PJ PALGO JOURNAL OF MEDICINE AND MEDICAL SCIENCE
ISSN 2476-8340
Volume 5 Issue 3, April 2018.Page 33- 42
http://www.palgojournals.org/PJMMS/Index.htm
Corresponding Authors Email:afolabikennyi@yahoo.co.uk

USING GENERALIZED ESTIMATING EQUATION (GEE) TO ANALYSE THE INFLUENCE OF SOME FACTORS ON THE STATE OF HEALTH OF DIABETES PATIENTS

A. Adedayo Adepoju[1*] and Kehinde Isreal Afolabi[2]

Department of Statistics,University of Ibadan,Ibadan, Nigeria[1*]
University Teaching Hospital, Ibadan, Nigeria[2]

Accepted 22 May, 2018

In longitudinal studies, observations measured repeatedly from the same subject over time are serially correlated. One objective of statistical analysis is to describe the marginal expectation of the outcome variable as a function of the covariates while accounting for the correlation among the repeated observations for a given subject. Generalized Estimating Equation (GEE) is a general statistical approach to fit a marginal model for longitudinal/clustered data analysis, and it has been popularly applied into clinical trials and biomedical studies. Generalized linear Model (GLM)on the other hand has been widely used in fitting a regression to a set of data of dependent variables depending solely on a/some set of covariates with the different set of distributions and their link function and its use has been extended to longitudinal data.This paper examines the effects of some factors; age, sex, Body Mass Index (BMI), blood pressure, exercise and glucose tolerance on the health status of 840 diabetes patients attending clinic over a period of five years using the generalized linear model and the generalized estimating equations methods. The GEE performs better than the GLM. The result reveals that glucose tolerance, blood pressure and BMI are the important factors that affect the state of health of these patients.

Key words: Generalized Estimating Equation, Longitudinal data; Serial correlation; Covariates; Diabetes patients

1.0 INTRODUCTION

Observations measured repeatedly from the same subject over time are serially correlated in longitudinal studies. When observations are measured on a continuous scale, the dependency structure between observations can be modelled in a covariance matrix with non-constant variances and non-zero covariances. There are different types of covariance matrices to model such dependency between observations from the same subject. The method of generalized estimating equations (GEE) is often used to analyze longitudinal and other correlated response data, particularly if responses are binary. The generalized estimating equations (GEE) method, an extension of the quasi-likelihood approach by Wedderburn, is being increasingly used to analyze longitudinal and other correlated data, especially when they are binary or in the form of counts [Burton, P., Gurrin, L. and Sly, P. (1998), Chan, Jennifer S.K. (2014)].

In statistics, GEE is used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes. Parameter estimates from the GEE are consistent even when the covariance structure is misspecified under mild regularity conditions. The focus of the GEE is on estimating the average response over the population ("population-averaged" effects) rather than the regression parameters that would enable prediction of the effect of changing one or more covariates on a given individual [Wang, Ming (2014,]. Diggle, P. J.,et al (1994) The GEEs are usually used in conjunction with Huber–White standard error estimates also known as "robust standard error" or "sandwich variance" estimates. In the case of a linear model with a working independence variance structure, these are known as "heteroscedasticity consistent standard error" estimators. Indeed, the GEE unified several independent formulations of these standard error estimators in a general framework [Chan, Jennifer S.K. (2014) Dolcini, M.M. and Adler, N.E. (1994).,

Generalized Estimating Equationx(GEEs) belong to a class of semi parametric regression techniques because they rely on specification of only the first two moments. They are a popular alternative to the likelihood–based generalized linear mixed model which is more sensitive to variance structure specification. They are commonly used in large epidemiological studies, especially multi-site cohort studies because they can handle many types of unmeasured

dependence between outcomes. Some examples of these studies include cross-over trials, cluster data and longitudinal data. The mixed logistic model is usually used to model the heterogeneity between the subjects and the correlation among the repeated observations [,Wang, Ming (2014), Hardin, J.W and Hilbe, J.M. (2003) , and Liang, K. Y. and Zeger, S. L. (1993)]. It is well known that using maximum likelihood to estimate the expected parameters and variance components is computational difficult. This approach does not require the complete specification of the joint distribution of the repeated responses but rather only thetwo moments [Thompson, A. M., Humbert, M. L. *et al* (2003]. The major advantages of GEE are: it provides a consistent estimate for the regression parameter even when the correlation matrix is mis-specified, it indicates that the efficiency loss relative to the maximum likelihood is small, it extends this approach for correlated binary data by specifying supplementary generalized estimating equations, it is based on the empirical pair wise covariances that permit estimation of the correlated parameters, it is used to estimate the regression parameters through the use of log it approximations by the probit function while the variance components are estimated empirically [Chan, Jennifer S.K. (2014), Shock. N. W. et al (1984) and Hay, J. L. and Pettit, A. N. (2001).].

Generalized Estimating Equations(GEEs) include an additional variance component to accommodate correlated data, and to allow for differences among clusters. GEEs have several favourable properties for ecological analyses; for example, parameter estimates and empirical standard errors are robust to misspecification of the correlation structure [Overall, J. E. and Tonidandel, S. (2004)]. GEEs have been used extensively in a variety of disciplines, such as epidemiology [Wu, Y. B. et al(1999)] and political science [Zorn, C. J. W. (2001).]. In ecology, they have been used to control for lack of independence among nests clustered within sites [Driscoll, M. J. L.,et al (2005).] and among related species [Duncan, R. P. (2004)].

Over the past two decades, GEE approach has been developed and advanced. For example, Zhao and Prentice [Zhao, L.P., Prentice, R.L. (1990)] and Lipsitz et al. and Lipsitz, S. R. and Fitzmaurice, G. M. (1994).] apply the GEE approach to analyze binary data, Kenward et al. consider ordinal data, Lipsitz et al. use categorical data, Prentice and Zhao focus on multivariate data, Park, T. (1993) compares GEE approach to maximum likelihood approach and Miller et al. to weighted least squares approach, etc. [Chan, Jennifer S.K. (2014).].

The aim of this research is to study the effects of age, sex, Body Mass Index (BMI), blood pressure, exercise and glucose tolerance on 840 diabetes patients who attended clinic for a period of five yearsusing the generalized estimating equation method. Since there is often some confusion between generalized estimation equation (GEE) and the generalized linear model (GLM), the results of both estimators are compared.

The rest of the paper is organised as follows: section 2 discusses the materials and methods used in this study, analysis and discussion of the results are presented in section 3 while conclusion is discussed in section 4.

2.0 MATERIALS AND METHODS

2.1.Generalized Estimating Equations

Generalized Estimating Equations (GEEs) are methods of parameter estimation for correlated data. They handle problems of longitudinal data analysis by assessing how the mean dependent on variable changes over time, while separately dealing with the nuisance covariance among the observations within subjects in order to get a better estimate and valid significance tests. When data are collected on the same units across successive points in time, these repeated observations are correlated over time. If this correlation is not taken into account, then the standard errors of the parameter estimates will not be valid and statistical inference made with such estimates will be inefficient [Fitzmaurice, G., *et al* (2004)].

Three components are important in the GEE; generalized estimating equations require a model for the mean response (as a function of covariates), the variance (often specified as a function of the mean), and a working correlation assumption. They are semi-parametric because estimates rely on parametric assumptions regarding the mean and variance/covariance, but they are not fully parametric (i.e. they require no other distributional assumptions). Most problems arise from the model construction, the standard regression model comprises of dependent variable, independent variable(s) and the disturbance term [Fitzmaurice, G., *et al* (2004]. Unlike the marginal model, the disturbance term is not included; the model depends on the first and the second moments which is why it is being referred to as semi-parametric model. However, this does not indicate the negligence of the disturbance term; it has been accounted for in the score equation that is used to estimate the parameters of interest [Koper, Nicola and Manseau, Micheline (2009)].

Liang and Zeger introduced the generalized estimating equations (GEE), which extend the use of generalized linear models (GLM) to longitudinal data. Generalized estimating equations are used in regression analysis of longitudinal data, where observations on the same subject are correlated [6].

A linear function of regressors is given by,

$$\psi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik} \quad \ldots (1)$$

The regressors X_{ij} are pre-specified functions of the explanatory variables, ψ_i is the linear predictor and $\beta's$ are the coefficients.

A smooth and invertible linearizing link function g(.) which transforms the expectation of the response variable

$$\mu_i = E(Y_i) \quad \ldots (2)$$

to the linear predictor

$$g(\mu_i) = \psi_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \quad \ldots (3)$$

$$\mu_i = g^{-1}(\psi_i) = g^{-1}(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik}) \quad \ldots (4)$$

The inverse link $g^{-1}(.)$ is also called the mean function where μ_i is the expected value of the response.

In most research papers, GEE are often regarded as extension of GLM which accounts for correlation [Zeger, S. L. and Liang, K. Y. (1986) and Wedderburn, R. W. (1974)] This requires providing some information used in fitting the desired regression model.

Stated below are the required information needed to fit a GEE regression model:

- The distribution of the explained variable.
- The form of relationship that exists between the explained and the explanatory variable (i.e. link function).
- The explained variable.
- The mean-model (if interested in population effect).
- The mean-variance model.
- The correlation model.
- The correlation structure of the repeated measurements.

The following sections illustrate the various forms of mean model, mean-variance model and correlation model that exist.

2.2 The GEE Mean Model

The mean model also known as the marginal population model is given as follows:

$$E(Y_{it} / X_{it}) = \mu_{it} = g^{-1}(x_{it}{}^T \beta) \quad \ldots (5)$$

Where:

$\mu =$ Marginal response

$g(\cdot) =$ Link function

$x_{it} =$ Covariates

$\beta =$ Regression parameter

The link function helps in defining the model as it relates the predictors to the outcome. The choice of link function depends on the probability distribution. Examples of common link functions are:

Table 1: Some Common Link functions and Their Inverses

Link	$\psi_i = g(\mu_i)$	$\mu_i = g^{-1}(\psi_i)$
Identity	μ_i	ψ_i
Log	$\log_e \mu_i$	$\ell \psi_i$
Inverse	μ_i^{-1}	ψ_i^{-1}
Inverse-square	μ_i^{-2}	$\psi_i^{\frac{-1}{2}}$
Square=root	$\sqrt{\mu_i}$	ψ_1^{2}
Logit	$\log_\ell \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+\ell^{-\psi_i}}$

Probit $\Phi^{-1}(\mu_i)\ \Phi(\psi_i)$

Log-log $-\log_\ell[-\log_\ell(\mu_i)]\ \exp[-\exp(-\psi_i)]$

Complementary log-log $\log_\ell[-\log_\ell(1-\mu_i)]\ 1-\exp[-\exp(\psi_i)]$

B. Generalized Linear Models

Normal Linear Model

$$E[Y_i] = \mu_i = X_i^T\beta;\ Y_i \sim N(\mu_i, \sigma^2) \quad \text{... (6)}$$

Where $Y_1,...,Y_N$ are independent

Here the link function is the identify function

$g(\mu_i) = \mu_i$

The model is usually written in the form

$y = X\beta + \ell$

Where

$e = \begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}$ andthe $e_i's$ are the independent, identically distributed random variables with

$e_i \sim N(0, \sigma^2)$ For $i = 1,...,N$

C. Maximum Likelihood (MLE) Estimation of Generalized Linear Models

Consider independent random variables $Y_1, \ldots, Y_N$ satisfying the properties of a generalized linear model.

We wish to estimate parameters β which are related to the $Y_i's$ through $E[Y_i] = \mu_i$ and

$g(\mu_i) = X_i^T\beta$.

For each Y_i, the log-likelihood function is

$$l_i = y_i b(\theta_i) + c(\theta_i) + d(y_i) \quad \ldots(7)$$

Also

$$E[Y_i] = \mu_i = -\frac{c'(\theta_i)}{b'(\theta_i)} \quad \ldots(8)$$

$$Var[Y_i] = \frac{\left[b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)\right]}{\left[b'(\theta_i)\right]^3} \quad \ldots(9)$$

And $$g(\mu_i) = X_i^T\beta = \eta_i \quad \ldots(10)$$

Where X_i is a vector with elements x_{ij}, $j = 1,..., p.$ The log-likelihood function for all the $Y_i's$ is

$$l = \sum_{i=1}^{N} l_i = \sum y_i b(\theta_i) + \sum c(\theta_i) + \sum d(y_i)$$

To obtain the MLE for the parameter β_j we need

$$\frac{\partial l}{\partial \beta_j} = S_j = \sum_{i=1}^{N}\left[\frac{\partial l_i}{\partial \beta_j}\right] = \sum_{i=1}^{N}\left[\frac{\partial l_i}{\partial \beta_j}.\frac{\partial \theta_i}{\partial \mu_i}.\frac{\partial \mu_i}{\partial \beta_j}\right] \quad \ldots(11)$$

Using the chain rule for differentiation, we will consider each term on the equation (9) separately.

First, differentiating (5) and substituting (6)

$$\frac{\partial l_i}{\partial \theta_i} = y_i b^{'}(\theta_i) + c^{'}(\theta_i) = b^{'}(\theta_i)\left(y_i - \mu_i\right)$$

Second, differentiate (6)

$$\frac{\partial \mu_i}{\partial \theta_i} = \frac{c^{''}(\theta_i)}{b^{'}(\theta_i)} + \frac{c^{'}(\theta_i) b^{''}(\theta_i)}{\left[b^{'}(\theta_i)\right]^2}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b^{'}(\theta_i)\,\mathrm{var}(Y_i)$$

Finally, from (4)

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta_j} = \frac{\partial \mu_i}{\partial \eta_i} x_{ij}$$

Hence, the score given in (9) is

$$S_j = \sum_{i=i}^{N}\left[\frac{\left(y_i - \mu_i\right)}{\mathrm{var}(Y_i)} xij \left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right] \quad \ldots (10)$$

The variance- covariance matrix of the S_j has terms

$$I_{jk} = E\left[S_j S_k\right]$$

which forms the information matrix I.
From (10),

$$I_{jk} = E\left\{\sum_{i=1}^{N}\left[\frac{\left(y_i - \mu_i\right)}{\mathrm{var}(Y_i)} x_{ik}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]\sum_{i=1}^{N}\left[\frac{\left(y_i - \mu_i\right)}{\mathrm{var}(Y_i)} x_{ik}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right]\right\}$$

$$I_{jk} = \sum_{i=1}^{N}\frac{E\left[\left(y_i - \mu_i\right)^2\right] x_{ij} x_{ik}}{\left[\mathrm{var}(Y_i)\right]^2}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \quad \ldots (11)$$

because $E\left[\left(y_i - \mu_i\right)\left(y_i - \mu_i\right)\right] = 0$ for $i \neq j$ as the $Y_i^{'}s$ are independent.

Using $E\left[\left(Y_i - \mu_i\right)^2\right] = \mathrm{var}(Y_i)$, (7) can be simplified to

$$I_{jk} = \sum_{i=1}^{N}\frac{x_{ij} x_{ik}}{\mathrm{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

$$I_{jk} = \sum_{i=1}^{N}\frac{x_{ij} x_{ik}}{\mathrm{var}(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \quad \ldots (12)$$

The estimating equation for the method of scoring generalizes to

$$b^{(m)} = b^{(m-1)} + \left[I^{(m-1)}\right]^{-1} S^{(m-1)} \quad \ldots (13)$$

Where

$\Rightarrow b^{(m)}$ Is the vector of estimates of the parameters $\beta_1, ..., \beta_p$

At the m[th] iteration

$\Rightarrow \left[I^{(m-1)} \right]^{-1}$ Is the inverse of the information matrix with elements I_{jk} given by (12)

$\Rightarrow S^{(m-1)}$ is the vector of elements given (10)

If both sides of equation (13) are multiplied by $I^{(m-1)}$ we obtain

$$I^{(m-1)}b^{(m)} = I^{(m-1)}b^{(m-1)} + S^{(m-1)} \quad \ldots(14)$$

From (12), I can be written as

$$I = X^T WX$$

Where

W is the N X N diagonal matrix with elements

$$w_u = \frac{1}{\mathrm{var}\left(Y_i\right)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \quad \ldots(15)$$

The expression on the right-hand side of (10) is the vector with elements

$$\sum_{k=1}^{P}\sum_{i=1}^{N}\frac{x_{ij}x_{ik}}{\mathrm{var}\left(Y_i\right)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 b_k^{(m-1)} + \sum_{i=1}^{N}\frac{\left(y_i-\mu_i\right)x_{ij}}{\mathrm{var}\left(Y_i\right)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$$

evaluated at $b^{(m-1)}$; this follows from equations (12) and (10) .Thus the right hand side of equation (14) can be written as

$$X^T WZ$$

Where Z has elements

$$Z_i = \sum_{k=1}^{P} x_{ik}b_k^{(m-1)} + \left(y_i - \mu_i\right)\left(\frac{\partial \eta_i}{\partial \mu_i}\right) \quad \ldots(16)$$

With μ_i and $\frac{\partial \eta_i}{\partial \mu_i}$ evaluated at $b^{(m-1)}$.

Hence, the iterative equation, can be written as

$$X^T WXb^{(m)} = X^T WZ \quad \ldots(17)$$

Thus for GLMs, MLEs are obtained by an iterative weighted least squares procedure.
Most statistical packages that include procedures for fitting GLM model have an efficient algorithm based on (17).

The following steps are considered;

(i) Begin by using some initial approximation $b^{(o)}$ to evaluate Z and W.

(ii) Solve (17) to give $b^{(1)}$ which in turn is used to obtain better approximations for Z and W, and so on until adequate convergence is achieved.

(iii) When the difference between successive approximations $b^{(m-1)}$ and $b^{(m)}$ is sufficiently small, $b^{(m)}$ is taken as the maximum likelihood estimate

3.0 DATA ANALYSIS

The data used in this study comprised 840 diabetes patients observed for a period of five years to examine the effects of the following factors; age, sex, Body Mass Index (BMI), blood pressure, exercise and glucose tolerance on their state of health using theGeneralised Linear Model (GLM) and Generalised Estimating Equations (GEE) methods.

Table 3 gives the results of the analysis using the generalized linear model. The intercept is the only significant coefficient while the other covariates are insignificant. The intercept however does not have any effect on the patients' health status. The Table shows that the following coefficients; age, BMI and glucose tolerance have negative effect to the patients' health status while blood pressure, sex and exercise have minimal effect on their health status. This result

Table 2: Correlation Model

Correlation type	Correlation formula	Working Correlation Structure
Independence	$Cor\left(Y_{ij}Y_{ik}=o\right), j\neq k$ or $Cor\left(Y_{ij}Y_{ik}=1\right), j=k$	$R(\alpha)=\begin{pmatrix}1 & \alpha & ... & 0\\ 0 & 1 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1\end{pmatrix}=I$
Exchangeable	$Cor\left(Y_{ij}Y_{ik}=\alpha\right), j\neq k$ $Cor\left(Y_{ij}Y_{ik}=1\right) i, j=1$	$R\left(\alpha\right)=\begin{pmatrix}1 & \alpha & \cdots & \alpha\\ \alpha & 1 & \cdots & \alpha\\ \vdots & \vdots & \ddots & \vdots\\ \alpha & \alpha & \cdots & \alpha\end{pmatrix}$
AR(1)	$Cor\left(Y_{ij}Y_{ik}=\alpha^{\left\|j-k\right\|}\right), j\neq k$ $Cor(Y_{ij},Y_{i+k})=e^{k}$	$R\left(\alpha\right)=\begin{pmatrix}1 & \alpha & \cdots & \alpha^{\left\|j-1\right\|}\\ \alpha & 1 & \cdots & \alpha^{\left\|j-2\right\|}\\ \vdots & \vdots & \ddots & \vdots\\ \alpha^{\left\|j-1\right\|} & \alpha^{\left\|j-2\right\|} & \cdots & 1\end{pmatrix}$
Unstructured	$Cor\left(Y_{ij}Y_{ik}=\alpha_{jk}\right), j\neq k$ $Cor\left(Y_{ij}Y_{ik}=1\right), j=k$	$R(\alpha)=\begin{pmatrix}1 & \alpha & \cdots & \alpha_{1j}\\ \vdots & \vdots & \ddots & \vdots\\ \alpha_{1j} & \alpha_{2j} & \cdots & 1\end{pmatrix}$

shows that GLM may not be an appropriate method for fitting this data, hence a more appropriate or robust method should be considered. Since the dependent variable is a dichotomous variable describing whether the covariates accumulate to improve the health of the patients or not, the GEE which is a more robust model is therefore used as follows:

Table 3: Fitted Generalized Linear Model

	Estimate	Std. Error	t- value	Pr (>|t|)
(Intercept)	1.5631263	0.1770657	8.828	<2e-16***
Age	-0.0002037	0.0015883	-0.128	0.898
Sex	0.0076264	0.0760843	0.100	0.920
BMI	-0.0039659	0.0061119	-0.649	0.517
Blood Pressure	0.0017394	0.0378850	0.046	0.963
Exercise	0.0037654	0.0760360	0.050	0.961
GT	-0.0033931	0.1314900	-0.026	0.979

3.1 Generalized Estimating Equations (GEE)

The dependent variable being a dichotomous variable of whether or not the covariates accumulated influence the health status, a more robust model of GEE model was used with the binary identity restriction of the dependent variable.
Table 4 shows that Body Mass Index, Blood Pressure and Glucose Tolerance are the significant covariates whereas the remaining factors; age, sex and exercise are not significant.

Table 5 compares the general linear statistics of the two methods considered and it shows that the result of GEE is better than GLM. The standard error of GEE is 0.03 compared to 0.506 for GLM.
The Wald test given in Table 6 suggests that age, blood pressure, exercise and glucose tolerance are Gaussian.
To interpret the group-related effects, we compare these models statistically to determine if the group by time interaction terms is jointly significant or not.

Table 4: Generalized Estimating Equation

	Estimated	Std. Error	Wald test	Pr(>| W|)
(Intercept)	**1.496e+00**	**1.138e-01**	**172.756**	**< 2e-16*** **
Age	**3.010e-09**	**1.053e-07**	**0.001**	**0.977**
Sex	**-4.141e-07**	**6.586e-07**	**0.395**	**0.530**
BMI	**1.9056e-07**	**3.77e-07**	**0.254**	**0.001***
BP	**-6.801e-08**	**5.078e-06**	**0.000**	**0.000****
Exercise	**-2.408e-07**	**5.270e-02**	**0.000**	**1.000**
GT	**2.478e-07**	**9.949e-07**	**0.062**	**4e-9*** **

Table 5: General Linear Model Statistic for GLM and GEE

	GLM	**GEE**
P-value	0.9948	0.0067
Residual standard error	**0.506**	**0.03**
Adjusted R-squared	**-0.02025**	**0.7890**

Table 6: Analysis of Wald Statistic Table

	Df	X^2	p(> |Chi|)
Age	1	0.235	0.63
Sex	1	0.663	0.42
BMI	1	0.663	0.42
BP	1	0.000	1.00
Exercise	1	0.000	1.00
GT	1	0.062	0.80

Table 7: Independence Correlation Structure

	[,1]	**[,2]**	**[,3]**	**[,4]**	**[,5]**	**[,6]**
[1,]	1	-0.5600	-0.9835	-0.5600	0.92575	0.34408
[2,]	-0.5600	1	0.2981	0.9524	-0.16521	-0.22008
[3,]	-0.9835	1.2981	1	0.2981	-0.20546	-0.34002
[4,]	-0.5600	0.9524	0.2981	1	-0.16521	-0.22008
[5,]	0.92575	-0.16521	-0.2055	-0.1652	1	0.92575
[6,]	0.34408	-0.22008	-0.3400	-0.2201	0.92575	1

From Table 7, it can be deduced that physical activeness and Glucose tolerance are correlated, blood pressure and exercise are correlated, and age and Blood pressure are correlated.

Table 8: Exchangeable Correlation Structure

	[,1]	**[,2]**	**[,3]**	**[,4]**	**[,5]**	**[,6]**
[1,]	1	-0.325396	-0.94432	-0.322396	0.076169	0.037863
[2,]	-0.32540	1	-0.01067	0.102057	-0.030771	0.001356
[3,]	-0.94432	-0.10667	1	-0.010667	0.016355	-0.026881
[4,]	-0.32540	0.102057	-0.01067	1	-0.030771	0.001356
[5,]	0.07617	-0.030771	0.01636	-0.030771	1	1
[6,]	0.03786	0.001356	-0.02688	0.001356	1	1

From Table 8, it can be deduced that physical activeness and Glucose Tolerance are exchangeable in terms of exchanging exercise for glucose tolerance.

Table 9: AR (1) Correlation Structure

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0.1781	-0.82553	-0.326551	-0.82553	0.123829	0.037863
[2,]	-0.8255	0.15229	0.160961	0.15229	-0.037879	0.001356
[3,]	-2.3266	0.16096	0.972572	0.16096	0.005093	-0.026881
[4,]	-0.8255	0.15229	0.160961	0.15229	-0.037879	0.001356
[5,]	0.1238	-0.03788	0.005093	-0.03788	0.012236	-0.002046
[6,]	0.2150	-0.01135	-0.099735	-0.01135	-0.002691	0.003641

From Table 9 above, it can be deduced that an approximately Auto-Regressive model of lag will be best fitted for the dichotomous dependent variable.

Table 10: Unstructured Correlation Structure

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0.78639	-0.286925	-0.97371	-0.286925	0.061120	0.044655
[2,]	-0.28693	0.085970	-0.00212	0.085970	-0.027275	0.003930
[3 ,]	-0.97371	-0.002120	0.54066	-0.002120	0.018353	-0.035924
[4,]	-0.28693	0.085970	-0.00212	0.085970	-0.027275	0.003930
[5,]	0.06112	-0.027275	0.01835	-0.027275	0.009733	-0.003533
[6,]	0.04466	0.003930	-0.03592	0.003930	-0.003533	0.005232

4.0 RESULTS AND DISCUSSION

In this study, a general linear model was first used determine the factors contributing to the health status of 840 diabetes patients and also Generalized Estimating Equation restricted to the same factors was also employed. The General Linear Model was unable to accommodate categorical covariates and dependent variables. While the Generalized Estimating Equation was able to make room for categorical covariates and dichotomous dependent variable.

5.0 CONCLUSION

Having considered the General linear Model and Generalized Estimating Equation, It is seen that GEE performed better than GLM with their P-values of 0.00673 and 0.9948 respectively. The results showed that Residual Standard Error of GEE is 0.03 compared to the Residual Standard Error of 0.506 for GLM. Lastly, GEE was able to explain the diabetes contributing factors with Adjusted R-Squared value of 0.789 compared to GLM with -0.02025.

The following coefficients; BMI, Blood pressure and Glucose Tolerance are highly significant to diabetes status while age, sex and exercise merely contributed.

REFERENCES

Burton, P., Gurrin, L. and Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Stat Med, 17:1261–91.

Chan, Jennifer S.K. (2014). Analysis of Correlation Structures using Generalized Estimating Equation Approach for Longitudinal Binary Data, Journal of Data Science 12, 293-305

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). Analysis of longitudinal data. Oxford, United Kingdom: Oxford University Press.

Dolcini, M.M. and Adler, N.E. (1994). Perceived Competencies, Peer Group Affiliation, and Risk Behavior among Early Adolescents Health Psychology, 13 (6), 496-506.

Driscoll, M. J. L., Donovan, T., Mickey, R., Howard, A. and Fleming, K. K. (2005). Determinants of wood thrush nest success: a multi-scale, model selection approach. Journal of Wildlife Management, 69, 699–709.

Duncan, R. P. (2004). Extinction and endemism in the New Zealand avifauna. Global Ecology and Biogeography, 13, 509–517.

Fitzmaurice, G., Larid, N. M. and Ware, J. H. (2004). Applied Longitudinal Data, John Wiley & Sons, 294–295.

Hay, J. L. and Pettit, A. N. (2001). Bayesian analysis of a time series of counts with covariates: an application to the control of an infectious disease, Biostatistics, vol. 2, no.4: 433 – 444.

Hardin, J.W and Hilbe, J.M. (2003). Generalized Estimating Equations. Boca Raton, Florida: CRC Press LLC.

Kenward, M.G., Lesaffre, E., Molenberghs, G. (1994). Application of maximum likelihood and generalized estimating equatins to the analysis of ordinal data from a

Koper, Nicola and Manseau, Micheline (2009). Generalized estimating equations and generalized linear mixed-effects models for

modelling resource selection. Journal of Applied Ecology, Vol. 46, 590–599
Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73:13–22.
Liang, K. Y. and Zeger, S. L. (1993). Regression Analysis for Correlated Data. Annual Review of Public Health, 14: 43 – 68.
Lipsitz, S.R., Laird, N.M., Harrington, D.P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association, Biometrika78, 153-160
Lipsitz, S. R. and Fitzmaurice, G. M.(1994). Sample size for repeated measures studies with binary responses," Statistics in Medicine, vol. 13, no. 12: 1233–1239.
Lipsitz, S. R and Fitzmaurice, G. M. (1996). Estimating equations for measures of association between repeated binary responses. Biometrics,vol.52,no.3:903–912,
longitudinal study with cases missing at random. Biometrics 50, 945-953.
Miller, M.E., Davis, S.C. and Landis, J.R. (1993). The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares, Biometrics 49, 1033-1044.
Overall, J. E. and Tonidandel, S. (2004). Robustness of generalized estimating equation (GEE) tests of significance against misspecification of the error structure model. Biometrical Journal, 46, 203–213.
Park, T. (1993). A comparison of the generalizing estimating equation approach with the maximum likelihood approach for repeated measurements. Statistics in Medicine12, 1723-1732.
Shock. N. W., Greulich, R. C., Costa, P. T., Andres, R., Lakatta, E. G., Arenberg, D. and Tobin, J. D. (1984). Normal Human Aging: The Baltimore Longitudinal Studyof Aging. NIH Publication No. 84-2450, 223 - 233
Thompson, A. M., Humbert, M. L.andMirwald, R. L. (2003). A Longitudinal Study of the Impact of Childhood and Adolescent Physical Activity Experiences on Adult Physical Activity Perceptions and Behaviors, Sage Journals, vol. 13,no. 3. 112-119
Wang, Ming (2014). Generalized Estimating Equations i n Longitudinal Data Analysis: A Review and Recent Developments. Advances in Statistics
Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika, 61:439–47.
Wu, Y. B., Clopper, R.R. and Wooldridge, P. J. (1999). A comparison of traditional approaches to hierarchical linear modeling when analyzing longitudinal data. Research in Nursing and Health, 22, 421–432
Zeger, S. L. and Liang, K. Y. (1986). The analysis of discrete and continuous longitudinal data. Biometrics, 42: 121–130.
Zhao, L.P., Prentice, R.L. (1990). Correlated binary regression using a generalized quadratic model. Wiley Interdisciplinary Reviews: Computational Statistics 77,642-648.
Zorn, C. J. W. (2001). Generalized estimating equation models for correlated data: a review with applications. American Journal of Political Science, 45, 470 – 490.