Empirical likelihood confidence intervals and hypothesis testing for multidimensional parameters when modelling complex survey data

Melike Oguz-Alper*
The University of Southampton, Southampton, United Kingdom - M.OguzAlper@soton.ac.uk

Yves G. Berger
The University of Southampton, Southampton, United Kingdom - Y.G.Berger@soton.ac.uk

Regression models are widely used in social sciences, biological sciences, econometrics and finance. It is a common practice to base these models on survey data, which may be collected for a sample selected from a finite population with respect to a probability sampling design which may involve unequal inclusion probabilities, stratification and/or clustering. Confidence intervals based on model based estimators may have poor coverages when the sampling design is not taken into account. Standard design based confidence intervals may not have the right coverages when the sampling distribution is skewed. We propose to use an empirical likelihood approach to construct design based confidence intervals and to test hypotheses for unknown population quantities under unequal probability sampling. We consider the situation when the parameter of interest depends on other unknown parameters. In this case, we have a vector of unknown parameters and we are interested in a subset of the parameter. This requires profiling out the parameters which are not of primary interest. We show that the resulting profile empirical log-likelihood ratio function follows asymptotically a chi-squared distribution. Based on this property, we can test hypotheses and construct confidence intervals. The proposed empirical likelihood confidence intervals may have better coverages and more balanced tail errors when the sampling distribution is not normal, the parameter of interest is not linear, and/or when the model is misspecified. The proposed approach intrinsically incorporates sampling weights and design variables. Population level information can be incorporated into the proposed approach. It is simple to implement and less computer intensive than bootstrap. The proposed approach can be applied to any finite population parameter which can be defined as the solution of an estimating equation. For example, it can be applied to generalised linear models, linear mixed effects models, and quantile (robust) regression models.

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