A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 255, 2011

Tomasz Żądło[*]

ON SOME PRACTICAL ISSUES IN PREDICTION OF DOMAIN MEAN AND FRACTION

Abstract. Survey research in practice is often based on non-random samples what implies that the model approach should be used. What is more, in many surveys no information is available on auxiliary variables for non-sampled elements of population. One of the key issues in this case may be a problem of model misspecification too. In this paper two proposals of predictors of domain means and fractions based on models without auxiliary variables are studied including the problem of accuracy in the case of model misspecification and estimating subpopulations sizes.

Key words: small area estimation, best linear unbiased predictors, model misspecification.

I. BASIC NOTATIONS

The population Ω of size N is divided into C disjoint subpopulations Ω_c $(c=1,2,...,C)$ called strata each of size N_c. From the population random or non-random sample s of size n is selected. Let samples in strata be denoted by $s_c = \Omega_c \cap s$ $(c=1,2,...,C)$ and their sizes by n_c. Let $\Omega_{rc} = \Omega_c - s_c$ and $N_{rc} = N_c - n_c$. Let us study another division of the population into subpopulations. Let the population be also divided into D subpopulations Ω_d $(d=1,2,...,D)$ called domains of sizes N_d. Let $s_d = \Omega_d \cap s$, $\Omega_{rd} = \Omega_d - s_d$ and $\Omega_{cd} = \Omega_c \cap \Omega_d$ $(d=1,2,...,D)$ and their sizes by denoted respectively by n_d, $N_{rd} = N_d - n_d$ and N_{cd}. Some of Ω_{cd} sets can be empty sets. The sample in the set Ω_{cd} is denoted by $s_{cd} = \Omega_{cd} \cap s$ and its size by n_{cd}. Let $\Omega_{rcd} = \Omega_{cd} - s_{cd}$ and $N_{rcd} = N_{cd} - n_{cd}$. What is more, elements of any domain may belong to many strata.

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

II. FIRST PROPOSAL

The following considerations are based on Royall's (1976) theorem. Let us consider the following special case of the General Linear Model (GLM). Let us assume that $Y_1, Y_2, ..., Y_N$ are independent and

$$\begin{cases} E_\xi(Y_{ci}) = \mu_c \\ D^2_\xi(Y_{ci}) = \sigma^2_c \end{cases} \tag{1}$$

for each $i = 1,2,...,N$ and $c = 1,2,...,C$. In the case of the first proposal no other assumptions (e.g. on the distribution function of random variables) are needed. The unbiased estimator of σ^2_c is given by:

$$\hat{\sigma}^2_c = \frac{1}{n_c - 1}\sum_{i=1}^{n_c}(Y_i - \overline{Y}_{sc})^2 \tag{2}$$

(where $\overline{Y}_{sc} = \frac{1}{n_c}\sum_{i=1}^{n_c} Y_i$) what simplifies in the case of the prediction of fraction (i.e. in the case of the prediction of the mean of zero-one variables $Y_1, Y_2, ..., Y_N$) to the following formula:

$$\hat{\sigma}^2_c = \frac{n_c}{n_c - 1}\overline{Y}_{sc}(1 - \overline{Y}_{sc}) \tag{3}$$

where $\overline{Y}_{sc}$ is in this case a sample fraction in the *c*-th strata.

Under (1) the Best Linear Unbiased Predictor (BLUP) of the domain mean or fraction $\overline{Y}_d = \frac{1}{N_d}\sum_{i=1}^{N_d} Y_i$ is given by:

$$\hat{\overline{Y}}_{d\,BLUP} = \frac{1}{N_d}\sum_{c=1}^{C}\left(n_{cd}\overline{Y}_{scd} + (N_{cd} - n_{cd})\overline{Y}_{sc}\right), \tag{4}$$

where $\bar{Y}_{scd} = \frac{1}{n_{cd}} \sum_{i=1}^{n_{cd}} Y_i$. Note that the predictor (4) may be used even in the case of zero domain sample size. Under (1) prediction MSE of $\hat{\bar{Y}}_{d\,BLUP}$ is given by:

$$MSE_{\xi}\left(\hat{\bar{Y}}_{d\,BLUP}\right) = \frac{1}{N_d^2} \sum_{c=1}^{C} \sigma_c^2 \left(\frac{(N_{cd} - n_{cd})^2}{n_{cd}} + (N_{cd} - n_{cd}) \right), \tag{5}$$

and its unbiased estimator by:

$$\hat{MSE}_{\xi}\left(\hat{\bar{Y}}_{d\,BLUP}\right) = \frac{1}{N_d^2} \sum_{c=1}^{C} \hat{\sigma}_c^2 \left(\frac{(N_{cd} - n_{cd})^2}{n_{cd}} + (N_{cd} - n_{cd}) \right), \tag{6}$$

where $\hat{\sigma}_c^2$ is given by (2). These equations can also be used in the case of the prediction of the domain fraction. In this case the equation (2) simplifies to the formula (3).

III. SECOND PROPOSAL

The following considerations are based on Royall's (1976) theorem. We assume that (similar assumptions but for domains instead of strata are considered by Chambers and Ayoub (2003) p. 12):

$$Y_{ic} = \mu + v_c + e_{ic}, \tag{7}$$

where $i = 1,2,...,N$ and $c = 1,2,...,C$, μ is unknown fixed parameter, $v_c \overset{iid}{\sim} (0, \sigma_v^2)$, $e_{ic} \overset{iid}{\sim} (0, \sigma_e^2)$ and v_c and e_{ic} are independent. In this case the additional assumption of normality of random components is made to derive the MSE and its estimator. Let $\boldsymbol{\delta} = \begin{bmatrix} \sigma_e^2 & \sigma_v^2 \end{bmatrix}^T$ be the vector of unknown parameters of the variance-covariance matrix. The model (7) is the special case of the General Linear Mixed Model (GLMM). It is worth noting that $\mu + v_c$ in (7) may be treated as a random mean in strata what may imply more flexible method of prediction.

From (7) we obtain that (Valliant *et al.*, 2000, s.256):

$$E_{\xi}(Y_{ci}) = \mu$$

$$Cov_{\xi}(Y_{ic}, Y_{i'c'}) = \begin{cases} \sigma_e^2 + \sigma_v^2 & \text{for} \quad i = i' \wedge c = c' \\ \sigma_v^2 & \text{for} \quad i \neq i' \wedge c = c' \\ 0 & \text{otherwise} \end{cases}. \tag{8}$$

From (8) we obtain that in the model (7) the correlation of random variables in strata is taken into consideration (in (1) independence of random variables is assumed). The BLUP under (7) is a function of unknown model parameters. If these parameters are replaced by some estimators, two-stage predictor called Empirical BLUP (EBLUP) is obtained. It remains unbiased under some week assumptions presented in some general case by Żądło (2004).

In this paper Restricted Maximum Likelihood (REML) method of estimation of $\boldsymbol{\delta}$ under normality assumption is considered. What is important, Jiang (1996) proves that this method is robust on nonnormality.

For (7) the BLUP of the domain mean and its MSE are given by:

$$\hat{\bar{Y}}_{d\,BLUP} = \frac{1}{N_d}\sum_{c=1}^{C}\left(n_{cd}\bar{Y}_{scd} + N_{rcd}\hat{\mu} + N_{rcd}n_{cd}\sigma_v^2(\sigma_e^2 + n_c\sigma_v^2)^{-1}(\bar{Y}_{scd} - \hat{\mu})\right), \tag{9}$$

where

$$\hat{\mu} = \left(\sum_{c=1}^{C} n_c(\sigma_e^2 + n_c\sigma_v^2)^{-1}\right)^{-1}\sum_{c=1}^{C}(\sigma_e^2 + n_c\sigma_v^2)^{-1}\sum_{i=1}^{n_c} Y_{ic},$$

and

$$MSE_{\xi}(\hat{\bar{Y}}_{d\,BLUP}) = \frac{1}{N_d^2}(g_{1d}(\boldsymbol{\delta}) + g_{2d}(\boldsymbol{\delta})), \tag{10}$$

where

$$g_{1d}(\boldsymbol{\delta}) = \sum_{c=1}^{C}\left(N_{rcd*}(\sigma_e^2 + N_{rcd*}\sigma_v^2) + n_{cd*}(N_{rcd*}\sigma_v^2)^2(\sigma_e^2 + n_c\sigma_v^2)^{-1}\right), \tag{11}$$

$$g_{2d}(\boldsymbol{\delta}) = \left(\sum_{c=1}^{C} N_{rcd*}\sigma_e^2(\sigma_e^2 + n_c\sigma_v^2)^{-1}\right)^2\left(\sum_{c=1}^{C} n_c(\sigma_e^2 + n_c\sigma_v^2)^{-1}\right)^{-1}. \tag{12}$$

Żądło (2009) derives the approximation of the MSE of the EBLUP and its approximately unbiased estimator and proves that they are correct to the term $o(D^{-1})$ for some general case under some assumptions including normality and independence between areas (which is not met under (7)). Based on these results we obtain (for REML estimates of $\boldsymbol{\delta}$):

$$MSE_{\xi}(\hat{\bar{Y}}_{d\,EBLUP}) \approx \frac{1}{N_d^2}(g_{1d}(\boldsymbol{\delta}) + g_{2d}(\boldsymbol{\delta}) + g_{3d}^{*}(\boldsymbol{\delta})), \tag{13}$$

and MSE estimator:

$$\hat{MSE}_{\xi}(\hat{\bar{Y}}_{d\,EBLUP}) = \frac{1}{N_d^2}(g_{1d}(\hat{\boldsymbol{\delta}}) + g_{2d}(\hat{\boldsymbol{\delta}}) + 2g_{3d}^{*}(\hat{\boldsymbol{\delta}})), \tag{14}$$

where

$$g_{3d}^{*}(\boldsymbol{\delta}) = \left(\sum_{c=1}^{C} N_{rcd*}^2 n_{cd*} a_c^{-4}(\sigma_e^2 + n_{cd*}\sigma_v^2)\right)\left(\sigma_v^4 I_{vv} - 2\sigma_e^2\sigma_v^2 I_{ev} + \sigma_e^4 I_{ee}\right), \tag{15}$$

where

$$I_{vv} = 2a^{-1}\sum_{c=1}^{C} n_c^2 a_c^{-2},\ I_{ve} = -2a^{-1}\sum_{c=C}^{C} n_c a_c^{-2},\ I_{ee} = 2a^{-1}\sum_{c=1}^{C}\left((n_c - 1)\sigma_e^{-4} + a_c^{-2}\right),$$

$$a_c = \sigma_e^2 + n_c\sigma_v^2,\ a = \left(\sum_{c=1}^{C}\left((n_c - 1)\sigma_e^{-4} + a_c^{-2}\right)\right)\left(\sum_{c=1}^{C} n_c^2 a_c^{-2}\right) - \left(\sum_{c=1}^{C} n_c a_c^{-2}\right)^2.$$

The equations presented above may be used for prediction of domain fraction (as domain mean of zero-one variable) but because of the introduced assumptions the accuracy of this method in the case of the prediction of fraction will be studied in the simulation study.

IV. SIMULATION ANALYSIS

In the Monte Carlo simulation analysis prepared using R (R Development Core Team (2009)) real data on the population of size 10093 elements divided into 21 geographical strata are used. From the population a sample of size 3000 elements was drawn. What is more, the population was (artificially) divided into

20 domains. Sizes of the first 19 domains equal 500 elements and the size of the last domain equals 593 elements. The domain labels are assigned to the population elements in the order of labels of population elements (and hence the order of labels of strata). Hence, elements of one domain belong from 1 to 3 strata. Two simulation analyses are made by generation of zero-one variable based on the model (1). In the first simulation study fractions used in simulation as values μ_c are real but sample fractions for the data. In this simulation strata are ordered in ascending order of μ_c. Artificial fractions used in the second simulation are presented in the table 1. Hence, in the first simulation random variables for elements of one domain (i.e. which belong from 1 to 3 neighboring strata) have similar expected values. In the case of the second simulation expected values of random variables for elements of one domain may differ significantly because expected values in neighboring strata may vary significantly (see table 1).

Table 1. Data used in the simulation

numer of stratum	size of stratum	sample size in stratum	fraction in stratum in simulation 1	fraction in stratum in simulation 2
1	517	160	0.5438	0.4050
2	241	75	0.5733	0.3100
3	295	91	0.6044	0.5950
4	261	81	0.6049	0.8325
5	875	269	0.6245	0.3575
6	752	234	0.6325	0.9275
7	293	90	0.6667	0.1200
8	481	154	0.6688	0.4525
9	431	82	0.6829	0.1675
10	310	51	0.6863	0.5000
11	333	104	0.6923	0.5475
12	366	43	0.6977	0.6900
13	305	95	0.7158	0.6425
14	582	184	0.7174	0.8800
15	394	126	0.7222	0.2625
16	604	204	0.7402	0.7850
17	696	219	0.7489	0.9750
18	759	238	0.7647	0.0725
19	631	197	0.7868	0.0250
20	824	259	0.7876	0.2150
21	143	44	0.8409	0.7375
sum	10093	3000		

In each simulation three cases connected with subpopulation sizes are considered:

- subpopulation sizes are known (in the tables presenting simulation results to the name of the predictor „r" is added),
- sizes of products of strata and domains are estimated (to the name of the predictor "e" is added). To mimic the problem of the size estimation in the simulation rounded values of random variables of 0 expected values and standard deviation equal 20% of the real size of the subpopulation are added to the real sizes. Respective sums of these values are treated as estimated domain and strata sizes in simulations,
- sizes of products of strata and domains are estimated (as in the previous case) but we assume that seven real sizes of 3 neighboring strata (sizes of subpopulations formed from strata 1–3, strata 4–6, etc.) are known. Estimated sizes of strata and domains are proportionally corrected to obtain real seven values of sums of the strata (to the name of the predictor "ec" is added).

In these cases the differences between real and estimated sizes of domains are between –33% and 24,8% and between real and corrected estimated sizes of domains between –30,6% and 29,4%. The differences between real and estimated sizes of strata are between –32,73% and 50,62% and between real and corrected estimated sizes of strata between –33% and 24,8%.

In both simulations two cases of superpopulation model are considered:

- the assumed superpopulation model is correct (to the name of the predictor "1" is added),
- the assumed superpopulation model is misspecified. Strata assumed in the prediction are sums of three neighboring strata assumed in the model (to the name of the predictor "2" is added).

The predictor presented as the proposal 1 is denoted in the tables below as BLUP. Tables with results of the predictor presented as the proposal 2 will be omitted as giving mainly worse and less sensitive on model misspecification results than the proposal 1. To the name of the predictor "r", "e" or "ec" is added due to the possible treatment of subpopulation sizes (as described above) and "1" or "2" due to the problem of specification of the model (as described above). The number of iterations equals 100 000.

It is known the BLUP presented as the proposal 1 if the model is correctly specified and if the subpopulation sizes are known (denoted as BLUP_1_r) gives predictions with the minimum prediction RMSE but in the class of model unbiased predictors. It means that BLUP_2 predictors (BLUPs for misspecified superpopulation model) theoretically can give smaller RMSE than BLUP_1_r because they do not belong to the class of unbiased predictors. For the case presented in the table 6 the RMSEs for the BLUP_1_r are the smallest (although the differences may be observed for further decimal place). In some cases of simulation 2 (table 3) model RMSE of BLUP_1_r is higher comparing to other predictors.

Let us compare results for the correct model specification and the real subpopulation sizes with other cases. In the case of similar fractions in strata (the simulation 1) the increase of the bias and the RMSE due to the model misspecification and the estimation of subpopulation sizes may be treated as acceptable. In many cases the bias of the MSE estimator may be treated as acceptable too (the MSE estimator usually underestimate the real MSE). In the case of high differences between fractions in strata (the simulation 2) the increase of the bias and the RMSE is not acceptable.

Table 2. Absolute model biases – proposal 1, simulation 1

		BLUP_1_r	BLUP_2_r	BLUP_1_e	BLUP_2_e	BLUP_1_ec	BLUP_2_ec
results for 20 domains	min	0.000*	–0.017	–0.003	–0.017	–0.003	–0.017
	Q1	0.000*	–0.005	0.000	–0.004	0.000	–0.004
	Me	0.000*	0.000	0.000	0.000	0.000	0.000
	mean	0.000*	0.000	0.000	0.000	0.000	0.000
	Q3	0.000*	0.003	0.000	0.003	0.000	0.003
	max	0.000*	0.017	0.001	0.015	0.001	0.015

* real value equals zero.

Table 3. Absolute model RMSE – proposal 1, simulation 1

		BLUP_1_r	BLUP_2_r	BLUP_1_e	BLUP_2_e	BLUP_1_ec	BLUP_2_ec
results for 20 domains	min	0.017	0.016	0.017	0.016	0.017	0.016
	Q1	0.026	0.021	0.026	0.021	0.026	0.021
	Me	0.030	0.023	0.030	0.023	0.030	0.023
	mean	0.029	0.024	0.029	0.024	0.029	0.024
	Q3	0.033	0.026	0.034	0.026	0.034	0.026
	max	0.039	0.033	0.046	0.033	0.046	0.033

Table 4. Relative biases of model MSE estimator (in %) – proposal 1, simulation 1

		BLUP_1_r	BLUP_2_r	BLUP_1_e	BLUP_2_e	BLUP_1_ec	BLUP_2_ec
results for 20 domains	min	–0.311*	–28.313	–44.904	–39.176	–40.220	–34.672
	Q1	–0.218*	–11.037	–5.656	–10.283	–3.258	–13.197
	Me	–0.051*	–3.775	–0.249	–5.133	–0.134	–4.335
	mean	–0.022*	–7.046	–2.765	–8.672	–2.754	–8.446
	Q3	0.085*	–0.434	0.941	–0.409	1.047	–0.200
	max	0.570*	0.090	12.062	3.371	8.212	3.907

* real value equals zero.

Table 5. Absolute model biases – proposal 1, simulation 2

		BLUP_1_r	BLUP_2_r	BLUP_1_e	BLUP_2_e	BLUP_1_ec	BLUP_2_ec
results for 20 domains	min	0.000*	–0.353	–0.034	–0.356	–0.034	–0.356
	Q1	0.000*	–0.096	–0.002	–0.107	–0.002	–0.102
	Me	0.000*	–0.022	0.000	–0.018	0.000	–0.021
	mean	0.000*	0.002	0.000	–0.007	0.000	–0.007
	Q3	0.000*	0.073	0.008	0.071	0.008	0.070
	max	0.000*	0.268	0.023	0.246	0.023	0.241

* real value equals zero.

Table 6. Absolute model RMSE – proposal 1, simulation 2

		BLUP_1_r	BLUP_2_r	BLUP_1_e	BLUP_2_e	BLUP_1_ec	BLUP_2_ec
results for 20 domains	min	0.012	0.029	0.012	0.024	0.012	0.028
	Q1	0.014	0.053	0.020	0.049	0.020	0.049
	Me	0.025	0.099	0.027	0.108	0.027	0.103
	mean	0.024	0.127	0.028	0.119	0.028	0.119
	Q3	0.031	0.175	0.032	0.162	0.032	0.164
	max	0.041	0.353	0.052	0.356	0.052	0.357

Table 7. Relative biases of model MSE estimator (in %) – proposal 1, simulation 2

		BLUP_1_r	BLUP_2_r	BLUP_1_e	BLUP_2_e	BLUP_1_ec	BLUP_2_ec
results for 20 domains	min	–0.456*	–99.517	–79.156	–99.461	–79.491	–99.471
	Q1	–0.194*	–98.445	–33.797	–98.278	–26.077	–98.289
	Me	0.218*	–96.847	–16.012	–96.884	–14.890	–96.927
	mean	0.169*	–82.802	–21.257	–80.327	–19.777	–81.736
	Q3	0.463*	–73.415	–7.997	–68.019	–7.260	–68.519
	max	0.752*	–38.054	1.389	–34.702	1.974	–34.782

* real value equals zero.

In practice, in the case of stratification of the population, strata should be formed by subpopulations with similar fractions. The subpopulation sizes should be known or estimated with high accuracy. The accuracy of prediction may also be acceptable (even if subpopulations sizes are estimated) for subpopulations but if fractions in strata, to which the subpopulation belongs, are similar.

V. CONCLUSION

In the paper the problem of prediction of domain fraction and mean is considered both for random and nonrandom samples. The problem of influence of model misspecification and replacing unknown subpopulation sizes by their estimates on prediction accuracy is analyzed in the simulation study.

BIBLIOGRAPHY

Chambers R., Ayoub S. (2003), Small area estimation: A review of methods based on the application of mixed models, Southampton Staistical Sciences Research Institute Methodology Working Paper M03/16, University of Southampton.

Jiang J. (1996), REML Estimation: Asymptotic Behavior and Related Topics, *Annals of Statistics*, 24, 255–286.

R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

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

Valliant, R., Dorfman, A.H., Royall, R.M. (2000), Finite population sampling and inference. A prediction approach, John Wiley & Sons, New York.

Żądło T. (2004), On unbiasedness of some EBLU predictor. In: *Proceedings in Computational statistics 2004*, Antoch J. (red.), Physica-Verlag, Heidelberg-New York, 2019–2026.

Żądło T. (2009), On MSE of EBLUP, *Statistical Papers,* Springer-Verlag, 50, 101–118.

Tomasz Żądło

O PEWNYCH PRAKTYCZNYCH ASPEKTACH PREDYKCJI ŚREDNIEJ I FRAKCJI W DOMENIE

W opracowaniu jest analizowany problem predykcji frakcji i średniej w domenie z wykorzystaniem modeli nadpopulacji bez zmiennych dodatkowych uwzględniających podział populacji na warstwach. W rozważaniach symulacyjnych uwzględniono problem wpływu złej specyfikacji modelu nadpopulacji i szacowania liczebności populacji na dokładność predykcji.