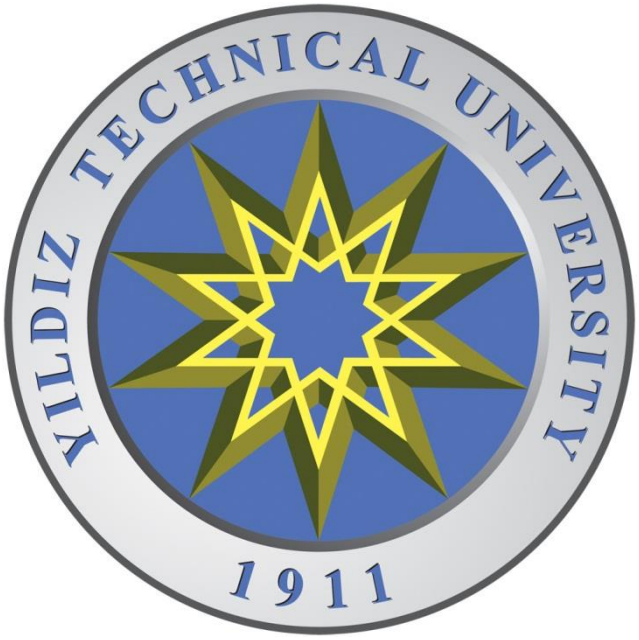


**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

**MSC. THESIS**

**JUN, 2015**

**REPUBLIC OF TURKEY**
**YILDIZ TECHNICAL UNIVERSITY**
**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES**

# APPLICATION OF GENERALIZED ESTIMATING EQUATIONS (GEE) ON REAL LONGITUDINAL DATA

**FATIMA YASSIN**

**MSc. THESIS**
**DEPARTMENT OF STATISTIC**
**PROGRAM OF STATTISTIC**

**ADVISER**
**Assoc. Prof. Dr. FİLİZ KARAMAN**

**İSTANBUL, 2015**

## ACKNOWLEDGEMENTS

This phase is the most important phases in my life , where does not mean just scientific degree but has allowed me to know new country and new people, I would like to thank Turkish government for giving me this scholarship that enabled me to study in Yildiz Technical university one of the most discriminate universities.

Also I would to thank my advisor Assoc. Prof. Dr. Filiz KARAMAN for helping me and very upscale conducting with me.

My family I like to thank you for supporting and encouragement me, especially my lovely mother.

My lovely daughter Reem thank you for your patient, I love you very much.

June, 2015

Fatima YASSIN

## TABLE OF CONTENTS

## LIST OF SYMBOLS

| | |
|---|---|
| $l$ ( ) | Likelihood function |
| $\xi$(y) | Sufficient statistic |
| $\psi$ ( ) | Estimating Equation |
| a($\emptyset$) | Scale parameter |
| $\chi^2$ | Chi statistic |
| $R(\alpha)$ | Correlation matrix |
| $\beta^{PA}$ | Population-average GEE parameter |
| $\beta^{SS}$ | Subject-specific GEE parameter |
| $\Phi$ | Dispersion parameter |

## LIST OF ABBREVIATIONS

| | |
|---|---|
| AFT | Accelerated Failure Time |
| AIC | Akaike Information Criterion |
| AR(1) | Autoregressive Correlation Structure |
| CIC | Coefficient of Internal Consistency |
| FIML | Full Information Maximum Likelihood |
| GEE | Generalized Estimating Equation |
| GEE1 | Models Assume Orthognality |
| GEE2 | Models Doesn't Assumed Orthognality |
| GLM | Generalized Linear Model |
| GLMM | Generalized Linear Mixed Model |
| LIML | Limited Information Maximum Likelihood |
| LIMQL | Limited Information Maximum Quasi Likelihood |
| ML | Maximum Likelihood |
| MLE | Maximum likelihood estimation |
| PA-GEE | Population-average Generalized Estimating Equation |
| SS-GEE | Subject-Specific Generalized Estimating Equation |
| QIC | Quasi-AIC |
| QL | Quasi likelihood |

## LIST OF FIGURES

## LIST OF TABLES

## ABSTRACT

# Application of Generalized Estimating Equations (GEE) for Real Longitudinal Data

Fatima Yassin

Department of Statistic

MSc. Thesis

Adviser: Assoc. Prof. Dr. Filiz KARAMAN

Every statistician's goal is to obtain optimal statistical models and then estimate the parameters of these models to make inferences. Selecting an appropriate statistical method actually requires careful thinking about how data should be collected and what they measure, but normally raw data is sometimes overcome the classical assumptions of statistics which inhibit access to good estimates.

We addressed the problem of correlated data which is one of the most important assumptions in estimation methods. Many solutions were developed to solve this problem; however our goal is always to get sufficient solution. The Generalized Estimating Equations (GEE) have become a popular approach that provides a sufficient solution for the data that occur in correlation condition. An important feature of the GEE methodology takes into account the type of correlation effect ("working correlation structure"). In this thesis, initially we will give an overview of Generalized Linear Models (GLM) algorithms and their techniques of getting estimates; which the GEEs had presented in the GLM environment.

In addition GLMs have an important utility that is ability to derive statistics and properties for group of data which based on likelihood, that makes statisticians create methods within the GLMs framework, but it require independence where sometimes is not applicable. Additionally, we will make comparison for estimated parameter and standard errors between models which solved the correlated data and GEE. There are

many kinds of correlated data in social sciences; we used longitudinal data as an example.

**Key words**: Generalized Estimating Equations (GEE), Quasi Likelihood, Longitudinal Data, Correlation Effect

# ÖZET

## Genelleştirilmiş Tahmin Denklemlerinin (GEE) Gerçek Boylamsal Veri İçin Uygulaması

Fatima Yassin

İstatistik Anabilim Dalı

Yüksek Lisans Tezi

Tez Danışmanı: Doç.Dr. Filiz KARAMAN

Genellikle istatistikçilerin hedefi optimum istatistiksel modeller elde etmek ve bu modellerin parametrelerini tahmin ederek çıkarsamalar yapmaktır . Uygun bir istatistiksel yöntem seçimi aslında verinin nasıl toplanması gerektiği ve ne ölçtüğü konusunda dikkatli düşünmeyi gerektirir. Ama bazen ham veri, iyi tahminler elde etmemizi sağlayan klasik istatistik varsayımlarını sağlamayabilir.Tahmin yöntemlerinde, önemli varsayımlardan biri verinin korelasyonsuz olduğudur.

Çalışmamızda, bu varsayımı ihlal eden, korelasyonlu veriler incelenecektir. Bu problemi çözmek için birçok çözüm yöntemi geliştirilmiştir. Bunların arasından, her zaman yeterli çözümü veren yöntemleri kullanmak hedefimizdir. Bu yöntemlerden, Genelleştirilmiş Tahmin Denklemleri (GEE), verilerde korelasyon olduğu durumda yeterli çözüm bulmak için popüler bir yöntem haline gelmiştir. GEE metodolojisinin önemli bir özelliği, korelasyon etkisi çeşitlerinden "working kovaryans" etkisini dikkate almasıdır. GEE, geleneksel olarak Genelleştirilmiş Lineer Modeller (GLM) 'in bir uzantısı olarak sunulduğundan dolayı, bu çalışmada, başlangıç olarak GLM algoritmaları ve tahmin elde etme tekniklerinden bahsedilecektir. Bunlara ek olarak GLM'in önemli bir faydasıda tahminleri olabilirlik olasılığına dayalı veri grupları için istatistikler ve özellikler türetme kabiliyetidir ki; bu sebeple istatistikçiler GLM kapsamında tahmin yöntemleri oluşturur.

Bağımsızlık olmadığı durumlarda GLM uygulamada bağımsızlık gerektirdiğinden dolayı uygulanamaz. Çalışmamızda, GEE ve korelasyonlu veriyi çözen diğer yöntemlerden bulunan modeller için parametre tahminleri, standart hataları ve özet

istatistiklerinin karşılaştırılması yapılacaktır. Sosyal bilimlerde korelasyon içeren birçok veri çeşidi vardır. Bu çalışmamızda örnek olarak boylamsal veri kullanılacaktır.

**Anahtar Kelimeler**: Genelleştirilmiş Tahmin Denklemleri (GEE), Boylamsal Veri, Quasi Olabılirlik, Korelasyon Etkisi

# CHAPTER 1

# INTRODUCTION

## 1.1 Literature Review

Statistical modelling is a part of the foundation of statistical inference to deduct properties of underlying distribution using data drawn from the population via some form of sampling. The model assumptions describe a set of probability distribution. So we can describe the modelling process as:

1. Determining models in two parts: equations connecting the outcome and predictor variable, the probability distribution of the outcome variable.
2. Estimating parameters used in the model.
3. Exam how well the models fit the actual data.
4. Applying inferences.

Models according to their assumptions are: fully-parametric, Non-parametric, and semi-parametric. These types determine how we can estimate the parameters and which estimating equations we can use. To select an appropriate statistical method actually requires careful thought how data were collected and what they are measured. Data are not "just numbers" but are more complicated and full of problems. In the presence of problems and insufficient resources of data we must find ways to overcome these problems.

## 1.2 Objective of the Thesis

One of the common problems that we find it in the normal life is correlated data that take many forms. The correlation between data impedes the process of analysing data, that all of traditional methods of analysis require independency. Many methods were developed to solve this problem, many of adjustments were introduced.

The correlation between data takes many of forms; one of these forms repeated measure or longitudinal data where the correlation within between subjects that makes simple adjustments doesn't solve the problem of correlated data. Most traditional methods are trying to dispose of the effect of error terms from the model or make adjustment on coefficients of the model. None of them describes the type of correlation.
We discuss Generalized Estimating Equations (GEEs) as consistent estimation for longitudinal data.

### 1.3 Hypothesis

In this thesis we will discuss longitudinal data that represent one kind of correlated data. We make a definition of longitudinal data, the way of modelling this kind of data in chapter 2. In chapter 3 we make overview on the estimation, definition and the methodology. Generalized Estimating Equation which is used to solve repeated data problem is the topic of chapter 4. Chapter 5 is about the application on GEE. We make comparison between GEE and other methods for analysis of longitudinal data.

# CHAPTER 2

# LONGITUDINAL DATA

## 2.1 Definition

A longitudinal data is one that follows a given sample of individuals over time, and thus provides multiple observations on each individual in the sample [38]. If our panels represent a level of data organization where the observations within panels come from different experimental units belonging to the same classification, the data are called clustered data [4]. If the observations within panels are from the same experimental unit over time, we called that longitudinal data. Longitudinal studies had a role in epidemiology, clinical research, and therapeutic rating. As an example of longitudinal studies are used to describe normal growth and advancing age, to estimate the effect of risk factors on human health, and assess to the effectiveness of treatments.

To see the nature of longitudinal data, define $y_{it}$ to be the outcome for the *i*th subject during the *t*th time period. A longitudinal data set contains observations of the *i*th subject over $t = 1\ldots T_i$ time periods, for each of $i = 1. . . n$ subjects. So, the form of data will be :

$$\text{first subject} - \{y_{11}, y_{12}, \ldots, y_{1T1}\}$$
$$\text{second subject} - \{y_{21}, y_{22}, \ldots, y_{2T2}\}$$
$$\vdots$$
$$\text{nth subject} - \{y_{n1}, y_{n2}, \ldots, y_{nTn}\} \quad [18]$$

Study design depends on the nature of the investigator's question. That means, knowing what type of information that study should collect is a first step in determining how the study will be implemented (methodology).That is a key of distinguish between longitudinal data and other kinds of datasets.

The main distinguish of longitudinal data and cross-sectional data that longitudinal data are repeated over time and their analysis concerned with estimating how individuals change during the length of the study and examining the factors that influence heterogeneity among individuals in how they change over time. Cross-sectional data is the same as longitudinal data except for the time element, that it can compare many population groups at one point in time.

At first look longitudinal data seem like time series, that both types of data include measurement of time; but they are not same thing. Time series have more repeated observations than subjects while longitudinal data have more subjects than repeated observations [15]. Time series that a single data come from combination of many data points throughout time from single subject, like from a person, country, firm,...etc. While, longitudinal data come from gathering a few observations throughout time from many subjects, such as a few blood pressure measurements from many people at level of individual.

There are many features that distinguish longitudinal data:

- Can be estimated within several estimators.
- The estimators differ based on whether they consider the between or within variation in the data.
- Their properties (consistency) differ based on which model appropriate.
- They provide information on the time-ordering of events.
- They allow controlling for subject unobserved heterogeneity.
- They provide studying individual dynamics (e.g. cohort effects).

### 2.2 Challenges of Longitudinal Studies

1. Participant follow-up. There is the risk of bias due to incomplete follow up, or "drop-out" of study participants. If subjects that are followed to the planned end of study differ from subjects who discontinue follow-up then a naive analysis may provide summaries that are not representative of the original target population.[12]
2. Analysis of correlated data. Statistical analysis of longitudinal data requires methods that can properly account for the intra-subject correlation of response measurements. If such correlation is ignored then inferences such as statistical tests or confidence intervals can be grossly invalid.[12]
3. Time-varying covariates. Although longitudinal designs offer the ability to associate changes in exposure with changes in the response of interest, the direction of

causality can be complicated by "feedback" between the response and the predictor. For example, in an observational study of the effects of a drug on specific indicators of health, a patient's current health status may influence the drug exposure or dosage received in the future. Although the investigator interest with the effect of medication on health, this example has reciprocal influence between exposure and outcome and poses analytical difficulty when trying to separate the effect of medication on health from the effect of health on drug exposure. [12]

Models play an important role in statistical inference. A model is a mathematical way of describing the relationships between an outcome variable and a set of predictor variables. Some models can be seen as a theory about how the data were generated. Other models are only intended to provide a convenient summary of the data.

We have many kind of modelling forms for longitudinal data enable us to put it in statistical forms which represent the foundation of statistical inferences.

### 2.3 Pooled Estimators

A simple method to modelling longitudinal data is forgetting the panel correlation that might be existing in the data. The result of this method is called a pooled estimator; since the data are simply pooled without observe to which panel the data naturally belong.

The resulting estimated coefficient vector, though consistent, is not efficient. A direct result of ignoring the within-panel correlation is that the estimated (naive) standard errors are not a reliable measure for testing purposes [4]. To address the standard errors, we should employ a modified sandwich estimate of variance, or another variance estimate that adjusts for the panel nature of the data [4].

There is no difference between pooled estimator and likelihood estimator.

Pooled estimating equation:-

$$\left[\left\{\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\sum_{t=1}^{ni}\frac{y_{it}-\mu_{it}}{a(\emptyset)V(\mu_{it})}\left(\frac{\partial \mu}{\partial \eta}\right)_{it} x_{it}\right\}_{j=1,\dots,P}\right]_{P*1} = [0]_{P*1} \qquad (2.1)$$

Where p is the column dimension of the matrix of covariates X[3]. If we consider that there is within-panel correlation, our using likelihood is not right. Instead, we can use the modified sandwich estimate of variance for testing and interpretation; but we should acknowledge the fact that employing a pooled estimator with a modified variance estimate is a declaration that the underlying likelihood is not correct [3].

## 2.4 Fixed-effect Models

To treat the panel structure in the data, we may involve an effect for each panel in our estimating equation. We can divide these effects to fixed effects or random effects. The fixed effects models mean that the researchers analysing data conditionally on the effects. Fixed-effect models divided to conditional fixed effects and unconditional fixed effects.

### 2.4.1 Conditional Fixed-effects Model

A conditional fixed-effects model is formed by conditioning out the fixed effects from the estimation [3]. These models come from specific distributions and not available for all kind of distributions that belongs to exponential family distribution.

Generally, some special distributions for a one outcome on $y_{it}$ where we called $f_1(y_{it})$we find the joint distribution for all of the observations for a specific panel $f_1(y_i) = \prod_{t=1}^{ni} f_1(y_{it})$ and get the sufficient statistic $\xi(y_i)$ for the fixed effect $v_i$. The next step we find the distribution of the sufficient statistic$f_2(\xi(y_i))$. Eventually, we obtain the conditional distribution of the outcomes according to distribution of the sufficient statistic as$f_3(y_i;\ \beta/\xi(y_i)\ )$.We note that the Obtained distribution is free of the fixed effect $v_i$. So, the conditional log-likelihood for all of the panels is given by

$$l = \ln \prod_{i=1}^{n} f_3(y_i;\ \beta/\xi(y_i)\ ) \qquad (2.2)$$

with estimating equation for $\Theta = (\beta)$ and $j = 1, \ldots, p$

$$\Phi(\Theta) = \left[\left\{\frac{\partial l}{\partial \beta_j}\right\}\right]_{p*1} [0]_{p*1} \qquad (2.3)$$

We know that the estimating equation can be made full information maximum likelihood or limited information maximum likelihood depending on whether they have additional parameters from $f_1$therefore we can identify $\Theta$ as ancillary or not.

### 2.4.2 Unconditional Fixed-effects Models

If we have a limited number of panels in population and every one are represented in our sample we would use an unconditional fixed-effects model. In the other side if we have unlimited number of panels (uncountable) we would use a conditional fixed-effects model, that using of unconditional models would give biased estimators.

The estimating equation for unconditional fixed-effects models that belong to exponential family is given by accepting the fixed effect $v_i$ in the linear equation

$\eta_{it} = x_{it}\beta + v_i$ .We want to get estimation for (p+n)*1 parameter vector $\theta = (\beta, v)$ the estimating equation for unconditional fixed-effects GLM given by

$$\begin{bmatrix} \left\{\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\sum_{t=1}^{ni} \frac{y_{it}-\mu_{it}}{a(\emptyset)V(\mu_{it})}\left(\frac{\partial \mu}{\partial \eta}\right)_{it} x_{jit}\right\} \\ \left\{\frac{\partial l}{\partial v_k} = \sum_{t=1}^{nk} \frac{y_{kt}-\mu_{kt}}{a(\emptyset)V(\mu_{kt})}\left(\frac{\partial \mu}{\partial \eta}\right)_{kt}\right\} \end{bmatrix}_{(p+n)*1} = [0]_{(p+n)*1} \qquad (2.4)$$

## 2.5 Random-effects Models

A random effect models that the researchers analysing data unconditionally on the effects. Random effect models parameterize the random effects according to an assumed distribution for which the parameters of the distribution are estimated [3]. Also are called Subject-Specific models where the likelihood models are for the individual observations instead of marginal distribution of the panels.

The log-likelihood for a random –effect model is

$$l = \ln \prod_{i=1}^{n} \int_{-\infty}^{\infty} f(v_i)\left\{\prod_{t=1}^{ni} f_y(x_{it}\beta + v_i)\right\} dv_i \qquad (2.5)$$

Where $f_y$ is the assumed density function for the overall model (response) and f is the density function for the iid random effects $v_i$.The estimating equation is the derivative of the log-likelihood in terms of β and parameters of the model that assumed as random-effect models.

We can chose between fixed-effects or random-effects models when we have a well known about the panel and what the investigators want from data. The inference follows the nature of the models. Estimating equation given by

$$\begin{bmatrix} \left\{\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\sum_{t=1}^{ni} \frac{y_{it}-\mu_{it}}{a(\emptyset)V(\mu_{it})}\left(\frac{\partial \mu}{\partial \eta}\right)_{it} x_{jit}\right\} \\ \left\{\frac{\partial l}{\partial v_k} = \sum_{t=1}^{nk} \frac{y_{kt}-\mu_{kt}}{a(\emptyset)V(\mu_{kt})}\left(\frac{\partial \mu}{\partial \eta}\right)_{kt}\right\} \end{bmatrix}_{(p+n)*1} = [0]_{(p+n)*1} \qquad (2.6)$$

Which model we will choice to present our data fixed-effects models or random-effects models? The choice depends on the nature of panels. The inference will be according to the nature of the model. When there is no compulsory choice between the two models, sometimes random-effects model is preferred if there are covariates that are constant within panels. Parameters for these covariates cannot be estimated by applying fixed-effects models since the covariate is collinear with the fixed effect.

## 2.6 Population-averaged and Subject-Specific Models

There are two categories of models that treating the panel structure of data, as general: Population-averaged model (marginal model) and Subject-Specific model.

Population-averaged models that the outcome can be modelled as a function of covariates without clearly accounting for subject to subject heterogeneity. Population-averaged models concern with the within-panel dependence by taking average effects across all panels.

A population-averaged model is derived through introducing a parameterization for a panel-level covariance.

The second type Subject-specific models obtained through introduction of individual effect. While this implies an individual -level covariance, each panel effect is estimated using information only from the specific panel [25].

## 2.7 Cohort Study

Longitudinal study had many forms one of it cohort study (a type of observational study) used in medical areas, social science, actuarial science, business analytics, and ecology. As an example in medicine, it is means analysis of risk factors and follow up group of subjects (people) who do not infected, and use dependency to determine the absolute risk of subject contraction. Cohort study is one type of clinical study design and must be compared with a cross-sectional. They are interest about the life histories of sectors of populations, and the individual people who represent these sectors.

A cohort is a group of people who share a common characteristic or experience within a defined period (e.g., are born, are exposed to a drug or vaccine or pollutant, or undergo a certain medical procedure) [11]. Example for cohort study, group of people who born in the same day. Thus the group under investigation would be the whole population that the cohort taken , or would be other cohort of individuals considered had no infection to the same group under examination, otherwise the same. Instead, subgroups inside the cohort may be compared with each other. As an example for cohort study design figure (2.1) :

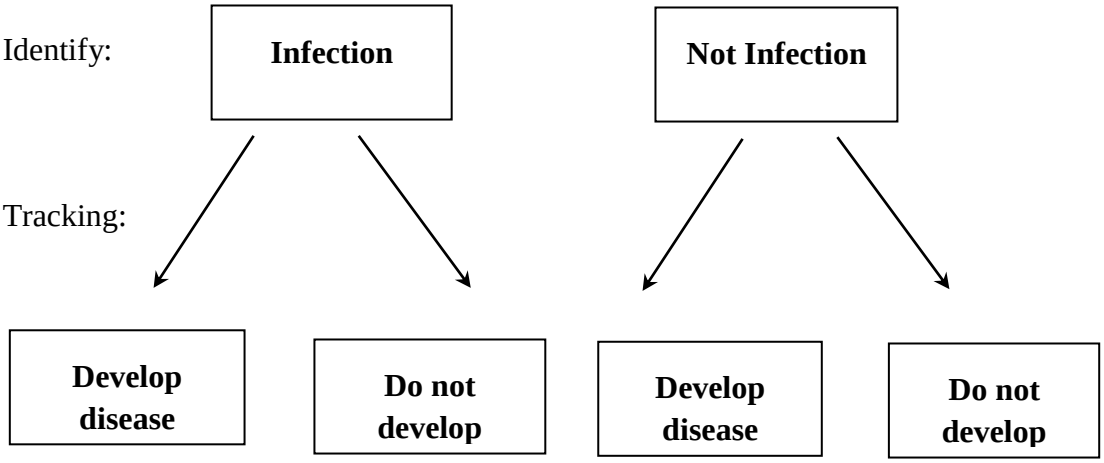


Figure 2.1 : Design of cohort study

The groups who do not have infection where followed up until specific time to observe who develop the infection (new patients). We cannot consider group of people who have the diseases cohort study. Prospective (longitudinal) cohort studies between initial infection and develop the disease effectively help addressing causal associations even though getting true description for causal need more verification from any other kind of experimental trials.

The advantage of prospective cohort study data is that it can help determine risk factors for contracting a new disease because it is a longitudinal observation of the individual through time, and the collection of data at regular intervals, so recall error is reduced [35]. But, cohort studies are expensive to apply, and take a long time to get useful data. Although more effectively than results taken from cross-sectional studies. In observational epidemiology prospective cohort studies had most reliable results.

## 2.8 Joint Modelling of Longitudinal and Survival Data

Joint modelling means models that mixed survival analysis and longitudinal studies. Significantly had increased use in medical fields that sometimes we need both technique, we are consider with the time until death or recovery occur for some subjects at same time we need to take repeated measures from these subjects. Oftenlongitudinal data and survival data are related in some ways.So if we apply one of longitudinal data or survival data separately that may be we get insufficient or biased estimation .So Joint models (longitudinal and survival data) integrate all available information and give unbiased and effective inferences.

### 2.8.1 Survival Data

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest [40]. We mean by the event death, birth, recovery from particular disease ...etc. And the time to occur this event may be hours, days, years ... etc.

Classical survival analysis, which developed in the second half of the 20th century, is generally concerned with estimating effects of fully observed variables on the time to event distribution. Such variables are often called covariates. There are several well established methods to model survival with covariates. Accelerated failure time (AFT) models, for example, are built in generalized linear models environment to obtain the survival time estimation. The effect of a covariate is thus to multiply the expected event time by some constant.

Modern survival analysis goes further in exploiting subject specific information to predict survival. Several studies have shown that effects of aging may be seen in changes of certain characteristics over time. For example longitudinal study, where the variable of interest is measured several times for each subject.

Recent years have seen a substantial advancement in joint modelling of longitudinal and survival data. A lack of modelling techniques in the field leaves most of the data unexplored. Thus, the approach of random senescence and random covariates seems very attractive and worth developing.

It is generally not feasible to collect time dependent data at more than a few time points per subject. In such situations the random effects models are often used in place of the more technical and cumbersome continuous diffusions. To describe this important class of joint models, we give a brief overview of the random effects approach.

Suppose we obtain measurements $Y_{ij}$on subject $i$ at time $j$. A linear random effects model is given by

$$Y_{ij} = X_i\beta + Z_i\beta_i + \epsilon_{ij} \quad (2.7)$$

$\beta$ are fixed effects and $\beta i$are random effects. The latter are usually assumed to have normal distribution with zero mean. The error term $\epsilon_{ij}$ may be allowed to have serial correlation (that is, along the time index j). Because both fixed and random effects are present, such models are also called “mixed models”. The role of $\beta_i$ is very similar to that of frailty to model heterogeneity across the subjects. This class of models is a very powerful tool and is widely used in applications.

In joint analysis it is common to model the longitudinal component with a mixed model and the survival component with either PH (Proportional Hazard) or AFT. The two parts have to be linked together to ensure that it is possible for the longitudinal data to carry information about survival.

The approach shown that longitudinal data may possess richer information on survival than a cross-sectional measurement called 'joint models'. The joint models had several features:

1. The data are continuous measurements.
2. If measurements are taken longitudinally, there are only a few to a moderate number of time points.
3. The hidden variable or process involved in the joint model is generally not of interest to practicing statisticians.

# CHAPTER 3

# ESTIMATION

Estimation is defined as forming a summary of the collected incomplete information (the data) with the purpose of getting as close as possible to the complete information quantity of interest (the target).

After finding appropriate modelling form next step is to find estimating equations, to find parameter estimation. There are many types of estimating equation for longitudinal data differ depend on its modelling assumption. The most commonly used methods to get estimators if the data are independent (there are other restriction on response distribution and error terms) Generalized linear models (GLMs) that use maximum likelihood estimator (MLE).

As a general linear model (GLM), the response variable y for number i (i = 1, 2, ..., n) observation is modelled as a linear combination of (p − 1) called it predictor variables $x_1, x_2, \ldots, x_{p-1}$ as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i(p-1)} + e_i \qquad (3.1)$$

GLMs enable descriptive and predictive models to be built that are effectively general to be applicable for much social science data. They can be used to model data collected from survey and experimental studies and can replace many of the more traditional hypothesis tests that are still in common use. Of specific importance is the unified theoretical framework that the method offers, as this enables particular 'economies of scale' to be realised that allow a whole range of data to be analysed using similar techniques. The use of the techniques will be described using a modelling procedure whereby a particular variable can be modelled (or predicted) using information about other variables.

Peter McCullagh of the University of Chicago and John Nelder of the Imperial College of Science and Technology, London, authored the seminal work on Generalized Linear Models in 1983, in a text with the same name. Major revisions were made in McCullagh and Nelder (1989), which is still the most current edition [3].

GLMs are depend on likelihood models restrection, one of the important assumption is the individual must be uncorrelated. That we called it iid requirement (independent and identically distributed). Actually this assumption is lacking in many common situations.

Also generalized linear models are built in exponential family of distributions (all available distributions are belonging to exponential family of distributions). Members of this family are Gaussian or normal, binomial, gamma, inverse Gaussian, Poisson, geometric, and the negative binomial for a specified ancillary parameter. As we will see first GEE models extension of GLM depended on classical distribution (Gaussian, binomial, gamma, and Poisson),this does not mean it stop at these distributions, also their application clearly extends to other members.

All members of classical form of generalized linear models are referring to one the previous mentioned probability functions. The object of likelihood function in a simple form is re-parameterization of the probability function. Where the probability function estimates a probability according to given position and scale parameters.The idea is that the likelihood estimates parameters that make the observed data most probable or likely[3]. To simplify it we use log transform for likelihood, which is more applicable in computer estimation. Then we use the maximization for likelihood estimation to provide it feature of generalization for whole distribution not on the conditional expectation.

Members of the exponential family of distributions have general form for their distribution functions that their likelihood equation expressed as

$$\exp\left\{\frac{y\theta - b(\theta)}{a(\emptyset)} - c(y, \emptyset)\right\} \qquad (3.2)$$

The main power of GLMs in applicability to provide techniques and statistics for specific group according to likelihood estimation forms. The one of the properties for exponential family distribution is the expected value of their members related to the outcome variable in interest. We have what we called the natural connection between distribution function and response variable to provide covariates into the model instead of expected value that is the $\theta$ parameter.When a particular distribution is written in exponential family form, the $\theta$ parameter is represented by some monotonic

differentiable function of the expected value $\mu$ [17].This function links the response variable y to the expected value $\mu$. Canonical link function come from writing a distribution in exponential form. As general, the models can provide covariates through any differentiable link function, so we can faced numeric difficulties in the case the function fails to enforce range assumptions that define the particular distribution of the exponential family. We can conclude that the link function put the systematic component as random component. Link function have many forms can be "one of identity for normally distributed random components, or one of a number of non-linear links when the random component is not normally distributed" [21].

There is a general link function for any member distribution of the exponential family, that is canonical link, that relates the linear predictor $\eta = X\beta$ to the expected value $\mu$. These canonical links occur when $\theta = \eta$ of the variance in terms of mean. As a consequence, the class of GLMs allows not only a identification of the link function relating the outcome to the covariates, but also a identification of the form of the variance in terms of the mean. These two options are not limited to specific distributions in the exponential family [21].

Table 3.1 Standard link and inverse link functions

| Link Name | Link function $\eta = g(\mu)$ | inverse link $\mu = g^{-1}(\eta)$ |
|---|---|---|
| Complementary log-log | $\ln\{-\ln(1-\mu)\}$ | $1 - exp\{-\exp(\eta)\}$ |
| Identity | $\mu$ | $\eta$ |
| Inverse square | $\frac{1}{\mu^2}$ | $\frac{1}{\sqrt{\eta}}$ |
| Log | $\ln(\mu)$ | $\exp(\eta)$ |
| Log-Log | $-\ln\{-\ln(\mu)\}$ | $exp\{-\exp(-\eta)\}$ |
| Logit | $\ln\{\mu/(1-\mu)\}$ | $e^n/(1+e^n)$ |
| Probit | $\Phi^{-1}(\mu)$ | $\Phi(\eta)$ |
| Reciprocal | $\frac{1}{\mu}$ | $\frac{1}{\eta}$ |

Table 3.2 Variance functions for distributions in the exponential family

| Distribution | Variance V($\mu$) |
|---|---|
| Bernoulli | $\mu(1-\mu)$ |
| Binomial | $\mu(1-\mu/k)$ |
| Gamma | $\mu^2$ |
| Gaussian | 1 |
| Inverse Gaussian | $\mu^3$ |
| Negative binomial | $\mu + k\mu^2$ |
| Poisson | $\mu$ |

The estimates of the GLM regression parameter vector $\psi$ are the solution of the estimating equation given by

$$\frac{\partial l}{\partial \beta} = 0 \tag{3.3}$$

Where $l$ is the log-likelihood for the exponential family. A Taylor series expansion about the solution β* is given by

$$\frac{\partial l}{\partial \beta}(\beta^*) - (\beta - \beta^*)\frac{\partial^2}{\partial \beta \partial \beta^T} + \ \ldots\ldots\ldots = 0 \tag{3.4}$$

such that a recursive relationship for finding the solution is

$$\beta^{(r)} = \beta^{(r-1)} + \left[-\frac{\partial^2 l}{\partial \beta \partial \beta^T} + \left(\beta^{(r-1)}\right)\right]^{-1} \frac{\partial l}{\partial \beta}\left(\beta^{(r-1)}\right) \tag{3.5}$$

The method of Fisher scoring substitutes the expected value of the Hessian matrix, resulting in

$$\beta^{(r)} = \beta^{(r-1)} + \left[E\left(\frac{\partial l \partial l}{\partial \beta \partial \beta^T}\right) + \left(\beta^{(r-1)}\right)\right]^{-1} \frac{\partial l}{\partial \beta}\left(\beta^{(r-1)}\right) \tag{3.6}$$

Filling in the details, the updating steps are defined by

$$\left\{\sum_{i=1}^{n} \frac{1}{v(\mu_i)a(\emptyset)}\left(\frac{\partial \mu}{\partial \eta}\right)^2 x_{ji}\, x_{ki}\right\}_{p*p} \beta_{p*1}^{(r)} = \left\{\sum_{i=1}^{n} \frac{1}{v(\mu_i)a(\emptyset)}\left(\frac{\partial \mu}{\partial \eta}\right)^2_i \left\{(y_i - \mu_i)\left(\frac{\partial \eta}{\partial \mu}\right)_i + \eta_i\right\} x_{ji}\right\}_{p\text{x}1} \tag{3.7}$$

for j 1, . . . , p. Here, we note that

$$w = Diag\left\{\sum_{i=1}^{n} \frac{1}{v(\mu)a(\emptyset)}\left(\frac{\partial \mu}{\partial \eta}\right)^2\right\}_{n\text{x}n} \tag{3.8}$$

$$Z = \left\{(y - \mu)\left(\frac{\partial \eta}{\partial \mu}\right)_i + \eta\right\}_{n\text{x}1} \quad (3.9)$$

so that we may rewrite the updating step in matrix notation (with $X_{nxp}$) as the weighted OLS equation

$$(X^T W X)\beta^{(r)} = X^T W Z \quad (3.10)$$

As we mentioned that GLMs is likelihood-based models so must fallow these constrictions involve several standard steps:

1. Select a distribution for the response variable.
2. Drafting of the joint distribution of the data.
3. Get the likelihood from joint distribution.
4. Make generalization for likelihood through a linear combination of covariates and related coefficients.
5. Parameterize the linear combination of covariates to compel range assumptions on the mean and variance implied by the distribution.
6. Find the estimating equation for the solution of unknown parameters.

In case the model was selected, we have two technique to estimate model parameters . First technique estimate all parameter in the model even including extra parameters called full information maximum likelihood (FIML).Second technique consider some parameter as ancillary to the analysis, so we estimate some parameters called limited information maximum likelihood (LIML).Estimation may then be carried out using an optimization method. The most common technique is that of Newton-Raphson, or a modification of the Newton-Raphson estimating algorithm. We either derive an estimating equation in terms of $\Theta = (\beta, \sigma^2)$ (a FIML model), or we specify an estimating equation in terms of $\Theta = (\beta)$ where $\sigma^2$ is ancillary (a LIML model). The ancillary parameters in a LIML model are either estimated separately, or specified. The resulting estimates for the parameters are conditional on the ancillary parameters being correct [3].

### 3.1 Full information Maximum Likelihood (FIML)

Is a technique to get full parameter estimation in the model under analysis. The estimating equation for FIML after applying link function to linear combination for our covariates is $\psi(\Theta) = 0$ for $\left(\beta_{p*1}, \sigma^2\right)$is given by

$$\begin{bmatrix} \left\{ \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} x_{ji} \frac{1}{\sigma^2} (y_i - x_i\beta) \right\}_{j=1,\ldots..,P} \\ \frac{\partial l}{\partial \sigma^2} = -\sum_{i=1}^{n} \frac{1}{2\sigma^2} + \frac{(y_i - x_i\beta)^2}{2\sigma^4} \end{bmatrix}_{(P+1)*1} = [0]_{(P+1)*1} \quad (3.11)$$

We can write in term of μ rather than xβ . To include the parameterization, we use the chain rule

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \beta} \quad (3.12)$$

In this more general notation, the estimating equation $\psi(\Theta) = 0$ is given by

$$\begin{bmatrix} \left\{ \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n} \frac{1}{\sigma^2} (y_i - \mu_i) \frac{\partial \mu}{\partial \eta} x_{ji} \right\}_{j=1,\ldots..,P} \\ \frac{\partial l}{\partial \sigma^2} = -\sum_{i=1}^{n} \frac{1}{2\sigma^2} + \frac{(y_i - \mu_i)^2}{2\sigma^4} \end{bmatrix}_{(P+1)*1} = [0]_{(P+1)*1} \quad (3.13)$$

and we must specify the relationship (parameterization) of the expected value it to the linear predictor$\eta = X\beta$. In the case of linear regression $\eta = \mu$.

### 3.2 Limited Information Maximum Likelihood (LIML)

Limited information maximum likelihood differs from FLIM that does not treat all parameter in the model there is some parameters consider as ancillary.

The estimating equation for the LIML model $\psi(\theta) = (\beta) = 0$, treating $\sigma^2$ as ancillary, it take just first row from (3.13).

If we back to general form for exponential family of distributions what contain a location parameter $\theta$, a scale parameter a(∅), and a normalizing term $c\,(y, \emptyset)$with probability density

$$f\,(y\,;\,\theta, \emptyset) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\emptyset)} + c\,(y\,, \emptyset) \right\} \quad (3.14)$$

Where

$$E(y) = b(\theta) = \mu \quad (3.15)$$

$$v(y) = b''(\theta) a\,(\emptyset) \quad (3.16)$$

The normalizing term doesnot contain $\theta$.

The log-likelihood for the exponential family is

$$l(\theta, \emptyset \backslash y_1, \ldots.., y_n) = \sum_{i=1}^{n} \left\{ \frac{y_i\,\theta_i - b(\theta_i)}{a(\emptyset)} + c(y_i, \emptyset) \right\} \quad (3.17)$$

LIML is common approach to get parameter estimation deal with all models as general not for specific type of distribution as in FIML. Where our focus only on , dispersion parameter a(∅) consider as ancillary parameter.

Using the chain rule to obtain a more useful form of the LIML estimating equation $\psi(\Theta) = 0$ for $\Theta = \beta_{(P*1)}$

$$\frac{\partial l}{\partial \beta} = \left\{\left(\frac{\partial l}{\partial \theta}\right)\left(\frac{\partial \theta}{\partial \mu}\right)\left(\frac{\partial \mu}{\partial \eta}\right)\left(\frac{\partial \eta}{\partial \beta_j}\right)\right\} = \left[\sum_{i=1}^{n}\left(\frac{y_i - b'(\theta_i)}{a(\emptyset)}\right)\left(\frac{1}{V(\mu_i)}\right)\left(\frac{\partial \mu}{\partial \eta}\right)_i (x_{ij})\right]_{P*1}$$

$$= \left[\sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{V(\mu_i)a(\emptyset)}\left(\frac{\partial \mu}{\partial \eta}\right)_i (x_{ij})\right]_{P*1} \quad (3.18)$$

so that the general LIML estimating equation for the exponential family is

$$\frac{\partial l}{\partial \beta} = \left[\sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{V(\mu_i)a(\emptyset)}\left(\frac{\partial \mu}{\partial \eta}\right)_i (x_{ij})\right]_{P*1} = [0]_{P*1} \quad (3.19)$$

Where j = 1,.....,P

As we see GLMs require independence cannot deal with correlated data. In the late 1970s, John Nelder designed the first commercial software developed exclusively for GLMs. Called GLIM, for Generalized Linear Interactive Modelling, the software was manufactured and distributed by Numerical Algorithms Group in Great Britain [3].

Later, Nelder and GLIM team members introduced capabilities into GLIM that allowed adjustment the variance-covariance or Hessian matrix so that the effects of extra correlation in the data would be taken into account with respect to standard errors. This was accomplished through estimation of the dispersion statistic [3].

In GLM approach we have two kinds of dispersion statistics, differ according to the way of estimation this statistics. The first obtain by using deviance statistic, the other use the Pearson $\chi^2$ statistic. The two kinds are summary measure for the model which appear in the output of fitting process.

The value of dispersion statistic determines the kind of correlation effect for example underdispersed or overdispersed. The traditional methods interest with adjustment on the standard errors according to calculated correlation effects, this approach called scaling the standard errors. This approach is applying by divide all parameters error by the square root of dispersion parameter. Scaling method does not affect on the estimated value of regression parameter. The scaling approach is a simple way to make models more applicable and comprehensive for the data that are marginally correlated. This approach still in use especially when we want to take a first look on our data.

As we note the Scaling standard errors give general conception about the structure of correlation in the whole model, it does not has ability to definite the intra correlation inside the clusters. That we get its value after fitting model ($posthoc$ method), has no

effect on the estimated parameters. That is not enough to treat problem of correlated data because sometimes the dispersion statistic is not real and not explain the correlation structure between subjects.

In the mid-1980s, researchers at Johns Hopkins Hospital in Baltimore developed methods to deal with longitudinal and cluster data using the GLM format. In so doing, they created a 2-step algorithm that first estimates a straightforward GLM, and then calculates a matrix of scaling values [3]. The scaling matrix adjusts the Hessian matrix at the next algorithm iteration. For each iteration in algorithm updates the parameter estimates the same thing for adjusted Hessian matrix and a matrix of scales. Liang and Zeger (1986) provided further exposition of how the matrix of scales could be parameterized to allow user control over the structure of the dependence in data [15].

Even though this weak description of their method, but may be give a conception about the logic of initial versions of the extended GLMs which it represent the base of generalized estimating equations.

The reason that complicates longitudinal data analysis is heterogeneous variability, because:

1. Repeated measures on the same individual are usually positively correlated.
2. Variability is often heterogeneous across measurement occasions

For that we need another technique to solve these problems which was GEE.

# CHAPTER 4

## Generalized Estimating Equation (GEE)

The GEE approach has its roots in the quasi-likelihood methods introduced by Wedderburn (1974) and Nelder and Wedderburn (1972) and developed and extended by McCullagh and Nelder (1983, 1989) While standard maximum-likelihood analysis specification of the full conditional distribution of the response variable, quasi-likelihood requires only that we postulate the relationship between the expected value of the response variable and the covariates and between the conditional mean and variance of the response variable. The quasilikelihood is a general form for the likelihood.

The quasi-likelihood (QL) technique is used for estimating regression coefficients ignoring fully specifying the distribution of the observations. It thus provides a more flexible approach to estimation than ML (maximum likelihood) estimation.

### 4.1 When to Use QL

There are two common scenarios where QL is used:

1. To allow for over- or under-dispersion relative to the Poisson or binomial distribution usually while maintaining the same linear predictor and link function chosen for the preliminary Poisson/binomial model.
2. To estimate regression coefficients when it's not clear how to specify the distribution of the observed data.

GEE method is general and robust approach to obtain point estimators, which considered as special case for methods, such as the method of moments, the last squares, the maximum likelihood, M-estimator, quasi-likelihood ,…etc. That the term generalized estimating equations indicates that an estimating equation is not the result

of a likelihood-based derivation, but that it is obtained by generalizing another estimating equation [3]. GEE's were first introduced by Liang and Zeger (1986).To understand how we can use GEE we begin with the LIMQL estimating equation for GLMs

$$\psi(\beta) = \left[\left\{\sum_{i=1}^{n}\sum_{t=1}^{ni}\frac{y_{it}-\mu_{it}}{a(\emptyset)\,V(\mu_{it})}\left(\frac{\partial\mu}{\partial\eta}\right)_{it}x_{jit}\right\}_{j=1,\ldots\ldots,p}\right]_{p*1} = \qquad (4.1)$$

And rewrite it in matrix terms of the panels

$$\psi(\beta) = \left[\left\{\sum_{i=1}^{n}X_{ij}^{T}\;D\left(\frac{\partial\mu}{\partial\eta}\right)\;[V(\mu_i)]^{-1}\left(\frac{y_i-\mu_i}{a(\emptyset)}\right)\right\}_{j=1,\ldots,p}\right]_{p*1} = [0]_{p*1} \qquad (4.2)$$

Where D( ) denotes a diagonal matrix. $V(\mu_i)$ is clearly a diagonal matrix which can be decomposed into

$$V(\mu_i) = \left[D(V(\mu_{it}))^{1/2}I_{ni*ni}D(V(\mu_{it}))^{1/2}\right]_{ni*ni} \quad (4.3)$$

From equation (4.3) we understated that the estimating equation each panel as independent regarding the possibility of correlation existence. This kind of estimating equation is effective for independent models.

Liang and Zeger use a modification of the LIMQL estimating equation for GLMs to introduced their approach (GEE) through identify general matrix instead of identity matrix called correlation matrix, where the correlated data have full matrix form not a diagonal form

$$V(\mu_i) = \left[D(V(\mu_{it}))^{1/2}\,I_{ni*ni}\,D(V(\mu_{it}))^{1/2}\right]_{ni*ni} \qquad (4.3)$$

We write $R(\alpha)$to conformation that the correlation matrix is to be estimated through the parameter vector$\alpha$,that the most important characteristic for GEE over other methods for estimating parameter for correlated data.

## 4.2 Fitting Generalized Estimating Equations

To fit generalized estimating equations we must specify:

- The link function
- The distribution of the dependent variable
- The correlation structure of the dependent variable

### 4.2.1 Link Function

In the first step the user must specify a link transformation function that will allow the dependent variable to be expressed as a vector of parameter estimates (β) in the form of

an additive model (McCullagh & Nelder, 1989) [18]. Harrison (2002) noted that the link function is what "makes [generalized linear modelling] techniques part of a larger family of log-linear models; nonlinear and distinct from multiple linear regressions in the link function but linear and familiar in terms of the string of regression parameters". There is link function for any distribution privet for it linking response variable with covariates. The process of choosing link function for a specific model is depending on the range of the response variable. The canonical link is common used. Although we chosen canonical link or some other link functions in most time has no effect on the result of analysis, otherwise may has effect on calculation of the sandwich estimate of variance.

### 4.2.2 The Distribution of the Dependent Variable

After determining link function the second step is identifying the distribution of the response variable. GEE models take their distribution proprieties from the exponential family, the same classical distribution member of the exponential family. GEEs built inside GLMs environment, many of time carry the same feature of their function. Such as the variance provided as a function of the mean, that will use then in the calculation of the covariance matrix by multiplying the components against an N × N matrix (W) with a value $W_i$ on the diagonal that is determined by the variance function. For Poisson distributions, that figure is μ; for binary data, it is μ (1 – μ); and for normally distributed data, it is 1 (Gardner et al., 1995; McCullagh&Nelder, 1989). As Gardner and colleagues (1995) indicated that the link function or the variance functions have big consequences that specify distribution instead of other can lead to wrong statistical conclusions.

Although the specification of the distribution is important, users do not need to be careful in the determination of the variance functions for the coefficient estimates that they have sampling distribution is approximately normal distribution (Liang&Zeger, 1986). In applying a GEE (or any generalized linear model), the user should make every acceptable effort to correctly determine the distribution of the response variable so that the variance can be efficiently calculated as a function of the mean and the regression coefficients can be properly interpreted (McCullagh & Nelder, 1989).

As general, the user will have some initial information about the outcome variable. Such as, if the outcome is binary data, determined distribution will be binomial. If the outcome is counted the distribution should be Poisson distribution. We can know if we

fit a suitable distribution or not that by examines the dispersion parameter in the outcomes.

### 4.2.2 The Correlation Structure of the Dependent Variable

This step represent a most important feature for GEE models that is determination of the forms of correlation structure for the outcome variable between subjects within groups of the sample. What we called it working correlation matrix that enable us to estimate correlated models which the correlation is between subjects. In GEE models we have several forms represent the correlation matrix. The forms of the correlation matrix differ according to the nature collecting data. "The goal of selecting a working correlation structure is to estimate β more efficiently" (Pan, 2001, p. 122), where any wrong determination of the correlation structure may be affect the efficiency of the parameter estimates. Although GEE models in general robust to misspecification of the correlation structure (Liang & Zeger, 1986), the inefficient estimators will appear in case of the specified structure does not include all possible information about correlated measure between subjects. There are many forms for correlation structure that we will see in the coming sections.

## 4.3 Population-averaged (PA)- GEE Models for GLMs

The most common derived model from GEE group models described in the landmark paper of Liang and Zeger (1986). Where the authors are give the first introduction to generalized estimating equations. Also they describe the theoretical justification in addition the asymptotic properties for getting estimators. The most researchers who refer to GEE models are sign to these models.

As we said the GEEs equations derived from LIMQL where the equation

$$\psi(\beta) = \left[\left\{\sum_{i=1}^{n} X_{ij}^{T}\ D\ \left(\frac{\partial\mu}{\partial\eta}\right)\ [V(\mu_i)]^{-1}\ \left(\frac{y_i-\mu_i}{a(\emptyset)}\right)\right\}_{j=1,\dots,p}\right]_{p*1} = [0]_{p*1} \qquad (4.5)$$

And

$$V(\mu_i) = \left[D(V(\mu_{it}))^{1/2}\ R(\alpha)\ D(V(\mu_{it}))^{1/2}\right]_{ni*ni} \qquad (4.6)$$

Where $R(\alpha)$is the correlation structure.

In these models we are interested in the marginal distribution of the outcome, for which the expected value and variance functions are averaged over the panels. $R(\alpha)$ the within-panel correlation matrix.

### 4.3.1 The Working Correlation Matrix

An important feature distinction GEE from other derived methods to solve the correlation problem is the ability of GEE models to describe the correlation structure of between subjects correlation. There are several forms to specify the appropriate correlation structure for our data.

In case of we believe that the correlation between data into subjects is not following any order form we need only one additional scalar parameter that the correlation between subjects is equally. Otherwise we can obtain more complicated forms if we belief that the correlation follow specific order (not equally).

The estimating equation for the estimator of for the ancillary association parameters Take this formula

$$\psi(\alpha) = \sum_{i=1}^{n} \left(\frac{\partial \xi_i}{\partial \alpha_i}\right)^T H_i^{-1}(W_i - \xi_i) = [0]_{q*1} \quad (4.7)$$

Where

$$W_i = (r_{i1}\, r_{i2}, r_{i1}\, r_{i3}, \dots\dots, r_{ini-1}\, r_{ini})^T_{q*1} \quad (4.8)$$

$$H_i = D\left(V(W_{ij})\right)_{q*q} \quad (4.9)$$

$$\xi_i = E(W_i)_{q*1} \quad (4.10)$$

Such that $r_{ij}$ the *ij*th Pearson residual, and $\hat{r}_{ij}$ obtained from the current estimate $\hat{\beta}$ $H_i$ is a diagonal matrix, and $q = \binom{n_i}{2}$.

The following subsections provide standard approaches to determining a structure for the estimated within group correlation.

#### 4.3.1.1 Exchangeable Correlation

The common applied structure that used to describe the correlation between subjects. Where all subjects had the same correlation in panel data (equal correlation). In this case α is a scalar and the working correlation matrix has the structure

$$\begin{bmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{bmatrix} \quad (4.11)$$

Which we can write succinctly as

$$R_{uv} = \begin{cases} 1 & if\ u = v \\ \alpha & otherwise \end{cases} \quad (4.12)$$

This means in our repeated measurements in the data set any changing between subjects is valid. A GEE with an exchangeable correlation structure uses the estimated Pearson residuals $\hat{r}_{it=(y_{it}-\hat{\mu}_{it})/\sqrt{V(\hat{\mu}_{it})}}$ from the current fit of the model to estimate the common correlation parameter. The estimate of $\alpha$ using these residuals is

$$\hat{\alpha} = \frac{1}{\hat{\emptyset}} \sum_{i=1}^{n} \frac{\sum_{u=1}^{ni}\sum_{v=1}^{ni}\hat{r}_{iu}\hat{r}_{iv} - \sum_{u=1}^{ni}\hat{r}^2{}_{iu}}{ni\ (ni-)} \quad (4.13)$$

$$\hat{\emptyset} = \frac{1}{\sum_{i=1}^{n} ni} \sum_{i=1}^{n}\sum_{t=1}^{ni}\hat{r}_{it}^2 \quad (4.14)$$

**4.3.1.2 Autoregressive Correlation**

When there is there is time dependence in data and repeated observations in the panels have order form. The correlation between subjects is seemed to be exponential function for the time lag. The correlation structure is described as $corr(y_{it}, y_{i\acute{t}}) = \alpha^{|t-\acute{t}|}$.For $y_{it}$ normally distributed is the same as a continuous time autoregressive (AR) process.

In this case using Pearson residuals $\hat{r}_{it=(y_{it}-\hat{\mu}_{it})/\sqrt{V(\hat{\mu}_{it})}}$ from the current fit of the model

$$\hat{\alpha} = \frac{1}{\emptyset}\left[\sum_{i=1}^{n}\left(\frac{\sum_{t=1}^{ni-0}\hat{r}_{i,t}\hat{r}_{i,t+0}}{ni}, \ldots\ldots, \frac{\sum_{t=1}^{ni-k}\hat{r}_{i,t}\hat{r}_{i,t+k}}{ni},\right)\right] \quad (4.15)$$

After that the correlation matrix structure is built as autoregressive structure what we called AR correlations. Selecting correct order for autoregressive process represents a challenge for this structure. We can use the QIC information criterion as tool to help which correlation model represent the correlation order and more appropriate.

**4.3.1.3 Stationary Correlation**

That summed the correlation between subjects is constant into equal time intervals. Used as alternative for the time series autocorrelation which that correlation exists in small time units. Maximum time difference was specified for this situation which observation might be correlated and that the correlation matrix is banded, α is a vector of the correlations for up to user-specified k lags. Also by using Pearson residuals from fitted model, we can obtained the vector of correlation parameters α as the same method as for the autoregressive correlation.

The estimated correlation matrix is banded with 1s down the diagonal, $\hat{\alpha}_1$ down the first band, $\hat{\alpha}_2$ down the second band, and so forth.

The correlation matrix may be succinctly described as

$$R_{uv} = \begin{cases} \alpha_{|u-v|} & if\ |u-v| \le k \\ 0 & otherwise \end{cases} \quad (4.16)$$

The difference between the stationary model and the autoregressive model is that the correlation in stationary modelneither or nor assumed to be nonzero after the specified order.

#### 4.3.1.4 Non stationary Correlation

In this structure the correlation assumed a diagonal of ones the other elements where outside lag of specific time equal to zero, there is no constriction for remaining elements. We specify a matrix of parameter $\alpha$ as

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} ni}{\sum_{i=1}^{n}\sum_{t=1}^{ni}\widehat{r^2}_{i,t}/n_i}\ G \quad (4.17)$$

$$G = \begin{pmatrix} g_{1,1}\widehat{r^2}_{i,1} & g_{1,2}\hat{r}_{i,1}\hat{r}_{i,2} & \cdots & g_{1,ni}\hat{r}_{i,1}\hat{r}_{i,ni} \\ g_{2,1}\hat{r}_{i,2}\hat{r}_{i,1} & g_{2,2}\widehat{r^2}_{i,2} & \cdots & g_{2,ni}\hat{r}_{i,2}\hat{r}_{i,ni} \\ \vdots & \vdots & \ddots & \vdots \\ g_{ni,1}\hat{r}_{i,ni}\hat{r}_{i,1} & g_{ni,2}\hat{r}_{i,ni}\hat{r}_{i,2} & \cdots & g_{ni,ni}\hat{r}_{i,ni}{}^2 \end{pmatrix} \quad (4.18)$$

$$g_{uv} = \left(\sum_{i=1}^{n} I(i,u,v)\right)^{-1} \quad (4.19)$$

$$I(i,u,v) = \begin{cases} 1 & if\ panel\ i\ has\ observations\ at\ indexes\ u\ and\ v \\ 0 & otherwise \end{cases}$$

The same as the stationary structure where the nonstationary correlation matrix is uses the fitted correlation for a particular number of bands g for the matrix. The working correlation matrix is specified as

$$R_{uv} = \begin{cases} 1 & if\ u = v \\ \alpha_{uv} & if\ 0 < |u-v| \le g \\ 0 & otherwise \end{cases} \quad (4.20)$$

The nonstationarystructure differs from stationary structure in that the correlation matrix is not assumed to have constant correlations down the diagonals.

#### 4.3.1.5 Unstructured correlation

Considered as most general correlation structure where we assume a saturated free construction for correlation structure and is equal to the nonstationary matrixfor the maximum lag of time. The working correlation matrix is defined as

$$R = \alpha \quad (4.21)$$

Where α is the same as $\hat{\alpha}$ in non stationary structure

#### 4.3.1.6 Fixed Correlation

In this structure we have not standard form for our correlation which the correlation structure in this case is a user- specified matrix. The matrix can be fitted if we have information about the structure of the correlation matrix from another source. This approach does not estimate the working correlation at each step, but rather it takes the supplied correlation matrix as given [3]. The feature of this structure is provide estimation of a structure that is not have directly ability to derived by specific software program.

#### 4.3.1.7 Toeplitz Working Correlation Matrix

The Toeplitz working correlation structure is a more flexible structure (relying on more parameters).

$$R(\alpha) = \begin{bmatrix} 1 & i = j \\ \alpha_t & i = 1,2,3\ , \dots \dots , ni - t \end{bmatrix} \quad (4.22)$$

As example to this structure

$$\begin{bmatrix} 1 & \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1 & 1 & \alpha_1 & \alpha_2 \\ \alpha_2 & \alpha_1 & 1 & \alpha_1 \\ \alpha_3 & \alpha_2 & \alpha_1 & 1 \end{bmatrix} \quad (4.23)$$

### 4.3.2 PA-GEE Estimating Equation

As we see that the estimating equation for PA-GEE contains two components; estimating equation for the regression parameters and estimating equation for the variance.

Integrating the estimating equations for the regression parameters (4.5) and the ancillary parameters (4.7) the overall PA-GEE is given by

$$\psi(\beta, \alpha) = \left(\psi_\beta(\beta, \alpha), \psi_\alpha(\beta\ , \alpha)\right) = \begin{pmatrix} \sum_{i=1}^{n} X_{ij}^T\ D\ \left(\frac{\partial \mu_i}{\partial \eta}\right)\ [V(\mu_i)]^{-1}\ \left(\frac{y_i - \mu_i}{a(\emptyset)}\right) \\ \sum_{i=1}^{n} \left(\frac{\partial \xi_i}{\partial \alpha}\right)^T\ H_i^{-1}\ (W_i - \xi_i\ ) \end{pmatrix} \quad (4.24)$$

$$V(\mu_i) = \ D(V(\mu_{it}))^{1/2}\ \boldsymbol{R}(\boldsymbol{\alpha})\ D(V(\mu_{it}))^{1/2} \quad (4.25)$$

In PA-GEE model estimation means estimation for regression parameter $\beta$ and treats α as ancillary. Also at each step in the usual GLM algorithm, we first estimate R, and then use it to estimate$\beta$.

## 4.4 The SS-GEE for GLMs

The subject-specific models are second form for the GEE models, same as PA-GEE derived from LIMQL estimating equation for GLMs. They have the same base as the

population averaged models. However, we hypothesize that there is some underlying distribution for random effects in the model that serves as the genesis of the within-panel correlation [3]. Also there are three important point we must address it to create models and estimated their parameter the same as PA-GEE models.

1. Choose a distribution for the random effect.
2. Derive the expected value which depends on the link function and the distribution of the random effect.
3. We must derive the variance-a function of the usual variance and the random effect.

To determined the estimated value of the expected values from SS-GEE models we should use the same forms of the expected value for the random-effects models method.

$$\mu_{it}^{SS} = E(y) = E[E(y_{it}/v_i)] = \int f(v_i) g^{-1}(X\beta_i + v_i) dv_i \quad (4.26)$$

For a single random effect, for many several effects

$$\mu_{it}^{SS} = E(y) = E[E(y_{it}/v_i)] = \int f(v_i) g^{-1}(X\beta_i + z_{it}v_i) dv_i \quad (4.27)$$

Where $z$ is a vector of covariate associated with the random effects, and $f$ is the multivariate density of the random effect vector $vi$.

If we take one single random effect vi, we will find that for the PA-GEES, we have

$$g(\mu_{it}^{PA}) = X_{it}\beta^{PA} \quad (4.28)$$

$$\mu_{it}^{PA} = E(y_{it}) = g^{-1}(X_{it}\,\beta^{PA}) \quad (4.29)$$

$$V^{PA}(y_{it}) = V(\mu_{it}^{PA})\, a \quad (4.30)$$

and for the SS-GEES we have

$$g(\mu_{it}^{SS}) = X_{it}\beta^{SS} + v_i \quad (4.31)$$

$$\mu_{it}^{SS} = \int f(v_i) g^{-1}(x_{it}\beta^{SS} + v_i)\, dv_i \quad (4.32)$$

$$V^{SS}(y_{it}) = \int [g^{-1}(x_{is}\beta^{SS} + v_i) - \mu_{is}^{SS}][g^{-1}(x_{it}\beta^{SS} + v_i) - \mu_{it}^{SS}] df v_i + a(\emptyset) I(s=t) \int V(\mu_{it}^{SS})\, df(v_i) \quad (4.32)$$

where $f$ the distribution of the random effect $v$. The variance matrix for the $ith$ subject is defined in terms of the (s,t) entry.

We have an important note that $g^{-1}(\mu_{it}^{PA})$ is not equal to $g^{-1}(\mu_{it}^{SS})$ except if g() is the identity function and the mean value for the random effect equal to zero ( that we will find it in the Gaussian random effects).

As we mentioned the SS-GEE models for GLMs derived using the same LIMQL estimating equation for the pooled GLMs but we treat equation (4.31) for individual

level for $\mu_i$and equation (4.32) also for $V(\mu_{it})$. The difficulty of this approachrepresentin finding the link function and random-effects distribution f. Some, not for all , options lead to an analytic solution of the integral equations.

The estimating equation for SS-GEE given by

$$\psi(\beta,\alpha) = \left(\sum_{i=1}^{n} x_{ji}^{T}\ D\left(\frac{\partial \mu_i}{\partial \eta}\right)\left(V(\mu_i)\right)^{-1}\left(\frac{y_i-\mu_i}{a(\emptyset)}\right)\right) \quad (4.34)$$

$$\mu_i = \int f(v_i) g^{-1}(X\beta_i + v_i) dv_i \quad (4.35)$$

$$\tilde{V}(\mu_i) \approx D\left(\frac{\partial \mu^{SS}}{\partial \eta^{SS}}\right) V_i\, \Sigma_v(\alpha) V_i^T D\left(\frac{\partial \mu^{SS}}{\partial \eta^{SS}}\right) + \emptyset V(\mu^{SS}) \quad (4.36)$$

$\Sigma_v(\alpha)$ =Parameterized variance matrix

As an example for subject-specific models is mixed model where the subject-specific effects are assumed to follow a parametric distribution across the population.

## 4.5 Efficiency Considerations in GEE

The robustness property of the sandwich variance estimator might tempt the analyst to care little about correctly specifying the within-subject association among repeated measures because consistent $\hat{\beta}$estimates are obtained. However, there are three reasons why an appropriate choice of working correlation matrix is important, especiallyin terms of statistical efficiency.

First, the powerful feature of the sandwich variance estimator to avoid misspecification of the form of working correlation matrix is an asymptotic property, and cannot be assumed to hold in all situations. Use of the sandwich variance estimator tacitly depending on there being independent replications of the vector of outcomes with enough data so that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_i)(Y_i - \hat{\mu}_i)' \to cov(Y_i) \quad as\ n \to \infty \quad (4.37)$$

So, when there are too few replications, highly irregularly spaced observations, or too many missing components of $Y_i$(even if missing completely at random) to estimate the true value for under studying covariance matrix well, the sandwich variance estimator is less appealing. If the number of individuals ($n$) is small but the number of repeated measures ($m$) on each individual is large, sandwich variance estimator is not recommended (Liang and Zeger, 1986; Fitzmaurice et al., 2004).

Second, a working correlation structure which closely approximates the true correlation matrix yields greater efficiency of $\hat{\beta}$ (Albert and McShane, 1995; Wang and Carey, 2003; Wang and Lin, 2005). Wang and Carey (2003) showed that the asymptotic

relative efficiency depended on the three features: 1) disparity between the working correlation structure and the true underlying correlation structure, 2) the estimation method of the correlation parameters, and 3) the design (e.g., the structures of the predictor matrices and the sizes of clusters). Even when conditions 2) and 3) are not problematic, misspecification of the working correlation structure leads to a loss of efficiency and inflated variance estimates.

Third, misspecified working correlation structure sometimes yields inconsistent $\hat{\alpha}$ estimates. Crowder (1995) addressed some issues regarding inconsistency of $\hat{\alpha}$ under a misspecified working correlation structure. If, for example, the true correlation matrix is AR-1, he showed that $\hat{\alpha}$ of the exchangeable working correlation structure does not exist or does not have unique solution in certain cases. For occasions where $\hat{\alpha}$ is not a consistent estimate, that is, it is a violation of an assumption of the Liang and Zeger's theorem, the consistency property of the regression parameter estimates (along with their impact on the variance estimates), is no longer assured.

Even though $\hat{\beta}$ is consistent, misspecified working correlation structure leads the sandwich variance estimate to be inefficient. The most efficient statistical inference in GEE would be to use the model-based variance estimator under the correctly specified working correlation structure. But, of course, one does not know a priori which correlation structure is correct. An appropriate choice of the working correlation matrix can result in large efficiency gains and bias reduction or elimination. Therefore, it is important to try and choose a working correlation matrix that is close to the true correlation matrix, and to do so, the analyst needs a decision tool that performs well.

### 4.6 The GEE2 for GLMs

As we discussed the PA-GEE models for GLMs involve two estimating equations; one use to estimate β (regression parameters) and the other for estimating α (Correlation parameter). Where in PA-GEE we are not concerned about correlation parameter, and we suppose that the coefficient vectors were orthogonal, our interest only on estimating equation for β assuming α as ancillary. Efficiently, our GEE2 models for GLMs interest with both of the parameter vectorsand their associated estimating equations. So there are no constrictions that the regression parameter and correlation parameter estimating equations are orthogonal.

The GEE2 estimating equation be written

$$\sum_{i=1}^{n} \begin{pmatrix} \frac{\partial \mu_i}{\partial \beta} & \frac{\partial \mu_i}{\partial \alpha} \\ \frac{\partial \sigma_i}{\partial \beta} & \frac{\partial \sigma_i}{\partial \alpha} \end{pmatrix}^{T} \begin{pmatrix} v(y_i, y_i) & v(y_i, s_i) \\ v(s_i, y_i) & v(s_i, s_i) \end{pmatrix}^{-1} \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix} = [0] \qquad (4.38)$$

What GEE1 does differently from GEE2 is to assume that the first two terms in the estimating equation are block diagonal (assume zero matrices in the off diagonal positions). It is therefore clear that GEE2 is a generalization of GEE1.

So in GEE2 we are not only have to give a working correlation matrix for the regression parameters, but also we should give a working covariance matrix for the correlation parameters. That make GEE2 model makes constrictions for the first four moments instead of constrictions for the first two moments. These assumptions make GEE2 approach more complicated to interpret the mean vector that involves a correlation on the association parameters. In most applications the excepted value is only defined by the regression parameters; but the assumption that $\partial \mu_i / \partial \sigma \neq 0$ implies that the correlation is a function of the regression parameters β.

Estimating equations in GEE2 are to interpret the estimating equations to model collection of the covariance in terms of the regression parameters and the correlation parameters. The GEE2 in not probably used the reason for that to get consistent estimator of the regression parameters, we should correctly determine the link function as well as the covariance function.

Although, if we want to estimate just for a bloack of diagnal paramterrization and get the correct values, the procedure provides consistent estimation of the regression parameters– even the association is incorrectly specified

As we mentioned that in PA-GEE models the sandwich estimate of variance is powerful to a void misspecification of the correlation structure, however this properties it dies not appropriate by GEE2 models. That the GEE2 models do not support the orthogonality between regression parameter and correlation parameter.

The PA-GEE models give the same value for the association parameters as GEE2 models value when the GEE2 assumes a Gaussian distribution for the random component.

### 4.7 Choosing an Appropriate Model

Referring to the previously mentioned about the likelihood-based models and clarified model restrictions. Then we illustrated the techniques and assumption of GEE models given repeated measurements, the question is which model can take to provide

estimation value for our model? The answer is driven by a combination of factors: the scientific questions of interest, the size and nature of the panel dataset, and the nature of the covariates [21].If the investigator interested in the individual effects of covariates on the outcome variable, so the more appropriate model is a subject-specific likelihood-based model or a subject-specific GEE model to describe subject-specific assumption including unconditional fixed-effects models, conditional fixed-effects models, and random-effects models.

The other option is if the investigator interested about the marginal effects of covariates, in this case the more appropriate model is a population-averaged model. In addition chosen GEE models include the PA-GEE model (using two techniques moment estimators for count data or ALR for binary data), PA-EGEE models, PA-REGEE models, or GEE2 models. The sufficiently of subject-specific and population-averaged GEE models are based on the accessibility of a sufficient number of panels in the dataset to be analyzed. In the general a fixed-effects model is a common model if there are a small number of panels.We cannot include covariates with values that are constant within panels in a fixed-effects (unconditional or conditional) model[17]. Where the covariates are called panel level covariates. Even if our focus is on interpreting the subject-specific effects in our experimental data, we cannot separate the effects of panel level covariates from the fixed effect-they are collinear [17].

We noted that after chosen of appropriate model such as a population averaged model, we have another options to choose between moments estimators for the correlation or the ALR approach estimating correlations that built upon log odds ratios. To make a decision which approach will be used, we will look to the nature of our data. If the data is binary so ALR more appropriate especially if the focus of the analysis involves translation of the correlation coefficients. If the data are not binary and we are interested in analysis includes interpretation of the dependency, then a GEE2 model is more appropriate over a GEE1 model.

### 4.8 Generalized Linear Mixed Models (GLMM)

GLMMs are an extension of linear mixed models to allow response variables from different distributions, such as binary responses. Alternatively, you could think of GLMMs as an extension of generalized linear models (e.g., logistic regression) to include both fixed and random effects (hence mixed models). The concept mixed model refers to the use fixed and random effects together in the same analysis; where the fixed

effects have particular levels that are ineptest in it and may be repeat if we repeat the experiment. The opposite in the random effects model that we have unconditionally interested level, but instated are random selection from large number of levels.

Almost of time the Subject effects are random effects, while treatment levels are usually fixed effects. To apply mixed model, we should have many options involving the nature of the following: hierarchy, the fixed effects and the random effects.

GLMM method is founded on computing a weighted average of all of the subject within-cluster models regressing Y on X in that cluster.

# CHAPTER 5

## APPLICATION

Many statistical programs are support GEEs , SAS, S-plus, STATA, SPSS,R all of these programs had it own functions to fit GEEs. In this thesis we use R program to make analysis on data.

In R program we have two packages geepack package and gee package have several functions fit the GEEs.

### 5.1 Applied Case (1)

We faced a big problem in access to real data to apply on it; that most of data no available and follow-up new cases require a lot of time.

As attempt to apply on data that had never use GEEs as analysis method, we take subset of data from study of women's Health Across the Nation (SWAN). SWAN is a multi-site, longitudinal, epidemiologic study designed to examine the health of women during their middle years. The study examines the physical, biological, psychological and social changes during this transitional period. The study is co-sponsored by the National Institute on Aging (NIA), the National Institute of Nursing Research (NINR), the National Institutes of Health (NIH), Office of Research on Women's Health, and the National Center for Complementary and Alternative Medicine. The data collected from seven designated research centres, as call interview. The fallow-up is annual divided as visits (1998-2007).

Our sub dataset took three visits ((1998-199) , (1999-2000) , (2000 – 2001)). Assuming that we are interest to know affecting factor on menopausal status variable across the time. We assumed menopausal status as dependent variable, age , ethnic and overall health status. All of these variables depend on Self-Administered Questionnaire. In the menopausal status we took two levels Pre-menopausal and Post-menopausal, ethnic defined as the language of questionnaire, overall health determined by the cases as:

excellent, very good, good, poor, fail. We want to see how longitudinal study can address change of menopausal status over time and what variables predict differences among subjects (taking three predict variables).

The model will be

$$Status = f(ethnic, overallhealth, age) \quad (5.1)$$

We took 396 subjects form the datasets for three visits (1998 – 2001 ) (N = 1185 observations). Using R program and applying glm in the beginning we get the following output

glm(formula = nstat ~ ethnic + overallhealth + age, family = "binomial", data = my)

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -1.864 | -0.608 | -0.417 | -0.251 | 2.636 |

Table 5.1 The output for applied case (1) after applying GLM

| Coefficients | Estimate | Std. Error | z | value Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -19.6353 | 1.6739 | -11.73 | 2e-16 *** |
| ethnic English | -0.7513 | 0.4319 | -1.74 | 0.08191 |
| ethnic Japanese | -0.9739 | 0.635 | -1.53 | 0.12509 |
| ethnic Spanish | -0.2775 | 0.5762 | -0.48 | 0.6301 |
| overallhealth Fair | 1.4274 | 0.3164 | 4.51 | 6.4e-06 *** |
| overallhealth Good | 0.6288 | 0.2808 | 2.24 | 0.02515 * |
| overallhealth Poor | 1.6164 | 0.4357 | 3.71 | 0.00021 *** |
| overVeryallhealth good | 0.3244 | 0.2743 | 1.18 | 0.23705 |
| Age | 0.3693 | 0.0316 | 11.68 | < 2e-16 *** |
| Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 | | | | |

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1163.43 on 1184 degrees of freedom

Residual deviance: 941.17 on 1176 degrees of freedom

AIC: 959.2

Number of Fisher Scoring iterations: 5

Then by using geeglm ( ) function to apply GEE method , we get the following :

Call:

geeglm(formula = nstat ~ ethnic + overallheath + age, family = binomial", data = my, id = id, corstr = "exchangeable")

Table 5.2 The Output for applied case (1) after Applying GEE

| Coefficients | Estimate | Std. Error | Wald | Pr(>\|W\|) |
|---|---|---|---|---|
| (Intercept) | -19.635 | 1.623 | 146.37 | 2e-16 *** |
| ethnic English | -0.751 | 0.392 | 3.68 | 0.055 . |
| ethnic Japanese | -0.974 | 0.606 | 2.58 | 0.108 |
| ethnic Spanish | -0.278 | 0.595 | 0.22 | 0.641 |
| overallhealth Fair | 1.427 | 0.325 | 19.31 | 1.1e-05 *** |
| overallhealth Good | 0.629 | 0.28 | 5.06 | 0.024 * |
| overallhealth Poor | 1.616 | 0.389 | 17.24 | 3.3e-05 *** |
| overVeryallhealth good | 0.324 | 0.278 | 1.37 | 0.242 |
| Age | 0.369 | 0.031 | 142.24 | < 2e-16 *** |
| Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 | | | | |

Estimated Scale Parameters:

Estimate Std.err

(Intercept) 0.997 0.257

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

Estimate Std.err

alpha 0 0

Number of clusters: 1185 Maximum cluster size: 1

We note that there is no different between the two methods in estimated coefficients, where the overall health status and age is significant. We note that ethnic English in GEE method is 0.055 close to be significant more than in GLM.

The reason for that may be our predictor variables are ordinal variable type and the correlation measure in GEE depends on Pearson method that is on true values of variables. So we need to follow the data for long time to determine the true value of correlation. Another challenge is that availability sampling designs because the correlation structure for combinations of used and available points are not likely to follow common parametric forms.

One of the properties of GEEs that if the correlation parameter is not significant it give the same result as GLM methods.

### 5.2 Applied Case (2)

To help clarify vision behind GEE we took another dataset to make analysis on it. The dataset is from university of North Carolina library; about Wildlife telemetry data using GPS. The goal of this study was to relate nest occupancy by mallards (a binary response) to surrounding land use characteristics and the type of nesting structure that was provided. Individual nests were observed repeatedly over time and at each time point the nest was classified as occupied (nesting2 = 1) or unoccupied (nesting2 = 0) by mallards. If the nest was occupied by a species other than mallard, no data were recorded or nesting2 was assigned a missing value, NA. The available predictors include the following:

- Year : the year of the observation (1997, 1998, 1999).
- Periods : the seasonal observation period (1, 2, 3, 4).
- Deply1 : nest structure type (1 for single cylinder, 2 for double cylinder).
- Size. : the size of the wetland in which the nest structure was located.
- Rblval..: a measure of visual obstruction around the nest.
- Stratno.: id variable.

The model choose to fit is one that includes main effects for the five predictors years, period, rblval, deply, and size, and interactions between period and year as well as between rblval and period. The variables year, period, and deply are recorded as numeric variables in the data set but need to be treated as factors in the model.

$$nesting2 = f(year, period, deply, rblval, year * period, rblval * period, size)$$

First we fitted GLM on the data the output will be :

Call:

glm(formula = nesting2 ~ factor(year) + factor(period) + factor(deply) +
rblval + factor(year) * factor(period) + rblval * factor(period) +
size, family = binomial, data = sm)

Deviance Residuals:

| Min | 1Q | Median | 3Q | Max |
|---|---|---|---|---|
| -1.369 | -0.564 | -0.413 | -0.299 | 2.636 |

Table 5.3 Output applied case (2) using glm function

| Coefficients | Estimate | Std. Error | z | valuePr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -3.5145 | 0.5248 | -6.7 | 2.1e-11 *** |
| factor(year)1998 | 0.9951 | 0.3864 | 2.58 | 0.01002* |

Table 5.3 Output applied case (2) using glm function(cont'd)

| | | | | |
|---|---|---|---|---|
| factor(year)1999 | 0.95 | 0.4021 | 2.36 | 0.01814 * |
| factor(period)2 | 1.1643 | 0.7153 | 1.63 | 0.1036 |
| factor(period)3 | 2.7722 | 1.9902 | 1.39 | 0.16364 |
| factor(period)4 | 5.3477 | 2.3792 | 2.25 | 0.02460 * |
| factor(deply)2 | 0.1974 | 0.1896 | 1.04 | 0.2978 |
| rblval | 4.1176 | 1.3785 | 2.99 | 0.00282 ** |
| size | 0.0539 | 0.0111 | 4.87 | 1.1e-06 *** |
| factor(year)1998:factor(period)2 | -2.0035 | 0.6212 | -3.23 | 0.00126 ** |
| factor(year)1999:factor(period)2 | -1.2902 | 0.5773 | -2.23 | 0.02543 * |
| factor(year)1998:factor(period)3 | -1.3525 | 0.9563 | -1.41 | 0.15724 |
| factor(year)1999:factor(period)3 | -0.9565 | 0.9363 | -1.02 | 0.30701 |
| factor(year)1998:factor(period)4 | -1.4774 | 0.7067 | -2.09 | 0.03657 * |
| factor(year)1999:factor(period)4 | -1.1883 | 0.6694 | -1.78 | 0.07589 . |
| factor(period)2:rblval | -2.2513 | 2.2175 | -1.02 | 0.30998 |
| factor(period)3:rblval | -5.9158 | 2.6325 | -2.25 | 0.02463 * |
| factor(period)4:rblval | -6.5256 | 1.8606 | -3.51 | 0.00045 *** |
| Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 | | | | |

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 880.20 on 1139 degrees of freedom

Residual deviance: 803.19 on 1122 degrees of freedom

(96 observations deleted due to missingness)

AIC: 839.2

Number of Fisher Scoring iterations: 5

The purpose behind these analyses is to make comparison between different methods for analyses of correlated data. To apply GEE method we used geeglm () function to fit the data.

Call:

geeglm(formula = nesting2 ~ factor(year) + factor(period) + factor(deply) +rblval + factor(year) * factor(period) + rblval * factor(period) +size, family = binomial, data = sm, id = strtno, corstr = "exchangeable", scale.fix = T)

Table 5.4 Output applied case (2) of geeglm function

| | Estimate | Std. Error | z | valuePr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -3.3454 | 0.5891 | 32.25 | 1.4e-08 *** |
| factor(year)1998 | 0.984 | 0.3482 | 7.99 | 0.00471 ** |

Table 5.4 Output applied case (2) of geeglm function(cont'd)

| | | | | |
|---|---|---|---|---|
| factor(year)1999 | 0.9367 | 0.2959 | 10.02 | 0.00155 ** |
| factor(period)2 | 1.2491 | 0.6713 | 3.46 | 0.06276 . |
| factor(period)3 | 2.6739 | 1.7904 | 2.23 | 0.13532 |
| factor(period)4 | 5.0561 | 2.6043 | 3.77 | 0.05220 . |
| factor(deply)2 | 0.054 | 0.2905 | 0.03 | 0.85246 |
| rblval | 3.805 | 1.6471 | 5.34 | 0.02088 * |
| size | 0.0532 | 0.015 | 12.64 | 0.00038 *** |
| factor(year)1998:factor(period)2 | -1.8274 | 0.5464 | 11.18 | 0.00082 *** |
| factor(year)1999:factor(period)2 | -1.1794 | 0.4353 | 7.34 | 0.00674 ** |
| factor(year)1998:factor(period)3 | -1.3463 | 0.8127 | 2.74 | 0.09758 . |
| factor(year)1999:factor(period)3 | -0.9768 | 0.7712 | 1.6 | 0.20534 |
| factor(year)1998:factor(period)4 | -1.5372 | 0.6499 | 5.59 | 0.01802 * |
| factor(year)1999:factor(period)4 | -1.1972 | 0.5979 | 4.01 | 0.04526 * |
| factor(period)2:rblval | -2.3446 | 2.1158 | 1.23 | 0.26779 |
| factor(period)3:rblval | -5.334 | 2.5268 | 4.46 | 0.03478 * |
| factor(period)4:rblval | -6.0642 | 2.057 | 8.69 | 0.00320 ** |
| 0.02463 * | | | | |

Scale is fixed.

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

| | Estimate | Std.err |
|---|---|---|
| alpha | 0.13 | 0.0228 |

Number of clusters: 109 Maximum cluster size: 12

We note that parameter alpha is significant that means there is correlation between subjects. Although there is no big difference between GLM and GEE estimated coefficients, but in exist of alpha GEE is more appropriate. It is known that the focus of the GEE in on estimating average response over population rather than regression parameters that would enable prediction of the effect of changing our covariates on the individuals. This explain appear of a little significant on period factor; that means the factor period had effect on changing of our dependent variable over time. Where does not appear on GLM that the focus on estimating the coefficients not on the effect of changing over time.

As we know GEE method doesn't stop on finding the estimated value for the predictors but also we have many structures describe the correlation within subjects that is sometimes give different values for our coefficients. We can choose the appropriate structure to describe the nature of correlation by QIC measure which the structure has a smallest value is the best structure to define correlation between subjects.

In the following we applied different correlation structure on our data Table (5.3) summarize the outputs

Table 5.5: Summarize output of applying exchangeable, unstructured and autoregressive structure

| Coefficients | Exchangeable | | Autoregressive (1) | | Unstructured | |
|---|---|---|---|---|---|---|
| | Estimate | Std.err | Estimate | Std.err | Estimate | Std.err |
| (Intercept) | -3.3454 | 0.58913 | -3.5471 | 0.6211 | -3.1459 | 0.4948 |
| factor(year)1998 | 0.984 | 0.3482 | 0.9294 | 0.3448 | 1.2724 | 0.2854 |
| factor(year)1999 | 0.9367 | 0.29589 | 0.9061 | 0.3056 | 0.6846 | 0.3442 |
| factor(period)2 | 1.2491 | 0.67125 | 1.3209 | 0.7089 | 0.7457 | 0.6888 |
| factor(period)3 | 2.6739 | 1.7904 | 3.2361 | 1.9664 | 3.3484 | 1.6316 |
| factor(period)4 | 5.0561 | 2.6043 | 5.1727 | 3.0548 | 5.5634 | 2.1854 |
| factor(deply)2 | 0.054 | 0.29049 | 0.1836 | 0.2935 | 0.2015 | 0.2634 |
| Rblval | 3.805 | 1.64706 | 4.2881 | 1.7141 | 3.3533 | 1.3543 |
| size | 0.0532 | 0.01495 | 0.0533 | 0.0149 | 0.0603 | 0.0147 |
| factor(year)1998:factor(period)2 | -1.8274 | 0.54644 | -1.8079 | 0.5641 | -2.6512 | 0.6005 |
| factor(year)1999:factor(period)2 | -1.1794 | 0.43531 | -1.1945 | 0.4529 | -0.8355 | 0.4835 |
| factor(year)1998:factor(period)3 | -1.3463 | 0.81265 | -1.1735 | 0.8833 | -1.6968 | 0.7825 |
| factor(year)1999:factor(period)3 | -0.9768 | 0.77123 | -0.8528 | 0.829 | -1.2122 | 0.8236 |
| factor(year)1998:factor(period)4 | -1.5372 | 0.64991 | -1.5233 | 0.7157 | -1.5179 | 0.816 |
| factor(year)1999:factor(period)4 | -1.1972 | 0.59792 | -1.1964 | 0.6595 | -0.9501 | 0.6503 |
| factor(period)2:rblval | -2.3446 | 2.1158 | -2.7369 | 2.2156 | -1.0816 | 2.0681 |
| factor(period)3:rblval | -5.33395 | 2.52679 | -6.4411 | 2.7823 | -5.6149 | 2.1306 |
| factor(period)4:rblval | -6.0642 | 2.05699 | -6.6059 | 2.2612 | -6.1187 | 1.7976 |

For stationary and non-stationary structure we cannot apply it in geeglm () function we used gee () function to get the results :

Table 5.6 Estimated coefficients using stationary and nonstationary structure

| Coefficients | Estimate (stationary) | Estimate (non-stationary ) |
|---|---|---|
| (Intercept) | -3.5145 | -3.5145 |
| factor(year)1998 | 0.9951 | 0.9951 |
| factor(year)1999 | 0.9500 | 0.9500 |
| factor(period)2 | 1.1643 | 1.1643 |
| factor(period)3 | 2.7722 | 2.7722 |
| factor(period)4 | 5.3477 | 5.3477 |
| factor(deply)2 | 0.1974 | 0.1974 |
| Rblval | 4.1176 | 4.1176 |
| size | 0.0539 | 0.0539 |
| factor(year)1998:factor(period)2 | -2.0035 | -2.0035 |
| factor(year)1999:factor(period)2 | -1.2902 | -1.2902 |
| factor(year)1998:factor(period)3 | -1.3525 | -1.3525 |
| factor(year)1999:factor(period)3 | -0.9565 | -0.9565 |
| factor(year)1998:factor(period)4 | -1.4774 | -1.4774 |
| factor(year)1999:factor(period)4 | -1.1883 | -1.1883 |
| factor(period)2:rblval | -2.2513 | -2.2513 |
| factor(period)3:rblval | -5.9158 | -5.9158 |
| factor(period)4:rblval | -6.5256 | -6.5256 |

We note that the output for stationary and non-stationary does not include Std-err that the structures run on gee ( ) function.

To see which structure more appropriate from another we use QIC measure which the smallest value is the best. We test exchangeable, AR(1) and unstructured structures the output is :

| Measure | Exchangeable | AR(1) | Unstructured |
|---|---|---|---|
| QIC | 844 | 847.4 | 856.4 |
| CIC | 19 | 21.7 | 20.5 |

So exchangeable is a suitable structure to define our data within correlation.

The robust or Huber sandwich estimator for GEE provides reasonable variance estimates even when the specified correlation model is incorrect. If the correlation model is nearly correct, so will be the naive standard errors, and robust speciation is unlikely to help much. So, unless you have strong reason to believe that the chosen correlation matrix is correct, it is recommended to always report the robust standard errors.

An alternative way to account for repeated measures is to fit a mixed effects model with random intercepts. This yields to a conditional model instead of the marginal model that we obtain with GEE also what we called Subject-specific models.

The following table show the estimated coefficients value using three methods (GEE,GLM,GLMM).

Table 5.7 Values estimated coefficients using GLM, GEE and GLMM

| Coefficients | GLM | GEE | GLMM |
|---|---|---|---|
| (Intercept) | -3.5145 | -3.3454 | -4.4397 |
| factor(year)1998 | 0.9951 | 0.984 | 1.214 |
| factor(year)1999 | 0.95 | 0.9367 | 1.1412 |
| factor(period)2 | 1.1643 | 1.2491 | 1.4737 |
| factor(period)3 | 2.7722 | 2.6739 | 3.1248 |
| factor(period)4 | 5.3477 | 5.0561 | 6.0263 |
| factor(deply)2 | 0.1974 | 0.054 | 0.1628 |
| Rblval | 4.1176 | 3.805 | 4.6914 |
| size | 0.0539 | 0.0532 | 0.0773 |
| factor(year)1998:factor(period)2 | -2.0035 | -1.8274 | -2.2292 |
| factor(year)1999:factor(period)2 | -1.2902 | -1.1794 | -1.3829 |
| factor(year)1998:factor(period)3 | -1.3525 | -1.3463 | -1.7355 |
| factor(year)1999:factor(period)3 | -0.9565 | -0.9768 | -1.2231 |
| factor(year)1998:factor(period)4 | -1.4774 | -1.5372 | -1.8611 |
| factor(year)1999:factor(period)4 | -1.1883 | -1.1972 | -1.4234 |
| factor(period)2:rblval | -2.2513 | -2.3446 | -2.6646 |

Table 5.7 Values estimated coefficients using GLM, GEE and GLMM(cont'd)

| | | | |
|---|---|---|---|
| factor(period)3:rblval | -5.9158 | -5.33395 | -6.3562 |
| factor(period)4:rblval | -6.5256 | -6.0642 | -7.3751 |

We note the parameter estimates from a conditional model tend to be larger in magnitude (more positive or more negative) than the corresponding parameter estimates from a marginal model. That the logistic regression uses a nonlinear link function, the magnitudes of the estimated effects in marginal models will differ from those same effects estimated in conditional models. These are not inconsistent results but reflect the fact that the models are estimating different parameters (marginal and conditional).

# CHAPTER 6

## RESULTS AND DISCUSSION

Longitudinal data provide unique opportunities for inference regarding the effect of an intervention or a predictor. Changes in predictor conditions can be correlated with changes in response conditions. However, analysis of longitudinal data requires methods that take into account the within-subject correlation of repeated measures.

GEE is application treat correlated data use the same integrated method of GLM but use a quasi-likelihood instead of full likelihood method. Where correct specification of the expected value and variance functions is sufficient for unbiased estimates, the used model does not fully determine the distribution of the outcomes in each panel.

The GEE model is differs in a substantial logical way from the approaches involved under what we called “random-effects,” “multilevel,” and “hierarchical” models. In addition to our seeking of more efficient estimators of regression confidents, the important feature for GEE model is providing reasonably accurate standard errors, thus will affect on confidence intervals that will be with the correct coverage rates. For the other method the process different, that these approaches explicitly model and fit the intra variation between subject and gathering this, also the residual variance, into standard errors. We can say that the GEE method does not provide exactly model for between penal or group variation; but fit its counterpart, between subjects or within panel variance then use fitted value for correlation to re-estimate the regression parameters and to calculate standard errors. This estimated correlation takes many forms known as correlation structures. Choosing between correlation structures depends on the nature of data; additionally we use QIC measure to choose appropriate structure; the smallest value is the best. The main objective of using GEE is determined

correlation structure that helps to avert spurious explanatory variables sometimes or includes important explanatory variables to our model.

GEE have two important powerful features. The first one is the fitted value of regression parameters $\widehat{\beta}$ that were derived approaching GEE are widely valid estimates that provide correct estimation value with increasing sample size whatever was the choice of correlation structure. Lies the importance of the correlation structure in that making analysis of weight observations more simple also choosing appropriate correlation structure will make estimation of regression parameters more efficient. According to optimal estimation theory (e.g. Gauss-Markov theory) the appropriate correlation structure option for efficiency of estimation is the true correlation structure.

Second, the correlation option is used to obtain model-based standard errors and these do require that the correlation model option is correct in order to use the standard errors for inference. A standard feature of GEE is the additional reporting of empirical standard errors which provide valid estimates of the uncertainty in $\hat{\beta}$even if the correlation model is not correct.

The scale parameter in GEE models is the same as the dispersion parameter in GLMs. For Gaussian model, the scale parameter is sigma-squared, for binomial or Poisson model scale parameter fixed at unity.

Comparing GEE with GLM:

1. Typically, GEE returns parameter estimates that are fairly close to those returned by GLM when the models are being compared assume the same mean-variance relationship, i.e., use the same value of the family argument.
2. When there is a hypothesized correlation structure in a GEE model, the estimated variance of the parameter estimates tends to be larger in GEE than it is for GLM.
3. Because GEE is not a likelihood-based method, GEE output does not include a log-likelihood, likelihood ratio tests, or AIC. Reported significance tests are the Wald tests for individual parameter estimates.

The generalized linear mixed models (GLMMs) allow subject-specific inferences and variance decomposition between and within groups with a single model fit. Accommodates hierarchical correlation structures. In contrast requires more assumptions and specification of correlation structure problematic with use–availability designs.

## REFERENCES


[1] Edwards, L.J.,(2000).”Modern Statistical Techniques for the Analysis of Longitudinal Data in Biomedical Research”, Pediatric pulmonology, 30(4): 330-344.

[2] Chiou, J.-M., H.-G. Muller, and J.-L. Wang,(2004).” Functional Response Models”, Statistica Sinica, 14(3): 675-694.

[3] Hardin, J.W.( 2005), Generalized Estimating Equations (GEE), Chapman & Hall/CRC, New York

[4] Liang, K.-Y. and S.L. Zeger,(1986).”Longitudinal Data Analysis Using Generalized Linear Models”, Biometrika, 73(1): 13-22.

[5] Lindsey, J.K. and P. Lambert, (1998).”On the Appropriateness of Marginal Models for Repeated Measurements in Clinical Trials”, Statistics in medicine, 17(4): 447-469.

[6] Pan, W.,(2001).”Akaike's Information Criterion in Generalized Estimating Equations. Biometrics”, 57(1): 120-125.

[7] Tirthapura, S. and D.P. Woodruff,(2012). “A general Method for Estimating Correlated Aggregates Over A data Stream in Data Engineering (ICDE)”, IEEE 28th International Conference. IEEE.

[8] Tu, X., et al.,(2004).” Power Analyses for Longitudinal Trials and other Clustered Designs”, Statistics in Medicine, 23(18): 2799-2815.

[9] Yan, J. and J. Fine,(2004).” Estimating Equations for Association Structures”, Statistics in medicine, 23(6): 859-874.

[10] Zeger, S.L., K.-Y. Liang, and P.S. Albert,(1988).” Models for Longitudinal Data: A generalized Estimating Equation Approach”, Biometrics: 1049-1060.

[11] Diggle, P., et al.,(2002), Analysis of Longitudinal Data, Oxford University Press, London

[12] Hedeker, D. and R.D. Gibbons,(2006).” Longitudinal Data Analysis”, John Wiley & Sons. 451

[13] Horton, N.J. and S.R. Lipsitz,(1999) .”Review of Software to Fit Generalized Estimating Equation Regression Models”, The American Statistician, 53(2): 160-169.

[14] Liang, K.-Y. and S.L. Zeger,(1986).” Longitudinal Data Analysis Using Generalized Linear Models”, Biometrika, 73(1): 13-22.

[15] Lipsitz, S.R., N.M. Laird, and D.P. Harrington,(1991).” Generalized Estimating Equations for Correlated Binary Data: Using the Odds Ratio as A measure of Association”, Biometrika, 78(1): 153-160.

[16] McCulloch, C.E. and J.M. Neuhaus,(2001) Generalized Linear Mixed Models, Wiley Online Library, New work

[17] Twisk, J.W.,(2013), Applied Longitudinal Data analysis for Epidemiology: A Practical Guide, Cambridge University Press, London

[18] Verbeke, G. and G. Molenberghs,( 2009), Linear mixed models for longitudinal data, Springer Science & Business Media, New work

[19] Williams, R.L.,(2000).” A note on Robust Variance Estimation for Cluster Correlated Data”, Biometrics, 56(2): 645-646.

[20] Wu, C.-F.,(1981) .”Asymptotic Theory of Nonlinear Least Squares Estimation. The Annals of Statistics”. : 501-513.

[21] Ballinger, G.A.,(2004).” Using Generalized Estimating Equations for Longitudinal Data Analysis”. Organizational research methods, 7(2): 127-150.

[22] Cho, H. and A. Qu,(2013). “Model Selection for Correlated Data with Diverging Number of Parameters”, Statistica Sinica, 23: 901-927.

[23] Feddag, M.-L., I. Grama, and M. Mesbah,(2003).” Generalized Estimating Equations (GEE) for Mixed Logistic Models”, Communications in Statistics-Theory and methods, 32(4): 851-874.

[24] Fitzmaurice, G.M., N.M. Laird, and J.H. Ware,(2012).” Applied Longitudinal Analysis”. 998. 2012

[25] Ghisletta, P. and D. Spini,( 2004).” An introduction to Generalized Estimating Equations and An application to Assess Selectivity Effects in A longitudinal Study on Very Old Individuals”, Journal of Educational and Behavioral Statistics, 29(4): 421-437.

[26] Hansen, L.P.,(2007),Generalized Method of Moments Estimation. The New Palgrave Dictionary of Economics, Spring, New work

[27] Hothorn, T. and B.S. Everitt,( 2014), A handbook of statistical analyses using R, CRC Press.

[28] Lange, C., et al., (2003).”A multivariate Family Based Association Test Using Generalized Estimating Equations” FBAT GEE. Biostatistics, 4(2): 195-206.

[29] Mátyás, L.,(1999) ,Generalized Method of Moments Estimation, Cambridge University Press, London.

[30] Singer, J.D. and J.B. Willett,(2003), Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence, Oxford university press,London.

[31]. Frees, E.W.,(2004), Longitudinal and Panel Data: Analysis and Applications in the Social Sciences, Cambridge University Press,London.

[32] Hsiao, C.,( 2014), Analysis of Panel Data, Cambridge university press, London.

[33] Twisk, J.W.,(2013), Applied Longitudinal Data analysis for Epidemiology: a Practical Guide, Cambridge University Press, London.

[34] Blommaert, A., N. Hens, and P. Beutels,(2014).” Data Mining for Longitudinal Data Under Multicollinearity and Time Dependence Using Penalized Generalized Estimating Equations”, Computational Statistics & Data Analysis, 71: 667-680.

[35] Frees, E.W.,(2004), Longitudinal and Panel Data: Analysis and Applications in the Social Sciences, Cambridge University Press, London.

[36] Hanley, J.A., A. Negassa, and J.E. Forrester,(2003).” Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An orientation”, American journal of epidemiology, 157(4): 364-375.

[37] Hu, F.B., et al.,(1998).” Comparison of Population-averaged and Subject-specific Approaches for Analyzing Repeated Binary Outcomes”, American Journal of Epidemiology, 147(7): 694-703.

[38] Likelihood, R.,( 1997).” Modelling Longitudinal and Spatially Correlated Data”, Lecture Notes in Statistics, 122.

[39] Tsai, M.-Y., J.-F. Wang, and J.-L. Wu,(2011).” Generalized Estimating Equations with Model Selection for Comparing Dependent Categorical Agreement Data”, Computational Statistics & Data Analysis, 55(7): 2354-2362.

[40] Cox, D.R. and D. Oakes,(1984) Analysis of Survival Data. , CRC Press, New work.

[41] McCullagh, P., J.A. Nelder, and P. McCullagh,(1989), Generalized Linear Models, Chapman and Hall. London.

[42] Chiou, J.-M., H.-G. Muller, and J.-L. Wang,(2004)." Functional Response Models", Statistica Sinica, 14(3): 675-694.

[43] The university of North Carolina at Chapel Hel ,Courses spring (2010),

www.unc.edu/courses/2010spring/ecol/562/001/data/lab7/sm003.txt

[44] Inter-university Consortium for political and social research, Data analysis http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies

## APPENDIX-A

### CODES

In this section we provide R code for our examples (1) and (2)

```
# program code for example (1)
# data loaded from text file
my <- read.delim("C:/Users/Fatma/Desktop/insa4.txt", header=TRUE)
data(my)
my$nstat <- as.numeric(my$stat == "Post-menopausal")
# glm function
resp_glm <- glm(= nstat ~ ethnic + overallhealth + age
          , data = my, family = "binomial")
# GEE using geeglm () function
resp_gee <- geeglm(= nstat ~ ethnic + overallhealth + age
          , data = my, family = "binomial",id = id,
          corstr="exchangeable")
summary(resp_glm)
summary(resp_gee)
```

```
# code for example (2)
# data loaded from webside
sm<-read.table(
'http://www.unc.edu/courses/2010spring/ecol/562/001/data/lab7/sm003.txt',    header=T,
sep=',')
# using glm to find estimators
result_glm <-glm( nesting2 ~ factor(year) + factor(period) + factor(deply) + rblval+
factor(year)*factor(period) + rblval*factor(period) + size , family=binomial, data=sm)
summary(result_glm)
# using GEE moethod bu gee() function
result_geeind <-gee( nesting2 ~ factor(year) + factor(period) + factor(deply) + rblval+
                     factor(year)*factor(period) + rblval*factor(period) + size ,
                   family=binomial,corstr="independence",
id=strtno,data=sm,scale.fix=T)
summary(result_geeind)
# Using GEE method by geeg () function using exchangeable correlation strucure
result_geeexch <-gee( nesting2 ~ factor(year) + factor(period) + factor(deply) + rblval+
factor(year)*factor(period) + rblval*factor(period) + size ,
family=binomial,corstr="exchangeable",  id=strtno,data=sm,scale.fix=T)
summary(result_geeexch)
#Using GEE method by geeg () function using AR(1) correlation strucure
result_geear <-gee( nesting2 ~ factor(year) + factor(period) + factor(deply) + rblval+
                     factor(year)*factor(period) + rblval*factor(period) + size ,
                   family=binomial,corstr="ar1",  id=strtno,data=sm,scale.fix=T)
summary(result_geear)
#Using GEE method by geeg () function using unstructured correlation strucure
result_geeuns <-gee( nesting2 ~ factor(year) + factor(period) + factor(deply) + rblval+
                     factor(year)*factor(period) + rblval*factor(period) + size ,
                  family=binomial,corstr="unstructured",  id=strtno,data=sm,scale.fix=T)
summary(result_geeuns)
#Using GEE method by geeg () function using stationary correlation strucure
result_geesta <-gee( nesting2 ~ factor(year) + factor(period) + factor(deply) + rblval+
                     factor(year)*factor(period) + rblval*factor(period) + size ,
```

```
                    family=binomial,corstr="stat_M_dep",Mv=1,
id=strtno,data=sm,scale.fix=T)
##Using GEE method by geeg () function using nonstationary correlation strucure
result_geenosta <-gee( nesting2 ~ factor(year) + factor(period) + factor(deply) + rblval+
                    factor(year)*factor(period) + rblval*factor(period) + size ,
                  family=binomial,corstr="non_stat_M_dep",Mv=1,
id=strtno,data=sm,scale.fix=T)
# calcoulate QIC
QIC.binom.geeglm <- function(model.geeglm, model.independence)
{
  #calculates binomial QAIC of Pan (2001)
  #obtain trace term of QAIC
  AIinverse <- solve(model.independence$geese$vbeta.naiv)
  V.msR <- model.geeglm$geese$vbeta
  trace.term <- sum(diag(AIinverse%*%V.msR))
  #estimated mean and observed values
  mu.R <- model.geeglm$fitted.values
  y <- model.geeglm$y
  #scale for binary data
  scale <- 1
  #quasilikelihood for binomial model
  quasi.R <- sum(y*log(mu.R/(1-mu.R))+log(1-mu.R))/scale
  QIC <- (-2)*quasi.R + 2*trace.term
  output <- c(QIC,trace.term)
  names(output) <- c('QIC','CIC')
  output
}
########
QIC.binom.geeglm<-function(model.geeglm,model.independence)
{
  #calculates binomial QAIC of Pan (2001)
  #obtain trace term of QAIC
  AIinverse<-solve(model.independence$geese$vbeta.naiv)
  V.msR<-model.geeglm$geese$vbeta
```

```
  trace.term<-sum(diag(AIinverse%*%V.msR))
  #estimated mean and observed values
  mu.R<-model.geeglm$fitted.values
  y<-model.geeglm$y
  #scale for binary data
  scale <- 1
  #quasilikelihood for binomial model
  quasi.R<-sum(y*log(mu.R/(1-mu.R))+log(1-mu.R))/scale

  QIC<-(-2)*quasi.R + 2*trace.term
  output <- c(QIC,trace.term)
  names(output)<-c('QIC','CIC')
  output
}
sapply(list(result_geeexch, result_geear, result_geeuns,result_geesta), function(x)
QIC.binom.geeglm(x,result_geeind))
# glmer () function for applying GLMM
lmerfit <- glmer(nesting2~factor(year)+ factor(period)+ factor(deply)+ rblval+
factor(year)*factor(period)+ rblval*factor(period)+ size+ (1|strtno), family=binomial,
data=sm)
summary(lmerfit)
```

# CURRICULUM VITAE

## PERSONAL INFORMATION

**Name Surname** : Fatima YASSIN

**Date of birth and place** : May,23, 1985 - Suadn

**Foreign Languages** : English - Turkish

**E-mail** : fatima_batch26@hotmail.com

## EDUCATION

| Degree | Department | University | Date of Graduation |
|---|---|---|---|
| Undergraduate | Statistic/computer | Gezira University- Sudan | June , 2008 |
| High School | Scientific | Medani - Sudan | March, 2003 |

## WORK EXPERIENCE

| Year | Corporation/Institute | Enrollment |
|---|---|---|
| 2009-2015 | Central Bank Of Sudan | Employ |
| 2008- 2009 | Gezira University | Teaching assistant |