Title
A Drug-Dependence Treatment Medication Analysis based on Longitudinal Data with Missing Values using Multiple-Imputation Generalized Estimating Equations

Permalink
https://escholarship.org/uc/item/94k5p5mx

Author
Yi, Yi

Publication Date
2014-01-01

Peer reviewed|Thesis/dissertation
A Drug-Dependence Treatment Medication Analysis based on Longitudinal Data with Missing Values using Multiple-Imputation Generalized Estimating Equations

A thesis submitted in partial satisfaction of the requirements for the degree Master of Science in Statistics

by

Yi Yi

2014
ABSTRACT OF THE THESIS

A Drug-Dependence Treatment Medication Analysis based on Longitudinal Data with Missing Values using Multiple-Imputation Generalized Estimating Equations

by

Yi Yi

Master of Science in Statistics
University of California, Los Angeles, 2014
Professor Yingnian Wu, Chair

Repeated-Measures longitudinal data is common in drug research, where every patient is repeatedly measured across time. Responses could either be continuous variables such as blood pressure or binary variables such as drug test positive/negative. Issues to be addressed are within-subject observation dependence, as well as the between-subject differences (mixed effects). Another important problem to address is the missingness and dropouts.

Full likelihood-based models such as the generalized linear mixed model (GLMM) together with EM algorithm could be utilized, given simplified parametric correlation structure between random components. If the interest is only the mean parameters, little in subject effects, the non-likelihood-based generalized estimating equations (GEE) is a good alternative. GEE circumvents the structural and computational complexities of likelihood-based models, and it is robust to misspecification of the working correlation structures of marginal observations. However, as a non-likelihood frequentist marginal model, GEE itself lacks strength dealing with missing data mechanism beyond missing completely at random (MCAR).
Therefore, integration of multiple imputation and GEE (MI-GEE) is a great solution to longitudinal data with dropouts.

Real data application is performed on MI-GEE and GLMM. Our results for the bupropion study dataset show effective but not significant strength of Bupropion for treating methamphetamine dependence, which is consistent with previous studies and biological sense.
The thesis of Yi Yi is approved.

Rick Schoenberg

Hongquan Xu

Yingnian Wu, Committee Chair

University of California, Los Angeles

2014
TABLE OF CONTENTS

1 Introduction .................................................. 1

2 Modeling ....................................................... 4
  2.1 Generalized Estimating Equation (GEE) ....................... 4
      2.1.1 Intuition from Linear Models in Gaussian Framework .... 5
      2.1.2 Overdispersed Exponential Family ........................ 6
      2.1.3 The Estimating Equations ................................. 6
      2.1.4 Asymptotics of GEE ...................................... 8
      2.1.5 Bias of using GEE under MCAR when the real model is under MAR .............................................. 9
  2.2 Multiple Imputation (MI) .................................... 10
  2.3 Generalized Estimating Equation with Multiple Imputation (MI-GEE) ................................................ 11
  2.4 Generalized Linear Mixed Model (GLMM) ...................... 11

3 Bupropion Data Analysis ......................................... 13
  3.2 Exploratory Data Visualization ............................... 15
  3.3 GEE ......................................................... 21
      3.3.1 GEE on One Imputation .................................. 22
      3.3.2 MI-GEE ................................................. 25
      3.3.3 Exclusion of Missing Data ................................. 26
  3.4 GLMM ....................................................... 28
LIST OF FIGURES

3.1 Proportion of Methamphetamine test positives against Week, grouped by Treatment and Week ........................................ 15

3.2 Proportion of Methamphetamine test positives against Week, grouped by Treatment, Week, and Gender ............................. 16

3.3 Proportion of Methamphetamine test positives against MA_{30D}, grouped by Subject ID, denoted by Treatment ...................... 17

3.4 Proportion of Methamphetamine test positives against MA_{30D}, grouped by MA_{30D} and Treatment, denoted by Number of Methamphetamine test positives .................................................. 18

3.5 Proportion of Methamphetamine test positives against MA_{30D}, grouped by Week and MA_{30D}, denoted by Number of Methamphetamine test positives .................................................. 19

3.6 Proportion of Methamphetamine test positives against MA_{30D}, grouped by Week, MA_{30D} and Treatment, denoted by Number of Methamphetamine test positives .................................................. 20

3.7 Number of missing observations grouped by subject .................. 21

3.8 GEE fitted test positive probabilities for each observation against Time. Missing data imputed only once by subject historical proportion of Methamphetamine test positives ........................................ 24

3.9 GEE Fitted test positive probabilities for each observation against Time. (Complete data): all missing slots were deleted. (Imputed): missing data imputed only once by subject historical proportion of Methamphetamine test positives (Table 3.3) ........................................ 26
3.10 GEE fitted number of cigarettes for each observation against Time (days), grouped by Time. (Complete data): all missing slots were deleted ................................. 31

3.11 GEE fitted HRSD depression scores for each observation against Time (weeks), grouped by Time. (Complete data): all missing slots were deleted ................................. 33
**List of Tables**

3.1 Demographics of 84 participants (Heinzerling, [2]). The p-value shows the dependence between Bupropion/Placebo assignment and demographical variables. It is best to assign treatments so that the treatment is not confounded with any observable demographical variables. ................................................................. 14

3.2 Variable coding. Subtracting the mean from Time and MA_30D is intended to make the intercept and Treatment:Bupropion coefficient in the model represent their effects where Time and MA_30D are controlled at an average level, instead of 0, for easier interpretation. ................................................................. 21

3.3 GEE based on one-time imputation ................................................. 23

3.4 GEE with one imputation and MI-GEE ............................................. 25

3.5 Sensitivity Analysis. This model and the model in table 3.3 are basically consistent with each other, as we would draw the similar conclusions based on these two models. Plus the robust nature against correlation structure misspecification of GEE, we could say this modeling process is not sensitive to inclusion or exclusion of missing values. ................................................................. 27

3.6 GLMM under the data with one-time imputation .............................. 29

3.7 GEE on cigarette use with all missing data deleted .......................... 30

3.8 GEE on HRSD depression scores with all missing data deleted ....... 32
ACKNOWLEDGMENTS

This is part of the work published in the bupropion study paper (Heinzerling, [2]).
Thanks for Dr. Keith Heinzerling’s study design and collaboration on this work.
CHAPTER 1

Introduction

Bupropion is an old medication for treating methamphetamine dependence. Previous randomized trials found an effect of Bupropion in reducing methamphetamine use. The study in (Heinzerling, [2]) aims to replicate these effects.

This dataset consists of 84 subjects, case-control assignment is well adjusted to best randomize it with all covariates, and correlation between covariates are also controlled by screening participants from a pool of total 294 participants. Each subject comes in the lab to take samples on a scheduled basis for 12 weeks. The response is binary variable methamphetamine test positive or negative across time. Missing values or dropouts are common. The format of this dataset is suitable for mixed-effect model or non-likelihood based estimating equations. Previous analyses on Bupropion do not have a strong modeling scheme behind the analysis. We would like to do a more adequate analysis on this dataset.

Likelihood-based model is most common for regression analysis, from the ordinary least square to generalized linear model. For data with subject effects, mixed models (Laird, [5]) are widely used. As the generalization of gaussian linear models, generalized linear mixed models (GLMM) extends the distribution to a non-gaussian framework (see Stiratelli [13], Williams, [14]).

Besides the advantages of likelihood-based models such as retrieving full informa-
tion and flexibility dealing with missing data, their complicated structures might end up causing intensive computations, and there is no omnipotent approximating method suitable for every situation. This motivates us to use a model that is easy to deal with and fulfills our goal, say, only the coefficients in regression are of interest, especially in our Bupropion study.

Generalized Estimating Equation (GEE, Liang, [6]) is a non-likelihood based model which aims to find the true coefficients and treat correlation of within-subject responses as a nuisance factor. The idea of GEE is to use marginal distributions of every observation to simplify the their joint distribution by focusing only on marginals and the correlation between marginals, just like in a “faked” multivariate gaussian framework, where correlation matrix plus marginals determines the joint distribution. However, GEE is a frequentist model, which can not deal with complicated missing mechanism. Its strict rule requires the missingness to be completely independent of both observed and unobserved data a.k.a missing completely at random (MCAR, Rubin, [10]). Most problems in longitudinal data analysis will not guarantee such a missingness mechanism. We would like to build models on missing at random (MAR) which means missingness is only required to be independent of unobserved data given observed data. To served the purpose, two improved models were proposed, weight generalized estimating equations (WGEE, Robins, [3]) and multiple-imputation generalized estimating equations (MI-GEE). WGEE gives the inverse probability weight (e.g. the probability a participant drops out when he/she drops out). The lower chance an event happens, the higher credibility we should give it. However, MI-GEE is a combination of multiple imputation (MI, Rubin, [12]) and GEE, which serves our purpose under the weaker assumption MAR and gives a convenient way of combining multiple imputed modeling results.
In our study, we mainly use GEE with one imputation and MI-GEE for data modeling. Chapter 2 basically illustrates the models GLMM, GEE, MI, MI-GEE. Chapter 3 is application of GEE on our Bupropion dataset, our conclusion from the analysis shows effects of bupropion treating methamphetamine dependence, but they are not significant, which is consistent with other studies.
CHAPTER 2

Modeling

In 1986, GEE was proposed for regression problems where dependency of responses is of less interest than sizes of predictor coefficients. However, if there is missing data, it requires the strong assumption that missingness is independent of both unobserved and observed data, which isn’t the case for our longitudinal data, because we want to believe the dropouts should be related to previous observations (e.g., bad treatment effects may cause participants lose the enthusiasm to stay in the study). To improve the applicability of GEE, multiple imputation (MI) was later incorporated into GEE, which works under weaker assumption that missingness of data is independent of unobserved data given observed data. Section 2.1 generally describes the GEE model scheme; section 2.2 introduces the multiple imputation; section 2.3 briefly introduces how to combine MI with GEE; section 2.4 discusses the general model scheme for generalized linear mixed model (GLMM), which is another option to deal with our data and used in comparison with GEE in the data analysis section.

2.1 Generalized Estimating Equation (GEE)

GEE (Liang, [6]) is motivated when the modeling interest is only in the regression coefficients, instead of the within-in subject dependence. In our case, it is the main Bupropion effect that is of great interest. Although, the within-in subject dependence is also very important in some other cases, e.g., the correlation of the drug effect across the time for each subject, which is not explained by a...
fixed regression coefficient $\beta \ast t$. Therefore, “less” detailed and accurate specification of the dependence is chosen for the goal of estimating regression coefficients.

### 2.1.1 Intuition from Linear Models in Gaussian Framework

Intuition from linear regression: in the ordinary linear models with Gaussian errors, we have observations $y_{n \times 1}$ and predictors $X_{n \times p}$. With the true coefficient $\beta$, the underlying model is assumed to be $y|X \sim N(X\beta, \sigma^2 I)$. The maximum likelihood estimator for the $\beta$ part ends up as the solution of: $\min_\beta (y - X\beta)^T(y - X\beta)$. By taking derivative w.r.t $\beta$ we have $X^T(y - X\beta) = 0$, through which we have the general solution of linear regression $\beta = (X^TX)^{-1}X^Ty$. To generalize the model in Gaussian case a little bit, we assume all of the observations are correlated, then the true underlying model is $y|X \sim N(X\beta, \Sigma)$, where $\Sigma$ does not contain $\beta$. We use $y_{n \times 1}$ as the density notations as well as the observations.

$$f(y|X) = \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}exp\left(-\frac{1}{2}(y - X\beta)^T\Sigma^{-1}(y - X\beta)\right)$$

The log likelihood of observation $y$ is:

$$l(\beta, \Sigma) = -\frac{1}{2}(y - X\beta)^T\Sigma^{-1}(y - X\beta) - \frac{1}{2}\log(|\Sigma|) + C$$

Estimating equations are $\frac{\partial l}{\partial \beta} = 0$, $\frac{\partial l}{\partial \Sigma} = 0$:

$$\begin{cases}
\frac{\partial l}{\partial \beta} = 0 \Rightarrow X^T\Sigma^{-1}(y - X\beta) = 0 \\
\frac{\partial l}{\partial \Sigma} = 0
\end{cases} \quad (2.1)$$

Different specification of covariance structure $\Sigma$ will result in different solutions. Independence is one special case where $\Sigma = \sigma^2 I$

$$\begin{cases}
\frac{\partial l}{\partial \beta} = 0 \Rightarrow X^T(y - X\beta) = 0 \\
\frac{\partial l}{\partial \sigma^2} = 0 \Rightarrow \frac{\partial l}{\partial \sigma^2} = 0 \Rightarrow \frac{(y - X\beta)^T(y - X\beta)}{(\sigma^2)^2} + \log(\sigma^2) = 0
\end{cases} \quad (2.2)$$

The estimating equation in Equation 2.1 is similar to the one in GEE. When a multivariate distribution is normal, their dependence structure is all contained in
the correlation matrix, whereas this is not the case for other general distributions. GEE takes advantage of the estimating equation in Gaussian assumption, and extend it to an estimator for more general distribution forms.

2.1.2 Overdispersed Exponential Family

The natural form of overdispersed exponential family:

\[ f(y) = h(y, \tau) \exp \left( \frac{\eta^T T(y) - b(\eta)}{\tau} \right) \]  
\hspace{1cm} (2.3)

Their marginal distribution (for every \( y_k \)) is of the form:

\[ f(y) = h(y, \tau) \exp \left( \frac{\eta^T T(y) - b(\eta)}{\tau} \right) = h(y, \tau) \exp \left( \frac{\theta y - a(\theta) + \gamma^T T_2(y) - c(\gamma)}{\tau} \right) \]  
\hspace{1cm} (2.4)

\( T(y) = [y, T_2(y)^T]^T \) could be a vector of any dimension, each element is a function of \( y \). \( \eta = [\theta, \gamma]^T \), \( b(\eta) = a(\theta) + c(\gamma) \). The estimating equation is interested in using \( \exp\left(\frac{\theta y - a(\theta)}{\tau}\right) \) as a part of the marginal likelihood.

\[ E(Y) = a'(\theta), \quad Var(Y) = a''(\theta)/\tau \]  
\hspace{1cm} (2.5)

2.1.3 The Estimating Equations

In our case, \( y \) is \( n \)-dimensional, and GEE is based on the assumption that marginal distributions of \( f(y_1, ..., y_n) \) follow the form of (2.4). For regression purpose, we include a suitable link function \( s(\mu) = \zeta \), scalar \( \zeta = x\beta \). From the overdispersed exponential family, we have scalar \( \mu = a'(\theta) = s^{-1}(\zeta) \), where scalar \( \theta = (a')^{-1}(a'(\theta)) = (a')^{-1}(\mu) = (a')^{-1}(s^{-1}(\zeta)) = g(\zeta) \), \( x_{1\times p} \) is the predictors for each observation. Denote \( \Theta = [\theta_1, ..., \theta_n]^T = [g(\zeta_1), ..., g(\zeta_n)]^T \), \( Z = [\zeta_1, ..., \zeta_n]^T \).
Let $G_{n \times n}'$ be a diagonal matrix, define

$$G'(Z)_{i,j} = \begin{cases} g'(z_i) & i = j \\ 0 & \text{o.w.} \end{cases} \Rightarrow G'(X\beta)_{i,j} = \begin{cases} g'(X_{(i,\cdot)}\beta) & i = j \\ 0 & \text{o.w.} \end{cases}$$

Similar notations apply to $G$. $X_{(i,\cdot)}$ is the $i$-th row of $X$.

Let $A''_{n \times n}$ be a diagonal matrix, define

$$A''(\Theta)_{i,j} = \begin{cases} a''(\theta_i) & i = j \\ 0 & \text{o.w.} \end{cases} \Rightarrow [A''(G(X\beta) \star 1)]_{i,j} = \begin{cases} a''(g(X_{(i,\cdot)}\beta)) & i = j \\ 0 & \text{o.w.} \end{cases}$$

Similar notations apply to $A'$ and $A$. Diagonal of $A'$ represents the mean of every observation, as the property of overdispersed exponential family.

Let $R(\alpha)$ represent the correlation structure of $n$ observations fully characterized by a vector of unknown parameters $\alpha$. The variance-covariance $V_{n \times n}$ between them would be expressed as:

$$V = A'' R(\alpha) A''$$

Suppose we try to minimize objective similar to the Gaussian case, the problem would be to

$$\min_{\beta} p = (y - A'(G(X\beta)1)1)^TV^{-1}(y - A'(G(X\beta)1)1)$$

$$\downarrow$$

$$\frac{\partial p}{\partial \beta} = 0 \Rightarrow X^TA''(G(X\beta)1)G'(X\beta) V^{-1} (y - A'(G(X\beta)1)1) = 0$$

For multiple subjects, we have $(y_k)_{n_k \times 1}$ and $(X_k)_{n_k \times p}$, $k = 1, \ldots, K$. We might as well assume $n_1 = n_2 = \ldots = n_K = n$ for simplicity. Combine the estimating equations for all subjects, we have the formal GEE formulation:

$$\sum_{k=1}^K \{ X_k^TA''(G(X_k\beta)1)G'(X_k\beta) V^{-1} (y_k - A'(G(X_k\beta)1)1) \} = 0$$
For simplicity, we write it in a shorter form:
\[
\sum_{k=1}^{K} \left\{ X_k^T A'' G' V^{-1} (y_k - \text{diag of } A') \right\} = 0 \tag{2.12}
\]
When \( y \) is multivariate distribution, this estimating equation is exactly the same as MLE estimator. Take a look at Equation (2.10), usually, in the model of Gaussian independent observations with same variance, the marginal distribution for every observation \( y_k \):
\[
f(y|X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left( \frac{\theta y - \frac{\theta^2}{2} - \frac{1}{2}y^2}{\sigma^2} \right) \tag{2.13}
\]
\[a(\theta) = \frac{\theta^2}{2},\]
\[A'(\Theta)_{n \times n} = \text{diag}(\Theta), A''(\Theta)_{n \times n} = I,\]
\[\mu = \frac{\partial(\frac{\theta^2}{2})}{\partial \theta} = \theta, \theta = \mu, \mu = s^{-1}(\zeta) = \zeta = x\beta, \theta = g(\zeta) = \zeta = x\beta,\]
\[\Theta = X\beta, G(Z) = \text{diag}(Z), G'(Z) = I,\]
\[\tau = \sigma^2,\]
\[R = I\]
The estimating equation is
\[
\frac{1}{\sigma^2} \sum_{k=1}^{K} \left\{ X_k^T (y_k - X_k\beta) \right\} = 0 \Rightarrow \sum_{k=1}^{K} \left\{ X_k^T (y_k - X_k\beta) \right\} = 0 \tag{2.14}
\]
This is exactly the objective function of MLE in the i.i.d Gaussian case (all subjects are pulled together as in the assumption that all subjects are independent).

### 2.1.4 Asymptotics of GEE

Asymptotics of GEE (refer to Liang, [6] for more details):
For Equation (2.12): If \( \alpha \) in \( R(\alpha) \) and \( \tau \) is known, based on estimating equations, one could show the asymptotic normality of beta estimated from this set of equations.
\[
E \left( \frac{1}{K} \sum_{k=1}^{K} h(y_k|X_k) \right) = E \left( \frac{1}{K} \sum_{k=1}^{K} \left\{ X_k^T A'' G' V^{-1} (y_k - \text{diag of } A') \right\} \right) = 0 \tag{2.15}
\]
\[
\Sigma_\beta = \text{Var}(h(y_k|X_k)) = (X_k^T A'' G' V^{-1}) \text{Cov}(y_k, y_k) (X_k^T A'' G' V^{-1})^T
\] (2.16)

\[
S_\beta = E\left( \frac{\partial h(y_k|X_k)}{\partial \beta} \right) = (X_k^T A'' G') V^{-1} (X_k^T A'' G')^T
\] (2.17)

\[
\sqrt{K}(\hat{\beta} - \beta) \xrightarrow{K \to \infty} N(0, S_\beta^{-1} \Sigma_\beta S_\beta^{-1})
\] (2.18)

However, \( R(\alpha) \) is exactly something we need to assume and estimate, this corresponds to several common working correlation assumption, “Independence”, “Exchangeable”, and “First-order autoregressive”. To get our asymptotic estimate of \( \beta \), we need \( \sqrt{K} \) consistent estimators \( \hat{\alpha} \), and \( \hat{\tau} \) first, and plug them into the estimating equation of \( \beta \), together we would get the estimator \( \hat{\beta} \), under mild regularity conditions ([6]). \( \hat{\beta} \) is consistent and asymptotically following a normal distribution as shown above. When \( R \) is misspecified, which is often the case, \( \hat{\beta} \) is still consistent, as long the missing is MAR ([6]). We can also intuitively get it from the asymptotic property of \( \hat{\beta} \), (Equation 2.18), as long as \( R \) asymptotically converges to a fixed matrix, we still have a consistent estimator, just with a different variance and worse performance. This is why GEE is robust to misspecification of the correlation matrix.

### 2.1.5 Bias of using GEE under MCAR when the real model is under MAR

Standard GEE above is purely frequentist inference, when missing occurs, a strong condition missing completely at random (MCAR[10]) is required to ensure ignorability. MCAR means the missingness is independent of both observed and unobserved data. Thus, improved approach is needed under weaker condition missing at random (MAR), meaning missingness is independent of unobserved data given observed data. Our longitudinal study dataset should fall into the MAR category. Two approaches are available to deal with MAR longitudinal data [1], weighted GEE (WGEE, Robins, [3]) and multiple-imputation based GEE (MIGEE). The
former includes a dropout model, weighting the estimating scores of each subject by their inverse probability weights (the inverse of the probability that a subject drops out at the time he/she drops out), giving less probable events more weights. The later imputes missing data given observed data multiple times and “takes average”. We focus on MI-GEE in this context. Performance differences between GEE and WGEE using same imputation model is discussed in Jansen, [4].

2.2 Multiple Imputation (MI)

Multiple imputation was proposed by Rubin [12], to extend the single imputation for better results.

\[ h(Q|Y_{obs}) = \int g(Q|Y_{obs}, Y_{mis})f(Y_{mis}|Y_{obs})dY_{mis} \]  

(2.19)

Here Q represents any quantity, by sampling from \( f(Y_{mis}|Y_{obs}) \) we are actually using the sample average to approximate the true integral. Each sample is one imputation. Here our interested quantities are true regression coefficients \( \beta \). If no missing data, we should expect (2.18), let \( U = S^{-1}_\beta S^{-1}_\beta / K \) to be the estimated variance of the normal approximation (\( \hat{\beta} - \beta \)), approximately for K subjects

\[ (\hat{\beta} - \beta) \sim N(0, U) \]  

(2.20)

if we impute missing \( y_{mis} \) M times, and get estimates of \( \hat{\beta} \) M times, each \( \hat{\beta}^m \), we have \( (\hat{\beta}^m - \beta) \) approximated by \( N(0, U^m) \), \( m \in 1, ..., M \).

\[ (\hat{\beta}^m - \beta) \sim N(0, U^m) \]  

(2.21)

The multiple imputation suggest using the estimator \( \hat{\beta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\beta}^m \). Approximately [12],

\[ (\hat{\beta}_M - \beta) \sim N(0, \frac{1}{M} \sum_{m=1}^{M} U^m + \frac{m+1}{m} \sum_{m=1}^{M} (\hat{\beta}^m - \hat{\beta}_M)(\hat{\beta}^m - \hat{\beta}_M)^T (m-1)) \]  

(2.22)
First part is the within-imputation variance $\hat{W} = \frac{1}{M} \sum_{m=1}^{M} U^m$, and second part is the between-imputation variance $\hat{B} = \frac{m+1}{m} \sum_{m=1}^{M} (\hat{\beta}^m - \hat{\beta}_M)(\hat{\beta}^m - \hat{\beta}_M)^T$.

2.3 Generalized Estimating Equation with Multiple Imputation (MI-GEE)

To implement MI in GEE, we only need to specify an imputation model. Let $y_{k,j}$ denote $j$-th element (row) of $y_k$, and $x_{k,j}$ denote $j$-th row of $X_k$, $y_{k,j}$ is missing.

$$p(y_{k,j}|y_{k,1}, \ldots, y_{k,j-1}, x_{k,1}, \ldots, x_{k,j})$$

The model we specify for imputation might not be the actual one, but the mis-specification of imputation model only affects the missing data, and the missing data is “ignorable” under missing at random (MAR). Results show that MI-GEE performs well, as long as the imputation model is not too off (Beunckens, [1] and Meng, [7]).

2.4 Generalized Linear Mixed Model (GLMM)

This is not our focus in this study, but if the interest is to characterize all information from the responses $y_1, y_2, \ldots, y_k$, including subject effects and within-subject dependence, where the response is binary, a generalized version of linear mixed model is a good option.

In Gaussian:

$$y_k = X_k\beta + Zu_k + \epsilon_k \quad (2.23)$$

where $\beta$ is fixed and $u \sim N(0, D)$, $\epsilon \sim N(0, R)$. $y_k$ is a vector. The MLE for fitting this model is presented in Laird [5] as a two-stage random-effects model fitted by combination of empirical Bayes and MLE using EM algorithm.
In the generalized case where distribution is non-gaussian, generalized linear model (GLM) and linear mixed model are combined as generalized linear mixed model (GLMM, Stiratelli [13], Williams, [14], etc.) where the link function connects $g(\mu) = X_k\beta + Zu_k$. The estimating process for GLMM often involves working on integrals not expressible in analytical form, methods like numerical quadrature or MCMC could serve as computational solutions (Breslow [9]).

Results from GLMM is also presented in next chapter, just for a simple comparison with MI-GEE.
CHAPTER 3

Bupropion Data Analysis

Bupropion is a traditional medication for reducing methamphetamine use. In this study, our variables include binary response (test positive or negative), binary predictor of interest treatment (bupropion or placebo), and other covariate predictors include time of observation, methamphetamine use days in previous 30 days, smoking (yes or no), age and gender. In GEE, the model is fitted through estimating equations, and in GLMM, the random effect is each patient’s base effect in the logit function for all his/her observations across time.

3.1 Study Design Protocol [2]

Our study data consists of 84 methamphetamine-dependent participants. The design was a double-blind, randomized clinical trial comparing bupropion twice daily to placebo twice daily for 12 weeks. In the modeling part, this denoted as “bupropion” and “placebo”. 294 participants were initially screened in order to randomize the 84 eligible participants. Participants visited the clinic three times a week to provide urine samples and complete study assessments. Study treatment was free of charge and participants received gift cards for every visit (Heinzerling, [2]). The goal of this study is to see if bupropion serves as an effective treatment for methamphetamine dependence. The covariates included are age, gender, days of methamphetamine use past 30 days, and smoking.
<table>
<thead>
<tr>
<th></th>
<th>Bupropion (N = 41)</th>
<th>Placebo (N = 43)</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>38.6(10.1)</td>
<td>38.1(10.3)</td>
<td>0.81</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td></td>
<td>0.65</td>
</tr>
<tr>
<td>Male</td>
<td>83%(34)</td>
<td>79%(34)</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>17%(7)</td>
<td>21%(9)</td>
<td></td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td>0.65</td>
</tr>
<tr>
<td>Hispanic</td>
<td>44%(18)</td>
<td>40%(17)</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>37%(15)</td>
<td>30%(13)</td>
<td></td>
</tr>
<tr>
<td>African American</td>
<td>17%(7)</td>
<td>23%(10)</td>
<td></td>
</tr>
<tr>
<td>Asian / Pacific Islander</td>
<td>2%(1)</td>
<td>7%(3)</td>
<td></td>
</tr>
<tr>
<td>Days of methamphetamine use past 30 days (MA&lt;sub&gt;30D&lt;/sub&gt;)</td>
<td>10.3(6.8)</td>
<td>9.9(6.1)</td>
<td>0.78</td>
</tr>
<tr>
<td>Cigarette Smoker</td>
<td></td>
<td></td>
<td>0.48</td>
</tr>
<tr>
<td>Smoker</td>
<td>63%(26)</td>
<td>56%(24)</td>
<td></td>
</tr>
<tr>
<td>Non-smoker</td>
<td>37%(15)</td>
<td>44%(19)</td>
<td></td>
</tr>
<tr>
<td>Hamilton Rating Scale for Depression (HRSD)</td>
<td>6.3 (4.8)</td>
<td>6.8 (5.1)</td>
<td>0.67</td>
</tr>
</tbody>
</table>

Table 3.1: Demographics of 84 participants (Heinzerling, [2]). The p-value shows the dependence between Bupropion/Placebo assignment and demographical variables. It is best to assign treatments so that the treatment is not confounded with any observable demographical variables.
3.2 Exploratory Data Visualization

Figure 3.1: Proportion of Methamphetamine test positives against Week, grouped by Treatment and Week

We can see roughly a small advantage of bupropion in reducing drug use from this exploratory plot.
Figure 3.2: Proportion of Methamphetamine test positives against Week, grouped by Treatment, Week, and Gender

This picture shows the interaction of bupropion effect and gender. The reduction of drug use for women is more significant than men. However, this might be affected by the inequality of gender distribution of this dataset.
Proportions of MA use against MA_30D subject by subject

Figure 3.3: Proportion of Methamphetamine test positives against MA_30D, grouped by Subject ID, denoted by Treatment

This picture visually shows the effect of methamphetamine use in previous 30 days, we can see that this covariate has a great impact on our outcome.
Figure 3.4: Proportion of Methamphetamine test positives against MA_30D, grouped by MA_30D and Treatment, denoted by Number of Methamphetamine test positives

We aggregates the test positive counts for all participants and plot that against methamphetamine use in previous 30 days to visualize the overall effect of MA_30 has on test positive proportions.
Figure 3.5: Proportion of Methamphetamine test positives against MA_30D, grouped by Week and MA_30D, denoted by Number of Methamphetamine test positives

Here it visualizes the test result against time for groups categorized by MA_30D to help view MA_30D effect on time trends of test positive proportions.
Figure 3.6: Proportion of Methamphetamine test positives against MA_30D, grouped by Week, MA_30D and Treatment, denoted by Number of Methamphetamine test positives.

To provide an bigger overview, we use aggregated proportion of methamphetamine test positives of all participants, and include the MA_30, treatment and time as predictors in this picture.
3.3 GEE

Modeling is performed under R package *geepack* version 1.1-6. Correlation structure is specified to “exchangeable”. However, “AR(1)” was also used but results did not show much difference, thus not presented in the context.

![Missing Observations](image)

Figure 3.7: Number of missing observations grouped by subject

<table>
<thead>
<tr>
<th>Name</th>
<th>Content</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>y</td>
<td>test results of methamphetamine</td>
<td>Binary: 0 (Negative), 1 (Positive)</td>
</tr>
<tr>
<td>Age</td>
<td>Age of participants</td>
<td>Continuous</td>
</tr>
<tr>
<td>Gender</td>
<td>Gender of participants</td>
<td>Binary: 0 (Female), 1 (Male)</td>
</tr>
<tr>
<td>Treatment</td>
<td>Treatment condition</td>
<td>Binary: 0 (Placebo), 1 (Bupropion)</td>
</tr>
<tr>
<td>Time.Std</td>
<td>Time (days) of participant visits subtracted by their average</td>
<td>Continuous</td>
</tr>
<tr>
<td>MA_30D.Std</td>
<td>Days of methamphetamine use past 30 days subtracted by their average</td>
<td>Continuous</td>
</tr>
<tr>
<td>Smoker</td>
<td>Whether smoker or not</td>
<td>Binary: 0 (non-smoker), 1 (smoker)</td>
</tr>
</tbody>
</table>

Table 3.2: Variable coding. Subtracting the mean from Time and MA_30D is intended to make the intercept and Treatment:Bupropion coefficient in the model represent their effects where Time and MA_30D are controlled at an average level, instead of 0, for easier interpretation.
3.3.1 GEE on One Imputation

The data collection process (Heinzerling [2]) did not keep all missing slots, instead, when a participant stopped showing up for a few consecutive times, the hospital personnel would stop record empty slots in the data file. Therefore, the number of observation across time for each subject will then be different (see Figure 3.7). Our model is based on the data where slots already presented during collection, instead of adding empty slots after them for uniformity of the missing structure.

We specify a simple imputation model. If subject $i$ and observation $j$ is missing, impute the missing by

$$p(y_{k,j}|y_{k,1}, \ldots, y_{k,j-1}, x_{k,1}, \ldots, x_{k,j}) = \left\{ \begin{array}{ll} 1 & \frac{\sum_{i=1}^{j-1} 1(y_{k,i}=1)}{j-1} \\ 0 & \frac{\sum_{i=1}^{j-1} 1(y_{k,i}=0)}{j-1} \end{array} \right.$$

Results presented in Table 3.3. The most significant coefficient is MA$_{30D}$, which is the days of methamphetamine use past 30 days. Our interest in the bupropion treatment effect given average visit time and MA$_{30D}$, which is -0.43199, and P-value 0.233, not significant. In a logistic regression setup, this is the odds ratio $P(+|Bup)/P(-|Bup)$ for the main effect of a binary variable Treatment, meaning odds ratio of methamphetamine test results under Bupropion to Placebo is $e^{-0.43199} = 0.64$, given same other conditions. The closer this value to zero, the bigger Bupropion effect there is. Although prediction of individual observations is not interesting here and not well captured under GEE, we could use it to visualize our results, especially for the trend differences under Bupropion and Placebo in Figure 3.8.
GEE on One Imputation.

Marginal prediction $\hat{P}(y_{i,j} = 1)$ calculated.

Call:
geeglm(formula = y ~ Age + Gender + Treatment + Time.Std + MA_30D.Std + Smoker + MA_30D.Std:Treatment + Time.Std:Treatment, family = "binomial", data = data.final.extended, id = data.final.extended$RID, corstr = "exchangeable")

Coefficients:

|                      | Estimate | Std. err | Wald  | Pr(>|W|) |
|----------------------|----------|----------|-------|----------|
| (Intercept)          | −0.25212 | 0.88188  | 0.08  | 0.775    |
| Age                  | 0.00572  | 0.02092  | 0.07  | 0.784    |
| GenderMale           | −0.94853 | 0.48896  | 3.76  | 0.052    |
| TreatmentBupropion   | −0.43199 | 0.36183  | 1.43  | 0.233    |
| Time.Std             | 0.00345  | 0.00260  | 1.76  | 0.185    |
| MA_30D.Std           | 0.19700  | 0.04497  | 19.19 | 1.2e−05 ***|
| Smoker               | 0.21105  | 0.39394  | 0.29  | 0.592    |
| TreatmentBupropion:MA_30D.Std | −0.10369 | 0.06130  | 2.86  | 0.091    |
| TreatmentBupropion:Time.Std | −0.00610 | 0.00391  | 2.43  | 0.119    |

Signif. codes:  0 ***  0.001 **  0.01 *  0.05 .  0.1  1

Estimated Scale Parameters:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. err</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.03</td>
<td>0.279</td>
</tr>
</tbody>
</table>

Correlation: Structure = exchangeable  Link = identity

Estimated Correlation Parameters:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. err</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>0.408</td>
<td>0.142</td>
</tr>
</tbody>
</table>

Number of clusters: 84  Maximum cluster size: 36

Table 3.3: GEE based on one-time imputation
Figure 3.8: GEE fitted test positive probabilities for each observation against Time. Missing data imputed only once by subject historical proportion of Methamphetamine test positives
3.3.2 MI-GEE

We apply the same imputation scheme for our multiple imputation purpose. The imputation is repeated 20 times, all of them pooled together to perform MI-GEE (Table 3.4). The combined estimates are close to the model with one imputation in previous section. We draw a similar conclusion based on MI-GEE.

### MI-GEE:

```r
> model.impute

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. err</th>
<th>z-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>−0.11220</td>
<td>0.89593</td>
<td>−0.125</td>
<td>0.900</td>
</tr>
<tr>
<td>Age</td>
<td>0.00230</td>
<td>0.02079</td>
<td>0.111</td>
<td>0.900</td>
</tr>
<tr>
<td>GenderMale</td>
<td>−0.93743</td>
<td>0.47948</td>
<td>−1.955</td>
<td>0.051</td>
</tr>
<tr>
<td>TreatmentBupropion</td>
<td>−0.40067</td>
<td>0.36188</td>
<td>−1.107</td>
<td>0.268</td>
</tr>
<tr>
<td>Time.Std</td>
<td>0.00288</td>
<td>0.00380</td>
<td>0.758</td>
<td>0.448</td>
</tr>
<tr>
<td>MA30D.Std</td>
<td>0.19178</td>
<td>0.04484</td>
<td>4.276</td>
<td>1.9e−05 ***</td>
</tr>
<tr>
<td>Smoker</td>
<td>0.19541</td>
<td>0.39706</td>
<td>0.492</td>
<td>0.623</td>
</tr>
<tr>
<td>TreatmentBupropion:MA30D.Std</td>
<td>−0.09997</td>
<td>0.06114</td>
<td>−1.635</td>
<td>0.102</td>
</tr>
<tr>
<td>TreatmentBupropion:Time.Std</td>
<td>−0.00619</td>
<td>0.00497</td>
<td>−1.246</td>
<td>0.213</td>
</tr>
</tbody>
</table>
```

### Previous result of GEE with one imputation:

Model from one imputation:

|             | Estimate | Std.err | Wald   | Pr(>|W|) |
|-------------|----------|---------|--------|---------|
| (Intercept) | −0.25212 | 0.88188 | 0.08   | 0.775   |
| Age         | 0.00572  | 0.02092 | 0.07   | 0.784   |
| GenderMale  | −0.94853 | 0.48896 | 3.76   | 0.052   |
| TreatmentBupropion | −0.43199 | 0.36183 | 1.43   | 0.233   |
| Time.Std    | 0.00345  | 0.00260 | 1.76   | 0.185   |
| MA30D.Std   | 0.19700  | 0.04497 | 19.19  | 1.2e−05 ***|
| Smoker      | 0.21105  | 0.39394 | 0.29   | 0.592   |
| TreatmentBupropion:MA30D.Std | −0.10369 | 0.06130 | 2.86   | 0.091   |
| TreatmentBupropion:Time.Std | −0.00610 | 0.00391 | 2.43   | 0.119   |

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Table 3.4: GEE with one imputation and MI-GEE
3.3.3 Exclusion of Missing Data

We also want to see how the conclusion drawn is sensitive to exclusion of missing values. The model GEE where all missing data are deleted (Table 3.5 is used to compare with the model GEE with one imputation (Table 3.3). The shows that GEE here is not sensitive to inclusion of imputed data. Again, the predicted probabilities for visualization for the results is presented in Figure 3.9.

![Figure 3.9: GEE Fitted test positive probabilities for each observation against Time. (Complete data): all missing slots were deleted. (Imputed): missing data imputed only once by subject historical proportion of Methamphetamine test positives (Table 3.3)](image-url)
Table 3.5: Sensitivity Analysis. This model and the model in table 3.3 are basically consistent with each other, as we would draw the similar conclusions based on these two models. Plus the robust nature against correlation structure misspecification of GEE, we could say this modeling process is not sensitive to inclusion or exclusion of missing values.
3.4 GLMM

Random intercept is used for each subject, prediction \( \hat{y} \) not presented for GLMM. The model is run under R package \textit{glmmML}. Results shown in Table 3.6. For simplicity, the missing data were imputed using the same scheme as in GEE, although more formal approach is to use likelihood-based method.

The coefficients in GLMM here are generally bigger than GEE, but the p-value are similar. Thus our conclusion still holds that Bupropion has a small but not significant effect on treating methamphetamine, while days of methamphetamine use past 30 days (MA\(_{30D}\)) is still a strong indicator for methamphetamine test results. Since our interest in this study is not to fit full likelihood, but in the overall effects of Treatment, non-likelihood based model GEE is more robust to misspecification of missing imputation models and correlation structures.
**Table 3.6: GLMM under the data with one-time imputation**

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z value</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.00227</td>
<td>1.57061</td>
<td>-0.638</td>
<td>5.2e-01</td>
</tr>
<tr>
<td>Age</td>
<td>0.02433</td>
<td>0.03584</td>
<td>0.679</td>
<td>5.0e-01</td>
</tr>
<tr>
<td>GenderMale</td>
<td>-1.84661</td>
<td>0.89768</td>
<td>-2.057</td>
<td>4.0e-02</td>
</tr>
<tr>
<td>TreatmentBupropion</td>
<td>-0.81671</td>
<td>0.66607</td>
<td>-1.226</td>
<td>2.2e-01</td>
</tr>
<tr>
<td>Time.Std</td>
<td>0.00546</td>
<td>0.00441</td>
<td>1.236</td>
<td>2.2e-01</td>
</tr>
<tr>
<td>MA30D.Std</td>
<td>0.37774</td>
<td>0.08997</td>
<td>4.198</td>
<td>2.7e-05</td>
</tr>
<tr>
<td>Smoker</td>
<td>0.60561</td>
<td>0.74270</td>
<td>0.815</td>
<td>4.1e-01</td>
</tr>
<tr>
<td>TreatmentBupropion:MA30D.Std</td>
<td>-0.25564</td>
<td>0.11318</td>
<td>-2.259</td>
<td>2.4e-02</td>
</tr>
<tr>
<td>TreatmentBupropion:Time.Std</td>
<td>-0.01086</td>
<td>0.00628</td>
<td>-1.729</td>
<td>8.4e-02</td>
</tr>
</tbody>
</table>

Scale parameter in mixing distribution: 2.7 gaussian
Std. Error: 0.342
LR p-value for H_0: sigma = 0: 1.27e-162
Residual deviance: 1550 on 2181 degrees of freedom  AIC: 1570
3.5 Additional: Bupropion on Smoking and Depression

There is also interest in Bupropion on smoking and depression, the related data were collected along the process.

3.5.1 Smoking

56 smokers are included before missing deleting, and only data of daily cigarette use from week 1 to 12 are included. Covariates include Age, Gender, Treatment, MA_30D, and Time. The result shows slight but not significant effect of Treatment on reducing cigarette use (Table 3.7). Figure 3.10 gives an idea of the results.

```
GEE on smoking, all missing data deleted:

Call:
geeglm(formula = Cig ~ Age + Gender + Treatment + MA_30D + Time,
       family = "gaussian", data = data.smoking.model, id = data.smoking.
       model$SID,
corstr = "exchangeable")

Coefficients:
                Estimate Std. Err Wald Pr(>|W|)
(Intercept) 1.22975 5.22605  0.06  0.814
Age 0.02950  0.10430  0.08  0.777
GenderMale 3.18926  1.59028  4.02  0.045 *
TreatmentBupropion -0.29122  1.65383  0.03  0.860
MA_30D 0.12771  0.14768  0.75  0.387
Time 0.00337  0.00876  0.15  0.701
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Estimated Scale Parameters:
  Estimate Std.err
(Intercept) 52.9  13.5

Correlation: Structure = exchangeable  Link = identity
Estimated Correlation Parameters:
  Estimate Std.err
alpha 0.798  0.0485

Number of clusters: 50  Maximum cluster size: 90
```

Table 3.7: GEE on cigarette use with all missing data deleted
Figure 3.10: GEE fitted number of cigarettes for each observation against Time (days), grouped by Time. (Complete data): all missing slots were deleted
3.5.2 Depression

All 84 participants are included before missing deleting, and only data from week 1 to 12 are considered in the model. The model shows slightly more severe but not significant effect of Bupropion on Depression. Results shown in Table 3.8, and visualization shown in Figure 3.11.

GEE on Depression scores:

Call: geeglm(formula = Depression ~ Age + Gender + Treatment + MA30D + Week,
family = "gaussian", id = data.depression.model$SID, corstr = "exchangeable")

Coefficients:

|                | Estimate | Std.err | Wald  | Pr(>|W|) |
|----------------|----------|---------|-------|----------|
| (Intercept)    | 4.764610 | 1.942822| 6.01  | 0.01419  |
| Age            | 0.000316 | 0.032429| 0.00  | 0.99221  |
| GenderMale     | -0.687530| 0.985329| 0.49  | 0.48532  |
| TreatmentBupropion | 0.554727 | 0.645026| 0.74  | 0.38978  |
| MA30D          | 0.000326 | 0.059081| 0.00  | 0.99559  |
| Week           | -0.158128| 0.040897| 14.95 | 0.00011  |

Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Estimated Scale Parameters:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std.err</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>14.6</td>
<td>2.06</td>
</tr>
</tbody>
</table>

Correlation: Structure = exchangeable Link = identity

Estimated Correlation Parameters:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std.err</th>
</tr>
</thead>
<tbody>
<tr>
<td>alpha</td>
<td>0.432</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Number of clusters: 83 Maximum cluster size: 12

Table 3.8: GEE on HRSD depression scores with all missing data deleted
Figure 3.11: GEE fitted HRSD depression scores for each observation against Time (weeks), grouped by Time. (Complete data): all missing slots were deleted
CHAPTER 4

Discussion

Repeated-measures data is very common in clinical trials, drug manufacturing and epidemiological studies. We have discussed different models based on different interests. No model is omnipotent. GLMM is more accurate to picture the underlying data structure, and it could answer most common questions in the research. However, as mentioned before, when the assumption is not Gaussian, GLMM might be hard to fit correctly. Under various circumstances, different approaches for approximating integrals in GLMM are needed, leaving a potential to misestimate the quantities. GEE sacrifices full likelihood information to estimate coefficients only, losing the ability to find the within-subject dependence, but it is fast and robust. Depends on the goal of the study, we might choose the model wisely for the best efficiency.

The missingness is always a problem, if the case is missing non random (MNR), GEE definitely fails to serve our purpose, however, most models would suffer from it.

As for the scientist problem we started with, whether using Bupropion reduces methamphetamine use? The result shows mild effect on the treatment. Nowadays, there is no adequate medication to completely suppress the drug use. Hopefully, statistical models along with biological efforts could work together towards finding something phenomenal.
References


