UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Division of Social Statistics and Demography

Variance Estimation of Change in Poverty Rates and Empirical Likelihood Inference in the Presence of Nuisance Parameters under Complex Sampling Designs

by

Melike Oguz-Alper

Thesis for the degree of Doctor of Philosophy

March 14, 2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Division of Social Statistics and Demography

Doctor of Philosophy

VARIANCE ESTIMATION OF CHANGE IN POVERTY RATES AND EMPIRICAL LIKELIHOOD INFERENCE IN THE PRESENCE OF NUISANCE PARAMETERS UNDER COMPLEX SAMPLING DESIGNS

by Melike Oguz-Alper

This thesis includes three papers. The first paper demonstrates how to estimate variance of change in poverty rates under rotating complex sampling designs. Measuring variance of change enables practitioners to judge whether or not the observed changes over time are statistically significant. The main difficulty in estimation of variance of change under rotating designs arises in the estimation of correlations between cross sectional estimates. This paper addresses a multivariate linear regression approach that provides a valid correlation estimator. Furthermore, poverty rate is a complex statistic that depends on a poverty threshold, which is estimated from the survey data. The paper mainly contributes by taking into account the variability of the poverty threshold in variance estimation of change. The approach is applied to the Turkish EU-SILC survey data. The second paper presents a design based inference in the presence of nuisance parameters by using an empirical likelihood approach. The main contribution of the paper is to develop an asymptotic theory to support the approach. The approach proposed can be used for testing and confidence intervals for finite population parameters such as (non)linear (generalised) regression parameters. For example, when comparing two nested models, the additional parameters are the parameters of interest, and the common parameters are the nuisance parameters. Sampling design and population level information are taken into account with the approach. Confidence intervals do not rely on resampling, linearisation, variance estimation, or design effects. The third paper shows how the empirical likelihood approach proposed in the second paper is applied to make inferences for regression coefficients when modelling hierarchical data collected from a two-stage sampling design where the first stage units may be selected with unequal probabilities. Multilevel regressions are often employed in social sciences to analyse data with hierarchical structure. This paper considers fixed effect regression parameters that can be defined through 'general estimating equations'. We use an 'ultimate cluster approach' by treating the first stage sample units as the units of interest.

Contents

Declaration of Authorship **ix**

Acknowledgements **xi**

1 General Introduction **1**
1.1 Contribution of the papers . . . 2
1.2 Authors' contributions to the papers . . . 4
1.3 Literature review on the first paper . . . 5
1.4 Literature review on the second paper . . . 7
1.5 Literature review on the third paper . . . 11

2 First Paper **17**
2.1 Introduction . . . 18
2.2 Rotating sampling designs . . . 19
2.3 Estimation of change of a poverty rate . . . 19
2.4 Allowing for the variability of the poverty threshold . . . 21
2.5 Estimation of change within domains . . . 23
2.6 Simulation study . . . 25
2.7 An application to the Turkish EU-SILC survey . . . 27
2.8 Conclusion . . . 30

3 Second Paper **33**
3.1 Introduction . . . 34
3.2 Parameters and estimating equations . . . 37
3.2.1 Examples: regression parameters . . . 37
3.3 Empirical log-likelihood function for unequal probabilities . . . 38
3.4 Maximum empirical likelihood point estimator . . . 39
3.5 Profile ELLR function in the presence of nuisance parameters . . . 40
3.5.1 Hypothesis testing and confidence intervals . . . 41
3.6 An algorithm to compute the profile empirical log-likelihood ratio function 41
3.7 Asymptotic distribution of the profile ELLR function . . . 43
3.8 Incorporating stratification . . . 45
3.9 Incorporating known population level information . . . 45
3.9.1 Maximum EL point estimator under population level information . 46
3.9.2 Testing and confidence intervals under population level information 47
3.9.3 Stratified and clustered population . . . 48
3.10 Simulation study . . . 49
3.10.1 Linear regression with the Hansen, Madow and Tepping population 51

3.10.2 Testing the significance of the intercept . . . 52
3.10.3 Linear regression with outlying values . . . 52
3.10.4 Logistic regression . . . 53
3.10.5 Logistic regression with population level information . . . 54
3.11 Conclusion . . . 56

4 Third Paper **59**
4.1 Introduction . . . 60
4.2 Two stage sampling design . . . 62
4.3 Working model . . . 63
4.4 Target finite population parameter and general estimating equation . . . 64
4.5 Empirical log-likelihood function . . . 65
4.6 Maximum EL point estimator under two stage sampling design . . . 67
4.7 Estimation within clusters . . . 68
4.7.1 Scaling . . . 69
4.7.2 Estimation of variance components . . . 70
4.8 EL inference under two stage sampling design . . . 71
4.9 Simulation study . . . 72
4.9.1 Population with outlying values . . . 75
4.10 Conclusion . . . 78

5 General Conclusion **81**

A Supplementary Material for the First Paper **87**
A.1 Derivation of the influence function of the poverty rate over a domain . . . 87
A.2 Generation of random variables . . . 89
A.3 R code for the first paper . . . 90

B Supplementary Material for the Second Paper **97**
B.1 Proof of expression (3.47) . . . 97
B.2 Proofs of the asymptotic results . . . 98
B.3 R code for the second paper . . . 110

C Supplementary Material for the Third Paper **123**
C.1 R code for the third paper . . . 123

Bibliography **133**

List of Tables

2.1 Empirical RB (%) of the variance and correlation estimators. 27
2.2 Empirical RRMSE (%) of the variance and correlation estimators. 28
2.3 Estimates when the poverty threshold is treated as fixed. 29
2.4 Estimates when the sampling variation of the poverty threshold is taken into account. Bandwidth parameter is based on standard deviation. . . . 30
2.5 Estimates when the sampling variation of the poverty threshold is taken into account. Bandwidth parameter is based on inter quartile range. . . . 31
2.6 Estimates when the sampling variation of the poverty is threshold taken into account. Bandwidth parameter is based on parameter A (see definition (2.8)). 31

3.1 95% confidence intervals for the slope of linear regression (3.59). HMT population. 52
3.2 Observed powers (in %) for testing the hypothesis that the intercept is equal to zero. 53
3.3 95% confidence intervals for the slope of linear regression (3.5). Population with outliers. 53
3.4 95% confidence intervals for the slope of logistic regression (3.6). 54
3.5 95% confidence intervals for the slope of the logistic regression (3.6). With population level information. 55
3.6 95% confidence intervals for the intercept of logistic regression (3.6). With and without population level information. 56

4.1 Finite population values of the regression coefficients in working model (4.25) and the relative bias (%) of their estimators. 75
4.2 95% confidence intervals for the estimates of regression coefficients in working model (4.25). 76
4.3 Relative bias (%) of the point estimators and the standard error estimators. Population with outlying values. 77
4.4 95% confidence intervals for the estimates of regression coefficients in working model (4.25). Population with outlying values. 77

Declaration of Authorship

I, Melike Oguz-Alper , declare that the thesis entitled *Variance Estimation of Change in Poverty Rates and Empirical Likelihood Inference in the Presence of Nuisance Parameters under Complex Sampling Designs* and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Part of this work has been published as:

   Oguz-Alper, M. and Berger, Y.G. (2015). Variance estimation of change in poverty rates: an application to the Turkish EU-SILC survey. *Journal of Official Statistics*, 31(2), 155-175.
   Oguz-Alper, M. and Berger, Y.G. (2014). Empirical likelihood confidence intervals and significance test for regression parameters under complex sampling designs. In, *Proceedings of the Survey Research Method Section, JSM 2014 - Joint Statistical Meetings - American Statistical Association*, Boston, US, 02-07 August 2014, 2070-2079.

Signed:.....

Date:.....

Acknowledgements

I owe my sincere gratefulness to my primary supervisor, Dr. Yves G. Berger, for his valuable support, useful comments and suggestions during the planning, development and writing of this thesis. I am thankful to Dr Berger for encouraging and motivating me during the whole period of my Doctor of Philosophy (PhD) training. His technical support and critique helped me to substantially improve my insight about the subject matter, to make progress and complete this thesis. I am also very grateful to him for training me on writing journal articles, encouraging me to attend international conferences and introducing me to people specialised in the research field. I have well trained in terms of many aspects during my PhD thanks to him.

I also would like to offer my special thanks to my secondary supervisor, Prof. Dr. Li-Chun Zhang, for providing very helpful comments on my second project and suggesting useful ideas regarding my third project. I appreciate the time he allocated me to discuss the technical aspects of my work.

I would like to express my sincere appreciation to my internal examiner Paul Smith and my external examiner Dr. Maria-Giovanna Ranalli for reviewing of my PhD thesis and providing very useful comments that certainly improved the thesis. I am very glad to had the opportunity to discuss my research with them from the theory to the practical implications of the methods presented in this thesis.

I wish to thank to Dr. Solange Correa-Onel for reviewing of my upgrade document and providing useful comments.

This research was supported by the Economic and Social Research Council (ESRC), United Kingdom. Some research regarding the first paper was done in the support of the Jean Monnet Scholarship Programme, European Union. I wish to thank to Turkish Statistical Institute (TurkStat), Turkey, for providing the datasets used in the first paper. I wish to thank to the guest editor, the associate editor and the two anonymous referees for their helpful comments which helped to improve the first paper. I am also thankful to Prof. Dr. J.N.K. Rao for his helpful comments on the second paper.

I would like to show my gratitude to my spouse, Ufuk Alper, for his patience, support and encouragement throughout my study.

Chapter 1

General Introduction

This thesis consists of three papers. The first paper demonstrates how to estimate variance of change in poverty rates under rotating complex sampling designs. The second paper presents a design based inference in the presence of nuisance parameters by using an empirical likelihood approach. The third paper investigates how the empirical likelihood approach proposed in the second paper is applied to make inferences for regression coefficients when modelling hierarchical data collected from a two-stage sampling design. The first paper is not directly connected to the other two papers that are based on the use of an empirical likelihood approach even though there are some common points among all.

The papers in this thesis approach the problems in the design based point of view (Neyman, 1934). In this framework, the sample s is selected from a finite population U with respect to a probability sampling design. The finite population values are treated as fixed. The target population quantities are known functions of these fixed values such as population totals, means, proportions, ratios or quantiles. We assume that all the units included in the sample s are respondents. The randomness arises only due to the random selection of the sample. Thus the sampling distribution is solely driven by the sampling design. Inferences for the finite population quantities are made based on the *probability distribution* denoted by $\mathcal{P}(s)$. This approach is called the *design based approach*. The inference may refer to point estimation, confidence intervals or hypothesis testing.

Inference for complex parameters is the main consideration of all the three papers. These parameters are defined as known functions of totals. The first paper deals with a poverty indicator, the poverty rate, that is computed as a ratio of two totals. It is a complex statistic not only because of being a ratio but also depending on the poverty threshold that is often computed from the median of the income distribution. The second paper focuses on linear and logistic regression parameters. The main interest of the third paper is the regression coefficients estimated under a correlated error structure when

modelling hierarchical survey data. Special treatment is required to make inference for those complex statistics. For example, linearisation (e.g. Deville, 1999) or resampling techniques (e.g. Bruch et al., 2011) are often used for variance estimation or the estimation of the *design effect* (Kish, 1965) may be required for testing (e.g. Rao and Scott, 1987) or confidence intervals (e.g. Wu and Rao, 2006). The last two papers present an empirical likelihood approach for testing and confidence intervals that does not require linearisation, resampling, variance estimation or design effect.

The first paper was published in the *Journal of Official Statistics* (see Oguz Alper and Berger, 2015). The part of the second paper was published in the *Proceedings of the Joint Statistical Meeting 2014* (see Oguz-Alper and Berger, 2014). The second paper has been submitted to a peer reviewed scientific statistical journal.

In Section 1.1, the main contributions of the papers are provided. In Section 1.2, the contributions of the authors are laid out. In Sections 1.3–1.5, a detailed literature review is provided respectively for each paper. The first paper is given in Chapter 2. The second paper is presented in Chapter 3. The third paper is demonstrated in Chapter 4. The supplementary materials including R (R Development Core Team, 2014) code for the papers are respectively provided in Chapters A–C.

1.1 Contribution of the papers

The main contribution of the first paper is that the variability of the poverty threshold is taken into account in the variance estimation of change in poverty rates. The poverty rate is often computed from a proportion of the median of the income distribution (e.g. Eurostat, 2003). As the median income is unknown and estimated from the survey data, it is subject to sampling errors. There are numerous papers that consider the variability of the threshold when estimating the cross sectional variances of the poverty rates (e.g. Berger and Skinner, 2003; Verma and Betti, 2005; Osier, 2009). The estimators may be biased when the fact that the median income is estimated is ignored. Variance estimates are conservative in this case (e.g. Preston, 1995; Berger and Skinner, 2003; Verma and Betti, 2011; Berger and Priam, 2016). However, this result may not necessarily hold for the variance estimates of change (see Section 2.6). In this paper, we investigate the effect of the variability of the poverty threshold on the variance estimates of change in poverty rates. We use the multivariate linear regression approach proposed by Berger and Priam (2016) to estimate the variance of change. The randomness of the median income is captured through the *generalised linearisation* technique proposed by Deville (1999). We demonstrate that the bias may be ignorable for the variance of change (see Section 2.6), unlike cross sectional variances. Furthermore, variance estimates of change may not be conservative. Another contribution of the first paper arises in derivation of the linearised variables for the poverty rates over domains (see Appendix A.1). Osier

(2009) derived an expression, by using the generalised linearisation technique, for the linearised variable of the poverty rate over the whole data. This expression cannot be applied directly to domain level poverty rates as the poverty rate is defined at domain level while the poverty threshold is computed from the whole data (e.g. Osier, 2009). A modification is required. We extend Osier's (2009) approach to domain level poverty rates. We apply the approach to the Turkish EU-SILC data to estimate the variance of change in poverty rates over several domains (see Section 2.7).

The main contribution of the second paper is to develop an asymptotic theory to support the empirical likelihood approach proposed for the inference in the presence of nuisance parameters. There are numerous works regarding the use of empirical likelihood inference in regression (e.g. Owen, 1991; Qin and Lawless, 1994; Wu and Sitter, 2001; Chaudhuri et al., 2008; Kim and Zhou, 2008; Zheng et al., 2012). The paper by Chen and Keilegom (2009) includes an elaborate review of the existing literature. These methods are based on the use of the empirical likelihood function given by Owen (1988) that assumes independent and identically distributed observations. These approaches cannot be used directly for the inference for regression parameters when modelling complex survey data as the identical distribution assumption is not valid due to the selection of the sample with unequal probabilities. We aim to fill this gap. We recall the empirical likelihood function proposed by Berger and De La Riva Torres (2016). This approach is limited to the single parameter case. It is not straightforward to extend this approach to the multidimensional parameter case. We propose profiling out the empirical log-likelihood ratio function provided by Berger and De La Riva Torres (2016) over the nuisance parameters. We show that the profile empirical log-likelihood ratio function asymptotically follows a χ^2 distribution under a set of regularity conditions. This property allows us to test hypotheses and construct confidence intervals for the subvector of parameters. The complexity of the sampling design such as stratification, clustering, unequal probabilities are taken into account with the approach proposed. We show that the population level information can be incorporated.

The main contribution of the third paper arises in the application of the profile empirical likelihood approach proposed in the second paper to make inference for the regression parameters when modelling hierarchical data. We assume that the hierarchy in the population is the same as the sampling hierarchy. We consider a two stage sampling design. We assume an ultimate cluster strategy (e.g. Hansen et al., 1953) so we consider that the ultimate cluster units are the units of interest. We use *general estimating equations* (GEE) to define the finite population parameters. We show how the estimating function is defined at the ultimate cluster level. We show how the design weights can be incorporated in to the estimation of the working variance covariance matrix under a uniform covariance structure.

1.2 Authors' contributions to the papers

My supervisor, Dr Yves G. Berger, and I both contributed to the first paper. The main idea of the first paper was suggested by Dr Berger. He advised me to read a few key papers to start with. He provided me the R (R Development Core Team, 2014) code. I made amendments in the code, wherever it requires, to apply the approach to the Turkish EU-SILC data and to implement the linearisation approach. I obtained the data from Turkish Statistical Institute (TurkStat) and prepared it for the analysis. I wrote my own code to carry out the simulation study. Dr Berger suggested to me that I could refer to the paper by Berger (2008) to generate the income variables for one wave. I found out how to generate correlated income variables for two waves (see Appendix A.2). I reviewed the relevant literature, conducted the numerical analysis and interpreted the results. I derived the expression for the linearised variables for domain level poverty rates (see Appendix A.1). I wrote the very first draft of the paper. Dr Berger revised it and put the written work into a journal paper structure. I amended notations and made further corrections in accordance with the comments and suggestions provided by Dr Berger. The first paper was also amended on the grounds of the referees' and the Associate Editor's comments. The version provided in this thesis is not the last version that was published in the *Journal of Official Statistics*. It is the version before the proof reading procedure.

Dr Berger and I have both participated in drafting the second paper. We both noticed that profiling is required to make inference for the subvector of parameters in the multidimensional parameter case. Dr Berger suggested to me that I should read the paper by Qin and Lawless (1994) and provided me a scribble of the proof to start with. I found out, by following Qin and Lawless's (1994) paper, how to derive the asymptotic distribution of the profile empirical log-likelihood ratio function. I made the derivation provided in Appendix B.1. I modified the asymptotic derivations provided by Berger and De La Riva Torres (2016) for the inference in the presence of nuisance parameters (see Appendix B.2). I did the literature review, conducted the simulation study and analysed the numerical results. I wrote functions in R (R Development Core Team, 2014) that are required to implement the empirical likelihood approach proposed and the other approaches used in the simulation study. Dr Berger recommended me to generate the Hansen et al.'s (1983) population. He also advised me to work with a population with outlying values. It was my idea to compare the approach that we propose with the 'pseudo likelihood' approaches (e.g. Binder, 1983; Binder and Patak, 1994) and the 'q-weighted' approach (e.g. Pfeffermann and Sverchkov, 1999). Prof Li-Chun Zhang, my second supervisor, suggested to me that I should perform a simulation on the significance of the intercept for the Hansen et al. (1983) data. I wrote the first draft of the paper including the asymptotic derivations. Dr Berger revised it and put it into a journal paper format. He added Sections 3.8 and 3.9.3, Lemma B.1 and Corollaries B.1 and B.2. He helped me in the proof of Lemma B.2. He suggested some amendments in

the notations. We both did proof reading and amended the paper several times until we have the version provided in this thesis.

I have fully participated in the production of the third paper. My supervisors, Dr Berger and Prof Zhang, provided me useful comments that inspired me to develop the concepts underlying the third paper. I did the literature review, drafted the paper, conducted the simulation study and analysed the numerical results.

1.3 Literature review on the first paper

Monitoring change in social and economic indicators may be a primary interest of data users to evaluate to what extent agreed policy targets are achieved. The poverty rate is an important policy indicator in the sense that it is one of the headline targets in the Europe 2020 strategy. This rate is defined as the proportion of people with an equivalised total net income below 60% of the national median income (Eurostat, 2003, p.2). It is calculated from the European Union Statistics on Income and Living Conditions (EU-SILC) surveys (Eurostat, 2012) collecting yearly information on income, poverty, social exclusion and living conditions from approximately 300,000 households across Europe.

Standard error estimation of change in poverty rates is required to judge whether or not the observed changes are statistically significant. Several methods for estimating the sampling variance of poverty rates like resampling and linearisation techniques have been discussed in the literature (e.g. Preston, 1995; Deville, 1999; Berger and Skinner, 2003; Demnati and Rao, 2004; Verma and Betti, 2005; Osier, 2009; Goedemé, 2010; Verma and Betti, 2011; Münnich and Zins, 2011; Osier et al., 2013; Berger and Priam, 2016). However, variance of change in poverty rate has been studied in only limited number of papers (e.g. Betti and Gagliardi, 2007; Münnich and Zins, 2011; Osier et al., 2013; Berger and Priam, 2016).

The estimation of variance of change requires estimation of covariances between two estimates measured at different time points (*waves*). This estimation is especially challenging when the sampling data has a rotational structure. As a trivial solution, covariances can be estimated based on the common sample by assuming the covariance in the common data is the same as the covariance between the two data at two different waves. Kish (1965, p.457-458) proposed estimating the covariance in a way that the correlation estimates based on the common sample is multiplied by the square roots of the cross sectional variances estimated from the whole data at corresponding waves. Population is treated as fixed. Simple random sampling with negligible sampling fractions was considered. Tam (1984) proposed estimating the covariance solely from the common sample unlike Kish (1965). Three sampling schemes that involve selections with simple random sampling are considered. Dynamic changes in the population are

not allowed. Tam's (1984) approach is applicable to designs with large sampling fractions. This approach can be extended to the case with unequal probabilities. However, this requires the use of the second order inclusion probabilities. Laniel (1987) extended Tam's (1984) approach by allowing dynamic changes in the population. In all of these approaches, sample sizes are assumed fixed. Nordberg (2000) introduced a rotation plan that is based on the use of *permanent random numbers*. Sample sizes are random with this approach. Dynamic changes in the population are taken into account. It is applicable to stratified simple random sampling design. Units can change their strata between two waves. The estimator is based on the conditional covariances. Holmes and Skinner (2000, p.29) proposed an approach for rotational stratified two stage sampling with unequal probabilities. Covariance is estimated from the variance of change and the cross sectional variances estimated from the common sample. They assumed that the primary sampling units are selected with replacement. They used ultimate cluster approach for variance estimation. Covariance estimators based on the common sample are biased and may provide negative variance estimates as noticed by Berger (e.g. 2004). A design consistent variance estimator of change was proposed by Berger (2004). The whole data at two waves is considered in variance estimation. The method is applicable under the sampling designs that can be approximated by the conditional poisson sampling. The covariance is estimated conditionally on the sample sizes that are fixed by the design. Finite population corrections are taken into account. It can be applied to the case with dynamic stratification. This method generally provides positive variance estimates. Qualité and Tillé (2008) extended Tam's (1984) approach by considering nonresponse and calibration under a stratified simple random sampling design. An unbiased estimator for the covariance was propose by Wood (2008). It can be applicable to many rotational designs. Unequal probabilities and finite population corrections are considered. However, his method involves determination of joint inclusion probabilities. Berger and Priam (2016) propose using a multivariate regression in estimation of correlations. Sampling fractions are assumed negligible. Sample sizes are fixed by the design. Berger and Escobar (2015) extended this approach by accommodating the effect of imputation.

The multivariate linear regression approach proposed by Berger and Priam (2016) provides a design consistent estimator for the correlation. They show that the variance estimates of change is always positive. Unequal probabilities and dynamic stratification are considered. It can be applicable to multi stage sampling designs by assuming an ultimate cluster approach. It can be used to estimate variances of change in complex statistics as long as the parameter of interest can be written as functions of totals. The method does not require calculation of second order inclusion probabilities unlike the approaches proposed by Nordberg (2000) and Wood (2008). It can be easily implemented in any statistical software. The effects of imputation and calibration can be taken into account (e.g. Berger and Escobar, 2015). Since the regression parameters are not of the primary interest, it is not required for the model to fit to the data for consistency.

In Chapter 2, we show how Berger and Priam's (2016) approach can be implemented by taking into account the variability of the poverty threshold. We consider two cases. In the first one, we treat the threshold as fixed, the *simple ratio approach* (see Section 2.3). In the second case, we consider the variability of the poverty threshold through the generalised linearisation technique (Deville, 1999). We call the second approach the *linearisation approach* (see Sections 2.4 and 2.5). We compare the variance estimators obtained from both approaches in terms of the relative bias and the relative mean square errors (see Section 2.6). Real data application is also provided (see Section 2.7).

1.4 Literature review on the second paper

Regression models are widely used in social sciences, biological sciences and econometrics. Models may be fitted to sample survey data that includes sample units selected with unequal probabilities from a clustered and/or stratified population. In this case, we do not have *independent and identically distributed* observations. Estimators based on the assumption of independent and identically distributed observations may be inconsistent and may produce invalid inferences when the sampling design is informative and the effect of the design is not taken into account (e.g. Pfeffermann and Sverchkov, 1999, 2003). We consider a general class of multidimensional parameters defined as the solution of a set of estimating equations. For example, this class includes complex parameters such as (non)linear regression parameters, generalised linear regression parameters, ratios, proportions or means. We consider a semiparametric approach, where the model is specified through estimating equations. We consider that the underlying distribution is unknown. Hence, parametric likelihood is not feasible.

We consider that the parameter of interest is a subvector of the parameters. The remaining parameters that are not of primary interest are called the '*nuisance*' parameters. The nuisance parameters are unknown by definition. There are numerous situations when nuisance parameters are used (e.g. Binder and Patak, 1994; Qin and Lawless, 1994; Godambe and Thompson, 2009; Zheng et al., 2012). In the parametric likelihood framework, the scale parameters are often treated as nuisance (Kim and Zhou, 2008). In a simple linear regression model, the intercept is the nuisance parameter when we want to construct a confidence interval for the slope. When comparing two nested models, the parameters of the more parsimonious model are the nuisance parameters while the additional parameters that we wish to test for significance are the parameters of interest. When we have heteroscedasticity and the variance function is unknown, the parameters of the variance model are the nuisance parameters (e.g. Owen, 1991, p.1740). The inference for the parameter of interest in the presence of nuisance parameters was first pointed out by Godambe and Thompson (1974). In the finite population context, this problem was discussed by Binder and Patak (1994) and Godambe and Thompson (1999, 2009). Owen (1990) demonstrated how to deal with the multidimensional parameter case under

an empirical likelihood framework. Qin and Lawless (1994) provided a profile empirical log-likelihood ratio test statistic that is used for hypothesis testing and confidence intervals when the observations are independent and identically distributed. Profiling requires maximisation of the empirical log-likelihood function over the nuisance parameters. This approach cannot be directly applied for complex survey data. We extend Qin and Lawless's (1994) approach by taking the sampling design and unequal probabilities into account.

The profile empirical likelihood approach proposed is an extension of the empirical likelihood approach proposed by Berger and De La Riva Torres (2016). Their approach can be used for a single parameter. We show that Berger and De La Riva Torres's (2016) approach can be extended for profiling by following Qin and Lawless's (1994) approach. This is not a trivial extension (see Appendix B.2). We show that the resulting profile empirical log-likelihood function asymptotically follows a χ^2-distribution. This property can be used for testing and confidence intervals in the presence of nuisance parameters. The approach proposed is different from Qin and Lawless's (1994) approach in the sense that the sampling design and unequal probabilities are considered with the approach proposed.

Empirical likelihood is a general, flexible, practical and a valid way of nonparametric inference. The leading triple works on empirical likelihood inference were provided by Owen (1988, 1990, 1991). It has flourished in terms of research and applications, especially in econometrics, since the review by Owen (2001). Rao and Wu (2009) provided an elaborate review on empirical likelihood inference in survey sampling. The use of empirical likelihood in survey sampling was motivated by Hartley and Rao (1968, 1969) with the well known 'scale-load' approach. This approach is applicable under simple random sampling and sampling with replacement with unequal probabilities. Auxiliary information can be incorporated. The scale-load approach is the same as the empirical likelihood approach proposed by Owen (1988) under simple random sampling when the sampling fraction is negligible. Chen and Qin (1993) formally used empirical likelihood in survey sampling. They were interested in point estimation for mean, median and cumulative distribution function by taking into account population level information. They assumed that the sampling fraction is negligible. Their approach can be used for the sampling with equal probabilities. Chen and Sitter (1999) introduced the '*pseudo empirical likelihood*' approach. They defined a population level empirical log-likelihood function by assuming that the finite population units are selected independently from the infinite population. Sample empirical log-likelihood was defined as the design based estimate of the population empirical log-likelihood. This approach can be used for point estimation under without replacement sampling with unequal probabilities. Zhong and Rao (2000) used the empirical log-likelihood function proposed by Chen and Qin (1993) for stratified simple random sampling. They used overall population totals and means

for point estimation. The sampling fraction within each stratum was assumed negligible. They also considered a confidence interval for the mean by adjusting the empirical log-likelihood ratio function. This adjustment is based on variance estimates that need to be computed. Wu and Sitter (2001) introduced a *model calibrated* pseudo empirical likelihood approach that takes into account the individual level population information (see also Chen et al., 2002). Wu and Rao (2006) reformulated the pseudo empirical log-likelihood function by using normalised design weights. They considered confidence intervals for the mean and cumulative distribution function. With this approach, the pseudo empirical log-likelihood ratio function is adjusted by the design effect to obtain a χ^2-distribution under unequal probabilities. Kim (2009) and Chen and Kim (2014) proposed an empirical likelihood under Poisson sampling. The population size should be known to apply this approach. Berger and De La Riva Torres (2016) propose a novel empirical likelihood approach that can be used under sampling with unequal probabilities. They propose using a *penalised empirical likelihood* to make inference under unequal probability sampling with large sampling fractions. They consider a wide range of finite population quantities such as totals, means, proportions and quantiles. Population level information can be incorporated with this approach.

Standard design based confidence intervals require variance estimates that are often computed through linearisation (e.g Deville, 1999; Demnati and Rao, 2004) or resampling techniques (e.g Rao et al., 1992). Confidence intervals are based on the assumption that the point estimator is normally distributed. However, this assumption may not hold when the data is skewed or includes outlying observations. Standard variance estimators ignore the fact that the nuisance parameters are estimated. This may yield a bias in the variance estimator. Standard confidence intervals may provide poor coverages and unbalanced tail errors when the point estimator is not normal or the variance estimator is biased or unstable. Empirical likelihood confidence intervals do require neither the normality of the point estimator nor the variance estimation. The approach proposed in this paper takes into account the randomness due to the nuisance parameters.

Binder (1983) proposed using estimating functions to make inference for complex parameters. Confidence intervals are based on the asymptotic normality of the point estimator. Variance estimates are functions of the point estimates of the parameters. Binder and Patak (1994) proposed a nonparametric version of the likelihood based score statistics that can be used in the presence of nuisance parameters (see also Godambe and Thompson, 1999, p.162). They proposed a method of *inverse testing* to construct confidence intervals. The sampling distribution of the pivotal statistics is inverted to obtain the lower and upper bounds. Binder and Patak (1994) demonstrated that the approach they proposed may provide better coverages than the approach based on Taylor linearisation (e.g. Binder, 1983). As pointed out by Godambe and Thompson (1999), model misspecification does not affect the performance of the inverse testing. Godambe and Thompson

(2009, p.92) mentioned that the solutions for the boundaries may not always exist. Pfeffermann and Sverchkov (1999, 2003) considered a semiparametric approach that may require modelling the survey weights. Variances of the model parameters are estimated through linearisation or resampling techniques (e.g. Pfeffermann and Sverchkov, 1999). In Section 3.10, we shall compare numerically the approach that we propose with the approaches proposed by Binder (1983), Binder and Patak (1994) and Pfeffermann and Sverchkov (1999, 2003).

Chen and Sitter (1999) proposed an algorithm based on profiling the pseudoempirical likelihood ratio function when strata totals of auxiliary variables are unknown. Zhong and Rao (2000) presented the same algorithm for both point estimation and confidence intervals under stratified simple random sampling. For confidence intervals, the pseudoempirical likelihood needs to be adjusted by variance estimates to obtain a χ^2-distribution. This approach is limited to estimation of totals. There is no general theory on profiling for the pseudoempirical likelihood approach.

The efficiency of the point estimators can be increased by incorporating population level information, which may be available from administrative data, census data and/or population projections (e.g. Deville and Särndal, 1992). The use of population level information in empirical likelihood inference has been demonstrated in various papers (e.g. Hartley and Rao, 1968; Chen and Qin, 1993; Qin and Lawless, 1994; Chen and Sitter, 1999; Zhong and Rao, 2000; Wu and Sitter, 2001; Wu and Rao, 2006; Chaudhuri et al., 2008; Kim, 2009; Berger and De La Riva Torres, 2016; Chen and Kim, 2014).

In econometrics, the *generalised method of moments* is commonly used to incorporate population level constraints (e.g. Hansen, 1982). However, this method requires estimation of the covariance matrix (e.g. Chen and Kim, 2014, p.18). In the parametric likelihood, the *constrained maximum likelihood estimator* may be used to take into account the population level information when estimating regression coefficients (e.g. Handcock et al., 2000). However, this method may be computationally cumbersome especially when the constraints are not linear (e.g. Chaudhuri et al., 2008). A two-step empirical likelihood approach was proposed by Chaudhuri et al. (2008) to estimate regression parameters. In the first step, calibration weights are obtained by incorporating population level information. In the second step, estimating equations that use these weights are solved to obtain point estimates. They demonstrated that a large gain can be achieved in the precision of the point estimators for the logistic regression parameters. This approach cannot be used directly for complex survey data as they assume independent and identically distributed observations.

The empirical likelihood approach proposed allows incorporating population level information in the presence of nuisance parameters. The effect of sampling design is taken into account. Confidence intervals do not rely on resampling, linearisation, variance estimation, or design effect. The approach proposed is simple to implement and less

computer intensive than the bootstrap. The empirical likelihood confidence intervals may provide better coverages than the standard methods especially when the sampling distribution is not normal or the parameter of interest is not linear.

In Chapter 3, we demonstrate how to extend Berger and De La Riva Torres's (2016) approach to make inference in the presence of nuisance parameters. We provide an algorithm to compute the profile empirical log-likelihood ratio function (see Section 3.6) and a set of regularity conditions under which the empirical likelihood inference proposed is valid (see Section 3.7). We show that the profile empirical log-likelihood ratio function can be used for hypothesis testing and confidence intervals (see Sections 3.5.1 and 3.9.2). We demonstrate that stratification, population level information and clustering can be incorporated with the approach proposed (see Sections 3.8–3.9.3). We compare the performance of the empirical likelihood confidence intervals and the power of the empirical likelihood test with the standard approaches (see Sections 3.10–3.10.5). We provide an asymptotic theory and show that the approach proposed produces a valid inference under a set of regularity conditions (see Appendix B.2).

1.5 Literature review on the third paper

The data used in social, behavioral, health or biological sciences may have a hierarchical structure due to the natural structure that occurs in the population of interest or due to the sampling or the experimental design itself. Multilevel models (Goldstein, 1986) or marginal models (e.g. Diggle et al., 2002) are often used to analyse such hierarchical data. The data may be collected from samples that are selected from a multi stage sampling design that may involve unequal probabilities at some or all stages of the selection. The sampling design is called informative when the selection probabilities are associated with the model outcome variable even after conditioning on the model covariates. Ignoring an informative sampling may result in invalid inference for regression parameters (e.g. Pfeffermann et al., 1998).

In standard, single level, regression models, sampling weights can be taken into account by using the *pseudo likelihood* approach (e.g. Binder, 1983; Skinner, 1989; Binder and Patak, 1994). With this approach, the population is fixed and population data is assumed independent. Finite population parameters are known functions of the population data. They can be defined as the solutions of the estimating equations. The population estimating equation that defines the parameter of interest can be estimated from sample data by using survey weights. In multilevel models, however, it is not straightforward to apply the pseudo likelihood approach as the observations within higher levels of the hierarchy are not independent. When this is the case, population totals cannot be written as a single summation of the individual units (e.g. Grilli and Pratesi, 2004).

Pfeffermann et al. (1998) proposed using probability weights, by relying on the pseudo likelihood principle, in the iterative generalised least squares (IGLS) algorithm to estimate multilevel regression parameters under a two stage sampling design. The regression coefficients and the variance components are estimated iteratively. Pfeffermann et al. (2006) proposed a model based approach involving Bayesian methods. They extended the *sample model* approach proposed by Pfeffermann and Sverchkov (1999, 2003) to multilevel models.

The IGLS estimation procedure may be computationally intensive as mentioned by Kovačević and Rai (2003). Alternatively, the *general estimating equations* (GEE) (e.g. Liang and Zeger, 1986; Diggle et al., 2002) that involve the use of *working correlation structure* can be used to estimate regression parameters. Variance components are treated fixed and replaced by their estimates when they are unknown. Liang and Zeger (1986) showed that the GEE estimator is fully efficient when the working correlation structure is correctly specified. They also provided some empirical evidence that the gain in efficiency obtained by using the GEE estimator rather than the *ordinary least squares* (OLS) estimator can be considerably increased for large values of the correlation coefficient. Crowder (1995) demonstrated that asymptotic properties of the GEE estimators may not always hold when the correlation structure depends on unknown quantities. They suggested using an estimating equation that would always give a solution. Sutradhar and Das (1999) showed that the consistency of the estimators for the regression parameters that are obtained by using Liang and Zeger's (1986) approach is usually valid. They demonstrated that the estimator obtained under the independence assumption may be more efficient than the GEE estimator when the true correlation structure is exchangeable and the working correlation structure is misspecified.

The general estimating equations provided by Liang and Zeger (1986) do not involve survey weights or the characteristics of the sampling design. Survey weights can be incorporated into the general estimating equations by following the pseudo likelihood approach (e.g. Binder, 1983; Skinner, 1989). The resulting approach is called the *multilevel pseudo likelihood* approach (e.g. Pfeffermann and La Vange, 1989; Kovačević and Rai, 2003; Grilli and Pratesi, 2004; Asparouhov, 2006; Skinner and De Toledo Vieira, 2007; De Toledo Vieira and Skinner, 2008). Asparouhov (2006) provided conditions under which the multilevel pseudo likelihood estimator is approximately unbiased. Skinner and De Toledo Vieira (2007) noticed that the weighted IGLS estimator (Pfeffermann et al., 1998) and the weighted GEE estimator are expected to provide identical point estimates under a working uniform correlation structure. Huang and Hidiroglou (2003) considered estimation of fixed and random effects parameters in a linear mixed effect model by taking the sampling design into account. Grilli and Pratesi (2004) showed that the multilevel pseudo maximum likelihood estimation can be extended to multilevel logistic regression (see also Rabe-Hesketh and Skrondal, 2006). They demonstrated how this approach can be implemented in SAS (SAS Institute Inc., 2011). The approach can

also be implemented in Mplus (Muthén and Muthén, 1998-2012) (e.g. Asparouhov and Muthén, 2006). Multilevel pseudo likelihood approach can be straightforwardly applied to other two-level generalised models (e.g. Asparouhov, 2006; Asparouhov and Muthén, 2006). Sutradhar and Kovačević (2000) used the GEE approach by incorporating survey weights to analyse longitudinal survey data with a polytomous response variable. La Vange et al. (2001) analysed several clinical trials data by using logistic, proportional hazards and proportional odds regression models. They considered the effect of clustering in estimation of regression coefficients by using the GEE approach.

A working covariance structure should be specified to apply the GEE approach. Liang and Zeger (1986) proposed using the estimates of the variance parameters when they are unknown. Several methods have been suggested to estimate variance components under a uniform covariance structure (e.g. Searle et al., 1992; Longford, 1995; Graubard and Korn, 1996; Huang and Hidiroglou, 2003; Korn and Graubard, 2003; De Toledo Vieira and Skinner, 2008). These methods are based on the *method of moments* estimation technique (Henderson, 1953). Graubard and Korn (1996) accommodated survey weights in the estimation of variance components under a simple multilevel model. Huang and Hidiroglou (2003) investigated model and design based properties of various estimators of the variance components including those that they developed. Korn and Graubard (2003) proposed new estimators for the variance components that are approximately design unbiased. Some empirical evidence showing that the estimators proposed performed well under small samples was provided for a simple random intercept model. However, joint inclusion probabilities should be known to apply these estimators. De Toledo Vieira and Skinner (2008) compared the multilevel pseudo maximum likelihood estimator and several GLS estimators of the variance components and their associated linearised variance estimators. Provided the empirical evidence by De Toledo Vieira and Skinner (2008), the pseudo maximum likelihood estimators performed well overall.

Pfeffermann et al. (1998) proposed scaling the survey weights of the first level units to reduce the sampling bias when estimating variance components under a two level model. Scaling is not required at the ultimate cluster or at the highest level as the estimates of the parameters are invariant to the scaling at this level (Pfeffermann et al., 1998). The sum of the scaled weights within clusters provides some cluster level characteristics. For example, weights can be scaled so that the sum of the scaled weights within a cluster reduces to the observed cluster sample size or the *effective* cluster sample size (e.g. Potthoff et al., 1992). Asparouhov (2006) compared the bias of the estimators of the parameters that are obtained from several scaling methods. Some empirical evidence was provided on whether or not the performance of the scaling methods was affected by cluster sample sizes or the informativeness of the sampling design. However, there is no theoretical evidence to support which scaling method is better for what kind of parameters and under which sampling designs.

The effect of multi stage sampling design can be accommodated by specifying the sampling clusters as the model hierarchy (e.g. Huang and Hidiroglou, 2003). In this case, the multilevel pseudo likelihood approach requires the knowledge of surveys weights of the sampling units at each stage of the sampling hierarchy. However, these weights may not be available to practitioners. Kovačević and Rai (2003) proposed using some proxy weights of the first and second stage sampling units. They suggested using a conditional *retrospective sampling* for the case when the design and model hierarchies are different. However, the properties of the estimators that are obtained by using the proxy weights have not been provided.

Standard confidence intervals relies on variance estimates. Variance estimators for the regression parameters can be computed through Taylor linearisation, the *sandwich type estimator*, (e.g. Binder, 1983; Pfeffermann et al., 1998; Skinner and De Toledo Vieira, 2007; Kovačević and Rai, 2003) or through bootstrap (e.g. Grilli and Pratesi, 2004). The latter is very computationally intensive for hierarchical data. Skinner and De Toledo Vieira (2007, p.5) showed how to incorporate stratification and clustering into the linearised variance estimation of regression coefficients when modelling longitudinal complex survey data. They demonstrated the effect of clustering on variance estimation for the regression coefficients (see also De Toledo Vieira and Skinner, 2008). When the parameter of interest is a subvector of the parameters, the approach proposed by Binder and Patak (1994) can be followed to compute the conditional variance of the parameter of interest. When there is a bias in the variance estimators, standard confidence intervals may provide poor coverages.

Standard methods require the normality of the point estimators. The inference for the parameters may not be valid when the normality assumption does not hold. We propose using the profile empirical likelihood approach (e.g. Oguz Alper and Berger, 2015), which is based on the empirical likelihood approach proposed by Berger and De La Riva Torres (2016) to make inferences for regression parameters when modelling hierarchical data. We incorporate the hierarchical structure through a general estimating equation. We use an *ultimate cluster approach* (Hansen et al., 1953). We treat the ultimate clusters as the units of interest. The empirical likelihood approach is applied at the ultimate cluster level. Estimating functions are defined as the sum of the individual observations within clusters. This summation takes into account the correlation between any two observations in a given cluster.

Empirical likelihood inference allows us to investigate the design performance of the estimators. We assume that the sampling distribution is specified by the sampling design. Hence, we use design based confidence intervals that do not require the specification of the underlying model which may not be known. The model is used to define the point estimators through general estimating equations. The sampling design is taken into account with the approach proposed. The resulting point estimators of the regression parameters are design consistent, which is a property often requested by the

survey practitioners (e.g. You and Rao, 2002). The resulting empirical likelihood confidence intervals may be better than the standard confidence intervals even when the point estimator is not normal, the variance estimators are biased or unstable or the individual error variances are heteroscedastic. The confidence intervals proposed do not rely on resampling, linearisation, variance estimation or design effect. Population level information can be accommodated to increase the precision of the estimators.

In Chapter 4, we demonstrate how the profile empirical likelihood approach proposed in the second paper (see Chapter 3) can be applied to make inference for hierarchical regression parameters. We consider a two stage of sampling design (see Section 4.2). We assume that the model and design hierarchies are the same (see Section 4.3). The parameter of interest is defined as the unique solution of the population GEE (see Section 4.4). The empirical likelihood approach proposed relies on an ultimate cluster approach (see Sections 4.5 and 4.8). We demonstrate how the cluster level general estimating function is estimated based on the sample data (see Section 4.7). We consider scaling of the weights of the first level units (see Section 4.7.1). We use the survey weights to estimate the variance components (see Section 4.7.2). We do not consider the variability due to the variance components estimated. We provide the profile empirical log-likelihood ratio function under two stage sampling design (see Section 4.8). We compare the performance of the empirical likelihood confidence intervals with the standard confidence intervals (see Sections 4.9 and 4.9.1).

Chapter 2

First Paper

Variance estimation of change in poverty rates: an application to the Turkish EU-SILC survey

Melike Oguz-Alper and Yves G. Berger
University of Southampton, SO17 1BJ, Southampton, U.K.
M.OguzAlper@soton.ac.uk Y.G.Berger@soton.ac.uk

Abstract

Interpreting changes between point estimates at different waves may be misleading, if we do not take the sampling variation into account. It is therefore necessary to estimate the standard error of these changes in order to judge whether or not the observed changes are statistically significant. This involves the estimation of temporal correlations between cross sectional estimates, because correlations play an important role in estimating the variance of a change in the cross sectional estimates. Standard estimators for correlations, based on common samples, cannot be used because of the rotation used in most panel surveys, such as the European Union Statistics on Income and Living Conditions (EU-SILC) surveys. Furthermore, as poverty indicators are complex functions of the data, they need a special treatment when estimating their variance. For example, poverty rates depend on poverty thresholds which are estimated from medians. We propose using a multivariate linear regression approach to estimate correlations by taking into account the variability of the poverty threshold. We apply the proposed approach to the Turkish EU-SILC survey data.

Keywords: Linearisation; multivariate regression; stratification; unequal inclusion probabilities.

2.1 Introduction

In order to monitor progress towards agreed policy goals, particularly in the context of the Europe 2020 strategy, there is an interest in evaluating the evolution of social indicators. For the purpose of interpreting changes between indicators at different waves, it is important to estimate the standard error of these changes, so that we can judge whether or not observed changes are statistically significant. The poverty rate is an important policy indicator, especially within the context of the Europe 2020 strategy. This rate is defined as the proportion of people with an equivalised total net income below 60% of the national median income (Eurostat, 2003, p.2). This indicator is calculated from the EU-SILC surveys (Eurostat, 2012) collecting yearly information on income, poverty, social exclusion and living conditions from approximately 300,000 households across Europe. The poverty rate is a complex statistic unlike population totals or means, since it is based on a poverty threshold computed from the median of the income distribution. Hence, there exist two sources of variability: one is due to the estimated threshold and the other one comes from the estimated proportion given the estimated threshold (e.g. Berger and Skinner, 2003; Verma and Betti, 2011).

Several methods to estimate the variance of the poverty rate like resampling and linearisation techniques have been discussed in the literature (e.g. Preston, 1995; Deville, 1999; Berger and Skinner, 2003; Demnati and Rao, 2004; Verma and Betti, 2005; Osier, 2009; Goedemé, 2010; Verma and Betti, 2011; Münnich and Zins, 2011; Osier et al., 2013; Berger and Priam, 2016). However, variance of change for the poverty rate has been studied in a limited number of papers (e.g. Betti and Gagliardi, 2007; Münnich and Zins, 2011; Osier et al., 2013; Berger and Priam, 2016). Berger and Priam (2010, 2016) proposed an estimator for the variance of change which takes into account the complexities of the sampling design such as stratification, unequal probabilities, clustering and rotation (see also Osier et al., 2013). The approach proposed relies neither on the second order inclusion probabilities nor on the resampling methods unlike its competitors (Betti and Gagliardi, 2007; Wood, 2008; Münnich and Zins, 2011, p.20). It is based on a multivariate linear regression (general linear regression) approach that can be easily implemented by any statistical software (Berger and Priam, 2016). Berger et al. (2013) show how it can be implemented in SPSS.

In Section 2.2, rotating sampling designs are briefly introduced. In Section 2.3, we recall the variance estimator of change proposed by Berger and Priam (2010, 2016). This estimator ignores the sampling variability due to the poverty threshold by treating the poverty rate as a ratio. In Sections 2.4 and 2.5, we show how this approach can be adjusted to take into account the sampling variability of the poverty threshold. In Section 2.6, we compare the approach proposed with the more simple approach proposed by Berger and Priam (2010, 2016) (see also Osier et al., 2013) via a series of simulation. In Section 2.7, we apply the approach proposed to the Turkish EU-SILC survey

data. The variance estimator proposed depends on a bandwidth used for the estimation of the density. We also show how sensitive the variance estimates are to the chosen bandwidth parameter by considering different bandwidth parameters. In Section 2.8, a brief synopsis of our findings is given and some extensions are suggested. The supplementary material including the R (R Development Core Team, 2014) code is provided in Appendix A.

2.2 Rotating sampling designs

With rotating panel surveys, it is common practice to select new units in order to replace old units that have been in the survey for a specified number of waves (e.g. Gambino and Silva, 2009; Kalton, 2009). The units sampled in both waves usually represent a large fraction of the first wave sample. This fraction is called the fraction of the common sample. For example, for the annual EU-SILC surveys, this fraction is 75%. For the monthly Canadian labour force survey and the quarterly British labour force survey, this fraction is 80%. For the quarterly Finnish labour force survey, this fraction is 60%. We consider that the sample design is such that the common sample has a fixed number of units. Throughout this paper, we assume that the sampling fraction is negligible, that is, $(1-\pi_{t;i}) \approx 1$, where $\pi_{t;i}$ denotes the inclusion probability of unit i at wave t. In the case of a multi stage sampling design, the sampling fraction is associated with the primary sampling units, which are the units selected at the first stage of sampling. We use an "ultimate cluster approach" (e.g. Hansen et al., 1953) in this case. Under this approach, the ultimate cluster variance dominates the total variance when the sampling fraction at the first stage of sampling is negligible.

2.3 Estimation of change of a poverty rate

Let s_1 and s_2 be the samples selected at wave 1 and wave 2 respectively. Suppose that we wish to estimate the absolute net change $\Delta = \theta_2 - \theta_1$ between two population poverty rates θ_1 and θ_2, from wave 1 and wave 2 respectively. Suppose that Δ is estimated by $\widehat{\Delta} = \widehat{\theta}_2 - \widehat{\theta}_1$; where $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are the cross sectional estimators of poverty rates defined by

$$\widehat{\theta}_1 = \frac{\widehat{\tau}_1}{\widehat{\tau}_2} = \frac{\sum_{i\in s_1} \delta\{y_{1;i} \leq 0.6\widehat{Y}_{1;0.5}\}\pi_{1;i}^{-1}}{\sum_{i\in s_1} \pi_{1;i}^{-1}} \quad \text{and} \quad \widehat{\theta}_2 = \frac{\widehat{\tau}_3}{\widehat{\tau}_4} = \frac{\sum_{i\in s_2} \delta\{y_{2;i} \leq 0.6\widehat{Y}_{2;0.5}\}\pi_{2;i}^{-1}}{\sum_{i\in s_2} \pi_{2;i}^{-1}},$$

where $y_{t;i}$ is the "net equivalised income" (see Eurostat, 2003, p.2) of individual i at wave t and $\widehat{Y}_{t;0.5}$ is the estimate of the median of the population income distribution at wave t $(t = 1, 2)$. The function $\delta\{A\} = 1$, when A is true, and $\delta\{A\} = 0$ otherwise.

The design-based variance of the estimator of change $\widehat{\Delta}$ is given by

$$\text{var}(\widehat{\Delta}) = \text{var}(\widehat{\theta}_1) + \text{var}(\widehat{\theta}_2) - 2\text{corr}(\widehat{\theta}_1, \widehat{\theta}_2)\sqrt{\text{var}(\widehat{\theta}_1)\text{var}(\widehat{\theta}_2)}\,. \tag{2.1}$$

Standard design-based estimators can be used to estimate the cross sectional variances $\text{var}(\widehat{\theta}_1)$ and $\text{var}(\widehat{\theta}_2)$ (e.g. Deville, 1999). The correlation $\text{corr}(\widehat{\theta}_1, \widehat{\theta}_2)$ is the most difficult part to estimate as $\widehat{\theta}_1$ and $\widehat{\theta}_2$ are estimated from different samples because of the rotation. Estimation of the covariance term has been discussed in several papers (Kish, 1965, p.457-458; Tam, 1984; Laniel, 1987; Nordberg, 2000; Holmes and Skinner, 2000; Berger, 2004; Qualité and Tillé, 2008; Wood, 2008; Münnich and Zins, 2011).

Berger and Priam (2010, 2016) proposed a multivariate approach to estimate the correlation between functions of totals by incorporating the information related to the whole sample, $s = s_1 \cup s_2$. This approach can be used to estimate the variance of change between poverty rates when we ignore the sampling variability due to the estimated poverty threshold $0.6\widehat{Y}_{t;0.5}$, that is, when we treat the poverty rates as simple ratios.

When we treat the threshold as fixed, the change becomes a smooth function of four totals, that is, $\widehat{\Delta} = \text{g}(\widehat{\boldsymbol{\tau}})$, where $\widehat{\boldsymbol{\tau}} = (\widehat{\tau}_1, \widehat{\tau}_2, \widehat{\tau}_3, \widehat{\tau}_4)^{\text{T}}$ is the estimator of the vector of population totals, $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3, \tau_4)^{\text{T}}$. Berger and Priam (2010, 2016) showed that using the first-order Taylor approximation, the design-based variance of $\widehat{\Delta}$ can be estimated by

$$\widehat{\text{var}}(\widehat{\Delta}) = \widehat{\mathbf{grad}}(\widehat{\boldsymbol{\tau}})^{\text{T}}\widehat{\mathbf{var}}(\widehat{\boldsymbol{\tau}})\,\widehat{\mathbf{grad}}(\widehat{\boldsymbol{\tau}}), \tag{2.2}$$

where $\widehat{\mathbf{grad}}(\widehat{\boldsymbol{\tau}})$ is the gradient of $\text{g}(\widehat{\boldsymbol{\tau}})$ evaluated at $\widehat{\boldsymbol{\tau}}$, that is,

$$\widehat{\mathbf{grad}}(\widehat{\boldsymbol{\tau}}) = \frac{\partial \text{g}(\widehat{\boldsymbol{\tau}})}{\partial \widehat{\boldsymbol{\tau}}} = \left(-\frac{1}{\widehat{\tau}_2}, -\frac{\widehat{\tau}_1}{\widehat{\tau}_2^2}, \frac{1}{\widehat{\tau}_3}, -\frac{\widehat{\tau}_3}{\widehat{\tau}_4^2}\right)^{\text{T}},$$

and $\widehat{\mathbf{var}}(\widehat{\boldsymbol{\tau}})$ is given by

$$\widehat{\mathbf{var}}(\widehat{\boldsymbol{\tau}}) = \widehat{\boldsymbol{D}}^{\text{T}}\widehat{\boldsymbol{\Sigma}}\widehat{\boldsymbol{D}},$$

with

$$\widehat{\boldsymbol{D}} = \text{diag}\left\{\sqrt{\widehat{\text{var}}(\widehat{\tau}_1)\widehat{\Sigma}_{11}^{-1}}, \sqrt{\widehat{\text{var}}(\widehat{\tau}_2)\widehat{\Sigma}_{22}^{-1}}, \sqrt{\widehat{\text{var}}(\widehat{\tau}_3)\widehat{\Sigma}_{33}^{-1}}, \sqrt{\widehat{\text{var}}(\widehat{\tau}_4)\widehat{\Sigma}_{44}^{-1}}\right\},$$

where $\widehat{\boldsymbol{\Sigma}}$ is the *ordinary least squares* (OLS) estimator of the residual covariance matrix $\boldsymbol{\Sigma}$ of the multivariate linear regression given in (2.3) proposed by Berger and Priam (2010, 2016); $\widehat{\text{var}}(\widehat{\tau}_k)$ is the design-based variance estimator of the Horvitz and Thompson (1952) estimator of total τ_k, and $\widehat{\Sigma}_{kk}^{-1}$ is the k-th diagonal element of $\widehat{\boldsymbol{\Sigma}}$ ($k = 1, 2, 3, 4$). Berger and Priam (2010, 2016) showed that the variance estimator (2.2) gives an approximately unbiased estimator for the variance of change.

Let $\breve{p}_{t;i} = \delta\{y_{t;i} \leq 0.6\widehat{Y}_{t;0.5}\}\pi_{t;i}^{-1}$ and $w_{t;i} = \pi_{t;i}^{-1}$. The multivariate linear regression is given as follows,

$$\begin{pmatrix} \breve{p}_{1;i} \\ w_{1;i} \\ \breve{p}_{2;i} \\ w_{2;i} \end{pmatrix} = \begin{pmatrix} \alpha_{1;1}z_{1;i} + \alpha_{1;2}z_{2;i} + \alpha_{1;3}z_{1;i} \times z_{2;i} \\ \beta_{1;1}z_{1;i} + \beta_{1;2}z_{2;i} + \beta_{1;3}z_{1;i} \times z_{2;i} \\ \alpha_{2;1}z_{1;i} + \alpha_{2;2}z_{2;i} + \alpha_{2;3}z_{1;i} \times z_{2;i} \\ \beta_{2;1}z_{1;i} + \beta_{2;2}z_{2;i} + \beta_{2;3}z_{1;i} \times z_{2;i} \end{pmatrix} + \boldsymbol{\epsilon}_i. \tag{2.3}$$

The vector of the residuals follows a multivariate distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$. The distribution of $\boldsymbol{\epsilon}_i$ does not have to be specified as the covariance $\boldsymbol{\Sigma}$ is estimated based on a least squares technique. Rotation of the sampling design is incorporated into the model through the model covariates: $z_{t;i} = \delta\{i \in s_t\}$ and $z_{1;i} \times z_{2;i} = \delta\{i \in s_1, i \in s_2\}$. It should be noted that the correlations $\widehat{\text{corr}}(\widehat{\tau}_k, \widehat{\tau}_\ell)$, with $(k, \ell = 1, 2, 3, 4)$, are obtained from the estimated residual covariance matrix $\widehat{\boldsymbol{\Sigma}}$. The covariance terms on the non-diagonal part of the matrix $\widehat{\mathbf{var}}(\widehat{\boldsymbol{\tau}})$ are based on those estimated correlations $\widehat{\text{corr}}(\widehat{\tau}_k, \widehat{\tau}_\ell)$ and the estimated cross sectional variance terms $\widehat{\text{var}}(\widehat{\tau}_k)$. Note that this approach also accounts for a multi stage sampling, using an "ultimate cluster approach" (e.g. Osier et al., 2013; Di Meglio et al., 2013).

Berger and Priam (2010, 2016) showed that the multivariate approach gives estimates which are approximately equal to the Hansen and Hurwitz (1943) variance estimator (e.g. Holmes and Skinner, 2000).

The approach proposed can be easily extended to stratified sampling. In this case, we assume that the sample sizes within each stratum are fixed (non-random) quantities. The model covariates $z_{t;i}$ are replaced by the stratum wave indicators $z_{th;i} = \delta\{i \in s_{th}\}$, where s_{th} is the sample for the stratum h at wave t. As the rotation is done within each stratum, we consider the interactions $z_{th;i} \times z_{(t+1)h;i}$.

2.4 Allowing for the variability of the poverty threshold

Note that in the variance estimator (2.2), the variability of the poverty threshold is not taken into account because we treat $\widehat{\theta}_1$ and $\widehat{\theta}_2$ as ratios. Treating the poverty threshold as fixed might lead to over-estimation of the variances (e.g. Preston, 1995; Berger and Skinner, 2003; Verma and Betti, 2011). Verma and Betti (2011) compared the ratio variance estimator (i.e. when the poverty threshold is treated as fixed) with linearisation and Jackknife repeated replication. They found that the ratio variance estimator over-estimated the standard errors for all the poverty measures and several complex statistics. However, these findings are related to the cross sectional estimators and do not necessarily hold for the variance of change.

Taking into account the whole variability means that the sampling variation of the poverty threshold is also considered. However, the poverty rate is more complex than a ratio and cannot be expressed as a function of totals. We propose using the linearisation approach proposed by Deville (1999). The implementation of this approach for the poverty rate and the inequality measures can be found in the literature (e.g. Berger and Skinner, 2003; Verma and Betti, 2005; Osier, 2009; Münnich and Zins, 2011; Verma and Betti, 2011).

The linearised variable $L_{t;i}$ for individual i at wave t for the poverty rate is given by (see Osier, 2009)

$$L_{t;i} = \frac{1}{\widehat{N}_t}\left(\delta\{y_{t;i} \leq 0.6\widehat{Y}_{t;0.5}\} - \widehat{\theta}_t\right) - \frac{0.6}{\widehat{N}_t}\frac{\widehat{f}_t(0.6\widehat{Y}_{t;0.5})}{\widehat{f}_t(\widehat{Y}_{t;0.5})}\left(\delta\{y_{t;i} \leq \widehat{Y}_{t;0.5}\} - 0.5\right), \quad (2.4)$$

where $\widehat{f}_t(\cdot)$ is an estimator of the density function, which is defined in (2.5). The second term in (2.4) is an additional term which reflects the sample variation originating from the randomness of the estimated median income.

The density functions can be estimated on the basis of the Gaussian kernel function as follows (e.g. Preston, 1995):

$$\widehat{f}_t(x) = \frac{1}{\widehat{N}_t\widehat{h}_t}\sum_{i\in s_t}\frac{1}{\pi_{t;i}}K\left(\frac{x - y_{t;i}}{\widehat{h}_t}\right) \quad (2.5)$$

where $K(\eta) = (\sqrt{2\pi})^{-1}\exp(-\eta^2/2)$ is the Gaussian kernel, $\widehat{N}_t = \sum_{i\in s_t}\pi_{t;i}^{-1}$ is the Horvitz and Thompson (1952) estimator of the population size at wave t ($t = 1, 2$), and $\widehat{h}_t$ is the bandwidth parameter that can be defined in several ways (Silverman, 1986, p.45-48). For a normally distributed population and smooth densities, the following bandwidth parameter is recommended by Silverman (1986, p.46).

$$\widehat{h}_t = 1.06\widehat{\sigma}_{t;\widehat{Y}}\widehat{N}_t^{-1/5}, \quad (2.6)$$

where

$$\widehat{\sigma}_{t;\widehat{Y}} = \sqrt{\frac{1}{\widehat{N}_t}\left\{\sum_{i\in s_t}\frac{1}{\pi_{t;i}}y_{t;i}^2 - \frac{1}{\widehat{N}_t}\left(\sum_{j\in s_t}\frac{1}{\pi_{t;j}}y_{t;j}\right)^2\right\}},$$

is the estimated standard deviation of the income distribution. However, for skewed and long tailed distributions, Silverman (1986, p.47) proposed using the inter quartile range instead of the standard deviation of the distribution; that is,

$$\widehat{h}_t = 0.79\widehat{Y}_{t;iqr}\widehat{N}_t^{-1/5}, \quad (2.7)$$

where $\widehat{Y}_{t;iqr} = \widehat{Y}_{t;0.75} - \widehat{Y}_{t;0.25}$ is the weighted inter quartile range of the income distribution. Another bandwidth, which is very suitable for many densities, even for the modest bimodal ones, was suggested by Silverman (1986, p.48) as follows:

$$\widehat{h}_t = 0.9\widehat{A}_t\widehat{N}_t^{-1/5}, \tag{2.8}$$

where $\widehat{A}_t = \min(\widehat{\sigma}_{t;\widehat{Y}}, \widehat{Y}_{t;iqr}/1.34)$. It should be noted that the bandwidth in (2.8) is smaller than the other bandwidths in (2.6) and (2.7). Thus we are likely to obtain less smooth densities with the bandwidth (2.8).

It is worth mentioning that choosing a bandwidth parameter is a crucial step in applications (e.g. Verma and Betti, 2005; Graf, 2013; Graf and Tillé, 2014). For example, Verma and Betti (2005) showed that probability density functions are sensitive to the chosen bandwidth parameter. A large value for the bandwidth parameter results in a smoother density. Graf (2013, p.26-28) pointed out the potential danger of using standard deviation when estimating densities that might arise from extreme values in the data observed (for example, with income data). In such cases, Graf (2013) proposed using the logarithm to reduce the adverse impact of extreme values. He also remarked a fixed-bandwidth parameter might be problematic when observations are heaped up around some values. To avoid this problem, a more robust technique to estimate density involving nearest neighbours with minimal bandwidth was suggested by Graf (2013). Opsomer and Miller (2005) proposed a design-based criterion for selecting the bandwidth parameter in the finite population estimation context. This approach provides a data-driven bandwidth parameter selection. These alternative bandwidth parameter selection methods are not considered in this paper.

2.5 Estimation of change within domains

In practice, we are often interested in change within domains of interest. For example, we may be interested in change in poverty within different age groups. According to the definition given by Eurostat (2003), the poverty threshold is calculated based on the overall estimated median income rather than the estimated median income within the domains. Hence, when we are interested in a domain, the threshold will be the same for all domains.
Consider $d_{t;i}$ to be a domain indicator for individual i at wave t defined by

$$d_{t;i} = \begin{cases} 1 & \text{if } i \in D \text{ at wave } t, \\ 0 & \text{if } i \notin D \text{ at wave } t, \end{cases}$$

where D refers to the domain of interest. The poverty rate over a domain is defined by

$$\widehat{\theta}_{Dt} = \frac{\sum_{i \in s_t} d_{t;i} \delta\{y_{t;i} \leq \widehat{Y}_{t;0.5}\} \pi_{t;i}^{-1}}{\sum_{i \in s_t} d_{t;i} \pi_{t;i}^{-1}}.$$

To estimate the variance of change within domains under the ratio approach (see expression (2.2)), we substitute $\breve{p}_{t;i}$ by $\breve{p}_{Dt;i} = d_{t;i}\breve{p}_{t;i}$, and $w_{t;i}$ by $w_{Dt;i} = d_{t;i}w_{t;i}$ into the model in (2.3). Note that the values of the response variables will be equal to zero for units not included in the domain of interest.

For the linearisation approach, the linearised variables $L_{Dt;i}$ for individual i in domain D at wave t derived in Appendix A.1 (see expression (A.5)) are given by

$$L_{Dt;i} = \frac{d_{t;i}}{\widehat{N}_{Dt}} \left(\delta\{y_{t;i} \leq 0.6\widehat{Y}_{t;0.5}\} - \widehat{\theta}_{Dt} \right) - \frac{0.6}{\widehat{N}_t} \frac{\widehat{f}_{Dt}(0.6\widehat{Y}_{t;0.5})}{\widehat{f}_t(\widehat{Y}_{t;0.5})} \left(\delta\{y_{t;i} \leq \widehat{Y}_{t;0.5}\} - 0.5 \right),$$

where

$$\widehat{N}_{Dt} = \sum_{i \in s_t} \frac{d_{t;i}}{\pi_{t;i}},$$

$$\widehat{f}_{Dt}(x) = \frac{1}{\widehat{N}_{Dt}\widehat{h}_{Dt}} \sum_{i \in s_t} \frac{d_{t;i}}{\pi_{t;i}} K_D \left(\frac{x - y_{t;i}}{\widehat{h}_{Dt}} \right).$$

Here, $\widehat{h}_{Dt}$ can be (2.6), (2.7), or (2.8) with $\widehat{N}_{Dt}$, $\widehat{Y}_{Dt;iqr} = \widehat{Y}_{Dt;0.75} - \widehat{Y}_{Dt;0.25}$,

$$\widehat{\sigma}_{Dt;\widehat{Y}} = \sqrt{\frac{1}{\widehat{N}_{Dt}} \left\{ \sum_{i \in s_t} \frac{d_{t;i}}{\pi_{t;i}} y_{t;i}^2 - \frac{1}{\widehat{N}_{Dt}} \left(\sum_{j \in s_t} \frac{d_{t;j}}{\pi_{t;j}} y_{t;j} \right)^2 \right\}},$$

and $\widehat{A}_{Dt} = \min(\widehat{\sigma}_{Dt;\widehat{Y}}, \widehat{Y}_{Dt;iqr}/1.34)$. Let $\widehat{\Delta}_D = \widehat{\theta}_{D2} - \widehat{\theta}_{D1}$ be the estimate of change in poverty rate in domain D from wave 1 to wave 2. Thus the variance of domain change is estimated by

$$\widehat{\text{var}}(\widehat{\Delta}_D) = \widehat{\text{var}}(\widehat{\theta}_{D1}^L) + \widehat{\text{var}}(\widehat{\theta}_{D2}^L) - 2\widehat{\text{corr}}(\widehat{\theta}_{D1}^L, \widehat{\theta}_{D2}^L) \sqrt{\widehat{\text{var}}(\widehat{\theta}_{D1}^L) \widehat{\text{var}}(\widehat{\theta}_{D2}^L)}, \tag{2.9}$$

with

$$\widehat{\theta}_{Dt}^L = \sum_{i \in s_t} \frac{L_{Dt;i}}{\pi_{t;i}}. \tag{2.10}$$

We use the approach proposed by Berger and Priam (2010, 2016) by treating $\widehat{\theta}_{D1}^L$ and $\widehat{\theta}_{D2}^L$ in (2.10) as the estimators of totals. The correlation term $\widehat{\text{corr}}(\widehat{\theta}_{D1}^L, \widehat{\theta}_{D2}^L)$ in (2.9) is

computed from the estimated residual covariance matrix of the following model, $\widehat{\boldsymbol{\Sigma}}$,

$$\begin{pmatrix} \breve{L}_{D1;i} \\ \breve{L}_{D2;i} \end{pmatrix} = \begin{pmatrix} \alpha_{1;1}z_{1;i} + \alpha_{1;2}z_{2;i} + \alpha_{1;3}z_{1;i} \times z_{2;i} \\ \alpha_{2;1}z_{1;i} + \alpha_{2;2}z_{2;i} + \alpha_{2;3}z_{1;i} \times z_{2;i} \end{pmatrix} + \boldsymbol{\epsilon}_i,$$

with $\breve{L}_{Dt;i} = L_{Dt;i}\pi_{t;i}^{-1}$.

It should be noted that the domain information is incorporated into the model through the response variables, in contrast to the stratification, which are defined via model covariates (see Section 2.3). Note that the approach proposed can be used for stratum domains and unplanned domains.

2.6 Simulation study

In this section, the variance estimators from the ratio and the linearisation approaches are compared in terms of the relative bias (RB) and the relative root mean square error (RRMSE), respectively defined by expressions (2.11) and (2.12). Additionally, we investigate whether the ratio approach gives more conservative estimates. The statistical software R (R Development Core Team, 2014) was used.

The income variables at wave 1 and wave 2 were generated according to different probability distributions (see Appendix A.2). For each wave, a gamma distribution (shape=2.5, rate=1), a lognormal distribution (mean=1.119, standard deviation=0.602) and a Weibull distribution (shape=0.8, scale=1) were used to generate populations with a size of N = 20,940. As stated by Salem and Mount (1974) and McDonald (1984), these distributions are good approximations of income distributions. For the Turkish EU-SILC survey data, the distribution of income variables are approximated by a lognormal distribution with mean of 8.626 and standard deviation of 0.736. The correlation coefficient between the variables of the first and the second wave is given by $\rho = 0.94$, which is the correlation observed from the common sample of the Turkish EU-SILC survey data. Note that this correlation and the correlation in (2.1) are different; in other words, the correlation $\rho = 0.94$ is the population correlation between income variables measured at two waves, whereas the correlation in (2.1) is the correlation between the point estimators of poverty rates.

The population is assumed fixed and the same sample size was used for both waves. We have 1047 primary sampling units in the Turkish EU-SILC survey data. For this reason, we used $n_1 = n_2 = 1047$ units for each wave. The fraction of the common sample is 75%. Hence, the number of units in the common sample is $n_c = 785$. Unequal and equal probabilities were used to select the samples. For unequal probability sampling (πps) design, the Chao (1982) sampling design was used. The first wave samples were selected without replacement with the inclusion probabilities proportional to a size variable x_i,

which was generated by the model $x_i = \alpha + \rho y_{1;i} + e_i$, with $e_i \sim N(0, (1-\rho^2)\sigma^2_{y_1})$, $\alpha = 5$, and $\rho = 0.7$, which represents the correlation between x_i and $y_{1;i}$. The inclusion probabilities at first wave are given by $\pi_{1;i} = n_1 x_i / \sum_{i \in U_1} x_i$, where U_1 is the population at first wave. As we consider a fixed size sampling design, we have $\sum_{i \in U_1} \pi_{1;i} = n_1$. For the second wave, a simple random sample of n_c units were selected from the sample s_1; and, $n_2 - n_c$ units are selected with the probabilities proportional to size $q_i = \pi_{1;i}/(1-\pi_{1;i})$ from the population $U \backslash s_1$. It can be shown that $\pi_{2;i} \approx \pi_{1;i}$ (Christine and Rocher, 2012). For equal probability sampling designs, $\pi_{2;i} = \pi_{1;i} = n_1/N$.

We did six simulation studies for three populations and two sampling designs. For each simulation, 10,000 samples were selected. For each sample, the RB and the RRMSE were computed for the cross sectional variance estimators, the variance estimator of change and the estimator of the correlation. The RB and the RRMSE are defined by

$$RB(\widehat{\sigma}) = \frac{E(\widehat{\sigma}) - \sigma}{\sigma} 100\%, \tag{2.11}$$

$$RRMSE(\widehat{\sigma}) = \frac{\sqrt{(B-1)^{-1} \sum_{b=1}^{B} (\widehat{\sigma}_b - \sigma)^2}}{\sigma} 100\%, \tag{2.12}$$

where $E(\widehat{\sigma}) = B^{-1} \sum_{b=1}^{B} \widehat{\sigma}_b$, with $B = 10{,}000$, is the empirical expectation; σ is either the empirical variances or the empirical correlation in (2.1); $\widehat{\sigma}$ is the estimator of the quantity σ; $\widehat{\sigma}_b$ is the estimate of the quantity σ for the bth sample. For the linearisation, we considered three bandwidth parameters (see expressions (2.6)–(2.8)). The linearisation based on (2.6), (2.7) and (2.8) are respectively labelled as Lin_Sd, Lin_Iqr, and Lin_A in Tables 2.1–2.2.

For a gamma distribution, the poverty rates are 24.2% and 23.6% for the first and the second wave respectively. Hence, we have -0.59% point change between two waves. For a lognormal distribution, the poverty rates are 19.4% and 19.9%. Thus there is a 0.54% point change for this case. For a Weibull distribution, we have the highest poverty rates, which are 36.6% and 37.3% respectively. Hence, the change is 0.66% point.

Table 2.1 shows the RB (%) of the variance and the correlation estimators for several distributions and sampling designs. Overall, the linearisation approach has lower RB than the ratio one. Thus we have more accurate estimates with the linearisation. Differences between the two approaches in terms of the RB is much more pronounced for the Weibull distribution, which is the most skewed distribution. For all situations except with the lognormal distribution, the ratio approach overestimates all the variances and the correlations. Therefore, the ratio approach may not always give more conservative estimates. However, note that whenever we have a positive bias, we obtain relatively

Table 2.1: Empirical RB (%) of the variance and correlation estimators for the poverty rates for three distributions and two sampling designs

	Relative Bias (%)							
	Gamma							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	41.3	2.4	2.6	3.1	50.9	7.2	7.4	7.8
Var Wave2	42.8	5.1	5.3	5.8	41.1	2.9	3.0	3.5
Var Change	8.1	1.0	1.2	1.8	13.0	2.1	2.4	2.9
Correlation	23.2	2.6	2.6	2.5	22.0	2.7	2.6	2.5
	Lognormal							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	15.6	0.9	2.2	2.9	22.7	-0.5	0.5	1.0
Var Wave2	24.1	6.4	7.6	8.2	28.9	4.2	5.1	5.6
Var Change	-14.1	1.3	2.6	3.4	-8.7	0.5	1.7	2.4
Correlation	38.1	3.1	2.9	2.8	35.5	1.6	1.3	1.1
	Weibull							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	140.1	4.3	6.5	6.7	132.9	2.9	4.8	5.1
Var Wave2	146.0	1.9	4.0	4.2	137.6	1.0	2.9	3.1
Var Change	26.6	0.9	4.2	4.4	28.3	2.0	5.2	5.5
Correlation	152.0	6.6	3.3	3.3	132.1	-0.4	-3.8	-3.8

larger variance estimates with the ratio approach. When we compare the three linearisation methods based on different bandwidth, we obtained the largest RB with the smallest bandwidth (see expression (2.8)).

As far as the RRMSE is concerned (see Table 2.2), we have more precise estimates with the linearisation approach. We observe the smallest RRMSE with the bandwidth (2.6) and the largest RRMSE with the bandwidth (2.8). The ratio approach gives less accurate point estimates. The differences between the two approaches are notable, especially with the Weibull distribution.

2.7 An application to the Turkish EU-SILC survey

The 2007 and 2008 cross sectional Turkish EU-SILC survey data was used. The Turkish EU-SILC survey has a stratified two-stage cluster probability sampling design. For the first stage, address blocks are selected within each stratum with a probability proportional to size (πps) without replacement sampling design. Each block is composed of

Table 2.2: Empirical RRMSE (%) of the variance and correlation estimators for the poverty rates for three distributions and two sampling designs

	Relative Root Mean Square Error (%)							
	Gamma							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	41.5	4.8	5.1	5.9	51.2	8.2	8.4	9.0
Var Wave2	36.8	6.8	7.1	7.9	41.4	4.9	5.1	5.9
Var Change	10.9	7.9	8.1	8.6	15.4	8.7	8.8	9.4
Correlation	20.0	6.0	6.5	6.5	22.7	7.3	7.3	7.3
	Lognormal							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	16.4	4.9	6.2	7.2	30.6	7.8	8.2	8.7
Var Wave2	24.6	8.1	9.8	10.8	35.2	8.9	9.8	10.5
Var Change	15.1	7.0	8.0	8.8	18.5	10.6	11.2	11.7
Correlation	38.5	7.1	7.0	7.0	37.4	11.1	11.0	11.0
	Weibull							
	SRS				πps			
	Ratio	Lin_Sd	Lin_Iqr	Lin_A	Ratio	Lin_Sd	Lin_Iqr	Lin_A
Var Wave1	140.1	6.5	8.5	9.0	133.0	6.5	8.0	8.5
Var Wave2	146.0	5.5	7.1	7.7	137.7	6.2	7.4	7.9
Var Change	27.1	5.7	7.8	8.4	29.1	7.0	9.3	9.9
Correlation	152.2	16.7	16.3	16.6	132.4	16.2	17.2	17.5

approximately 100 addresses. Households within the selected address blocks are selected with a systematic sampling design. All individuals within the selected households participate in the survey. The cross sectional survey weights in the "personal register" file (RB050) were used as inverses of the inclusion probabilities. The effect of calibration was not taken into account, because we did not have any information about the auxiliary variables. The effect of imputation was ignored for the same reason.

In Table 2.3, we have the estimates for several domains when the poverty threshold is treated as fixed (see expression (2.2)). We observe a significant change for the domain "tenant" at the 5% level.

In Table 2.4, we have the estimates obtained with the linearisation approach based on the bandwidth in (2.6) described in Section 2.4. We also observe a highly significant change for the domain "tenant". We do not observe major differences in the p-values between Table 2.3 and Table 2.4. We observe a slight decrease in the p-values when the sampling variation of the poverty threshold is taken into account. This is due to the fact that the variances of changes are larger in Table 2.3.

The correlations in Table 2.4 are smaller than in Table 2.3 overall. Hence, the estimated correlations are smaller when the variability of the poverty threshold is taken into account.

By comparing Table 2.3 and Table 2.4, we also found that all variances were estimated more conservatively when the threshold is treated as fixed. Preston (1995), Berger and Skinner (2003), and Verma and Betti (2011) demonstrated that the cross sectional variances are more conservative when the poverty threshold is treated as fixed. This finding was explained by Preston (1995) by the fact that the two sources of variability offset each other. This is more pronounced when the high fractions of the median are used. For the variance of change, we cannot anticipate an increase in the variance when

Table 2.3: Estimates when the poverty threshold is treated as fixed (see expression (2.2))

Domain	Pov'07(%)	Var'07	Pov'08	Var'08(%)	Change(in % point)	Var Change	Corr	p-value
Turkey	23.4	0.616	24.1	0.644	0.7	0.447	0.65	0.297
Male	23.0	0.650	23.7	0.665	0.7	0.494	0.62	0.328
Female	23.8	0.639	24.6	0.678	0.7	0.465	0.65	0.299
Owner	24.9	0.739	23.8	0.872	-1.1	0.593	0.63	0.140
Tenant	18.5	1.395	25.3	1.511	6.7	1.522	0.48	0.000
0_14	33.5	1.164	34.5	1.258	1.1	0.882	0.64	0.263
15_24	24.2	1.162	25.3	1.181	1.1	1.118	0.52	0.296
25_49	19.8	0.527	20.7	0.548	0.9	0.405	0.62	0.178
50_64	14.4	0.568	15.0	0.719	0.6	0.569	0.56	0.404
65+	17.7	1.077	16.2	0.929	-1.5	0.988	0.51	0.120

Source: 2007 and 2008 cross sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

the poverty threshold is treated as fixed for the following reason. Let us assume that the cross sectional variances are equal: $\widehat{\text{var}}(\widehat{\theta}_1) = \widehat{\text{var}}(\widehat{\theta}_2)$. Thus the variance estimator of change is given by $\widehat{\text{var}}(\widehat{\Delta}) = 2\,\widehat{\text{var}}(\widehat{\theta}_1)\,(1 - \widehat{\text{corr}}(\widehat{\theta}_1, \widehat{\theta}_2))$. Hence, the variance of change is affected in the same direction by the variance term, and in the opposite direction by the correlation term. Thus when both the variance and the correlation terms increase or decrease concurrently, the direction of the effect on the variance of change cannot be predicted. Therefore, we may not necessarily have more conservative estimates of the variance of change when the poverty threshold is treated as fixed. With the Turkish EU-SILC survey data, we found that the variances of changes were more conservative, although the differences between the two approaches were not as pronounced as the differences between the cross sectional variances (see Table 2.3 and Table 2.4).

In Table 2.4, the bandwidth parameter is given by expression (2.6). We also investigate the situations when the bandwidth parameter is given by expressions (2.7) and (2.8).

Table 2.4: Estimates when the sampling variation of the poverty threshold is taken into account (see Sections 2.4 and 2.5). The bandwidth parameter is based on the standard deviation of the income distribution (see definition (2.6)).

Domain	Pov'07(%)	Var'07	Pov'08	Var'08(%)	Change(in % point)	Var Change	Corr	p-value
Turkey	23.4	0.292	24.1	0.290	0.7	0.372	0.36	0.252
Male	23.0	0.314	23.7	0.306	0.7	0.416	0.33	0.287
Female	23.8	0.327	24.6	0.327	0.7	0.390	0.40	0.257
Owner	24.9	0.417	23.8	0.495	-1.1	0.527	0.42	0.117
Tenant	18.5	1.121	25.3	1.238	6.7	1.435	0.39	0.000
0_14	33.5	0.796	34.5	0.793	1.1	0.787	0.50	0.236
15_24	24.2	0.790	25.3	0.919	1.1	1.050	0.39	0.281
25_49	19.8	0.255	20.7	0.252	0.9	0.362	0.29	0.154
50_64	14.4	0.403	15.0	0.491	0.6	0.476	0.47	0.361
65+	17.7	0.929	16.2	0.807	-1.5	0.978	0.44	0.118

Source: 2007 and 2008 cross sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

The results are given in Table 2.5 and Table 2.6. By comparing Table 2.5 and Table 2.6 with Table 2.3, we also observed smaller cross sectional variances, variance of change and correlation with the bandwidth parameters (2.7) and (2.8). When we compare Tables 2.4–2.6 with each other, the estimates do not differ significantly between the three linearisation approaches based on different bandwidth parameters, although we observe slight differences between them in terms of the RB and the RRMSE in the simulation study (see Section 2.6).

2.8 Conclusion

We applied a approach, easy to implement, to estimate the variances of changes for the poverty rates over several domains by using the 2007-2008 Turkish EU-SILC survey data. It involves a multivariate linear regression proposed by Berger and Priam (2010, 2016), which can be easily applied. Survey characteristics such as rotation, stratification, and cluster sampling are all taken into account. The approach proposed is flexible and can be implemented for most of the EU-SILC surveys as long as sampling fractions are negligible. This assumption implies that the second order inclusion probabilities are not needed.

We have two ways to estimate the variances depending on whether we treat the poverty threshold as fixed or not. When treated as fixed, we obtained more conservative variance estimates of change with the Turkish EU-SILC survey data. However, our simulation study shows that treating the threshold as fixed does not necessarily give more conservative variance estimates of change. For the lognormal distribution, for example, variances

Table 2.5: Estimates when the sampling variation of the poverty threshold is taken into account (see Sections 2.4 and 2.5). The bandwidth parameter is based on the inter quartile range of the income distribution (see definition (2.7)).

Domain	Pov'07(%)	Var'07	Pov'08	Var'08(%)	Change(in % point)	Var Change	Corr	p-value
Turkey	23.4	0.292	24.1	0.290	0.7	0.372	0.36	0.252
Male	23.0	0.316	23.7	0.306	0.7	0.416	0.33	0.287
Female	23.8	0.325	24.6	0.328	0.7	0.391	0.40	0.257
Owner	24.9	0.418	23.8	0.497	-1.1	0.530	0.42	0.118
Tenant	18.5	1.117	25.3	1.226	6.7	1.428	0.39	0.000
0_14	33.5	0.802	34.5	0.814	1.1	0.805	0.50	0.241
15_24	24.2	0.787	25.3	0.907	1.1	1.038	0.39	0.278
25_49	19.8	0.256	20.7	0.251	0.9	0.361	0.29	0.154
50_64	14.4	0.403	15.0	0.491	0.6	0.476	0.47	0.361
65+	17.7	0.946	16.2	0.791	-1.5	0.976	0.44	0.118

Source: 2007 and 2008 cross sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

Table 2.6: Estimates when the sampling variation of the poverty is threshold taken into account (see Sections 2.4 and 2.5). The bandwidth parameter is based on the parameter A (see definition (2.8)).

Domain	Pov'07(%)	Var'07	Pov'08	Var'08(%)	Change(in % point)	Var Change	Corr	p-value
Turkey	23.4	0.291	24.1	0.291	0.7	0.372	0.36	0.253
Male	23.0	0.316	23.7	0.306	0.7	0.416	0.33	0.287
Female	23.8	0.324	24.6	0.329	0.7	0.392	0.40	0.258
Owner	24.9	0.419	23.8	0.498	-1.1	0.531	0.42	0.119
Tenant	18.5	1.114	25.3	1.223	6.7	1.425	0.39	0.000
0_14	33.5	0.802	34.5	0.823	1.1	0.812	0.50	0.243
15_24	24.2	0.787	25.3	0.903	1.1	1.034	0.39	0.277
25_49	19.8	0.255	20.7	0.251	0.9	0.361	0.29	0.154
50_64	14.4	0.403	15.0	0.491	0.6	0.476	0.47	0.361
65+	17.7	0.949	16.2	0.788	-1.5	0.977	0.44	0.118

Source: 2007 and 2008 cross sectional data of the EU-SILC survey for Turkey conducted by TurkStat.

of changes were underestimated with the ratio method. On the other hand, differences between the variance estimators of changes can be negligible in terms of the RB and the RRMSE even though we observed significant differences between the cross sectional variances and the correlations. For the latter, the linearisation approach gave less biased and more precise variance estimates. Thus based upon our results and due to the fact that linearisation involves complex numerical computations, the simple ratio approach

may sound preferable to estimate the variance of change for the poverty rates. Albeit, we should be careful with highly skewed distributions similar to a Weibull one. As in this case, the linearisation approach is significantly better. For the Turkish EU-SILC survey, the ratio approach can be used as where the distribution of income variables are approximated by lognormal distribution.

The approach proposed can also be used to estimate the variances of the other poverty and income inequality measures such as the relative median at risk of poverty gap (RMPG), the quantile share ratio (QSR) and the GINI coefficient, which are included in the "Laeken" indicators (Eurostat, 2003), by using linearisation (e.g. Berger, 2008). The RMPG and the GINI coefficient cannot be treated as a simple ratio, whereas the QSR can be. The linearised variables of many complex parameters are given by Verma and Betti (2005, 2011).

In this paper, we implemented the fixed-bandwidth kernel method for its simplicity (Silverman, 1986, p.95). The bandwidth given by expression (2.8) is a suitable choice for a wide range of densities as pointed out by Silverman (1986). If the distribution is heavily skewed, then, an adaptive kernel method can be applied (Silverman, 1986, chap.5). This method uses a variable bandwidth, that is, for each observed data point, a different bandwidth is computed. It would be interesting to check if an adaptive bandwidth improved the variance estimation in the presence of outliers.

Chapter 3

Second Paper

Modelling survey data under complex sampling designs and population level information: an empirical likelihood based approach

Melike Oguz-Alper and Yves G. Berger
University of Southampton, SO17 1BJ, Southampton, U.K.
M.OguzAlper@soton.ac.uk Y.G.Berger@soton.ac.uk

Abstract

Survey data are often collected with unequal probabilities from a stratified population. Suppose we wish to fit a model to such survey data. We consider a design-based inference for model parameters defined through population estimating equations. In many modelling situations, the parameter of interest is a subset of a set of parameters, with the others treated as nuisance parameters. For example, when comparing two nested models, the additional parameters are the parameters of interest, and the common parameters are the nuisance parameters. We propose using a profile empirical log-likelihood ratio function that minimises the empirical log-likelihood ratio function with respect to the nuisance parameters. We show that the profile empirical log-likelihood ratio function, follows a χ^2-distribution asymptotically. This can be used to make inference for the subparameter of interest. For example, it can be used to test if two nested models are significantly different and to construct confidence intervals. We show how the approach proposed can be generalised to incorporate population level information, stratification and clustering. The approach is simple to implement and less computer intensive than the bootstrap. The confidence interval proposed does not rely on resampling, linearisation, variance estimation, or design effect.

Keywords: Design based inference; estimating equations; empirical likelihood; regression parameters; unequal inclusion probabilities.

3.1 Introduction

Statistical models are widely used in social sciences, biological sciences, econometrics and finance. Suppose we wish to fit a model to sample data selected randomly with unequal probabilities. In this case, we do not have *independent and identically distributed* (i.i.d) observations. When the random selection of the sample (or sampling design) is ignored, estimators based on the assumption of i.i.d observations may be inconsistent and may produce invalid inferences especially when the sampling design is informative (e.g. Pfeffermann and Sverchkov, 1999, 2003). We consider a general class of multidimensional parameters defined as the solution of a set of estimating equations (see equation (3.2)). For example, this class includes complex parameters such as (non)linear regression parameters, generalised linear regression parameters, ratios, proportions or means. We consider a semiparametric approach, where the model is specified by estimating equations (see equation (3.2)). We consider that the underlying distribution is unknown. Hence, parametric likelihood is not feasible.

Let U be a finite population of N units labelled $i = 1, \dots N$. Consider that n units are selected independently with replacement with unequal probabilities p_i (e.g. Hansen and Hurwitz, 1943) from U, where $\sum_{i \in U} p_i = 1$. Let

$$\pi_i = np_i, \tag{3.1}$$

where n is the fixed number of draws. The sampling design may also be stratified (see Section 3.9.3). Note that n is different from the number of distinct units selected, because the units are selected with replacement. Let $\bar{\pi}$ denote the sampling fraction or the mean of the π_i: $\bar{\pi} = N^{-1} \sum_{i \in U} \pi_i = n/N$. The approach proposed is also valid under without-replacement sampling with first-order inclusion probabilities given by (3.1), when n/N is negligible, as sampling with and without-replacement are asymptotically equivalent in this case. This is usually the case with social surveys.

Let s denote the sample containing the labels of the units selected after n draws. The probability distribution of s is called *the sampling design* and is denoted by $\mathcal{P}(s)$. Let $\boldsymbol{v}_i$ be a vector that contains the values of a set of variables for a unit $i \in U$. The sample data, given by $\{\boldsymbol{v}_i : \ i \in s\}$, is a set of not identically distributed observations, because the sample is selected with unequal probabilities. We consider that the $\boldsymbol{v}_i$ are fixed (nonrandom) vectors. This setting is often called the *design based approach*, as the sampling distribution of the sample data is solely driven by the random selection of the sample (Neyman, 1934).

Consider that the parameter of interest is a sub-vector of the parameters. The remaining parameters, which are not part of the parameter of interest, are called '*nuisance*' parameters. The nuisance parameters are unknown by definition. There are numerous situations when nuisance parameters are used (e.g. Binder and Patak, 1994; Qin and

Lawless, 1994; Godambe and Thompson, 2009; Zheng et al., 2012). For example, with a regression estimator of a mean, the regression parameter is a nuisance parameter. When testing if the slope of a simple regression is significant, the intercept is the nuisance parameter. When comparing two nested models, we want to test if the additional parameters are significant. The nuisance parameters are the parameters of the more parsimonious model and the parameters of interest are the additional parameters. When we have heteroscedasticity, we may assume that the residual variance is a function of nuisance parameters (e.g. Owen, 1991, p.1740). Another example is the correlation coefficient, where the means and variances are the nuisance parameters (e.g. Owen, 1990, 2001).

Standard design based approaches involve linearisation (Binder, 1983; Deville, 1999) or resampling techniques for variance estimation. Standard variance estimators treat the nuisance parameters as fixed values given by their estimates. For example, standard variance estimators of the regression estimator ignore the randomness of the regression parameter, despite the fact that it is an estimate. Furthermore, variance estimators may be biased when the randomness of the nuisance parameter is ignored.

Confidence intervals are usually constructed by assuming that point estimators have a normal distribution. However, point estimators may not have a normal distribution, when we have outlying observations, and linearised variance estimators may be biased for a moderate sample size. Empirical likelihood inference does not rely on variance estimates or on the normality of the point estimator. It can also handle nuisance parameters; that is, empirical likelihood confidence intervals takes into account the randomness of the nuisance parameters.

The empirical likelihood approach is a well established topic under the classical i.i.d framework (e.g. Owen, 1991; Qin and Lawless, 1994; Wu and Sitter, 2001; Chaudhuri et al., 2008; Kim and Zhou, 2008; Zheng et al., 2012). The paper by Chen and Keilegom (2009) includes an elaborate review of the existing literature. Chen and Sitter (1999) proposed a pseudoempirical likelihood under unequal probability sampling. Wu and Rao (2006) showed that pseudoempirical likelihood confidence intervals can be constructed by using the design effect, which is estimated from variance estimates. Kim (2009) and Chen and Kim (2014) proposed an empirical likelihood approach under Poisson sampling. Berger and De La Riva Torres (2016) proposed an empirical likelihood approach for unequal probability sampling. This approach does not rely on design effects, variance estimates, linearisation or resampling. It deals with a wide range of nonlinear finite population parameters.

Profiling consists in minimising the empirical log-likelihood ratio function over the nuisance parameters. This technique allows us to test and derive confidence intervals for the sub-parameters of interest. Qin and Lawless (1994) showed that the profile empirical log-likelihood ratio function follows a χ^2-distribution with i.i.d observations. Qin

and Lawless's (1994) approach does not directly apply because it does not take into account the sampling design and the unequal probabilities. One of our aims is to fill this gap. Berger and De La Riva Torres's (2016) approach is limited to single parameters and cannot be straightforwardly extended for profiling. We show that the empirical log-likelihood ratio function that was proposed by Berger and De La Riva Torres (2016) can be profiled out over the nuisance parameter. We show that the profile empirical log-likelihood ratio function follows a χ^2-distribution asymptotically. The approach proposed is different from Qin and Lawless's (1994) approach, because we shall take the information about the sampling design into account, and we assume that the observations are selected with unequal probabilities.

Binder and Patak (1994) proposed a nonparametric version of the likelihood based score statistics that can be used with nuisance parameters (see also Godambe and Thompson, 1999, p.162). They proposed a method of inverse testing to construct confidence intervals. They used a pivotal statistic defined by the square of the estimating function divided by its variance. The bounds of a confidence interval are the solutions of a system of equations that can be solved numerically. Godambe and Thompson (2009, p.92) pointed out that solutions may not exist. Furthermore, this approach relies on variance estimates.

Chen and Sitter (1999) and Zhong and Rao (2000) proposed an algorithm based on profiling the pseudoempirical likelihood ratio function when stratum totals of auxiliary variables are unknown. For confidence intervals, the pseudoempirical likelihood needs to be adjusted by variance estimates. This approach is limited to estimation of totals. There is no general theory on profiling for the pseudoempirical likelihood approach.

Pfeffermann and Sverchkov (1999, 2003) considered a semiparametric approach that may require modelling the survey weights. The variances of the model parameters are estimated through linearisation or resampling techniques (e.g. Pfeffermann and Sverchkov, 1999). In Section 3.10, we shall compare numerically the approach that we propose with the semiparametric approach that was proposed by Pfeffermann and Sverchkov (1999, 2003).

Inferences about the population parameters can be improved by incorporating population level information, which may be available from administrative data, census data and/or population projections (e.g. Deville and Särndal, 1992; Chaudhuri et al., 2008). The empirical likelihood approach proposed allows incorporation of population level information in the presence of nuisance parameters.

In Section 3.2, we define the population parameter and provide some examples regarding regression parameters (see Section 3.2.1). In Section 3.3, we introduce the empirical log-likelihood function under unequal probability selection proposed by Berger and De La Riva Torres (2016). In Section 3.4, the maximum empirical likelihood estimator is defined. In Section 3.5, we define the profile empirical log-likelihood ratio function and

show how it can be used for testing and constructing confidence intervals (see Section 3.5.1). In Section 3.6, we provide an algorithm to compute the profile empirical log-likelihood ratio function. Section 3.7 describes the large sample properties of the profile empirical log-likelihood ratio function. In Section 3.8, stratification is incorporated with the approach proposed. In Sections 3.9–3.9.2, we show how to incorporate population level information. In Section 3.9.3, we provide a trivial extension of the approach proposed to stratified multi stage sampling designs in the presence of population level information. Simulation results are presented in Sections 3.10–3.10.5. The supplementary material including asymptotic derivations and the R (R Development Core Team, 2014) code is provided in Appendix B.

3.2 Parameters and estimating equations

The parameter $\boldsymbol{\psi}_N \in \boldsymbol{\Psi} \subset \mathbb{R}^b$ $(b = O(1))$ is the $b \times 1$ finite population vector that is the unique solution of the population estimating equation (3.2) (Godambe, 1960), where $\boldsymbol{\Psi}$ is compact.

$$\boldsymbol{G}(\boldsymbol{\psi}) = \sum_{i \in U} \boldsymbol{g}_i(\boldsymbol{v}_i, \boldsymbol{\psi}) = \mathbf{0}_b; \tag{3.2}$$

where $\boldsymbol{g}_i(\boldsymbol{v}_i, \boldsymbol{\psi})$ is a $b \times 1$ vector of estimating functions and $\boldsymbol{v}_i$ is the vector of variables for unit i. Here, $\mathbf{0}_b$ is a $b \times 1$ vector of zeros. For simplicity, we replace $\boldsymbol{g}_i(\boldsymbol{v}_i, \boldsymbol{\psi})$ by $\boldsymbol{g}_i(\boldsymbol{\psi})$. As we consider a design based approach, the parameter $\boldsymbol{\psi}_N$ is a fixed (nonrandom) unknown quantity. The parameter $\boldsymbol{\psi}_N$ shall be estimated from the sample data. The $\boldsymbol{g}_i(\boldsymbol{\psi})$ may also depend on some known population parameters (see Sections 3.9–3.9.3).

Most finite population parameters can be defined by estimating equations (e.g. Binder, 1983; Binder and Patak, 1994; Qin and Lawless, 1994; Godambe and Thompson, 2009). For example, $\boldsymbol{\psi}_N$ can be a vector of population means, totals, ratios, quantiles, low income measures or regression coefficients.

3.2.1 Examples: regression parameters

Consider $\boldsymbol{v}_i = (y_i, \boldsymbol{x}_i{}^{\mathrm{T}})^{\mathrm{T}}$, where y_i is a scalar response variable and $\boldsymbol{x}_i$ are some explanatory variables. Consider a nonlinear model with a smooth model (scalar) function $\mu(\cdot)$. The parameter $\boldsymbol{\psi}_N$ can be the nonlinear least squares parameter defined by

$$\boldsymbol{\psi}_N = \arg\min_{\boldsymbol{\psi} \in \boldsymbol{\Psi}} \sum_{i \in U} \sigma_i^{-1} \left\{ y_i - \mu(\boldsymbol{h}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\psi}) \right\}^2, \tag{3.3}$$

where $\boldsymbol{h}(\cdot)$ is a known vector function and σ_i is a known variance function. In this case, $\boldsymbol{\psi}_N$ is the solution of the estimating equation (e.g. Chen and Keilegom, 2009).

$$\sum_{i \in U} \frac{\partial(\boldsymbol{h}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\psi})}{\partial \boldsymbol{\psi}} \left\{ y_i - \mu(\boldsymbol{h}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\psi}) \right\} \sigma_i^{-2} = \mathbf{0}_b \tag{3.4}$$

For ordinary least squares parameters, we have $\boldsymbol{h}(\boldsymbol{x}_i) = \boldsymbol{x}_i$, $\mu(\boldsymbol{h}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\psi}) = \boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\psi}$ and $\sigma_i^2 = \sigma$, for some σ. Hence, equation (3.4) reduces to the normal equation.

$$\sum_{i \in U} \boldsymbol{x}_i (y_i - \boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\psi}) = \mathbf{0}_b \cdot \tag{3.5}$$

Equation (3.5) can be extended to include instrumental variables.

For generalised linear models, we have $\mu(\boldsymbol{h}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\psi}) = \mathcal{F}^{-1}(\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\psi})$, where $\mathcal{F}(\cdot)$ is a link function. For example, $\mathcal{F}(\mu) = \log(\mu(1-\mu)^{-1})$ with a logistic regression model. In this case, equation (3.4) reduces to (e.g. Binder, 1983, p.285)

$$\sum_{i \in U} \boldsymbol{x}_i \left\{ y_i - \frac{exp(\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\psi})}{1 + exp(\boldsymbol{x}_i{}^{\mathrm{T}}\boldsymbol{\psi})} \right\} = \mathbf{0}_b \cdot \tag{3.6}$$

3.3 Empirical log-likelihood function for unequal probabilities

In this Section, we recall the approach that was proposed by Berger and De La Riva Torres (2016). Consider the following *empirical likelihood function.*

$$L(\boldsymbol{m}) = \prod_{i \in s} m_i, \tag{3.7}$$

where the m_i are unknown scale-loads allocated to data points $i \in s$ (Hartley and Rao, 1968) and $\boldsymbol{m}$ is the $n \times 1$ vector of the m_i ($i \in s$). Hartley and Rao (1969) showed that expression (3.7) is the empirical likelihood function for unequal probability sampling with replacement (see also Kim, 2009; Berger and De La Riva Torres, 2016)

Let $\widehat{m}_i$ maximise the following *empirical log-likelihood function*

$$\ell(\boldsymbol{m}) = \sum_{i \in s} \log(m_i), \tag{3.8}$$

with respect to the constraints: $m_i > 0$ and

$$\sum_{i \in s} m_i \boldsymbol{c}_i = \boldsymbol{C}, \quad \text{with } \boldsymbol{C} = \sum_{i \in U} \boldsymbol{c}_i \cdot \tag{3.9}$$

Here, the $\boldsymbol{c}_i$ are r-vectors, where $r = O(1)$. When we have a single stratum and when we do not use any population level information, $\boldsymbol{c}_i$ is defined by $\boldsymbol{c}_i = \bar{\pi}^{-1}\pi_i$ and $r = 1$, where $\bar{\pi} = n/N$. The constant $\bar{\pi}^{-1}$ can be removed from $\boldsymbol{c}_i$ because it cancels out in equation (3.9). The $\boldsymbol{c}_i$ may also include population level information or stratification variables. In this case, $r > 1$ and the $\boldsymbol{c}_i$ are defined in a different way (see Sections 3.8–3.9.3).

We assume that the $\boldsymbol{C}$ is an inner point of the conical hull formed by $\sum_{i\in s} m_i\boldsymbol{c}_i$, so that the set of $\widehat{m}_i$ is unique. We assume that $\boldsymbol{c}_i$ and $\boldsymbol{C}$ satisfy the regularity conditions (3.28)–(3.33) given (see Section 3.7). We also assume that there exists an r-vector $\boldsymbol{t}$ such that $\boldsymbol{t}^{\mathrm{T}}\boldsymbol{c}_i = \pi_i$. By using (3.9), we have that $\sum_{i\in s} m_i\pi_i = n$, which specifies the fact that a sample of size n is selected.

Berger and De La Riva Torres (2016) showed that by using the method of Lagrange multipliers, we have

$$\widehat{m}_i = (\pi_i + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{c}_i)^{-1}, \tag{3.10}$$

where the vector $\boldsymbol{\eta}$ is such that (3.9) and $\widehat{m}_i > 0$ hold. A modified Newton-Raphson algorithm as in Chen et al. (2002) can be used to compute $\boldsymbol{\eta}$.

The $\boldsymbol{c}_i$ incorporate the information about the sampling design and the population level information (see Sections 3.9– 3.9.3). When we do not use any population level information, we use $\boldsymbol{c}_i = \bar{\pi}^{-1}\pi_i$. Then it can be shown that $\boldsymbol{\eta} = \boldsymbol{0}_r$ and $\widehat{m}_i = \pi_i^{-1}$ which is the standard Horvitz and Thompson (1952) weight for unit i. The definitions of $\boldsymbol{c}_i$ and $\boldsymbol{C}$ have to be modified with population level information (see Sections 3.9– 3.9.3) and under stratified sampling (see Sections 3.8 and 3.9.3).

3.4 Maximum empirical likelihood point estimator

Let $\widehat{m}_i^*(\boldsymbol{\psi})$ maximise $\ell(\boldsymbol{m})$ subject to the constraints $m_i > 0$ and

$$\sum_{i\in s} m_i\ \boldsymbol{c}_i^*(\boldsymbol{\psi}) = \boldsymbol{C}^* \tag{3.11}$$

with

$$\boldsymbol{c}_i^*(\boldsymbol{\psi}) = (\boldsymbol{c}_i{}^{\mathrm{T}}, \boldsymbol{g}_i(\boldsymbol{\psi})^{\mathrm{T}})^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{C}^* = (\boldsymbol{C}^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}}, \tag{3.12}$$

for a given vector $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}$; where $\boldsymbol{g}_i(\boldsymbol{\psi})$ is defined in Section 3.2. We assume that $\boldsymbol{c}_i^*(\boldsymbol{\psi})$ is differentiable with respect to $\boldsymbol{\nu}$ for all $i \in s$ in a neighbourhood around the true

population value $\boldsymbol{\nu}_N$. The maximum value of $\ell(\boldsymbol{m})$ under $m_i > 0$ and (3.11) is given by

$$\ell(\boldsymbol{\psi}) = \sum_{i \in s} \log\left(\widehat{m}_i^*(\boldsymbol{\psi})\right). \tag{3.13}$$

The *maximum empirical likelihood estimator* $\widehat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}_N$ is the vector that maximises $\ell(\boldsymbol{\psi})$ over $\boldsymbol{\psi}$. Berger and De La Riva Torres (2016) showed that $\widehat{\boldsymbol{\psi}}$ is the unique solution of the sample estimating equation.

$$\widehat{\boldsymbol{G}}(\boldsymbol{\psi}) = \mathbf{0}_b, \quad \text{where} \quad \widehat{\boldsymbol{G}}(\boldsymbol{\psi}) = \sum_{i \in s} \widehat{m}_i \; \boldsymbol{g}_i(\boldsymbol{\psi}), \tag{3.14}$$

where $\widehat{m}_i$ is given by (3.10). We assume that the $\boldsymbol{g}_i(\boldsymbol{\psi})$ is such that equation (3.14) has a unique solution.

When $\widehat{m}_i = \pi_i$, $\widehat{\boldsymbol{G}}(\boldsymbol{\psi})$ is the Horvitz-Thompson estimator of $\boldsymbol{G}(\boldsymbol{\psi})$, for a given $\boldsymbol{\psi}$, and the estimator $\widehat{\boldsymbol{\psi}}$ is the *pseudo likelihood estimator* that was proposed by Binder (1983). In this case, the sample estimate $\widehat{\boldsymbol{\psi}}$ is design consistent (e.g Godambe and Thompson, 2009).

3.5 Profile empirical log-likelihood ratio function in the presence of nuisance parameters

Suppose that we would like to make inference about a $p \times 1$ sub-parameter $\boldsymbol{\theta}_N \in \boldsymbol{\Theta} \subset \mathbb{R}^p$, where $p < b$. Consider $\boldsymbol{\psi}_N = ({\boldsymbol{\theta}_N}^{\mathrm{T}}, {\boldsymbol{\nu}_N}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\nu}_N$ is a $q \times 1$ sub-parameter ($\boldsymbol{\nu}_N \in \boldsymbol{\Lambda} \subset \mathbb{R}^q$) that are not of primary interest ($q = b - p$), where $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$ are compact set. The parameter $\boldsymbol{\nu}_N$ is assumed unknown and may need to be estimated when making inferences about $\boldsymbol{\theta}_N$. In this paper, the parameter $\boldsymbol{\nu}_N$ is called the *nuisance parameter*. In this Section, we assume that we do not use any population level information. In Sections 3.9–3.9.2, we extend this Section's approach for population level information. We propose to test and construct a confidence region for the parameter of interest $\boldsymbol{\theta}_N$ by using the *profile empirical log-likelihood ratio function* defined by

$$\widehat{r}(\boldsymbol{\theta}) = 2\left\{\ell(\widehat{\boldsymbol{\psi}}) - \max_{\boldsymbol{\nu} \in \boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}, \boldsymbol{\nu})\right\}, \tag{3.15}$$

where $\ell(\boldsymbol{\theta}, \boldsymbol{\nu}) = \ell(\boldsymbol{\psi})$ with $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}$. The $\widehat{r}(\boldsymbol{\theta})$ is a random function of $\boldsymbol{\theta}$. We assume that the $\boldsymbol{g}_i(\boldsymbol{\psi})$ are differentiable with respect to $\boldsymbol{\nu}$. In Section 3.6, we propose an algorithm to compute (3.15). It can be shown that

$$\ell(\widehat{\boldsymbol{\psi}}) = \sum_{i \in s} \log\left(\widehat{m}_i\right) \tag{3.16}$$

is the maximum value of $\ell(\boldsymbol{m})$ under the constraints $m_i > 0$ and (3.9), because (3.14) holds for $\widehat{\boldsymbol{\psi}}$. The maximum empirical likelihood estimator of $\boldsymbol{\theta}_N$ minimises the function (3.15).

In Section 3.7, we shall show that under a series of regularity conditions and for specific choices of $\boldsymbol{c}_i$, the random variable $\widehat{r}(\boldsymbol{\theta}_N)$ asymptotically follows a χ^2-distribution with p degrees of freedom under unequal probability sampling, where p denotes the dimension of $\boldsymbol{\theta}_N$; that is,

$$\widehat{r}(\boldsymbol{\theta}_N) \xrightarrow{d} \chi^2_{df=p}. \tag{3.17}$$

3.5.1 Hypothesis testing and confidence intervals

The $\widehat{r}(\boldsymbol{\theta}_N)$ is a pivotal statistic that can be used to make inference about the sub-parameter $\boldsymbol{\theta}_N$. Suppose we wish to test $H_0 : \boldsymbol{\theta}_N = \boldsymbol{\theta}_N^0$ versus $H_1 : \boldsymbol{\theta}_N \neq \boldsymbol{\theta}_N^0$, by using $\widehat{r}(\boldsymbol{\theta}_N^0)$. The *p-value* is $\int_{\widehat{r}(\boldsymbol{\theta}_N^0)}^{\infty} \chi^2_{df=p}(x)\,\mathrm{d}x$, where $\chi^2_{df=p}(x)$ is the density of a χ^2-distribution with p degrees of freedom. This *p-value* is obtained from a statistical table of the χ^2-distribution.

The pivotal statistic (3.15) can also be used to construct confidence intervals for a scalar ($p = 1$) sub-parameter θ_N of $\boldsymbol{\psi}_N$. In this case, $\boldsymbol{\nu}_N$ denotes the remaining parameters of $\boldsymbol{\psi}_N$. Hence, $\widehat{r}(\theta_N)$ follows asymptotically a χ^2-distribution with one degree of freedom. Thus the α% empirical likelihood Wilks's (1938) type confidence interval for θ_N is given by

$$\left\{\theta \,:\, \widehat{r}(\theta) \leqslant \chi^2_{df=1}(\alpha)\right\},$$

where $\chi^2_{df=1}(\alpha)$ is the upper α-quantile of the χ^2-distribution with one degree of freedom. The $\widehat{r}(\theta)$ is a convex function of θ with a minimum value when θ is equal to the empirical maximum likelihood estimator $\widehat{\theta}$. 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.

3.6 An algorithm to compute the profile empirical log-likelihood ratio function

As the $\widehat{m}_i^*(\boldsymbol{\psi})$ maximise $\ell(m)$ under the constraint (3.11), for a given $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}$, we have that $\widehat{m}_i^*(\boldsymbol{\psi}) = \{\pi_i + \overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi})^{\mathrm{T}} \boldsymbol{c}_i^*(\boldsymbol{\psi})\}^{-1}$ (see expression (3.10)), where $\overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi})$ is such that constraint (3.11) holds or equivalently $\overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi})$ is the solution of

$$\boldsymbol{\Gamma}_1(\boldsymbol{\eta}, \boldsymbol{\nu}) = \sum_{i \in s} \left\{\pi_i + \overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi})^{\mathrm{T}} \boldsymbol{c}_i^*(\boldsymbol{\psi})\right\}^{-1} \boldsymbol{c}_i^*(\boldsymbol{\psi}) - \boldsymbol{C}^* = \mathbf{0}_{r+b}. \tag{3.18}$$

Furthermore, by using (3.13) and $\ell(\boldsymbol{\psi}) = \ell(\boldsymbol{\theta}, \boldsymbol{\nu})$, we have

$$\ell(\boldsymbol{\theta}, \boldsymbol{\nu}) = -\sum_{i \in s} \log\left(\pi_i + \overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi})^{\mathrm{T}} \boldsymbol{c}_i^*(\boldsymbol{\psi})\right). \tag{3.19}$$

In order to compute (3.15), we need to maximise (3.19) over the nuisance parameter $\boldsymbol{\nu}$. Let $\overset{\circ}{\boldsymbol{\nu}}(\boldsymbol{\theta})$ be the vector $\boldsymbol{\nu}$ that maximises (3.19) for a given value of $\boldsymbol{\theta}$. Assuming that $\boldsymbol{c}_i^*(\boldsymbol{\psi})$ is differentiable with respect to $\boldsymbol{\nu}$, the vector $\overset{\circ}{\boldsymbol{\nu}}(\boldsymbol{\theta})$ is the solution of the equation.

$$\frac{\partial \ell(\boldsymbol{\theta}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}} = \frac{\partial \overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi})^{\mathrm{T}}}{\partial \boldsymbol{\nu}} \sum_{i \in s} \widehat{m}_i^*(\boldsymbol{\psi})\ \boldsymbol{c}_i^*(\boldsymbol{\psi}) + \boldsymbol{\Gamma}_2(\overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi}), \boldsymbol{\nu}) = \mathbf{0}_q, \tag{3.20}$$

with

$$\boldsymbol{\Gamma}_2(\overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi}), \boldsymbol{\nu}) \quad = \quad \overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi})^{\mathrm{T}} \sum_{i \in s} \widehat{m}_i^*(\boldsymbol{\psi}) \frac{\partial \boldsymbol{c}_i^*(\boldsymbol{\theta}, \boldsymbol{\nu})}{\partial \boldsymbol{\nu}}.$$

Here, $\boldsymbol{c}_i^*(\boldsymbol{\theta}, \boldsymbol{\nu}) = \boldsymbol{c}_i^*(\boldsymbol{\psi})$, with $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}$. Equation (3.20) reduces to

$$\boldsymbol{\Gamma}_2(\overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi}), \boldsymbol{\nu}) = \mathbf{0}_q, \tag{3.21}$$

because $\sum_{i \in s} \widehat{m}_i^*(\boldsymbol{\psi})\ \boldsymbol{c}_i^*(\boldsymbol{\psi}) = \boldsymbol{C}^*$, as the $\widehat{m}_i^*(\boldsymbol{\psi})$ satisfy the constraint (3.18) and $\overset{*}{\boldsymbol{\eta}}(\boldsymbol{\psi})^{\mathrm{T}} \boldsymbol{C}^* = 0$ (see Lemma B.1 in Appendix B).

Let $\overset{\circ}{\boldsymbol{\nu}} = \overset{\circ}{\boldsymbol{\nu}}(\boldsymbol{\theta})$ and $\overset{\circ}{\boldsymbol{\eta}} = \overset{*}{\boldsymbol{\eta}}(\overset{\circ}{\boldsymbol{\psi}})$ with $\overset{\circ}{\boldsymbol{\psi}} = (\boldsymbol{\theta}^{\mathrm{T}}, \overset{\circ}{\boldsymbol{\nu}}(\boldsymbol{\theta})^{\mathrm{T}})^{\mathrm{T}}$. By definition, the vectors $\overset{\circ}{\boldsymbol{\eta}}$ and $\overset{\circ}{\boldsymbol{\nu}}$ satisfy the equations (3.18) and (3.21). In other words, $\overset{\circ}{\boldsymbol{\eta}}$ and $\overset{\circ}{\boldsymbol{\nu}}$ are the solutions of

$$\boldsymbol{\Gamma}(\boldsymbol{\eta}, \boldsymbol{\nu}) = \mathbf{0}_{r+b+q}. \tag{3.22}$$

where

$$\boldsymbol{\Gamma}(\boldsymbol{\eta}, \boldsymbol{\nu}) = [\boldsymbol{\Gamma}_1(\boldsymbol{\eta}, \boldsymbol{\nu})^{\mathrm{T}}, \boldsymbol{\Gamma}_2(\boldsymbol{\eta}, \boldsymbol{\nu})^{\mathrm{T}}]^{\mathrm{T}}. \tag{3.23}$$

A root-search algorithm, such as the Newton-Raphson algorithm or the Levenberg (1944) and Marquardt (1963) algorithm, can be used to solve equation (3.22). These algorithms are based on the Taylor approximation of $\boldsymbol{\Gamma}(\boldsymbol{\eta}, \boldsymbol{\nu})$ in the neighbourhood of $(\boldsymbol{\eta}_t^{\mathrm{T}}, \boldsymbol{\nu}_t^{\top})^{\mathrm{T}}$:

$$\boldsymbol{\Gamma}(\boldsymbol{\eta}, \boldsymbol{\nu}) - \boldsymbol{\Gamma}(\boldsymbol{\eta}_t, \boldsymbol{\nu}_t) \doteq \widehat{\boldsymbol{\nabla}}(\boldsymbol{\eta}_t, \boldsymbol{\nu}_t) \begin{pmatrix} \boldsymbol{\eta} - \boldsymbol{\eta}_t \\ \boldsymbol{\nu} - \boldsymbol{\nu}_t \end{pmatrix}, \tag{3.24}$$

where

$$\widehat{\boldsymbol{\nabla}}(\boldsymbol{\eta}, \boldsymbol{\nu}) = \frac{\partial \boldsymbol{\Gamma}(\boldsymbol{\eta}, \boldsymbol{\nu})}{\partial (\boldsymbol{\eta}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}}. \tag{3.25}$$

The iterative Newton-Raphson algorithm consists in combining (3.22) and (3.24) to obtain the following recursive formula.

$$\widehat{\boldsymbol{\nabla}}(\boldsymbol{\eta}_t, \boldsymbol{\nu}_t) \begin{pmatrix} \boldsymbol{\eta}_{t+1} - \boldsymbol{\eta}_t \\ \boldsymbol{\nu}_{t+1} - \boldsymbol{\nu}_t \end{pmatrix} = -\boldsymbol{\Gamma}(\boldsymbol{\eta}_t, \boldsymbol{\nu}_t) \cdot \tag{3.26}$$

For the first iteration ($t = 0$), $\boldsymbol{\eta}_0 = \mathbf{0}$ and $\boldsymbol{\nu}_0 = \widehat{\boldsymbol{\nu}}$, where $\widehat{\boldsymbol{\nu}}$ is the maximum empirical likelihood estimate of $\boldsymbol{\nu}_N$. The solution $(\boldsymbol{\eta}_{t+1}, \boldsymbol{\nu}_{t+1})$ of the system of equations (3.26) gives a new set of vectors used for the next iteration. We repeat this process until convergence. The values of $\mathring{\boldsymbol{\eta}}$ and $\mathring{\boldsymbol{\nu}}$ are the values obtained at convergence.

Finally, by using the expression (3.19), we have

$$\max_{\boldsymbol{\nu} \in \boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}, \boldsymbol{\nu}) = \ell(\boldsymbol{\theta}, \mathring{\boldsymbol{\nu}}) = -\sum_{i \in s} \log \left(\pi_i + \mathring{\boldsymbol{\eta}}^{\mathrm{T}} \boldsymbol{c}_i^*(\boldsymbol{\theta}, \mathring{\boldsymbol{\nu}}) \right) \cdot \tag{3.27}$$

We obtain the value of $\widehat{r}(\boldsymbol{\theta})$ by substituting the expression (3.27) into expression (3.15).

3.7 Asymptotic distribution of the profile empirical log-likelihood ratio function

In this Section, we show that, under a set of regularity conditions, the property (3.17) holds. We assume that $n \to \infty$ and $N \to \infty$. Let $o_{\mathcal{P}}(\cdot)$ and $O_{\mathcal{P}}(\cdot)$ be the order of convergence in probability with respect to the sampling design $\mathcal{P}(s)$ (e.g. Isaki and Fuller, 1982). We assume that the sampling design is such that the following regularity conditions hold for $\boldsymbol{\psi}_N = (\boldsymbol{\theta}_N{}^{\mathrm{T}}, \boldsymbol{\nu}_N{}^{\mathrm{T}})^{\mathrm{T}}$.

$$\begin{aligned}
\max_{i \in s} \{ \bar{\pi} \pi_i^{-1} \} &= O_{\mathcal{P}}(1), && (3.28) \\
N^{-1} \| \widehat{\boldsymbol{C}}_{\pi}^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^* \| &= O_{\mathcal{P}}(n^{-1/2}), && (3.29) \\
\max_{i \in s} \| \boldsymbol{c}_i^*(\boldsymbol{\psi}_N) \| &= o_{\mathcal{P}}(n^{1/2}), && (3.30) \\
\| \widehat{\boldsymbol{S}}^*(\boldsymbol{\psi}_N) \| &= O_{\mathcal{P}}(1), && (3.31) \\
\| \widehat{\boldsymbol{S}}^*(\boldsymbol{\psi}_N)^{-1} \| &= O_{\mathcal{P}}(1), && (3.32) \\
\frac{\bar{\pi}^{\tau}}{n} \sum_{i \in s} \frac{1}{\pi_i^{\tau}} \| \boldsymbol{c}_i^*(\boldsymbol{\psi}_N) \|^{\tau} &= O_{\mathcal{P}}(1), \quad \text{with } \tau = 2, 3 \text{ and } 4, && (3.33)
\end{aligned}$$

where $\bar{\pi} = nN^{-1}$,

$$\begin{aligned}
\widehat{\boldsymbol{C}}_{\pi}^*(\boldsymbol{\psi}_N) &= \sum_{i \in s} \frac{1}{\pi_i} \boldsymbol{c}_i^*(\boldsymbol{\psi}_N), \\
\widehat{\boldsymbol{S}}^*(\boldsymbol{\psi}_N) &= -\frac{\bar{\pi}}{N} \sum_{i \in s} \frac{1}{\pi_i^2} \boldsymbol{c}_i^*(\boldsymbol{\psi}_N) \boldsymbol{c}_i^*(\boldsymbol{\psi}_N)^{\mathrm{T}} \cdot && (3.34)
\end{aligned}$$

Here, $\|\cdot\|$ denotes the Frobenius (Euclidean) norm. The quantities $\boldsymbol{c}_i^*(\boldsymbol{\psi})$ and $\boldsymbol{C}^*$ are defined in expressions (3.12). We assume that $\boldsymbol{c}_i^*(\boldsymbol{\psi})$ is differentiable with respect to $\boldsymbol{\nu}$ for all $i \in s$ in a neighbourhood around the true population value $\boldsymbol{\nu}_N$.

The condition (3.28) can be found in Krewski and Rao (1981, p.1014). It guarantees that the π_i and $\bar{\pi}$ are of the same order of magnitude. The condition (3.29) assumes that $\widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)$ is $\sqrt{n}$ design consistent. This can be justified by using Isaki and Fuller's (1982, p.91) sufficient conditions. Chen and Sitter (1999, Appendix 2) showed that condition (3.30) holds for most unequal probability sampling designs. It can be shown that conditions (3.31) and (3.32) hold when $-\widehat{\boldsymbol{S}}^*(\boldsymbol{\psi}_N)$ is positive definite and when there exists a positive definite matrix $-\boldsymbol{S}$ such that $\|\widehat{\boldsymbol{S}}^*(\boldsymbol{\psi}_N) - \boldsymbol{S}\| = o_{\mathcal{P}}(1)$ and $\|\boldsymbol{S}\| = O(1)$. We shall see that we need to include the constant $\bar{\pi}^{-1}$ within the definition of $\boldsymbol{c}_i$ to ensures that the conditions (3.31) and (3.32) hold. However, the constant $\bar{\pi}^{-1}$ can be omitted for the computation of the function (3.15), because this constant cancels out in the constraint (3.9). The last condition (3.33) is a Liapounov type condition for the existence of moments (e.g. Krewski and Rao, 1981, p.1014).

When we have a single stratum without population level information, we propose using $\boldsymbol{c}_i = \bar{\pi}^{-1}\pi_i$. This implies $\boldsymbol{C} = N$. By using Corollary B.1 in Appendix B, we have that, under the regularity conditions (3.28)-(3.33),

$$\widehat{r}(\boldsymbol{\theta}_N) = \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)^{\mathrm{T}} \left(\mathbf{I}_b - \widehat{\boldsymbol{A}}_{\boldsymbol{g}}\right) \widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{-1} \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N) + O_{\mathcal{P}}(n^{-1/2}), \tag{3.35}$$

where

$$\widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}) = \sum_{i \in s} \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}) \tag{3.36}$$

and $\breve{\boldsymbol{g}}_i(\boldsymbol{\psi}) = \boldsymbol{g}_i(\boldsymbol{\psi})\pi_i^{-1}$. Here, $\mathbf{I}_b$ is a $b \times b$ identity matrix, with $b = p + q$, and $\widehat{\boldsymbol{A}}_{\boldsymbol{g}}$ is a *symmetric idempotent* matrix, defined by

$$\widehat{\boldsymbol{A}}_{\boldsymbol{g}} = \widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{-1/2} \widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}} \left(\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}{}^{\mathrm{T}} \widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{-1} \widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}\right)^{-1} \widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}{}^{\mathrm{T}} \widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{-1/2}, \tag{3.37}$$

with

$$\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}} = \left.\frac{\partial \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi})}{\partial \boldsymbol{\nu}}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_N} = \sum_{i \in s} \left.\frac{\partial \breve{\boldsymbol{g}}_i(\boldsymbol{\psi})}{\partial \boldsymbol{\nu}}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_N}, \tag{3.38}$$

$$\widehat{\boldsymbol{V}}_{\boldsymbol{gg}} = \sum_{i \in s} \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)\breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)^{\mathrm{T}} - \frac{1}{n} \sum_{j \in s} \breve{\boldsymbol{g}}_j(\boldsymbol{\psi}_N) \sum_{k \in s} \breve{\boldsymbol{g}}_k(\boldsymbol{\psi}_N)^{\mathrm{T}}. \tag{3.39}$$

The estimator $\widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)$ is the design unbiased Hansen and Hurwitz (1943) estimator of $\boldsymbol{G}(\boldsymbol{\psi}_N)$. The matrix $\widehat{\boldsymbol{V}}_{\boldsymbol{gg}}$ is the Hansen and Hurwitz (1943) variance estimator of the variance of $\widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)$, which is design consistent under unequal probability sampling with replacement (e.g. Durbin, 1953). The $\breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)$ are independent and standard large sample

theory can be used to show the normality of $\widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}_N)$ (e.g. Prášková and Sen, 2009), i.e.

$$\widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{-1/2}\ \widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}_N) \xrightarrow{d} \boldsymbol{\mathcal{N}}(\mathbf{0}_b, \mathbf{I}_b) \tag{3.40}$$

holds, where $\boldsymbol{\mathcal{N}}(\mathbf{0}_b, \mathbf{I}_b)$ denotes the standardised multivariate normal distribution. The condition (3.40) is weaker than the conditions under which $\widehat{\boldsymbol{\psi}}$ is known to be asymptotically normally distributed.

The condition (3.40) and expression (3.35) imply that $\widehat{r}(\boldsymbol{\theta}_N)$ follows asymptotically a χ^2-distribution with p degrees of freedom, because $(\mathbf{I}_b - \widehat{\boldsymbol{A}}_{\boldsymbol{g}})$ is an idempotent matrix with a trace equal to p (e.g. Qin and Lawless, 1994). Hence, the property (3.17) holds.

3.8 Incorporating stratification

Suppose that a population U is stratified into H, strata denoted by $U_1, \ldots, U_h, \ldots, U_H$, where $\cup_{h=1}^{H} U_h = U$. Suppose that a with-replacement sample s_h of fixed size n_h is selected from U_h, with unequal probabilities. We assume that the number of strata H is bounded ($H = O(1)$). We propose using the functions (3.13) and (3.15), with $\boldsymbol{c}_i$ replaced by $\boldsymbol{c}_i = \bar{\pi}^{-1}\boldsymbol{z}_i$, where $\boldsymbol{z}_i$ are the values of the design (or stratification) variables defined by

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iH})^{\mathrm{T}}, \tag{3.41}$$

where $z_{ih} = \pi_i$ for $i \in U_h$ and $z_{ih} = 0$ otherwise. We have $\boldsymbol{C} = \bar{\pi}^{-1}\boldsymbol{n}$, where $\boldsymbol{n} = (n_1, \ldots, n_H)^{\mathrm{T}}$ denotes the vector of the stratum sample sizes. By using Theorem B.1 of the supplementary materials (see Appendix B), we obtain the consistent stratified Hansen and Hurwitz (1943) variance estimator (B.46) in the right hand side of (3.35). Hence, the property (3.17) holds under the condition (3.40).

3.9 Incorporating known population level information

Since Hartley and Rao (1968) first introduced population level information within the empirical likelihood framework, it became a key feature of empirical likelihood (e.g. Owen, 2001; Chaudhuri et al., 2008; Rao and Wu, 2009). Berger and De La Riva Torres (2016) proposed an empirical likelihood approach for a single parameter in the presence of population level information, under unequal probability sampling. In this Section, we extend this approach for the multiparameter case in which we have a nuisance parameter.

Let $\boldsymbol{\varphi}_N$ be a vector of known population level parameters, which is is not subject to any uncertainty. We consider that $\boldsymbol{\varphi}_N$ can be defined as the unique solution of the population

estimating equation

$$\sum_{i \in U} \mathbf{f}_i(\boldsymbol{v}_i, \boldsymbol{\varphi}) = \mathbf{0}. \tag{3.42}$$

We assume that $\mathbf{f}_i(\boldsymbol{v}_i, \boldsymbol{\varphi})$ is a function that does not depend on $\boldsymbol{\psi}_N$. For simplicity, we replace $\mathbf{f}_i(\boldsymbol{v}_i, \boldsymbol{\varphi})$ by $\mathbf{f}_i(\boldsymbol{\varphi})$ in what follows. For example, $\boldsymbol{\varphi}_N$ may be known population means, totals, ratios, proportions, variances, quantiles or distribution functions of some of the variables within $\boldsymbol{v}_i$. If the $\boldsymbol{\varphi}_N$ are the population means of an auxiliary variable $\boldsymbol{x}_i$ (a sub-vector of $\boldsymbol{v}_i$), we use $\mathbf{f}_i(\boldsymbol{\varphi}) = \boldsymbol{x}_i - \boldsymbol{\varphi}$. If $\boldsymbol{\varphi}_N$ is a vector of population totals, we use $\mathbf{f}_i(\boldsymbol{\varphi}) = \boldsymbol{x}_i - \boldsymbol{\varphi}\pi_i n^{-1}$. The $\mathbf{f}_i(\boldsymbol{\varphi})$ do not have to be differentiable. For example, $\mathbf{f}_i(\boldsymbol{\varphi})$ is not differentiable when $\boldsymbol{\varphi}_N$ contains population quantiles. The $\boldsymbol{g}_i(\boldsymbol{\psi})$ can be a function that depends on $\boldsymbol{\varphi}_N$. We assume that the $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ are such that the conditions (3.29)–(3.33) hold.

3.9.1 Maximum empirical likelihood point estimator under population level information

Let $\boldsymbol{c}_i = (\bar{\pi}^{-1}\pi_i, \mathbf{f}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}}$, with $\boldsymbol{C} = (N, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$. Let $\widehat{m}_i^*(\boldsymbol{\psi}, \boldsymbol{\varphi}_N)$ maximises $\ell(\boldsymbol{m})$ under the constraints $m_i > 0$ and (3.11), with $\boldsymbol{c}_i^*(\boldsymbol{\psi}) = (\boldsymbol{c}_i{}^{\mathrm{T}}, \boldsymbol{g}_i(\boldsymbol{\psi})^{\mathrm{T}})^{\mathrm{T}}$. The $\boldsymbol{g}_i(\boldsymbol{\psi})$ may also be a function of known population level parameters. The maximum value of $\ell(\boldsymbol{m})$ is

$$\ell(\boldsymbol{\psi} \mid \boldsymbol{\varphi}_N) = \sum_{i \in s} \log\left(\widehat{m}_i^*(\boldsymbol{\psi}, \boldsymbol{\varphi}_N)\right). \tag{3.43}$$

The *maximum empirical likelihood estimate* $\widehat{\boldsymbol{\psi}}$ is defined as the vector $\boldsymbol{\psi}$ which maximises expression (3.43). It can be shown that $\widehat{\boldsymbol{\psi}}$ is the solution of the estimating equation (see Section 3.4).

$$\widehat{\boldsymbol{G}}(\boldsymbol{\psi}, \boldsymbol{\varphi}_N) = \sum_{i \in s} \widehat{m}_i(\boldsymbol{\varphi}_N)\, \boldsymbol{g}_i(\boldsymbol{\psi}) = \mathbf{0}_b, \tag{3.44}$$

where $\widehat{m}_i(\boldsymbol{\varphi}_N)$ maximises $\ell(\boldsymbol{m})$ subject to $m_i > 0$ and (3.9). The $\widehat{m}_i(\boldsymbol{\varphi}_N)$ are given by

$$\widehat{m}_i(\boldsymbol{\varphi}_N) = (\pi_i + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{c}_i)^{-1}, \tag{3.45}$$

with $\boldsymbol{c}_i = (\bar{\pi}^{-1}\pi_i, \mathbf{f}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}}$ (see expression (3.10)).

Here, the $\widehat{m}_i(\boldsymbol{\varphi}_N)$ play the role of survey weights. The $\widehat{m}_i(\boldsymbol{\varphi}_N)$ are positive weights. They are also calibrated weights, because $\sum_{i \in s} \widehat{m}_i(\boldsymbol{\varphi}_N)\, \mathbf{f}_i(\boldsymbol{\varphi}_N) = \mathbf{0}$. The calibration property is the consequence of the maximisation of $\ell(\boldsymbol{m})$ and the fact that $\boldsymbol{\varphi}_N$ is known.

In Appendix B (see Section B.1), we show that under the regularity conditions (3.28)–(3.33) and for $\boldsymbol{\psi}$ such that

$$\frac{\bar{\pi}^2}{n}\sum_{i\in s}\frac{1}{\pi_i^2}\|\boldsymbol{g}_i(\boldsymbol{\psi})\|^2 = O_{\mathcal{P}}(1), \tag{3.46}$$

we have

$$\widehat{\boldsymbol{G}}(\boldsymbol{\psi},\boldsymbol{\varphi}_N) = \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi},\boldsymbol{\varphi}_N) + o_{\mathcal{P}}(Nn^{-1/2}), \tag{3.47}$$

where $\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi},\boldsymbol{\varphi}_N)$ is the following *regression estimator*.

$$\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi},\boldsymbol{\varphi}_N) = \widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}) - \widehat{\boldsymbol{B}}(\boldsymbol{\psi},\boldsymbol{\varphi}_N)^{\mathrm{T}}\,\widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N), \tag{3.48}$$

where $\widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N) = \sum_{i\in s}\mathbf{f}_i(\boldsymbol{\varphi}_N)\pi_i^{-1}$ and $\widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}) = \sum_{i\in s}\boldsymbol{g}_i(\boldsymbol{\psi})\pi_i^{-1}$ and $\widehat{\boldsymbol{B}}(\boldsymbol{\psi},\boldsymbol{\varphi}_N)$ is the regression coefficient

$$\widehat{\boldsymbol{B}}(\boldsymbol{\psi},\boldsymbol{\varphi}_N) = \widehat{\boldsymbol{V}}_{\mathbf{ff}}^{-1}\widehat{\boldsymbol{V}}_{\mathbf{f}\boldsymbol{g}}, \tag{3.49}$$

where

$$\begin{aligned}
\widehat{\boldsymbol{V}}_{\mathbf{ff}} &= \sum_{i\in s}\breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)\breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}} - \frac{1}{n}\sum_{j\in s}\breve{\mathbf{f}}_j(\boldsymbol{\varphi}_N)\sum_{k\in s}\breve{\mathbf{f}}_k(\boldsymbol{\varphi}_N)^{\mathrm{T}}, \\
\widehat{\boldsymbol{V}}_{\mathbf{f}\boldsymbol{g}} &= \sum_{i\in s}\breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)\breve{\boldsymbol{g}}_i(\boldsymbol{\psi})^{\mathrm{T}} - \frac{1}{n}\sum_{j\in s}\breve{\mathbf{f}}_j(\boldsymbol{\varphi}_N)\sum_{k\in s}\breve{\boldsymbol{g}}_k(\boldsymbol{\psi})^{\mathrm{T}}.
\end{aligned} \tag{3.50}$$

Here, $\breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N) = \mathbf{f}_i(\boldsymbol{\varphi}_N)\pi_i^{-1}$. The matrix $\widehat{\boldsymbol{V}}_{\mathbf{ff}}$ is the Hansen and Hurwitz (1943) variance estimator of $\widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N)$. The matrix $\widehat{\boldsymbol{V}}_{\mathbf{f}\boldsymbol{g}}$ is the covariance estimator between $\widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi})$ and $\widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N)$.

The expression (3.47) implies that the maximum empirical likelihood estimator of $\boldsymbol{\psi}_N$ is design consistent. The estimator (3.48) is the asymptotic *design optimal* regression estimator under with replacement sampling (Montanari, 1987; Rao, 1994; Berger et al., 2003).

3.9.2 Hypothesis testing and confidence intervals under population level information

The profile empirical likelihood log-likelihood ratio function in the presence of population level information is defined by

$$\widehat{r}(\boldsymbol{\theta} \mid \boldsymbol{\varphi}_N) = 2\left\{\ell(\widehat{\boldsymbol{\psi}} \mid \boldsymbol{\varphi}_N) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}}\ell(\boldsymbol{\theta},\boldsymbol{\nu} \mid \boldsymbol{\varphi}_N)\right\}. \tag{3.51}$$

where $\ell(\boldsymbol{\theta}, \boldsymbol{\nu} \mid \boldsymbol{\varphi}_N) = \ell(\boldsymbol{\psi} \mid \boldsymbol{\varphi}_N)$ with $\boldsymbol{\psi} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}$. It can be shown that

$$\ell(\widehat{\boldsymbol{\psi}} \mid \boldsymbol{\varphi}_N) = \sum_{i \in s} \log\left(\widehat{m}_i(\boldsymbol{\varphi}_N)\right),$$

where the $\widehat{m}_i(\boldsymbol{\varphi}_N)$ are given by (3.10), with $\boldsymbol{c}_i = (\bar{\pi}^{-1}\pi_i, \mathbf{f}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}}$.

In Appendix B, Corollary B.2 shows that under the regularity conditions (3.28)–(3.33), we have that

$$\widehat{r}(\boldsymbol{\theta}_N \mid \boldsymbol{\varphi}_N) = \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)^{\mathrm{T}} \left(\mathbf{I}_b - \widehat{\boldsymbol{A}}_{\boldsymbol{g}}^{\bullet}\right) \widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{\bullet -1} \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N) + O_{\mathcal{P}}(n^{-\frac{1}{2}}), \tag{3.52}$$

where $\widehat{\boldsymbol{A}}_{\boldsymbol{g}}^{\bullet}$ is a symmetric and idempotent matrix that is defined by

$$\widehat{\boldsymbol{A}}_{\boldsymbol{g}}^{\bullet} = \widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{\bullet -1/2} \widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}^{\bullet} \left(\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}^{\bullet \mathrm{T}} \widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{\bullet -1} \widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}^{\bullet}\right)^{-1} \widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}^{\bullet \mathrm{T}} \widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{\bullet -1/2}, \tag{3.53}$$

with

$$\begin{aligned}
\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}^{\bullet} &= \left.\frac{\partial \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}, \boldsymbol{\varphi}_N)}{\partial \boldsymbol{\nu}}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_N} = \sum_{i \in s} \left.\frac{\partial \breve{\boldsymbol{g}}_i^{\bullet}(\boldsymbol{\psi}, \boldsymbol{\varphi}_N)}{\partial \boldsymbol{\nu}}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_N}, \\
\widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{\bullet} &= \sum_{i \in s} \breve{\boldsymbol{g}}_i^{\bullet}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)\, \breve{\boldsymbol{g}}_i^{\bullet}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)^{\mathrm{T}} - \frac{1}{n} \sum_{j \in s} \breve{\boldsymbol{g}}_j^{\bullet}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N) \sum_{k \in s} \breve{\boldsymbol{g}}_k^{\bullet}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)^{\mathrm{T}}, \\
\breve{\boldsymbol{g}}_i^{\bullet}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N) &= \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N) - \widehat{\boldsymbol{B}}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)^{\mathrm{T}} \breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)
\end{aligned} \tag{3.54}$$

and $\widehat{\boldsymbol{B}}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)$ is given by expression (3.49).

The $\widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{\bullet}$ is the Hansen and Hurwitz (1943) variance estimator of the regression estimator $\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)$. The asymptotic normality of the regression estimator is shown by Scott and Wu (1981). Hence, we assume that

$$\widehat{\boldsymbol{V}}_{\boldsymbol{gg}}^{\bullet -1/2} \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N) \xrightarrow{d} \boldsymbol{\mathcal{N}}(\mathbf{0}_b, \mathbf{I}_b) \tag{3.55}$$

holds. By using expression (3.52) and condition (3.55), the random variable $\widehat{r}(\boldsymbol{\theta}_N \mid \boldsymbol{\varphi}_N)$ given by expression (3.51) follows asymptotically a χ^2-distribution with p degrees of freedom, because $(\mathbf{I}_b - \widehat{\boldsymbol{A}}_{\boldsymbol{g}}^{\bullet})$ is an idempotent matrix with a trace equal to p, where p is the dimension of $\boldsymbol{\theta}$. Hence, the property (3.17) holds. Thus $\widehat{r}(\boldsymbol{\theta} \mid \boldsymbol{\varphi}_N)$ can be used to test hypotheses and to construct confidence intervals (see Section 3.5.1).

3.9.3 Stratified and clustered population

We propose using the approaches of Sections 3.9.1 and 3.9.2, after replacing $\boldsymbol{c}_i$ by

$$\boldsymbol{c}_i = (\bar{\pi}^{-1} \boldsymbol{z}_i^{\mathrm{T}}, \mathbf{f}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}}, \tag{3.56}$$

where $\boldsymbol{z}_i$ are defined by expression (3.41). By using Theorem B.2 in Appendix B, we obtain the consistent variance estimator (B.50) in the right hand side of equation (3.52). Hence, the property (3.17) holds under the condition (3.55).

The population may be subdivided into a large number M of small disjoint subsets called clusters, denoted $\widetilde{U}_i$, where $i = 1, \ldots, M$. Suppose that a stratified with-replacement sample s of n clusters is sampled with unequal probabilities and a sample of units is sampled within each selected cluster. For example, the clusters may be selected with probabilities proportional to their size. Let $\boldsymbol{g}_{ij}(\boldsymbol{\psi}) = \boldsymbol{g}_{ij}(\boldsymbol{v}_{ij}, \boldsymbol{\psi})$ be the estimating function for a unit $j \subset \widetilde{U}_i$, where $\boldsymbol{v}_{ij}$ is the corresponding vector of variables. Let $\boldsymbol{\psi}_N$ be the unique solution of the following population estimating equation:

$$\sum_{i=1}^{M} \boldsymbol{g}_{i\bullet}(\boldsymbol{\psi}) = \mathbf{0}_b, \tag{3.57}$$

where $\boldsymbol{g}_{i\bullet}(\boldsymbol{\psi}) = \sum_{j \in \widetilde{U}_i} \boldsymbol{g}_{ij}(\boldsymbol{\psi})$. We propose using an *ultimate cluster approach* (e.g. Hansen et al., 1953) described as follows. Let $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\psi})$ be the Hansen and Hurwitz (1943) estimator of $\boldsymbol{g}_{i\bullet}(\boldsymbol{\psi})$ for a given $\boldsymbol{\psi}$. The approach proposed in Section 3.9.1 can be used by using expression (3.56) and by treating the clusters as sampling units. That is, we substitute $\boldsymbol{g}_i(\boldsymbol{\psi})$ by $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\psi})$ within (3.12). Now, $\pi_i = np_i$, where p_i is the selection probability of the i-th cluster. With population level information, the $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ in (3.56), are defined at cluster level. When $\boldsymbol{\varphi}_N$ is a function of unit level variables as in function (3.57), $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ are replaced by unbiased estimates (e.g Estevao and Särndal, 2006). We assume that the regularity conditions (3.29)–(3.33) hold with $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\psi})$. The result (3.52) shows that $\widehat{r}(\boldsymbol{\theta}_N \mid \boldsymbol{\varphi}_N)$ is approximated by a quadratic form with an ultimate cluster covariance matrix which is consistent as long as the sampling fraction is negligible $n/M = o(1)$ (e.g. Särndal et al., 1992, Ch.4). Hence, $\widehat{r}(\boldsymbol{\theta}_N \mid \boldsymbol{\varphi}_N)$ has a χ^2-distribution asymptotically and the property (3.17) holds.

3.10 Simulation study

In this Section, we present simulation studies for parameters of linear and logistic regression models. Population data are either generated from models or based on the first quarter of the 2011 UK Labour Force Survey (LFS) data. In all cases, we selected 1000 random samples by using the randomised systematic sampling design (Hartley and Rao, 1962).

We compare the Monte-Carlo performance of the empirical likelihood confidence intervals proposed with nonparametric confidence intervals based on pseudo likelihood (Binder and Patak, 1994), linearisation (Binder, 1983) and rescaled bootstrap (Rao et al., 1992). We also consider parametric and semiparametric confidence intervals respectively based on Wald's test statistics and the Q-weighted approach for informative

sampling (Pfeffermann and Sverchkov, 1999). We used the Hartley and Rao (1962) variance estimator for the approaches that require variance estimates.

The pseudo likelihood confidence intervals rely on a variance estimator of an estimating equation, for a given value of $\boldsymbol{\theta}$. Binder and Patak (1994) mentioned two versions, which are denoted by pseudo likelihood 1 and 2 in this paper. For the pseudo likelihood 1 confidence interval, we substitute $\boldsymbol{\theta}$ by its estimate within the variance estimator. For the pseudo likelihood 2 confidence interval, the variance estimator is kept as a function of $\boldsymbol{\theta}$. For the pseudo likelihood 1, $\boldsymbol{\nu}$ is replaced by $\widehat{\boldsymbol{\nu}}$. For the pseudo likelihood 2, the nuisance parameter is kept as a function of $\boldsymbol{\theta}$. Godambe and Thompson (2009, p.92) noticed that it may not always be possible to find the confidence interval bounds by using the pseudo likelihood 2 approach.

Linearisation is based on Binder's (1983) approach. The rescaled bootstrap (Rao et al., 1992, p.214) consists in selecting $B = 1000$ bootstrap samples of size $m = n - 1$. The quantiles of the 1000 bootstrap sample estimates are used to compute the confidence intervals.

The Q-weighted approaches are based on a Q-weighted estimator of the estimating equation (3.2) (Pfeffermann and Sverchkov, 1999, 2003). Q-weighted confidence intervals are based on linearised variances. We consider two versions, which are denoted by Q-weighted 1 and 2, in this paper. The Hartley and Rao (1962) variance estimator is used for the Q-weighted 1 approach. The conditional variance estimator (e.g. Pfeffermann and Sverchkov, 2003) is used for the Q-weighted 2 approach.

The Wald's approach is the standard model-based approach based on the assumption of normality and i.i.d observations. Least squares estimators and model-based variance estimators are used. The effect of the sampling design and the sampling weights are not taken into account with this method.

We consider a nominal level of 95% for the confidence intervals. 'Std. Length' is the standardised length given by $\text{AL}/(2 \times 1.96 \times \sqrt{\text{MSE}})$ (see Kovar et al., 1988, p.32), where MSE is the Monte-Carlo mean squared error of the point estimator and 'AL' is the average length of the confidence intervals. 'Ratio AL' is the AL divided by the AL of the empirical likelihood confidence intervals. 'SDL' is the standard deviation of the lengths of the confidence intervals. 'Ratio SDL' is the SDL divided by the SDL of the empirical likelihood confidence intervals. Shapiro and Wilk's (1965) test is used to test the normality of the point estimators. In Tables 3.1–3.6, we have the observed coverages, the observed lower and upper tail errors, the Std. Length, the Ratio AL, the Ratio SDL, and Shapiro and Wilk (1965) p-value, for each confidence interval. Significance of observed coverages and tail errors was tested by using a z-test for proportions given by $z^2 = (p - P_0)^2/(P_0(1 - P_0)/1000)$, where p is observed coverage or tail error rate and P_0 is the nominal value equal to 0.95 or 0.025. We have asymptotically $z^2 \to \chi^2_{df=1}$ in distribution as the distribution of z-statistic is approximated by a standard normal distribution under large

sample (De Moivre, 1733). The statistical software R (R Development Core Team, 2014) was used.

3.10.1 Linear regression with the Hansen, Madow and Tepping population

We generate a population of size $N = 10{,}000$ according to the model proposed by Hansen, Madow and Tepping (1983); that is, the values y_i are generated from the conditional gamma distribution

$$y_i|x_i \sim gamma(shape = 0.04x_i^{-3/2}(8+5x_i)^2, scale = 1.25x_i^{3/2}(8+5x_i)^{-1}), \quad (3.58)$$

where $x_i \sim gamma(shape = 2, scale = 5)$. The values generated are treated as fixed. We selected 1000 randomised systematic samples of size $n = 500$, from this population, with unequal probabilities. The π_i are proportional to the measure of size $z_i = 5+y_i+x_i+\epsilon_i$, where $\epsilon_i \sim exponential(rate = 1) - 1$.

Suppose we want to fit the linear regression model

$$y_i = \nu + \theta x_i + x_i^{3/4} e_i, \quad (3.59)$$

to the sample data, by using the least-square equation (3.3), where $\boldsymbol{\psi} = (\theta, \nu)^{\mathrm{T}}$ and the e_i are residuals with an unknown distribution with mean zero and a residual variance $\sigma^2 = 0.0625$ that does not depend on i. Let $\boldsymbol{\psi}_N = (\theta_N, \nu_N)^{\mathrm{T}}$ be the solution of the least-square equation (3.3) with $\sigma_i^2 \propto x_i^{3/2}$. The data population generated gives the finite population parameters $\theta_N = 0.25$ and $\nu_N = 0.4$.

Suppose that the parameter of interest is the slope θ_N. The intercept term ν_N is treated as the nuisance parameter. In Table 3.1, we observe that the coverage of θ_N with the Wald and bootstrap confidence intervals are significantly different from the nominal level (95%). The empirical likelihood confidence intervals are more stable than the pseudo likelihood approaches as the Ratio SDL is smaller. Shapiro & Wilk's test suggests that all the estimators do not deviate from normal distribution. The Wald confidence interval gives a poor coverage because we have biased point and variance estimators, as this approach does not take into account the unequal probabilities. The coverage is larger than 95% with the Q-weighted approaches. The pseudo likelihood 2 approach has better coverage than the pseudo likelihood 1 approach because the pivotal statistic of the pseudo likelihood 2 approach is normally distributed. For some samples, it may not be possible to construct confidence intervals with the pseudo likelihood 2 approach (see Section 3.10.3). With the rescaled bootstrap, we observe an over-coverage and the largest confidence intervals on average (see Ratio AL). In linear models, the pseudo likelihood 1 approach reduces to the linearisation approach (Binder, 1983). This is the reason why we observe the same results for these approaches.

Table 3.1: 95% confidence intervals for the slope of linear regression (3.59). $N = 10{,}000$. $n = 500$. HMT population.

Approaches	Observed coverages %	Lower tail err. rates %	Upper tail err. rates %	Std. Length	Ratio AL	Ratio SDL	Shapiro & Wilk p-value
Empirical likelihood	94.8	3.1	2.1	0.98	1.00	1.00	0.89
Wald	76.6*	23.8*	0.1*	0.63	0.96	0.53	0.64
Q-weighted 1	95.7	3.0	1.3*	0.99	0.86	0.63	0.49
Q-weighted 2	96.2	2.7	1.1*	1.03	0.89	0.64	0.49
Pseudo likelihood 1	94.0	3.5*	2.5	0.95	0.97	1.07	0.89
Pseudo likelihood 2	94.8	3.3	1.9	0.97	0.99	1.09	0.89
Bootstrap	96.5*	2.4	1.1*	1.03	1.05	0.91	0.89
Linearisation	94.0	3.5*	2.5	0.95	0.97	1.07	0.89

$*$ Coverages (or tail error rates) significantly different from 95% (or 2.5%). p-value ≤ 0.05.

3.10.2 Testing the significance of the intercept

Hansen et al. (1983) pointed out that model-based approaches may not detect that the slope in (3.59) is different from zero, when the data are generated from (3.58). We should reject with a large probability the null hypothesis that the intercept is equal to zero, as $\nu_N = 0.4$. We can test the significance (level $\alpha = 0.05$) of the intercept by considering that the intercept is the parameter of interest and the slope is the nuisance parameter. The empirical likelihood test statistics is given by expression (3.15). In Table 3.2, we have the observed powers (or rejection rate) of the empirical likelihood test, the model-based F-test, the Wald-test, the pseudo likelihood tests and the Q-weighted tests. Our results are different from Hansen et al. (1983), because Hansen et al. (1983) used a stratified simple random sampling design, which is different from the design used here.

For sample sizes larger than 300, we observe a power above 99%, except for the pseudo likelihood 2 test, which is the least powerful test. As expected, the power decreases with the sample size. We observe the largest power for the empirical likelihood test, followed by the pseudo likelihood 1 and the Q-weighted tests. The model-based tests, F-test and Wald-test, are less powerful.

3.10.3 Linear regression with outlying values

We consider a population of size $N = 10{,}000$. The population values y_i are generated from the model $y_i = 1 + x_i + \sigma e_i$, where $x_i \sim N(8, 1)$, $e_i \sim N(0, 1)$ and $\sigma = 0.75$. 5% of the y_i are replaced by very small values generated randomly from $\min_{i \in U}\{y_i\}$ to $\{Y_{0.25}(y) - 1.5 \times (Y_{0.75} - Y_{0.25})\}$, where $Y_{0.25}$ and $Y_{0.75}$ are the lower and upper quartiles of the generated values $\{y_i : i \in U\}$. 5% of the y_i are replaced by very large values generated randomly from $\{Y_{0.75} + 1.5 \times (Y_{0.75} - Y_{0.25}\}$ to $\max_{i \in U}\{y_i\}$.

Table 3.2: Observed powers (in %) for testing the hypothesis that the intercept is equal to zero, at the significance level $\alpha = 0.05$. The population size is $10,000$ in all cases.

Sample size	EL test	Wald-test	F-test	Pseudo Lik.1 test	Pseudo Lik.2 test	Q-weighted 1 test	Q-weighted 2 test
50	60.4	45.4	43.3	58.3	16.0	54.3	54.7
100	82.1	72.3	71.9	79.5	37.4	78.5	78.5
150	94.2	89.0	88.7	91.7	54.0	92.6	92.6
200	97.6	94.7	94.7	96.5	65.3	96.8	96.8
300	99.6	99.0	99.0	99.4	75.1	99.6	99.5
400	100.0	100.0	100.0	100.0	81.2	100.0	100.0

The parameter $\boldsymbol{\psi}_N$ is the solution of the least-square equation (3.5) with $\boldsymbol{x}_i = (1, x_i)^{\mathrm{T}}$. The slope is the parameter of interest and the intercept is the nuisance parameter. We selected 1000 simple random samples of size $n = 500$. The simulation results are given in Table 3.3. We omit the pseudo likelihood 2 approach because the confidence interval cannot be obtained with some samples. This is an issue that was pointed out by Godambe and Thompson (see 2009, p.92).

The empirical likelihood approach gives the best coverages and tail error rates. The distribution of the point estimator deviates from normal distribution as the Shapiro & Wilk p-value is 0.057. This explains the lower coverages of the alternative approaches. However, the alternative confidence intervals are slightly shorter and more stable.

Table 3.3: 95% confidence intervals for the slope of the model (3.5). $N =$ 10,000. $n = 500$. Simple random sampling. In all cases, the Shapiro & Wilk p-value is 0.057. Population with outliers.

Approaches	Observed coverages %	Lower tail err. rates %	Upper tail err. rates %	Std. Length	Ratio AL	Ratio SDL
Empirical likelihood	95.0	2.5	2.5	1.05	1.00	1.00
Wald	94.1	2.8	3.1	0.96	0.91	0.33
Q-weighted 1	93.3*	4.9*	1.8	0.97	0.93	0.69
Q-weighted 2	93.7	4.6*	1.7	1.00	0.95	0.71
Pseudo likelihood 1	93.3*	4.9*	1.8	0.97	0.93	0.69
Bootstrap	94.1	4.1*	1.8	1.00	0.95	0.71
Linearisation	93.3*	4.9*	1.8	0.97	0.93	0.69

$*$ Coverages (or tail error rates) significantly different from 95% (or 2.5%). p-value ≤ 0.05.

3.10.4 Logistic regression

In this Section, we use the 2011 UK Labour Force Survey (LFS) data on individuals who are of working age (i.e $16 - 60$ years old for females and $16 - 65$ years old for

Table 3.4: 95% confidence intervals for the slope of the logistic regression (3.6). $N = 13{,}048$. $n = 600$. LFS data. Systematic sampling. In all cases, the Shapiro & Wilk p-value is larger than 0.6 and the MSE of the point estimator is 0.0318

Approaches	Observed coverages %	Lower tail err. rates %	Upper tail err. rates %	Std. Length	Ratio AL	Ratio SDL
Empirical likelihood	94.9	2.4	2.7	1.01	1.00	1.00
Wald	94.2	3.4	2.4	1.00	0.97	0.92
Q-weighted 1	94.4	2.5	3.1	0.99	0.98	0.97
Q-weighted 2	95.1	2.2	2.7	1.01	1.00	0.99
Pseudo likelihood 1	94.2	3.0	2.8	0.99	0.98	1.01
Pseudo likelihood 2	94.3	2.9	2.8	0.99	0.98	1.04
Bootstrap	94.6	2.4	3.0	1.01	1.00	1.86
Linearisation	94.4	2.5	3.1	0.99	0.98	0.97

males). The missing observations are removed from the dataset. We quadrupled the dataset to create an artificial population of size $N = 13{,}048$. The variable y_i is the binary unemployment variable: $y_i = 1$ if the individual i is unemployed for one year or more; $y_i = 0$ otherwise. The variable x_i specifies the gender ($x_i = 1$ for male and $x_i = 0$ for female). We consider the logistic regression model with the response variable y_i and one explanatory variable, x_i. Let $\boldsymbol{\psi}_N$ be the solution of the least-square equation (3.6) with $\boldsymbol{x}_i = (1, x_i)^{\mathrm{T}}$. The parameter of interest is the slope. The intercept is the nuisance parameter. 1000 randomised systematic samples, of size $n = 600$, are selected with unequal probabilities proportional to the inverse of the survey weights (provided in the LFS dataset).

The simulation results are given in Table 3.4. We observe the same mean squared error (MSE), 0.0318, for all the point estimators. The observed coverages and tail errors are not significantly different from 95% with all approaches. The rescaled bootstrap confidence intervals are less stable, because its SDL is 1.86 times the SDL of the empirical likelihood confidence intervals. The Q-weighted 1 approach reduces to the linearisation approach (Binder, 1983), because the same variance estimator is used for both approaches and the point estimators are the same as x_i is a binary variable. This is the reason why we observe the same results for these approaches.

3.10.5 Logistic regression with population level information

In this Section, we consider the logistic regression and the LFS population introduced in Section 3.10.4. Suppose we have known population level information given by the population proportion $\varphi_N = 0.37$ of individuals unemployed for more than one year; that is, φ_N is the population mean of the binary response unemployment variable y_i (see

Table 3.5: 95% confidence intervals for the slope of the logistic regression (3.6). With population level information. $N = 13{,}048$. $n = 600$. LFS data. Systematic sampling. In all cases, the MSE of the point estimator is 0.0318 and the Shapiro & Wilk p-value is larger than 0.8.

Approaches	Observed coverages %	Lower tail err. rates %	Upper tail err. rates %	Std. Length	Ratio AL	Ratio SDL
Empirical likelihood	95.0	2.3	2.7	1.009	1.00	1.00
Pseudo likelihood 1	94.2	3.0	2.8	0.992	0.98	0.94
Pseudo likelihood 2	94.3	2.9	2.8	0.993	0.98	0.96
Bootstrap	94.6	2.4	3.0	1.011	1.00	1.73
Linearisation	94.4	2.5	3.1	0.988	0.98	0.91

Section 3.10.4). Hence, the associated estimating function is $\mathbf{f}_i(\boldsymbol{v}_i, \varphi_N) = y_i - \varphi_N$ (see equation (3.42)).

The regression weights (Deville and Särndal, 1992, p.377) are used for the linearisation, the pseudo likelihood, and the rescaled bootstrap approaches. The variance estimator proposed by Deville and Särndal (1992) is used. The regression weights are adjusted for each bootstrap sample. The Wald approach and the Q-weighted approaches are not considered because they do not take into account the population level information.

In Table 3.5, we have the results when the slope is the parameter of interest. We do not observe major differences between Tables 3.4 and 3.5. All the methods provide correct coverages and tail error rates. The rescaled bootstrap confidence intervals are less stable, as we observe a large Ratio SDL. The use of a population level information has not improved the point estimation of the slope, because we observe the same MSE (0.0318) with and without population level information (see Tables 3.4 and 3.5).

In Table 3.6, we have the results when the intercept is the parameter of interest. With population level information (values in parentheses), the empirical likelihood confidence interval has better coverage than the other methods. The bootstrap confidence intervals are less stable than the other confidence intervals. The population level information gives a slightly more precise point estimator, because the MSE is 0.0123 with population level information and 0.0195 without population level information.

Table 3.6: 95% confidence intervals for the intercept of the logistic model (3.6). N = 13,048. n = 600. LFS data. Systematic sampling. In parentheses, we have the values with population level information. Without population level information, the MSE of the point estimator is 0.0195. With population level information, the MSE of the point estimator is 0.0123. In all cases, the Shapiro & Wilk p-value is larger than 0.4.

Approaches	Observed coverages %	Lower tail err. rates %	Upper tail err. rates %	Std. Length	Ratio AL	Ratio SDL
Emp. lik.	94.5(94.7)	2.3(2.9)	3.2 (2.4)	1.00(1.01)	1.00(1.00)	1.00(1.00)
Pseudo lik. 1	93.8(93.9)	2.3(3.0)	3.9*(3.1)	0.99(0.99)	0.99(0.98)	1.02(0.99)
Pseudo lik. 2	93.9(94.0)	2.2(2.9)	3.9*(3.1)	0.99(1.00)	0.99(0.99)	1.03(1.00)
Bootstrap	94.1(94.4)	2.4(2.9)	3.5*(2.7)	1.01(1.01)	1.00(1.00)	1.29(1.17)
Linearisation	94.0(94.2)	2.8(3.5*)	3.2 (2.3)	0.98(0.99)	0.98(0.98)	0.97(0.96)

$*$ Coverages (or tail error rates) significantly different from 95% (or 2.5%). p-value ≤ 0.05.

3.11 Conclusion

There are numerous situations where the parameter of interest depends on nuisance parameters. A statistical test on the parameter of interest needs to take into account the estimation of the nuisance parameters. In Section 3.7, we show that the profile empirical log-likelihood ratio function (3.15) is a pivotal statistic that follows a χ^2-distribution asymptotically, under the sampling distribution specified by the sampling design. The function (3.15) can be used to test the parameter of interest, and to construct confidence intervals that take into account the estimation of the nuisance parameters (see Sections 3.5.1 and 3.9.2). These confidence intervals do not rely on variance estimates, linearisation (e.g Binder, 1983; Deville, 1999; Demnati and Rao, 2004) or resampling. The empirical likelihood confidence intervals do not rely directly on the normality of the point estimator. Our simulation studies show that the empirical likelihood confidence interval based on the function (3.15) achieves better coverages and tail error rates than standard approaches, which involve linearisation or resampling.

The approach proposed can be applicable to nonlinear models such as generalised linear models (see Sections 3.2.1 and 3.10.4. The approach proposed is not limited to regression parameters. It can be applied to any finite population parameters that are uniquely defined as the solution of a set of estimating equations (see equation (3.2)).

The approach proposed is less computer intensive than the bootstrap and simpler to implement than linearisation, because it does not involve the derivation of linearised variables. Standard confidence intervals based on variance estimates may give poor coverages, when normality does not hold. This can be the case with skewed data and outlying values. Even when the normality holds, heteroscedasticity or model misspecification may affect the coverage of standard confidence intervals (e.g Owen, 1991; Rao

and Wu, 2009). Furthermore, the coverage may also be affected by the bias of linearised or resampling variance estimators.

In Section 3.9, we show that population level information can be taken into account. The empirical likelihood survey weights (3.45) appear naturally because of the maximisation of the empirical log-likelihood function (3.43), and the fact that a know population parameter is fixed within the function (3.43). The survey weights (3.45) are always positive and calibrated. There are some analogies between empirical likelihood and calibration (Deville and Särndal, 1992), although they are different. First, the empirical likelihood approach does not always require population level information (see Sections 3.5, 3.7 and 3.8). Secondly, the calibration distance function is only used to derive calibration weights for point estimation, and plays no role in testing or constructing confidence intervals. The empirical log-likelihood ratio function (3.51) is used for point estimation, testing and confidence intervals. The empirical likelihood weights are also asymptotically optimal for the estimation of totals and means. Calibration weights (Deville and Särndal, 1992) can be negative and not asymptotically optimal.

Chapter 4

Third Paper

Modelling hierarchical data under complex sampling designs: inference using an empirical likelihood approach

Melike Oguz-Alper and Yves G. Berger
University of Southampton, SO17 1BJ, Southampton, U.K.
M.OguzAlper@soton.ac.uk Y.G.Berger@soton.ac.uk

Abstract

The data used in social, behavioral, health or biological sciences may have a hierarchical structure due to the natural structure that occurs in the population of interest or due to the sampling or the experimental design itself. Multilevel or marginal models are often used to analyse such hierarchical data. The data may include sample units that may be selected with unequal probabilities from a clustered and stratified population. Inferences for the regression coefficients may be invalid when the sampling design is informative. We apply the profile empirical likelihood approach proposed by Oguz Alper and Berger (2015) to the regression parameters under a correlated error structure. The effect of the sampling design is taken into account. This approach can be used for point estimation, hypothesis testing and confidence intervals for the subvector of parameters. It asymptotically provides valid inference for the finite population parameters under a set of regularity conditions. We consider a two stage sampling design, where the first stage units may be selected with unequal probabilities. We assume that the model and sampling hierarchies are the same. We use general estimating equations to define the regression parameters under correlated error structures. We treat the ultimate clusters as the units of interest by using an ultimate cluster approach.
Keywords: Design based inference; general estimating equations; empirical likelihood; two stage sampling; uniform correlation structure; regression coefficients; unequal inclusion probabilities.

4.1 Introduction

The data used in social, behavioral, health or biological sciences may have a hierarchical structure due to the natural structure that occurs in the population of interest or due to the sampling or the experimental design itself. Multilevel models (Goldstein, 1986) or marginal models (e.g. Diggle et al., 2002) are often used to analyse such hierarchical data. The data may be collected from samples that are selected from a multi stage sampling design that may involve unequal probabilities at some or all stages of the selection. The sampling design is called informative when the selection probabilities are associated with the model outcome variable even after conditioning on the model covariates. Ignoring an informative sampling may result in invalid inference for regression parameters (e.g. Pfeffermann et al., 1998).

In standard, single level, regression models, sampling weights can be taken into account by using the *pseudo likelihood* approach (e.g. Binder, 1983; Skinner, 1989; Binder and Patak, 1994). With this approach, the population is fixed and population observations are assumed to be independent. In multilevel models, however, it is not straightforward to apply the pseudo likelihood approach as the observations within higher levels of the hierarchy are not independent. When this is the case, population totals cannot be written as a single summation of the individual units (e.g. Grilli and Pratesi, 2004).

Pfeffermann et al. (1998) proposed using probability weights, by relying on the pseudo likelihood principle, in the iterative generalised least squares (IGLS) algorithm to estimate multilevel regression parameters under two stage sampling design. They proposed scaling the survey weights of first level units to reduce the sampling bias when estimating variance components (see also Clogg and Eliason, 1987; Potthoff et al., 1992; Longford, 1995; Graubard and Korn, 1996; Asparouhov, 2006).

The IGLS estimation procedure may be computationally intensive as mentioned by Kovačević and Rai (2003). Alternatively, the *general estimating equations* (GEE) (e.g. Liang and Zeger, 1986; Diggle et al., 2002) that involve the use of a *working correlation structure* can be used to estimate regression parameters. Variance components are treated as fixed and replaced by their estimates when they are unknown. Liang and Zeger (1986) showed that the GEE estimator is fully efficient when the working correlation structure is correctly specified. More discussion about the asymptotic properties of the GEE estimators together with some empirical evidence can be found in the literature (e.g. Liang and Zeger, 1986; Crowder, 1995; Sutradhar and Das, 1999).

The general estimating equations provided by Liang and Zeger (1986) do not involve survey weights or the characteristics of the sampling design. Survey weights can be incorporated into the general estimating equations by following the pseudo likelihood approach (e.g. Binder, 1983; Skinner, 1989). The resulting approach is called the *multilevel pseudo likelihood* approach (e.g. Pfeffermann and La Vange, 1989; Kovačević and Rai,

2003; Grilli and Pratesi, 2004; Asparouhov, 2006; Skinner and De Toledo Vieira, 2007; De Toledo Vieira and Skinner, 2008). Asparouhov (2006) provided conditions under which the multilevel pseudo likelihood estimator is approximately unbiased. Skinner and De Toledo Vieira (2007) noticed that the weighted IGLS estimator (Pfeffermann et al., 1998) and the weighted GEE estimator are expected to provide identical point estimates under a working uniform correlation structure. Multilevel pseudo maximum likelihood estimation can be straightforwardly extended to multilevel generalised regression models (e.g. Sutradhar and Kovačević, 2000; La Vange et al., 2001; Grilli and Pratesi, 2004; Asparouhov, 2006; Asparouhov and Muthén, 2006; Rabe-Hesketh and Skrondal, 2006).

Standard confidence intervals rely on variance estimates. Variances of the estimators for the regression parameters can be computed through Taylor linearisation, the *sandwich type estimator*, (e.g. Binder, 1983; Pfeffermann et al., 1998; Skinner and De Toledo Vieira, 2007; Kovačević and Rai, 2003) or through the bootstrap (e.g. Grilli and Pratesi, 2004). The latter is very computationally intensive for hierarchical data. When the parameter of interest is a subvector of the parameters, the approach proposed by Binder and Patak (1994) can be applied to compute the conditional variance of the parameter of interest. When there is a bias in the variance estimators, standard confidence intervals may provide poor coverages.

Standard methods require the normality of the point estimators. The inference for the parameters may not be valid when the normality assumption does not hold. We propose using the profile empirical likelihood approach (e.g. Oguz Alper and Berger, 2015), which is based on the empirical likelihood approach proposed by Berger and De La Riva Torres (2016) to make inferences for regression parameters when modelling hierarchical data. We incorporate correlated error structure through the GEE (see Section 4.4). We use an *ultimate cluster approach* (Hansen et al., 1953). We treat the ultimate clusters as the units of interest. The empirical likelihood approach is applied at the ultimate cluster level. Estimating functions are defined as the sum of individual observations within clusters. This summation takes into account the correlation between any two observations in a given cluster.

Empirical likelihood inference allows us to investigate the design performance of the estimators. We assume that the sampling distribution is specified by the sampling design. Hence, we use design based confidence intervals that do not require the specification of the underlying model which may not be known. The model is used to define the point estimators through GEE (see Section 4.4). The sampling design is taken into account with the approach proposed. The resulting point estimators of the regression parameters are design consistent, which is a property often requested by survey practitioners (e.g. You and Rao, 2002). The resulting empirical likelihood confidence intervals may be better than the standard confidence intervals even when the point estimator is not normal, the variance estimators are biased or unstable or the individual error variances are heteroscedastic. The confidence intervals proposed do not rely on resampling, linearisation,

variance estimation or design effect. Population level information can be accommodated with the approach proposed.

In Section 4.2, we describe the two stage sampling design that we consider. In Section 4.3, we define the working model. In Section 4.4, we define the parameter of interest through GEE. In Section 4.5, we introduce the empirical log-likelihood function that we aim to maximise under a set of constraints. In Section 4.6, we demonstrate how the maximum empirical likelihood estimators of the regression coefficients are obtained under two stage sampling design. In Section 4.7, we explain how the estimating function is computed at the ultimate cluster level under a uniform covariance structure. We provide an expression for the inverse of the within covariance matrix that involves survey weights. In Section 4.7, we demonstrate how the cluster level general estimating function is estimated based on the sample data. In Section 4.7.1, we provide a scaling factor to scale the weights of the first level units. In Section 4.7.2, we show how the variance components can be estimated by incorporating survey weights. In Section 4.8, we define the profile empirical log-likelihood ratio function. In Sections 4.9-4.9.1, we compare the performance of the empirical likelihood confidence intervals with the standard confidence intervals. In Section 4.10, we briefly present our findings and discuss about possible extensions. The R (R Development Core Team, 2014) code is provided in Appendix B.

4.2 Two stage sampling design

Let U be a finite population comprised of N disjoint clusters U_i of sizes K_i, with $i = 1, \ldots, N$. Let s be the sample of clusters, called the *primary sampling units* (PSUs), of size n selected *with replacement* with unequal probabilities p_i (e.g. Hansen and Hurwitz, 1943) from U, where $\sum_{i \in U} p_i = 1$. We assume that both n and N are large. When the sampling fraction denoted by n/N is negligible, without replacement sampling with first-order inclusion probabilities given by

$$\pi_i = np_i, \tag{4.1}$$

where n is the fixed number of draws, is asymptotically equivalent to with-replacement sampling. In this case, the approach proposed is valid under selection of the PSUs without replacement with unequal probabilities given by (4.1). Let s_i be the sample of *secondary sampling units* (SSUs), of size k_i, with $j = 1, \ldots, k_i$, selected with conditional probabilities $\pi_{j|i}$ within the ith PSU selected at the first stage. Thus the unconditional probability of unit j in cluster i is given by $\pi_{ij} = \pi_{j|i}\, \pi_i$.

Let $\boldsymbol{v}_{ij}$ be a vector of the values of a set of variables associated with unit $j \in U_i$. We consider a *design-based approach*, where the sampling distribution of the sample data, $\{\boldsymbol{v}_{ij} : \ j \in s_i \text{ and } i \in s\}$, is only specified by the sampling design $\mathcal{P}(s)$. Thus we consider that the $\boldsymbol{v}_{ij}$ are fixed, non-random, vectors.

The sample data does not contain *independent and identically distributed* observations under two-stage sampling design. The identical distribution assumption does not hold because of the selection of first stage units, PSUs, with unequal probabilities. The independence assumption is broken due to the correlation expected between the second stage sampling units, SSUs, within the PSUs selected.

4.3 Working model

Consider $\boldsymbol{v}_{ij} = (y_{ij}, \boldsymbol{x}_{ij}^{\mathrm{T}})^{\mathrm{T}}$, where the y_{ij} are the observed values of a scalar variable of interest and the $\boldsymbol{x}_{ij}$ is the vector of values of the explanatory variables associated with the jth unit within the ith cluster, where $j = 1, \ldots, K_i$ and $i = 1, \ldots, N$. Let $\boldsymbol{y}$ be the vectors obtained from stacking the vectors $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iK_i})^{\mathrm{T}}$ and $\boldsymbol{X}$ be a fixed known data matrix obtained from stacking the $K_i \times b$ dimensional data matrices $\boldsymbol{X}_i = (\boldsymbol{x}_{i\cdot}^{(1)}, \cdots, \boldsymbol{x}_{i\cdot}^{(b)})$, where $\boldsymbol{x}_{i\cdot}^{(l)} = (x_{i1}^{(l)}, \cdots, x_{iK_i}^{(l)})^{\mathrm{T}}$, with $l = 1, \ldots, b$. Here, b is the number of covariates. Consider the *general linear regression*

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.2}$$

where $\boldsymbol{\beta}$ is a $b \times 1$ vector of fixed unknown parameters and $\boldsymbol{\epsilon}$ is the vector of residuals with a vector of mean $\mathbf{0}$ and a block-diagonal covariance structure $\boldsymbol{\mathcal{V}}$, where the $K_i \times K_i$ covariance matrices $\boldsymbol{V}_i$ are on the main diagonal and zeros elsewhere. The vector of residuals $\boldsymbol{\epsilon}$ is obtained from stacking the vectors $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iK_i})^{\mathrm{T}}$. The random variables ϵ_{ij}, $j = 1, \cdots, K_i$, are expected to be correlated within clusters. We specify a covariance structure for $\boldsymbol{V}_i$ to estimate unknown parameters $\boldsymbol{\beta}$.

Suppose that the *uniform correlation model* (Diggle et al., 2002, p.55) holds for the population data, $\{\boldsymbol{v}_{ij} : \ j \in U_i \text{ and } i \in U\}$. Under this model, the same correlation coefficient, ρ, is assumed between any given two observations within the same cluster. The uniform correlation model is equivalent to the following two-level model (e.g. Goldstein, 2011).

$$y_{ij} = \boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + u_i + e_{ij}, \tag{4.3}$$

where the u_i and the e_{ij} are independent random variables with means zero and variances σ_u^2 and σ_e^2 respectively. Here, the subscripts i and j are associated respectively with *level-two* and *level-one* of model (4.3). We will consider that the level-two units correspond to clusters while the level-one units refer to the units within clusters. We call the model (4.3) the *working model* (e.g. Skinner and De Toledo Vieira, 2007). This model is a particular case of *mixed effects models*. It is commonly used for small area estimation of means and called the *nested error model* in that context (e.g. Battese et al., 1988). Park and Fuller (2009) proposed a survey regression estimator based on a mixed model. They provide the conditions under which the mixed model estimator is design consistent for

finite population mean. Here, we only consider the inference for the finite population value of $\boldsymbol{\beta}$. It would be interesting to extend this work to make inference for the finite population mean of y_{ij}.

The covariance matrices $\boldsymbol{V}_i$, under the uniform correlation structure, are given by

$$\boldsymbol{V}_i = \sigma_e^2 \, \boldsymbol{I}_{K_i} + \sigma_u^2 \, (\mathbf{1}_{K_i} \times \mathbf{1}_{K_i}^{\mathrm{T}}), \tag{4.4}$$

where $\boldsymbol{I}_{K_i}$ is the $K_i \times K_i$ identity matrix and $\mathbf{1}_{K_i}$ is the $K_i \times 1$ column vector of ones (e.g. Rao, 2003, p.135). The correlation coefficient, ρ, is given by $\sigma_u^2/(\sigma_u^2 + \sigma_e^2)$. The same covariances σ_u^2 appear on the non diagonal parts of the matrices $\boldsymbol{V}_i$. Because of this, the structure is also called the *uniform covariance structure* under model (4.3).

4.4 Target finite population parameter and general estimating equation

Suppose that the working model (4.3) holds for the population data. We consider the $b \times 1$ finite population parameter vector $\boldsymbol{\beta}_N$, with $b = O(1)$, that is the unique solution of the following population estimating equation (Godambe, 1960).

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{g}_{i\cdot}(\boldsymbol{v}_{ij}, \boldsymbol{\sigma}, \boldsymbol{\beta}) = \mathbf{0}_b,$$

where $\boldsymbol{g}_{i\cdot}(\boldsymbol{v}_{ij}, \boldsymbol{\sigma}, \boldsymbol{\beta})$ is a $b \times 1$ vector of estimating functions associated with the second-level units under model (4.3), clusters, $\boldsymbol{v}_{ij}$ is the vector of variables corresponding to the unit j in the ith cluster and $\boldsymbol{\sigma}$ is a vector of variance components, that is, $\boldsymbol{\sigma} = (\sigma_e^2, \sigma_u^2)^{\mathrm{T}}$. We will use the notation $\boldsymbol{g}_{i\cdot}(\boldsymbol{\beta})$ in substitution for $\boldsymbol{g}_{i\cdot}(\boldsymbol{v}_{ij}, \boldsymbol{\sigma}, \boldsymbol{\beta})$ for simplicity.

Under the uniform correlation model, equivalently model (4.3), the parameter $\boldsymbol{\beta}_N$ is defined as the unique solution of the *general estimating equation* (GEE) (e.g. Liang and Zeger, 1986) given by

$$\boldsymbol{G}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{g}_{i\cdot}(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{X}_i^{\mathrm{T}} \boldsymbol{V}_i^{-1} (\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}) = \mathbf{0}_b, \tag{4.5}$$

where the $\boldsymbol{X}_i$ are known data matrices and the $\boldsymbol{y}_i$ are the vectors of response variables that are defined in Section 4.3 and the covariances $\boldsymbol{V}_i$ are given by expression (4.4). The cluster level estimating functions $\boldsymbol{g}_{i\cdot}(\boldsymbol{\beta})$ are the functions of the first-level unit observations, $\boldsymbol{v}_{ij}$. The correlation between observations within a cluster is incorporated through the $\boldsymbol{V}_i$. The inverses of the covariance matrices $\boldsymbol{V}_i$ are given by Rao (2003, p.135) as follows.

$$\boldsymbol{V}_i^{-1} = \sigma_e^{-2} \, \{\boldsymbol{I}_{K_i} - K_i^{-1} \, \gamma_i \, (\mathbf{1}_{K_i} \times \mathbf{1}_{K_i}^{\mathrm{T}})\}, \tag{4.6}$$

where $\boldsymbol{I}_{K_i}$ and $\mathbf{1}_{K_i}$ are defined in Section 4.3 and the constants γ_i are given by

$$\gamma_i = \sigma_u^2/(\sigma_u^2 + \sigma_e^2/K_i)\cdot \tag{4.7}$$

Under a set of regularity conditions given by Liang and Zeger (1986), the solution of the GEE (4.5) provides a consistent estimator of the infinite population parameter $\boldsymbol{\beta}$. Furthermore, the resulting estimator is fully efficient when the working model is correctly specified. The estimator $\boldsymbol{\beta}_N$ is also called the *generalised least square* (GLS) estimator. The GLS estimator is the *maximum likelihood estimator* when the vector of response variables $\boldsymbol{y}$ in (4.2) follows a multivariate normal distribution with known block-diagonal covariance structure $\boldsymbol{\mathcal{V}}$. Under the independence assumption, that is, $\rho = 0$, the GLS estimator reduces to the *ordinary least square* (OLS) estimator of $\boldsymbol{\beta}$.

The parameter $\boldsymbol{\beta}_N$ can also be obtained from solving the *mixed model* equations (Henderson et al., 1959). As shown by Rao (2003, p.97), the solution of (4.5) and the solution of the mixed model equations are identical. The use of the GEE (4.5) may be more convenient than the latter in the sense that we may specify other covariance structures in equation (4.5) that would work for the data. The choice of the covariance structure does not affect the consistency but only the efficiency of the inference for the regression parameter $\boldsymbol{\beta}$ (Liang and Zeger, 1986; Diggle et al., 2002).

The variance components are replaced by their estimates in (4.7) when they are unknown. In this case, the asymptotic properties of the estimator $\boldsymbol{\beta}_N$ may not always hold (e.g. Crowder, 1995). However, the consistency will be usually valid as shown by Sutradhar and Das (1999).

4.5 Empirical log-likelihood function

We consider the *empirical likelihood approach* that was proposed by Berger and De La Riva Torres (2016). The *empirical log-likelihood function* is given by

$$\ell(\boldsymbol{m}) = \sum_{i=1}^{n} \log m_i, \tag{4.8}$$

where the m_i, with $i \in s$, are unknown scale-loads allocated to data points $i \in s$ (Hartley and Rao, 1968) and $\boldsymbol{m}$ is the $n\times 1$ vector of the scale-loads m_i. Here, the subscript i refers to the PSUs selected with inclusion probabilities π_i (4.1) from the finite population U, s includes the PSUs selected and n is the sample size of the PSUs (see Section 4.2). Hartley and Rao (1969) showed that expression (4.8) is the empirical log-likelihood function for unequal probability sampling with replacement (see also Kim, 2009; Berger and De La Riva Torres, 2016).

The *maximum empirical likelihood estimators* $\widehat{m}_i$ are obtained by maximising the empirical log-likelihood function (4.8) with respect to the constraints: $m_i > 0$ and

$$\sum_{i=1}^{n} m_i \boldsymbol{c}_i = \boldsymbol{C}, \quad \text{with} \quad \boldsymbol{C} = \sum_{i=1}^{N} \boldsymbol{c}_i, \tag{4.9}$$

where $\boldsymbol{c}_i = \{\bar{\pi}^{-1}\boldsymbol{z}_i{}^{\mathrm{T}}, \mathbf{f}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}}\}^{\mathrm{T}}$ and $\boldsymbol{C} = (\bar{\pi}^{-1}\boldsymbol{n}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$. Here, $\bar{\pi} = n/N$ is the sampling fraction at the first stage of selection. The solution to this maximisation is invariant to $\bar{\pi}$ as it is a constant. It is only required to justify the asymptotic results (e.g. Berger and De La Riva Torres, 2016). Hence the sampling fraction $\bar{\pi}$ can be removed from both $\boldsymbol{c}_i$ and $\boldsymbol{C}$ when applying the approach. The $\boldsymbol{z}_i$ are the values of the stratification variables.

Suppose that the population U is stratified into H strata denoted by $U_1, \ldots, U_H$, with $\cup_{h=1}^{H} U_h = U$. Let s_h be a sample of PSUs of fixed size n_h selected from U_h with unequal probabilities with replacement. We assume that the number of strata H is bounded: $H = O(1)$. The design variables $\boldsymbol{z}_i$ are defined by $(z_{i1}, \ldots, z_{iH})^{\mathrm{T}}$, where $z_{ih} = \pi_i$ for $i \in U_h$ and $z_{ih} = 0$ otherwise. The $\boldsymbol{n}$ is the vector of stratum sample sizes defined by $\boldsymbol{n} = (n_1, \ldots, n_H)^{\mathrm{T}}$. With this constraint, we take into account the fact that samples of fixed sizes n_h are selected from each stratum U_h.

The efficiency of the inference for $\boldsymbol{\beta}$ can be increased by incorporating population level information, which may be available from administrative data and/or census data (e.g. Deville and Särndal, 1992; Chaudhuri et al., 2008). We can put this information into the $\boldsymbol{c}_i$ through estimating functions $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ (e.g. Berger and De La Riva Torres, 2016). Here, $\boldsymbol{\varphi}_N$ is a vector of known population level parameters that is the unique solution of the population estimating equation $\sum_{i=1}^{N} \mathbf{f}_i(\boldsymbol{\varphi}) = \mathbf{0}$. We assume that the $\mathbf{f}_i(\boldsymbol{\varphi})$ do not depend on $\boldsymbol{\beta}_N$. When $\boldsymbol{\varphi}_N$ is a function of variables associated with the SSUs , the $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ are replaced by their unbiased estimates (e.g Estevao and Särndal, 2006). For example, suppose that $\boldsymbol{\varphi}_N$ is a population mean that is defined by $\boldsymbol{\varphi}_N = \sum_{i=1}^{N}\sum_{j=1}^{K_i} \boldsymbol{\zeta}_{ij} / \sum_{i=1}^{N} K_i$, where $\boldsymbol{\zeta}_{ij}$ is a vector of auxiliary variables. In this case, an unbiased estimate is given by $\widehat{\mathbf{f}}_i(\boldsymbol{\varphi}_N) = \sum_{j \in s_i} (\boldsymbol{\zeta}_{ij} - \boldsymbol{\varphi}_N)\pi_{j|i}^{-1}$.

We assume that the $\boldsymbol{C}$ is an inner point of the conical hull formed by $\sum_{i=1}^{n} m_i \boldsymbol{c}_i$. Hence the set of $\widehat{m}_i$ is unique. We assume that $\boldsymbol{c}_i$ and $\boldsymbol{C}$ satisfy the regularity conditions given by Berger and De La Riva Torres (2016). As shown by Berger and De La Riva Torres (2016), using the method of Lagrange multipliers, we have

$$\widehat{m}_i(\boldsymbol{\varphi}_N) = (\pi_i + \boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{c}_i)^{-1}, \tag{4.10}$$

where the vector $\boldsymbol{\eta}$ is such that (4.9) and $m_i > 0$ hold. A modified Newton–Raphson algorithm as in Chen et al. (2002) can be used to compute $\boldsymbol{\eta}$. When we have a single stratum and do not use any population level information, we have $\boldsymbol{c}_i = \bar{\pi}^{-1}\pi_i$. In this case, it can be shown that $\boldsymbol{\eta} = \mathbf{0}$ and $\widehat{m}_i(\boldsymbol{\varphi}_N) = \widehat{m}_i = \pi_i^{-1}$, which is the standard Horvitz and Thompson (1952) weight for the ith PSU.

4.6 Maximum empirical likelihood point estimator under two stage sampling designs

Let $\widehat{m}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}_N)$ maximize $\ell(\boldsymbol{m})$ subject to the constraints $m_i > 0$ and

$$\sum_{i=1}^{n} m_i \, \boldsymbol{c}_i^*(\boldsymbol{\beta}) = \boldsymbol{C}^* \tag{4.11}$$

with

$$\boldsymbol{c}_i^*(\boldsymbol{\beta}) = \{\boldsymbol{c}_i{}^{\mathrm{T}}, \widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta})^{\mathrm{T}}\}^{\mathrm{T}} \quad \text{and} \quad \boldsymbol{C}^* = (\boldsymbol{C}^{\mathrm{T}}, \boldsymbol{0}^{\mathrm{T}})^{\mathrm{T}},$$

where $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta})$ is a sample based estimate of the $\boldsymbol{g}_{i\cdot}(\boldsymbol{\beta})$ defined by expression (4.5) for a given parameter vector $\boldsymbol{\beta}$. The maximum value of $\ell(\boldsymbol{m})$ under $m_i > 0$ and (4.11) is given by

$$\ell(\boldsymbol{\beta} \mid \boldsymbol{\varphi}_N) = \sum_{i=1}^{n} \log \widehat{m}_i^*(\boldsymbol{\beta}, \boldsymbol{\varphi}_N) \cdot \tag{4.12}$$

The *maximum empirical likelihood estimator* $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}_N$ is the vector that maximizes expression (4.12) over $\boldsymbol{\beta}$ for a given value of $\boldsymbol{\varphi}_N$. Berger and De La Riva Torres (2016) showed that $\widehat{\boldsymbol{\beta}}$ is the unique solution of the following sample level GEE.

$$\widehat{\boldsymbol{G}}(\boldsymbol{\beta}, \boldsymbol{\varphi}_N) = \sum_{i=1}^{n} \widehat{m}_i(\boldsymbol{\varphi}_N) \, \widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta}) = \boldsymbol{0}_b, \tag{4.13}$$

where the $\widehat{m}_i(\boldsymbol{\varphi}_N)$ are given by (4.10). We assume that there exists a design consistent estimator, the $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta})$, for the $\boldsymbol{g}_{i\cdot}(\boldsymbol{\beta})$. We also assume that the $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta}_N)$ satisfy the regularity conditions given by Berger and De La Riva Torres (2016). As shown by Berger and De La Riva Torres (2016), $\widehat{\boldsymbol{G}}(\boldsymbol{\beta}, \boldsymbol{\varphi}_N)$ is approximately equivalent to the asymptotic *design optimal regression estimator*. Under a one stratum design where no population level information is used, we have $\widehat{m}_i(\boldsymbol{\varphi}_N) = \widehat{m}_i = \pi_i^{-1}$. This yields the Horvitz-Thompson estimator of $\boldsymbol{G}(\boldsymbol{\beta})$ when $\boldsymbol{\beta} = \boldsymbol{\beta}_N$. In this case, the estimator $\widehat{\boldsymbol{\beta}}$ is design-consistent (e.g Godambe and Thompson, 2009) and called the *multilevel pseudo likelihood estimator* (e.g. Pfeffermann and La Vange, 1989; Kovačević and Rai, 2003; Grilli and Pratesi, 2004; Asparouhov, 2006; Skinner and De Toledo Vieira, 2007). Skinner and De Toledo Vieira (2007) noticed that the weighted IGLS estimator (Pfeffermann et al., 1998) and the weighted GEE estimator, that is, the estimator $\widehat{\boldsymbol{\beta}}$, are expected to provide identical point estimates under a working uniform correlation structure. Both the number of PSUs selected, n, and the sample sizes of clusters, k_i, should be large for design consistency. Pfeffermann et al. (1998, p.29) showed that consistency of $\widehat{\boldsymbol{\beta}}$ can be obtained by only assuming that n is large.

4.7 Estimation within clusters

Let $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta})$ be the sample based estimates of $\boldsymbol{g}_{i\cdot}(\boldsymbol{\beta})$ that are defined by

$$\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta}) = \boldsymbol{X}_i^{\mathrm{T}}\, \widehat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \tag{4.14}$$

where the $\widehat{\boldsymbol{V}}_i$ are called the *working covariance matrices* (e.g. Liang and Zeger, 1986). The inverse matrices $\widehat{\boldsymbol{V}}_i^{-1}$ can be viewed as sort of weights. Based on the expression given by Rao (2003, p.135), we derive the $\widehat{\boldsymbol{V}}_i^{-1}$ as follows.

$$\widehat{\boldsymbol{V}}_i^{-1} = \widehat{\sigma}_e^{-2}\,\{\mathrm{diag}_{1\le j\le k_i}(w_{j|i}) - \widehat{\gamma}_i\, w_{i\cdot}^{-1}\,(\boldsymbol{w}_i\,\boldsymbol{w}_i^{\mathrm{T}})\}, \tag{4.15}$$

where $w_{j|i} = (\pi_{j|i}a_i)^{-1}$, $\boldsymbol{w}_i = (w_{1|i},\ldots,w_{k_i|i})^{\mathrm{T}}$, with $j = 1,\ldots,k_i$ and $i = 1,\ldots,n$, where the a_i are some constants defined at cluster level, $w_{i\cdot} = \sum_{j=1}^{k_i} w_{j|i}$ and the $\widehat{\gamma}_i$ are defined by

$$\widehat{\gamma}_i = \widehat{\sigma}_u^2/(\widehat{\sigma}_u^2 + \widehat{\sigma}_{ei}^2), \tag{4.16}$$

where $\widehat{\sigma}_{ei}^2 = \widehat{\sigma}_e^2/w_{i\cdot}$. Here, $\widehat{\sigma}_e^2$ and $\widehat{\sigma}_u^2$ are sample based estimates, which are respectively given by expressions (4.21) and (4.22), for the variance components σ_e^2 and σ_u^2. The first-level error variance σ_e^2 has a multiplicative factor $w_{j|i}^{-1}$ in equation (4.15). In the expression given by Rao (2003, p.135), the constants that specify the heterogeneity of error variances are used as a multiplicative factor.

We call the constants a_i the *scaling factors* (e.g. Pfeffermann et al., 1998). The $\widehat{\gamma}_i$ (4.16) is a design consistent estimator of γ_i (4.7) when $a_i = 1$ and $\widehat{\sigma}_e^2$ and $\widehat{\sigma}_u^2$ are both design consistent. Pfeffermann et al. (1998, p.29) suggested scaling to reduce the bias in the estimators of variance components when cluster sample sizes k_i are small. In this case, variance components' estimators are not consistent. Bias reduction is achieved when the sampling design at the selection stage of the SSUs is ignorable. Consistency of $\widehat{\boldsymbol{\beta}}$ is still valid under scaling provided n is large and the scaling factors a_i are independent from the response variable y_{ij} (Pfeffermann et al., 1998, p.29).

When we plug the estimate $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta})$ (4.14) into the sample level GEE (4.13) and solve it for $\boldsymbol{\beta}$, we obtain the estimator

$$\widehat{\boldsymbol{\beta}} = \left\{\sum_{i=1}^{n} \widehat{m}_i(\boldsymbol{\varphi}_N)a_i^{-1}\sum_{j=1}^{k_i}\boldsymbol{x}_{ij}(\boldsymbol{x}_{ij} - \widehat{\gamma}_i\,\bar{\boldsymbol{x}}_i)^{\mathrm{T}}\pi_{j|i}^{-1}\right\}^{-1}\left\{\sum_{i=1}^{n} \widehat{m}_i(\boldsymbol{\varphi}_N)a_i^{-1}\sum_{j=1}^{k_i}\boldsymbol{x}_{ij}(y_{ij} - \widehat{\gamma}_i\,\bar{y}_i)\pi_{j|i}^{-1}\right\}, \tag{4.17}$$

where $\widehat{m}_i$ is defined by (4.10) and $\bar{\boldsymbol{x}}_i$ and $\bar{y}_i$ are weighted sample means corresponding to the ith cluster that are respectively defined by $\bar{\boldsymbol{x}}_i = \sum_{j\in s_i}\boldsymbol{x}_{ij}\pi_{j|i}^{-1}/\sum_{j\in s_i}\pi_{j|i}^{-1}$ and $\bar{y}_i = \sum_{j\in s_i} y_{ij}\pi_{j|i}^{-1}/\sum_{j\in s_i}\pi_{j|i}^{-1}$.

Under the independence assumption and when $a_i = 1$, the inverse covariance matrices $\widehat{\boldsymbol{V}}_i^{-1}$ are proportional to conditional design weights of SSUs , that is, $\widehat{\boldsymbol{V}}_i^{-1} = (\widehat{\sigma}_e^2 \pi_{j|i})^{-1}$. In this case, the $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta})$ is the Horvitz and Thompson (1952) estimator of $\boldsymbol{g}_{i\cdot}(\boldsymbol{\beta})$ for a given value of $\boldsymbol{\beta}$. Together with the independence assumption, when we have $\widehat{m}_i(\boldsymbol{\varphi}_N) = \widehat{m}_i = \pi_i^{-1}$ in (4.13), the solution of the expression (4.13) gives the simple *probability weighted least squares estimator*.

$$\widehat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^{n} \sum_{j=1}^{k_i} \boldsymbol{x}_{ij} \boldsymbol{x}_{ij}^{\mathrm{T}} \pi_{ij}^{-1} \right\}^{-1} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{k_i} \boldsymbol{x}_{ij} y_{ij} \pi_{ij}^{-1} \right\}, \tag{4.18}$$

where the $\pi_{ij} = \pi_i \pi_{j|i}$ is the joint inclusion probability associated with the jth unit in the ith cluster. The estimator $\widehat{\boldsymbol{\beta}}$ is consistent for the finite population parameter $\boldsymbol{\beta}_N$ with respect to the sampling distribution (e.g. Isaki and Fuller, 1982).

4.7.1 Scaling

Several scaling methods have been proposed (e.g. Clogg and Eliason, 1987; Potthoff et al., 1992; Longford, 1995; Graubard and Korn, 1996; Pfeffermann et al., 1998). We consider the scaling method proposed by Potthoff et al. (1992), where the scaling factors are defined by

$$a_i = \sum_{j \in s_i} \pi_{j|i}^{-2} / \sum_{j \in s_i} \pi_{j|i}^{-1}. \tag{4.19}$$

This scaling method was motivated by You and Rao (2002) through linking a cluster level model, which is assumed to be valid for the sample data, and the direct design-based estimators $\bar{y}_i$ of cluster means. This model may be obtained from aggregating the unit level model (4.3) as follows (Rao, 2003, p.149).

$$\bar{y}_i = \bar{\boldsymbol{x}}_i^{\mathrm{T}} \boldsymbol{\beta} + u_i + \bar{\epsilon}_i \tag{4.20}$$

where $\bar{y}_i, \bar{\boldsymbol{x}}_i$ and $\bar{\epsilon}_i$ are the sample means that are respectively defined by $\bar{y}_i = \sum_{j \in s_i} y_{ij} \pi_{j|i}^{-1} / \sum_{j \in s_i} \pi_{j|i}^{-1}$, $\bar{\boldsymbol{x}}_i = \sum_{j \in s_i} \boldsymbol{x}_{ij} \pi_{j|i}^{-1} / \sum_{j \in s_i} \pi_{j|i}^{-1}$ and $\bar{\epsilon}_i = \sum_{j \in s_i} \epsilon_{ij} \pi_{j|i}^{-1} / \sum_{j \in s_i} \pi_{j|i}^{-1}$. Under model (4.20), $\mathrm{var}(\bar{\epsilon}_i) \equiv \sigma_{ei}^2 = \sigma_e^2 a_i / \widehat{K}_i$ (see also Potthoff et al., 1992, p.385).

The sum of the scaled weights, $w_{i\cdot}$, was considered as the *effective sample size* by Potthoff et al. (1992). They proposed adjusting survey weights by using the scaling coefficients (4.19) to take into account the effect of the sampling design when making inferences for infinite population parameters. They also considered the case in which the error variances differ and are proportional to some known constants. The scaling method based on (4.19) is called *scaling method 1* by Pfeffermann et al. (1998). They interpreted the scaling factor (4.19) as a *design effect*.

Under simple random sampling at both stages of the sampling design and when $\widehat{m}_i(\boldsymbol{\varphi}_N) = \widehat{m}_i = \pi_i^{-1}$, the estimator (4.17) reduces to the standard *empirical best linear unbiased prediction* (EBLUP) estimator of $\boldsymbol{\beta}_N$ after applying this scaling method. The standard estimator under simple random sampling is not obtained if the scaling is not applied (Huang and Hidiroglou, 2003).

Scaling is used as a tool to reduce the bias of the estimators when cluster sample sizes are small. Asparouhov (2006) compared the empirical biases of the estimators under several scaling methods. However, there has been no theoretical evidence supporting which scaling method is better for what kind of parameters and under what type of sampling designs. The consistency for $\boldsymbol{\beta}_N$ is achieved regardless of the scaling method provided the number of clusters n is large and the scaling factors a_i are independent from the response variable y_{ij} (Pfeffermann et al., 1998, p.29). However, how the efficiency of the inference about the regression parameters and the variance components are affected by the chosen scaling method has not been clearly specified in the literature. In this respect, it is worth developing a general theory for scaling as a future work.

4.7.2 Estimation of variance components

The assumption of known variance components, σ_e^2 and σ_u^2, may not be ideal in practice. Several approaches have been considered to estimate variance components (e.g. Prasad and Rao, 1990; Searle et al., 1992; Longford, 1995; Graubard and Korn, 1996; Pfeffermann et al., 1998; You and Rao, 2002; Huang and Hidiroglou, 2003; Korn and Graubard, 2003; Breidt et al., 2005; De Toledo Vieira and Skinner, 2008). Design and model properties of some of those estimators are discussed by Korn and Graubard (2003) and Huang and Hidiroglou (2003). Prasad and Rao (1990) (see also You and Rao, 2002) proposed variance estimators that are based on the method of moments. As noted by Rao (2003, p.138), these estimators are the same as those obtained from the method of *fitting-of-constants*, which is also known as Henderson's (1953) Method III. We incorporate survey weights into these estimators as seen in the equations (4.21) and (4.22).

$$\widehat{\sigma}_e^2 = \frac{1}{\sum_{i=1}^{n} \pi_i^{-1} \left(\sum_{j=1}^{k_i} w_{j|i} - 1 \right) - b + 1} \sum_{i=1}^{n} \pi_i^{-1} \sum_{j=1}^{k_i} w_{j|i}\, \widehat{\varepsilon}_{ij}^2, \tag{4.21}$$

where the $w_{j|i}$ are the scaled weights defined in Section 4.7, b is the dimension of the covariates $\boldsymbol{x}_{ij}$ in model (4.3), the $\widehat{\varepsilon}_{ij}$ are the residuals obtained from the OLS regression of $y_{ij} - \bar{y}_i$ on $\boldsymbol{x}_{ij} - \bar{\boldsymbol{x}}_i$ without intercept, where the $\boldsymbol{x}_{ij}$ is the $b \times 1$ vector of covariates for the jth unit in the ith cluster, $\bar{y}_i = \sum_{j \in s_i} y_{ij} w_{j|i} / \sum_{j \in s_i} w_{j|i}$ and $\bar{\boldsymbol{x}}_i = \sum_{j \in s_i} \boldsymbol{x}_{ij} w_{j|i} / \sum_{j \in s_i} w_{j|i}$.

$$\widehat{\sigma}_u^2 = \frac{1}{M^*} \left\{ \sum_{i=1}^{n} \pi_i^{-1} \sum_{j=1}^{k_i} w_{j|i}\, \widehat{v}_{ij}^2 - \left(\sum_{i=1}^{n} \pi_i^{-1} \sum_{j=1}^{k_i} w_{j|i} - b \right) \widehat{\sigma}_e^2 \right\}, \tag{4.22}$$

where the $\widehat{v}_{ij}$ are the residuals obtained from the OLS regression of y_{ij} on $\boldsymbol{x}_{ij}$ with intercept and

$$M^* = \sum_{i=1}^{n}\sum_{j=1}^{k_i} w_{ij} - \operatorname{tr}\left\{\left(\sum_{i=1}^{n}\pi_i^{-1}\sum_{j=1}^{k_i} w_{j|i}\,\boldsymbol{x}_{ij}\,\boldsymbol{x}_{ij}^{\mathrm{T}}\right)^{-1}\sum_{i=1}^{n}\pi_i^{-1}\left(\sum_{j=1}^{k_i} w_{j|i}\right)^2\overline{\boldsymbol{x}}_i\overline{\boldsymbol{x}}_i^{\mathrm{T}}\right\},$$

with $w_{ij} = \pi_i^{-1}w_{j|i}$. The estimators (4.21) and (4.22) are equivalent to the estimators that are given by Graubard and Korn (1996) under model (4.3) with $b = 1$ and $\boldsymbol{x} = 1$. They are approximately unbiased when both n and k_i are large (Korn and Graubard, 2003). Scaling reduces the bias when k_i is small (see Pfeffermann et al., 1998, p.29). Huang and Hidiroglou (2003) recommended using these estimators without scaling of weights under informative sampling at both stages of selection. Korn and Graubard (2003) proposed estimators that are approximately design unbiased. However, these estimators require second order inclusion probabilities. Huang and Hidiroglou (2003, p.1901) provided adjusted versions of the estimators (4.21) and (4.22). They showed that the adjusted versions are both design and model consistent. Breidt et al. (2005) used the restricted maximum likelihood sample-weighted estimating equations for the estimation of the variance components. The solutions to those estimating equations are design-consistent for σ_e^2 and σ_u^2 (Breidt et al., 2005, p.838).

Poor estimation of the variance components may result in some lost in efficiency of the inference for the finite population parameter $\boldsymbol{\beta}_N$. However, the consistency will still hold provided the number of PSUs n is large (e.g. Pfeffermann et al., 1998). Park and Fuller (2009) showed that the mixed model estimator they proposed is design consistent even when the variance component σ_u^2 is poorly estimated. The empirical study they conducted shows that the poor estimation of σ_u^2 causes a modest lost in efficiency of the inference for the finite population mean. It would be interesting to investigate how sensitive the inference for the regression parameter $\boldsymbol{\beta}_N$ would be to the chosen estimation method for variance components.

4.8 Empirical likelihood inference under two stage sampling design: testing and confidence intervals

Let $\boldsymbol{\beta}_N = (\boldsymbol{\theta}_N^{\mathrm{T}}, \boldsymbol{\nu}_N^{\mathrm{T}})^{\mathrm{T}}$ be the $(p+q)\times 1$ vector of parameters that is the unique solution of the population GEE (4.5). Let $\boldsymbol{\Theta}$ and $\boldsymbol{\Lambda}$ be compact sets. Suppose that $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^p$ is the $p\times 1$ vector of the parameter of interest and $\boldsymbol{\nu} \in \boldsymbol{\Lambda} \subset \mathbb{R}^q$ is the $q\times 1$ vector of parameters that are not of primary interest (Oguz Alper and Berger, 2015). The latter is called the *nuisance* parameter (e.g. Qin and Lawless, 1994). We recall the *profile empirical log-likelihood ratio function* proposed by Oguz Alper and Berger (2015) to make inferences for $\boldsymbol{\theta}_N$ in the presence of nuisance parameters $\boldsymbol{\nu}_N$. We treat the

estimates of variance components σ_e^2 and σ_u^2 as fixed. Hence we do not consider the fact that they are estimated. The profile empirical log-likelihood ratio function is defined by

$$\widehat{r}(\boldsymbol{\theta} \mid \boldsymbol{\varphi}_N) = 2\left\{\ell(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{\varphi}_N) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}, \boldsymbol{\nu} \mid \boldsymbol{\varphi}_N)\right\}, \tag{4.23}$$

where $\ell(\boldsymbol{\theta}, \boldsymbol{\nu} \mid \boldsymbol{\varphi}_N) = \ell(\boldsymbol{\beta} \mid \boldsymbol{\varphi}_N)$ with $\boldsymbol{\beta} = (\boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\nu}^{\mathrm{T}})^{\mathrm{T}}$ and $\ell(\widehat{\boldsymbol{\beta}} \mid \boldsymbol{\varphi}_N) = \sum_{i\in s} \log \widehat{m}_i(\boldsymbol{\varphi}_N)$, where the $\widehat{m}_i(\boldsymbol{\varphi}_N)$ are defined by (4.10). Under a set of regularity conditions given by Berger and De La Riva Torres (2016) and assuming that $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ and $\widehat{\boldsymbol{g}}_{i\cdot}(\boldsymbol{\beta})$ satisfy those conditions, Oguz Alper and Berger (2015) showed that the empirical log-likelihood ratio function $\widehat{r}(\boldsymbol{\theta} \mid \boldsymbol{\varphi}_N)$ asymptotically converges to an expression with a quadratic form that depends on the estimator (4.13) and the Hansen and Hurwitz (1943) estimator of the variance of (4.13). We consider applying an *ultimate cluster approach* (Hansen et al., 1953). In this approach, the variance estimator corresponds to the variance among the PSUs. It is consistent when the sampling fraction of the PSUs is negligible, that is, $n/N = o(1)$ (e.g. Särndal et al., 1992, Ch.4). The square root of the quadratic form follows a normal distribution as the estimator (4.13) is approximately equivalent to the regression estimator that is normally distributed as shown by Scott and Wu (1981) and the variance estimator is consistent when $n/N = o(1)$. Hence Oguz Alper and Berger (2015) showed that

$$\widehat{r}(\boldsymbol{\theta}_N \mid \boldsymbol{\varphi}_N) \sim \chi^2_{df=p}. \tag{4.24}$$

The property (4.24) can be used for testing and constructing confidence intervals. Suppose we wish to test $H_0 : \boldsymbol{\theta}_N = \boldsymbol{\theta}_N^0$ versus $H_1 : \boldsymbol{\theta}_N \neq \boldsymbol{\theta}_N^0$. The *p-value* is $\int_{\widehat{r}(\boldsymbol{\theta}_N^0)}^{\infty} \chi^2_{df=p}(x)\,\mathrm{d}x$, where $\chi^2_{df=p}(x)$ is the density of a chi-squared distribution with p degrees of freedom.

We consider construction of confidence intervals for a scalar, $p = 1$, by using the property (4.24). Let θ_N be a scalar subparameter of $\boldsymbol{\beta}_N$. In this case, $\widehat{r}\,(\theta_N \mid \boldsymbol{\varphi}_N)$ follows asymptotically a chi-squared distribution with one degree of freedom. Thus the $\alpha\%$ empirical likelihood Wilks's (1938) type confidence interval for θ_N is given by $\{\theta : \widehat{r}\,(\theta \mid \boldsymbol{\varphi}_N) \leqslant \chi^2_{df=1}(\alpha)\}$, where $\chi^2_{df=1}(\alpha)$ is the upper α-quantile of the chi-squared distribution with one degree of freedom. The $\widehat{r}\,(\theta \mid \boldsymbol{\varphi}_N)$ is a convex function of θ, for given value of $\boldsymbol{\varphi}_N$, with a minimum value when θ is equal to the empirical maximum likelihood estimator $\widehat{\theta}$ that is the unique solution of (4.13). The bisection method can be used to find the lower and upper bounds. This involves calculating $\widehat{r}\,(\theta \mid \boldsymbol{\varphi}_N)$ for several values of θ.

4.9 Simulation study

In this Section, we present some numerical results for the parameters of a hierarchical linear model defined by (4.3). Our simulation study shows the design performance of

the confidence intervals. We selected 1000 random samples with respect to a two-stage sampling design from a finite population that is a realisation of the infinite population model given by

$$y_{ij} = \beta_0 + \beta_1 x_{ij}^{(1)} + \beta_2 x_{ij}^{(2)} + u_i + e_{ij}, \tag{4.25}$$

where $\beta_0 = 20$, $\beta_1 = \beta_2 = 1$, $x_{ij}^{(1)} \sim \text{rgamma}(K_i, \text{shape} = 2, \text{scale} = \alpha_1)$ and $x_{ij}^{(2)} \sim \text{rgamma}(K_i, \text{shape} = 2, \text{scale} = \alpha_2)$, where α_1 and α_2 are selected randomly with replacement among the values $(1, 2, 3)$ and $(1, 2, 3, 4)$ respectively. The K_i are the cluster sizes generated from a lognormal distribution and defined by $K_i = 100\exp(\tau_i)$, with $\tau_i \sim N(0, 0.2)$. Here, the number of clusters is $N = 3000$. The minimum and the maximum values of the cluster sizes are respectively 47 and 207. The u_i are the random effects following a normal distribution with mean zero and standard deviation σ_u. The e_{ij} are the level one residuals that were generated from a chi-squared distribution, that is, $e_{ij} \sim \chi^2(\sigma_e^2/2) - \sigma_e^2/2$, where σ_e^2 is the variance of e_{ij}. The values of the variances σ_e^2 and σ_u^2 were chosen based on the intra-cluster correlation that is defined by $\rho = \sigma_u^2/\sigma^2$, where $\sigma^2 = \sigma_e^2 + \sigma_u^2$. The total variance, σ^2, was kept fixed at 12. We considered seven different values for the correlation coefficient that lie in the range $[0.04, 0.50]$. Population size is $\sum_{i=1}^{N} K_i = 305{,}305$. The finite population parameters, $(\beta_{0N}, \beta_{1N}, \beta_{2N})$ (see Table 4.1), under model (4.25) are obtained by solving the population GEE (4.5).

We selected samples with respect to a two-stage sampling design. At the first stage, a sample of clusters, PSUs,, s, of size $n = 150$ was selected with randomized systematic sampling with unequal probabilities π_i proportional to the measure of size $\delta_i = b_0 + u_i + b_1\epsilon_i$, where $\epsilon_i \sim exponential(rate = 1) - 1$ and $\text{corr}(\delta_i, u_i) \approx 0.85$. The constant b_0 was used to avoid very small inclusion probabilities for the PSUs. The constant b_1 was used to control the correlation between δ_i and u_i. At the second stage, samples of the SSUs, the s_i, of sizes $k_i = \alpha K_i$ were selected with simple random sampling without replacement within the PSUs selected at the first stage, where $\alpha = 0.25$ in Tables 4.1–4.2 and $\alpha = \{0.10, 0.25, 0.40\}$ in Tables 4.3–4.4. The ranges of sample sizes within clusters are given by $[5, 21]$ when $\alpha = 0.10$, $[12, 52]$ when $\alpha = 0.25$ and $[19, 83]$ when $\alpha = 0.40$.

We compare the Monte Carlo design-based performance of the empirical likelihood confidence interval with the standard parametric confidence interval and a non-parametric confidence interval that is based on the pseudo likelihood approach (e.g. Binder, 1983; Pfeffermann and La Vange, 1989; Binder and Patak, 1994; Skinner and De Toledo Vieira, 2007). The pseudo likelihood confidence interval relies on variance estimation. We used the (Hartley and Rao, 1962) variance estimator for this approach.

The pseudo likelihood confidence interval is obtained by solving a system of equations that rely on an estimating function and its variance estimator. The estimating function is defined by using Binder and Patak's (1994) approach that takes into account the nuisance parameters. Variance estimator is computed at point estimates of the parameters. Here,

the pseudo likelihood confidence interval is the same as the one that relies on a linearised variance estimator that is obtained by using the 'delta method' (e.g. Binder, 1983; Pfeffermann et al., 1998; Skinner and De Toledo Vieira, 2007). This approach assumes the normality of the point estimator.

Standard parametric confidence intervals involve maximum likelihood estimation. This method is based on the normality assumption. The hierarchical structure is considered by fitting a two level model with a uniform covariance structure. Survey weights are not taken into account with this approach. Point estimates of the parameters and their standard errors were obtained by using the 'lme' function in R (R Development Core Team, 2014).

We consider 95% confidence intervals for finite population parameters that are estimates of regression coefficients in working model (4.25). We used Shapiro and Wilk's (1965) test statistics to test the normality of the point estimators. Significance of observed coverages and tail errors was tested by using a z-test for proportions given by $z^2 = (p - P_0)^2/(P_0(1 - P_0)/1000)$, where p is observed coverage or tail error rate and P_0 is the nominal value equal to 0.95 or 0.025. We have asymptotically $z^2 \rightarrow \chi^2_{df=1}$ in distribution as the distribution of z-statistic is approximated by a standard normal distribution under large sample (De Moivre, 1733). We also considered the percentage relative bias (%) of the point estimators and the standard error estimators with respect to the sampling design. The percentage relative bias (%) is defined by $\text{RB}\% = [\{E(\widehat{\phi}) - \phi\}/\phi] * 100\%$, where $E(\widehat{\phi}) = M^{-1} \sum_{m=1}^{M} \widehat{\phi}_m$, with $M = 1000$, is the empirical expectation of the estimator $\widehat{\phi}$, where ϕ is the quantity of interest and $\widehat{\phi}_m$ is an estimate of the quantity ϕ in the mth sample. Here, the quantity ϕ may refer to the finite population parameter when $\widehat{\phi} = \widehat{\beta}_N$ or the empirical standard error (se) when $\widehat{\phi} = \widehat{se}(\widehat{\beta}_N)$, where $\widehat{\beta}_N = \{\widehat{\beta}_{0N}, \widehat{\beta}_{1N}, \widehat{\beta}_{2N}\}$. The empirical standard error is computed by $(M-1)^{-1} \sum_{m=1}^{M} (\widehat{\phi}_m - E(\widehat{\phi}))^2$, with $\widehat{\phi} = \widehat{\beta}_N$. We consider two estimators denoted by $\widehat{\boldsymbol{\beta}}_N^{el}$ and $\widehat{\boldsymbol{\beta}}_N^{ml}$ in Table 4.1, where $\widehat{\boldsymbol{\beta}}_N^{el} = (\widehat{\beta}_{0N}^{el}, \widehat{\beta}_{1N}^{el}, \widehat{\beta}_{2N}^{el})^{\text{T}}$ and $\widehat{\boldsymbol{\beta}}_N^{ml} = (\widehat{\beta}_{0N}^{ml}, \widehat{\beta}_{1N}^{ml}, \widehat{\beta}_{2N}^{ml})^{\text{T}}$. Weights are incorporated with the empirical likelihood approach. The estimator $\widehat{\boldsymbol{\beta}}_N^{el}$ is the solution of the sample GEE (4.13). We have the same estimator with the pseudo likelihood approach as we did not use population level information and the weights are equivalent to the Horvitz and Thompson (1952) weights in this case (see Section 4.6). The estimator $\widehat{\boldsymbol{\beta}}_N^{ml}$ does not involve weights. In Table 4.3, the point estimators are identical due to the equal selection probabilities.

In Table 4.1, we have the finite population values of regression coefficients and the relative bias (%) of their estimators. There is a bias in the estimator of the intercept, $\widehat{\beta}_{0N}^{ml}$, when the sampling weights are not incorporated. The relative bias (%) is negligible for the estimators of the fixed effects, β_{1N} and β_{2N}. It is not surprising to observe a bias in the estimator $\widehat{\beta}_{0N}^{ml}$ as the selection of the PSUs are proportional to random effects given in (4.25). Relative bias (%) increases with the intra-cluster correlation. We observe that the relative bias (%) is negligible for the estimators that involve weights.

Table 4.1: Finite population values of the regression coefficients in working model (4.25) and the relative bias (%) of their estimators. Two stage sampling design. Unequal probability selection of the PSUs. $N = 3000$, $n = 150$, $range(k_i) = [12, 52]$.

Intra-cluster correlation	Population value			Relative bias (%) *unweighted*			*weighted*		
ρ	β_{0N}	β_{1N}	β_{2N}	$\widehat{\beta}_{0N}^{ml}$	$\widehat{\beta}_{1N}^{ml}$	$\widehat{\beta}_{2N}^{ml}$	$\widehat{\beta}_{0N}^{el}$	$\widehat{\beta}_{1N}^{el}$	$\widehat{\beta}_{2N}^{el}$
0.04	20.001	1.002	1.002	0.64	-0.01	-0.09	0.03	-0.02	-0.07
0.08	20.013	1.003	1.000	0.86	0.00	-0.08	0.05	0.00	-0.07
0.17	20.030	1.004	0.999	1.55	0.14	0.02	-0.03	0.03	0.05
0.25	20.047	0.998	1.000	1.99	0.08	0.06	-0.05	0.08	0.02
0.33	20.038	1.003	1.000	2.09	0.08	0.07	-0.03	-0.02	0.01
0.42	20.056	1.002	1.001	2.40	-0.13	-0.16	-0.04	-0.06	-0.11
0.50	20.059	1.000	1.001	2.60	-0.02	-0.07	-0.02	-0.01	-0.08

In Table 4.2, we have the observed coverages of the 95% confidence intervals. All coverages for the intercept are significantly different from the nominal level (0.95%) with maximum likelihood approach. The Shapiro & Wilk's test of normality shows that the normality of point estimators are not rejected at significance level 0.05. This may not necessarily hold at significance level 0.10. We observe that coverages for the maximum likelihood estimate of the intercept, β_{0N}, decrease with the intra-cluster correlation. This may be because of the fact that the point estimator has higher bias for large values of the correlation (see Table 4.1). We observe slightly poor coverages with the pseudo likelihood approach in most cases. The empirical likelihood confidence intervals have better coverages overall.

4.9.1 Population with outlying values

In this Section, we investigate the performance of the empirical likelihood confidence intervals in the presence of outlying values in the population. We generated the same population as in Section 4.9, except that here, the random variables e_{ij} follow a normal distribution with mean zero and standard deviation σ_e. We replaced 20% of the y_{ij} randomly by very large values. Half of the new values was generated from $Y_{0.75} + 1.5 \times (Y_{0.75} - Y_{0.25}) + \tau_{ij}$, where $\tau_{ij} \sim gamma(shape = 2, scale = 2) - 4$. Here, $Y_{0.25}$ and $Y_{0.75}$ are the lower and upper population quartiles of the y_{ij}, where $j \in U_i$ and $i \in U$. The other half was obtained from $\max_{j \in U_i, i \in U}(y_{ij}) + \tau_{ij}$.

Sample units were selected with simple random sampling without replacement at both stages of the sampling design. We consider small and large sample sizes, varying from 10% to 40% of cluster sizes, K_i, for the SSUs in Table 4.4. Sample sizes respectively lie in the ranges $[5, 21]$, $[12, 52]$ and $[19, 83]$. In Table 4.4, we only present the results for $\rho = 0.50$ as we obtained very similar results for the other values of the intra-cluster

Table 4.2: 95% confidence intervals for the estimates of regression coefficients in working model (4.25). Two stage sampling design. Unequal probability selection of the PSUs. $N = 3000$, $n = 150$, $range(k_i) = [12, 52]$.

Intra-cluster correlation ρ	Parameter $\boldsymbol{\beta}_N$	Empirical likelihood (%)	Maximum likelihood (%)	Pseudo likelihood (%)	Shapiro Wilk p-value		
					EL	ML	Pseudo
0.04	β_{0N}	95.4	83.4*	94.7	0.83	0.85	0.83
	β_{1N}	94.8	95.8	94.4	0.77	0.73	0.77
	β_{2N}	94.8	95.6	94.0	0.95	0.95	0.95
0.08	β_{0N}	95.1	76.3*	94.8	0.30	0.44	0.30
	β_{1N}	96.1	97.1*	95.5	0.18	0.06	0.18
	β_{2N}	94.8	94.4	94.2	0.63	0.29	0.63
0.17	β_{0N}	94.1	47.3*	93.5*	0.71	0.90	0.71
	β_{1N}	94.8	94.9	93.9	0.72	0.80	0.72
	β_{2N}	95.0	95.7	94.5	0.23	0.35	0.23
0.25	β_{0N}	95.2	33.2*	94.7	0.17	0.76	0.17
	β_{1N}	94.1	94.2	93.2*	0.37	0.07	0.37
	β_{2N}	94.4	95.4	93.7	0.75	0.30	0.75
0.33	β_{0N}	94.7	37.6*	95.0	0.16	0.54	0.16
	β_{1N}	95.2	95.5	94.3	0.58	0.51	0.58
	β_{2N}	94.8	94.5	94.5	0.76	0.66	0.76
0.42	β_{0N}	95.2	32.1*	94.6	0.79	0.52	0.79
	β_{1N}	95.1	95.8	93.7	0.11	0.10	0.11
	β_{2N}	94.0	95.1	93.8	0.65	0.08	0.65
0.50	β_{0N}	95.2	28.5*	94.6	0.59	0.68	0.59
	β_{1N}	93.1*	93.8	92.6*	0.34	0.47	0.34
	β_{2N}	94.9	96.3	94.2	0.35	0.71	0.35

$*$ Coverages significantly different from 95%. p-value ≤ 0.05.

correlation that were used to generate the y_{ij} before introducing outlying values. This is because of the fact that the estimate of between cluster variance σ_u^2 was relatively too small comparing with the estimate of within cluster variance σ_e^2 after introducing the outlying values. Empirical expectation of the estimator of the intra-cluster correlation was approximately equal to 0.06 for all cases.

Table 4.3 presents the relative bias (%) of the point estimators and the standard error estimators. The point estimators are the same for all methods due to the equal selection probabilities. The estimator $\widehat{\beta}_{2N}$ slightly underestimates the population parameter β_{2N}. There is a large positive bias in the standard error estimators, $\widehat{se}(\widehat{\beta}_{1N})$ and $\widehat{se}(\widehat{\beta}_{2N})$, with

the maximum likelihood approach. We observe that the standard errors of the point estimators are slightly underestimated with the pseudo likelihood approach for all cases, except the case with the estimator $\widehat{\beta}_{1N}$ and $k_i = 0.40K_i$.

Table 4.3: Relative bias (%) of the point estimators and the standard error estimators. Population with outlying values. Two stage sampling design. Simple random sampling at both stages. $N = 3000$, $n = 150$, $range(k_i) = [5, 21]$ when $k_i = 0.10K_i$, $range(k_i) = [12, 52]$ when $k_i = 0.25K_i$ and $range(k_i) = [19, 83]$ when $k_i = 0.40K_i$.

Approach	Second stage sample size	Relative bias (%)					
	(k_i)	$\widehat{\beta}_{0N}$	$\widehat{\beta}_{1N}$	$\widehat{\beta}_{2N}$	$\widehat{se}(\widehat{\beta}_{0N})$	$\widehat{se}(\widehat{\beta}_{1N})$	$\widehat{se}(\widehat{\beta}_{2N})$
Maximum	$0.10K_i$	0.34	-0.07	-2.07	-1.15	14.75	17.15
likelihood	$0.25K_i$	0.15	0.41	-1.19	-0.06	15.13	19.01
	$0.40K_i$	0.11	0.12	-0.78	1.72	19.41	19.12
Pseudo	$0.10K_i$	0.34	-0.07	-2.07	-3.16	-0.81	-2.36
likelihood	$0.25K_i$	0.15	0.41	-1.19	-2.93	-0.75	-1.29
	$0.40K_i$	0.11	0.12	-0.78	-1.39	2.29	-1.33

Table 4.4: 95% confidence intervals for the estimates of regression coefficients in working model (4.25). Population with outlying values. Two stage sampling design. Simple random sampling at both stages. $N = 3000$, $n = 150$, $range(k_i) = [5, 21]$ when $k_i = 0.10K_i$, $range(k_i) = [12, 52]$ when $k_i = 0.25K_i$ and $range(k_i) = [19, 83]$ when $k_i = 0.40K_i$.

Second stage sample size (k_i)	Parameter $\boldsymbol{\beta}_N$	Empirical likelihood (%)	Maximum likelihood (%)	Pseudo likelihood (%)	Shapiro Wilk p-value
$0.10K_i$	β_{0N}	94.8	94.2	94.0	0.01
	β_{1N}	95.1	97.5*	94.2	0.98
	β_{2N}	93.8	97.2*	93.6*	0.16
$0.25K_i$	β_{0N}	94.7	94.6	94.2	0.37
	β_{1N}	94.9	97.3*	94.1	0.60
	β_{2N}	94.3	97.8*	93.7	0.64
$0.40K_i$	β_{0N}	94.1	94.5	93.5*	0.11
	β_{1N}	95.4	97.4*	95.0	0.48
	β_{2N}	95.0	97.7*	94.5	0.86

$*$ Coverages significantly different from 95%. p-value ≤ 0.05.

Table 4.4 demonstrates observed coverages of the 95% confidence intervals for the parameters of the population with outlying values. We have better coverages with the empirical likelihood approach comparing with the maximum likelihood and the pseudo

likelihood approaches. The coverages of maximum likelihood confidence intervals for the slope parameters are significantly different from the nominal level, 95%. Point estimates are the same for all the three approaches due to the equal probability selection. The Shapiro & Wilk p-value provides evidence that the point estimator of the parameter β_{0N} does not follow a normal distribution when $k_i = 0.10K_i$. This may be the reason for poor coverages of the maximum likelihood and the pseudo likelihood confidence intervals. The empirical likelihood approach proposed provides a better coverage in this case. The large positive bias in the standard error estimators (see Table 4.3) might be the reason for over coverage for the parameters β_{1N} and β_{2N} with the maximum likelihood approach. Under coverage with the pseudo likelihood approach may be due to the negative bias in the standard error estimators (see Table 4.3).

4.10 Conclusion

We proposed using an empirical likelihood approach to make design based inference for regression parameters when modelling hierarchical data. We considered two stage sampling (see Section 4.2). We assumed a working model with a uniform covariance structure for the population data (see Section 4.3). We used a general estimating equation to define the parameter of interest (see Section 4.4). We assumed that the population model and the sampling design had the same hierarchical structure. The profile empirical log-likelihood ratio function was defined at the first stage of selection by using an ultimate cluster approach (see Section 4.8). As demonstrated in Section 4.8, the inference can be made for the subvector of regression parameters by profiling out the empirical log-likelihood ratio function over the nuisance parameters, the parameters that are not of primary interest. Stratification and population level information can be easily incorporated (see Section 4.5).

The empirical likelihood approach proposed does not depend on variance estimation, resampling, linearisation and second order inclusion probabilities. Neither does it rely on the normality of the point estimator. It may provide better inference, even when the point estimator is not normal, the data is skewed or includes outlying values, than the standard approaches that are based on the normality assumption and variance estimates (see Sections 4.9 and 4.9.1). The empirical evidence shows that the performance of the empirical likelihood confidence intervals is not affected by the number of SSUs (see Section 4.9.1).

Standard confidence intervals for regression coefficients may have poor coverages when variance estimators are biased or not stable. This may be the case with linearised variance estimator when sample sizes are not large enough or data includes outlying values (see Section 4.9.1). Standard parametric confidence intervals may produce invalid inference when the sampling design is informative (see Section 4.9).

The approach proposed requires knowledge of the first order inclusion probabilities. When samples are selected with equal probabilities at the second stage, totals within clusters can be estimated without using the weights of the SSUs as the scaled weights reduce to one in this case (see Section 4.7.1).

In this work, we considered linear multilevel regression parameters. It is straightforward to extend the empirical likelihood approach proposed to generalised multilevel regression parameters by using *generalised estimating equations* (e.g. Grilli and Pratesi, 2004). We used cross sectional hierarchical data. In social sciences, data with repeated measures is also commonly used. In this case, it is expected that the measures belonging to the same unit at different time points are correlated. Individual units followed over time form the first level of hierarchy. These units may be selected with respect to a multi stage sampling design. This may form another level of hierarchy. It would be interesting to implement the empirical likelihood approach proposed to that sort of data. We assumed a uniform covariance structure. Longitudinal data may have more complex correlation structures. We assumed the same hierarchy for the model and the sampling design. This may not be the case in practice. It would be good to extend the approach proposed to these situations.

Chapter 5

General Conclusion

In this thesis, three papers were respectively presented in Chapters 2–4. A general introduction that includes the main contribution of the papers (see Section 1.1), the author's contributions to the papers (see Section 1.2) and detailed literature review for each paper (see Sections 1.3–1.5) were given in Chapter 1.

We considered design-based inference for finite population quantities throughout this thesis. In this context, we assumed a sample s is selected from a finite population U with respect to a probability sampling design. The sampling distribution is solely driven by the sampling design as we assume all units in the sample are respondent. The finite population values are assumed fixed, non-random, quantities. The target finite population quantities are defined as (non)linear functions of these fixed values. In the thesis, we considered poverty rates, linear regression, generalised linear regression and general linear regression coefficients. We proposed methods that provide valid design-based inferences for those complex statistics.

In the first paper (see Chapter 2), a multivariate regression approach proposed by Berger and Priam (2016) was introduced to estimate the variance of change in poverty rates. This approach provides a valid way to estimate the correlation between two samples at different waves. The whole sample data is considered unlike the standard approaches that only rely on the common sample.

Two methods were considered to estimate the variance of change. In the first approach, called the 'ratio approach', the poverty threshold was assumed fixed. The poverty rate was treated as a simple ratio (see Section 2.3). Thus the sampling variability that arises from the estimation of the poverty threshold was ignored. As demonstrated in the numerical results (see Section 2.6), this may yield biased variance estimators. In the second approach, namely the 'linearisation approach', the randomness due to the estimated poverty threshold was accommodated (see Sections 2.4–2.5). This approach relies on a bandwidth parameter that needs to be estimated. We used several fixed bandwidth

parameters. In the numerical results, we observe that the variance estimators of change in poverty rates are not significantly affected by the chosen bandwidth parameter (see Sections 2.6–2.7). This may be because of the fact that the bandwidth parameter plays a role in both numerator and denominator in the expression of the linearised variable for the poverty rate (2.4) (see Section 2.4). This may remove the effect of the bandwidth parameter on variance estimates. However, this may not be the case for complex statistics other than the poverty rate.

Treating the poverty threshold as fixed may yield serious bias in cross sectional variance estimators. However, the bias considerably decreases for the variance estimator of change although it may not be negligible for highly skewed income distributions (see Section 2.6). Thus the ratio method should be used with caution. Our research also reveals that the variance estimates of change may not be conservative with the ratio approach, contrary to cross sectional variance estimates.

The multivariate regression can be used to estimate the variances of the other poverty and income inequality measures such as relative median at risk of poverty gap (RMPG), quantile share ratio (QSR) and GINI coefficient by applying the linearisation approach (e.g. Verma and Betti, 2005; Berger, 2008; Verma and Betti, 2011). The approach proposed can be easily implemented in any statistical software.

In the second paper, we proposed a profile empirical likelihood approach to make inference for the subvector of parameters. We recalled the empirical likelihood approach proposed by Berger and De La Riva Torres (2016) (see Section 3.3). We extended Berger and De La Riva Torres's (2016) approach to the multidimensional parameter case. The inference for the subvector of parameters requires profiling that relies on maximisation of the empirical log-likelihood function over the nuisance parameters, which are not of primary interest. We showed that the profile empirical log-likelihood ratio function (see expressions (3.15) and (3.51)) asymptotically follows a χ^2-distribution under a set of regularity conditions (see Section 3.7). The asymptotic derivations are provided in the supplementary material (see Appendix B.2). This property allows us to test hypotheses and construct confidence intervals. The empirical likelihood confidence interval does not depend on the normality of the point estimator. It may provide better coverages and tail errors than the standard approaches that rely on symmetric confidence intervals or resampling methods (see Sections 3.10-3.10.5). Standard confidence intervals based on variance estimates may give poor coverages especially when the sampling distribution is skewed or the data includes outlying values (see Section 3.10.3).

In the third paper, we implemented the profile empirical likelihood approach proposed in the second paper to regression parameters under correlated error structure. As demonstrated in the numerical work (see Sections 4.9 and 4.9.1), empirical likelihood confidence intervals may provide better coverages than the standard approaches even when the point estimator is not normal, the data is skewed or includes outlying values. We

observed that the standard parametric confidence intervals may provide considerably lower coverages when the design is informative.

The empirical likelihood confidence interval demonstrated in the second and the third papers does not rely on normality, resampling, linearisation, variance estimation or design effect. Variance estimator may not be reliable with heavily skewed data including outlying values even it is asymptotically unbiased. The normality may not hold either with such data. Our simulation studies show that the empirical likelihood approach proposed outperforms the alternative approaches based on variance estimates (see Sections 3.10.3 and 4.9.1). Bootstrap can be an alternative approach and may be preferred from practical point of view. Our simulation studies show that bootstrap confidence intervals may be more unstable (see Sections 3.10-3.10.5). Besides, the empirical likelihood approach proposed is easier to implement and less computer intensive than bootstrap.

Population level information can be incorporated with the empirical likelihood approach proposed. The approach proposed provides survey weights (3.45), which are intuitively calibrated as a result of the maximisation of the empirical log-likelihood function (3.43) with a fixed known population parameter φ_N. They are always positive and asymptotically design optimal. Calibration weights (Deville and Särndal, 1992) can be negative and not necessarily asymptotically optimal. Standard calibration weights are derived for point estimation. The empirical log-likelihood ratio function (3.51) is used for point estimation, testing and confidence intervals. Population level information does not have to be in the form of totals or means. It belongs to a wider class of parameters that can be defined as the unique solutions to the population estimating equations (3.42).

The approaches proposed in this thesis satisfy the properties usually required by survey practitioners. Design consistency is one of them. In the first paper, we proposed using a design consistent estimator for correlation when estimating variance of change under rotating panel surveys. In the second and the third papers, we show that the empirical likelihood point estimator is design consistent. For survey practitioners, it is also important to consider the characteristics of the complex survey data to make a valid inference. With the approaches proposed in the thesis, unequal probabilities, stratification and clustering are all taken into account. A panel rotation plan is also considered in the first paper when estimating variance of change in the poverty rates. We show that the empirical log-likelihood statistic provides a valid design-based inference for finite population parameters. Our simulation studies show that the standard confidence interval may provide poor coverage when the design is ignored under informative sampling (see Table 3.1 and 4.2).

Second order inclusion probabilities are required for without replacement sampling to obtain unbiased variance estimator. However, these probabilities are usually unknown to survey practitioners or hard to compute. Second order inclusion probabilities are not required for variance estimation or confidence intervals with the approaches proposed in

this thesis. The asymptotic variance estimators in the pivotal statistics (3.40) and (3.55) are design consistent when the sampling fraction is negligible. In this case, sampling with replacement and large entropy sampling without replacement are equivalent (e.g. Hájek, 1981; Berger, 2011). In all three papers, we assume that the sampling design has a large entropy. Most sampling designs used in practice can be approximated by a large-entropy sampling design, except non-randomised systematic sampling (e.g. Berger, 2011).

The approaches proposed in the thesis provide valid inferences under sampling designs with negligible sampling fractions. In the case of a multi-stage sampling design, we assume negligible sampling fraction at the first stage of selection. This assumption usually hold with household surveys. However, this may not be valid, for example, if the number of strata is large, which is usually the case with business surveys. The multivariate regression approach in the first paper cannot be used in this case. Other approaches should be used instead (e.g. Nordberg, 2000; Berger, 2004; Wood, 2008). It would be very useful to generalise the profile empirical likelihood approach to designs with large sampling fractions.

We ignored the effect of nonresponse and calibration in the first paper. When the auxiliary information is correlated with the variable of interest, a huge gain can be obtained in the precision of the estimator. Ignoring the fact that the estimated totals are calibrated against known population totals may result in the overestimation of the variance. On the contrary, variances may be underestimated when the effect of imputation is ignored. It is important to investigate how the variance estimation of change might be affected under these circumstances. It would be also very useful to extend the multivariate regression approach when calibration is performed at both waves. Berger and Escobar (2015) showed that the approach can be implemented under nonresponse when hot-deck imputation is used for unit nonresponse.

The empirical likelihood approach in the third paper can be straightforwardly extended to generalised multilevel regression parameters by using the generalised estimating equations (e.g. Grilli and Pratesi, 2004). We used a cross sectional hierarchical data by assuming that the model and the design hierarchies are the same. It would be interesting to extend the empirical likelihood approach to the circumstance when the hierarchies differ. It would also be good to implement it on data with repeated measures. This sort of data may bring about dealing with correlation structures other than the uniform correlation structure. Hence, it would be good to extend the approach proposed to other correlation structures.

The empirical likelihood approach proposed in the thesis requires the knowledge of the first order inclusion probabilities. However, these probabilities may not be available to survey practitioners. It would be interesting to develop the theory in such a way that

the empirical log-likelihood statistic (3.52) is valid when the survey weights are available but not the inclusion probabilities.

The linearisation approach is proposed in the first paper to take into account the fact that the poverty threshold is estimated, thereby allowing approximately unbiased variance estimator. However, the empirical likelihood approach can be a better alternative to measure the design uncertainty of the poverty rate as it does not rely on any analytic derivation. Even linearised variables are analytically derived, the variance estimator may not be reliable if the data contains outlying values. With the empirical likelihood approach proposed, the poverty rate will be the parameter of interest and the median of the income distribution will be the nuisance parameter. However, we assumed that the empirical log-likelihood function is differentiable with respect to the nuisance parameters (see Sections 3.6 and 3.7). Thus the approach cannot be directly applied to poverty rates as the median income is not a smooth statistic. It would be very useful to develop the theory in the second paper to consider the parameters defined through non-smooth estimating equations. This would allow us to make inference for a subset of a set of more general class of parameters including quantiles. The poverty rate defined in the first paper is a very good example to this situation.

Empirical likelihood is a cutting-edge approach in survey sampling. Hence, the theory could be taken forward in to many directions. Some of them have been mentioned above. Small domain estimation, estimation in the presence of nonresponse, estimation and inference from combined data sources are some of the other areas to which the empirical likelihood approach proposed can be extended. The working model (4.3) assumed in the third paper responds to many practical situations. For example, it is mostly used in small domain estimation. It would be very useful to extend the empirical likelihood approach proposed to make inference for small domain means. A synthetic type of regression estimator (e.g. Rao, 2003, p.46) based on the regression parameter (4.3) may be used by assuming that the domain specific characteristics are the same as those of the whole population or a larger group of domains (e.g. Gonzales, 1973). In this case, the domain means will the parameter of interest while the regression parameters will be the nuisance parameters, thereby taking into account the fact that the regression parameters are estimated.

Appendix A

Supplementary Material for the First Paper

By Melike Oguz-Alper
University of Southampton, SO17 1BJ, Southampton, U.K.
M.OguzAlper@soton.ac.uk

A.1 Derivation of the influence function of the poverty rate over a domain

Let M be a measure that assigns a unit mass to each unit i in the population U. For example, the population size N can be written as $N = \int d\mathrm{M} = \sum_{i\in U} 1$ and the total of a variable y can be expressed as $Y = \int y d\mathrm{M} = \sum_{i\in U} y_i$ (Deville, 1999). Let $\mathrm{F}(\mathrm{M}, x)$ be the income distribution function at x over the population U, that is,

$$\mathrm{F}(\mathrm{M}, x) = \frac{1}{N}\sum_{i\in U} \delta\{y_i \leq x\}.$$

Then, the income distribution function at the median of the income distribution is given by $\mathrm{F}(\mathrm{M}, Med(\mathrm{M})) = 0.5$. Thus the influence function of the functional $\mathrm{F}(\mathrm{M}, Med(\mathrm{M}))$ is equal to 0 for all i, that is, $\mathrm{IF}_i(\mathrm{M}, Med(\mathrm{M})) = 0$. By using the "Rule 7" in Deville (1999, p.198), the influence function of F at i (see also Osier, 2009, p.181-183) can be derived as follows:

$$\mathrm{IF}_i(\mathrm{M}, Med(\mathrm{M})) = \mathrm{IF}_i(\mathrm{M}, Med(\mathrm{M}) \mid Med(\mathrm{M}) \text{ fixed}) + \frac{\partial \mathrm{F}(\mathrm{M}, x)}{\partial x}|_{x=Med(\mathrm{M})}\mathrm{IMed}_i(\mathrm{M}) = 0, \tag{A.1}$$

where $\mathrm{IMed}_i(\mathrm{M})$ is the influence function of the median income defined by (A.2).

The influence function of $\mathrm{F}(\mathrm{M}, Med(\mathrm{M})$ at i, when the median is fixed, is given by

$$\mathrm{IF}_i(\mathrm{M}, Med(\mathrm{M}) \mid Med(\mathrm{M}) \text{ fixed}) = \frac{1}{N}[\delta\{y_i \leq Med\} - 0.5].$$

Thus the influence function of the functional $Med(\mathrm{M})$ is obtained as

$$\mathrm{IMed}_i(\mathrm{M}) = -\frac{1}{N}\frac{1}{f(Med)}[\delta\{y_i \leq Med\} - 0.5], \tag{A.2}$$

where

$$f(Med) = \frac{\partial \mathrm{F}(\mathrm{M}, x)}{\partial x}|_{x=Med(\mathrm{M})}$$

is the probability density function at the median of the income distribution.

Now define the income distribution function at x over a domain D as follows:

$$\mathrm{F}_D(\mathrm{M}, x) = \frac{1}{N_D}\sum_{i \in U} d_i \delta\{y_i \leq x\}.$$

Hence, the income distribution function over a domain D at the poverty threshold T is defined by

$$\mathrm{F}_D(\mathrm{M}, T(\mathrm{M})) = \frac{1}{N_D}\sum_{i \in U} d_i \delta\{y_i \leq T(\mathrm{M})\},$$

where $T(\mathrm{M}) = 0.6Med(\mathrm{M})$ and d_i is the domain indicator, that is, 1 when $i \in D$, and 0 otherwise. $\mathrm{F}_D(\mathrm{M}, T(\mathrm{M}))$ is equivalent to the povery rate over a domain D (i.e. R_D). Thus we can obtain the influence function of the poverty rate analogously to (A.1), that is,

$$\mathrm{IF}_{D;i}(\mathrm{M}, T(\mathrm{M})) = \mathrm{IF}_{D;i}(\mathrm{M}, T(\mathrm{M})) \mid T(\mathrm{M}) \text{ fixed}) + \frac{\partial \mathrm{F}_D(\mathrm{M}, x)}{\partial x}|_{x=T(\mathrm{M})}\mathrm{IT}_i(\mathrm{M}) = \mathrm{IR}_{D;i}.$$

The influence function of F_D, when the threshold is fixed, is given by

$$\mathrm{IF}_{D;i}(\mathrm{M}, T(\mathrm{M})) \mid T(\mathrm{M}) \text{ fixed}) = \frac{d_i}{N_D}[\delta\{y_i \leq T\} - R_D].$$

Hence, the influence function of the poverty rate is obtained as follows:

$$\mathrm{IR}_{D;i} = \frac{d_i}{N_D}[\delta\{y_i \leq T\} - R_D] + f_D(T)\mathrm{IT}_i(\mathrm{M}), \tag{A.3}$$

where

$$f_D(T) = \frac{\partial \mathrm{F}_D(\mathrm{M}, x)}{\partial x}|_{x=T(\mathrm{M})}$$

is the probability density function at the poverty threshold. The influence function of the functional $T(\mathrm{M})$ at i is given by

$$\mathrm{IT}_i(\mathrm{M}) = 0.6\,\mathrm{IMed}_i(\mathrm{M}). \tag{A.4}$$

If we substitute $\mathrm{IMed}_i(\mathrm{M})$ in (A.2) into (A.4), we obtain the following:

$$\mathrm{IT}_i(\mathrm{M}) = -\frac{0.6}{N}\frac{1}{f(Med)}[\delta\{y_i \leq Med\} - 0.5].$$

Therefore, the influence function of the poverty rate at i over a domain D given in (A.3) can be rewritten as follows:

$$\mathrm{IR}_{D;i} = \frac{d_i}{N_D}[\delta\{y_i \leq T\} - R_D] - \frac{0.6}{N}\frac{f_D(T)}{f(Med)}[\delta\{y_i \leq Med\} - 0.5]. \tag{A.5}$$

Note that we assume the derivatives of F and F_D exist and are strictly non-negative for all x.

A.2 Generation of random variables

For the gamma random variables, we used the algorithm proposed by Schmeiser and Lal (1982, p.358). First, three independent random variables were generated by a gamma distribution as follows:

$$\begin{aligned}
Y_1 &\sim gamma(\alpha_1 - \rho\sqrt{\alpha_1}\sqrt{\alpha_2}, 1),\\
Y_2 &\sim gamma(\alpha_2 - \rho\sqrt{\alpha_1}\sqrt{\alpha_2}, 1),\\
Y_3 &\sim gamma(\rho\sqrt{\alpha_1}\sqrt{\alpha_2}, 1),
\end{aligned}$$

with $\alpha_1 = 2.5$, $\alpha_2 = 2.6$, and $\rho = 0.94$. Then, the income variables were obtained by the following expressions: $y_{1;i} = Y_1 + Y_3$ and $y_{2;i} = Y_2 + Y_3$, so that $y_{1;i} \sim gamma(2.5, 1)$, $y_{2;i} \sim gamma(2.6, 1)$, and $\rho(y_{1;i}, y_{2;i}) \approx 0.94$.

The Cholesky decomposition was used to generate the correlated lognormal variables. Hence, the log income variables with the correlation of $\rho = 0.95$, a mean of $\mu = 1.119$ and a standard deviation of $\sigma = 0.602$ were generated by

$$\begin{aligned}
\log(y_{1;i}) &= \mu + \sigma X_1,\\
\log(y_{2;i}) &= \mu + \rho\sigma X_1 + \sqrt{1-\rho^2}\sigma X_2,
\end{aligned}$$

where X_1 and X_2 are independent standard normal variables. The correlation coefficient between the income variables is approximately 0.94.

For correlated Weibull variables, we followed the algorithm proposed by Feiveson (2002, p.117). First, two correlated standard normal variables Y_1 and Y_2 with a correlation of $\rho = 0.95$ were generated by using the Cholesky decomposition: $Y_1 = X_1$ and $Y_2 = \rho X_1 + \sqrt{1-\rho^2}X_2$, where X_1 and X_2 are independent standard normal variables. Secondly, correlated uniform variables were obtained by the standard normal cumulative distribution function transformation; such that $U_1 = \Phi(Y_1)$ and $U_2 = \Phi(Y_2)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Finally, uniform random variables were transformed by the inverse of the Weibull cumulative distribution function to achieve the correlated income variables as follows: $y_{1;i} = \mathrm{F}_U^{-1}(U_1) = (-\ln(1-U_1))^{5/4}$ and $y_{2;i} = \mathrm{F}_U^{-1}(U_2) = (-\ln(1-U_2))^{5/4}$, so that $y_{1;i}, y_{2;i} \sim Weibull(0.8, 1)$ and $\rho(y_{1;i}, y_{2;i}) \approx 0.94$.

A.3 R code for the first paper

By Melike Oguz-Alper and Yves G. Berger
University of Southampton, SO17 1BJ, Southampton, U.K.
M.OguzAlper@soton.ac.uk Y.G.Berger@soton.ac.uk

```
rm(list=ls())
memory.size(max=TRUE)
#---------------------------------------------------------------------
# Load the 'sampling_YGB' package from the library
#---------------------------------------------------------------------
Path <- paste(Sys.getenv("R_HOME"),"\\library\\sampling_YGB\\",sep="")
source(paste(Path,"initsampling",sep=""))
load.sampling(Path)
#=====================================================================
# Load packages required
#=====================================================================
library(lpSolve)
library(sampling)
library(boot)
library(MASS)
library(laeken)
library(mixtools)
#=====================================================================
# Generate random variables
#=====================================================================
N<-20940
#----------------------
# Gamma
#----------------------
rho.Income<-0.94
set.seed(15)
Y1<-rgamma(N,2.5-rho.Income*sqrt(2.5)*sqrt(2.6),1)
set.seed(6)
Y2<-rgamma(N,2.6-rho.Income*sqrt(2.5)*sqrt(2.6),1)
set.seed(15)
Y3<-rgamma(N,rho.Income*sqrt(2.5)*sqrt(2.6),1)
Income.Wave1<-Y1+Y3
```

```
Income.Wave2<-Y2+Y3
#---------------------
# Lognormal
#---------------------
rho.Income<-0.95
set.seed(6)
Z1<-rnorm(N,0,1)
set.seed(1)
Z2<-rnorm(N,0,1)
Y1<-0.602*Z1+1.119
Y2<-rho.Income*0.602*Z1+sqrt(1-ro^2)*0.602*Z2+1.119
Income.Wave1<-exp(Y1)
Income.Wave2<-exp(Y2)
#---------------------
# Weibull
#---------------------
rho.Income<-0.95
set.seed(6)
X1<-rnorm(N,0,1)
set.seed(3)
X2<-rnorm(N,0,1)
cor(X1,X2)
Y1<-X1
Y2<-rho.Income*X1+sqrt(1-rho.Income^2)*X2
U1<-pnorm(Y1)
U2<-pnorm(Y2)
Income.Wave1<-(-log(1-U1))^(1/0.8)
Income.Wave2<-(-log(1-U2))^(1/0.8)
#=========================================================================
# Generate measure of size variable
#=========================================================================
rho<-0.7
var.ei<-(1-rho^2)*var(Income.Wave1)
ei<-rnorm(N,0,sqrt(var.ei))
xi<-5+ro*Income.Wave1+ei
#=========================================================================
# Inclusion probabilities
#=========================================================================
n<-1047
Inc.Prob<-PI.PROPORTIONAL(xi,n)
Weight<-1/Inc.Prob
#=========================================================================
# START SIMULATION
#=========================================================================
#-------------------------------------------------------------------
# Output matrices (Only includes ratio and linearisation.SD cases)
#-------------------------------------------------------------------
NB.Simulations<-10000

NB.Col.Ratio <- 7
NB.Row.Ratio <- NB.Simulations
Result.Ratio <- data.frame(matrix(rep(0,times=NB.Col.Ratio*NB.Row.Ratio),
ncol=NB.Col.Ratio,nrow=NB.Row.Ratio))
colnames(Result.Ratio) <- c("%Pov 1","Var 1","%Pov 2","Var 2","Change",
"Var Change","Corr")
Label.Row.Ratio <- 0

NB.Col.Lin.SD <- 7
NB.Row.Lin.SD <- NB.Simulations
```

```
Result.Lin.SD <- data.frame(matrix(rep(0,times=NB.Col.Lin.SD*NB.Row.Lin.SD),
ncol=NB.Col.Lin.SD,nrow=NB.Row.Lin.SD))
colnames(Result.Lin.SD) <- c("%Pov 1","Var 1","%Pov 2","Var 2","Change",
"Var Change","Corr")
Label.Row.Lin.SD <- 0
#---------------------
# Start loop
#---------------------
for (Simulation in 1:NB.Simulations)
{
#------------------------
# Simple random sampling
#------------------------
# (Adjust the code for Chao (1982) sampling by using the function
# CHAO.SAMPLING in package 'sampling_YGB'
#-----------------------------------------------------------------
#--------------------------------
# Sample selection at first wave
#--------------------------------
Sample.Wave1<-srswor(n,N)
Labels.Wave1<-as.numeric(labels(Sample.Wave1))[Sample.Wave1==1]
#--------------------------------
# Sample selection at second wave
#--------------------------------
Sample.Wave2<-rep(0,N)
Fraction<-0.75
Sample.Common<-srswor(round(n*Fraction),n)
Labels.Common<-Labels.Wave1[Sample.Common==1]

Sample.Not.Common<-srswor(n-round(n*Fraction),N-n)
Labels.Not.Common<-as.numeric(labels(Sample.Wave1[-Labels.Wave1]))
[Sample.Not.Common==1]
Sample.Wave2[c(Labels.Common,Labels.Not.Common)]<-1
#----------------------------
# Weights of the sample units
#----------------------------
Weight.Wave1<-Weight*Sample.Wave1
Weight.Wave2<-Weight*Sample.Wave2
#-------------
# Sample data
#-------------
Select<-Sample.Wave1==1 | Sample.Wave2==1
Data.Sample<-data.frame(cbind(Income.Wave1[Select],Income.Wave2[Select],
Weight.Wave1[Select],Weight.Wave2[Select],Sample.Wave1[Select],
Sample.Wave2[Select]))
names(Data.Sample)<-c("Income.Wave1","Income.Wave2","Weight.Wave1","Weight.Wave2",
"Sample.Wave1","Sample.Wave2")
Data.Sample$Income.Wave1[Data.Sample$Sample.Wave1==0]<-0
Data.Sample$Income.Wave2[Data.Sample$Sample.Wave2==0]<-0
#---------------------------------
# Weighted median and poverty rate
#---------------------------------
Median.Wave1<-weightedMedian(Data.Sample$Income.Wave1,weights=Data.Sample$
Weight.Wave1)
Median.Wave2<-weightedMedian(Data.Sample$Income.Wave2,weights=Data.Sample$
Weight.Wave2)
Poverty.Wave1<-as.numeric(Data.Sample$Income.Wave1<=0.6*Median.Wave1)
Poverty.Wave2<-as.numeric(Data.Sample$Income.Wave2<=0.6*Median.Wave2)
Data.Sample$Poor.Wave1<-Data.Sample$Weight.Wave1*Poverty.Wave1
```

```
Data.Sample$Poor.Wave2<-Data.Sample$Weight.Wave2*Poverty.Wave2
#---------------------------------
# Point estimate of poverty rate (%)
#---------------------------------
Poor1<-sum(Data.Sample$Poor.Wave1)
Total1<-sum(Data.Sample$Weight.Wave1)
Point.Estimate1<-100*Poor1/Total1

Poor2<-sum(Data.Sample$Poor.Wave2)
Total2<-sum(Data.Sample$Weight.Wave2)
Point.Estimate2<-100*Poor2/Total2
#-----------------------------------------
# Point Estimate of change in poverty rates
#-----------------------------------------
Point.Estimate.Change <- Point.Estimate2 - Point.Estimate1
#-----------------------------------------------------
# Derivation of linearised variables for poverty rate
#-----------------------------------------------------
# Indicator at median income
#===========================
Iinc.MED1<-as.numeric(Data.Sample$Income.Wave1<=Median.Wave1)
Iinc.MED2<-as.numeric(Data.Sample$Income.Wave2<=Median.Wave2)
#-------------------------------------------------------------------
# Weighted standard deviation
# (Adjust for IQR and A by using:
# IQR<-wIQR(W,Y)) and A<-min(SD,IQR/1.34)
#============================
Y1<-Data.Sample[Data.Sample$Sample.Wave1==1,]$Income.Wave1
W1<-Data.Sample[Data.Sample$Sample.Wave1==1,]$Weight.Wave1
var1<-(sum(W1*Y1*Y1)-sum(W1*Y1)*sum(W1*Y1)/sum(W1))/sum(W1)
SD1<-sqrt(var1)

Y2<-Data.Sample[Data.Sample$Sample.Wave2==1,]$Income.Wave2
W2<-Data.Sample[Data.Sample$Sample.Wave2==1,]$Weight.Wave2
var2<-(sum(W2*Y2*Y2)-sum(W2*Y2)*sum(W2*Y2)/sum(W2))/sum(W2)
SD2<-sqrt(var2)
#-------------------------------------------------------------------
# Smoothing parameter (bandwidth) for kernel density function
# Adjust for IQR and A by using:
# h.IQR<-0.79*IQR/exp(0.2*log(sum(W)))
# h.A<-0.9*A/exp(0.2*log(sum(W)))
#===========================================================
h1.SD<-1.06*SD1/exp(0.2*log(sum(W1)))
h2.SD<-1.06*SD2/exp(0.2*log(sum(W2)))
#-------------------------------------------------------------------
# Estimation of F'(quantile=MED(M)) for quantile=0.5
#===================================================
F1.MED.SD<-sum(exp(-(Median.Wave1-Y1)^2/h1.SD/h1.SD/2)*W1)/sum(W1)/h1.SD/sqrt(2*pi)
F2.MED.SD<-sum(exp(-(Median.Wave2-Y2)^2/h2.SD/h2.SD/2)*W2)/sum(W2)/h2.SD/sqrt(2*pi)
#-------------------------------------------------------------------
# Linearised variable for at-risk-of-poverty threshold
#=====================================================
IARPT1.SD<--0.6*(Iinc.MED1-0.5)/F1.MED.SD/sum(W1)
IARPT2.SD<--0.6*(Iinc.MED2-0.5)/F2.MED.SD/sum(W2)
#-------------------------------------------------------------------
# Computation of F'(beta*quantile=ARPT) for beta=0.6 and quantile=0.5
#====================================================================
F1.ARPT.SD<-sum(exp(-(0.6*Median.Wave1-Y1)^2/h1.SD/h1.SD/2)*W1)/sum(W1)/h1.SD/
sqrt(2*pi)
```

```
F2.ARPT.SD<-sum(exp(-(0.6*Median.Wave2-Y2)^2/h2.SD/h2.SD/2)*W2)/sum(W2)/h2.SD/
sqrt(2*pi)
#-----------------------------------------------------------------
# Poverty rate
#==============
ARPR1<-Point.Estimate1/100
ARPR2<-Point.Estimate2/100
#-----------------------------------------------------------------
# Linearised variable for at-risk-of-poverty rate
#================================================
IARPR1.SD<-100*((Poverty.Wave1-ARPR1)/sum(W1)+F1.ARPT.SD*IARPT1.SD)
IARPR2.SD<-100*((Poverty.Wave2-ARPR2)/sum(W2)+F2.ARPT.SD*IARPT2.SD)
#-----------------------------------------------------------------
# Merge linearised variables with the sample data
#================================================
Data.Sample$Pseudo.Wave1.SD<-Data.Sample$Weight.Wave1*IARPR1.SD
Data.Sample$Pseudo.Wave2.SD<-Data.Sample$Weight.Wave2*IARPR2.SD
#------------
# RATIO CASE
#------------
# Variance estimation at wave1
#=============================
Data.Sample1 <- Data.Sample[Data.Sample$Sample.Wave1==1,]
Y.Ratio <- cbind(Data.Sample1$Poor.Wave1,Data.Sample1$Weight.Wave1)
Fit.Ratio <- lm(formula=Y.Ratio~-1+Sample.Wave1, data=Data.Sample1)
Var.Cov.Matrix1.Ratio <- (n - 1) * as.matrix(estVar(Fit.Ratio))
Var1.Ratio <- diag(Var.Cov.Matrix1.Ratio)
Gradient.Ratio <- matrix(c( (1/Total1) , (-Poor1*Total1^(-2))),ncol=1)
Variance1.Ratio <- 100^2 * t(Gradient.Ratio)%*%Var.Cov.Matrix1.Ratio%*%
Gradient.Ratio
#-----------------------------------------------------------------
# Variance estimation at wave2
#=============================
Data.Sample2 <- Data.Sample[Data.Sample$Sample.Wave2==1,]
Y.Ratio <- cbind(Data.Sample2$Poor.Wave2,Data.Sample2$Weight.Wave2)
Fit.Ratio <- lm(formula=Y.Ratio~-1+Sample.Wave2, data=Data.Sample2)
Var.Cov.Matrix2.Ratio <- (n - 1) * as.matrix(estVar(Fit.Ratio))
Var2.Ratio <- diag(Var.Cov.Matrix2.Ratio)
Gradient.Ratio <- matrix(c( (1/Total2) , (-Poor2*Total2^(-2))),ncol=1)
Variance2.Ratio <- 100^2 * t(Gradient.Ratio)%*%Var.Cov.Matrix2.Ratio%*%
Gradient.Ratio
#-----------------------------------------------------------------
#  Variance estimation for change
#=======================
Y.Ratio <- cbind(Data.Sample$Poor.Wave1,Data.Sample$Weight.Wave1,Data.Sample$
Poor.Wave2,Data.Sample$Weight.Wave2)
Fit.Ratio <- lm(formula=Y.Ratio~-1+Sample.Wave1*Sample.Wave2, data=Data.Sample)
Var.Cov.Matrix.Reg.Ratio <- Fit.Ratio$df * as.matrix(estVar(Fit.Ratio))
Vect.D.Ratio <- sqrt(c(Var1.Ratio,Var2.Ratio) / diag(Var.Cov.Matrix.Reg.Ratio))
Matrix.D.Ratio <- diag(Vect.D.Ratio)
Var.Cov.Matrix.Ratio<-t(Matrix.D.Ratio)%*%Var.Cov.Matrix.Reg.Ratio%*%Matrix.D.Ratio
Gradient.Ratio<-matrix(c((-1/Total1),(Poor1*Total1^(-2)),(1/Total2),(-Poor2*
Total2^(-2))),ncol=1)
Variance.Change.Ratio<-100^2*t(Gradient.Ratio)%*%Var.Cov.Matrix.Ratio%*%
Gradient.Ratio
#-----------------------------------------------------------------
# Correlation
#============
Correlation.Ratio <- ( Variance1.Ratio + Variance2.Ratio - Variance.Change.Ratio )/
```

```
( 2 * sqrt(Variance1.Ratio * Variance2.Ratio) )
#--------------------
# LINEARISATION CASE
#--------------------
# Variance estimation at wave1
#=============================
Y.Lin.SD <- Data.Sample1$Pseudo.Wave1.SD
Fit.Lin.SD <- lm(formula=Y.Lin.SD~-1+Sample.Wave1, data=Data.Sample1)
Variance1.Lin.SD <- anova(Fit.Lin.SD)[2,2]
#-----------------------------------------------------------------
# Variance estimation at wave1
#=============================
Y.Lin.SD <- Data.Sample2$Pseudo.Wave2.SD
Fit.Lin.SD <- lm(formula=Y.Lin.SD~-1+Sample.Wave2, data=Data.Sample2)
Variance2.Lin.SD <- anova(Fit.Lin.SD)[2,2]
#-----------------------------------------------------------------
# Variance estimation for change
#======================
Y.Lin.SD <- cbind(Data.Sample$Pseudo.Wave1.SD,Data.Sample$Pseudo.Wave2.SD)
Fit.Lin.SD <- lm(formula=Y.Lin.SD~-1+Sample.Wave1*Sample.Wave2, data=Data.Sample)
Var.Cov.Matrix.Reg.Lin.SD <- Fit.Lin.SD$df * as.matrix(estVar(Fit.Lin.SD))
Vect.D.Lin.SD<-sqrt(c(Variance1.Lin.SD,Variance2.Lin.SD)/
diag(Var.Cov.Matrix.Reg.Lin.SD))
Matrix.D.Lin.SD<-diag(Vect.D.Lin.SD)
Var.Cov.Matrix.Lin.SD<-t(Matrix.D.Lin.SD)%*%Var.Cov.Matrix.Reg.Lin.SD%*%
Matrix.D.Lin.SD
Gradient.Lin<-matrix(c(-1,1),ncol=1)
Variance.Change.Lin.SD <-  t(Gradient.Lin) %*% Var.Cov.Matrix.Lin.SD %*%
Gradient.Lin
#-----------------------------------------------------------------
# Correlation
#============
Correlation.Lin.SD<-( Variance1.Lin.SD+Variance2.Lin.SD-Variance.Change.Lin.SD )/
( 2 * sqrt(Variance1.Lin.SD * Variance2.Lin.SD) )
#-----------------------------------------------------------------
# Save output for the ratio case
#===============================
Label.Row.Ratio <- Label.Row.Ratio + 1
j <- 1
Result.Ratio[Label.Row.Ratio,j] <- Point.Estimate1
j <- j + 1
Result.Ratio[Label.Row.Ratio,j] <- Variance1.Ratio
j <- j + 1
Result.Ratio[Label.Row.Ratio,j] <- Point.Estimate2
j <- j + 1
Result.Ratio[Label.Row.Ratio,j] <- Variance2.Ratio
j <- j + 1
Result.Ratio[Label.Row.Ratio,j] <- Point.Estimate.Change
j <- j + 1
Result.Ratio[Label.Row.Ratio,j] <- Variance.Change.Ratio
j <- j + 1
Result.Ratio[Label.Row.Ratio,j] <- Correlation.Ratio

# Save output for the linearisation.SD case
#==========================================
Label.Row.Lin.SD <- Label.Row.Lin.SD+ 1
j <- 1
Result.Lin.SD[Label.Row.Lin.SD,j] <- Point.Estimate1
j <- j + 1
```

```
Result.Lin.SD[Label.Row.Lin.SD,j] <- Variance1.Lin.SD
j <- j + 1
Result.Lin.SD[Label.Row.Lin.SD,j] <- Point.Estimate2
j <- j + 1
Result.Lin.SD[Label.Row.Lin.SD,j] <- Variance2.Lin.SD
j <- j + 1
Result.Lin.SD[Label.Row.Lin.SD,j] <- Point.Estimate.Change
j <- j + 1
Result.Lin.SD[Label.Row.Lin.SD,j] <- Variance.Change.Lin.SD
j <- j + 1
Result.Lin.SD[Label.Row.Lin.SD,j] <- Correlation.Lin.SD
}
#----------------------
# End loop
#----------------------
#=====================================================================
# END SIMULATION
#=====================================================================
```

Appendix B

Supplementary Material for the Second Paper

By Melike Oguz-Alper and Yves G. Berger
University of Southampton, SO17 1BJ, Southampton, U.K.
M.OguzAlper@soton.ac.uk Y.G.Berger@soton.ac.uk

B.1 Proof of expression (3.47)

Under the regularity conditions (3.28)–(3.33), and for $\boldsymbol{\psi}$ such that condition (3.46) holds, the result of the Appendix A in Berger and De La Riva Torres (2016) can be used to show that

$$\widehat{\boldsymbol{G}}(\boldsymbol{\psi}) = \widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}) - \sum_{i\in s} \frac{\boldsymbol{g}_i(\boldsymbol{\psi})\boldsymbol{c}_i{}^{\mathrm{T}}}{\pi_i^2} \left(\sum_{i\in s} \frac{\boldsymbol{c}_i\boldsymbol{c}_i{}^{\mathrm{T}}}{\pi_i^2} \right)^{-1} \begin{pmatrix} \mathbf{0} \\ \widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N) \end{pmatrix} + o_{\mathcal{P}}(Nn^{-1/2}). \quad \text{(B.1)}$$

The inverse of the block matrix $\sum_{i\in s} \boldsymbol{c}_i\boldsymbol{c}_i{}^{\mathrm{T}}\pi_i^{-2}$ is given by

$$\left(\sum_{i\in s} \frac{\boldsymbol{c}_i\boldsymbol{c}_i{}^{\mathrm{T}}}{\pi_i^2} \right)^{-1} = \begin{pmatrix} \bullet & -n^{-1}\,\widehat{\boldsymbol{V}}_{\mathbf{ff}}^{-1}\,\widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N) \\ \bullet & \widehat{\boldsymbol{V}}_{\mathbf{ff}}^{-1} \end{pmatrix}, \quad \text{(B.2)}$$

where $\widehat{\boldsymbol{V}}_{\mathbf{ff}}$ is defined by (3.50). We do not need to compute the submatrices denoted by $\bullet$, since we multiply those components by zero in equation (B.1). Replacing equation (B.2) into equation (B.1), we obtain equation (3.47). This completes the proof.

B.2 Proofs of the asymptotic results

Lemma B.1. Let $\widehat{m}_i^* = (\pi_i + \overset{*}{\boldsymbol{\eta}}{}^{\mathrm{T}} \boldsymbol{c}_i^*)^{-1}$, where $\overset{*}{\boldsymbol{\eta}}$ is such that $\sum_{i\in s} \widehat{m}_i^* \boldsymbol{c}_i^* = \boldsymbol{C}^*$ holds, with $\boldsymbol{c}_i^* = (\bar{\pi}^{-1}\boldsymbol{z}_i{}^{\mathrm{T}}, \boldsymbol{g}_i{}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{C}^* = (\bar{\pi}^{-1}\boldsymbol{n}^{\mathrm{T}}, \boldsymbol{0}_b{}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{z}_i$ and $\boldsymbol{n}$ are $H \times 1$ vectors that specify the stratification information (see definition (3.41)). The vector $\boldsymbol{g}_i$ could be any $b \times 1$ vector. Then we have that $\overset{*}{\boldsymbol{\eta}}{}^{\mathrm{T}} \boldsymbol{C}^* = 0$.

With a single stratum, $\boldsymbol{z}_i = \pi_i$ and $\boldsymbol{n} = n$ (see Section 3.9.3). Hence, the Lemma B.1 holds with $\boldsymbol{c}_i^* = (\bar{\pi}^{-1}\pi_i, \boldsymbol{g}_i{}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{C}^* = (\bar{\pi}^{-1}n, \boldsymbol{0}_b{}^{\mathrm{T}})^{\mathrm{T}}$.

Proof of Lemma B.1. We have that $\widehat{m}_i^* = (\pi_i(1 + v_i^*))^{-1}$, where $v_i^* = \pi_i^{-1}\boldsymbol{c}_i^{*\mathrm{T}} \overset{*}{\boldsymbol{\eta}}$. Thus $\widehat{m}_i^* = \pi_i^{-1} - \{\pi_i(1 + v_i^*)\}^{-1} v_i^* = \pi_i^{-1} - \left\{\pi_i^2(1 + v_i^*)\right\}^{-1} \boldsymbol{c}_i^{*\mathrm{T}} \overset{*}{\boldsymbol{\eta}}$. Hence, we have that

$$\sum_{i\in s} \widehat{m}_i^* \boldsymbol{c}_i^* = \widehat{\boldsymbol{C}}_\pi^* - \widetilde{\boldsymbol{\Sigma}}^* \overset{*}{\boldsymbol{\eta}}, \tag{B.3}$$

where $\widehat{\boldsymbol{C}}_\pi^* = \sum_{i\in s} \pi_i^{-1}\boldsymbol{c}_i^*$ and

$$\widetilde{\boldsymbol{\Sigma}}^* = \sum_{i\in s} \frac{\boldsymbol{c}_i^* \boldsymbol{c}_i^{*\mathrm{T}}}{\pi_i^2(1 + v_i^*)}. \tag{B.4}$$

As $\sum_{i\in s} \widehat{m}_i^* \boldsymbol{c}_i^* = \boldsymbol{C}^*$, the equation (B.3) can be re-written as

$$\widetilde{\boldsymbol{\Sigma}}^* \overset{*}{\boldsymbol{\eta}} = \widehat{\boldsymbol{C}}_\pi^* - \boldsymbol{C}^* = \left(\boldsymbol{0}_r{}^{\mathrm{T}}, \sum_{i\in s} \frac{\boldsymbol{g}_i{}^{\mathrm{T}}}{\pi_i}\right)^{\mathrm{T}}, \tag{B.5}$$

where $\boldsymbol{0}_r$ is an $r \times 1$ vector of zeros, where $r = H$. Consider $\boldsymbol{L} = (\boldsymbol{1}_r{}^{\mathrm{T}}, \boldsymbol{0}_b{}^{\mathrm{T}})^{\mathrm{T}}$; where $\boldsymbol{1}_r$ is an $r \times 1$ vector of ones and $\boldsymbol{0}_b$ is a $b \times 1$ vector of zeros. We have

$$\boldsymbol{L}^{\mathrm{T}} \widetilde{\boldsymbol{\Sigma}}^* \overset{*}{\boldsymbol{\eta}} = 0, \tag{B.6}$$

because $\boldsymbol{L}$ and $\widetilde{\boldsymbol{\Sigma}}^* \overset{*}{\boldsymbol{\eta}}$ are orthogonal (see expression (B.5)). As $\widehat{m}_i^* = (\pi_i(1 + v_i^*))^{-1}$, the equation (B.6) can be re-written as

$$\boldsymbol{1}_r{}^{\mathrm{T}} \bar{\pi}^{-1} \sum_{i\in s} \frac{\widehat{m}_i^*}{\pi_i} \boldsymbol{z}_i \boldsymbol{z}_i{}^{\mathrm{T}} \boldsymbol{\eta}_r^* + \boldsymbol{1}_r{}^{\mathrm{T}} \sum_{i\in s} \frac{\widehat{m}_i^*}{\pi_i} \boldsymbol{z}_i \boldsymbol{g}_i{}^{\mathrm{T}} \boldsymbol{\eta}_b^* = 0, \tag{B.7}$$

where $\boldsymbol{\eta}_r^*$ is sub-vector of the first r components of $\overset{*}{\boldsymbol{\eta}}$ and $\boldsymbol{\eta}_b^*$ is the vector of the remaining components of $\overset{*}{\boldsymbol{\eta}}$; that is, $\overset{*}{\boldsymbol{\eta}} = (\boldsymbol{\eta}_r^{*\mathrm{T}}, \boldsymbol{\eta}_b^{*\mathrm{T}})^{\mathrm{T}}$. As $\boldsymbol{1}_r{}^{\mathrm{T}} \boldsymbol{z}_i = \pi_i$, the equation (B.7) reduces to

$$\bar{\pi}^{-1} \sum_{i\in s} \widehat{m}_i^* \boldsymbol{z}_i{}^{\mathrm{T}} \boldsymbol{\eta}_r^* + \sum_{i\in s} \widehat{m}_i^* \boldsymbol{g}_i{}^{\mathrm{T}} \boldsymbol{\eta}_b^* = 0 \tag{B.8}$$

or equivalently $\bar{\pi}^{-1} \boldsymbol{n}^{\mathrm{T}} \boldsymbol{\eta}_r^* = 0$, as $\sum_{i\in s} \widehat{m}_i^* \boldsymbol{c}_i^* = \boldsymbol{C}^*$ implies $\sum_{i\in s} \widehat{m}_i^* \boldsymbol{z}_i = \boldsymbol{n}$ and $\sum_{i\in s} \widehat{m}_i^* \boldsymbol{g}_i = 0$, by definition of $\boldsymbol{z}_i$ and $\boldsymbol{C}^*$. As the last b components of $\boldsymbol{C}^*$ are equal to zero, we have

that $\boldsymbol{C}^{*\mathrm{T}}\overset{*}{\boldsymbol{\eta}} = \bar{\pi}^{-1}\boldsymbol{n}^{\mathrm{T}}\boldsymbol{\eta}_r^* = 0$. Thus $\boldsymbol{C}^{*\mathrm{T}}\overset{*}{\boldsymbol{\eta}} = 0$, because $\bar{\pi}^{-1}\boldsymbol{n}^{\mathrm{T}}\boldsymbol{\eta}_r^* = 0$. This completes the proof. □

Lemma B.2. Let $\mathring{\boldsymbol{\nu}}_N$ and $\mathring{\boldsymbol{\eta}}_N$ be the solution of (3.22), with $\boldsymbol{\theta} = \boldsymbol{\theta}_N$. In other words, $\mathring{\boldsymbol{\nu}}_N = \mathring{\boldsymbol{\nu}}(\boldsymbol{\theta}_N)$ and $\mathring{\boldsymbol{\eta}}_N = \overset{*}{\boldsymbol{\eta}}(\mathring{\boldsymbol{\psi}}_N)$, with $\mathring{\boldsymbol{\psi}}_N = (\boldsymbol{\theta}_N{}^{\mathrm{T}}, \mathring{\boldsymbol{\nu}}(\boldsymbol{\theta}_N)^{\mathrm{T}})^{\mathrm{T}}$. Assuming that $\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)$ and $\boldsymbol{C}^*$ are such that the regularity conditions (3.28)–(3.32) hold, we have

$$\mathring{\boldsymbol{\eta}}_N = (\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}(\widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^*) + \bar{\pi}\widehat{\boldsymbol{e}}, \tag{B.9}$$

where $\widehat{\boldsymbol{e}}$ is such that $\|\widehat{\boldsymbol{e}}\| = O_{\mathcal{P}}(n^{-1})$, where $\mathbf{I}_{r+b}$ is an $(r+b)\times(r+b)$ identity matrix with $b = p+q$, and $\widehat{\boldsymbol{A}}^*$ is an $(r+b)\times(r+b)$ symmetric and idempotent matrix defined by

$$\widehat{\boldsymbol{A}}^* = \widehat{\boldsymbol{\Sigma}}^{*-1/2}\widehat{\boldsymbol{\nabla}}_{\boldsymbol{C}}^*(\widehat{\boldsymbol{\nabla}}_{\boldsymbol{C}}^{*\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{*-1}\widehat{\boldsymbol{\nabla}}_{\boldsymbol{C}}^*)^{-1}\widehat{\boldsymbol{\nabla}}_{\boldsymbol{C}}^{*\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{*-1/2}, \tag{B.10}$$

where

$$\widehat{\boldsymbol{\Sigma}}^* = -N\bar{\pi}^{-1}\widehat{\boldsymbol{S}}^*(\boldsymbol{\psi}_N) = \sum_{i\in s}\frac{1}{\pi_i^2}\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)^{\mathrm{T}} \tag{B.11}$$

$$\widehat{\boldsymbol{\nabla}}_{\boldsymbol{C}}^* = \sum_{i\in s}\frac{1}{\pi_i}\frac{\partial \boldsymbol{c}_i^*(\boldsymbol{\psi})}{\partial \boldsymbol{\nu}}\bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}_N}. \tag{B.12}$$

Proof of Lemma B.2. This proof has two parts. In part 1, we derive the first order approximation of $\mathring{\boldsymbol{\eta}}_N$ (the right hand side of (B.9)). In part 2, we show that $\|\widehat{\boldsymbol{e}}\| = O_{\mathcal{P}}(n^{-1})$.

Part 1:

Consider the function $\boldsymbol{\Gamma}(\boldsymbol{\eta},\boldsymbol{\nu})$ defined by expression (3.23). A first order Taylor approximation of $\boldsymbol{\Gamma}(\boldsymbol{\eta},\boldsymbol{\nu})$ around $\boldsymbol{\Gamma}(\mathbf{0},\boldsymbol{\nu}_N)$ gives

$$\boldsymbol{\Gamma}(\boldsymbol{\eta},\boldsymbol{\nu}) \doteq \boldsymbol{\Gamma}(\mathbf{0},\boldsymbol{\nu}_N) + \widehat{\boldsymbol{\nabla}}(\mathbf{0},\boldsymbol{\nu}_N)\left((\boldsymbol{\eta}-\mathbf{0})^{\mathrm{T}}, (\boldsymbol{\nu}-\boldsymbol{\nu}_N)^{\mathrm{T}}\right)^{\mathrm{T}},$$

or equivalently

$$\begin{pmatrix} \boldsymbol{\eta} \\ \boldsymbol{\nu}-\boldsymbol{\nu}_N \end{pmatrix} \doteq \widehat{\boldsymbol{\nabla}}(\mathbf{0},\boldsymbol{\nu}_N)^{-1}\left(\boldsymbol{\Gamma}(\boldsymbol{\eta},\boldsymbol{\nu}) - \boldsymbol{\Gamma}(\mathbf{0},\boldsymbol{\nu}_N)\right), \tag{B.13}$$

where $\widehat{\boldsymbol{\nabla}}(\mathbf{0},\boldsymbol{\nu}_N)$ is the value of the derivative (3.25), with $\boldsymbol{\eta} = \mathbf{0}$ and $\boldsymbol{\nu} = \boldsymbol{\nu}_N$. It can be shown that

$$\widehat{\boldsymbol{\nabla}}(\mathbf{0},\boldsymbol{\nu}_N) = \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^* & \widehat{\boldsymbol{\nabla}}_{\boldsymbol{C}}^* \\ \widehat{\boldsymbol{\nabla}}_{\boldsymbol{C}}^{*\mathrm{T}} & \mathbf{0}_q \end{pmatrix},$$

where $\widehat{\boldsymbol{\Sigma}}^*$ and $\widehat{\boldsymbol{\nabla}}^*_{\boldsymbol{C}}$ are respectively defined by (B.11) and (B.12). We also have that

$$\boldsymbol{\Gamma}(\mathbf{0}, \boldsymbol{\nu}_N) = \begin{pmatrix} \widehat{\boldsymbol{C}}^*_\pi(\boldsymbol{\psi}_N) - \boldsymbol{C}^* \\ \mathbf{0}_q \end{pmatrix}, \tag{B.14}$$

$$\boldsymbol{\Gamma}(\mathring{\boldsymbol{\eta}}_N, \mathring{\boldsymbol{\nu}}_N) = \mathbf{0}_{r+b+q}, \tag{B.15}$$

as $\mathring{\boldsymbol{\eta}}_N$ and $\mathring{\boldsymbol{\nu}}_N$ are the solutions of (3.22). Thus when $\boldsymbol{\eta} = \mathring{\boldsymbol{\eta}}_N$ and $\boldsymbol{\nu} = \mathring{\boldsymbol{\nu}}_N$, the expression (B.13) reduces to

$$\begin{pmatrix} \mathring{\boldsymbol{\eta}}_N \\ \mathring{\boldsymbol{\nu}}_N - \boldsymbol{\nu}_N \end{pmatrix} \simeq \begin{pmatrix} \widehat{\boldsymbol{\Sigma}}^* & \widehat{\boldsymbol{\nabla}}^*_{\boldsymbol{C}} \\ \widehat{\boldsymbol{\nabla}}^{*\mathrm{T}}_{\boldsymbol{C}} & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{C}^* - \widehat{\boldsymbol{C}}^*_\pi(\boldsymbol{\psi}_N) \\ \mathbf{0}_q \end{pmatrix}, \tag{B.16}$$

by using (B.14) and (B.15). By taking the inverse of the block matrix in (B.16), we obtain the right hand side of equation (B.9) excluding the remainder term $\bar{\pi}\widehat{\boldsymbol{e}}$.

Part 2:

Now, we show that $\|\widehat{\boldsymbol{e}}\| = O_{\mathcal{P}}(n^{-1})$. Let

$$\|\widehat{\boldsymbol{e}}\| = O_{\mathcal{P}}(n^{-t}) \tag{B.17}$$

for some t. The Lemma is proven if $t = 1$. In the rest of this proof, we show that $t = 1$.

By multiplying both sides of the equation (B.9) by $(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)$, we obtain

$$(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\mathring{\boldsymbol{\eta}}_N = (\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}(\widehat{\boldsymbol{C}}^*_\pi(\boldsymbol{\psi}_N) - \boldsymbol{C}^*) + \bar{\pi}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{e}}, \tag{B.18}$$

as $(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*) = (\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)$. It can be shown that

$$\widehat{\boldsymbol{C}}^*_\pi(\boldsymbol{\psi}_N) - \boldsymbol{C}^* = \sum_{i \in s} \frac{1}{\pi_i} \boldsymbol{c}^*_i(\boldsymbol{\psi}_N) v^*_i - \sum_{j \in s} \frac{1}{\pi_j} \boldsymbol{c}^*_j(\boldsymbol{\psi}_N) \gamma_j, \tag{B.19}$$

where

$$v^*_i = \pi_i^{-1} \boldsymbol{c}^*_i(\boldsymbol{\psi}_N)^{\mathrm{T}} \mathring{\boldsymbol{\eta}}_N, \tag{B.20}$$

$$\begin{aligned} \gamma_j &= (1 + v^*_j)^{-1} - 1 + v^*_j \\ &= (\pi_j + \mathring{\boldsymbol{\eta}}_N{}^{\mathrm{T}} \boldsymbol{c}^*_j(\boldsymbol{\psi}_N))^{-1} \pi_j - 1 + v^*_j. \end{aligned} \tag{B.21}$$

By using the definition (B.11) and the expression for v^*_i, it can be shown that

$$(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1} \sum_{i \in s} \frac{1}{\pi_i} \boldsymbol{c}^*_i(\boldsymbol{\psi}_N) v^*_i = (\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\mathring{\boldsymbol{\eta}}_N. \tag{B.22}$$

By using equation (B.19) and (B.22), we obtain the following expression.

$$(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}(\widehat{\boldsymbol{C}}^*_\pi(\boldsymbol{\psi}_N) - \boldsymbol{C}^*) = (\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\mathring{\boldsymbol{\eta}}_N - \bar{\pi}\widehat{\boldsymbol{e}}_1, \tag{B.23}$$

where

$$\widehat{\boldsymbol{e}}_1 = \bar{\pi}^{-1}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}\sum_{i\in s}\boldsymbol{c}^*_i(\boldsymbol{\psi}_N)\frac{\gamma_i}{\pi_i}\cdot \tag{B.24}$$

By substituting expression (B.23) within equation (B.18), we obtain

$$\widehat{\boldsymbol{e}}_1 = (\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{e}},$$

which implies

$$\begin{aligned}
\|\widehat{\boldsymbol{e}}_1\|^2 &= \|(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{e}}\|^2 \\
&= \mathrm{tr}(\widehat{\boldsymbol{e}}^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{e}}) \\
&= \mathrm{tr}(\widehat{\boldsymbol{e}}^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{e}}) \\
&= \mathrm{tr}(\widehat{\boldsymbol{e}}^{\mathrm{T}}\widehat{\boldsymbol{e}}) - \mathrm{tr}((\widehat{\boldsymbol{A}}^*\widehat{\boldsymbol{e}})^{\mathrm{T}}(\widehat{\boldsymbol{A}}^*\widehat{\boldsymbol{e}})) \\
&= \|\widehat{\boldsymbol{e}}\|^2 - \|\widehat{\boldsymbol{A}}^*\widehat{\boldsymbol{e}}\|^2\cdot
\end{aligned} \tag{B.25}$$

We have that $\|\widehat{\boldsymbol{A}}^*\| = [\mathrm{tr}(\widehat{\boldsymbol{A}}^*)]^{1/2} = \dim(\boldsymbol{\nu}_N)^{1/2} = O(1)$, as the nuisance parameter $\boldsymbol{\nu}_N$ is of finite dimension. Thus equation (B.17) implies that $\|\widehat{\boldsymbol{A}}^*\widehat{\boldsymbol{e}}\| \leqslant \|\widehat{\boldsymbol{A}}^*\|\,\|\widehat{\boldsymbol{e}}\| = O_{\mathcal{P}}(n^{-t})$, or equivalently

$$\|\widehat{\boldsymbol{A}}^*\widehat{\boldsymbol{e}}\| = O_{\mathcal{P}}(n^{-t}), \tag{B.26}$$

Thus by combining (B.17), (B.25) and (B.26), we obtain

$$\|\widehat{\boldsymbol{e}}_1\| = O_{\mathcal{P}}(n^{-t})\cdot \tag{B.27}$$

Now, we derive the order of $\widehat{\boldsymbol{e}}_1$ to find the value of t. The equation (B.24) implies that

$$\|\widehat{\boldsymbol{e}}_1\| \leqslant \frac{1}{N}\|\widehat{\boldsymbol{D}}\|\sum_{i\in s}\frac{|\gamma_i|}{\pi_i}\|\boldsymbol{c}^*_i(\boldsymbol{\psi}_N)\|, \tag{B.28}$$

where $\|\widehat{\boldsymbol{D}}\| = N\bar{\pi}^{-1}\|\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*\|\,\|\widehat{\boldsymbol{\Sigma}}^{*-1}\| = \|\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*\|\,\|\widehat{\boldsymbol{S}}^*(\boldsymbol{\psi}_N)^{-1}\|$. As $(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)$ is a symmetric and idempotent matrix, we have that

$$\begin{aligned}
\|(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\| &= [\mathrm{tr}((\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*))]^{1/2} = [\mathrm{tr}((\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*))]^{1/2} \\
&= [r + b - \dim(\boldsymbol{\nu}_N)]^{1/2} = O(1),
\end{aligned} \tag{B.29}$$

because $\dim(\boldsymbol{\nu}_N) = O(1)$, $r = O(1)$ and $b = O(1)$. Thus the condition (3.32) and the equation (B.29) imply that

$$\|\widehat{\boldsymbol{D}}\| = O_p(1)\cdot \tag{B.30}$$

The definition (B.21) implies that

$$|\gamma_j| \leqslant |(1+v_j^*)^{-1} - 1 + v_j^*| \leqslant v_j^{*2} + |\varepsilon_j|,$$

where ε_j is such that $Pr(|\varepsilon_i| \leqslant \kappa |v_j^*|^3, j \in s) \to 1$, for some finite constant $\kappa > 0$. Thus the definition (B.20) implies that

$$|\gamma_j| \leqslant \frac{1}{\pi_j^2}\|\boldsymbol{c}_j^*(\boldsymbol{\psi}_N)\|^2 \|\mathring{\boldsymbol{\eta}}_N\|^2 + |\varepsilon_j|\cdot \tag{B.31}$$

Under the regularity conditions given by (3.28)-(3.31), Berger and De La Riva Torres (2016) showed that

$$\bar{\pi}^{-1}\|\mathring{\boldsymbol{\eta}}_N\| = O_{\mathcal{P}}(n^{-1/2})\cdot \tag{B.32}$$

By combining equation (B.28) and (B.31), and by using $|v_i^*|^3 \leqslant \pi_i^{-3}\|\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)\|^3 \|\mathring{\boldsymbol{\eta}}_N\|^3$ we have

$$\begin{aligned}
\|\widehat{\boldsymbol{e}}_1\| &\leqslant \|\widehat{\boldsymbol{D}}\|\|\mathring{\boldsymbol{\eta}}_N\|^2 \frac{1}{N}\sum_{i\in s}\frac{1}{\pi_j^3}\|\boldsymbol{c}_j^*(\boldsymbol{\psi}_N)\|^3 + \frac{1}{N}\|\widehat{\boldsymbol{D}}\|\sum_{i\in s}\frac{|\varepsilon_j|}{\pi_i}\|\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)\| \\
&\leqslant \|\widehat{\boldsymbol{D}}\|\|\mathring{\boldsymbol{\eta}}_N\|^2 \frac{1}{N}\sum_{i\in s}\frac{1}{\pi_j^3}\|\boldsymbol{c}_j^*(\boldsymbol{\psi}_N)\|^3 + \|\widehat{\boldsymbol{D}}\|\|\mathring{\boldsymbol{\eta}}_N\|^3 \frac{\kappa}{N}\sum_{i\in s}\frac{1}{\pi_j^4}\|\boldsymbol{c}_j^*(\boldsymbol{\psi}_N)\|^4 \\
&\leqslant \|\widehat{\boldsymbol{D}}\|\frac{N^2}{n^2}\|\mathring{\boldsymbol{\eta}}_N\|^2 \frac{\bar{\pi}^3}{n}\sum_{i\in s}\frac{1}{\pi_j^3}\|\boldsymbol{c}_j^*(\boldsymbol{\psi}_N)\|^3 + \|\widehat{\boldsymbol{D}}\|\frac{N^3}{n^3}\|\mathring{\boldsymbol{\eta}}_N\|^3 \kappa \frac{\bar{\pi}^4}{n}\sum_{i\in s}\frac{1}{\pi_j^4}\|\boldsymbol{c}_j^*(\boldsymbol{\psi}_N)\|^4 \\
&= O_{\mathcal{P}}(n^{-1}) + O_{\mathcal{P}}(n^{-3/2}) = O_{\mathcal{P}}(n^{-1}),
\end{aligned} \tag{B.33}$$

by using the condition (3.33), and the expressions (B.30) and (B.32). Thus the expressions (B.27) and (B.33) imply $t = 1$. The Lemma follows from the expression (B.17). □

Lemma B.3. Under the regularity conditions (3.28)–(3.33), we have that

$$2\left\{\ell(\boldsymbol{\pi}) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}}\ell(\boldsymbol{\theta}_N, \boldsymbol{\nu})\right\} = (\widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^*)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}(\widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^*) + O_{\mathcal{P}}(n^{-1/2}), \tag{B.34}$$

where

$$\ell(\boldsymbol{\pi}) = -\sum_{i\in s}\log(\pi_i), \tag{B.35}$$

$\widehat{\boldsymbol{\Sigma}}^*$ is defined by equation (B.11) and $\widehat{\boldsymbol{A}}^*$ is defined by equation (B.10).

Proof of Lemma B.3. From Section 3.6, we have that

$$\max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}} \ell\left(\boldsymbol{\theta}_N, \boldsymbol{\nu}\right) = \ell\left(\boldsymbol{\theta}_N, \mathring{\boldsymbol{\nu}}_N\right) = -\sum_{i\in s} \log\left(\pi_i + \mathring{\boldsymbol{\eta}}_N{}^{\mathrm{T}}\, \boldsymbol{c}_i^*(\mathring{\boldsymbol{\psi}}_N)\right), \tag{B.36}$$

where $\mathring{\boldsymbol{\psi}}_N = (\boldsymbol{\theta}_N{}^{\mathrm{T}}, \mathring{\boldsymbol{\nu}}_N{}^{\mathrm{T}})^{\mathrm{T}}$. Here, $\mathring{\boldsymbol{\nu}}_N$ and $\mathring{\boldsymbol{\eta}}_N$ be the solutions of (3.22), with $\boldsymbol{\theta} = \boldsymbol{\theta}_N$. The equation (B.36) implies

$$2\left\{\ell(\boldsymbol{\pi}) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}} \ell\left(\boldsymbol{\theta}_N, \boldsymbol{\nu}\right)\right\} = 2\sum_{i\in s} \log\left(1+ \mathring{\varrho}_i\right), \tag{B.37}$$

with

$$\mathring{\varrho}_i = \pi_i^{-1} \boldsymbol{c}_i^*(\mathring{\boldsymbol{\psi}}_N)^{\mathrm{T}} \mathring{\boldsymbol{\eta}}_N.$$

The expression (B.37) is a function of $(\mathring{\boldsymbol{\eta}}_N{}^{\mathrm{T}}, \boldsymbol{\theta}_N{}^{\mathrm{T}}, \mathring{\boldsymbol{\nu}}_N{}^{\mathrm{T}})^{\mathrm{T}}$. A multivariate Taylor expansion of this function around $(\mathbf{0}_r{}^{\mathrm{T}}, \boldsymbol{\theta}_N{}^{\mathrm{T}}, \boldsymbol{\nu}_N{}^{\mathrm{T}})^{\mathrm{T}}$ is given by

$$2\left\{\ell(\boldsymbol{\pi}) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}} \ell\left(\boldsymbol{\theta}_N, \boldsymbol{\nu}\right)\right\} = 2\sum_{i\in s} \varrho_i - \sum_{i\in s} \varrho_i^2 + 2\sum_{i\in s} \varphi_i, \tag{B.38}$$

where

$$\varrho_i \;=\; \pi_i^{-1} \boldsymbol{c}_i^*(\boldsymbol{\psi}_N)^{\mathrm{T}} \mathring{\boldsymbol{\eta}}_N \tag{B.39}$$

and φ_i is such that $Pr\{|\varphi_i| \leqslant \kappa|\varrho_i^3|, i \in s\} \to 1$ for some finite $\kappa > 0$. By using the expressions (B.9) and (B.39); and the fact that $\mathring{\boldsymbol{\eta}}_N{}^{\mathrm{T}}\boldsymbol{C}^* = 0$ (see Lemma B.1), we obtain

$$\begin{aligned}
\sum_{i\in s} \varrho_i \;&=\; \sum_{i\in s} \frac{\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)^{\mathrm{T}}}{\pi_i} \mathring{\boldsymbol{\eta}}_N = (\widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^*)^{\mathrm{T}} \mathring{\boldsymbol{\eta}}_N + \mathring{\boldsymbol{\eta}}_N{}^{\mathrm{T}} \boldsymbol{C}^* \\
&=\; \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}} (\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*) \widehat{\boldsymbol{\Sigma}}^{*-1} \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) + \bar{\pi}\, \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}} \widehat{\boldsymbol{e}},
\end{aligned} \tag{B.40}$$

where

$$\widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) = \widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^*. \tag{B.41}$$

Furthermore, by using the expressions (B.9) and (B.11), we obtain

$$\begin{aligned}
\sum_{i\in s} \varrho_i^2 \;&=\; \mathring{\boldsymbol{\eta}}_N{}^{\mathrm{T}} \sum_{i\in s} \frac{1}{\pi_i^2} \boldsymbol{c}_i^*(\boldsymbol{\psi}_N) \boldsymbol{c}_i^*(\boldsymbol{\psi}_N)^{\mathrm{T}} \mathring{\boldsymbol{\eta}}_N \\
&=\; \mathring{\boldsymbol{\eta}}_N{}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}^* \mathring{\boldsymbol{\eta}}_N \\
&=\; \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}} \widehat{\boldsymbol{E}}^* \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{E}}^* \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) + 2\bar{\pi}\, \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}} \widehat{\boldsymbol{E}}^* \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{e}} + \bar{\pi}^2 \widehat{\boldsymbol{e}}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}^* \widehat{\boldsymbol{e}},
\end{aligned} \tag{B.42}$$

where $\widehat{\boldsymbol{E}}^*$ is a symmetric matrix such that $\widehat{\boldsymbol{E}}^* = (\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}$. We have $\widehat{\boldsymbol{E}}^*\widehat{\boldsymbol{\Sigma}}^*\widehat{\boldsymbol{E}}^* = \widehat{\boldsymbol{E}}^*$ as $(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)$ is an idempotent matrix. Thus equation (B.42) reduces to

$$\sum_{i\in s} \varrho_i^2 = \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}\widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) + 2\bar{\pi}\widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{e}} + \bar{\pi}^2\widehat{\boldsymbol{e}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^*\widehat{\boldsymbol{e}}. \tag{B.43}$$

We also have that

$$|\sum_{i\in s}\varphi_i| \leqslant \kappa\sum_{i\in s}|\varrho_i|^3 \leqslant \kappa\|\mathring{\boldsymbol{\eta}}_N\|^3\sum_{i\in s}\frac{1}{\pi_i^3}\|\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)\|^3 = O_{\mathcal{P}}(n^{-1/2}), \tag{B.44}$$

by using (3.33) and (B.32). Thus by using the expressions (B.40) and (B.43), and inequality (B.44), the equation (B.38) reduces to

$$\begin{aligned} 2\left\{\ell(\boldsymbol{\pi}) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}}\ell(\boldsymbol{\theta}_N, \boldsymbol{\nu})\right\} &= \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}\widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) + 2\bar{\pi}\widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)\widehat{\boldsymbol{A}}^*\widehat{\boldsymbol{e}} - \bar{\pi}^2\widehat{\boldsymbol{e}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^*\widehat{\boldsymbol{e}} \\ &\quad + O_{\mathcal{P}}(n^{-1/2}) \\ &= \widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}\widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) + O_{\mathcal{P}}(n^{-1/2}), \end{aligned} \tag{B.45}$$

because $|\widetilde{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N)^{\mathrm{T}}\widehat{\boldsymbol{A}}^*\widehat{\boldsymbol{e}}| = O_{\mathcal{P}}(Nn^{-3/2})$ and $|\widehat{\boldsymbol{e}}^{\mathrm{T}}\widehat{\boldsymbol{S}}^*(\boldsymbol{\psi}_N)\widehat{\boldsymbol{e}}| = O_{\mathcal{P}}(n^{-2})$, by using $\|\widehat{\boldsymbol{A}}^*\| = O(1)$, the regularity conditions (3.29) and (3.31), and $\|\widehat{\boldsymbol{e}}\| = O_{\mathcal{P}}(n^{-1})$ (see Lemma B.2). The Lemma follows from (B.41) and (B.45). □

Theorem B.1. Let $\boldsymbol{c}_i = \bar{\pi}^{-1}\boldsymbol{z}_i$, where the $\boldsymbol{z}_i$ are the stratification variables defined by (3.41). Suppose that the regularity conditions (3.28)–(3.33) hold for $\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)$ defined by expression (3.12), under an unequal probability (with replacement) stratified sampling design. We have that

$$\widehat{r}(\boldsymbol{\theta}_N) = \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)^{\mathrm{T}}(\mathbf{I}_b - \widehat{\boldsymbol{A}}_{\boldsymbol{g}|\boldsymbol{z}})\widehat{\boldsymbol{V}}_{\boldsymbol{gg}|\boldsymbol{z}}^{-1}\,\widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N) + O_{\mathcal{P}}(n^{-1/2}),$$

where $\widehat{r}(\boldsymbol{\theta}_N)$ and $\widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)$ are respectively defined by (3.15) and (3.36), and $\widehat{\boldsymbol{A}}_{\boldsymbol{g}|\boldsymbol{z}}$ is given by

$$\widehat{\boldsymbol{A}}_{\boldsymbol{g}|\boldsymbol{z}} = \widehat{\boldsymbol{V}}_{\boldsymbol{gg}|\boldsymbol{z}}^{-1/2}\,\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}\left(\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}{}^{\mathrm{T}}\widehat{\boldsymbol{V}}_{\boldsymbol{gg}|\boldsymbol{z}}^{-1}\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}\right)^{-1}\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}{}^{\mathrm{T}}\,\widehat{\boldsymbol{V}}_{\boldsymbol{gg}|\boldsymbol{z}}^{-1/2},$$

where $\widehat{\boldsymbol{\nabla}}_{\boldsymbol{G}}$ is defined by (3.38). Here, $\widehat{\boldsymbol{V}}_{\boldsymbol{gg}|\boldsymbol{z}}$ is the Hansen and Hurwitz (1943) stratified variance estimator given by

$$\widehat{\boldsymbol{V}}_{\boldsymbol{gg}|\boldsymbol{z}} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{gg}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}\boldsymbol{g}}{}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}\boldsymbol{g}}, \tag{B.46}$$

with

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{gg}} &= \sum_{i\in\boldsymbol{s}} \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)\ \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)^{\mathrm{T}},\\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}} &= \sum_{i\in\boldsymbol{s}} \breve{\boldsymbol{z}}_i\ \breve{\boldsymbol{z}}_i{}^{\mathrm{T}},\\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zg}} &= \sum_{i\in\boldsymbol{s}} \breve{\boldsymbol{z}}_i\ \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)^{\mathrm{T}}.\end{aligned} \tag{B.47}$$

Here, $\breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N) = \boldsymbol{g}_i(\boldsymbol{\psi}_N)\pi_i^{-1}$ and $\breve{\boldsymbol{z}}_i = \boldsymbol{z}_i\pi_i^{-1}$.

Berger and De La Riva Torres (2016) show that $\widehat{\boldsymbol{V}}_{\boldsymbol{gg|z}}$ is an alternative expression for the Hansen and Hurwitz (1943) stratified variance estimator.

Proof of Theorem B.1. As $\boldsymbol{c}_i = \bar{\pi}^{-1}\boldsymbol{z}_i$, the vector $\boldsymbol{\eta}$ in expression (3.10), which satisfies the constraint (3.9), is given by $\boldsymbol{\eta} = \boldsymbol{0}_r$. Hence, the unique solution is $\widehat{m}_i = \pi_i^{-1}$. The expression (3.16) implies that $\ell(\widehat{\boldsymbol{\psi}}) = \ell(\boldsymbol{\pi})$ where $\ell(\boldsymbol{\pi})$ is defined by expression (B.35). Thus $\widehat{r}(\boldsymbol{\theta}_N) = 2\{\ell(\boldsymbol{\pi}) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}_N, \boldsymbol{\nu})\}$ (see definition (3.15)) and by using (B.34), we have that

$$\widehat{r}(\boldsymbol{\theta}_N) = (\widehat{\boldsymbol{C}}^*_\pi(\boldsymbol{\psi}_N) - \boldsymbol{C}^*)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}(\widehat{\boldsymbol{C}}^*_\pi(\boldsymbol{\psi}_N) - \boldsymbol{C}^*) + O_{\mathcal{P}}(n^{-1/2}), \tag{B.48}$$

by using Lemma B.3, where $\widehat{\boldsymbol{A}}^*$ is defined by expression (B.10) and $\widehat{\boldsymbol{\Sigma}}^*$ is defined by expression (B.11). It can be shown that

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}^* &= \begin{pmatrix} \bar{\pi}^{-2}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}} & \bar{\pi}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zg}} \\ \bar{\pi}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{gz}} & \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{gg}} \end{pmatrix},\\ \widehat{\boldsymbol{C}}^*_\pi(\boldsymbol{\psi}_N) - \boldsymbol{C}^* &= (\boldsymbol{0}_r{}^{\mathrm{T}}, \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

Hence, expression (B.48) implies that

$$\begin{aligned}\widehat{r}(\boldsymbol{\theta}_N) &= (\boldsymbol{0}_r{}^{\mathrm{T}}, \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)^{\mathrm{T}})(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}(\boldsymbol{0}_r{}^{\mathrm{T}}, \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)^{\mathrm{T}})^{\mathrm{T}} + O_{\mathcal{P}}(n^{-1/2})\\ &= \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)^{\mathrm{T}}(\mathbf{I}_b - \widehat{\boldsymbol{A}}_{\boldsymbol{g|z}})(\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{gg}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{gz}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zg}})^{-1}\ \widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N) +\ O_{\mathcal{P}}(n^{-1/2}),\end{aligned}$$

by using the Schur complement of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}}$ in $\widehat{\boldsymbol{\Sigma}}^*$. The Theorem follows. □

Corollary B.1. Let $\boldsymbol{c}_i = \bar{\pi}^{-1}\pi_i$. Suppose that the regularity conditions (3.28)–(3.33) hold for $\boldsymbol{c}^*_i(\boldsymbol{\psi}_N)$ defined by expression (3.12), under an unequal probability (with replacement) sampling design, with a single stratum. Then the expression (3.35) holds.

Proof of Corollary B.1. When we have a single stratum, we have $\boldsymbol{z}_i = \pi_i$. This implies $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}} = n$, $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zg}} = \sum_{i\in s}\breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)^{\mathrm{T}}$ and $\widehat{\boldsymbol{A}}_{\boldsymbol{g|z}} = \widehat{\boldsymbol{A}}_{\boldsymbol{g}}$. Thus the expression (B.46) reduces to expression (3.39). Hence, the result (3.35) holds, when we have a single stratum. □

Theorem B.2. Let $\boldsymbol{c}_i = (\bar{\pi}^{-1}\boldsymbol{z}_i{}^{\mathrm{T}}, \mathbf{f}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}}$, where the $\boldsymbol{z}_i$ are the stratification variables defined by (3.41) and the $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ are defined in Section 3.9. Suppose that the regularity conditions (3.28)–(3.33) hold for $\boldsymbol{c}_i^*(\boldsymbol{\psi}_N)$ defined by expression (3.12), under an unequal probability (with replacement) stratified sampling design. We have that

$$\widehat{r}(\boldsymbol{\theta}_N \mid \boldsymbol{\varphi}_N) = \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z})^{\mathrm{T}}(\mathbf{I}_b - \widehat{\boldsymbol{A}}^{\circ}_{\boldsymbol{g}|\boldsymbol{z}})\widehat{\boldsymbol{V}}^{\circ -1}_{\boldsymbol{gg}|\boldsymbol{z}}\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z}) + O_{\mathcal{P}}(n^{-1/2}), \quad \text{(B.49)}$$

where $\widehat{r}(\boldsymbol{\theta}_N \mid \boldsymbol{\varphi}_N)$ is defined by expression (3.51) and

$$\widehat{\boldsymbol{V}}^{\circ}_{\boldsymbol{gg}|\boldsymbol{z}} = \widehat{\boldsymbol{\Sigma}}^{\circ}_{\boldsymbol{gg}} - \widehat{\boldsymbol{\Sigma}}^{\circ\mathrm{T}}_{\boldsymbol{zg}}\widehat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{zz}}\widehat{\boldsymbol{\Sigma}}^{\circ}_{\boldsymbol{zg}}, \quad \text{(B.50)}$$

where $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}}$ is defined by (B.47),

$$\begin{aligned}
\widehat{\boldsymbol{\Sigma}}^{\circ}_{\boldsymbol{gg}} &= \sum_{i\in s} \breve{\boldsymbol{g}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)\breve{\boldsymbol{g}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)^{\mathrm{T}}, \\
\widehat{\boldsymbol{\Sigma}}^{\circ}_{\boldsymbol{zg}} &= \frac{1}{\bar{\pi}}\sum_{i\in s} \breve{\boldsymbol{z}}_i\ \breve{\boldsymbol{g}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)^{\mathrm{T}}, \\
\breve{\boldsymbol{g}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N) &= \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N) - \widehat{\boldsymbol{B}}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z})^{\mathrm{T}}\breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N), \quad \text{(B.51)} \\
\widehat{\boldsymbol{B}}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z}) &= \widehat{\boldsymbol{V}}^{-1}_{\mathbf{ff}|\boldsymbol{z}}\widehat{\boldsymbol{V}}_{\mathbf{f}\boldsymbol{g}|\boldsymbol{z}}, \\
\widehat{\boldsymbol{V}}_{\mathbf{ff}|\boldsymbol{z}} &= \widehat{\boldsymbol{\Sigma}}_{\mathbf{ff}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}\mathbf{f}}{}^{\mathrm{T}}\ \widehat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{zz}}\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}\mathbf{f}}, \quad \text{(B.52)} \\
\widehat{\boldsymbol{V}}_{\mathbf{f}\boldsymbol{g}|\boldsymbol{z}} &= \widehat{\boldsymbol{\Sigma}}_{\mathbf{f}\boldsymbol{g}} - \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}\mathbf{f}}{}^{\mathrm{T}}\ \widehat{\boldsymbol{\Sigma}}^{-1}_{\boldsymbol{zz}}\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zg}}, \\
\widehat{\boldsymbol{\Sigma}}_{\mathbf{ff}} &= \sum_{i\in s} \breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)\ \breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}}, \\
\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}\mathbf{f}} &= \sum_{i\in s} \breve{\boldsymbol{z}}_i\ \breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}}, \\
\widehat{\boldsymbol{\Sigma}}_{\mathbf{f}\boldsymbol{g}} &= \sum_{i\in s} \breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)\ \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)^{\mathrm{T}}, \\
\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zg}} &= \frac{1}{\bar{\pi}}\sum_{i\in s} \breve{\boldsymbol{z}}_i\breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N)^{\mathrm{T}}, \\
\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z}) &= \widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}_N) - \widehat{\boldsymbol{B}}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z})^{\mathrm{T}}\ \widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N), \\
\widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N) &= \sum_{i\in s} \breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N), \\
\widehat{\boldsymbol{G}}_{\pi}(\boldsymbol{\psi}_N) &= \sum_{i\in s} \breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N), \\
\widehat{\boldsymbol{A}}^{\circ}_{\boldsymbol{g}|\boldsymbol{z}} &= \widehat{\boldsymbol{V}}^{\circ -1/2}_{\boldsymbol{gg}|\boldsymbol{z}}\widehat{\boldsymbol{\nabla}}^{\circ}_{\boldsymbol{G}}\left(\widehat{\boldsymbol{\nabla}}^{\circ\mathrm{T}}_{\boldsymbol{G}}\widehat{\boldsymbol{V}}^{\circ -1}_{\boldsymbol{gg}|\boldsymbol{z}}\widehat{\boldsymbol{\nabla}}^{\circ}_{\boldsymbol{G}}\right)^{-1}\widehat{\boldsymbol{\nabla}}^{\circ\mathrm{T}}_{\boldsymbol{G}}\widehat{\boldsymbol{V}}^{\circ -1/2}_{\boldsymbol{gg}|\boldsymbol{z}}, \\
\widehat{\boldsymbol{\nabla}}^{\circ}_{\boldsymbol{G}} &= \sum_{i\in s} \left.\frac{\partial\breve{\boldsymbol{g}}^{\circ}_i(\boldsymbol{\psi}, \boldsymbol{\varphi}_N)}{\partial\boldsymbol{\nu}}\right|_{\boldsymbol{\psi}=\boldsymbol{\psi}_N}.
\end{aligned}$$

Here, $\breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N) = \mathbf{f}_i(\boldsymbol{\varphi}_N)\pi_i^{-1}$, $\breve{\boldsymbol{g}}_i(\boldsymbol{\psi}_N) = \boldsymbol{g}_i(\boldsymbol{\psi}_N)\pi_i^{-1}$.

Proof of Theorem B.2. The regularity conditions (3.29)–(3.33) imply

$$\begin{aligned} N^{-1}\|\widehat{\boldsymbol{C}}_\pi - \boldsymbol{C}\| &= O_{\mathcal{P}}(n^{-1/2}), && \text{(B.53)}\\ \max_{i\in s}\|\boldsymbol{c}_i\| &= o_{\mathcal{P}}(n^{1/2}), && \text{(B.54)}\\ \|\widehat{\boldsymbol{S}}\| &= O_{\mathcal{P}}(1), && \text{(B.55)}\\ \|\widehat{\boldsymbol{S}}^{-1}\| &= O_{\mathcal{P}}(1), && \text{(B.56)}\\ \frac{\bar{\pi}^\tau}{n}\sum_{i\in s}\|\breve{\boldsymbol{c}}_i\|^\tau &= O_{\mathcal{P}}(1), \quad \text{with } \tau = 2, 3 \text{ and } 4, && \text{(B.57)} \end{aligned}$$

where $\breve{\boldsymbol{c}}_i = \boldsymbol{c}_i\pi_i^{-1}$ and

$$\begin{aligned} \widehat{\boldsymbol{C}}_\pi &= \sum_{i\in s}\breve{\boldsymbol{c}}_i,\\ \widehat{\boldsymbol{S}} &= -\frac{\bar{\pi}}{N}\sum_{i\in s}\breve{\boldsymbol{c}}_i\breve{\boldsymbol{c}}_i{}^{\mathrm{T}}. \end{aligned} \qquad \text{(B.58)}$$

Berger and De La Riva Torres (2016) showed that the conditions (3.28) and (B.53)–(B.57) imply

$$2\{\ell(\boldsymbol{\pi}) - \ell(\widehat{\boldsymbol{\psi}} \mid \boldsymbol{\varphi}_N)\} = (\widehat{\boldsymbol{C}}_\pi - \boldsymbol{C})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{cc}}^{-1}(\widehat{\boldsymbol{C}}_\pi - \boldsymbol{C}) + O_{\mathcal{P}}(n^{-1/2}), \qquad \text{(B.59)}$$

where

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{cc}} = \sum_{i\in s}\breve{\boldsymbol{c}}_i\breve{\boldsymbol{c}}_i{}^{\mathrm{T}} = \begin{pmatrix} \bar{\pi}^{-2}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}} & \bar{\pi}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}\mathbf{f}}\\ \bar{\pi}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{z}\mathbf{f}}{}^{\mathrm{T}} & \widehat{\boldsymbol{\Sigma}}_{\mathbf{ff}} \end{pmatrix}.$$

As $\widehat{\boldsymbol{C}}_\pi - \boldsymbol{C} = (\mathbf{0}_H{}^{\mathrm{T}}, \widehat{\mathbf{f}}_\pi(\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}}$, the expression (B.59) gives

$$2\{\ell(\boldsymbol{\pi}) - \ell(\widehat{\boldsymbol{\psi}} \mid \boldsymbol{\varphi}_N)\} = (\mathbf{0}_H{}^{\mathrm{T}}, \widehat{\mathbf{f}}_\pi(\boldsymbol{\varphi}_N)^{\mathrm{T}})\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{cc}}^{-1}(\mathbf{0}_H{}^{\mathrm{T}}, \widehat{\mathbf{f}}_\pi(\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}} + O_{\mathcal{P}}(n^{-1/2}),$$

which implies that

$$2\{\ell(\boldsymbol{\pi}) - \ell(\widehat{\boldsymbol{\psi}} \mid \boldsymbol{\varphi}_N)\} = \widehat{\mathbf{f}}_\pi(\boldsymbol{\varphi}_N)^{\mathrm{T}}\widehat{\boldsymbol{V}}_{\mathbf{f}|\boldsymbol{z}}^{-1}\,\widehat{\mathbf{f}}_\pi(\boldsymbol{\varphi}_N) + O_{\mathcal{P}}(n^{-1/2}), \qquad \text{(B.60)}$$

by using the Schur complement $\widehat{\boldsymbol{V}}_{\mathbf{f}|\boldsymbol{z}}$ of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}}$ in $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{cc}}$.

Lemma B.3 gives

$$2\{\ell(\boldsymbol{\pi}) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}}\ell(\boldsymbol{\theta}_N, \boldsymbol{\nu} \mid \boldsymbol{\varphi}_N)\} = (\widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^*)^{\mathrm{T}}(\mathbf{I}_{r+b} - \widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}(\widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^*) + O_{\mathcal{P}}(n^{-1/2}), \qquad \text{(B.61)}$$

where $\widehat{\boldsymbol{A}}^*$ and $\widehat{\boldsymbol{\Sigma}}^*$ are respectively defined by expressions (B.10) and (B.11). We have

$$\widehat{\boldsymbol{C}}_\pi^*(\boldsymbol{\psi}_N) - \boldsymbol{C}^* = (\mathbf{0}_H{}^{\mathrm{T}}, \widehat{\boldsymbol{W}}_\pi{}^{\mathrm{T}})^{\mathrm{T}}, \qquad \text{(B.62)}$$

where

$$\begin{aligned}\widehat{\boldsymbol{W}}_\pi &= \sum_{i\in s}\breve{\boldsymbol{w}}_i(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N) = (\widehat{\mathbf{f}}_\pi(\boldsymbol{\varphi}_N)^{\mathrm{T}},\widehat{\boldsymbol{G}}_\pi(\boldsymbol{\psi}_N)^{\mathrm{T}})^{\mathrm{T}},\\ \breve{\boldsymbol{w}}_i(\boldsymbol{\psi},\boldsymbol{\varphi}_N) &= (\breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}},\breve{\boldsymbol{g}}_i(\boldsymbol{\psi})^{\mathrm{T}})^{\mathrm{T}}.\end{aligned}$$

Thus expressions (B.61) and (B.62) imply that

$$2\{\ell(\boldsymbol{\pi})-\max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}}\ell(\boldsymbol{\theta}_N,\boldsymbol{\nu}\mid\boldsymbol{\varphi}_N)\} = (\mathbf{0}_H^{\mathrm{T}},\widehat{\boldsymbol{W}}_\pi^{\mathrm{T}})^{\mathrm{T}}(\mathbf{I}_{r+b}-\widehat{\boldsymbol{A}}^*)\widehat{\boldsymbol{\Sigma}}^{*-1}(\mathbf{0}_H^{\mathrm{T}},\widehat{\boldsymbol{W}}_\pi^{\mathrm{T}})+O_{\mathcal{P}}(n^{-1/2}), \tag{B.63}$$

which reduces to

$$2\{\ell(\boldsymbol{\pi})-\max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}}\ell(\boldsymbol{\theta}_N,\boldsymbol{\nu}\mid\boldsymbol{\varphi}_N)\} = \widehat{\boldsymbol{W}}_\pi^{\mathrm{T}}(\mathbf{I}_{r-H+b}-\widehat{\boldsymbol{A}}_{\boldsymbol{w}})\widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}}^{-1}\widehat{\boldsymbol{W}}_\pi+O_{\mathcal{P}}(n^{-1/2}). \tag{B.64}$$

Here, $\widehat{\boldsymbol{A}}_{\boldsymbol{w}}$ is a symmetric and idempotent matrix defined by

$$\widehat{\boldsymbol{A}}_{\boldsymbol{w}} = \widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}}^{-1/2}\,\widehat{\boldsymbol{\nabla}}_{\boldsymbol{W}}(\widehat{\boldsymbol{\nabla}}_{\boldsymbol{W}}^{\mathrm{T}}\,\widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}}^{-1}\,\widehat{\boldsymbol{\nabla}}_{\boldsymbol{W}})^{-1}\,\widehat{\boldsymbol{\nabla}}_{\boldsymbol{W}}^{\mathrm{T}}\,\widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}}^{-1/2}, \tag{B.65}$$

with

$$\begin{aligned}\widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}} &= \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ww}}-\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zw}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zw}}, \qquad (\text{B.66})\\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ww}} &= \sum_{i\in s}\breve{\boldsymbol{w}}_i(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N)\breve{\boldsymbol{w}}_i(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N)^{\mathrm{T}},\\ \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zw}} &= \sum_{i\in s}\breve{\boldsymbol{z}}_i\;\breve{\boldsymbol{w}}_i(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N)^{\mathrm{T}},\\ \widehat{\boldsymbol{\nabla}}_{\boldsymbol{W}} &= \sum_{i\in s}\frac{\partial\breve{\boldsymbol{w}}_i(\boldsymbol{\psi},\boldsymbol{\varphi}_N)}{\partial\boldsymbol{\nu}}\bigg|_{\boldsymbol{\psi}=\boldsymbol{\psi}_N},\end{aligned}$$

It can be shown that expression (B.64) is equivalent to

$$2\{\ell(\boldsymbol{\pi})-\max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}}\ell(\boldsymbol{\theta}_N,\boldsymbol{\nu}\mid\boldsymbol{\varphi}_N)\} = \widehat{\boldsymbol{W}}_\pi^{\circ\mathrm{T}}(\mathbf{I}_{r-H+b}-\widehat{\boldsymbol{A}}_{\boldsymbol{w}}^{\circ})\widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}}^{\circ-1}\widehat{\boldsymbol{W}}_\pi^{\circ}+O_{\mathcal{P}}(n^{-1/2}), \tag{B.67}$$

where

$$\begin{aligned}\widehat{\boldsymbol{W}}_\pi^{\circ} &= \sum_{i\in s}\breve{\boldsymbol{w}}_i^{\circ}(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N) = (\widehat{\mathbf{f}}_\pi(\boldsymbol{\varphi}_N)^{\mathrm{T}},\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N,\boldsymbol{z})^{\mathrm{T}})^{\mathrm{T}},\\ \breve{\boldsymbol{w}}_i^{\circ}(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N) &= (\breve{\mathbf{f}}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}},\breve{\boldsymbol{g}}_i^{\circ}(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}}. \qquad (\text{B.68})\end{aligned}$$

The matrix $\widehat{\boldsymbol{A}}_{\boldsymbol{w}}^{\circ}$ is defined by expression (B.65), after substituting $\boldsymbol{w}_i(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N)$ by $\boldsymbol{w}_i^{\circ}(\boldsymbol{\psi}_N,\boldsymbol{\varphi}_N)$, and $\widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}}$ by $\widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}}^{\circ}$ defined by

$$\widehat{\boldsymbol{V}}_{\boldsymbol{ww}|\boldsymbol{z}}^{\circ} = \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{ww}}^{\circ}-\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zw}}^{\circ\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zz}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{zw}}^{\circ},$$

where

$$\begin{aligned}\widehat{\boldsymbol{\Sigma}}^{\circ}_{\boldsymbol{ww}} &= \sum_{i\in s} \breve{\boldsymbol{w}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)\ \breve{\boldsymbol{w}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)^{\mathrm{T}},\\ \widehat{\boldsymbol{\Sigma}}^{\circ}_{\boldsymbol{zw}} &= \sum_{i\in s} \breve{\boldsymbol{z}}_i\ \breve{\boldsymbol{w}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)^{\mathrm{T}}.\end{aligned}$$

As $\breve{\boldsymbol{g}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)$ and $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ are orthogonal, $\widehat{\boldsymbol{V}}^{\circ}_{\boldsymbol{ww}|\boldsymbol{z}}$ reduces to a block-diagonal matrix; that is,

$$\widehat{\boldsymbol{V}}^{\circ}_{\boldsymbol{ww}|\boldsymbol{z}} = \begin{pmatrix} \widehat{\boldsymbol{V}}_{\mathbf{ff}|\boldsymbol{z}} & \mathbf{0} \\ \mathbf{0} & \widehat{\boldsymbol{V}}^{\circ}_{\boldsymbol{gg}|\boldsymbol{z}} \end{pmatrix}, \tag{B.69}$$

where $\widehat{\boldsymbol{V}}_{\mathbf{ff}|\boldsymbol{z}}$ is defined by expression (B.52) and $\widehat{\boldsymbol{V}}^{\circ}_{\boldsymbol{gg}|\boldsymbol{z}}$ is defined by expression (B.50). Hence, by substituting the expressions (B.68) and (B.69) into (B.67), we obtain

$$\begin{aligned}2\{\ell(\boldsymbol{\pi}) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}_N, \boldsymbol{\nu} \mid \boldsymbol{\varphi}_N)\} &= \widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N)^{\mathrm{T}} \widehat{\boldsymbol{V}}^{-1}_{\mathbf{ff}|\boldsymbol{z}}\ \widehat{\mathbf{f}}_{\pi}(\boldsymbol{\varphi}_N) + \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z})^{\mathrm{T}}(\mathbf{I}_b - \widehat{\boldsymbol{A}}^{\circ}_{\boldsymbol{g}|\boldsymbol{z}})\\ &\quad \widehat{\boldsymbol{V}}^{\circ -1}_{\boldsymbol{gg}|\boldsymbol{z}} \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z}) + O_{\mathcal{P}}(n^{-1/2}),\end{aligned} \tag{B.70}$$

The definition (3.51) implies

$$\widehat{r}(\boldsymbol{\theta}_N \mid \boldsymbol{\varphi}_N) = 2\left\{\ell(\boldsymbol{\pi}) - \max_{\boldsymbol{\nu}\in\boldsymbol{\Lambda}} \ell(\boldsymbol{\theta}_N, \boldsymbol{\nu} \mid \boldsymbol{\varphi}_N)\right\} - 2\left\{\ell(\boldsymbol{\pi}) - \ell(\widehat{\boldsymbol{\psi}} \mid \boldsymbol{\varphi}_N)\right\} \tag{B.71}$$

Finally, by substituting the expressions (B.60) and (B.70) into (B.71), we obtain (B.49). The Theorem follows. □

Corollary B.2. Let $\boldsymbol{c}_i = (\bar{\pi}^{-1}\pi_i, \mathbf{f}_i(\boldsymbol{\varphi}_N)^{\mathrm{T}})^{\mathrm{T}}$, where the $\mathbf{f}_i(\boldsymbol{\varphi}_N)$ are defined in Section 3.9. Suppose that the regularity conditions (3.28)–(3.33) hold for $\boldsymbol{c}^*_i(\boldsymbol{\psi}_N)$ defined by expression (3.12), under an unequal probability (with replacement) sampling design, with a single stratum. We have that the expression (3.52) holds.

Proof of Corollary B.2. With a single stratum, we have $\boldsymbol{z}_i = \pi_i$. It can be shown that this implies $\widehat{\boldsymbol{B}}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z}) = \widehat{\boldsymbol{B}}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)$, $\widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N, \boldsymbol{z}) = \widehat{\boldsymbol{G}}_{reg}(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)$, $\widehat{\boldsymbol{A}}^{\circ}_{\boldsymbol{g}|\boldsymbol{z}} = \widehat{\boldsymbol{A}}^{\bullet}_{\boldsymbol{g}}$, $\breve{\boldsymbol{g}}^{\circ}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N) = \breve{\boldsymbol{g}}^{\bullet}_i(\boldsymbol{\psi}_N, \boldsymbol{\varphi}_N)$ and $\widehat{\boldsymbol{V}}^{\circ}_{\boldsymbol{gg}|\boldsymbol{z}} = \widehat{\boldsymbol{V}}^{\bullet}_{\boldsymbol{gg}}$. Hence, the expression (B.49) reduces to expression (3.52). The Corollary follows. □

B.3 R code for the second paper

By Melike Oguz-Alper
University of Southampton, SO17 1BJ, Southampton, U.K.
M.OguzAlper@soton.ac.uk

```
#-----------------------------------------------------------------
# Clear the workspace
#-----------------------------------------------------------------
rm(list=ls(all=TRUE))
#-----------------------------------------------------------------
# Load packages required
#-----------------------------------------------------------------
library(rootSolve)
library(MASS)
library(lpSolve)
library(sampling)
library(minpack.lm)
#-----------------------------------------------------------------
# FUNCTIONS FOR LINEAR REGRESSION
#-----------------------------------------------------------------
# Estimating functions for linear regression parameters
#=================================================================
# DEFINITIONS OF THE PARAMETERS USED IN THE FOLLOWING FUNCTIONS
# (unless otherwise stated)
#-----------------------------------------------------------------
# p:factor of heteroskedasticity
# Constant residual variance when p=1
# Theta: a given value of Theta
# Lambda: a given value of Lambda
# Here, Lambda: is the nuisance (intercept); Theta: is the parameter
# of interest (slope)
# n: sample size
# xi.Sample: nx1 vector of covariate
# yi.Sample: nx1 vector of response variable
# W: design weights
# Inc.Prob: nx1 vector of inclusion probabilities
# LambdaDot: a given initial value of Lambda
# Tol1: tolerance value for the difference between lower and upper bounds
# Tol2: tolerance value for the difference between the ELLR value and
# the table value of the chi-squared distribution
# a: significance level
# Pr: first-order inclusion probabilities
# f: sum(Inclusion.Probabilities^2)/n: Here, summation is taken over the
# population.
# Gamma: known population value (population mean)
#=================================================================
Fun.EF.Linear<-function(Theta,Lambda,p)
{
Wi<-xi.Sample^(-p)
gi.Lambda<-(yi.Sample-Lambda-Theta*xi.Sample)*Wi
gi.Theta<-xi.Sample*(yi.Sample-Lambda-Theta*xi.Sample)*Wi
matrix(c(gi.Lambda,gi.Theta),nrow=length(yi.Sample),ncol=2,byrow=FALSE)
}
#-----------------------------------------------------------------
# Estimating equations for linear regression parameters
#=================================================================
```

```
# W: nx1 vector of weights
#-----------------------------------------------------------------
Fun.EE.Linear<-function(X,W,p)
{
gi<-Fun.EF.Linear(X[2],X[1],p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]
Fun1<-sum(gi.Lambda*W)
Fun2<-sum(gi.Theta*W)
c(Fun1,Fun2)
}
#-------------------------------------------------------------------------
# Function for solving Lambda for a given value of Theta (in linear regression)
#=========================================================================
Fun.LHat.Linear<-function(X,Theta,W,p)
{
gi<-Fun.EF.Linear(Theta,X,p)
gi.Lambda<-gi[,1]
Fun<-sum(gi.Lambda*W)
c(Fun)
}
#-----------------------------------------------------------------
# FUNCTIONS FOR EMPIRICAL LIKELIHOOD
#=================================================================
#-------------------------------------------------------------------------
# Function for solving lagrange coefficients for given values of parameters
#=========================================================================
Fun.Eta<-function(Eta,Theta,Lambda,p)
{
gi<-Fun.EF.Linear(Theta,Lambda,p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]

Fun1<-sum(Inc.Prob/(Inc.Prob+Eta[1]*Inc.Prob+Eta[2]*gi.Lambda+Eta[3]*gi.Theta))
-length(Inc.Prob)
Fun2<-sum(gi.Lambda/(Inc.Prob+Eta[1]*Inc.Prob+Eta[2]*gi.Lambda+Eta[3]*gi.Theta))
Fun3<-sum(gi.Theta/(Inc.Prob+Eta[1]*Inc.Prob+Eta[2]*gi.Lambda+Eta[3]*gi.Theta))
c(Fun1,Fun2,Fun3)
}
#-------------------------------------------------------------------------
# Function for finding the value of the empirical log-likelihood ratio (ELLR)
# funtion when all parameters are given
#=========================================================================
ELLR<-function(Theta,Lambda,p)
{
Roots.Eta<-nls.lm(par=c(0,0,0),lower=NULL,upper=NULL,fn=Fun.Eta,jac=NULL,Theta=
Theta,Lambda=Lambda,p=p)$par

gi<-Fun.EF.Linear(Theta,Lambda,p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]

mi.Hat<-1/(Inc.Prob+Roots.Eta[1]*Inc.Prob+Roots.Eta[2]*gi.Lambda+Roots.Eta[3]*
gi.Theta)

if(range(mi.Hat)[1]<0)
{
s<-5e-05
a<-1
```

```
while(range(mi.Hat)[1]<0)
{
Roots.Eta<-nls.lm(par=c(a*s,a*s,a*s),lower=NULL,upper=NULL,fn=Fun.Eta,jac=NULL,
Theta=Theta,Lambda=Lambda,p=p)$par
mi.Hat<-1/(Inc.Prob+Roots.Eta[1]*Inc.Prob+Roots.Eta[2]*gi.Lambda+Roots.Eta[3]*
gi.Theta)
a<-a+1
}
}
ELLR.Function<-2*sum(log(mi)-log(mi.Hat))
return(c(ELLR.Function,Roots.Eta,range(mi.Hat)))
}
#---------------------------------------------------------------------------
# PROFILING
#===========================================================================
#---------------------------------------------------------------------------
# Function for solving lagrange coefficients for a given value of Theta
# Nuisance parameter Lambda is unknown (PROFILING)
#===========================================================================
Fun.X<-function(X,Theta,p)
{
gi<-Fun.EF.Linear(Theta,X[4],p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]

Wi<-xi.Sample^(-p)

Fun1<-sum(Inc.Prob/(Inc.Prob+X[1]*Inc.Prob+X[2]*gi.Lambda+X[3]*gi.Theta))-
length(Inc.Prob)
Fun2<-sum(gi.Lambda/(Inc.Prob+X[1]*Inc.Prob+X[2]*gi.Lambda+X[3]*gi.Theta))
Fun3<-sum(gi.Theta/(Inc.Prob+X[1]*Inc.Prob+X[2]*gi.Lambda+X[3]*gi.Theta))
Fun4<-sum((X[2]+X[3]*xi.Sample)*Wi/(Inc.Prob+X[1]*Inc.Prob+X[2]*gi.Lambda+X[3]*
gi.Theta))
c(Fun1,Fun2,Fun3,Fun4)
}
#---------------------------------------------------------------------------
# Function for finding the value of empirical log-likelihood ratio (ELLR)
# funtion for a given value of Theta
# Nuisance parameter Lambda is unknown (PROFILING)
#===========================================================================
# mi: design weights
#---------------------------------------------------------------------------
ELLR.Nuisance<-function(Theta,LambdaDot,p)
{
Roots.X<-nls.lm(par=c(0,0,0,LambdaDot),lower=NULL,upper=NULL,fn=Fun.X,jac=NULL,
Theta=Theta,p=p)$par

gi<-Fun.EF.Linear(Theta,Roots.X[4],p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]

mi.Hat<-1/(Inc.Prob+Roots.X[1]*Inc.Prob+Roots.X[2]*gi.Lambda+Roots.X[3]*gi.Theta)

if(range(mi.Hat)[1]<0)
{
s<-5e-05
a<-1
while(range(mi.Hat)[1]<0)
{
```

```
Roots.X<-nls.lm(par=c(a*s,a*s,a*s,LambdaDot),lower=NULL,upper=NULL,fn=Fun.X,
jac=NULL,Theta=Theta,p=p)$par
gi<-Fun.EF.Linear(Theta,Roots.X[4],p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]
mi.Hat<-1/(Inc.Prob+Roots.X[1]*Inc.Prob+Roots.X[2]*gi.Lambda+Roots.X[3]*gi.Theta)
a<-a+1
}
}
ELLR.Function<-2*sum(log(mi)-log(mi.Hat))
return(c(ELLR.Function,Roots.X[4],range(mi.Hat)))
}
#-------------------------------------------------------------------------
# EMPIRICAL LIKELIHOOD CONFIDENCE INTERVALS
#=========================================================================
# Theta: a given initial value of Theta
# Lambda: a given initial value of Lambda
#-------------------------------------------------------------------------
# Lower bound
#=========================================================================
EL.LowerBound<-function(Theta,Lambda,Tol1,Tol2,a,p)
{
k<-0.05
q<-qchisq(1-a,1)

Boundary.Lower<-Theta-2*k*abs(Theta)
Boundary.Upper<-Theta
Difference1<-abs(Boundary.Upper-Boundary.Lower)
Difference2<-1

while (Difference1>Tol1 | Difference2>Tol2 )
{
Thetadot<-(Boundary.Lower+Boundary.Upper)/2
Vector.ELLR<-ELLR.Nuisance(Thetadot,Lambda,p)
ELLR.Function<-Vector.ELLR[1]
if(ELLR.Function>q) Boundary.Lower<-Thetadot
if(ELLR.Function<=q) Boundary.Upper<-Thetadot
Difference1<-abs(Boundary.Upper-Boundary.Lower)
Difference2<-q-ELLR.Function
if(Difference1<=Tol1 & Difference2>Tol2) {k<-k+0.05 ; Boundary.Lower<-
Boundary.Lower-2*k*abs(Boundary.Lower)}
Difference1<-abs(Boundary.Upper-Boundary.Lower)
Lambda<-Vector.ELLR[2]
}
Lower.Bound<-(Boundary.Lower+Boundary.Upper)/2
return(Lower.Bound)
}
#-------------------------------------------------------------------------
# Upper bound
#=========================================================================
EL.UpperBound<-function(Theta,Lambda,Tol1,Tol2,a,p)
{
k<-0.05
q<-qchisq(1-a,1)

Boundary.Lower<-Theta
Boundary.Upper<-Theta+2*k*abs(Theta)
Difference1<-abs(Boundary.Upper-Boundary.Lower)
Difference2<-1
```

```
while (Difference1>Tol1 | Difference2>Tol2 )
{
Thetadot<-(Boundary.Lower+Boundary.Upper)/2
Vector.ELLR<-ELLR.Nuisance(Thetadot,Lambda,p)
ELLR.Function<-Vector.ELLR[1]
if(ELLR.Function>q) Boundary.Upper<-Thetadot
if(ELLR.Function<=q) Boundary.Lower<-Thetadot
Difference1<-abs(Boundary.Upper-Boundary.Lower)
Difference2<-q-ELLR.Function
if(Difference1<=Tol1 & Difference2>Tol2) {k<-k+0.05 ; Boundary.Upper<-
Boundary.Upper+2*k*abs(Boundary.Upper)}
Difference1<-abs(Boundary.Upper-Boundary.Lower)
Lambda<-Vector.ELLR[2]
}
Upper.Bound<-(Boundary.Lower+Boundary.Upper)/2
return(Upper.Bound)
}
#-------------------------------------------------------------------------
# FUNCTIONS FOR PSEUDO LIKELIHOOD METHODS
#=========================================================================
#-------------------------------------------------------------------------
# HARTLEY & RAO (1962) VARIANCE ESTIMATOR
#=========================================================================
# y: variable of interest
#-------------------------------------------------------------------------
Var.Sys<-function(y,n,Pr,f)
{
Total<-0
for(k in 1:n-1)
{
wi<-1/Pr
Inc.Prob.Trun<-Pr[(k+1):n]
mi.Trun<-wi[(k+1):n]
y.Trun<-y[(k+1):n]
Sum.k<-sum((1-Pr[k]-Inc.Prob.Trun+f)*(wi[k]*y[k]-mi.Trun*y.Trun)^2)
Total<-sum(Total+Sum.k)
}
return(Total/(n-1))
}
#-------------------------------------------------------------------------
# JACOBIAN FOR LINEAR REGRESSION PARAMETERS
#=========================================================================
Fun.Jacobian.Linear<-function(p)
{
Wi<-xi.Sample^(-p)
J1.i<--xi.Sample*Wi
J2.i<--Wi
matrix(c(J1.i,J2.i),nrow=length(xi.Sample),ncol=2,byrow=FALSE)
}
#-------------------------------------------------------------------------
# Variance estimation of new estimating function (gi.Star)
# (Pseudo likelihood 1)
#=========================================================================
# Theta: estimate of Theta
# Lambda: estimate of Lambda
#-------------------------------------------------------------------------
Varyans.Pseudo1<-function(Theta,Lambda,W,f,p)
{
```

```
gi<-Fun.EF.Linear(Theta,Lambda,p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]

J.i<-Fun.Jacobian.Linear(p)
J1.i<-J.i[,1]
J2.i<-J.i[,2]
J1<-sum(J1.i*W)
J2<-sum(J2.i*W)

gi.Star<-gi.Theta-J1/J2*gi.Lambda
Var.gi.Star<-Var.Sys(gi.Star,length(W),1/W,f)
return(Var.gi.Star)
}
#-------------------------------------------------------------------------
# PSEUDO LIKELIHOOD 1 CONFIDENCE INTERVALS
#=========================================================================
#-------------------------------------------------------------------------
# Lower bound (replace -1.96 by +1.96 for upper bound)
#=========================================================================
# Theta: estimate of Theta
# Lambda: estimate of Lambda
#-------------------------------------------------------------------------
Fun.X.LB.Pseudo1<-function(X,Theta,Lambda,W,f,p)
{
gi<-Fun.EF.Linear(X[2],X[1],p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]

Var.gi.Star<-Varyans.Pseudo1(Theta,Lambda,W,f,p)
Fun1<-sum(gi.Lambda*W)
Fun2<-sum(gi.Theta*W)/sqrt(Var.gi.Star)-1.96
c(Fun1,Fun2)
}
#-------------------------------------------------------------------------
# PSEUDO LIKELIHOOD 2 CONFIDENCE INTERVALS
#=========================================================================
#-------------------------------------------------------------------------
# Lower bound (replace -1.96 by +1.96 for upper bound)
#=========================================================================
Fun.X.LB.Pseudo2<-function(X,W,f,p)
{
gi<-Fun.EF.Linear(X[2],X[1],p)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]

J.i<-Fun.Jacobian.Linear(p)
J1.i<-J.i[,1]
J2.i<-J.i[,2]
J1<-sum(J1.i*W)
J2<-sum(J2.i*W)
gi.Star<-gi.Theta-J1/J2*gi.Lambda
Var.gi.Star<-Var.Sys(gi.Star,length(W),1/W,f)

Fun1<-sum(gi.Lambda*W)
Fun2<-sum(gi.Theta*W)/sqrt(Var.gi.Star)-1.96
c(Fun1,Fun2)
}
#-------------------------------------------------------------------------
```

```
# RESCALED BOOTSTRAP (Rao et al. 1992)
#=========================================================================
# m: bootstrap sample size
# B: number of bootstrap samples
#-------------------------------------------------------------------------
Rescaled.Boot<-function(m,n,W,B,p)
{
Vector.Boot.Theta<-NULL
for (k in 1:B)
{
Boot<-srswr(m,n)
mi.Boot<-((1-(m/(n-1))^0.5)+((m/(n-1))^0.5)*n/m*Boot)*W
Temp<-nls.lm(par=c(0,0),lower=NULL,upper=NULL,fn=Fun.EE.Linear,jac=NULL,
W=mi.Boot,p=p)$par[2]
Vector.Boot.Theta<-c(Vector.Boot.Theta,Temp)
}
return(Vector.Boot.Theta)
}
#-------------------------------------------------------------------------
#                         START EXAMPLES
#-------------------------------------------------------------------------
# APPLICATION OF THE FUNCTIONS PROVIDED ABOVE
# A PART OF THE MAIN CODE
# (Excludes generation of population and sample selection)
#=========================================================================
# Beta1.Pop: population value of the slope (OLS estimate)
#-------------------------------------------------------------------------
# Point estimation, ELLR value and EL confidence interval
#=========================================================================
Beta.Est<-nls.lm(par=c(0,0),lower=NULL,upper=NULL,fn=Fun.EE.Linear,jac=NULL,
W=W,p=p)$par
Beta0.Est<-Beta.Est[1]
Beta1.Est<-Beta.Est[2]
ELLR.Value<-ELLR.Nuisance(Beta1.Pop,Beta0.Est,p)[1]
LB.EL<-EL.LowerBound(Beta1.Est,Beta0.Est,1e-08,1e-06,0.05,p)
UB.EL<-EL.UpperBound(Beta1.Est,Beta0.Est,1e-08,1e-06,0.05,p)
#-------------------------------------------------------------------------
# Pseudo likelihood confidence intervals
#=========================================================================
LB.Initial<-Beta1.Est-2*0.05*abs(Beta1.Est)
UB.Initial<-Beta1.Est+2*0.05*abs(Beta1.Est)
LB.Pseudo1<-nls.lm(par=c(Beta0.Est,LB.Initial),lower=NULL,upper=NULL,fn=
Fun.X.LB.Pseudo1,jac=NULL,W=W,f=f,Theta=Beta1.Est,Lambda=Beta0.Est,p=p)$par[2]
UB.Pseudo1<-nls.lm(par=c(Beta0.Est,UB.Initial),lower=NULL,upper=NULL,fn=
Fun.X.UB.Pseudo1,jac=NULL,W=W,f=f,Theta=Beta1.Est,Lambda=Beta0.Est,p=p)$par[2]
LB.Pseudo2<-nls.lm(par=c(Beta0.Est,LB.Initial),lower=NULL,upper=NULL,fn=
Fun.X.LB.Pseudo2,jac=NULL,W=W,f=f,p=p)$par[2]
UB.Pseudo2<-nls.lm(par=c(Beta0.Est,UB.Initial),lower=NULL,upper=NULL,fn=
Fun.X.UB.Pseudo2,jac=NULL,W=W,f=f,p=p)$par[2]
#-------------------------------------------------------------------------
# Bootstrap confidence interval
#=========================================================================
Vector.Boot.Beta1<-Rescaled.Boot(n-1,n,W,1000,p=p)
CI.ResBoot<-Percentile.CI(Vector.Boot.Beta1,0.05)
#-------------------------------------------------------------------------
#                          END EXAMPLES
#-------------------------------------------------------------------------
#-------------------------------------------------------------------
# FUNCTIONS FOR LOGISTIC REGRESSION (WITH AUXILIARY INFORMATION)
```

```
#=================================================================
# Estimating functions for logistic regression parameters
#=================================================================
Fun.EF.Logistic<-function(Theta,Lambda)
{
gi.Lambda<-yi.Sample-exp(Lambda+Theta*xi.Sample)/(1+exp(Lambda+Theta*xi.Sample))
gi.Theta<-xi.Sample*(yi.Sample-exp(Lambda+Theta*xi.Sample)/(1+exp(Lambda+Theta*
xi.Sample)))
matrix(c(gi.Lambda,gi.Theta),nrow=length(yi.Sample),ncol=2,byrow=FALSE)
}
#-----------------------------------------------------------------
# Estimating equations for logistic regression parameters
#=================================================================
# W: nx1 vector of weights
#-----------------------------------------------------------------
Fun.EE.Logistic<-function(X,W)
{
gi<-Fun.EF.Logistic(X[2],X[1])
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]
Fun1<-sum(gi.Lambda*W)
Fun2<-sum(gi.Theta*W)
c(Fun1,Fun2)
}
#-------------------------------------------------------------------------
# Function for solving Lambda for a given value of Theta (in logistic regression)
#=========================================================================
# W: nx1 vector of weights
#-----------------------------------------------------------------
Fun.LHat.Logistic<-function(X,Theta,W)
{
gi<-Fun.EF.Logistic(Theta,X)
gi.Lambda<-gi[,1]
Fun<-sum(gi.Lambda*W)
c(Fun)
}
#-----------------------------------------------------------------
# Estimating function for known population mean
#=================================================================
# vi.Sample: sample values of the auxiliary variable V
#-----------------------------------------------------------------
Fun.EF.Aux<-function(Gamma)
{
gi.Gamma<-vi.Sample-Gamma
}
#-------------------------------------------------------------------------
# FUNCTIONS FOR EMPIRICAL LIKELIHOOD (WITH AUXILIARY INFORMATION)
#=========================================================================
#-------------------------------------------------------------------------
# Function for solving lagrange coefficients given population
# level information
#=========================================================================
Fun.Eta.Aux<-function(Eta,Gamma)
{
gi.Gamma<-Fun.EF.Aux(Gamma)
Fun1<-sum(Inc.Prob/(Inc.Prob+Eta[1]*Inc.Prob+Eta[2]*gi.Gamma))-length(Inc.Prob)
Fun2<-sum(gi.Gamma/(Inc.Prob+Eta[1]*Inc.Prob+Eta[2]*gi.Gamma))
c(Fun1,Fun2)
}
```

```
#-------------------------------------------------------------------------
# EL weights in the presence of population level information
#=========================================================================
EL.Calib.Weight<-function(Gamma)
{
Roots.Eta<-nls.lm(par=c(0,0),lower=NULL,upper=NULL,fn=Fun.Eta.Aux,jac=NULL,
Gamma=Gamma)$par
gi.Gamma<-Fun.EF.Aux(Gamma)
mi.Hat<-1/(Inc.Prob+Roots.Eta[1]*Inc.Prob+Roots.Eta[2]*gi.Gamma)
if(range(mi.Hat)[1]<0)
{
s<-5e-05
a<-1
while(range(mi.Hat)[1]<0)
{
Roots.Eta<-nls.lm(par=c(a*s,a*s),lower=NULL,upper=NULL,fn=Fun.Eta.Aux,jac=NULL,
Gamma=Gamma)$par
mi.Hat<-1/(Inc.Prob+Roots.Eta[1]*Inc.Prob+Roots.Eta[2]*gi.Gamma)
a<-a+1
}
}
return(mi.Hat)
}
#-------------------------------------------------------------------------
# PROFILING
#=========================================================================
#-------------------------------------------------------------------------
# Function for solving lagrange coefficients for given value of Theta and
# population level information Gamma
# Nuisance parameter Lambda is unknown (PROFILING)
#=========================================================================
Fun.X<-function(X,Theta,Gamma)
{
gi<-Fun.EF.Logistic(Theta,X[5])
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]
gi.Gamma<-Fun.EF.Aux(Gamma)
Fun1<-sum(Inc.Prob/(Inc.Prob+X[1]*Inc.Prob+X[2]*gi.Lambda+X[3]*gi.Theta+X[4]*
gi.Gamma))-length(Inc.Prob)
Fun2<-sum(gi.Lambda/(Inc.Prob+X[1]*Inc.Prob+X[2]*gi.Lambda+X[3]*gi.Theta+X[4]*
gi.Gamma))
Fun3<-sum(gi.Theta/(Inc.Prob+X[1]*Inc.Prob+X[2]*gi.Lambda+X[3]*gi.Theta+X[4]*
gi.Gamma))
Fun4<-sum(gi.Gamma/(Inc.Prob+X[1]*Inc.Prob+X[2]*gi.Lambda+X[3]*gi.Theta+X[4]*
gi.Gamma))
Fun5<-sum((X[2]+X[3]*xi.Sample)*(exp(X[5]+Theta*xi.Sample)/(1+exp(X[5]+Theta*
xi.Sample))/(1+exp(X[5]+Theta*xi.Sample)))/(Inc.Prob+X[1]*Inc.Prob+X[2]*
gi.Lambda+X[3]*gi.Theta+X[4]*gi.Gamma))
c(Fun1,Fun2,Fun3,Fun4,Fun5)
}
#-------------------------------------------------------------------------
# Function for finding the value of the empirical log-likelihood ratio (ELLR)
# funtion for a given value of Theta and population level information Gamma
# Nuisance parameter Lambda is unknown (PROFILING)
#=========================================================================
# mi.Star: EL weights in the presence of auxiliary information
# and computed by using function 'EL.Calib.Weight'
#-------------------------------------------------------------------------
ELLR.Nuisance<-function(Theta,LambdaDot,Gamma)
```

```
{
Roots.X<-nls.lm(par=c(0,0,0,0,LambdaDot),lower=NULL,upper=NULL,fn=Fun.X,jac=
NULL,Theta=Theta,Gamma=Gamma)$par

gi<-Fun.EF.Logistic(Theta,Roots.X[5])
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]
gi.Gamma<-Fun.EF.Aux(Gamma)

mi.Hat<-1/(Inc.Prob+Roots.X[1]*Inc.Prob+Roots.X[2]*gi.Lambda+Roots.X[3]*gi.Theta+
Roots.X[4]*gi.Gamma)

if(range(mi.Hat)[1]<0)
{
s<-5e-05
a<-1
while(range(mi.Hat)[1]<0)
{
Roots.X<-nls.lm(par=c(a*s,a*s,a*s,a*s,LambdaDot),lower=NULL,upper=NULL,fn=Fun.X,
jac=NULL,Theta=Theta,Gamma=Gamma)$par
gi<-Fun.EF.Logistic(Theta,Roots.X[5])
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]
gi.Gamma<-Fun.EF.Aux(Gamma)
mi.Hat<-1/(Inc.Prob+Roots.X[1]*Inc.Prob+Roots.X[2]*gi.Lambda+Roots.X[3]*gi.Theta+
Roots.X[4]*gi.Gamma)
a<-a+1
}
}
ELLR.Function<-2*sum(log(mi.Star)-log(mi.Hat))
return(c(ELLR.Function,Roots.X[5]))
}
#-----------------------------------------------------------------------------
# FUNCTIONS FOR PSEUDO LIKELIHOOD METHODS UNDER POPULATION LEVEL INFORMATION
#=============================================================================
#-----------------------------------------------------------------------------
# CALIBRATION WEIGHTS
#=============================================================================
# W: design weights or bootstrap weights
#-----------------------------------------------------------------------------
Calib.Weight<-function(Gamma,W)
{
gi.Gamma<-Fun.EF.Aux(Gamma)
L<--sum(gi.Gamma*W)/sum(gi.Gamma*gi.Gamma*W)
W.Calib<-(1+gi.Gamma*L)*W
return(W.Calib)
}
#-----------------------------------------------------------------------------
# HARTLEY & RAO (1962) VARIANCE ESTIMATOR under population level information
#=============================================================================
# y: variable of interest
# W: calibration weights (use function 'Calib.Weight' provided  above)
#-----------------------------------------------------------------------------
Var.Sys.Aux<-function(y,n,Pr,W,f)
{
Total<-0
for(k in 1:n-1)
{
wi<-W
```

```
Inc.Prob.Trun<-Pr[(k+1):n]
wi.Trun<-wi[(k+1):n]
y.Trun<-y[(k+1):n]
Sum.k<-sum((1-Pr[k]-Inc.Prob.Trun+f)*(wi[k]*y[k]-wi.Trun*y.Trun)^2)
Total<-sum(Total+Sum.k)
}
return(Total/(n-1))
}
#-------------------------------------------------------------------------
# JACOBIAN FOR LINEAR REGRESSION PARAMETERS
#=========================================================================
Fun.Jacobian.Logistic<-function(Theta,Lambda)
{
J1.i<--xi.Sample*exp(Lambda+Theta*xi.Sample)/(1+exp(Lambda+Theta*xi.Sample))/
(1+exp(Lambda+Theta*xi.Sample))
J2.i<--exp(Lambda+Theta*xi.Sample)/(1+exp(Lambda+Theta*xi.Sample))/(1+exp(Lambda+
Theta*xi.Sample))
matrix(c(J1.i,J2.i),nrow=length(xi.Sample),ncol=2,byrow=FALSE)
}
#-------------------------------------------------------------------------
# Variance estimation of new estimating function (gi.Star)
# (Pseudo likelihood 1) in the presence of auxiliary information
#=========================================================================
# Theta: estimate of Theta
# Lambda: estimate of Lambda
# W: calibration weights
#-------------------------------------------------------------------------
Varyans.Pseudo1<-function(Theta,Lambda,Gamma,W,Pr,f)
{
gi<-Fun.EF.Logistic(Theta,Lambda)
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]
gi.Gamma<-Fun.EF.Aux(Gamma)

J.i<-Fun.Jacobian.Logistic(Theta,Lambda)
J1.i<-J.i[,1]
J2.i<-J.i[,2]
J1<-sum(J1.i*W)
J2<-sum(J2.i*W)

gi.Star<-gi.Theta-J1/J2*gi.Lambda
Beta<-sum(gi.Star*gi.Gamma*W)/sum(gi.Gamma*gi.Gamma*W)
ei<-gi.Star-Beta*gi.Gamma
Var.ei<-Var.Sys.Aux(ei,length(Pr),Pr,W,f)
return(Var.ei)
}
#-------------------------------------------------------------------------
# PSEUDO LIKELIHOOD 2 CONFIDENCE INTERVALS under POPULATION LEVEL INFORMATION
#=========================================================================
#-------------------------------------------------------------------------
# Lower bound (replace -1.96 by +1.96 for upper bound)
#=========================================================================
# W: calibration weights
#-------------------------------------------------------------------------
Fun.X.LB.Pseudo2<-function(X,Gamma,W,Pr,f)
{
gi<-Fun.EF.Logistic(X[2],X[1])
gi.Lambda<-gi[,1]
gi.Theta<-gi[,2]
```

```
gi.Gamma<-Fun.EF.Aux(Gamma)

J.i<-Fun.Jacobian.Logistic(X[2],X[1])
J1.i<-J.i[,1]
J2.i<-J.i[,2]
J1<-sum(J1.i*W)
J2<-sum(J2.i*W)

gi.Star<-gi.Theta-J1/J2*gi.Lambda
Beta<-sum(gi.Star*gi.Gamma*W)/sum(gi.Gamma*gi.Gamma*W)
ei<-gi.Star-Beta*gi.Gamma
Var.ei<-Var.Sys.W(ei,length(Pr),Pr,W,f)

Fun1<-sum(gi.Lambda*W)
Fun2<-sum(gi.Theta*W)/sqrt(Var.ei)-1.96
c(Fun1,Fun2)
}
#-----------------------------------------------------------------------------
# RESCALED BOOTSTRAP (Rao et al. 1992) under POPULATION LEVEL INFORMATION
#=============================================================================
# m: bootstrap sample size
# B: number of bootstrap samples
#-----------------------------------------------------------------------------
Rescaled.Boot<-function(m,n,Gamma,W,B)
{
Vector.Boot.Theta<-NULL
for (k in 1:B)
{
Boot<-srswr(m,n)
mi.Boot<-((1-(m/(n-1))^0.5)+((m/(n-1))^0.5)*n/m*Boot)*W
mi.Boot.Calib<-Calib.Weight(Gamma,mi.Boot)
Temp<-nls.lm(par=c(0,0),lower=NULL,upper=NULL,fn=Fun.EE.Logistic,jac=NULL,W=
mi.Boot.Calib)$par[2]
Vector.Boot.Theta<-c(Vector.Boot.Theta,Temp)
}
return(Vector.Boot.Theta)
}
#-----------------------------------------------------------------------------
#                        START EXAMPLES
#-----------------------------------------------------------------------------
# APPLICATION OF THE FUNCTIONS PROVIDED FOR LOGISTIC REGRESSION
# IN THE PRESENCE OF AUXILIARY INFORMATION
#=============================================================================
# Beta1.Pop: population value of the slope (GLS estimate)
#-----------------------------------------------------------------------------
# EL point estimation and ELLR value in the presence of
# auxiliary information
#=============================================================================
mi.Star<-EL.Calib.Weight(Gamma)
Beta.Est<-nls.lm(par=c(0,0),lower=NULL,upper=NULL,fn=Fun.EE.Logistic,jac=NULL,
W=mi.Star)$par
Beta0.Est<-Roots.Beta[1]
Beta1.Est<-Roots.Beta[2]
Vector.ELLR<-ELLR.Nuisance(Beta1.Pop,Beta0.Est,Gamma)[1]
#-----------------------------------------------------------------------------
# The following functions for the EL confidence interval in the presence of
# auxiliary information can be written in a similar manner as what has been
# provided for linear regression (see above)
#-----------------------------------------------------------------------------
```

```
LB.EL<-EL.LowerBound(Beta1.Est,Beta0.Est,Gamma,1e-08,1e-06,0.05)
UB.EL<-EL.UpperBound(Beta1.Est,Beta0.Est,Gamma,1e-08,1e-06,0.05)
#-------------------------------------------------------------------------
# Pseudo likelihood point estimation and confidence intervals
# in the presence of population level information
#=========================================================================
wi<-Calib.Weight(Gamma,W)
Beta.Pseudo.Est<-nls.lm(par=c(0,0),lower=NULL,upper=NULL,fn=Fun.EE.Logistic,jac=
NULL,W=W)$par
LB.Initial<-Beta.Pseudo.Est[2]-2*0.05*abs(Beta.Pseudo.Est[2])
UB.Initial<-Beta.Pseudo.Est[2]+2*0.05*abs(Beta.Pseudo.Est[2])
#-------------------------------------------------------------------------
# The following functions for the pseudo likelihood 1 confidence interval
# in the presence of auxiliary information can be written in a similar
# manner as what has been provided for linear regression (see above)
#-------------------------------------------------------------------------
LB.Pseudo1<-nls.lm(par=c(Beta.Pseudo.Est[1],LB.Initial),lower=NULL,upper=NULL,
fn=Fun.X.LB.Pseudo1,jac=NULL,Theta=Beta.Pseudo.Est[2],Lambda=Beta.Pseudo.Est[1],
Gamma=Gamma,W=wi,Pr=Inc.Prob,f=f)$par[2]
UB.Pseudo1<-nls.lm(par=c(Beta.Pseudo.Est[1],UB.Initial),lower=NULL,upper=NULL,
fn=Fun.X.UB.Pseudo1,jac=NULL,Theta=Beta.Pseudo.Est[2],Lambda=Beta.Pseudo.Est[1],
Gamma=Gamma,W=wi,Pr=Inc.Prob,f=f)$par[2]
#=========================================================================
LB.Pseudo2<-nls.lm(par=c(Beta.Pseudo.Est[1],LB.Initial),lower=NULL,upper=NULL,
fn=Fun.X.LB.Pseudo2,jac=NULL,Gamma=Gamma,W=wi,Pr=Inc.Prob,f=f)$par[2]
UB.Pseudo2<-nls.lm(par=c(Beta.Pseudo.Est[1],UB.Initial),lower=NULL,upper=NULL,
fn=Fun.X.UB.Pseudo2,jac=NULL,Gamma=Gamma,W=wi,Pr=Inc.Prob,f=f)$par[2]
#-------------------------------------------------------------------------
# Bootstrap confidence interval in the presence of auxiliary information
#=========================================================================
# wi: calibration weights
#-------------------------------------------------------------------------
Vector.Boot.Beta1<-Rescaled.Boot(n-1,n,Gamma,wi,1000)
CI.ResBoot<-Percentile.CI(Vector.Boot.Beta1,0.05)
#-------------------------------------------------------------------------
#                          END EXAMPLES
#-------------------------------------------------------------------------
```

Appendix C

Supplementary Material for the Third Paper

By Melike Oguz-Alper
University of Southampton, SO17 1BJ, Southampton, U.K.
M.OguzAlper@soton.ac.uk

C.1 R code for the third paper

```
#-----------------------------------------------------------------
# Clear the workspace and maximise the memory
#-----------------------------------------------------------------
rm(list=ls(all=TRUE))
memory.size(TRUE)
memory.limit(size=4095)
#-----------------------------------------------------------------
# Load packages required
#-----------------------------------------------------------------
library(rootSolve)
library(MASS)
library(lpSolve)
library(sampling)
library(minpack.lm)
library(nlme)
library(lattice)
library(Matrix)
#-----------------------------------------------------------------
# FUNCTIONS FOR TWO-LEVEL REGRESSION UNDER UNIFORM COVARIANCE
#-----------------------------------------------------------------
#=================================================================
# DEFINITIONS OF THE PARAMETERS USED IN THE FOLLOWING FUNCTIONS
# (unless otherwise stated)
#-----------------------------------------------------------------
# data: sample data frame including nine columns of variables: individual ID,
# cluster ID, response variable, intercept (column vector of 1s),covariate 1,
# covariate 2, inclusion probabilities of the PSUs,inclusion probabilities of
```

```
# the SSUs, joint inclusion probabilities (Inc.Prob.PSU*Inc.Prob.SSU)
# d: number of sample PSUs
# size: dx1 size vector where sizes are the number of secondary sample units
# (SSUs) within associated sample cluster (PSU)
# id: total sample sizex1 vector of sample cluster ID that take values id=1,...,d.
# Units in the same cluster take the same id.
# adj.w: total sample sizex1 vector of scaled weights of SSUs
# psu.w: dx1 vector of design weights of PSUs
# n: number of sample PSUs
# ni: nx1 size vector where sizes are the number of SSUs within associated
# sample PSU
# u: estimate (true value if known) of second level residual variance (between
# PSU variance)
# W: inverse of scaled weights of SSUs ('effective' sample size)
# e: estimate (true value if known) of first level residual variance (within PSU
# variance)
# PsuID: total sample sizex1 vector of sample cluster ID that take values
# id=1,...,n. Units in the same cluster take the same id.
# Beta0: a given value of Beta0
# Beta1: a given value of Beta1
# Beta2: a given value of Beta2
# DomainData0: first object of the output of function 'TwoStage.EF'.
# DomainData1: first object of the output of function 'TwoStage.EF'.
# DomainData2: third object of the output of function 'TwoStage.EF'.
# Theta: a given value of Theta (Here, Theta is the parameter of interest)
# a: 1,2 or 3 depending on which regression coefficient is the parameter
# of interest. For example, a=1 when the par. of int. is the 'intercept'.
# Nu: vector of initial values for the nuisance parameters
# f: sum(Inclusion.Probabilities^2)/n: Here, summation is taken over population
# clusters. Inclusion probabilities are assosiacated to PSUs.
#-------------------------------------------------------------------------
# Function for estimating variance components (weights are incorporated)
#=========================================================================
EBLUP.VarEst<-function(data,d,size,id,adj.w,psu.w)
{

psu.w.i<-1/data[,7]

y<-data[,3]
Intercept<-data[,4]
x1<-data[,5]
x2<-data[,6]

W<-psu.w.i*adj.w

yd.bar<-NULL
x1d.bar<-NULL
x2d.bar<-NULL
Nd.Hat<-NULL

for(i in 1:d){
temp.y<-sum(y[id==i]*adj.w[id==i])/sum(adj.w[id==i])
temp.x1<-sum(x1[id==i]*adj.w[id==i])/sum(adj.w[id==i])
temp.x2<-sum(x2[id==i]*adj.w[id==i])/sum(adj.w[id==i])
temp.Nd<-sum(adj.w[id==i])
yd.bar<-c(yd.bar,temp.y)
x1d.bar<-c(x1d.bar,temp.x1)
x2d.bar<-c(x2d.bar,temp.x2)
Nd.Hat<-c(Nd.Hat,temp.Nd)
```

```
if(is.nan(yd.bar[i])) yd.bar[i]<-0
if(is.nan(x1d.bar[i])) x1d.bar[i]<-0
if(is.nan(x2d.bar[i])) x2d.bar[i]<-0
}

ydi.bar<-rep(yd.bar,size)
x1di.bar<-rep(x1d.bar,size)
x2di.bar<-rep(x2d.bar,size)

Temp.yi<-(y-ydi.bar)
Temp.x1i<-(x1-x1di.bar)
Temp.x2i<-(x2-x2di.bar)

fit.ols.nointercept<-lm(Temp.yi~-1+Temp.x1i+Temp.x2i)
k0<-length(fit.ols.nointercept$coef)
VarEst.e<-sum(fit.ols.nointercept$resid^2*W)/(sum(W)-sum(psu.w)-k0)

fit.ols<-lm(y~x1+x2)
k1<-k0+1
X.Mat=cbind(Intercept,x1,x2)
XW.Mat=cbind(Intercept*W,x1*W,x2*W)

xd.bar.Mat<-matrix(c(Nd.Hat,x1d.bar*Nd.Hat,x2d.bar*Nd.Hat),nrow=k1,ncol=d,byrow=
TRUE)
xdW.bar.Mat<-matrix(c(Nd.Hat*psu.w,x1d.bar*Nd.Hat*psu.w,x2d.bar*Nd.Hat*psu.w),
nrow=k1,ncol=d,byrow=TRUE)
Sum.xd.bar<-xd.bar.Mat%*%t(xdW.bar.Mat)
n.Star<-sum(W)-sum(diag(solve(t(X.Mat)%*%XW.Mat)%*%Sum.xd.bar))
VarEst.u<-(sum(fit.ols$resid^2*W)-(sum(W)-k1)*VarEst.e)/n.Star
if (VarEst.u<0) VarEst.u<-0
return(c(VarEst.e,VarEst.u))
}
#-----------------------------------------------------------------------
# Function for computing totals from SSUs to form estimating functions at first
# level
#=======================================================================
# data: sample data frame including columns of variables: response variable,
# covariate 1, covariate 2.
#-----------------------------------------------------------------------
TwoStage.EF<-function(n,ni,u,W,e,PsuID,data)
{
gj1.Beta0<-rep(0,n)
gj2.Beta0<-rep(0,n)
gj3.Beta0<-rep(0,n)
gj4.Beta0<-rep(0,n)

gj1.Beta1<-rep(0,n)
gj2.Beta1<-rep(0,n)
gj3.Beta1<-rep(0,n)
gj4.Beta1<-rep(0,n)

gj1.Beta2<-rep(0,n)
gj2.Beta2<-rep(0,n)
gj3.Beta2<-rep(0,n)
gj4.Beta2<-rep(0,n)

y<-data[,1]
x1<-data[,2]
x2<-data[,3]
```

```
for(i in 1:n)
{
x0i.Sample.Temp<-rep(1,ni[i])
VarMat<-x0i.Sample.Temp%*%t(x0i.Sample.Temp)*u+diag(e*W[PsuID==i])
InvVarMat<-solve(VarMat)
yi.Sample.Temp<-y[PsuID==i]
x1i.Sample.Temp<-x1[PsuID==i]
x2i.Sample.Temp<-x2[PsuID==i]

Part1<-InvVarMat%*%yi.Sample.Temp
Part2<-InvVarMat%*%x0i.Sample.Temp
Part3<-InvVarMat%*%x1i.Sample.Temp
Part4<-InvVarMat%*%x2i.Sample.Temp

gj1.Beta0[i]<-as.vector(x0i.Sample.Temp%*%Part1)
gj2.Beta0[i]<-as.vector(x0i.Sample.Temp%*%Part2)
gj3.Beta0[i]<-as.vector(x0i.Sample.Temp%*%Part3)
gj4.Beta0[i]<-as.vector(x0i.Sample.Temp%*%Part4)

gj1.Beta1[i]<-as.vector(x1i.Sample.Temp%*%Part1)
gj2.Beta1[i]<-as.vector(x1i.Sample.Temp%*%Part2)
gj3.Beta1[i]<-as.vector(x1i.Sample.Temp%*%Part3)
gj4.Beta1[i]<-as.vector(x1i.Sample.Temp%*%Part4)

gj1.Beta2[i]<-as.vector(x2i.Sample.Temp%*%Part1)
gj2.Beta2[i]<-as.vector(x2i.Sample.Temp%*%Part2)
gj3.Beta2[i]<-as.vector(x2i.Sample.Temp%*%Part3)
gj4.Beta2[i]<-as.vector(x2i.Sample.Temp%*%Part4)
}
gj.Data.Beta0<-cbind(gj1.Beta0,gj2.Beta0,gj3.Beta0,gj4.Beta0)
gj.Data.Beta1<-cbind(gj1.Beta1,gj2.Beta1,gj3.Beta1,gj4.Beta1)
gj.Data.Beta2<-cbind(gj1.Beta2,gj2.Beta2,gj3.Beta2,gj4.Beta2)

return(list(gj.Data.Beta0,gj.Data.Beta1,gj.Data.Beta2))
}
#-----------------------------------------------------------------
# Estimating functions associated with the first stage sample units (PSUs)
#=================================================================
Fun.EF.Linear.GEE<-function(Beta0,Beta1,Beta2,DomainData0,DomainData1,DomainData2)
{
m<-length(DomainData0[,1])
gi.Beta0<-DomainData0[,1]-Beta0*DomainData0[,2]-Beta1*DomainData0[,3]-Beta2*
DomainData0[,4]
gi.Beta1<-DomainData1[,1]-Beta0*DomainData1[,2]-Beta1*DomainData1[,3]-Beta2*
DomainData1[,4]
gi.Beta2<-DomainData2[,1]-Beta0*DomainData2[,2]-Beta1*DomainData2[,3]-Beta2*
DomainData2[,4]
matrix(c(gi.Beta0,gi.Beta1,gi.Beta2),nrow=m,ncol=3,byrow=FALSE)
}
#-----------------------------------------------------------------
# Estimating equations associated with the first stage sample units (PSUs)
#=================================================================
# W: nx1 vector of weights
#-----------------------------------------------------------------
Fun.EE.Linear.GEE<-function(X,W,DomainData0,DomainData1,DomainData2)
{
gi<-Fun.EF.Linear.GEE(X[1],X[2],X[3],DomainData0,DomainData1,DomainData2)
gi.Beta0<-gi[,1]
```

```
gi.Beta1<-gi[,2]
gi.Beta2<-gi[,3]
Fun1<-sum(gi.Beta0*W)
Fun2<-sum(gi.Beta1*W)
Fun3<-sum(gi.Beta2*W)
c(Fun1,Fun2,Fun3)
}
#-----------------------------------------------------------------------------
# FUNCTIONS FOR EMPIRICAL LIKELIHOOD
#=============================================================================
#-----------------------------------------------------------------------------
# PROFILING
#=============================================================================
#-----------------------------------------------------------------------------
# Function for solving lagrange coefficients for a given value of the
# parameter of interest
# Nuisance parameters are unknown (PROFILING)
#=============================================================================
# W: design weights of PSUs
#-----------------------------------------------------------------------------
Fun.X.No.GEE<-function(X,Theta,DomainData0,DomainData1,DomainData2,W,a)
{
m<-length(DomainData0[,1])
pi<-1/W

if(a==1) gi<-Fun.EF.Linear.GEE(Theta,X[5],X[6],DomainData0,DomainData1,DomainData2)
if(a==2) gi<-Fun.EF.Linear.GEE(X[5],Theta,X[6],DomainData0,DomainData1,DomainData2)
if(a==3) gi<-Fun.EF.Linear.GEE(X[5],X[6],Theta,DomainData0,DomainData1,DomainData2)

gi.Beta0<-gi[,1]
gi.Beta1<-gi[,2]
gi.Beta2<-gi[,3]
Denom<-pi+pi*X[1]+gi.Beta0*X[2]+gi.Beta1*X[3]+gi.Beta2*X[4]

Fun1<-sum(pi/Denom)-m
Fun2<-sum(gi.Beta0/Denom)
Fun3<-sum(gi.Beta1/Denom)
Fun4<-sum(gi.Beta2/Denom)
Fun5<-sum(-(DomainData0[,2]*X[2]+DomainData1[,2]*X[3]+DomainData2[,2]*X[4])/Denom)
Fun6<-sum(-(DomainData0[,3]*X[2]+DomainData1[,3]*X[3]+DomainData2[,3]*X[4])/Denom)
Fun7<-sum(-(DomainData0[,4]*X[2]+DomainData1[,4]*X[3]+DomainData2[,4]*X[4])/Denom)
Fun.Der<-c(Fun5,Fun6,Fun7)
Fun.Der<-Fun.Der[-a]
c(Fun1,Fun2,Fun3,Fun4,Fun.Der)
}
#-----------------------------------------------------------------------------
# Function for finding the value of empirical log-likelihood ratio (ELLR)
# funtion for a given value of Theta
# Nuisance parameter Lambda is unknown (PROFILING)
#=============================================================================
# W: design weights fo the primary sampling units (PSUs)
#-----------------------------------------------------------------------------
ELLR.Nuisance.GEE<-function(Theta,Nu,DomainData0,DomainData1,DomainData2,W,a)
{
k<-length(Nu)+2
pi<-1/W
X<-nls.lm(par=c(rep(0,k),Nu),lower=NULL,upper=NULL,fn=Fun.X.No.GEE,jac=NULL,Theta=
Theta,DomainData0=DomainData0,DomainData1=DomainData1,DomainData2=DomainData2,W=W,
a=a)$par
```

```
if(a==1) gi<-Fun.EF.Linear.GEE(Theta,X[5],X[6],DomainData0,DomainData1,DomainData2)
if(a==2) gi<-Fun.EF.Linear.GEE(X[5],Theta,X[6],DomainData0,DomainData1,DomainData2)
if(a==3) gi<-Fun.EF.Linear.GEE(X[5],X[6],Theta,DomainData0,DomainData1,DomainData2)
gi.Beta0<-gi[,1]
gi.Beta1<-gi[,2]
gi.Beta2<-gi[,3]

Denom<-pi+pi*X[1]+gi.Beta0*X[2]+gi.Beta1*X[3]+gi.Beta2*X[4]
mi.Hat<-1/Denom

if(range(mi.Hat)[1]<0.01)
{
s<-5e-09
l<-1
while(range(mi.Hat)[1]<0.01)
{
X<-nls.lm(par=c(s,rep(0,(k-1)),X[5],X[6]),lower=NULL,upper=NULL,fn=Fun.X.No.GEE,
jac=NULL,Theta=Theta,DomainData0=DomainData0,DomainData1=DomainData1,DomainData2=
DomainData2,W=W,a=a)$par

if(a==1) gi<-Fun.EF.Linear.GEE(Theta,X[5],X[6],DomainData0,DomainData1,DomainData2)
if(a==2) gi<-Fun.EF.Linear.GEE(X[5],Theta,X[6],DomainData0,DomainData1,DomainData2)
if(a==3) gi<-Fun.EF.Linear.GEE(X[5],X[6],Theta,DomainData0,DomainData1,DomainData2)
gi.Beta0<-gi[,1]
gi.Beta1<-gi[,2]
gi.Beta2<-gi[,3]
Denom<-pi+pi*X[1]+gi.Beta0*X[2]+gi.Beta1*X[3]+gi.Beta2*X[4]
mi.Hat<-1/Denom
l<-l+1
}
}
ELLR.Function<-2*sum(log(W)-log(mi.Hat))
return(ELLR.Function)
}
#-------------------------------------------------------------------------
# FUNCTIONS FOR PSEUDO LIKELIHOOD METHOD
#=========================================================================
#-------------------------------------------------------------------------
# HARTLEY & RAO (1962) VARIANCE ESTIMATOR
#=========================================================================
# y: variable of interest
# n: sample size
# Pr: first-order inclusion probabilities
# f: sum(Inclusion.Probabilities^2)/n. Here, summation is taken over the population.
#-------------------------------------------------------------------------
Var.Sys<-function(y,n,Pr,f)
{
Total<-0
for(k in 1:n-1)
{
wi<-1/Pr
Inc.Prob.Trun<-Pr[(k+1):n]
mi.Trun<-wi[(k+1):n]
y.Trun<-y[(k+1):n]
Sum.k<-sum((1-Pr[k]-Inc.Prob.Trun+f)*(wi[k]*y[k]-mi.Trun*y.Trun)^2)
Total<-sum(Total+Sum.k)
}
return(Total/(n-1))
```

```
}
#-------------------------------------------------------------------------
# Variance estimation of new estimating function (gi.Star) that is computed
# at PSU level
#=========================================================================
# Beta0: estimate of Beta0
# Beta1: estimate of Beta1
# Beta2: estimate of Beta2
# W: design weights fo the primary sampling units (PSUs)
#-------------------------------------------------------------------------
Varyans.Pseudo.TwoStage<-function(Beta0,Beta1,Beta2,DomainData0,DomainData1,
DomainData2,W,f,a)
{
gi<-Fun.EF.Linear.GEE(Beta0,Beta1,Beta2,DomainData0,DomainData1,DomainData2)
gi.Beta0<-gi[,1]
gi.Beta1<-gi[,2]
gi.Beta2<-gi[,3]

Sum.DomainData0<- -c(sum(DomainData0[,2]*W),sum(DomainData0[,3]*W),
sum(DomainData0[,4]*W))
Sum.DomainData1<- -c(sum(DomainData1[,2]*W),sum(DomainData1[,3]*W),
sum(DomainData1[,4]*W))
Sum.DomainData2<- -c(sum(DomainData2[,2]*W),sum(DomainData2[,3]*W),
sum(DomainData2[,4]*W))

if(a==1) Der.Theta<- Sum.DomainData0[-a]
if(a==2) Der.Theta<- Sum.DomainData1[-a]
if(a==3) Der.Theta<- Sum.DomainData2[-a]
if(a==1) Der.Lambda<-matrix(c(Sum.DomainData1[-a],Sum.DomainData2[-a]),nr=2,nc=2,
byrow=TRUE)
if(a==2) Der.Lambda<-matrix(c(Sum.DomainData0[-a],Sum.DomainData2[-a]),nr=2,nc=2,
byrow=TRUE)
if(a==3) Der.Lambda<-matrix(c(Sum.DomainData0[-a],Sum.DomainData1[-a]),nr=2,nc=2,
byrow=TRUE)

J<-as.vector(Der.Theta%*%solve(Der.Lambda))
if(a==1) gi.Star<-gi.Beta0-J[1]*gi.Beta1-J[2]*gi.Beta2
if(a==2) gi.Star<-gi.Beta1-J[1]*gi.Beta0-J[2]*gi.Beta2
if(a==3) gi.Star<-gi.Beta2-J[1]*gi.Beta0-J[2]*gi.Beta1
Var.gi.Star<-Var.Sys(gi.Star,length(W),1/W,f)
return(Var.gi.Star)
}
#-------------------------------------------------------------------------
# PSEUDO LIKELIHOOD CONFIDENCE INTERVALS
#=========================================================================
#-------------------------------------------------------------------------
# Lower bound (replace -1.96 by +1.96 for upper bound)
#=========================================================================
# Beta0: estimate of Beta0
# Beta1: estimate of Beta1
# Beta2: estimate of Beta2
# W: design weights fo the primary sampling units (PSUs)
#-------------------------------------------------------------------------
Fun.X.LB.Pseudo.TwoStage<-function(X,Beta0,Beta1,Beta2,DomainData0,DomainData1,
DomainData2,W,f,a)
{
gi<-Fun.EF.Linear.GEE(X[1],X[2],X[3],DomainData0,DomainData1,DomainData2)
gi.Beta0<-gi[,1]
gi.Beta1<-gi[,2]
```

```
gi.Beta2<-gi[,3]

Var.gi.Star<-Varyans.Pseudo.TwoStage(Beta0,Beta1,Beta2,DomainData0,DomainData1,
DomainData2,W,f,a)

Fun1<-sum(gi.Beta0*W)
Fun2<-sum(gi.Beta1*W)
Fun3<-sum(gi.Beta2*W)
Fun.Lambda<-c(Fun1,Fun2,Fun3)
Fun.Lambda<-Fun.Lambda[-a]

if(a==1) Fun4<-sum(gi.Beta0*W)/sqrt(Var.gi.Star)-1.96
if(a==2) Fun4<-sum(gi.Beta1*W)/sqrt(Var.gi.Star)-1.96
if(a==3) Fun4<-sum(gi.Beta2*W)/sqrt(Var.gi.Star)-1.96
c(Fun.Lambda,Fun4)
}
#---------------------------------------------------------------------------
#                       START EXAMPLES
#---------------------------------------------------------------------------
# APPLICATION OF THE FUNCTIONS PROVIDED FOR TWO-LEVEL MODEL WITH UNIVARIATE
# COVARIANCE STRUCTURE
#===========================================================================
# Beta.Pop: vector of population values of regression coefficients
#---------------------------------------------------------------------------
# Point estimation for regression coefficients and computing ELLR function
#===========================================================================
VarEst<-EBLUP.VarEst(data,d,size,id,adj.w,psu.w)
e.Value<-VarEst[1]
u.Value<-VarEst[2]
Cluster.EFs<-TwoStage.EF(d,size,u.Value,W,e.Value,id,data)
Beta.Est<-nls.lm(par=c(rep(0,3)),lower=NULL,upper=NULL,fn=Fun.EE.Linear.GEE,
jac=NULL,DomainData0=Cluster.EFs[[1]],DomainData1=Cluster.EFs[[2]],DomainData2=
Cluster.EFs[[3]],W=psu.w)$par
Vector.ELLR<-NULL
for(a in 1:3)
{
Temp<-ELLR.Nuisance.GEE(Beta.Pop[a],Beta.Est[-a],Cluster.EFs[[1]],Cluster.EFs[[2]],
Cluster.EFs[[3]],psu.w,a)
Vector.ELLR<-c(Vector.ELLR,Temp)
}
#---------------------------------------------------------------------------
# Computing observed coverage of EL confidence interval for a given value of ELLR
# by using p-value
#===========================================================================
# test: a given value for ELLR function
# Y: true population value of the parameter of interest
# y: estimate of the parameter of interest
# a: significance level
#---------------------------------------------------------------------------
CP.TailErrors.PValue<-function(test,Y,y,a)
{
if ((1-pchisq(test,1))<a) {CP<-0} else {CP<-1}
if (CP==0 && Y<y) {LE<-1} else {LE<-0}
if (CP==0 && Y>y) {UE<-1} else {UE<-0}
return(c(CP,LE,UE))
}
EL.CP<-NULL
for(a in 1:3)
{
```

```
Temp<-CP.TailErrors.PValue(Vector.ELLR[a],Beta.Pop[a],Beta.Est[a],0.05)[1]
EL.CP<-c(EL.CP,Temp)
}
#---------------------------------------------------------------------------
# Observed coverage of pseudo likelihood confidence interval
#===========================================================================
# Y: true population value of the parameter of interest
# lb: lower bound
# ub: upper bound
#---------------------------------------------------------------------------
CP.TailErrors<-function(Y,lb,ub)
{
if(Y<lb) LE<-1 else LE<-0
if(Y>ub) UE<-1 else UE<-0
if(LE+UE==0) CP<-1 else CP<-0
return(c(CP,LE,UE))
}
Pseudo.CP<-NULL

for (a in 1:3)
{
LB.Initial<-Beta.Est[a]-2*0.05*abs(Beta.Est[a])
UB.Initial<-Beta.Est[a]+2*0.05*abs(Beta.Est[a])
LB.Initial.Par<-Beta.Est
LB.Initial.Par[a]<-LB.Initial
LB.Pseudo<-nls.lm(par=LB.Initial.Par,lower=NULL,upper=NULL,fn=
Fun.X.LB.Pseudo.TwoStage,jac=NULL,Beta0=Beta.Est[1],Beta1=Beta.Est[2],Beta2=
Beta.Est[3],DomainData0=Cluster.EFs[[1]],DomainData1=Cluster.EFs[[2]],DomainData2=
Cluster.EFs[[3]],W=psu.w,f=f,a=a)$par[a]

UB.Initial.Par<-Beta.Est
UB.Initial.Par[a]<-UB.Initial
UB.Pseudo<-nls.lm(par=UB.Initial.Par,lower=NULL,upper=NULL,fn=
Fun.X.UB.Pseudo.TwoStage,jac=NULL,Beta0=Beta.Est[1],Beta1=Beta.Est[2],Beta2=
Beta.Est[3],DomainData0=Cluster.EFs[[1]],DomainData1=Cluster.EFs[[2]],DomainData2=
Cluster.EFs[[3]],W=psu.w,f=f,a=a)$par[a]

Temp<-CP.TailErrors(Beta.Pop[a],LB.Pseudo,UB.Pseudo)[1]
Pseudo.CP<-c(Pseudo.CP,Temp)
}
#---------------------------------------------------------------------------
#                          END EXAMPLES
#---------------------------------------------------------------------------
```

Bibliography

Asparouhov, T. (2006), "General multi-level modelling with sampling weights," *Communication in Statistics - Theory and Methods*, 35(3), 439–460.

Asparouhov, T., and Muthén, B. (2006), "Multilevel modelling of complex survey data," *Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meeting, Seattle*, pp. 2718–2726.

Battese, G., Harter, R., and Fuller, W. (1988), "An error-components model for prediction of county crop areas using survey and satellite data," *Journal of the American Statistical Association*, 83(401), 28–36.

Berger, Y. G. (2004), "Variance estimation for measures of change in probability sampling," *Canadian Journal of Statistics*, 32(4), 451–467.

Berger, Y. G. (2008), "A note on the asymptotic equivalence of Jackknife and linearization variance estimation for the Gini coefficient," *Journal of Official Statistics*, 24(4), 541–555.

Berger, Y. G. (2011), "Asymptotic consistency under large entropy sampling designs with unequal probabilities," *Pakistan Journal of Statistics, Festschrift to honour Ken Brewer's 80th birthday*, 27(4), 407–426.

Berger, Y. G., and De La Riva Torres, O. (2016), "An empirical likelihood approach for inference under complex sampling design," *Journal of the Royal Statistical Society, Series B*, p. 22pp.

Berger, Y. G., and Escobar, E. L. (2015), "Variance estimation of Hot-Deck imputed estimators of change for repeated rotating surveys,", Southampton Statistical Sciences Research Institute.

Berger, Y. G., Goedemé, T., and Osier, G. (2013), *Handbook on standard error estimation and other related sampling issues in EU-SILC* Second Network for the Analysis of EU-SILC, EuroStat. `http://www.cros-portal.eu/content/handbook-standard-error-estimation-and-other-related-sampling-issues-ver-29072013`. [Online; accessed 06 February 2013].

Berger, Y. G., and Priam, R. (2010), "Estimation of correlations between cross-sectional Estimates from Repeated Surveys - an Application to the Variance of Change," *Proceeding of the 2010 Symposium of Statistics Canada*, p. 10pp. 26-29 October, 2010.

Berger, Y. G., and Priam, R. (2016), "A Simple Variance Estimator of Change for Rotating Repeated Surveys: an Application to the EU-SILC Household Surveys,"

Journal of the Royal Statistical Society, Series A, p. 22pp. Available at: DOI: http://dx.doi.org/10.1111/rssa.12116. [Online; accessed 09 June 2015].

Berger, Y. G., and Skinner, C. J. (2003), "Variance Estimation of a Low-Income Proportion," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52, 457–468.

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.

Betti, G., and Gagliardi, F. (2007), Jackknife variance estimation of differences and averages of poverty measures,, Working Paper 68, Siena: Dipartimento di Metodi Quantitativi, Universit degli Studi.

Binder, D. A. (1983), "On the variance of asymptotically normal estimators from complex surveys," *International Statistical Review*, 51(427), 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.

Breidt, F., Claeskens, G., and Opsomer, J. (2005), "Model-assisted estimation for complex surveys using penalised splines," *Biometrika*, 92, 831–846.

Bruch, C., Münnich, R., and Zins, S. (2011), "Variance Estimation for Complex Surveys,", Work-package of the European project on Advanced Methodology for European Laeken Indicators (AMELI) http://www.uni-trier.de/index.php?id=24676. [Online; accessed 4 Jan. 2013].

Chao, M. T. (1982), "A General Purpose Unequal Probability Sampling Plan," *Biometrika*, 69, 653–656.

Chaudhuri, S., Handcock, M. S., and Rendall, M. S. (2008), "Generalized Linear Models Incorporating Population Level Information: An Empirical-Likelihood-Based Approach," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 70(2), pp. 311–328.

Chen, J., and Qin, J. (1993), "Empirical likelihood estimation for finite populations and the effective usage of auxiliary information," *Biometrika*, 80(1), 107–116.

Chen, J., and Sitter, R. R. (1999), "A pseudo empirical likelihood approach to the effective use of auxiliary information in complex surveys," *Statistica Sinica*, 9, 385–406.

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(1), 230–237.

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

Chen, S., and Kim, J. K. (2014), "Population empirical likelihood for nonparametric inference in survey sampling," *Statistica Sinica*, 24, 335–355.

Christine, M., and Rocher, T. (2012), "Construction d'échantillons astreints à des conditions de recouvrement par rapport un échantillon antérieur et à des conditions d'équilibrage par rapport à des variables courantes," *Proceeding of the 10th Journée de Méthodologie Statistique de l'INSEE (Paris, 24-26 January 2012)*, p. 41pp.

Clogg, C., and Eliason, S. (1987), "Some common problems in log-linear analysis," *Sociological Methods and Research*, 16, 8–44.

Crowder, M. (1995), "On the Use of a Working Correlation Matrix in Using Generalized Linear Models for Repeated Measurements," *Biometrika*, 82(2), 407–410.

De Moivre, A. (1733), "Approximatio ad summam terminorum binomii (a+b)n in seriem expansi," *Self-published pamplet*, . 7pp.

De Toledo Vieira, M., and Skinner, C. J. (2008), "Estimating Models for Panel Survey Data under Complex Sampling," *Journal of Official Statistics*, 24(3), 343–364.

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.

Di Meglio, E., Osier, G., Goedemé, T., Berger, Y. G., and Di Falco, E. (2013), *Standard Error Estimation in EU-SILC - First Results of the Net-SILC2 Project*, Brussels, 5-7 March, 2013: Proceeding of the conference on New Techniques and Technologies for Statistics, Brussels, 10pp. `http://www.cros-portal.eu/sites/default/files/NTTS2013fullPaper_144.pdf`. [Online; accessed, 30 April 2015].

Diggle, P., Heagerty, P., Liang, K., and Zeger, S. (2002), *Analysis of longitudinal data (2nd ed.)*, Oxford: Oxford University Press.

Durbin, J. (1953), "Some results in sampling theory when the units are selected with unequal probabilities," *Journal of the Royal Statistical Society Series B*, 15(2), 262–269.

Estevao, V. M., and Särndal, C.-E. (2006), "Survey estimates by calibration on complex auxiliary information," *International Statistical Review*, 74(2), 127–147.

Eurostat (2003), "'Laeken' Indicators-Detailed Calculation Methodology,", `http://www.cso.ie/en/media/csoie/eusilc/documents/Laeken%20Indicators%20-%20calculation%20algorithm.pdf`. [Online; accessed 4 Feb. 2014].

Eurostat (2012), "European Union Statistics on Income and Living Conditions (EU-SILC),", `http://epp.eurostat.ec.europa.eu/portal/page/portal/microdata/eu_silc`.

Feiveson, A. H. (2002), "Power by simulation," *The STATA Journal*, 2, 107–124.

Gambino, J. G., and Silva, P. L. N. (2009), "Sampling and estimation in household surveys," *Handbook of Statistics: Design, Method and Applications: D. Pfeffermann and C.R. Rao.(editors). Elsevier*, 29A, 407–439.

Godambe, V. P. (1960), "An Optimum Property of Regular Maximum Likelihood Estimation," *The Annals of Mathematical Statistics*, 31(4), pp. 1208–1211.

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

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

Goedemé, T. (2010), The standard error of estimates based on EU-SILC. An exploration through the Europe 2020 poverty indicators.,, Working paper 10/09. [Online; accessed 30 April 2015].

Goldstein, H. (1986), "Multilevel mixed linear model analysis using iterative generalised least squares," *Biometrika*, 73, 43–56.

Goldstein, H. (2011), *Multilevel Statistical Models*, 4th ed., Chichester: Wiley.

Gonzales, M. E. (1973), "Use and Evaluation of Synthetic Estimates," *Proceedings of the Social Statistics Section of the American Statistical Association*, pp. 33–36.

Graf, E. (2013), *Variance estimation by linearization for indicators of poverty and social exclusion in a person and household survey context* Presented at New Techniques and Technologies for Statistics, Brussels. `http://www.cros-portal.eu/content/14a01ericgraf`. [Online; accessed 5 Feb. 2014].

Graf, E., and Tillé, Y. (2014), "Variance estimation using linearization for poverty and social exclusion indicators," *Survey Methodology*, (40), 61–79.

Graubard, B., and Korn, E. (1996), "Modelling the sampling design in the analysis of health surveys," *Statistical Methods in Medical Research*, 5, 263–281.

Grilli, L., and Pratesi, M. (2004), "Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs," *Survey Methodology*, 30, 93–103.

Hájek, J. (1981), *Sampling from a Finite Population*, New York: Marcel Dekker.

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, L. P. (1982), "Large sample propeties of generalized method of moments estimators," *Econometrica*, 50(4), 1029 – 1054.

Hansen, M. H., and Hurwitz, W. N. (1943), "On the Theory of Sampling from Finite Populations," *The Annals of Mathematical Statistics*, 14(4), pp. 333–362.

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.

Hansen, M., Hurwitz, W., and Madow, W. (1953), *Sample Survey Methods and Theory, volume I*, New York: John Wiley and Sons.

Hartley, H. O., and Rao, J. N. K. (1962), "Sampling with unequal probabilities without replacement," *Ann. math. Statist. Assoc.*, 33, 350–374.

Hartley, H. O., and Rao, J. N. K. (1968), "A new estimation theory for sample surveys," *Biometrika*, 55(3), 547–557.

Hartley, H. O., and Rao, J. N. K. (1969), "A new estimation theory for sample surveys, II," *New Developments in survey Sampling (Johnson, N.L., and Smith, H.Jr., Eds.) Wiley, New York*, pp. 147–169.

Henderson, C. R. (1953), "Estimation of variance and variance components," *Biometrics*, 9, 226–252.

Henderson, C. R., Kempthorne, O., Searle, S. R., and von Krosigk, C. M. (1959), "The Estimation of Environmental and Genetic Trends from Records Subject to Culling," *Biometrics*, 15(2), 192–218.

Holmes, D. J., and Skinner, C. J. (2000), "Variance estimation for Labour Force Survey estimates of level and change," *Government Statistical Service Methodology Series*, . The Office for National Statistics, London, United Kingdom, 21, 40pp.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47(260), 663–685.

Huang, R., and Hidiroglou, M. (2003), "Design consistent estimators for a mixed linear model on survey data," *Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meetings, San Francisco*, pp. 1897–1904.

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

Kalton, G. (2009), "Design for surveys over time," *Handbook of Statistics: Design, Method and Applications: D. Pfeffermann and C.R. Rao.(editors). Elsevier*, 29A, 89–108.

Kim, J. K. (2009), "Calibration estimation using empirical likelihood in survey sampling," *Statistica Sinica*, 19, 145–157.

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.

Kish, L. (1965), *Survey Sampling* Wiley.

Korn, E., and Graubard, B. (2003), "Estimating variance components by uisng survey data," *Journal of the Royal Statistical Society. Series B*, 65, 175–190.

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.

Kovačević, M., and Rai, S. (2003), "A pseudo maximum likelihood approach to multilevel modelling of survey data," *Communication in Statistics - Theory and Methods*, 32(1), 103–121.

Krewski, D., and Rao, J. N. K. (1981), "Inference from stratified sample: properties of linearization jackknife, and balanced repeated replication methods," *The Annals of*

Statistics, 9, 1010–1019.

La Vange, L., Koch, G., and Schwartz, T. (2001), "Applying sample survey methods to clinical trials data," *Statistics in Medicine*, 20, 2609–2623.

Laniel, N. (1987), "Variances for a rotating sample from a changing population," *Proceedings of the Survey Research Methods Section, American Statistical Association*, . 17-20 August 1987, pp. 496-500.

Levenberg, K. (1944), "A Method for the Solution of Certain Problems in Least Squares," *Quart. Appl. Math.*, 2, 164–168.

Liang, K., and Zeger, S. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Longford, N. (1995), *Models for uncertainty in educational testing*, New York: Springer.

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

McDonald, J. B. (1984), "Some generalized functions for the size distribution of income," *Econometrica*, 52, 647–664.

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

Münnich, R., and Zins, S. (2011), "Variance Estimation for Indicators of Poverty and Social Exclusion,", Work-package of the European project on Advanced Methodology for European Laeken Indicators (AMELI) `http://www.uni-trier.de/index.php?id=24676`. [Online; accessed 4 Jan. 2013].

Muthén, L., and Muthén, B. (1998-2012), M*plus User's Guide* Seventh Edition. Los Angeles, CA: Muthén and Muthén.

Neyman, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 97(4), 558–625.

Nordberg, L. (2000), "On variance estimation for measures of change when samples are coordinated by the use of permanent random numbers," *Journal of Official Statistics*, 16, 363–378.

Oguz-Alper, M., and Berger, Y. G. (2014), "Empirical likelihood confidence intervals and significance test for regression parameters under complex sampling designs," *Proceedings of the Survey Research Method Section of the American Statistical Association, Joint Statistical Meetings, Boston*, pp. 2070–2079.

Oguz Alper, M., and Berger, Y. G. (2015), "Profile empirical likelihood in the presence of nuisance parameters and population level information under unequal probability sampling,", Southampton Statistical Sciences Research Institute `http://eprints.soton.ac.uk/376699/`.

Opsomer, J., and Miller, C. (2005), "Selecting the amount of smoothing parameter in nonparametric regression estimation for complex surveys," *Journal of Nonparametric Statistics*, 17(5), 593–611.

Osier, G. (2009), "Variance estimation for complex indicators of poverty and inequality using linearization techniques," *Survey Research Method*, 3(3), 167–195.

Osier, G., Berger, Y. G., and Goedemé, T. (2013), "Standard Error Estimation for the EU-SILC Indicators of Poverty and Social Exclusion," *Eurostat Methodologies and Working Papers series*, . Available at: `http://ec.europa.eu/eurostat/documents/3888793/5855973/KS-RA-13-024-EN.PDF`. [Online; accessed, 30 April 2015].

Owen, A. B. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75(2), 237–249.

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

Owen, A. B. (1991), "Empirical Likelihood for Linear Models," *The Annals of Statistics*, 19(4), 1725–1747.

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

Park, M., and Fuller, W. A. (2009), "The mixed model for survey regresion estimation," *Journal of Statistical Planning and Inference*, 139, 1320–1331.

Pfeffermann, D., Da Silva Moura, F. A., and Do Nascimento Silva, P. L. (2006), "Multilevel modelling under infromative sampling," *Biometrika*, 93(4), 943–959.

Pfeffermann, D., and La Vange, L. (1989), "Regression models for stratified multi-stage cluster samples," In *Analysis of Complex Surveys. C.J. Skinner, D. Holt and T.M.F. Smith (editors). Chichester: Wiley.*, pp. 237–260.

Pfeffermann, D., Skinner, C., Holmes, D., Goldstein, H., and Rasbash, J. (1998), "Weighting for unequal selection probabilities in multilevel models," *Journal of the Royal Statistical Society. Series B*, 60, 23–40.

Pfeffermann, D., and Sverchkov, M. (1999), "Parametric and semi-parametric estimation of regression models fitted to survey data," *The Indian Journal of Statistics* Special Issue on Sample Surveys, 61, 166–186.

Pfeffermann, D., and Sverchkov, M. (2003), "Fitting generalized linear models under informative probability sampling," In *Analysis of Survey Data. R.L. Chambers and C.J. Skinner (editors). New York: John Wiley and Sons, Inc.*, pp. 175–195.

Potthoff, R., Woodbury, M., and Manton, K. (1992), ""Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models," *Journal of the American Statistical Association*, 87, 383 – 396.

Prasad, N., and Rao, J. (1990), "The estimation of the mean square error of small-area estimators," *Journal of American Statistical Association*, 85, 163–171.

Prášková, Z., and Sen, P. K. (2009), "Asymptotic in Finite Population Sampling," *Handbook of Statistics 29: Sample Surveys: Inference and Analysis. Elsevier. Danny Pfeffermann and C.R. Rao eds*, pp. 489–522.

Preston, I. (1995), "Sampling distributions of relative poverty statistics," *Appl. Statist.*, 44, 91–99.

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

Qualité, L., and Tillé, Y. (2008), "Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added," *Survey Methodology*, 34(2), 173–181.

R Development Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. `http://www.R-project.org`, Vienna, Austria.

Rabe-Hesketh, S., and Skrondal, A. (2006), "Multilevel modelling of complex survey data," *Journal of Royal Statistical Society: Series A*, 169(4), 805–827.

Rao, J. (2003), *Small Area Estimation* Wiley, Hoboken, NJ.

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., and Scott, A. J. (1987), "On simple adjustments to chi-square tests with sample survey data," *The Annals of Statistics*, 15, 385–397.

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.

Rao, J. N. K., and Wu, W. (2009), "Empirical Likelihood Methods," *Handbook of statistics: Sample Surveys: Inference and Analysis, D. Pfeffermann and C. R. Rao eds. The Netherlands (North-Holland)*, 29B, 189–207.

Salem, A. B. Z., and Mount, T. D. (1974), "A convenient descriptive model of income distribution: the Gamma density," *Econometrika*, 42, 1115–1127.

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

Schmeiser, B. W., and Lal, R. (1982), "Bivariate gamma random vectors," *Operations Research*, 30, 355–374.

Scott, A., and Wu, C. F. (1981), "On the asymptotic distribution of ratio and regression estimators," *Journal of the American Statistical Association*, 76, 98–102.

Searle, S., Casella, G., and McCulloch, C. (1992), *Variance Components*, New York: Wiley.

Shapiro, S. S., and Wilk, M. B. (1965), "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrika*, 52(3/4), pp. 591–611.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis* London: Chapman and Hall.

Skinner, C. (1989), "Domain means, regression and multivariate analysis," In *Analysis of Complex Surveys. C.J. Skinner, D. Holt and T.M.F. Smith (editors). Chichester: Wiley.*, pp. 59–87.

Skinner, C. J., and De Toledo Vieira, M. (2007), "Variance estimation in the analysis of clustered longitudinal survey data," *Survey Methodology*, 33(1), 3–12.

Sutradhar, B., and Das, K. (1999), "On the efficiency of regression estimators in generalised linear models for longitudinal data," *Biometrika*, 86(2), 459–465.

Sutradhar, B., and Kovačević, M. (2000), "Analysing ordinal longitudinal survey data: generalized estimating equations approach," *Biometrika*, 87(4), 837–848.

Tam, S. M. (1984), "On covariances from overlapping samples," *American Statistician*, 38(4), 288–289.

SAS Institute Inc. (2011), *SAS/STAT 9.3 User's Guide* Cary, NC: SAS Institute Inc.

Verma, V., and Betti, G. (2005), Sampling errors and design effects for poverty measures and other complex statistics,, Working Paper 53, Siena: Dipartimento di Metodi Quantitativi, Universit degli Studi.

Verma, V., and Betti, G. (2011), "Taylor linearisation sampling errors and design effects for poverty measures and other complex statistics," *Journal of Applied Statistics*, 38, 1549–1576.

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

Wood, J. (2008), "On the Covariance Between Related Horvitz-Thompson Estimators," *Journal of Official Statistics*, 24(1), 53–78.

Wu, C., and Rao, J. N. K. (2006), "Pseudo-empirical likelihood ratio confidence intervals for complex surveys," *The Canadian Journal of Statistics*, 34(3), 359–375.

Wu, C., and Sitter, R. (2001), "A Model-Calibration Approach to Using Complete Auxiliary Information from Survey Data," *Journal of the American Statistical Association*, 96(453), 185–193.

You, Y., and Rao, J. (2002), "A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights," *Canadian Journal of Statistics*, 30(3), 431–439.

Zheng, M., Zhao, Z., and Yu, W. (2012), "Empirical likelihood methods based on influence functions," *Statistics and Its Interface*, 5, 355–366.

Zhong, B., and Rao, J. N. K. (2000), "Empirical Likelihood Inference under Stratified Random Sampling Using Auxiliary Population information," *Biometrika*, 87(4), 929–938.