ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 286, 2013

Tomasz Żądło[*]

ON MSE ESTIMATION OF SOME MISSPECIFIED PREDICTOR

Abstract. The problem of prediction of subpopulation (domain) total is studied as in Rao (2003). The problem is inspired by results obtained by Żądło (2012) who considered two predictors – empirical best linear unbiased predictor (EBLUP) under some correct model and some simpler misspecified predictor. In the simulation study he showed that the misspecified predictor may be in some cases more accurate than the EBLUP derived under the correct model what resulted from the decrease of accuracy of the EBLUP due to the estimation of unknown parameters of the correct model. But the problem occurred in the case of MSE estimation – under the correct model the bias of the MSE estimator derived under the misspecified model was very large. Hence, in the paper we consider a predictor based on some misspecified model and we derive some MSE estimator under the correct model and we propose usage of two other MSE estimators.

Key words: small area estimation, MSE estimation, model misspecification.

I. INTRODUCTION

The finite population Ω consists of *N* units. The population vector of the variable of interest is $\mathbf{y}=\left[y_1,y_2,...,y_N\right]^T$ and it is treated as a realization of a random vector $\mathbf{Y}=\left[Y_1,Y_2,...,Y_N\right]^T$. The joint distribution of **Y** is denoted by ξ. From the population of *N* units, a sample *s* of *n* units is selected. For any sample *s* we can reorder the population vector **y** so that the first *n* elements are those in the sample: $\mathbf{y}=\left[\mathbf{y}_\mathbf{s}^\mathbf{T},\mathbf{y}_\mathbf{r}^\mathbf{T}\right]^T$ where $\mathbf{y}_\mathbf{s}$ is the *n* - dimensional vector of observed values and $\mathbf{y}_\mathbf{r}$ is the N_r - dimensional vector of unobserved values where $N_r=N-n$. The set of unsampled elements is denoted by $\Omega_r=\Omega-s$. Hence, the vector **Y** can be reordered as follows: $\mathbf{Y}=\left[\mathbf{Y}_\mathbf{s}^\mathbf{T},\mathbf{Y}_\mathbf{r}^\mathbf{T}\right]^T$. The population is divided into *D* domains Ω_d (*d*=1,...,*D*), each of size N_d (*d*=1,...,*D*). The set of sampled elements which belong to the *d*-th domain denoted by $s_d=\Omega_d\cap s$ consists of

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

n_d elements (where n_d may be random). Let us introduce additional notations: $\Omega_{rd} = \Omega_d - s_d$ and $N_{rd} = N_d - n_d$.

II. MODELS AND PREDICTORS

In the paper we consider two special cases of the General Linear Model (GLM) and the General Linear Mixed Model (GLMM). Models considered in the paper mentioned in the abstract (Żądło (2012)) were quite complicated (i.e. spatial and temporal correlation in the longitudinal surveys) what implies very time-consuming computations. In the paper the considered superpopulation models will be quite simple because of large number of iterations in the Monte Carlo simulation study (and bootstrap and jackknife within each Monte Carlo iteration).

Firstly, let the correct superpopulation model denoted by ξ be given by:

$$Y_{id} = \mathbf{x_{id}}\boldsymbol{\beta} + v_d + e_{id} \ (i\text{=}1,...,N;\ d\text{=}1,...,D) \tag{1}$$

where $\mathbf{x_{id}}$ is a $1 \times p$ vector of values of p auxiliary variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, $v_d \overset{iid}{\sim} (0, \sigma_v^2)$, $e_{id} \overset{iid}{\sim} (0, \sigma_e^2)$, and v_d and e_{id} are independent.

Secondly, let the misspecified superpopulation model denoted by M be given by

$$Y_{id} = \mathbf{x_{id}}\boldsymbol{\beta} + e_{id} \ (i\text{=}1,...,N;\ d\text{=}1,...,D) \tag{2}$$

where $e_{id} \overset{iid}{\sim} (0, \sigma_e^2)$, and v_d and e_{id} are independent. Under the misspecified model (2) the BLUP of the domain total is given by

$$\hat{\theta}_d^{(miss)} = \sum\nolimits_{i \in s_d} Y_{id} + \left(\sum\nolimits_{i \in \Omega_{rd}} \mathbf{x_{id}} \right) \hat{\boldsymbol{\beta}} \tag{3}$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X_s^T X_s})^{-1} \mathbf{X_s^T Y_s}$, $\mathbf{X_s}$ is $n \times p$ matrix of values of auxiliary variables in the sample. Under the misspecified model (3) the MSE of (3) is given

$$MSE_M(\hat{\theta}_d^{(miss)}) = \sigma_e^2 \left(N_{rd} + \left(\sum\nolimits_{i \in \Omega_{rd}} \mathbf{x_{id}} \right) \left(\mathbf{X_s^T X_s} \right)^{-1} \left(\sum\nolimits_{i \in \Omega_{rd}} \mathbf{x_{id}} \right)^T \right) \tag{4}$$

where $\mathbf{x_{id}}$ is a $p \times 1$ vector of values of auxiliary variables for the *i*th population element in the *d*th domain.

Let us introduce additional notations. Let $\mathbf{A_s^{-1}} = (\mathbf{X_s^T X_s})^{-1}$, $\tilde{\mathbf{x}}_{\mathbf{rd}} = \sum_{i \in \Omega_{rd}} \mathbf{x_{id}}$, $\mathbf{Y_{rd}} = \left[Y_{id} \right]_{N_{rd} \times 1}$. Under the correct model the MSE of the predictor (3) is given by:

$$MSE_{\xi}(\hat{\theta}_d^{(miss)}) = Var_{\xi}(\hat{\theta}_d^{(miss)} - \theta_d) =$$

$$= Var_{\xi}\left(\sum\nolimits_{i \in s_d} Y_{id} + \left(\sum\nolimits_{i \in \Omega_{rd}} \mathbf{x_{id}}\right)(\mathbf{X_s^T X_s})^{-1} \mathbf{X_s^T Y_s} - \sum\nolimits_{i \in \Omega_d} Y_{id}\right) =$$

$$= Var_{\xi}\left(\tilde{\mathbf{x}}_{\mathbf{rd}} \mathbf{A_s^{-1} X_s^T Y_s} - 1_{N_{rd}}^T \mathbf{Y_{rd}}\right) =$$

$$= \mathbf{A_s^{-1} A_s^{-1} X_s^T} D_{\xi}^2(\mathbf{Y_s}) \mathbf{X_s A_s^{-1}} \tilde{\mathbf{x}}_{\mathbf{rd}}^{\mathbf{T}} + \mathbf{1}_{N_{rd}}^T D_{\xi}^2(\mathbf{Y_{rd}}) \mathbf{1}_{N_{rd}} + \tilde{\mathbf{x}}_{\mathbf{rd}} \mathbf{A_s^{-1} X_s^T} Cov_{\xi}(\mathbf{Y_s}, \mathbf{Y_{rd}}) \mathbf{1}_{N_{rd}} =$$

$$= \tilde{\mathbf{x}}_{\mathbf{rd}} \mathbf{A_s^{-1} X_s^T} \left(diag_{1 \le d \le D}(\sigma_e^2 \mathbf{I}_{n_d} + \sigma_v^2 \mathbf{1}_{n_d} \mathbf{1}_{n_d}^T) \right) \mathbf{X_s A_s^{-1}} \tilde{\mathbf{x}}_{\mathbf{rd}}^{\mathbf{T}} +$$

$$+ \sigma_e^2 N_{rd} + \sigma_v^2 N_{rd}^2 - 2 N_{rd} \sigma_v^2 \tilde{\mathbf{x}}_{\mathbf{rd}} \mathbf{A_s^{-1} X_{sd}^T} \mathbf{1}_{n_d} . \tag{5}$$

III. MSE ESTIMATORS

Some results of MSE estimation are presented in the literature but for optimal predictors. MSE estimators of EBLUPs are presented inter alia by Prasad and Rao (1990), Datta and Lahiri (2000), Żądło (2009). Some results for MSE estimators of empirical best predictors are presented by Jiang (2003), Jiang and Lahiri (2001), Jiang, Lahiri and Wan (2002), Molina and Rao (2010). In the paper MSE estimators will be studied not for a predictor optimal in some sense but for a predictor derived under some misspecified model.

Let us consider the misspecified predictor (3). Let us note that the predictor does not depend on unknown model parameters. Its MSE under the correct model (1) is given by (5) and its formula depends on unknown model parameters. In the paper four estimators of $MSE_{\xi}(\hat{\theta}_d^{(miss)})$ given by (5) will be considered.

Firstly, we will consider MSE estimator given by the formula (4) where σ_e^2 is replaced by $\hat{\sigma}_e^2 = (n-p)^{-1} \sum_{i=1}^{n} \sum_{d=1}^{D} (Y_{id} - \mathbf{x_{id}} \hat{\boldsymbol{\beta}})$. The MSE estimator is unbiased but under the misspecified model (2). We will be interested in its properties under the correct model (1). The MSE estimator in the simulation study will be denoted by *miss*.

Secondly, we will consider naive MSE estimator, given by (5), where σ_e^2 and σ_v^2 are replaced by restricted maximum likelihood (REML) estimates under

(5) and normality of random components. Simulation results for the naive MSE estimator presented on the figures below will be denoted by *naive*.

Thirdly, we will correct the naive estimator due to the bias. Let $\boldsymbol{\delta} = \begin{bmatrix} \sigma_e^2 & \sigma_v^2 \end{bmatrix}$. Using our notations the general formula of jackknife MSE estimator of Jiang, Lahiri, Wan (2002) is given by:

$$M\hat{S}E_{\xi}^{jack}(\hat{\theta}_d^{(miss)}) = b_d(\hat{\boldsymbol{\delta}}) - \frac{D-1}{D}\sum_{d=1}^{D}\left(b_d(\hat{\boldsymbol{\delta}}_{-d}) - b_d(\hat{\boldsymbol{\delta}})\right) + \\ + \frac{D-1}{D}\sum_{d=1}^{D}\left(\hat{\theta}_d^{(miss)}(\hat{\boldsymbol{\delta}}_{-d}) - \hat{\theta}_d^{(miss)}(\hat{\boldsymbol{\delta}})\right)^2 \tag{6}$$

where $\hat{\boldsymbol{\delta}}_{-d}$ is given by the same formula as $\hat{\boldsymbol{\delta}}$ but it is based on the set $s - s_d$ instead of s, $b(\hat{\boldsymbol{\delta}})$ is given by (5) where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}$, $b(\hat{\boldsymbol{\delta}}_{-d})$ is given by (5) where $\boldsymbol{\delta}$ is replaced by $\hat{\boldsymbol{\delta}}_{-d}$. What is more, $\hat{\theta}_d^{(miss)}(.)$ is given by (3) and it does not depend on unknown parameters. Hence, in our case the last (third) term of the right hand side of (6) equals zero. Simulation results for the jackknife MSE estimator presented on the figures below will be denoted by *jack*.

Fourthly, we will us parametric bootstrap method presented inter alia by Gonzalez-Manteiga et al. (2008) and Molina and Rao (2010). The procedure will be as follows:

(i) based on the sample we estimate parameters of the model (1) using REML method under normality and we obtain estimates $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_e^2$, $\hat{\sigma}_v^2$,

(ii) then, we construct bootstrap superpopulation model ξ^*: $Y_{id}^* = \mathbf{x_{id}}\hat{\boldsymbol{\beta}} + v_d^* + e_{id}^*$, where i=1,...,N, d=1,....,D, $v_d^* \overset{iid}{\sim} N(0,\hat{\sigma}_v^2)$, $e_{id}^* \overset{iid}{\sim} N(0,\hat{\sigma}_e^2)$,

(iii) based on B realizations of $Y_{id}^{*(b)}$, where b=1,...,B, we compute B values of (3) which will be denoted by $\hat{\theta}_d^{(miss)(b)}$ (based on the sample with the same indices as in the original population) and B values of domain totals $\theta_d^{(b)} = \sum_{i\in\Omega_d} Y_{id}^{*(b)}$,

(iv) finally, we compute bootstrap MSE estimator as follows: $M\hat{S}E_{\xi}^{boot}(\hat{\theta}_d^{(miss)}) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}_d^{(miss)(b)} - \theta_d^{(b)})^2$.

What is important, in the bootstrap procedure normality is assumed both in the step (i) of parameter's estimation and step (ii) of generating data. It means that in the simulation study for cases when the assumption of normality components is not met the presented procedure will be studied in the case of bootstrap model misspecification. Simulation results for the bootstrap MSE estimator presented on the figures below will be denoted by *boot*.

IV. SIMULATION STUDY

In each simulation study 2000 realizations of superpopulation model (1) were generated using R package (R Development Core Team (2012)). The population of size *N*=1500 elements was divided into *D*=30 domains each of size $N_d = 50$ elements. The sample size in 10 domains was 2 elements, in the next 10 domains 3 elements and in the last 10 domains 4 elements what means that the overall sample size equaled *n*=90 elements. For each out of 2000 realizations of superpopulation model *B*=200 iterations of bootstrap procedure and *D*=30 iterations of jackknife procedure were conducted.

It was assumed that $\forall_{i,d}\ \mathbf{x_{id}\beta} = \beta = 0$ (but β was estimated in the simulations), $\sigma_e^2 = 1$, three scenarios of values of σ_v^2: $\sigma_v^2 = 10$ or $\sigma_v^2 = 1$ or $\sigma_v^2 = 0{,}1$, nine scenarios of distributions of random components $v_d \overset{iid}{\sim} (0, \sigma_v^2)$ and $e_{id} \overset{iid}{\sim} (0, \sigma_e^2)$: normal-normal (i.e. normal distribution of v_d and normal distribution of e_{id}), normal-uniform (i.e. normal distribution of v_d and uniform distribution of e_{id}), normal-shifted exponential (i.e. normal distribution of v_d and shifted exponential of e_{id}), uniform-normal, uniform-uniform, uniform-shifted exponential, shifted exponential-normal, shifted exponential-uniform, shifted exponential-shifted exponential.

In each simulation we compute values of the misspecified predictor (3), its simulation MSE but under the correct model (1), values of four MSE estimators proposed in the section 3 and their biases. In all of the considered cases parameters σ_e^2 and σ_v^2 are estimated using restricted maximum likelihood method assuming normality of random components even in the cases of different distributions of random components. It means that in all of the cases except normal-normal case the problem of misspecification of distribution of random components in the model is considered.

In the following figures values of biases of the misspecified MSE estimator *miss* are omitted because of the unacceptable biases – for different domains and different distributions of random components its relative biases obtained in the simulation were as follows: for $\sigma_v^2 = 10$ from –96,90% to –95,85; for $\sigma_v^2 = 1$ from –94,32% to –92,47; and for $\sigma_v^2 = 0{,}1$ from –76,05% to –67,40%.

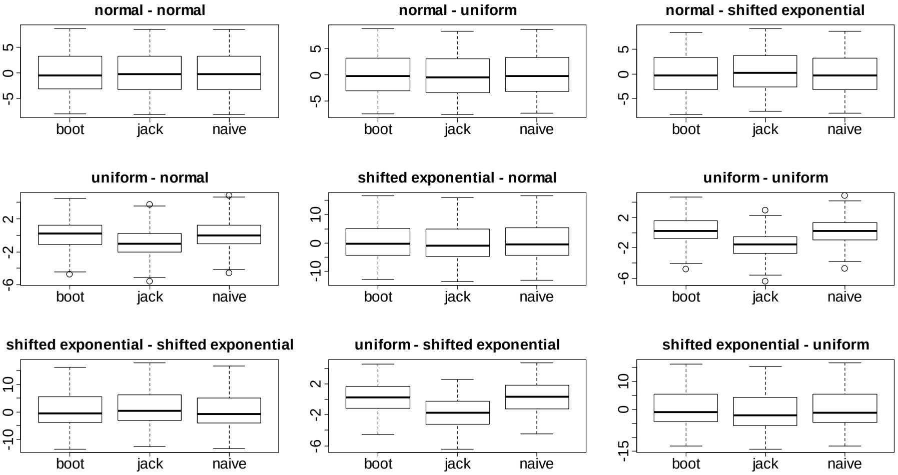

Fig. 1. Relative biases of MSE estimators for $\sigma_v^2 = 10$

Source: author's calculations.

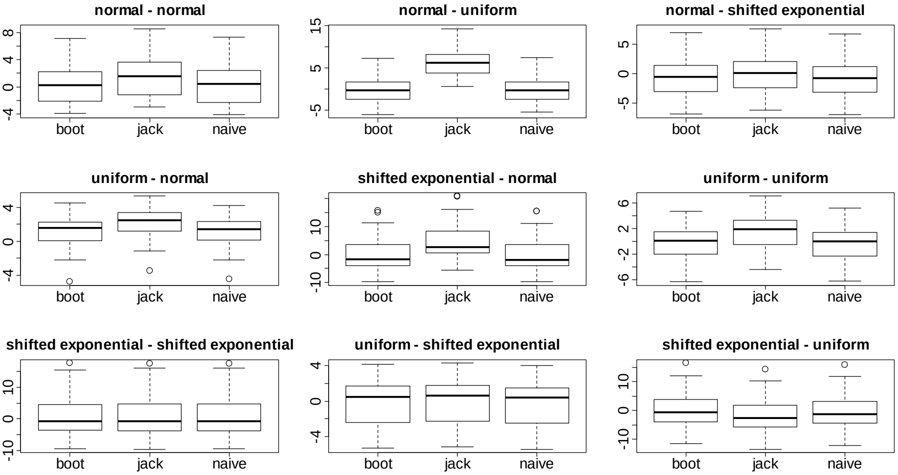

Fig. 2. Relative biases of MSE estimators for $\sigma_v^2 = 1$

Source: author's calculations

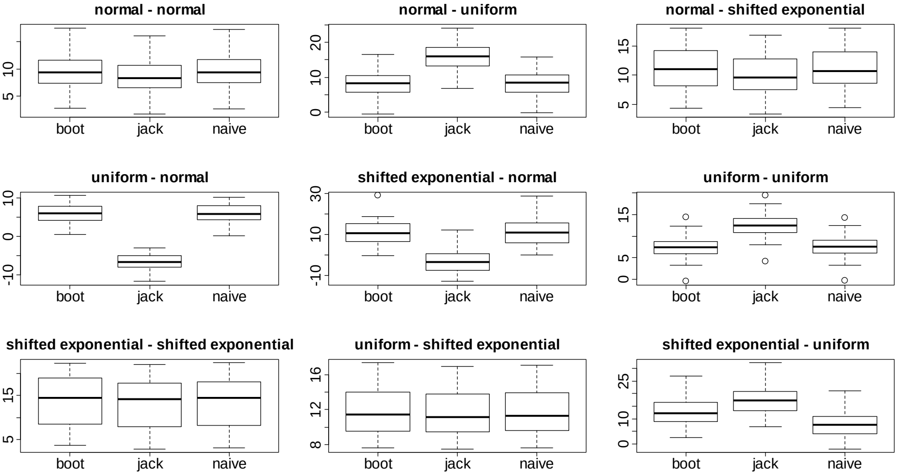

Fig. 3. Relative biases of MSE estimators for $\sigma_v^2 = 0,1$

Source: author's calculations.

In the Figures 1-3 distributions of relative model biases of naive, parametric bootstrap and jackknife MSE estimators for *D*=30 domains and different distributions of random components are presented. What is interesting, the biases of these three MSE estimators are very similar within three considered scenarios of different values of σ_v^2. What is more, for $\sigma_v^2 = 10$ and $\sigma_v^2 = 1$ (see Figure 1 and Figure 2) average over domains relative biases are close to zero even if the distribution of random components is misspecified (i.e. if other cases then normal-normal case are considered). Relative biases of MSE estimators for the case when $\sigma_v^2 = 0,1$ (see Figure 3) are higher what results from the fact that in this case area effect v_d may be treated as negligible what implies low accuracy of σ_v^2 estimates.

Results obtained for three estimators are similar and it is difficult to indicate which is better in terms of relative bias. But it is possible to improve the results obtained for bootstrap MSE estimator. It is possible using double bootstrap method proposed by Hall and Maiti (2006) but testing the method in the Monte Carlo simulation study especially for large population may not be computationally feasible.

V. CONCLUSION

In the paper we analyse the problem of MSE estimation of some predictor which is EBLUP under some misspecified model. The naive MSE estimator is derived and it is compared in different cases with other MSE estimators. Every of the proposed MSE estimators performs well even in the cases of model misspecification.

ACKNOWLEDGEMENTS

The research was supported by National Science Centre grant 2011/03/B/HS4/00954.

REFERENCES

Datta, G. S., Lahiri, P. (2000), A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems, *Statistica Sinica*, 10, 613-627.

Gonzalez-Manteiga, W., Lombardia, M. J., Molina, I., Morales, D., Santamaria, L. (2008), Bootstrap mean squared error of a small-area EBLUP, *Journal of Statistical Computation and Simulation*, 78, 443-462.

Hall, P., Maiti, T. (2006), On Parametric Bootstrap Methods for Small Area Prediction, *Journal Royal Statistical Society Series B*, 68, 221-238.

Jiang, J. (2003), Empirical best prediction for small-area inference based on generalized linear mixed models, *Journal of Statistical Planning and Inference*, 111, 117-127.

Jiang, J., Lahiri, P., (2001), Empirical best prediction for small area inference with binary data, *Ann. Inst. Statist. Math.*, 53, 217-243.

Jiang J., Lahiri P, Wan S.-M. (2002), Unified jackknife theory for empirical best prediction with M-estimation, *The Annals of Statistics*, 30, 6, 1782-1810.

Molina, I. and Rao, J.N.K (2010). Small area estimation of poverty indicators, *The Canadian Journal of Statistics,* 38 (3), 369-385.

Prasad, N. G. N, Rao, J. N. K. (1990), The estimation of mean the mean squared error of small area estimators, *Journal of the American Statistical Association*, 85, 163-171.

R Development Core Team (2012), A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna.

Rao J.N.K (2003), *Small area estimation*, John Wiley and Sons, New Jersey.

Żądło T. (2009), *On prediction of domain totals based on unbalanced longitudinal data*, [in:] Wywiał J., Żądło T. (eds.) Survey Sampling in Economic and Social Research, University of Economic in Katowice, Katowice.

Żądło T. (2012), On accuracy of two predictors for spatially and temporally correlated longitudinal data, *Studia Ekonomiczne*, 120, 97-105.

Tomasz Żądło

O ESTYMACJI MSE DLA PEWNEGO PREDYKTORA W PRZYPADKU ZŁEJ SPECYFIKACJI MODELU

Rozważany jest problem predykcji wartości globalnej w podpopulacji (domenie) jak w Rao (2003). Analizowane jest wykorzystanie predyktora, który jest empirycznym najlepszym liniowym nieobciążonym predyktorem, ale przy założeniu błędnego modelu. Dla rozważanego predyktora wyprowadzono postać naiwnego estymatora MSE dla prawidłowego modelu nadpopulacji oraz zaproponowano wykorzystanie estymatorów MSE typu jackknife i parametryczny bootstrap. W badaniu symulacyjnym analizowano względne obciążenia zaproponowanych estymatorów MSE.