UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert Meeting on Statistical Data Editing

3-7 October 2022, (virtual)

Comparison between Clark and Kokic and Bell approaches

Romain Lesauvage (Insee, France)

romain.lesauvage@insee.fr

I. Introduction

Economic variables with highly skewed distribution are very usual in business survey. In this context, we often face influential units problems. In this work, 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 of influential values treatment is to limit their impact, which leads to estimators that are more stable but potentially biased.

Thus, the determination of atypical units is an main issue in business statistics. To do this, we proceed by winsorization, which is the transformation of a variable of interest Y into another variable called winsorized one, noted Y^*.There are two main ways to compute Y^*, we mainly use the so-called type II winsorization which is such that, for each stratum h,

$$Y^* = \begin{cases} Y \quad \text{si} \quad Y \leqslant K_h \\ \dfrac{n_h}{N_h} Y + \left(1 - \dfrac{n_h}{N_h}\right) K_h \quad \text{si} \quad Y > K_h \end{cases}$$

The idea is to fix a threshold K_h from which we modify the value of Y. The question is to determine this threshold K_h.

We worked here on the comparison between two methods and their implementation in two different programming languages: on the one hand, Kokic and Bell showed in a 1994 paper [1] that the optimal threshold for a stratified simple random survey is obtained by $K_h = -\dfrac{B}{\dfrac{N_h}{n_h} - 1} + \mu_h$ where B is the opposite of the bias of the optimal winsorized estimator and μ_h is the average of Y in the h stratum obtained by an independent survey (for example the former edition of the survey), n_h is the number of units of the stratum h in the survey and N_h is the number of units in the stratum h in the database. INSEE has developed a macro-SAS implementing this method.

On the other hand, Clark in 1995 [2], proposed to compute the L value, defined as -B where B is the bias of the Kokic and Bell formula, in another way, in a more general context (not only stratified simple random samples), and this is implemented in the package *surveyoutliers*. The threshold is then computed in an analogous way by replacing μ_h by the mean of the winsorized estimator obtained by linear regression.

We propose in this work to show how the formula proposed by Clark can be put in touch with Kokic and Bell version in the calculation of the optimal bias and to compare the results obtained by the macro-SAS and the R package, by focusing on the differences in results. Another goal is to see whether or not, it could be simple to switch from SAS to R in the process of adjustmen in business survey. To illustrate the differences between the two methods, we use the data from ESA, a French annual business survey [3].

II. The two methods

A. Kokic and Bell

Kokic and Bell (1994) propose to calculate the thresholds K_h in such a way that they are independent of the sample to which they are applied and in such a way as to minimise the squared error of the winsorized estimator, this error being calculated by taking into account the randomness resulting from both the sample design and the distribution of the variable *Y* in the population.

Thus, on average over all the samples and over the possible values that Y can take in these samples, i.e. over all the situations to which the survey data can confront us, the winsorized estimator has the lowest possible mean square error.

Kokic and Bell thus seek to calculate a winsorized estimator whose properties extend those of the Horvitz-Thompson estimator. The latter is indeed unbiased insofar as, if all possible samples of the population were drawn and the Horvitz-Thompson estimator of the total of Y were calculated in each of them, the average of these estimators weighted by the probability that each sample has of being selected would be exactly equal to the total of Y.

The winsorized estimator calculated with the Kokic and Bell thresholds is no longer unbiased, but it has the smallest error in estimating the total of Y on average over all possible samples and over all possible values of Y in those samples.

Kokic and Bell then show that in this case the K_h thresholds are asymptotically equivalent in each stratum to $-\frac{B}{\frac{N_h}{n_h}-1}+\mu_h$, where μ_h is the expectation of Y in stratum h and B is the bias of the minimum winsorized estimator.

In the following formulas, the notation Y_{hi} representing the value of Y for individual i in stratum h will be simplified by Y_h . The same will apply to other quantities.

This bias is calculated as the point that cancels the function *F* defined by :

$$F(B)=-B[1+\sum_h n_h E_h(J_h^*)]-\sum_h n_h E_h(Y_h^* J_h^*)$$

with E_h the expectation according to the distribution of Y in stratum h, $Y_h^*=(\frac{N_h}{n_h}-1)(Y_h-\mu_h)$ and J_h^* the indicator which is 1 if Y_h is greater than K_h.

To calculate the optimal value of B in practice, it can be noted that by positing L=-B and estimating the necessary quantities by the values on an independent sample, the function F is re-written $\hat{F}(L)=L[1+\sum_h(\frac{n_h}{m_h})\sum_j^{m_h} I(\tilde{Y}_j^h>L)]-\sum_h(\frac{n_h}{m_h})\sum_j^{m_h}\tilde{Y}_j^h I(\tilde{Y}_j^h>L)$ where m_h is the number of individuals in stratum h from a previous survey, $\tilde{Y}_j^h$ is the value of Y* for the unit j in stratum h in the same survey, and I represents the indicator.

One can then notice that the function is affine and continuous by pieces with jumps for each value of Y_k observed and one can thus estimate L (and B) optimum by linear interpolation between the last rank where F is negative and the first where it becomes positive.

This method is implemented at INSEE by two macro-SAS which take as input a table corresponding to the sample, and a table corresponding to a set of Y values supposed to be independent of the sample.

B. Clark

The surveyoutliers package allows the calculation of optimal cut-offs for a variable in a more general framework than the case of the stratified simple random survey studied by Kokic and Bell. It is based on the paper by R. Clark (1995), which is based on the case where auxiliary variables are available and where, in reality, a regression relationship between our variable of interest *Y* and the auxiliary variables is given as input. The results of Kokic and Bell are thus extended to the case of a GREG (generalized regression) estimator, which allows us to obtain results for other designs than stratabased SAS.

The hypothesis of Kokic and Bell can in fact also be integrated in this case: it can be rewritten as the hypothesis that within each stratum, the value of *Y* follows a law of the type : $Y_h=\mu_h+\epsilon_h$ where ϵ_h is white Gaussian noise.

By noting $Y^*=min(K_h,Y)$, Clark shows that the values K_h verify asymptotically $-\frac{B}{\frac{N_h}{n_h}-1}+\mu_h^*$ with $\mu_h^*=E[min(K_h,Y)]$ which in practice we will try to estimate. In addition, the bias *B* is calculated as $B=\sum_i(\omega_i-1)(\mu_i^*-\mu_i)$ where $\mu_i^*=E(Y_i^*)$, $\mu_i=E(Y_i)$ and ω_i is the weight of the unit *i*, which is assumed to be equal to $\frac{N_h}{n_h}$ in the case of a stratified simple random survey.

Since μ_i^* is difficult to calculate, Clark proposes a rewriting of the problem, posing $D_i=(Y_i-\mu_i)(\omega_i-1)$ and *L=-B*, we can show that we have $B(L)=-E[\sum_{i\in s} max(D_i-L,0)]$ which can be estimated by $\hat{B}(L)=-E[\sum_{i\in s} max(\hat{D}_i-L,0)]$ with $\hat{D}_i=(Y_i-\hat{\mu}_i)(\omega_i-1)$.

Finally, we can solve the problem by finding *L* that satisfies $\psi(L)=L+\hat{B}(L)=0$

In fact, this formula is similar to the method detailed by Kokic and Bell in the case of a stratified simple random survey, for them we have :

$$\hat{F}(L)=L[1+\sum_h(\frac{n_h}{m_h})\sum_j^{m_h} I(\tilde{Y}_j^h>L)]-\sum_h(\frac{n_h}{m_h})\sum_j^{m_h}\tilde{Y}_j^h I(\tilde{Y}_j^h>L)$$

This becomes with Clark's notations :

$$\hat{F}(L)=L[1+\sum_h(\frac{n_h}{m_h})\sum_j^{m_h} I(\hat{D}_i-L\geq 0)]-\sum_h(\frac{n_h}{m_h})\sum_j^{m_h}\hat{D}_i I(\hat{D}_i-L\geq 0)$$

Putting the terms together, we have :

$$\hat{F}(L)=L+\sum_h(\frac{n_h}{m_h})\sum_j^{m_h}(L-\hat{D}_i)I(\hat{D}_i-L\geq 0)$$

This gives us :

$$\hat{F}(L)=L-\sum_h\left(\frac{n_h}{m_h}\right)\sum_j^{m_h}(\hat{D}_i-L)I(\hat{D}_i-L\geq 0)=L-\sum_h\left(\frac{n_h}{m_h}\right)\sum_j^{m_h} max(\hat{D}_i-L,0)$$

With the assumption $m_h=n_h$, we therefore find $\hat{F}(L)=\psi(L)$: the same function must therefore be cancelled.

The package therefore proposes an implementation of this method via two functions optimal.onesided.cutoff and optimal.onesided.cutoff.bygroup depending on whether you wish to work on one or several domains. These functions take as parameters :

- A formula that explains Y as a function of auxiliary variables, here we take Y ~ id_strate
- A dataset that must necessarily contain two variables named piwt, which is the inverse of the selection probability, and gregwt which is the weight to be used in the regression. In our case, we will assume that the two variables are equal to the weight $\frac{N_h}{n_h}$
- Possibly the name of the domain variable (for the multi-domain version)
- In the event that we do not have $m_h=n_h$, a historical.reweight parameter to weight the formula by $\frac{n_h}{m_h}$ where needed.
- Other optional parameters that can be modified in particular if we have succeeded in estimating the μ_i^*

Note: Unlike SAS macros, the R package takes as input a single data set corresponding to the sample. To winsorise as recommended in the paper by Kokic and Bell, it is therefore necessary to launch the calculation of the thresholds on the set of independent observations and to deduce, via a code (external to the package) corresponding to the formula mentioned in the first part, the value of the Y* on our sample or to use the historical.reweight parameter.

There are two major differences between the package and the macro-SAS from a programming point of view:

1. The estimate of $\mu_h^*=E[min(K_h,Y)]$ which is assimilated to $\bar{y}_h$ under SAS while a linear regression under R is used.
2. The interpolation method: simple linear interpolation in SAS versus using the *uniroot* function in R.

III. Main results

The two methods are compared[1] on the dataset from ESA 2020. Winsorization is implemented on the one hand on independent legal units and on the other hand on legal units that belong to profiled enterprises.

The aim is to compare the number of winsorized units in the two methods and to analyse the differences obtained. We will test three scenarios:

- Calculate the thresholds on an independent data set.
- Calculate thresholds on the sample.
- Calculating the thresholds on the sample with the historical.reweight parameter.

We have at our disposal 2,309,714 independent legal units (LU), distributed in 2,122 strata and 140,969 legal units belonging to profiled enterprises (PE LU) distributed in 2,213 strata for the independent data set. In the sample we have 24,291 independent legal units in 1,556 strata and 5,198 legal units belonging to profiled enterprises in 590 strata.

The results are as follows:

	SAS	**R – independent data**	**R - sampling**	**R – sampling + hist.reweight**
Independent LU	283	35	1 616	1 448
PE LU	158	28	459	340
Total	441	63	2 075	1 788

It can therefore be seen that there are 7 times fewer winsorised units in the surveyoutliers package than in the SAS macro currently used when using the independent data set.

In detail, the thresholds are lower in the R package for only 3 strata of independent legal units and 4 strata of legal units of profiled enterprises. The threshold therefore tends to be higher in the case of the package (on average, it is 4 times higher), compared to the SAS macro.

On the contrary, when using the sample in the R package, we have 5 times more winsorised units than with the SAS macro. Using a correction factor for the weights based on the independent sample reduces this to only 4 times more, but there is still a significant difference between the two methods (SAS and R).

IV. Conclusion

To conclude, the switch from the SAS macro to the R package cannot be done automatically without precaution. Changing the tool changes the results significantly, especially in terms of the number of winsorised units.

In order to determine which method is better, the study should be continued by comparing the results obtained in each case in terms of bias and variance and looking at what provides the best results.

[1] Each time, a winsorization is performed by group (APE 3 positions) and the thresholds are calculated from the Y of the sampling frame. In the SAS macro, the Nh correspond to $\hat{N}_h = \sum_{i \in h} w_i$ and n_h is the number of respondents in stratum h. In the R function, the same procedure is followed by putting this information as input.

Bibliography

[1] Kokic, P. and Bell, P. (1994), "Optimal winsorizing cutoffs for a stratified finite population estimator," J. Off. Stat., 10, 419-435

[2] Clark, R. G. (1995), "Winsorisation methods in sample surveys," Masters thesis, Australian National University

[3] Enquête sectorielle annuelle – ESA, https://www.insee.fr/fr/metadonnees/source/serie/s1269