Unbiased Survey Estimation of Highly Skewed Populations

Forough Karlberg
National Social Insurance Board
Department of Research, Analysis and Statistics
SE-103 51 STOCKHOLM, Sweden
Forough.Karlberg@rfv.sfa.se

1. Introduction

There are many surveys of populations that are very skew and thus contain a number of extreme values. This is particularly true in surveys of business enterprises and agriculture, as well as for surveys of personal income and fortune. Although the values are extreme, they need not necessarily be false; extremely large observations are a natural component in e.g. establishment survey populations, which often have distributions that are skewed to the right. In this paper, we will address the estimation of the population total of a highly skewed survey variable, for which the occurrence of outliers lies in the nature of the variable.

Previous efforts in the field of outlier treatment include outlier robust inference (see Barnett and Lewis, 1994), and weight and value modification strategies (see Chambers and Kokic, 1993); an example of the latter strategy is Winsorization (see e.g. Rivest, 1994). Robust methods, as well as weight and value modification strategies, are most suitable for situations where the outliers encountered are non-representative. Otherwise, negative bias is introduced if one merely focuses on the reduction of the impact of outliers that are present in the sample, without attempting to compensate for outliers that are present in the population, but not in the sample. This problem generally occurs when one-sided Winsorization is performed; the variance decreases at the cost of a large negative bias.

To reduce the moderate negative impact of the absence of outliers from the sample (along with the large positive impact of their presence), Karlberg (2001) develops a lognormal superpopulation model. The model estimator of the population total, which is applicable to strictly positive survey variables, assumes that outliers are likely to be present in the population, even when there are no outliers in the sample. The lognormal model has been extended by Karlberg (2000) to a lognormal-logistic model, that allows for survey variables that, while being highly skewed, have zero-valued units as well.

The estimator of Karlberg (2001) is only approximately model unbiased, as the lognormal shape parameter σ is unknown and has to be estimated. Here, we present a truly model unbiased estimator of the population total of a highly skewed survey variable; this estimator is applicable when $\sigma>0$ is known.

2. The lognormal model estimator

We are interested in estimating

$$T = \sum_{i=1}^{N} Y_i = \sum_{i\in s} Y_i + \sum_{i\in r} Y_i ,$$

the population total of the survey variable Y, which is assumed to be highly skewed to the right. We also assume that for each unit *i*, Y_i is strictly positive, and that the values of this survey variable are known only for a sample *s* consisting of *n* units out of the *N* units of the population. We denote the logarithm of Y_i by Z_i. Further, we assume that we have access to *k* auxiliary variables; *k* may be any non-negative integer (including 0) such that $n > k$. The auxiliary variable vector $\mathbf{X}_i$ is known for all *i*. The survey variable values for the *N*-*n* units belonging to *r*, the non-sample part of the population are estimated by

$$\hat{T} = \sum_{i\in s} Y_i + \sum_{i\in r} \hat{Y}_i \qquad \text{where} \qquad \hat{Y}_i = \exp\left(\hat{Z}_i\right) C_i .$$

Here $\hat{Z}_i = \hat{\boldsymbol{\beta}}\mathbf{X}_i$ is a model unbiased estimator of Z_i, while C_i is a bias correction factor.

3. Application of the model estimator

A simulation study has been performed on data from the Australian Agricultural and Grazing Industries Survey (AAGIS) described by Chambers (1996). The AAGIS data have previously been analyzed in this context by Karlberg (2000, 2001). We have used the two study variables *DSE* (Dry Sheep Equivalent, number of sheep + 8 × number of beef cattle + 12 × crops area in hectares) and *Beef cattle* (Number of beef cattle on farm). For these variables, 1000 simple random samples have been drawn for each n. For each sample, we have estimated T, with (k=1) and without (k=0) the auxiliary variable *Farm area* (in hectares). The model estimator bias and mean square error are presented in Table 1, along with the estimator efficiency (vs. a design-based regression estimator).

We see that the lognormal superpopulation model estimator generally works well, (as the relative efficiency exceeds 1.0 in most cases) when applied to survey variables that are known to be skewed to the right (e.g. economic variables). In contrast to design-based and traditional downweighting/deflating estimators, the estimator presented here has the particular advantage that, even when extremely large values are absent from the sample, this is compensated for by the assumed lognormal structure (which implies the presence of extremely large values).

Table 1. Properties of the model estimator when applied to *DSE* and *Beef cattle.*

Study variable(*Y*)	***T***	***k***	s	***n***	**Relative bias (%)**	**Relative MSE (%)**	**Relative efficiency vs. regression estimator**
DSE	$2.52\cdot 10^7$	0	1.740	50	-8.2	15.9	4.96
N=1652				100	-8.6	13.4	3.98
(*all farms in*		1	1.423	50	5.4	23.8	3.34
survey)				100	4.0	15.9	1.27
Beef cattle	$1.50\cdot 10^6$	0	1.069	50	-4.0	24.2	5.94
N=1149				100	-5.0	16.7	6.29
(*farms with one or*		1	0.678	50	8.7	40.2	0.68
more beef cattle)				100	9.8	27.8	0.93

REFERENCES

Barnett, V. and Lewis, T.

1994 *Outliers in statistical data*, 3*rd edition*. New York: John Wiley.

Chambers, R. L.

1996 Robust case-weighting for multipurpose establishment surveys, *Journal of Official Statistics* 12, 3-32.

Chambers, R. L. and Kokic, P. N.

1993 Outlier robust sample survey inference, Bulletin of the International Statistical Institute 55, Proceedings of the 49th session of the International Statistical Institute, Firenze, 55-72.

Karlberg, F.

2000 Survey estimation of highly skewed populations in the presence of zeroes, *Journal of Official Statistics* 16, 229-241.

2001 Population total prediction under a lognormal superpopulation model, *Metron* 58, 53-80.

Rivest, L. P.

1994 Statistical properties of Winsorized mean for skewed distributions, *Biometrika* 81, 373-383.

RÉSUMÉ

La population totale d'un variable orienté est estimée (en utilisant des variables auxilliares) suivant l'assomption de superpopulation de lognormalité. On obtient une prédiction non biaisée d'un modèle de la population totale, ainsi qu'une estimation non biaisée de la variation de l'erreur de sa prédiction.