Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2021, 2021:41
https://doi.org/10.28919/cmbn/5596
ISSN: 2052-2541

AN EVALUATION OF THE LOG-TRANSFORMED STRATEGY FOR COUNT DATA IN ECOLOGICAL STUDIES

ANNA CHADIDJAH[1,*], I.G.N.M. JAYA[1,2]

[1]Department Statistics, Universitas Padjadjaran, Indonesia

[2]Faculty Spatial Science, Groningen University, The Netherlands

Copyright © 2021 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract. Count data are found in Ecological studies. The log-transformed strategy is commonly used in the count or rate data. The rate data are defined as the count data divided by a scale variable such as population at risk or an expected count. The log-transformed strategy is used to satisfy the parametric approach and simplify the model estimation. However, this strategy is not correct. The parameter estimation based on the log-transformed strategy could produce a biased estimate with a high standard error estimate. In this study, we are interested in evaluating the bias of parameter estimates based on the log-transformed strategy on the linear regression model. The generalized linear models have better performance in dealing with count data. However, some practitioners who are more familiar with the linear regression model prefer to use a log-transformed strategy and handle the zero cases by adding small values to zero observations. Simulation data from a Poisson distribution were used to compare the Poisson regression model and the linear regression model combined with the log-transformed strategy. The models were evaluated based on the bias and the root-mean-squared error statistics. We found that the linear regression with log-transformation strategy provided a high bias and a small value of root-mean-squared error, especially for small sample size and a

*Corresponding author
E-mail address: anna.chadidjah@unpad.ac.id
Received February 23, 2021

small value of the count data. We also use real data set to explore more detail the uses of log-transformed strategy and compare it with Poisson regression.

Keywords: count data; log-transformed; Poisson; bias; simulation.

2010 AMS Subject Classification: 37M05.

1. INTRODUCTION

Epidemiological and ecological data often deal with discrete count data, for example the number of disease incidence. The incidence rate is basically counted data by including a scale variable so that the scaled count data becomes the rate [1]. The population at risk or expected count are the scale variables that commonly considered in epidemiology [2-4]. The log-transformed strategy was commonly recommended to normalize count data as a part of the analysis with the parametric model [5, 6]. However, there is no clear explanation of the log-transformed method so that the normal linear regression analysis is often applied. The argument only focuses on correcting the variance and obtains the linear relationship between the response and predictors variables. They did not address the bias estimate that might appear if this approach was implemented. However, this strategy is not totally correct. The parameter estimation based on the log-transformed strategy may be biased and the research conclusions may be misleading. In this study, we are interested to evaluate the bias of parameter estimate based on the log-transformed strategy under the linear regression model framework. The generalized linear models are known to be suitable models for dealing with count data. However, some practitioners often use the linear regression model for analyzing count data by applying the log-transformation and adding a small number on zero value. To address this problem, simulation data from a Poisson distribution are used to compare the performances between the Poisson regression model and the linear regression model with a log-transformed strategy. We evaluate the average count and rate models. The models are then evaluated based on the bias and the root-mean-squared error criteria.

2. METHOD

2.1 A Generalized linear model (GLM): Poisson regression

A GLM is an extension of the standard linear model [7, 8]. The basic idea of the GLM is a modelling the expected value of the response variable with the predictor linear using a specific link function. A GLM extends the standard linear regression model, in which the response variable does not follow the normal distribution again but follows other distributions, such as the count response variable with the Poisson distribution. Then a function of the linear predictor is applied on the expectation of the response variable.

2.2. Average count data model

Here we assume the count data y_i follow Poisson distribution with mean and variance λ_i and can be defined as [2-4]:

$$y_i \sim Poisson(\lambda_i) \quad (1)$$

Using a log linear link function, the simple Poisson regression of the average count data is presented as:

$$\log(\lambda_i) = \alpha + \beta x_i \quad (2)$$

where α is an intercept and β denotes the slope of covariate. This is equivalent to a multiplicative model for λ_i:

$$\lambda_i = e^{\alpha+\beta x_i} = e^{\alpha}(e^{x_i})^{\beta} \quad (3)$$

2.3 Rate model

If we are interested in estimating the rate of individual, then we can scale the count data y_i using the scale variable (N_i) such as population or expected count [1]:

$$\lambda_i = r_i N_i \quad (4)$$

where the log linear model is:

$$\log(\lambda_i) = \alpha + \beta x_i + offset(\log(N_i)) \quad (5)$$

where $r_i = y_i/N_i$. An $offset(\log(N_i))$ is used to change a scale and defined as a correction factor in the model specification. It is assumed to have a regression coefficient of 1. The offset represents the denominator of the rate and it is included to the regression on the logarithmic scale.

In model (5), the slope β can be interpreted as the risk scale rather than the absolute scale. The exponentiating intercept (i.e., $\exp(\alpha)$) explains the overall relative risk and the exponential of β represents the change in the relative risk for one unit change in the predictor x. In many studies of epidemiology, the interest focuses on the rates or relative risks rather than on the average number of counts. Some papers were found using a normal linear regression approach to modeling the rate variables [9, 10].

3. SIMULATION STUDY

Data generating process (DGP) was based on a Poisson regression model with a single covariate of x. The covariate values were generated from a normal distribution with μ={0, 5, 10}. The variation of μ controls the size of λ. A small value of μ is corresponding to a small value of λ and y. We consider four different sample size with n = {10, 50, 100, 500} represent to small, medium, large, and extra-large sample size. In order to evaluate the rate model, we simulated the number of populations N_i between 1000 until 10,000 inhabitants. One thousand replicates simulations were carried out for each parameter value. Based on this simulation design, we have 12 data sets for each count and rate models. The regression coefficients were fixed to $\alpha = 1$ and $\beta = 1$.

The data were modeled using the Poisson and ordinary least square regression. The performances of the two models were then compared with the bias and root mean square error (RMSE) of the parameter estimate. Here we assume there are no non linearity, heteroskedasticity and overdispersion problems. In this simulation study, we focus on the evaluation of the relationship between the covariate and the response variable. The simulation evaluated by comparing the mean of biases, B:

$$B = \frac{1}{S}\sum_{s=1}^{S}\left(\hat{\beta}_1 - \beta_1\right) \tag{6}$$

and the root mean-squared error (RMSE) is given as follows

$$RMSE = \sqrt{\left(\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\beta}_1 - \beta_1\right)^2\right)} \tag{7}$$

where $\hat{\beta}_1$ denotes the estimated slope regression coefficient from the Poisson or linear regression models, β_1 is the true value of the slope coefficient and S denotes the number of simulation. Note that, we estimated the model on the log scale model, for count models: $\log(\lambda) = \alpha + \beta x$ and $\log(y) = \alpha + \beta x + \varepsilon$ and for the rate models: $\log(\lambda) = \alpha + \beta x + offset(\log(N))$ and $\log(y/N) = \alpha + \beta x + \varepsilon$. Simulations and analyses were done in the R statistical program using stats packages. The codes used are available by request.

The biases and root-mean-squared error for the different models are presented in Tables 1 and 2. The surface plots display in Fig.1 and Fig.2. The Poisson regression with the average count data model has a lower bias and root mean squared error (RMSE) compared to the normal linear regression, especially for small average data values and sample size. For the rate model, the Poisson regression model performs better for all conditions. The normal regression model performs a good result only for a large average of count data. Increasing sample size does not reduce the bias and mean squared error estimates.

Table 1. Bias estimates of the regression coefficient (slope) for the average count model

n	μ	Poisson Regression		Log Transformation	
		$B(\beta_0)$	$B(\beta_1)$	$B(\beta_0)$	$B(\beta_1)$
10	0	-0.033	0.125	-0.425	1.151
50	0	-0.009	-0.017	-0.446	0.803
100	0	-0.004	-0.015	-0.442	0.677
500	0	-0.002	0.003	-0.446	0.754
10	5	-0.009	0.002	-0.010	0.002
50	5	-0.027	0.005	-0.032	0.006
100	5	0.011	-0.002	0.004	-0.001
500	5	0.000	0.000	-0.008	0.001
10	10	0.000	0.000	0.000	0.000
50	10	-0.003	0.000	-0.003	0.000
100	10	0.002	0.000	0.002	0.000
500	10	0.000	0.000	0.000	0.000

Table 2. Bias estimates of the regression coefficient (slope) for the rate model

n	μ	Poisson Regression		Log Transformation	
		$B(\beta_0)$	$B(\beta_1)$	$B(\beta_0)$	$B(\beta_1)$
10	0	-8.618	0.199	-5.539	-0.926
50	0	-8.617	0.000	-5.540	-0.938
100	0	-8.615	0.026	-5.540	-0.936
500	0	-8.615	0.009	-5.540	-0.938
10	5	-8.701	0.024	-7.666	-0.134
50	5	-8.713	0.022	-7.762	-0.115
100	5	-8.608	0.000	-7.784	-0.111
500	5	-8.647	0.007	-7.784	-0.111
10	10	-10.398	0.181	-9.969	0.150
50	10	-8.446	-0.016	-8.026	-0.044
100	10	-8.658	0.005	-8.416	-0.005
500	10	-8.745	0.013	-8.611	0.015

Tables 1 and 2 clearly show that the Poisson regression model work outperforms in comparison with the normal linear regression model for the average count and rate models. The idea of dividing the count data with the scale variable of the population size in order to be the continuous data before applying the normal linear regression is not a good idea. The normal linear regression model has high bias except for very large average count data. The clear presentation for this result can be seen in Figure 1(b). The interesting point is the negative bias for the rate model which indicates that the regression coefficient based on the normal linear regression for the rate model tends is underestimate.

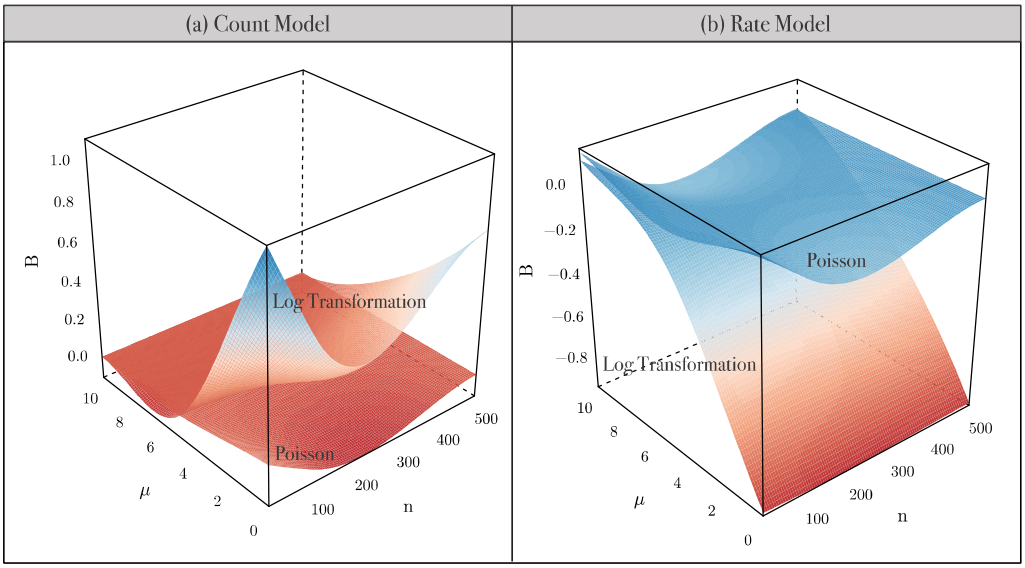

Figure 1. Bias estimates of the regression coefficient (slope) for (a) count model and (b) rate model

Table 3. Root mean squared error estimate (RMSE) of the regression coefficient (slope) for count model

n	μ	Poisson Regression		Log Transformation	
		$RMSE(\beta_0)$	$RMSE(\beta_1)$	$RMSE(\beta_0)$	$RMSE(\beta_1)$
10	0	0.2184	2.3325	0.6366	5.8183
50	0	0.0905	0.8956	0.4892	2.3524
100	0	0.0584	0.6224	0.4655	1.7043
500	0	0.0278	0.2722	0.4510	1.0095
10	5	0.9508	0.1900	0.9573	0.1913
50	5	0.3746	0.0748	0.3766	0.0751
100	5	0.2617	0.0523	0.2630	0.0525
500	5	0.1111	0.0222	0.1131	0.0226
10	10	0.1528	0.0153	0.1536	0.0153
50	10	0.0589	0.0059	0.0593	0.0059
100	10	0.0409	0.0041	0.0412	0.0041
500	10	0.0179	0.0018	0.0182	0.0018

Table 4. Root mean squared error estimate (RMSE) of the regression coefficient (slope) for rate model

n	μ	Poisson Regression		Log Transformation	
		$RMSE(\beta_0)$	$RMSE(\beta_1)$	$RMSE(\beta_0)$	$RMSE(\beta_1)$
10	0	8.6221	3.2199	5.5394	0.9577
50	0	8.6174	1.1971	5.5397	0.9428
100	0	8.6157	0.8026	5.5398	0.9381
500	0	8.6147	0.3435	5.5398	0.9384
10	5	12.9845	1.9289	12.3301	1.9367
50	5	9.4714	0.7438	8.6761	0.7846
100	5	8.9333	0.4772	8.2128	0.5351
500	5	8.7125	0.2138	7.8722	0.2600
10	10	22.4285	1.9961	23.9013	2.1776
50	10	11.0353	0.7100	11.6966	0.8516
100	10	9.9708	0.4942	10.3693	0.6053
500	10	9.0109	0.2179	9.0091	0.2652

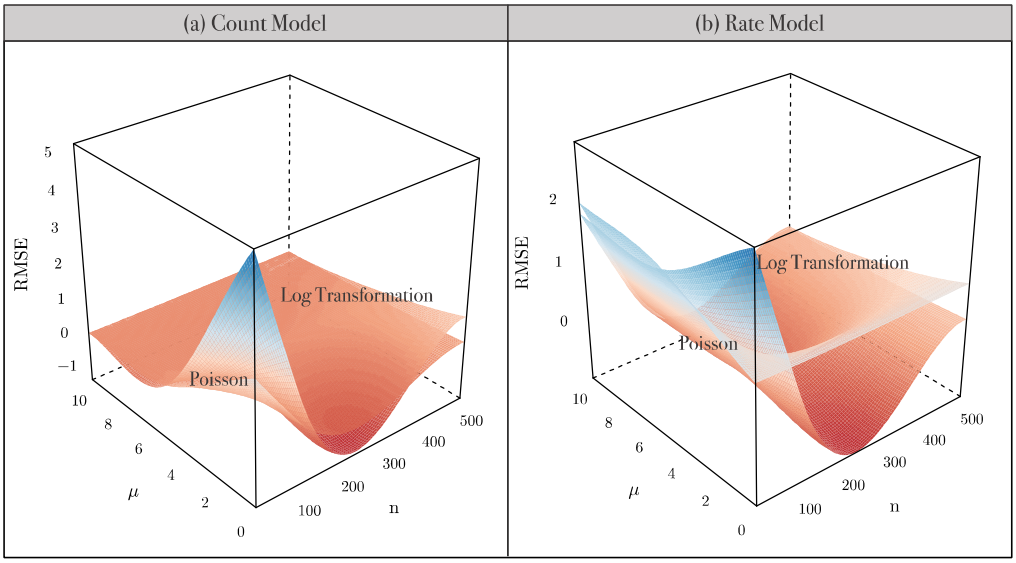

Figure 2. Root mean squared error estimate (RMSE) of the regression coefficient (slope) for (a) count model and (b) rate model

Figure 2 shows Poisson model has lowest RMSE than log-transformation. It indicates that the Poisson model provides a better predictive performance.

4. APPLICATION

Modeling lymphatic filariasis in West Java, Indonesia

Lymphatic filariasis (LF) is one of serious health problem in West Java Indonesia especially in Bogor municipality. LF is an infectious disease caused by filarial worms *Wuchereria bancrofti*, *Brugia malayi*, and *B. timori* [11]. In order to evaluate log-transformation approach and Poisson log-linear model, we model number of cases and rate of LF as dependent variables and health behavior index as predictor. The data LF for period 2016-2018 were obtained from West Java official health with 27 districts. Figure 3 shows the spatial distribution of variables interest and Table 5 shows the estimation results.

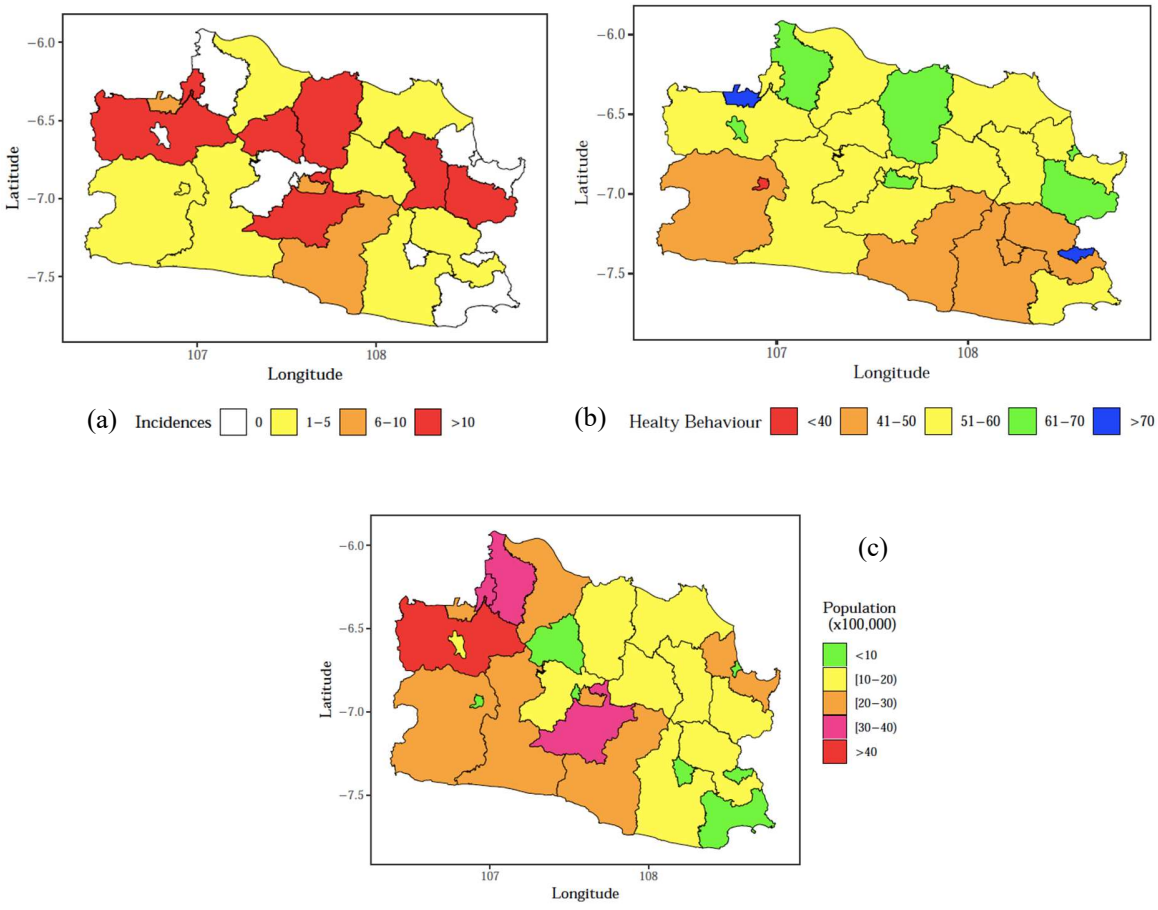

Figure 3. Spatial distribution of (a) Total incidences of FL, (b) Healthy Behavior Index, and (c) Population at Risk, 2016-2018

Table 5. Log-transformation versus Poisson Regression

Model	Estimate	Log Transformation		Poisson Regression	
		Estimate	SE	Estimate	SE
Count	Intercept	2.179	1.869	2.721	0.427
	Slope	-0.019	0.033	-0.011	0.008
Rate	Intercept	-7.820	1.601	-11.678	0.482
	Slope	-0.011	0.028	-0.011	0.009

Table 5 shows the log-transformation approach and Poisson regression have different slop for count model and similar for rate model. However, the intercept of log-transformation and Poisson regressions are different for both models. It indicates that log-transformation can be used for rate model for explain the effect of the risk factors. However, we have to be aware if the objective of the study is evaluate the rate prediction because it has different intercept. The log-transformation approach may be provide the under or overestimate prediction.

5. DISCUSSION AND CONCLUSION

A log-transformed strategy is commonly used when the error structure of data is simple. We can improve the ability of a model fits to the data by correcting variance and make the relationship closes to linear [12, 13] before applying a simple linear regression. However, we have to be more careful especially for count data that were characterized by small values with a lot of zero and small sample size. Different models could be used, and Generalized Linear Model with Poisson or Negative Binomial models can be the best alternatives [7, 14]. For count data, which is modeled by original data (i.e., on the average number of counts) or rate (i.e., scaling by an offset to correct the heterogeneity in the data) [1], our simulation results suggest that the log-transformed strategy on the average count model performs poorly for a small value of count data and for small sample size. It is getting better for a large average of count data and large sample size. For the rate model, an increasing number of observations does not reduce bias. It works to be good if the data have a large average. An additional problem with the regression of transformed variables is that it can

lead to impossible predictions such as the negative number of individuals.

It is may not be really surprising that fitting the model using Poisson regression gives the best result; what is more interesting is that the log-transformed strategy provides a similar result with Poisson regression in terms of the slope of the regression coefficient for a large value of λ. However, we already knew that, for Poisson distribution phenomenon, the λ tends to small. We therefore suggest to fit the count data with the Poisson regression model. The Poisson model provide a more accurate foundation for the model due to the ecological data follow Poisson process.

This is a simple study to answer the question " do we have to use Poisson regression for any kind of count data?". Our study showed that the Poisson regression model is more appropriate than the linear regression model based on the valuation of bias and root mean squared error estimates for all parameters simulation. However, modeling is not only about bias and mean squared error estimates. In practice, we have to consider the linearity, heteroskedasticity, overdispersion, and also autocorrelation assumptions that might cause the Poisson regression model is hard to be applied and needs some modification. This situation can be fixed if there is a user-friendly software that can help practitioners to apply the right model. Now, the R statistical program facilitates the complex model computations as the new packages are developed rapidly by many scientists around the world. However, for many practitioners, working on the syntax is less convenient compare to the GUI program.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] M. Blangiardo, M. Cameletti, Spatial and spatio-temporal Bayesian models with R-INLA. Wiley, Hoboken, (2015).

[2] I.G.N.M. Jaya, H. Folmer, B.N. Ruchjana, F. Kristiani, Y. Andriyana, Modeling of Infectious Diseases: A Core Research Topic for the Next Hundred Years, in: R. Jackson, P. Schaeffer (Eds.), Regional Research Frontiers - Vol. 2, Springer International Publishing, Cham, 2017: pp. 239–255.

[3] I.G.N.M. Jaya, H. Folmer, Bayesian spatiotemporal mapping of relative dengue disease risk in Bandung, Indonesia, J. Geogr. Syst. 22 (2020), 105–142.

[4] I.G.N.M. Jaya, H. Folmer, Identifying Spatiotemporal Clusters by Means of Agglomerative Hierarchical Clustering and Bayesian Regression Analysis with Spatiotemporally Varying Coefficients: Methodology and Application to Dengue Disease in Bandung, Indonesia, Geogr Anal. (2020), https://doi.org/10.1111/gean.12264.

[5] J. Maindonald, J. Braun, Data analysis and graphics using R—an example-based approach. Cambridge University Press, Cambridge, (2007).

[6] J.H. Zar, Biostatistical Analysis, 4th edn. Prentice Hall, New Jersey, (1999).

[7] R.B. O'Hara, How to Make Models Add Up - A Primer on GLMMs, Ann. Zool. Fenn. 46 (2009), 124–137.

[8] R. O'Hara, J. Kotze, Do not log-transform count data, Nat Prec. (2010). https://doi.org/10.1038/npre.2010.4136.1.

[9] A.A. Godana, S.M. Mwalili, G.O. Orwa, Dynamic spatiotemporal modeling of the infected rate of visceral leishmaniasis in human in an endemic area of Amhara regional state, Ethiopia, PLoS ONE. 14 (2019), e0212934.

[10] H. Mokhort, Multiple Linear Regression Model of Meningococcal Disease in Ukraine: 1992–2015, Comput. Math. Meth. Med. 2020 (2020), 5105120.

[11] P. Ginandjar, L.D. Saraswati, D. Suparyanto, M. Sakundaro, T. Supali, The prevalence of lymphatic filariasis in elementary school children living in endemic areas: a baseline survey prior to mass drug administration in Pekalongan district-Indonesia, Iran. J. Public Health, 47 (2018), 1484-1492.

[12] D.L. Miller, R. Glennie, A.E. Seaton, Understanding the Stochastic Partial Differential Equation Approach to Smoothing, J. Agric. Biol. Environ. Stat. 25 (2020), 1–16.

[13] H.-P. Piepho, Data transformation in statistical analysis of field trials with changing treatment variance, Agron. J. 101 (2009), 865–869.

[14] P. McCullagh, J. Nelder, Generalized linear models, 2nd edn, Chapman & Hall, London, 1989.