Survey Research Methods (2012)
Vol.6, No.3, pp. 137–143
ISSN 1864-3361
http://www.surveymethods.org

Robust Lavallée-Hidiroglou stratified sampling strategy

M. Caterina Bramati
Sapienza University of Rome

There are several reasons why robust regression techniques are useful tools in sampling design. First of all, when stratified samples are considered, one needs to deal with three main issues: the sample size, the strata bounds determination and the sample allocation in the strata. Since the target variable Y, the objective of the survey, is unknown, some auxiliary information X known for the entire population from which the sample is drawn, is used. Such information is helpful as it is typically strongly correlated with the target Y. However, some discrepancies between these variables may arise. The use of auxiliary information, combined with the choice of the appropriate statistical model to estimate the relationship between Y and X, is crucial for the determination of the strata bounds, the size of the sample and the sampling rates according to a chosen precision level for the estimates, as has been shown by Rivest (2002). Nevertheless, this regression-based approach is highly sensitive to the presence of contaminated data. Since the key tool for stratified sampling is the measure of scale of Y conditional on the knowledge of the auxiliary X, a robust approach based on the S-estimator of the regression is proposed in this paper. The aim is to allow for robust sample size and strata bounds determination, together with optimal sample allocation. Simulation results based on data from the Construction sector of a Structural Business Survey illustrate the advantages of the proposed method.
Keywords: robust regression, stratified design, auxiliary data

1 Introduction

The word 'robust' has been extensively employed in survey sampling referring to resistance to the bias induced by misspecification in model-based inference (see Nedyalkova and Tillé 2012). To this extent, several authors (Royall and Herson 1973, Scott et al. 1978) have proposed methods to protect inference against misspecification. In what follows, by the term 'robust' design we mean a sampling design which is insensitive to the occurrence of gross errors in the data. In other words, a robust design is not altered by removing or modifying a small percentage of the data set.

It is important to realize that outliers occur frequently in real data. Outlying observations can be present in a sample because of errors in recording observations, they can be due to transcription or transmission errors, or they may be caused by an exceptional occurrence in the observed phenomenon. Rousseeuw and Leroy (1987) present many real data sets in which the Least Squares residuals are of little help in identifying outliers. Sensitivity analysis is often used to detect the influential cases deleting observations one by one and assessing the effects of such deletions on the regression output. Unfortunately, outliers are not necessarily influential observations. Moreover, when outliers are clustered, they 'mask' each other and sensitivity analysis fails to detect them (see Rousseeuw and Leroy 1987). In low dimensional data sets visual inspection could be effective for locating outliers.However, there are no simple methods for visually detecting outliers in high dimensional data sets – see Rousseeuw and Van Zomeren (1990) for a discussion. In practice therefore one needs an objective procedure which is able to diminish the impact of outliers, as an alternative technique to the deletion of observations, which can be very subjective.

To illustrate, consider a stratified design where the stratification variable X includes some low quality data. Then aberrant observations in X will affect both the location and scale measures for each stratum, attributing higher dispersion to the units belonging to them. This in turn implies an overestimation of the sample size which depends on the dispersion of the data, thus affecting the sampling design and hence the survey results.

Ratio-type estimators for robust regression have been proposed by Chambers (1986) and more recently by Kadilar et al. (2007), using simple random sampling. However, they use the M-estimator, which is known to be vulnerable to outliers in X (see Rousseeuw and Leroy 1987) and are therefore not suitable in a stratified design where the auxiliary variable X is contaminated.

This paper deals with the model-based approach to sample survey design. For a comprehensive description of this approach, see Valliant, Dorfman and Royall (2000). The aim is to estimate the population total t_y of a Y variable using an estimator $\hat{t}_y$. Under the model-based approach, the properties of this estimator are determined by the distribution of its sample error under the assumed model for the population (see Brewer 1963 and Royall 1970). In this paper, the expression for the conditional expectation and variance of Y, given that the statistical unit is classified in stratum h, de-

Contact information: M. Caterina Bramati, Sapienza University of Rome, Dpt Methods and Models for Economics, Territory and Finance, e-mail: mariacaterina.bramati@uniroma1.it

pends on the model specified. As a consequence, we speak of the model bias of the estimator of the total as $E(\hat{t}_y - t_y)$, where the expectation is with respect to the assumed model for Y.

Note that the concept of model bias has a different interpretation from the usual concept of design bias. For example, Nedyalkova and Tillé (2012) refer to model-unbiased and design-unbiased estimators. These authors stress that the Horvitz-Thompson (HT) estimator can be model-biased (due to model misspecification for example) thus inflating its mean square error under the model. However, it still remains design-unbiased.

Here we focus on the stratified sampling design, which has been proven to be a very efficient surveying technique for skewed populations, as pointed out in Lavallée and Hidiroglou (1988). For this reason, stratified samples are often employed in business surveys carried out by National Statistical Offices.

Lavallée and Hidiroglou (1988) propose an iterative procedure, the LH algorithm, which stratifies skewed populations into a take-all stratum and a number of take-some strata. Given a particular allocation rule, the stratum bounds are then chosen in order to minimize the overall sample size subject to a specified level of precision for the target variable. Outliers can strongly affect the outcome of the LH algorithm since the sample size is inflated when observations appear more extreme than they really are. Moreover, the stratum bounds and the sample allocation might be both affected. This is clear when we consider Neyman allocation based on the within-stratum dispersion of X. Since the allocation's rationale is to survey more units in strata where the auxiliary variable is more dispersed, such outliers might have the effect of enormously and unduly increasing the sample size in each contaminated stratum.

In this paper, two robust versions of the LH method are suggested: the 'naive robust' and the 'robust' LH sampling strategy, which we compare through a simulation study.

The LH method assumes use of the Horviz-Thompson estimator. However, precision can be improved by taking advantage of the available auxiliary information about the target population. In particular, we consider the estimator derived from a linear regression model based on the relationship between the values of Y and a set of auxiliary variables X for which the totals in the finite target population are known. Given such an assumed relationship, a generalized regression estimator can be derived. If the linear model underlying this generalized regression estimator explains the variation of the target parameter reasonably well, then using it in the optimal design will result in a reduction of the design variance relative to that of the Horvitz-Thompson estimator. If, however, this model is misspecified, then there could be an increase in the design variance, even though the generalized regression estimator remains approximately design-unbiasedness.

1.1 Stratified Design and the LH algorithm

In what follows we focus on simple stratified sample designs with one take-all stratum and several take-some strata. This strategy is suitable in the presence of populations with skewed distributions (a few units account for a large share of the study variable), as pointed out in Tillé (2001), and when the statistical units composing the universe are available in a list together with some auxiliary information from administrative sources (i.e. tax declaration, social security registers and so on). Moreover, stratified sampling is required for EU countries to be compliant with the recommendations of the European statistical office (Eurostat) concerning designs for business surveys. As a consequence, the quality of the data held by the administrative sources is a key issue for the efficient use of these data in sample design.

In a stratified sample, the population is divided into subgroups (or strata), which are mutually exclusive (i.e. 1 unit can belong to 1 stratum only) and collectively exhaustive (i.e. no population unit excluded).

The 'statistical precision' is the constraint under which choices of sample sizes and allocation are made. In a seminal paper, Lavallée and Hidiroglou (1988) suggest an optimal solution to the choice of the stratum bounds, the sample size and allocation subject to the constraint of a fixed precision for the target variable. In particular, their algorithm allows for the simultaneous determination of the minimum sample size, the strata bounds and the sample allocation in order to satisfy a specified level of statistical precision.

Consider a stratified random sampling scheme with L strata for a variable of interest Y defined over a target population U of size N. Denoting by U_h, $h = 1, \ldots, L$, the component of size N_h of the target population making up stratum h, and by s_h the random sample of size n_h taken from this stratum, with $f_h = \frac{n_h}{N_h}$ the corresponding sampling fraction, the Horvitz-Thompson estimator $\hat{t}_{\text{ystrat}} = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{k \in s_h} y_k$ then has design variance

$$\text{Var}(\hat{t}_{\text{ystrat}}) = \sum_{h=1}^{L} N_h \frac{(1-f_h)}{f_h} S^2_{yh} \quad (1)$$

where

$$S^2_{yh} = \frac{1}{N_h - 1} \sum_{k \in U_h} (y_k - \overline{Y}_h)^2,$$

and $\overline{Y}_h$ is the mean of Y within stratum h.

In the procedure the L-th stratum is the take-all stratum, i.e. all the enterprises belonging to it are sampled. Random sampling is then used to select the enterprises in the remaining $L-1$ strata. Thus, for the take-all stratum $n_L = N_L$, whilst for $h < L$, the sample size n_h in the take-some stratum can be written as $(n - N_L)a_h$, where $\sum_{h=1}^{L-1} a_h = 1$.

By straightforward calculations (1) can be rewritten as

$$\text{Var}(\hat{t}_{\text{ystrat}}) = \frac{1}{n - N_L} \sum_{h=1}^{L-1} \frac{N_h^2 S^2_{yh}}{a_h} - \sum_{h=1}^{L-1} N_h S^2_{yh} \quad (2)$$

from which, solving for n,

$$n_{\hat{t}_{\text{ystrat}}} = N_L + \frac{\sum_{h=1}^{L-1} \frac{W_h^2}{a_h} S^2_{yh}}{(c\overline{Y}/N)^2 + \sum_{h=1}^{L-1} \frac{W_h}{N} S^2_{yh}}, \quad (3)$$

where $W_h = \frac{N_h}{N}$, c is the target coefficient of variation (the precision level, which often ranges between 1% to 10% in business surveys) and $\overline{Y}$ is the population mean of Y.

The idea is to find the optimal stratum boundaries $b_1, \ldots, b_{L-1}$ which minimize $n_{\hat{t}_{\text{ystrat}}}$ given an appropriate sample allocation method (Neyman, proportional and so on). For instance, under Neyman allocation

$$a_h = \frac{W_h S_{yh}}{\sum_{k=1}^{L-1} W_k S_{yk}},$$

which means that

$$n_{\hat{t}_{\text{ystrat}}} = N_L + \frac{(\sum_{h=1}^{L-1} W_h S_{yh})^2}{(c\overline{Y}/N)^2 + \sum_{h=1}^{L-1} \frac{W_h}{N} S_{yh}^2}. \quad (4)$$

1.2 The Regression Model

In practice, implementing the design described in the previous section requires knowledge of the stratum variances S_{yh}^2 and therefore of the population Y, while the strata are defined in terms of the values of an auxiliary variable X, e.g. a size variable, which is known for all statistical units in the target population. In this situation Lavallée and Hidiroglou (1988) suggest replacing the stratum variances S_{yh}^2 in equation (4) by the stratum variances of the auxiliary variable, S_{xh}^2.

However, the auxiliary variable X used for stratification is only a proxy for the survey variable Y. To account for the discrepancies existing between Y and X, Rivest (2002) suggests a generalized LH algorithm which uses a regression model linking the target and the auxiliary variable(s) to define appropriate population and stratum moments of Y. Such a model-based optimal stratification method can be useful in very long-tailed populations, as encountered in business surveys, for example. In particular, the relationship between Y and X is often characterized by a log-linear regression relationship.

In what follows we consider variables X and Y as continuous random variables and we denote by $f(x)$, $x \in \mathbb{R}$ the density of X. The data $x_1, \ldots, x_N$ are considered as N independent realizations of the random variable X.

Since stratum h consists of the population units with an X-value in the interval $(b_{h-1}, b_h]$, the stratification process replaces the stratum mean and variance of Y by the values of $E(Y|b_h \geq X > b_{h-1})$ and $\text{Var}(Y|b_h \geq X > b_{h-1})$, the conditional mean and variance of Y given that the unit falls in stratum h, for $h = 1, \ldots, L-1$.

In particular, the model assumes that the regression relationship between Y and X can be expressed as

$$\log Y = \alpha + \beta \log X + \varepsilon, \quad (5)$$

where ε is assumed to be a zero-mean random variable, normally distributed with variance σ^2 and independent from X, whereas α and β are the parameters to be estimated.

The conditional moments of Y are obtained using the basic properties of the log-normal distribution. They are

$$E(Y|X \in (b_{h-1}, b_h]) = e^{\alpha+\sigma^2/2} E(X^\beta | X \in (b_{h-1}, b_h])$$

and

$$\begin{aligned} \text{Var}(Y|X \in (b_{h-1}, b_h]) &= e^{\alpha+\sigma^2/2}\{e^{\sigma^2} E(X^{2\beta}|X \in (b_{h-1}, b_h]) \\ &- E(X^\beta|X \in (b_{h-1}, b_h])^2\}. \end{aligned} \quad (6)$$

Plugging the expression for the variance above into (4), it is clear that strata bounds, sample size and allocation then depend on the first and second-order conditional moments of the auxiliary variable.

1.3 Outliers and Robust Design

The main weak point in the algorithm proposed by Rivest (2002) is that, since S_{yh}^2 is unknown, the design is in practice based on the auxiliary information provided by administrative records. This is particularly true in the case of business surveys, where business registers with information on firm activities are used. Such sources often suffer from low data quality.

In general, several types of outliers can occur in the data which might affect the LH sampling algorithm. Using the same notion of outliers as in Rousseeuw and Leroy (1987), we define three main types of anomalies depending on the data contaminated

- outliers in the survey (Y) data (vertical outliers)
- outliers in the auxiliary (X) data (leverage points)
- outliers in both variables (X, Y) (good/bad leverage points).

The presence of such anomalies makes the estimation of the conditional mean and variance of $Y|X$ unreliable, therefore affecting the sample size and strata bounds determination, as well as the sample allocation.

In what follows we focus on contamination in the administrative data (leverage points) and we propose two alternatives to the Rivest (2002) Generalized LH algorithm (GLH). Of course, errors in survey data (vertical outliers) might also occur, but they are unknown in advance. Correction for such outliers can be applied only after the data collection, for instance using a post-stratified robust estimator, which is beyond the scope of this paper. Note, however, that the approach here is based on using a regression estimator which is robust to vertical outliers and to bad leverage points, allowing for robustness when the design process uses contaminated data from previous surveys.

In the first approach, robust regression estimators of the parameters of the log-linear regression model are used. Then, the estimated robust parameters are plugged in the LH objective function used for the computation of the strata bounds, the minimum sample size and the sample allocation which satisfy a fixed statistical precision. We call this the naive robust approach (NR-GLH).

In the second approach, strata bounds and sizes are derived after re-weighting the auxiliary information according

to the degree of outlyingness. This approach, which we refer as the 'Robust GLH' algorithm, is presented in the next section.

2 The Robust GLH algorithm

When outliers arise in the auxiliary variable X they might affect the strata bounds $b_1, \ldots, b_{L-1}$, the overall sample size n and the sample allocation a_h, $h = 1, \ldots, L-1$. Let $\omega(\cdot)$ be some weighting function which assigns values between $[0, 1]$ according to the degree of reliability of the data. Given the log-linear relationship (5) between the survey variable Y and the auxiliary variable X, we can then replace the conditional variance (6) by the weighted conditional variance

$$\begin{aligned} & \mathrm{Var}_\omega(Y|X \in (b_{h-1}, b_h]) \\ = \; & \exp^{\alpha+\sigma^2/2}\{e^{\sigma^2} E_\omega(X^{2\beta}|X \in (b_{h-1}, b_h]) \\ - \; & E_\omega(X^\beta|X \in (b_{h-1}, b_h])^2\} \\ = \; & \exp^{\alpha+\sigma^2/2}(e^{\sigma^2}\psi_h/W_h - (\phi_h/W_h)^2), \end{aligned} \tag{7}$$

where

$$\begin{aligned} W_h &= \int_{b_{h-1}}^{b_h} \omega(x^\beta) f(x) dx, \\ \phi_h &= \int_{b_{h-1}}^{b_h} x^\beta \omega(x^\beta) f(x) dx, \\ \psi_h &= \int_{b_{h-1}}^{b_h} x^{2\beta} \omega(x^\beta) f(x) dx. \end{aligned}$$

Note that the sample design approach proposed in this paper assumes knowledge of a sample of Y values from the target population. This is usually from previous surveys or pilot studies. Since the auxiliary information X is generally available at different points in time from administrative sources, it should be possible to estimate the regression of Y on X in a time-coherent manner. The estimated scale and regression coefficients obtained from the sample from the previous survey are then used in the Generalized LH algorithm. This is the situation that we assume throughout this paper.

When no data are available on the variable of interest Y, stratification is done using only the information on the auxiliary variable X in the LH algorithm (i.e. there are no regression adjustments). In this case, the presence of outlying observations in X could inflate the stratum variances, resulting in an unduly large take-all stratum. In the univariate case (one scalar stratification variable), it is possible to identify outliers by visual inspection. However, this is much more difficult if X is a vector of auxiliary variables. In that case, one could use multivariate outlier detection methods based on robust location and covariance estimators – see Rousseeuw and Van Zomeren (1990).

2.1 Choice of $\omega(\cdot)$

Several weighting functions might be considered. Our choice focuses on $\omega(x) = \rho'(x)/x$, the weighting function associated with the S-estimator of the regression of Y on X, see Rousseeuw and Yohai(1984). This estimator is the solution to

$$S(x, y) = \arg\min_\beta s(r_1(\beta), ..., r_N(\beta)) \tag{8}$$

where the $r_i(\beta)$ are the regression residuals and s is a scale measure defined by

$$\frac{1}{N}\sum_{i=1}^{N} \rho(\frac{r_i(\beta)}{s}) = K \tag{9}$$

for $K = E_\Phi[\rho]$, Φ being the Gaussian distribution, and a conveniently chosen ρ function satisfying

- $\rho(\cdot)$ symmetric and continuously differentiable, with $\rho(0) = 0$,
- there exists $c > 0$ such that $\rho(\cdot)$ is strictly increasing on $[0, c]$ and constant on $[c, +\infty)$.
- $K = \delta\rho(c)$, where δ is the breakdown point of the procedure as $n \to \infty$.

The S-estimator, whose name is derived from the fact that it is implicitly defined in terms of a scale statistic, has several statistical properties. It is regression, scale and affine equivariant. It is robust with respect to both vertical outliers and leverage points up to a 50% breakdown point (see Donoho and Huber 1983, for a definition of breakdown point), and it is highly efficient. In particular, it is possible to tune the efficiency level of the estimator and its degree of resistance to outliers by means of the choice of the breakdown point δ. When δ is set to 0, the efficiency level of the S-estimator is highest and coincides with that of the classical least squares estimator.

Some options for $\rho(x)$ and $\omega(x) = \rho'(x)/x$ are presented in Figures 1 and 2. Given choice of a particular specification, we denote the resulting S-estimates of the model parameters (α, β, σ) by $(\hat{\alpha}_{ROB}, \hat{\beta}_{ROB}, \hat{\sigma}_{ROB})$ in what follows. Of course, other robust regression estimators can be used, e.g. the Least Median of Squares, the Least Trimmed Squares and the GM estimator. However, these are usually less efficient than the S-estimator (see Leroy and Rousseeuw 1984).

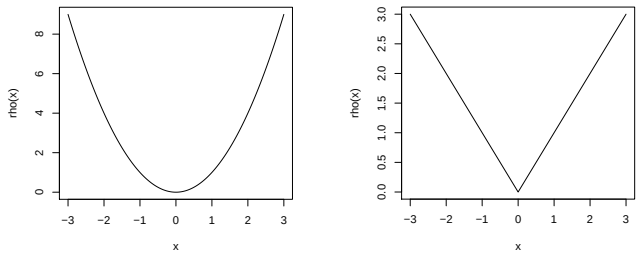
2.2 The algorithm

The problem is essentially one of solving for bounds $b_1, \ldots, b_h, \ldots, b_L$ which minimize n. Several allocation criteria can be considered in this context: here we focus on the Neyman allocation scheme. Replacing S^2_{yh} in (4) by (7), the log-linear specification for the objective function is then

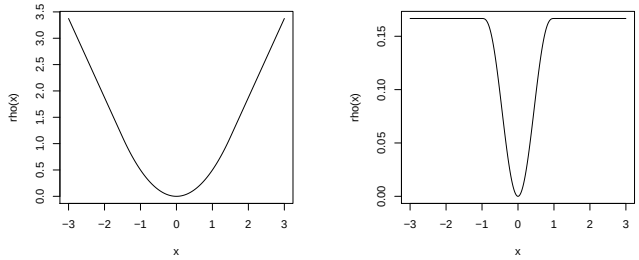
$$n_{\hat{f}_{\text{ystrat}}} = N_L + \frac{(\sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2})^2}{(c * \mathrm{med}_i|x_i^\beta|/N)^2 + \sum_{h=1}^{L-1}\frac{(e^{\sigma^2}\psi_h - \phi_h^2/W_h)}{N}} \tag{10}$$

with the moments W_h, ϕ_h and ψ_h replaced by their robust estimates. These are defined by substituting the S-estimates $\hat{\beta}_{ROB}$ and $\hat{\sigma}_{ROB}$ for β and σ respectively in (8).

As in Rivest (2002), the iterative scheme proposed by Sethi (1963) is then implemented for a given L and precision c, and the optimal strata bounds and sample size computed.

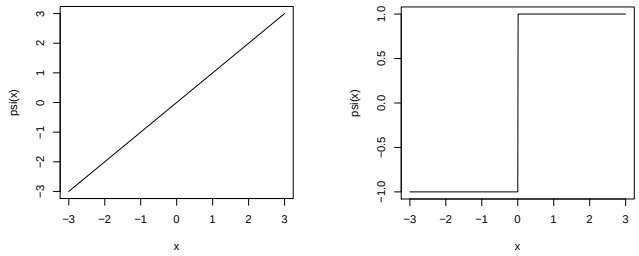


(a) Squared errors (b) Absolute errors

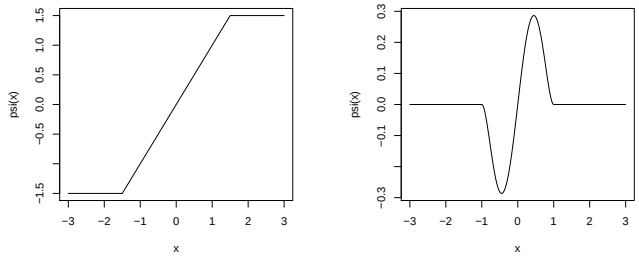


(c) Winsorizing at 1.5 (d) Biweight

Figure 1. Possible options for rho



(a) Squared errors (b) Absolute errors



(c) Winsorizing at 1.5 (d) Biweight

Figure 2. Possible options for psi

In order to define this iterative scheme, let F be the anti-derivative of the integrable function $\omega(x^\beta)f(x)$, then from the fundamental theorem of calculus $W_h = F(b_h) - F(b_{h-1})$. Similarly, $W_{h+1} = F(b_{h+1}) - F(b_h)$.

Therefore, $\partial W_h/\partial b_h = F'(b_h) = \omega(b_h^\beta)f(b_h)$ and $\partial W_{h+1}/\partial b_h = -F'(b_h) = -\partial W_h/\partial b_h$. The same argument applies for definite integrals ϕ_h and ψ_h. Thus, from

$$\begin{aligned}
\frac{\partial W_h}{\partial b_h} &= -\frac{\partial W_{h+1}}{\partial b_h} = \omega(b_h^\beta)f(b_h) \\
\frac{\partial \phi_h}{\partial b_h} &= -\frac{\partial \phi_{h+1}}{\partial b_h} = b_h^\beta \omega(b_h^\beta)f(b_h) \\
\frac{\partial \psi_h}{\partial b_h} &= -\frac{\partial \psi_{h+1}}{\partial b_h} = b_h^{2\beta} \omega(b_h^\beta)f(b_h)
\end{aligned}$$

one obtains

$$\begin{aligned}
\frac{\partial n}{\partial b_h} &= \omega(b_h^\beta)f(b_h) \\
&\times \left\{\left(\frac{\partial n}{\partial W_h} - \frac{\partial n}{\partial W_{h+1}}\right) + \left(\frac{\partial n}{\partial \phi_h} - \frac{\partial n}{\partial \phi_{h+1}}\right)b_h^\beta + \left(\frac{\partial n}{\partial \psi_h} - \frac{\partial n}{\partial \psi_{h+1}}\right)b_h^{2\beta}\right\}
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial n}{\partial b_{L-1}} &= \omega(b_{L-1}^\beta)f(b_{L-1}) \\
&\times \left\{-N + \frac{\partial n}{\partial W_{L-1}} + \frac{\partial n}{\partial \phi_{L-1}}b_{L-1}^\beta + \frac{\partial n}{\partial \psi_{L-1}}b_{L-1}^{2\beta}\right\}.
\end{aligned}$$

The largest solution to the estimating equation $\frac{\partial n}{\partial b_h} = 0$ is the updated value of b_h in the iterative scheme. In order to calculate the partial derivative on left hand side of this equation we therefore need to calculate the partial derivatives of n with respect to W_h, ϕ_h and ψ_h. Under Neyman allocation these are given by

$$\begin{aligned}
\frac{\partial n}{\partial W_h} &= \frac{Pe^{\sigma^2}\psi_h/(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2}}{Q} - \frac{P^2(\phi_h/W_h)^2/N}{Q^2} \\
\frac{\partial n}{\partial \phi_h} &= \frac{-2P\phi_h/(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2}}{Q} + \frac{2P^2(\phi_h/(W_h N)}{Q^2} \\
\frac{\partial n}{\partial \psi_h} &= \frac{Pe^{\sigma^2}/(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2}}{Q} - \frac{e^{\sigma^2}P^2/N}{Q^2}
\end{aligned}$$

with

$$\begin{aligned}
P &= \sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h W_h - \phi_h^2)^{1/2} \\
Q &= (c * \mathrm{med}_i|x_i^\beta|/N)^2 + \frac{\sum_{h=1}^{L-1}(e^{\sigma^2}\psi_h - \phi_h^2/W_h)}{N}.
\end{aligned}$$

3 Simulation Study

A simulation study was carried out to compare the performance of the two robust sampling strategies with the GLH strategy of Rivest (2002) under several distributions for the data. Simulations were based on the business sampling frame of the Structural Business Survey (SBS) in 2005, where the target variable Y is the value added for enterprises in the Construction industry, stratified by economic-size class. The number of strata was set to 6, as Cochran (1977) recommends, with one take-all stratum and 5 take-some strata. The auxiliary information X is turnover which is available from the VAT register.

Table 1: MSE comparisons of NR-GLH and R-GLH methods versus GLH (Rivest 2002), target precision: 1%

Design	$MSE(\bar{y}_{NR-GLH})/MSE(\bar{y}_{GLH})$	$MSE(\bar{y}_{R-GLH})/MSE(\bar{y}_{GLH})$
No outliers	4.52	1.10
Long-tailed Cauchy	0.07	0.00
Long-tailed t	6.79	0.08
Vertical outliers 15%	1.01	0.99
Leverage points 15%	4.52	0.00
Vertical outliers 30%	0.99	0.99
Leverage points 30%	0.11	0.00

In each simulation a population was generated from the equation

$$\log y_i = \beta \log x_i + \varepsilon_i$$

with $\beta = .75$, and with the distribution of ε_i specified as follows:

1. no outliers: $\varepsilon_i \sim \mathcal{N}(0, 1)$
2. long-tailed errors: $\varepsilon_i \sim$ Cauchy
3. long-tailed errors: $\varepsilon_i \sim t_3$
4. vertical outliers: $\delta\%$ of $\varepsilon_i \sim \mathcal{N}(5\sqrt{\chi^2_{1;0.99}}, 1.5)$
5. bad leverage points: $\delta\%$ of $\varepsilon_i \sim \mathcal{N}(10, 10)$ and corresponding $X \sim \mathcal{N}(-10, 10)$.

The contamination level, i.e. the percentage of outliers in the data, was set to $\delta = 15\%$ and 30%, whilst the number of replications was set to 1000. Then the GLH algorithm (Rivest 2002), the naive robust GLH (NR-GLH) algorithm and the robust GLH (R-GLH) algorithm were used to compute the strata bounds, sizes and allocation for a 1% precision level. The resulting designs were then evaluated by comparing the averages over the simulations of the Mean Squared Error (MSE) of the Horvitz-Thompson estimator for the population mean of Y generated by these different designs. Table 1 displays the main results.

From the simulations we observe that the R-GLH is more efficient than the GLH algorithm mainly when data are long-tailed and when leverage points occur. When compared, the NR-GLH is less efficient than the R-GLH in the case of long-tailed distributions, whereas in the case of small percentages of vertical outliers it is as efficient as the GLH approach. Of course, when data are not contaminated and drawn from a symmetric distribution, both robust approaches are less efficient than the non-robust approach. In other words, the low quality of the auxiliary information used for the stratified design can dramatically influence upwards the sample size, the sample allocation and the strata bounds determination. The presence of vertical outliers alone does not much affect the entire procedure, as expected.

4 Conclusions

This work suggests a robust approach to the sample stratification which modifies the generalized Lavallée-Hidiroglou algorithm (see Rivest 2002) introducing a robust regression estimator which is not vulnerable to low quality auxiliary data. In particular, stratified sampling is very sensitive to outliers especially in business surveys where administrative sources with poor data quality are used. Sample size and strata bounds are then obtained under a chosen precision and sample frequencies computed using the Neyman allocation. A simulation study illustrates the performance of the sampling strategy hereby proposed, showing the superior efficiency of the R-GLH estimator compared to the GLH stratification when data are long-tailed distributed and when the auxiliary information is affected by outliers.

Acknowledgements

The author acknowledges Ray Chambers, Monica Pratesi, the participants of the 2011 ITACOSM Conference in Pisa and the participants of the 2011 NTTS Conference in Brussels for their comments and many insightful discussions. The author also thanks the two anonymous reviewers for their comments. All remaining errors are my own.

References

Brewer, K. R. W. (1963). Ratio estimation and finite populations: Some results deducible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, *5*, 93-105.

Chambers, R. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, *81*, 1063-1069.

Donoho, D. L., & Huber, P. J. (1983). The notion of breakdown point. In P. Bickel, K. Doksum, & J. L. Hodges (Jr.) (Eds.), *A Festschrift for Erich Lehmann.* Wadsworth, Belmont, CA.

Kadilar, C., Candan, M., & Cingi, H. (2007). Ratio estimators using robust regression. *Journal of Mathematics and Statistics*, *36*, 181-188.

Lavallée, P., & Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, *14*, 33-43.

Nedyalkova, D., & Tillé, Y. (2012). Bias robustness and efficiency in model-based inference. *Statistica Sinica*, *22*, 777-794.

Rivest, L. P. (2002). A generalization of Lavallée and Hidiroglou algorithm for stratification in business surveys. *Techniques d'enquêtes*, *28*, 207-214.

Rousseeuw, P. J., & Leroy, V. J. (1987). *Robust regression and outlier detection*. New York: J. Wiley.

Rousseeuw, P. J., & Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, *85*, 633-651.

Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators, in Robust and Nonlinear Time Series Analysis. In J. Franke, W. Hrdle, & R. D. Martin (Eds.), *Lecture Notes in Statistics* (p. 256-272). New York: Springer Verlag.

Royall, R. M. (1970). On finite population sampling under certain linear regression models. *Biometrika*, *57*, 377-387.

Royall, R. M., & Herson, J. H. (1973). Robustness estimation in finite populations I. *Journal of the American Statistical Association*, *68*, 880-889.

Scott, A., Brewer, K. R. W., & Ho, E. W. H. (1978). Finite population sampling and robust estimation. *Journal of the American Statistical Association*, *73*, 359-361.

Sethi, V. K. (1963). A note on the optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, *5*, 20-33.

Tillé, Y. (2001). *Théorie des sondages*. Paris: Dunod.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons.