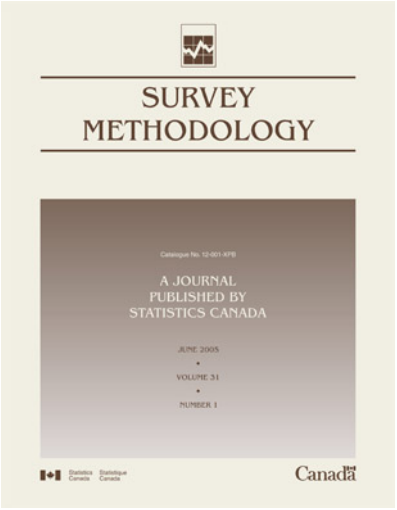
Catalogue no. 12-001-XIE

Survey Methodology

December 2005



Statistics Canada Statistique Canada

Canada

How to obtain more information

Specific inquiries about this product and related statistics or services should be directed to: Business Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6 (telephone: 1 800 263-1136).

For information on the wide range of data available from Statistics Canada, you can contact us by calling one of our toll-free numbers. You can also contact us by e-mail or by visiting our website.

National inquiries line	1 800 263-1136
National telecommunications device for the hearing impaired	1 800 363-7629
Depository Services Program inquiries	1 800 700-1033
Fax line for Depository Services Program	1 800 889-9734
E-mail inquiries	infostats@statcan.ca
Website	www.statcan.ca

Information to access the product

This product, catalogue no. 12-001-XIE, is available for free. To obtain a single issue, visit our website at www.statcan.ca and select Our Products and Services.

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner and in the official language of their choice. To this end, the Agency has developed standards of service that its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll free at 1 800 263-1136. The service standards are also published on www.statcan.ca under About Statistics Canada > Providing services to Canadians.

Statistics Canada
Business Survey Methods Division

Survey Methodology

December 2005

Published by authority of the Minister responsible for Statistics Canada

May 2006

Catalogue no. 12-001-XIE
ISSN 1492-0921

Frequency: semi-annual

Ottawa

Cette publication est disponible en français sur demande (nº 12-001-XIF au catalogue).

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued cooperation and goodwill.

Vol. 31, No. 2, pp. 239-243
Statistics Canada, Catalogue No. 12-001

Algorithms and R Codes for the Pseudo Empirical Likelihood Method in Survey Sampling

Changbao Wu [1]

Abstract

We present computational algorithms for the recently proposed pseudo empirical likelihood method for the analysis of complex survey data. Several key algorithms for computing the maximum pseudo empirical likelihood estimators and for constructing the pseudo empirical likelihood ratio confidence intervals are implemented using the popular statistical software R and S-PLUS. Major codes are written in the form of R/S-PLUS functions and therefore can directly be used for survey applications and/or simulation studies.

Key Words: Confidence interval; Bi-section algorithm; Empirical likelihood; Newton-Raphson procedure; Stratified sampling; Unequal probability sampling.

1. Introduction

One of the major challenges in applying advanced and often sophisticated statistical methods for real world surveys is the computational implementation of the method. Practical considerations often rule out the use of methods which are theoretically sound and attractive but are computationally formidable.

The empirical likelihood method first proposed by Owen (1988) is one of the major advances in statistics during the past fifteen years. In addition to its data driven and range respecting feature in estimation and testing, its nonparametric and discrete nature is particularly appealing for finite population problems. Indeed an early version of the method, the so-called scale-load estimators, was used in survey sampling by Hartley and Rao back in 1968. The more recent investigation of the method in survey sampling has resulted in a series of research papers and generated noticeable interests among survey statisticians to further explore the method. Wu and Rao (2004) contains a brief summary on the recent development of the pseudo empirical likelihood (PEL) method in survey sampling.

Progress on algorithmic development for the PEL method has also been made. A modified Newton-Raphson procedure for computing the maximum PEL estimators under non-stratified sampling was proposed by Chen, Sitter and Wu (2002). The procedure was further modified by Wu (2004a) to handle stratified sampling designs.

In this article we present computational algorithms for computing the maximum PEL estimators and for constructing the related PEL ratio confidence intervals for complex surveys under a unified framework, with particular interest in implementing those algorithms using R and S-PLUS. The software package R, a friendly programming environment and compatible to the popular commercial statistical software S-PLUS, is attracting more and more users from the statistical community. What is advantageous about using R is that it is available free for research use and the package may be easily downloaded from the web. It is hoped that this article will bridge the current gap between theoretical developments and practical applications of the PEL method and will generate more research activities in this direction to make fully practical use of the PEL method a reality.

The algorithm for computing the maximum PEL estimator under non-stratified sampling and some notes on its implementation in R/S-PLUS are presented in section 2. The algorithm of Wu (2004a) for stratified sampling is discussed in section 3. Construction of the PEL ratio confidence intervals involves profiling the pseudo empirical likelihood ratio statistic and is detailed in section 4. All R functions or sample codes are included in the Appendix. They can also be downloaded from the author's personal homepage http://www.stats.uwaterloo.ca/~ cbwu/paper. html. These functions and codes had been tested in the simulation study reported in Wu and Rao (2004) and were observed to perform very well.

2. Non-Stratified Sampling

Consider a finite population consisting of N identifiable units. Associated with the i^{th} unit are values of the study variable, y_i, and a vector of auxiliary variables, $\boldsymbol{x}_i$. The vector of population means $\overline{\boldsymbol{X}} = N^{-1}\sum_{i=1}^{N} \boldsymbol{x}_i$ is known. Let $\{(y_i, \boldsymbol{x}_i), i \in s\}$ be the sample data where s is the set of units selected using a complex survey design. Let $\pi_i = P(i \in s)$ be the inclusion probabilities and $d_i = 1/\pi_i$ be the design weights.

1. Changbao Wu, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada. E-mail: cbwu@uwaterloo.ca.

The pseudo empirical maximum likelihood estimator of the population mean $\bar{Y}=N^{-1}\sum_{i=1}^{N} y_i$ is computed as $\hat{\bar{Y}}_{\mathrm{PEL}}=\sum_{i\in s}\hat{p}_i\, y_i$ where the weights $\hat{p}_i$ are obtained by maximizing the pseudo empirical log likelihood function

$$l_{ns}(\boldsymbol{p})=n^{*}\sum_{i\in s} d_i^{*}\log(p_i) \qquad (2.1)$$

subject to the set of constraints

$$0<p_i<1,\ \sum_{i\in s}p_i=1 \ \text{ and } \ \sum_{i\in s}p_i\,\boldsymbol{x}_i=\bar{\boldsymbol{X}}. \qquad (2.2)$$

The original pseudo empirical likelihood function proposed by Chen and Sitter (1999) is $l(\boldsymbol{p})=\sum_{i\in s} d_i\,\log(p_i)$. The pseudo empirical likelihood function $l_{ns}(\boldsymbol{p})$ given by (2.1) was used by Wu and Rao (2004), where $d_i^{*}=d_i/\sum_{i\in s}\,d_i$ are the normalized design weights and n^{*} is the effective sample size. The point estimator $\hat{\bar{Y}}_{\mathrm{PEL}}=\sum_{i\in s}\hat{p}_i\,y_i$ remains the same for either version of the likelihood function. The rescaling used in $l_{ns}(\boldsymbol{p})$ facilitates the construction of the PEL ratio confidence intervals.

Using a standard Lagrange multiplier argument it can be shown that

$$\hat{p}_i=\frac{d_i^{*}}{1+\boldsymbol{\lambda}'(\boldsymbol{x}_i-\bar{\boldsymbol{X}})} \quad \text{for} \quad i\in s, \qquad (2.3)$$

where the vector-valued Lagrange multiplier, $\boldsymbol{\lambda}$, is the solution to

$$g_1(\boldsymbol{\lambda})=\sum_{i\in s}\frac{d_i^{*}(\boldsymbol{x}_i-\bar{\boldsymbol{X}})}{1+\boldsymbol{\lambda}'(\boldsymbol{x}_i-\bar{\boldsymbol{X}})}=0.$$

The major computational task here is to find the solution to $g_1(\boldsymbol{\lambda})=0$. This can be done using the modified Newton-Raphson procedure proposed by Chen *et al.* (2002). The modification involves checking at each updating stage that the constraint $1+\boldsymbol{\lambda}'(\boldsymbol{x}_i-\bar{\boldsymbol{X}})>0$ (*i.e.*, $p_i>0$) is always satisfied. Without loss of generality, we assume $\bar{\boldsymbol{X}}=0$ (if not, replace $\boldsymbol{x}_i$ by $\boldsymbol{x}_i-\bar{\boldsymbol{X}}$ throughout). The modified procedure is as follows.

Step 0: Let $\boldsymbol{\lambda}_0=\boldsymbol{0}$. Set $k=0$, $\gamma_0=1$ and $\varepsilon=10^{-8}$.

Step 1: Calculate $\boldsymbol{\Delta}_1(\boldsymbol{\lambda}_k)$ and $\boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)$ where

$$\boldsymbol{\Delta}_1(\boldsymbol{\lambda})=\sum_{i\in s} d_i^{*}\frac{\boldsymbol{x}_i}{1+\boldsymbol{\lambda}'\boldsymbol{x}_i}$$

and

$$\boldsymbol{\Delta}_2(\boldsymbol{\lambda})=\left\{-\sum_{i\in s} d_i^{*}\frac{\boldsymbol{x}_i\,\boldsymbol{x}_i'}{(1+\boldsymbol{\lambda}'\boldsymbol{x}_i)^2}\right\}^{-1}\boldsymbol{\Delta}_1(\boldsymbol{\lambda}).$$

If $\|\boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)\|<\varepsilon$, stop the algorithm and report $\boldsymbol{\lambda}_k$; otherwise go to Step 2.

Step 2: Calculate $\boldsymbol{\delta}_k=\gamma_k\,\boldsymbol{\Delta}_2(\boldsymbol{\lambda}_k)$. If $1+(\boldsymbol{\lambda}_k-\boldsymbol{\delta}_k)'\,\boldsymbol{x}_i\le 0$ for some i, let $\gamma_k=\gamma_k/2$ and repeat Step 2.

Step 3: Set $\boldsymbol{\lambda}_{k+1}=\boldsymbol{\lambda}_k-\boldsymbol{\delta}_k$, $k=k+1$ and $\gamma_{k+1}=(k+1)^{-1/2}$. Go to Step 1.

In the original algorithm presented by Chen *et al.* (2002), their step 2 also checks a related dual objective function. While this is necessary for the theoretical proof of convergence of the algorithm, it is not really required for practical applications.

The R function Lag2(u,ds,mu) can be used for finding the solution to $g_1(\boldsymbol{\lambda})=0$ when the vector of auxiliary variables $\boldsymbol{x}$ is of dimension m and $m\ge 2$. When $\boldsymbol{x}$ is univariate, an extremely simple and stable bi-section method to be described shortly should be used. Let n be the sample size. The three required arguments are the $n\times m$ data matrix u, the $n\times 1$ vector of design weights ds and the $m\times 1$ population mean vector mu. The output of the function Lag2(u,ds,mu) returns the value of $\boldsymbol{\lambda}$ which is the solution to $g_1(\boldsymbol{\lambda})=0$.

The function Lag2(u,ds,mu) will fail to provide a solution if (i) the mean vector $\bar{\boldsymbol{X}}$ is not an inner point of the convex hull formed by $\{\boldsymbol{x}_i,\ i\in s\}$, or (ii) the matrix $\sum_{i\in s} d_i\,\boldsymbol{x}_i\,\boldsymbol{x}_i'$ is not of full rank. In case (i) the pseudo empirical maximum likelihood estimator does not exist. This happens with probability approaching to zero as the sample size n goes to infinity; in case (ii) one may consider to remove some components of the $\boldsymbol{x}$ variables from the set of constraints (2.2) to eliminate the collinearity problem.

When the $\boldsymbol{x}$ variable is univariate, so is the involved Lagrange multiplier $\boldsymbol{\lambda}$. In this case we need to solve $g_2(\lambda)=\sum_{i\in s} d_i^{*}\,x_i/(1+\lambda x_i)=0$ for a scalar λ, assuming $\bar{X}=0$. A unique solution exists if and only if $\min\{x_i,\ i\in s\}<0<\max\{x_i,\ i\in s\}$. The solution, if exists, lies between $L=-1/\max\{x_i,\ i\in s\}$ and $U=-1/\min\{x_i,\ i\in s\}$. Noting that $g_2(\lambda)$ is a monotone decreasing function for $\lambda\in(L,\,U)$, the most efficient and reliable algorithm for solving $g_2(\boldsymbol{\lambda})=0$ is the bi-section method. The function Lag1(u,ds,mu) does exactly this, where the required arguments are u $=(x_1,\ldots,x_n)$, ds $=(d_1,\ldots,d_n)$ and mu $=\bar{\boldsymbol{X}}$. The output returns the solution to $g_2(\lambda)=0$.

The function Lag1(u,ds,mu) can be used in conjunction with the model-calibrated pseudo empirical likelihood (MCPEL) approach of Wu and Sitter (2001) to handle cases where the $\boldsymbol{x}$ variable is high dimensional. The MCPEL approach involves only a single dimension reduction variable derived from a multiple linear regression model and the related Lagrange multiplier problem is always of dimension one.

3. Stratified Sampling

Let $\{(y_{hi},\,\boldsymbol{x}_{hi}),\ i\in s_h,\ h=1,\ldots,H\}$ be the sample data from a stratified sampling design. Let $d_{hi}^{*}=d_{hi}/\sum_{i\in s_h} d_{hi}$ be the normalized design weights for stratum h, $h=1,\ldots,H$. The pseudo empirical likelihood function

under stratified sampling defined by Wu and Rao (2004) is given by

$$l_{st}(\boldsymbol{p}_1, ..., \boldsymbol{p}_H) = n^* \sum_{h=1}^{H} W_h \sum_{i \in s_h} d_{hi}^* \log(p_{hi}), \quad (3.1)$$

where $W_h = N_h / N$ are the stratum weights and n^* is the total effective sample size as defined in Wu and Rao (2004). The value of n^* is not required for point estimation but this scaling constant is needed for the construction of confidence intervals. Let $\overline{\boldsymbol{X}}$ be the known vector of population means for auxiliary variables. The maximum pseudo empirical likelihood estimator of the population mean $\overline{Y} = \sum_{h=1}^{H} W_h \overline{Y}_h$ is defined as $\hat{\overline{Y}}_{\text{PEL}} = \sum_{h=1}^{H} W_h \sum_{i \in s_h} \hat{p}_{hi} y_{hi}$ where the $\hat{p}_{hi}$ maximize $l_{st}(\boldsymbol{p}_1, ..., \boldsymbol{p}_H)$ subject to the set of constraints

$$p_{hi} > 0, \sum_{i \in s_h} p_{hi} = 1, \ h = 1, ..., H$$

and

$$\sum_h W_h \sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}. \quad (3.2)$$

The major computational difficulty under stratified sampling is caused by the fact that the subnormalization of weights (*i.e.*, $\sum_{i \in s_h} p_{hi} = 1$) occurs at the stratum level while the benchmark constraints (*i.e.*, $\sum_h W_h \sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}$) and the constrained maximization of the PEL function are taken at the the population level. The algorithm proposed by Wu (2004a) for computing the $\hat{p}_{hi}$ proceeds as follows: let $\boldsymbol{x}_{hi}$ be augmented to include the first $H-1$ stratum indicator variables and $\overline{\boldsymbol{X}}$ be augmented to include $(W_1, ..., W_{H-1})$ as its first $H-1$ components. In the case of no benchmark constraints involved, the augmented $\boldsymbol{x}$ variable will consist of the $H-1$ stratum indicator variables only and $\overline{\boldsymbol{X}} = (W_1, ..., W_{H-1})$. It follows that the set of constraints (3.2) is equivalent to

$$p_{hi} > 0, \sum_{h=1}^{H} W_h \sum_{i \in s_h} p_{hi} = 1$$

and

$$\sum_{h=1}^{H} W_h \sum_{i \in s_h} p_{hi} \boldsymbol{x}_{hi} = \overline{\boldsymbol{X}}, \quad (3.3)$$

where the $\boldsymbol{x}$ variable is now augmented. Let $\boldsymbol{u}_{hi} = \boldsymbol{x}_{hi} - \overline{\boldsymbol{X}}$. It is straightforward by using a standard Lagrange multiplier argument to show that

$$\hat{p}_{hi} = \frac{d_{hi}^*}{1 + \boldsymbol{\lambda}' \boldsymbol{u}_{hi}},$$

with the vector-valued $\boldsymbol{\lambda}$ being the solution to

$$g_3(\boldsymbol{\lambda}) = \sum_h W_h \sum_{i \in s_h} \frac{d_{hi}^* \boldsymbol{u}_{hi}}{1 + \boldsymbol{\lambda}' \boldsymbol{u}_{hi}} = 0.$$

The modified Newton-Raphson procedure of section 2 for solving $g_1(\boldsymbol{\lambda}) = \boldsymbol{0}$ can be used for solving $g_3(\boldsymbol{\lambda}) = \boldsymbol{0}$. The key computational step under stratified sampling designs is to prepare the data file into suitable format so that the R function Lag2(u,ds,mu) for non-stratified sampling can directly be called. Sample R codes for doing this are included in the Appendix.

4. Construction of PEL Ratio Confidence Intervals

While the computational algorithms for the maximum PEL estimator under non-stratified and stratified sampling designs are somewhat different, the search for the lower and the upper boundary of the pseudo empirical likelihood ratio confidence interval for $\overline{Y}$ involves the same type of profile analysis. Under non-stratified sampling designs, the $(1-\alpha)-$level PEL ratio confidence interval of $\overline{Y}$ is constructed as

$$\{\theta \mid r_{ns}(\theta) < \chi_1^2(\alpha)\}, \quad (4.1)$$

where $\chi_1^2(\alpha)$ is the $1-\alpha$ quantile from a χ^2 distribution with one degree of freedom. The pseudo empirical log likelihood ratio statistic $r_{ns}(\theta)$ is computed as

$$r_{ns}(\theta) = -2\{l_{ns}(\tilde{\boldsymbol{p}}) - l_{ns}(\hat{\boldsymbol{p}})\},$$

where the $\hat{\boldsymbol{p}}$ maximize $l_{ns}(\boldsymbol{p})$ subject to the set of "standard constraints" such as (2.2) and the $\tilde{\boldsymbol{p}}$ maximize $l_{ns}(\boldsymbol{p})$ subject to the "standard constraints" plus an additional one induced by the parameter of interest, $\overline{Y}$, *i.e.*

$$\sum_{i \in s} p_i y_i = \theta. \quad (4.2)$$

To compute $\tilde{\boldsymbol{p}}$ one needs to treat (4.2) as an additional component of the "standard constraints" for each fixed value of θ so that the maximization process is essential the same as before.

Let $(\hat{L}, \hat{U})$ be the interval given by (4.1). Our proposed bi-section method in searching for $\hat{L}$ and $\hat{U}$ is based on following observations:

i) The minimum value of $r_{ns}(\theta)$ is achieved at $\theta = \sum_{i \in s} \hat{p}_i y_i = \hat{\overline{Y}}_{\text{PEL}}$. In this case $\tilde{\boldsymbol{p}} = \hat{\boldsymbol{p}}$ and $r_{ns}(\theta) = 0$.

ii) The interval $(\hat{L}, \hat{U})$ is bounded by $(y_{(1)}, y_{(n)})$ where $y_{(1)} = \min\{y_i, i \in s\}$ and $y_{(n)} = \max\{y_i, i \in s\}$.

iii) The pseudo empirical likelihood ratio function $r_{ns}(\theta)$ is monotone decreasing for $\theta \in (y_{(1)}, \hat{\overline{Y}}_{\text{PEL}})$ and monotone increasing for $\theta \in (\hat{\overline{Y}}_{\text{PEL}}, y_{(n)})$.

Conclusion iii) can be reached by noting that $l_{ns}(\hat{\boldsymbol{p}})$ does not involve θ and $l_{ns}(\tilde{\boldsymbol{p}}) = n^* \sum_{i \in s} d_i^* \log(\tilde{p}_i)$ is typically a concave function of θ. It is also possible to show this by directly checking $dr_{ns}(\theta)/d\theta$. For instance, in the case of no auxiliary information involved, the "standard constraints" are $p_i > 0$ and $\sum_{i \in s} p_i = 1$. The $\hat{p}_i$ are given by d_i^* and $\hat{\overline{Y}}_{\text{PEL}} = \sum_{i \in s} d_i^* y_i$. The $\tilde{p}_i$ are computed as

$$\tilde{p}_i = \frac{d_i^*}{1+\lambda(y_i-\theta)}, \qquad (4.3)$$

where the λ is the solution to

$$\sum_{i\in s} \frac{d_i^*(y_i-\theta)}{1+\lambda(y_i-\theta)} = 0. \qquad (4.4)$$

Using (4.3) and (4.4), and noting that $\sum_{i\in s} d_i^* / (1+\lambda(y_i-\theta)) = 1$, it is straightforward to show that

$$\frac{d}{d\theta} r_{ns}(\theta) = 2n^* \sum_{i\in s} \frac{d_i^* \{(d\lambda/d\theta)(y_i-\theta)-\lambda\}}{1+\lambda(y_i-\theta)} = -2n^*\lambda.$$

By re-writing $d_i^*(y_i-\theta)$ as $d_i^*(y_i-\theta)\,[\{1+\lambda(y_i-\theta)\} - \lambda(y_i-\theta)]$ and after some re-grouping in (4.4) we get

$$\lambda \sum_{i\in s} \frac{d_i^*(y_i-\theta)^2}{1+\lambda(y_i-\theta)} = \sum_{i\in s} d_i^* y_i - \theta.$$

It follows that $dr_{ns}(\theta)/d\theta = -2n^*\lambda < 0$ if $\theta < \sum_{i\in s} d_i^* y_i = \hat{\bar{Y}}_{\mathrm{PEL}}$ and $dr_{ns}(\theta)/d\theta > 0$ otherwise.

Sample codes for finding $(\hat{L}, \hat{U})$ where no auxiliary variable is involved are included in the Appendix. In this case $\hat{p}_i = d_i^*$ and $\hat{\bar{Y}}_{\mathrm{PEL}} = \sum_{i\in s} d_i^* y_i = \hat{\bar{Y}}_H$ is the Hajek estimator for $\bar{Y}$. The profiling process involves finding λ for each chosen value of θ and evaluating the PEL ratio statistic $r_{ns}(\theta)$ against the cut-off value from the χ_1^2 distribution under the desired confidence level $1-\alpha$. With auxiliary information, one needs to modify the computation of $r_{ns}(\theta)$ for each fixed θ. The bi-section search algorithm for finding $\hat{L}$ and $\hat{U}$ remains the same.

The value of the effective sample size n^* is required for computing the PEL ratio statistic $r_{ns}(\theta)$. For non-stratified sampling designs it is computed as $n^* = \hat{S}_y^2 / \hat{V}(y)$ where

$$\hat{S}_y^2 = \frac{1}{N(N-1)} \sum_{i\in s} \sum_{j>i} \frac{(y_i - y_j)^2}{\pi_{ij}},$$

and

$$\hat{V}(y) = \frac{1}{N^2} \sum_{i\in s} \sum_{j>i} \frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \left(\frac{e_i}{\pi_i} - \frac{e_j}{\pi_j} \right)^2,$$

where $e_i = y_i - \hat{\bar{Y}}_{\mathrm{HT}}$ and $\hat{\bar{Y}}_{\mathrm{HT}} = N^{-1} \sum_{i\in s} d_i y_i$. See Wu and Rao (2004) for further detail. Computation of n^* involves the second order inclusion probabilities π_{ij} which may impose a real challenge if a πps sampling scheme is used. In the simulation study reported in Wu and Rao (2004), the Rao-Sampford πps sampling method was used. R functions for selecting a πps sample using this method as well as for computing the related second order inclusion probabilities can be found in Wu (2004b). Similar R functions are also available in an add-on R package called "pps", written by J. Gambino (2003), which can be downloaded from the R homepage http://cran.r-project.org/ by clicking the packages option.

Acknowledgements

This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The author thanks an associate editor for helpful comments which lead to improvement of the paper.

Appendix: R/S-PLUS Codes

A1. R Function for solving $g_1(\lambda) = 0$.

Let m be the number of auxiliary variables involved and $m \geq 2$. There are three required arguments in the function Lag2(u,ds,mu):

(1) u: the $n \times m$ data matrix with $\boldsymbol{x}_i$ as its i^{th} row, $i = 1, \ldots, n$.

(2) ds: the $n \times 1$ vector of design weights consisting of $d_1, \ldots, d_n$.

(3) mu: the $m \times 1$ population mean vector $\bar{\boldsymbol{X}}$.

The output of the function is the solution to $g_1(\lambda) = 0$.

```
Lag2<-function(u,ds,mu)
{
  n<-length(ds)
  u<-u-rep(1,n)%*%t(mu)
  M<-0*mu
  dif<-1
  tol<-1e-08
  while(dif>tol){
    D1<-0*mu
    DD<-D1%*%t(D1)
    for(i in 1:n){
      aa<-as.numeric(1+t(M)%*%u[i,])
      D1<-D1+ds[i]*u[i,]/aa
      DD<-DD-ds[i]*(u[i,]%*%t(u[i,]))/aa^2
            }
    D2<-solve(DD,D1,tol=1e-12)
    dif<-max(abs(D2))
    rule<-1
    while(rule>0){
      rule<-0
      if(min(1+t(M-D2)%*%t(u))<=0) rule<-rule+1
      if(rule>0) D2<-D2/2
            }
    M<-M-D2
  }
  return(M)
}
```

A2. R Function for solving $g_2(\lambda) = 0$.

When the x variable is univariate, the solution to $g_2(\lambda) = 0$ can be found through a simple and reliable bi-section method. The three required arguments for the function Lag1(u,ds,mu) are $\mathrm{u} = (x_1, \ldots, x_n)$, $\mathrm{ds} = (d_1, \ldots, d_n)$ and $\mathrm{mu} = \bar{\boldsymbol{X}}$. The output is the solution to $g_2(\lambda) = 0$.

```
Lag1<-function(u,ds,mu)
{
  L<--1/max(u-mu)
  R<--1/min(u-mu)
  dif<-1
  tol<-1e-08
  while(dif>tol){
    M<-(L+R)/2
    glam<-sum((ds*(u-mu))/(1+M*(u-mu)))
    if(glam>0) L<-M
    if(glam<0) R<-M
    dif<-abs(glam)
  }
  return(M)
}
```

A3. Sample code for stratified sampling.

We need to call the function Lag2(u,ds,mu) from nonstratified sampling. The key step is to prepare the data file into suitable format. Let

(1) n = $(n_1, ..., n_H)$ be the vector of stratum sample sizes.

(2) x be the data matrix with $\boldsymbol{x}_{hi}$ as row vectors, $i = 1, ..., n_h, h = 1, ..., H$.

(3) ds = $(d_{11}^*, ..., d_{1n_1}^*, ..., d_{H1}^*, ..., d_{Hn_H}^*)$, where d_{hi}^* are the normalized initial design weights for stratum h.

(4) X be the vector of known population means.

(5) W = $(W_1, ..., W_H)$ be the vector of stratum weights (*i.e.*, $W_h = N_h / N$).

The following sample codes show how the solution to $g_3(\boldsymbol{\lambda}) = \mathbf{0}$ is found (M from the second last line of the following code) and how the $\hat{p}_{hi}$'s are computed (phi from the last line).

```
###
nst<-sum(n)
k<-length(n)-1
ntot<-rep(0,k)
  ntot[1]<-n[1]
  for(j in 2:k) ntot[j]<-ntot[j-1]+n[j]
ist<-matrix(0,nst,k)
  ist[1:n[1],1]<-1
  for(j in 2:k) ist[(ntot[j-1]+1):ntot[j],j]<-1
uhi<-cbind(ist,x)
mu<-c(W[1:k],X)
whi<-rep(W[1],n[1])
  for(j in 2:(k+1)) whi<-c(whi,rep(W[j],n[j]))
dhi<-whi*ds
M<-Lag2(uhi,dhi,mu)
phi<-as.vector(ds/(1+(uhi-rep(1,nst)%*%t(mu))%*%M))
###
```

A4. Sample code for finding the PEL ratio confidence interval.

The search for the lower boundary (LB) and the upper boundary (UB) of the PEL ratio confidence interval needs to be carried out separately. The following codes show how this is done for the case of no auxiliary information. With auxiliary information, one needs to modify the computation of the involved pseudo empirical likelihood ratio statistic (elratio) accordingly. Let

(1) a = $1-\alpha$ be the confidence level of the desired interval.

(2) ys = $(y_1, ..., y_n)$ be the sample data.

(3) ds = $(d_1^*, ..., d_n^*)$ be the normalized design weights.

(4) YEL = $\sum_{i \in s} \hat{p}_i \, y_i$ (in this case $\hat{p}_i = d_i^*$).

(5) nss be the estimated effective sample size n^*.

```
###
tol<-1e-08
cut<-qchisq(a,1)
###
t1<-YEL
t2<-max(ys)
dif<-t2-t1
while(dif>tol){
    tau<-(t1+t2)/2
    M<-Lag1(ys,ds,tau)
    elratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
    if(elratio>cut) t2<-tau
    if(elratio<=cut) t1<-tau
    dif<-t2-t1
        }
UB<-(t1+t2)/2
###
t1<-YEL
t2<-min(ys)
dif<-t1-t2
while(dif>tol){
    tau<-(t1+t2)/2
    M<-Lag1(ys,ds,tau)
    elratio<-2*nss*sum(ds*log(1+M*(ys-tau)))
    if(elratio>cut) t2<-tau
    if(elratio<=cut) t1<-tau
    dif<-t1-t2
        }
LB<-(t1+t2)/2
###
```

References

Chen, J., Sitter, R.R. and Wu, C. (2002). Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys. *Biometrika*, 89, 230-237.

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

Owen, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75, 237-249.

Wu, C. (2004a). Some algorithmic aspects of the empirical likelihood method in survey sampling. *Statistica Sinica*, 14, 1057-1067.

Wu, C. (2004b). R/S-PLUS Implementation of pseudo empirical likelihood methods under unequal probability sampling. Working paper 2004-07, Department of Statistics and Actuarial Science, University of Waterloo.

Wu, C., and Sitter, R.R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the Americal Statistical Association*, 96, 185-193.

Wu, C., and Rao, J.N.K. (2004). Pseudo empirical likelihood ratio confidence intervals for complex surveys. Working paper 2004-06, Department of Statistics and Actuarial Science, University of Waterloo.