Bangladesh J. Sci. Res. **29**(1): 1-9, 2016 (June)

GENERALIZED QUASI-LIKELIHOOD APPROACH FOR ANALYZING LONGITUDINAL COUNT DATA OF NUMBER OF VISITS TO A DIABETES HOSPITAL IN BANGLADESH

Kanchan K. Sen* and Taslim S. Mallick

Department of Statistics, Biostatistics & Informatics, University of Dhaka, Dhaka-1000, Bangladesh

Abstract

The generalized quasi-likelihood (GQL) estimation approach has been used to analyze the longitudinal data of four repeated count responses of 872 registered diabetic patients. The data on variables such as age, sex, body mass index, family history of diabetes (heredity), area of residence, education level and physical exercise are obtained. It was aimed at proposing the GQL approach for analyzing longitudinal count data and to determine the factors related to the visits of diabetic patients at hospital. The heredity, gender, area of residence, physical exercise and age < 40 years are the potential factors to visit the hospital. It reveals that the patients who are below 40 years old, do physical exercise and whose ancestors have or had diabetes visit more to the hospital than the patients who are between 40 and 60 years old, do not exercise and whose ancestors did not have diabetes, respectively but the patients who are male and live in urban area visit less to the hospital than the patients who are female and live in rural area, respectively.

Key words: Diabetes mellitus, longitudinal count responses, consistent and efficient estimates, generalized quasi-likelihood

Introduction

Diabetes mellitus is a major public health problem worldwide. It is the most common metabolic disorder and non-communicable disease. The number of people with diabetes was 108 million in 1980 but it has been raised to 422 million in 2014. In 2012, estimated deaths were 1.5 million which were directly caused by diabetes and another 2.2 million deaths were attributable to high blood glucose. World Health Organization (WHO) projects that diabetes will be the seventh leading cause of death by 2030. The global prevalence of diabetes among adults over 18 years of age has risen from 4.7% in 1980 to 8.5% in 2014 (WHO 2016). There are three quarters of people with diabetes live in low and middle income countries and by 2040, 1 adult in 10 (642 million) will have diabetes in the world. In 2015, there were 7.1 million cases of diabetes and it has the second largest number of adults with diabetes (5.1 million adults, 6.31%) in Bangladesh (IDF 2015). Therefore, visiting to the hospital for checkup or treatment or controlling their diabetes condition may reveal opportunities to reduce premature death, disability, and household economic shock.

Imam and Hossain (2012) carried out a comparison study between urban and rural areas of Bangladesh for the prevalence of diabetes based on the 14789 patients using the diabetes data of BIRDEM. They showed that men were more prone to developing diabetes as compared to women

*Author for correspondence: <kksen.du@gmail.com>.

and patients whose either or both the parents were diabetic experience diabetes more than the others. Again the highly educated person with high annual income had the tendency to experience diabetes. They also revealed that the disease is more common for people who are mostly physically inactive, and higher blood pressure and excess body weight also contribute to incidence of diabetes. Khanam *et al.* (2014) used a competing risk hazard model for complications of diabetes mellitus based on the 2887 patients of BIRDEM who have at least two follow-up visits and who are free from complications at the first visit during the follow-up period of 1984-1997. They reported that increase in blood pressure is a major risk factor for coronary heart disease (CHD) and nephropathy in type 2 diabetes mellitus (T2DM). They also revealed that urban participants were more affected by CHD whereas, rural population was the most vulnerable for developing nephropathy. They found that male and illiterate patients are more affected by nephropathy and female and illiterate patients are also more influenced by cataract. Tareque *et al.* (2015) studied about hypertension and diabetes using the data of BDHS (2011). They revealed that people from the highest wealth quintile were significantly more likely to have hypertension, diabetes and the coexistence of hypertension and diabetes than people from the lowest wealth quintile. They also revealed that the odds of having hypertension, diabetes, and their coexistence were higher for older people, women, people who are engaged in less physical labor, and people who were overweight and obese. Rahman *et al.* (2015) used multilevel logistic regression models to identify the risk factors for diabetes awareness using the data of BDHS (2011). They found that participants who had a lower education and lower economic condition were less likely to be aware of their diabetes. Again the people in higher socio-economic status and those living in urban areas have higher rates of diabetes. They also reported that people with no education, lower socio-economic status, and those who lived in disadvantaged regions in terms of education and economic profile (north-western part of Bangladesh) were found lacking of diagnosis, treatment, and control of diabetes.

In longitudinal data analysis, one should take into account the correlation between observations from the same subject. The regression estimates (βs) are less efficient i.e. they are more widely scattered around the true population value than they would be if the within subject correlation was incorporated in the analysis (Diggle *et al*. 2002, Fitzmaurice 1995). The generalized estimating equation (GEE) approach was developed by Liang and Zeger (1986) and Zeger and Liang (1986) to produce more efficient and unbiased regression estimates for analyzing longitudinal or repeated measurements. This is an extension of generalized linear models, which facilitates regression analyses on dependent variables that are not normally distributed (McCullagh and Nelder 1989, Nelder and Wedderburn 1972). Sutradhar (2003) showed through a simulation study that GQL performs the best in estimating both the regression and the true correlation parameters, even though the longitudinal correlations are estimated separately by the method of moments.

Regular visit in a hospital ensures the up-to-date diabetes status and hence patient's well-being depends on his/her number of visits to the hospital. This is because, depending on his/her current

diabetes status, the patient will be able to control it by appropriate actions such as controlling diet, doing physical exercise etc. Therefore, risk of deteriorating health status increases for those who do not have regular visit. We, therefore, would like to identify the factors that are associated with the number of visits to the hospital, BIRDEM.

Model for Longitudinal Count Data

Consider that T repeated count responses are collected from each of K independent individuals. Let $y_i = (y_{i1}, ..., y_{it}, ..., y_{iT})$ denote the T repeated count responses obtained from the i^{th} individual, $i = 1, 2, ..., K$ and $x_{it} = (x_{it1}, ..., x_{itj}, ..., x_{itp})$ be the $p \times 1$ vector of covariates associated with response y_{it}. Let $\beta = (\beta_1, ..., \beta_j, ..., \beta_p)$ be the $p \times 1$ vector of regression coefficients which we want to estimate and $\mu_i = (\mu_{i1}, ..., \mu_{it}, ..., \mu_{iT})$ be the $T \times 1$ vector of mean of response y_i, with $\mu_{it} = E(Y_{it})$; $i = 1,2, ..., K$ and $t = 1, 2, ..., T$. Also let Let Σ_i be the $T \times T$ variance - covariance matrix of Y_t i.e. $\Sigma_i = Var\ (Y_i) = \sigma_{itt}$ with $Var\ (Yit) = \sigma_{itt}'$. Furthermore, suppose that the marginal density of the response y_{it} is of the exponential family form

$$f(y_{it}) = exp[y_{it}\theta_{it} - \alpha(\theta_{it})\phi + b(y_{it}, \phi)] \qquad (2.1)$$

(Liang and Zeger 1986, Sutradhar 2003), where $\theta_{it} = h(\eta_{it})$ with $\eta_{it} = x'_{it}\ \beta$; $a(.)$, $b(.)$ and $h(.)$ of known functional forms and ϕ is a possibly unknown scale parameter and β is the $p \times 1$ vector of parameters of interest. In many important situations, for example, for binary and Poisson data, one may use $\phi = 1$. Consequently, for Poisson data, we use $\phi = 1$ in (2.1) and write the mean and the variance of y_{it} as

$$E(Y_{it}) = \alpha'(\theta_{it}) \text{ and } Var\ (Y_{it}) = \alpha''(\theta_{it})$$

Under regression setup, the most common approach assumes that the count responses follow a Poisson distribution. Note, however, that for rare events, Poisson regression model is also used as generalization of binomial distribution (Cameron *et al*. 1998). Under longitudinal count model, we assume that the response variable, number of visits, y_{it} follows Poisson distribution with mean μ_{it}. Therefore,

$$E(Y_{it}) = \mu_{it} = Var(Y_{it}) = \exp\ (X'_{it}\ \beta)$$

Furthermore, in the longitudinal setup, the components of the vector y_i are repeated responses, which are likely to be correlated. Let $C(\rho)$ be the $T \times T$ true correlation matrix of y_i, which is unknown in practice. Here ρ is, say, a $s \times 1$ vector of correlation parameters which fully characterizes $C(\rho)$. It is of primary interest to estimate β after taking the longitudinal correlation structure $C(\rho)$ into account.

Estimation of Parameters

Sutradhar and Das (1999) showed that even though the Liang and Zeger (1986) approach in many situations yields consistent estimators for the regression parameters, these estimators are usually

inefficient as compared to the regression estimators obtained by using the independence estimating equation approach. In this aspect, a recently developed methodology is generalized quasi-likelihood (GQL) approach which was introduced by Sutradhar (2003). In the GQL approach, the quasi-likelihood estimator of β is the root of the score equation

$$\sum_{i=1}^{K} X_i^{'} A_i \Sigma_i(\rho)^{-1}(y_i - \mu_i) = 0, \qquad (2.1.1)$$

where $\Sigma_i(\rho)$ is the true covariance matrix of Y_i that can be expressed as $\Sigma_i(\rho) = A_i^{½} V(\rho) A_i^{½}$ with $A_i = diag(\sigma_{i11}, ..., \sigma_{itt}, ..., \sigma_{iTT})$ and $C(\rho)$ as the true correlation matrix of Yi, ρ being a correlation index parameter. To overcome the difficulty of unknown $C(\rho)$ in practice, Sutradhar (2003) has suggested a general stationary auto-correlation structure given by

$$C(\rho) = C(\rho_1, \dots, \rho_l, \dots, \rho_{T-1}) = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{T-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{T-2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho_{T-1} & \rho_{T-2} & \rho_{T-3} & \cdots & 1 \end{pmatrix}, \qquad (2.1.2)$$

where for $l = 1, 2, ..., T-1$, ρ_l, represents the lag l autocorrelation. The GQL estimate of β is then computed by solving the estimating equation

$$\sum_{i=1}^{K} X_i^{'} A_i \Sigma_i(\hat{\rho})^{-1}(y_i - \mu_i) = 0, \qquad (2.1.3)$$

where $\Sigma_i(\rho) = A_i^{½} C(\rho_1, ..., \rho_l, ..., \rho_{T-1}) A_i^{½}$ with $C(\rho_1, ..., \rho_l, ..., \rho_{T-1})$ as the true stationary correlation structure for any of the AR(1), MA(1) or equicorrelation models. It is, however, not necessary to know the specific form for the correlation matrix $C(\rho)$, as this form in (2.1.2) is general which is valid under any of the three correlation structures. In practice ρ is unknown, therefore the lag correlations can be consistently estimated by using the well known method of moments. For $l = |t - t'|$, $t \neq t'$ and $t, t' = 1, 2, ..., T$ the autocorrelation of lag l, ρ_l, is estimated by the method of moments as

$$\hat{\rho}_l = \frac{\sum_{i=1}^{K}\sum_{t=1}^{T-l} \tilde{y}_{it}\tilde{y}_{i,t+l}/K(T-l)}{\sum_{i=1}^{K}\sum_{t=1}^{T-l} \tilde{y}_{it}^{\,2}/KT} \qquad (2.1.4)$$

[Sutradhar and Kovacevic (2000), Sutradhar (2003)], where $\tilde{y}_{it}$ is the standardized residual, defined as $\tilde{y}_{it} = (y_{it} - \mu_{it}) / \{\sigma_{itt}\}^{½}$.

The GQL estimating equation (2.1.3) for β and the moment estimate of ρ_l by (2.1.4) are solved imperatively by an iterative process until convergence. The final estimate of β obtained from the iterative process is referred to as the GQL estimate of β and may be denoted by $\hat{\beta}_{GQL}$. We may solve the estimating equation (2.1.3) for β by using Newton - Raphson iterative procedure.

Data and Variables

We have used a follow-up data of registered patients collected by BIRDEM hospital where the patients visit at least two years but must visit at least one of last two years during the follow-up period of 1993 to 1996. In the follow-up period, we took 872 individuals (patients) with their various characteristics such as body mass index (BMI**), age, heredity, area of residence, education level, physical exercise etc. As the responses (number of visits of patients per year) are counts, it is appropriate to assume that the response variable marginally follows the Poisson distribution and the repeated counts recorded for four years will be longitudinally correlated. It is of scientific interest to take the longitudinal correlations into account. In the study we treat all the covariates as categorical variables. The covariate age is used as a categorical variable with three categories- age < 40, age 40 - 60 and age > 60 years. Again, gender is also a categorical variable with two categories- male and female, education level is used as three categories- pre-secondary, secondary and higher, area of residence has two categories- rural and urban, physical exercise has two categories-exercised and non-exercised and heredity has also two categories- heredity and non-heredity. We have considered rural patients as combined of rural and semi urban patients in the study. We treat the covariate body mass index as two categories- under-weight and over-weight.

Results

Bivariate analysis is used to analyze the association between two variables. Thus we have used the one way ANOVA (analysis of variance) for bivariate analysis to know the association of the different characteristics of patients with their number of visits. Again the longitudinal count model has been used for multivariate analysis in the study to know the significant factors of number of visits of the patients.

Table 1 represents the mean and standard deviation (SD) of the visits of patients per year to the hospital for several categories of the selected covariates. Table 1 also incorporates p-values obtained by one way ANOVA F-test and p-values for pairwise comparisons (wherever applicable) obtained by t-test. From the p-values, we identify six variables that have significant associations with the number of visits. These are: heredity, gender, BMI, area of residence, education level and age of patients. Only one covariate physical exercise does not give the significant association with the number of visits.

It is seen that the average visits of male patients is 1.991 times per year and the average visits of female patients is 2.491 to BIRDEM hospital. Thus a female patient visits, on an average, more than a male patient to the BIRDEM hospital, which is highly significant (p-value < 0.001).

**BMI is calculated using the formula = $BMI \frac{weight\ (kg)}{[height\ (m)]^2}$

Table 1. Summary of the visits of patients for each category of the selected covariates.

Variables	Category	Average visits	SD	p-value	Multiple comparisons	p-value
Gender	Male	1.991	1.842	0.000***	-	-
	Female	2.491	2.092			
	< 40	2.571	2.206		Age < 40 vs. Age 40-60	0.000***
Age	40-60	2.151	1.918	0.002***	Age < 40 vs. Age > 60	0.003***
	> 60	2.191	1.985		Age 40-60 vs. Age > 60	0.628
	Pre-secondary	2.361	2.079		Pre-secondary vs. Secondary	0.413
Education	Secondary	2.281	2.038	0.002***	Pre-secondary vs. Higher	0.001***
	Higher	2.091	1.873		Secondary vs. Higher	0.028**
Area	Rural	2.231	1.980	0.016**		
	Urban	2.001	1.847			
BMI	Under-weight	2.251	1.958	0.053*		
	Over-weight	2.121	1.971			
Exercise	Yes	2.301	1.790	0.561		
	No	2.191	1.970			
Heredity	Yes	2.251	1.961	0.058*		
	No	2.121	1.967			

***, ** and * indicate significance at 1, 5 and 10% levels, respectively.

Similarly, we may say that, on an average, the number of visits for patient who is 40 years old is significantly large than the patients of other age groups and the pre-secondary educated patient visits more than the secondary and highly educated patients. Again a rural patient visits more than an urban patient and a patient who is under-weight visits more than a patient who is over-weight. Similarly, an exercised patient visits, on an average, more than a non-exercised patient but not significantly and a heredity patient (whose ancestors has or had diabetes) visits, on the average, more than a non-heredity patient (whose ancestors did not have diabetes) to the BIRDEM hospital.

Table 2 represents the regression parameter estimates ($\hat{\beta}$), standard errors and corresponding p-values (based on Wald test) obtained by Generalized Quasi-likelihood (GQL) estimation approach with lag correlation estimates for the longitudinal count data obtained by BIRDEM. Based on Table 2, the covatiates gender, area of residence, age < 40 and heredity have highly significant effects (p-value < 0.01) on the response variable and the physical exercise also has significant effect (p-value < 0.10) on the response variable. On the other hand, age > 60, body mass index, secondary and higher education levels do not have significant effects (p-value < 0.1) on the response variable. From the regression coefficient of the covariate gender, we see that it has significantly negative effect on the number of visits. The negative value of $\hat{\beta}_1$ (effect of gender) = –0.185 suggests that the male patients visit less to the hospital than the female patients. To be

specific, the mean visit of a male patient to the hospital is $[1 - exp(\hat{\beta}_1)] \times 100\% = 16.9\%$ lower than that of a female patient. The variable age <40 shows the positive association with the number of visits. The positive value of $\hat{\beta}_2$ (effect of age < 40) = 0.178 suggests that the patients who are 40 years old visit more to the hospital most significantly than the patients who are between 40 and

Table 2. Regression estimates of the selected covariates for number of visits of patients by general autocorrelation based GQL approach with lag correlation estimates.

Variables	Categories	$\hat{\beta}$	$S.E.(\hat{\beta})$	p-value
(Intercept)	-	0.854	0.043	0.000***
Gender	Female			
	Male	–0.185	0.034	0.000***
	40-60			
Age	< 40	0.178	0.050	0.000***
	> 60	0.029	0.034	0.394
	Pre-secondary			
Education	Secondary	–0.059	0.042	0.160
	Higher	–0.058	0.040	0.147
Area	Rural			
	Urban	–0.135	0.047	0.004***
BMI	Under-weight			
	Over-weight	0.016	0.031	0.606
Exercise	No			
	Yes	0.132	0.075	0.078*
Heredity	No			
	Yes	0.081	0.030	0.007***
Lag correlations	Estimates			
$\hat{\rho}_1$	0.331			
$\hat{\rho}_2$	0.121			
$\hat{\rho}_3$	0.000			

*** and * indicate significance at 1 and 10% levels, respectively.

60 years old. To be specific, we may say that, on an average, a patient who is 40 years old has 19.5% higher rate of visiting the hospital as compared to that of a patient who is between 40 and 60 years old. Again the variable age > 60 shows the positive association with the response variable and the positive value suggests that the age > 60 patients visit more to the hospital than the age 40-60 patients but not significantly. From the covariate education level, we created two dummy variables secondary and higher education with pre-secondary as reference category. The variable secondary education shows the negative association with the response variable and it may be said that the pre-secondary educated patients visit hospital more than the secondary educated patients but not significantly. Again the variable higher education level shows the negative association with the response variable and the negative value suggests that the pre-secondary educated patients visit hospital more than the higher educated patients but not significantly.

The covariate area of residence is negatively associated with the response variable. We may say that the rural patients visit more to the hospital most significant than the urban patients. By

percentage, it may be said that the mean visit of an urban patient to the hospital is 12.6 lower as compared to that of a rural patient. The covariate BMI does not give the significant effect on the number of visits of patients but it is positively associated and the positive value suggests that the over-weight patients visit hospital more than the under-weight patients. The covariate physical exercise is positively associated with the response variable and it may be said that the mean visits of a patient who does physical exercise is 14.1% higher as compared to that of a patient who does not exercise. It is found that heredity has highly significant effect (p-value < 0.01) on the response variable, number of visits, to the BIRDEM hospital. From the regression coefficient, we see that it is positively associated with the response variable. The positive value suggests that heredity patients made more visit to the hospital significantly as compared to non-heredity patients. To be specific, we may say that, on an average, a patient whose ancestors have/had diabetes have 8.4% higher rate of visiting the hospital as compared to that of a patient whose ancestors did not have diabetes. Again, from Table 2, we notice that first lag correlation estimate of the general autocorrelation structure is 0.331 which is moderately large. Thus avoiding the lag correlations will result in inefficient regression estimates.

Conclusion

We use longitudinal count model for diabetes related data during the follow-up period of 1993-1996. It is clear that generalized quasi-likelihood approach can be an ideal choice for analyzing the longitudinal data which consider the general autocorrelation structure. In longitudinal data analysis, estimating the effect of covariates on a response variable is often of interest while longitudinal correlations are typically considered as nuisance parameters. In the study, we have discussed the generalized quasi-likelihood (GQL) approach. The approach has been applied to the longitudinal count data of BIRDEM to get consistent as well as highly efficient estimates of the regression parameters. From the resulting regression estimates using GQL approach under a general autocorrelation structure among the responses of the individuals, we have found that heredity, gender, area of residence, physical exercise and age < 40 have significant effects on the response variable, number of visits, to the BIRDEM hospital and all other covariates do not have significant effects. So these significant factors are the potential factors to visit to the BIRDEM hospital. We revealed that the patients who are below 40 years old, do physical exercise and whose ancestors have or had diabetes visit more to the hospital than the patients who are between 40 and 60 years old, do not exercise and whose ancestors did not have diabetes, respectively but the patients who are male and live in urban area visit less to the hospital than the patients who are female and live in rural area, respectively. Thus, we recommend to increase awareness for the groups who are male, live in urban area, patients who do not exercise, patients who are between 40 and 60 years old and the patients whose ancestors did not have diabetes, so that they increase their visits i.e. follow-up visits to the hospital for controlling diabetes.

Acknowledgement

The authors would like to thank Professor Dr. Wasimul Bari, Department of Statistics, Biostatistics & Informatics, University of Dhaka for his help and valuable comments on the preparation of this manuscript.

References

Bangladesh Demographic and Health Survey (BDHS). 2011. National Institute of Population Research and Training (NIPORT). Calverton, Maryland and Dhaka, NIPORT, Mitra and Associates and Macro International inc.

Cameron C. and P.K. Trivedi. 1998. Regression Analysis of Count Data, Econometric Society Monograph No. 30, Cambridge University Press.

Diggle, P.J., P. Heagerty, K.Y. Liang and S.L. Zeger. 2002. Analysis of Longitudinal Data. Oxford University Press, Oxford.

Fitzmaurice, G.M. 1995. A caveat concerning independence estimating equations with multivariate binary data. *Biometrics* **51**: 309-317.

Imam, T. and M. B. Hossain. 2012. Diabetes prevalence : A comparison between urban and rural areas of Bangladesh. Scholars, Volume 1, Issue 4 Article - 10.

International Diabetes Federation (IDF). 2015. IDF Diabetes Atlas 7th Edition. https://www.idf.org/idf-diabetes-atlas-seventh-edition

Khanam, P.A., M.A. Islam, M.A. Sayeed, T. Begum, M.G. Rabbani, S. Choudhury and H. Mahtab. 2014. A competing risk hazard model for complications of diabetes mellitus. *J. Biosci. and Medicines* **2**: 1-11.

Liang, K.Y. and S.L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika* **73**: 13-22.

McCullagh, P. and J.A. Nelder. 1989. Generalized Linear Models. Second ed. London: Chapman and Hall.

Nelder, J. and R. W. M. Wedderburn. 1972. Generalized linear models. *J. Roy. Statist. Soc*. **135**: 370-384.

Rahman, M.S., S. Akter, S.K. Abe, M.R. Islam, M.N.I. Mondal, J.A.M.S. Rahman, and M.R. Rahman. 2015. Awareness, Treatment, and Control of Diabetes in Bangladesh: A Nationwide Population-Based Study. PLOS ONE 10(2): e0118365. Doi: 10.1371/journal. pone.0118365.

Sutradhar, B.C. 2003. An overview on regression models for discrete longitudinal responses. *Statistical Sci*. **18**: 377-93.

Sutradhar, B.C. and K. Das. 1999. On the efficiency of regression estimators in generalized linear models for longitudinal data. *Biometrika* **86**: 459-65.

Sutradhar, B.C. and M. Kovacevic. 2000. Analyzing ordinal longitudinal survey data: Generalized estimating equations approach. *Biometrika* **87**: 837-848.

Tareque, M.I., A. Koshio, A.D. Tiedt, T. Hasegawa. 2015. Are the Rates of Hypertension and Diabetes Higher in People from Lower Socioeconomic Status in Bangladesh? Results from a Nationally Representative Survey. PLOS ONE 10(5): e0127954. doi:10.1371/journal.pone.0127954.

World Health Organization (WHO). 2016. Global Report on Diabetes. http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf?ua=1

Zeger, S.L. and K.Y. Liang. 1986. The analysis of discrete and continuous longitudinal data. *Biometrics* **42**: 121-30.

(Manuscript received on 10 May, 2016; revised on 9 August, 2016)