ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 175, 2004

*Janusz Wywial**, *Tomasz Żądło***

# ON SOME ROBUST AGAINST OUTLIERS PREDICTOR OF THE TOTAL VALUE IN SMALL DOMAIN

**Abstract**. The problem of prediction of the total value in a domain based on simple regression superpopulation model (with one auxiliary variable and no intercept) is considered. The problem of robust estimation against outliers of regression function's parameter is shown. The presented robust estimator is median value of gradients of all straight lines each determined by the origin and one of n points $(x, y)$, where n is sample size, $y$ – the variable of interest and $x$ – auxiliary variable. This estimator is simplified form of the estimator presented by H. Theil (1979). The equation of the mean square error of the robust predictor based on the robust estimator of regression's parameter is derived for asymptotic assumptions. The best linear predictor based on the considered superpopulation model is presented. The equation of mean square error of the BLU predictor is derived. The accuracy of these predictors is compared for the assumption of normal distribution of variables of interest.

**Key words:** small area statistics, model approach, robust estimation.

## 1. INTRODUCTION

In survey sampling, including small area statistics, two approaches are considered – random and model approach. There is also a problem of robust estimation, extremely important in respect of practical aspects of sample surveys especially supported by model approach. The reason is, that the statistician in the case of model approach must assume some superpopulation model and estimate its parameters. In this paper, some robust predictor of total value in domain will be proposed, and it will be compared with BLU predictor for assumed, presented below, superpopulation model. Robustness is considered in the context of the presence of outliers.

* Prof., Department of Statistics, University of Economics, Katowice.
** MA, Department of Statistics, University of Economics, Katowice.

## 2. SUPERPOPULATION MODEL

Following considerations are based on simple regression model assumed for the entire population. It is assumed, that values of auxiliary variable are known for all elements of the population. With regard to $\xi$ distribution describing superpopulation model it is assumed, that $Y_1, \ldots, Y_N$ are independent and $Y_i = \mu_i + \varepsilon_i$, $\mu_i = \mathrm{E}_\xi(Y_i) = \beta x_i$, $\mathrm{E}_\xi(\varepsilon_i) = 0$, $\sigma_i^2 = \mathrm{D}_\xi^2(Y_i) = \mathrm{D}_\xi^2(\varepsilon_i) = \sigma^2 v(x_i)$, where $\beta, \sigma^2$ are unknown and $x_1, \ldots, x_N$ are known for every $i$ $(i = 1, \ldots, N)$. In following consideration it will be additionally assumed, that $v(x_i) = x_i^2$.

## 3. CONSTRUCTION OF PREDICTOR

Considerations are conducted for any sample design. It is assumed, that the sample $s$ is drawn from the entire population by sample design $P(S)$ with first order inclusion probabilities $\pi_i$, where $i = 1, \ldots, N$. For any sample $s$ with size $n$ drawn from population $\varsigma$ with size $N$, $\Omega = S \cup \overline{S}$, where $\overline{S}$ denotes elements of the population, which were not drawn to the sample. Let $S_d = S \cap \Omega_d$, where the $d$-th domain is denoted by $\varsigma_d$. The size of $S_d$ equals $n_d$ (random variable) and the size of $\varsigma_d$ equals $N_d$. The set of elements of populations which belong to $d$-th domain $\varsigma_d$ could be written as $\Omega_d = S_d \cup \overline{S}_d$, where $\overline{S}_d$ denotes elements of the $d$-th domain, which were not drawn to the sample.

In following considerations notations presented below will be used. Gradients of all straight lines, each determined by the origin and one of n points $(x, y)$, where n is sample size, $y$ – the value of the variable of interest and $x$ – the value of auxiliary variable, are considered:

$$h_u = \frac{Y_i}{x_i} = \frac{\mu_i}{x_i} + \frac{Y_i - \mu_i}{x_i} = \beta + \frac{\varepsilon_i}{x_i} \tag{1}$$

where $i = 1, \ldots, n$.

Based on assumed superpopulation model it is known, that:

$$\begin{aligned} \mathrm{E}_\xi(h_i) &= \beta \\ \mathrm{D}_\xi^2(h_i) &= \frac{\sigma^2 x_i^2}{x_i^2} = \sigma^2 \end{aligned} \tag{2}$$

Let us discuss two predictors of the total value in the domain:

$$T_{1S_d} = n_d \overline{Y}_{S_d} + b_{1s} \sum_{i \in S_d} x_i \tag{3}$$

where:

$$b_{1S} = \frac{1}{n} \sum_{i \in S} h_i, \quad \overline{Y}_{s_d} = \frac{1}{n_d} \sum_{i \in S_d} Y_i, \quad T_{2S_d} = n_d \overline{Y}_{S_d} + b_{2S} \sum_{i \in S_d} x_i \tag{4}$$

where:

$$b_{2s} = \mathrm{Me}\{h\} \tag{5}$$

The estimator $b_{2s}$ is the median of a sequence of random variables $\{h_1, ..., h_s\}$. It is particular form of the estimator considered by H. T h e i l (1979). Let us note that the estimator $b_{2S}$ is robust against possible outliers. From the theorem presented by R. M. R o y a l l (1976) it is known, that $T_{1S_d}$ statistic is BLU predictor for assumed in section one superpopulation model. It means, that it is $\xi$ – unbiased predictor of the total value in small area $Y_d = \sum_{i \in S_d} Y_i$ and it minimises $\xi$ – variance for assumed superpopulation model.

Let us additionally assume that random variables $\varepsilon_i$, $i = 1, ..., N$ has continues distributions with different variances. Hence, from the equation (2) it is known, that $\{h_1, ..., h_S\}$ are sequence of independent random variables with the same distributions given by density function $f(.)$ with the same expected values and variances. Finally, from known results on distributions of sample quantiles (e.g. F i s z 1976) it results, that $b_{2S}$ is consistent estimator of $\beta$ and for large sample size $b_{2S}$ statistic is well approximated by normal distribution with following parameters $N\left(\beta, \frac{1}{4n \cdot f^2(\beta)}\right)$.

## 4. MEAN SQUARE ERROR (MSE) OF PREDICTION

First, the mean square error of $T_{1S_d}$ statistic (given by the equation (3)) will be analysed assuming, that superpopulation model $G_R$ is true. From Royall theorem (R o y a l l 1979) it results, that

$$E_\zeta E_p(T_{1S_d} - Y_d)^2 = \sigma^2 \frac{1}{n} E_p\left(\sum_{i\in S_d} x_i\right)^2 + \sigma^2 E_p\left(\sum_{i\in S_d} x_i^2\right) \tag{6}$$

Second, mean square error of $T_{2S_d}$ statistic (given by equation (4)) will be analysed assuming, that superpopulation model $G_R$ is true. Let us notice, that because of parameters of asymptotic distribution of $b_{2S}$, for large sample size $E_\zeta(b_{2S}) \approx \beta$. Hence it is easy to prove, that predictor $T_{2S_d}$ is approximately $\xi$ – unbiased predictor of the total value in small domain:

$$E_\zeta\left[\sum_{i\in S_d} Y_i + b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in\Omega_d} Y_i\right] = E_\zeta\left[\sum_{i\in S_d} Y_i + b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d} Y_i - \sum_{i\in\bar{S}_d} Y_i\right] =$$

$$= E_\zeta\left[b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d} Y_i\right] = E_\zeta(b_{2S})\sum_{i\in S_d} x_i - E_\zeta\left(\sum_{i\in S_d} Y_i\right) \approx \beta\sum_{i\in S_d} x_i - \beta\sum_{i\in S_d} x_i = 0 \tag{7}$$

The mean square error of predictor $T_{2S_d}$ for noninformative sample design is as follows:

$$E_\zeta E_p(T_{2S_d} - Y_d)^2 = E_p E_\zeta\left(\sum_{i\in S_d} Y_i + b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in\Omega_d} Y_i\right)^2 =$$

$$= E_p E_\zeta\left(b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d} Y_i\right)^2 = E_p E_\zeta\left(b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d} Y_i - \sum_{i\in S_d}\mu_i + \sum_{i\in S_d}\mu_i\right)^2 =$$

$$= E_p E_\zeta\left[\left(b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d}\mu_i\right)^2 + \left(\sum_{i\in S_d} Y_i - \sum_{i\in S_d}\mu_i\right)^2 - 2\left(b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d}\mu_i\right)\left(\sum_{i\in S_d} Y_i - \sum_{i\in S_d}\mu_i\right)\right] =$$

because $b_{2S}$ i $\sum_{i\in S_d} Y_i$ are independent random variables, we receive:

$$= E_\zeta\left(b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d}\mu_i\right)\left(\sum_{i\in S_d} Y_i - \sum_{i\in S_d}\mu_i\right) = E_{S_d}\left(b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d}\mu_i\right) E_{\zeta_{n-s}}\left(\sum_{i\in S_d} Y_i - \sum_{i\in S_d}\mu_i\right) = 0$$

and then:

$$E_\zeta E_p(T_{2S_d} - Y_d)^2 = E_p E_\zeta\left[\left(b_{2S}\sum_{i\in S_d} x_i - \sum_{i\in S_d}\mu_i\right)^2 + \left(\sum_{i\in S_d} Y_i - \sum_{i\in S_d}\mu_i\right)^2\right] =$$

$$= E_p E_\xi\left[\left(\sum_{i\in S_d} x_i\right)^2 (b_{2S} - \beta)^2 + \left(\sum_{i\in S_d} Y_i - \sum_{i\in S_d}\mu_i\right)^2\right] =$$

$$= \mathrm{E}_p\left[\left(\sum_{i\in S_d} x_i\right)^2 \mathrm{MSE}_\zeta(b_{2S}) + \sum_{i\in S_d} D_\zeta^2(Y_i)\right] =$$

$$= \mathrm{E}_p\left\{\left(\sum_{i\in S_d} x_i\right)^2 \mathrm{MSE}_\zeta(b_{2S}) + \sigma^2 \sum_{i\in S_d} x_i^2\right\},$$

where $\mathrm{MSE}_\zeta(b_{2S}) = D_\zeta^2(b_{2S}) + E_\zeta(b_{2S}) - \beta)^2 \approx D_\zeta^2(b_{2S})$

with $\mathrm{D}_\zeta^2(b_{2S}) \approx \dfrac{1}{4n \cdot f^2(\beta)}$.

In particular, if random variables $Y_i$ have normal distributions, then $\mathrm{D}_\zeta^2(b_{2S}) \approx \dfrac{\pi\sigma^2}{2n}$ and

$$\mathrm{E}_\zeta \mathrm{E}_p(T_{2S_d} - Y_d)^2 \approx \sigma^2 \frac{\pi}{2b} \mathrm{E}_p\left(\sum_{i\in S_d} x_i\right)^2 + \sigma^2 \mathrm{E}_p\left(\sum_{i\in S_d} x_i^2\right).$$

Let us compare mean square errors of both predictors assuming, that $Y_i$ have normal distributions. One should remember, that outliers can occur in the sample because of some disturbances in distribution or errors connected with data edition. This situation can imply significant bias of estimates in the case of usage of non-robust predictor. But there is also problem of a value of the difference of robust predictor and BLU predictor's mean square errors. It equals:

$$\mathrm{E}_\zeta \mathrm{E}_p(T_{2S_d} - Y_d)^2 - \mathrm{E}_\zeta \mathrm{E}_p(T_{1S_d} - Y_d)^2 \approx \frac{1}{n}\sigma^2\left(\frac{\pi}{2} - 1\right)\mathrm{E}_p\left(\sum_{i\in S_d} x_i\right)^2 > 0 \qquad (8)$$

This difference is positive. Therefore the BLU predictor is more accurate than the robust one. The difference is of order $O(n^{-1})$. It means, that it decreases due to the increase of sample size. Hence in large samples the accuracy of both predictors is similar but $T_{2S_d}$ is additionally robust. It should be noticed, that the difference (8) is the smaller, the smaller is $p$ – expected value of the total value of auxiliary variable for non-sampled small domain's elements. Proposed predictor should give good results for sample design proportional to the total value of auxiliary variable executed by i.e. Lahiri sampling scheme.

## 5. CONCLUSION

Summing up, we would like to state, that by analogy one can receive similar results assuming different superpopulation models for domains or strata, but one cannot forget, that asymptotic conditions must be met to use distribution parameters of discussed regression coefficient. Although possibility of usage of the predictor can be seemed as limited, its strong advantages must be underlined. It should be stressed, that robust predictors different from presented predictor one can find in the book of R. Valliant, *et al.* (2000). Their idea is based on the proposal of excluding information on outliers for estimation purposes. This approach requires subjective assessment, which of sampled elements are outliers. The construction of the predictor proposed in this paper does not require to take such a subjective decision.

## REFERENCES

Cassel C. M., Särndal C. E., Wretman J. H. (1977), *Foundations of Inference in Survey Sampling*, John Wiley & Sons, New York–London–Sydney–Toronto.

Fisz M. (1967), *Rachunek prawdopodobieństwa i statystyka matematyczna* [in Polish], PWN, Warszawa.

Royall R. M. (1976), *The Linear Least Squares Prediction Approach to Two-Stage Sampling*, "Journal of the American Statistical Association", 71, 657–664.

Theil H. (1979), *Zasady ekonometrii* [in Polish], PWN, Warszawa.

Valliant R., Dorfman A. H., Royall R. M. (2000), *Finite Population Sampling and Inference. A Prediction Approach*, John Wiley & Sons, New York–Chichester–Weinheim–Brisbane–Singapore–Toronto.

*Janusz Wywiał, Tomasz Żądło*

## O PEWNYM ODPORNYM NA WARTOŚCI ODDALONE PREDYKTORZE WARTOŚCI GLOBALNEJ W MAŁYM OBSZARZE

Rozważany jest problem predykcji wartości globalnej w domenie przy założeniu prostego modelu regresyjnego nadpopulacji (model regresyjny z jedną zmienną objaśniającą i bez stałej). Podjęty zostaje problem odpornej na wartości oddalone estymacji parametru funkcji regresji. Zaprezentowany estymator parametru funkcji regresji jest medianą wszystkich współczynników kierunkowych prostych przechodzących przez początek układu współrzędnych i jeden z $n$ punktów $(x, y)$, gdzie $n$ oznacza liczebność próby, $x$ – zmienną dodatkową, a $y$ – zmienną badaną. Estymator ten jest uproszczoną formą estymatora prezentowanego w: H. Theil (1979).

Autorzy przy asymptotycznych założeniach wyprowadzają wzór na błąd średniokwadratowy predykcji rozważanego predyktora odpornego. Przedstawiony zostaje także predyktor typu BLU dla zakładanego modelu nadpopulacji wraz z błędem średniokwadratowym predykcji. Dokładność obu predyktorów zostaje porównana przy założeniu normalności rozkładu badanych zmiennych losowych.