{U}_N = (\bar{X}_N, W_1, \ldots, W_L)^T$ is known. Letting $u(z_i) = U_i - \bar{U}_N$ and applying Theorem 1 amounts to maximizing the pseudo empirical log-likelihood function (7) subject to

$$\sum_{i \in s_h} p_{hi} = W_h \quad \text{for } h = 1, \ldots, L \quad \text{and} \quad \sum_h \sum_{i \in s_h} p_{hi} x_{hi} = 0,$$

and using the resulting $\hat{p}_{hi}$ to get $\hat{F}_N(z) = \sum_h \sum_{i \in s_h} \hat{p}_{hi} \delta_{y_{hi}}$ and thus $\hat{\theta} = \theta(\hat{F}_N)$, a PEMLE of $\theta_N = \theta(F_N)$.

Remark. Constructing the pseudo empirical likelihood for each stratum will result in the same combined pseudo empirical likelihood with the same constraints.

When y_i is a scalar quantity, and no auxiliary information beyond the stratum sizes is available, the resulting PEMLE of the population mean is the usual unbiased estimator of the mean, i.e. under stratified srswor it is the usual stratified mean, $\bar{y}_{st}$. Suppose, instead, that $\bar{X}_N = \sum_h \sum_j x_{hj}/N$ is also known. Then the pseudo empirical likelihood should be maximized with restriction

$$\sum_{h=1}^{L} \sum_{j \in s_h} p_{hj} x_{hj} = \sum_{h=1}^{L} W_h \tilde{x}_h = \bar{X}_N, \tag{9}$$

with $W_h \tilde{x}_h = \sum_{j \in s_h} p_{hj} x_{hj}$. Viewing this maximization problem, two questions arise: (a) when does a unique solution exist? and (b) how do we solve it numerically?

In Appendix 3, we develop the following simple numerical method. The method involves finding t such that $\sum_{h=1}^{L} W_h \tilde{x}_h = \bar{X}_N$, where, for a given t, $\tilde{x}_h$ for $h = 1, \ldots, L$ are the solutions to

$$\sum_{i \in s_h} \frac{d_{hi}(x_{hi} - \tilde{x}_h)}{d_h + tW_h(x_{hi} - \tilde{x}_h)} = 0, \tag{10}$$

where $d_h = \sum_{i \in s_h} d_{hi}$. Since the $\tilde{x}_h$ are functions of t through (10) and one can show that $\sum_h W_h \tilde{x}_h$ is monotonic in t, we need only increase or decrease the size of t to determine the existence of the solution, while the uniqueness is a simple consequence of the monotonicity. Once we obtain the correct t and thus $\tilde{x}_h$ for $h = 1, \ldots, L$, $\hat{p}_{hi} = W_h d_{hi}/[d_h + tW_h(x_{hi} - \tilde{x}_h)]$.

Large sample results can be obtained in the same way as in Theorem 1 by including stratum indicators as part of z_i as above, or plainly speaking, by making the auxiliary variable vector-valued by adding stratum indicator variables. This also indicates that both the GREG and the ORE might be improved in the light of this pseudo empirical likelihood approach, by including stratum indicator variables in their respective derivations. The potential usefulness of this usually ignored information, and how to incorporate it, becomes obvious when the problem is viewed in an empirical likelihood framework. This is not the case for the other two methods. There has been no discussion in the literature on utilizing the stratum size information in this way.

To offer some insight in comparing the pseudo empirical likelihood approach and the optimal regression approach, we offer the following result for stratified srswor. Assume that there is a sequence of finite populations indexed by ν such that when $\nu \to \infty$: (i) $0 \le c_1 \le \sum_{h=1}^{L} W_h \sigma_h^2 \le c_2 \le \infty$; (ii) $\max\{n_h^{-1} W_h\} =$

$O(n^{-1})$; (iii) $N^{-1}\sum_{h=1}^{L}\sum_{i=1}^{N_h}|x_{hi}|^3 = O(1)$; and (iv) $N^{-1}\sum_{h=1}^{L}\sum_{i=1}^{N_h}|y_{hi}|^3 = O(1)$. See Rao and Wu (1985) for discussion of the first two assumptions. Note that the second assumption allows for the two most common situations: n_h bounded with $L \to \infty$; and L bounded with $n_h \to \infty$ for each h. Assumption (iv) is used to control the size of the remainder in the next theorem. It is not needed for the expansion.

Corollary 1. *Under stratified srswor and conditions* (i)-(iv) *above, the PEMLE of* $\bar{Y}_N$*, when* $\bar{X}_N$ *is known and the stratum size information is incorporated, is asymptotically equivalent to*

$$\hat{\bar{Y}}_N = \bar{y}_{st} - \frac{\sum_{h=1}^{L} W_h \sum_{i\in s_h}(x_{hi}-\tilde{x}_h)y_{hi}/n_h}{\sum_{h=1}^{L} W_h \sum_{i\in s_h}(x_{hi}-\tilde{x}_h)^2/n_h}(\bar{x}_{st}-\bar{X}_N) + o_p(n^{-1/2}), \qquad (11)$$

where $\tilde{x}_h$ *for* $h = 1, \ldots, L$ *are defined in equation* (A.3) *of Appendix* 3.

The proof is given in Appendix 4.

From the discussion in Appendix 4, it is known that, when L remains finite, $\tilde{x}_h - \bar{x}_h = O_p(n^{-1/2})$. Hence, in that case, the above estimator is asymptotically equivalent to the optimal linear estimator given by Rao (1994). Zhong and Rao (1996) extend the scale-load approach of Hartley and Rao (1968) to stratified srswor and thus derives an empirical likelihood estimator for this situation which is of similar form to (11) and is also asymptotically equivalent to the ORE.

When other sampling designs are used inside each stratum, closed-form comparison between the two approaches becomes tedious. The comparison will be done by simulation in Section 7.

5.2. Stratified multi-stage sampling

Many large scale surveys use a stratified multistage design. The population is stratified into L strata with the hth stratum consisting of N_h clusters. A sample of $n_h \geq 2$ clusters is drawn from stratum h, independently across strata. A subsample is then drawn from each obtained cluster. We assume that subsampling within cluster is performed to ensure unbiased estimation of cluster totals, Y_{hi}. The usual unbiased estimator of the population total Y_N is of the form $\hat{Y}_N = \sum_{hij\in s} d_{hij}(s)y_{hij}$, where s is the sample and y_{hij} refers to the value of interest of the jth unit in the ith cluster within the hth stratum. If there is no auxiliary information, the PEMLE of $\bar{Y}_N$ would be $\hat{Y}_N/\sum d_{hij}$, a ratio estimator. This follows directly from the result in Section 3 with the subscript i replaced by hij.

As before, if the $u_{hij} = x_{hij} - \bar{X}_N$ are known, the pseudo empirical likelihood can be maximized with an additional restriction $\sum_{hij\in s} p_{hij}u_{hij} = 0$. If the

conditions in Theorem 1 are satisfied, the PEMLE of $\bar{Y}_N$ will be

$$\hat{\bar{Y}}_N = \sum_{hij} \tilde{w}_{hij}(s) y_{hij} + o_p(n^{-1/2}) = \bar{y}_{GREG} + o_p(n^{-1/2}), \tag{12}$$

where $\tilde{w}_{hij}(s) = w_{hij}(s)[1 - (u_{hij} - \bar{u}_w)\bar{u}_w / \sum_{hij} w_{hij}(s)(u_{hij} - \bar{u}_w)^2]$ and $\bar{u}_w = \sum_{hij \in s} w_{hij}(s) u_{hij}$.

The conditions of Theorem 1, unfortunately, have to be verified case by case. However, as in the general PPS case, if we consider the outcome of u as a random variable with finite or "not so large" fourth moment, we can use a Chebyshev type inequality to show $u^* = o_p(n^{1/2})$. The second condition of Theorem 1 is usually satisfied and is often required by other methods as well.

If the number of psu's within each stratum of the population is known, it is possible to incorporate this additional information in exactly the same way as in the previous section. The numerical algorithm for solving the resulting maximization problem is the same as that discussed for single-stage sampling with the hi subscript replaced by hij throughout and $d_h = \sum_i \sum_j d_{hij}$.

6. Variance Estimation and Central Limit Theorem

For the purposes of variance estimation, it is clear from Theorem 1 that any consistent variance estimator, $\hat{\sigma}^2$, for $\bar{y}_{GREG}$ will remain consistent for the PEMLE, $\hat{\bar{Y}}_N$. This includes the stratified cases if one uses a vector-valued auxiliary variate which includes the strata indicators. Consider, for example, stratified multistage sampling as described in Section 5.2. In this case, many such variance estimators are available if we treat the sample as if the clusters were sampled with replacement. This is common practice for the purpose of variance estimation. The approximation leads to overestimation of the variance but the relative bias is likely to be small if the first stage sampling fractions are small. For example, the linearization-substitution method (Rao (1988)) for $\bar{y}_{GREG}$ could be used.

Though asymptotically valid, it is not attractive to use a variance estimator of a GREG to estimate the variance of the PEMLE. Instead, one might apply resampling variance estimators such as the jackknife, bootstrap and balanced repeated replications (see Shao and Wu (1989) and (1992), Chen and Qin (1993), Shao (1994)) directly to $\hat{\bar{Y}}_N$, recalculating the $\hat{p}_{hij}$ for each resample. These may perform better for finite samples since they are applied directly to $\hat{\bar{Y}}_N$ and not to the GREG which approximates it. As an example, we consider the jackknife for stratified srswor. Let $\hat{\theta} = \hat{\bar{Y}}_N$ (PEMLE) with vector valued auxiliary variable and define $v_J = \sum_{k=1}^{L} (1 - f_k) n_k^{-1} (n_k - 1) \sum_{j \in s_k} (\hat{\theta} - \hat{\theta}_{-kj})^2$, where $f_k = n_k / N_k$ and $\hat{\theta}_{-kj}$ is $\hat{\theta}$ recalculated with the jth sample unit from stratum k removed, i.e., the usual

delete-1 jackknife variance estimator (note that the results hold for sufficiently smooth functions of $\hat{\bar{Y}}_N$). Using similar arguments to Chen and Qin (1993), Appendix 5 proves the consistency. This will extend quite naturally to stratified multi-stage sampling, if we treat the clusters as being sampled with replacement. It is also clear from Theorem 1 that any central limit results available for GREG's apply to the PEMLE.

7. Simulation

To study the properties of the proposed PEMLE relative to the GREG and the ORE, we conducted a limited simulation study. For this purpose, we created various stratified finite populations. Each population consisted of $L = 4$ strata with stratum sizes $N_h = 8000 - 300h$ and stratum sample sizes $n_h = 100 - 9h$ for $h = 1, 2, 3, 4$. For the ith unit within the hth stratum, the characteristics x_{hi} were generated by adding $h/2$ to a χ^2_{2h} variate and the y_{hi} were generated using the model

$$y_{hi} = \alpha_h + \beta_h x_{hi} + \gamma_h x_{hi}^2 + \xi_h x_{hi}^a \epsilon_{hi} \tag{13}$$

for specific values of α_h, β_h, γ_h, a and ξ_h, where ϵ_{hi} are random variables, independent and identically distributed over i, from either a chi-square distribution with b_h degrees of freedom, $\chi^2_{b_h}$, or a standard normal distribution, $N(0, 1)$.

For each parameter combination, we first generated the stratified finite population of values $\{x_{hi}, y_{hi}\}$ using model (13). The six parameter combinations used to generate six finite populations are given in Table 1. For each of 1-6 in Table 1, the stratified finite population was created and $B = 1000$ independent stratified srswor samples were drawn as were $B = 1000$ stratified pps with replacement (ppswr) samples with probabilities proportional to x. The simulation mean square errors of the three estimators were then calculated as $MSE_j = \sum_{b=1}^{B} \{\hat{\bar{Y}}_{Nj}^{(b)} - \bar{Y}_N\}^2/B$, where $\hat{\bar{Y}}_{Nj}^{(b)}$ is the value of $\hat{\bar{Y}}_{Nj}$ for the bth simulation run and $j = 1, 2, 3$ refer to the PEMLE, the GREG and the ORE, respectively. The random generations were done using the NAG fortran library functions.

The choice of model and parameter settings was somewhat artificial but some factors were taken into consideration when selecting them. The theoretical development in Section 5 suggests the PEMLE and the ORE should be more efficient than the GREG under stratified srswor because they make better use of the strata size information. The strata size information is most useful when the between strata variation is larger than the within stratum variation. (Note, in the extreme case of zero within stratum variation, the PEMLE and ORE have zero mean square error. This is not true for the GREG.) Populations 1-6 were chosen

to have this property with Population 6 having the smallest ratio of between to within strata variation.

Table 1. Parameter settings for generated finite populations

h	α_h	β_h	γ_h	ξ_h	a	ϵ_{hi}
			Population 1			
1	2	0	0	0.2	-0.5	χ^2_3
2	6	0	0	0.2	-0.5	χ^2_4
3	10	0	0	0.2	-0.5	χ^2_5
4	14	0	0	0.2	-0.5	χ^2_6
			Population 2			
1	2	0.5	0	0.2	-0.5	χ^2_3
2	6	1.0	0	0.2	-0.5	χ^2_4
3	10	-0.5	0	0.2	-0.5	χ^2_5
4	14	-1.0	0	0.2	-0.5	χ^2_6
			Population 3			
1	2	0.5	0	0.1	0.5	χ^2_1
2	6	1.0	0	0.1	0.5	χ^2_2
3	10	-0.5	0	0.2	0.5	χ^2_3
4	14	-1.0	0	0.1	0.5	χ^2_4
			Population 4			
1	2	0.5	0	0.1	0.5	$N(0,1)$
2	6	1.0	0	0.2	0.5	$N(0,1)$
3	10	-0.5	0	0.2	0.5	$N(0,1)$
4	14	-1.0	0	0.2	0.5	$N(0,1)$
			Population 5			
1	2	0.5	0.05	0.1	0.5	$N(0,1)$
2	6	1.0	-0.05	0.2	0.5	$N(0,1)$
3	10	-0.5	0.05	0.2	0.5	$N(0,1)$
4	14	-1.0	-0.05	0.2	0.5	$N(0,1)$
			Population 6			
1	4	0.5	0.05	0.1	0.5	$N(0,1)$
2	6	1.0	-0.05	0.2	0.5	$N(0,1)$
3	8	-0.5	0.05	0.2	0.5	$N(0,1)$
4	10	-1.0	-0.05	0.2	0.5	$N(0,1)$

Another consideration in selecting the parameter settings was the strength and form of the resulting dependence of y on x. In Population 1, y depends mildly on x and only through the variance. In Populations 2, 3 and 4, y and x are linearly related, while in Populations 5 and 6 a quadratic term is added to create departures from linearity. We also considered different error structures using both chi-square (skewed) and normal (symmetric) errors and variances proportional to x and x^{-1}.

Table 2. Comparing MSE's of the PEMLE, GREG and ORE under Stratified SRSWOR

Population	MSE(PEMLE)/MSE(GREG)	MSE(PEMLE)/$MSE(ORE)$
1	0.01	1.00
2	0.53	1.01
3	0.52	1.01
4	0.52	1.01
5	0.66	1.07
6	0.89	1.02

Table 3. Comparing MSE's of the PEMLE, GREG and ORE under Stratified PPSWR

Population	MSE(PEMLE)/MSE(GREG)	MSE(PEMLE)/$MSE(ORE)$
1	0.03	0.05
2	0.40	0.41
3	0.37	0.38
4	0.40	0.41
5	0.50	0.56
6	1.19	1.18

Tables 2 and 3 report the ratios of MSE(PEMLE) to MSE(GREG) and MSE(ORE) for the six populations under stratified srswor and under stratified ppswr sampling with probability proportional to x, respectively. For stratified srswor, in all six populations the PEMLE and the ORE perform similarly and significantly outperform the GREG in terms of MSE. For stratified pps, the PEMLE significantly outperforms both the GREG and the ORE in populations 1-5 while performing slightly worse in population 6.

As was discussed earlier, the probability that there is no solution for the empirical likelihood method converges to zero as sample size goes to infinity. In our simulation, we did not find any cases when the solutions did not exist. If this does occur, the practitioner may want to use the ORE or the GREG. The above simulation may then shed some light on how to make such a choice.

8. Some Concluding Remarks

A pseudo empirical likelihood approach to the use of auxiliary information in complex surveys was introduced and shown to be asymptotically equivalent to a GREG when making use of known population mean of some auxiliary variables in estimating the population mean of a characteristic of interest. The method allows the inclusion of more complex auxiliary information to advantage, as was demonstrated in the simple case of stratified sampling. In principle, one could

include very complex auxiliary information into the estimation through the use of this method. A simple example is when the median of the x's is known.

A more fundamental question (raised by one referee) is what one should do when all the x-values in the population are known. Most commonly used methods, though appearing to incorporate the individual x_i's from the entire population, result in estimators which essentially adjust to the population mean of the x's. An exception is the method proposed by Pfeffermann and Krieger (1991). Though their context is slightly more complicated, their estimator essentially incorporates the x information by partitioning the population units into groups and adjusting to the population group means of the x's. The pseudo empirical likelihood could be extended in a similar fashion. Of course, this raises many interesting theoretical and practical questions for future investigation.

Appendix 1. Proof of Theorem 1

We will assume that the solution to the pseudo empirical likelihood exists in probability. Under the conditions of Theorem 1, its proof is very simple. From

$$0 = \sum_{i\in s} w_i u_i/(1+\lambda u_i) = \sum_{i\in s} w_i u_i - \lambda \sum_{i\in s} w_i u_i^2/(1+\lambda u_i),$$

we conclude that $\lambda > 0$ when $\sum_{i\in s} w_i u_i > 0$. In this case,

$$\frac{\lambda}{1+\lambda u^*} \le \frac{\sum_{i\in s} w_i u_i}{\sum_{i\in s} w_i u_i^2} = O_p(n^{-1/2}),$$

where $u^* = \max\{|u_i| : i \in s\} = o_p(n^{1/2})$ and the last equality is from the second assumption. Hence, we must have

$$\lambda = \frac{\sum_{i\in s} w_i u_i}{\sum_{i\in s} w_i u_i^2} + o_p(n^{-1/2}).$$

The case when $\sum_{i\in s} w_i u_i < 0$ can be proved in a similar fashion. The expansion for $\hat{\bar{Y}}_N$ is then straightforward.

Appendix 2. Verifying Conditions for Theorem 1 in Three Common Designs

A2.1. PPS sampling with replacement

Let a_i be some known measure of size attached to the units and which is positively correlated with y_i. Suppose we sample n units with replacement with the probability of selecting the ith unit equal to $\alpha_i = a_i/A$, where $A = \sum_{i=1}^N a_i$. Letting $d_i = 1/(n\alpha_i)$ and $w_i(s) = d_i/\sum_{i\in s} d_i$, we can apply the pseudo empirical

likelihood method of Section 3. With no auxiliary information the PEMLE of $\bar{Y}_N$ will be $\hat{Y}_{ppz}/\sum_{i\in s} d_i$, a ratio estimator, where $\hat{Y}_{ppz} = \sum_{i\in s} d_i y_i$ (see Cochran (1977), p.252).

If $u_i = x_i - \bar{X}_N$ is known, to apply Theorem 1 we need Conditions 1 and 2. If we assume some conditions on the moments of the u_i's like $\sum_{i=1}^N \alpha_i u_i^4 = O(1)$, we have $EU^4 = \sum_{i=1}^N \alpha_i u_i^4 = O(1)$ where U denotes the u-values of the first sampled unit. So

$$P[U \geq n^{1/2}(\log n)^{-1/2}] \leq (\log n)^2 n^{-2} EU^4 = O((\log n)^2 n^{-2}).$$

Therefore,

$$P(u^* \geq n^{1/2}(\log n)^{-1/2}) \leq nP(U \geq n^{1/2}(\log n)^{-1/2}) = O((\log n)^2 n^{-1}),$$

and $u^* = o_p(n^{1/2})$. That is, the first condition of Theorem 1 is satisfied.

Routine calculations (see Cochran (1977), Theorem 9A.3), the conditions $\alpha_i \geq cN^{-1}$ for all i (i.e. none of the selection probabilities are too small), and $N^{-1}\sum_{i=1}^N u_i^2 \geq c > 0$ clearly imply the second condition of Theorem 1.

A2.2. The Rao-Hartley-Cochran method of PPS sampling without replacement

Suppose we have some known measure of size, a_i, $i = 1, \ldots, N$, and we wish to sample n units without replacement with the probability of selecting the ith unit approximately proportional to a_i. The Rao, Hartley and Cochran (1962) method (see Cochran (1977)) first partitions the population into n random groups of units with sizes $N_1, \ldots, N_n$. Then one unit is selected from each group. If A_g is the total measure of size of group g, the ith unit in group g is given selection probability a_i/A_g. The estimate of the population total of characteristic y used is then $\hat{Y}_{RHC} = \sum_{g=1}^n \frac{A_g}{a_g} y_g$, where y_g and a_g refer to the unit drawn from group g.

Let $d_g(s) = A_g/a_g$ and $w_g(s) = d_g(s)/\sum_{g\in s} d_g(s)$. Without auxiliary information, the pseudo empirical likelihood methodology of Section 3 gives the PEMLE for $\bar{Y}_N$ as $\hat{\bar{Y}}_N = \sum_{g\in s} w_g(s) y_g = \sum_{g\in s} d_g(s) y_g / \sum_{g\in s} d_g(s)$, a ratio estimator of $\bar{Y}_N$.

To apply Theorem 1 when the $u_i = x_i - \bar{X}_N$ are known, assume $N^{-1}\sum_{i=1}^N u_i^4 = O(1)$, $N^{-1}\sum_{i=1}^N u_i^2 \geq c$ and $\max\{a_i/a_j\} \leq cn^{1/2}$, for some absolute constant c. Let U_1 be the u-value of the sampled unit from the first random group A_1. Then for any i,

$$P(U_1 = u_i) = E[\frac{a_i I(i \in A_1)}{\sum_{i\in A_1} a_j}] \leq cn^{1/2} N^{-1}.$$

Hence $E(U_1^4) = O(n^{1/2})$. From the symmetry of sampled units, we get

$$P(u^* \geq n^{1/2}(\log n)^{-1}) \leq nP(U \geq n^{1/2}(\log n)^{-1}) \leq n^{-1}(\log n)^4 EU_1^4 = o(1).$$

That is, $u^* = o_p(n^{1/2})$ and the first condition of Theorem 1 holds.

As before, the second condition of Theorem 1 can be verified by routine calculations.

A2.3. Cluster sampling

Suppose we have N clusters and M_i is the number of elements in the ith cluster, where the clusters are sampled with probability proportional to size and with replacement. Suppose we take the sample, denoted s, of n clusters and completely enumerate them. If we assume that cluster i consists of M_i distinct Z_{ij}, themselves independently distributed from superpopulations F_i, independent for $i = 1, \ldots, N$, then the pseudo empirical log-likelihood will be $\sum_{i\in s}^n \sum_{j=1}^{M_i} d_{ij} \log(p_{ij})$, where $d_{ij} = d_i = 1/(n\alpha_i)$, $\alpha_i = M_i/M_0$ and $M_0 = \sum_i M_i$. That is, the pseudo empirical likelihood method of Section 3 can be applied by replacing i with ij throughout and letting $w_{ij}(s) = w_i(s) = d_i/(\sum_{i=1}^n d_i M_i)$ since $\sum_{i=1}^n \sum_{j=1}^{M_i} d_{ij} = \sum_{i=1}^n d_i M_i$. In this case, $\theta = \theta(F_N)$ is estimated by its PEMLE $\hat{\theta} = \theta(\hat{F}_n)$, where $\hat{F}_n = \sum_{i\in s}^n \sum_{j=1}^{M_i} \hat{p}_{ij}\delta_{y_{ij}}$ in obvious notation. Note that, if there is no auxiliary information, the PEMLE of $\bar{Y}_N$ will be $\hat{\bar{Y}}_N = \sum_{i\in s} d_i Y_i / \sum_{i\in s} d_i M_i$. In the common situation when all M_i's are bounded as $n \to \infty$, Condition 1 of Theorem 1 holds if we assume

$$N^{-1} \sum_{i=1}^N \sum_{j=1}^{M_i} u^4(x_{ij}) = O(1); \; N^{-1} \sum_{i=1}^N \sum_{j=1}^{M_i} u^2(x_{ij}) \geq c > 0.$$

To see this, let $u_i^* = \max_j\{|u(x_{ij})|\}$ and I be the cluster index obtained in the first draw. Clearly, $P(I = i) = \alpha_i$. Thus,

$$E(u_I^*)^4 = \sum_{i=1}^N \alpha_i u_i^* \leq \sum_{i=1}^N \sum_{j=1}^{M_i} \alpha_i u^4(x_{ij}) \leq \frac{\max\{M_i\}}{N} \sum_{i=1}^N \sum_{j=1}^{M_i} u^4(x_{ij}) = O(1).$$

Obviously, u^* given in Condition 1 of Theorem 1 is the largest observation of n independent and identically distributed u_I. Therefore,

$$P(u^* > n^{1/2}(\log n)^{-1}) \leq nP(u_I^* > n^{1/2}(\log n)^{-1}) \leq n^{-1}(\log n)^4 E(u_I^*)^4 = o(1).$$

That is, $u^* = o_p(n^{1/2})$.

Again, Condition 2 of Theorem 1 can be verified through straightforward calculations.

Appendix 3. Numerical Method for Stratified Sampling

Let $\tilde{x}_h$ be a group of numbers such that $\sum W_h \tilde{x}_h = \bar{X}_N$. Hence, the maximum of

$$\sum_{h=1}^{L} \sum_{i \in s_h} d_{hi} \log p_{hi} \tag{A.1}$$

subject to restrictions $\sum_{i \in s_h} p_{hi} = W_h$, $\sum_{i \in s_h} p_{hi} x_{hi} = W_h \tilde{x}_h$, $h = 1, \ldots, L$, is no larger than the original maximum of (7) under restriction (9). However, the maximum with these new restrictions equals the original maximum of (7) under restriction (9) for a specific group of $\tilde{x}_h$ values.

The maximum of (A.1) with the new restrictions can be obtained by using the ordinary Lagrange multiplier method. The solution is $p_{hi} = W_h d_{hi} / \{d_h + \lambda_h (x_{hi} - \tilde{x}_h)\}$, with λ_h satisfying

$$\sum_{i \in s_h} \frac{d_{hi}(x_{hi} - \tilde{x}_h)}{d_h + \lambda_h (x_{hi} - \tilde{x}_h)} = 0. \tag{A.2}$$

Clearly, the maximum of the original likelihood (7) equals

$$-\sum \sum_i d_{hi} \log[d_h + \lambda_h (x_{hi} - \tilde{x}_h)] + \sum_h \sum_i d_{hi} [\log(d_{hi}) + \log(W_h)]$$

with the best choice of feasible values of $\tilde{x}_h$. Hence, the problem reduces to maximizing

$$-\sum \sum_i d_{hi} \log[d_h + \lambda_h (x_{hi} - \tilde{x}_h)]$$

with respect to $\tilde{x}_h$ under the restriction $\sum_{h=1}^{L} W_h \tilde{x}_h = \bar{X}_N$. Note that, in this problem, λ_h is a function of $\tilde{x}_h$ defined by (A.2).

Using the Lagrange multiplier method, we get the function

$$l(\tilde{x}_1, \ldots, \tilde{x}_L, t) = -\sum \sum_i \log d_{hi} [d_h + \lambda_h (x_{hi} - \tilde{x}_h)] - t(\sum_{h=1}^{L} W_h \tilde{x}_h - \bar{X}_N).$$

Taking derivatives with respect to $\tilde{x}_h$ and setting to zero, we get

$$-\sum_{i \in s_h} \frac{d_{hi}[\lambda_h'(x_{hi} - \tilde{x}_h) - \lambda_h]}{d_h + \lambda_h (x_{hi} - \tilde{x}_h)} - tW_h = -\lambda_h - tW_h = 0,$$

where $\lambda_h' = \partial \lambda_h / \partial \tilde{x}_h$. Hence, we obtain $\lambda_h = tW_h$ and

$$\sum_{i \in s_h} \frac{d_{hi}(x_{hi} - \tilde{x}_h)}{d_h + tW_h (x_{hi} - \tilde{x}_h)} = 0 \tag{A.3}$$

for $h = 1, \ldots, L$. The other equation is $\sum_{h=1}^{L} W_h \tilde{x}_h = \bar{X}_N$. For each given t, $\tilde{x}_h$ takes a different value and thus we denote it as $\tilde{x}_h(t)$. It can be obtained from (A.3) easily. In addition, it is simple to show that $\sum_{h=1}^{L} W_h \tilde{x}_h(t)$ is a monotone function of t. Hence, numerically, we need only increase or decrease the size of t to determine the existence of the solution and the uniqueness is a simple consequence of the monotonicity.

Appendix 4. Proof of Corollary 1

Before we prove Corollary 1, we first state and prove the following lemma.

Lemma 1. *Under the conditions of Corollary* 1*, a solution to the pseudo empirical likelihood equations exists with probability tending to one as the sample size goes to infinity.*

Proof of Lemma 1. First, assume $\max\{W_h\} \to 0$. Let $x_{h1} \le x_{h2}$ be any two randomly selected observations in stratum h. Note that a solution exists if

$$\sum_{h=1}^{L} W_h x_{h2} \ge \bar{X}_N \quad \text{and} \quad \sum_{h=1}^{L} W_h x_{h1} \le \bar{X}_N.$$

This ensures that $\bar{X}_N$ falls in the convex hull of x values in the sample. From the moment conditions in Corollary 1, it can be shown $E[\sum_{h=1}^{L} W_h x_{h2}] - \bar{X}_N \ge c > 0$ for some c independent of the index ν (Chen and Sitter (1996)). Hence, $\sum_{h=1}^{L} W_h x_{h2} > \bar{X}_N$ with probability tending to 1 since its variance goes to zero as a result of $\mathrm{Var}\,(x_{h2}) \le 2\sigma_h^2$. Similarly, $\sum_{h=1}^{L} W_h x_{h1} < \bar{X}_N$ with probability tending to 1. That is, the solution of the pseudo empirical likelihood equations exists with probability approaching one when $\max\{W_h\} \to 0$.

Next, assume $m = \min_h\{n_h\}$ goes to infinity. Let x_{hi}, $i = 1, \ldots, m$, be the first m observations from the hth stratum. The moment conditions will then imply

$$P(\sum_{h=1}^{L} W_h x_{hi} > \bar{X}_N) \ge c > 0$$

for some c for all $i = 1, \ldots, m$. Therefore, if x_h^+ is the largest observation from the hth stratum, we have

$$P(\sum_{h=1}^{L} W_h x_h^+ \le \bar{X}_N) \le (1 - c)^m$$

which goes to zero as $m \to \infty$. Similarly for the smallest observation x_h^- from the hth stratum

$$P(\sum_{h=1}^{L} W_h x_h^- \ge \bar{X}_N) \to 0.$$

Together, these imply the existence of a solution with probability approaching 1.

For any other case, let $L_1 = \{h : W_h \leq n^{-1/2}\}$ and L_2 be its complement. Strata in L_1 satisfy the condition $\max\{W_h\} \to 0$ and strata in L_2 satisfy the condition $\min\{n_h\} \to \infty$, by condition (ii) of the theorem. However, the moment conditions have to be revised. Note that under the moment conditions of the theorem $N^{-1}\sum_{h\in L_j}\sum_i |x_{hi}|^3 = O(1)$ for both $j = 1, 2$, and $0 < c_1/2 \leq \sum_{h\in L_j} W_h\sigma_h^2$ is true for either $j = 1$ or $j = 2$ (possibly both). If it is true for $j = 1$, we can show $\sum_{h\in L_1} W_h x_{h2} + \sum_{h\in L_2} W_h\bar{x}_h - \bar{X}_N$ has a mean which is larger than some $c > 0$ and a variance which goes to zero. Therefore, it is larger than $\bar{X}_N$ with probability approaching one. Similarly, $\sum_{h\in L_1} W_h x_{h1} + \sum_{h\in L_2} W_h\bar{x}_h - \bar{X}_N$ has a mean which is smaller than some $-c < 0$ and a variance which goes to zero. Therefore, it is smaller than $\bar{X}_N$ with probability approaching one. If it is true for $j = 2$, we can show $\sum_{h\in L_1} W_h\bar{x}_h + \sum_{h\in L_2} W_h x_h^+$ is larger than $\bar{X}_N$ with probability approaching one. Similarly replacing x_h^+ by x_h^-, we can show it is smaller than $\bar{X}_N$ with probability approaching one. Hence, a solution exists with probability approaching one in general.

Proof of Corollary 1. When the solution exists, note that for any $h = 1, \ldots, L$, we have

$$\sum_{i\in s_h}(x_{hi} - \tilde{x}_h) = t\sum_{i\in s_h}\frac{(x_{hi} - \tilde{x}_h)^2}{1 + t(x_{hi} - \tilde{x}_h)},$$

and hence

$$\sum_{h=1}^{L}\frac{W_h}{n_h}\sum_{i\in s_h}(x_{hi} - \tilde{x}_h) = t\sum_{h=1}^{L}\frac{W_h}{n_h}\sum_{i\in s_h}\frac{(x_{hi} - \tilde{x}_h)^2}{1 + t(x_{hi} - \tilde{x}_h)}.$$

Similar to the proof of Theorem 1, we have

$$|t| \leq (1 + |t|u^*)\frac{|\sum_{h=1}^{L}\frac{W_h}{n_h}\sum_{i\in s_h}(x_{hi} - \tilde{x}_h)|}{\sum_{h=1}^{L}\frac{W_h}{n_h}\sum_{i\in s_h}(x_{hi} - \tilde{x}_h)^2},$$

where $u^* = \max\{x_{hi} : i \in s_h\}$, since

$$0 = \sum_{h=1}^{L} W_h\tilde{x}_h - \bar{X}_N = [\sum_{h=1}^{L} W_h(\tilde{x}_h - \bar{x}_h)] + \sum_{h=1}^{L} W_h(\bar{x}_h - \bar{X}_N).$$

Note that the second term has mean zero and variance $\sum_{h=1}^{L} n^{-1}W_h^2\sigma_h^2 = O(\max\{n_h^{-1}W_h\}) = O(n^{-1})$ by assumption. Thus the second term is of order $n^{-1/2}$ and, consequently, the first term is of the same order. Applying this to the inequality for $|t|$, we find $t = O_p(n^{-1/2})$. Now, with the simple random sampling plan,

the third moment condition implies $u^* = o_p(n^{1/2})$. Hence, $t(x_{hi} - \tilde{x}_h) = o_p(1)$ uniformly over sampled units. We therefore get

$$t = \frac{\sum_{h=1}^{L} W_h(\bar{x}_h - \tilde{x}_h)}{\sum_{h=1}^{L} W_h n_h^{-1} \sum_{i\in s_h}(x_{hi} - \tilde{x}_h)^2} + o_p(n^{-1/2}).$$

With this expansion for t and the relation $\hat{p}_{hi} = \{n_h[1 + t(x_{hi} - \tilde{x}_h)]\}^{-1}$, it is straightforward to expand $\hat{\bar{Y}}_N = \sum_{h=1}^{L} W_h \sum_{i\in s_h} \hat{p}_{hi} y_{hi}$ to obtain the required result.

Appendix 5. Proof of Consistency of the Jackknife Variance Estimator

The Lagrange multiplier with the jth unit from the kth strata removed, λ_{-kj}, solves

$$\sum_{h=1}^{L} \frac{W_h}{n_h} \sum_{i\in s_h} \frac{u_{hi}}{1+\lambda_{-kj}^T u_{hi}} + \frac{W_k}{n_k(n_k-1)} \sum_{i\in s_k} \frac{u_{ki}}{1+\lambda_{-kj}^T u_{ki}} - \Big(\frac{W_k}{n_k-1}\Big)\Big(\frac{u_{kj}}{1+\lambda_{-kj}^T u_{kj}}\Big) = 0,$$

and so

$$\sum_{h=1}^{L} \frac{W_h}{n_h} \sum_{i\in s_h} \frac{u_{hi}u_{hi}^T}{(1+\lambda_{-kj}^T u_{hi})(1+\lambda^T u_{hi})}(\lambda_{-kj} - \lambda)$$
$$= \frac{W_k}{n_k(n_k-1)} \sum_{i\in s_k} \frac{u_{ki}}{1+\lambda_{-kj}^T u_{ki}} - \left(\frac{W_k}{n_k-1}\right)\left(\frac{u_{kj}}{1+\lambda_{-kj}^T u_{kj}}\right). \quad \text{(A.4)}$$

Similarly,

$$\hat{\theta} - \hat{\theta}_{-kj} = \sum_{h=1}^{L} \frac{W_h}{n_h} \sum_{i\in s_h} \frac{y_{hi}u_{hi}^T}{(1+\lambda_{-kj}^T u_{hi})(1+\lambda^T u_{hi})}(\lambda_{-kj} - \lambda)$$
$$- \frac{W_k}{n_k(n_k-1)} \sum_{i\in s_k} \frac{y_{ki}}{1+\lambda_{-kj}^T u_{ki}} + \left(\frac{W_k}{n_k-1}\right)\left(\frac{y_{kj}}{1+\lambda_{-kj}^T u_{kj}}\right). \quad \text{(A.5)}$$

Note that $\lambda = O_p(n^{-1/2})$ by condition 1 of Theorem 1, and similarly it can be shown that $\lambda_{-kj} = O_p(n^{-1/2})$ uniformly (see the proof of Theorem 1 in Appendix 1).

Letting $A_{uu} = \sum_{h=1}^{L} W_h \sum_{i\in s_h} u_{hi}u_{hi}^T/n_h$ and $A_{uy} = \sum_{h=1}^{L} W_h \sum_{i\in s_h} y_{hi}u_{hi}^T/n_h$, we get

$$\sum_{h=1}^{L} \frac{W_h}{n_h} \sum_{i\in s_h} \frac{u_{hi}u_{hi}^T}{(1+\lambda_{-kj}^T u_{hi})(1+\lambda^T u_{hi})} = A_{uu}(1 + o_p(1)) \quad \text{(A.6)}$$

and

$$\sum_{h=1}^{L} \frac{W_h}{n_h} \sum_{i\in s_h} \frac{y_{hi}u_{hi}^T}{(1+\lambda_{-kj}^T u_{hi})(1+\lambda^T u_{hi})} = A_{uy}(1 + o_p(1)) \quad \text{(A.7)}$$

uniformly in kj. It is not difficult to show that replacing $\sum_{i\in s_k} y_{ki}/(1+\lambda_{-kj}^T u_{kj})$ by $\sum_{i\in s_k} y_{hi}$ and $y_{kj}/(1+\lambda_{-kj}^T u_{kj})$ by y_{kj} in the expression of $\hat{\theta}-\hat{\theta}_{-kj}$ has negligible effect on the jackknife variance estimator.

By ignoring these higher order terms and using (A.4) and (A.6) we get $A_{uu}(\lambda_{-kj}-\lambda) \doteq -(n_k-1)^{-1}W_k(u_{kj}-\bar{u}_k)$, where $\bar{u}_k=\sum_{i\in s_k} u_{ki}/n_k$, and thus by (A.5) and (A.7), $\hat{\theta}-\hat{\theta}_{-kj} \doteq (n_k-1)^{-1}W_k[(y_{kj}-\bar{y}_k)-A_{uy}A_{uu}^{-1}(u_{kj}-\bar{u}_k)]$, where $\bar{y}_k=\sum_{j\in s_k} y_{kj}/n_k$.

Thus

$$v_J \doteq \sum_{k=1}^{L}(1-f_k)^2\frac{W_k^2}{n_k(n_k-1)}\sum_{j\in s_k}[(y_{kj}-\bar{y}_k)-A_{uy}A_{uu}^{-1}(u_{kj}-\bar{u}_k)]^2$$

which implies the desired result.

References

Basu, D. (1971). *Foundations of Statistical Inference, A Symposium* (Edited by V. P Godambe and D. A Sprott). Holt Rinehart and Winston of Canada, Limited. Toronto.

Bethlehem, J. G. and Keller, W. J. (1987). Linear weighting of sample survey data. *J. Off. Statist.* **3**, 141-153.

Binder, D. (1983). On the variances of asymptotically normal estimators from complex surveys. *Internat. Statist. Rev.* **51**, 279-292.

Cassel, C., Sarndal, C. and Wretman, J. H. (1977). *Foundations of Inference in Survey Sampling.* Wiley, New York.

Chen, J. and Qin, J. (1993). Empirical likelihood estimation for finite populations and the effective usage of auxiliary information. *Biometrika* **80**, 107-116.

Chen, J. and Sitter, R. R. (1996). A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys. Technical Report, Department of Mathematics and Statistics, Simon Fraser University.

Cochran, W. G. (1977). *Sampling Techniques.* Third Edition. John Wiley, New York.

Deville, J. and Sarndal, C. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Assoc.* **87**, 376-382.

Godambe, V. P. and Thompson, M. E. (1986). Parameters of superpopulation and survey population: Their relationships and estimation. *Internat. Statist. Rev.* **54**, 127-138.

Hartley, H. O. and Rao, J. N. K. (1968). A new estimation theory for sample surveys. *Biometrika* **55**, 547-557.

Jagers, P. (1986). Post-stratification against bias in sampling. *Internat. Statist. Rev.* **54**, 159-167.

Owen, A. (1990). Empirical likelihood confidence regions. *Ann. Statist.* **18**, 90-120.

Pfeffermann, D. and Krieger, A. M. (1991). Poststratification using regression estimators when information on strata means and sizes is missing. *Biometrika* **78**, 409-419.

Rao, J. N. K. (1966). Alternative estimators in PPS sampling for multiple characteristics. *Sankhya Ser. A.* **28**, 47-60.

Rao, J. N. K. (1988). Variance estimation in sample surveys. *Handbook of Statistics* (Edited by P. K. Krishnaiah and C. R. Rao) **6**, 427-447. Elsevier, North-Holland, Amsterdam.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *J. Off. Statist.* **10**, 153-165.

Rao, J. N. K., Hartley, H. O. and Cochran, W. G. (1962). A simple procedure of unequal probability sampling without replacement. *J. Roy. Statist. Soc. Ser. B* **24**, 482-491.

Rao, J. N. K. and Wu, C. F. J. (1985). Inference from stratified samples: Second-order analysis of three methods for nonlinear statistics. *J. Amer. Statist. Assoc.* **80**, 620-630.

Sarndal, C. E. (1980). On π-inverse weighting versus best linear weighting in probability sampling. *Biometrika* **67**, 639-650.

Shao, J. (1994). L-statistics in complex survey problems. *Ann. Statist.* **22**, 946-967.

Shao, J. and Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *Ann. Statist.* **17**, 1176-1197.

Shao, J. and Wu, C. F. J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Ann. Statist.* **20**, 1571-1593.

Zhong, C. X. B. and Rao, J. N. K. (1996). Empirical likelihood of inference for finite populations with auxiliary information using stratified random sampling. Preprint.

Department of Statistics & Actuarial Science, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada.

E-mail: jhchen@math.uwaterloo.ca

Department of Mathematics and Statistics, Simon Fraser University, Burnaby, B. C. V5A 1S6, Canada.

E-mail: sitter@math.sfu.ca

(Received November 1997; accepted January 1999)