Variance Estimation for Complex Surveys in the Presence of Outliers

Beat Hulliger[1], Ralf Münnich[2]
Swiss Federal Statistical Office, Espace de l'Europe 10, CH-2010 Neuchâtel[1]
University of Trier, Universitätsring 15, D-54286 Trier, e-mail: muennich@uni-trier.de[2]

Abstract

Quantitative variables in surveys often have a markedly skew distribution and, in addition, contain outliers. Robust estimators, which may be used in this situation, generally are biased. In addition linearized variance estimators tend to underestimate the true variance considerably. Alternatives are Bootstrap variance estimators or estimators based on multiple imputation. A simulation study with data from the Swiss Household Budget Survey shows the effects of outliers on estimators and their variance estimators. Three poverty measures and proposals for the estimation of their variance are included in the simulation study.

Keywords: Sampling, Robust estimator, Gini-coefficient, Quintile Share Ratio, Simulation.

1 Introduction

Outliers are a frequent concern in surveys with quantitative variables like household budget surveys or business surveys on production or turnover. A relatively small fraction of the data has extreme values in one or several variables. Often these extreme values occur when the bulk of the data has already a markedly skew distribution. Figure 1 shows the expenditures and incomes of a sample of the Swiss Household Budget Survey.

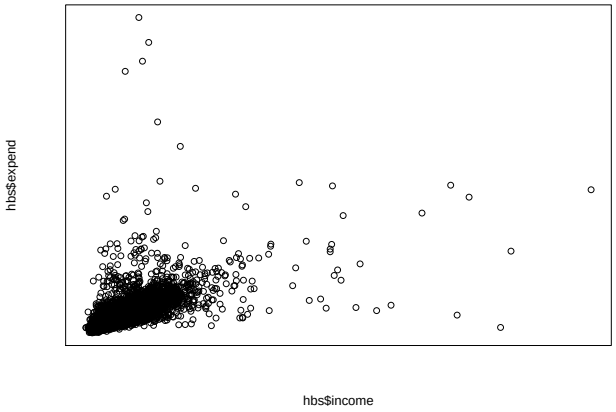

Figure 1: Sample of data on expenditures versus income

The traditional approach is to detect these extreme values on the basis of fixed univariate or bivariate limits, to review these observations manually and either "correct" the outliers, dismiss them from analysis or to leave them unchanged. Today the outlier detection or nomination methods use robust estimators, imputations based on robust models are used to replace outliers and, finally, traditional non-robust estimators may be used on the data treated beforehand. The alternative to this editing and imputation approach is to use robust estimators directly on the raw data.

Univariate robust estimators which are adapted to survey sampling have been introduced e.g. by Searls, Fuller, Chambers, Rivest and Hulliger [Hulliger 1995]. Variance estimators for these robust estimators are based on approximations. However, there is a basic problem when estimating the variance of robust estimators which are applied to skew data. Carrol [Carrol 1979] showed that the effect of the scale estimator in the defining equation for M-estimators is neglected in the usual variance estimators and this may lead to a large bias. The scale estimator is needed to make the M-estimator scale equivariant. Jackknife and Bootstrap variance estimators may be able to account at least partially for the variance of the scale estimator [Gwet and Lee 2000].

The asymptotic variance of an M-estimator with the MAD as a preliminary scale estimator involves the influence of the M-estimator, the influence of the MAD and a covariance term (see Appendix). Neglecting the two latter terms may result in a large bias. These results are valid for infinite populations but should hold also for finite populations. One purpose of this article is to show this with the help of Monte-Carlo simulations. We use the simple robust estimators from [Hulliger 1999], medians and winsorised means.

Multivariate outliers often are handled by first detecting them, then replace the values of the outliers by imputation, and subsequently apply standard linear estimators. Multivariate outlier detection needs sophisticated algorithms which usually do not allow for direct variance estimators. Multiple imputation is a possible solution. We investigate outlier detection with the Transformed Rank Correlation Algorithm [Béguin and Hulliger 2004] combined with regression imputation and multiple imputation. To observe differences between univariate and multivariate situations, also estimators of ratios of two variables are investigated.

Income is an important example of a positive, skew variable with outliers. Inequality measures like the Gini-coefficient and the recently defined set of Laeken indicators ([Dennis and Guio 2004]) usually are very non-robust. It is difficult to define alternatives because the characteristic of interest to estimate is highly non-robust. Nevertheless the influence of the outliers on the estimator and its variance estimator are of high interest and some simulations are shown in Section 3.4.

The following example shows that influential units may have a high impact on the shape of the point and variance estimation distribution even if the outliers are representative and moderate. In a simulation study a rare sub-population, here unemployed women of age 65 and higher, was estimated for Saarland within the German Microcensus, a 1% stratified cluster sample. Alltogether 128 cases from approximately 1 mio. people were observed. Even for samples of more than 10'000 individuals the fact that some of these rare observations are collected within small clusters may spoil the approximation to normality which could be expected from the central limit theorem.

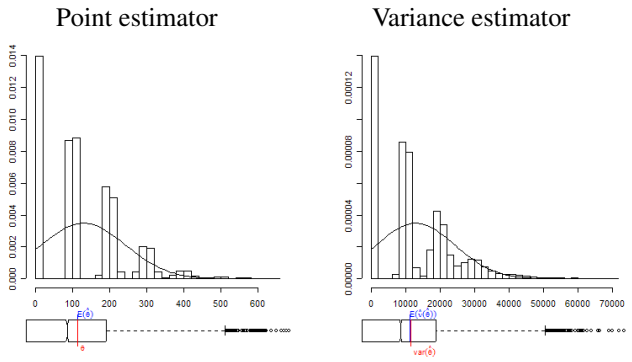

Figure 2: Distributions of point and variance estimator for *Unemployed Women with Age ≥ 65 in Saarland*

Figure 2 shows that approximately 35% of the distribution is close to the true value 128 in the universe but the mode is zero and some estimates are close to 600. The variance estimation distribution gives a similar picture. On the other hand, the application of the law of large numbers seems appropriate. The expected value of the point estimator is rather close to the true value. Analogously, the expected value of the variance estimator is close to the variance of the point estimator. However, the fact that averages of estimates behave appropriately does not mean that particular estimates behave well.

2 Estimation methods

2.1 Characteristics to estimate

The population characteristics we want to estimate always include the representative outliers and do not consider missing values. Thus we calculate them with the complete and uncontaminated universe. We consider the population mean, the Gini-coefficient and the Quintile Share Ratio for income and the ratio ψ of mean expenditure and income.

The Gini-coefficient for a variable of interest y can be expressed as

$$G = \frac{1}{\tau_Y} \cdot \sum_{i=1}^{N} \big(2F(y_i) - 1\big) \cdot y_i \quad .$$

The population quantile α of a variable y is $q_y(\alpha) = \min\{t : \sum_{i\in U} \mathbf{1}\{y_i \le t\}/N \ge \alpha\}$. Thus the Quintile Share Ratio is

$$Q(y) = \frac{\sum_{i\in U} y_i \mathbf{1}\{y_i > q_{0.8}(y)\}}{\sum_{i\in U} y_i \mathbf{1}\{y_i \le q_{0.2}(y)\}}$$

2.2 Point estimators for the parameters of interest

The estimators assume a sampling weight w_i per observation which could simply be the inverse of the inclusion probabilities π_i. The estimators are applied to the whole samples. For linear estimators they may be applied to a domain B by the usual replacement of a variable y_i by $y_i \mathbf{1}\{i \in B\}$. In the case of item-non-response the response indicator $\mathbf{1}\{i \in R(y)\}$ is used in the same way as a domain indicator. The robust estimators are expressed with robustness weights u_i per observation. Estimators then take the form $\theta(y) = \sum_{i\in S} w_i u_i y_i / \sum_{i\in S} w_i u_i$. The definition of the robustness weights involves the weighted empirical cumulative distribution function $F_{S,y}(t) = \sum_{i\in S} w_i \mathbf{1}\{y_i \le t\} / \sum_{i\in S} w_i$. We denote a quantile with respect to $F_{S,y}(t)$ by $q_{S,y}(\alpha)$. We define the quantile by weighted interpolation if the set $\{t : \sum_{i\in S} w_i \mathbf{1}\{y_i \le t\} / \sum_{i\in S} w_i = \alpha\}$ contains more than one value.

The main estimators are described in table 1.

Taking into consideration the weighted empirical distribution function $F_{S,y}(t)$, one can express the estimated Gini-coefficient as

$$\widehat{G} = \frac{1}{\widehat{\tau}_Y} \cdot \sum_{i\in\mathcal{S}} w_i \cdot \Big(2 \cdot \underbrace{\frac{1}{\widehat{N}} \sum_{j\in\mathcal{S}} w_j \cdot \mathbf{1}\{y_j \le y_i\}}_{F_{S,y}(t)} - 1\Big) \cdot y_i$$

where $\widehat{\tau}_y$ denotes the Horvitz-Thompson estimate for the quantity y and $\widehat{N} = \sum_{\mathcal{S}} w_i$ the estimated number of individuals.

For the Quintile Share Ratio we will have to estimate the ratio

$$\widehat{Q} = \frac{\widehat{\mu}_{\mathrm{R}}}{\widehat{\mu}_{\mathrm{P}}} = \frac{\widehat{\tau}_1}{\widehat{\tau}_2} \Big/ \frac{\widehat{\tau}_3}{\widehat{\tau}_4}$$

of the mean of the upper 20% $\widehat{\mu}_{\mathrm{R}}$ and the mean of the lowest 20% values $\widehat{\mu}_{\mathrm{P}}$ of the variable of interest.

$$\widehat{\mu}_{\mathrm{R}} = \sum_i w_i \cdot \big(y_i - y_i \cdot \mathbf{1}\{y_i \le \widehat{y}_{0,8}\}\big) \Big/ \sum_i w_i \cdot (1 - 0{,}8)$$

$$\widehat{\mu}_{\mathrm{P}} = \sum_i w_i \cdot y_i \cdot \mathbf{1}\{y_i \le \widehat{y}_{0,2}\} \Big/ \sum_i w_i \cdot 0{,}2 \quad .$$

The estimated Quintile Share Ratio then can be expressed as a function of four totals which will have to be used for linearized variance estimation.

2.3 Variance estimators

Variance estimators for the robust estimators are derived from their estimating equation as in [Hulliger 1999]. The variance can be approximated by

$$V(\hat{\theta}) \approx V\left(\sum_{i\in S} w_i u_i e_i\right),$$

Table 1: Estimators in the study

Name	Formula	Comments
Horvitz-Thompson	$\hat{\theta}_{HT}(y) = \sum_{i\in S} w_i y_i / N$	
Winsorized HTE	$u_i = \begin{cases} q_{S,\alpha}(y_i)/y_i, & y_i \le q_{S,\alpha}(y_i); \\ q_{S,1-\alpha}(y_i)/y_i, & y_i \ge q_{S,1-\alpha}(y_i); \\ 1, & \text{otherwise.} \end{cases}$	$\alpha = 0.02$
Hajek estimator	$N\hat{\theta}_{HT}(y)/\sum_{i\in S} w_i$	
Median	$q_{S,y}(0.5)$	
One-step Huber M	$u_i = \begin{cases} c\hat{\sigma}/\lvert y_i - q_{S,y}(0.5)\rvert, & \lvert y_i - q_{S,y}(0.5)\rvert < c\hat{\sigma}; \\ 1, & \text{otherwise.} \end{cases}$	$c = 5$, $\hat{\sigma} = 1.4826\, q_{S,\lvert y-q_{S,y}(0.5)\rvert}(0.5)$ is the weighted median absolute deviation (mad)
Ratio	$\hat{\psi} = \bar{x}_U \hat{\theta}_{HT}(y)/\hat{\theta}_{HT}(x)$	x the auxiliary variable
One-step ratio	$u_i = \begin{cases} c\hat{\sigma}\sqrt{x_i}/\lvert y_i - \beta_0 x_i\rvert, & \lvert y_i - \beta_0 x_i\rvert < c\hat{\sigma}\sqrt{x_i}; \\ 1, & \text{otherwise.} \end{cases}$	$\beta_0 = q_{S,y}(0.5)/q_{S,x}(0.5)$, $\hat{\sigma}$ is the mad of $\lvert y_i - \beta_0 x_i\rvert/\sqrt{x_i}$
Robust ratio	$u_i = \begin{cases} c\hat{\sigma}\sqrt{x_i}/\lvert y_i - \hat{\beta}_M x_i\rvert, & \lvert y_i - \hat{\beta}_M x_i\rvert < c\hat{\sigma}\sqrt{x_i}; \\ 1, & \text{otherwise.} \end{cases}$	$\beta_0 = q_{S,y}(0.5)/q_{S,x}(0.5)$, $\hat{\sigma}$ is the mad of $\lvert y_i - \beta_0 x_i\rvert/\sqrt{x_i}$

where e_i is the residual $y_i - \hat{\theta}$, considered fixed. In the case of a ratio the variance can be approximated by

$$V(\hat{\theta}) \approx V\left(\sum_{i\in S} w_i u_i e_i\right) \frac{1}{(\sum_{i\in S} w_i u_i x_i)^2},$$

where $e_i = y_i - \hat{\theta} x_i$. To estimate $V(\sum_{i\in S} w_i u_i e_i)$ we use the same estimators as for a weighted total of a variable $u_i e_i$, taking into account the sample design. In our simulation study we assume the weights w_i given and only take into account the stratification.

In addition to the above estimator which is based on linearization, a Bootstrap variance estimator and a balanced repeated replication variance estimator was applied for the poverty measures (Section 3.4).

In the case of poverty measurement, we apply the estimating equation approach as described in [Binder and Kovačević 1995] or [Deville 1999]. Taking the influence values

$$u_i^* = \frac{1}{\widehat{\tau}_Y} \cdot \left(2 \cdot y_i \cdot \widehat{F}(y_i) - (\widehat{G} + 1) \cdot y_i\right) \quad .$$

for the Gini-coefficient, we obtain the variance estimate

$$\widehat{V}(\widehat{G}_{\alpha,p}) \approx \widehat{V}(\widehat{\tau}_{u^*}) = \sum_{h=1}^{H} N_h^2 \cdot \frac{s^2_{u^*_{(h)}}}{n_h} \cdot \left(1 - \frac{n_h}{N_h}\right) \quad .$$

$s^2_{u^*_{(h)}}$ denotes sampling variance of the influence values in stratum h.

The Quintile Share Ratio becomes slightly more complicated due to the function of four totals. Applying Woodruff-linearization (cf. [Andersson 1994]), one can derive the values

$$\begin{aligned}
u_{1i} &= y_i - \left((y_i - y_{0.8}) \cdot \mathbf{1}\{y_i \le \widehat{y}_{0.8}\} + 0.8 \cdot y_{0.8}\right) \\
u_{2i} &= 0.2 \\
u_{3i} &= (y_i - y_{0.2}) \cdot \mathbf{1}\{y_i \le \widehat{y}_{0.2}\} + 0.2 \cdot y_{0.2} \\
u_{4i} &= 0.2 \\
u_{5i} &= (u_{1i} - \frac{\widehat{\tau}_1}{\widehat{\tau}_2} \cdot u_{2i}) \cdot \frac{1}{\widehat{N} \cdot 0.2} = \widehat{\mu}_{\mathrm{R}} \\
u_{6i} &= (u_{3i} - \frac{\widehat{\tau}_3}{\widehat{\tau}_4} \cdot u_{4i}) \cdot \frac{1}{\widehat{N} \cdot 0.2} = \widehat{\mu}_{\mathrm{P}}
\end{aligned}$$

where $\widehat{\tau}_2 = \widehat{\tau}_4 = \widehat{N} \cdot 0.2$. Finally, we get

$$z_i = (u_{5i} - \widehat{Q} \cdot u_{6i}) \cdot \frac{\widehat{\tau}_4}{\widehat{\tau}_3} \quad .$$

The variance is estimated as before with z_i instead of u_i^*.

Instead of applying classical design weights one can also use calibrated weights. The necessary corrections are described in [Deville 1999] and tested with only little success in the simulation study.

3 The simulation study

3.1 The simulation environment

The simulation universe of the Swiss Houshold Budget Survey was created under the DACSEIS project [Münnich et al. 2003]. The synthetic Swiss HBS universe contains data on income and expenditure of $N = 3'179'231$ fictive households modeled according to data from the Swiss HBS of 1998 (see Figure 3.1). From this universe, $R = 1'000$ stratified samples of size $n = 9'302$ were drawn consequently. The sample size within the strata is $\mathbf{n} = (2093, 1734, 1645, 1258, 1337, 795, 440)^\top$; applying proportional allocation with stratum sample sizes $\mathbf{n} = (2107, 1472, 1708, 1298, 1193, 789, 735)^\top$ generally did not influence the results considerably.

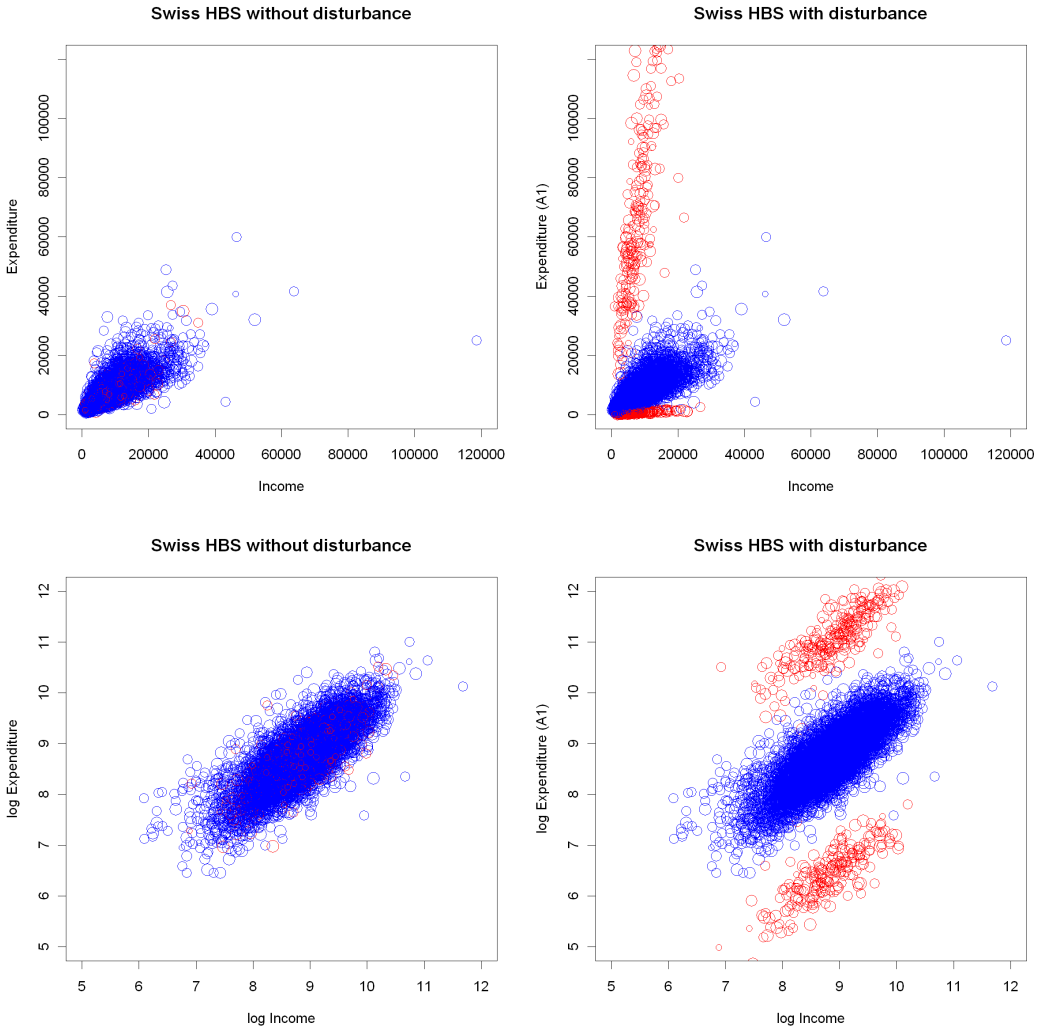

Figure 3: Expenditure versus income in the Swiss Household and Budget Survey universe without (left) and with (right) nonrepresentative outliers (logarithmic scale below)

The universe contains outliers which we consider representative. A random selection of the universe was contaminated. Two proportions of contamination were considered: 5% for scheme A1 and 10% for scheme A2. The contaminated part of the population is referred to as the nonrepresentative outliers. The contamination was created to simulate errors in units, e.g. Cents instead of Franks. In order not to spoil the result too much, a factor a of 10 and 1/10 instead of 100 and 1/100 was applied to the expenditure variable according to the following distributions:

Table 2: Distribution of factor a under contamination A1 and A2

Factor a	0.1	1	10
Proportion for A1	0.0250	0.9502	0.0248
Proportion for A2	0.0502	0.8999	0.0499

The baseline were the estimates from the uncontaminated population. The observable values of the quantities of interest were drawn from the contaminated universes.

The main interest of the simulation below is to elaborate differences between the two main procedures:

1. Application of robust estimates without any correction of the samples with respect to outliers;

2. Identification and elimination of outliers and application of multiple regression imputation.

In both cases, adequate settings had to be investigated, either the correct tuning constants for robust estimators, or the tuning constants for identification of outliers with TRC.

3.2 Estimation of means

Table 3 shows the results of the mean estimators for income for the population without a contamination by nonrepresentative outliers. The median in the population is 6501.01. The Monte-Carlo expectation, denoted $\mathrm{E}\,\widehat{\mu}$ of the

weighted medians of the replicates is 7427.65052 and thus the weighted median is a good estimate of the population median. But of course its bias as an estimator of the population mean is huge. The MC-standard deviation, denoted $\sqrt{\mathrm{V}\,\widehat{\mu}}$, of the weighted median is only slightly larger than the MS-standard deviation of the mean. This is due to the skewness of the variable income. Under a normally distributed variable the median has $\sqrt{\pi/2}$ times the standard deviation of the mean. The variance estimator for the median and the mean are calculated with the R-package `survey` from Lumley [R Development Core Team 2006, Lumley 2006]. The root MC-mean of the estimates for the variance are close to the MC-standard deviation. Note that the some variability is left in the MC-means due to the moderate number of replicates ($R = 1'000$) that could be realised.

The set of one-step M-estimators shows a large bias for small tuning constants (Mean.onestep1 has tuning constant $c = 1$) while for $c = 10$ the bias is negligible. The variance estimator for $c = 1$ underestimates heavily (34% in standard deviation) while for $c = 10$ the bias is negligible. The winsorized means behave very similar. For a winsorisation of $\alpha = 0.005$ the estimator is close to the mean, while for $\alpha = 0.2$ it is closer to the one-step estimator and the median. The Boostrap variance estimator is better than the linearised variance estimator of the winsorized mean.

Table 3: Mean Estimators: No Contamination

Mean μ (true value: 7425.627):

Estimator	$\mathrm{E}\,\widehat{\mu}$	$\sqrt{\mathrm{V}\,\widehat{\mu}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{\mu})}$
MEDIAN.svy	6503.35639	47.28961	45.45671
MEAN.svy	7427.65052	46.03809	44.43854
MEAN.onestep1	6682.20488	42.38378	28.07277
MEAN.onestep3	7270.29677	44.21736	38.61414
MEAN.onestep5	7387.12481	44.90683	42.17518
MEAN.onestep10	7424.26159	45.72493	44.08277
MEAN.wins005	7412.34583	45.63254	43.40769
MEAN.wins01	7400.09689	45.39059	42.86870
MEAN.wins05	7320.99267	44.82976	40.42399
MEAN.wins1	7237.75097	44.25192	38.48452
MEAN.wins2	7092.63914	43.09582	35.57720
MEAN.wins1boot	7237.92980	44.26231	42.43745

Table 4 shows the effect of non-representative outliers in the population. The contamination pushes up the observed population mean to $8'913$. The survey mean estimates the observed population mean well but has a large bias for the true, uncontaminated mean. The median stays close to its former value while the mean is attracted to the contaminated population mean. The variance of the mean is inflated heavily. A tuning constant between $c = 3$ and $c = 5$ would lead to an unbiased estimator of the true uncontaminated population mean. The variance of the onestep estimators is much smaller than the one of the mean. However the variance is underestimated and again the underestimation increases with smaller tuning constants. The winsorised means with $\alpha < 0.05$ are not robust enough. Note the strong bias of the variance estimate for $\alpha = 0.05$. Under contamination scheme A1 the proportion of non-representative outliers equals the winsorizing proportion. Due to sampling variability the proportion of non-representative outliers in the sample may be above or below the winsorizing proportion. This induces a large variance which is not captured by the variance estimator. Only with large enough α the variance drops. The underestimation still seems to be somewhat heavier for these winsorised means than for comparable one-step estimators. Again the Bootstrap variance estimator for the winsorized mean with tuning constant $\alpha = 0.1$ is much better than the closed form variance estimate.

Table 4: Mean Estimators: A1 Contamination

Mean μ, A1 (true value: 7425.627, obs. value: 8913.214):

Estimator	$\mathrm{E}\,\widehat{\mu}$	$\sqrt{\mathrm{V}\,\widehat{\mu}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{\mu})}$
MEDIAN.svy	6503.15672	50.15796	47.83504
MEAN.svy	8918.75352	137.15254	135.90658
MEAN.onestep1	6701.84914	45.18774	30.34084
MEAN.onestep3	7391.04511	50.80556	43.70581
MEAN.onestep5	7664.16278	57.12105	51.66995
MEAN.onestep10	8080.75000	75.60090	70.22488
MEAN.wins005	8819.17023	135.03663	122.67243
MEAN.wins01	8727.79952	133.72136	114.27362
MEAN.wins05	7804.68491	112.42370	57.94730
MEAN.wins1	7445.96521	56.55716	45.07021
MEAN.wins2	7205.67767	49.42340	39.34871
MEAN.wins1boot	7446.75059	57.10231	56.20454

Table 5: Mean Estimators: A1 Contamination and MI

Mean μ, A1 (true value: 7425.627, obs. value: 8913.214):

Estimator	$\mathrm{E}\,\widehat{\mu}$	$\sqrt{\mathrm{V}\,\widehat{\mu}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{\mu})}$
TRC+MEDIAN.svy	6501.55588	44.20731	46.79200
TRC+MEAN.svy	7426.12761	45.63716	45.60892
TRC+onestep1	6679.19898	39.83891	29.13693
TRC+onestep3	7266.48651	42.59433	39.66428
TRC+onestep5	7384.33334	44.02728	43.26460
TRC+onestep10	7422.73912	45.42544	45.24991
TRC+wins005	7410.75969	45.24826	44.59416
TRC+wins01	7398.34344	45.03672	44.02896
TRC+wins05	7318.32959	43.94134	41.48531
TRC+wins1	7235.29494	43.04974	39.52186
TRC+wins2	7090.27262	41.86248	36.59510
TRC+wins1boot	7235.43537	43.16541	43.66767

Table 5 shows the result, when outlier identification with the TRC algorithm, elimination and multiple regression imputation is applied. In practice one would prefer to use a simple mean once the outliers are detected and replaced by imputations. Here we nevertheless apply robust estimators to see, what effect a further robustification has. In this case the TRC tuning constants were very close to the optimum which results in a standard survey mean with negligible bias for the true population mean. The variance estimator based on multiple imputation performs very well. The behavior of the robust es-

timators is very close to the case without contamination. However, it seems that the true variance is recovered much better by multiple imputation in spite of the linearized variance estimator that is used in the multiple imputation procedure. Again the bootstrap variance estimator yields better estimates than the linearized variance estimators.

Table 6: Mean Estimators: A2 Contamination

Mean μ, A2 (true value: 7425.627, obs. value: 10422.15):

Estimator	$\mathrm{E}\,\widehat{\mu}$	$\sqrt{\mathrm{V}\,\widehat{\mu}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{\mu})}$
MEDIAN.svy	6501.42997	52.82870	50.45562
MEAN.svy	10429.74035	187.74815	186.40202
MEAN.onestep1	6723.91477	49.27236	32.92910
MEAN.onestep3	7549.73755	61.93205	49.33963
MEAN.onestep5	8006.54247	75.62242	61.74152
MEAN.onestep10	8842.52609	109.73591	92.86426
MEAN.wins005	10323.10320	185.77687	174.64751
MEAN.wins01	10229.84130	184.56436	167.46096
MEAN.wins05	9449.50262	184.92682	123.74193
MEAN.wins1	8081.58256	170.15373	65.93043
MEAN.wins2	7373.53165	63.43871	44.65481
MEAN.wins1boot	8093.65515	170.56634	169.09158

Table 6 shows some results for the mean with contamination scheme A2. The results are similar results as for contamination A1. However, the optimal tuning constant would be different, i.e. would lead to heavier robustification. Again the variance estimator for the winsorized mean has a heavy downward bias when the contamination proportion equals the winsorizing proportion ($\alpha = 0.1$).

3.3 Estimation of ratios

Table 7 gives an overview of the results from the ratio estimators `onerat` and `robrat` with tuning constants 3, 5, and 10. The estimator `robrat` is a fully iterated M-estimator with preliminary scale estimat MAD of the residuals while `onerat` just is the result of one iteration from the median. Thus `onerat` is the analogue to the one-step estimator for univariate estimation. The estimator `onerat` with a tuning constant $c = 10$ is close to the non-robust ratio estimator, which is optimal when no contamination is present. The estimator `robrat` has a bias even for tuning constant $c = 10$. This shows that already the uncontaminated population has outliers and is not distributed symmetrically. Thus for `robrat` the tuning constant would have to be chosen much larger to make the estimates approximate the true ratio. The variance estimates for `onerat` are slightly biased while for `robrat` the bias is moderate.

Table 8 shows the result under contamination A1. The choice of the optimal tuning constant is again of major importance due to the sensitivity of the bias of the ratio estimates to the tuning constant. Contrary to mean estimation, `onerat` ratio estimation seems generally to underestimate the true variances but to a less extent than before. The optimal tuning constants are now for both onerat and robrat in the range 3 to 10. And the behavior of the two estimators is quite similar. The

Table 7: Ratio Estimators: No Contamination

Ratio ψ (true value: 0.8828425):

Estimator	$\mathrm{E}\,\widehat{\psi}$	$\sqrt{\mathrm{V}\,\widehat{\psi}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{\psi})}$
RATIO.svy	0.88273351	0.00381331	0.00370994
RATIO.onerat3	0.87712272	0.00356503	0.00336276
RATIO.onerat5	0.88121103	0.00371831	0.00358449
RATIO.onerat10	0.88261516	0.00379548	0.00369381
RATIO.robrat3	0.85265838	0.00369364	0.00329352
RATIO.robrat5	0.85244642	0.00390663	0.00351199
RATIO.robrat10	0.85204325	0.00419021	0.00362364

underestimation of the linearised variance estimators behaves similar to the univariate case.

Table 8: Ratio Estimators: A1 Contamination

Ratio ψ, A1 (true value: 0.8828425, observable value: 1.059704):

Estimator	$\mathrm{E}\,\widehat{\psi}$	$\sqrt{\mathrm{V}\,\widehat{\psi}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{\psi})}$
RATIO.svy	1.05994435	0.01573079	0.01551722
RATIO.onerat3	0.88067212	0.00426282	0.00393057
RATIO.onerat5	0.89636871	0.00491184	0.00461583
RATIO.onerat10	0.93208672	0.00668066	0.00629847
RATIO.robrat3	0.85484357	0.00417392	0.00384607
RATIO.robrat5	0.86089502	0.00466531	0.00449412
RATIO.robrat10	0.88330402	0.00601805	0.00618228

Figure 4 gives an overview of the variance distributions of the ratio estimators. One can observe a dependency of the smoothness and normality on the tuning constant and a more skewed distribution of the survey ratio linearized variance estimator. This can be attributed to the fact that robust estimators downweight the influence of some observations which may yield skew variance estimation distributions with outliers.

Applying multiple imputation after outlier nomination and removal yields similar results as for mean estimation. The results are not reported here.

3.4 Estimation of poverty measures

Poverty measures, like the Gini coefficient (GINI or simply G) or the Quintile Share Ratio (QSR or simply Q) in general are highly sensitive to outliers in the income distribution. One exception is the At-risk-of-poverty rate which only takes the proportion of poor to all individuals into consideration and hence not extreme income values. The sensitivity of poverty measures certainly has an impact on the variance estimates. In contrast to linear statistics applied to less skewed distributions, the linearization of highly non-linear statistics in very skewed distributions such as the Quintile Share Ratio, may yield biased point estimates and most notably biased variance estimates. Hence, one has to pay considerably more attention to the choice of the variance estimator. The delete-1-jackknife

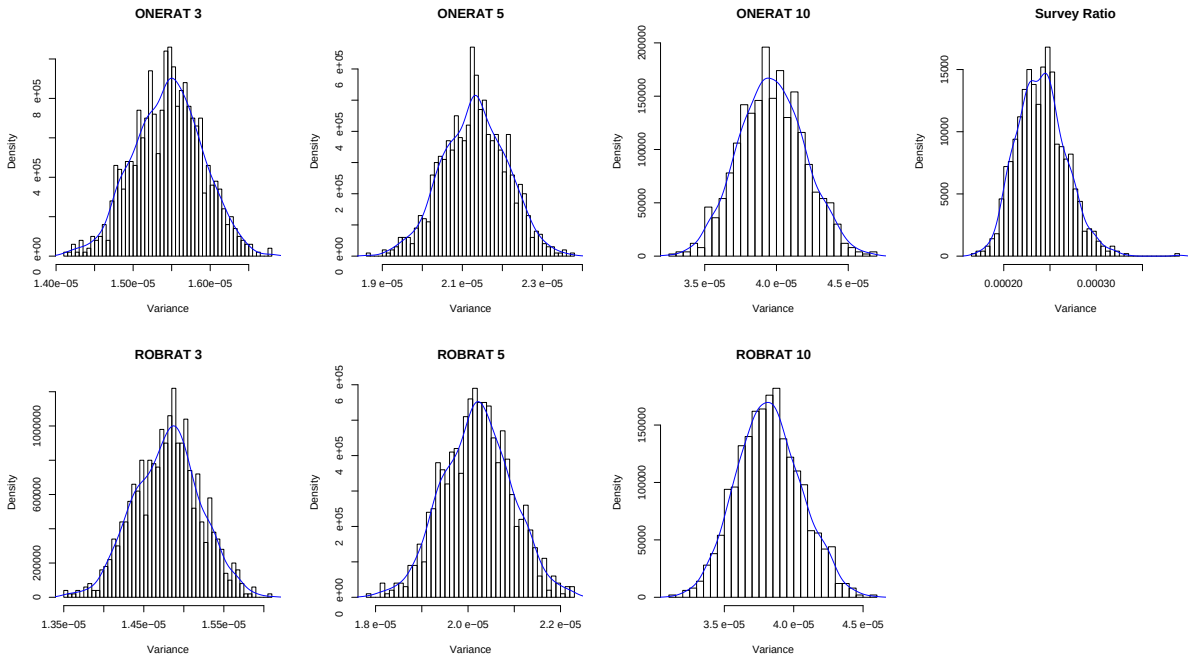

Figure 4: Ratio Estimators: Variance distributions under contamination scheme A1

estimator is omitted due to its insuffiency when applied to non-smooth statistics.

Table 9: Poverty Measures without contamination

G (true value: 0.2972914):

Estimator	$\mathrm{E}\,\widehat{G}$	$\sqrt{\mathrm{V}\,\widehat{G}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{G})}$
G.linearized	0.29734148	0.00233619	0.00622152
G.cal.lin	0.29733250	0.00232330	0.00420530
G.brr	0.29740425	0.00254017	0.00149393
G.boot99	0.29730517	0.00234937	0.00233850

Q (true value: 4.664377):

Estimator	$\mathrm{E}\,\widehat{Q}$	$\sqrt{\mathrm{V}\,\widehat{Q}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{Q})}$
Q.linearized	4.66684881	0.05649330	0.07395028
Q.cal.linearized	4.66666569	0.05629651	0.07396798
Q.brr	4.66870652	0.06273145	0.03684399
Q.boot99	4.66626426	0.05674634	0.05738790

Table 9 shows the possible impact of the selected variance estimator on the accuracy measurement for the Gini-coefficient G and the Quintile Share Ratio Q. In general, when the sample proportions are small in all strata, the bootstrap variance estimator yields very good results without contamination or with contamination (Tables 9 and 10). In case of moderate sampling fractions, however, the bootstrap may tend to underestimate the true variances. In this example, Balanced Repeated Replications (BRR) suffers from the small number of strata which results in unacceptable variance estimates. A synthetic enlargement of the number of strata would easily improve the results. Similar simulations on the German EU-SILC dataset show that under more adequate settings the BRR may deliver good variance estimates [Münnich 2006]. Slightly unexpectedly, the linearization variance estimators tend to overestimate the true variances in the case of the QSR (Q.linearized) and especially the Gini-coefficient (G.linearized). This phenomenon was also observed in the above mentioned EU-SILC based simulation. Applying calibration weights (G.cal.lin and Q.cal.lin) may help to reduce this bias a little but not to the extent needed.

Table 10: Poverty Measures with contamination A1

G (true value: 0.2972914, observable value: 0.4245081):

Estimator	$\mathrm{E}\,\widehat{G}$	$\sqrt{\mathrm{V}\,\widehat{G}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{G})}$
G.linearized	0.42465619	0.00762868	0.00981564
G.cal.lin	0.42465005	0.00762768	0.00918457
G.brr	0.42474723	0.00821429	0.00468553
G.boot99	0.42452150	0.00764980	0.00747598

Q (true value: 4.664377, observable value: 8.20626):

Estimator	$\mathrm{E}\,\widehat{Q}$	$\sqrt{\mathrm{V}\,\widehat{Q}}$	$\sqrt{\mathrm{E}\,\widehat{\mathrm{V}}\,(\widehat{Q})}$
Q.linearized	8.21900561	0.25791294	0.25611228
Q.cal.linearized	8.21894439	0.25789861	0.25619141
Q.brr	8.21632122	0.28483321	0.16084159
Q.boot99	8.21875738	0.25813909	0.25347326

The results in Table 10 show that the non-robustness of the Gini-coefficient and the Quintile Share Ratio yield unacceptable point estimates. However, the variance estimates seem better than in the case without contamination. A solution may be to detect and remove outliers, e.g. via TRC, and then im-

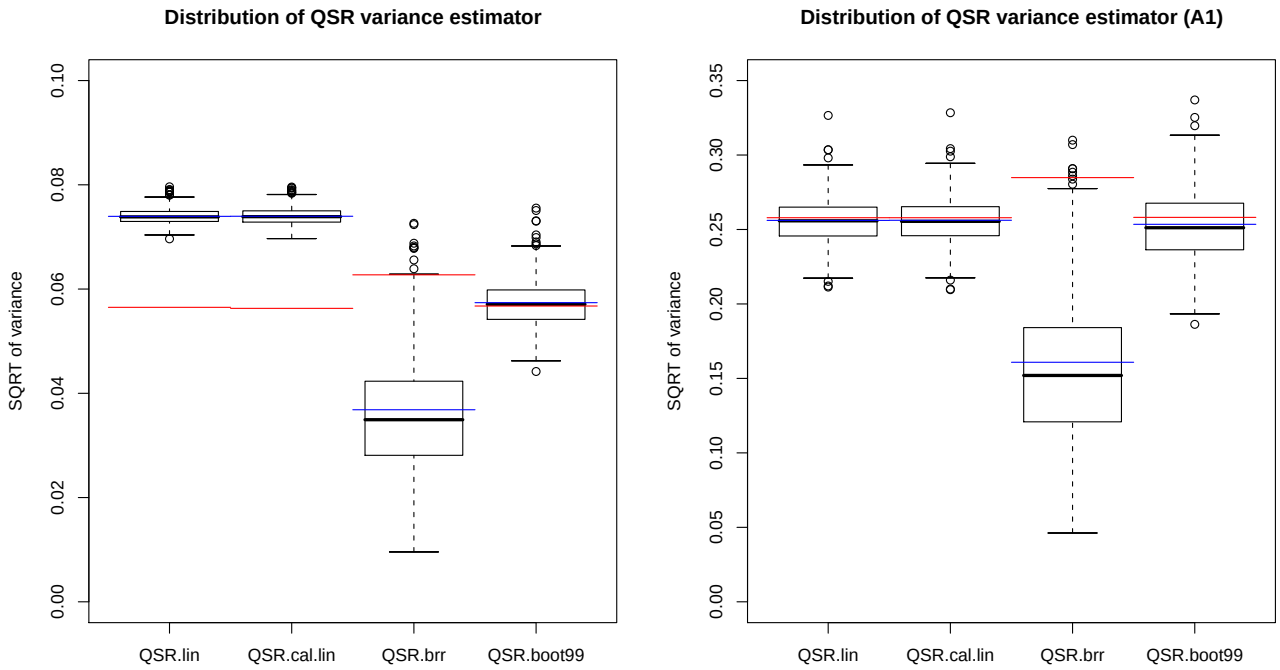

Figure 5: Distribution of QSR Variance Estimators

pute the missing values.

Figure 5 allows to compare the variance estimation distributions of the Quintile Share Ratio. The blue line gives the average estimate, whereas the red line the value to be estimated. Very typical in this situation is the very good performance of the Bootstrap variance estimator with respect to unbiasedness. However, the slightly biased linearized variance estimator generally have smaller variances of the variance estimation distributions.

4 Summary and outlook

The simulation study with a synthetic universe of households showed the well known high impact of contamination on estimators of means, ratios and in particular poverty measures. Robust estimators resist to outliers depending on the tuning constant chosen. The choice of the tuning constant is difficult in practice. Several tuning constants should always be tested. In particular median and non-robust classical estimators should be calculated as extremes. The point and variance estimates and the mean of the robustness weights, which indicates the degree of robustification, may help in the choice of an appropriate tuning constant.

Outlier detection and regression imputation followed by either classical non-robust estimators or mild robustification seems to perform well. The complicated procedure needs resampling variance estimation. The multiple imputation variance estimators in the study and the Bootstrap variance estimators performed well.

Nevertheless, instead of finding the right tuning constant of a robust estimator one has to determine the right tuning constant for outlier detection which is of analogue difficulty.

Linearized variance estimators of robust estimators underestimate the true variance in general. The underestimation depends on the tuning constant. Usually the bias of the linearized variance estimators grows with the degree of robustification. Thus linearized variance estimators can only serve as a first approximation when the robustification is mild. Bootstrap estimators perform well even for stronger robustification, except when the sampling fraction is moderate or large.

For poverty measures the linearized variance estimators moderately overestimated the true variance but had a remarkably low variance themselves. Resampling based variance estimators, especially the bootstrap, seem to outperform the linearized variance estimators.

More situations should be investigated, in particular multivariate outliers and more skew distributions which occur in business surveys. Bias corrections for linearized variance estimators seem promising but may be complicated.

Appendix: Variance of Huber M-estimator with MAD

The Huber M-estimator with the MAD as preliminary scale estimate is the solution of

$$\sum_{i \in S} \psi_c \left(\frac{X_i - \theta}{\delta} \right) = 0.$$

where $\psi_c(x) = \max(-c, \min(c, x))$ for the tuning constant $c > 0$ and δ the median absolute deviation MAD.

Denote $r = (x - \theta)/\delta$ and

$$\begin{aligned} A &= \int_0^\infty \psi'(r)\, f(x;\gamma)\, dx = \int_{\theta - c\delta}^{\theta + c\delta} f(x;\gamma)\, dx \\ B &= \int_0^\infty \psi'(r)\, r\, f(x;\gamma)\, dx = \int_{\theta - c\delta}^{\theta + c\delta} \frac{x-\theta}{\delta} f(x;\gamma)\, dx \end{aligned}$$

for the densitiy $f(x;\gamma)$ of the variable X. If the distribution of X is symmetric around θ then $B = 0$.

The asymptotic variance of the Huber M-estimator is

$$\begin{aligned} V(\theta(F)) &= \frac{\delta^2}{A^2} \int_0^\infty \psi(r)^2\, f(x;\gamma)\, dx \\ &\quad + \frac{B^2}{A^2} \int_0^\infty IF(x;F,\delta)^2 f(x;\gamma)\, dx \\ &\quad + \frac{B\,\delta}{A^2} \int_0^\infty \psi(r)\ IF(x;F,\delta)\, f(x;\gamma)\, dx, \end{aligned}$$

where $IF(x;F,\delta)$ is the influence function of the MAD. The usual variance estimator for M-estimators estimates only the first summand of this asymptotic variance. For asymmetric distributions the second and third summand may be large and variance estimates which neglect them may be heavily biased.

References

[Andersson 1994] Andersson, C., Nordberg, L., (1994), "A Method for Variance Estimation of Non-Linear Functions of Totals in Surveys – Theory and Sorftware Implementation" *Journal of Official Statistics*, **10** (4), S. 395 – 405.

[Béguin and Hulliger 2004] Béguin, C. and Hulliger, B. (2004), "Multivariate oulier detection in incomplete survey data: The epidemic algorithm and transformed rank correlations" *Journal of the Royal Statistical Society A*,**67**, 275–294.

[Binder and Kovačević 1995] Binder, D. A.; Kovačević, M. S., (1995), "Estimatin some measures of income inequality from survey data: An application of the estimating equation approach" *Survey Methodology*, **21**, S. 137 – 145.

[Carrol 1979] Carroll, R. L. (1979), "On Estimating Variances of Robust Estimators When the Errors Are Asymmetric", *Journal of the American Statistical Association*, **74**, 674–679.

[Dennis and Guio 2004] Dennis, I. und Guio, A.-C. (2004) "Poverty and social exclusion in the EU", Statistics in Focus 16, Eurostat, Luxembourg, catalogue number: KS-NK-04-016-ENN. http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-NK-04-016/EN/KS-NK-04-016-EN.PDF

[Deville 1999] Deville, J.-C., (1999), "Variance Estimation for Complex Statistics and Estimators: Linearization and Residual Techniques" *Survey Methodology*, **25** (2), S. 193 – 203.

[Gwet and Lee 2000] Gwet, J.-P., Lee, H. (2000), "An Evaluation of Oultier-Resistant Procedures in Establishment Surveys", Proc. of the Second International Conference on Establishment Surveys, June 2000, Buffalo NY, 707–716.

[Hulliger 1995] Hulliger, B. (1995)"Outlier Robust Horvitz-Thompson Estimators", *Survey Methodology*, **21**, 79–87.

[Hulliger 1999] Hulliger, B. (1999) "Simple and Robust Estimators for Sampling", Proceedings of the Section on Survey Research Methods, American Statistical Association, 54–63.

[Lumley 2006] Lumley, T. (2006), Survey Package for R, URL http://faculty.washington.edu/tlumley/survey/

[Münnich et al. 2003] Münnich, R., Schürle, J., Bihler, W., Boonstra, H.-J., Knottnerus, P., Nieuwenbroek, N., Haslinger, A., Laaksonen, S., Wiegert, R., Eckmair, D., Quatember, A., Wagner, H., Renfer, J.-P. and Oetliker, U., (2003), "Monte Carlo Simulation Study of European Surveys". DACSEIS deliverables. `http://www.dacseis.de`.

[Münnich 2006] Münnich, R. (2006) "Genauigkeit der Armutsberichterstattung in der EU", Presentation at the annual pentecost meeting of the German Statistical Society in Hamburg.

[R Development Core Team 2006] R Development Core Team (2006), "R: A language and environment for statistical computing", R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.