The estimated proportion of chronic disease sufferer in Duren Sawit district, East Jakarta, using hierarchical bayes method in small area estimation (SAE)

I Fajar*, T Siswantining and S M Soemartojo

Department of Mathematics, Universitas Indonesia, Indonesia

*Corresponding author: ibrahimfajar173@gmail.com

Abstract. The purpose of this research is to observe the proportion of chronic disease sufferer in the Duren Sawit district, East Jakarta. The data that used in this research is the primary data in the form of survey data directly and the secondary data in the form of census data from Dinas Kesehatan (Dinkes) 2017 and Badan Pusat Statistik (BPS) 2017. The sampling method used is simple random sampling with a sample size of 1% of the total heads of families living in the Duren Sawit district, that is 1229 the heads of the family. On direct estimation, estimating parameter only based on survey data of subpopulations is inappropriate action, because the sample size obtained relatively few or there are subpopulations that is not selected as the sample.

To overcome this, indirect estimation is used with small area estimation (SAE) method, which borrowed extra information such as administrative data or census data from other areas or area itself and there’s an addition random area effect into the model. In this research, the proportion of chronic disease sufferer in Duren Sawit district is obtained used direct estimation and indirect estimation with hierarchical bayes method in SAE. The results of the estimates proportion from direct estimation is 18.87% with Mean Square Error (MSE) is 0.000130741, whereas the estimates of the proportion from indirect estimation is 18.37% with MSE 0.0000922915. Based on the MSE that obtained, the estimate proportion of chronic disease sufferer from indirect estimation is more reliable than direct estimation.

1. Introduction

Chronic disease causing inability on its victims, and to heal the disease sufferers need to do treatments in a long period of time [1]. Chronic disease is a disease that takes up most of the victim in the world which the chronic diseases that cause the most of death is cardiovascular disease, chronic respiratory disease, diabetes mellitus and cancer [2]. Cardiovascular disease in general consist heart failure, coronary heart disease, and stroke [3], whereas chronic respiratory disease in general consists of asthma and Chronic Obstruk pulmonary disease (COPD) [4]. For Diabetes mellitus (DM). There are two primary forms of diabetes, insulin-dependent DM and non-insulin-dependent DM [5]. According to the Departemen Kesehatan (2011), there is a 59.5% of all deaths in Indonesia are caused by chronic disease. It indicates that it needs a special attention to control a chronic disease. Availability of complete, accurate, and very routine data is really needed to assist the Government in making policies that are effective in controlling chronic diseases.

Survey data are commonly used to estimate the parameters of a population in a large scope like a state, province, or county [6]. In this paper, the scope is in the form of district. Estimating the parameters required a sufficient sample size in order to represent the population, so the valuation of the parameters
can be made directly on the basis of survey data that obtained. The last few years, the need for the estimation of parameters on a smaller scope (subpopulations) is increased. It is causing a reduced survey data that are available for the population. Using the survey data from subpopulation to estimate population parameters is not an appropriate action because the sample size that obtained is relatively little and/or there are some unselected subpopulations as a sample, so the direct estimation can't be done. To resolve the issue, the estimation is done indirectly by small area estimation (SAE) method.

2. Small Area Estimation (SAE)

Small area estimation is a method for estimating parameters on small areas with added information from other areas or from that area itself like an earlier survey [7]. Small area in this paper is an area where a survei data that are owned is not big enough to do direct estimation with a good precision [7]. Based on the availability of the additional information (auxilliary variable), the small area models are classified into 2 model, namely: (i) area level model, where additional information is available only for area level. (ii) unit level model, where additional information is available to the unit level [7].

Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes (EB), and the Hierarchical Bayes (HB) was the usual method that used on SAE. EBLUP is designed for a continue variable, while the EB or HB method is suitable for binary, count, or catorgical data [8]. HB method has some advantages compared to the EB method. From the posterior distribution, inference of HB can be directly used for the whole of the inference. The other advantage is HB able to resolve a complex model such as the unmatched sampling and linking model where it is diffic ult to be handled by EB [9]. Based on the primacy of HB described earlier and the owned survey data is count, then HB method is used in this research to find the proportion of chronic diseases sufferer in Duren Sawit district. Several studies using HB have already been done before by ([8],[10],[11]), while for further knowledge related SAE, readers are advised to read [12].

In HB, the observation y is given from a conditional distribution against the parameter 0, where 0 is given from a conditional distribution against an additional parameter called λ which is known as the hyperparameter [10]. The prior distribution of R (λ) of the model parameters λ is set in advance and the posterior distribution of R (θ/y) with θ is the parameter of interest, given the data y will be obtained.

Some model that can be used for count variable response data is Poisson-Gamma and Log-Normal [12]. Suppose \( y_i \) is parameter interest that we want to search, it is proportion of chronic disease sufferer in area-i. On direct estimation, the estimate proportion of chronic disease sufferer is obtained as follows:

\[
\theta_i = \frac{n_i}{n_i} \tag{1}
\]

where \( y_i \) is the number of people who suffered with chronic diseases in the area-i and \( n_i \) is the number of samples in the area-i. For indirect estimation by using log-normal model, a model for HB approach gained as follows [12]:

1. \( y_i | \theta_i \sim \text{Poisson}(n_i \theta_i) \)
2. \( \xi_i = \log(\theta_i) | \beta, \sigma^2 \sim \text{iid}\{N(x_i^T \beta, \sigma^2) \} \)
3. \( f(\beta, \sigma^2) \propto f(\beta) f(\sigma^2) \) dengan \( f(\beta) \propto 1; \sigma^2 \sim \text{gamma}(a, b); a \geq 0, b > 0 \)

The estimate parameters of interest will be obtained by finding the mean values of posterior [12]. However, this is difficult to obtained because it involves a complicated calculation in analytic, so that will do the expectation value calculation in numerically with the help of the method Markov Chain Monte Carlo (MCMC). MCMC method is commonly used when is diffic ult to generate samples \( \theta \) directly from the posterior distribution of \( R(\theta | y) \). Instead, the sample \( \theta \) is generated in iterative such that each sample withdrawal is expected to come from the posterior distribution that approximates the posterior distribution of \( R(\theta | y) \) [13]. At MCMC, algorithm that commonly used is a Gibbs Sampling and Metropolis-Hasting. To further knowledge about MCMC, readers are advised to read ([13],[14],[15]). Based on the model of HB at equation (2), full gibbs conditional posterior distribution for each parameter is given as follows [12]:

![](image)
1. \[ f(\theta_i|\beta, \sigma^2, y) \propto \theta_i^{y_i-1} \exp \left[ -\frac{1}{2\sigma^2} (\xi_i - x_i^T \beta)^2 \right] \]

2. \[ [\beta, \theta, \sigma^2, y] \sim N \left( \left( \sum x_i^2 x_i \right)^{-1} (\sum x_i \xi_i), \sigma^2 \left( \sum x_i x_i^T \right)^{-1} \right) \] (3)

3. \[ [\sigma^2, \theta, \beta, y] \sim G \left[ \frac{m}{\lambda} + a, \frac{1}{\lambda} \sum (\xi_i - x_i^T \beta)^2 + b \right] \]

MCMC sample for \( \beta \) and \( \sigma^2 \) can be generated directly from point 2 and 3 in equation (3) using gibbs sampling algorithm because it has a clear form of distribution that is Multivariate Normal and Gamma.

To see which is the better estimates, compared the value of Mean Squared Error (MSE) from direct estimation and indirect estimation. The smaller value of MSE is a better parameter estimation. In this paper, the value of MSE can be obtained as follows:

\[ [13]: \]

Here is a gibbs sampling algorithm to find the parameter model \( \beta \) and \( \sigma^2 \) : [12]:

1. Take an arbitrary initial value \( \theta^{(0)}(\beta), \sigma^{(0)}(\beta) \)
2. Generate \( \beta^{(1)}(\beta), \sigma^{(2)}(\beta) \) with information \( \theta^{(0)}, \sigma^{(0)} \) from equation (3) point 2
3. Generate \( \sigma^{(2)}(\beta) \) with information \( \beta^{(1)}, \theta^{(0)} \) from equation (3) point 3
4. When in k-iteration, generate sample random \( \beta^{(k)} \) with information \( \theta^{(k-1)}, \sigma^{(k-1)} \), \( y \) and generate sample random \( \sigma^{(2)}(k) \) with information \( \beta^{(k)}, \theta^{(k-1)} \), \( y \) \( k = 1, 2, \ldots, t + T \)
5. Generated \( \beta \) and \( \sigma^2 \) as much as k sample
6. Do the “burn in” by way of disposing the first t iteration to remove the effect of initial value, so obtained \( \left[ \beta_1^{(k)}, \ldots, \beta_m^{(k)} \right], \sigma^{(2)}(k); k = t + 1, \ldots, t + T \).

in M-H algorithm, it takes a candidate density for generating a candidate value \( \theta^*_1 \). Equation (3) point 1 can be modified as follows:

\[ f(\theta_i|\beta, \sigma^2) \propto \theta_i^{y_i} \exp \left[ (-\frac{(\xi_i - x_i^T \beta)^2}{2\sigma^2}) \right] \]

with \( g'(\theta_i) = \frac{\partial g(\theta_i)}{\partial (\theta_i)} \) and \( g(\theta_i) = log(\theta_i) \). Here is a M-H algorithm to find the parameter interest \( \theta|_{[13]} \):

1. Generated a candidate value \( \theta^* \) from \( h(\theta|\beta, \sigma^2) \)
2. Calculate and acceptance probability \( a(\theta^*, \theta(k)) = \min \left( \frac{\theta^*}{\theta(k)} \right) \)
3. Generate \( u \sim U(0, 1) \)
4. If \( u \leq a(\theta^*, \theta(k)) \), \( \theta^{(k+1)} = \theta^* \), and \( \theta^{(k+1)} = \theta(k) \) for the other.

After running the MCMC method, it will obtained a value of \( \beta_1^{(k)}, \ldots, \beta_m^{(k)} \), \( \sigma^{2(k)}, k = t + 1, \ldots, t + T \). The proportion of chronic disease sufferer is estimated from posterior mean \( E[\theta_i|y] \) as follows:

\[ E[\theta_i|y] = \hat{\theta}_i^{(HB)} \approx \frac{1}{T} \sum_{k=t+1}^{t+T} \theta_i^{(k)} \]

To see which is the better estimates, compared the value of Mean Squared Error (MSE) from direct estimation and indirect estimation. The smaller value of MSE is a better parameter estimation. In this paper, the value of MSE can be obtained as follows:
\[
MSE = \frac{\sum_{i=1}^{n} (\hat{\theta}_i - \theta)^2}{n}
\]

where \(\theta\) is estimation the proportion of chronic diseases sufferer for overall Duren Sawit district.

3. Application of Hierarchical Bayes SAE

In this paper, HB SAE is used to find the estimates proportion of chronic disease sufferer in Duren Sawit district, East Jakarta. The observe \(y_i\) is the number of people who suffer a chronic disease in \(i\)-subdistricts and \(n_i\) is the number of sample in \(i\)-subdistricts where \(i = 1,2,..,7\). In this paper, the parameter that want to be obtained is the proportion of chronic disease sufferer in Duren Sawit district which is estimated directly and indirectly.

Data of the number of chronic disease sufferer in each subdistricts is obtained by conducting the interview directly, while additional data (auxiliary variable) is obtained from the Department of health and the Central Bureau of statistics of the year 2017. Based on the availability of data, the independent variables that are used is:

| No | Independent Variables | Operational Definition |
|----|-----------------------|------------------------|
| 1  | Age                   | According to WHO, the increasing of age can increase the risk of chronic disease. According to Riskesdas 2013, the prevalence of chronic diseases sufferer such as cardiovascular disease, COPD, diabetes, and cancer is increase when entered at the age of 45 years and over. Therefore, in this paper the variable of age is a percentage of the number of people aged 45 years and over in each subdistricts at Duren Sawit. |
| 2  | The intensity of sport| In this study, the intensity of sports represented by the number of sports facilities in each subdistricts at Duren Sawit. According to WHO, the lack of exercise can increase the risk of chronic diseases. WHO also advised to exercise 30 minutes a day. Physical activities contribute to reduces the risk of DM type 2 around 30-50% [18]. |
| 3  | Total visits of Pos Pembinaan Terpadu (Posbindu) | Represented as the number of people who visited the Posbindu in one year. The activities of prevention and early detection can be implemented through community empowerment in Posbindu [5]. |

The independent variables that are mentioned in Table 1 are used as auxiliary variable on indirect estimation. By using excel for direct estimation and program R for indirect estimation, the estimates proportion of chronic disease sufferer is obtained as shown in Table 2.

According Table 2, the estimates value \(\hat{\theta}_i\) on direct estimation is greater than the estimates with indirect estimation for all subdistricts and the MSE values on indirect estimation are smaller compared to the value of the MSE on direct estimation. This indicate that the proportion of chronic diseases sufferer with indirect estimation is better than the estimates of proportion with direct estimation.
Table 2. Comparison the estimates proportion of chronic disease sufferer with direct estimation and indirect estimation

| Subdistricts   | Estimates of $\hat{\theta}_i$ for each subdistricts | The estimates proportion of chronic diseases sufferer for overall Duren Sawit district | MSE       |
|----------------|---------------------------------------------------|--------------------------------------------------------------------------------------|-----------|
|                | $\hat{\theta}_{iDE}$ | $\hat{\theta}_{iHB}$ | $\hat{\theta}_{DE}$ | $\hat{\theta}_{HB}$ | $\bar{\theta}_{DE}$ | $\bar{\theta}_{HB}$ |
| Duren Sawit    | 0.1961722 | 0.1916978 |                                      |          |
| Pondok Bambu   | 0.1745283 | 0.1708422 |                                      |          |
| Klender         | 0.1897233 | 0.1867422 |                                      |          |
| Malaka Jaya     | 0.2086957 | 0.1998860 | 0.1887180 | 0.1837  | 0.000130741 | 0.0000922915 |
| Malaka Sari     | 0.1980198 | 0.1887180 |                                      |          |
| Pondok Kopi     | 0.1946903 | 0.1865123 |                                      |          |
| Pondok Kelapa   | 0.1769912 | 0.1730971 |                                      |          |

4. Conclusion
Upon the results of MSE indirect estimation that obtained is smaller than MSE direct estimation on Table 2, the best estimates proportion of chronic disease sufferer in Duren Sawit district is generated from indirect estimation where in this paper using the HB SAE method. The results of the estimates that obtained is 0.1837 with MSE is 0.0000922915, which means that around 18.37% from the number of population that living in Duren Sawit district suffer a chronic diseases.

5. Acknowledgments
This research was supported by Pitta UI 2018 Research Grant.

References
[1] Bestari B K and W D N Kusuma 2016 J. Keperawatan Indonesia 19 49
[2] World Health Organization (WHO) 2017 Noncommunicable Disease Progress Monitor 2017
[3] Wong N D, Zhao Y, Patel R, Patao C, Malik S, Bertoni A G, Correa A, Folsom A R, Kachroo S, Mukerjee J, Taylor H and Selvin E 2016 Diabetes Care 39(5) 668
[4] Viswanathan K, Rakesh P S, Balakrishnan S, Shanavas A and Dharman V 2017 Indian J. Tuberc. 65(1) 48
[5] Yanling W, Yanping D, Yoshimasa T and Wen Z 2014 Int. J. Med. Sci. 11 1185
[6] Handayani D, Folmer H, Kurnia A and Notodiputro K A 2018 J. Inst. Adv. Stud. 55 146
[7] Ramadhan S, Siswanti T and Soemartojo S M 2017 J. Phys.: Conf. Ser. 909 012043
[8] Siswanti T 2013 IndoMS J. Stat. 1 7
[9] Wulansari I Y, Pawitan G, Sunengsih N 2014 Conference: Seminar Nasional IV Universitas Padjajaran (Universitas Padjajaran: Bandung)
[10] Kubacki J 2012 Stat. Transmition 13 261
[11] Trevisani M and Torrelli N 2017 J. Stat. 7 521
[12] Rao JNK. 2003 Small Area Estimation (New York: John Wiley and Sons)
[13] Wang H, Wang C, Wang Y, Gao X and Yu C 2017 J. Hydrol. 549 476
[14] Ottobre M 2016 Rep. Math. Physics 77 267
[15] Brooks S P 1998 Statistician 47 69