Generalized estimating equation (GEE) on binary longitudinal data

Devita Putri Mardyanti, Rohmatul Fajriyah
Statistic Department of Universitas Islam Indonesia, Yogyakarta, Indonesia

E-mail: 13611012@students.uii.ac.id

Abstract. Binary logistic regression is a regression in which the response variable is binary with one or more predictor variables being both categorical and continuous. It can be applied to longitudinal data, where the observations are cross-sectional units over several periods of time. These repeated observations cause autocorrelation and needs to be addressed by implementing the Generalized Estimating Equation (GEE) method. The study aims to apply GEE and choose the best correlation structure based on the QIC value, where the data is the sputum status of pulmonary tuberculosis patients at PKU Muhammadiyah Hospital at Bantul, Yogyakarta. Based on the analysis then the model of sputum status is $\text{logit}[\pi_i] = 2{,}017 + 1{,}491 job - 0{,}025$ time, with the correlation structure is unstructured.

1. Introduction

Binary logistic regression is used to determine the model of the relationship between a binary response variable with one or more predictor variables being categorical and continuous. This model can be applied to longitudinal data, i.e. data obtained from the observation of several cross-sectional units over a period of time. Repeated observation leads to autocorrelation and can be overcome by Generalized Estimating Equation (GEE) method. GEE is able to overcome the problem of autocorrelation by forming a correlation structure that describes the correlation in the data. This model is applied to sputum status of patients with pulmonary tuberculosis at PKU Muhammadiyah Hospital of Bantul, Yogyakarta, where 82 patients were observed during three time of treatments. The predictor variables are duration of treatment, age, sex, education, job and the response variables is patient sputum status. The purpose of this study is to model the data above and choose the best correlation structure based on the value of QIC.

2. Methodology

2.1 Longitudinal Data

Longitudinal data (balanced and unbalanced) is the result of observation on several cross-sectional units over a period of time (Twisk, 2003). The cross-sectional data comes from observations made on different individuals at any given time.

2.2 Generalized Linear Models (GLM)

GLM was first introduced by Nelder and Wedderburn in 1972, an expansion of the classical linear model, which is able to overcome the problem of response abnormalities (Kleinbaum and Klein, 2010). The three main components in GLM are:

1. The random component, that the response variable Y is mutually free.
2. A fixed component called a linear predictor in the form of:

$$\eta_i = \sum_{j=1}^{p} \beta x_{ij}, i = 1, \dots, n \tag{1}$$

Where x_{ij} is the predictor variable value of the *i* subject on the *j* predictor variable and β is the parameter of the *j* predictor variable.

3. The connecting function $g(\mu_i)$ connects a function of the mean value of a random component with a fixed component η_i that $\eta_i = g(\mu_i)$.

2.3 Binary Logistic Regression

Binary logistic regression is a data analysis method used to find the relationship between binary response modifiers (two categories) with predictor variables that are polychotomous (having nominal or ordinal scales with more than two categories) (Hosmer and Lemeshow in Octaviana, 2017). Binary logistic regression model can be written as follows:

$$\pi(x) = \frac{\exp(g(x))}{1 + \exp(g(x))} \tag{2}$$

$$g(x) = ln \frac{\pi(x)}{1 - \pi(x)} = \beta^0 + \beta^1 x_1 + \cdots + \beta x_p \tag{3}$$

The parameter estimation method used in binary logistic regression is Maximum Likelihood Estimator (MLE). If a certain X value produces more than one observation of Y and is classified into k category, then n_k is the number of observations in the k category. Further the likelihood for Y observation is:

$$\ell(\alpha, \beta) = \prod (P\ )^{yk} \tag{4}$$

Agresti (2002) explains that the Newton-Raphson iteration method for determining parameter estimators repeatedly until convergent at a certain value. The Newton-Raphson iteration method can complete the likelihood function, to determine the values of̂ and β by the iteration formula:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - (\boldsymbol{H}^{(t)})^{-1}\boldsymbol{g}^{(t)} \tag{5}$$

Simultaneous and partial test parameters are performed to determine whether the predictor variable value determines the response variable value. Partial test using the Wald test statistic while the simultaneously use the likelihood ratio test.

2.4 Generalized Estimating Equation (GEE)

Liang and Zeger (1986) introduced Generalized Estimating Equation as a method of GLM development to estimate model parameters based on data containing autocorrelation and did not spread normally. The general GEE equation is:

$$g(\mu) = \boldsymbol{X'\beta} \tag{6}$$

Hedeker and Gibbons (2006) explain that the parameter estimator β is obtained by completing the quasi score function (*S(β)*) called Generalized Estimating Equation:

$$\boldsymbol{S(\beta)} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{D^T V_i^{-1}(Y_i - \mu_i) = 0} \tag{7}$$

In GEE, the correlation is formed in a $\boldsymbol{R(\alpha)}$ correlation matrix sized $n \times n$, where the correlation structure is unknown and should be expected. Kleinbaum and Klein (2010) present five correlation structures namely independent, exchangeable, autoregressive (1), m-dependent, and unstructured.

Hedeker and Gibbons (2006) describe the matrix of cov(β) parameters to select the correlation structure and parameter testing. The matrix estimator cov(β) has two types, namely Naive/model-based estimator and Robust/empirical/sandwich estimator.

Parameter testing on GEE is done simultaneously or partially. Simultaneous testing of parameters used Generalized Wald test and partially by Wald test.

Swan (2006) provides guidance on the selection of correlation structures, if more than one possibility exists, then Quasi-likelihood under the independence Information Criterion (QIC) is used:

$$QIC = -2Q(\beta) + 2(\boldsymbol{V_m^-(\beta) V_e(\beta)}) \tag{8}$$

The model with the smallest QIC value is the model with the best correlation structure.

3. Results and Discussion

3.1 Binary Logistic Regression Modeling

The results of partial logistic regression model test shows that intercept, duration of treatment and patient's job had an effect on the sputum status of patients with pulmonary tuberculosis, while age, sex and education were not the determinants. The simultaneous test show that the three predictor variables have an effect on the sputum status of pulmonary tuberculosis patients. The logistics model is:

$$logit[\pi_i] = 2{,}402 - 0{,}005\ age + 0{,}579\ sex - 0{,}328\ education + 3{,}693\ job - 0{,}038\ time \quad (9)$$

3.2 Selection the Best Correlation Structure

The purpose of selecting the correlation structure is to obtain the most efficient parameter estimator, then use Quasi-likelihood under the independence Information Criterion to obtain the best correlation structure that describes the correlation in the data. The use of QIC criteria for Quasi likelihood function is able to overcome the overdispersion problem that may occur in Binomial and Poisson spread data. The QIC value of each correlation structure is presented in Table 1.

Table 1. QIC Value

Correlation Structure	QIC
Independent	242,14
Exchangeable	241,32
Unstructured	241,08
Autoregressive (1)	242,17

Table 1 shows the smallest value of QIC is 241,08 which is the unstructured correlation structure. The unstructured correlation structure shows that the model have no assumptions about the correlation.

3.3 Generalized Estimating Equation (GEE) Modeling

The GEE parameter estimation is done through 3 steps, estimation of the starting point parameter, the formation of the correlation matrix structure and the convergence parameter approximation iteration. The selection of correlation structures is based on the smallest QIC values resulting in an autoregressive (1) correlation structure that best describes the correlation of the data. Estimation of GEE parameter starting point is presented in the following table:

Table 2. Initial Point Estimation and GEE Parameters

Parameter	Coefficient (point estimation)	Coefficient (GEE parameter)
Intercept	2,402	2,338
Age	-0,005	-0,003
Sex	0,579	0,706
Education	-0,328	-0,117
Job	3,693	3,541
Time	-0,038	-0,038

Partial test of GEE parameter estimation using Wald test statistic shows that intercept, job and duration of patient's treatment have an effect on patient's recovery based on patient's sputum status, while age, sex and education are not.

After all the analysis steps are performed, a GEE model based on the Unstructured correlation structure is:

$$logit[\pi_i] = 2{,}017 + 1{,}491 - 0{,}025 time \quad (10)$$

Table 3. Estimation Result of Odds Ratio of Sputum Status

Predictors Variables	Odds Ratio	Interpretation
Job	4,442	Patients suffering from pulmonary tuberculosis with job, 4,442 times more at risk of positive sputum containing mycobacterium tuberculosis bacteria compared with patients who have not or do not have job.
Time	0,975	Each additional one-day duration of patient treatment would be 0,975 times more at risk of positive sputum containing mycobacterium tuberculosis bacteria with other variables considered constant.

4. Conclusion

Generalized Estimating Equation model for longitudinal data, where binary response is the sputum status of patients with pulmonary tuberculosis at PKU Muhammadiyah Hospital, Bantul Yogyakarta, is $logit[\pi_i] = 2{,}017 + 1{,}491 job - 0{,}025 time$ with Unstructured correlation structure based on the smallest QIC value.

5. References

[1] Agresti A 2002 *An Introduction to Categorical Data Analysis* (New York: John Wiley and Sons)

[2] Hedeker D and Gibbons R D 2006 *Longitudinal Data Analysis* (New York: John Wiley and Sons)

[3] Kleinbaum D G and Klein M 2010 *Logistic Regression A Self-Learning Text 3rd editions* (New York: Springer)

[4] Octaviana F A 2017 *Pemodelan Status Bekerja Ibu Rumah Tangga Menggunakan Model Multilevel dengan Respon Biner Tesis* (Surabaya: Institut Teknologi Sepuluh Nopember)

[5] Swan T 2006 *Generalized Estimating Equation when The Respons Variable Has a Tweedle Distribution*: *In Application for Multi-site Rainfall Modelling* (Toowoomba : Department of Mathematics and Computing of The University of Southern Queensland). Retreived https://core.ac.uk/download/pdf/11036965.pdf access date Aug, 13th 2017

[6] Twisk J 2003 *Applied Longitudinal Data Analysis for Epidemiology* (New York: Cambridge University Press)