One-Sided Winsorization in Sample Surveys

Robert Clark \(^1\)  Phil Kokic \(^1\)  Paul Smith \(^2\)

\(^1\)National Institute for Applied Statistics Research (NIASRA), University of Wollongong

\(^2\)Southampton Statistical Sciences Institute (S3RI)

ICORS Workshop on Robust Inference for Sample Surveys
UOW 7 July 2017
Outline

- Sample surveys and "representative outliers"
- Winsorization
- Optimal cutoffs for weighted estimation
- Problems with winsorization
- A competing approach (Beaumont and Alavi)
- Simulation study
In business surveys, many financial variables are non-negative, skewed and heavy tailed. There may also be numerous zero values, e.g. capital expenditure.

A common aim is to estimate finite population quantities such as population totals \( t_{yU} = t_{ys} + t_{yr} \), where the subscripts \( U, s \) and \( r \) refer to the population, sample and non-sampled units respectively. The aim is very clearly to estimate the totals including these outliers.

Chambers (JASA, 1986) coined the term “representative outlier”. All sample observations should be used to “represent” themselves in \( t_{ys} \), with \( \hat{t}_y = t_{ys} + \hat{t}_{yr} \). But we must also attempt to estimate \( t_{yr} \) including the likely outliers in the non-sample.

We pay a penalty in bias due to outlier treatment, but if the variance is sufficiently reduced, \( MSE = E \left[ (\hat{t}_yU - t_{yU})^2 \right] \) will improve.
• First considered for surveys by Searls (1966).

• Suppose that the population values of $Y$ were generated independent but not identically from a superpopulation model.

• Suppose that $\hat{t}_y = t_y + \hat{t}_r = t_y + \sum_{i \in s} w_i y_i$ is unbiased.

• Winsorized estimator is
  
  $\hat{t}_y^* = t_y + \sum_{i \in s} w_i y_i^* = t_y + \sum_{i \in s} w_i \min(y_i, K_i)$.

• Searls (1966) considered the special case that $y_i$ are i.i.d. exponentials, with $K_i = K$, in which case $K_{\text{opt}} \propto \mu$ and $K_{\text{opt}} \to \infty$ as $n \to \infty$. 
Kokic and Bell (1994) considered the case of stratified simple random sampling without replacement is used, with simple expansion weights $N_h/n_h$ in each stratum $h$, and a different weight in each stratum. They showed in a design-based framework that

$$K_{h(\text{opt})} = \mu_h^* + L \left( \frac{N_h}{n_h} - 1 \right)^{-1},$$

where $L$ is the solution to $L = -B(L)$ where $B(L)$ is the bias associated with the winsorized estimator. In practice, $\mu_h^*$ is negligible and can be replaced by 0 or a robust estimator independent of location. Estimate $B(L)$ using historical data.
Optimal Cutoffs for Weighted Estimation: General Case

- Suppose population values were generated independently but not identically from a model such that

\[ E \left[ \hat{t}_{yU} - t_{yU} \right] = \sum_{i \in s} w_i \mu_i - \sum_{i \in r} \mu_i = 0. \]  

(1)

- Suppose the winsorized estimator is

\[ \hat{t}_{yU}^* = \sum_{i \in s} y_i + \sum_{i \in s} w_i \min(y_i, K_i). \]

Then the MSE is minimised by:

\[ K_{i(\text{opt})} = \mu_i^* + Lw_i^{-1} \]

\[ L = -B(L) = \sum_{i \in s} w_i (\mu_i^* - \mu_i) \]

(Clark, 1995) where \( \mu_i = E[y_i] \) and \( \mu_i^* = E[\min(y_i, K_i)]. \)
Proof of General Result I

Let \( \sigma_i^{*2} = \text{var} \left[ \min(y_i, K_i) \right] \). Straightforward to show that

\[
MSE = E \left[ \hat{t}_{yU}^* - t_{yU} \right]
= \sum_{i \in s} w_i^2 \sigma_i^{*2} + \sum_{i \in r} \sigma_i^2 + \left\{ \sum_{i \in s} (w_i \mu_i^* - \mu_i) \right\}^2
= \sum_{i \in s} w_i^2 \sigma_i^{*2} + B^2 + \text{constants}
\]
Proof of General Result II

Let \( p_i = P[y_i > K_i] \). Then

\[
\frac{\partial \mu_i^*}{\partial K_i} = \frac{\partial}{\partial K_i} \left\{ \int_{-\infty}^{K_i} y f_i(y) \, dy + \int_{K_i}^{\infty} K_i f_i(y) \, dy \right\} \\
= K_i f_i(K_i) - K_i f_i(K_i) + \int_{K_i}^{\infty} f_i(y) \, dy = a_{0i}
\]

\[
\frac{\partial \sigma_i^*}{\partial K_i} = \frac{\partial}{\partial K_i} \left\{ \int_{-\infty}^{K_i} y^2 f_i(y) \, dy + \int_{K_i}^{\infty} K_i^2 f_i(y) \, dy - \mu_i^* \right\} \\
= K_i^2 f_i(K_i) - K_i^2 f_i(K_i) + 2K_i \int_{K_i}^{\infty} f_i(y) \, dy - 2\mu_i^* a_{0i} \\
= 2K_i a_{0i} - 2\mu_i^* a_{0i} = 2a_{0i} (K_i - \mu_i^*)
\]
Hence

\[
\frac{\partial \text{MSE}}{\partial K_i} = \frac{\partial}{\partial K_i} \left( \sum_{i \in s} w_i^2 \sigma_i^* + B^2 \right)
\]

\[
= w_i^2 \frac{\partial \sigma_i^*}{\partial K_i} + 2Bw_i \frac{\partial \mu_i^*}{\partial K_i}
\]

\[
= 2w_i^2 a_{0i} \left( K_i - \mu_i^* \right) + 2Bw_i a_{0i}
\]

Setting to zero gives \( K_i = \mu_i^* - Bw_i^{-1} \).
Figure 1: $L + \hat{B}(L)$ vs $L$. The points are the sample values of $w_i y_i$. 
The optimal value of $L$ depends on the level at which estimates are produced. If the focus is national total, $L$ is larger (less outlier treatment), than if the focus is industry totals.

How to manage derived variables. E.g. Net Capex = Capital Acquisitions - Capital Disposals. Should we winsorize two of these variables, thereby determining the third? Which two? Or winsorize all three, which will lead to inconsistencies. The inconsistencies can be resolved in several separate ways (Cruddas and Kokic, 1996), but due to their complexity these methods are not used in practice at present.

In many repeated surveys, the focus is (or is seen to be) on change over time, not in estimates of level.
• Model-assisted approach. Linear model relating $y$ and a series of auxiliary variables $x$: $E[y_i] = \beta^T x_i$ with $\text{var}[y_i] = v_i$.

• Generalized regression estimator (GREG):

$$\hat{t}_{yU} = \hat{\beta}^T t_{xU}$$

where $\hat{\beta}$ is the probability-weighted least squares estimator:

$$\sum_{i \in s} \pi_i^{-1} v_i^{-1} \left( y_i - \hat{\beta}^T x_i \right) x_i = 0.$$
One of the proposed alternatives: \( \hat{t}_{yU(M)} = \hat{\beta}_M^T t_x U \) where

\[
\sum_{i \in S} v_i^{-1} Q \psi \left( Q^{-1} \pi_i^{-1} (y_i - \hat{\beta}^T x_i) \right) x_i = 0.
\]

and \( \psi \) is a modification of the Huber function

\( \psi(u) = sgn(u) min(|u|, 1) \).

\( Q \) is a tuning parameter. \( Q \to \infty \) means no outlier treatment, \( Q \to 0 \) means drastic outlier treatment.

\( Q \) is chosen to minimise the estimated MSE of \( \hat{t}_{yU(M)} \). Simplest MSE estimator is (5.3):

\[
\widehat{MSE} \left[ \hat{t}_{yU(M)} \right] = \widehat{\text{var}} \left[ \hat{t}_{yU(M)} \right] + \left\{ \hat{t}_{yU(M)} - \hat{t}_{yU} \right\}^2.
\]
Another proposal was to replace $\pi_i$ by $\pi_i^\alpha$ where $\alpha \in [0, 1]$ is also chosen to minimize MSE.

A weight modification version was also described - this deals with multiple variables more naturally.
• A monthly survey covering all retail businesses with 10 or more employees. Sample size approximately 5000.

• Outcome variable is average weekly sales (y). Auxiliary variable is register turnover derived from value-added tax returns (x).

• Stratified by 27 industries and 6 size categories.

• The top one or two size strata are completely enumerated. Stratum-by-stratum ratio estimation used in the remaining strata. This is equivalent to using a linear regression model with a separate slope for each stratum and zero intercept.
Simulation Study

- Simulation was based on a linear regression model for $\log(y)$ with $\log(x)$ as covariate, with separate models for each of 6 industry groups.
- The model was fitted to 5 months of survey data. Estimated $\sigma$ between 0.43 and 0.91.
- $x$ available from the whole population. $y|x$ simulated from the model to form a population, with observations multiplied by 10 with probability 0.002 to simulate gross outliers.
- In each of 1000 simulations, 29 partially overlapping monthly samples selected by stratified simple random sampling without replacement.
- Months 1-12 used as training data. Various estimates of $t_{yU}$ calculated for months 28 and 29.
- Tuning parameters calculated to minimise MSE overall, and by industry group.
Figure 2: Raw Data
# Table 1: RRMSE% of Estimators of Level (Historical Training Data)

<table>
<thead>
<tr>
<th>Industry Group</th>
<th>no treatment</th>
<th>winsorization optimal for overall</th>
<th>BA method optimal for overall</th>
<th>winsorization optimal for ind.grp</th>
<th>BA method optimal for ind.grp</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>2.14</td>
<td>1.97</td>
<td>2.02</td>
<td>2.38</td>
<td>2.48</td>
</tr>
<tr>
<td>521</td>
<td>3.08</td>
<td>2.87</td>
<td>2.86</td>
<td>2.53</td>
<td>2.55</td>
</tr>
<tr>
<td>522</td>
<td>3.27</td>
<td>2.99</td>
<td>2.98</td>
<td>2.77</td>
<td>2.75</td>
</tr>
<tr>
<td>523</td>
<td>10.25</td>
<td>9.37</td>
<td>9.35</td>
<td>8.38</td>
<td>9.45</td>
</tr>
<tr>
<td>524</td>
<td>3.41</td>
<td>2.99</td>
<td>3.07</td>
<td>3.04</td>
<td>3.23</td>
</tr>
<tr>
<td>526-527</td>
<td>7.13</td>
<td>7.09</td>
<td>7.06</td>
<td>6.78</td>
<td>6.86</td>
</tr>
</tbody>
</table>
Table 2: RRMSE% of Estimators of Level (Historical vs Live Training Data) (optimal for industry group level)

<table>
<thead>
<tr>
<th>Industry Group</th>
<th>no treatment</th>
<th>winsorization historical</th>
<th>winsorization live</th>
<th>BA method historical</th>
<th>BA method live</th>
</tr>
</thead>
<tbody>
<tr>
<td>total</td>
<td>2.14</td>
<td>2.38</td>
<td>2.26</td>
<td>2.48</td>
<td>2.31</td>
</tr>
<tr>
<td>521</td>
<td>3.08</td>
<td>2.53</td>
<td>2.73</td>
<td>2.55</td>
<td>2.70</td>
</tr>
<tr>
<td>522</td>
<td>3.27</td>
<td>2.77</td>
<td>2.96</td>
<td>2.75</td>
<td>2.93</td>
</tr>
<tr>
<td>523</td>
<td>10.25</td>
<td>8.38</td>
<td>9.26</td>
<td>9.45</td>
<td>9.03</td>
</tr>
<tr>
<td>524</td>
<td>3.41</td>
<td>3.04</td>
<td>3.18</td>
<td>3.23</td>
<td>3.22</td>
</tr>
<tr>
<td>525</td>
<td>9.90</td>
<td>9.05</td>
<td>9.60</td>
<td>9.39</td>
<td>9.05</td>
</tr>
<tr>
<td>526-527</td>
<td>7.13</td>
<td>6.78</td>
<td>7.05</td>
<td>6.86</td>
<td>6.91</td>
</tr>
</tbody>
</table>
Table 3: Relative Bias (%) of Estimators of Level (Historical vs Live Training Data) (optimal for industry group level)

<table>
<thead>
<tr>
<th>Industry Group</th>
<th>no treatment</th>
<th>winsorization</th>
<th>BA method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>historical</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>live</td>
<td></td>
</tr>
<tr>
<td>total</td>
<td>-0.10</td>
<td>-1.75</td>
<td>-1.83</td>
</tr>
<tr>
<td>521</td>
<td>0.11</td>
<td>-1.06</td>
<td>-0.93</td>
</tr>
<tr>
<td>522</td>
<td>0.04</td>
<td>-1.27</td>
<td>-1.11</td>
</tr>
<tr>
<td>523</td>
<td>-0.11</td>
<td>-4.31</td>
<td>-4.95</td>
</tr>
<tr>
<td>524</td>
<td>-0.18</td>
<td>-1.66</td>
<td>-1.99</td>
</tr>
<tr>
<td>525</td>
<td>-0.73</td>
<td>-4.54</td>
<td>-2.92</td>
</tr>
<tr>
<td>526-527</td>
<td>0.05</td>
<td>-2.79</td>
<td>-1.71</td>
</tr>
</tbody>
</table>
Table 4: Non-Coverage (%) of Asymptotic and Percentile Bootstrap Confidence Intervals (estimators of level, historical training data at industry group level)

<table>
<thead>
<tr>
<th>Industry Group</th>
<th>no treatment</th>
<th>winsorization</th>
<th>BA method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>asymp</td>
<td>percentile</td>
<td>asymp</td>
</tr>
<tr>
<td>total</td>
<td>7.0</td>
<td>6.4</td>
<td>24.2</td>
</tr>
<tr>
<td>521</td>
<td>7.0</td>
<td>7.5</td>
<td>11.1</td>
</tr>
<tr>
<td>522</td>
<td>7.9</td>
<td>7.3</td>
<td>11.5</td>
</tr>
<tr>
<td>523</td>
<td>10.2</td>
<td>7.0</td>
<td>15.8</td>
</tr>
<tr>
<td>524</td>
<td>10.0</td>
<td>7.8</td>
<td>14.7</td>
</tr>
<tr>
<td>525</td>
<td>9.5</td>
<td>6.1</td>
<td>15.4</td>
</tr>
<tr>
<td>526-527</td>
<td>6.6</td>
<td>4.1</td>
<td>11.4</td>
</tr>
</tbody>
</table>
Conclusions

- When outliers are predominantly positive, both the BA method and one-sided optimal winsorization give useful improvements in MSE for levels, with winsorization doing better (20-40% reductions in industry group MSEs).
- Conflict between best cutoffs for estimating overall vs group.
- Lose about half the gain if live data is used rather than historical training data.
- Asymptotic confidence intervals have poor coverage even for the raw estimator, and particularly for the biased winsorized and BA estimators. Percentile confidence intervals much better, though still far from perfect for the biased estimators.
- “Representative outliers” are a relevant concept whenever the aim is to infer about the population of both outliers and inliers, or to make predictions of new values including outliers.