Empirical Likelihood Confidence Intervals and Significance Test for Regression Parameters under Complex Sampling Designs

Melike OGUZ ALPER* Yves G. BERGER †

Abstract
Confidence intervals based on ordinary least squares may have poor coverages for regression parameters when the effect of sampling design is ignored. Standard confidence intervals based on design variances may not have the right coverages when the sampling distribution is skewed. Berger and De La Riva Torres (2012) proposed an empirical likelihood approach which can be used for point estimation and to construct confidence intervals under complex sampling designs for a single parameter. We show that this approach can be extended to test the significance of a subset of model parameters and to derive confidence intervals. The proposed approach is not a straightforward extension of Berger and De La Riva Torres (2012) approach, because we consider the situation when the parameter is multidimensional and the parameter of interest is a subset of the parameter. This requires profiling which is not covered by Berger and De La Riva Torres (2012). The proposed approach intrinsically incorporates sampling weights, design variables, and auxiliary information. It may yield to more accurate confidence intervals when the sampling distribution of the regression parameters is not normal, the point estimator is biased, or the regression model is not linear. The proposed approach is simple to implement and less computer intensive than bootstrap. The proposed approach does not rely on re-sampling, linearisation, variance estimation, or design-effect.

Key Words: Design-based inference, estimating equations, empirical likelihood, regression parameters, unequal inclusion probabilities

1. Introduction

Regression models are widely used in social sciences, biological sciences, econometry and finance. Models are usually fitted to survey data, which may be collected for samples selected from finite populations. Sample units may be drawn from a complex sampling design which involves unequal inclusion probabilities, stratification and/or clustering. When the sampling design is informative (e.g. Skinner 1994; Pfeffermann 1993; Pfeffermann and Sverchkov 2009; Pfeffermann 2011), model-based estimators may be inconsistent and produce invalid inferences (see Binder and Roberts 2009). The standard design-based approaches assume the normality of the point estimators (e.g. Binder 1983; Deville 1999; Demnati and Rao 2004), which may not hold with moderate sample sizes or skewed data.

In this context, we propose to use an empirical likelihood approach to make inferences about model parameters and/or functions of them under unequal probability sampling. This is a non-parametric approach, where the sampling distribution is completely specified by the sampling design. Berger and De La Riva Torres (2012) proposed an empirical likelihood approach which can be used for point estimation and to construct confidence intervals under complex sampling designs. We show that this approach can be extended to the multidimensional parameter case, in the sense that we can derive confidence intervals and test the significance of a subset of model parameters while taking the sampling design into account.

*Funded by the Economic and Social Research Council, United Kingdom. University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, United Kingdom, M.OguzAlper@soton.ac.uk

†University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, United Kingdom, Y.G.Berger@soton.ac.uk

2. Parameters of Interest and Estimating Functions

Let s be a random sample of size n which is selected from the finite population U of size N with respect to a probability sampling design $p(s)$. Let y_i and $\boldsymbol{x}_i$ be some variables of interest. We assume that the values of y_i and $\boldsymbol{x}_i$ are known for all $i \in s$. Note that $\boldsymbol{x}_i$ is not necessarily a vector of auxiliary variables. The auxiliary variables will be denoted by $\boldsymbol{\xi}_i$ (see Section 4). We assume that all the units in the sample are respondents. Consider an unknown finite population parameter vector $\boldsymbol{\psi}_N$, which is the solution of the following population estimating equation.

$$\boldsymbol{G}(\boldsymbol{\psi}) = \sum_{i \in U} \boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\psi}) = \boldsymbol{0},$$

where $\boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\psi})$ is a vector of estimating functions. Most finite population parameters can be formulated through estimating functions (e.g. Binder 1983; Binder and Patak 1994; Qin and Lawless 1994; Godambe and Thompson 2009).

We assume that the parameter $\boldsymbol{\psi}_N$ converges to a model parameter $\boldsymbol{\psi}_0$. Let $\widehat{\boldsymbol{\psi}}$ be a design-consistent estimator of $\boldsymbol{\psi}_N$ based on the sample data. The estimator $\widehat{\boldsymbol{\psi}}$ is also an estimator of $\boldsymbol{\psi}_0$. When the sampling fraction is negligible (i.e. $n/N \to 0$), the variability of $\widehat{\boldsymbol{\psi}}$ is driven by the sampling design. Hence, the confidence intervals proposed in this paper can be viewed as confidence intervals of $\boldsymbol{\psi}_0$ or $\boldsymbol{\psi}_N$.

The proposed empirical likelihood approach is valid under a sampling design with replacement selection (*pps sampling*). However, the πps sampling design with fixed sample sizes is commonly used in practice. In this paper, we assume that the sampling fraction is negligible. This allows to approximate the actual design by the *pps* sampling design. Hence, the proposed approach is valid under the πps sampling as long as $n/N \to 0$.

2.1 Example: Regression Parameters

A generic expression for estimating equation for unknown model parameter $\boldsymbol{\beta}_0$ can be written as (e.g. Chen and Keilegom 2009)

$$\sum_{i \in U} \frac{\partial (h(\boldsymbol{x}_i)^\top \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{(y_i - \mu(h(\boldsymbol{x}_i)^\top \boldsymbol{\beta}))}{v_i} = \boldsymbol{0}, \tag{1}$$

where $\mu(\cdot)$ is a smooth function.

For a simple linear regression model, we have $h(\cdot) = \boldsymbol{x}_i$, $\mu(\cdot) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, and $v_i \propto 1$. In this case, the estimating equation (1) reduces to the following.

$$\sum_{i \in U} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}) = \boldsymbol{0}.$$

For generalised linear models, we have $\mu(\boldsymbol{x}_i^\top \boldsymbol{\beta}) = f^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta})$, and $v_i = v(\mu(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$, where $f(\cdot)$ is a link function. For example, the estimating equation for a logistic regression model under homoscedasticity; that is, $v_i \propto 1$, is given by

$$\sum_{i \in U} \boldsymbol{x}_i \left(y_i - \frac{exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})}) \right) = \boldsymbol{0},$$

where the link function is $f(\mu) = \text{logit}(\mu) = \log(\mu/(1-\mu))$ (see also Binder 1983, p.285; Owen 2001, ch.4.7; Chaudhuri et al. 2008, p.322).

2.2 Parameters of Interest and Nuisance Parameters

When dealing with regression, we are often interested in making inference about a subset of parameters. Let $\boldsymbol{\psi}_N = (\boldsymbol{\theta}_N^\top, \boldsymbol{\lambda}_N^\top)^\top$ where $\boldsymbol{\theta}_N$ is a $p \times 1$ vector of parameters of interest and $\boldsymbol{\lambda}_N$ is a $q \times 1$ vector of parameters which are not of primary interest. The parameter $\boldsymbol{\lambda}_N$ will be called as '*nuisance*' in this paper. Under a parametric likelihood framework, scale parameters are usually treated as nuisance (Kim and Zhou 2008). However, we call any unknown parameters as nuisance if they are not of primary interest, but are necessary to make inferences for the parameters of interest.

The parameter $\boldsymbol{\lambda}_N$ is assumed unknown and may need to be estimated in order to make inferences about $\boldsymbol{\theta}_N$ (e.g. Godambe and Thompson 1974; Binder and Patak 1994; Godambe and Thompson 1999, 2009). Owen (1990) proposed an approach to deal with the multidimensional parameters under an empirical likelihood framework. Qin and Lawless (1994) formally defined a profile empirical likelihood ratio test statistics to test hypotheses and to construct confidence intervals in the presence of a nuisance parameter. We propose to extend Qin and Lawless (1994) work in the context of a design-based inference.

3. Empirical Likelihood Approach

We use the *empirical log-likelihood function* given by Berger and De La Riva Torres (2012). It is defined as follows.

$$\ell(m) = \sum_{i \in s} \log(m_i), \tag{2}$$

where the m_i are unknown scale loads. The empirical log-likelihood function in (2) can be used for the sampling with replacement with unequal probability selection (i.e. pps sampling) designs as shown by Hartley and Rao (1969).

The *maximum empirical likelihood estimators* $\widehat{m}_i$ maximise the empirical log-likelihood in (2) with respect to the constraints $m_i \geq 0$ and

$$\sum_{i \in s} m_i \boldsymbol{c}_i = \boldsymbol{C}, \tag{3}$$

where the $\boldsymbol{c}_i$ and $\boldsymbol{C}$ are vectors defined in Sections 3.1 and 4. We assume that $\boldsymbol{c}_i$ and $\boldsymbol{C}$ satisfy with a set of regularity conditions given by Berger and De La Riva Torres (2012) and the following condition.

$$\left\| \frac{\partial \boldsymbol{c}_i}{\partial \boldsymbol{\lambda}} \right\| = O(1), \quad \text{for all } i \in s \text{ and } \boldsymbol{\lambda} \in \boldsymbol{\Lambda}, \tag{4}$$

where $\|\cdot\|$ denotes the Euclidean norm, $O(\cdot)$ defines the order of convergence, and $\boldsymbol{\Lambda}$ is a neighbourhood around the true population value $\boldsymbol{\lambda}_N$. The condition (4) implicitly implies that the $\boldsymbol{c}_i$ are differentiable with respect to $\boldsymbol{\lambda}$ in a neighbourhood of $\boldsymbol{\lambda}_N$ and the $\|\partial \boldsymbol{c}_i / \partial \boldsymbol{\lambda}\|$ is bounded in this neighbourhood (e.g. Godambe and Thompson 1974; Binder 1983; Qin and Lawless 1994; Owen 2001).

The maximum empirical likelihood estimators $\widehat{m}_i$ can be found using the method of Lagrange multipliers. Berger and De La Riva Torres (2012) showed that the solution of this maximisation is given by

$$\widehat{m}_i = (\pi_i + \boldsymbol{\eta}^\top \boldsymbol{c}_i)^{-1},$$

where $\boldsymbol{\eta}$ is such that the constraint (3) is satisfied.

3.1 Maximum Empirical Likelihood Estimator

Let $\ell(\widehat{m})$ be the maximum value of the empirical log-likelihood function $\ell(m)$ under the constraints $m_i \geq 0$ and (3) with $c_i = \pi_i$ and $C = n$. This implies that $\widehat{m}_i = \pi_i^{-1}$. Assume that $\widehat{m}_i^*$ maximises $\ell(m)$ subject to the constraints $m_i \geq 0$ and $\sum_{i\in s} m_i \boldsymbol{c}_i^* = \boldsymbol{C}^*$ with $\boldsymbol{c}_i^* = (c_i, \boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\psi})^\top)^\top$ and $\boldsymbol{C}^* = (C, \boldsymbol{0}^\top)^\top$, for a given vector $\boldsymbol{\psi} = (\boldsymbol{\theta}^\top, \boldsymbol{\lambda}^\top)^\top$. We assume that the condition (4) holds when $\boldsymbol{c}_i$ is substituted by $\boldsymbol{c}_i^*$ in (4). The *empirical log-likelihood ratio function* is defined by

$$\widehat{r}(\boldsymbol{\psi}) = 2\{\ell(\pi) - \ell\left(\widehat{m}^*(\boldsymbol{\psi})\right)\}, \tag{5}$$

where $\ell(\pi) = -\sum_{i\in s} \log(\pi_i)$ is the maximum value of (2) under the constraint (3) when $c_i = \pi_i$ and $C = n$.

The *maximum empirical likelihood estimates* $\widehat{\boldsymbol{\psi}}$ of the population parameters $\boldsymbol{\psi}_N$ is defined by the vector which minimises the empirical log-likelihood ratio function (5) over $\boldsymbol{\psi}$. The minimum value of $\widehat{r}(\boldsymbol{\psi})$ is obtained when $\widehat{r}(\boldsymbol{\psi}) = 0$; that is, when $\widehat{m}_i^* = \widehat{m}_i = \pi_i^{-1}$. Thus, the maximum empirical likelihood estimator of $\boldsymbol{\psi}_N$ is the solution of the following sample estimating equations.

$$\widehat{\boldsymbol{G}}(\boldsymbol{\psi}) = \sum_{i\in s} \boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\psi})\pi_i^{-1} = \boldsymbol{0}\cdot \tag{6}$$

3.1.1 Example: Simple Linear Regression

Consider a simple linear regression model with an intercept β_1 and a slope term β_2; that is, $\mu(x_i) = \beta_1 + \beta_2 x_i$. The sample estimating equations (6) for β_2 and β_1 are respectively given by

$$\sum_{i\in s} x_i(y_i - \beta_1 - \beta_2 x_i)\pi_i^{-1} = 0 \quad \text{and} \quad \sum_{i\in s} (y_i - \beta_1 - \beta_2 x_i)\pi_i^{-1} = 0\cdot \tag{7}$$

When solving the equations (7) for β_1 and β_2, we obtain the following point estimators.

$$\widehat{\beta}_1 = \bar{y}_H - \widehat{\beta}_2 \bar{x}_H \quad \text{and} \quad \widehat{\beta}_2 = \frac{\sum_{i\in s}(x_i - \bar{x}_H)(y_i - \bar{y}_H)\pi_i^{-1}}{\sum_{i\in s}(x_i - \bar{x}_H)^2\pi_i^{-1}},$$

where $\bar{x}_H = \sum_{i\in s} \breve{x}_i/\widehat{N}$ and $\bar{y}_H = \sum_{i\in s} \breve{y}_i/\widehat{N}$, with $\breve{x}_i = x_i\pi_i^{-1}$, $\breve{y}_i = y_i\pi_i^{-1}$, and $\widehat{N} = \sum_{i\in s} \pi_i^{-1}$. The random variables $\bar{y}_H$ and $\bar{x}_H$ are the Hájek (1971) estimators of the population means $\overline{X} = \sum_{i\in U} x_i/N$ and $\overline{Y} = \sum_{i\in U} y_i/N$ respectively. Under an equal probability sampling design, the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are the ordinary least square estimators.

3.2 Hypothesis Testing

Hypothesis testing for regression parameters is necessary for model comparison. For example, we may want to test if an additional regression parameter is significant.

Suppose we wish to test $H_0 : \boldsymbol{\theta}_N = \boldsymbol{\theta}_N^0$. Consider the *profile empirical log-likelihood ratio function* defined by

$$\widehat{r}(\boldsymbol{\theta}_N^0) = 2\left\{\ell(\pi) - \max_{\boldsymbol{\lambda}} \ell\left(\widehat{m}^*(\boldsymbol{\theta}_N^0, \boldsymbol{\lambda})\right)\right\}, \tag{8}$$

where the set of $\widehat{m}_i^*$ maximises $\ell(m)$ subject to the constraints $m_i \geq 0$ and $\sum_{i\in s} m_i \boldsymbol{c}_i^* = \boldsymbol{C}^*$ with $\boldsymbol{c}_i^* = (\pi_i, \boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_N^0, \boldsymbol{\lambda})^\top)^\top$ and $\boldsymbol{C}^* = (n, \boldsymbol{0}^\top)^\top$.

Note that in (8), we maximise $\ell(\widehat{m}^*(\boldsymbol{\theta}_N^0, \boldsymbol{\lambda}))$ over the nuisance parameter $\boldsymbol{\lambda}$ for a given value of $\boldsymbol{\theta}_N = \boldsymbol{\theta}_N^0$. The value of $\boldsymbol{\lambda}$ that maximises $\ell(m(\boldsymbol{\theta}_N^0, \boldsymbol{\lambda}))$ can be found by taking the derivative of $\ell(\widehat{m}^*(\boldsymbol{\theta}_N^0, \boldsymbol{\lambda}))$ with respect to $\boldsymbol{\lambda}$. Because (3) holds for $\widehat{m}_i^*$, $\boldsymbol{c}_i^*$, and $\boldsymbol{C}^*$, we obtain the following set of equations.

$$\dot{\ell}(\boldsymbol{\eta}, \boldsymbol{\lambda}) = \frac{\partial \ell(\widehat{m}^*(\boldsymbol{\theta}_N^0, \boldsymbol{\lambda}))}{\partial \boldsymbol{\lambda}} = \boldsymbol{\eta}^\top \sum_{i \in s} \widehat{m}_i \frac{\partial \boldsymbol{c}_i^*}{\partial \boldsymbol{\lambda}} = \mathbf{0}_q, \tag{9}$$

where q is the dimension of $\boldsymbol{\lambda}$. The Lagrange coefficients $\boldsymbol{\eta}$ and $\boldsymbol{\lambda}$ satisfying $\sum_{i \in s} \widehat{m}_i^* \boldsymbol{c}_i^* = \boldsymbol{C}^*$ and the equation (9) can be computed through an iterative procedure such as the Levenberg-Marquardt algorithm (e.g. Levenberg 1944; Marquardt 1963).

Under H_0, it can be shown that the profile empirical log-likelihood ratio function $\widehat{r}(\boldsymbol{\theta}_N^0)$ given by (8) asymptotically follows a limited *chi-squared distribution* with a p degree of freedom, where p is the dimension of the vector of parameters of interest $\boldsymbol{\theta}_N$. Thus, the *p-value* is given by

$$p - value = \int_{\widehat{r}(\boldsymbol{\theta}_N^0)}^{\infty} \chi^2_{df=p}(x)\, \mathrm{d}x,$$

where $\chi^2_{df=p}(\cdot)$ is the density of a chi-squared distribution with a p degree of freedom. Note that lack of fit would not affect the performance of the proposed empirical likelihood test (e.g. Owen 2001).

3.3 Confidence Intervals

We can construct confidence intervals for each parameter individually by treating the other parameters as nuisance. In this case, $p = 1$ and the vector $\boldsymbol{\theta}_N$ is the scalar θ_N. Then, based on the asymptotic chi-squared distribution of $\widehat{r}(\theta_N^0)$ under the null hypothesis $H_0 : \theta_N = \theta_N^0$, the $(1-\alpha)\%$ empirical likelihood Wilks (1938) type confidence interval for θ_N is given by

$$\left[\min\left\{\theta \,:\, \widehat{r}(\theta) \leq \chi^2_{df=1}(\alpha)\right\},\ \max\left\{\theta \,:\, \widehat{r}(\theta) \leq \chi^2_{df=1}(\alpha)\right\}\right],$$

where $\chi^2_{df=1}(\alpha)$ is the upper α - *quantile* of the chi-squared distribution with one degree of freedom. Note that $\widehat{r}(\theta)$ is a convex function of θ with a minimum value when θ is the empirical maximum likelihood estimator. Based on this property, the bisection method can be used to find the lower and upper bounds. This involves calculating $\widehat{r}(\theta)$ for several values of θ.

4. Incorporating Population Level Information

The efficiency of the estimators can be increased by using population level information. Handcock et al. (2000) and Chaudhuri et al. (2008) demonstrated that a large gain can be obtained in the precision of the estimators of model parameters when an auxiliary information is used in the estimation procedure. A population level information can be easily taken into account with the proposed approach.

Let $\boldsymbol{\xi}_i$ be a vector of auxiliary variables. Let $\boldsymbol{\vartheta}_N$ be a set of known population quantities which are functions of those variables. For example, these quantities may be in the form of means, totals, proportions, variances, quantiles, and/or distribution functions. Suppose $\mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}_N)$ be the vector of estimating functions which is used to define known population parameters $\boldsymbol{\vartheta}_N$. That is, $\boldsymbol{\vartheta}_N$ is the solution of the census estimating equation $\sum_{i \in U} \mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}) = \mathbf{0}$. For example, if the $\boldsymbol{\vartheta}_N$ are the population means, we use

$\mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}_N) = \boldsymbol{\xi}_i - \boldsymbol{\vartheta}_N$. In fact, $\mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}_N)$ does not have to be differentiable with respect to the nuisance parameter $\boldsymbol{\lambda}$. Because, we assume that $\mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}_N)$ does not depend on $\boldsymbol{\lambda}$. We also assume that $\boldsymbol{\vartheta}_N$ is not subject to any uncertainty.

Let $\boldsymbol{c}_i = (\pi_i, \mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}_N)^\top)^\top$ and $\boldsymbol{C} = (n, \mathbf{0}^\top)^\top$, where $\boldsymbol{\vartheta}_N$ is the known population value of $\boldsymbol{\vartheta}$. Then, the maximum value of $\ell(m)$ is $\ell(\widehat{m}) = \sum_{i\in s} \log(\widehat{m}_i)$, where the set of $\widehat{m}_i$ maximises $\ell(m)$ subject to $\widehat{m}_i \geq 0$ and $\sum_{i\in s} m_i \boldsymbol{c}_i = \boldsymbol{C}$. Let $\widehat{m}_i^*$ maximise $\ell(m)$ under the constraints $m_i \geq 0$ and $\sum_{i\in s} m_i \boldsymbol{c}_i^* = \boldsymbol{C}^*$, with $\boldsymbol{c}_i^* = (\boldsymbol{c}_i^\top, \boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\psi})^\top)^\top$ and $\boldsymbol{C}^* = (\boldsymbol{C}^\top, \mathbf{0}^\top)^\top$. The empirical likelihood log-likelihood ratio function is defined by

$$\widehat{r}(\boldsymbol{\psi}) = 2\left\{\ell(\widehat{m}(\boldsymbol{\vartheta}_N)) - \ell\left(\widehat{m}^*(\boldsymbol{\psi}, \boldsymbol{\vartheta}_N)\right)\right\}\cdot \tag{10}$$

The maximum empirical likelihood estimate $\widehat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}_N$ minimises (10). The minimum value of $\widehat{r}(\boldsymbol{\psi})$ is obtained when $\widehat{m}_i^* = \widehat{m}_i$ such that $\widehat{r}(\boldsymbol{\psi}) = 0$. Thus, the maximum empirical likelihood estimator of $\boldsymbol{\psi}_N$ will be the solution of the following sample estimating equations.

$$\widehat{\boldsymbol{G}}(\boldsymbol{\psi}) = \sum_{i\in s} \widehat{m}_i\, \boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\psi}) = \mathbf{0}\cdot \tag{11}$$

It can be shown that $\widehat{\boldsymbol{G}}(\boldsymbol{\psi})$ can be approximated by

$$\widehat{\boldsymbol{G}}(\boldsymbol{\psi}) = \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}) + o_p(Nn^{-1/2}),$$

where $\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi})$ is a *regression estimator* defined by

$$\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}) = \widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}) + \widehat{\boldsymbol{B}}_{opt}^\top \left(\mathbf{f}(\boldsymbol{\vartheta}_N) - \widehat{\mathbf{f}}_{\pi}(\boldsymbol{\vartheta}_N)\right), \tag{12}$$

where $\widehat{\mathbf{f}}_{\pi}(\boldsymbol{\vartheta}_N) = \sum_{i\in s} \mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}_N)\pi_i^{-1}$ is the Horvitz-Thompson estimator of $\mathbf{f}(\boldsymbol{\vartheta}_N) = \sum_{i\in U} \mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}_N)$, $\widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}) = \sum_{i\in s} \boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\psi})\pi_i^{-1}$, and $\widehat{\boldsymbol{B}}_{opt}$ is the regression coefficient given by

$$\widehat{\boldsymbol{B}}_{opt} = \widehat{\mathbf{var}}_{pps}(\widehat{\mathbf{f}}_{\pi}(\boldsymbol{\vartheta}_N))^{-1}\widehat{\mathbf{cov}}_{pps}(\widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}), \widehat{\mathbf{f}}_{\pi}(\boldsymbol{\vartheta}_N)),$$

where $\widehat{\mathbf{var}}_{pps}(\cdot)$ and $\widehat{\mathbf{cov}}_{pps}(\cdot)$ are the variance and covariance estimators with respect to the pps sampling design. Note that the variance and covariance estimators under the πps sampling design can be approximated to $\widehat{\mathbf{var}}_{pps}(\cdot)$ and $\widehat{\mathbf{cov}}_{pps}(\cdot)$ respectively when the sampling fraction is negligible (e.g. Särndal et al. 1992, p.422). The estimator (12) is asymptotically a design-optimal regression estimator under the pps sampling design (e.g. Isaki and Fuller 1982; Montanari 1987; Rao 1994; Berger et al. 2003).

Suppose we wish to test $H_0 : \boldsymbol{\theta}_N = \boldsymbol{\theta}_N^0$. Let the set of $\widehat{m}_i^*$ maximises $\ell(m)$ subject to the constraints $m_i \geq 0$ and $\sum_{i\in s} m_i \boldsymbol{c}_i^* = \boldsymbol{C}^*$, with $\boldsymbol{c}_i^* = (\boldsymbol{c}_i^\top, \boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\theta}_N^0, \boldsymbol{\lambda})^\top)^\top$ and $\boldsymbol{C}^* = (\boldsymbol{C}^\top, \mathbf{0}^\top)^\top$, where $\boldsymbol{c}_i = (\pi_i, \mathbf{f}_i(\boldsymbol{\xi}_i, \boldsymbol{\vartheta}_N)^\top)^\top$ and $\boldsymbol{C} = (n, \mathbf{0}^\top)^\top$. The profile empirical likelihood log-likelihood ratio function in the presence of population level information is given by

$$\widehat{r}(\boldsymbol{\theta}_N^0) = 2\left\{\ell(\widehat{m}(\boldsymbol{\vartheta}_N)) - \max_{\boldsymbol{\lambda}} \ell\left(\widehat{m}^*(\boldsymbol{\theta}_N^0, \boldsymbol{\lambda}, \boldsymbol{\vartheta}_N)\right)\right\}\cdot \tag{13}$$

It can be shown that the profile empirical log-likelihood function (13) asymptotically follows a chi-squared distribution with a p - degree of freedom under H_0. This allows us to test hypotheses and to construct confidence intervals using the approach described in Sections 3.2 and 3.3.

5. Simulation Study

In this §, we present some numerical results for a linear regression model with one intercept and one slope. We generated the Hansen, Madow and Tepping (HMT) population (Hansen et al. 1983). The population size is $N = 10\,000$. The values x_i are generated from a gamma distribution with a shape parameter equal to two and a scale parameter equal to five. Auxiliary variables are not considered in this section. We generate the y_i from a conditional gamma distribution with a shape parameter equal to $0.04x_i^{-3/2}(8+5x_i)^2$ and a scale parameter equal to $1.25x_i^{3/2}(8+5x_i)^{-1}$; that is,

$$y_i = 0.4 + 0.25x_i + x_i^{3/4}e_i,$$

where e_i are independent and identically distributed random variables with a zero mean and a standard deviation 0.25. We selected 1000 random samples of size $n = 500$ from this population using the randomised systematic sampling. The π_i are proportional to a measure of size $z_i = 5 + y_i + x_i + \epsilon_i$, where $\epsilon_i \sim exp(rate = 1) - 1$. The linear regression model of interest is defined by

$$y_i = \beta_1 + \beta_2 x_i + u_i, \quad \text{with} \quad v_i = x_i^{3/2}. \tag{14}$$

The parameter of interest is the slope term β_2. The intercept term β_1 is treated as a nuisance parameter. We take into account the heteroskedasticity by introducing the variance function, $v_i = x_i^{3/2}$.

We compare the Monte-Carlo performance of the proposed empirical likelihood confidence intervals for β_2 with the Wald type of confidence interval and design-based confidence intervals. For the latter, we used the Pseudo likelihood (e.g. Binder 1983; Binder and Patak 1994; Godambe and Thompson 2009) and the rescaled bootstrap methods (Rao et al. 1992). The Wald confidence intervals are based on the normality of the ordinary least squares estimator of the slope parameter. We considered two Pseudo likelihood approaches. With the '*pseudo likelihood 1*', we substitute β_2 with its design-based estimator $\widehat{\beta}_2$ in the variance term in the pivot given by Godambe and Thompson (2009, p.99) (see also Binder and Patak 1994, p.1039). For a linear regression, this is equivalent to the method of linearisation of estimating functions (e.g. Binder 1983; Deville 1999; Demnati and Rao 2004). With the '*pseudo likelihood 2*' approach, we solve the same pivot for β_2 without substituting β_2 with $\widehat{\beta}_2$ in the variance expression. The percentile method was used to obtain the rescaled bootstrap confidence intervals (e.g. Rao et al. 1992).

In Table 1, we compare several statistics regarding the performance of confidence intervals. The standardised lengths are computed based on the expression given by Kovar et al. (1988, p.32).

$$Standardised\ Length = \frac{Average\ Length}{2z_{\alpha/2}\sqrt{\text{MSE}}},$$

where $\text{MSE} = (B-1)^{-1}\sum_{b=1}^{B}(\widehat{\beta}_{2b} - \beta_2^N)^2$, where $\widehat{\beta}_{2b}$ is the point estimate of the population parameter β_2^N for the b-th sample, $B = 1000$, and $z_{\alpha/2}$ is the $\alpha/2$ - quantile of the standard normal distribution.

Ratio of average length (AL) is defined as the ratio of the average length of the empirical likelihood confidence intervals to the average length of the confidence intervals produced by other methods. Ratio of standard deviation (SD) of length is the ratio of the standard deviation of length of the empirical likelihood confidence intervals to the standard deviation of length of the confidence intervals produced by other methods.

Table 1 gives the observed coverages of the 95% confidence intervals constructed based on different methods. Standard confidence intervals are based on the normality of the point

estimator. However, when the sampling distribution is skewed, the normality assumption may not hold. This explains the poor coverages of the Wald and the pseudo likelihood 1 approaches. The poor coverage of the Wald type of confidence intervals is also due to the fact that this method ignores the sampling design. We have an overcoverage with the rescaled bootstrap.

The coverage probabilities of the empirical likelihood and the pseudo likelihood 2 confidence intervals are not significantly different from the nominal level (i.e. 95%). The pivot of the pseudo likelihood 2 approach is closer to normality than the standard t-statistics that is obtained from the pseudo likelihood 1 approach. However, for some samples, the pseudo likelihood 2 does not produce two solutions for the confidence boundaries. This is an issue that was pointed out by Godambe and Thompson (2009, p.92).

The least square estimators of the slope parameter and its standard error are not design-unbiased. We observed an overestimation of the parameter of interest. This explains the high level of lower tail error. The tail errors for all the methods, except the upper tail of the pseudo likelihood 1 and the lower tail of the rescaled bootstrap, significantly differ from the nominal level (i.e. 2.5%).

In terms of the standardised lengths, the rescaled bootstrap method has the largest confidence intervals on average compared to the other methods. Although the empirical likelihood and the pseudo likelihood 2 confidence intervals have good coverage probabilities, the former is more reliable than the latter with regards to the standard deviation of length (see the last column of Table 1).

Table 1: Coverages of the 95% Confidence intervals for the slope parameter (β_2^N) of the linear regression model in (14). The Hansen-Madow-Tepping population size is N=10 000. The sample size is $n = 500$.

$n = 500$	Coverage Probability	Lower Error	Upper Error	Standardised Length	Ratio AL	Ratio SD Length
Wald	76.6*	23.8*	0.1*	0.632	0.96	0.53
Empirical Likelihood	94.8	3.1*	2.1*	0.980	1.00	1.00
Pseudo Likelihood 1	94.0*	3.5*	2.5	0.951	0.97	1.07
Pseudo Likelihood 2	94.8	3.3*	1.9*	0.973	0.99	1.09
Rescaled Bootstrap	96.5*	2.4	1.1*	1.030	1.05	0.91

* Significantly different from the nominal levels (95% and 2.5% for coverage probability and tail errors respectively) at the 5% significance level (i.e. $p - value \leq 0.05$).

6. Conclusion

We proposed an empirical likelihood approach which can be used to make inferences for regression parameters. This approach takes the sampling design into account. It can be applicable to generalised linear models (see Section 2.1). It is also possible to apply it to quantile (robust) regression models (Huber 1981). In Section 3.2, we propose to profile out the parameters which are not of primary interest. The resulting profile empirical log-likelihood ratio function follows asymptotically a chi-squared distribution. Based on this property, we can test hypotheses and construct confidence intervals.

In Section 4, we showed how the population level information can be incorporated into the proposed approach. In addition, the empirical maximum likelihood estimator is asympotically design optimal. Unlike the usual calibration approach (Deville and Särndal

1992), the proposed approach can be used for testing and constructing confidence intervals. Moreover, the auxiliary information does not have to be in the form of totals or means, and empirical likelihood weights are always positive.

The proposed approach can be easily extended to stratified sampling designs by incorporating the strata information into the $\boldsymbol{c}_i$. It can also be used for two-stage sampling designs when the sampling fraction of the primary sampling units are negligible. In this case, primary sampling units play the role of units.

REFERENCES

Berger, Y. G., and De La Riva Torres, O. (2012), *Empirical likelihood confidence intervals for complex sampling designs*, Southampton: Southampton Statistical Sciences Research Institute http://eprints.soton.ac.uk/337688.

Berger, Y. G., Tirari, M. E. H., and Tillé, Y. (2003), "Towards Optimal Regression Estimation in Sample Surveys," *Australian and New Zealand Journal of Statistics*, 45, 319–329.

Binder, D. A. (1983), "On the variances of asymptotically normal estimators from complex surveys," *International Statistical Review*, 51, 279–292.

Binder, D. A., and Patak, Z. (1994), "Use of estimating functions for estimation from complex surveys," *Journal of the American Statisticsl Association*, 89(427), 1035–1043.

Binder, D. A., and Roberts, G. (2009), "Design and model-based inference for model parameters," *Handbook of Statistics 29: Sample Surveys: Inference and Analysis. Elsevier. Danny Pfeffermann and C.R. Rao eds*, pp. 33 – 54.

Chaudhuri, S., Handcock, M., and Rendall, M. (2008), "Generalized linear models incorporating population level information: an empirical likelihood based approach," *Journal of the Royal Statistical Society Series B*, 70, 311–328.

Chen, S., and Keilegom, I. V. (2009), "A review on empirical likelihood methods for regression," *Test*, 18, 415–447.

Demnati, A., and Rao, J. N. K. (2004), "Linearization variance estimators for survey data," *Survey Methodology*, 30, 17–26.

Deville, J. C. (1999), "Variance estimation for complex statistics and estimators: linearization and residual techniques," *Survey Methodology*, 25, 193–203.

Deville, J. C., and Särndal, C. E. (1992), "Calibration Estimators in Survey Sampling," *Journal of the American Statistical Association*, 87(418), 376–382.

Godambe, V. P., and Thompson, M. E. (1999), "A new look at confidence intervals in survey sampling," *Survey Methodology*, 25(2), 161–173.

Godambe, V. P., and Thompson, M. E. (2009), "Estimating functions and survey sampling," *Handbook of Statistics: Design, Method and Applications: D. Pfeffermann and C.R. Rao.(editors). Elsevier*, 29B, 83–101.

Godambe, V., and Thompson, M. E. (1974), "Estimating equations in the presence of a nuisance parameter," *The Annals of Statistics*, 2(3), 568–571.

Hájek, J. (1971), "Comment on a paper by D. Basu. In Foundations of Statistical Inference. Toronto: Holt, Rinehart and Winston,".

Handcock, M. S., Huovilainen, S. M., and Rendall, M. S. (2000), "Combining registration system and survey data to estimated birth probabilities," *Demography*, 37, 187–192.

Hansen, M. H., Madow, W. G., and Tepping, B. J. (1983), "An evaluation of model-dependent and probability-sampling inferences in sample surveys," *Journal of the American Statistical Association*, 78(384), 776–793.

Hartley, H. O., and Rao, J. N. K. (1969), *A new estimation theory for sample surveys, II*, A Symposium on the Foundations of Survey Sampling held at the University of North Carolina, Chapel Hill, North Carolina: Wiley-Interscience, New York.

Huber, P. J. (1981), *Robust Statistics* Wiley, New York.

Isaki, C. T., and Fuller, W. A. (1982), "Survey design under the regression super-population model," *Journal of the American Statistical Association*, 77, 89–96.

Kim, M.-O., and Zhou, M. (2008), "Empirical likelihood for linear models in the presence of nuisance parameters," *Statistics and Probability Letters*, 78, 1445–1451.

Kovar, J. G., Rao, J. N. K., and Wu, C. F. J. (1988), "Bootstrap and other methods to measure errors in survey estimates," *The Canadian Journal of Statistics*, 16, 25–45.

Levenberg, K. (1944), "A method for the solution of certain non-linear problems in least squares," *The Quarterly of Applied Mathematics*, 2, 164–168.

Marquardt, D. (1963), "An algorithm for least-squares estimation of nonlinear parameters," *Journal of the Society for Industrial and Applied Mathematics*, 11(2), 431–441.

Montanari, G. (1987), "Post sampling efficient QR-prediction in large sample survey," *International StatistiCalReview*, 55, 191–202.

Owen, A. B. (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18(1), 90–120.

Owen, A. B. (2001), *Empirical Likelihood*, New York: Chapman & Hall.

Pfeffermann, D. (1993), "The role of sampling weights when modelling survey data," *International Statistical Review*, 61, 317 – 337.

Pfeffermann, D. (2011), "Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?," *Survey Methodology*, 37(2).

Pfeffermann, D., and Sverchkov, M. (2009), "Inference under informative sampling," *Handbook of Statistics 29: Sample Surveys: Inference and Analysis: Danny Pfeffermann and C.R. Rao (editors). Elsevier*, pp. 455–487.

Qin, J., and Lawless, J. (1994), "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22(1), pp. 300–325.

Rao, J. N. K. (1994), "Estimating total and distribution function using auxiliary information at the estimation stage," *Journal of Official Statistics*, 10(2), 153–165.

Rao, J. N. K., Wu, C. F. J., and Yue, K. (1992), "Some recent work on resampling methods for complex surveys," *Survey Methodology*, 18, 209–217.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Skinner, C. J. (1994), "Sample models and weights," *ASA Proceedings of the Section on Survey Research Methods*, pp. 133–142.

Wilks, S. S. (1938), "Shortest Average Confidence Intervals from Large Samples," *The Annals of Mathematical Statistics*, 9(3), 166–175.