Design-based confidence intervals and significance test for regression parameters using an empirical likelihood approach

Melike Oguz-Alper* (M.OguzAlper@soton.ac.uk)[1], Yves G. Berger (Y.G..Berger@soton.ac.uk)[1]

Keywords: Design-based inference, estimating equations, nuisance parameter, unequal inclusion probabilities.

1. INTRODUCTION

Confidence intervals based on least squares may have poor coverages for regression parameters when the effect of sampling design is ignored. In addition, confidence intervals obtained from the standard design-based approaches [e.g. 1, 2, 3, 4] may not have the right coverages when the sampling distribution is skewed.

We propose to use an empirical likelihood approach to construct design-based confidence intervals and to test hypotheses for regression parameters under unequal probability sampling. Berger and De La Riva Torres [5] 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 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. This requires profiling which is not covered by Berger and De La Riva Torres [5].

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. It does not rely on re-sampling, linearisation, variance estimation, or design-effect.

1.1. Parameter of interest and estimating equations

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 $p(s)$. Let y_i and $\boldsymbol{x}_i$ be some variables of interest. Suppose that $\boldsymbol{\psi}_N$ is an unknown finite population parameter, 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 [e.g. 1, 2, 4, 6]. For example, for a simple linear regression, we have $\boldsymbol{g}_i(y_i, \boldsymbol{x}_i, \boldsymbol{\psi}) = \boldsymbol{x}_i(y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})$.

We assume that the finite population parameter $\boldsymbol{\psi}_N$ converges to the model parameter $\boldsymbol{\psi}_0$. If $\widehat{\boldsymbol{\psi}}$ is a design-consistent estimator of $\boldsymbol{\psi}_N$ based on a sample data (see Section 2.1), the estimator $\widehat{\boldsymbol{\psi}}$ is also an estimator of $\boldsymbol{\psi}_0$. Assuming that the sampling fraction is negligible, the variability of $\widehat{\boldsymbol{\psi}}$ is driven by the sampling design. Hence, design-based

[1] University of Southampton, Southampton Statistical Sciences Research Institute, Southampton, SO17 1BJ, United Kingdom. * Funded by the Economic and Social Research Council (ESRC), United Kingdom.

confidence intervals proposed in this paper can be viewed as confidence intervals of $\boldsymbol{\psi}_N$ or $\boldsymbol{\psi}_0$.

2. EMPIRICAL LIKELIHOOD INFERENCE

We use the *empirical log-likelihood function* given by Berger and De La Riva Torres [5]. It is defined as follows.

$$\ell(m) = \sum_{i \in s} \log(m_i), \tag{1}$$

where the m_i are unknown scale loads. The empirical log-likelihood function in (1) can be used for the sampling with replacement with unequal probability designs as shown by Hartley and Rao [7]. In this paper, we assume that the sampling fraction is negligible. Hence, the proposed approach is valid under the πps sampling as $n/N \to 0$.

The *maximum empirical likelihood estimators* $\widehat{m}_i$ maximise the empirical log-likelihood in (1) with respect to the constraints $m_i \geq 0$ and

$$\sum_{i \in s} m_i \boldsymbol{c}_i = \boldsymbol{C}, \tag{2}$$

where the $\boldsymbol{c}_i$ and $\boldsymbol{C}$ are vectors defined in Section 2.1. We assume that $\boldsymbol{c}_i$ and $\boldsymbol{C}$ satisfy with a set of regularity conditions given by Berger and De La Riva Torres [5] and the condition $\|\partial \boldsymbol{c}_i/\partial \boldsymbol{\lambda}\| = O(1)$, for all $i \in s$ and $\boldsymbol{\lambda} \in \boldsymbol{\Lambda}$, where $\|.\|$ denotes the Euclidean norm, $O(\cdot)$ defines the order of convergence, and $\boldsymbol{\Lambda}$ is a neighbourhood around the true population value $\boldsymbol{\lambda}_N$. This condition implicitly implies that the $\boldsymbol{c}_i$ are differentiable with respect to $\boldsymbol{\lambda}$ in a neighbourhood of $\boldsymbol{\lambda}_N$ [e.g. 1, 6, 8].

Berger and De La Riva Torres [5] showed that the maximum empirical likelihood estimators $\widehat{m}_i$ are given by $\widehat{m}_i = (\pi_i + \boldsymbol{\eta}^\top \boldsymbol{c}_i)^{-1}$, where $\boldsymbol{\eta}$ is such that the constraint (2) is satisfied.

2.1. Point estimation

Let $\ell(\widehat{m})$ be the maximum value of the empirical log-likelihood function $\ell(m)$ under the constraints $m_i \geq 0$ and (2) 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 $\boldsymbol{\psi}$. The *empirical log-likelihood ratio function* is defined by

$$\widehat{r}(\boldsymbol{\psi}) = 2\{\ell(\widehat{m}) - \ell(\widehat{m}^*(\boldsymbol{\psi}))\}. \tag{3}$$

The *maximum empirical likelihood estimate* $\widehat{\boldsymbol{\psi}}$ of the population parameter $\boldsymbol{\psi}_N$ is defined by the vector which minimises (3). The minimum value of (3) 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 equation.

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

2.2. Hypothesis testing

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

$$\widehat{r}(\boldsymbol{\theta}_N^0) = 2\{\ell(\widehat{m}) - \max_{\boldsymbol{\lambda}} \ell\left(\widehat{m}^*(\boldsymbol{\theta}_N^0, \boldsymbol{\lambda})\right)\}, \tag{4}$$

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 (4), we maximise $\ell(\widehat{m}^*(\boldsymbol{\theta}_N^0, \boldsymbol{\lambda}))$ over the parameter $\boldsymbol{\lambda}$ for a given value of $\boldsymbol{\theta}_N = \boldsymbol{\theta}_N^0$.

Under H_0, it can be shown that the profile empirical log-likelihood ratio function $\widehat{r}(\boldsymbol{\theta}_N^0)$ given by (4) follows asymptotically a *chi-squared distribution* with a p degree of freedom. Based on this, we can compute the *p-value*. Note that lack of fit would not affect the performance of the proposed empirical likelihood test [e.g. 8].

2.3. Confidence region

We can obtain confidence region for each parameter individually profiling out over the other parameters. In this case, $p = 1$ and we have the scalar θ_N instead of the vector $\boldsymbol{\theta}_N$. 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 confidence region for θ_N is given by the set $\{\theta : \widehat{r}(\theta) \leq \chi^2_{df=1}(\alpha)\}$, where $\chi^2_{df=1}(\alpha)$ is the upper α - *quantile* of the chi-squared distribution with one degree of freedom.

3. RESULTS

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 [see 9]. The population size is $N = 10\,000$. We selected 1000 random samples of size $n = 500$ from this population using the randomised systematic sampling with unequal probabilities.

The linear regression model of interest is defined by $y_i = \lambda + \theta x_i + u_i$, where the u_i are independent random variables with $var(u_i|x_i) \propto x_i^{3/2}$. The parameter of interest is the slope θ. We profile out over the intercept λ when minimising (4).

Table 1 gives the observed coverages of the 95% confidence intervals constructed based on several methods. We considered two Pseudo likelihood approaches which are given by Binder and Patak [2] and Godambe and Thompson [4] [see also 1].

Standard confidence intervals are based on the normality of the point estimator. Note that, 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 (see Table 1). 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 [e.g. 10]. Moreover, it has the largest confidence intervals on average compared to the other methods (see the ratio of average lengths in Table 1).

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%). However, the former is more reliable than the latter with regards to the standard deviation of length (see the last column of Table 1).

4. CONCLUSIONS

We proposed an empirical likelihood approach which can be used to make inferences for regression parameters incorporating the sampling design. The proposed approach can be used for generalised linear models.

It can be easily shown that the population level information can be taken into account with the proposed approach. Unlike the usual calibration approach [11], 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 [5].

The proposed approach can be easily extended to stratified sampling designs by incorporating the strata information into the $\boldsymbol{c}_i$.

Table 1. Observed coverages of the 95% confidence intervals for the slope θ_N.

N=10 000, n=500	Coverage probability	Lower error	Upper error	Ratio average length	Ratio SD length
Wald	76.6*	23.8*	0.1*	0.96	0.53
Empirical likelihood	94.8	3.1*	2.1*	1.00	1.00
Pseudo likelihood 1	94.0*	3.5*	2.5	0.97	1.07
Pseudo likelihood 2	94.8	3.3*	1.9*	0.99	1.09
Rescaled bootstrap	96.5*	2.4	1.1*	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. $p - value \leq 0.05$).

REFERENCES

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

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

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

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

[5] Y. G. Berger and O. De La Riva Torres, *Empirical likelihood confidence intervals for complex sampling designs*. S3RI, http://eprints.soton.ac.uk/337688, (2012).

[6] J. Qin and J. Lawless, Empirical likelihood and general estimating equations, *The Annals of Statistics*, 22, (1994), 300–325.

[7] H. O. Hartley and J. N. K. Rao, *A new estimation theory for sample surveys*, II. Wiley-Interscience, New York, 1969.

[8] A. B. Owen, *Empirical Likelihood*, Chapman & Hall, New York, 2001.

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

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

[11] J. C. Deville and C. E. Särndal, Calibration estimators in survey sampling, *Journal of the American Statistical Association*, 87, (1992), 376–382.