Estimation of Small Area Means for Subsample Repeated Measurement Data

Erwin Tanur\textsuperscript{1,2}, Anang Kurnia\textsuperscript{2}, Khairil Anwar Notodiputro\textsuperscript{2}, Agus M Soleh\textsuperscript{2}

\textsuperscript{1}Statistics Indonesia (BPS), Jakarta, Indonesia
\textsuperscript{2}Department of Statistics, Bogor Agricultural University, Indonesia

E-mail: anangk@apps.ipb.ac.id

Abstract. The National Socioeconomic Survey (SUSENAS) has been regularly conducted by BPS (Statistics Indonesia) to collect data related to socio-economic conditions of society such as health, education, fertility, housing, and other social and economic conditions. In 2015 important changes were made in the implementation of SUSENAS, such as (1) data collection were conducted two times a year, respectively in March and September and (2) the sample size in March was extremely larger than the sample sizes in September implying that more precise estimates were expected from data collected in March. Moreover, the sampling units enumerated in September were basically a subset of sampling units used in March survey. In another word, the measurement in a particular year was repeated for several sampling units but most of the sampling units was measured only in March. Hence, it is interesting to investigate how small area estimation can improve the precision of estimates resulted from data collected in September by borrowing strength from other areas as well as from previous survey.

Keywords: small area estimation, empirical best prediction, measurement error

1. Introduction

The general survey design challenge is to draw conclusions at the precision level determined at minimal cost, or to achieve the best precision at a fixed cost. Another challenge is the decline in response rates in recent years that may increase the risk of non-responsible bias. A good example has an accurate response to the object under study. When the response in the survey is different from the true value, a measurement error has occurred. The bias of measurement occurs when the response has a tendency to differ from its true value in one direction. Measurement errors and biases should be considered and minimized in the design phase of the survey.

In 1938, Neyman proposed an example design which came to be known as Double Sampling. Hansen and Hurwitz (1946) applied the idea to a non-respondent sub-sample in a special handling of nonresponse cases. Early data collection methods, applied to larger examples, were relatively inexpensive, while follow-up methods used the sub-sample data, although more effective, were also more expensive. For example, the initial data collection model may be by mail, with individual interviews cultivated for sub-samples of no respondent cases.

In 1984 Bellhouse reviewed the optimal design for sampling and his theory provided useful guidelines for designing surveys. There are many ways to "try" different sample sizes before doing a survey. One way is to adjust the model using preliminary data or information from other studies and
compare costs. Another way is to experiment and collect data on relative costs and diversity with different psu (primary sampling units) size. After the sample size is specified and the sub-sample fraction is designed, then look at the number of instances, n. Like other survey designs, the cluster example design is a recurring process: (1) Find the desired precision, (2) select the sample unit and sub-sample size, (3) guess the range to be achieved with the design used, (4) Define n to achieve precision, and (5) iterations (adding additional stratifications and variables for use in ratio estimation) until survey costs are included within the limits in the existing budget.

BPS (Statistics Indonesia) regularly conducts the National Socio-Economic Survey (Susenas), through Susenas collected data relating to socio-economic conditions of the community. Data and indicators from Susenas have been widely used and is seen as one of the most important evidence that can be useful for the planning, monitoring and evaluation of government development programs. For the implementation of Susenas in 2015, there were several important changes compared to the previous, such as enumeration conducted twice annually, each in March and September. The March enumeration was carried out with a large number of samples to produce representative data to the district/city level, while the September enumeration was conducted with a small sample size and produced representative data only at the provincial and national levels. Sample on September is a subsample of Susenas on March selected by using two stages stratified sampling method.

Battese et al. (1988) and Prasad & Rao (1990) used a unit-level nested error linear regression model where the covariates are not subject to measurement errors. Ghosh et al. (2006), henceforth abbreviated as GSK, proposed a nested error linear regression population model with an area-level covariate, x, subject to measurement error. Torabi et al. (2008) derive the fully efficient Bayes (Best) predictor by utilizing all the available data and then estimate the regression and variance component parameters in the model to get an empirical Bayes/Best (EB) predictor and show that the EB predictor is asymptotically optimal. In this paper, will trying to apply Best Prediction method which has been done by Torabi et al. (2008) with the Susenas data 2015, to estimate the average per capita consumption at the district level based on data collected in September (stage 2) using the average information on per capita consumption in March (stage 1) as explanatory variables containing errors measurement.

The measurement error model discussed in this paper is a model that is prepared when the free variable contains measurement error. However, the true value of free variables is unknown. The purpose of measurement error modeling is to obtain the indirect estimator of indirect parameters by matching the y model to the actual x value (unknown) and presumably from the x* value containing the measurement error. Fuller (1987) specifies a measurement error model based on the nature of the unobserved real value. The measurement error model used is additive. The model with variable x as a fixed value, called the functional model. When the model with the random value x is called the structural model. Here is an explanation of both models (Carroll et al., 2006):

a. Functional Model
In this model we have a random variable $X_i^\star$ and $y_i$ with $E(X_i^\star) = x_i$, $E(y_i) = \alpha + \beta x_i$. The functional model is denoted by $y_i = \alpha + \beta x_i + e_i$, $e_i \sim N(0, \sigma_e^2)$, $X_i^\star = x_i + U_i$, $U_i \sim N(0, \sigma_U^2)$ with $x_i$ being fixed but unknown, also $e_i$ and $U_i$ are independent.

b. Structural Model
The structural model is the development of the functional model. As in the functional model we have random variables $X_i^\star$ and $y_i$ with $E(X_i^\star) = x_i$, $E(y_i) = \alpha + \beta x_i$ but in this model, it is assumed that $x_i$ is a random variable. The structural model is denoted by $y_i = \alpha + \beta x_i + e_i$, $e_i \sim N(0, \sigma_e^2)$, $X_i^\star = x_i + U_i$, $U_i \sim N(0, \sigma_U^2)$ and $x_i \sim N(\mu_x, \sigma_x^2)$ with $x_i$ being the random variable and $e_i$, $U_i$ and $x_i$ are independent.

Furthermore, the measurement error model has two types (Carroll et al., 2006): a. Classic Error Model
The statistical model for this case with $X^\star$ is the size of $X$ (measurement error) is $X^\star = X + U$ with $E(U|X) = 0$ and independent with $X$, $Y$, $Z$, $X^\star$ can be observed while $X$ no. Since $E(X^\star|X) = X$ then $X^\star$ is the unbiased measure of $X$.  

2
b. Berkson Error Model

In the case of a Berkson error, X changes around $X^*$ so the statistical model is $X = X^* + U$ with $E(U|X^*) = 0$ and freely with $X^*, Y, Z$. This model is used when the assumption between $U$ and $X$ is mutually unfulfilled. In this model because $E(X|X^*) = X^*$ then $X^*$ is called the unbalanced predictor of Berkson.

In this paper will be used primary data obtained from the Statistics Indonesia (BPS), were obtained from SUSENAS in March and September 2015 in West Java Province. Data on March will be used as an auxiliary variable containing measurement error to estimate average per capita consumption by regency/city with condition on September as the focus of the study. The variables used include monthly per capita household expenditure as a response variable and monthly per capita household expenditure as an explanatory variable with measurement error. The kind of measurement error model, that used in this paper, is structural model with classical error.

2. Methods

Battese et al. (1988) and Prasad -Rao (1990) used a unit-level nested error linear regression model where the covariates are not subject to measurement errors. Ghosh et al. (2006), henceforth abbreviated GSK, proposed a nested error linear regression population model with an area-level covariate, $x$, subject to measurement error. The model is given by:

$$y_{ij} = b + b_1x_i + v_i + e_{ij}, \quad (j = 1, ..., N_i; i = 1, ..., m) \quad (3.1)$$
$$X_{ij} = x_i + \eta_{ij}, \quad (j = 1, ..., N_i; i = 1, ..., m) \quad (3.2)$$

where $N_i$ is the known population size of the $i$th area ($i = 1, ..., m$), $y_{ij}$ is the value of study variable associated with the $j$th unit in the $i$th area and $x_i$ is the unknown true area-specific covariate associated with $y_{ij}$. The random errors $e_{ij}$, measurement error $\eta_{ij}$ and the area-level random effects $v_i$ are assumed to be mutually independent with $e_{ij} \sim iidN(0, \sigma^2_e)$, $\eta_{ij} \sim iidN(0, \sigma^2_\eta)$ and $v_i \sim iidN(0, \sigma^2_v)$. GSK considered the case of structural measurement error model (3.2), where $x_i \sim iidN(\mu_x, \sigma^2_x)$ and are independent of $e^i$, $v^i$ and $\eta^i$ (Fuller, 1987). The vector of model parameter is denoted by $\phi = (b_0, b_1, \mu_x, \sigma^2_v, \sigma^2_\eta, \sigma^2_x)^T$. In Torabi (2008), a sample of size $n_i$ is selected from the $i$th area and the sample data, without loss of generality, is denoted by $(y_{ij}, X_{ij}; j = 1, ..., N_i; i = 1, ..., m)$. Then write $y_i = (y_{i1}, ..., y_{in_i})^T$ as $y_i = (y^{(1)}_i, y^{(2)}_i)^T$ with $y^{(1)}_i = (y_{i1}, ..., y_{in_i})^T$ and $y^{(2)}_i = (y_{in_i+1}, ..., y_{in_i})^T$ also $X_i = (X_{i1}, ..., X_{in_i})^T$. As in GSK, Torabi (2008) assumed that the small area model given by (3.1) and (3.2) holds for the sample data $(y^{(3)}_i, X^{(3)}_i; i = 1, ..., m)$. The goal is to estimate the small area mean $y_i = N_i^{-1} \sum_{j=1}^{n_i} y_{ij}$, $(i = 1, ..., m)$, from the sample data.

GSK obtained the best predictor of $y_i$ under squared error loss using observed data on the response variable only. It is given by:

$$\tilde{y}_i^B = \tilde{y}_i^B(\phi) = E(y_i|y_i) = (1 - f_i D_i) \tilde{y}_i + f_i D_i (b_0 + b_1 \mu_x) \quad (3.3)$$

where $\phi = (b_0, b_1, \mu_x, \sigma^2_v, \sigma^2_\eta, \sigma^2_x)^T$ and $f_i = (N_i - n_i)/N_i$ is the finite population correction factor, and $\tilde{y}_i = n_i^{-1} \sum_{j=1}^{n_i} y_{ij}$, $D_i = \sigma^2_x / (\sigma^2_x + n_i (\sigma^2_v + \sigma^2_\eta))$ $(i = 1, ..., m)$.

$$\text{MSPE}(\tilde{y}_i^B) = V (y_i|y^{(1)}_i, \phi) = \sigma^2_v f_i \left( \frac{1}{N_i} + \frac{f_i (\sigma^2_v + \sigma^2_\eta)}{\sigma^2_x + n_i (\sigma^2_v + \sigma^2_\eta)} \right) \quad (3.4)$$
An empirical best predictor of $y_i$ is obtained as $\hat{\gamma}^{EB}_i = \hat{\gamma}^{\mathcal{B}}(\hat{\phi})$, where $\hat{\phi}$ is a method of moments estimator of $\phi$. GSK also established the asymptotic optimality of $\{\hat{\gamma}^{EB}_i\}$ by showing that $m^{-1} \sum_{i=1}^{m} E(\hat{\gamma}^{EB}_i - \gamma_i)^2 \rightarrow 0$ as $m \rightarrow \infty$.

The Best predictor (Torabi, 2008) of $y_i$ is given by:

$$\hat{\gamma}^{\mathcal{B}} = (1 - f_i A_i) \hat{y}_i + f_i A_i (b_0 + b_1 \mu_x) + f_i A_i \left( \frac{n_i \sigma_x^2}{\sigma_0^2 + n_i \sigma_x^2} \right) b_1 (\hat{X}_i - \mu_x) \quad (3.5)$$

$$V(\hat{\gamma}^{EB}_i) = f_i^2 A_i \left( \frac{n_i \sigma_x^2}{\sigma_0^2 + n_i \sigma_x^2} \right)^2 + \frac{1}{N_i} f_i \sigma^2_0 \quad (3.6)$$

$$\text{MSPE}(\hat{\gamma}^{\mathcal{B}}) = E(\hat{\gamma}^{EB}_i - \gamma_i)^2 = B_{11}(\delta) = V(\hat{\gamma}^{EB}_i, \mathbf{X}^{(1)}_i, \phi) \quad (3.7)$$

where $A_i = \frac{\sigma_0^2}{n_i \sigma_x^2 + (n_i \sigma_x^2 + \sigma_0^2)(\sigma_0^2 + n_i \sigma_x^2)}$.

The EB predictor (Torabi, 2008), $\hat{\gamma}^{EB}_i$, of $y_i$ is given by:

$$\hat{\gamma}^{EB}_i = (1 - f_i A_i) \hat{y}_i + f_i A_i (\hat{b}_0 + \hat{b}_1 \hat{X}_i) + f_i A_i \left( \frac{n_i \sigma_x^2}{\sigma_0^2 + n_i \sigma_x^2} \right) b_1 (\hat{X}_i - \hat{X}) \quad (3.8)$$

$$\hat{\delta}_i = [(\text{MSB}_X - \text{MSW}_X)(m - 1)]^{-1} \sum_i n_i (\hat{y}_i - \hat{X}_i) \quad (3.9)$$

$$\hat{\delta} = n^{-1} \sum_i n_i \hat{\delta}_i, \quad \hat{b}_0 = \hat{y} - \hat{\delta}_i \hat{X}$$

$$\text{MSB}_X = \frac{\text{SSW}_X}{n_{y,m}} = \frac{\sum_i (x_{i} - \hat{X})^2}{\sum_i n_i - m}, \quad \hat{\sigma}_0^2 = \text{MSW}_y = \frac{\text{SSW}_y}{n_{y,m}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i n_i - m}$$

$$\hat{\delta}_i = \text{max}(0, (\text{MSB}_X - \text{MSW}_X)(m - 1)/g_m), \quad \hat{\delta}_0 = \text{max}(0, g_m (\text{MSB}_y - \text{MSW}_y) - \hat{b}_0^2 \hat{\sigma}_0^2)$$

$$\text{MSB}_y = (m - 1)^{-1} \sum_i n_i (\hat{y}_i - \hat{y})^2, \quad g_m = n_{y,m} - \sum_i n_i^2 / n_T$$

Torabi (2008) obtain a nearly unbiased estimator of $\text{MSPE}(\hat{\gamma}^{EB}_i) = E(\hat{\gamma}^{EB}_i - \gamma_i)^2$, using the jackknife methods proposed by Jiang et al. (2002) and Chen & Lahiri (2002). The following orthogonal decomposition is $\text{MSPE}(\hat{\gamma}^{EB}_i) = E(\hat{\gamma}^{EB}_i - \gamma_i)^2 + E(\hat{\gamma}^{EB}_i - \hat{\gamma}^{\mathcal{B}}_i)^2 = M_{1i} + M_{2i}$, where $M_{1i} = g_i(\hat{\delta})$ is given by (7). A plug-in estimator of $g_i(\hat{\delta})$ is $g_i(\hat{\delta})$. Let $\hat{\phi}_{-i}$ be the estimator of $\phi$ obtained by deleting the $i$th area data set $(\mathbf{y}^{(1)}_i, \mathbf{X}^{(1)}_i)$ from the full data set and then applying the method-of-moments. This calculation is done for each $i$ in turn to get $m$ estimators. A weighted jackknife estimator of $M_{1i}$ is given by:

$$\hat{M}_{1ijw} = g_i(\hat{\delta}) - \sum_{i=1}^{m} \hat{w}_i (g_i(\hat{\delta}_{-i}) - g_i(\hat{\delta})) \quad (9)$$

$$\hat{M}_{1ijw} = 1 - \hat{X}_i^T (\hat{\Sigma}_{i=1}^{m} \hat{X}_i - \hat{X}_i^T)^{-1} \hat{X}_i, \quad M_{1} = (1, \hat{X}_i)^T$$

An unweighted jackknife estimator of $M_{1i}$, denoted by $\hat{M}_{1ij}$, is obtained by letting $\hat{w}_i = (m - 1)/m$ in (9); (Jiang et al. 2002). Turning to jackknife estimation of the last term, $M_{2i}$, let $\hat{\gamma}^{EB}_i = k_i (\mathbf{y}^{(1)}_i, \mathbf{X}^{(1)}_i, \phi)$ be the Best predictor expressed as a function of $\mathbf{y}^{(1)}_i, \mathbf{X}^{(1)}_i$ and $\phi$, so that $\hat{\gamma}^{EB}_i = k_i (\mathbf{y}^{(1)}_i, \mathbf{X}^{(1)}_i, \phi)$. Now replace $\phi$ by $\hat{\phi}_{-i}$ to get $\hat{\gamma}^{EB}_{i,-l} = k_i (\mathbf{y}^{(1)}_i, \mathbf{X}^{(1)}_i, \hat{\phi}_{-l})$.\ 

An unweighted jackknife estimator of $M_{2i}$ is then given by $\hat{M}_{2ijw} = \sum_{i=1}^{m} \hat{w}_i (\hat{\gamma}^{EB}_{i,-l} - \hat{\gamma}^{EB}_i)^2$. An unweighted jackknife estimator of $M_{2i}$, denoted by $\hat{M}_{2ij}$, is obtained by letting $\hat{w}_i = (m - 1)/m$. A weighted and unweighted jackknife estimator of $\text{MSPE}(\hat{\gamma}^{EB}_i)$ is obtained as:

$$\text{mspe}_{jw}(\hat{\gamma}^{EB}_i) = \hat{M}_{1ijw} + \hat{M}_{2ijw} \quad \text{and} \quad \text{mspe}_{j}(\hat{\gamma}^{EB}_i) = \hat{M}_{1ij} + \hat{M}_{2ij} \quad (3.9)$$
3. Results and Discussion

Figure 1 reports that the value of MSPE-GSK ($\tilde{y}_i^G$), based on (3.4) and MSPE-EB, based on (3.7). From Figure 1 shows that MSPE-EB value in each area is substantially smaller than MSPE-GSK value. So from these results illustrate that the estimation using EB method is better than GSK method.

Table 1 shows the estimation of average per capita consumption by regency/city in Provinsi Jawa Barat. It shows that estimates using the EB method with Susenas data collection in September, previously representative only to estimate up to provincial/national level, taking into account measurement errors can provide estimates that approximate direct predictors from March to district level. Although the estimation results are still not as expected but this has provided an idea that a better estimate result will be obtained, one of which is how to select / determine the information direct estimator that has the element of measurement error as one of the elements that determine the value of the estimator. Also in this paper while using only 1 auxiliary variables that have measurement errors, for further research will be done by adding explanatory variable that is fixed. It will be seen whether it will have a more positive impact on the model or vice versa.
Table 1. Direct Estimation, GSK, EB and Naive EB Predictor for Mean Consumption

| Area | Number of Population | Number of Samples | Direct Estimation (Susenas on March) | GSK         | EB          | Naive EB     |
|------|----------------------|-------------------|--------------------------------------|-------------|-------------|--------------|
| 1    | 40                   | 23                | 986,480                              | 1,270,533   | 1,171,346   | 1,278,145    |
| 2    | 47                   | 18                | 717,212                              | 909,704     | 836,119     | 905,392      |
| 3    | 32                   | 17                | 573,118                              | 645,821     | 643,040     | 631,655      |
| 4    | 31                   | 20                | 867,852                              | 927,194     | 935,275     | 925,476      |
| 5    | 42                   | 19                | 518,647                              | 681,469     | 631,923     | 668,014      |
| 6    | 39                   | 20                | 492,494                              | 543,307     | 546,830     | 526,963      |
| 7    | 26                   | 17                | 588,534                              | 727,280     | 692,530     | 719,338      |
| 8    | 32                   | 16                | 697,998                              | 841,407     | 794,455     | 834,378      |
| 9    | 40                   | 22                | 642,430                              | 675,524     | 709,374     | 665,841      |
| 10   | 26                   | 17                | 703,357                              | 882,374     | 912,878     | 879,096      |
| 11   | 32                   | 14                | 795,177                              | 1,019,935   | 939,577     | 1,021,509    |
| 12   | 31                   | 20                | 630,340                              | 755,749     | 744,113     | 749,524      |
| 13   | 30                   | 16                | 891,444                              | 850,053     | 899,185     | 843,927      |
| 14   | 17                   | 12                | 1,043,068                            | 879,600     | 969,557     | 875,717      |
| 15   | 30                   | 20                | 853,489                              | 896,085     | 907,958     | 893,711      |
| 16   | 23                   | 15                | 1,283,377                            | 1,450,665   | 1,435,881   | 1,466,083    |
| 17   | 16                   | 13                | 631,329                              | 632,239     | 637,464     | 624,892      |
| 18   | 10                   | 8                 | 802,512                              | 720,383     | 739,228     | 711,257      |
| 19   | 6                    | 5                 | 1,436,318                            | 1,888,828   | 1,879,889   | 1,925,601    |
| 20   | 7                    | 7                 | 1,040,490                            | 1,175,989   | 1,175,989   | 1,176,010    |
| 21   | 30                   | 18                | 1,521,069                            | 1,841,177   | 1,939,088   | 1,869,039    |
| 22   | 5                    | 5                 | 892,487                              | 1,092,585   | 1,092,585   | 1,092,600    |
| 23   | 12                   | 12                | 1,507,285                            | 1,760,403   | 1,760,403   | 1,760,404    |
| 24   | 11                   | 9                 | 1,594,316                            | 2,297,999   | 2,249,553   | 2,333,767    |
| 25   | 3                    | 3                 | 1,188,187                            | 1,411,049   | 1,411,049   | 1,411,054    |
| 26   | 10                   | 10                | 1,077,540                            | 992,032     | 992,032     | 992,077      |
| 27   | 4                    | 4                 | 800,758                              | 934,931     | 934,931     | 934,948      |

4. References

[1] Battese, G. E., Harter, R. M. & Fuller, W. A. (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. *J. Amer. Statist. Assoc.* 83, 28-36

[2] BPS. 2015. Buku 1 Susenas Maret 2015. Pedoman Kepala BPS Provinsi, Kepala Bidang Statistik Sosial dan Kepala BPS Kabupaten/Kota. Jakarta (ID).

[3] BPS. 2015. Buku 1 Susenas September 2015. Pedoman Kepala BPS Provinsi, Kepala Bidang Statistik Sosial dan Kepala BPS Kabupaten/Kota. Jakarta (ID).

[4] Carroll, R.J, Ruppert D, Stefanski LA. (2006). Measurement Error in Nonlinear Models: A Modern Perspective (Second Edition. Boca Raton: Chapman and Hall.)
[5] Chen, S. & Lahiri, P. (2002). On Mean Squared Prediction Error Estimation in Small Area Estimation Problems. In Proceedings of The Survey Research Methods Section, 473,477. American Statistical Association, Washington

[6] Fuller, W. (1987). Measurement Error Models. Wiley, New York, NY.

[7] Ghosh, M. Sinha, K. & Kim, D. (2006). Empirical and Hierarchical Bayesian Estimation in Finite Population Sampling Under Structural Measurement Error Models. Scand. J. Statist. 33, 591-608.

[8] Jiang, J., Lahiri, P. (2006). Mixed Model Prediction and Small Area Estimation. Test 15, 1-96.

[9] Prasad, N. G. N. & Rao, J. N. K. (1990). The Estimation of Mean Squared Error of Small Area Estimators. J. Amer. Statist. Assoc. 85, 163-171.

[10] Torabi M, Datta GS, Rao JNK. (2009). Empirical Bayes Estimation of Small Area Means under a Nested Error Linear Regression Model with Measurement Errors in the Covariates. Scandinavian Journal of Statistics doi: 10.1111/j.1467-9469.2008.00623.x