Austrian Journal of Statistics
October 2015, Volume 44, 3–15.
http://www.ajs.or.at/
doi:10.17713/ajs.v44i3.17

On the Selection of Relevant Covariates and Correlation Structure in Longitudinal Binary Models: Analysing the Impact of the Height of Type II Diabetes

Md Erfanul Hoque
University of Dhaka

Mahfuzur Rahman Khokan
University of Dhaka

Wasimul Bari
University of Dhaka

Abstract

To examine the impact of height on the occurrence of Type II diabetes, a longitudinal binary data set has been analyzed. The relevant covariates were selected by using quasi-likelihood based information criteria (QIC) and correlation information criteria (CIC) was used to select the correlation structure appropriate for the repeated binary responses. The consistent and efficient estimates of regression parameters were obtained from the generalized estimating equations (GEE). With the selected covariates height, education level, gender and unstructured correlation structure, it is found that there exists a statistically significant inverse relationship between height of an individual and the development of Type II diabetes. Risk Ratios for different covariates along with standard errors and confidence intervals are also given.

Keywords: correlation information criteria, generalized estimating equations, longitudinal binary data, quasi-likelihood based information criteria, risk ratio.

1. Introduction

The prevalence of diabetes particularly type 2 diabetes is increasing day by day and it becomes an emerging epidemic in the world. Among the regions, Southeast Asia region is affected markedly by this and according to WHO report approximately 79.5 million diabetic patients will live in this area,which is more than 26% of the world's total diabetic population (e.g. IDF, 1998). The prevalence rates in India, Pakistan and China are 12.1%, 11.1%, and 6.1% respectively; where in Bangladesh this rate is 8.1% in urban and 2.3% in rural. That is, Bangladesh as a developing country is facing a high prevalence of diabetes.

It has been well established that the increased prevalence of metabolic syndrome components and resultant increased risk of type 2 diabetes mellitus are associated with obesity (see, e.g. Janghorbani *et al*., 2010, WHO, 2000). Epidemiological studies have demonstrated that different anthropometric measures of obesity such as body mass index (BMI), waist circumference (WC), waist-height ratio (WHtR), waist-hip ratio (WHR) are strong and consistent predictors of type 2 diabetes (see, e.g. Janghorbani *et al*., 2010, Schulze *et al*, 2006). The relationship between increased BMI, WHtR and type 2 diabetes mellitus risk may be due to a direct or to inverse effect of height. It implies that height may play an important role for the incidence of diabetes. The association between height

of respondent and risk of type 2 diabetes mellitus has been investigated by several epidemiological studies but it is still unclear whether height affects the association. Also, the role of height as risk factor for type 2 diabetes mellitus remains uncertain. In most but not all studies, height appears to be inversely related with diabetes. There have been contravening reports about possible association of height and diabetes (see, e.g. Sicree et al, 2008, Snijder *et al*., 2003, Bozorgmanesh *et al*., 2011, Wang *et al*., 1997, Njolstand et al., 1998): a positive association was found in a studies (e.g. Wang *et al*.,1997), whereas no association (e.g. Lorenzo *et al*., 2009) or an inverse relation was reported in others (see, e.g. Snijder *et al*.,2003, Njolstand et al., 1998). Also, there was an association only in women (e.g. Bozorgmanesh *et al*., 2011) or men (e.g. Schulze *et al*., 2006). Hence, it would be interesting to find out a relationship between height and risk of type 2 diabetes mellitus. In this paper, we try to investigate this relationship in the context of Bangladesh using BIRDEM data.

The studies mentioned in the literature to explore the relationship between height and diabetes is based on the cross-sectional or follow-up designs. There exists no literature that deals with this relationship in the context of repeated observations obtained from an individual over a short period of time under a longitudinal study setup. Now-a-days, analysis of repeated observations has been extensively used in the biomedical studies. For example, the disease status of the patients may vary from time to time and covariates related with the disease behave differently with the changes in disease status. To analyze these types of data, observation at a single point provides misleading inferences about the disease status or the disease risk factor relationship. To overcome this problem longitudinal analysis plays an important role to draw valid inference. Note that repeated responses are likely to be correlated as these are collected from an individual. Therefore, it is necessary to take this correlation into account to estimate regression parameters consistently and efficiently. Using quasi-likelihood function, Liang and Zeger (1986) proposed the 'working' correlation based generalized estimating equations (GEE) for the purpose of estimation of regression parameters as well as the correlation parameters. In this paper, an attempt has been made to examine how height of an individual affects his/her diabetes status controlling relevant socioeconomic and demographic factors using longitudinal binary data. For the purpose of analysis, data have been obtained from Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic Disorders (BIRDEM).

One of the important features of any regression analysis is the model selection. In a longitudinal study, repeated responses along with a large number of covariates are collected from each individual of the study. Including all covariates in the regression analysis may result in a complex model and may lead to less precise estimates of parameters of interest. To overcome this problem, a subset of important covariates needs to be considered for the regression analysis so that model predictability and parsimony increase. There exist a number of subset selection criteria and procedures for linear regression models. Among them, likelihood function based Akaike's Information Criterion (AIC) (see, e.g. Akaike, 1973) is widely used. Since the construction of likelihood function is very much complicated in the longitudinal setup, Pan (2001a) proposed a modification of AIC based on the GEE, which is known as quasi-likelihood under the independent model information criterion (QIC). The other non-likelihood function bases criteria for model selection are: bootstrap smoothed cross-validation (BCV) [see, e.g. Pan (2001b)] that minimizes the expected predictive bias (EPB); bias-corrected bootstrap approaches to estimate the predictive mean squared error (PMSE) of a model and use the PMSE for model selection [see, e.g. Pan and Lee (2001)]; a generalized version of Mallows's C_p (GC_p) suitable for both parametric and non-parametric models [see, e.g. Cantoni et al. (2005)]; a cross-validation Markov Chain Monte Carlo (MCMC) procedure [see, e.g. Cantoni et al. (2008)].

Another issue that needs to address in the longitudinal setup is to select an appropriate correlation structure for the repeated responses. The QIC (e.g. Pan, 2001a) can also be used to select the appropriate 'working' correlation structure. Hin and Wang (2009) argued that the QIC measures are more sensitive to changes in the mean structure than changes in the covariance structure. As a remedy, Hin and Wang (2009) proposed correlation information criterion (CIC) for selecting the appropriate correlation structure.

Since the main focus of this paper is to measure the impact of height on the occurrence of diabetes, other covariates along with height are selected by using QIC (Pan, 2001a) and the correlation structure

for the repeated responses is selected by the CIC (e.g. Hin and Wang, 2009). Finally, longitudinal model is fitted by GEE (e.g. Liang and Zeger, 1986). In Section 2, a longitudinal binary model, GEE, QIC, CIC, and risk ratio estimation are described mathematically. A longitudinal binary model with selected covariates and correlation structure is illustrated to the data obtained from BIRDEM to determine the potential determinants of diabetes in Section 3. This paper concludes in Section 4 with a short discussion.

2. Methods

2.1. Longitudinal binary model

Suppose that y_{it} is the binary response obtained from individual i, $i = 1, \cdots\cdots, N$ at time point $t = 1, \cdots, \cdots, T$. Also, suppose that $\mathbf{x_{it}} = (x_{it1}, \cdots, x_{itj}, \cdots x_{itp})^{'}$ is the $p \times 1$ vector of covariates associated with the response y_{it}. Furthermore, suppose that the marginal probability distribution of y_{it} is a number of exponential family of distributions, i.e.,

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it})\}\varphi + b(y_{it}\varphi)], \tag{2.1}$$

(Liang and Zeger, 1986), where $a(\cdot)$ and $b(\cdot)$ are of known functional form. It can be shown that $\theta_{it} = \mathbf{x_{it}^{'}}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_j, \cdots, \beta_p)^{'}$ is the $p \times 1$ vector of regression coefficients. In equation (2.1), φ is the scale parameter and for binary response $\varphi = 1$. The marginal mean and variance of Y_{it} can be expressed as $\mu_{it} = E(Y_{it}) = a^{'}(\theta_{it})$ and $\sigma_{itt} = var(Y_{it}) = a^{''}(\theta_{it})$. For binary response, $\mu_{it} = [1 + exp(-\mathbf{x_{it}^{'}}\boldsymbol{\beta})]^{-1}$ and $\sigma_{itt} = \mu_{it}(1 - \mu_{it})$ The response vector for individual i is given by $\mathbf{Y_i} = (Y_{i1}, \cdots, Y_{it}, \cdots, Y_{iT})^{'}$ with mean $\boldsymbol{\mu_i} = (\mu_{i1}, \cdots, \mu_{it}, \cdots, \mu_{iT})^{'}$. Under a longitudinal set up, the repeated responses $Y_{i1}, \cdots Y_{it}, \cdots, Y_{iT}$ are likely to be correlated. Here, variance of $\mathbf{Y_i}$ can be expressed as

$$\Sigma_i = var\ (\mathbf{Y_i}) = A_i^{\frac{1}{2}} C(\rho)\ A_i^{\frac{1}{2}},$$

where $C(\rho)$ is the correlation matrix for response vector $\mathbf{Y_i}$ and $A_i = diag[\sigma_{i11}, \cdots, \sigma_{itt}, \cdots, \sigma_{iTT}]$. Note that the correlation matrix $C(\rho)$ is usually unknown. Here, the main parameter of interest is regression parameter $\boldsymbol{\beta}$ and the correlation parameter ρ is known as nuisance parameter. To obtain consistent as well as efficient estimates for $\boldsymbol{\beta}$, one needs to take the correlation parameter ρ into account. Since the probability distribution of $\mathbf{Y_i}$ is cumbersome, it would be difficult to obtain the maximum likelihood estimates of regression parameter $\boldsymbol{\beta}$ and correlation parameter ρ. As a remedy, Liang and Zegar (1986) proposed quasi-likelihood function based estimating equation for $\boldsymbol{\beta}$, which is well known as GEE. Note that GEE is constructed assuming 'working' correlation for response $\mathbf{Y_i}$. Liang and Zegar (1986) also proposed method of moments estimates for correlation parameters under different working correlation structures.

2.2. GEE for regression parameter

For known correlation parameter ρ, the GEE for regression parameter $\boldsymbol{\beta}$ is given by

$$\sum_{i=1}^{N} U_i(\boldsymbol{\beta}, \mathbf{y_i}, C) = 0, \tag{2.2}$$

[Liang and Zeger, (1986)] with $U_i(\boldsymbol{\beta}, \mathbf{y_i}, C) = \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu_i}^{'} \Sigma_i^{-1}(\mathbf{y_i} - \boldsymbol{\mu_i})$, where $\boldsymbol{\mu_i}$ and Σ_i are defined in Section 2.1. The estimating equations given in (2.2) can be solved for $\boldsymbol{\beta}$ by using Newton-Raphson iterative procedure. The estimate, denoted by $\widehat{\boldsymbol{\beta}_C}$ under working correlation $C(\rho)$, obtained at the m^{th} $(m = 1, 2, 3, \cdots, \cdots, \cdots)$ iteration is given by

$$\widehat{\boldsymbol{\beta}_C}^{(m)} = \widehat{\boldsymbol{\beta}}^{(m-1)} + [A]^{-1}_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(m-1)}} \left[\sum_{i=1}^{N} U_i(\boldsymbol{\beta}, \mathbf{y_i}, C)\right]_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}^{(m-1)}}$$

where $A = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu}_{\boldsymbol{i}}^{'} \Sigma_i^{-1} \frac{\partial}{\partial \boldsymbol{\beta}^{'}} \boldsymbol{\mu}_{\mathbf{i}}$. Note that $\widehat{\boldsymbol{\beta}_C}$ is asymptotically distributed as normal with mean $\boldsymbol{\beta}$ and the variance $V_{\widehat{\boldsymbol{\beta}_C}}$. The sandwich or robust estimate of $V_{\widehat{\boldsymbol{\beta}_C}}$ is given by

$$\widehat{V_{\widehat{\boldsymbol{\beta}_C}}} = A^{-1} \left\{ \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\mu}_{\mathbf{i}}^{'} \Sigma_i^{-1} \left(\mathbf{y_i} - \boldsymbol{\mu}_{\mathbf{i}}\right) \left(\mathbf{y_i} - \boldsymbol{\mu}_{\mathbf{i}}\right)^{'} \Sigma_i^{-1} \frac{\partial}{\partial \boldsymbol{\beta}^{'}} \boldsymbol{\mu}_{\mathbf{i}} \right\} A^{-1}, \tag{2.3}$$

with replacing $\boldsymbol{\beta}$ and ρ with their respective estimates. The estimation of ρ depends on the 'working' correlation structure. One may assume independence, exchangeable, first-order autoregressive (AR-1), or unstructured correlation structure for the repeated responses. For independence structure, $Corr\left(Y_{it}, Y_{it'}\right) = 0$; exchangeble, $Corr\left(Y_{it}, Y_{it'}\right) = \rho$; AR-1, $Corr\left(Y_{it}, Y_{it'}\right) = \rho^{\left|t-t^{'}\right|}$; and unstructured, $Corr(Y_{it}, Y_{it'}) = \rho_{itt'}$ with $t \neq t^{'}$. The estimator of ρ for different correlation structure is given by Liang and Zeger, (1986, section 3.3). The main purpose of this paper is to determine the effects of height on the occurrence of type-II diabetes along with other relevant covariates. For selecting relevant covariates from the available covariates, one may use QIC (e.g. Pan, 2001a). A short discussion on QIC is given below.

2.3. Quasi-likelihood based information criterion (QIC)

Pan (2001a) proposes QIC by modifying Akaike's information criterion (AIC) [e.g. Akaike, (1973)]. When the formulation of likelihood function is tractable, one may use AIC for the purpose of model selection. Akaike (1973) defined AIC as $AIC = -2 \ln L\left(\widehat{\boldsymbol{\beta}}\right) + 2p$, where $L\left(\widehat{\boldsymbol{\beta}}\right)$ is the likelihood function evaluated at $\widehat{\boldsymbol{\beta}}$ and p is number of regression parameters. In longitudinal setup, it may not be possible to construct the likelihood function. In this case, following AIC, Pan (2001a) proposed QIC, which is based on quasi-likelihood function under independent correlation structure. Mathematically, QIC may be defined as

$$QIC\left(C\right) = -2 \sum_{i=1}^{N} Q_i\left(\widehat{\boldsymbol{\beta}_C}, \mathbf{y_i}, I\right) + 2\, trace\left(\widehat{\Omega_I}\, \widehat{V_{\widehat{\boldsymbol{\beta}_C}}}\right), \tag{2.4}$$

where $Q_i\left(\widehat{\boldsymbol{\beta}_C}, \mathbf{y_i}, I\right)$ is the quasi-likelihood function under independence correlation structure, I, evaluated at estimated regression coefficient obtained under a 'working ' correlation structure C. Note that for binary repeated responses, one can express

$$\sum_{i=1}^{N} Q_i\left(\widehat{\boldsymbol{\beta}_C}, \mathbf{y_i}, I\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} \left[y_{it} ln \frac{\mu_{it,}}{1-\mu_{it}} + \ln(1-\mu_{it}) \right]$$

$$= \sum_{i=1}^{N} \sum_{t=1}^{T} \left[y_{it} \mathbf{x}_{\mathbf{it}}^{'} \widehat{\boldsymbol{\beta}_C} + \ln(1 - e^{\mathbf{x}_{\mathbf{it}}^{'} \widehat{\boldsymbol{\beta}_C}}) \right]$$

In (2.4), the expression for $\widehat{V_{\widehat{\boldsymbol{\beta}_C}}}$ is given in (2.3) and $\widehat{\Omega_I}$ is defined as

$$\widehat{\Omega_I} = -\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{'}} \sum_{i=1}^{N} Q_i\left(\widehat{\boldsymbol{\beta}_C}, \ \mathbf{y_i}, \ I\right) = \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{x_{it}}\, \mu_{it}\left(1-\mu_{it}\right) \mathbf{x}_{\mathbf{it}}^{'}.$$

Like AIC, a model with minimum QIC is chosen to be the best model. Pan (2001a) also proposed to use QIC for selecting a correlation structure appropriate for the repeated responses. But, Hin and Wang (2009) argued that QIC cannot be used for correlation structure selection because of the following reasons. The first term of QIC depends neither on 'working' correlation nor on the correlation structure. In addition, the quasi-likelihood function is constructed assuming an independence correlation structure. Therefore, the first term has not contributed in selecting correlation structure. Though second term of QIC reflects the 'working' correlation through sandwich estimator $\widehat{V_{\widehat{\boldsymbol{\beta}_C}}}$, QIC is heavily

influenced by the first term. Hence, QIC is not an appropriate tool to select the correlation structure. Hin and Wang (2009) proposed to use only the second term for selecting the correlation structure and this measure is known as correlation information criterion (CIC). That is,

$$CIC = tr(\widehat{\Omega_I}\,\widehat{V_{\beta_C}}) \quad . \tag{2.5}$$

The correlation structure for which the binary longitudinal model provides the minimum CIC will be chosen to analyze the data.

2.4. Risk ratio estimation

In this paper, the adjusted risk ratio (RR) is used to compare the rate of incidence of diabetes among the categories of a covariate. Mathematically, the RR for a covariate x_j having values x_j^1 and x_j^2 can be defined as

$$RR = \frac{R\left(\bar{x}_1, \cdots, x_j = x_j^1, \cdots, \bar{x}_p\right)}{R\left(\bar{x}_1, \cdots, x_j = x_j^2, \cdots, \bar{x}_p\right)}, \tag{2.6}$$

$$\text{with} \quad R\left(x_1, \cdots, x_j, \cdots, x_p\right) = \left[1 + exp\left(-\sum_{j=1}^{p} x_j\beta_j\right)\right]^{-1}$$

where all other covariates will be considered at respective mean values (e.g. Kleinbaum and Klein, 2005). The estimated RR can be computed from (2.6) by replacing β_j's with their corresponding estimates obtained from GEE. The $100\ (1-\alpha)\,\%$ confidence interval for RR is

$$\widehat{RR} \pm\ Z_{\alpha/2}\sqrt{var\left(\widehat{RR}\right)},$$

where $var(\widehat{RR}) = RR^2[(x_j^1)^2(1 - R(\bar{x}_1, \cdots, x_j = x_j^1, \cdots, \bar{x}_p))^2 - (x_j^2)^2(1 - R(\bar{x}_1, \cdots, x_j = x_j^2, \cdots, \bar{x}_p))^2\,]var(\widehat{\beta}_j)$.

3. Analysis of impact of height on type II diabetes

The main objective of this paper is to examine the impact of height of an individual on the occurrence of Type II diabetes controlling selected important factors by using repeated observations obtained from each individual considered in the study. For this purpose, the longitudinal data collected by BIRDEM has been used.

3.1. Data and variables

The data set consists of 2297 individuals each having 4 observations. An individual is defined whether diabetic or not by observing the glucose level after two hours of 75 gms glucose load at each visit. If the observed glucose level is less than 11.1 mmol/liter, then the patient is categorized as non-diabetic and otherwise diabetic (WHO, 2007; WHO/IDF, 2006). The main covariate of interest in this paper is height of an individual. It is found that mean height is 158.88 cm with standard deviation 8.8, and maximum and minimum heights are 193 cm and 109 cm, respectively. Besides height, age, heredity [HRD: Yes, No (ref)], education level [EDU: Yes, No (ref)], gender [Male, Female (ref)], physical exercise [PHEX: Yes, No (ref)], place of residence [AREA: Urban, Rural (ref)], and other complications [COM: Yes, No (ref)] are taken into consideration as these covariates are found to have significant impact on the occurrence of Type II diabetes in other studies (Njostad *et al*. 1998, Lorenzo *et al*. 2009, Schulze *et al*. 2006).

Among the individuals, the mean age is 53.64 years with standard deviation 11.85 and the maximum and minimum age are 106.4 and 13.3, respectively. Parents of 40.6% of individuals have diabetes.

Most of the individuals (89.5%) have at least primary education. It is observed that 65.3% of individuals are male and 34.7% are female. Most of the individuals are not involved with physical exercise (96.6%). This data set is based on urban as 97.5% of individuals are from urban. Regarding complications other than diabetes, 96.5% of individuals have no other complications.

3.2. Selection of the best set of covariates

For the purpose of selection of covariates for the occurrence of Type II diabetes, we consider longitudinal binary models under different correlation structures. Since four responses from each individual are collected, they are likely to be correlated. Therefore, in this analysis, we do not consider independence as a 'working' correlation structure. The correlation structures considered are exchangeable, AR-1, and unstructured. Since height is the main covariate of interest, we consider this covariate in all possible models. The values of QIC are calculated using equation (2.4) for all possible models under different correlation structures. This result is shown in Table 1. It is clear from Table 1 that Model 20 produces minimum value under all three correlation structures. Hence, the selected covariates for the analysis are height, education level and gender.

3.3. Selection of the best correlation structure

To obtain estimates for the regression coefficients of the selected covariates, one may need a 'working' correlation that is appropriate for the repeated responses. To choose the appropriate correlation structure, one can compute the values of CIC for different correlation structures using equation (2.5) and then select the correlation that produces the minimum value of CIC. The values of CIC under exchangeable, AR-1, and unstructured correlations with the previously selected covariates are given in Table 2. It is clear from the table that unstructured correlation is appropriate for the longitudinal binary data obtained from BIRDEM.

3.4. Estimation of regression parameters using GEE

For the consistent and efficient estimates of the regression coefficients, one may solve the estimating equation given in equation (2.2) with the selected covariates and unstructured correlation structure by using Newton-Raphson iterative process. Estimates along with standard error, p-value, and 95% confidence interval are given in Table 3. From this table, it reveals that height is negatively associated with the occurrence of Type II diabetes and this effect is found to be statistically significant as p-value is 0.004. Education and gender have also negative significant effects on the diabetes with p-values 0.00 and 0.074, respectively.

The correlation parameters in a longitudinal setup are considered as the nuisance parameters and these parameters can be estimated by the method of moments (e.g. Liang and Zeger, 1986). The moment estimates of correlation parameters are given below. Note that the values in the parentheses are the standard errors of the estimators.

$$\widehat{C(\rho)} = \begin{bmatrix} 1 & 0.853\ (0.017) & 0.800\ (0.018) & 0.759\ (0.018) \\ & 1 & 0.889\ (0.016) & 0.835\ (0.017) \\ & & 1 & 0.909\ (0.016) \\ & & & 1 \end{bmatrix}$$

Since all the estimates of correlation parameters are more than 0.75, there exists a high correlation among the binary responses. Therefore, it is essential to take the correlation structure into account for the estimation of regression parameters.

Note that maximum and minimum heights are found to be 193 cm and 109 cm, whereas the mean height is 158.88 cm with standard deviation 8.8 cm. It indicates that the data contain outliers with respect to height. To examine the impact of height controlling other covariates, the GEE estimates are also obtained after deleting outliers. The outliers were detected by the 'robust three sigma' rule (Maronna *et al.*, 2006). After deleting outliers, the GEE estimates under selected model using the

unstructured correlation matrix for constant, height, education, and gender are 3.914, -0.0179, -0.574, and -0.213 with standard errors 1.02, 0.007, 0.144, and 0.121, respectively. The corresponding p-values are 0.00, 0.008, 0.00, and 0.057. It is observed that there is a little difference in the values of estimates before and after deleting outliers.

3.5. Estimation of risk ratio

To examine the association of a covariate with the occurrence of Type II diabetes controlling other covariates in the model, one may compute the adjusted risk ratio (RR) using equation (2.6). The adjusted RR and its standard error with 95% confidence interval for the selected covariates are given in Table 4. Since height is considered as continuous, we compute the quartile values first [e.g. first quartile (Q_1), second quartile (Q_2), and third quartile (Q_3)] and then RRs for Q_3 versus Q_1, Q_3 versus Q_2, and Q_2 versus Q_1. The values of first, second, and third quartiles are 152, 160, and 165 cm, respectively. From Table 4, it is found that the RR for Q_3 versus Q_1 is 0.78. It implies that an individual with height 165 cm is 22% less to have Type II diabetes compared to an individual with height 152 cm. The RR for Q_3 versus Q_2 is found to be 0.91, which implies that an individual with median height is 10% [((1/0.91)-1)$\times$100%] more like to develop Type II diabetes than an individual with height 165 cm. Finally, while comparing the second and first quartiles, an individual with second quartile height is 14% less likely to be a Type II diabetic patient compared to an individual with first quartile height. Note that all the RRs for height are statistically highly significant as p-values are 0.00. Therefore, an individual with shorter height is substantially at a higher risk of developing of Type II diabetes.

Education and gender are also found to have statistically significant impact on the occurrence of Type II diabetes with p-values 0.00 for both cases. Educated individuals are at 42% less risk for developing diabetes than their counterparts. On the other hand, male is 19% less likely to have diabetes than female.

4. Discussion

Generally, risk factors of Type II diabetes are modifiable and preventable. Therefore, early identification and preventive behavior for these risk factors can reduce the risk of developing Type II diabetes by 90% (see e.g. CDC, 2009). In this paper, an attempt has been made to identify the potential risk factors for Type II diabetes and to establish a relationship between height of an individual and the occurrence of diabetes by analyzing the longitudinal binary model obtained from BIRDEM. No study has been conducted in Bangladesh to identify the risk factors of diabetes by considering the longitudinal data. For the purpose of analysis, along with height, we first chose the important factors from the available covariates by using QIC (Pan, 2001a), which is appropriate for variable selection when the response is multivariate discrete variable and formulation of full likelihood function is mathematically involved. To obtain the efficient estimates for the regression parameters, one needs to consider a correlation structure appropriate for the repeated responses. After selecting the relevant covariates, we select the correlation structure using CIC (see e.g. Hin and Wang, 2009). Finally, estimates of regression parameters are obtained by solving the GEE (e.g. Liang and Zeger, 1986).

The selected covariates in this analysis are height, education level and gender and the appropriate correlation structure selected for the repeated responses is unstructured. It is found that education plays an important role for preventing the occurrence of Type II diabetes and male is at more risk of developing diabetes compared to female. One of main objectives of this paper is to examine the relationship between height and diabetes. This analysis reveals the fact that the probability of occurring diabetes deceases as the height of individual increases. That is, shorter height is associated with a higher occurrence of diabetes. One of the explanations of this inverse relation is that taller individuals have more muscle mass and muscle is the major tissue involved in uptake of glucose, against the fixed glucose load of 75 grams (see e.g. Sicree *et al.*, 2008). The dilution effect of total body water may contribute in establishing the results (e.g. Sicree *et al.*,2008).

In this study, a severe metabolic disturbance is identified in a shorter individual than a taller one regarding the occurrence of Type II diabetes. Therefore, developing diabetes may be reduced by controlling the factors that may influence the height. The height may be controlled by genetic and non-genetic (early-life and childhood) factors (e.g. Hirschhorn *et al*., 2001; Park *et al.*, 2004; Li *et al.*, 2006). Naturally, the next generation is likely to have shorter height, if most of the family members of the family are of short height. Note that genetic factors are totally beyond the control of human. The non-genetic factors that may affect the height are maternal smoking during pregnancy, birth weight, ill health during childhood and adolescence, and mental condition during childhood and adolescence. Non-genetic factors can be controlled to some extent by leading a healthy life style from childhood.

Acknowledgements We would like to thank Bangladesh Institute of Research and Rehabilitation in Diabetes, Endocrine and Metabolic Disorders (BIRDEM), Bangladesh to make data available for analysis. We also thank anonymous reviewer and editor for their valuable comments and suggestions that led to significant improvements in the presentation.

References

1. International Diabetes Federation (1998). Diabetes around the World.

2. Janghorbani M., and Amini M. (2010). Comparison of Body Mass Index with Abdominal Obesity Indicators and Waist-to-stature Ratio for Prediction of Type 2 Diabetes: the Isfahan Diabetes Prevention Study. Obesity Research & Clinical Practice **4**: e25-e32.

3. WHO (2000). Obesity: Preventing and Managing the Global Epidemic. Report of a WHO consultation. World Health Organization Technical Report 2000; **894: i-xii**, 1-253.

4. Schulze M. B, Heidemann C, Schienkiewitz A. Bergmann M. M, Hoffmann K, and Boeing H (2006). Comparison of Anthropometric Characteristics in Predicting the Incidence of Type 2 Diabetes in the EPIC-Potsdam Study. Diabetes Care **29:** 1921-1923

5. Sicree, R. A., Zimmet, P. Z., Dunstan, D. W., Cameron, A. J., Wel-born, T. A., and Shaw, J. E. (2008). Differences in Height Explain Gender Differences in the Response to the Oral Glucose Tolerance Test the Aus Diab Study. Diabetic Medicine **25(3):**296-302

6. Snijder M. B., Dekker J. M, Visser, M, Bouter L. M, Stehouwer C. D. A, Kostense P. J, Yudkin J. S, Heine R. J, Nijpels G, and Seidell J. C (2003). Association of Hip and Thing Circumferences Independent of Waist Circumference with the Incidence of Type 2 Diabetes: the Hoorn Study. The American Journal of Clinical Nutrition **77:** 1192-1197

7. Bozorgmanesh M., Hadaegh F, Zabetian A, and Azizi F. (2011). Impact of Hip Circumference and Height on Incident Diabetes: Result from 6-year Follow-up in the Tehran Lipid and Glucose Study. Diabetic Medicine **28:** 1330-1336

8. Wang S. L., Pan W. H, Hwu C. M, Ho L. T, Lo C. H, Lin S. L, and Jong Y. S. (1997). Incidence of NIDDM and the Effects of Gender, Obesity, and Hyperinsulinaemia in Taiwan. Diabetologia **40:** 1431-1438

9. Njolstad I., Amesen E, and Lund-Larsen P. G (1998). Sex-differences in Risk Factors for Clinical Diabetes Mellitus in a General Population**:** a 12-years Follow-up of the Finnmark Study. American journal of Epidemiology **147:** 49-58.

10. Lorenzo C., Williams K, Stern M. P, and Haffner S. M. (2009). Height, Ethnicity and the Incidence of Diabetes: the San Antonio Heart Study. Metabolism **58:** 1530-1535.

11. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika **73:** 13–22.

12. Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). Robust Statistics: Theory and Methods. Wiley Series in Probability and Statistics.

13. Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. Akademiai Kiado, Budapest: 267-281.

14. Pan, W. (2001a). Akaike's Information Criterion in Generalized Estimating Equations. Biometrics **57:** 120-125.

15. Pan, W. (2001b). Model Selection in Estimating Equations. Biometrics **57:** 529-534.

16. Pan, W., and Lee, C. T. (2001). Bootstrap Model Selection in Generalized Linear Models. Journal of Agricultural, Biological & Environmental Statistics **6:** 49-61.

17. Cantoni, E., Flemming, J. M., and Ronchetti, E. (2005). Variable Selection for Marginal Longitudinal Generalized Linear Models. Biometrics **61:** 507-514.

18. Cantoni, E., Flemming, J. M., and Ronchetti, E. (2008). Longitudinal Variable Selection by Cross-validation in the Case of Many Covariates. Statistics in Medicine **26:** 919–930.

19. Hin, L. Y., and Wang, Y. G. (2009). Working-correlation-structure Identification in Generalized Estimating Equations. Statistics in Medicine **28(4):** 642-658.

20. Kleinbaum D. G., and Klein M. (2005). Survival Analysis: A Self-Learning Text, 2^{nd} edition. ISBN: Springer-Verlag New York, Inc; 105-127.

21. WHO. (2007). World Health Organization. "Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications: Report of a WHO Consultation. Part 1. Diagnosis and classification of diabetes mellitus".

22. WHO/IDF. (2006). Definition and Diagnosis of Diabetes Mellitus and Intermediate Hyperglycemia: Report of a WHO/IDF Consultation. Geneva: World Health Organization. p. 21. ISBN 978-92-4-159493-6.

23. Centers for Disease Control and Prevention (CDC) and National Center for Chronic Disease Prevention and Health Promotion. (2009). The Power of Prevention: Chronic Disease: The Public Health Challenge of the 21st Century. Atlanta, GA:CDC, http://www.cdc.gov/chronicdisease/pdf/2009-power-of-prevention.pdf

24. Hirschhorn J. N., Lindgren C. M., Daly M. J. *et al.* (2001). Genomewide Linkage Analysis of Stature in Multiple Populations Reveals Several Regions with Evidence of Linkage to Adult Height. American Journal of Human Genetics **69:** 106-116.

25. Park H. S., Yim K. S., and Cho S. I. (2004). Gender Differences in Familial Aggregation of Obesity-related Phenotypes and Dietary Intake Pattern in Korean Families. Annals of Epidemiology **14:** 486-491.

26. Li J. K., Ng M. C., So W. Y. *et al.* (2006). Phenotype and Genetic Clustering of Diabetes and Metabolic Syndrome in Chinese Families with Type 2 Diabetes Mellitus. Diabetes/Metabolism Research and Reviews **22:** 46-52

Table 1: Values of QIC for all possible models, keeping variable Height fixed under different correlation structures

Model	covariates								QIC values		
no.	height	age	hrd	edu	gender	phex	area	com	exchange	AR-1	unstructured
1	x	-	-	-	-	-	-	-	12260.9	12261.5	12261.1
2	x	x	-	-	-	-	-	-	12263.5	12279.4	12270.8
3	x	-	x	-	-	-	-	-	12261.9	12262.2	12261.8
4	x	-	-	x	-	-	-	-	12216.8	12217.6	12217.1
5	x	-	-	-	x	-	-	-	12256.3	12257.0	12256.5
6	x	-	-	-	-	x	-	-	12265.2	12265.8	12265.3
7	x	-	-	-	-	-	x	-	12265.4	12266.2	12265.5
8	x	-	-	-	-	-	-	x	12265.4	12264.7	12265.0
9	x	x	x	-	-	-	-	-	12264.5	12280.1	12271.5
10	x	x	-	x	-	-	-	-	12218.9	12234.5	12226.3
11	x	x	-	-	x	-	-	-	12259.8	12275.5	12267.0
12	x	x	-	-	-	x	-	-	12268.0	12283.5	12274.8
13	x	x	-	-	-	-	x	-	12268.0	12284.0	12275.3
14	x	x	-	-	-	-	-	x	12267.8	12282.3	12274.4
15	x	-	x	x	-	-	-	-	12218.0	12218.3	12217.8
16	x	-	x	-	x	-	-	-	12257.3	12257.7	12257.0
17	x	-	x	-	-	x	-	-	12266.2	12266.5	12266.0
18	x	-	x	-	-	-	x	-	12266.0	12267.0	12266.3
19	x	-	x	-	-	-	-	x	12266.4	12265.5	12265.8
20	**x**	-	-	**x**	**x**	-	-	-	**12211.9**	**12212.6**	**12212.0**
21	x	-	-	x	-	x	-	-	12221.9	12222.6	12222.0
22	x	-	-	x	-	-	x	-	12221.0	12222.0	12220.9
23	x	-	-	x	-	-	-	x	12221.3	12220.7	12221.0
24	x	-	-	-	x	x	-	-	12260.8	12261.4	12260.8
25	x	-	-	-	x	-	x	-	12260.7	12261.5	12260.8
26	x	-	-	-	x	-	-	x	12261.0	12260.0	12260.5
27	x	-	-	-	-	x	x	-	12269.7	12270.5	12269.7
28	x	-	-	-	-	x	-	x	12269.7	12269.0	12269.0
29	-	-	-	-	-	-	x	x	12269.9	12269.4	12269.5
30	x	x	x	x	-	-	-	-	12219.8	12235.0	12227.0
31	x	x	x	-	x	-	-	-	12260.8	12276.2	12268.0
32	x	x	x	-	-	x	-	-	12268.6	12284.2	12276.0
33	x	x	x	-	-	-	x	-	12268.9	12284.8	12276.0
34	x	x	x	-	-	-	-	x	12268.8	12283.0	12275.1
35	x	x	-	x	x	-	-	-	12214.9	12230.3	12222.0
36	x	x	-	x	-	x	-	-	12223.8	12239.5	12231.2
37	x	x	-	x	-	-	x	-	12222.7	12238.6	12230.2
38	x	x	-	x	-	-	-	x	12223.0	12237.4	12229.9
39	x	x	-	-	x	x	-	-	12264.1	12279.9	12271.2
40	x	x	-	-	x	-	x	-	12264.2	12280.1	12271.4
41	x	x	-	-	x	-	-	x	12264.2	12278.5	12270.7
42	x	x	-	-	-	x	x	-	12272.1	12288.2	12279.4
43	x	x	-	-	-	x	-	x	12272.0	12286.4	12278.5
44	x	x	-	-	-	-	x	x	12272.3	12287.0	12278.9
45	x	-	x	x	x	-	-	-	12212.8	12213.4	12212.8
46	x	-	x	x	-	x	-	-	12222.9	12223.4	12222.8
47	x	-	x	x	-	-	x	-	12221.7	12222.4	12221.7
48	x	-	x	x	-	-	-	x	12222.3	12221.5	12221.7
49	x	-	x	-	x	x	-	-	12261.8	12262.2	12261.6
50	x	-	x	-	x	-	x	-	12260.7	12261.5	12260.8
51	x	-	x	-	x	-	-	x	12261.0	12260.0	12260.5
52	x	-	x	-	-	x	x	-	12269.7	12270.5	12269.7
53	x	-	x	-	-	x	-	x	12269.7	12269.0	12269.0
54	x	-	x	-	-	-	x	x	12269.9	12269.4	12269.5
55	x	-	-	x	x	x	-	-	12217.1	12217.8	12217.2
56	x	-	-	x	x	-	x	-	12215.6	12216.5	12215.8
57	x	-	-	x	x	-	-	x	12216.4	12215.8	12216.1
58	x	-	-	x	-	x	x	-	12225.8	12226.7	12226.0
59	x	-	-	x	-	x	-	x	12226.3	12225.8	12226.0

60	x	-	-	x	-	-	x	x	12225.2	12224.8	12224.9
61	x	-	-	-	x	x	x	-	12265.1	12266.0	12265.2
62	x	-	-	-	x	x	-	x	12265.3	12264.7	12264.8
63	x	-	-	-	x	-	x	x	12265.3	12264.8	12264.9
64	x	-	-	-	-	x	x	x	12274.1	12273.7	12273.7
65	x	x	x	x	x	-	-	-	12215.9	12231.1	12222.9
66	x	x	x	x	-	x	-	-	12224.8	12240.2	12231.9
67	x	x	x	x	-	-	x	-	12223.7	12239.3	12230.9
68	x	x	x	x	-	-	-	x	12224.1	12238.1	12230.6
69	x	x	x	-	x	x	-	-	12265.0	12280.6	12271.9
70	x	x	x	-	x	-	x	-	12265.2	12280.8	12272.1
71	x	x	x	-	x	-	-	x	12265.2	12279.2	12271.4
72	x	x	x	-	-	x	x	-	12273.1	12288.9	12280.1
73	x	x	x	-	-	x	-	x	12272.9	12287.2	12279.2
74	x	x	x	-	-	-	x	x	12273.3	12287.7	12279.6
75	x	x	-	x	x	x	-	-	12220.0	12235.4	12227.2
76	x	x	-	x	x	-	x	-	12218.7	12234.3	12225.9
77	x	x	-	x	x	-	-	x	12219.2	12233.2	12225.9
78	x	x	-	x	-	x	x	-	12227.7	12243.6	12235.1
79	x	x	-	x	-	x	-	x	12228.1	12242.3	12234.8
80	x	x	-	x	-	-	x	x	12227.1	12241.4	12233.8
81	x	x	-	-	x	x	x	-	12268.5	12284.5	12275.6
82	x	x	-	-	x	x	-	x	12268.5	12282.8	12274.9
83	x	x	-	-	x	-	x	x	12268.6	12283.1	12275.1
84	x	x	-	-	-	x	x	x	12276.5	12291.1	12283.0
85	x	-	x	x	x	x	-	-	12218.0	12219.0	12217.9
86	x	-	x	x	x	-	x	-	12217.0	12217.3	12216.6
87	x	-	x	x	x	-	-	x	12217.4	12216.6	12216.8
88	x	-	x	x	-	x	x	-	12226.8	12227.5	12226.7
89	x	-	x	x	-	x	-	x	12227.0	12226.5	12226.7
90	x	-	x	x	-	-	x	x	12226.2	12225.6	12225.6
91	x	-	x	-	x	x	x	-	12266.1	12266.7	12266.0
92	x	-	x	-	x	x	-	x	12266.3	12265.4	12265.6
93	x	-	x	-	x	-	x	x	12266.2	12265.5	12265.6
94	x	-	x	-	-	x	x	x	12275.0	12274.4	12274.4
95	x	-	-	x	x	x	x	-	12220.8	12222.0	12221.0
96	x	-	-	x	x	x	-	x	12221.6	12221.0	12221.2
97	x	-	-	x	x	-	x	x	12220.2	12219.8	12219.9
98	x	-	-	x	-	x	x	x	12230.2	12229.9	12229.9
99	x	-	-	-	x	x	x	x	12270.0	12269.0	12269.2
100	x	x	x	x	x	x	-	-	12221.0	12236.2	12227.9
101	x	x	x	x	x	-	x	-	12219.6	12235.0	12226.7
102	x	x	x	x	x	-	-	x	12220.0	12234.0	12226.6
103	x	x	x	x	-	x	x	-	12228.7	12244.3	12235.8
104	x	x	x	x	-	x	-	x	12229.1	12243.1	12235.5
105	x	x	x	x	-	-	x	x	12228.0	12242.2	12234.5
106	x	x	x	-	x	x	x	-	12269.5	12285.2	12276.4
107	x	x	x	-	x	x	-	x	12269.5	12284.0	12275.6
108	x	x	x	-	x	-	x	x	12269.6	12284.0	12275.8
109	x	x	x	-	-	x	x	x	12277.4	12292.0	12283.7
110	x	x	-	x	x	x	x	-	12223.8	12239.4	12231.0
111	x	x	-	x	x	x	-	x	12224.3	12238.3	12230.9
112	x	x	-	x	x	-	x	x	12223.0	12237.2	12229.6
113	x	x	-	x	-	x	x	x	12232.0	12246.4	12238.7
114	x	x	-	-	x	x	x	x	12272.9	12287.4	12279.3
115	x	-	x	x	x	x	x	-	12221.8	12222.6	12221.7
116	x	-	x	x	x	x	-	x	12222.6	12221.8	12221.9
117	x	-	x	x	x	-	x	x	12221.1	12220.6	12220.6
118	x	-	x	x	-	x	x	x	12231.2	12230.6	12230.6
119	x	-	x	-	x	x	x	x	12270.7	12270.0	12269.9
120	x	-	-	x	x	x	x	x	12225.4	12225.0	12225.0
121	x	x	x	x	x	x	x	-	12224.8	12240.2	12231.7
122	x	x	x	x	x	x	-	x	12225.3	12239.1	12231.6
123	x	x	x	x	x	-	x	x	12224.0	12238.0	12230.0
124	x	x	x	x	-	x	x	x	12233.0	12247.2	12239.4

125	x	x	x	-	x	x	x	x	12273.9	12288.2	12280.0
126	x	x	-	x	x	x	x	x	12228.1	12242.3	12234.7
127	x	-	x	x	x	x	x	x	12226.3	12225.8	12225.7
128	x	x	x	x	x	x	x	x	12229.0	12243.1	12235.4

Table 2: Correlation Information Criterion (CIC)values under different correlation structures

Correlation Structures	Exchangeable	AR-1	Unstructured
Value of CIC	14.1	14.1	14.0

Table 3: GEE estimates of regression coefficients under selected model using unstructured correlation with standard errors, p-values and 95 % confidence intervals

Variables	Estimates	Standard Error	p-value	95 % Confidence Interval
Constant	4.1501	0.988	0.000	(2.21, 6.09)
Height	-0.020	0.007	0.004	(-0.03, -0.01)
Education	-0.569	0.144	0.000	(-0.85, -0.29)
Gender	-0.214	0.120	0.074	(-0.45, 0.02)

Table 4: Risk Ratios for covariates under selected model with standard errors, p-values and 95 % confidence intervals

Variables	Risk Ratio	Standard Error	p-value	95 % Confidence Interval
Height				
Q3 vs. Q1	0.78	0.021	0.000	(0.74, 0.82)
Q3 vs. Q2	0.91	0.015	0.004	(0.88, 0.94)
Q2 vs. Q1	0.86	0.018	0.000	(0.82, 0.89)
Education	0.58	0.081	0.000	(0.42, 0.73)
Gender	0.81	0.095	0.000	(0.62, 0.99)

Affiliation:

Md Erfanul Hoque
Department of Statistics, Biostatistics & Informatics,
University of Dhaka
Dhaka-1000,
Bangladesh
E-mail: imerfan49@yahoo.com

Mahfuzur Rahman Khokan
Department of Statistics, Biostatistics & Informatics,
University of Dhaka
Dhaka-1000,
Bangladesh
E-mail: mahfuz_sbi34@yahoo.com

Wasimul Bari
Department of Statistics, Biostatistics & Informatics,
University of Dhaka
Dhaka-1000,
Bangladesh
E-mail: w_bari@yahoo.com

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 44
October 2015

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2014-03-03
Accepted: 2014-09-04