Aligning estimates from different surveys using Empirical Likelihood methods

Ewa Kabzinska (ejk1g12@soton.ac.uk)[1], Yves G. Berger (Y.G.Berger@soton.ac.uk)[1]

Keyword: Empirical Likelihood, survey estimation, aligning estimates, auxiliary variables

1. INTRODUCTION

It is often the case that several surveys carried out independently in the same population measure some common variables. The population level parameters associated with these common variables are often unknown. Whether the common variable is of interest itself or is treated as an auxiliary information for estimation of other parameters, it may be beneficial to combine information gathered separately in different surveys. Combining information will usually increase precision and ensure that estimates are consistent across surveys. By consistency we mean a requirement that both samples give the same point estimate for the unknown population level parameter associated with the common variable. Typically there are also other side variables measured in the surveys, for which population level parameters, such as totals or means, are known. These variables are used to create benchmark constraints.

Aligning estimates from two surveys in presence of benchmark constraints was first addressed by Zieschang [1] in relation to the American Consumer Expenditure Survey. His method was extended later on by Renssen and Nieuwenbroek [2]. These authors propose to estimate the unknown population totals of the common variables using a pooled sample from two surveys and then include them as additional regressors in a GREG-type estimator. One of the drawbacks of this estimator is an increased probability of obtaining negative weights, especially when the number of regressors is large, which may be inconvenient from the practical point of view. Use of GREG-type estimators to combine information from different surveys was also investigated by Merkouris [3].

Wu [4] used Pseudo Empirical Likelihood methods to combine information from two independent surveys and obtained an estimator for a mean which is asymptotically equivalent to a GREG-type estimator.

Berger and de La Riva Torres [5] proposed an Empirical Likelihood based approach that may be used to estimate more complex parameters than means and totals in complex sampling designs. They obtain confidence intervals which may be calculated without relying on variance estimation or on unknown population parameters such as the design effect or the population size.

We extend the approach presented by Berger and de La Riva Torres [5] so that it can be used to combine multiple samples and to ensure that the estimates based on the common variable are equal across samples. We propose a method to obtain point estimators and confidence intervals for a wide class of parameters which are defined by estimating equations. Our approach allows to easily incorporate constraints constructed around the common variables as well as benchmark constraints. It is relatively computationally simple and does not require the intermediate step of estimating the unknown population level parameters associated with the common variables. It also produces weights that are

[1] The University of Southampton

always positive. We measure the relative bias of the proposed estimator in a series of simulations on a real dataset as well as on some purposively created data.

2. METHODS

2.1. Empirical Likelihood approach

Empirical Likelihood (EL) is a non-parametric method that uses the likelihood ratio function for inference. In this section, we briefly present how we use the EL approach to obtain point estimators and confidence intervals for population level parameters.

Suppose that two surveys are carried out independently in the same population. In each survey *t* the following variables are measured: a study variable y_t, an auxiliary variable x_t, for which a population level parameter is known and a common variable z, for which no population level parameters are known. Suppose that we wish to estimate some fixed, unknown population level parameters of interest, θ_1^N and θ_2^N, solutions to the following estimating equations:

$$\sum_{i\in U} g_{1i}(y_{1i},\theta_1) = 0,\ \sum_{i\in U} g_{2i}(y_{2i},\theta_2) = 0. \tag{1}$$

Consider the following combined empirical log-likelihood function for two samples:

$$l(m_1,m_2) = \sum_{i\in s_1} \log(m_{1i}) + \sum_{j\in s_2} \log(m_{2j}), \tag{2}$$

where $\boldsymbol{m}_t = (m_{t1}, m_{t2}, \dots, m_{tn_t})^T$. The values m_{ti} are unknown positive scale loads which need to be estimated [5].

2.2. Estimation of scale loads

The scale loads m_{ti} are estimated by the values which maximize (2) under a set of constraints, including benchmark and consistency constraints as well as a requirement that the estimated scale loads are positive. The constraints incorporate also the inclusion probabilities. Adding the inclusion probabilities to the system of constraints rather than putting them in the likelihood function is a key difference between our estimator and the method proposed by Wu [4]. One of the benefits of our approach is that it makes it possible to obtain Empirical Likelihood confidence regions for the point estimator, as explained in section 2.4.

2.3. Point estimation

The point estimators for θ_1^N and θ_2^N are obtained as the values which maximise the following log likelihood ratio function

$$\hat{r}(\theta_1,\theta_2) = 2\big(\ell(\hat{\boldsymbol{m}}_{\boldsymbol{1}},\hat{\boldsymbol{m}}_{\boldsymbol{2}}) - \ell(\hat{\boldsymbol{m}}^*_{\boldsymbol{1}},\hat{\boldsymbol{m}}^*_{\boldsymbol{2}},\theta_1,\theta_2)\big), \tag{3}$$

where $\ell(\hat{\boldsymbol{m}}^*_{\boldsymbol{1}},\hat{\boldsymbol{m}}^*_{\boldsymbol{2}},\theta_1,\theta_2) = \sum_i \log(\hat{m}^*_{1i}(\theta_1)) + \sum_i \log(\hat{m}^*_{2i}(\theta_2))$ and $\hat{m}^*_{ti}(\theta_t)$ are the values which maximise (2) subject to the same constraints as those imposed on $\hat{m}_{ti}$ and two additional constraints:

$$\textstyle\sum_{i\in s_1} \hat{m}_{1i}\, g_{1i}(y_{1i},\theta_1) = 0,\ \sum_{i\in s_2} \hat{m}_{2i}\, g_{2i}(y_{2i},\theta_2) = 0, \tag{4}$$

for given values of $\hat{\theta}_1$ and $\hat{\theta}_2$.

2.4. Confidence regions

Under some regularity conditions, the log likelihood ratio function (3) follows a χ^2 distribution asymptotically under $H_0: \theta_1 = \theta_1^N,\ \theta_2 = \theta_2^N$. This property allows to construct the (1−α) Wilk type confidence regions for θ_1 and θ_2 by selecting the values (θ_1, θ_2) which satisfy the following condition:

$$\hat{\mathrm{r}}(\theta_1, \theta_2) \leq \chi^2_{df=2,\alpha}\ . \tag{5}$$

3. RESULTS

Finite population performance of the proposed point estimator is compared with other existing methods: the GREG estimators of Zieschang [1] and Renssen and Nieuwenbroek [2] and the pseudo EL estimator presented by Wu [4]. We design two scenarios, one relying on artificial (generated) data and one using a real dataset. In the first scenario, we generate a dataset according to a model proposed by Wu and Rao [6]. In each of the samples, there is a different variable of interest, which follows a skewed distribution. We treat the generated dataset as a population. We select two independent samples and estimate parameters of interest using the proposed EL estimator, the GREG estimators of Zieschang [1] (ZG) and Renssen and Nieuwenbroek [2] (RN) and the pseudo EL estimator presented by Wu [4] (WU). Sampling and estimation is repeated 10 000 times. Samples are selected using random systematic sampling design. The relative bias is calculated for each estimator.

In the second set of simulations we use data from the 2006 British Expenditure and Food Survey [7]. The simulation process is the same as described above, i.e., in each of the 10 000 iterations, two independent samples are selected by systematic random sampling and estimates are calculated using the four estimators. In all the simulations, the number of people living in the household and the number of rooms in the household are used as auxiliary information with known population totals. Gross weekly income is the common variable with unknown population total. The study variables differ in each simulation. In simulation 7, the total gross expenditure is estimated from both samples. In simulation 8, the total expenditure on clothing and the total expenditure on housing are estimated from the first and the second samples respectively. In simulation 9, the total expenditure on clothing and the total expenditure on food are the parameters of interest. The following table shows relative biases of the estimators considered.

Table 1. Relative biases of the proposed Empirical Likelihood estimator (EL), Wu's Pseudo Empirical Likelihood estimator [3] (WU), GREG estimators proposed by Zieschang [1] (ZG) and Renssen and Nieuwenbroek [2] (RN).

	N	n_1	n_2	$\hat{\theta}_1^{(EL)}$	$\hat{\theta}_1^{(WU)}$	$\hat{\theta}_1^{(RN)}$	$\hat{\theta}_1^{(ZG)}$	$\hat{\theta}_2^{(EL)}$	$\hat{\theta}_2^{(WU)}$	$\hat{\theta}_2^{(RN)}$	$\hat{\theta}_2^{(ZG)}$
Generated data											
1	100000	1000	1000	0.01%	-0.02%	0.19%	-0.16%	-0.03%	-0.06%	-0.16%	-0.17%
2	100000	200	400	0.01%	0.01%	-0.99%	-0.76%	-0.01%	-0.11%	-0.37%	-0.53%
3	100000	200	200	0.01%	0.13%	-0.76%	-0.64%	0.02%	-0.06%	-0.62%	-0.68%
4	2500	160	160	0.00%	-0.04%	-1.14%	-0.98%	-0.02%	-0.12%	-0.97%	-1.09%
5	2500	140	260	-0.01%	0.15%	-1.28%	-0.98%	0.00%	-0.13%	-0.51%	-0.72%
6	2500	240	240	0.01%	0.13%	-0.76%	-0.64%	0.02%	-0.06%	-0.62%	-0.68%
Expenditure and Food Survey data											
7	6645	500	500	-0.11%	0.07%	-0.57%	-0.31%	-0.05%	0.21%	-0.56%	-0.20%

8	6645	500	500	0.38%	0.44%	-0.07%	0.03%	0.06%	0.06%	-0.38%	-0.35%
9	6645	500	500	0.07%	0.07%	-0.38%	-0.30%	0.01%	0.01%	-0.36%	-0.32%

The table presented above shows that in all scenarios, the relative bias of the proposed estimator is of an acceptable size. In most cases, the proposed estimator has smaller relative bias than the alternative estimators, especially the GREG estimators. Note that when the sample size is small, the GREG estimators show relative bias close to 1%, while the relative bias of the proposed EL estimator remains lower. We conclude that the EL point estimator is asymptotically unbiased.

The main advantage of the proposed method is not in the performance of the point estimator, but in the possibility of obtaining asymmetric EL confidence regions, defined by the shape of the log likelihood ratio function (3). An example of such a confidence region is presented in Figure 1. Note that in each survey there is a different parameter of interest.

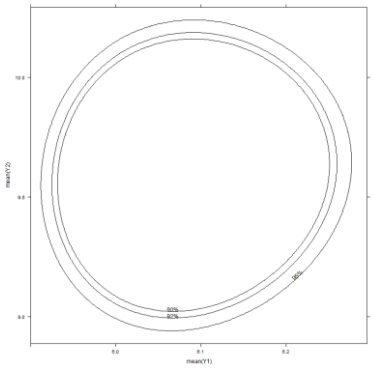

Figure 1. An example of a confidence region for two parameters
Data generated according to a model proposed in [6]

4. CONCLUSIONS

The proposed method allows to easily combine different datasets when common variables are measured in both of them and to ensure that the point estimates for the common variable are consistent across surveys. Additional benchmark constraints may also be incorporated. The method allows to obtain point estimators for a wide class of parameters which may be expressed as solutions to estimating equations, such as means, ratios or quantiles. The confidence regions are constructed using the χ^2 approximation of the log likelihood ratio function. Under the tested scenarios, the proposed point estimator shows satisfactory performance compared to the other available estimators in terms of relative bias.

REFERENCES

[1] K. D. Zieschang, Sample weighting methods and estimation of totals in the consumer expenditure survey. Journal of the American Statistical Association, 85(412), (1990), 986–1001.

[2] R.H. Renssen and N.J. Nieuwenbroek. Aligning estimates for common variables in two or more sample surveys. Journal of the American Statistical Association, 92(437), (1997), 368–374.

[3] Takis Merkouris. Combining independent regression estimators from multiple surveys. Journal of the American Statistical Association, 99(468), (2004), 1131-1139.

[4] Ch. Wu, Combining information from multiple surveys through the empirical likelihood method, Canadian Journal of Statistics, 32(1) (2004), 15–26.

[5] Y.G. Berger and O. De La Riva Torres. Empirical likelihood confidence intervals for complex sampling designs. Southampton Statistical Sciences Research Institute, (S3RI Methodology Working Papers), (2012).

[6] Ch. Wu and J.K. Rao, Pseudo Empirical Likelihood Ratio Confidence Intervals for Complex Surveys, The Canadian Journal of Statistics, 34, (2006), 359-375.
[7] Office for National Statistics and Department for Environment, Food and Rural Affairs, Expenditure and Food Survey, 2006 [computer file]. 3rd Edition. Colchester, Essex: UK Data Archive [distributor], July 2009. SN: 5986.