Study on the performance of four variance estimators for logistic GREG estimator for domains

Mikko Myrskylä[1]

Let $U=\{1,2,\ldots,N\}$ be the population under study. We estimate frequencies of class A in domains $U^{(d)}, d=1,2,\ldots,D$; these are defined by $\sum_{U^{(d)}} I_{\{i\in A\}} \equiv T^{(d)}$, where $I_{\{i\in A\}}=1$ if i belongs to A and 0 otherwise. Sampling vector $\underline{\mathbf{I}}=(I_1,I_2,\ldots,I_N)$ has distribution $p(\mathbf{I})$, and realisation $\mathbf{I}=(k_1,k_2,\ldots,k_N)$ of $\underline{\mathbf{I}}$ is the sample so that unit i is sampled k_i times. Sample set and sample set in domain are $s=\{i:k_i>0\}$ and $s\cap U^{(d)}\equiv s^{(d)}$, respectively. Sampling weights are $w_i=k_i/E(I_i)$. Under this notation, generalised regression (GREG) estimator for $T^{(d)}$ is $\hat{T}^{(d)}=\sum_{U^{(d)}}\hat{y}_i-\sum_{s^{(d)}} w_i\hat{e}_i$, where $\hat{e}_i=I_{\{i\in A\}}-\hat{y}_i$ and $\hat{y}_i$ is prediction from some assisting statistical model. If the assisting model is linear with fixed effects, we call this estimator GREG-lin, if logistic with fixed effects, GREG-log, respectively.

The accuracy of GREG estimator with respect to functional form of the model has been studied in [4], [5], [2] and [6]. Results indicate that the choice of model form is important: for class frequencies, GREG-log is more accurate than GREG-lin. However, the Sen-Yates-Grundy (SYG) variance estimator $\sum\sum_{U^{(d)}} w_i\hat{e}_i w_k\hat{e}_k$, which is often used for GREG-lin, seems to underestimate the variance of GREG-log, and especially if i) domains are minor[4], ii) if the assisting model is complex in the sense that it has a large number of covariates[6], and iii) if auxiliary information is very strong[6]. In addition, variance of SYG often becomes unbearably large in these cases[6]. Thus, the need for better variance estimator is evident.

Using Monte Carlo simulation, I study the performance of four variance estimators for logistic GREG estimator for domains under simple random sampling without replacement (SRSWOR). The baseline estimator is SYG, which under the SRSWOR design is $N^2n^{-1}(1-n/N)\hat{S}^2_{e^{(d)}}$. The reason for this estimator failing as well as the performance of three alternative variance estimators are studied. These estimators are the standard iid bootstrap[7], bootstrap without replacement[1],[3], and delete-one jackknife[8]. These estimators are externally scaled in a standard way so that the linear condition (unbiasedness in the case of linear estimator) is fulfilled. Performance of the variance estimators is then compared by means of bias, MSE, and coverage rate.

References

[1] Bickel, P. J. – Freedman, D. A. (1984): Asymptotic Normality and the Bootstrap in Stratified Sampling. The Annals of Statistics, Vol. 12, No. 2, 470-482.

[2] Duchesne, P. (2003). Estimation of a Proportion with Survey Data. Journal of Statistics Education Vol. 11, No. 3.

[1] Statistics Finland, Box 5V FI-00022, email: mikko.myrskyla@stat.fi

[3] Gross, S.T. (1980). Median estimation in sample surveys, ASA Proc. of Survey Research Methods Sect.: 181-4.
[4] Lehtonen R. and Veijanen A. (1998). Logistic generalized regression estimators. Survey Methodology 24, 51-55.
[5] Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. Survey Methodology 29, 33-44
[6] Myrskylä, M. (2004). Estimation of class frequencies with micro level auxiliary information. An application to Finnish Labour Force Survey. Master's thesis (unpublished), University of Jyväskylä.
[7] Sitter, R.R. (1992). A Resampling Procedure For Complex Survey Data, Journal of the American Statistical Association, 87, 755-765.
[8] Wolter, K. M.(1985). Introduction to Variance Estimation. New York: Springer-Verlag.