Multilevel models approach on longitudinal studies for disease data

Bertho Tantular¹, I Gede Nyoman Mindra Jaya¹, Zulhanif⁴

¹Department of Statistics, Universitas Padjadjaran
Jl. Raya Bandung-Sumedang KM 21, Sumedang, Indonesia

E-mail: bertho@unpad.ac.id

Abstract. A study of a disease in an area is periodically measured. In this study, time is nested within the individual so it derives a data with nested structure. Ordinary regression model is not appropriate used because it does not consider the existence of nesting effects. Multilevel model usually used for hierarchical data, because it is not ignore the effect of the data structure. According to Sneijder and Bosker, nested data can be viewed as hierarchical data, so it can be analyzed using multilevel model. Diarrhea case in an area is affected by clean water, hand washing habits, healthy latrines and healthy house. This data is measured periodically from 2012 to 2016. Multilevel model is applied to diarrhea case data in Bandung city which has nested structure. The results obtained from the analysis is that the data used has a hierarchy structure so that it can be analyzed using a multilevel model. In addition, the random intercept model used is very fit with three significant variables. In general, the result for this case is that the occurrence of cases of diarrhea in Bandung influenced by clean water and hand wash habits with the variation between districts.

1. Introduction

In the survey research there are times when the data obtained from repeated measures. Recurrent measurement data has longitudinal data structures. Longitudinal data is commonly used when looking to investigate the correlation of individual changes. Longitudinal data is obtained when the units determine the response at different times. One of the things that can be obtained from longitudinal data is the chronology of a response in sequence. Longitudinal data also have the property that the data often consists of a large number of clusters of small size each.

The analysis of longitudinal data has the primary purpose of examining the covariate effects of each level in general on responses and effects in response changes over time. One of the important benefits of longitudinal data analysis is that it allows the separation of cross-sectional and repeated measurement effects. In addition, in the analysis of longitudinal data can also be examined the diversity between units both within the level of the response and in changes between time. The variation not obtained from the observed variation results in a dependency between responses as well after the covariates are controlled. This violates assumptions in ordinary linear regression models and must be overcome to avoid fallacy in inference.

In general, the measurement data repeatedly differentiated into balanced data and unbalanced data. A repeatable measurement data is called balanced data when all units are measured at the same time. Instead a data is called unbalanced data when units are measured at different times.
Logistic regression model is a regression model for binary response. The parameter estimation for the logistic regression model using the Maximum Likelihood Estimation (MLE) method is based on sample data that meets independent and identical properties. In the longitudinal data, the independent and identical properties cannot be satisfied because of the correlation arising from repeated measurements. So, the logistic regression model for longitudinal data cannot be estimated using the maximum likelihood method because this method does not accommodate the correlation between observations.

Some researchers have made several approaches to analyzing logistic regression models for longitudinal data. A common approach is the generalized linear mixed model. In this model, the assumption of the response distribution, so that for correlated responses cannot be used because the correlation elements are not taken into account.

Another approach for analyzing longitudinal data is the multilevel model [7]. Multilevel models introduced by Goldstein (1995) are mentioned to address various problems arising from data with hierarchical structures including longitudinal data. In the multilevel model, the hierarchical structure is defined as the level. In general, the level used is not limited but in the repeated measurement data, the level used is only two that is at the lowest level of time called Level 1 and higher level that is called individual Level 2. Multilevel model in addition can determine the diversity between groups can also show the correlation between the two observations which on the other model is assumed does not exist. In addition, multilevel models can also measure interactions that may occur between variables at different levels. Like the generalized linear mixed model, multilevel models also require strict assumptions about the distribution of responses.

2. Longitudinal Analysis
The parameter estimation method for data containing correlation structures such as longitudinal data without requiring strict distribution assumption can use Generalized Estimating Equation. The advantage of this method is that the computational method for this method is simpler than the generalized linear model. In addition, this method also does not require the assumption of multivariate distribution. Another advantage is that the resulting estimator is consistent despite the specification errors in the correlation structure used and can be applied to incomplete data (missing data), with the missing observations being MCAR [4].

2.1 Generalized Estimating Equation
Generalized Estimating Equations (GEE), first introduced by Liang and Zeger in 1986, this method became popular in various fields of science. Hardin and Hilbe [4] suggest that although the Generalized Linear Model (GLM) is a powerful and flexible method to use, but the analysis is limited, the assumption of GLM is that each individual must be independent with the other individual. But on the longitudinal data, this assumption may be unfulfilled. The main advantage of using GEE lies in the results of consistent and unbiased parameter estimation, even when miss-specified correlation structures [3]. GEE is an adjustment to GLM when the data is correlated, then this method can be applied to various response variables that are often encountered in empirical applications (e.g., continuous data, polychotomous, dichotomous, and enumeration). Although GEE does not require data to follow a certain distribution, it assumes the variance of the response variable is described as an expectation function.

2.2 Working Correlation Matrix
One of the important things that differentiate GEE from other methods is the correlation structure that is included in its parameter estimation. This correlation structure is called the Working Correlation Matrix (WCM). Determination of WCM is very dependent on the correlation conditions that occur in the data used. According to Liang and Zeger (1986), GEE remains robust even though WCM used is not appropriate. Some of the WCM options that can be used are the Independent Structure,
Exchangeable Structure, Autoregresive Structure and Unstructure Correlation. (Swan, 2006). For the right longitudinal WCM data is the Autoregresive Structure.

\[
\mathbf{R}_i(\alpha) = \begin{bmatrix}
1 & \alpha & \alpha^2 & \ldots & \alpha^{N-1} \\
\alpha & 1 & \alpha & \ldots & \alpha^{N-2} \\
\alpha^2 & \alpha & 1 & \ldots & \alpha^{N-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\alpha^{N-1} & \alpha^{N-2} & \ldots & \alpha & 1
\end{bmatrix}
\] (1)

dengan: \( \alpha_t = \sum_{i=1}^{N} e_{it} e_{i,t+1}' / (NT - K) \) dan \( \alpha = \sum_{t=1}^{T-1} \beta_t / (T - 1) \)

2.3 Estimation Methods

The logistics model used for longitudinal data is as follows

\[
\logit Pr(Y_{hij=1}) = \logit \pi_{ij} = X_{ij}' \beta
\] (2)

For \( i = 1, 2, \ldots, n \) and \( j = 1, 2, \ldots, m_i \). With many repeatable measurements of unit \( i \), \( n \) are many units of observation, \( ij \) is the probability of success of unit \( i \) at time \( j \) and \( \beta \) is the regression coefficient.

To estimate the parameter in Eq. (2) must take into account the existence of WCM as in Equation (1) so that its parameter estimation cannot use Maximum Likelihood method. The method used is to maximize log-likelihood function based on the normal distribution even without the normality assumption of the response. The resulting likelihood function is called quasi-likelihood. GEE modifies the log-likelihood function by entering WCM into equations that take into account the correlation in the data [4]. The likelihood quasi-score equation is expressed as:

\[
U(\alpha, \beta) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \omega_{ij} D_{ij} D_{ij}' S_{ij} = 0
\] (3)

where: \( D_i = \frac{\partial \pi_{ij}}{\partial \beta} = A_{ij} X_{ij} \) dan \( V_{ij} = A_{ij}^{-1} R(\alpha) A_{ij}^{-1} \)

\[
A_{ij} = diag\{\pi_{ij}(1 - \pi_{ij})\} \text{ dan } S_{ij} = Y_{ij} - \pi_{ij}
\]

With \( R(\alpha) \) is WCM of \( Y_{i1}, \ldots, Y_{ij} \), the correlation structure for the longitudinal data uses the correlation structure of the Autoregression Structure as in Equation (1). If Equation (3) is lowered to \( \beta \) it will produce the equation:

\[
I = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \omega_{ij} D_{ij} D_{ij}' V_{ij}^{-1} D_{ij}
\]

where:

\[
\widehat{\mathbf{V}} = \frac{n}{n - 1} \sum_{i=1}^{n} (B_i - \overline{B})(B_i - \overline{B})'
\]

\[
B_i = \sum_{j=1}^{m_i} \omega_{ij} X_{ij}' \hat{\mathbf{A}}_{ij}^{-1/2} \hat{R}(\alpha)^{-1} \hat{A}_{ij}^{1/2} X_{ij} \text{ dan } \overline{B} = \frac{1}{n} \sum_{i=1}^{n} B_i
\]

Equation (1) is not linear in the \( \beta \) parameter, so it requires an optimization method to obtain the solution. The optimization method used to calculate \( \beta \) is the Fisher Scoring Method. Based on Dobson [2] Fisher scoring method equations for quasi-likelihood are as follows:

\[
\beta^{(t+1)} = \beta^{(t)} - \left( \sum_{i,j} \omega_{ij} D_{ij}' (Y_{ij} - \pi_{ij})^{-1} D_{ij} \right)^{-1} \left( \sum_{i,j} \omega_{ij} D_{ij}' V_{ij} (Y_{ij} - \pi_{ij})^{-1} (Y_{ij} - \pi_{ij}) \right)
\] (5)

The process is lit up until it is convergent. According to Agresti [1] the convergence measures that can be used are:
\[ |\beta(t+1) - \beta(t)| \leq |\beta(t)| + 10^{-6} < \epsilon = 10^{-8} \]  

(6)

For the iteration process the initial estimated value of \( \beta (t = 0) \) is obtained from the estimation using the usual logistic regression model using Maximum Likelihood Estimation without involving the correlation elements. Then in the next stage the estimated value is calculated by involving WCM.

3. Application

Diarrhea case in an area is affected by clean water, hand washing habits, healthy latrines and healthy house. The data used are secondary data sourced from Bandung Health Department at 30 District from 2012 until 2016. The results obtained from the analysis is that the data used has a hierarchy structure so that it can be analyzed using a multilevel model. The variables studied are the number of cases of Diarhea as a response, while the predictor is Clean water, Hand wash habits, Healthy latrines and healthy house in every district. Data on the number of cases of diarrhea obtained by the Bandung City Health Office came from the existing hospital in Bandung. The result of analysis this data refer to following tables.

| Table 1. Hierarchical structure test |
|-------------------------------------|
| Model | Log-lik | LRT  | DF  | p.value |
|-------|---------|------|-----|---------|
| Regression | -64141.287 |      |     |         |
| Multilevel  | -10937.497 | 106405.94 | 1 | 0       |

Table 1 shows hierarchical structure test of this data. Multilevel model has smaller deviance than regression model, this indicates that multilevel model appropriate to be use for this data.

| Table 2. Fixed effect and random effect estimate |
|-----------------------------------------------|
| Parameter | Estimate | Std.Error | Z   | p-value |
| Fixed      |          |           |     |         |
| Intercept  | 119.86390 | 23.72935  | 5.05| 0.000*  |
| Clean water | -0.80816 | 0.319270  | -2.53| 0.011*  |
| Hand wash habits | 1.38165 | 0.214311 | 6.45| 0.000*  |
| Healthy latrines | 0.09419 | 0.121140 | 0.78| 0.437   |
| Healthy house | -0.26497 | 0.154393 | -1.72| 0.086   |
| Random     |          |           |     |         |
| Sub-district | 105.4224 | 1.757040 |     |         |

Table 2 shows that only two variables are significance. They are Clean Water and Hand wash habits, it means that if clean water increase one unit so number of cases is decrease -0.80816 units, and if hand wash habits increase one unit so number of cases is decrease 1.38165. The variation between district is 105.4224 with standard error 1.757.

| Table 3. Goodness of fit test |
|------------------------------|
| Parameter | Estimate | Std.Error | Z   | p-value |
| Log-lik   |          |           |     |         |
| No Cases  | 1800     | 49.79     | 4   | 0       |

Table 3 shows Goodness of fit test and the result is quietly significant so multilevel model suitable to be use for this data.
4. Conclusion

Diarhea case data of Bandung City which have structured data or nested data so can be analysed using multilevel model. The result of analysis this data is that diarhea data has multilevel structure so multilevel model appropriate to be use for this data. The results show that only two variables significance are Clean Water and Hand wash habits. Variable clean water has negative coefficients that means if it is increase one unit so number of cases is decrease -0.80816 units. Variable hand wash habits has positive coefficients that means if its increase one unit so number of cases is decrease 1.38165. In general, the result shows that Goodness of fit test and the result is quietly significant so multilevel model suitable to be use for this data.

Acknowledgments

Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff or financial support from organizations should do so in an unnumbered Acknowledgments section immediately following the last numbered section of the paper.

References

[1] Alan Agresti. 2007. An introduction to categorical data analysis. 2nd edition. New York: John Wiley & Sons
[2] Dobson, Annette J. 2008. An Introduction to Generalized Linear Models 3rd edition. London. Chapman & Hall/CRC.
[3] Ghisletta, Paolo, and Spini, Dario. 2004. “An Introduction to Generalized Estimating Equations and an Application to Asses Selectivity Effects in a Longitudinal Study on Very Old Individuals”. Journal of Educational and Behavioral Statistics Winter, Vol.29, No.4, pp 421 – 437. University of Geneva and University of Laussane.
[4] Hardin, J. W., and Hilbe, J. M. 2012. Generalized Estimating Equations. Second Edition. United States of America: Chapman & Hall/CRC.
[5] Paul, Sudhir., Zhang, Xuemao., Xu, Jianwen. 2013. Estimation of Regression Parameters for Binary Longitudinal Data Using GEE: Review, Extension and an Application to Environmental Data. Journal of Environmental Statistics, Vol 4, Issue 1.
[6] Snijder, Tom A. B., Bosker, Roel J. 2012. Multilevel Analysis: An introduction to basic and advance multilevel modelling 2nd edition. SAGE Publications, London.
[7] Tantular, Bertho. 2012. Pendekatan Model Multilevel untuk Data Repeated Measures. Paper dipresentasikan pada Seminar Nasional Matematika dan Pendidikan FMIPA UNY, 10 November 2012.
[8] Tantular, Bertho, Mindra Jaya, I Gede Nyoman, Zulhanif. 2016. Pendekatan Model Multilevel pada Analisis Regresi Poisson untuk Data Longitudinal (Studi Kasus DBD Kota Bandung). Prosiding Seminar Nasional Matematika dan Pendidikan Jurusan Pendidikan Matematika FMIPA UNY, ISBN: 978-602-73403-1-2.