M – 21

MODELLING CASES OF LOW BIRTH-WEIGHT INFANTS WITH GENERALIZED LINEAR MIXED MODEL

Antonius Benny Setyawan, Khairil Anwar Notodiputro, Indahwati
Bogor Agriculture University, Indonesia
abenny@bps.go.id; bennsetyawan@gmail.com

Abstract

Low Birth-Weight (LBW) is defined as a birth weight of a live-born infant of less than 2.500 grams regardless of gestational age. The causes of LBW cases can be grouped into two main causes: premature birth and case of small for gestational age (SGA). There are many risk factors that can induce directly or indirectly so that these causes may occur. Case of LBW is associated with infant mortality, infant morbidity, inhibited growth and slow cognitive development, also chronic diseases in later life.

To suppress rate of LBW first we must estimate the rate correctly. Data of LBW comes from Indonesian Health and Demographic Survey (IDHS) 2012 which is divided into 3 groups: written (measured accurately), recall (measured inaccurately) and not weighed (not measured). Published national rate of LBW is 7.3% with provincial rates fall between 4.7-15.7 %. The estimation came from only 2 former groups without consideration of assumed difference accuracy on second group.

To estimate the difference and the rate of the third group, Generalized Linear Mixed Model (GLMM) is used with live-born infants as observation units because observations from the same sampling unit tends to correlate due to multistage sampling design.

The result of the model at $\alpha = 0.05$ is highly-significant, with fixed effect variables that are statistically significant to the case of LBW are Estimated Size, Preceding Interval, Pregnancy Complication, Mother's Age, Province and Education. Higher portion of variance component is on the G-side as a result of multistage sampling, with Household level has highest within variance. On the R-side, recall group data has higher variance than written group. It is an indication of lower accuracy of the birth weight data on this group. Based on the model, estimation of LBW rate including not weighed group result 7.96% slightly higher than direct estimate.

Keywords: Low Birth-Weight, GLMM, Logistic Regression, IDHS 2012

Introduction

Low Birth-Weight Case is defined as a birth weight of a live-born infant of less than 2500 grams (WHO, 2011) regardless of gestational age measured on first hours after birth. During early days of life, babies may suffer significant weight loss due to feeding adjustment so that measurement after several days after birth tends to result lower value. Value cut-off point at 2.500 grams is based on epidemiological observation that infants weighing less than 2.500 grams are approximately 20 times more likely lead to case of infant mortality (Kramer, 1987). Hence, reducing LBW case becomes an important effort because reducing Infant Mortality Rate (IMR) is one of eight Millennium Development Goals (MDGs). Reducing LBW case to relatively 30% is also one of Six Global Nutrition Targets 2025 declared by WHO (WHO, 2014). Besides infant mortality, LBW case is closely related to infant morbidity, inhibited growth and slow cognitive development, also chronic diseases in later life.

LBW case is commonly caused by two reason: premature birth (below 37 weeks) and intra uterine growth restriction (IUGR) (Kramer, 1987). Normal birth weight is ranged between 10th and 90th percentile of 40 weeks gestational age. Although the distribution differs from one country to another, the standard value is ranged from 2.500-4.000 grams and termed with appropriate for gestational age (AGA). Measurement below 2.500 grams on 40th week is termed small for the gestational age (SGA) and over 4.000 grams is termed large for gestational age (LGA) (Hutcheon, et al. 2010).

Risk factors for the LBW case are factors that induced premature birth and IUGR. These factors are genetic factors, maternal characteristics, diseases and pregnancy complications, nutrition intakes, lifestyles and environmental factors. Some of these factors have been researched are: sex, firstborn, twins, age of mother, precedence birth, parity, stillbirth/abortion history, malaria, HIV, anemia, diabetes, hypertension, poverty, education, alcohol, tobacco and drug abuse, altitude, pollution, etc. (Kramer, 1987; UNICEF and WHO, 2004).

In Indonesia, complete data of LBW case only available at provincial level based on five-yearly Indonesia Health and Demographic Survey held by BPS in collaboration with BKKBN and Ministry of Health. Based on latest IDHS 2012 estimation of LBW national prevalence is 7.3% with Province of East Nusa Tenggara has the highest prevalence (15.7%) and Province of DKI Jakarta the lowest (4.7%) (BPS et al. 2013). The estimation is calculated from data of live-born infants on 5 years period which are weighed after birth. Live-born infants on 5 years period observed by the survey are grouped into three: (1) weighed and written (18.4%), (2) weighed on recall (65.5%) and (3) not weighed (16.1%). So we can say that the estimation based only on two former groups of data (83.9%) with only small proportion (18.4%), which are group 1, is measured accurately.

Problem that may arise is the sufficiency of the prevalence estimates, which are based on incomplete sample, which is not taking group of not weighed infants (group 3) into consideration, and high proportion of inaccurate measurement which are based on mother's recall (group 2). The objective of the study is to estimate LBW prevalence based on a model containing all of the three groups and compare it with the published result.

Research Method

IDHS 2012 is a multistage sampling designed survey with clusters (called census blocks) as primary sampling units and households as ultimate sampling units. The observational unit, however, is all infants born in five-year periods from every woman in the sampled households. From this data we practically can assume that there will be a relationship from every infant came from same woman, from same household and from same cluster. Hence, the observational units are not independent one to another. This condition violates assumption of classical Linear Model of independent observations. Fisher (1918) proposed *random effect model* to study correlations of trait between relatives. Combination of classical linear model which contains fixed effects and random effects to the model results into what so called Linear Mixed Model (LMM) with general form:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \qquad ....(1)$$

with $\boldsymbol{X\beta}$ represents fixed effects to the model and $\boldsymbol{Z\gamma}$ as random effects and $\boldsymbol{\varepsilon}$ is random errors. Random effect, on a contrary to the fixed effect, is an effect which are assumed to be drawn randomly from a population of effect. Thus, we no longer focus on estimation of parameter $\boldsymbol{\gamma}$, instead we are more interested on the composition of variance component Var($\boldsymbol{\gamma}$) and Var($\boldsymbol{\varepsilon}$). Which each component $\boldsymbol{X\beta}$, $\boldsymbol{Z\gamma}$ and $\boldsymbol{\varepsilon}$ are independent one another we can derive.

$$\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{R}) \quad \Rightarrow \quad \boldsymbol{y}|\gamma \sim N(\boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \boldsymbol{R})$$
$$\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{G}) \quad \Rightarrow \quad \boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$$
$$Cov[\boldsymbol{\varepsilon}, \boldsymbol{\gamma}] = \boldsymbol{0} \quad \Rightarrow \quad \boldsymbol{V} = Var(\boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}) = \boldsymbol{ZGZ}' + \boldsymbol{R}$$

The composition of variance of random effects ($\boldsymbol{\gamma}$) is called *G-side* and the composition of variance of random effects ($\boldsymbol{\varepsilon}$) is called *R-side.* Zero value on the G-side variance indicates that no variance can be explained by the random effect, the random effects thus has no effect on the model.

Another violation to the classical Linear Model occurs because LBW case is a binary response variable which its values is divided into two: 1(LBW case) and 0 (non-LBW case). Because of this binary property, the variable fits Bernoulli distribution, and violates assumption of Normality. Consequences of the Bernoulli distribution is that the variance of the response are dependent to the expected value. For every Bernoulli case y_i the expected value is E(y_i) = p_i and the variance is Var(y_i) = p_i (1 – p_i), violating another assumption of homoscedasticity. Nelder and Wedderburn (1972) introduced Generalized Linear Model (GLM) to accommodate variables with non-Normal distribution so they can fit into Linear Model via so-called *link function*. Distribution of these variables, however, should belong to *exponential family distribution*.

General form of exponential family distribution is:

$$f(y_i, \eta_i, \phi_i) = exp\left\{\frac{y_i\eta_i - a(\eta_i)}{b(\phi_i)} + c(y_i, \phi_i)\right\} \qquad ....(2)$$

with η_i = g(E(y_i)) is the link function, a function of the expected value which linked random component y_i to the linear predictor of Linear Model constructing the GLM form below:

$$g(\boldsymbol{y}) = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \qquad ....(3)$$

Data with Bernoulli, Binomial, Poisson, Exponential, Normal, Chi-Square, Gamma and Beta distributions are members of exponential family distribution, thus can be applied into GLM. Generalized Linear Mixed Model (GLMM) is generalization of Linear Mixed Model (1) in a way of Generalized Linear Model (3) which takes form as below (McCulloch and Searle, 2001):

$$g(\boldsymbol{y}) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \qquad ....(4)$$

Assumptions of GLMM thus are slightly different:

$$\begin{aligned} \boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \boldsymbol{R}) \quad &\Rightarrow \quad g(\boldsymbol{y}|\gamma) \sim N(\boldsymbol{X\beta} + \boldsymbol{Z\gamma}, \boldsymbol{R}) \\ \boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{G}) \quad &\Rightarrow \quad g(\boldsymbol{y}) \sim N(\boldsymbol{X\beta}, \boldsymbol{V}) \\ Cov[\boldsymbol{\varepsilon}, \boldsymbol{\gamma}] = \boldsymbol{0} \quad &\Rightarrow \quad \boldsymbol{V} = Var(\boldsymbol{Z\gamma} + \boldsymbol{\varepsilon}) = \boldsymbol{ZGZ'} + \boldsymbol{R} \end{aligned}$$

Because the case of LBW is a binary distribution of Bernoulli, and the link function of Bernoulli distribution is logit function, GLMM model for LBW case is Logistic Regression with random effect as follows:

$$logit(\boldsymbol{y}) = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{\varepsilon} \quad ; logit\ (y_i) = \ln\left(\frac{y_i}{1 - y_i}\right) \qquad ....(5)$$

Estimation of the parameters are obtained using method of Residual Pseudo-Likelihood with optimization via Newton-Raphson with ridging. The method is used because in general exponential family distribution (except Normal) the variance of the distribution is a function of the mean (Jiang, 2007). In the case of Bernoulli distribution Var(y) = p(1 – p) = E(y)(1 – E(y)). The method maximizes approximation of joint distribution of the random error which has zero mean so that the estimation of variance component can be obtained. Estimation of fixed effect then can be computed with Generalized Least Square (GLS) with estimated variance component.

All the data which are used in the model comes from IDHS 2012. Total number of observation units are 18,021 births from 15,262 mothers, 14,742 households and 1,827 census blocks. Variables from IDHS 2012 which are included in the model are all categorical. The dependent variable are binary with LBW case considered as *event*. Independent variables provided by the survey which are assumed to affect the case as fixed effects are (categories in parentheses are reference categories):

Table 1.
Fixed Independent Variables included in the Model

Variable	**Categories**	**Para-meters**	**Code**
Intercept	-	1	INT
Weight Status	(Written), Recall	1	FLAG
Estimated Size	Very Small, Smaller than Average, (Average), Larger than Average, Very Large	4	SIZE
Twin	(Singleton), Twin or more	1	TWIN
Preceding Birth	Firstborn,< 2 years, (≥ 2 years)	2	PREC
Birth Order	(1st -3rd), 4th or more	1	BORD
Pregnancy Complication	Premature, Other Pregnancy Complication, (No Pregnancy Complication), No Information	3	PREM
Termination History	(No Terminated History), Terminated History	1	TERM
Mother's Age	< 20 years old, (20-34 years old), 35-49 years old	2	AGE
Sex	Male, (Female)	1	SEX
Urban/Rural	Urban, (Rural)	1	UR
Province	(DKI Jakarta), other 32 Provinces	32	SPROV
Mother's Education	(No or Primary), Secondary or Higher	1	EDU
Family Wealth Index	Poor, Middle, (Wealthy)	2	WEALTH
Mother's Physical Work	Non Physical Work, Physical Work, (Not Working)	2	WORK
Mother's Smoking Habit	Active, Passive, (Not Smoking)	2	SMOKE
Water Source	Protected, (Unprotected)	1	WATER
Total	**73 categories**	**58**	

Province and Urban-Rural play a role as strata on IDHS sampling design, while Weight Status is our other key variable besides LBW case. Missing values in independent variables is categorized as No Information so that there is not too many discarded observations. Estimated size is a *proxy* variable for birth weight provided by the survey.

Besides all the independent variables considered as fixed effects, there are also random effects variables as a result of multistage sampling design which are included as nested random effects: Census Block (CB_i), Household ($HH_{j(i)}$), Mother ($M_{k(ij)}$). Therefore, the hypothesized model are as follows:

$$\ln\left(\frac{y_{ijkl}}{1-y_{ijkl}}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_{57} X_{[57]ijkl} + CB_i + HH_{j(i)} + M_{k(ij)} + \varepsilon_{ijkl} \qquad \text{........(6)}$$

Our point of interest here is to estimate parameter of fixed effects $\boldsymbol{\beta}'$ = $(\beta_0, \beta_1, \dots, \beta_{57})$ and variance of random effects Census Block (**G₁**), House Hold (**G₂**), Mother (**G₃**) and error (**R**). The variance error is separated into two group: Written (**R**$_W$), and Recall (**R**$_R$). Computation procedure to estimate the parameters is done by SAS Program using PROC GLIMMIX.

.

Results and Discussion

The result model are overall statistically significant, with fixed parameter estimates are as follows (significant categories are in **bold letters**, slightly not significant categories are in **grey bold letters**).

Table 2.
Parameter Estimation for Fixed Effects

Effect	Categories	Estimate	Pr > |t|	Exp(Est.)
INT	**-**	**-4.1214**	**<.0001**	**0.02**
FLAG	Recall	0.1479	0.146	1.16
SIZE	**Very Large**	**-3.0564**	**0.0017**	**0.05**
	Larger than Average	**-1.6687**	**<.0001**	**0.19**
	Smaller than Average	**3.1915**	**<.0001**	**24.32**
	Very Small	**5.0305**	**<.0001**	**153.01**
TWIN	**Twin +**	**2.7973**	**<.0001**	**16.40**
PREC	**< 2 yrs**	**0.6402**	**<.0001**	**1.90**
	Firstborn	**0.2425**	**0.0174**	**1.27**
BORD	4th +	-0.04607	0.7273	0.95
PREM	No Information	0.02491	0.8398	1.03
	Other Pregnancy Complication	**0.2785**	**0.0337**	**1.32**
	Premature	0.3925	0.1151	1.48
TERM	History	-0.03478	0.7728	0.97
AGE	**35-49 yrs**	**0.2276**	**0.0716**	**1.26**
	< 20 yrs	0.09444	0.5091	1.10
SEX	Male	0.05851	0.4797	1.06
UR	Urban	-0.1116	0.2663	0.89
SPROV	Aceh	-0.05505	0.8716	0.95
	Bali	0.2851	0.4026	1.33
	Bangka Belitung	-0.01042	0.9754	0.99
	Banten	**0.9187**	**0.0012**	**2.51**
	Bengkulu	-0.04165	0.9151	0.96
	Central Java	0.1778	0.5598	1.19
	Central Kalimantan	-0.1259	0.7375	0.88
	Central Sulawesi	**1.1169**	**0.0004**	**3.06**
	East Java	0.2007	0.5012	1.22
	East Kalimantan	-0.00282	0.9935	1.00
	East Nusa Tenggara	**1.2005**	**0.0001**	**3.32**
	Gorontalo	**0.6381**	**0.0529**	**1.89**
	Jambi	-0.4351	0.2477	0.65
	Lampung	0.2389	0.4837	1.27
	Maluku	0.3108	0.4387	1.36
	North Maluku	0.1718	0.648	1.19
	North Sulawesi	-0.1183	0.7187	0.89
	North Sumatera	-0.2482	0.4344	0.78
	Papua	0.4916	0.2853	1.63
	Riau	-0.3659	0.2871	0.69
	Riau Islands	0.2679	0.4501	1.31
	South Kalimantan	0.2298	0.4844	1.26
	South Sulawesi	0.0496	0.868	1.05
	South Sumatera	0.08045	0.8087	1.08

	Southeast Sulawesi	**-0.6896**	**0.067**	**0.50**
	West Java	0.02697	0.9252	1.03
	West Kalimantan	**0.8307**	**0.0082**	**2.29**
	West Nusa Tenggara	**0.6315**	**0.04**	**1.88**
	West Papua	0.3639	0.2913	1.44
	West Sulawesi	**0.8466**	**0.0125**	**2.33**
	West Sumatera	-0.1586	0.6482	0.85
	Yogyakarta	**0.9464**	**0.0024**	**2.58**
EDU	**Secondary or Higher**	**-0.4484**	**<.0001**	**0.64**
WEALTH	Middle	0.05993	0.5824	1.06
	Poor	0.2101	0.1332	1.23
WORK	**Non Physical Work**	**0.1864**	**0.0647**	**1.20**
	Physical Work	-0.02852	0.8002	0.97
SMOKE	Active	0.1083	0.6985	1.11
	Passive	0.07442	0.4351	1.08
WATER	Protected	-0.1419	0.1665	0.87

Fixed effect variables that are statistically significant to the case of LBW are Intercept, Estimated Size, Twin, Preceding Interval, Pregnancy Complication, Mother's Age, Province and Education. Positive coefficients means that the categories have greater chance of LBW case and vice versa.

Interpretation of regression coefficients of the fixed effects in logistic regression based model is odds ratio. Odds ratio is a ratio of odd of a certain category to odd of reference category, where odd is a ratio of probability of an event compared to the probability that event is not happening. Thus, definition of odds ratio by itself is rather complicated. In a simple way, odds ratio of an event of category A to category B with value of 2 simply means that in category A the event are twice more likely to happen than in category B (reference category). The estimate of odds ratio are exponential of estimated fixed effect as follows:

$$\hat{\theta} = \frac{\left(\frac{\pi_A}{1-\pi_A}\right)}{\left(\frac{\pi_B}{1-\pi_B}\right)} = e^{\widehat{\beta_{AB}}} \qquad ....(7)$$

All categories in Estimated Size as a proxy variable is highly significant with correct signs and order and great magnitude ($|\beta_j|>1$) for each respective case. It indicates that it is a very good proxy for LBW case. A mother reporting very small baby 153 times, while mother reporting very large baby is 1/20 times, more likely to have a LBW baby than the mother reporting average size.

Variables which are theoretically relate to case of LBW as studied by Kramer (1987) turn out not all of them significantly influence the chance of LBW. It is because the proxy dominating the model. The high correlation between birth weight and estimated birth size seems to cancel out the effect on some proposed categories such as: high birth order, premature birth, and poor wealth index. Firstborns and < 2 years interval babies, twins, babies with pregnancy complication still has significantly higher chance of LBW. Firstborns are 1.27 times, while babies with birth interval <2 years are 1.9 times, more likely than babies with birth interval more than 2 years. Twins are 16.4 times more likely to have LBW than singletons and babies with pregnancy complication 1.32 more likely. While baby from a mother with secondary or higher education has significant lower chance of LBW, 0.64 times more likely than mother with no or primary education. It is all with assumption that other variables are considered constant.

While province with significant higher chance of LBW case are: D.I. Yogyakarta, Banten, West Kalimantan, West Nusa Tenggara, East Nusa Tenggara, Central Sulawesi and West Sulawesi compared to DKI Jakarta as reference given other variables included in the model remains the same. The estimated odds ratios of LBW case for those provinces fall between 1.88 – 3.32 times more likely than in DKI Jakarta with East Nusa Tenggara as the highest.

Table 3.
Parameter Estimates of Variance Components

Covariance Parameter	Subject	Group	Estimate	SE
Intercept	CB		0.2449	0.1047
Intercept	HH(CB)		0.781	0.2116
Intercept	M(CB*HH)		0	.
Residual (VC)		Recall	0.9486	0.01424
Residual (VC)		Written	0.7296	0.0201

Table 4.
Tests of Covariance Parameters Based on the Residual Pseudo-Likelihood

Label	DF	-2 Res Log P-Like	ChiSq	Pr > ChiSq	Note
Homogeneity	1	105659	77.46	<.0001	DF

The inclusion of random effects in the model results to parameter estimation of variance components below (Table 3.). Higher portion of variance component is on the G-side as a result of multistage sampling, with Household level has highest within variance. On the R-side, recall group data has higher variance than written group and homogeneity test results that there is a significant variance difference between groups. It is an indication of lower precision of the birth weight data on recall group. Although the recall group has a lower precision, as the residual variance is higher, the accuracy of both group are statistically same, as parameter estimate of fixed effect on recall category is insignificant (Table 2.). It means estimated rate of LBW on recall group does not tend to underestimate or overestimate.

Using the model above, now we can calculate prediction of probability to the case of LBW on not weighed group data. The result of the calculation for the total of three groups, estimated national LBW rate from the model is 7.96% slightly higher than the direct estimation based only from two groups 7.3%.

Conclusion and Suggestion

Twin, firstborn and narrow preceding interval, pregnancy complication and low mother’s education are categories which are highly significant to increase the chance of LBW case. While birth order, mother’s age, premature case stillbirth/miscarriage/abortion history, urban-rural area, sex of infant, occupation, wealth index, smoking habit, and water source are not statistically significant to LBW case. The significance of risk factors of LBW case suppressed by the magnitude of proxy variable estimated size at birth. Because our objective is to obtain good prediction for not weighed group, this doesn’t seem to matter. But if we interested in studying effect of factors to LBW case including proxy variables into the model is not a good point.

Provinces with significantly higher LBW case indicate the lack of area effect included in the model. Unique characteristics of every region haven’t been portrayed enough in the model. Inclusion of several area effects presumably related may reduce province effect significantly. Not all these province with higher chance of LBW are also underdeveloped. Then the issue is not always poverty or access to health facility but may fall in socio-cultural area. The higher chance of LBW case in those provinces can’t be explained entirely by variables proposed in the model. Other factors which trigger higher case of LBW in those provinces must be studied more extensively.

There is not enough proof for bias from recall data, which shown by insignificant coefficient parameter of weight status. It means estimated rate of LBW on recall group has the same accuracy and does not tend to underestimate or overestimate. The difference of mean of the two groups are already explained by variables included in the model. But the significant variance difference between groups indicate that weight information of the recall group is less precise.

Total estimation via model using all information provided by the data should be taken into consideration in the case of incomplete data. Ignorance to the missing data may bring to biased estimation if there is a pattern for the incompletion. For example, if mother with no record of birth weight is a result of absence of health facility and in turns also result to higher chance of LBW, estimation ignoring the incomplete data will be underestimated. However, modelling of the total data should be done with caution to take every aspects of the data into consideration.

Bibliography

BPS, BKKBN, Ministry of Health, ICF International. 2013. *Indonesia Demographic and Health Survey 2012*. Jakarta (ID). BPS.

Fisher, R.A. 1918. The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh*, 52 (2), 399–433.

Hutcheon, J.A. Walker M. and Platt, R.W. 2010. Assessing the Value of Customized Birth Weight Percentiles. *American Journal of Epidemiology,* 173(4), 459-467.

Jiang J. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. New York (US): Springer.

Kramer, MS. 1987. Determinants of Low Birth Weight: Methodological Assessment and Meta-Analysis. *Bulletin of World Health Organization*, 65(5), 663–737.

McCulloch, C. E and Searle S. R. 2001. *Generalized, Linear, and Mixed Models*. New York (US): John Wiley and Sons Inc.

Nelder, J. and Wedderburn, R. W. M. 1972. Generalized Linear Models. *Journal of the Royal Statistical Society,* A, 135, 370-384.

United Nations Children’s Fund, World Health Organization. 2004. *Low Birth Weight: Country, Regional and Global Estimates*. New York (US). UNICEF.

World Health Organization. 2011. *International Statistical Classification of Diseases and Health Related Problems*. 10th Revision. Geneva (CH). WHO.

World Health Organization. 2014. *Global Nutrition Targets 2025: Low Birth Weight Policy Brief*. Geneva (CH). WHO.