Journal of Official Statistics, Vol. 16, No. 3, 2000, pp. 229–241

Survey Estimation for Highly Skewed Populations in the Presence of Zeroes

Forough Karlberg[1]

Estimation of the population total of a highly skewed survey variable from a small sample using straightforward methods is problematic for two reasons: (i) when there are no extreme values in the sample, too small estimates will be obtained, and (ii) if extreme values are sampled, the estimates will become grotesquely large. Traditional methods for outlier treatment will usually compensate for outliers in the sample, thereby avoiding (ii), whereas the small negative bias of (i) will persist. Here, an estimator based on a lognormal-logistic superpopulation model is proposed.

A particular strength of the model estimator is that the lognormal structure of the survey variable is used for estimation – even in the absence of extremely large values in the sample. Another advantage of the model estimator is that it can be applied to situations in which the survey variable, while highly skewed, may assume the value zero for a number of units.

The model estimator is applied to an agricultural survey variable in a simulation study, in which it is compared to a design-based (regression) estimator as well as a Winsorization-based estimator specifically constructed for outlier treatment. The simulation results indicate that the lognormal-logistic model estimator constitutes a sensible alternative to the other estimators, in particular when the sample size is small.

Key words: Extreme values; model-based inference; superpopulation; lognormal distribution.

1. Introduction

There are many surveys of very skewed populations which contain a few extreme values. This is particularly true of surveys of business enterprises and agriculture, as well as of surveys of personal income and fortune. Although the values are extreme, they need not necessarily be false; extremely large observations constitute a natural ingredient in e.g., establishment survey populations, which often are skewed to the right. In this article, 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.

In the Australian Agricultural and Grazing Industries Survey (data from which has previously been used by e.g., Chambers 1996), the variable *Beef cattle* has many extreme values. From the symmetric distribution of the logarithms of the positive *Beef cattle* values (see Figure 1), it is evident, however, that the extreme values fit rather well into the overall distribution of the variable; in the terminology of Chambers (1986) they are referred to as

[1] Biostatistics and Data Management, R&D Sweden, Pharmacia Corporation, SE-112 87 Stockholm, Sweden. E-mail: forough.karlberg@eu.pnu.com

representative outliers. In this article, we will study estimation when the extreme values are in line with the overall distribution of the data, i.e., representative.

Previous efforts within outlier treatment include outlier robust inference (see e.g., Chambers 1986 and Lee 1995 for applications), which is discussed by Barnett and Lewis (1994), and weight modification (see for instance Hidiroglou and Srinath 1981) or value modification (see for instance Kokic and Smith 1998a, who discuss one-sided Winsorization, in which the largest observations are replaced by a smaller value) strategies. The two latter approaches are compared by Chambers and Kokic (1993). Robust methods, as well as weight and value modification strategies, are most suitable for situations where the outliers encountered are nonrepresentative.

Thorburn (1991) suggested the use of a lognormal superpopulation model for estimation in this context. Karlberg (2000) has extended the applicability of the model-based estimator of Thorburn to situations in which there are a number of auxiliary variables available. Still, under this lognormal superpopulation model, the survey variable only can have strictly positive values. There are many survey variables like *Beef cattle* that, while having extremely large values for some units, also are equal to 0 for other units.

A conceivable extension of the lognormal superpopulation model is to introduce (in addition to the scale and shape parameters) a third shift parameter γ. Initially, this approach seems appealing, since if $\gamma < 0$, there is a positive probability density for 0;

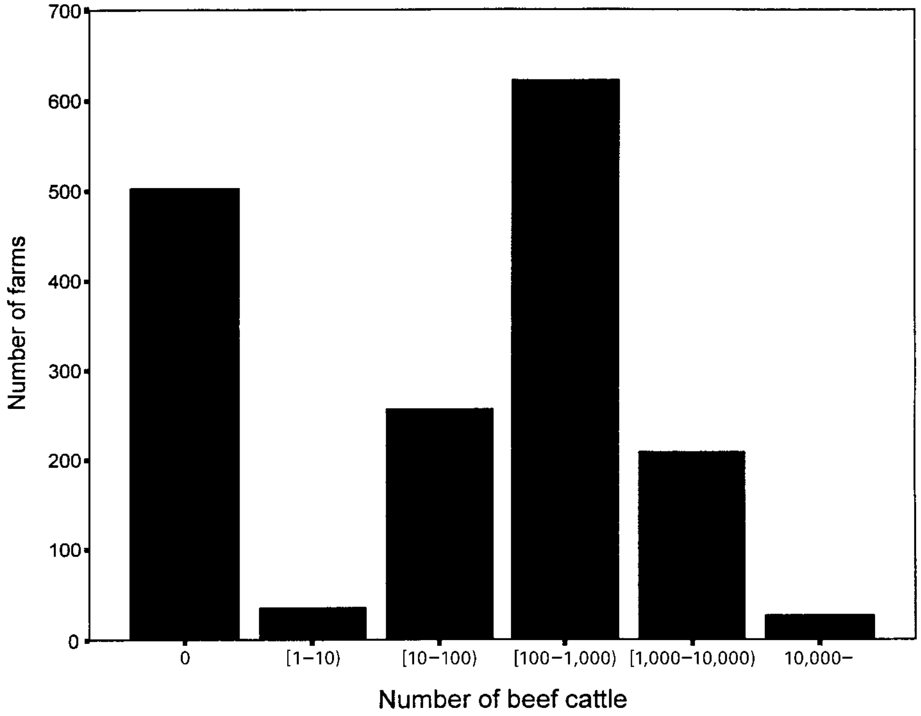

Fig. 1. Distribution of the variable Beef cattle for 1,652 Australian farms.

the zero-valued observations appear to be accommodated by this distribution. However, while the positive values of *Beef cattle* can be modeled rather well by a lognormal distribution (with the mode somewhere in the interval 100–1,000), the entire empirical distribution of *Beef cattle* (including the zeroes) seems to be *bimodal*, and thus not very well accommodated by a (unimodal) 3-parameter lognormal distribution. In addition, the qualitative difference between units with positive and zero-valued observations (e.g., farms with and without beef cattle) is missed out by this approach. From Figure 1, we see that many farms have no cattle at all; *Beef cattle* is thus equal to 0 for those farms. Such variables should be modeled by a distribution that has a large *point probability* at the value 0, rather than by a distribution that has a positive *probability density* within the interval $(\gamma, 0)$. Here, we consequently address this by combining the lognormal model of Karlberg (2000) with a logistic component, which determines whether the survey variable is positive or zero.

In Section 2, we derive the approximately model unbiased population total estimator. We use model-based inference under a lognormal-logistic superpopulation model, and we attempt to demonstrate that this is a sensible alternative to traditional methods. In a simulation study, presented in Section 3, two relevant alternative estimators are compared to the model estimator. A concluding discussion is offered in Section 4.

2. The Lognormal-Logistic Model Estimator

2.1. Preliminaries

The approach used for inference here would, in the terminology of Brewer and Mellor (1973), be regarded as model-based *descriptive* inference, since the target population for inference is the finite population from which the sample was drawn. This is to be contrasted with model-based *analytic* inference (cf. Deming 1950), where the target is some other population (and even a perfect census thus leaves uncertainty).

When model-based inference is used, it is assumed that the survey variable has been generated according to some superpopulation model. If we can find a distribution model that closely resembles the distribution of the survey variable, a model-based estimator will generally have a smaller mean squared error than a design-based one. In essence, the model-based approach draws its strength from the knowledge of the survey variable distribution in such situations.

Strictly speaking, the unobserved values of the survey variable should be viewed as stochastic, and being distributed according to the assumed superpopulation model. What we refer to as estimation in this article, is therefore generally referred to as model-based predictive inference in the terminology of Smith (1999). However, we will use the more common term ‘‘estimation’’ throughout.

Throughout this article, we are (like Brewer 1995) going to use the notation of Särndal et al. (1992), letting $E_\xi(\bullet)$ and $Var_\xi(\bullet)$ denote the expected value and variance, respectively, of the quantity $(\bullet)$ under the superpopulation model ξ. We use the terms *moments* and *probabilities* to refer to model moments and model probabilities; they are thus not related to the sampling design. Similarly, by referring to an estimator as being unbiased, we indicate that it is model unbiased.

2.2. The superpopulation model

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 assume that for each unit i, Y_i is non-negative, 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. Further, we assume that for $i = 1, \ldots, N$, we have access to the auxiliary variable vector $\mathbf{X}_i = (1, X_{i1}, \ldots, X_{ik})'$; k, the number of auxiliary variables, may be any non-negative integer (including 0).

According to our lognormal-logistic superpopulation model ξ, the survey variable

$$Y_i = \tilde{Y}_i \Delta_i$$

is a product of a lognormal component $\tilde{Y}_i$ and a logistic component Δ_i.

For the lognormal model component, we adopt the lognormal model of Karlberg (2000). For all i, we assume that $\tilde{Z}_i = \ln(\tilde{Y}_i)$, conditional on $\mathbf{X}_i$, follows a normal distribution with mean

$$\mu_i = \mu(\mathbf{X}_i) = \boldsymbol{\alpha}\mathbf{X}_i$$

and variance

$$\sigma_i^2 = \sigma^2 v(\mathbf{X}_i)$$

Here $v_i = v(\mathbf{X}_i)$ is a known function of $\mathbf{X}_i$.

For the logistic model component, we assume that Δ_i, conditional on $\mathbf{X}_i$, are independently Bernoulli (p_i) distributed for all i, where

$$P(Y_i > 0 \big| \mathbf{X}_i) = P(\Delta_i = 1 \big| \mathbf{X}_i) = p_i = \frac{\exp(\boldsymbol{\beta}\mathbf{X}_i)}{1 + \exp(\boldsymbol{\beta}\mathbf{X}_i)}$$

Here $\boldsymbol{\beta}$ is a vector of unknown logistic model parameters. The three unknown model parameters to be estimated are thus $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and σ.

A central assumption is that the sampled values of the survey variable follow the superpopulation model and that the sampling is uninformative. The estimation method used here holds as long as the sampling design only depends on the survey variable through the auxiliary variables. It must not depend on the survey variable in any other way. Generally, the preferred sampling design thus depends on the known auxiliary information. Here, however, we only focus on the model and not on the sample design; as a consequence, the complexities in the calculation process stem from the assumed superpopulation distribution.

Throughout this article, we regard the auxiliary variable values as fixed, and in what follows we will condition implicitly on those auxiliary variables: that is, we will use $E_\xi(\bullet)$ to mean $E_\xi(\bullet \big| \mathbf{X}_1, \ldots, \mathbf{X}_N)$.

2.3. Estimating the lognormal component

Letting $s_+ = \{i \in s : Y_i > 0\}$ denote the subset of the sample for which the survey variable is positive, and letting $n_+ = |s_+| = \sum_{i \in s} \Delta_i$ denote the number of positive sample units, we assume that $n_+ > k + 1$.

We let $\mathbf{Z}_{s+}$, $\mathbf{X}_{s+}$ and $\mathbf{V}_{s+}$ denote the vector of the logarithmed values of the strictly positive survey variable values of the sample, the matrix of auxiliary variables and the diagonal matrix containing the variance coefficients raised to the -1:th power, respectively. We assume that $\mathbf{X}_{s+}\mathbf{V}_{s+}\mathbf{X}'_{s+}$ is positive definite (and thus non-singular) and may thus define

$$a_{ij} = \mathbf{X}'_i(\mathbf{X}_{s+}\mathbf{V}_{s+}\mathbf{X}'_{s+})^{-1}\mathbf{X}_j$$

We further assume that when $n, N \to \infty$ in a way such that $(N - n) \to \infty$, then $n_+ \xrightarrow{p} \infty$ and

$$a_{ij} \xrightarrow{p} 0 \tag{1}$$

for each pair of nonsample units, $i, j \in r$. For simplicity, we will regard the sample vector of Δ_i-values, $\boldsymbol{\Delta}_S$, as fixed in the rest of this section; we will thus use $E_\xi(\bullet)$ when referring to $E_\xi(\bullet|\boldsymbol{\Delta}_s)$.

We define, for each nonsample unit $i \in r$, an unbiased estimator $\hat{\tilde{Z}}_i$ of $\tilde{Z}_i$ (if $\Delta_i = 1$, i.e., if Z_i exists) as

$$\hat{\tilde{Z}}_i = \hat{\boldsymbol{\alpha}}\mathbf{X}_i$$

where the parameter $\boldsymbol{\alpha}$ is ML estimated by

$$\hat{\boldsymbol{\alpha}} = (\mathbf{Z}_{s+}\mathbf{V}_{s+}\mathbf{X}'_{s+})(\mathbf{X}_{s+}\mathbf{V}_{s+}\mathbf{X}'_{s+})^{-1}$$

We now have that

$$E_\xi(\exp(\hat{\tilde{Z}}_i)|\Delta_i = 1) = \exp\left(\boldsymbol{\alpha}\mathbf{X}_i + \frac{\sigma^2 a_{ii}}{2}\right) \neq \exp\left(\boldsymbol{\alpha}\mathbf{X}_i + \frac{\sigma^2 v_i}{2}\right) = E_\xi(\exp(\tilde{Z}_i)|\Delta_i = 1) \tag{2}$$

$\text{Exp}(\hat{\tilde{Z}}_i)$ is thus a severely biased estimator of $\tilde{Y}_i$ and the bias is a function of the shape parameter σ^2. We estimate the lognormal shape parameter by

$$\hat{\sigma}^2 = \frac{\mathbf{Z}_{s+}\mathbf{V}_{s+}\mathbf{Z}'_{s+} - \hat{\boldsymbol{\alpha}}(\mathbf{X}_{s+}\mathbf{V}_{s+}\mathbf{X}'_{s+})\hat{\boldsymbol{\alpha}}'}{n_+ - k - 1}$$

which is a bias-corrected estimator of σ^2. If we were to use $\exp(\frac{1}{2}\hat{\sigma}^2(v_i - a_{ii}))$ to compensate for the bias demonstrated in (2), we would obviously introduce a new bias stemming from the randomness of $\hat{\sigma}^2$. In order to compensate for this bias, we will use the moment generation function of $\hat{\sigma}^2$; an approximation of this function is given in Lemma 1. We then use this approximation in Lemma 2 for deriving an approximately ξ-unbiased estimator of Y_i; this estimator has a bias that is an order of approximation smaller than the simple ''plug-in'' estimator $\exp(\hat{\boldsymbol{\alpha}}\mathbf{X}_i + \frac{1}{2}\hat{\sigma}^2 v_i)$.

Lemma 1. *The moment generating function of the estimated variance can be written as*

$$E_\xi(\exp(\hat{\sigma}_i^2 t)) \approx \exp\left(\sigma_i^2 t + \frac{\sigma_i^4 t^2}{n_+}\right)$$

for large n_+.

Proof. It is well known (see e.g., Casella and Berger 1990, p. 569) that the variance estimator $\hat{\sigma}^2$ follows a χ^2-distribution

$$\frac{(n_+ - k - 1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n_+ - k - 1)}$$

when the observations (i.e., $\tilde{Z}_i$) follow a normal distribution. Using the moment generating function of the χ^2-distribution, we have that

$$E_\xi(\exp(\hat{\sigma}_i^2 t)) = \left(1 - \frac{2\sigma_i^2 t}{n_+ - k - 1}\right)^{-\frac{n_+ - k - 1}{2}}$$

Using both binomial and exponential expansion, it can be proven that

$$\left(1 - \frac{2\sigma_i^2 t}{n_+ - k - 1}\right)^{\frac{n_+ - k - 1}{2}} \approx \exp\left(-\sigma_i^2 t - \frac{\sigma_i^4 t^2}{n_+}\right)$$

for large n_+. (Q.E.D.)

Lemma 2. *Conditional on $\boldsymbol{\Delta}_s$, we have*

$$\hat{\tilde{Y}}_i = \exp(\hat{\tilde{Z}}_i) \exp\left(\frac{\hat{\sigma}^2}{2}(v_i - a_{ii}) - \frac{\hat{\sigma}_i^4}{4n_+}\right)$$

is an approximately ξ-unbiased estimator of $\tilde{Y}_i$ for all nonsample units $i \in r$.

Proof. Using the fact that $\hat{\sigma}_i^2$ and $\hat{\tilde{Z}}_i$ are ξ-independent (see Casella and Berger 1990, p. 569), it follows that

$$\begin{aligned}
&E_\xi(\hat{\tilde{Y}}_i - \tilde{Y}_i \big| \Delta_i = 1) \\
&= E_\xi(\exp(\hat{\tilde{Z}}_i) \big| \Delta_i = 1) E_\xi\left(\exp\left(\frac{\hat{\sigma}^2}{2}(v_i - a_{ii}) - \frac{\hat{\sigma}_i^4}{4n_+}\right) \big| \Delta_i = 1\right) - E_\xi(\exp(\tilde{Z}_i) \big| \Delta_i = 1) \\
&\approx E_\xi(\exp(\hat{\tilde{Z}}_i) \big| \Delta_i = 1) \exp\left(-\frac{\sigma_i^4}{4n_+}\right) E_\xi\left(\exp\left(\frac{\hat{\sigma}^2}{2}(v_i - a_{ii})\right) \big| \Delta_i = 1\right) \\
&- E_\xi(\exp(\tilde{Z}_i) \big| \Delta_i = 1)
\end{aligned} \tag{3}$$

In the sequel, we are going to use the approximation

$$E_\xi\left(\exp\left(\frac{\hat{\sigma}_i^4}{n_+}\right)\right) \approx \exp\left(\frac{\sigma_i^4}{n_+}\right)$$

Using (1) and Lemma 1 with $t = (v_i - a_{ii})/2$, we have that

$$E_\xi\left(\exp\left(\frac{\hat{\sigma}^2}{2}(v_i - a_{ii})\right)\right) \approx \exp\left(\frac{\sigma^2}{2}(v_i - a_{ii}) + \frac{\sigma^4}{4n_+} v_i^2\right)$$

From this, and by using the right-hand side of the inequality of (2) in (3), we have that

$$\begin{aligned}
E_\xi(\hat{\tilde{Y}}_i - \tilde{Y}_i \big| \Delta_i = 1) &\approx \exp\left(\boldsymbol{\alpha}\mathbf{X}_i + \frac{\sigma^2 a_{ii}}{2}\right) \exp\left(\frac{\sigma^2}{2}(v_i - a_{ii})\right) - \left(\exp\left(\boldsymbol{\alpha}\mathbf{X}_i + \frac{\sigma^2 v_i}{2}\right)\right) \\
&= 0.
\end{aligned}$$

For each nonsample unit $i \in r$, it follows that if the survey variable Y_i is positive, then $\hat{Y}_i$ is an approximately model unbiased estimator of Y_i.

(Q.E.D.)

2.4. *Estimating the logistic component*

Using $\hat{\boldsymbol{\beta}}$, the estimator of the logistic parameter $\boldsymbol{\beta}$, obtained using the ML method as described in e.g., McCullagh and Nelder (1989, p. 116), we compute

$$\hat{p}_i = \frac{\exp(\hat{\boldsymbol{\beta}}\mathbf{X}_i)}{1 + \exp(\hat{\boldsymbol{\beta}}\mathbf{X}_i)}$$

the estimated probability of a positive value of Y_i for all nonsample units $i \in r$. From McCullagh and Nelder (1989, p. 119) we have

$$E_\xi(\hat{p}_i) - E_\xi(\Delta_i) = E_\xi(\hat{p}_i) - p_i = O(n^{-1}) \tag{4}$$

Thus, $\hat{p}_i$ is a biased estimator of p_i. In the sequel, we will however work under the assumption that the ξ-bias of $\hat{p}_i$, i.e., the residual term of (4), is negligible.

2.5. *An approximately unbiased population total estimator*

To obtain an unbiased estimator of the survey variable, we assume that $\hat{p}_i$ and $\hat{\hat{Y}}_i$ are uncorrelated. This is not exactly true, but as can be seen from Section 3, the estimator performs well in reality, so the error made by this assumption is in no way crucial. From the zero correlation, we have for all $i \in r$, that

$$\begin{aligned} E_\xi(\hat{p}_i\hat{\hat{Y}}_i) &= E_\xi(E_\xi(\hat{p}_i\hat{\hat{Y}}_i \big| \boldsymbol{\Delta}_s)) \\ &= E_\xi(\hat{p}_i E_\xi(\hat{\hat{Y}}_i \big| \boldsymbol{\Delta}_s)) \\ &= E_\xi(\hat{p}_i)E_\xi(E_\xi(\hat{\hat{Y}}_i \big| \boldsymbol{\Delta}_s)) \end{aligned} \tag{5}$$

From the conditional approximate unbiasedness of $\hat{\hat{Y}}_i$ proven in Lemma 2, we have that

$$E_\xi[E_\xi(\hat{\hat{Y}}_i - \tilde{Y}_i \big| \Delta_i = 1, \boldsymbol{\Delta}_s) \big| \Delta_i = 1] \approx E_\xi(0) = 0$$

Hence

$$E_\xi(\hat{\hat{Y}}_i \big| \Delta_i = 1) \approx E_\xi(\tilde{Y}_i \big| \Delta_i = 1) = \exp\left(\boldsymbol{\alpha}\mathbf{X}_i + \frac{\sigma_i^2}{2}\right) \tag{6}$$

Using (4) and (6) in (5), we obtain

$$E_\xi(\hat{p}_i\hat{\hat{Y}}_i) \approx p_i \exp\left(\boldsymbol{\alpha}\mathbf{X}_i + \frac{\sigma_i^2}{2}\right) = E_\xi(Y_i)$$

An approximately model unbiased estimator of the population total of the survey variable Y_i is thus given by

$$\hat{T} = \sum_{i \in s} Y_i + \sum_{i \in r} \hat{p}_i \hat{\hat{Y}}_i$$

2.6. *Estimation error variance of the population total estimator*

Letting $\delta_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma)$ denote the covariance between Y_i and Y_j, it is obvious that $\delta_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) = 0$ for $i \neq j$. Using the properties of the lognormal and Bernoulli distributions, we obtain

$$\begin{aligned}\delta_{ii}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) &= Var_{\xi}(\Delta_i \tilde{Y}_i)\\ &= E_{\xi}(\Delta_i^2 \tilde{Y}_i^2) - (E_{\xi}(\Delta_i \tilde{Y}_i))^2\\ &= P(\Delta_i = 1)E_{\xi}(\tilde{Y}_i^2 | \Delta_i = 1) - (P(\Delta_i = 1)E_{\xi}(\tilde{Y}_i | \Delta_i = 1))^2\\ &\approx p_i \exp(2\boldsymbol{\alpha}\mathbf{X}_i + 2\sigma_i^2) - p_i^2 \exp(2\boldsymbol{\alpha}\mathbf{X}_i + \sigma_i^2)\\ &= p_i \exp(2\boldsymbol{\alpha}\mathbf{X}_i + \sigma_i^2)(\exp(\sigma_i^2) - p_i)\end{aligned}$$

Defining

$$\begin{aligned}\Omega(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma) = \sum_{i,j \in r} & p_i p_j \exp\left(\boldsymbol{\alpha}(\mathbf{X}_i + \mathbf{X}_j) + \frac{\sigma_i^2 + \sigma_j^2}{2}\right)\\ & \times \left(b_{ij} \exp\left(\frac{\sigma^2}{2}(A_{ij}(\boldsymbol{\beta}) + A_{ji}(\boldsymbol{\beta})) + \frac{\sigma_i^2 \sigma_j^2}{2n_{.}(\boldsymbol{\beta})}\right) - 1\right)\\ & + \delta_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma)\end{aligned}$$

where

$$A_{ij}(\boldsymbol{\beta}) = E_{\xi}(a_{ij})$$

and

$$n_{\bullet}(\boldsymbol{\beta}) = E_{\xi}(n_{+})$$

it can be proven that $Var_{\xi}(\hat{T} - T) \approx \Omega(\boldsymbol{\alpha}, \boldsymbol{\beta}, \sigma)$. The estimation error variance can thus be estimated using the ''plug-in'' expression $\Omega(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}, \hat{\sigma})$. One could also estimate the variance by using bootstrap technology (see Shao and Tu 1995), or some version of the robust variance estimator of Royall and Cumberland (1978).

3. Application of the Model Estimator

3.1. *Alternative estimators*

As the first alternative estimator to compare to the model estimator we use a standard linear regression (LR) estimator (see Särndal et al. 1992)

$$\hat{T}_{LR} = \sum_{i \in s} Y_i + \hat{\boldsymbol{\beta}}_{LR} \sum_{i \in r} \mathbf{X}_i$$

where

$$\hat{\boldsymbol{\beta}}_{LR} = \mathbf{Y}_s \mathbf{X}_s' (\mathbf{X}_s \mathbf{X}_s')^{-1}$$

is the ML estimator of the regression coefficient vector $\boldsymbol{\beta}_{LR}$. Here, the vector $\mathbf{Y}_s$ contains the sample values of the survey variable. For obvious reasons, we have to assume that

$\mathbf{X}_s'\mathbf{X}_s$ is nonsingular. Since $\hat{T}_{LR}$ is virtually design unbiased in many applications (as noted by e.g., Särndal et al. 1992), and thus does not depend on the superpopulation model ξ, it will not be biased when the survey variable is not distributed according to ξ; on the other hand, it is likely to be inefficient (although unbiased) in comparison to a model-based estimator when ξ holds.

A comparison of the model estimator to estimators specially designed for dealing with outliers would be of interest. The second alternative estimator is therefore the estimator based on one-sided Winsorization as defined as by Kokic and Smith (1998a). For $k = 0$ and 1, we denote the ratio between the population and sample totals of the auxiliary variable by

$$R_X = \sum_{i=1}^{N} X_{1i} \Bigg/ \sum_{i \in s} X_{1i}$$

The population total estimator based on one-sided Winsorization is

$$\hat{T}_W = R_X \sum_{i \in s} W_i$$

where

$$W_i = \begin{cases} Y_i & \text{if } Y_i < K_i \\ K_i + (Y_i - K_i)R_X^{-1} & \text{otherwise} \end{cases}$$

Here the upper bound K_i, which is a function of (among other quantities) X_{1i}, is defined as in Kokic and Smith (1998a). We have estimated the Winsorization cutoff parameter by sampling from the survey variable *Beef cattle* itself. This is obviously a more favorable situation than in practice, when only historic data is available; the potential bias resulting from this will most likely be in favor of the Winsorization-based estimators. We have used five replicate samples in this first estimation step; this is in accordance with the recommendations of Kokic and Smith (1998b).

3.2. Results

We have used the number of *Beef cattle* for $N = 1{,}652$ farms that have participated in the Australian Agricultural and Grazing Industries Survey (AAGIS) carried out by the Australian Bureau of Agricultural and Resource Economics. Data from this survey has previously been used by, for instance, Chambers (1996). As the auxiliary variable, we have used the *Farm area* (in hectares). We have used the logarithmed farm area for the model estimator and the ‘‘raw’’ unlogarithmed farm area for the other estimators. In all, there were 1.5 million beef cattle on the 1,652 farms, thus $T = 1.50 \cdot 10^6$. The correlation between the logarithmed *Beef cattle* and the logarithmed *Farm area* is 0.575.

Although the sampling is assumed to be uninformative under the model assumptions, the practical feasibility of an estimator is often evaluated using repeated sampling from one population. We comply with this general practice when assessing the applicability of the estimator to survey data; there is no real choice, since in principle only one population is available. The presence of the constant auxiliary variable X_{0i} (which is equal to 1 for all i) is implied without comment. We have used $v_i = 1$ (implying a constant coefficient of variation) for the model estimator.

Table 1. Relative efficiency of the model estimator when applied to Beef cattle

	n				
Relative efficiency of $\hat{T}$ vs	50	100	150	200	250
$\hat{T}_{LR}$	1.74(0.07)	1.88(0.04)	1.86(0.04)	1.94(0.04)	1.94(0.03)
$\hat{T}_W$	1.51(0.04)	1.10(0.03)	1.03(0.02)	1.03(0.02)	1.00(0.02)

The auxiliary variable is *Farm area*. The cell entries are mean values over 50 replicate runs of 250 samples each (the estimated standard errors are shown within brackets).

Table 2. Relative bias of the estimators when applied to Beef cattle

	n				
Relative bias of	50	100	150	200	250
$\hat{T}$	2.90(0.46)	4.48(0.29)	4.83(0.22)	5.01(0.16)	4.89(0.18)
$\hat{T}_{LR}$	−5.09(0.57)	−2.68(0.35)	−1.64(0.29)	−0.63(0.27)	−0.77(0.23)
$\hat{T}_W$	8.11(0.49)	0.46(0.33)	−0.82(0.26)	−1.00(0.21)	−1.37(0.18)

The relative bias is expressed in percent of the true population total $T = 1.50 \cdot 10^6$. The auxiliary variable is *Farm area*. The cell entries are mean values over 50 replicate runs of 250 samples each (the estimated standard errors are shown within brackets).

Table 3. Relative RMSE of the estimators when applied to Beef cattle

	n				
Relative *RMSE* of	50	100	150	200	250
$\hat{T}$	45.05(0.62)	31.10(0.26)	25.15(0.22)	21.50(0.16)	18.83(0.13)
$\hat{T}_{LR}$	58.54(0.99)	42.50(0.45)	34.13(0.26)	29.79(0.27)	26.13(0.16)
$\hat{T}_W$	54.86(0.81)	32.40(0.33)	25.34(0.22)	21.75(0.14)	18.78(0.12)

The relative *RMSE* is the root of the *MSE* expressed in percent of the true population total $T = 1.50 \cdot 10^6$. The auxiliary variable is *Farm area*. The cell entries are mean values over 50 replicate runs of 250 samples each (the estimated standard errors are shown within brackets).

We have assessed the model estimator performance by computing (for both of the alternative estimators) the relative efficiency, which is defined as the ratio between the *MSE* of the alternative estimator and that of the model estimator. A relative efficiency exceeding 1.0 thus implies that the model estimator is more efficient than the alternative estimator is. In addition, we have (for all estimators) computed the relative bias, i.e., the bias of the estimator divided by the true population total.

In order to assess the precision in the relative efficiency and relative bias estimates, we have made 50 replicate runs. In each run, we have drawn 250 simple random samples (without replacement) of each of the sizes $n = 50$, 100, 150, 200 and 250. Over all 50 runs, the average (and standard error) of the relative efficiency (see Table 1), the average (and standard error) of the relative bias (see Table 2), as well as the average (and standard error) of the relative *RMSE* (see Table 3), have been computed for all four alternative estimators.

As we see from Table 1, the model estimator is always the most efficient one for *Beef cattle*, but the relative efficiency versus the Winsorization-based estimators decreases with increasing sample size. For all three estimators, we see from Tables 2 and 3 that the bias is small in relation to the overall *MSE*.

4. Discussion

4.1. Possible generalizations of the model estimator

For the applied survey statistician, it is of little significance that the model estimator is better than a design-based alternative under the simple random sampling design. In many situations, a sampling design more conducive to efficiency, e.g., stratified sampling, is a natural choice. This modified model estimator may use the sampling design when estimating $\boldsymbol{\alpha}$ and σ and yet retain the efficiency advantages of model-based inference. A practical ‘‘plug-in’’ approach in this case would replace the simple random sampling based ML estimates of the mean and variance of the logarithms of the survey variables by their corresponding ML estimates under the actual sampling design used (Breckling et al. 1994) or by pseudo-ML estimates calculated using inverse sample inclusion probability weighted versions of the simple random sampling estimating equations for these parameters (Särndal et al. 1992).

For skewed data with extreme outliers, we could follow Brewer and Ferrier (1966) and replace the lognormal assumption used in this article by one where the logarithms of the positive valued survey variable have a t-distribution. However, this will have the effect of complicating the theory considerably.

Many populations have negative extreme values as well as positive ones. Assume that the survey variable Y_i is the sum of two lognormal-logistic components. The parameters of the two lognormal distributions can be estimated separately (using the positive and negative values of Y_i). The proportions of positives, negatives and zeros could all be estimated simultaneously using logistic methods (cf. McCullagh and Nelder 1989, p.149). The estimation of T can then be carried out as in Section 2.

If we have access to two auxiliary variables, and the first auxiliary variable is highly correlated with the positive values of the survey variable while the other auxiliary variable is a good predictor of whether the survey variable is positive or zero, we will use the first variable to estimate $\boldsymbol{\alpha}$ and σ while using the second variable for the estimation of $\boldsymbol{\beta}$. If the model of Section 2 is slightly generalized, different auxiliary variables may be used for the two model components. The generalization required is straightforward, and will increase the applicability of the model estimator to situations similar to the one described above.

4.2. Conclusions

In this article, we have developed a population total estimation method for non-negative skew variables; the practical applicability of the estimator, which is derived in Section 2, is enhanced by the possibility of using any number of auxiliary variables.

As we have seen from the simulation study in Section 3, the model estimators are most useful when the sample size is small. If one can afford drawing a large sample, the potential benefit of the model-based estimator (relative to the alternative estimators) does not outweigh the risk of bias due to model misspecification.

One might argue that the model estimators are useless, since we never know for certain that the survey variable is lognormal distributed. The estimator is of course not of any use in situations when the survey variable has a nice symmetric distribution with no outliers. In such a situation, even the arithmetic mean of the survey variable may be more efficient than the model estimator is. Some prior knowledge about the type of survey variable is thus required.

As can be seen from Section 3, the lognormal estimator works well when applied to survey variables that are known to be skewed to the right (e.g., economic variables). In applications to such variables, a particular strength of the model estimator is that, even in the absence of extremely large values in the sample, the assumed lognormal structure (which implies the presence of extremely large values) of the survey variable is used for estimating the population total. Neither the design-based estimators nor the traditional weight or value modification estimators have this feature.

5. References

Barnett, V. and Lewis, T. (1994). Outliers in Statistical Data, (3rd edition). New York: John Wiley.

Breckling, J.U., Chambers, R.L., Dorfman, A.H., Tam, S.M., and Welsh, A.H. (1994). Maximum Likelihood Inference from Sample Survey Data. International Statistical Review, 62, 349–363.

Brewer, K.R.W. (1995). Combining Design-based and Model-based Inference. In Business Survey Methods, eds. B.G. Cox, D.A. Binder, B.N. Chinappa, A. Christianson, M.J. Colledge, and P.S. Kott, 589–606. New York: John Wiley.

Brewer, K.R.W. and Ferrier, A.E. (1966). Optimum Weighting of Outlying Observations When Sampling From Economic Populations. Unpublished manuscript, Commonwealth Bureau of Census and Statistics, Canberra, Australia.

Brewer, K.R.W. and Mellor, R.W. (1973). The Effect of Sample Structure on Analytical Surveys. Australian Journal of Statistics, 15, 145–152.

Casella, G. and Berger, R. (1990). Statistical Inference. Belmont, California: Duxbury Press.

Chambers, R.L. (1986). Outlier Robust Finite Population Estimation. Journal of the American Statistical Association, 81, 1063–1069.

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, Tome LV, 55–72.

Deming, W.E. (1950). Some Theory of Sampling. New York: John Wiley.

Hidiroglou, M.A. and Srinath, K.P. (1981). Some Estimators of a Population Total from Simple Random Samples Containing Large Units. Journal of the American Statistical Association, 76, 690–695.

Karlberg, F. (2000). Population Total Prediction Under a Lognormal Superpopulation Model. Metron, forthcoming.

Kokic, P.N. and Bell, P.A. (1994). Optimal Winsorizing Cutoffs for a Stratified Finite Population Estimator. Journal of Official Statistics, 10, 419–435.

Kokic, P.N. and Smith, P.A. (1998a). Winsorization of Outliers in Business Surveys. Submitted to Journal of the Royal Statistical Society Series D.

Kokic, P.N. and Smith, P.A. (1998b). Outlier-Robust Estimation in Sample Surveys Using Two-Sided Winsorization. Submitted to Journal of the American Statistical Association.

Lee, H. (1995). Outliers in Business Surveys. In Business Survey Methods, eds. B.G. Cox, D.A. Binder, B.N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott, 503–526. New York: John Wiley.

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, 2nd edition. London: Chapman & Hall.

Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. Journal of the American Statistical Association, 73, 351–358.

Shao, J. and Tu, D. (1995). The Jacknife and Bootstrap. New York: Springer-Verlag.

Smith, T.M.F. (1999). Recent Developments in Sample Survey Theory and Their Impact on Official Statistics. Bulletin of the International Statistical Institute. Proceedings of the 52nd session of the International Statistical Institute. Helsinki, Tome LVIII, Book 1, 7–10.

Särndal, C.-E., Swensson, B. and Wretman, J.H. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Thorburn, D. (1991). Model-Based Estimation in Survey Sampling of Lognormal Distribution. A Spectrum of Statistical Thought, Essays in Statistical Theory, Economics and Population Genetics in Honor of Johan Fellman, 228–243. Helsinki: The Swedish School of Economics and Business Administration.

Received July 1997
Revised May 2000