ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Tomasz Żądło**

ON SYNTHETIC RATIO ESTIMATOR BASED ON SUPERPOPULATION APPROACH

Abstract

In the paper properties of a predictor of the form of synthetic ratio estimator of domain total, known from randomisation approach, are considered. The proof of its ξ-unbiasedness for simple regression superpopulation model in strata is shown. For the model BLU predictor is also presented. Equations of prediction variances of both predictors are derived. For considered predictors the problem of model misspecification is considered and equations of prediction mean square errors are derived. The comparison of accuracy is supported by simulation study.

Key words: small area statistics, superpopulation approach, model misspecification, ξ-bias.

I. INTRODUCTION

Let population Ω of size N be divided into C strata denoted by Ω_c each of size N_c (where $c = 1, ..., C$) and D domains Ω_d each of size N_d (where $d = 1, ..., D$). One domain can be a part of more than one stratum. Sets $\Omega_c \cap \Omega_d$ will be denoted by Ω_{cd} and their sizes by N_{cd}. From each strata sample s_c of size n_c is drawn. Let sets $s_c \cap \Omega_d$ be denoted by s_{cd} and their sizes by n_{cd}. Let us introduce additional symbols: $\bigcup\limits_{c=1}^{C} s_c = s$, $\sum\limits_{c=1}^{C} n_c = n$, $\Omega_{rc} = \Omega_c - s_c$, $N_{rc} = N_c - n_c$, $\Omega_{rd} = \Omega_d - s_d$, $N_{rd} = N_d - n_d$, $\Omega_{rcd} = \Omega_{cd} - s_{cd}$, $N_{rcd} = N_{cd} - n_{cd}$. Let us stress that subscript d^* will denote domain of interest, which total value $T_{d^*} = \sum\limits_{i \in \Omega_{d^*}} Y_i$ is estimated.

[*] MSc, Department of Statistics, University of Economics in Katowice.

II. SIMPLE REGRESSION SUPERPOPULATION MODEL IN STRATA

Let us consider simple regression superpopulation model in strata with assumption:

$$Y_{ci} = \mu_{ci} + \varepsilon_{ci}, \tag{1}$$

where

$$\mu_{ci} = \mathrm{E}_{\xi}(Y_{ci}) = \beta_c x_{ci}, \quad \mathrm{E}_{\xi}(\varepsilon_{ci}) = 0$$

Let us add that β_c is unknown and $x_1, ..., x_N$ are known. What is more, for considered superpopulation model and for other superpopulation models assumed for strata, which will be discussed in following parts of the paper, it is assumed that random variables $Y_1, ..., Y_N$ are independent and:

$$\sigma_{ci}^2 = D_{\xi}^2(Y_{ci}) = \mathrm{D}_{\xi}^2(\varepsilon_{ci}) = \sigma_c^2 v(x_{ci}) \tag{2}$$

where $v(.)$ denotes values of known function of auxiliary variable.

Let us introduce predictor of domain total value of the form of ratio synthetic estimator known from randomization approach. For considered stratified random sampling it is as follows (e.g. Bracha, 1994; Bracha, 1996; Getka-Wilczyńska, 2000; Wywiał, Żądło, 2003):

$$\hat{T}_{d*}^{SYN-rat} = \sum_{c=1}^{C} \frac{X_{cd_*}}{\hat{X}_c} \hat{Y}_c, \tag{3}$$

where

$$\hat{Y}_c = \sum_{i \in s_c} \frac{Y_i}{\pi_i}, \; \hat{X}_c = \sum_{i \in s_c} \frac{x_i}{\pi_i}, \; X_{cd*} = \sum_{i \in \Omega_{cd*}} x_i.$$

Let us notice that for assumed superpopulation model:

$$\mathrm{E}_{\xi}(\hat{T}_{d*}^{SYN-ilor} - T_{d*}) = \sum_{c=1}^{C} \frac{X_{cd*}}{\hat{X}_c} \mathrm{E}_{\xi}(\hat{Y}_c) - \sum_{c=1}^{C} \sum_{i \in \Omega_{cd*}} \mathrm{E}_{\xi}(Y_i) =$$

$$= \sum_{c=1}^{C} \left(\frac{X_{cd*}}{\sum_{i \in s_c} \frac{x_i}{\pi_i}} \sum_{i \in s_c} \frac{\beta_c x_i}{\pi_i} - \beta_c X_{cd*} \right) = 0$$

It was proved that predictor of the form of synthetic ratio estimator is ξ-unbiased for simple regression superpopulation model assumed for strata.

What should be stressed is that predictor of the form of synthetic ratio estimator (3) does not have minimal prediction variance among all linear ξ-unbiased predictors for simple regression superpopulation model assumed for strata. From Royall's theorem (1976) it is known that BLU predictor for the considered superpopulation model with assumptions (1) and (2) is as follows:

$$\hat{T}_{d*}^{BLU-rat} = \sum_{c=1}^{C} (Y_{scd*} + \hat{\beta}_c X_{rcd*}) \tag{4}$$

where

$$\hat{\beta}_c = \frac{\sum_{i \in s_c} \frac{x_i Y_i}{v(x_i)}}{\sum_{i \in s_c} \frac{x_i^2}{v(x_i)}}, \quad X_{rcd*} = \sum_{i \in \Omega_{rcd*}} x_i, \quad Y_{scd*} = \sum_{i \in s_{cd*}} Y_i$$

Let inclusion probabilities in strata be constant (e.g. simple random sample without replacement is drawn from strata) and $\forall_i v(x_i) = x_i$. Hence:

$$\hat{T}_{d*}^{SYN-rat} = \sum_{c=1}^{C} \frac{X_{cd*}}{X_{sc}} Y_{sc}, \tag{5}$$

where

$$Y_{sc} = \sum_{i \in s_c} Y_i, \quad X_{sc} = \sum_{i \in s_c} x_i \quad \text{and}$$

$$\hat{T}_{d*}^{BLU-rat} = \sum_{c=1}^{C} \left(Y_{scd*} + \frac{Y_{sc}}{X_{sc}} X_{rcd*} \right) \tag{6}$$

It easy to notice that if above-mentioned assumptions and the following conditions are fulfilled:

- none of elements of domain $d*$ are drawn to the sample,
- for each strata from which elements of $d*$-th domain were drawn following equation holds $\frac{Y_{sc}}{X_{sc}} = \frac{Y_{scd*}}{X_{scd*}}$,
- for each strata from which elements of $d*$-th domain were drawn following equation holds $s_c = s_{cd*}$,

then

$$\hat{T}_{d*}^{BLU-rat} = \hat{T}_{d*}^{SYN-rat} = \sum_{c=1}^{C} \frac{X_{cd*}}{X_{sc}} Y_{sc}. \tag{7}$$

Let us derive equations of prediction variances of predictors (3) and (4) assuming that condition (2) is fulfilled. It should be stressed that they are correct even when condition (1), which defines simple regression superpopulation model, is not fulfilled.

After some algebra prediction variance of the predictor of the form of synthetic ratio estimator is as follows:

$$\mathrm{Var}_{\xi}(\hat{T}_{d*}^{SYN-rat} - T_{d*})^2 = \sum_{c=1}^{C} \sigma_c^2 \left[\left(\frac{X_{cd*}}{\hat{X}_c} \right)^2 \sum_{i \in s_c} \frac{v(x_i)}{\pi_i^2} - 2 \frac{X_{cd*}}{\hat{X}_c} \sum_{i \in s_{cd*}} \frac{v(x_i)}{\pi_i} + \sum_{i \in \Omega_{cd*}} v(x_i) \right]. \tag{8}$$

If first order inclusion probabilities are constant is strata and if $\forall_i v(x_i) = x_i$, then prediction variance of the predictor of the form of synthetic ratio estimator will be given by following equation:

$$\mathrm{Var}_{\xi}(\hat{T}_{d*}^{SYN-rat} - T_{d*})^2 = \sum_{c=1}^{C} \sigma_c^2 \left[\frac{X_{cd*}^2}{X_{sc}} - 2 \frac{X_{cd*}}{X_{sc}} X_{scd*} + X_{cd*} \right], \tag{9}$$

where

$$X_{scd*} = \sum_{i \in s_{cd*}} x_i$$

Prediction variance of predictor (4) for superpopulation model with assumption (2) can be derived using Royall's theorem (1976). Let us stress that it is correct even when condition (1), which defines simple regression superpopulation model, is not fulfilled. Prediction variance of predictor (4) is as follows:

$$\mathrm{Var}_{\xi}(\hat{T}_{d*}^{BLU-rat} - T_{d*})^2 = \sum_{c=1}^{C} \sigma_c^2 \left[X_{rcd*}^2 \left(\frac{x_i^2}{v(x_i)} \right)^{-1} - \sum_{i \in \Omega_{rcd*}} v(x_i) \right]. \tag{10}$$

If $\forall_i v(x_i) = x_i$, then prediction variance will simplify to the following form:

$$\mathrm{Var}_{\xi}(\hat{T}_{d*}^{BLU-rat} - T_{d*})^2 = \sum_{c=1}^{C} \sigma_c^2 \left[\frac{X_{rcd*}^2}{X_{sc}} + X_{rcd*} \right]. \tag{11}$$

Let us compare prediction variances of both predictors when $\forall_i v(x_i) = x_i$ and for constant first order inclusion probabilities in strata.

$$\mathrm{Var}_\zeta(\hat{T}_{d*}^{BLU-rat} - T_{d*})^2 - \mathrm{Var}_\zeta(\hat{T}_{d*}^{SYN-rat} - T_{d*})^2 = -\sum_{c=1}^{C} \sigma_c^2 \frac{X_{scd*}}{X_{sc}}[X_{sc} - X_{scd*}] \tag{12}$$

Let us notice, that the value of X_{scd*} is closer to zero (what holds when n_{cd*} decreases), the smallest precision difference of both predictors is. In discussed case, the maximum value of equation (12) is received for $\frac{X_{scd*}}{X_{sc}} = 0.5$. The difference (12) equals 0 for $\frac{X_{scd*}}{X_{sc}} = 0$ and for $\frac{X_{scd*}}{X_{sc}} = 1$. For small area statistics purposes considerations can be limited to $0 < \frac{X_{scd*}}{X_{sc}} < 0.5$. In this case, the lower value of $\frac{X_{scd*}}{X_{sc}}$ is, the lower value of precision difference (12) is observed. Prediction variances of the considered predictors are equal when equation (7) holds.

III. SIMPLE REGRESSION SUPERPOPULATION MODEL IN DOMAINS

Synthetic estimators use assumption that some relationships which occur in population (or in strata) hold in domains (or domains and strata products) too. In the previous part of the paper two ξ-unbiased predictors for simple regression superpopulation model in strata were presented. Let us add that predictor (4) have minimal prediction variance among all ξ-unbiased predictors (hence its more precise than predictor (3)). Assumption that simple regression superpopulation model in strata is true can be incorrect. For example simple regression superpopulation model in domains can be true. In the following part of the paper accuracy of the predictors (3) and (4) for simple regression superpopulation model in domains will be considered. It will be proved that both predictors are ξ-biased and equations of their ξ-biases and prediction MSEs will be derived.

Let us assume that simple regression superpopulation model in domains is true. The assumption is as follows:

$$\mathrm{E}_\xi(Y_{di}) = \beta_d x_{di} \tag{13}$$

Let us consider two additional alternative assumptions. It is assumed that random variables $Y_1, \ldots, Y_N$ are independent and:

$$\sigma_{ci}^2 = \mathrm{D}_\xi^2(Y_{ci}) = \mathrm{D}_\xi^2(\varepsilon_{ci}) = \sigma_c^2\, v(x_{ci}) \tag{14}$$

as in equation (2) or

$$\sigma_{di}^2 = \mathrm{D}_\xi^2(Y_{di}) = \mathrm{D}_\xi^2(\varepsilon_{di}) = \sigma_d^2\, v(x_{di}). \tag{15}$$

In previous paragraph it was stressed that if assumption given by equation (2) (the same is presented by equation (14)) is true, then $\mathrm{Var}_\xi(\hat{T}_{d*}^{BLU-rat} - T_d) < \mathrm{Var}_\xi(\hat{T}_{d*}^{SYN-rat} - T_d)$. Let us consider prediction variances of both predictors when equation (15) is true.

Prediction variance of the predictor of the form of synthetic ratio estimator for assumption (15) after some algebra is received as follows:

$$\mathrm{Var}_\xi(\hat{T}_{d*}^{SYN-rat} - T_d) = \sum_{c=1}^{C}\left[\left(\frac{X_{cd*}}{\hat{X}_c}\right)^2 \sum_{d=1}^{D}\sigma_d^2 \sum_{i\in s_{cd}} \frac{v(x_i)}{\pi_i^2} + \sigma_{d*}^2 \sum_{i\in\Omega_{cd*}} v(x_i) - \right.$$
$$\left. + 2\frac{X_{cd*}}{\hat{X}_c}\sigma_{d*}^2 \sum_{i\in s_{cd*}} \frac{v(x_i)}{\pi_i}\right]. \tag{16}$$

If $\forall_d v(x_i) = x_i$ and first order inclusion probabilities will be constant in strata, then above equation simplifies to the following form:

$$\mathrm{Var}_\xi(\hat{T}_{d*}^{SYN-rat} - T_d)^2 = \sum_{c=1}^{C}\left(\frac{X_{cd*}^2}{X_{sc}} \sum_{d=1}^{D}\sigma_d^2 X_{scd} + \sigma_{d*}^2 X_{cd*} - 2\frac{X_{cd*}}{X_{sc}}\sigma_{d*}^2 X_{scd*}\right). \tag{17}$$

Let us derive prediction variance of predictor (4) for assumption (15). The following result can be received:

$$\mathrm{Var}_\xi(\hat{T}_{d*}^{BLU-rat} - T_d) = \sum_{c=1}^{C}\left(X_{rcd*}^2\left(\sum_{i\in s_c}\frac{x_i^2}{v(x_i)}\right)^{-2}\left(\sum_{d=1}^{D}\sigma_d^2\sum_{i\in s_{cd}}\frac{x_i^2}{v(x_i)}\right) + \sigma_{d*}^2\sum_{i\in\Omega_{rcd*}} v(x_i)\right). \tag{18}$$

If $\forall_d v(x_i) = x_i$, then above equation simplifies to the following form:

$$\mathrm{Var}_\xi(\hat{T}_{d*}^{BLU-rat} - T_d) = \sum_{c=1}^{C}\left(X_{rcd*}^2 X_{sc}^{-2}\sum_{d=1}^{D}\sigma_d^2 X_{scd} + \sigma_{d*}^2 X_{rcd*}\right). \tag{19}$$

If $\forall_d v(x_i) = x_i$ and first order inclusion probabilities are constant in strata, then for assumption (15):

$$\mathrm{Var}_\xi(\hat{T}_{d*}^{BLU-rat} - T_d) - \mathrm{Var}_\xi(\hat{T}_{d*}^{SYN-rat} - T_d) =$$

$$= \sum_{c=1}^{C}\left[\sigma_{d*}^2 \frac{X_{scd*}}{X_{sc}^2}(X_{sc} - X_{scd*})(X_{cd*} + X_{rcd*} - X_{sc}) - \sum_{d \neq d*=1}^{D} \sigma_d^2 \frac{X_{scd}}{X_{sc}^2} X_{scd*}(X_{cd*} + X_{rcd*})\right]. \tag{20}$$

Let us notice, that the value of X_{scd*} is closer to zero (what holds when n_{cd*} decreases), the smallest precision difference of both predictors is. Above equation is sum for strata of sums of two elements. Let us assume that $\forall_d x_i > 0$.

For each strata second element is negative. The first element is negative for every strata if and only if $X_{cd*} + X_{rcd*} \leqslant X_{sc}$. Hence,

$$\forall_c X_{cd*} + X_{rcd*} \leqslant X_{sc} \Rightarrow \mathrm{Var}_\xi(\hat{T}_{d*}^{BLU-rat} - T_d) < \mathrm{Var}_\xi(\hat{T}_{d*}^{SYN-rat} - T_d).$$

Based on equation (20) it can also be proved that

$$\forall_c \sigma_{d*}^2 \leqslant \sum_{d=1}^{D} \sigma_d^2 \frac{X_{scd}}{X_{sc}} \Rightarrow \mathrm{Var}_\xi(\hat{T}_{d*}^{BLU-rat} - T_d) < \mathrm{Var}_\xi(\hat{T}_{d*}^{SYN-rat} - T_d).$$

It was shown that predictor (4) can be more precise than predictor (3) for assumption (15).

Let us derive equation of ξ-bias of the predictor of the form of synthetic ratio estimator (3) for the superpopulation model with assumption (13). After some algebra it is obtained that:

$$\mathrm{E}_\xi(\hat{T}_{d*}^{SYN-rat} - T_{d*}) = \sum_{c=1}^{C} \frac{X_{cd*}}{\hat{X}_c} \sum_{d=1}^{D} (\beta_d - \beta_{d*})\hat{X}_{cd} \tag{21}$$

where

$$\hat{X}_{cd} = \sum_{i \in s_{cd}} \frac{x_i}{\pi_i}.$$

What was expected, the predictor of the form of synthetic ratio estimator is ξ-unbiased, when simple regression superpopulation model is true in strata to which domain of interest belongs (superpopulation model with assumption (1)).

Let us derive equation of ξ-bias of the predictor (4) for superpopulation model assumed in this part of the paper.

$$E_{\xi}(\hat{T}_{d*}^{BLU-rat}-T_{d*})=\sum_{c=1}^{C}X_{rcd*}\left(\sum_{i\in s_c}\frac{x_i^2}{v(x_i)}\right)^{-1}\sum_{d=1}^{D}(\beta_d-\beta_{d*})\sum_{i\in s_{cd}}\frac{x_i^2}{v(x_i)}. \qquad (22)$$

Similarly to the predictor of the form of synthetic ratio estimator, the predictor (4) is ξ-unbiased if simple regression superpopulation model in domains becomes simple regression superpopulation model in strata (simple regression superpopulation model in strata with assumption (2) is true).

Let us assume that $\forall_i v(x_i)=x_i$ and that first order inclusion probabilities are constant is strata. Then, equations (21) and (22) of ξ-bias of predictors (3) and (4) simplify to the following forms:

$$E_{\xi}(\hat{T}_{d*}^{SYN-rat}-T_{d*})=\sum_{c=1}^{C}\frac{X_{cd*}}{X_{sc}}\sum_{d=1}^{D}(\beta_d-\beta_{d*})X_{scd} \qquad (23)$$

$$E_{\xi}(\hat{T}_{d*}^{BLU-rat}-T_{d*})=\sum_{c=1}^{C}\frac{X_{rcd*}}{X_{sc}}\sum_{d=1}^{D}(\beta_d-\beta_{d*})X_{scd}. \qquad (24)$$

Hence,

$$E_{\xi}(\hat{T}_{d*}^{BLU-rat}-T_{d*})-E_{\xi}(\hat{T}_{d*}^{SYN-rat}-T_{d*})=-\sum_{c=1}^{C}\frac{X_{scd*}}{X_{sc}}\sum_{d=1}^{D}(\beta_d-\beta_{d*})X_{scd}. \qquad (25)$$

First, let us remind that if both predictors are ξ-unbiased (i.e. simple regression superpopulation model in strata is true) or if equality (7) holds, then difference given by equation (25) will equal zero. Let us notice, that the value of X_{scd*} is closer to zero (what holds when n_{cd*} decreases), the smallest difference of ξ-biases of both predictors is.

Let us consider two cases with additional assumptions that $\forall_i x_i>0$ and $\hat{T}_{d*}^{BLU-rat}\neq\hat{T}_{d*}^{SYN-ilor}$. Let in the first case for each strata to which elements of $d*$ domain belong following inequality occurs $\frac{1}{X_{sc}}\sum_{d=1}^{D}(\beta_d-\beta_{d*})X_{scd}>0$, what can hold when $\forall_{\substack{d\\ d\neq d*}}\beta_d>\beta_{d*}$. Hence, $E_{\xi}(\hat{T}_{d*}^{SYN-rat}-T_{d*})>0$ and $E_{\xi}(\hat{T}_{d*}^{BLU-rat}-T_{d*})>0$ and finally $E_{\xi}(\hat{T}_{d*}^{BLU-rat}-T_{d*})-E_{\xi}(\hat{T}_{d*}^{SYN-rat}-T_{d*})<0$. Let in the second case for each strata to which elements of $d*$ domain belong following inequality occurs $\frac{1}{X_{sc}}\sum_{d=1}^{D}\beta_d-\beta_{d*})X_{scd}<0$, what can hold when $\forall_{\substack{d \\ d \neq d^*}} \beta_d < \beta_{d^*}$. Hence, $\mathrm{E}_\xi(\hat{T}_{d^*}^{SYN-rat} < 0$ and $\mathrm{E}_\xi(\hat{T}_{d^*}^{BLU-rat} - T_{d^*}) < 0$ and finally $\mathrm{E}_\xi(\hat{T}_{d^*}^{BLU-rat} - T_{d^*}) - E_\xi(\hat{T}_{d^*}^{SYN-rat} - T_{d^*}) > 0$. In both cases absolute value of ξ-bias of $\hat{T}_{d^*}^{BLU-rat}$ predictor is lower then absolute value of $\hat{T}_{d^*}^{SYN-rat}$. Let us stress that when elements of d^* domain were drawn to the sample only from one strata, only one of these two situations can hold.

Prediction MSE of the predictor of the form of synthetic ratio estimator for simple regression superpopulation model in domains is obtained by summation of prediction variance (8) for assumption (14) or prediction variance (16) for assumption (15) and squared ξ-bias (21). Prediction MSE of predictor (4) for simple regression superpopulation model in domains is received by summation of prediction variance (10) for assumption (14) or prediction variance (18) for assumption (15) and squared ξ-bias (22).

Because analytical results of MSE comparison are quite modest, in part V simulation study will additionally be conducted.

IV. POLYNOMIAL SUPERPOPULATION MODEL IN STRATA

In the previous section the misspecification of superpopulation model was considered in the case when simple regression superpopulation model in domains is true. In the following section polynomial superpopulation model in strata is assumed.

It is assumed that

$$\mathrm{E}_\xi(Y_{ci}) = \sum_{j=0}^{J} \beta_c^{(j)} x_{ci}^j. \tag{26}$$

Particular form of polynomial superpopulation model with assumption (26) is regression superpopulation model with following assumption:

$$\mathrm{E}_\xi(Y_{ci}) = \beta_{ci}^{(1)} x_{ci} + \beta_c^{(0)}. \tag{27}$$

What should be reminded is that for models assumed for strata equation (2) holds. It implies that, prediction variances of both predictors are given by equations (8) and (10) and

$$\mathrm{Var}_\xi(\hat{T}_{d^*}^{BLU-rat} - T_{d^*}) < Var_\xi(\hat{T}_{d^*}^{SYN-rat} - T_{d^*}).$$

Let us derive equation of ξ-bias of the predictor of the form of synthetic ratio estimator for polynomial superpopulation model in strata (superpopulation model with assumption (26)). After some algebra it is obtained, that

$$E_{\xi}(\hat{T}_{r*}^{SYN-rat}-T_{d*})=\sum_{c=1}^{C}\frac{\sum_{i\in s_c}\sum_{j=0}^{J}\beta_c^{(j)}\frac{x_i^j}{\pi_i}}{\hat{X}_c}\sum_{i\in\Omega_{cd*}}\sum_{j=0}^{J}\beta_c^{(j)}x_i^j\left(\frac{\hat{X}_{cd*}}{\sum_{i\in\Omega_{cd*}}\sum_{j=0}^{J}\beta_c^{(j)}x_i^j}-\frac{\hat{X}_c}{\sum_{i\in s_c}\sum_{j=0}^{J}\beta_c^{(j)}\frac{x_i^j}{\pi_i}}\right).\tag{28}$$

If regression superpopulation model is assumed for strata (superpopulation model with assumption (27)) and if first order inclusion probabilities are constant in strata, the equation will simplify to the following form:

$$E_{\xi}(\hat{T}_{d*}^{SYN-rat}-T_{d*})=\sum_{c=1}^{C}\beta_c^{(0)}\frac{n_c}{X_{sc}}N_{cd*}\left(\frac{X_{cd*}}{N_{cd*}}-\frac{X_{sc}}{n_c}\right).\tag{29}$$

In the considered case if for each strata the mean value of auxiliary variable for domain d^* and stratum products equals the mean value of auxiliary variable for sampled elements from stratum, the predictor of the form of synthetic ratio estimator will be ξ-unbiased.

Let us derive equation of ξ-bias of predictor (4) for polynomial superpopulation model in strata (superpopulation model with assumption (26)). The result is as follows:

$$E_{\xi}(\hat{T}_{d*}^{BLU-rat}-T_{d*})=$$

$$=\sum_{c=1}^{C}\left\{\left(\sum_{i\in\Omega_{rcd*}}\sum_{j=0}^{J}\beta_c^{(j)}x_i^j\right)\frac{\sum_{i\in s_c}\sum_{j=0}^{J}\beta_c^{(j)}\frac{x_i^{j+1}}{v(x_i)}}{\sum_{i\in s_c}\frac{x_j^2}{v(x_i)}}\left[\frac{X_{rcd*}}{\sum_{i\in\Omega_{cd*}}\sum_{j=0}^{J}\beta_c^{(j)}x_i^j}-\frac{\sum_{i\in s_c}\frac{x_i^2}{v(x_i)}}{\sum_{i\in s_c}\sum_{j=0}^{J}\beta_c^{(j)}\frac{x_i^{j+1}}{v(x_i)}}\right]\right\}.\tag{30}$$

If regression superpopulation model in strata is true (superpopulation model with assumption (27)) and if $\forall_i v(x_i)=x_i$, then the equation will simplify to the following form:

$$E_{\xi}(\hat{T}_{d*}^{BLU-rat}-T_{d*})=\sum_{c=1}^{C}\beta_c^{(0)}\left[N_{rcd*}\frac{n_c}{X_{sc}}\left(\frac{X_{rcd*}}{N_{rcd*}}-\frac{X_{sc}}{n_c}\right)\right].\tag{31}$$

In the considered case if for each strata the auxiliary variable mean value for non-sampled elements of intersection of domain d^* and stratum equals the mean value of auxiliary variable for sampled elements from stratum, predictor (4) will be ξ-unbiased.

Let us compare ξ-biases of both predictors for regression superpopulation model in strata when $\forall_i v(x_i) = x_i$ and first order inclusion probabilities are constant in strata. Let us assume that equality (7) does not occure. Hence,

$$\mathrm{E}_\xi(\hat{T}^{BLU-rat}_{d*} - T_{d*}) - \mathrm{E}_\xi(\hat{T}^{SYN-rat}_{d*} - T_{d*}) = -\sum_{c=1}^{C} \beta_c^{(0)} \frac{n_c}{X_{sc}} n_{cd*} \left[\frac{X_{scd*}}{n_{cd*}} - \frac{X_{sc}}{n_c} \right]. \tag{32}$$

Let us notice, that the value of n_{cd*} is closer to zero, the smallest difference of ξ-biases of both predictors is. If for each strata the auxiliary variable mean value for sampled elements of intersection of domain d^* and stratum equals the mean value of auxiliary variable for sampled elements from stratum, values of ξ-bias for both predictors will be equal.

Let us consider two cases assuming that $\forall_i x_i > 0$ and $\forall_c \beta_c^{(0)} > 0$. Let in the first case for each strata from which elements of d^*-th domain were drawn following inequalities appear $\frac{X_{cd*}}{N_{cd*}} > \frac{X_{sc}}{n_c}$, $\frac{X_{rcd*}}{N_{rcd*}} > \frac{X_{sc}}{n_c}$, $\frac{X_{scd*}}{n_{cd*}} > \frac{X_{sc}}{n_c}$. It can hold for example when domain of interest consists of elements with the highest values of auxiliary variable. Hence, $\mathrm{E}_\xi(\hat{T}^{SYN-rat}_{d*} - T_{d*}) > 0$ and $\mathrm{E}_\xi(\hat{T}^{BLU-rat}_{d*} - T_{d*}) > 0$ and finally $\mathrm{E}_\xi(\hat{T}^{BLU-rat}_{d*} - T_{d*}) - \mathrm{E}_\xi(\hat{T}^{SYN-rat}_{d*} - T_{d*}) < 0$. Let in the second case for each strata from which elements of d^*-th domain were drawn following inequalities appear $\frac{X_{cd*}}{N_{cd*}} < \frac{X_{sc}}{n_c}$, $\frac{X_{rcd*}}{N_{rcd*}} < \frac{X_{sc}}{n_c}$, $\frac{X_{scd*}}{n_{cd*}} < \frac{X_{sc}}{n_c}$. It can hold for example when domain of interest consists of elements with the lowest values of auxiliary variable. Hence, $\mathrm{E}_\xi(\hat{T}^{SYN-rat}_{d*} - T_{d*}) < 0$ and $\mathrm{E}_\xi(\hat{T}^{BLU-rat}_{d*} - T_{d*}) < 0$ and finally $\mathrm{E}_\xi(\hat{T}^{BLU-rat}_{d*} - T_{d*}) - \mathrm{E}_\xi(\hat{T}^{SYN-rat}_{d*} - T_{d*}) > 0$. In both cases absolute value of ξ-bias of $\hat{T}^{BLU-rat}_{d*}$ predictor is lower than absolute value of ξ-bias of $\hat{T}^{SYN-rat}_{d*}$, what implies lower value of prediction MSE of $\hat{T}^{BLU-rat}_{d*}$ predictor (because value of prediction variance of $\hat{T}^{BLU-rat}_{d*}$ is lower). Let us add that the same conclusions can be received for both cases for assumptions $\forall_i x_i > 0$ and $\forall_c \beta_c^{(0)} < 0$.

Prediction MSE's of predictors (3) and (4) for simple regression superpopulation model in domains are received by summation of prediction variances (8) and (10) and squared ξ-biases given by equations (28) and (30) adequately.

V. SIMULATION STUDY

Simulation study is conducted based on artificial population which consists of 200 elements divided into 3 strata and 6 domains. First stratum, which consists of 80 elements, includes 20 elements from first domain, 20 elements from second domain and 40 elements from third domain. Second stratum, which consists of 70 elements, includes 30 elements from first domain, 30 elements from fourth domain and 10 elements from fifth domain. Third stratum, which consists of 50 elements, includes 20 elements from second domain, 10 elements from fifth domain and 20 elements from sixth domain. Values of auxiliary variable were generated using normal distributions with following parameters set arbitrarily: in first stratum distribution $N(100, 20)$, in second stratum – $N(120, 30)$ and in third stratum – $N(150, 40)$. Elements in strata are assigned to domains at random.

Three predictors are considered: predictor given by equation (3) (in tables denoted by synt), predictor given by equation (4) with $v(x_i) = \sqrt{x_i}$ for every $i = 1, ..., N$ (in tables denoted by BLU 1) and predictor given by equation (4) with $v(x_i) = 1$ for every $i = 1, ..., N$ (in tables denoted by BLU 2). Accuracy of the three predictors is considered for four superpopulation models with following parameters set arbitrarily. Let us add, that for all following superpopulation models random components ε_i are generated by using $N(0, 1)$ distribution. First model is simple regression superpopulation model in strata as follows: $Y_{ci} = \beta_c x_{ci} + \varepsilon_{ci}\sqrt{x_{ci}}$, where $\beta_1 = 1$, $\beta_2 = 2$, $\beta_3 = 3$. Second model is regression superpopulation model in strata as follows: $Y_{ci} = \beta_c^{(1)} x_{ci} + \beta_c^{(0)} + \varepsilon_{ci}\sqrt{x_{ci}}$, where $\beta_1^{(1)} = 1$, $\beta_2^{(1)} = 2$, $\beta_3^{(1)} = 3$, $\beta_1^{(0)} = 200$, $\beta_2^{(0)} = 250$, $\beta_3^{(0)} = 300$. Third model is polynomial superpopulation model in strata as follows: $Y_{ci} = \sum_{j=0}^{2} \beta_c^{(j)} x_{ci}^{j} + \varepsilon_{ci}\sqrt{x_{ci}}$, where $\beta_1^{(2)} = 1.5$, $\beta_2^{(2)} = 1$, $\beta_3^{(2)} = 0.5$, $\beta_1^{(1)} = 1$, $\beta_2^{(1)} = 2$, $\beta_3^{(1)} = 3$, $\beta_1^{(0)} = 200$, $\beta_2^{(0)} = 250$, $\beta_3^{(0)} = 300$. Fourth model is simple regression superpopulation model in domains as follows: $Y_{di} = \beta_d x_{di} + \varepsilon_{di}\sqrt{x_{di}}$, where $\beta_1 = 1$, $\beta_2 = 3$, $\beta_3 = 5$, $\beta_4 = 7$, $\beta_5 = 9$, $\beta_6 = 11$.

It should be underlined, that although model approach is conditional approach, results in simulation study are averaged by taking sampling design distribution into consideration. Symbol E_p denotes expected value of sampling design distribution. In the following tables bias (in %) denotes approximated in simulation study value of $\frac{\mathrm{E}_p \mathrm{E}_\xi(\hat{T}_{d*} - T_{d*})}{\mathrm{E}_\xi(T_{d*})} \times 100$, root variance (in %) approximated in simulation study value of

$\frac{\sqrt{\mathrm{E}_p\mathrm{E}_\xi((\hat{T}_{d*} - T_{d*}) - \mathrm{E}_\xi(\hat{T}_{d*} - T_{d*}))^2}}{\mathrm{E}_\xi(T_{d*})} \times 100$ and root MSE (in %) denotes approximated in simulation study value of $\frac{\sqrt{\mathrm{E}_p\mathrm{E}_\xi(\hat{T}_{d*} - T_{d*})^2}}{\mathrm{E}_\xi(T_{d*})} \times 100$. It is worth stressing that ξ p-bias, p-expected value of prediction variance and p-expected value of prediction MSE are computed instead of p ξ-bias, ξ-expected value of p-variance and ξ-expected value of p-MSE. Values of above-mentioned statistics are equal because sampling design is noninformative.

Stratified random sampling with proportional allocation is considered. Results received in simulation are based on 500 random samples and are additionally averaged with respect to 1000 realizations of superpopulation model. This way for simulation purposes 500 000 values of each predictor are generated. Three sizes of sample are considered: 40, 60 and 80 elements which amount to 20%, 30% and 40% of population size. High fractions of drawn elements are considered because it was proved, for cases discussed in previous parts of the paper, that for small sample sizes precision difference of both predictors is small.

Let us compare accuracy of analysed predictors when simple regression superpopulation model in strata is true.

Results presented in the Table 1 show that root ξ-expected values of p-MSEs for all of predictors in all domains except of domain three equal less than 1% of ξ-expected domain total. In domain three they does not exceed 3%. It is worth stressing that although accuracies of the considered predictors are similar, root ξ-expected value of p-MSE of the predictor of the form of synthetic ratio estimator is higher comparing to predictor (4) with misspecification of variance structure (in table denoted by BLU 2). If statistician specifies correct form of ξ-expected value of random variables (i.e. he decides that simple regression superpopulation model in strata is true) and incorrect form of their ξ-variance (i.e. he decides that model is homoscedastic), the choice of BLU predictor with wrong specification of variance structure will be better than choice of the predictor of the form of synthetic ratio estimator. Interesting is that in simulation study the decrease of root ξ-expected p-MSEs for synthetic estimator due to the increase of sample size is slower comparing with other predictors. Let us add, that the highest values of root ξ-expected p-MSEs are observed in domain three, because it is the only domain which belongs only to first strata – strata with the lowest β_c coefficient. Because distributions of auxiliary variable in strata are similar, in the first strata the higher dispersion of variable of interest with respect to ξ distribution is observed. Notice that the smaller is sample size the smaller is difference in accuracy of synthetic estimator and BLU predictor (denoted by BLU 1) what was proofed for different assumptions in part 2 of the paper.

Table 1. Accuracy of predictors for simple regression superpopulation model in strata

Domain	Predictor	Bias (in %)			Root variance and root MSE (in %)		
		Sample size			Sample size		
		40	60	80	40	60	80
1	synt	0.00	0.00	0.00	0.86	0.72	0.65
	BLU 1	0.00	0.00	0.00	0.84	0.68	0.57
	BLU 2	0.00	0.00	0.00	0.85	0.68	0.58
2	synt	0.00	0.00	0.00	0.63	0.54	0.48
	BLU 1	0.00	0.00	0.00	0.61	0.50	0.43
	BLU 2	0.00	0.00	0.00	0.61	0.50	0.43
3	synt	0.00	0.00	–0.01	2.52	2.04	1.77
	BLU 1	0.00	0.00	0.00	2.48	1.95	1.63
	BLU 2	0.00	0.00	0.00	2.50	1.97	1.64
4	synt	0.00	0.00	0.00	0.85	0.70	0.62
	BLU 1	0.00	0.00	0.00	0.83	0.67	0.56
	BLU 2	0.00	0.00	0.00	0.84	0.68	0.57
5	synt	0.00	0.00	0.00	0.58	0.52	0.49
	BLU 1	0.00	0.00	0.00	0.56	0.47	0.41
	BLU 2	0.00	0.00	0.00	0.56	0.47	0.41
6	synt	0.00	0.00	0.00	0.55	0.46	0.40
	BLU 1	0.00	0.00	0.00	0.54	0.44	0.37
	BLU 2	0.00	0.00	0.00	0.55	0.44	0.37

Let us consider results for regression superpopulation model in strata which are presented in the Table 2. Accuracy of the considered predictors will be discussed in the case of model misspecification. Let us notice that values of root ξ-expected p-MSEs do not exceed 3,5% of ξ-expected domain totals and they are determined by values of ξ-p-bias. It should be underlined that in this case none of predictors have better accuracy in comparison with others. For polynomial model in strata (result are not presented) values of root ξ-expected p-MSEs exceed 6% of ξ-expected domain totals only in few cases for sample size 40 elements. These results are determined by ξ p-bias, values of root ξ-expected p-variances do not exceed 0.04% of ξ-expected domain totals. It should be stressed that in some cases ξ-expected p-MSEs of synthetic ratio estimator increase due to the increase of sample size, what for p-MSEs was discussed earlier by Wywiał, Żądło (2003). The same property can be observed for ξ-expected p-MSEs, because sampling design is noninformative.

Table 2. Accuracy of predictors for regression superpopulation model in strata

Domain	Predictor	Bias (in %)			Root variance (in %)			Root MSE (in %)		
		Sample size			Sample size			Sample size		
		40	60	80	40	60	80	40	60	80
1	synt	−1.75	−1.87	−1.91	0.44	0.37	0.33	1.81	1.90	1.94
	BLU 1	−2.42	−2.30	−1.90	0.43	0.35	0.30	2.46	2.33	1.92
	BLU 2	−3.33	−3.10	−2.60	0.43	0.35	0.30	3.36	3.12	2.62
2	synt	−1.34	−1.43	−1.50	0.38	0.32	0.29	1.39	1.47	1.53
	BLU 1	−1.93	−1.84	−1.61	0.37	0.30	0.26	1.97	1.86	1.63
	BLU 2	−2.76	−2.73	−2.26	0.37	0.30	0.26	2.78	2.74	2.27
3	synt	1.73	1.52	1.50	0.84	0.68	0.59	1.93	1.67	1.62
	BLU 1	0.38	0.12	0.10	0.83	0.66	0.55	0.91	0.67	0.55
	BLU 2	−1.65	−0.77	−0.62	0.83	0.66	0.55	1.84	1.01	0.83
4	synt	0.42	0.30	0.27	0.50	0.41	0.36	0.66	0.51	0.45
	BLU 1	−0.87	−0.67	−0.50	0.49	0.39	0.33	1.00	0.78	0.60
	BLU 2	−1.57	−1.49	−1.21	0.50	0.40	0.33	1.65	1.55	1.25
5	synt	2.14	2.06	2.03	0.39	0.35	0.33	2.17	2.09	2.06
	BLU 1	0.74	0.58	0.46	0.37	0.32	0.28	0.83	0.66	0.53
	BLU 2	−0.12	−0.12	−0.12	0.38	0.32	0.28	0.39	0.34	0.31
6	synt	1.71	1.57	1.56	0.40	0.33	0.29	1.76	1.60	1.59
	BLU 1	0.50	0.33	0.31	0.39	0.31	0.26	0.63	0.46	0.40
	BLU 2	−0.40	−0.40	−0.34	0.40	0.31	0.26	0.56	0.51	0.43

Finally, in the Table 3 results of simulation study for simple regression superpopulation model in domains are presented. At the beginning it must be stressed that prediction accuracy is not sufficient mainly because of high values of the bias. It should be noticed that predictor (4) (both in cases of correct and incorrect specification of variance structure) has better accuracy comparing to the predictor of the form of synthetic ratio estimator. The highest values of ξ p-bias and ξ-expected p-MSE are observed in first and second domain. It results form fact, that elements of these domains belong to strata in which most of elements are from domains with higher β_d than in the first and second domain. It should be stressed that, as in Table 2, in some cases ξ-expected p-MSEs of the predictor of the form of synthetic ratio estimator increase due to the increase of sample size.

Table 3. Accuracy of predictors for simple regression superpopulation model in domains

Domain	Predictor	Bias (in %)			Root variance (in %)			Root MSE (in %)		
		Sample size			Sample size			Sample size		
		40	60	80	40	60	80	40	60	80
1	synt	336.89	336.39	336.06	1.97	1.66	1.49	336.89	336.40	336.06
	BLU 1	227.68	241.19	206.71	1.92	1.56	1.32	276.69	241.20	206.71
	BLU 2	280.01	244.17	209.23	1.94	1.57	1.33	280.02	244.17	209.23
2	synt	93.09	95.69	95.86	0.68	0.59	0.53	93.09	95.69	95.86
	BLU 1	76.90	70.01	59.05	0.66	0.55	0.46	76.90	70.01	59.05
	BLU 2	78.37	71.16	60.05	0.67	0.55	0.47	78.38	71.17	60.05
3	synt	–28.55	–28.89	–28.99	0.50	0.41	0.35	28.55	28.89	28.99
	BLU 1	–23.58	–20.16	–17.36	0.50	0.39	0.33	23.58	20.16	17.37
	BLU 2	–23.23	–19.89	–17.11	0.50	0.39	0.33	23.24	19.89	17.12
4	synt	–31.06	–31.37	–31.41	0.36	0.30	0.27	31.06	31.37	31.41
	BLU 1	–24.66	–21.86	–18.90	0.36	0.29	0.24	24.66	21.86	18.90
	BLU 2	–24.07	–21.32	–18.45	0.36	0.29	0.24	24.07	21.32	18.45
5	synt	–30.41	–29.82	–29.67	0.26	0.24	0.22	30.41	29.82	29.67
	BLU 1	–24.06	–20.56	–17.74	0.25	0.21	0.19	24.06	20.57	17.74
	BLU 2	–23.48	–20.10	–17.33	0.25	0.21	0.19	23.48	20.10	17.33
6	synt	–31.79	–30.72	–30.54	0.25	0.21	0.18	31.79	30.72	30.54
	BLU 1	–25.42	–21.60	–18.77	0.25	0.20	0.17	25.43	21.60	18.77
	BLU 2	–24.86	–21.18	–18.40	0.25	0.20	0.17	24.86	21.18	18.40

VI. CONCLUSION

In the paper properties of the predictor of the form of synthetic ratio estimator based on superpopulation approach were studied. It was proved that it is ξ-unbiased for simple regression superpopulation model in strata. For the model BLU predictor was presented and situations when both predictors are equal were shown. Properties of both predictors were additionally studied in the case of superpopulation model misspecification. Analytical considerations were supported by simulation study. It was shown that for discussed data both predictors gives similar results both for correct and incorrect model specification. For correct model specification and for simple regression model assumed in domains, accuracy of the BLU predictor is higher comparing to accuracy of the predictor of the form of synthetic ratio estimator in simulation study. When problem of model misspecification for analysed artificial population is discussed, both predictors gives better results for incorrect models assumed for strata than for incorrect models assumed for domains.

REFERENCES

Bolfarine H., Zacks S. (1992), *Prediction Theory for Finite Populations*, Springer-Verlag, New York.

Bracha Cz. (1994), *Metodologiczne aspekty badania małych obszarów*, Studia i Materiały, Z Prac Zakładu Badań Statystyczno-Ekonomicznych, 43, GUS, Warszawa.

Bracha Cz. (1996), *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa.

Domański Cz., Pruska K. (2001), *Metody statystyki małych obszarów*, Wyd. Uniwersytetu Łódzkiego, Łódź.

Getka-Wilczyńska E. (2000), Estimation of total domain in finite population, *Statistics in Transition*, **4**, 4, 711–728.

Royall R.M. (1976), The linear least squares prediction approach to two-stage sampling, *Journal of the American Statistical Association*, **71**, 473–657.

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

Wywiał J., Żądło T. (2003), *On Mean Square Error of Synthetic Ratio Estimator*, Studia Ekonomiczne, AE Katowice, 2003.

Tomasz Żądło

O SYNTETYCZNYM ESTYMATORZE ILORAZOWYM Z PUNKTU WIDZENIA PODEJŚCIA MODELOWEGO

Streszczenie

W opracowaniu rozważane są z punktu widzenia podejścia modelowego własności predyktora postaci syntetycznego estymatora ilorazowego wartości globalnej w domenie znanego z podejścia randomizacyjnego. Przedstawiony jest dowód jego ξ-nieobciążoności dla prostego regresyjnego modelu nadpopulacji w warstwach. Dla tego modelu zaprezentowany jest także predyktor typu BLU. Wyprowadzone są wzory opisujące wariancje predykcji obu predyktorów dla wspomnianego modelu nadpopulacji. Dla obu predyktorów rozważany jest także problem nieprawidłowej specyfikacji modelu nadpopulacji i dla tego przypadku wyprowadzone są błędy średniokwadratowe predykcji. Porównanie dokładności obu predyktorów wsparte jest analizą symulacyjną.