Bayesian Hierarchical Random Intercept Model Based on Three Parameter Gamma Distribution

Ika Wirawati, Nur Iriawan*, and Irhamah
Departement of Statistics, Institut Teknologi Sepuluh Nopember Surabaya, Indonesia
E-mail: wirawati_ika15@mhs.statistika.its.ac.id, nur_i@statistika.its.ac.id, irhamah@statistika.its.ac.id

ABSTRACT. Hierarchical data structures are common throughout many areas of research. Beforehand, the existence of this type of data was less noticed in the analysis. The appropriate statistical analysis to handle this type of data is the hierarchical linear model (HLM). This article will focus only on random intercept model (RIM), as a subclass of HLM. This model assumes that the intercept of models in the lowest level are varied among those models, and their slopes are fixed. The differences of intercepts were suspected affected by some variables in the upper level. These intercepts, therefore, are regressed against those upper level variables as predictors. The purpose of this paper would demonstrate a proven work of the proposed two level RIM of the modeling on per capita household expenditure in Maluku Utara, which has five characteristics in the first level and three characteristics of districts/cities in the second level. The per capita household expenditure data in the first level were captured by the three parameters Gamma distribution. The model, therefore, would be more complex due to interaction of many parameters for representing the hierarchical structure and distribution pattern of the data. To simplify the estimation processes of parameters, the computational Bayesian method couple with Markov Chain Monte Carlo (MCMC) algorithm and its Gibbs Sampling are employed.

1. Introduction

Often encountered in many researches that the data used has a hierarchical structure. For example, in a social research, human populations are structured by hierarchical groupings. Whereby, individuals are grouped within households, households are grouped within community, and at the top level could be a district/city, province or country [9]. Grouping on hierarchical data is built on the principle of the similarities between members of a group, so that the members in the group have similar characteristics [8]. While on the other hand, among the members from a group to the other group members are different, so-called there is variation between groups. Therefore, the hierarchical structure of the data cannot be ignored in the analysis process.

Beforehand, the existence of a hierarchical structure in data was less noticed in the analysis. Conventional statistical analysis will not provide the proper inference and will not be able to capture the real phenomenon if they ignore the existence of this data type [13]. When the observation data has a hierarchical structure, then the appropriate analysis is hierarchical or multilevel models. The hierarchical model has two main advantages [12]. First, it can be used to analyze at several different levels simultaneously in a single statistical analysis. Secondly, it considers the variance at each level of the response variance. In addition, a study had proven that the hierarchical model was better than
the classical methods (unilevel model) [15]. This study was performed using hierarchical data simulation.

The hierarchical model was developed for the data analysis involving two or more levels of correlation among variables and parameters. Thus the hierarchical model can be used to examine the relationship between the measured variables at different levels in the hierarchical data structure [12]. Basically, the hierarchical model constructed by two sub-models, micro and macro models. In the two-level hierarchical model, micro models are the regression model between the observed response and predictor in the first/lowest level. Meanwhile macro models are the model that show the relationship between the regression coefficients from micro models to the predictor variables on the second/upper level [13].

This article will only focus on random intercept model (RIM), as a subclass of the hierarchical model. This model assumes that the intercept of models in the lowest level are varied among those models, and their slopes are fixed. The differences of intercepts were suspected affected by some variables in the upper level. These intercepts, therefore, are regressed against those upper level variables as predictors. This paper would demonstrate the work of this two level RIM on the household expenditure in Maluku Utara, which has five characteristics in the first level and three characteristics of districts/cities in the second level.

Distribution of household expenditure has a unique characteristic, which it has a positive value and a long right tail (skewed distribution). Some distribution that represents these properties well are log-normal distribution, log-logistic distribution, gamma distribution, etc. Several previous studies have used these distributions to explain the pattern of household expenditure data, for example [1], [11], [19], and [14]. The data in this study were captured by the three parameters Gamma (Gamma3) distribution. The household expenditure in nine of ten districts/cities in Maluku Utara shown to have the Gamma3 distribution.

The model, therefore, would be more complex due to interaction of many parameters for representing the hierarchical structure and distribution pattern of the data. For the complex hierarchical model, parameter estimation using classical approach become very difficult to be done. Bayesian modeling approach on this issue going to provide a better solution. Bayesian method is very flexible and easy to estimate the parameters of a complex hierarchical models [17]. To simplify the estimation processes of parameters, the computational Bayesian method couple with Markov Chain Monte Carlo (MCMC) algorithm and its Gibbs Sampling are employed.

2. Three Parameter Gamma (Gamma3) Distribution
In probability theory and statistics, the Gamma3 distribution is a three parameter family of continuous probability distributions. Consider that \( Y_1, Y_2, \ldots, Y_n \) is a random sample from a population with the Gamma3 distribution. The probability distribution function (PDF) can be written as follows [16]:

\[
f(y|\alpha, \phi, \lambda) = \frac{\alpha^\phi (y-\lambda)^{1-\alpha}}{\Gamma(\alpha)} \exp[-\phi(y-\lambda)], \alpha > 0, \phi > 0, y > \lambda,
\]

where \( \alpha \) is the shape parameter, \( \phi \) is the scale parameter of the distribution. Meanwhile \( \lambda \) is the location or threshold parameter. If \( \lambda = 0 \), then \( Y \) would be a gamma distribution with two parameters, are denoted by Gamma(\( \alpha, \phi \)). The mean and variance of the Gamma3 distribution is [2]:

\[
E(Y) = \lambda + \frac{\alpha}{\phi}
\]

\[
V(Y) = \frac{\alpha}{\phi^2}.
\]

Gamma distribution has been widely used in many fields, such in the science of Physics, Meteorology, Ecology, and other fields. Even in 1974, Salem and Mount had showed that the gamma
distribution is more appropriate than the log-normal distribution to illustrate the income data of people in the United States in 1960 to 1969 [16].

3. Random Intercept Model

Basically, hierarchical model is formed by two sub-models, it is micro models (model at the lowest level of the hierarchical structure) and macro models (model at the upper level of the hierarchical structure) [12]. In principle, the depth level of the models hierarchy is not limited, however, for further elaboration is used two-level hierarchical model. The micro models illustrate association between household expenditure as the response and household characteristics as predictors in the first level. While the macro models represent the relationship between coefficients in the micro models with district characteristics as the predictors in the second level.

The two-level hierarchical model can be described as follows [20] and [8]:

1. Micro Model

Let \( N \) is number of household in Maluku Utara Province, \( n_j \) is number of household in \( j^{th} \) district, \( m \) is number of districts/cities. The model at the first level of each district/city can be expressed as below:

\[
Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \cdots + \beta_{kj}X_{kij} + e_{ij}, \quad i = 1,2,\ldots,n_j \text{ dan } j = 1,2,\ldots,m, \tag{2}
\]

or it can also be written in vector as,

\[
y_j = X_j\beta_j + e_j, \tag{3}
\]

where,

\[
y_j = (y_{1j} \quad y_{2j} \quad \cdots \quad y_{nj})^T,
\]

\[
X_j = \begin{pmatrix} 1 & X_{11j} & X_{21j} & \cdots & X_{k1j} \\ 1 & X_{12j} & X_{22j} & \cdots & X_{k2j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1nj} & X_{2nj} & \cdots & X_{knj} \end{pmatrix},
\]

\[
\beta_j = (\beta_{0j} \quad \beta_{1j} \quad \cdots \quad \beta_{kj})^T,
\]

\[
e_j = (e_{1j} \quad e_{2j} \quad \cdots \quad e_{nj})^T.
\]

\( X \) is predictors matrix indicated a household characteristics, \( \beta \) is regression coefficients vector of the micro models and \( e \) is residual vector of the micro models with assumption \( e_j \sim N(0, \sigma_j^2 I_{n_j}) \).

2. Macro Model

Regression coefficients of the micro model have varying values between districts/cities. Therefore, these variations will be explained by macro model. Modelling conducted for each regression coefficient in micro models using the predictors at the second level (macro predictors). These predictors represented district characteristic, noted by \( W \). The macro model is described in this following equations,

\[
\beta_{rj} = \gamma_{0r} + \sum_{l=1}^{q} \gamma_{lr}W_{lj} + u_{rj}, \quad r = 0,1,2,\ldots,k \tag{4}
\]

or it can also be written in vector as,

\[
\beta_r = W\gamma_r + u_r, \tag{5}
\]

where,
\[ \mathbf{y}_j = \mathbf{X}_j \mathbf{W} \mathbf{y} + \mathbf{X}_j \mathbf{u}_j + \mathbf{e}_j, \] (6)

Hierarchical model is built by regress all regression coefficients in micro models with predictors in macro model. When the hierarchical model only use intercept in micro models for regressed at the upper level, this model is known as random intercept model (RIM). The micro models of RIM are same with equation (2) and (3), rewritten for more simply as follow:

\[ Y_{ij} = \beta_{0j} + \sum_{r=1}^{k} \beta_{rj} X_{rij} + e_{ij}, \] (7)

while the macro model is,

\[ \beta_{0j} = \sum_{l=0}^{q} y_{l0} W_{ij} + u_{0j}, \] (8)

and the single equation of RIM is written as follow:

\[ Y_{ij} = \sum_{l=0}^{q} y_{l0} W_{ij} + u_{0j} + \sum_{r=1}^{k} \beta_{rj} X_{rij} + e_{ij}. \] (9)

4. Bayesian Methods

Conceptually, Bayesian methods was developed based on Bayes’ theorem, which combines formally prior distribution and data information (likelihood function) into the posterior distribution [18]. Consider Bayes’ theorem [3]:

\[ p(\Theta | \mathbf{y}) = \frac{p(\mathbf{y} | \Theta) p(\Theta)}{p(\mathbf{y})}, \] (10)

where \( \Theta \) is parameter vector and \( \mathbf{y} \) is vector of observation from the sample. \( p(\mathbf{y}) \) is defined as normalized constant with respect to \( \Theta \). Then, the posterior can be explained as a proportional form,

\[ p(\Theta | \mathbf{y}) \propto p(\mathbf{y} | \Theta) \times p(\Theta), \] (11)

Refer to equation (11), the posterior distribution of hierarchical Bayesian models could be specified as follow [20]:

\[ p(\beta, \mathbf{y}, \Omega, T | \mathbf{y}) \propto p(\mathbf{y} | \beta, \Omega) p_1(\beta | \mathbf{y}, T) p_2(\mathbf{y}, \Omega, T), \] (12)
where $\Omega = \text{Var}(y) = \sigma^2_N I_N$, $p_1(\beta|y, T)$ and $p_2(y, \Omega, T)$ denotes first stage prior and second stage prior.

In parameter estimation of complex model, the process of determining the posterior distribution is usually difficult. This process involves integral equation with large dimensions and needs a long time. A numerical approach as a solution to solve this problem is a Markov Chain Monte Carlo (MCMC) [4]. MCMC approach involves the simulation process by taking a random sample of the posterior distribution. This process is executed following due Markov chain process using Monte Carlo simulation iteratively to obtain convergence condition on posterior [18]. Furthermore, sample parameters of Markov chain are taken after a stationary condition reached. Thus, it can be guaranteed that the drawn sample is a posterior distribution sample of the parameter.

Implementation of MCMC methods in Bayesian analysis requires a proper sampling algorithm to obtain a sample of a distribution. Gibbs sampler is a highly efficient generator, so it is often used as a generator of random variables in data analysis using MCMC [7]. Gibbs sampler can be defined as a simulation technique to generate random variables of the marginal distribution functions without having to calculate the density function [5].

5. Data

The data used in this study is sourced from Badan Pusat Statistik (BPS), included data result of National Socioeconomic Survey (Susenas) 2015 and publications of BPS Maluku Utara Province. Data coverage in this study includes 3,749 household samples of Susenas and 10 districts/cities in Maluku Utara Province. The response used in the model is per capita household expenditure ($y$), while the used predictors are household characteristics as predictors at the first level ($X$) and district characteristics as predictors at the second level ($W$). The list of predictors is presented in Table 1.

| First Level Predictors | Second Level Predictors |
|------------------------|-------------------------|
| $X_1$ : Head of household age | $W_1$ : Per Capita GRDP (million rupiahs) |
| $X_2$ : Household dependency ratio | $W_2$ : Percentage of villages had roads drivable by four-wheel vehicle all year |
| $X_3$ : Head of household employment status | |
| $X_4$ : Head of household educational level | $W_3$ : Percentage of original local government revenues to total regional revenue |
| $X_5$ : Asset ownership | |

6. Identification of The Model Structure

Goodness of fit test on per capita household expenditure obtained the result that the data in nine districts/cities in Maluku Utara are following the three parameters Gamma distribution. Therefore, this distribution will be based on the hierarchical model analysis in this study. Figure 1 illustrate the three parameters Gamma distribution pattern of response data in Maluku Utara.
Based on the figure, it can be shown that the three parameters Gamma distribution can explain clearly the difference distribution pattern of household expenditure data among the districts/cities. These differences may be caused by the varied parameter values of the three parameters Gamma distribution. In this study, the third parameter (threshold parameter) was assumed varying, while the others are fixed. Variation of the threshold parameter was suspected be affected by household and district characteristics. Thus, the structure of models RIM using Bayesian approach in this study can be formulated as follows:

\[ Y_{ij} \sim \text{Gamma3}(\alpha_j, \phi_j, \lambda_{ij}) \]

\[ \lambda_{ij} = \beta_{0j} + \sum_{r=1}^{5} \beta_{rj}X_{rij} + e_{ij}, \]

where,

\[ \beta_{0j} = \gamma_{00} + \sum_{l=1}^{3} \gamma_{l0}W_{lj} + \gamma_{0j}, \]

Then, the single equation model of RIM can be written as,

\[ Y_{ij} \sim \text{Gamma3}(\alpha_j, \phi_j, \lambda_{ij}) \]

\[ \lambda_{ij} = \gamma_{00} + \sum_{l=1}^{3} \gamma_{l0}W_{lj} + \sum_{r=1}^{5} \beta_{rj}X_{rij} + e_{ij} + u_{0j}. \] (13)

Parameters of the equation (13) was estimated computationally using WinBUGS 1.4. The software was developed using the Gibbs sampler and MCMC. Therefore, it is very suitable to achieve the objectives of this study.

7. Main Results

The model structure has been formulated previously are implemented to model the per capita household expenditure data in Maluku Utara. According to the Bayesian hierarchical RIM concept, the parameters of micro and macro models were estimated simultaneously. Modeling was calculated...
based on the three parameter Gamma using five household predictors and three district predictors. Parameter estimation of model (13) using WinBUGS were conducted by 10,000 iterations, 10 iterations of thin and additional discarded 400 burn in. Thus, the samples used to estimate the parameters of micro and macro models were 9,600 samples. The estimation results are given in Table 2.

The estimation results of the WinBUGS also displayed a credible interval of the estimated posterior. This interval is analogous to the confidence intervals in frequentist. Credible interval is used as the interval of estimated parameter. Thus, the credible interval could also be used as a diagnostic model, which parameter values are zero for the null hypothesis. The estimated parameters are decided significant when credible interval were non zero. Based on Table 2 can be seen that almost all of the coefficients in models of micro significant at 95% credible interval. Only predictor of asset ownership in three districts/cities were not significant, i.e. Halmahera Timur, Pulau Taliabu, and Ternate.

The five household predictors were proven effected on the rate of per capita household expenditure. Four of the five predictors have a positive effect, such head of household age (X_1), head of household employment status (X_3), head of household educational level (X_4), and assets ownership (X_5). While the household dependency ratio (X_2) has negative effect on per capita household expenditure.
Table 2. Estimated Coefficients of Bayesian Hierarchical Random Intercept Model

| Parameters  | Estimated Value | Credible Interval (95%) | Parameters  | Estimated Value | Credible Interval (95%) |
|-------------|-----------------|-------------------------|-------------|-----------------|-------------------------|
|             |                 | Lower Bounds            | Upper Bounds|                 |                         |
| Estimated  Coefficients of Micro Model |                 |                         |             |                 |                         |
| $\beta_{01}$ | 109200          | 91510                  | 124900      | $\beta_{31}$   | 4003                  | 3039                    | 4953                    |
| $\beta_{02}$ | 96520           | 73960                  | 115300      | $\beta_{32}$   | 4021                  | 3026                    | 4998                    |
| $\beta_{03}$ | 34400           | 18940                  | 48750       | $\beta_{33}$   | 4079                  | 3112                    | 5071                    |
| $\beta_{04}$ | -78830          | -85440                 | -72590      | $\beta_{34}$   | 3996                  | 3000                    | 4986                    |
| $\beta_{05}$ | 89090           | 75780                  | 101100      | $\beta_{35}$   | 3997                  | 3022                    | 4992                    |
| $\beta_{06}$ | -175000         | -189300                | -161200     | $\beta_{36}$   | 4001                  | 3023                    | 4963                    |
| $\beta_{07}$ | 72830           | 54570                  | 90300       | $\beta_{37}$   | 4115                  | 3135                    | 5105                    |
| $\beta_{08}$ | -133400         | -148200                | -120700     | $\beta_{38}$   | 3992                  | 2999                    | 4972                    |
| $\beta_{09}$ | -138600         | -150900                | -126500     | $\beta_{39}$   | 3966                  | 2967                    | 4968                    |
| $\beta_{10}$ | 89670           | 65070                  | 112900      | $\beta_{310}$  | 4020                  | 3030                    | 4992                    |
| $\beta_{11}$ | 2816            | 2620                   | 3016        | $\beta_{41}$   | 8649                  | 7911                    | 9379                    |
| $\beta_{12}$ | 2815            | 2619                   | 3012        | $\beta_{42}$   | 8652                  | 7928                    | 9397                    |
| $\beta_{13}$ | 2794            | 2603                   | 2996        | $\beta_{43}$   | 8656                  | 7924                    | 9388                    |
| $\beta_{14}$ | 3028            | 2825                   | 3229        | $\beta_{44}$   | 8670                  | 7946                    | 9406                    |
| $\beta_{15}$ | 2817            | 2627                   | 3015        | $\beta_{45}$   | 8664                  | 7897                    | 9419                    |
| $\beta_{16}$ | 2727            | 2528                   | 2928        | $\beta_{46}$   | 8670                  | 7919                    | 9415                    |
| $\beta_{17}$ | 2651            | 2470                   | 2837        | $\beta_{47}$   | 8658                  | 7915                    | 9404                    |
| $\beta_{18}$ | 2848            | 2646                   | 3048        | $\beta_{48}$   | 8646                  | 7903                    | 9387                    |
| $\beta_{19}$ | 2824            | 2621                   | 3025        | $\beta_{49}$   | 8619                  | 7883                    | 9354                    |
| $\beta_{20}$ | 2840            | 2641                   | 3040        | $\beta_{410}$  | 8655                  | 7900                    | 9403                    |
| $\beta_{21}$ | -3515           | -4227                  | -2792       | $\beta_{51}$   | 1364                  | 140.3                   | 2594                    |
| $\beta_{22}$ | -3448           | -4168                  | -2732       | $\beta_{52}$   | 1329                  | 81.64                   | 2582                    |
| $\beta_{23}$ | -3437           | -4148                  | -2728       | $\beta_{53}$   | 1337                  | 105.2                   | 2587                    |
| $\beta_{24}$ | -3459           | -4180                  | -2750       | $\beta_{54}$   | 1452                  | 225.8                   | 2697                    |
| $\beta_{25}$ | -3420           | -4135                  | -2698       | $\beta_{55}$   | 1374                  | 114.1                   | 2636                    |
| $\beta_{26}$ | -3421           | -4152                  | -2704       | $\beta_{56}$   | 1192                  | -42.74                  | 2451                    |
| $\beta_{27}$ | -3478           | -4199                  | -2770       | $\beta_{57}$   | 1348                  | 107.4                   | 2589                    |
| $\beta_{28}$ | -3401           | -4130                  | -2684       | $\beta_{58}$   | 1221                  | -55.67                  | 2482                    |
| $\beta_{29}$ | -3414           | -4118                  | -2698       | $\beta_{59}$   | 1217                  | -25.06                  | 2451                    |
| $\beta_{30}$ | -3462           | -4178                  | -2746       | $\beta_{510}$  | 1289                  | 24.01                   | 2557                    |

Estimated Coefficients of Macro Model

| Parameters  | Estimated Value | Credible Interval (95%) |
|-------------|-----------------|-------------------------|
|             |                 | Lower Bounds            | Upper Bounds|                 |                         |
| $\gamma_{00}$ | 23730           | 22250                   | 25200       | $\gamma_{20}$   | 325.4                  | 103.3                   | 551                     |
| $\gamma_{10}$ | 1249            | 983.4                   | 1517        | $\gamma_{30}$   | 4312                  | 4058                    | 4570                    |

* Not significant coefficient

The results showed that the intercept of micro models have different values among the districts. This condition was significantly affected by the district characteristics. It was shown by all estimated parameter of the district predictors were significant at 95% credible interval. This means that the three districts predictor effect on per capita household expenditure. the third predictors, especially, had positive effects on per capita household expenditure.
8. Conclusions
Household and district characteristics had been proven to give significant effects on per capita household expenditure in Maluku Utara. This situation had been explained by the random intercept model based on three parameters Gamma distribution using Bayesian approach. Thus, this model had demonstrated its work to capture the hierarchical structure in the data. Another interesting future perspective is to model all of the coefficient parameters regression at the first level with predictors at level the second level. This will be performed in the author’s final project.

References
[1] Alanko T 1997 Statistical Models for Estimating the Distribution Function of Alcohol Consumption: A Parametric Approach (Vol. 44). Finnish Foundation for Alcohol Studies; Piscataway, NJ, USA: Rutgers University Center of Alcohol Studies
[2] Basak I and Balakrishnan, N 2012 Estimation for the three-parameter gamma distribution based on progressively censored data. Statistical Methodology, 9(3), 305-319
[3] Box G E and Tiao G C 1973 Bayesian Inference in Statistical Analysis (United State: Addison-Wesley)
[4] Carlin B P and Chib, S 1995 Bayesian Model Choice via Markov Chain Monte Carlo Methods. Journal of the Royal Statistical Society, Series B (Methodological), 473-484
[5] Casella G and George E I 1992 Explaining the Gibbs Sampler The American Statistician, 46(3), 167-174
[6] De Leeuw J, Meijer E, and Goldstein H 2008 Handbook of Multilevel Analysis (New York: Springer)
[7] Gelman A, Carlin J B, Stern H S, and Rubin D B 2014 Bayesian Data Analysis, Third Edition (USA: Chapman & Hall/CRC)
[8] Goldstein H 1995 Multilevel Statistical Models (London: Edward Arnold)
[9] Goldstein H, Rasbash J, Browne W, Woodhouse G, and Poulain M 2000 Multilevel Models in The Study of Dynamic Household Structures European Journal of Population/Revue européenne de Démographie, 16(4), 373-387
[10] Ghosh A, Gangopadhyay K and Basu B 2011 Consumer Expenditure Distribution in India, 1983–2007: Evidence of A Long Pareto Tail. Physica A: Statistical Mechanics and its Applications, 390(1), 83-97
[11] Hox J J, Moerbeek M, and Van de Schoot R 2010 Multilevel analysis: Techniques and applications. (New York: Routledge)
[12] Iriawan N 2012 Modelling and Data-Driven Analysis, Volume 1 (Surabaya: ITS Press)
[13] Ismartini P 2013 Development in Hierarchical Linear Model with Bayesian Approach: Modelling of per Capita Household Expenditure Data (Dissertation), Institut Teknologi Sepuluh Nopember, Surabaya
[14] Ismartini P, Iriawan N, and Ulama B S S 2013 Comparison of Unilevel and Multilevel Model for Hierarchical Structure Data Analysis Using Bayesian Approach. Paper Presented at The Proceeding of the Fourth National Mathematics Conference.
[15] Johnson N L, Kotz S, and Balakrishnan N 1995 Continuous Univariate Distribution, Volume 1 Second Edition (New York: John Wiley and Sons)
[16] Kruschke J K and Vanpaemel W 2015 Bayesian Estimation in Hierarchical Models. The Oxford Handbook of Computational and Mathematical Psychology, 279
[17] Ntzoufras I 2009 Bayesian Modeling in WinBugs (New Jersey, USA: John Wiley & Sons, Inc)
[18] Pertwi R 2012 Modelling per Capita Expenditure of Regencies Kalimantan Barat Using Hierarchical Bayesian Method (Surabaya: Institut Teknologi Sepuluh Nopember)
[19] Raudenbush S W and Bryk A S 2002 Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition (Vol. 1). (London: Sage).