Forum Statistika dan Komputasi : *Indonesian Journal of Statistics*
ISSN : 0853-8115
Vol. 20 No.2 Oktober 2015, p: 28-33

available online at:
journal.ipb.ac.id/index.php/statistika

COMPARISON OF LOW BIRTH WEIGHT RATE ESTIMATES BASED ON DIFFERENT AGGREGATE LEVELS DATA USING LOGISTIC REGRESSION MODEL

Antonius Benny Setyawan[1], Khairil Anwar Notodiputro[2], Indahwati[3]

[1,2] Department of Statistics, Bogor Agricultural University, Indonesia
E-mail : bennsetyawan@gmail.com[1]

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. Case of LBW is associated with infant mortality, infant morbidity, inhibited growth and slow cognitive development, also chronic diseases in later life. It is vital because with high LBW rate the generation hardly grow into its full potential. There are many risk factors, whether direct or indirect, can cause a birth as a high risk of Low Birth Weight case. These factors are genetics, obstetrics, nutrition intakes, diseases, toxic exposures, pregnancy care and social factors. With these factors measured, statistical modelling can be used to estimate rate on group level or probability on individual level of the Low Birth Weight event. As the case is a binary response, Logistic Regression Model is commonly used.

Data of LBW case and the risk factors came from Indonesian Demographic and Health Survey (IDHS) 2012. Published national rate of LBW was 7.3% with provincial rates fell between 4.7-15.7 %. Although the national rate was considered low, the wide variation of provincial rates showed that the problem was not handled so well. However, these rates cannot be measured yearly due to 5 year period of the survey. With the availability of risk factors data a model can be built to estimate the LBW rates. But, another problem for the model is the case when aggregate level data is available instead of individual level data. So, the purpose of this study was to compare models based on different aggregate levels and theirs estimated provincial rates. Comparison was done among individual birth level, mother level, household level and census block (cluster) level. Models from three former levels were quite similar with adequate significant parameters, while cluster level model was resulted only a few significant parameters. But instead, LBW rate estimates from cluster level model were the closest to the direct estimates. But the variance of these estimates was still higher than the other models.

Key words **:** *Low Birth-Weight, IDHS, Logistic Regression, GLM, Aggregate Data*

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 of life. During early days of life, babies may suffer significant weight loss due to feeding adjustment so that measurement several days after birth tends to result lower value. 2.500 grams cut off point is globally used based on 10th percentile of 40 weeks gestational age which are considered as small for gestational age (SGA) category (Hutcheon *et al*, 2010). Epidemiological observation shows 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 rate becomes an important effort that indirectly reduces Infant Mortality Rate (IMR) and a result of improvement of Maternal Health, two of eight Millennium Development Goals (MDGs) (UN, 2014). Reducing LBW case to relatively 30% is also declared by WHO as one of Six Global Nutrition Targets 2025 (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 (Barker, 1995). These long term effects will affects individual quality of life. It becomes crucial because a generation with high LBW rate hardly grows into its full potential as labor force and human resources, especially in Indonesia which in period of 2005-2040 is on what so-called as *Demographic Window*. In this period, with labor force at full capability and low dependency rate, a developing country will be able to grow to a developed country (Bloom *et al.* 2003).

LBW rate is measured by Indonesian Demographic and Health Survey (IDHS) held every five years by BPS in collaboration with BKKBN and Ministry of Health. The latest was IDHS 2012. Published national rate of LBW was 7.3% with provincial rates fell between 4.7-15.7 % with the lowest is DKI Jakarta (4.7%) and the highest is NTT (15.7%) (BPS *et al*, 2013). Although the national rate was considered low, the wide variation of provincial rates showed that the problem is quite serious in some provinces. With data availability is only every 5 years, it is hard to monitor if a policy can be considered as effective. Therefore, building a statistical model to estimate LBW rates is a necessity. A theoretical ground to build such model must be considered well. There are many risk factors, whether direct or indirect, can cause a birth as a high risk of Low Birth Weight case. These factors are grouped into: genetics, obstetrics, nutrition intakes, diseases, toxic exposures, pregnancy care and social factors (Kramer, 1987). Some of these factors may be available yearly from other sources beside IDHS itself for estimation purposes.

For the model based on IDHS was built in individual level, it would not be applicable if the data was only available on higher aggregate level, which was commonly happened. A model built based on the same aggregate level are more appropriate as a tool to estimate. In doing so, performance of each model must be measured and compared to conclude whether at a certain aggregate level, estimation based on the respective model can be statistically justified. This paper is a result of the research by aggregating and modelling data (response and explanatory variables) from IDHS 2012 on four aggregate levels: individual birth, mother, household and census block (cluster) level.

LITERATURE REVIEW

Low Birth Weight case, as response *Y*, can be considered as a binary variable. Infant born with LBW considered as *event* and coded as 1 and the counterpart considered as *non-event* and coded 0. Therefore modelling the variable can be seen as measuring probability of the *event* case. One of the models commonly used for this case is Logistic Regression Model (LRM). Logistic model is preferred because of its simple interpretation in relation to concept of *odds ratio*. Unlike Classical Regression Model which is based on Normal Distribution, LRM is based on discrete Binomial Distribution (or Bernoulli on single trial case). Which all of them are forms of Generalized Linear Model (GLM) proposed by Nelder and Wedderburn (1972) for Exponential Family Distribution below:

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

Thus, a binomial distribution (n, π) can be presented as exponential family:

$$f(y_i) = \binom{n_i}{y_i}\pi_i{}^{y_i}(1-\pi_i)^{n_i-y_i} = \exp\left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \ln(1-\pi_i) + \ln\binom{n_i}{y_i}\right]$$

$$\eta_i = g(\mu_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$

The model generally connects random component $\boldsymbol{Y}$, to linear predictor $\mathbf{X}\boldsymbol{\beta}$, via *link function* of η_i = $g(\mathrm{E}(Y_i))$ a function which mapped Y_i into $\mathbb{R}$. On binomial case the model becomes:

$$\begin{aligned} g(\mathrm{E}(Y_i)) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) \\ = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \\ + \varepsilon_i \end{aligned}$$

$$Y_i = 1,2,\ldots,n_i \quad ; -\infty < g\big(\mathrm{E}(Y_i)\big) < \infty$$

The value $\theta_i = \frac{\pi_i}{1-\pi_i}$ is termed as *odds* and the link function $g(\mathrm{E}(Y_i)) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ is known as *logit* function, therefore, GLM based on binomial distribution is also called Logistic Regression Model (LRM).

On individual level the event an i^{th} infant was a LBW case is distributed Bernoulli (π_i). On aggregate level, from n_i birth that level, the number of LBW cases is the aggregation of number of events on individual level. With the assumption that every case is independent one another the distribution is Binomial (n_i, π_i). Based on data from IDHS 2012, both response and explanatory variables were available on individual level. To model the aggregate data, the explanatory variables must also be presented as aggregates. Because LRM is modelling probability and the size of aggregates n_i was unbalanced, means of the explanatory variables ($\bar{X}_{ij}$) was preferred on modelling aggregate data. In the case of categorical variable, and dummy variables were used on individual level, the proportions of each category (except the reference category) were used as means. Thus the models proposed became:

Individual Model:

$$g(\mathrm{E}(Y_i)) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$$
$$= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

Aggregate Model:

$$g(\mathrm{E}(Y_i)) = \ln\left(\frac{y_i/n_i}{1-y_i/n_i}\right)$$
$$= \beta_0 + \beta_1 \bar{X}_{i1} + \cdots + \beta_p \bar{X}_{ip} + \varepsilon_i$$

Method of Maximum Likelihood Estimation (MLE) was used to estimate parameter $\boldsymbol{\beta}$ by maximizing likelihood function $l(\boldsymbol{\beta})$ as solution to equation $\mathbf{U} = l'(\boldsymbol{\beta}) = \mathbf{0}$. The solution is obtained numerically with iterative method similar to Iterated Weight Least Squares (IWLS) as follows (Dobson, 2002):

$$\mathfrak{J}^{(m-1)}\boldsymbol{b}^{(m)} = \mathfrak{J}^{(m-1)}\boldsymbol{b}^{(m-1)} + \mathbf{U}\left(\boldsymbol{b}^{(m-1)}\right)$$

where,

$E(\mathbf{U}) = \mathbf{0}$; $Var(\mathbf{U}) = E(\mathbf{U}^2) = -E(\mathbf{U}') = \mathfrak{J}$

$\mathbf{U}$ is called as vector score and $\mathfrak{J}$ is called as Fisher's information matrix. Thus the method is also simply known as Fisher scoring method.

Research Variables

IDHS is a survey with complex sampling design consists of stratified and multistage sampling. The stratification was used to reduce variance as each strata was relatively homogeneous subpopulation. While the multistage design was used to reduce cost by suppressing the spread of samples in large geographic area and to solve unavailability of sampling frame on unit level (Scheaffer *et al*, 2006).

Provinces and urban/rural category play roles as strata (65 strata) with sample allocation for each strata is proportional to its population on Population Census 2010. With total sample of 1.840 census blocks (clusters), each strata is sampled based on its allocation with systematic sampling. From each cluster, households are listed and 25 households sampled systematically. Every member from sampled household are enumerated with respective type of questionnaire (BPS *et al*, 2013).

Data from IDHS 2012 consisted 15.124 weighed births observed from past 5 years period, which came from 13.224 observed mothers, 12.809 households and 1.801 census blocks. Explanatory variables were all categorical and also came from the IDHS data. Based on Kramer (1987) those variables and their categories (categories inside brackets (.) are reference categories) proposed to the model were:

Table 1. Explanatory Variables included in the Model

Variable	Categories	Parameter
Intercept	-	1
Weight Status	(Written), Recall	1
Estimated Size	Very Small, Smaller than Average, (Average) Larger than Average, Very Large	4
Twin	(Singleton), Twin or more	1
Preceding Birth and Birth Orde	Firstborn, < 2 years and 2nd -3rd, (≥ 2 years an 2nd -3rd), < 2 years and 4th, ≥ 2 years and 4th.	4
Pregnancy Complication	Premature, Other Pregnancy Complication, (No Pregnancy Complication), No Informatic	3
Termination History	(No Terminated History), Terminated History	1
Mother's Age	< 20 years old, (20-34 years old), 35-49 years old	2
Sex	Male, (Female)	1
Mother's Education	(No or Primary), Secondary or Higher	1
Household Wealth Index	Poor, Middle, (Wealthy)	2
Mother's Physical Work	Non-Physical Work, Physical Work, (Not Working)	2
Mother's Smoking Habit	Active, Passive, (Not Smoking)	2
Iron Supplement	Iron Supp., (No Iron Supp.), No Information	1
Antenatal Care	Medic, Traditional, (No Antenatal), No Information	2
Water Source	Protected, (Unprotected)	1
Total	**43 categories**	**29**

Weight Status and Estimated Size was proxy variables which were considered had an association to LBW cases. Weight status is information from where the weight measurement of the baby is acquired, birth record or mother's recall. Estimated size is depend solely from mother's verdict about the baby's size at birth. Category of 'No Information' on variables of Pregnancy Complication, Iron Supplement and Antenatal Care is identical. It was a result from questionnaire design that those information was only inquired for the last child.

RESULT AND DISCUSSION

The result for each model's parameters estimations is compared below, with significant categories presented in **bold** value. For all models, Weight status, Terminated History, Sex, Mother's Smoking Habit, and Antenatal Care were not significantly affected the case of LBW. Aggregation to mother level and household level did not have much effect to the model. It can be seen that significance from each category is quite similar. It seemed that patterns of data from individual, mother to household levels did not change much, as the number of observation from one level to one above also only decreased slightly. Significant variables on these levels were Estimated Size, Twin, Preceding Birth and Birth Order, Pregnancy Complications, Mother's Age, Education, and Water Source. While Mother's Work and Household Wealth were only significant on individual and mother level.

On the other hand, model from census block level was different with fewer significant estimates. It seemed that the aggregation which decreased the number of observation rapidly (which also decreased degrees of freedom of the model) distorted the effects of the variables to the LBW case as well. Significant variables on census block level were only Estimated Size, Twin, Education and Water Source.

Based on model from each level, estimation of provincial LBW rates can be calculated from each observation predicted probability. Contrast to the model, estimates from model on census block level produced most accurate estimates almost on every province (closest estimates to the direct are **bolded**). However, the variances of these estimates were consistently increasing along with the process of aggregation. The precision of the estimates thus become more unreliable. Variance of estimates on Census Block aggregate level are significantly higher than the others. The precision of the estimates thus became quite unreliable. But still, in the case of the data, the closest estimates that produced by census block level must be taken into consideration.

Table 2. Comparison of Parameters Estimates

Categories	Individual	Mother	Household	Census Block
Intercept	-3.8938	-3.8275	-3.6765	-3.0397
Recall	0.06	0.0666	0.0902	-0.0128
Larger than Average	-1.6544	-1.1695	-1.0935	-0.2832
Smaller than Average	3.0905	3.1872	3.1602	4.2496
Very Large	-2.9413	-2.2785	-1.9509	-0.0377
Very Small	4.9246	5.1432	5.144	6.8085
Twin +	2.6168	2.3595	2.358	2.9353
2 yrs + and 4th +	-0.0389	-0.0565	-0.0357	0.0531
< 2 yrs and 2nd - 3rd	0.6522	0.9313	0.9643	0.459
< 2 yrs and 4th +	0.5926	0.6188	0.5253	0.3085
Firstborn	0.2495	0.2434	0.2583	0.3296
No Information	-0.0476	0.412	0.2276	0.2981
Other Pregnancy Complication	0.348	0.3043	0.2932	0.2004
Premature	0.4871	0.4406	0.4309	-0.189
Terminated History	-0.0701	-0.0527	-0.0971	-0.0171
35-49 yrs	0.2472	0.2807	0.2854	0.2557
< 20 yrs	0.1381	0.173	0.1862	0.3404
Male	0.0555	0.0444	0.05	-0.0504
Secondary or Higher	-0.4606	-0.4354	-0.4502	-0.3806
Middle	0.0529	0.0238	-0.0116	-0.00438
Poor	0.3033	0.2775	0.2317	0.2512
Unprotected Water	0.2364	0.232	0.2621	-0.2675
Non Physical Work	0.2369	0.2242	0.2064	-0.1555
Physical Work	0.0333	0.00617	-0.00086	-0.1532
Active	0.0184	0.0694	0.1184	-0.4953
Passive	0.0539	0.0508	0.0767	-0.198
Iron Suppl.	-0.1502	-0.1336	-0.0874	0.3586
No Information	0			
Medic Antenatal Care	-0.0063	-0.1353	-0.2776	-0.1705
No Information	0			
Traditional Antenatal Care	0.6556	0.2477	-0.0102	-0.5973

Table 3. Comparison of Provincial Rate Estimates and Variances

Province	Obs.	Direct	Indivi-dual	Mother	House-hold	Census Block
Aceh	507	7.10	8.05	8.18	8.19	**7.76**
North Sumatera	670	5.07	6.83	6.69	6.69	**6.01**
West Sumatera	507	4.73	5.99	5.83	5.77	**5.74**
Riau	574	5.05	6.88	6.77	6.82	**6.39**
Jambi	362	5.25	**8.14**	8.24	8.21	8.32
South Sumatera	523	6.31	6.80	6.60	6.77	**6.35**
Bengkulu	323	5.57	6.58	6.52	6.58	**6.43**
Lampung	461	6.29	6.03	6.15	**6.25**	6.14
Bangka Belitung	423	5.67	6.87	6.93	7.06	**6.60**
Riau Islands	421	5.46	5.30	**5.33**	5.24	5.26
Jakarta	777	4.63	5.72	5.89	5.97	**5.56**
West Java	753	6.77	7.64	7.58	7.61	**7.47**
Central Java	626	6.87	**6.94**	7.01	7.01	6.69
Yogyakarta	441	9.52	6.61	6.75	6.80	**7.39**
East Java	596	8.56	**8.54**	8.79	8.78	9.41
Banten	660	8.79	6.15	6.19	6.19	**6.43**
Bali	484	6.40	6.24	6.09	**6.26**	6.19
West Nusa Tenggara	491	10.79	9.12	9.20	9.25	**9.50**
East Nusa Tenggara	388	15.21	**9.83**	9.77	9.68	9.61
West Kalimantan	451	8.43	6.37	6.42	6.46	**6.69**
Central Kalimantan	349	5.73	7.51	7.42	7.34	**7.05**
South Kalimantan	426	7.04	7.14	6.93	**6.98**	7.62
East Kalimantan	416	5.53	**6.38**	6.55	6.45	6.62
North Sulawesi	435	7.82	**9.61**	9.82	9.71	10.37
Central Sulawesi	377	13.53	**8.91**	8.35	8.22	8.31
South Sulawesi	547	8.59	9.43	9.47	9.24	**9.19**
Southeast Sulawesi	358	5.03	9.00	8.93	8.99	**8.61**
Gorontalo	334	12.57	10.18	9.89	9.77	**10.80**
West Sulawesi	306	10.78	7.99	8.27	8.18	**8.57**
Maluku	292	5.48	5.64	**5.45**	5.43	6.63
North Maluku	298	6.71	6.59	6.78	**6.74**	6.51
West Papua	358	8.66	8.55	**8.66**	8.50	8.26
Papua	190	6.84	6.11	6.00	6.13	**6.27**

Because each process of aggregation produce a different dataset with a different size, comparison of the model cannot be don straightforward. Some goodness of fit tests results are not comparable. A comparable measurement can be used in this condition is area under curve from Receiver Operating Characteristics (ROC) curve. The result showed that on Census Block level the area coverage decreased drastically which means that the model became much less unreliable (Figure 1.).

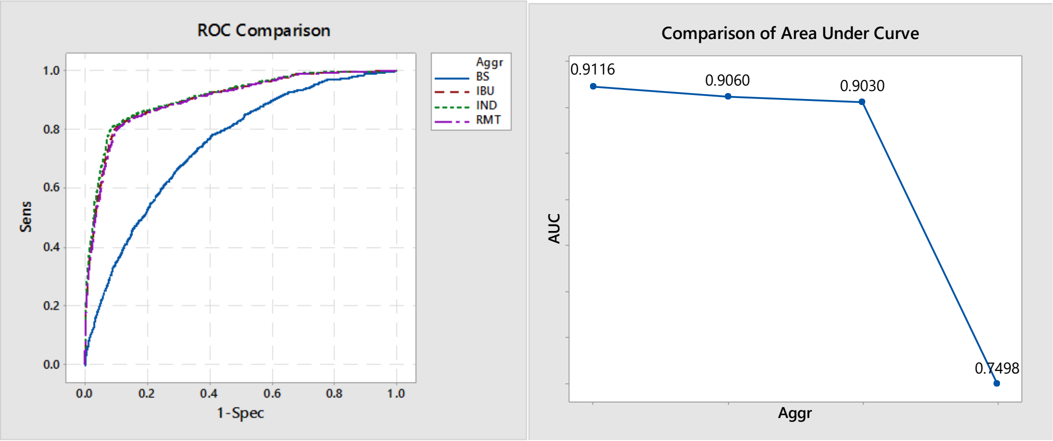

Figure 1. Comparison of ROC and Area under Curve

CONCLUSION

Effects of a variable to LBW case on the model can be distorted due to process of aggregation. The direct effect on individual level systematically decreases along with the increasing of aggregate size, thus decreasing of number of observation which causing higher variances to the estimates and lower significant level. Due to aggregation direct effect of a category was also mixed up with other categories on the same variables as presented in a form of proportion, which could be distorted the effect even further. A study on how much an effect will be distorted due to aggregation on some different certain conditions is recommended.

The insignificance of some theoretically associated categories such as birth order, sex, terminated history and young mother and active smoker may be a result of addition of some proxy variables which are put to bring more estimating power to the models. The effect of these proxies in the models suppressed explained variances of these categories into insignificance. It is because the purpose of the study is to produce best estimates. To describe the relationship between these variables better these proxies should not be included to the models.

Although the cluster level model estimated provincial rates of LBW case more accurately, compared to those lower aggregate level models, the estimates was not reliable enough for their low precision measured by respective variances. It is certain as consequences of increasing of aggregate size, thus decreasing of number of observation (also degrees of freedom). However, addition of area effects which theoretically related to the LBW case besides the aggregates explanatories may add up extra estimating power to the model. Furthermore, dropping insignificant explanatories may give more degree of freedom which decrease the variance. Moreover it is a trade-off between accuracy and precision so the decision must be made with caution.

REFERENCES

Bloom D., Canning D., and Sevilla J. 2003. *The Demographic Dividend: A New Perspective on the Economic Consequences of Population Change*. Santa Monica, CA (US): RAND Corporation.

Barker D.J.P. 1995. The Fetal and Infant Origins of Disease. *European Journal of Clinical Investigation*. 25: 457-463.

[BPS] Badan Pusat Statistik, [BKKBN] Badan Kependudukan dan Keluarga Berencana Nasional, Kementerian Kesehatan and ICF International. 2013. *Indonesia Demographic and Health Survey 2012*. Jakarta (ID). BPS.

Dobson A.J., 2002. *An Introduction to Generalized Linear Models*. 2nd Ed. New York (US). Chapman & Hall/CRC.

Firebaugh G. 1978. A Rule of Inferring Individual-Level Relationship from Aggregate Data. *American Sociological Review*. 43(4):557-572

Hosmer D.W., and Lemeshow S. 2000. *Applied Logistic Regression*. 2nd Ed. New York (US). John Wiley & Sons, Inc.

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.

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

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

Scheaffer R.L., Mendenhal W., Ott R.L. and Gerow K.G. 2006. *Elementary Survey Sampling*. Ed. ke-6. Belmont, CA (US). Duxbury Press

Steel D.G., Tranmer M., and Holt D. 1997. Logistic regression analysis with aggregate data: Tackling the ecological fallacy. *Proceedings of the Survey Method Research Section of the ASA*. 324-329.

[UN] United Nations. 2014. *The Millennium Development Goals Report 2014*. New York (US). United Nations.

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

[WHO] World Health Organization. 2011. *International Statistical Classification of Diseases and Health Related Problems*. Ed ke-10. Geneva (CH). WHO.

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