Adaptations of Winsorization caused by profiling

Arnaud Fizzala – Insee, Methodology department, expert on post-collect treatments for business surveys

Abstract

The sample design of the French Structural Business Statistics survey has changed for the 2016 edition. Now we no longer sample « legal units » but « enterprises ». Data is still collected on legal units, with the following rule : when we select an enterprise, then all legal units within this enterprise will be surveyed. This paper develops the adaptation of winsorization (influential values treatments) to this new context. We show that winsorization with the Kokic and Bell thresholds, applied as if the sampling were a stratified sampling of legal units, seems to be the best option to deal with influential values. We also test alternative methods based on conditional bias, but it leads to poorer results with some problems to solve to be operational.

Introduction

The sample design of the Structural Business Statistics (SBS) French survey has changed for the 2016 edition. Now we no longer sample « legal units » but « enterprises » that is « economic units ». An enterprise is defined as the smallest combination of legal units that is an organisational unit producing goods or services with a certain degree of autonomy. Data is still collected on legal units, with the following rule: when we select an enterprise, then all legal units within this enterprise will be surveyed. Since the statistical units (enterprises) differ from the data collection units (legal units), the sample design can be seen as a two-stage cluster sampling. Enterprises are randomly selected and then all legal units within those enterprises are included in the sample.

Until now and for the 2016 edition again[1], SBS results are computed on the population of legal units, based on data produced at the legal unit level. SBS surveys post-collect treatments (non-response, influential units and calibration treatments) have to be adapted to handle with the new sampling design.

This paper develops the adaptation of influential units treatments with winsorization to this new context.

At this stage of the study, winsorization with the Kokic and Bell [4] threshold, applied as if the sampling were a stratified sampling of legal units, seems to be the best option to deal with the problem of influential values. Alternative methods based on conditional bias [1 and 2] are also tested, but it leads to poorer results with some problems to solve to be operational[2].

Current treatment of influential values

Economic variables with highly skewed distribution are very usual in business survey. In this context, we often face influential units problems. In this paper, we assume that measurement errors (gross error, unity error...) have already been detected and corrected at the editing stage. Influential values are typically very large but “true”, and their presence in the sample tends to make classical estimators very unstable. The aim

1 For the next editions, SBS results will be computed on the population of enterprises, based on “enterprises data”.

2 See annex for details.

of influential values treatment is to limit their impact, which leads to estimators that are more stable but potentially biased.

Winsorization is the method used in the French SBS survey to treat influential values. It is based on the determination of thresholds in sampling strata in the case of stratified random sampling above which large values are reduced. Kokic and Bell [4] determined the value of the thresholds which minimize the winsorized estimator's mean square error.

Until the 2015 edition, we applied winsorization on the legal units' turnover available for all legal units in the sampling frame thanks to administrative information (especially fiscal data) [3]. Precisely, we applied Winsorization for the estimation of the total turnover by activity[3] and threshold were obtained by the Kokic and Bell method. After that, we modified the weights of the winsorized units so that effect of winsorization could be "transferred" to the other variables.

Kokic and Bell method was developed in a stratified sampling framework. As we use now a two-stage cluster sampling, there is no guarantee that the method holds.

To appreciate how far of the theoretical hypothesis we are, we compute some descriptive statistics on the sampling weights of legal units by stratum[4] (table 1).

Table 1 : Distribution of weights by stratum of legal units

Quantile	CV (%)	Range	Frequency of the mode
100%	401,7	797	100%
99%	118,1	265	100%
95%	59,5	153	100%
90%	38,7	96	100%
75%	20,7	50	99%
50%	11,9	20	97%
25%	6,1	9	91%
10%	2,0	1	84%
5%	0,0	0	80%
1%	0,0	0	68%
0%	0,0	0	47%

Note : Table 1 is computed by stratum of legal units (1 observation = 1 stratum). 95% of the activities have a coefficient of variation of the legal units weights less than 59,5%. 75% of the activities have a range of the legal units weights less than 50. In each activity, at least 47% of the legal units have the same weights.

We see that the majority of the units in a stratum have the same weights. We also see that the remaining units can have very different weights. As we have a big proportion of units with the same weights within each stratum, we can expect good results with Kokic and Bell method applied as if it was a stratified sample of legal units.

To confirm this intuition, we conduct a simulation study based on 1000 replications of the new sampling design. We compare « Kokic and Bell estimator »[5] to the classical Horvitz-Thompson estimator and some robust estimators based on conditional bias.
The study is based on the most recent complete information we have, referring to the year 2015.

3 NACE, 3 positions.

4 Meaning stratum if the sampling design still were a stratified sampling of legal units, but weights resulting of the new sampling design. *Remark : If the sampling design still were a stratified sampling of legal units, each legal unit in a stratum would have the same weight.*

5 Below, we write "Kokic and Bell estimator" for winsorized estimator with Kokic and Bell thresholds computed as if the sampling design was a stratified sampling of legal units.

Robust estimators based on conditional bias

The formal framework of robust estimation based on conditional bias is described in [1] and [2]. We simply remind here some elements for the understanding of the study.

Robust estimator minimize[6] the conditional bias of the most influential unit in the respondent population. Formally, Robust estimator for a variable Y is :

$$\hat{t_{yR}} = \hat{t_y} - \frac{1}{2}(B_{min} + B_{max})$$

With :

- $\hat{t_y}$ the Horvitz-Thompson estimator of the total of Y;

- B_{min} and B_{max}, the minimum and maximum conditional bias in the respondent population.

The conditional bias associated with the unit i is a measure of influence. It is the deviation from the population total we would observe if we were computing the mean of Horvitz-Thompson estimators on each sample containing the unit i.

$$B_i = E_p(\hat{t_y} / I_i = 1) - t_y$$

Conditional bias takes into account the sample process (sampling design and non-response modelling) used.

For the study, we consider two phases in the sample process : the first phase is the sampling of the legal units we send a questionnaire to, the second phase is the "selection" of the legal units answering. The second phase is modelled as a Poisson sampling. It is a classical modelling of non-response phenomenon in survey methodology studies.

We test two versions of the first phase which lead to two different robust estimators :

1 – Poisson sampling of enterprises ;

2 – Stratified sampling of enterprises.

Version 1 does not correspond to the « real » sampling design, but is easier to implement in operational phase.

We adopt the following notation:

π_{1i} and π_{1ij} denote the first-order and second-order probabilities of legal units in the first phase ;

π_{2i} and π_{2ij} denote the first-order and second-order probabilities of legal units in the second phase ;

π_{1E} denote the first-order probabilities of enterprises in the first phase ;

m_h : number of enterprises selected in the sample in the stratum h in the first phase ;

M_h : number of enterprises in the sampling frame in the stratum h in the first phase ;

y_E : sum of the turnover of legal units with the same activity than i and contained by the enterprise E;

t_{yh} : sum of the turnover of legal units with the same activity than i and contained by enterprises in the sampling frame in the stratum h.

Conditional bias of the unit i, for an arbitrary design in the first phase and Poisson sampling in the second phase, is [2] :

---

6 In the class of estimators of the form $\hat{t_{yR}} = \hat{t_y} + \delta$

$$B_i = \sum_{j \in U} \left(\frac{\pi_{1ij}}{\pi_{1i}\pi_{1j}} - 1 \right) y_j + \frac{1}{\pi_{1i}} \left(\frac{1}{\pi_{2i}} - 1 \right) y_i$$

In the version 1, we distinguish 3 situations[7] :

a) $j = i$ so $\pi_{1ij} = \pi_{1i} = \pi_{1E} = \frac{m_h}{M_h}$;

b) $j \neq i$ and $j \in E$ so $\pi_{1ij} = \pi_{1i} = \pi_{1E} = \frac{m_h}{M_h}$;

c) $j \neq i$ and $j \notin E$ so $\pi_{1ij} = \pi_{1i}\pi_{1j}$.

So we have :

$$B_i^1 = \left(\frac{M_h}{(m_h \cdot r_i)} - 1 \right) y_i + \left(\frac{M_h}{m_h} - 1 \right) (y_E - y_i)$$

In version 2, we distinguish 4 situations :

a) $j = i$ so $\pi_{1ij} = \pi_{1i} = \pi_{1E} = \frac{m_h}{N_h}$

b) $j \neq i$ and $j \in E$ so $\pi_{1ij} = \pi_{1i} = \pi_{1E} = \frac{m_h}{N_h}$

c) $j \neq i$ and $j \notin E_i$ and $E_j \in h$ so $\pi_{1ij} = \frac{m_h}{M_h} \frac{(m_h - 1)}{(M_h - 1)}$

d) $j \neq i$ and $j \notin E_i$ and $E_j \notin h$ so $\pi_{1ij} = \pi_{1i}\pi_{1j}$.

So we have :

$$B_i^2 = \left(\frac{M_h}{(m_h \cdot r_i)} - 1 \right) y_i + \left(\frac{M_h}{m_h} - 1 \right) (y_E - y_i) + \left(\frac{M_h}{m_h} \frac{(m_h - 1)}{(M_h - 1)} - 1 \right) (t_{yh} - y_E)$$

Remark : We can see that $B_i^2 = B_i^1 + \left(\frac{M_h}{m_h} \frac{(m_h - 1)}{(M_h - 1)} - 1 \right) (t_{yh} - y_E)$ *. In Version 2, conditional bias depends on the level of* y_E *in his stratum. With stratified sampling, selecting E is reducing the odds to select the other enterprises in the stratum. That is not the case with Poisson sampling.*

7 E is the enterprise containing i.

Simulations

To evaluate the quality of the estimators, we select 1000 samples with the new sampling design and estimate for each of the 207 activities the total of turnover with the estimators previously presented. As the turnover is available in the sampling frame, the true value of the total is also known. In the real SBS results however, the answers to the survey are used to compute a more accurate value for each legal unit's activity. In the simulations presented here, we use the value of activity available for all units in the sampling frame.

Next, we calculate the mean square error (MSE) of an estimator X for an activity :

$$MSE = \frac{1}{1000} \sum_{k=1}^{1000} (\hat{t_{yX}} - t_y)^2$$

To make the value of MSE more interpretable, we divide it by the MSE of the Horvitz-Thompson estimator using the same units to obtain a relative mean square error. If superior to 100 %, robust estimates are less accurate than the usual expansion estimator. Otherwise, they are able to increase estimates precision :

$$MSER = \frac{\left(\frac{1}{1000} \sum_{k=1}^{1000} (\hat{t_{yX}} - t_y)^2\right)}{\left(\frac{1}{1000} \sum_{k=1}^{1000} (\hat{t_{yHT}} - t_y)^2\right)}$$

Table 2 : Distribution of MSER by activity

Quantile	Kokic and Bell	Robust V1	Robust V2
100 %	100 %	131 %	141 %
99 %	100 %	108 %	100 %
95 %	88 %	101 %	95 %
90 %	84 %	98 %	92 %
75 %	77 %	93 %	87 %
50 %	67 %	83 %	78 %
25 %	43 %	61 %	59 %
10 %	16 %	39 %	39 %
5 %	10 %	31 %	29 %
1 %	1 %	24 %	22 %
0 %	1 %	22 %	19 %

Note : Table 2 is computed by activity (1 observation = 1 activity). In half of the activities, MSER is less than 67% with the Kokic and Bell estimator, less than 83% with the robust estimator V1, less than 78% with the robust estimator V2.

We see that Kokic and Bell estimator obtains the best results. Robust estimators have good results too : better than Horvitz-Thompson in more than 95 % of the activities for the Version 2.

To evaluate if Kokic and Bell estimator is systematically better than Robust estimators, we compute in table 3 the ratio between the robust estimators MSE and the Kokic and Bell estimator MSE.

Table 3 : Distribution of the ratio between the robust estimators MSE and the Kokic and Bell estimator MSE by activity

Quantile	Robust V1 / Kokic et Bell	Robust V2 / Kokic et Bell
100 %	27,6	27,5
99 %	23,0	22,8
95 %	3,6	3,5
90 %	2,3	2,2
75 %	1,5	1,4
50 %	1,3	1,2
25 %	1,2	1,1
10 %	1,1	1,0
5 %	1,0	1,0
1 %	0,5	0,6
0 %	0,3	0,4

Note : Table 3 is computed by activity (1 observation = 1 activity). The ratio between the robust estimator V1 MSE and the Kokic and Bell estimator MSE is more than 1,1 in 90% of the activities.

We see that Kokic and Bell estimator is better than the Robust Estimators (V1 or V2) in more than 90 % of the activities. The reasons of these are probably two folds :
- As we have already seen, the majority of the legal units in a stratum have the same weights ;
- Aim of robust estimation based on conditional bias is to minimize the influential of the most influential unit whereas aim of winsorization is to minimize the MSE which is also the indicator of quality that we use in the study.

In the lights of these results, we consider that winsorization with the Kokic and Bell threshold, calculated as if the sampling were a stratified sampling of legal units is the best option to deal with influential values.

To evaluate the impact of this winsorization on other variables than turnover, we compute the MSER for the estimator of totals of other variables with the weights after winsorization on turnover as described in the previous section. The other variables are :

- Value added ;
- Investments ;
- Number of legal units.

Table 4 : Distribution of MSER of estimators of total with winsorized weights by activity

Quantile	Turnover	Value added	Investments	Number of legal units
100 %	100 %	100 %	100 %	124 %
99 %	100 %	100 %	100 %	120 %
95 %	88 %	99 %	100 %	108 %
90 %	84 %	97 %	100 %	105 %
75 %	77 %	90 %	99 %	102 %
50 %	67 %	81 %	92 %	100 %
25 %	43 %	64 %	68 %	99 %
10 %	16 %	31 %	28 %	96 %
5 %	10 %	20 %	12 %	93 %
1 %	1 %	3 %	4 %	90 %
0 %	1 %	0 %	0 %	82 %

Note : Table 4 is computed by activity (1 observation = 1 activity). In half of the activities, MSER is less than 67% for turnover, less than 81% for value added, less than 92% for investments, less than 100% for number of legal units.

We see that even on investment which is a variable with low correlation with turnover, winsorization improves the estimators. On the number of legal units, winsorization has a « neutral » effect : MSER is better in half of the activities and worse in the other half, for 25 % maximum.

Conclusion

Winsorization with the Kokic and Bell threshold, applied as if the sampling were a stratified sampling of legal units appears to be the best option to deal with influential values. It has the better results on estimation of turnover and other variables like value added and investments.

Bibliography

[1] J-F. Beaumont, C. Favre Martinoz, D. Haziza*, A method of determining the winsorization threshold, with an application to domain estimation.* Survey Methodology, vol. 41, n°1, June 2015 ;

[2] J-F. Beaumont, C. Favre Martinoz, D. Haziza*, Robust Inference in Two-phase Sampling Designs with Application to Unit Non-response.* Scandinavian journal of statistics vol. 43, 2016 ;

[3] T. Deroyon, *Traitement des valeurs atypiques d'une enquête par winsorization - application aux enquêtes sectorielles annuelles*. Acte des JMS 2015.

[4] P.N. Kokic, P.A. Bell, Optimal winsorizing cut-offs for a stratified finite population estimator, Journal of Official Statistics, vol 10, n° 4, pp 419-435, 1994.

Annex – Practical problems with robust estimators

The form of robust estimators based on conditional bias $\hat{t}_R=\hat{t}-\frac{1}{2}(B_{min}+B_{max})$, Horvitz-Thompson estimator with an additional term, is not easy to use in production, because current computer programs work with estimators with a "linear form", that is a form of this type : $\hat{t}_R=\sum_{i\in S} w_i\, y_{iR}$

The paper [2] mentions a method to transform the robust estimator in a linear form, that is transfer the adjustment on the estimates on adjustments on weights or variables of interest values. It is based on a constant c, as :

$$y_{iR}=y_i-\frac{B_i-\psi_c(B_i)}{w_i} \quad \text{with} \quad \psi_c(B_i)=sign(B_i)\times min(|B_i|,c)$$

An algorithm to calculate y_{iR} for each units of the data is described in the paper [2]. We can see, with ψ_c form, that only units with the more important conditional bias (greater in absolute value than c) will obtain an y_{iR} different than y_i.

We currently work in SBS with winsorized weights. Doing so, winsorization built on turnover is transferred to estimators of other variables. The rationale of this method is two fold :

- Winsorization on turnover will have benefits on estimators of variables correlated with turnover ;

- Accounting links are preserved[8].

To pass from y_{iR} to w_{iR}, we can use the relation below :

$$w_{iR}=w_i\frac{y_{iR}}{y_i}$$

But this is not possible when $y_i=0$. This problem appears at least one time in our 1000 replications in a third of the activities.

An alternative method is described in paper [1]. The aim is to determinate the value K as :

$$\begin{array}{ll} y_{iR}=y_i & w_i\, y_i\le K \\ y_{iR}=\dfrac{K}{w_i} & w_i\, y_i>K \end{array} \quad \text{and} \quad \hat{t}_R=\sum_{i\in S} w_i\, y_{iR}$$

The advantage of this method is that the targets are the highest values of $w_i y_i$, so the units with $y_i=0$ will not be concerned. But there is another side to the coin because of the cases where $\hat{t}_R>\hat{t}$. With this method, we have, by construction, $y_{iR}\le y_i$ and so $\hat{t}_R\le\hat{t}$.

If $-\frac{1}{2}(B_{min}+B_{max})>0$ K is not calculable and the method does not work. It happens :

- Never with version 1 (conditional bias with Poisson sampling are positive) ;

- At least one time on the 1000 replications for half of the activities with version 2.

We are still working on ways to overcome this limit of conditional bias methods and to develop a method to translate robust estimates adjustments into new estimation weights for each sampled unit.

8 It is not the case if each variable is winsorized separately.