Journal of Data Science **12**(2014), 293-305

# Analysis of Correlation Structures using Generalized Estimating Equation Approach for Longitudinal Binary Data

Jennifer S.K. Chan
*The University of Sydney*

*Summary*: Longitudinal binary data often arise in clinical trials when repeated measurements, positive or negative to certain tests, are made on the same subject over time. To account for the serial correlation within subjects, we propose a marginal logistic model which is implemented using the Generalized Estimating Equation (GEE) approach with working correlation matrices adopting some widely used forms. The aim of this paper is to seek some robust working correlation matrices that give consistently good fit to the data. Model-fit is assessed using the modified expected utility of Walker & Gutiérrez-Peña (1999). To evaluate the effect of the length of time series and the strength of serial correlation on the robustness of various working correlation matrices, the models are demonstrated using three data sets containing respectively all short time series, all long time series and time series of varying length. We identify factors that affect the choice of robust working correlation matrices and give suggestions under different situations.

*Key words*: Longitudinal data; Serial correlation; Robustness; Generalized Estimating Equation approach

## 1. Introduction

In longitudinal studies, observations measured repeatedly from the same subject over time are serially correlated. Time series with rather pronounced serial correlation are common in clinical trials and stock markets. When observations are measured on a continuous scale, the dependency structure

between observations can be modelled in a covariance matrix with non-constant variances and non-zero covariances. There are different types of covariance matrices to model such dependency between observations from the same subject.

On the other hand, when the outcomes are binary such as the positive or negative results to certain tests, we usually consider a generalized linear model (GLM) within the exponential family of sampling distributions with different link functions. In this paper, we propose a marginal logistic model and the parameters are estimated using the first order Generalised Estimating Equation (GEE) approach (Liang & Zeger 1986, Zeger *et al.* 1988 and Hardin & Hilbe 2003). Marginal model is easy to interpret and forecast. While the dependency structure can be explicitly modelled by a working correlation matrix adopting some popular and general correlation structures, the dependency structure of a conditional model is often limited to an auto-regressive type only. Moreover marginal model allows the study of population-averaged effects of covariates on the response variable. Although random effects models allow within subject correlation, it fails to account for the serial correlation over time.

Over the past two decades, GEE approach has been developed and advanced. For example, Zhao and Prentice (1990) and Lipsitz *et al.* (1991) apply the GEE approach to analyse binary data, Kenward *et al.* (1994) consider ordinal data, Lipsitz *et al.* (1994) use categorical data, Prentice and Zhao (1991) focus on multivariate data, Park (1993) compares GEE approach to maximum likelihood approach and Miller *et al.* (1993) to weighted least squares approach, etc. Despite the emergence of new techniques in recent years, for example, the Bayesian and semi-parametric approaches, GEE approach is still popular and often offers an easy alternative. See Horton and Lipsitz (1999) for its implementation in various softwares and the new `geepack` module in R. Recent literature on GEE approach includes Ballinger (2004) for counts and continuous data and Copas and Seaman (2010) on the asymptotic bias using GEE under the missing at random (MAR) dropout environment. Although it has been known that parameter estimates using the GEE approach are robust to misspecification of the true correlation matrix, our experience with real data shows disagreement. Instead, we found that parameter estimates change considerably across models adopting different correlation matrices. In this regard, sensitivity analyses through simulation experiments are often performed to evaluate the per-

formances of different models. However, it is difficult to simulate binary data with a specific working correlation matrix when the actual correlation structure is often unknown in practice. Hence we propose, in this paper, to fit models adopting different working correlation matrices and assess the model-fit using the modified expected utility (Walker & Gutiérrez-Peña, 1999). Working correlation matrix of the best model reveals valuable information regarding the actual dependency structure between observations. Since the length of the time series may affect the choice of robust correlation matrices, we investigate three data sets containing all short time series ($N$=3), all long time series ($N$=14) and time series of varying length ($N$ ranges from 4 to 26).

In section 2, we describe the marginal logistic model and introduce some common working correlation matrices, namely, the completely independent (CI), the equal correlation (EC), the first order autoregressive (AR1), and the Toeplitz with 2 (TOEP2) and 3 (TOEP3) bands. We interpret these matrices in terms of the relationship between observations. Estimation procedures using the first order GEE approach are described in section 3. In section 4, three data sets are analyzed: the blood glucose data consisting of short time series ($N$=3), the plasma citrate concentration data consisting of long time series ($N$=14) and the methadone clinic data consisting of time series of varying length. The first two data sets are obtained from dichotomizing some continuous variables so that the transformed data can be fitted to a logistic regression model. Transformation of data is necessary here to study the robustness of different correlation structures because of the following reasons. Firstly, it has been increasingly popular to transform data to fulfill certain model assumptions. Secondly, simulation studies to evaluate the performance of marginal models adopting different correlation matrices is impossible because it is infeasible to simulate binary data with different working correlation structures. Furthermore, we resist the idea of simulation because arbitrary setting true model parameter values in a simulation experiment fails to reflect the complicated correlation structure in real data. In section 5, results from model fitting using GEE approach are analyzed and compared. Finally, in section 6, a conclusion is drawn on the best correlation structures that give consistently the best fit regardless of the length of time series and strength of autocorrelation.

## 2. Marginal logistic model

Longitudinal binary data are common and logistic regression is often used to model such data. Let $Y_{mn}$ denote the outcome of the $n$-th measurement from subject $m$ where $m = 1, \ldots, M$, $n = 1, \ldots, N_m$, $M$ is the number of subjects and $N_m$ is the number of observations for subject $m$. There are two approaches to handle the serial correlation in $Y_{mn}$. The first approach uses a conditional AR1 model where the previous observation $Y_{m,n-1}$ is entered as a covariate in the mean model of $Y_{mn}$. The conditional probabilities $P_{c,mn} = \Pr(Y_{mn} = 1|Y_{m,n-1})$ are modelled as logit-linear in some covariates, that is,

$$\text{logit}(P_{c,mn}) = \beta_0 + \beta_1\ X_{mn1} + \beta_2\ X_{mn2} + \cdots + \beta_p\ X_{mnp} + \beta_\rho\ Y_{m,n-1}.$$

However such approach limits the dependency structure to be an autoregressive type and the model is incapable of predicting future events for a given set of covariates. The second approach adopts a marginal model and uses a GEE approach with a working correlation matrix $\mathbf{\Sigma}_0$ to describe the dependency structure between observations.

Here the marginal probabilities $P_{mn} = \Pr(Y_{mn} = 1)$ are modelled as logit-linear in some covariates, that is,

$$\text{logit}(P_{mn}) = \eta_{mn} = \beta_0 + \beta_1\ X_{mn1} + \beta_2\ X_{mn2} + \cdots + \beta_p\ X_{mnp}$$

where $\eta_{mn}$ is a linear function of covariates such that $P_{mn} = \dfrac{e^{\eta_{mn}}}{1 + e^{\eta_{mn}}}$. The association across time of the repeated outcomes for a subject is treated as nuisance and is entered only in a working correlation matrix that appears in the estimating equations. Although it is known that parameter estimates are robust to misspecification of the true correlation matrix, our experience with data fitting shows disagreement. Discrepancy between certain significant parameter estimates can be as large as 30% among models adopting different working correlation matrices. See the parameter estimates with '*' in Table 2. Moreover as the actual correlation structure is often unknown in practice, we propose to adopt different working correlation matrices to model the dependency structure between observations explicitly and compare between results. The working correlation matrix of the best fitted model reveals valuable information regarding the dependency between observations. For $N = 4$, the working correlation can be modelled by the following $J = 5$ types:

| **Structure** | | **Properties** |
|---|---|---|
| Completely independent (CI) | $\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ | Zero correlation $\text{Corr}(\varepsilon_i, \varepsilon_j) = 0$<br>$p_\rho = 0$ |
| Equal correlation (EC) | $\begin{bmatrix} 1 & \rho_c & \rho_c & \rho_c \\ \rho_c & 1 & \rho_c & \rho_c \\ \rho_c & \rho_c & 1 & \rho_c \\ \rho_c & \rho_c & \rho_c & 1 \end{bmatrix}$ | Constant correlation $\text{Corr}(\varepsilon_i, \varepsilon_j) = \rho_c$<br>$p_\rho = 1$ |
| First-order autoregressive (AR1) | $\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$ | Equal correlation for given time lag $k$<br>$\text{Corr}(\varepsilon_i, \varepsilon_j) = \rho^k,\ \text{abs}(i-j) = k$<br>$p_\rho = 1$ |
| Toeplitz 2 bands (TOEP2) | $\begin{bmatrix} 1 & \rho_1 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 0 \\ 0 & \rho_1 & 1 & \rho_1 \\ 0 & 0 & \rho_1 & 1 \end{bmatrix}$ | Equal correlation for given time lag=1<br>$\text{Corr}(\varepsilon_i, \varepsilon_j) = \rho_1,\ \text{abs}(i-j) = 1$<br>$p_\rho = 1$ |
| Toeplitz 3 bands (TOEP3) | $\begin{bmatrix} 1 & \rho_1 & \rho_2 & 0 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_2 & \rho_1 & 1 & \rho_1 \\ 0 & \rho_2 & \rho_1 & 1 \end{bmatrix}$ | Equal correlation for given time lag=1 & 2<br>$\text{Corr}(\varepsilon_i, \varepsilon_j) = \rho_k,\ \text{abs}(i-j) = k,\ k = 1, 2$<br>$p_\rho = 2$ |

where $p_\rho$ is the number of correlation parameters which is fixed except TOEP ($p_\rho = N - 1$) and the error $\varepsilon_i = Y_i - E(Y_i)$ in general. For a larger $N$, the $J = 5$ types of correlation matrices can be similarly defined. Some of these correlation structures exhibit special properties. For example, EC corresponds to a *random intercept model* and AR1 an *exponential correlation model*. For equally spaced measurement times such that $t_{n+1} - t_n$ is a constant for all $n$, $\sigma_{nn'} = \sigma^2 \rho^{|n-n'|}$ implies that the covariance between a pair of measurements on the same subject decays to zero as the time separation between the measurements increases. AR1, TOEP2 and TOEP3 assign equal correlation to observations of equal time lag. Practically, correlation between observations with large time lag is very low. Hence TOEP is usually set to contain just a few bands and this will not eliminate some highly possible correlation structures. For more details about the interpretation of some of these correlation structures, see Diggle *et al.* (1996).

## 3. Estimation using GEE approach

Prentice (1988) proposed the GEE approach using a Taylor series expansion to approximate the likelihood function. The estimating equation

is

$$U(\boldsymbol{\beta}) = \sum_{m=1}^{M} \frac{\partial \boldsymbol{P}_m^T}{\partial \boldsymbol{\beta}} \, [\mathrm{Cov}^*(\boldsymbol{Y}_m)]^{-1} \, (\boldsymbol{Y}_m - \boldsymbol{P}_m) = \mathbf{0} \tag{1}$$

where

$$\frac{\partial \boldsymbol{P}_m^T}{\partial \boldsymbol{\beta}} = \boldsymbol{X}_m^T \, \mathrm{Diag}(S_{m1}^2, \ldots, S_{mN_m}^2), \tag{2}$$

$\boldsymbol{X}_m$ is the $N_m \times (p+1)$ design matrix for patient $m$ with row vectors $(1, x_{mn1}, \ldots, x_{mnp})$, $\boldsymbol{P}_m = (P_{m1}, \ldots, P_{mN_m})^T$ and $\boldsymbol{Y}_m = (Y_{m1}, \ldots, Y_{mN_m})^T$. We set the working covariance matrix of $\boldsymbol{Y}_m$ to be

$$\mathrm{Cov}^*(\boldsymbol{Y}_m) = \boldsymbol{V}_m = \mathrm{Diag}(S_{m1}, \ldots, S_{mN_m}) \, \boldsymbol{\Sigma}_0 \, \mathrm{Diag}(S_{m1}, \ldots, S_{mN_m}).$$

where $\boldsymbol{\Sigma}_0$ is the working correlation matrix and the variance estimates $S_{mn}^2 = P_{mn}(1 - P_{mn}) = e^{\eta_{mn}}/(1 + e^{\eta_{mn}})^2$. We estimate different correlation coefficients by

$$\widehat{\rho}_c = \frac{1}{\sum_{m=1}^{M} N_m(N_m - 1)/2} \sum_{m=1}^{M} \sum_{n=2}^{N_m} \sum_{n'=1}^{n-1} \frac{(Y_{mn} - P_{mn})(Y_{mn'} - P_{mn'})}{[P_{mn} \, (1 - P_{mn}) \, P_{mn'} \, (1 - P_{mn'})]^{\frac{1}{2}}} \quad \text{for EC,} \tag{3}$$

$$\widehat{\rho},\ \widehat{\rho}_1 = \frac{1}{N - M} \sum_{m=1}^{M} \sum_{n=1}^{N_m - 1} \frac{(Y_{mn} - P_{mn})(Y_{m\,n+1} - P_{m\,n+1})}{[P_{mn} \, (1 - P_{mn}) \, P_{m\,n+1} \, (1 - P_{m\,n+1})]^{\frac{1}{2}}} \quad \text{for AR1 \& TOEP2,3,} \tag{4}$$

$$\widehat{\rho}_2 = \frac{1}{N - 2M} \sum_{m=1}^{M} \sum_{n=1}^{N_m - 2} \frac{(Y_{mn} - P_{mn})(Y_{m\,n+2} - P_{m\,n+2})}{[P_{mn} \, (1 - P_{mn}) \, P_{m\,n+2} \, (1 - P_{m\,n+2})]^{\frac{1}{2}}} \quad \text{for TOEP3.} \tag{5}$$

since $\rho_c$, $\rho_1$, $\rho_2$ are correlations between any pairs, only lag-1 pairs and only lag-2 pairs of observations respectively and they are estimated by correlations based on corresponding pairs in the sample. Note that $\rho$ is correlation between any pairs allowing for the number of lag and it is estimated by sample correlation between successive (lag-1) pairs only. With estimates $\rho_c^{(l)}$, $\rho^{(l)}$, $\rho_1^{(l)}$, $\rho_2^{(l)}$ in iteration $l$, we can update $\boldsymbol{\beta}^{(l)}$ to $\boldsymbol{\beta}^{(l+1)}$ by solving (1) using the Newton-Raphson method. Then $\rho_c^{(l)}, \rho^{(l)}, \rho_1^{(l)}, \rho_2^{(l)}$ are subsequently updated to $\rho_c^{(l+1)}$, $\rho^{(l+1)}$, $\rho_1^{(l+1)}$, $\rho_2^{(l+1)}$ using $\boldsymbol{\beta}^{(l+1)}$ in (3) to (5) and the cycle repeats again until convergence is reached. Finally, the covariance matrix

of the proposed estimator $\widehat{\boldsymbol{\beta}}$ (Prentice, 1988) equals

$$\begin{aligned}\mathrm{Cov}(\widehat{\boldsymbol{\beta}}) &= \Big[\sum_{m=1}^{M}\Big(\frac{\partial \boldsymbol{P}_m^T}{\partial \boldsymbol{\beta}}\boldsymbol{V}_m^{-1}\frac{\partial \boldsymbol{P}_m}{\partial \boldsymbol{\beta}^T}\Big)\Big]^{-1}\Big[\sum_{m=1}^{M}\Big(\frac{\partial \boldsymbol{P}_m^T}{\partial \boldsymbol{\beta}}\boldsymbol{V}_m^{-1}\widehat{\mathrm{Cov}}(\boldsymbol{Y}_m)\boldsymbol{V}_m^{-1}\frac{\partial \boldsymbol{P}_m}{\partial \boldsymbol{\beta}^T}\Big)\Big] \\ &\quad \Big[\sum_{m=1}^{M}\Big(\frac{\partial \boldsymbol{P}_m^T}{\partial \boldsymbol{\beta}}\boldsymbol{V}_m^{-1}\frac{\partial \boldsymbol{P}_m}{\partial \boldsymbol{\beta}^T}\Big)\Big]^{-1} \qquad (6)\end{aligned}$$

where $\widehat{\mathrm{Cov}}(\boldsymbol{Y}_m)$ is given by $(\boldsymbol{Y}_m - \hat{\boldsymbol{P}}_m)(\boldsymbol{Y}_m - \hat{\boldsymbol{P}}_m)^T$.

## 4. Empirical studies

### 4.1 Blood glucose data

Data of inter and intra individual variation of blood glucose levels (Andrews and Herzberg 1985, P.211) obtained from registrants for pre-natal care at Boston City Hospital, USA, are used to estimate the variation of blood sugars on pregnant and non-pregnant women. There are 53 non-pregnant women, each receiving annual glucose tolerance test over a period of six years of which six fasting blood glucose tests were conducted and their one hour post blood glucose concentrations were measured. There are also 52 pregnant women, each having three fasting blood glucose tests and their one hour post blood glucose concentrations measurements. Measurements are in mg/100 ml. The variables in the data set are

Dependent variable:
$Y$: the indicator of whether the fasting glucose level $X_1$ is less than the 1-hour post blood glucose level $Y_1$ by more than 10 mg/100 ml, i.e. $I(X_1 - Y_1 < 10)$
Independent variables:
$X_1$: the fasting blood glucose level,
$X_2$: the indicator of pregnancy.

This data set is used in the analysis of short time series and it contains $K = 315$ $(3 \times (53 + 52))$ observations coming from $M = 105$ subjects each repeatedly measured $N = 3$ times. For the 53 non-pregnant women, only the first three fasting and one hour post blood glucose levels are used in the analysis. The counts of $M$ and $K$ and the means of $Y_1$, $X_1$ and $Y$ across the non-pregnant and pregnant groups of women are presented in Table 1.

Table 1: Summary of the three data sets

| | Blood glucose data | | | Plasma citrate con. data | | | Methadone data | | |
|---|---|---|---|---|---|---|---|---|---|
| Var. | Non-preg. | Pregnant | Total | Non-meal | Meal | Total | Non-drop | Drop | Total |
| $M$ | 53 | 52 | 105 | - | - | 10 | 85 | 51 | 136 |
| $K$ | 159 | 156 | 315 | 110 | 30 | 140 | 2210 | 652 | 2872 |
| $Y_1$ | 87.95 | 107.82 | 97.79 | 115.10 | 120.51 | 119.35 | - | - | - |
| $X_1$ | 79.21 | 72.88 | 76.08 | - | - | - | 64.1 | 65.3 | 64.4 |
| $Y$ | 43% | 84% | 63% | 37% | 47% | 45% | 13% | 26% | 16% |

Note: '-' refers to information either not available or not necessary to report.

The marginal probabilities $P_{mn} = \Pr(Y_{mn} = 1)$ are modelled as logit-linear in the covariates, that is,

$$\text{logit}(P_{mn}) = \beta_0 + \beta_1 X_{mn1} + \beta_2 X_{mn2}$$

where $(\beta_0, \beta_1, \beta_2)$ are the parameters for the intercept, $X_1$ and $X_2$ respectively. For model comparison, the conditional AR1 model (CAR1)

$$\text{logit}(P_{c,mn}) = \beta_0 + \beta_1 X_{mn1} + \beta_2 X_{mn2} + \beta_\rho Y_{m\,n-1},$$

where the conditional probabilities $P_{c,mn} = \Pr(Y_{mn} = 1|Y_{m,n-1})$ and $Y_{m,0} = 0$, is also considered together with the marginal models with 5 different working correlation matrices. Table 2 shows that all parameters are significant and that a decrease in fasting blood glucose level and the state of pregnancy are significantly associated with a higher probability of increasing the one-hour post blood glucose level by more than 10 mg/100 ml.

Table 2: Parameter estimates with SE in italic for the three data sets

| | $\beta_0$ | | $\beta_1$ | | $\beta_2$ | | $\rho$ or $\beta_\rho$ | | $U$ |
|---|---|---|---|---|---|---|---|---|---|
| Blood glucose data ($N = 3$) | | | | | | | | | |
| CAR1 | 1.9004 | *1.0310* | -0.0285 | *0.0131* | 1.7001 | *0.2885* | 0.2830 | *0.2804* | -0.55521 |
| CI | 1.7774 | *1.0744* | -0.0259 | *0.0136* | 1.7821 | *0.2993* | - | | -0.55682 |
| EC | 2.4270 | *1.1060* | -0.0341 | *0.0141* | 1.7444 | *0.3020* | 0.1388 | | -0.55747 |
| AR1 | 2.1176 | *1.0802* | -0.0301 | *0.0137* | 1.7595 | *0.3009* | 0.0986 | | -0.55700 |
| TOEP2 | 2.1027 | *1.0794* | -0.0299 | *0.0137* | 1.7602 | *0.3009* | 0.0983 | | -0.55699 |
| TOEP3 | 2.4388 | *1.1235* | -0.0343 | *0.0143* | 1.7467 | *0.3022* | 0.1040 | 0.2149 | -0.55751 |
| Plasma citrate concentration data ($N = 14$) | | | | | | | | | |
| CAR1 | -0.3413 | *0.5101* | -0.1176 | *0.0554* | -0.5981 | *0.5344* | 2.6297 | *0.4311* | -0.50189 |
| CI | 0.6953 | *0.6216* | -0.0999 | *0.0603* | -0.7274 | *0.2902* | - | | -0.66648 |
| EC | 0.6700 | *0.6370* | -0.1000 | *0.0624* | -0.7286* | *0.2960* | 0.4324 | | -0.66656 |
| AR1 | 0.4359 | *0.5624* | -0.0719 | *0.0581* | -0.5555* | *0.2774* | 0.5967 | | -0.66803 |
| TOEP3 | 0.4511 | *0.6872* | -0.0902 | *0.0858* | -0.2419 | *0.2776* | 0.6046 | 0.4909 | -0.67134 |
| Methadone clinic data ($4 \leq N_m \leq 26$) | | | | | | | | | |
| CAR1 | -0.8423 | *0.2189* | -0.0088 | *0.00282* | -0.4049 | *0.0628* | 2.3960 | *0.1196* | -0.73692 |
| CI | -0.1332 | *0.4011* | -0.0113 | *0.00620* | -0.3710 | *0.0765* | - | | -0.78251 |
| EC | -0.1892 | *0.3795* | -0.0118 | *0.00517* | -0.2911 | *0.0771* | 0.2360 | | -0.67457 |
| AR1 | -0.2197 | *0.3768* | -0.0108 | *0.00569* | -0.3434 | *0.0736* | 0.4310 | | -0.74800 |
| TOEP2 | -0.2614 | *0.4017* | -0.0107 | *0.00605* | -0.3282 | *0.0745* | 0.4424 | | -0.72970 |
| TOEP3 | -0.1900 | *0.3636* | -0.0104 | *0.00549* | -0.3683 | *0.0752* | 0.4422 | 0.3533 | -0.77870 |

Note: The larger parameter estimate $\beta_2$ with '*' is 30% more than the smaller parameter estimate for the plasma citrate concentration data. '-' refers to parameter not existing in the model.

### 4.2 Plasma Citrate Concentration data

An experiment involving 10 subjects (Andrews and Herzberg, 1985, P.237) was carried out to study the variation of plasma citrate concentration during a day. For each subject, the concentration of citrate in plasma (in $\mu$mol per litre) was measured hourly at 14 time points from 8am to 9pm during a day, with a total of $K = 140$ observations. Meals were given at 8am, at noon and at 5pm. The binary outcome of the logistic model is $Y$, an indicator of whether the plasma citrate concentration ($Y_1$) is more than 120 $\mu$mol per litre. Covariates include the time of the day ($X_1$) taking values from 1 to 14 and an indicator of meal time ($X_2$). A summary for the counts $M$ and $K$ and the means of $Y_1$ and $Y$ across observations taken at non-meal and meal times are reported in Table 1. Model parameters are $(\beta_0, \beta_1, \beta_2)$ for the intercept and the 2 covariates.

Again the CAR1 model with an autoregressive parameter $\beta_\rho$ for the previous outcome $Y_{m\,n-1}$ and the marginal models with 5 different working correlation matrices are fitted to the data. However model with TOEP2 does not have stable parameter estimates. Table 2 shows that for all except the CAR1 models, only the indicator of meal time ($X_2$) is significantly associated with a higher probability of having the concentration of citrate in plasma in excess of 120 $\mu$mol per litre.

### 4.3 Methadone clinic data

The methadone clinic data set consists of records of drug users under methadone maintenance treatment program at a clinic in Western Sydney in 1986. Outcomes $Y$ are the weekly urine test results which are positive or negative for morphine, a biological marker for heroin use. Methadone dose $d$ in mg ($X_1$) at the time of urine test is included as a predictor variable, as is the log of treatment duration $\ln t$ in weeks ($X_2$) because previous research revealed that the methadone dosage and the duration of treatment are significant treatment factors. There were $M = 136$ heroin users, submitting a total of $K = 2872$ urine screens and each urine screen result serves as the unit of analysis. The average number of treatment weeks per heroin user is 21.1 with each user submitting 4 to 26 weekly outcomes ($4 \leq N_m \leq 26$). 51 of them dropped out prematurely and the rest having 26 outcomes are regarded as having completed the program. Table 1 displays the counts of $M$ and $K$ and the means of $Y$ and $X_1$ across the non-dropout and dropout heroin user groups. For more detailed description of the data set, see Chan *et al.* (1998).

The marginal logistic regression model is:

$$\text{logit}(P_{mn}) = \beta_0 + \beta_d \; d_{mn} + \beta_t \;\; \ln n$$

where $d_{mn}$ is the dosage administered to patient $m$ at time $n$. Table 2 gives the results for the CAR1 model and 5 marginal models. In general, the dose effect is only marginally significant. The conclusion is qualitatively the same as other researches conducted on this data (Chan *et al.*, 1998), namely, increase in methadone dose is associated with reduced heroin use and heroin use is also reduced over time. Moreover the significant autoregressive parameter $\beta_\rho$ in the CAR1 model indicates that a strong and positive association between the present $Y_{it}$ and previous $Y_{i\,t-1}$ outcomes, suggesting that some patients in treatment tend to use heroin continuously while others do not.

## 5. Results

In order to compare the performance of marginal logistic models with different correlation matrices, the posterior expected utility

$$U = \frac{1}{K} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \ln \widehat{P}_{mn} \tag{7}$$

where $\widehat{P}_{mn} = \frac{e^{\hat{\eta}_{mn}}}{1+e^{\hat{\eta}_{mn}}}$ and

$$\hat{\eta}_{mn} = \begin{cases} \hat{\beta}_o + \hat{\beta}_1 \; X_{mn1} + \hat{\beta}_2 \; X_{mn2} & \text{for marginal models,} \\ \hat{\beta}_o + \hat{\beta}_1 \; X_{mn1} + \hat{\beta}_2 \; X_{mn2} + \hat{\beta}_\rho \; Y_{m\,n-1} & \text{for CAR1 model,} \end{cases}$$

proposed by Walker and Gutiérrez-Peña (1999) is evaluated for each model.

The $U$ is the average of the natural logarithm of the probability densities of data. Obviously, a less negative $U$ indicates a higher likelihood of the model given the data. Note

that the nuisance parameters in the working correlation matrix of the marginal models do not enter explicitly into the calculation of $\widehat{P}_{mn}$. However, as they appear in the estimating equations, they affect the parameter estimates $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$.

For the blood glucose data, Table 2 shows that the estimates of $\rho$ are very small for all marginal models and the autoregressive parameter $\beta_\rho$ in the CAR1 model is also insignificant. This shows that the strength of serial correlation is weak possibly because a short time series cannot adequately reveal the dependency structure for binary data. Hence the $U$ values are similar across conditional and marginal models. Marginal model adopting a CI correlation matrix has the least negative $U$ value showing that observations within patients are essentially independent.

For the plasma citrate concentration data, the $\rho$(s) are much larger for all marginal models and the autoregressive parameter $\beta_\rho$ in the CAR1 model is strongly significant. This shows that the strength of serial correlation is much stronger. Clearly the $\rho_c$ in the correlation matrix of EC type is generally smaller than the $\rho$ in the AR1 matrix or the $\rho_1$ in the TOEP2 and TOEP3 matrices because EC assumes equal correlation independent of the time-lag between any two observations. Moreover, $\rho_1$ in the TOEP3 matrix which describes the correlation between observations of 1 time-lag is generally larger than $\rho_2$ which reveals the correlation between observations of 2 time-lag. Since the CAR1 model gives the least negative $U$ value, it again confirms the strong serial correlation between observations within patients. However, across marginal models, the $U$ value is least negative for model with CI correlation matrix but this value is very close to that of model with EC matrix. For this data, we believe that the EC matrix is the best correlation matrix to describe the dependency structure between observations from the same subject.

Lastly, for the methadone clinic data, Table 2 shows that the $\rho$(s) are moderate for all marginal models and the autoregressive parameter $\beta_\rho$ in the CAR1 model is also significant. The $U$ values differ substantially across models possibly because of the generally longer time series and their specific dependency structures. The marginal model adopting an EC correlation matrix gives the least negative $U$ value again confirming the strong serial correlation between observations within patients but such serial correlation does not decrease as the time-lag between observations increases. Again the EC matrix is the best correlation matrix for this data to describe the dependency structure between observations from the same patient.

## 6. Conclusion

Although the GEE approach has been proposed for more than twenty years, few studies investigate choices of working correlation matrices for data with different correlation structures, possibly because parameter estimates for marginal models using a GEE approach are known to be robust against misspecification of working correlation matrix. However our experience with real data analyses shows the contrary: parameter estimates $(\beta_0, \beta_1, \beta_2)$ as well as the model fit measure $U$ differ substantially across models adopting different working correlation matrices. Results from three real data analyses suggest that the EC matrix is an appropriate choice of working correlation matrix if the time series is

short or the serial correlation does not decrease sharply with time. This is perhaps due to the limited information in binary data that favors simple correlation structures. If computationally feasible, fitting models with all the five working correlation matrices and choosing one with the least negative posterior expected utility $U$ is also a good practice.

## References


Andrews, D.F., Herzberg, A.M. (1985). *Data*. Springer-Verlag, New York.

Ballinger, G.A. (2004) Using Generalized Estimating Equations for Longitudinal Data Analysis. *Organizational Research Methods* **7**, 127-150.

Chan, J.S.K., Kuk, A.Y.C., Bell, J., McGilchrist, C. (1998) The analysis of methadone clinic data using marginal and conditional logistic models with mixture or random effects. *Australian and New Zealand Journal of Statistics* **40**, 1-40.

Copas, A.J., Seaman, S.R. (2010) Bias from the use of generalized estimating equations to analyze incomplete longitudinal binary data. *Journal of Applied Statistics* **37**, 911-922.

Diggle, P.J ., Liang, K. Y., Zeger, S.L. (1996). *Analysis of Longitudinal Data*. Oxford Science Publications.

Hardin, J., Hilbe, J. (2003). *Generalized Estimating Equations*. London: Chapman and Hall.

Horton, N.J., Lipsitz, S.R. (1999) Review of software to fit generalized estimating equation regression models. *The American Statistician* **53**, 160-169.

Kenward, M.G., Lesaffre, E., Molenberghs, G. (1994) Application of maximum likelihood and generalized estimating equatins to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* **50**, 945-953.

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

Lipsitz, S.R., Kim, K., Zhao, L.P. (1994) Analysis of repeated categorical data using generalized estimating equations, *Statistics in Medicine* **13**, 1149-1163.

Lipsitz, S.R., Laird, N.M., Harrington, D.P. (1991) Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association, *Biometrika* **78**, 153-160.

Miller, M.E., Davis, S.C. and Landis, J.R. (1993) The analysis of longitudinal polytomous data: Generalized estimating equations and connections with weighted least squares, *Biometrics* **49**, 1033-1044.

Park, T. (1993) A comparison of the generalizing estimating equation approach with the maximum likelihood approach for repeated measurements. *Statistics in Medicine* **12**, 1723-1732.

Prentice, R.L. (1988) Correlated binary regression with covariates specific to each binary observation. *Biometrics* **44**, 1033-1048.

Prentice, R.L., Zhao, L.P. (1991) Estimating equations for parameters in means and covariance of multivariate dicrete and continuous responses, *Biometrics* **47**, 825-839.

Walker, S. G., Gutiérrez-Peña, E. (1999). Robustifying Bayesian Procedures. *Bayesian Statistics* **6**, 685-710.

Zeger, S.L., Liang, K.Y. Liang, Albert, P.S. (1988) Models for Longitudinal Data: A Generalized Estimating Equation Approach, *Biometrics* **44**, 1049- 1060.

Zhao, L.P., Prentice, R.L. (1990) Correlated binary regression using a generalized quadratic model. *Wiley Interdisciplinary Reviews: Computational Statistics* **77**, 642-648.

Received March 14, 2013; accepted August 15, 2013.

Jennifer S.K. Chan
School of Mathematics and Statistics
The University of Sydney
NSW 2006, Australia
jchan@maths.usyd.edu.au