LOF OF LOGISTIC GEE MODELS AND COST EFFICIENT BAYESIAN OPTIMAL DESIGNS FOR NONLINEAR COMBINATIONS OF PARAMETERS IN NONLINEAR REGRESSION MODELS

by

ZHONGWEN TANG

M.S. , Kansas State University, 2004

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics

College of Arts and Science

KANSAS STATE UNIVERSITY

Manhattan, Kansas

2008

Abstract

When the primary research interest is in the marginal dependence between the response and the covariates, logistic GEE (Generalized Estimating Equation) models are often used to analyze clustered binary data. Relative to ordinary logistic regression, very little work has been done to assess the lack of fit of a logistic GEE model. A new method addressing the LOF of a logistic GEE model was proposed. Simulation results indicate the proposed method performs better than or as well as other currently available LOF methods for logistic GEE models. A SAS macro was developed to implement the proposed method.

Nonlinear regression models are widely used in medical science. Before the models can be fit and parameters interpreted, researchers need to decide which design points in a prespecified design space should be included in the experiment. Careful choices at this stage will lead to efficient usage of limited resources. We proposed a cost efficient Bayesian optimal design method for nonlinear combinations of parameters in a nonlinear model with quantitative predictors. An R package was developed to implement the proposed method.

LOF OF LOGISTIC GEE MODELS AND COST EFFICIENT BAYESIAN OPTIMAL DESIGNS FOR NONLINEAR COMBINATIONS OF PARAMETERS IN NONLINEAR REGRESSION MODELS

by

ZHONGWEN TANG

M. S., Kansas State University, 2004

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics

College of Arts and Science

KANSAS STATE UNIVERSITY

Manhattan, Kansas

2008

Approved by:

Major Professor

Shie-Shien Yang

Abstract

When the primary research interest is in the marginal dependence between the response and the covariates, logistic GEE (Generalized Estimating Equation) models are often used to analyze clustered binary data. Relative to ordinary logistic regression, very little work has been done to assess the lack of fit of a logistic GEE model. A new method addressing the LOF of a logistic GEE model was proposed. Simulation results indicate the proposed method performs better than or as well as other currently available LOF methods for logistic GEE models. A SAS macro was developed to implement the proposed method.

Nonlinear regression models are widely used in medical science. Before the models can be fit and parameters interpreted, researchers need to decide which design points in a prespecified design space should be included in the experiment. Careful choices at this stage will lead to efficient usage of limited resources. We proposed a cost efficient Bayesian optimal design method for nonlinear combinations of parameters in a nonlinear model with quantitative predictors. An R package was developed to implement the proposed method.

Contents

List of Figures

List of Tables

Acknowledgments

I consider myself a very lucky man. I get lots of help from many people while getting my dissertation done. First I would like to thank my advisor Dr. Shie-Shien Yang. He helps me a lot. My optimal design project was developed from my internship project in Amgen. I would like to thank Dr. Hugh Rand, Dr. Guang Chen, and Dr. Cheng Su for mentoring me on the optimal design project. I would like to thank Dr. Paul Nelson, Dr. Karen Garrett, Dr. Gabriel Nagy, and Dr. Haiyan Wang for serving in my committee.

I would like to thank Dr. John Boyer, Dr. Tom Loughin, Dr. James Neill, Dr. Jeffrey Pontius, Dr. Kenneth Kemp for helping me admitted into statistics department. I got tremendous pleasure by learning statistics. I would like to thank all the wonderful teachers in the department for their terrific teaching.

I would like to thank Dr. Ron Klein (University of Wisconsin) for providing me the diabetic retinopathy data. Their work is supported in part by research grants EY03083 and EY06594 from the National Institutes of Health.

1 Lack of Fit for Logistic GEE Model

1.1 Notation

- $\mathbf{0}$: a vector with all elements equal to 0.
- $\mathbf{1}$: a vector with all elements equal to 1.
- GEE: generalized estimating equation.
- GLM: generalized linear model.
- LOF: Lack-Of-Fit.
- I_n: An $n \times n$ identity matrix with diagonal elements equal to 1 and off diagonal elements equal to 0.
- $I(.)$: An indicator function, which equals to 1 with true argument; equals to 0 with false argument.
- J: a matrix with all elements equal to 1.
- $logit(.)$: logit function, $logit(x) = \log\left(\frac{x}{1-x}\right)$.

1.2 Introduction

Studies involving binary outcomes are quite common in medical science. Berkson (1944) first proposed using logistic regression for the analysis of binary bioassay data while Cox (1970) described its application for a variety of problems. The logistic regression model has become a widely used and accepted method for the analysis of binary outcome variables. This popularity stems from the availability of easily used software (such as SAS, R, Minitab) and the ease of interpretation of the results of the fitted model. Commensurate with this increase in application has been an increase in statistical research on the model. One area of current research is the development of methods to assess the adequacy of the fitted model (Lack-Of-Fit). Many Lack-Of-Fit methods for logistic regression models have been developed. An incomplete list of Lack-Of-Fit methods for logistic regression includes Prentice's goodness of link test based on generalized logistic regression (1976), Hosmer and Lemeshow chi-square test (1980) based on data partitioning, Tsiatis' score test based on data partitioning using covariate patterns (1980), Copas's graphical approach comparing parametric and nonparametric estimates (1980), Brown's score test extending Prentice work to multiple regressor (1982), Stukel's score test based on generalized logistic regression

model (1988), Azzalini et al. pseudo likelihood approach comparing parametric and nonparametric estimates (1989), le Cessie and Van Howelingen's smoothed residual tests (1991, 1995), Osius and Rojek's approximate normal test (1992), Royston's cusum method for detecting departures from linearity in the logit link function (1992), Farrington's generalized Pearson statistics (1996), Pulkstenis and Robinson's Pearson and Deviance chi-square tests (2002), and Liu and Yang's partitioned model test (2007).

If binary outcomes are correlated, ordinary logistic regression is not appropriate. These data can be analyzed using the generalized estimating equation (GEE) method proposed by Liang and Zeger (1986). Assessment of the adequacy of a fitted GEE model could be problematic since no likelihood exists and the residuals within a cluster are correlated. Compared to ordinary logistic regression model, relatively little work has been done to assess the adequacy of a logistic GEE model. Barnhart and Williamson (1998), Horton et al. (1999), and Evans and Li (2004) proposed some LOF (Lack Of Fit) methods for logistic GEE model based on data partitioning and model comparison. Williamson et al. (2003) proposed a kappa-like classification statistic for assessing the adequacy of a logistic GEE model. Evans (1998) and Pan (2002) proposed some residuals based LOF statistics for logistic

GEE model. In our work, we developed a LOF method for logistic GEE model which performs as well as or better than other currently available methods.

1.3 Literature Review

Our goal is to develop a LOF method for assessing the adequacy of a logistic GEE model. Since a logistic GEE model can be thought of as an extension of ordinary logistic regression models, it is useful to review ordinary logistic regression and its LOF methods first.

1.3.1 Logistic Regression Model

In our work, we follow the notation by McCullagh and Nelder (1983) to describe ordinary logistic regression. The responses $y_1, \cdots, y_n$ are assumed to be the observed values of independent random variables $Y_i's$, $i = 1, \cdots, n$ such that Y_i has a binomial distribution with m_i trials and probability of success π_i. The logit of π_i depends on covariate vector $\mathbf{x}_i$ in a linear manner, i.e.,

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^{'}\boldsymbol{\beta}, \tag{1}$$

where $\mathbf{x}_i^{'} = (1, x_{1i}, \cdots, x_{pi})$ denotes a set of $p+1$ dimensional fixed covariates for observation i. Parameter estimates for $\boldsymbol{\beta}$ are usually obtained by maximum likelihood approach and denoted by $\hat{\boldsymbol{\beta}}^{'} = \left(\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\right)$. We denote

the fitted value of π_i by $\hat{\pi}_i$.

Note that a Binomial experiment with m_i trials consists of m_i independent Bernoulli experiments. Equivalently, the model may be written in terms of the odds of a positive response in Bernoulli experiment, giving

$$\frac{\pi_i}{1-\pi_i} = \exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}\right)$$

Finally the probability of a positive response is

$$\pi_i = \frac{\exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}\right)}$$

Specifically a linear logistic regression model with two covariates can be expressed as

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Assuming x_1 and x_2 are functionally independent of each other, the conclusion based on such a model may be stated as follows. The effect of a unit change in x_2 is to increase the log odds by an amount β_2. Equivalently, we may say that the effect of a unit change in x_2 is to increase the odds of a positive response multiplicatively by a factor of $\exp(\beta_2)$. It is important here that x_1 is held fixed and is not permitted to vary as a consequence of the change in x_2.

The log likelihood may be written in the form

$$l(\boldsymbol{\beta};\mathbf{y}) = \sum_{i=1}^{n} y_i \mathbf{x}_i^{'}\boldsymbol{\beta} - \sum_{i=1}^{n} m_i \log\left(1+\exp\left(\mathbf{x}_i^{'}\boldsymbol{\beta}\right)\right)$$

The score function is: $\frac{\partial l}{\partial \boldsymbol{\beta}} = \mathbf{U}(\boldsymbol{\beta}) = X'(\mathbf{y}-\boldsymbol{\pi})$

where $X = [\mathbf{x}_1, \cdots, \mathbf{x}_n]'$, $\boldsymbol{\pi} = [\pi_1, \cdots, \pi_n]'$, $\mathbf{y} = [y_1, \cdots, y_n]'$.

The Fisher information for $\boldsymbol{\beta}$ is $X'WX$, where

$W = diag\{m_1\pi_1(1-\pi_1), \cdots, m_n\pi_n(1-\pi_n)\}$. The maximum likelihood estimates of the parameters are obtained by solving the following equations (McCullagh and Nelder, 1983).

$$\mathbf{U}(\boldsymbol{\beta}) = X'(\mathbf{y}-\boldsymbol{\pi}(\boldsymbol{\beta})) = \mathbf{0}$$

A logistic regression model provides accurate description and inference for a data set only if it fits that data set well. After fitting the assumed model, it is very important to check the adequacy of the model before drawing any conclusions. If the assumed model is not adequate to model the data, it is necessary to modify the original analysis or the assumed model. Otherwise misleading results may be produced.

In principle, there are two different approaches for assessing LOF of a logistic regression model. The first, known as residual analysis, investigates the model on the level of individual observations and looks for those ob-

servations which are not adequately described by the model or which are highly influential on the model fit. This approach is most useful in detecting outliers.

The second approach seeks to combine the information on lack of fit in a single number, a test statistic. Statistical tests, so called overall LOF tests, are performed based on this test statistic to judge if the observed Lack-Of-Fit is statistically significant or due to random chance. Overall lack-of-fit tests have application in model building. Modeling building is the process of searching for a good model from some candidate models. In this process, different knowledge and tools can be used to assess how good the models are. For example, prior knowledge may require some variables to be included in the model even the data don't support this. Hypothesis tests may show that some variables have significant relationship with the response and therefore are good to be kept. Some other tools such as Akeike information criteria of (AIC) or the Bayesian information criteria (BIC) can also be used to select a good model. Our goal is to develop a new overall lack-of-fit test.

According to Hosmer et al. (1997), in the context of a logistic regression model, evidence of Lack-Of-Fit may come from violation of one or more of the following 3 assumptions for a logistic regression model.

1. the logit transformation is the correct function linking some linear function of the covariates $\mathbf{x}$ with the conditional mean $E(Y|\mathbf{x}) = \pi(\mathbf{x})$, i.e. $\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \mathbf{x}'\boldsymbol{\beta}$;

2. the linear predicting function, $\mathbf{x}'\boldsymbol{\beta}$, is correct.

3. the conditional variance of the response is Binomial, i.e. $Var(Y|\mathbf{x}) = m\pi(\mathbf{x})[1-\pi(\mathbf{x})]$.

1.3.2 Lack-Of-Fit Methods for Logistic Regression

Lack-Of-Fit methods for logistic regression can be roughly divided into 3 categories. The first group of Lack-Of-Fit methods are Pearson and Deviance chi-square tests and their derivatives. The second group of methods embed the logistic regression model in a more general model. Then the lack of fit is assessed by comparing the fitness of the more general model with that of the assumed logistic regression model. The third group of methods are based on comparing the parameter estimates of the assumed logistic regression model and a robust nonparametric regression model.

Pearson and Deviance Chi-square Tests The first group of Lack-Of-Fit test statistics for the assumed logistic regression model defined in page 5 are based on the comparison between the fitted values and the observed

Table 1: Contingency table of binary data for considering LOF of a logistic regression model

	$\mathbf{x}_1$	$\mathbf{x}_2$	$\cdots$	$\mathbf{x}_{n-1}$	$\mathbf{x}_n$
y=0					
y=1					

values. If the discrepancy between the observed and fitted values is small, it indicates good fit, otherwise it shows evidence of Lack-Of-Fit.

Note that each binomial variable Y_i is the sum of m_i independent binary variables. Hence the n binomial variables $Y_i's$ can be considered as $\sum_{i=1}^{n} m_i$ independent binary observations. A useful conceptual framework for thinking about assessment of the model fit for binary data is to consider the data as described by a $2 \times n$ contingency table (Table 1). The 2 rows are defined by the values of the dichotomous outcomes of the variable Y and the n columns are defined by the assumed number of possible distinct values taken on by the covariates in the model.

One important measure of the discrepancy between the observed and predicted number of successes is the generalized Pearson chi-square statistic. The generalized Pearson chi-square test statistic (McCullagh and Nelder, 1983) is defined as the sum of the squared discrepancy scaled by the estimated variance. In the context of a logistic regression, it can be expressed as

$$X^2 = \sum_{i=1}^{n} \left(y_i - m_i\hat{\pi}_i\right)^2 / \left[m_i\hat{\pi}_i\left(1-\hat{\pi}_i\right)\right]$$

Another measure of the discrepancy between the fitted values and the observed values is the deviance test statistic (D). Given n observations we can fit models containing up to n parameters. This model is not informative, because it doesn't summarize the data but merely repeats them in full. However this model gives us a baseline for measuring the discrepancy for an intermediate model. The maximum log likelihood achievable in a model with n parameters $l\left(\tilde{\boldsymbol{\beta}};\mathbf{y}\right)$ is ordinarily finite. The deviance test statistic (D) is defined to be twice the difference between the maximum achievable log likelihood and that attained under the assumed model (McCullagh and Nelder, 1983). Under any given model with fitted probabilities $\hat{\pi}$, the log likelihood is

$$l\left(\boldsymbol{\beta};\mathbf{y}\right) = \sum_{i=1}^{n}\left\{y_i \log\left(\hat{\pi}_i\right) + \left(m_i - y_i\right)\log\left(1-\hat{\pi}_i\right)\right\}$$

The maximum achievable log likelihood is attained at the points $\tilde{\pi}_i = y_i/m_i$ (McCullagh and Nelder, 1983). Therefore the deviance function is

$$D = 2[l\left(\tilde{\boldsymbol{\beta}};\mathbf{y}\right) - l\left(\hat{\boldsymbol{\beta}};\mathbf{y}\right)]$$

$$= 2\sum_{i=1}^{n}\{y_i \ln\left(y_i/m_i\hat{\pi}_i\right) + \left(m_i - y_i\right)\ln\left[\left(m_i - y_i\right)/\left(m_i - m_i\hat{\pi}_i\right)\right]\}$$

Evidence for model Lack-Of-Fit occurs when the values of these statistics (D or X^2) are large. Asymptotically X^2 and D have a χ^2 distribution with $n - p - 1$ degrees of freedom, where $p + 1$ is the number of parameters in the assumed logistic regression model. This result applies when n is fixed and the fitted counts in the contingency tables are large (Aggresti, 1996). These conditions are violated in two ways if the covariates are continuous and most of the $m_i's$ are small (sparseness). First when most $m_i's$ are very small (sparseness), the fitted cell counts will be small. Second, when more data are collected, additional covariate values could occur, so n is not fixed. Because of this, X^2 and D for logistic regression models fitted with continuous covariates usually do not have approximate chi-squared distribution. These Lack-Of-Fit statistics are more properly applied when the explanatory variables are categorical, and relatively few fitted cell counts are small.

Many authors have proposed methods to check the adequacy of a logistic regression model with continuous covariates with sparseness. One school of thought is based on grouping the data using the similarity of the data. Then the discrepancy between the fitted values and the observed values for all the groups is summarized in a test statistic. The Pearson and Deviance chi-square test statistics similar to that for checking Poisson models are then used

to check the Lack-Of-Fit for the assumed logistic regression model (Aggresti, 1996).

The data can be pooled based on partitioning the covariate space into g distinct regions ($g < n$ and g is fixed) to increase the number of counts for each cell in the contingency table. The original $2 \times n$ contingency table is shrunk to a $2 \times g$ contingency table with the columns consisting of g covariate patterns and the rows consisting of binary response values. Then the discrepancy between the fitted and observed values for all the cells is summarized into one test statistic, and a chi-square test with $g - p - 1$ degrees of freedom can be performed to test the overall Lack-Of-Fit. Let E_{i1} be the sum of N_i estimated fitted probabilities for subjects in covariate pattern i, and $E_{i2} = N_i - E_{i1}$. The Deviance and Pearson chi-square statistics are expressed as follows.

$$X^2 = \sum_{i=1}^{g} \sum_{j=1}^{2} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \text{ and } D = 2 \sum_{i=1}^{g} \sum_{j=1}^{2} (O_{ij}) \log \left(\frac{O_{ij}}{E_{ij}} \right)$$

where O_{ij} and E_{ij} are the observed and expected cell counts for the cell in column i row j respectively. These test statistics were described by Aggresti (1996). One disadvantage of this method is that the choice of the partition of the covariate space is subjective.

Hosmer and Lemeshow (1980) proposed a test which first orders all the responses according to their fitted probabilities and then classifies them into 10 (or about 10) groups (decile of risk) with approximately equal sizes. Then the Hosmer-Lemeshow Lack-Of-Fit test statistic for the assumed logistic regression model can be expressed as:

$$\chi^2_{HL} = \sum_{i=1}^{g} \sum_{j=1}^{2} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where g is the number of decile of risk groups. Through simulation Hosmer and Lemeshow (1980) showed that this test statistic has approximate chi-square distribution with $g-2$ degrees of freedom.

Hosmer and Lemeshow used 2 methods to group the data based on the fitted probabilities. The first method divides the data into g groups with approximately equal sizes. The resulted chi-square test statistic is called Hosmer-Lemeshow $\hat{C}$. The second method put the responses with fitted probabilities between 0 and 0.1 in the first group; responses with fitted probabilities between 0.1 to 0.2 in the second group; and so on. The resulted chi-square test statistic is called Hosmer-Lemeshow $\hat{H}$.

le Cessie and van Houwelingen (1991, 1995) noted that because the Hosmer-Lemeshow tests are based on a grouping strategy in the 'y' space, they

lack power to detect departures from the model in regions of the '$\mathbf{x}$' space that yield the same estimated probabilities. For example, a model with a quadratic term may have widely different '$\mathbf{x}$' values with the same estimated probability.

Pulkstenis and Robinson (2002) developed two hybrid test statistics which combine the ideas of covariate partitioning and using groups based on ranked estimated probabilities. First the data are partitioned into g covariate patterns, where the covariate patterns are determined only by categorical explanatory variables. Then the data are split into two subcategories based on the sorted fitted probabilities within each covariate pattern. The medians of the sorted fitted probabilities are generally used as the cutoff values to separate the data. The additional stratification doubles the number of covariate patterns. The original $2 \times n$ contingency table is shrunk to a $2 \times (2g)$ contingency table. The model-based expected values are computed exactly as before. The proposed Lack-Of-Fit test statistics for assumed logistic regression model are given by

$$\chi^2_{PR} = \sum_{i=1}^{g} \sum_{h=1}^{2} \sum_{j=1}^{2} \frac{\left(O_{ihj} - E_{ihj}\right)^2}{E_{ihj}} \text{ and } D_{PR} = 2 \sum_{i=1}^{g} \sum_{h=1}^{2} \sum_{j=1}^{2} O_{ihj} \log\left(\frac{O_{ihj}}{E_{ihj}}\right)$$

where i indexes covariate patterns, h indexes substratification based on or-

dering of fitted probabilities, and j indexes binary response values. These two test statistics have approximate chi-square distribution if the assumed logistic regression model is correct. The degrees of freedom for χ^2_{PR} and D_{PR} are obtained by modifying the degrees of freedom for regular Pearson and Deviance chi-square test statistics and degrees of freedom for Hosmer-Lemeshow test statistics. The degrees of freedom used by Pulkstenis and Robinson are $2g-p-1$, where $2g$ is the number of columns in the new stratification, $p+1$ is the number of parameters in the assumed logistic regression model. Pulkstenis and Robinson empirically confirmed this approximate null distribution through simulation. The simulation indicates their method has higher power for detecting omission of an interaction between a continuous covariate and a dichotomous covariate. Their simulation indicates that both their method and Hosmer-Lemeshow methods have low power detecting link function misspecification.

Parametric Models Comparison Many authors seek to test the Lack-Of-Fit of a logistic regression model by embedding the logistic regression model in a more general parametric model. The fitness of the logistic regression model is assessed by comparing the fitted logistic regression model

and the more general model. If the assumed logistic regression model fits the data well, the more general model shouldn't explain significantly more variation than the logistic regression model. This can be tested by several methods such as likelihood ratio tests, score tests or Wald tests.

Tsiatis (1980) proposed a Lack-Of-Fit test statistic based on partitioning data using covariate patterns. The space of covariates $(X_1, X_2, \cdots, X_p)$ is partitioned into G distinct regions in p-dimensional space denoted by $R_1, \cdots, R_G$.

Define G group indicators

$$I_{ig} = \begin{cases} 1 \text{ if subject } i \text{ is in group } g, \\ \\ 0 \text{ otherwise} \end{cases} \quad g = 1, \cdots, G$$

Consider the model

$$\log\left\{\pi\left(\mathbf{x}_i\right) / \left(1 - \pi\left(\mathbf{x}_i\right)\right)\right\} = \boldsymbol{\beta}^{'}\mathbf{x}_i + \boldsymbol{\gamma}^{'}\mathbf{I}_i,$$

where $\mathbf{I}_i^{'} = (I_{1i}, \cdots, I_{Gi})$ and $\boldsymbol{\gamma}^{'} = (\gamma_1, \cdots, \gamma_G)$. The Lack-Of-Fit test is equivalent to testing the hypothesis $H_0 : \gamma_1 = \cdots = \gamma_G = 0$. Tsiatis' test is based on the efficient score test (Rao, 1973). The Tsiatis' test statistic for the score test can be expressed as follows.

$$T = \mathbf{s}' V^{-} \mathbf{s},$$

where $\mathbf{s}'$ is the G dimensional vector $(\partial l/\partial\gamma_1, \cdots, \partial l/\partial\gamma_G)$ and where l denotes the log likelihood.

The $G \times G$ matrix V is equal to

$V = A - BC^{-1}B'$,

where

$A_{jj'} = -\partial^2 l/\partial\gamma_j\partial\gamma_{j'}$ $(j, j' = 1, \cdots, G)$,

$B_{jj'} = -\partial^2 l/\partial\gamma_j\partial\beta_{j'}$ $(j = 1, \cdots, G;\ j' = 0, \cdots, p)$,

$C_{jj'} = -\partial^2 l/\partial\beta_j\partial\beta_{j'}$ $(j, j' = 0, \cdots, p)$.

All the terms are evaluated at $\boldsymbol{\gamma} = \mathbf{0}$ and $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of the parameters $\boldsymbol{\beta}$ when H_0 is true. Under the null hypothesis, the test statistic T is asymptotically distributed as chi-square with degrees of freedom equal to the rank of V. When T is large, the presence of Lack-Of-Fit is claimed.

Liu and Yang (2007) proposed a Lack-Of-Fit method based on approximating the true model by a partitioned logistic model. The "true" model is a logistic model with linear predictors.

$$\log\left(\frac{\pi}{1-\pi}\right) = \mathbf{x}'\boldsymbol{\beta} + \mathbf{w}'\boldsymbol{\delta} \tag{2}$$

First the data are fitted with the assumed logistic regression model de-

fined on page 5. Then the data are partitioned into G mutually disjoint groups $(R_1, \cdots, R_G)$ based on the ranked fitted probabilities. The following partitioned logistic model is used to approximate the unknown true model.

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^{'}\boldsymbol{\beta} + \mathbf{z}_{ig}^{'}\boldsymbol{\alpha}_g \tag{3}$$

where $g = 1, \cdots, G$ is used to index the partitioned groups; $i = 1, \cdots, n$ is used to index the observations; $\mathbf{z}_{ig} = \mathbf{x}_i I\left((\mathbf{x}_i, y_i) \in R_g\right)$.

They extended their model by allowing $\mathbf{x}_{ig}$ in the definition of $\mathbf{z}_{ig}$ be replaced by a vector whose components are any function of the components of $\mathbf{x}_{ig}$. They restrict their model by requiring that $\mathbf{z}_{ig}$ contains '1' function as a covariate which corresponds to the intercept for g^{th} group.

They assume that when the group number G is big enough, the full model (3) should approximate the unknown linear true model (2).

Let $X = \begin{bmatrix} \mathbf{x}_1^{'} \\ \mathbf{x}_2^{'} \\ \cdots \\ \mathbf{x}_n^{'} \end{bmatrix}$; $Z = \begin{bmatrix} \mathbf{z}_1^{'} \\ \mathbf{z}_2^{'} \\ \cdots \\ \mathbf{z}_n^{'} \end{bmatrix}$.

Let $l\left(X, Z\right)$ and $l\left(X\right)$ be the maximum log likelihood achievable from the approximate full model (3) and the assumed logistic regression model

defined in page 5, respectively. If the assumed logistic regression model is true, then $\boldsymbol{\delta} = \mathbf{0}$ in the true model (2). Therefore the Lack-Of-Fit test is equivalent to testing $H_0 : \boldsymbol{\delta} = \mathbf{0}$. When H_0 is true, $2l(X, Z) - 2l(X)$ follows an asymptotic chi-square distribution with degrees of freedom equal to $rank([X, Z]) - rank([X])$. If $2l(X, Z) - 2l(X)$ is large, the presence of Lack-Of-Fit is claimed for the assumed logistic regression model defined in page 5. Liu and Yang (2007) used simulation to show that their method has high power for detecting omission of a quadratic term, or an interaction term between continuous and discrete covariates.

Nonparametric Methods Copas (1983) proposed a goodness-of-fit method by comparing the estimates of the assumed parametric regression function with a non-parametric kernel estimate graphically. Azzalini (1989) proposed a more formal goodness-of-fit method by comparing these two regression estimates using pseudo-likelihood ratio test. One problem with these methods is that the nonparametric curve is biased (le Cessie and Houwelingen, 1995). The bias problem is avoided by Firth et al. (1991) using a local likelihood estimation to obtain a non-parametric estimate of the regression function. Their method uses Monte Carlo methods, which makes it computationaly

expensive.

Le Cessie and Houwelingen (1995) proposed a Lack-Of-Fit method using smoothed residuals. Hosmer et al. (1997) show that this method has low power for detecting Lack-Of-Fit.

1.3.3 Logistic GEE Model

Clustered data consisting of a set of multivariate responses $\mathbf{y}_i = (y_{i1}, \cdots, y_{iT_i})'$, for $i = 1, \cdots, K$, together with a $T_i \times p$ matrix X_i of covariates associated with each response $\mathbf{y}_i$ are common in medical science. The multivariate responses $\mathbf{y}_i's$ are independent of each other. Measurements $y_{i1}, \cdots, y_{iT_i}$within each cluster could be correlated. This data structure covers the well known longitudinal data, in which each cluster is defined by a set of repeated measures on the same subject. This data structure also covers so-called familial data with a set of observations on different subjects which are grouped into clusters sharing some common features, e.g. animal litters, families or geographical regions. For ease of presentation, without loss of generality we assume the clustered data are longitudinal data and each cluster is a subject.

In many practical applications, the purpose of the analysis is to construct a regression model as a function of the marginal mean of the response y_{it}, where $i = 1, \cdots, K$ and $t = 1, \cdots, T_i$ (Rotnitzky and Jewell, 1990) and the dependence within clusters is a nuisance (Liang and Zeger, 1986). For example, the presence of a disease along with some covariates (such as nutritional status, age, sex and family income) of children might be observed every year for 3 year period. Observations from each child form a cluster.

The dependence of the outcome variable, presence of disease, on the covariates is of interest. The correlations among the 3 years within the cluster are nuisance parameters. With Gaussian data, multivariate methods can be used to analyze cluster data due to the flexibility of multivariate normal distribution. With binary data, when the focus is on the marginal probability of the individual outcomes, Prentice (1988) showed that fully parametric approaches can be cumbersome and computationally prohibitive. Alternatively, mixed effects models (Stiratelli, Laird and Ware, 1984) can be used to analyze clustered binary data. However, in mixed effects binary model, the fixed effects regression parameters have their natural interpretation for the individual clusters rather than describing covariate effects on the marginal mean (Zeger, Liang and Albert, 1988).

Generalized estimating equation (GEE) models, described in Liang and Zegger's landmark paper (1986), can be used to analyze clustered binary data. This model essentially extends generalized linear models (GLM, McCullagh and Nelder, 1983) to the situation of correlated data. GEE models can be used to analyze binary data, count data, and continuous data. In our work we will focus on logistic GEE models for analyzing clustered binary data. Therefore in our discussion, the response y_{it}, $i = 1, \cdots, K$ and

$t = 1, \cdots, T_i$ is binary with values 0 or 1.

GEE models can be divided into GEE1 and GEE2 based on the relationship between the regression parameters and correlation parameters. In our work, we focus on GEE1, which assumes that the regression parameters and the correlation parameters don't depend on each other (Hardin and Hilbe, 2003).

In our work, without specification, the default GEE model is GEE1 model. Actually, most researchers who refer to a GEE model are referring to this special collection of models.

A basic feature of GEE models is that the joint distribution of a subject's response vector $\mathbf{y}_i$ does not need to be specified. Instead, only the marginal distribution of y_{it} at each time point is specified. For example, suppose that there are 2 time points and we observe 2 binary outcomes y_{i1} and y_{i2}. GEE only requires that y_{i1} and y_{i2} are two univariate binary variables rather than assuming that y_{i1} and y_{i2} form a (joint) bivariate distribution.

A related feature of GEE models is that the covariance structure is treated as nuisance. The focus is on the relationship between the conditional mean of y_{it} given the covariate value and the covariate $\mathbf{x}_{it}$. A GEE model yields consistent and asymptotically normal estimates for the regression parame-

ters, even with mis-specified covariance structure of the clustered data (Liang and Zeger, 1986). The logistic GEE model can be thought of as an extension of the logistic regression model to correlated data. The logistic GEE model specification involves those of logistic regression with one additional specification.

First the dependence between y_{it} and $\mathbf{x}_{it}$ is defined as

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \mathbf{x}_{it}^{'}\boldsymbol{\beta}; \qquad i = 1, \cdots, K; t = 1, ..., T_i \tag{4}$$

where $\mathbf{x}_{it} = [1, x_{it1}, \cdots, x_{itp}]'$ is a $(p+1) \times 1$ covariate vector for subject i at time t and $\boldsymbol{\beta}$ is a $(p+1) \times 1$ parameter vector. $\pi_{it} = P(y_{it} = 1|\mathbf{x}_{it})$ is the conditional mean of the positive response given the covariate values. The conditional variance of y_{it} is $\pi_{it}(1-\pi_{it})$.

This model can also be expressed for each subject.

$$logit(\boldsymbol{\pi}_i) = X_i\boldsymbol{\beta}$$

where $logit(\boldsymbol{\pi}_i) = [logit(\pi_{i1}), \cdots, logit(\pi_{iT_i})]'$, $X_i = [\mathbf{x}_{i1}, \cdots, \mathbf{x}_{iT_i}]'$ is a $T_i \times (p+1)$ design matrix.

The additional specification in a GEE model is the "working" correlation structure of the repeated measures. This working correlation matrix is of

size $T \times T$ assuming that there are T fixed number of time points for each subject. For a given subject, it doesn't have to be measured at all T time points. Therefore $T_i \leq T$. Each individual's correlation matrix R_i is of size $T_i \times T_i$ which could be obtained by removing appropriate rows and columns in R. It is assumed that the correlation matrix R, and hence R_i, depends on a vector of correlation parameters denoted by $\boldsymbol{\alpha}$. These parameters are shared by all subjects. They represent the average dependence among the repeated observations across subjects.

The simplest correlation structure is independence, i.e. $R = I_T$. This is equivalent to saying that the longitudinal data are not correlated.

The next simplest structure is to assume that all the correlations in R are the same, or "exchangeable". This exchangeable structure is specified as $R(\boldsymbol{\alpha}) = \rho J_T + (1-\rho) I_T$, where J_T is a $T \times T$ matrix with all elements equal to 1; I_T is a $T \times T$ identity matrix.

Another useful one parameter model for the longitudinal data is the AR(1) covariance structure. This correlation structure is specified as $R(\boldsymbol{\alpha}) = \left\{\rho_{ij}\right\}_{T\times T} = \left\{\rho^{|i-j|}\right\}_{T\times T}$, where $\left\{\rho_{ij}\right\}_{T\times T}$ is a $T \times T$ matrix with elements ρ_{ij} in the i^{th} row and j^{th} column. Here the within-subject correlation over time is an exponential function of the lag between the two time points.

Define A_i to be the $T_i \times T_i$ diagonal matrix with $\pi_{it}(1-\pi_{it})$ as the t^{th} diagonal element. Then the working covariance matrix for $\mathbf{y}_i$ can be expressed as

$$V_i(\boldsymbol{\alpha}) = A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2} \tag{5}$$

Then the GEE estimator of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, is found by solving the following estimating equations (Liang and Zeger, 1986):

$$\sum_{i=1}^{K} D_i' \left[V_i(\hat{\boldsymbol{\alpha}})\right]^{-1} (\mathbf{y}_i - \boldsymbol{\pi}_i) = 0 \tag{6}$$

where $\mathbf{y}_i = [y_{i1}, \cdots, y_{iT_i}]'$, $\boldsymbol{\pi}_i = [\pi_{i1}, \cdots, \pi_{iT_i}]'$, $D_i = \frac{\partial \boldsymbol{\pi}_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, and $\hat{\boldsymbol{\alpha}}$, which can be estimated from the Pearson residuals $(\frac{y_{it}-\hat{\pi}_{it}}{[\hat{\pi}_{it}(1-\hat{\pi}_{it})]^{1/2}})$, is a consistent estimate of $\boldsymbol{\alpha}$ (Liang and Zeger, 1986). Assuming that missing data are missing completely at random in the sense of Rubin (1976), $\hat{\boldsymbol{\beta}}$ estimated from these equations are consistent and asymptotically normal even when the correlation matrix structure is misspecified as long as the marginal mean model is correct (Liang and Zeger, 1986). Because the estimating equation

only depends on the mean and variance of $\mathbf{y}$, the precise distribution of $\mathbf{y}$ is not required for estimating $\boldsymbol{\beta}$. The solution from the estimating equation is also called the quasi-likelihood estimate (Wedderburn, 1974).

If the marginal mean model is correct and the working correlation is correctly specified, then the model-based estimator of the covariance matrix of $\hat{\boldsymbol{\beta}}$, $Cov\left(\hat{\boldsymbol{\beta}}\right)$, is given by

$$\left[\sum_{i=1}^{K} D_i' V_i(\hat{\boldsymbol{\alpha}})^{-1} D_i\right]^{-1}$$

A more robust estimate of $Cov\left(\hat{\boldsymbol{\beta}}\right)$ can be made without assuming the structure of the working correlation is correctly specified. It is given by the following expression.

$$\left[\sum_{i=1}^{K} D_i' V_i(\hat{\boldsymbol{\alpha}})^{-1} D_i\right]^{-1} \left[\sum_{i=1}^{K} D_i' V_i(\hat{\boldsymbol{\alpha}})^{-1} Cov\left(\mathbf{y}_i\right) V_i(\hat{\boldsymbol{\alpha}})^{-1} D_i\right] \left[\sum_{i=1}^{K} D_i' V_i(\hat{\boldsymbol{\alpha}})^{-1} D_i\right]^{-1}$$

This estimator is often referred to as the "sandwich" estimator. The outer pieces of the sandwich are $Cov\left(\hat{\boldsymbol{\beta}}\right)$ when the structure of the working correlation is correctly specified and the center terms depend on the true correlation of the responses. $Cov\left(\mathbf{y}_i\right)$ can be consistently estimated by $\left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right)\left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right)'$ (Liang and Zeger, 1986).

1.3.4 Lack-Of-Fit Methods for Logistic GEE Model

Currently available LOF methods for logistic GEE models are mainly based on parametric model comparison. The assumed logistic GEE model is embedded in a more general parametric model. The general model reduces to the assumed logistic GEE model if some parameters are set equal to 0. If the general model is significantly different from the assumed logistic GEE model, namely the additional parameters in the general model are significantly different from 0, then the presence of Lack-Of-Fit for the assumed logistic GEE model is claimed. In addition, statistics based on residuals and kappa-like classification statistic have been previously proposed.

Statistics based on covariate partitioning Barnhart and Williamson (1998) developed a Lack-Of-Fit test statistic for logistic GEE model based on partitioning data using covariate patterns. This method can be thought of as an extension of Tsiatis (1988) for the ordinary logistic regression models.

Suppose a random sample of multivariate data $\mathbf{y}_i = [y_{i1}, \cdots, y_{iT_i}]'$, $i = 1, \cdots, K$, are observed. $\mathbf{y}_i's$ are independent of each other, but the T_i measurements within a cluster could be correlated with each other. $y_{it}'s$, $t = 1, \cdots, T_i$ and $i = 1, \cdots, K$ are binary data with values 0 or 1. The

assumed model is the logistic GEE model defined on page 25. For simplicity, assume $T_i = T$ for all i. The authors proposed to partition the covariate space $X = (x_1, \cdots, x_p)'$ into G distinct regions in $p-$dimensional space.

Define G group indicator variables as follows.

$$I_{itg} = \begin{cases} 1 \text{ if covariates at time } t \text{ for subject } i \text{ are in group } g, \\ 0 \text{ otherwise} \end{cases} \quad g = 1, ..., G$$

Let $\mathbf{I}_{it} = (I_{it1}, \cdots, I_{itG})'$ be a $G \times 1$ vector and $I_i = [\mathbf{I}_{i1}, \cdots, \mathbf{I}_{iT}]'$ be a $T \times G$ matrix. Let Z_T be a $T \times (T-1)$ matrix where the first row has entries 0 and the remaining $(T-1)$ rows form a $(T-1) \times (T-1)$ identity matrix, i.e. $Z_T = [\mathbf{0}\ I_{T-1}]'$. Let $S_i = [\mathbf{0}, diag\,(\mathbf{I}_{i1}, \cdots, \mathbf{I}_{iT})]'$ be a $T \times (T-1)\,G$ matrix and $\mathbf{0}$ be a $(T-1)\,G \times 1$ vector of zeros.

Consider the following general model.

$$\text{logit}(\boldsymbol{\pi}_i) = X_i\boldsymbol{\beta} + Z_T\boldsymbol{\tau} + I_i\boldsymbol{\gamma} + S_i\boldsymbol{\rho}$$

Note that $\boldsymbol{\tau}$ is the $(T-1) \times 1$ vector of time effects (the first time effect is set to be 0). $\boldsymbol{\gamma}$ is a $G \times 1$ region effect vector. $\boldsymbol{\rho}$ is a $(T-1)\,G \times 1$ time and region interaction vector because each column of S_i results from componentwise multiplication of two column vectors, one column from Z_T and the other from I_i. Note that the assumed logistic GEE model defined on page 25 is embedded

in this general model. A Lack-Of-Fit statistic consists of testing $H_0 : \boldsymbol{\theta} = \mathbf{0}$, where $\boldsymbol{\theta} = [\boldsymbol{\tau}', \boldsymbol{\gamma}', \boldsymbol{\rho}']'$ is a $J \times 1$ vector with $J = (T-1) + G + (T-1)\,G$.

Let $\mathbf{U}$ be the $L = (p + 1 + J)$ vector with l^{th} component

$$U_l = \sum_{i=1}^{K} \hat{\mathbf{D}}_{il} \hat{V}_i^{-1} \left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right) \; l = 1, ..., L$$

where $\hat{\mathbf{D}}_{il} = \frac{\partial \hat{\boldsymbol{\pi}}_i}{\partial \beta_l}$ for $l \leqslant p + 1$, $\hat{\mathbf{D}}_{il} = \frac{\partial \hat{\boldsymbol{\pi}}_i}{\partial \theta_{l-p-1}}$ for $l > p + 1$, where $\hat{\boldsymbol{\pi}}_i = \text{logit}^{-1}\left(X_i\hat{\boldsymbol{\beta}} + Z_T\boldsymbol{\tau} + I_i\boldsymbol{\gamma} + S_i\boldsymbol{\rho}\right)$, and $\hat{\boldsymbol{\beta}}$ is the GEE estimator obtained from the assumed logistic regression model defined in page25. Then under H_0 : $\boldsymbol{\theta} = \mathbf{0}$, the asymptotic distribution of $\mathbf{U}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix (Liang and Zeger, 1986)

$$W_R = \sum_{i=1}^{K} \hat{D}_i \hat{V}_i^{-1} Cov\left(\mathbf{y}_i\right) V_i^{-1} \hat{D}_i$$

where $\hat{D}_i = \left[\hat{\mathbf{D}}_{i1}, \cdots, \hat{\mathbf{D}}_{iL}\right]$ is a $T \times L$ matrix. $Cov\left(\mathbf{y}_i\right)$ can be consistently estimated by $\left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right)\left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right)'$ (Liang and Zeger, 1986). If the correlation matrix $R\left(\boldsymbol{\alpha}\right)$ is correctly specified, then the asymptotic covariance of $\mathbf{U}$ reduces to

$$W = \sum_{i=1}^{K} \hat{D}_i \hat{V}_i^{-1} \hat{D}_i$$

Let $\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \\ \mathbf{U}_2 \end{pmatrix}$, $\quad W_R = \begin{pmatrix} A_R & B_R' \\ \\ B_R & C_R \end{pmatrix}$, and $\quad W = \begin{pmatrix} A & B' \\ \\ B & C \end{pmatrix}$

be the partitioning for $\mathbf{U}$, W_R, and W, where $\mathbf{U}_2$ is a $J \times 1$ vector and C_R and C are $J \times J$ matrices. Under $H_0 : \boldsymbol{\theta} = \mathbf{0}$, both the proposed model-based LOF test statistic

$$Q_M = \mathbf{U}_2' \left(C - BA^{-1}B'\right)^{-} \mathbf{U}_2$$

and the proposed robust LOF test statistic

$$Q_R = \mathbf{U}_2' \left(C_R - B_R A_R^{-1} B_R'\right)^{-} \mathbf{U}_2$$

have asymptotic chi-square distribution with degrees of freedom equals to

$$rank\left(\left(C - BA^{-1}B'\right)^{-}\right) = rank\left(\left(C_R - B_R A_R^{-1} B_R'\right)^{-}\right)$$

where M^{-} is any generalized inverse of matrix M.

The authors reported the statistics had high power for detecting omission of a quadratic term but low power for detecting omission of an interaction term. This method has also been shown to have low power for detecting omitted covariates (Evans and Li, 2005).

A statistic using groups based on ranked estimated probabilities

Horton et al. (1999) developed a statistic using ranked estimated probabilities. This method can be thought of as an extension of Hosmer and Lemeshow (1980) for ordinary logistic regression.

Suppose a random sample of multivariate data $\mathbf{y}_i = [y_{i1}, \cdots, y_{iT}]'$, $i = 1, \cdots, K$, is observed. $\mathbf{y}_i's$ are independent of each other, but the T_i measurements within the cluster could be correlated with each other. $y_{it}'s$, $t = 1, \cdots, T_i$ and $i = 1, \cdots, K$ are binary data with values 0 or 1. The assumed model is logistic GEE model defined on page 25.

The proposed LOF test method forms G groups of approximately equal sizes (decile of risk groups) in the following manner, where G is 10 or an integer close to 10.

1. The first group contains $\sum_{i=1}^{K} T_i/G$ observations with the smallest values of fitted probabilities $\hat{\pi}_{it}$.

2. The second group contains $\sum_{i=1}^{K} T_i/G$ observations with the next smallest values of fitted probabilities $\hat{\pi}_{it}$.

$\vdots$

G. The last group contains $\sum_{i=1}^{K} T_i/G$ observations with the biggest values of fitted probabilities $\hat{\pi}_{it}$.

In general, we could form G groups, with approximately $\sum_{i=1}^{K} T_i/G$ observations in each group. Since subject i can have different $\hat{\pi}_{it}'s$ for each of the T_i observations, a subject's group membership, g, can change for different t.

Then $(G-1)$ group indicators are defined as

$$I_{itg} = \begin{cases} 1 \text{ if } \hat{\pi}_{it} \text{ is in group } g, & g = 1, \cdots, G-1 \\ 0 \text{ otherwise} \end{cases}$$

Consider the following general model.

$$logit\left(\pi_{it}\right) = \boldsymbol{\beta}'\mathbf{x}_{it} + \boldsymbol{\theta}'\mathbf{I}_{it}$$

where $\mathbf{I}_{it} = [I_{it1}, \cdots, I_{itG-1}]'$. Note that $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_{G-1}]'$ represents region effects. When $\boldsymbol{\theta} = \mathbf{0}$, the general model reduces to the assumed logistic GEE model. A Lack-Of-Fit test of the assumed model is equivalent to test $H_0 : \boldsymbol{\theta} = \mathbf{0}$. The proposed LOF test statistic is shown below. Under H_0, it has asymptotic chi-square distribution with $G-1$ degrees of freedom.

$$X_H^2 = \mathbf{u}_2\left(\hat{\boldsymbol{\beta}}, \mathbf{0}\right)' \left\{\hat{V}ar\left[\mathbf{u}_2\left(\hat{\boldsymbol{\beta}}, \mathbf{0}\right)\right]\right\}^{-1} \mathbf{u}_2\left(\hat{\boldsymbol{\beta}}, \mathbf{0}\right)$$

where

$$\mathbf{u}\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right) = \begin{bmatrix} \mathbf{u}_1\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right) \\ \mathbf{u}_2\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right) \end{bmatrix} = \sum_{i=1}^{K} \begin{bmatrix} D_{1i}' V_i^{-1}\left[\mathbf{y}_i - \boldsymbol{\pi}_i\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right)\right] \\ D_{2i}' V_i^{-1}\left[\mathbf{y}_i - \boldsymbol{\pi}_i\left(\boldsymbol{\beta}, \boldsymbol{\theta}\right)\right] \end{bmatrix}$$

with $D_{1i} = \frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}}$, $D_{2i} = \frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\theta}}$ and

$\hat{\boldsymbol{\beta}}$ is obtained by solving $\mathbf{u}_1\left(\boldsymbol{\beta}, \mathbf{0}\right) = \mathbf{0}$ and

$$\hat{V}ar\left[\mathbf{u}_2\left(\hat{\boldsymbol{\beta}}, \mathbf{0}\right)\right] = \left[-AB^{-1}, I\right] C \left[-AB^{-1}, I\right]$$

where

$$A = E\left\{\frac{\partial \mathbf{u}_2(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right\}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{\theta}=\mathbf{0}},\ B = E\left\{\frac{\partial \mathbf{u}_1(\boldsymbol{\beta},\boldsymbol{\theta})}{\partial \boldsymbol{\beta}}\right\}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\boldsymbol{\theta}=\mathbf{0}},$$

$$C = K^{-1}\sum_{i=1}^{K}\begin{bmatrix} \mathbf{u}_{1i}\left(\hat{\boldsymbol{\beta}},\mathbf{0}\right) \\ \mathbf{u}_{2i}\left(\hat{\boldsymbol{\beta}},\mathbf{0}\right) \end{bmatrix}\begin{bmatrix} \mathbf{u}_{1i}\left(\hat{\boldsymbol{\beta}},\mathbf{0}\right) \\ \mathbf{u}_{2i}\left(\hat{\boldsymbol{\beta}},\mathbf{0}\right) \end{bmatrix}'$$

This method has been reported to have high power detecting omitted quadratic term, but low power detecting an omitted interaction term or additional covariates (Evans and Li, 2005).

Hybrid Statistics Evans and Li (2005) proposed a Lack-Of-Fit method for logistic GEE model based on partitioning data using both covariates and ranked fitted probabilities. This method can be thought of as an extension of Pulkstenis and Robinson (2002) for ordinary logistic regression.

Suppose a random sample of multivariate data $\mathbf{y}_i = [y_{i1}, \cdots, y_{iT_i}]$, $i = 1, \cdots, K$, is observed. $\mathbf{y}_i's$ are independent of each other, but the T_i measurements within the cluster could be correlated with each other. $y_{it}'s$, $t = 1, \cdots, T_i$ and $i = 1, \cdots, K$ are binary data with values 0 or 1. The assumed model is logistic GEE model defined in page25.

The proposed method first partitions the covariate space into G distinct regions using categorical covariates. Then each region is further partitioned into 2 parts based on the fitted probabilities from logistic GEE model defined

in page25 within each region. Then each observation $(y_{it}, \mathbf{x}_{it})$ belongs to one of $2G$ distinct regions.

Define $(2G-1)$ group indicators.

$$I_{itg} = \begin{cases} 1 \text{ if } \hat{\pi}_{it} \text{ is in group } g, & g = 1, \cdots, 2G-1 \\ 0 \text{ otherwise} \end{cases}$$

Let $\mathbf{I}_{it} = [I_{it1}, \cdots, I_{it2G-1}]'$ be a $(2G-1) \times 1$ vector.

Consider the following general model.

$$logit\,(\boldsymbol{\pi}_i) = X_i\boldsymbol{\beta} + I_i\boldsymbol{\theta}$$

where $I_i = [\mathbf{I}_{i1}, ..., \mathbf{I}_{iT}]'$ is a $T \times (2G-1)$ design matrix. Note that $\boldsymbol{\theta}$ is a $(2G-1) \times 1$ vector of region effects.

Let $\mathbf{U}$ be the $L = (p + 1 + 2G - 1)$ vector with l^{th} component

$$U_l = \sum_{i=1}^{K} \hat{\mathbf{D}}_{il} \hat{V}_i^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i) \; l = 1, ..., L$$

where $\hat{\mathbf{D}}_{il} = \frac{\partial \hat{\boldsymbol{\pi}}_i}{\partial \beta_l}$ for $l \leqslant p+1$, $\hat{\mathbf{D}}_{il} = \frac{\partial \hat{\boldsymbol{\pi}}_i}{\partial \theta_{l-p-1}}$ for $l > p+1$, where $\hat{\boldsymbol{\pi}}_i = \log it^{-1}\left(X_i\hat{\boldsymbol{\beta}} + I_i\boldsymbol{\theta}\right)$, and $\hat{\boldsymbol{\beta}}$ is the GEE estimator obtained from the assumed logistic regression model defined in page25. Then under $H_0 : \boldsymbol{\theta} = \mathbf{0}$, the asymptotic distribution of $\mathbf{U}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix (Liang and Zeger, 1986)

$$W_R = \sum_{i=1}^{K} \hat{D}_i \hat{V}_i^{-1} Cov\left(\mathbf{y}_i\right) V_i^{-1} \hat{D}_i$$

where $\hat{D}_i = \left[\hat{\mathbf{D}}_{i1}, \cdots, \hat{\mathbf{D}}_{iL}\right]$ is a $T \times L$ matrix. $Cov\left(\mathbf{y}_i\right)$ can be consistently estimated by $\left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right)\left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right)'$ (Liang and Zeger, 1986). If the correlation matrix $R\left(\boldsymbol{\alpha}\right)_i$ is correctly specified, then the asymptotic covariance of $\mathbf{U}$ reduces to

$$W = \sum_{i=1}^{K} \hat{D}_i \hat{V}_i^{-1} \hat{D}_i$$

Let

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \\ \mathbf{U}_2 \end{pmatrix} \quad W_R = \begin{pmatrix} A_R & B_R' \\ \\ B_R & C_R \end{pmatrix} \quad W = \begin{pmatrix} A & B' \\ \\ B & C \end{pmatrix}$$

be the partitioning for $\mathbf{U}$, W_R, and W, where $\mathbf{U}_2$ is a $J \times 1$ vector and C_R and C are $J \times J$ matrices. Under $H_0 : \boldsymbol{\theta} = \mathbf{0}$, both the proposed model-based LOF test statistic

$$N_M = \mathbf{U}_2' \left(C - BA^{-1}B'\right)^{-} \mathbf{U}_2$$

and the proposed robust LOF test statistic

$$N_R = \mathbf{U}_2' \left(C_R - B_R A_R^{-1} B_R'\right)^{-} \mathbf{U}_2$$

have asymptotic chi-square distribution with degrees of freedom equals to

$$rank\left(\left(C - BA^{-1}B'\right)^{-}\right) = rank\left(\left(C_R - B_R A_R^{-1} B_R'\right)^{-}\right)$$

where M^{-} is any generalized inverse of matrix M.

The methods were reported to have low power for detecting an omitted quadratic term, interaction term or additional covariates (Evans and Li, 2005).

A Classification Statistic Suppose a random sample of multivariate data $\mathbf{y}_i = [y_{i1}, \cdots, y_{iT_i}]'$, $i = 1, \cdots, K$, is observed. $\mathbf{y}_i's$ are independent of each other, but the T_i measurements within a cluster could be correlated with each other. $y_{it}'s$, $t = 1, \cdots, T_i$ and $i = 1, \cdots, K$ are binary data with values 0 or 1. The assumed model is logistic GEE model defined in page25.

Williamson et al. (2003) developed a kappa-like classification statistic for assessing LOF for logistic GEE models. Historically, kappa has been used to determine the agreement of binary (Cohen, 1960) and categorical (Fleiss, 1971) outcomes between raters. Kappa corrects the percentage of agreement between raters by taking into account the proportion of agreement expected by chance.

The general expression for the kappa statistic is

$\kappa = \frac{P_o - P_e}{1 - P_e}$

where P_o is the observed proportion of agreement and P_e is the proportion of agreement expected by chance alone (Fleiss, 1971). A value of 0 for κ indicates no agreement beyond chance and a value of 1 indicates perfect agreement. Larger values of κ indicate greater agreement between the outcomes (Fleiss, 1971). Willamson et al. estimated P_e by fitting an intercept-only model proposed by Cox and Snell (1989) and Nagelkeke (1991). For simplicity, assume $T_i = T$ for all i. The estimated probabilities from the intercept-only model will be the same for all subjects and all time points, and can be denoted as

$$\hat{P}_{itg} = \hat{P}_g = n_g/TK$$

where n_g is the number of observations equals to 0 or 1 for $g = 0, 1$. All n_g observations with response values 0 or 1 will each be correctly predicted with probability $\hat{P}_g$; accordingly the estimate of P_e is

$$\hat{P}_e = \sum_{g=0}^{1}\sum_{i=1}^{K}\sum_{j=1}^{T} I\left(Y_{it} = g\right) = \sum_{g=0}^{1} n_g \hat{P}_g/TK = \left[\left(\sum_{i=1}^{K}\sum_{j=1}^{T} y_{ij}\right)/TK\right]^2 + \left[1 - \left(\sum_{i=1}^{K}\sum_{j=1}^{T} y_{ij}\right)/TK\right]^2$$

Define P_{oit} to be the probability that the predicted response from the assumed model for t^{th} observation in subject i is equal to the observed response, i.e. $y_{it} = \hat{y}_{it}$. A natural estimate of P_{oit} is $\hat{\zeta}_{it} = \hat{\pi}_{it}^{y_{it}}\left(1 - \hat{\pi}_{it}\right)^{1-y_{it}}$. Let

$\mathbf{P}_{oi}$ and $\mathbf{U}_i$ denote the $T \times 1$ vectors $[P_{oi1}, \cdots, P_{oiT}]'$ and $[\zeta_{i1}, \cdots, \zeta_{iT}]'$. They estimate an overall κ to assess the fit of the assumed logistic GEE model. Here "overall" means $\kappa = \kappa_{it}$ for $i = 1, \cdots, K$ and $t = 1, \cdots, T$. Noting that $P_{oit} = P_e + \kappa(1 - P_e)$, they estimate κ by solving the following estimating equation together with the generalized estimating equations (page27)

$$\sum_{i=1}^{K} \mathbf{C}_i' W_i^{-1} \{\mathbf{U}_i(\boldsymbol{\beta}) - \mathbf{P}_{oi}(\kappa)\} = 0$$

where $\mathbf{C}_i = d\mathbf{P}_{oi}/d\kappa = \left[1 - \hat{P}_e, ..., 1 - \hat{P}_e\right]'$ is a $T \times 1$ vector and $W_i \simeq Var(\mathbf{U}_i)$ is the $T \times T$ working covariance matrix of $\mathbf{U}_i$.

The kappa measure intuitively estimates the probability of being correctly predicted by the fitted model and this probability is corrected for chance. An advantage of this statistic is that no subjective decision need to be made concerning partitioning. However, interpretation of the statistic is not trivial since no distribution of the statistic is given. As in Landis and Koch (1977), Williamson et al. recommended that: a κ value from 0 to 0.2 represents poor fit, a value from 0.21 to 0.4 represents fair fit, a value from 0.41 to 0.6 represents good fit, and a value from 0.61 to 1 represents excellent fit.

Statistics based on residuals Suppose a random sample of multivariate data $\mathbf{y}_i = [y_{i1}, \cdots, y_{iT_i}]'$, $i = 1, \cdots, K$, is observed. The $\mathbf{y}_i's$ are independent

of each other, but the T_i measurements within the cluster could be correlated with each other. $y'_{it}s$, $t = 1, \cdots, T_i$ and $i = 1, \cdots, K$ are binary data with values 0 or 1. The assumed model is logistic GEE model defined in page 25.

Pan (2002), and Evans (1998) developed two statistics to assess the LOF of the assumed logistic GEE model based on comparing observed versus predicted values. For simplicity, assume $T_i = T$ for all i.

Let $\mathbf{Y} = \left[\mathbf{Y}'_1, \cdots, \mathbf{Y}'_K\right]'$ be the $TK \times 1$ response vector for all the measurements in all K subjects, $\boldsymbol{\pi} = \left[\boldsymbol{\pi}'_1, \cdots, \boldsymbol{\pi}'_K\right]'$ be the $TK \times 1$ probability of positive response vector corresponding to $\mathbf{Y}$, $A = diag\,(A_1, \cdots, A_K)$, $V = diag\,(V_1, \cdots, V_K)$ where A_i and V_i are defined as in the definition of working covariance matrix on page27, $X = \left[\left(1\ \mathbf{x}'_{11}\right)',\ \left(1\ \mathbf{x}'_{12}\right)',\ \cdots,\ \left(1\ \mathbf{x}'_{K,T}\right)'\right]'$ be the $(TK) \times (p+1)$ design matrix including the intercept,

$H = AX\left(X'AV^{-1}AX\right)^{-1}X'AV^{-1}$, and $\mathbf{e} = \mathbf{Y} - \boldsymbol{\pi}$.

Pan shows that the residuals can be approximated by using the following expression.

$$\mathbf{Y} - \hat{\boldsymbol{\pi}} \simeq (I - H)(\mathbf{Y} - \boldsymbol{\pi})$$

The Pearson weighted sums of square test statistic is

$$G = \sum_{i=1}^{K}\sum_{t=1}^{T}\frac{(y_{it}-\hat{\pi}_{it})^2}{\hat{\pi}_{it}(1-\hat{\pi}_{it})} = KT + (\mathbf{1} - 2\hat{\boldsymbol{\pi}})'\hat{A}^{-1}\hat{\mathbf{e}} \simeq KT + (\mathbf{1} - 2\hat{\boldsymbol{\pi}})'\hat{A}^{-1}(I - H)\mathbf{e}$$

with $\widehat{E(G)} = KT$, and

$\widehat{Var(G)} = (\mathbf{1} - 2\hat{\boldsymbol{\pi}})' \hat{A}^{-1} \left(I - \hat{H}\right) \widehat{cov(\mathbf{Y})} \left(I - \hat{H}'\right) \hat{A}^{-1} (\mathbf{1} - 2\hat{\boldsymbol{\pi}})$. Pan used 2 estimates of $Cov(\mathbf{Y})$. The first is the empirical covariance estimator: $\widehat{Cov(\mathbf{Y})}_e = diag\left(\widehat{Cov(\mathbf{Y}_1)}, ..., \widehat{Cov(\mathbf{Y}_K)}\right)$, where $\widehat{Cov(\mathbf{Y}_i)} = (\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i)'$. The second is, $\widehat{Cov(\mathbf{Y})}_u = \hat{A}^{1/2} diag\left(\hat{R}_u, ..., \hat{R}_u\right) \hat{A}^{1/2}$, where $\hat{R}_u$ is the unstructured correlation matrix estimate, specifically,

$$\hat{R}_u = \frac{1}{K} \sum_{i=1}^{K} \hat{A}_i^{-1/2} (\mathbf{Y}_i - \hat{\boldsymbol{\pi}}_i) (\mathbf{Y}_i - \hat{\boldsymbol{\pi}}_i)' \hat{A}_i^{-1/2}$$

The corresponding test statistics are denoted as G1 and G2 respectively.

The unweighted sums of squares statistic is defined as

$$\begin{aligned} U &= \sum_{i=1}^{K} \sum_{t=1}^{T} (y_{it} - \hat{\pi}_{it})^2 = \hat{\boldsymbol{\pi}}' (\mathbf{1} - \hat{\boldsymbol{\pi}}) + (\mathbf{1} - 2\hat{\boldsymbol{\pi}})' \hat{\mathbf{e}} \\ &\simeq \hat{\boldsymbol{\pi}}' (\mathbf{1} - \hat{\boldsymbol{\pi}}) + (\mathbf{1} - 2\hat{\boldsymbol{\pi}})' (I - H) \mathbf{e} \end{aligned}$$

Its mean and variance are approximately

$$\begin{aligned} &\widehat{E(U)} = \hat{\boldsymbol{\pi}}' (\mathbf{1} - \hat{\boldsymbol{\pi}}), \\ &\widehat{Var(U)} = (\mathbf{1} - 2\hat{\boldsymbol{\pi}})' \left(I - \hat{H}\right) \widehat{Cov(\mathbf{Y})} \left(I - \hat{H}'\right) (\mathbf{1} - 2\hat{\boldsymbol{\pi}}) \end{aligned}$$

and $Cov(\mathbf{Y})$ is estimated by either $\widehat{Cov(\mathbf{Y})}_e$ or $\widehat{Cov(\mathbf{Y})}_u$ described above. The test statistics are denoted by U1 and U2 respectively. Both $G1$, $G2$ and

$U1$, $U2$ have approximately standard normal distributions upon standardization.

1.4 Proposed LOF Test

Although logistic GEE model has been widely used and accepted as a standard method for analyzing clustered binary data, relatively little work has been done to assess the adequacy of the fitted model. The purpose of this research is to develop a new test statistic for assessing the Lack-Of-Fit for a logistic GEE model. This method is designed to enable one to determine overall Lack-Of-Fit of a logistic GEE model.

There are 5 explicit objectives for the proposed research.

1. Develop a new method for assessing LOF of a logistic GEE model based on comparing the assumed logistic GEE model with a more general model.

2. Develop the asymptotic null distribution of the proposed test statistic and verify the distribution using extensive simulation in a number of different scenarios.

3. Investigate the power of the test to detect a variety of departures from the logistic GEE model via simulation.

4. Compare the performance of the proposed method with some currently available LOF methods for logistic GEE models.

5. Demonstrate the application of the proposed method using real examples.

1.4.1 Description of the Proposed Method

Suppose a random sample of multivariate data $\mathbf{y}_i = [y_{i1}, \cdots, y_{iT_i}]'$, $i = 1, \cdots, K$, are observed. $\mathbf{y}_i's$ are independent of each other, but the T_i measurements within the cluster could be correlated with each other. $y_{it}'s$, $t = 1, \cdots, T_i$ and $i = 1, \cdots, K$ are binary data with values 0 or 1. If missing values are present, they are assumed to be missing completely at random in the sense of Rubin (1976). The primary interest is in the dependence of the marginal mean of y_{it} on $(p+1) \times 1$ covariate vector $\mathbf{x}_{it} = [1, x_{it1}, \cdots, x_{itp}]$. The assumed model is logistic GEE1 model. The marginal regression of the model can be expressed as

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \mathbf{x}_{it}^{'}\boldsymbol{\beta}, \quad t = 1, \cdots, T_i \text{ and } i = 1, \cdots, K \tag{7}$$

or more concisely for each subject

$$logit\,(\boldsymbol{\pi}_i) = X_i\boldsymbol{\beta}$$

where $logit\left(\boldsymbol{\pi}_i\right) = \left[logit\left(\pi_{i1}\right), \cdots, logit\left(\pi_{iT_i}\right)\right]'$, $X_i = \left[\mathbf{x}_{i1}, \cdots, \mathbf{x}_{iT_i}\right]'$.

$\boldsymbol{\beta} = \left[\beta_0, \beta_1, \cdots, \beta_p\right]'$ is a $(p+1) \times 1$ population parameters vector.

We assume that all the subjects are set out to be measured at the same T time points. For each subject, measurements are available at $T_i \left(T_i \leq T\right)$ time points. The joint distribution of $\mathbf{y}_i$ is not specified. All the clusters share the same correlation for the measurements within clusters. A working correlation structure is specified, such as exchangeable, or AR(1). The working correlation R depends on a vector of parameters $\boldsymbol{\alpha}$. These parameters $\boldsymbol{\alpha}$ are assumed to be the same for all subjects. $\boldsymbol{\alpha}$ contains nuisance parameters. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ do not depend on each other. The working correlation for each cluster R_i is obtained by removing corresponding rows and columns from R according to which time points are not measured for the subject. Defining $A_i = diag\left[\pi_{i1}\left(1-\pi_{i1}\right), \cdots, \pi_{iT}\left(1-\pi_{iT_i}\right)\right]$, then the working covariance matrix V_i can be expressed as $A_i^{1/2} R_i A_i^{1/2}$.

The parameters $\boldsymbol{\beta}$ are estimated by solving the following generalized estimating equations.

$$\sum_{i=1}^{K} D_i V_i(\hat{\boldsymbol{\alpha}})^{-1} \left(\mathbf{y}_i - \boldsymbol{\pi}_i\right) = \mathbf{0}$$

where $D_i = \frac{\partial \boldsymbol{\pi}_i}{\partial \boldsymbol{\beta}}$, $\hat{\boldsymbol{\alpha}}$ is a consistent estimate of $\boldsymbol{\alpha}$, which can be estimated

using Pearson residuals ($\frac{y_{it}-\hat{\pi}_{it}}{[\hat{\pi}_{it}(1-\hat{\pi}_{it})]^{1/2}}$) by method of moments (Liang and Zeger, 1986).

This model may not be adequate to model the collected data. We define Lack-Of-Fit of a logistic GEE model by extending the method of Hosmer et al (1997) for standard logistic regression model. Evidence of Lack-Of-Fit may come from violation of one or more of the following 3 assumptions for marginal logistic regression model.

(A1) the logit transformation is the correct function linking the covariates $\mathbf{x}$ with the marginal conditional mean $E(Y|\mathbf{x}) = \pi(\mathbf{x})$, i.e. $\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \mathbf{x}'\boldsymbol{\beta}$;

(A2) the linear predicting function, $\mathbf{x}'\boldsymbol{\beta}$, is correct;

(A3) the marginal conditional variance of the response is the variance of Bernoulli distribution, i.e. $Var(Y|\mathbf{x}) = \pi(\mathbf{x})[1-\pi(\mathbf{x})]$.

In the proposed method, the data are divided into 2 groups using the median of the fitted probabilities. The first group contains the observations with the fitted probabilities $\hat{\pi}_{it}$ smaller than the median. The second group contains the observations with the fitted probabilities $\hat{\pi}_{it}$ equal to or bigger than the median.

In general, we could form 2 groups with approximately equal numbers of

observations. Since subject i can have different $\hat{\pi}'_{it}s$ for each of the T_i observations, each subject i can have multiple group memberships at T_i different time points.

Define one group indicator variable as follows.

$$I_{it} = \begin{cases} 1 \text{ if the covariates at time } t \text{ for subject } i \text{ is in group 1,} \\ 0 \text{ otherwise} \end{cases}$$

Consider the following piecewise logistic regression model (**??**).

$$\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \mathbf{x}'_{it}\boldsymbol{\beta} + \mathbf{x}'_{it}\boldsymbol{\theta}I_{it}; \quad i = 1, \cdots, K; t = 1, \cdots, T_i \tag{8}$$

where $\boldsymbol{\theta}$ is a vector of unknown parameters; The piecewise models for different groups are linear with different slopes and different intercepts. Notice that the general model reduces to the fitted model if $\boldsymbol{\theta} = \mathbf{0}$. Therefore the LOF test is equivalent to test if $\boldsymbol{\theta} = \mathbf{0}$ or not. For simplicity, we assume $T_i = T, i = 1, 2, \cdots, K$.

Let $\mathbf{U}$ be the the $L = (2p+2) \times 1$ vector with l^{th} component

$$U_l = \sum_{i=1}^{K} \hat{\mathbf{D}}_{il} \hat{V}_i^{-1} \left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right) \; l = 1, \cdots, L$$

where $\hat{\mathbf{D}}_{il} = \frac{\partial \hat{\boldsymbol{\pi}}_i}{\partial \beta_l}$ for $l \leqslant p+1$, $\hat{\mathbf{D}}_{il} = \frac{\partial \hat{\boldsymbol{\pi}}_i}{\partial \theta_{l-p-1}}$ for $l > p+1$, where $\hat{\boldsymbol{\pi}}_i = \left[\text{logit}^{-1}\left(\mathbf{x}_{i1}^{'}\hat{\boldsymbol{\beta}} + \mathbf{x}_{i1}^{'}\boldsymbol{\theta} I_{i1}\right), \text{logit}^{-1}\left(\mathbf{x}_{i2}^{'}\hat{\boldsymbol{\beta}} + \mathbf{x}_{i2}^{'}\boldsymbol{\theta} I_{i2}\right), \cdots, \text{logit}^{-1}\left(\mathbf{x}_{iT}^{'}\hat{\boldsymbol{\beta}} + \mathbf{x}_{iT}^{'}\boldsymbol{\theta} I_{iT}\right)\right]^{'}$, and $\hat{\boldsymbol{\beta}}$ is the GEE estimator obtained from the general logistic regression model (8) with $\boldsymbol{\theta} = \mathbf{0}$. Then under $H_0 : \boldsymbol{\theta} = \mathbf{0}$, the asymptotic distribution of $\mathbf{U}$ is multivariate normal with mean $\mathbf{0}$ and covariance matrix (Liang and Zeger, 1986).'

$$W_R = \sum_{i=1}^{K} \hat{D}_i \hat{V}_i^{-1} Cov\left(\mathbf{y}_i\right) \hat{V}_i^{-1} \hat{D}_i$$

where $\hat{D}_i = \left[\hat{\mathbf{D}}_{i1}, ..., \hat{\mathbf{D}}_{iL}\right]$ is a $T \times L$ matrix. $Cov\left(\mathbf{y}_i\right)$ can be consistently estimated by $\left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right)\left(\mathbf{y}_i - \hat{\boldsymbol{\pi}}_i\right)^{'}$ (Liang and Zeger, 1986).

Partition $\mathbf{U}$ and W_R respectively into the form:

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_1 \\ \\ \mathbf{U}_2 \end{pmatrix} \qquad W_R = \begin{pmatrix} A_R & B_R^{\prime} \\ \\ B_R & C_R \end{pmatrix}$$

where $\mathbf{U}_2$ is a $(P+1) \times 1$ vector and C_R is a $(P+1) \times (P+1)$ matrix. The proposed LOF test statistic is

$$Y_R = \mathbf{U}_2^{'}\left(C_R - B_R A_R^{-1} B_R^{'}\right)^{-} \mathbf{U}_2$$

Under $H_0 : \boldsymbol{\theta} = \mathbf{0}$, as $K \to \infty$, it has chi-square distribution with degrees of freedom equal to

$$rank\left(\left(C_R - B_R A_R^{-1} B_R^{'}\right)^{-}\right)$$

where M^{-} is any generalized inverse of matrix M.

When Y_R is large, the presence of Lack-Of-Fit for the assumed logistic regression model (7) is claimed.

The proposed method can be extended in the following ways. First the number of partitioned groups can be more than 2. Second the piecewise models can be other than linear models. The proposed method can be extended by adopting a data partitioning scheme similar to Pulkstenis and Robinson (2002). That is to partition the data based on categorical covariates (including time points within each subject when the number of time points are small) first. Then within each covariate pattern the data are further partitioned into two subgroups using ranked fitted probabilities from the assumed logistic GEE model defined in page 45.

The proposed method has connections with some currently available LOF methods. This method can be thought of as an extension of Hosmer and Lemeshow (1980) and Liu and Yang (2007) for ordinary logistic regression model. Our approach differs from Liu and Yang (2007) in that we utilize Liang and Zeger's generalized score test statistic whereas Liu and Yang use a likelihood ratio test statistic and that our method is applicable to clustered

data, whereas Liu and Yang applies to independent data. The proposed approach is also an extension of Horton (1999)'s method for logistic GEE model. Horton's method can be thought of as a special case of the proposed method. In Horton's method, the submodels for different groups, apart from a group specific intercept, are the same. However, in the proposed method, submodels for different groups can be quite different in forms.

1.4.2 Computation

In order to conduct simulation study about logistic GEE model, we need to generate pseudo random clustered binary data.

Let Y denote an $K \times 1$ vector of Bernoulli random variables $(Y_1, \cdots, Y_K)'$, with $E(Y) = (\mu_1, \cdots, \mu_K)' = \boldsymbol{\mu}$, $corr(Y) = \{r_{ij}\} = R$ and $cov(Y) = \{v_{ij}\} = V$. For Bernoulli variables $var(Y_i) = \mu_i(1 - \mu_i)$, therefore specifying ($\boldsymbol{\mu}$,$V$) is equivalent to specifying ($\boldsymbol{\mu}, R$). Suppose there are $2^K = m$ possible configurations of Y, and let p_j denote the probability of configuration j. The vector $\mathbf{p}$ takes values in the following set

$$\left\{ \mathbf{p} : \sum_{j=1}^{m} p_j = 1, p_j \geq 0 \right\}$$

We denote the mean of the Y as $g(\mathbf{p})$ and the covariance matrix of Y as $h(\mathbf{p})$. $g(\mathbf{p})$ and $h(\mathbf{p})$ are functions of $\mathbf{p}$.

If $\mathbf{p}$ is specified, it is straightforward to simulate Y. However specifying or computing $\mathbf{p}$ so that $g(\mathbf{p})$ and $h(\mathbf{p})$ are equal to some prespecified mean vetror and covariance matrix becomes impractical for a big cluster size.

Many methods have been proposed to tackle this problem. Emrich and Piedmonte (1991) proposed a method based on threadshoding multivariate normal variables. Lee (1993) developed a method based on copulas. Park et al. (1996) proposed a method based on generating correlated Poisson random variables. Lunn and Davies (1998) proposed a method based on a finite mixture of Bernoulli variables. Oman and Zucker (2001) proposed a method based on a finite mixture of continuous variables. Kang and Jung (2001) proposed a method based on computing joint distribution of the multivariate binary variable. Leish et al. (1998) proposed a method based on threasholding multivariate continuous variables. Qaqish (2003) proposed a method based on a conditional linear family of multivariate binary data. Qaqish's method is attractive in simulation study because of the following reasons. It allows unequal means for observations within the same cluster and both positive and negative correlation between observations within the same cluster. It avoids computing the joint distribution of the multivariate variable. The computation time for this task grows exponentially as the size

of the cluster increases. It has been adopted in some simulation study such as Evans and Li (2004). We used this method in our work. The method is described below.

For $i = 2, \cdots, K$ define $X_i \equiv (Y_1, \cdots, Y_{i-1})'$, $\boldsymbol{\theta}_i$=$E(X_i)$, $G_i = cov(X_i)$ and $s_i = cov(X_i, Y_i)$.

The conditional linear family is obtained by letting

$$E(Y_i|X_i = \mathbf{x}_i) = \mu_i + \mathbf{b}_i'(\mathbf{x}_i - \boldsymbol{\theta}_i) \qquad i = 2, , \cdots, K$$

This happens to be just enough restrictions to determine a unique $\mathbf{p}$ such that $g(\mathbf{p}) = \boldsymbol{\mu}$ and $h(\mathbf{p}) = V$ are equal to some prespecified mean vector and covariance matrix . Furthermore, for given $(\boldsymbol{\mu}, V)$, the parameters are given in closed form as $\mathbf{b}_i = G_i^{-1}\mathbf{s}_i$, for $i = 2, \cdots, K$. $Y_i|X_i = \mathbf{x}_i$ has conditional Bernoulli distribution and the corresponding conditional mean is given by

$$\lambda_i = \lambda_i(\mathbf{x}_i; \boldsymbol{\mu}, V) = E(Y_i|X_i = \mathbf{x}_i) = \mu_i + \mathbf{b}_i'(\mathbf{x}_i - \boldsymbol{\theta}_i) = \mu_i + \sum_{j=1}^{i-1} b_{ij}(y_j - \mu_j)$$

$(i = 2, \cdots, K)$.

If the correlation structure is exchangeable, namely $r_{ij} = \alpha$ for $i \neq j$ and $-1/(n-1) < \alpha < 1$, the j^{th} element of $\mathbf{b}_i$ is

$$b_{ij} = \frac{\alpha}{1+(i-2)\alpha}\left(\frac{v_{ii}}{v_{jj}}\right)^{1/2} \ (j = 1, \cdots, i-1).$$

If the correlation structure is ar(1), namely $r_{ij} = \alpha^{|i-j|}$, for $i \neq j$ and $|\alpha| < 1$, then

$$\lambda_i = \mu_i + \alpha \left(y_{i-1} - \mu_{i-1}\right) \left(\frac{v_{ii}}{v_{i-1,i-1}}\right)^{1/2} \ (i = 2, \cdots, K) \, .$$

1.5 Simulation Study

Since the proposed method is based on asymptotic results, we did extensive simulation study to check its performance for finite sample sizes.

1.5.1 Null distribution

In order to verify that the null distribution of the proposed test statistic can be approximated by a chi-square distribution with appropriate degrees of freedom, we considered a number of different scenerioes to examine whether the type I error rate are controlled for given levels of significance. Table 2 provides the list of models investigated. These scenarios provide the opportunity to assess the effects of several factors including: magnitudes of the correlation, number of covariates, number of observations within a cluster, correlation structures, cluster-level versus time-varying covariates, covariate distributions.

For each scenario, we generated random data from each of two linear logistic GEE models with different coefficients. The first model has intercept 0 and slopes 0.8 for all predictors. The second model has intercept 1 and

Table 2: Simulation Setting for Checking Type I Error Rate Control

				Correlation	
M[6]	Covariate dist[1]	x Level[2]	Dim[3]	ρ[4]	Structure[5]
1	U[-1,1],U[-1,1]	T,T	100X2	0.2	exchange
2	U[-1,1],U[-1,1]	T,T	100X2	0.6	exchange
3	B(0.5), B(0.5)	T,T	100X2	0.2	exchange
4	B(0.2), B(0.2)	T,T	100X2	0.2	exchange
5	U[-1,1],U[-1,1]	C,C	100X2	0.2	exchange
6	U[-1,1],B(0.5)	T,T	100X2	0.2	exchange
	U[-1,1],B(0.5)	C,C			
7	U[-1,1],B(0.2)	T,T	100X2	0.2	exchange
	U[-1,1],B(0.2)	C,C			
8	U[-1,1],U[-1,1]	T,T	250X2	0.2	exchange
9	U[-1,1],U[-1,1]	T,T	100X5	0.2	exchange
10	U[-1,1],U[-1,1]	T,T	100X5	0.2	ar1
11	U[-1,1],U[-1,1]	T,T	100X5	0.6	ar1
12	N(0,1),χ^2_3	T,T	100X2	0.2	exchange
13	χ^2_3,χ^2_3	T,T	100X2	0.2	exchange
14	U[-1,1],U[-1,1]	T,T	25X2	0.2	exchange
15	U[-1,1],U[-1,1]	T,T	50X2	0.2	exchange
16	U[-1,1],B(0.5)	T,T	500X2	0.2	exchange
	U[-1,1],B(0.5)	C,C			
17	U[-1,1],U(-3,3),N(0,1)	T,C,T	700X2	0.2	exchange
	N(0,2),N$(0,1)^2$,U$(-1,1)^2$	T,T,T			

1. U: uniform distribution; χ^2: chi-square distribution; B: Bernoulli distribution.

2. C: cluster-level (covariate values are the same for different time points within a cluster); T: time-varying (covariate values may differ for different time points within a cluster).

3. Number of clusters by number of observations within a cluster.

4. ρ value in the definition of exchangeable and AR(1) correlation structure.

5. In AR(1), the time points are equally spaced; exchange: exchangeable.

6. Model number

slopes 0.2. After fitting the correct models, the proposed LOF test statistics Y_R was calculated to evaluate model adequacy at 5% and 10% significance levels. Th type I error rate was estimated from 1000 simulation replicates. Table 3 displays the simulation results.

In some of the replications, the parameters can't be estimated. It is because some of the predicted values are 1, and therefore the Pearson residuals are not defined and hence the correlation parameters, the regression parameters can't be estimated. The proposed LOF method controls type I error rate pretty well except for model 13 with intercept 0 and slopes 0.8. In this case, the parameters are estimated for only about 20% of the replications, the result doesn't tell us much about the truth.

1.5.2 Power

We investigated the power of the proposed LOF methods for detecting various departures from the assumed logistic GEE model. We investigated 4 different departures: omitted covariates, omitted quadratic terms, omitted interaction terms, and incorrect link functions. For each situation, we studied 10 different models (Table 4), which allows us to investigate the effect of the following factors: magnitudes of correlation, sample sizes, number of

Table 3: Checking Type I Error Rate Control

intercept 0 and slopes 0.8[1]				intercept 1 and slopes 0.2[2]			
M	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]	model	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]
1	0.066	0.130	1000	1	0.042	0.103	1000
2	0.050	0.104	1000	2	0.058	0.119	1000
3	0.044	0.092	999	3	0.049	0.109	1000
4	0.035	0.075	960	4	0.035	0.082	961
5	0.058	0.123	1000	5	0.041	0.098	1000
6	0.061	0.126	992	6	0.065	0.120	962
7	0.047	0.101	1000	7	0.044	0.093	997
8	0.043	0.096	1000	8	0.054	0.112	1000
9	0.041	0.098	1000	9	0.050	0.094	1000
10	0.066	0.126	1000	10	0.053	0.105	1000
11	0.060	0.123	999	11	0.050	0.103	998
12	0.039	0.090	919	12	0.055	0.114	1000
13	**0.000**	**0.015**	**203**	13	0.046	0.119	994
14	0.041	0.106	989	14	0.037	0.098	964
15	0.046	0.113	1000	15	0.057	0.111	1000
16	0.046	0.096	1000	16	0.049	0.102	997
17	0.057	0.106	1000	17	0.050	0.103	1000

1. Data generated from logistic GEE model with linear predictor with intercept 0 and slopes 0.8;
2. Data generated from logistic GEE model with linear predictor with intercept 1 and slopes 0.2;
3. Estimated type I error rate based on analysis of n data sets and significance level of 0.05.
4. Estimated type I error rate based on analysis of n data sets and significance level of 0.1.
5. n is the number of replicates with successful GEE analysis.

observations within a cluster. Each departure is repeated for a model with intercept 0, regression coefficients 0.8 and a model with intercept 1, regression coefficients 0.2. This allows us to investigate the effect of coefficients in the model.

Table 4: Simulation Setting for Checking Power

M	Correlation[1]	Dim[2]	M	Correlation[1]	Dimension[2]
1	0.2	50X2	6	0.6	50X2
2	0.2	100X2	7	0.6	100X2
3	0.2	250X2	8	0.6	250X2
4	0.2	100X5	9	0.6	100X5
5	0.2	100X20	10	0.6	100X20

1. *ρ value in the definition of exchangeable correlation structure.*

2. *Number of clusters by number of observations within a cluster.*

Power for detecting omitted covariates We generated random data from the following 2 logistic GEE models: $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1} + 0.8x_{it,2} + 0.8x_{it,3}$, $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 1 + 0.2x_{it,1} + 0.2x_{it,2} + 0.2x_{it,3}$ where the last subscript indexes the covariate variables. There are three covariate variables in the model x_1, x_2, and x_3. The data are analyzed using the following GEE model:

Table 5: Simulation Setting for Checking Power of Detecting a Missing Covariate

				Correlation	
M	Covariate dist[1]	$\mathbf{x}$ Level[2]	Dimension[3]	ρ[4]	Structure[5]
1	B(0.5), U(-3,3), N(0,1)	C, T, T	50X2	0.2	Exchange
2	B(0.5), U(-3,3), N(0,1)	C, T, T	100X2	0.2	Exchange
3	B(0.5), U(-3,3), N(0,1)	C, T, T	250X2	0.2	Exchange
4	B(0.5), U(-3,3), N(0,1)	C, T, T	100X5	0.2	Exchange
5	B(0.5), U(-3,3), N(0,1)	C, T, T	100X20	0.2	Exchange
6	B(0.5), U(-3,3), N(0,1)	C, T, T	50X2	0.6	Exchange
7	B(0.5), U(-3,3), N(0,1)	C, T, T	100X2	0.6	Exchange
8	B(0.5), U(-3,3), N(0,1)	C, T, T	250X2	0.6	Exchange
9	B(0.5), U(-3,3), N(0,1)	C, T, T	100X5	0.6	Exchange
10	B(0.5), U(-3,3), N(0,1)	C, T, T	100X20	0.6	Exchange

1. B: Bernoulli distribution; U: Uniform distribution; N: Normal distribution; .
2. C: cluster-level (covariate values are the same for different time points within a cluster); T: time-varying (covariate values may differ for different time points within a cluster).
3. Number of clusters by number of observations within cluster.
4. ρ value in the definition of exchangeable correlation structure.
5. Exchange: exchangeable.

$\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2}$. We tested the LOF using Y_R. Table 5 displays the detailed simulation setting for detecting missing covariates.

Table 6 shows the simulation results for detecting a missing covariate.

The results indicate that proposed method has no power for detecting a missing covariate which is independent of covariates in the fitted model.

Table 6: Power of Detecting Missing Covariate

			intercept 0, slopes 0.8[1]			intercept 1, slopes 0.2[2]		
M	Dim[6]	ρ[7]	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]
1	50X2	0.2	0.061	0.119	994	0.048	0.092	947
2	100X2	0.2	0.045	0.101	1000	0.044	0.099	978
3	250X2	0.2	0.041	0.097	1000	0.049	0.098	997
4	100X5	0.2	0.051	0.103	1000	0.050	0.102	999
5	100X20	0.2	0.050	0.092	1000	0.050	0.093	998
6	50X2	0.6	0.059	0.117	992	0.028	0.078	932
7	100X2	0.6	0.046	0.100	1000	0.041	0.093	988
8	250X2	0.6	0.055	0.109	1000	0.056	0.118	998
9	100X5	0.6	0.048	0.091	996	0.041	0.093	998
10	100X20	0.6	0.058	0.112	996	0.046	0.100	998

1. Data generated from logistic GEE model
$\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1} + 0.8x_{it,2} + 0.8x_{it,3}$,
2. Data generated from logistic GEE model
$\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 1 + 0.2x_{it,1} + 0.2x_{it,2} + 0.2x_{it,3}$
3. Estimated power based on analysis of n data sets and significance level of 0.05.
4. Estimated power based on analysis of n data sets and significance level of 0.1.
5. n is the number of replicates with successful GEE analysis.
6. Number of clusters by number of observations within a cluster.
7. ρ value in the definition of exchangeable correlation structure.

Table 7: Setting for Power of Detecting Missing Quadratic Term

				Correlation	
M	Covariate dist[1]	**x** Level[2]	Dimension[3]	ρ[4]	Structure[5]
1	B(0.5), U(-3,3)	C, T	50X2	0.2	Exchange
2	B(0.5), U(-3,3)	C, T	100X2	0.2	Exchange
3	B(0.5), U(-3,3)	C, T	250X2	0.2	Exchange
4	B(0.5), U(-3,3)	C, T	100X5	0.2	Exchange
5	B(0.5), U(-3,3)	C, T	100X20	0.2	Exchange
6	B(0.5), U(-3,3)	C, T	50X2	0.6	Exchange
7	B(0.5), U(-3,3)	C, T	100X2	0.6	Exchange
8	B(0.5), U(-3,3)	C, T	250X2	0.6	Exchange
9	B(0.5), U(-3,3)	C, T	100X5	0.6	Exchange
10	B(0.5), U(-3,3)	C, T	100X20	0.6	Exchange

1. B: Bernoulli distribution; U: Uniform distribution; N: Normal distribution; .
2. C: cluster-level (covariate values are the same for different time points within a cluster); T: time-varying (covariate values may differ for different time points within a cluster).
3. Number of clusters by number of observations within cluster.
4. ρ value in the definition of exchangeable correlation structure.
5. Exchange: exchangeable.

Power for detecting omitted quadratic terms We generated data from the following 2 logistic GEE models: $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1}+0.8x_{it,2}+0.8x_{it,2}^2$, $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 1+0.2x_{it,1}+0.2x_{it,2}+0.2x_{it,2}^2$, but analyzed the data using the following logistic GEE model: $\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0+\beta_1 x_{it,1}+\beta_2 x_{it,2}$. Then we tested the LOF using Y_R. Table 7 shows the simulation setting for detecting a missing quadratic term.

Table 8 shows the simulation results of detecting a missing quadratic

term.

The results indicate that the proposed method has power for detecting a missing quadratic term; the power increases as the sample size and/or the number of observations within a cluster increase; the power increases when the missing quadratic term's coefficient getts bigger.

Power for detecting omitted interaction terms We generated data from the following 2 logistic GEE models: $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1} + 0.8x_{it,2} + 0.8x_{it,1}x_{it,2}$, $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 1 + 0.2x_{it,1} + 0.2x_{it,2} + 0.2x_{it,1}x_{it,2}$, where the last subscript indexes the covariate variables. There are two covariates x_1 and x_2 in the model. x_1x_2 represents the interaction term. The data were analyzed using the following GEE model: $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2}$. Then we tested the LOF using Y_R. Table 9 shows the simulation setting for detecting a missing interaction term.

Table 10 shows the simulation results for detecting a missing interaction term.

The results indicate that the proposed method has power for detecting missing interaction terms; the power increases as the sample sizes and/or the number of observations within a cluster increase; the power increases when

Table 8: Power of Detecting Missing Quadratic Term

			intercept 0, slopes 0.8[1]			intercept 1, slopes 0.2[2]		
M	Dim[6]	ρ[7]	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]
1	50X2	0.2	0.283	0.378	706	0.050	0.126	802
2	100X2	0.2	0.402	0.468	707	0.112	0.203	887
3	250X2	0.2	0.562	0.650	692	0.288	0.392	964
4	100X5	0.2	0.864	0.915	895	0.321	0.439	964
5	100X20	0.2	1.000	1.000	995	0.856	0.895	999
6	50X2	0.6	0.349	0.468	745	0.057	0.124	801
7	100X2	0.6	0.501	0.572	755	0.181	0.288	888
8	250X2	0.6	0.706	0.791	783	0.447	0.546	954
9	100X5	0.6	0.954	0.971	986	0.672	0.754	980
10	100X20	0.6	1.000	1.000	985	0.979	0.986	1000

1. Data generated from logistic GEE model
$\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1} + 0.8x_{it,2} + 0.8x_{it,2}^2$,
2. Data generated from logistic GEE model
$\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 1 + 0.2x_{it,1} + 0.2x_{it,2} + 0.2x_{it,2}^2$
3. Estimated power based on analysis of n data sets and significance level of 0.05.
4. Estimated power based on analysis of n data sets and significance level of 0.1.
5. n is the number of replicates with successful GEE analysis.
6. Number of clusters by number of observations within cluster.
7. ρ value in the definition of exchangeable correlation structure.

Table 9: Simulation Setting for Checking Power of Detecting Missing Interaction

				Correlation	
M	Covariate dist[1]	$\mathbf{x}$ Level[2]	Dimension[3]	ρ[4]	Structure[5]
1	B(0.5), U(-3,3)	C, T	50X2	0.2	Exchange
2	B(0.5), U(-3,3)	C, T	100X2	0.2	Exchange
3	B(0.5), U(-3,3)	C, T	250X2	0.2	Exchange
4	B(0.5), U(-3,3)	C, T	100X5	0.2	Exchange
5	B(0.5), U(-3,3)	C, T	100X20	0.2	Exchange
6	B(0.5), U(-3,3)	C, T	50X2	0.6	Exchange
7	B(0.5), U(-3,3)	C, T	100X2	0.6	Exchange
8	B(0.5), U(-3,3)	C, T	250X2	0.6	Exchange
9	B(0.5), U(-3,3)	C, T	100X5	0.6	Exchange
10	B(0.5), U(-3,3)	C, T	100X20	0.6	Exchange

1. B: Bernoulli distribution; U: Uniform distribution; N: Normal distribution; .
2. C: cluster-level (covariate values are the same for different time points within a cluster); T: time-varying (covariate values may differ for different time points within a cluster).
3. Number of clusters by number of observations within cluster.
4. ρ value in the definition of exchangeable correlation structure.
5. Exchange: exchangeable.

Table 10: Power of Detecting Missing Interaction Term

			intercept 0, slopes 0.8[1]			intercept 1, slopes 0.2[2]		
M	Dim[6]	ρ[7]	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]
1	50X2	0.2	0.134	0.232	990	0.063	0.112	975
2	100X2	0.2	0.260	0.416	999	0.094	0.182	999
3	250X2	0.2	0.701	0.793	1000	0.174	0.280	1000
4	100X5	0.2	0.672	0.792	982	0.195	0.295	1000
5	100X20	0.2	0.962	0.968	873	0.519	0.661	998
6	50X2	0.6	0.089	0.171	974	0.081	0.162	971
7	100X2	0.6	0.206	0.329	999	0.168	0.266	1000
8	250X2	0.6	0.493	0.637	1000	0.351	0.473	1000
9	100X5	0.6	0.342	0.481	830	0.364	0.479	993
10	100X20	0.6	0.836	0.888	725	0.611	0.739	998

1. Data generated from logistic GEE model
$\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1} + 0.8x_{it,2} + 0.8x_{it,1}x_{it,2}$,
2. Data generated from logistic GEE model
$\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 1 + 0.2x_{it,1} + 0.2x_{it,2} + 0.2x_{it,1}x_{it,2}$.
3. Estimated power based on analysis of n data sets and significance level of 0.05.
4. Estimated power based on analysis of n data sets and significance level of 0.1.
5. n is the number of replicates with successful GEE analysis.
6. Number of clusters by number of observations within cluster.
7. ρ value in the definition of exchangeable correlation structure.

the missing interaction term's coefficient getts bigger.

Power for detecting an incorrect link function Logit link function is often used for analyzing binary data. An alternative link function for binary data is the log log link (Figure 1). We generated data from the following 2 GEE model with a log-log link function: $\log(-\log(\pi_{it})) = 0.8x_{it,1} + 0.8x_{it,2}$, $\log(-\log(\pi_{it})) = 1 + 0.2x_{it,1} + 0.2x_{it,2}$.

The generated data were analyzed using the following GEE model with a logit link function: $\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2}$. Then we tested the LOF using Y_R. Table 11 shows the simulation setting for detecting an incorrect link function.

Table 12 shows the simulation results of detecting an incorrect link function.

The results indicate the proposed method has power detecting the incorrect link function for model $\log(-\log(\pi_{it})) = 1+0.8x_{it,1}+0.8x_{it,2}$;. the power increases as the sample sizes and/or the number of observations within the cluster increases. The results indicate the proposed method has very weak power for model $\log(-\log(\pi_{it})) = 1+0.2x_{it,1}+0.2x_{it,2}$, . Note that the slopes in the latter model is smaller, therefore the covariates are likely to cover a

Figure 1: Comparison of Link Functions. The black curve represents log log link function. The red curve represents logit link function.

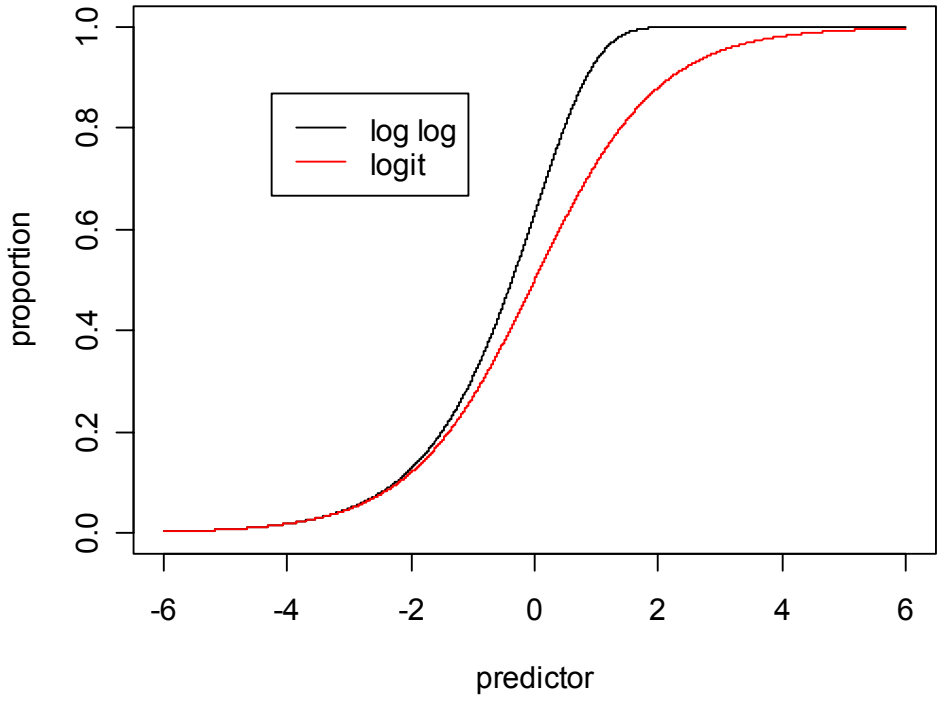


Table 11: Simulation Setting for Detecting an Incorrect Link Function

				Correlation	
M	Covariate dist[1]	$\mathbf{x}$ Level[2]	Dimension[3]	ρ[4]	Structure[5]
1	B(0.5), U(-3,3)	C, T	50X2	0.2	Exchange
2	B(0.5), U(-3,3)	C, T	100X2	0.2	Exchange
3	B(0.5), U(-3,3)	C, T	250X2	0.2	Exchange
4	B(0.5), U(-3,3)	C, T	100X5	0.2	Exchange
5	B(0.5), U(-3,3)	C, T	100X20	0.2	Exchange
6	B(0.5), U(-3,3)	C, T	50X2	0.6	Exchange
7	B(0.5), U(-3,3)	C, T	100X2	0.6	Exchange
8	B(0.5), U(-3,3)	C, T	250X2	0.6	Exchange
9	B(0.5), U(-3,3)	C, T	100X5	0.6	Exchange
10	B(0.5), U(-3,3)	C, T	100X20	0.6	Exchange

1. B: Bernoulli distribution; U: Uniform distribution; N: Normal distribution; .
2. C: cluster-level (covariate values are the same for different time points within a cluster); T: time-varying (covariate values may differ for different time points within a cluster).
3. Number of clusters by number of observations within cluster.
4. ρ value in the definition of exchangeable correlation structure.
5. Exchange: exchangeable.

Table 12: Power of Detecting Incorrect Link Function

			intercept 0, slopes 0.8[1]			intercept 1, slopes 0.2[2]		
M	Dim[6]	ρ[7]	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]	$\alpha = 0.05$[3]	$\alpha = 0.1$[4]	n[5]
1	50X2	0.2	0.062	0.143	804	0.005	0.030	395
2	100X2	0.2	0.239	0.383	967	0.014	0.049	659
3	250X2	0.2	0.637	0.766	1000	0.067	0.148	914
4	100X5	0.2	0.467	0.624	973	0.053	0.109	879
5	100X20	0.2	0.668	0.755	784	0.088	0.155	908
6	50X2	0.6	0.061	0.128	799	0.013	0.040	446
7	100X2	0.6	0.175	0.283	966	0.030	0.072	670
8	250X2	0.6	0.514	0.648	999	0.065	0.134	949
9	100X5	0.6	0.256	0.382	865	0.092	0.158	922
10	100X20	0.6	0.605	0.692	710	0.069	0.133	961

1. Data generated from logistic GEE model
$\log(-\log(\pi_{it})) = 0.8x_{it,1} + 0.8x_{it,2}$,
2. Data generated from logistic GEE model
$\log(-\log(\pi_{it})) = 1 + 0.2x_{it,1} + 0.2x_{it,2}$.
3. Estimated power based on analysis of n data sets and significance level of 0.05.
4. Estimated power based on analysis of n data sets and significance level of 0.1.
5. n is the number of replicates with successful GEE analysis.
6. Number of clusters by number of observations within cluster.
7. ρ value in the definition of exchangeable correlation structure.

smaller range than the first model. Since locally the log log link function is similar to a shifted logit link function (Figure 1), when the range of the predictor is small, the two link functions are hardly differentiable.

1.5.3 Comparison with Other Currently Available Methods

Currently there are several LOF methods for logistic GEE models available. We compared the performance of the proposed method with these methods.

If a method controls type I error rate, the simulation estimated error rates will locate in the following range with about 99% chance.

$\alpha \pm z_{0.005}\sqrt{\frac{\alpha(1-\alpha)}{n}}$

where $z_{0.005}$ is the value of a standard normal variable with upper tail area 0.005, α is the nominal error rate, and n is the replication number of simulation. For α =5% and n=1000, the simulation estimated error rate is 99% likely to locate in the range of $(0.034, 0.066)$ if the method controls type I error rate. The performance data for the methods other than the proposed method are obtained from Evans and Li (2004). The significance level is 0.05. κ-like method uses the recommended 'pool fit' cut-off value of 0.2.

Table 13 displays the simulation estimated type I error rate for all the methods currently available. The data were generated using the setting in

table 2 with model intercept 0 and slopes 0.8. Any values in the table outside the range $(0.034, 0.066)$ are bolded. Bold values indicate the methods significantly inflate or deflate the type I error rate.

The following conclusions are indicated in the simulation results. The kappa-like method doesn't control type I error rate at all. The proposed method performs better than the 4 statistics (G, G2, U, U2) based on residuals when the magnitude of the correlation is high (model 2). Compared to other robust test statistics (Q_R, N_R), the proposed test statistic controls type I error rate better when the covariate distribution is highly skewed (model 12) or when there are many covariates in the model (model 17). Compared to another test statistic based on partitioning data using fitted probabilities (X^2), the proposed method performs better when the covariates have highly skewed distribution (model 12) or small sample size (model 14). Overall, the results indicate the proposed method has better performance than other currently available methods in controlling type I error rate.

Table 14 displays the power of detecting a missing covariate for different methods. The data are generated in the same way as the proposed method using the model with intercept 0 and slopes 0.8. See table 2 for details of the model specification.

Table 13: Comparisons of Methods for Controlling Type I Error Rate

M	G	G2	U	U2	X^2	Q	Q_R	N_m	N_R	κ	Y_R
1	4.1	4.2	**3.1**	4.3	6.0	4.8	5.0	4.1	4.9	**98.0**	6.6
2	**6.8**	**7.2**	**7.4**	**7.2**	3.8	5.1	6.2	4.3	**6.8**	**98.3**	5.0
3	5.1	5.0	5.4	6.1	6.0	4.6	**8.2**	4.1	4.7	**100.0**	4.4
4	5.8	4.8	5.3	5.3	4.6	5.6	**14.3**	5.2	5.6	**100.0**	3.5
5	5.2	4.6	4.8	4.3	4.1	4.9	5.8	5.3	6.5	**97.8**	5.8
6	**6.7**	3.9	6.0	5.9	5.3	4.8	**8.4**	**6.9**	**11.2**	**81.5**	6.1
7	5.4	4.9	5.0	4.6	5.8	5.2	**10.2**	5.2	**7.4**	**88.7**	4.7
8	4.3	4.0	4.9	5.1	4.5	4.8	4.6	**3.2**	3.5	**100.0**	4.3
9	4.9	5.1	5.0	5.1	4.2	5.1	5.4	4.3	5.0	**100.0**	4.1
10	4.8	4.9	5.2	5.1	3.9	6.2	5.4	4.7	4.2	**100.0**	6.6
11	6.3	6.3	6.3	6.4	3.6	**7.1**	5.2	4.0	4.5	**100.0**	6.0
12	**0.2**	**0.7**	5.0	5.2	**2.9**	5.7	**63.3**	4.6	**14.7**	**30.2**	3.9
14	5.0	4.6	5.8	6.1	**1.2**	5.8	6.5	4.9	**7.5**	**81.2**	4.1
15	5.2	5.9	5.5	5.3	4.8	4.9	5.9	3.5	4.6	**92.3**	4.6
16	**7.9**	4.5	4.9	4.7	4.4	3.9	5.2	5.4	6.1	**99.8**	4.6
17	**19.7**	**0.9**	5.5	5.8	**23.9**	4.8	**29.6**	6.1	**15.1**	**0.0**	5.7

G, G2: estimated type I error rate (%) for Pan (2002) using Pearson sums of square of residuals. G uses empirical covariance estimate. G2 uses unstructured correlation matrix estimate.
U, U2: estimated type I error rate (%) for Pan (2002) using unweighted sums of square of residuals. U uses empirical covariance estimate. U2 uses unstructured correlation matrix estimate.
X^2: estimated type I error rate (%) for Horton et al. (1999) based on partitioning data using fitted probabilities.
Q, Q_R: estimated type I error rate (%) for Barnhart and Williamson (1998) based on partitioning data using covariates patterns. Q is computed using model based estimated covariance matrix. Q_R is computed using robust estimated covariance.
N, N_R: estimated type I error rate (%) for Evans and Li (2005) based on partitioning data using both fitted probabilities and covariates patterns. N is computed using model based estimated covariance matrix. N_R is computed using robust estimated covariance matrix.
κ: estimated type I error rate (%) for Williamson et al. (2003).
Y_R: estimated type I error rate (%) for proposed method.
The type I error rates are estimated from about 1000 simulated data sets. The simulated data are generated using logistic GEE model with linear predictor with intercept 0 and slopes 0.8. See table 2 for details of the model specification.

Table 14: Comparisons of Methods for Detecting a Missing Covariate

M	G	G2	U	U2	X^2	Q	Q_R	N_m	N_R	Y_R
1	6.5	4.8	4.9	5.5	2.6	4.8	7.2	5.4	7.0	6.1
2	9.5	5.0	7.6	6.2	6.5	3.8	10.4	6.3	6.1	4.5
3	8.1	5.4	4.9	4.1	5.6	4.1	5.9	4.7	5.0	4.1
4	6.2	6.2	5.5	6.2	5.5	6.3	10.1	5.6	7.0	5.1
5	9.4	9.8	9.6	10.2	3.9	10.3	0.0	6.0	6.2	5.0
6	7.6	6.8	6.6	6.6	2.8	5.8	6.4	5.5	8.1	5.9
7	9.7	5.5	7.7	5.3	5.8	3.0	10.9	11.9	7.7	4.6
8	8.8	8.0	6.6	7.4	6.2	3.7	6.0	5.0	5.5	5.5
9	7.6	8.0	7.4	7.9	4.4	6.6	8.2	6.8	7.8	4.8
10	9.9	13.5	10.0	13.4	3.4	17.5	0.0	7.1	5.5	5.8

G, G2: estimated power (%) for Pan (2002) using Pearson sums of square of residuals. G uses empirical covariance estimate. G2 uses unstructured correlation matrix estimate.
U, U2: estimated power (%) for Pan (2002) using unweighted sums of square of residuals. U uses empirical covariance estimate. U2 uses unstructured correlation matrix estimate.
X^2: estimated power for Horton et al. (1999) based on partitioning data using fitted probabilities.
Q, Q_R: estimated power (%) for Barnhart and Williamson (1998) based on partitioning data using covariates patterns. Q is computed using model based estimated covariance matrix. Q_R is computed using robust estimated covariance.
N, N_R: estimated power (%) for Evans and Li (2005) based on partitioning data using both fitted probabilities and covariates patterns. N is computed using model based estimated covariance matrix. N_R is computed using robust estimated covariance matrix.
κ: estimated power (%) for Williamson et al. (2003).
The power is estimated from about 1000 simulated data sets. The simulated data are generated using logistic GEE model $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1} + 0.8x_{it,2} + 0.8x_{it,3}$, *The data are analyzed using the following GEE model:* $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2}$. *See table 5 for details of simulation setting.*

No methods have decent power detecting a missing covariate which is independent of covariates in the fitted model. Especially the two methods (X^2 and Y_R) based on partitioning data using fitted probabilities have no power at all detecting missing independent covariates.

Table 15 displays the power of detecting a missing quadratic term, the data are generated in the same way as that for the proposed method using the model with intercept 0 and slopes 0.8.

The following conclusions are indicated in the simulation results. Almost every method has some power detecting missing quadratic terms. But when the cluster sizes are big, only a few methods have high power. The proposed method is one of them.

Table 16 displays the power of detecting missing an interaction term, the data are generated in the same way as that for the proposed method using the model with intercept 0 and slopes 0.8. In detecting missing interaction terms, a few methods perform satisfactorily in most of the scenarioes. The proposed method is one of them.

Table 17 displays the power of detecting an incorrect link function, the data are generated the in the same way as the proposed method using the model with intercept 0 and slopes 0.8.

Table 15: Comparisons of Methods for Detecting Missing a Quadratic Term

M	G	G2	U	U2	X^2	Q	Q_R	N_m	N_R	Y_R
1	44.0	25.8	40.7	30.0	22.7	5.5	14.9	5.7	17.9	28.3
2	74.4	52.0	76.8	58.8	90.3	30.9	48.1	9.6	13.2	40.2
3	94.9	84.0	94.6	87.8	99.9	68.2	57.9	10.4	11.9	56.2
4	44.9	36.8	12.6	13.1	100.0	46.4	71.2	4.5	7.6	86.4
5	11.7	10.1	14.0	14.1	100.0	87.2	0.1	7.3	7.4	100.0
6	30.3	18.0	23.0	17.8	15.0	4.6	14.2	6.1	23.1	34.9
7	71.9	50.1	73.9	57.1	86.2	25.9	45.5	15.2	14.8	50.1
8	88.2	73.0	60.6	54.1	100.0	67.8	56.5	8.0	9.4	70.6
9	27.8	21.0	6.8	6.8	99.8	47.5	68.1	6.1	10.5	95.4
10	7.2	6.2	12.1	11.4	100.0	86.5	0.6	8.9	5.7	100.0

G, G2: estimated power (%) for Pan (2002) using Pearson sums of square of residuals. G uses empirical covariance estimate. G2 uses unstructured correlation matrix estimate.
U, U2: estimated power (%) for Pan (2002) using unweighted sums of square of residuals. U uses empirical covariance estimate. U2 uses unstructured correlation matrix estimate.
X^2: estimated power for Horton et al. (1999) based on partitioning data using fitted probabilities.
Q, Q_R: estimated power (%) for Barnhart and Williamson (1998) based on partitioning data using covariates patterns. Q is computed using model based estimated covariance matrix. Q_R is computed using robust estimated covariance.
N, N_R: estimated power (%) for Evans and Li (2005) based on partitioning data using both fitted probabilities and covariates patterns. N is computed using model based estimated covariance matrix. N_R is computed using robust estimated covariance matrix.
κ: estimated power (%) for Williamson et al. (2003).
Y_R: estimated power (%) for the proposed method.
The power is estimated from about 1000 simulated data sets. The simulated data are generated using logistic GEE model $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1} + 0.8x_{it,2} + 0.8x_{it,2}^2$. *The data are analyzed using the following GEE model:* $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2}$. *See table 7 for details of simulation setting.*
**: Approximately 300 out of 1000 data sets analysis fail for Y_R in model 1, 2, 3, 6, 7, and 8.*

Table 16: Comparisons of Methods for Detecting Missing Interaction

M	G	G2	U	U2	X^2	Q	Q_R	N_m	N_R	κ	Y_R
1	3.9	3.5	4.4	5.7	1.3	5.7	9.00	8.5	15.2	0.3	13.4
2	5.4	2.8	4.3	4.4	4.0	23.2	47.70	13.3	15.6	0.0	26.0
3	4.4	4.4	4.7	5.8	12.6	63.4	71.00	18.7	23.1	0.0	70.1
4	4.7	5.2	6.1	7.1	8.3	44.0	60.70	7.5	8.5	0.0	68.0
5#	18.4	22.4	10.2	13.0	13.5	89.3	1.57	9.6	5.5	0.0	96.1
6	5.1	6.1	6.3	9.3	1.4	6.1	8.60	7.7	13.0	0.4	8.9
7	4.8	4.4	5.1	5.6	4.3	21.7	46.90	18.0	14.8	0.0	20.6
8	7.1	5.3	9.2	12.5	10.3	70.4	73.70	15.9	18.5	0.0	49.3
9&	7.3	10.7	8.9	10.1	7.1	57.0	53.20	7.4	7.7	0.0	33.5
10*	48.7	52.5	16.4	18.7	17.9	93.1	16.50	40.4	7.7	0.0	84.4

G, G2: estimated power (%) for Pan (2002) using Pearson sums of square of residuals. G uses empirical covariance estimate. G2 uses unstructured correlation matrix estimate.
U, U2: estimated power (%) for Pan (2002) using unweighted sums of square of residuals. U uses empirical covariance estimate. U2 uses unstructured correlation matrix estimate.
X^2: estimated power for Horton et al. (1999) based on partitioning data using fitted probabilities.
Q, Q_R: estimated power (%) for Barnhart and Williamson (1998) based on partitioning data using covariates patterns. Q is computed using model based estimated covariance matrix. Q_R is computed using robust estimated covariance.
N, N_R: estimated power (%) for Evans and Li (2005) based on partitioning data using both fitted probabilities and covariates patterns. N is computed using model based estimated covariance matrix. N_R is computed using robust estimated covariance matrix.
κ: estimated power (%) for Williamson et al. (2003).
Y_R: estimated power (%) for the proposed method.
The power are estimated from about 1000 simulated data sets. The simulated data are generated using logistic GEE model $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = 0.8x_{it,1} + 0.8x_{it,2} + 0.8x_{it,1}x_{it,2}$. The data are analyzed using the following GEE model: $\log\left(\frac{\boldsymbol{\pi}_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2}$. *See table 9 for details of simulation setting.*
#: Approximately 350 out of 1000 replicates did not converge for all statistics except for X^2, κ, and Y_R. Approximately 150 out of 1000 data sets analysis fail for Y_R.
*: Approximately 700 out of 1000 replicates did not converge for all statistics except for X^2, κ, and Y_R. Approximately 300 out of 1000 data sets analysis fail for Y_R.
&: Approximately 200 out of 1000 data sets analysis fail for Y_R.

Table 17: Comparisons of Methods for Detecting an Incorrect Link Function

M	G	G2	U	U2	X^2	Q	Q_R	N_m	N_R	κ	Y_R^*
1	2.4	0.0	7.2	7.7	1.9	5.3	8.10	4.1	9.4	0.0	6.2
2	25.8	0.0	25.1	15.4	4.2	3.0	89.50	7.3	8.4	0.0	24.0
3	87.5	1.4	32.0	28.2	38.2	6.2	64.50	6.0	6.5	0.0	63.7
4	74.6	3.9	8.7	11.1	28.4	3.2	87.20	7.6	6.8	0.0	46.6
5	92.2	65.0	14.8	16.2	98.9	10.2	1.30	14.3	5.4	0.0	67.2
6	4.1	0.1	3.8	6.5	0.9	5.4	9.20	4.5	8.4	0.3	6.1
7	22.3	0.1	22.8	14.8	4.0	1.7	85.00	9.4	9.0	0.0	17.6
8	82.0	6.1	4.4	6.6	32.1	4.2	56.90	5.4	5.7	0.0	51.5
9	65.3	8.6	5.8	8.4	21.3	3.0	65.80	9.4	6.7	0.0	25.2
10	85.7	51.0	21.3	23.3	96.3	18.1	1.53	22.8	7.6	0.0	59.6

G, G2: estimated power (%) for Pan (2002) using Pearson sums of square of residuals. G uses empirical covariance estimate. G2 uses unstructured correlation matrix estimate.
U, U2: estimated power (%) for Pan (2002) using unweighted sums of square of residuals. U uses empirical covariance estimate. U2 uses unstructured correlation matrix estimate.
X^2: estimated power for Horton et al. (1999) based on partitioning data using fitted probabilities.
Q, Q_R: estimated power (%) for Barnhart and Williamson (1998) based on partitioning data using covariates patterns. Q is computed using model based estimated covariance matrix. Q_R is computed using robust estimated covariance.
N, N_R: estimated power (%) for Evans and Li (2005) based on partitioning data using both fitted probabilities and covariates patterns. N is computed using model based estimated covariance matrix. N_R is computed using robust estimated covariance matrix.
κ: estimated power (%) for Williamson et al. (2003).
Y_R: estimated power (%) for the proposed method.
The power are estimated from about 1000 simulated data sets. The simulated data are generated using logistic GEE model $\log(-\log(\pi_{it})) = 0.8x_{it,1} + 0.8x_{it,2}$. *The data are analyzed using the following GEE model:* $\log\left(\frac{\pi_{it}}{1-\pi_{it}}\right) = \beta_0 + \beta_1 x_{it,1} + \beta_2 x_{it,2}$. *See table 7 for details of simulation setting.*
**: Approximately 200 out of 1000 data sets analysis fail for Y_R in model 1, 5, and 6. Approximately 300 out of 1000 data sets analysis fail for Y_R in model 10. Approximately 140 out of 1000 data sets analysis fail for all statistics except for X^2 and κ in model 10.*

In detecting an incorrect link function, several methods perform satisfactorily in most of the scenarios. The proposed method is one of them.

1.6 Examples

Next we demonstrate the application by applying the proposed method to some real examples.

1.6.1 Respiratory Disease

A clinical trial comparing two treatments for a respiratory disease was done by Koch et al. (1990). In each of the two centers, eligible patients were randomly assigned to active treatment or placebo. During the treatment, the respiratory status was determined at each of 4 visits and recorded on a 5-point scale as 0 for terrible to 4 for excellent. Potential explanatory variables are center, gender, baseline respiratory status (these 3 variables are dichotomous), and age (in years) at the time of entry to the study. There were 111 patients (54 active, 57 placebo) participating the clinical trial. There are no missing data for responses or covariates. The response is converted into a dichotomous variable by dividing the response to good outcome (response of 3 or 4) and bad outcome (response less than 3). Four visits of each patient form a cluster.

The data were analyzed using the following logistic GEE model.

logit(probability of good outcome) $= \beta_0 + \beta_1 center + \beta_2 treatment +$

Table 18: Lack of Fit for Logistic GEE Model analyzing Respiratory Data

G	G2	U	U2	X^2	Q	Q_R	N_m	N_R	Y_R
0.63	0.63	0.33	0.41	0.62	0.29	.003	0.13	0.20	0.76

G, G2: LOF p value for Pan (2002) using Pearson sums of square of residuals. G uses empirical covariance estimator. G2 uses unstructured correlation matrix estimate.
U, U2: LOF p value for Pan (2002) using unweighted sums of square of residuals. U uses empirical covariance estimator. U2 uses unstructured correlation matrix estimate.
X^2: LOF p value for Horton et al. (1999) based on partitioning data using fitted probabilities.
Q, Q_R: LOF p value for Barnhart and Williamson (1998) based on partitioning data using covariates patterns. Q is computed using model based estimated covariance matrix. Q_R is computed using robust estimated covariance.
N, N_R: LOF p value for Evans and Li (2005) based on partitioning data using both fitted probabilities and covariates patterns. N is computed using model based estimated covariance matrix. N_R is computed using robust estimated covariance matrix.
Y_R: LOF p value for the proposed method.

$\beta_3 sex + \beta_4 baseline + \beta_5 age$

where outcome, center, treatment, sex, baseline are all coded with binary values 0 and 1. The correlation structure within the cluster is assumed to be unstructured. We use all the methods currently available except for kappa-like method to assess the LOF of this model. The p values for different methods are shown in table 18.

None of the methods except for Barnhart's robust test statistic indicate

there is LOF present in the fitted model. Since in the simulation study, Barnhart's method has inflated type I error rate in 6 different scenerioes, Barnhart's correctness is a suspect.

1.6.2 Diabetic Retinopathy

In a study in southern Wisconsin, 996 insulin-taking, younger-onset diabetic patients were examined using standard protocols to determine the association between diabetic retinopathy and some risk factors. Both eyes of the patients are examined. The presence or absence of retinopathy together with some covariates are recorded. Thirty-two observations with missing covariate values are removed from the analysis.

Table 19 contains the p values for different LOF methods for 2 different models. The first logistic GEE model has 4 linear covariates: duration of diabetes, glycosylated haemoglobin level, diastolic blood pressure, and body mass index. Most of the LOF methods suggest inadequacy of this model. The second logistic GEE model has 4 linear covariates plus two quadratic terms. None of the methods indicate the model has LOF except for Q_R which has been shown to inflate the type I error rate when the number of covariates is big.

Table 19: Lack of Fit for Logistic GEE Model analyzing Diabetic Retinopathy Data

Model*	G	G2	U	U2	X^2	Q	Q_R	N_m	N_R	Y_R
1	0.07	0.00	0.00	0.01	0.00	0.13	0.00	0.00	0.00	0.03
2	0.26	0.32	0.18	0.19	0.46	0.51	0.00	0.12	0.34	0.12

Notes:
** Model 1 has 4 linear terms:* X_1 *:duration of diabetes,* X_2*: body mass index,* X_3 *:glycosylated haemoglobin level,* X_4*: diastolic blood pressure. Model 2 has 4 linear terms* X_1, X_2, X_3, X_4 and two quadratic terms: X_1^2 and X_2^2.
G, G2: LOF p value for Pan (2002) using Pearson sums of square of residuals. G uses empirical covariance estimator. G2 uses unstructured correlation matrix estimate.
U, U2: LOF p value for Pan (2002) using unweighted sums of square of residuals. U uses empirical covariance estimator. U2 uses unstructured correlation matrix estimate.
X^2*: LOF p value for Horton et al. (1999) based on partitioning data using fitted probabilities.*
Q, Q_R*: LOF p value for Barnhart and Williamson (1998) based on partitioning data using covariates patterns. Q is computed using model based estimated covariance matrix.* Q_R *is computed using robust estimated covariance.*
N, N_R*: LOF p value for Evans and Li (2005) based on partitioning data using both fitted probabilities and covariates patterns. N is computed using model based estimated covariance matrix.* N_R *is computed using robust estimated covariance matrix.*
Y_R*: LOF p value for the proposed method.*

1.7 Conclusions and Discussion

The simulation results indicate that the proposed method controls type I error rate well; it has power for detecting missing interaction terms and quadratic terms; the power increases as the sample sizes and/or the number of observations within a cluster increase. The proposed method has power detecting incorrect link functions if the predictor covers a reasonable range. The proposed method doesn't have power detecting missing covariates. The simulation results indicate the proposed method has better or similar performance compared to other currently available LOF methods for logistic GEE models.

Intuitively the proposed method should work better than other model comparison based method because its general model is a more flexible piecewise model. The general model in the proposed method allows both slopes and intercept to vary for different data groups, while the other model comparison methods restrict the slopes for different data groups to be the same, which limits the general model's ability to approximate the true model, and hence limit the performance of the method. Since the genral model in the proposed method approximates the true model well, when LOF is present in the fitted model, the piecewise logistic GEE model can be used as an

alternative model to analyze the data.

Several SAS macros have been developed to implement the proposed method. See appendix for the detailed SAS code.

1.8 Reference

1. Aggresti, A. (1996), *Introduction to Categorical Data Analysis*, Wiley, Hoboken, NJ.

2. Azzalini, A., Bowman, A., Hardel, W. (1989), "On the Use of Non-parametric Regression for Model Checking," *Biometrika*, 76: 1-11.

3. Barnhart, H. X., Williamson, J. M. (1998). "Goodness-of-fit Tests for GEE Modeling with Binary Responses," *Biometrics*, 54: 720-729.

4. Berkson, J. (1944). "Application of the Logistic Function to Bioassay," *Journal of the American Statistical Association*, 39: 357-365.

5. Brown, C. C. (1982), "On a Goodness-Of-Fit test for the Logistic Model Based on Score Statistics," *Communications in Statistics*, 11: 1087-1105.

6. Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, 20: 37-46.

7. Copas, J. (1980), "Plotting p Against x," *Applied Statistics*, 32: 25-31.

8. Cox, D. R. (1970). *The Analysis of Binary Data*, London, Methuen.

9. Cox, D. R. and Snell, E. J. (1989). *The Analysis of Binary Data*, 2nd ed. Chapman and Hall, London.

10. Emrich, L. J. and Piedmonte, M. R. (1991), "A Method for Generating High-dimensional Multivariate Binary Variates," *American Statistician*, 45: 302-304.

11. Evans, S. R. (1998), *Lack-Of-Fit in Two Models for Clustered Binary Data*, PhD Dissertation, University of Massachusetts.

12. Evans, S. R. and Li, L. (2005), "A Comparison of Lack-Of-Fit tests for the Logistic GEE Model," *Statistics in Medicine*, 24: 1245-1261.

13. Farrington, C. P., (1996), "On assessing Lack-Of-Fit of Generalized Linear Models to Sparse Data," *Journal of the royal statistical society*, B, 58: 349-360.

14. Firth, D., Glosup, J., and Hinkley, D. V. (1991), "Model Checking with Non-parametric Curves," *Biometrika*, 78: 245-252.

15. Fleiss, J. L. (1971), "Measuring Nominal Scale Agreement Among Many Raters," *Psychological Bulletin*, 76: 378-382.

16. Gange, S. J. (1995), "Generating Multivariate Categorical Variates Using the Iterative Proportional Fitting Algorithm," *American Statistician*, 49: 134-138.

17. Hardin, J. W. and Hilbe, J. M. (2003), *Generalized Estimating Equations*, Chapman & Hall/CRC.

18. Horton, N. J., Bebchuk, J. D. , Jones, C. L., Lipsitz, S. R., Catalano, P. J., Zahner, G. E. P., and Fitzmaurice, G. M. (1999), "Goodness-of-fit for GEE: an Example with Mental Health Service Utilization," *Statistics in Medicine*, 18: 213-222.

19. Hosmer, D. W., Hosmer, T., le Cessie, S., and Lemeshow. S., (1997), "A Comparison of Lack-Of-Fit Tests for Logistic Regression Model," *Statistics in Medicine*, 16, 965-980.

20. Hosmer, D. W., Lemeshow, S. (1980), "A Goodness-Of-Fit Test for the Multiple Logistic Regression Model," *Communications in Statistics (Theory and Methods)*, 9: 1043-1069.

21. Kang, S. H. and Jung, S. H. (2001), "Generating Correlated Binary Variables with Complete Specification of the Joint Distribution," *Biometrical Journal*, 43: 263-269.

22. Koch, G. G., Carr, G. J., Amara, I. A., Stokes, M. E., and Uryniak, T. J. (1990). "Categorical Data Analysis," *Statistical Methodology in the Pharmaceutical Sciences*, ed. D. A. Berry, New York: Marcel Dekker Inc. 391-475.

23. Landis, J. R., Koch, G. G. (1977), "The measurement of observer agreement for categorical data," *Biometrics*, 33: 159-174.

24. Lee, A. J. (1993), "Generating Random Binary Deviates Having Fixed Marginal Distributions and Specified Degrees of Association," *American Statistician*, 47: 209-215.

25. le Cessie, S. and van Houwelingen, J. C. (1991), "A Goodness-of-fit test for Binary Data Based on Smoothing Effects Models," *Biometrics*, 47: 1267-1282.

26. le Cessie, S. and van Houwellingen, J. C. (1995), "Testing the Fit of a Regression Model Via Score Tests in Random Effects Models," *Biometrics*, 51: 600-614.

27. Leisch, F., Weingessel, A. and Hornik, K. (1998), "On the Generation of Correlated Artificial Binary Data," Working paper series, SFB, "Adap-

tive Information Systems and Modeling in Economics and Management Science", Vienna University of Economics.

28. Liang, K., Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73: 13-22.

29. Liu, Y., (2007), *On Goodness-of-fit of Logistic Regression Model*, PhD thesis (adviser: Yang, S. S.), Kansas State University, 2007.

30. Lunn, A. D. and Davies, S. J. (1998), "A Note on Generating Correlated Binary Variables," *Biometrika*, 85: 487-490.

31. McCullagh, P. and Nelder, J. A. (1983), *Generalized Linear Models*, 115-117.

32. Nagelkerke, N. J. D. (1991). "A note on a general definition of the coefficient of determination," *Biometrika*, 78(3): 691-692.

33. Nelder, J. A., and Wedderburn, R. W. M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, Series A, 135: 370-384.

34. Oman, S. D. and Zucker, D. M. (2001). "Modeling and Generating Correlated Binary Variables," *Biometrika*, 88: 287-290.

35. Osius, G. and Rojek, D. (1992), "Normal Goodness-of-fit tests for multinomial models with large degees of freedom," *The Journal of American Statistical Association*, 87: 1145-1152.

36. Pan, W. (2002), "Lack-Of-Fit Tests for GEE with Correlated Binary Data," *Scandinavian Journal of Statistics*, 29(1):101-110.

37. Park, C. G., Park, T. and Shin, D. W. (1996), "A Simple Method for Generating Correlated Binary Variates," *American Statistician*, 50: 306-310.

38. Prentice, R. L. (1976), "Generalization of the probit and logit methods for dose response curves," *Biometrics*, 32: 761-768.

39. Prentice, RL (1988), "Correlated Binary Regression with Covariates Specific to Each Binary Observation," *Biometrics*, 44: 1033-1048.

40. Pulkstenis, E, Robinson, T. J. (2002), "Two Lack-Of-Fit tests for Logistic Regression Models with Continuous Covariates," *Statistics in Medicine*, 21: 79-93.

41. Qaquish, B. F. (2003), "A Family of Multivariate Binary Distributions for Simulating Correlated Binary Variables with Specified Mar-

ginal Means and Correlations," *Biometrika*, 90: 455-463.

42. Rao, C. R. (1973), *Linear Statistical Inference and its Applications*, 2nd edition, Wiley, New York, NY, p418.

43. Rotnitzky, A. and Jewell, N. P. (1990), "Hypothesis Testing of Regression Parameters in Semiparametric Generalized Linear Models for Cluster Correlated Data," *Biometrika*, 77(3): 485-497.

44. Royston, P. (1992), "The Use of Cusums and other Techniques in Modeling Continuous Covariates in Logistic Regression," *Statistics in Medicine*, 11: 1115-1129.

45. Rubin, D. B. (1976). "Inference and Missing data," *Biometrika*, 63: 581-592.

46. Stiratelli, R., Larid, N. and Ware, J. H. (1984), "Random-effects Models for Serial Observations with Binary Response," *Biometrics*, 40: 961-971.

47. Stukel, T. A. (1988), "Generalized Logistic Models", *Journal of American Statistical Association*, 83: 426-431.

48. Tsiatis, A. A. (1980), "A Note on a Lack-Of-Fit Test for the Logistic Regression Model, *Biometrika*, 67: 250-251.

49. Wedderburn, R. W. M. (1976), "On the Existence and the Uniqueness of the Maximum Likelihood Estimates for Certain Generalized Linear Models," *Biometrika*, 63: 27-32.

50. Williamson, J. M., Lin, H. M., and Barnhart, H. X. (2003), "A Classification Statistic for GEE Categorical Response Models," *Journal of Data Science*, 1: 149-165.

2 Optimal Designs for Nonlinear Combination of Parameters in Nonlinear Regression Model

2.1 Introduction

Applied researchers often find nonlinear regression models useful to describe phenomena in their study. Physical, biological, chemical or other theoretical consideration often lead to mechanistic nonlinear models (Bates and Watts, 2007). Therefore nonlinear model parameters are usually easier to interpret (O'Brien, 2005). Before the model can be fitted and parameters interpreted, researchers need to make decisions about at which design points the experiment should be carried out to take measurements. Careful choices of design points at this stage will lead to efficient usage of limited resources (such as time, money, patients) and maximize the information desired (such as precisely estimated parameters of interest). To accomplish this in an optimal way is called an optimal design.

Motivated by an application in medical science, we proposed a cost-efficient Bayesian optimal design for precisely estimating nonlinear combina-

tions of parameters in a nonlinear regression model with quantitative factors with respect to some design variables.

Like many areas of Bayesian statistics. applicaton of Bayesian optimal design lags behind its theory due to lack of appropriate software. In order to make the proposed method attractive to researchers, we developed an R package to implement the proposed method and make related graphs to visualize the results.

2.2 Background

An optimal design often involves choosing a design with n design points, denoted by ξ, to estimate the parameters of interest.

In traditional literature, such as Atkinson and Donev (1992), a design ξ is expressed as

$$\xi = \left\{ \begin{array}{cccc} \mathbf{x}_1, & \mathbf{x}_2, & \cdots, & \mathbf{x}_k \\ w_1, & w_2, & \cdots, & w_k \end{array} \right\}$$

where the k ($k \leq n$) distinct design support "points" (or vector) $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_k$ are elements of the design space X, and $n\omega_i$ design points are assigned to $\mathbf{x}_i$, $i = 1, \cdots, k$. Although this design scheme is commonly used in statistics literature, it is rarely used in medical application. Instead researchers

in medical application often use other design schemes such as dilution series designs. One example of dilution series schemes is that of Rocke and Jones (1997) which is detailed later in an example.

More generally a design ξ dependent on control variables, which determines the design points (Chaloner and Verdinelli, 1995), is denoted by $\xi = \xi(\boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a vector of control variables.

Let M denote the expected information matrix associated with the full parameter vector $\boldsymbol{\theta}$ $(p \times 1)$. M depends on both the design points and the true values of the parameters. Strictly speaking, we need to express the information matrix as $M(\xi, \boldsymbol{\theta})$. In our work, we use M, $M(\xi)$, $M(\xi, \boldsymbol{\theta})$ interchangeably when there is no confusion with the context. When ξ is determined by a control variable vector $\boldsymbol{\delta}$, ξ is replaced by $\boldsymbol{\delta}$.

Different experiments have different objectives. The goals of the experiments could be precisely estimate parameters of interest, precise calibration, or precise prediction the response of future observations. Different objectives lead to different optimal designs. In the proposed research, we assume the goal of the experiment is to estimate nonlinear combinations of parameters precisely.

2.2.1 Criteria of Optimality

When the research objective is precisely estimating parameters, in current practice, an optimal design typically involves minimizing a function of the information matrix of the parameters of interest (O'Brien, 2005). The most important design criterion in applications is that of D-optimality, in which the generalized variance, or its logarithm $-\log|M(\xi)|$, is minimized. Two other criteria which have a statistical interpretation in terms of the information matrix $M(\xi)$ are A- and E-optimality. In A-optimality $tr\{M^{-1}(\xi)\}$, the total variance of the parameter estimate, is minimized. In E-optimality the variance of the least well-estimated contrast $\boldsymbol{\alpha}'\boldsymbol{\theta}$ is minimized subject to the constraint $\boldsymbol{\alpha}'\boldsymbol{\alpha}$ =1. A contrast is a linear combination of 2 or more factor level means whose coefficients (elements of $\boldsymbol{\alpha}$) add up to 0.

When the interest is in the full parameter vector $\boldsymbol{\theta}$ $(p \times 1)$. The above ideas can be put more formally by considering the eigenvalues $\lambda_1, \cdots, \lambda_p$ of $M(\xi)$. The eigenvalues of $M^{-1}(\xi)$ are $1/\lambda_1, \cdots, 1/\lambda_p$. In terms of these eigenvalues the 3 criteria are as follows.

A-optimality minimizes the sum of the variances of the parameter estimates:

$$\min \sum_{j=1}^{p} \frac{1}{\lambda j} = tr\left\{M^{-1}\left(\xi\right)\right\}.$$

D-optimality minimizes the generalized variance of the parameter estimates:

$$\min \prod_{j=1}^{p} \frac{1}{\lambda_j} = \left|M\left(\xi\right)\right|.$$

E-optimality minimizes the variance of the least well-estimated contrasts:

$$\min \max \left(1/\lambda_j\right).$$

D-optimal design for quantitative factors do not depend on the scale of the variables (Atkison and Doev, 1992), which is not in general the case for A- and E-optimal designs. This property makes D-optimal design much more popular than the other two design criterion.

$\mathbf{D}_A-$optimality Sometimes the research interest is not in all p parameters, but only in s independent linear combinations of the parameters $\boldsymbol{\theta}$ which are the elements of $A'\boldsymbol{\theta}$, where A is $p \times s$ of rank $s < p$. The covariance matrix for these linear combinations is $A'M^{-1}\left(\xi\right)A$. If $s = 1$, A-, D-, E-optimal designs all reduce to minimizing the variance of the estimated linear combination. When $s > 1$, the A-optimal design minimizes the trace of $A'M^{-1}\left(\xi\right)A$. The D-optimal designs minimize $\left|A'M^{-1}\left(\xi\right)A\right|$. To emphasize

the dependence of the design on the matrix of coefficients A, this criterion is called D_A-optimality (Sibson, 1974).

$\mathbf{D}_S-$optimality D_S-optimal designs are appropriate when interest is in estimating a subset of s parameters as precisely as possible (Box, 1971). We can partition $\boldsymbol{\theta}$ into two sub-parameter vectors, i.e. $\boldsymbol{\theta} = \left[\boldsymbol{\theta}_1', \boldsymbol{\theta}_2'\right]'$, where $\boldsymbol{\theta}_1$ $(s \times 1)$ is the parameters of interest. The $(p-s) \times 1$ parameter vector $\boldsymbol{\theta}_2$ is treated as nuisance parameters.

To obtain expressions for the design criterion and related variance function, partition the information matrix as

$$M(\xi) = \begin{bmatrix} M_{11}(\xi) & M_{12}(\xi) \\ M_{21}(\xi) & M_{22}(\xi) \end{bmatrix}.$$

The covariance matrix for the least squares estimates of $\boldsymbol{\theta}_1$ is $M^{11}(\xi)$, the $s \times s$ upper left submatrix of $M^{-1}(\xi)$. It is easy to verify, from the results on the inverse of a partitioned matrix, that

$$M^{11}(\xi) = \left\{M_{11}(\xi) - M_{12}(\xi) M_{22}^{-1}(\xi) M_{21}'(\xi)\right\}^{-1}$$

The D_S-optimum design for $\boldsymbol{\theta}_1$ accordingly maximizes the following determinant

$$|M_{11}(\xi) - M_{12}(\xi) M_{22}^{-1}(\xi) M_{21}'(\xi)|.$$

c-optimality In c-optimality the research interest is in estimating the linear combination of the parameters $\mathbf{c}'\boldsymbol{\theta}$ with minimum variance. The design criterion to be minimized is thus

$$var\left(\mathbf{c}'\hat{\boldsymbol{\theta}}\right) \propto \mathbf{c}' M^{-1}(\xi)\,\mathbf{c}$$

where $\mathbf{c}$ is a $p \times 1$ vector.

Linear optimality (C- and L-optimality) Let L be a $p \times q$ matrix of coefficients. Then minimization of the criterion function

$$tr\left\{M^{-1}(\xi) L\right\}$$

leads to a linear, or L-optimal, design.

If L is of rank $s \leq q$ it can be expressed in the form $L = AA'$ where A is a $p \times s$ matrix of rank s. Then

$$tr\left\{M^{-1}(\xi) L\right\} = tr\left\{M^{-1}(\xi) AA'\right\} = tr\left\{A' M^{-1}(\xi) A\right\}$$

This form stresses the relationship with the D_A-optimal designs, where the determinant, rather than the trace, of $A' M^{-1}(\xi) A$ was minimized.

When $s = 1$, so that A becomes the $\mathbf{c}$ of c-optimality. If several contrasts are of interest, these can be written as the rows of $p \times s$ matrix C, then the criterion function is $tr\left\{C'M^{-1}\left(\xi\right)C\right\}$, whence the name C-optimality.

2.2.2 Optimal Design for Nonlinear Regression Models

A nonlinear model is a model such that at least one of the derivatives of the expected function with respect to the parameters depends on at least one of the parameters (Bates and Watts, 2007). Therefore, a nonlinear model's information matrix depends on at least one of the model parameters, which are not known exactly. If a previous estimate of parameters is plugged in to $M\left(\xi, \boldsymbol{\theta}\right)$ to find the optimal design, then the resulted optimal design is called a local optimal design (Atkinson, 1992). An alternative to this is a Bayesian optimal design, when a prior probability distribution for $\boldsymbol{\theta}$ is available. This extra information is incorporated into the design by taking expectation of the design criterion with respect to the prior distribution (Atkinson, 1992).

2.2.3 Bayesian Optimal Design

Because nonlinear models information matrix depends on unknown parameters values, people have been using Bayesian optimal designs for nonlinear

models for a long time.

Lindley (1956) proposed choosing an optimal design by maximizing the expected gain in Shannon information (Shannon, 1948) or, equivalently, maximizing the expected Kullback-Leiber distance between the posterior and prior distributions:

$$\int \log \frac{f(\boldsymbol{\theta}|\mathbf{y},\boldsymbol{\xi})}{f(\boldsymbol{\theta})} f\left(\mathbf{y},\boldsymbol{\theta}|\xi\right) d\boldsymbol{\theta} d\mathbf{y}$$

where $f\left(\boldsymbol{\theta}\right)$ is the prior density of the parameters $\boldsymbol{\theta}$, $f\left(\boldsymbol{\theta}|\mathbf{y},\xi\right)$ is the posterior density of the parameters $\boldsymbol{\theta}$. This design criterion is appropriate when the goal of the experiment is about the inference about the parameters.

The prior distribution does not depend on the design ξ, so the design ξ maximizing the expected gain in Shannon information is the one that maximizes:

$$\int \log\left\{f\left(\boldsymbol{\theta}|\mathbf{y},\xi\right)\right\} f\left(\mathbf{y},\boldsymbol{\theta}|\xi\right) d\boldsymbol{\theta} d\mathbf{y}$$

This is the expected Shannon information of the posterior density.

Corresponding to different local optimal designs, there are different Bayesian optimal designs which optimize the corresponding expected objective functions with respect to the prior distribution of the parameters (Atkinson, 1992). An optimal design which minimizes the expected value of the de-

terminant of the information matrix is referred to as a Bayesian-D optimal design. An optimal design which minimizes the expected value of the trace of the information matrix is referred to as a Bayesian-A optimal design.

When the researchers are not only interested in the inference of parameters, but also obtaining large values of outcome (response values), the following objective function proposed by Verdinelli and Kadane (1992) is appropriate:

$$\int \left[\rho \mathbf{y}'\mathbf{1} + \beta \log f\left(\boldsymbol{\theta}|\mathbf{y},\xi\right)\right] f\left(\mathbf{y},\boldsymbol{\theta}|\xi\right) d\mathbf{y} d\boldsymbol{\theta}$$

where ρ and β are nonnegative weights representing the relative contributions that the experimenter is willing to attach to the two components of the research objective.

Verdinelli (1992) suggested to maximize the following objective function to find an optimal design when the goal of the experiment is both inference about the parameters and prediction about the future observation.

$$\gamma \int \log f\left(y_{n+1}|\mathbf{y},\xi\right) f\left(\mathbf{y},y_{n+1}|\xi\right) d\mathbf{y} dy_{n+1} + \omega \int \log f\left(\boldsymbol{\theta}|\mathbf{y},\xi\right) f\left(\mathbf{y},\boldsymbol{\theta}|\xi\right) d\mathbf{y} d\boldsymbol{\theta}$$

where γ and ω express the relative contributions of the predictive and the inferential components of the objectives; y_{n+1} represents the future observation.

2.3 Methods

Although nonlinear models have been widely used in applied science, because of the limitation of currently available optimal designs methods for nonlinear models, many researchers are still picking design points based on intuition. The design picked in this way could be off the right track a lot. Traditional optimal designs in literatures, such as Silvey (1980), Atkinson and Donev (1992), assume a k point design scheme. This is rarely used in medical application since this design scheme usually requires more pipettes (measuring and moving of small amount of liquid) than design scheme such as dilution series. This could increase the measurement error. Second traditional optimal designs focus on linear combinations of parameters. To the best of our knowledge, optimal design for nonlinear combination of parameters hasn't been addressed. Objective functions of traditional optimal design methods could generate a design which requires high concentration of expensive reagents. The additional gain of information may not be worth the cost paid. There are 5 explicit objectives in the proposed research.

1. Develop a cost-efficient Bayesian optimal design method for precisely estimating nonlinear combinations of the parameters in nonlinear regression models with quantitative factors and design schemes deter-

mined by a few control variables.

2. Develop some R functions to implement the proposed optimal design method.

3. Illustrate the application of the proposed optimal design method using real examples.

4. Create R functions for visualizing and comparing experimental designs.

5. Compile the R functions to an R package and submit it to cran project.

2.3.1 Proposed Optimal Design Method

Consider the following nonlinear regression model.

$$Y = f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon^*$$

where Y is a continuous response; $\mathbf{x}$ is a $k \times 1$ quantitative predictor vector; $\boldsymbol{\theta} = [\theta_1, \cdots, \theta_p]$ is a $p \times 1$ parameter vector; ε^* is a random error with mean 0. $E(Y) = f(\mathbf{x}, \boldsymbol{\theta})$ where $f(.)$ is a nonlinear function with parameters $\boldsymbol{\theta}$.

The goal of the optimal design is to precisely estimate nonlinear combinations of the parameters in a nonlinear regression model with quantitative predictors. The nonlinear combination(s) could be of size 1 or more than 1. When the combinations have size more than 1, one problem may arise. A

design good for one parameter combination may be bad for another parameter combination. There is no clear cut solution to this problem. Currently, the optimal design typically maximizes some one dimensional function of the Fisher information.

Information Matrix of a Nonlinear Model Before deriving the information matrix, it is necessary to address the dependency between the mean and the variance of the response.

There are two major approaches for this problem. The first approach is called "transform both sides" approach (Carroll and Ruppert, 1982), which assumes the existence of a transformation ensuring the approximately equal variance across the range of the values of the predictors $\mathbf{x}$. We denote the transformation by $t(.)$. With this transformation, the model equation can be rewritten as

$$Z = t(Y) = t(f(\mathbf{x}, \boldsymbol{\theta})) + \varepsilon$$

where $\varepsilon \sim$Normal$(0, \sigma^2)$. The other approach to handle this problem is to assume the existence of a variance function $v(.)$ such that

$$Y \sim \text{Normal}(\mu, v(\mu)\sigma^2)$$

where $\mu = f(\mathbf{x}, \boldsymbol{\theta})$

(Raab, 1981; Carroll and Ruppert, 1982; Davidian and Carroll, 1987; Davidian, Carroll and Smith, 1988). In our work we focus on the first approach.

Let $f_i = f(\mathbf{x}_i, \boldsymbol{\theta})$ where $\mathbf{x}_i$ is the predictor value for the i^{th} design point. The design points $\mathbf{x}_i, i = 1, \cdots, n$ are determined by a few control variables $\boldsymbol{\delta}$. The optimal design is found with respect to these control variables.

Let $r_i = t(y_i) - t(f_i)$ where $t(.)$ is the transformation function.

The log likelihood for the i^{th} data point is

$$l_i = -1/2 \log(2\pi\sigma^2) - 1/2\sigma^{-2} (t(y_i) - t(f_i))^2$$

The gradient $\mathbf{g}_i = [g_{1i}, \cdots, g_{pi}]'$ of l_i with respect to the parameters vector $\boldsymbol{\theta}$ is given by

$$g_{di} = \frac{\partial l_i}{\partial \theta_d} = \sigma^{-2} r_i \frac{\partial t(f_i)}{\partial \theta_d} \; d = 1, \cdots, p$$

The information matrix for the i^{th} data point is $r_i^2 \sigma^{-4} \boldsymbol{\mu}_i \boldsymbol{\mu}_i'$, where

$\boldsymbol{\mu}_i = \frac{\partial t(f_i)}{\partial \boldsymbol{\theta}}$ is a $p \times 1$ vector.

Then the information matrix of the regression parameters for the whole data set is $\sum_{i=1}^{n} r_i^2 \sigma^{-4} \boldsymbol{\mu}_i \boldsymbol{\mu}_i'$, where n is the sample size.

The expected information matrix is $I = \sigma^{-2} \sum_{i=1}^{n} \boldsymbol{\mu}_i \boldsymbol{\mu}_i'$.

The asymptotic covariance matrix for the least square estimate of the parameters $\boldsymbol{\theta}$ is I^{-1} (Jennrich, 1969). Note that it depends on the control variables vector $\boldsymbol{\delta}$.

Optimal Design for Estimating Nonlinear Combinations of Parameters In practice researchers are sometimes interested in precisely estimating nonlinear combination(s) of parameters. We denote it as $\boldsymbol{\phi}$, which is a $s \times 1$ vector ($s \leq p$). When $\boldsymbol{\phi}$ is linear combination(s) of $\boldsymbol{\theta}$, the optimal design has been studied by Sibson (1974). When $\boldsymbol{\phi}$ is nonlinear combination(s) of the parameters, if the parameter estimates $\hat{\boldsymbol{\theta}}$ has normal distribution, applying the delta method (Cassela and Berger, 2002), the asymptotic covariance matrix of the estimates of $\boldsymbol{\phi}$ can be expressed as

$$AI^{-1}A' \tag{9}$$

where $A = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{\theta}}$ is a $s \times p$ matrix.

The most popular design criterion in application is D-optimality, because the D-optimal design is scale invariant (Atkinson and Donev, 1992) to quan-

titative factors. This is generally not the case for A- and E-optimality. It is not desirable that the optimal design should depend on the scale that the factor is measured at, such as meter or centimeter. Since the factors in the model of interest are quantitative, we choose to minimize the log determinant of $AI^{-1}A'$ with respect to the control variables vector $\boldsymbol{\delta}$ to obtain the D-optimal design for precisely estimating $\boldsymbol{\phi}$.

For a nonlinear model, its information matrix depends on the parameters. If $\boldsymbol{\phi}$ is a nonlinear function of the parameters, the A matrix in the asymptotic covariance matrix of the estimate of $\boldsymbol{\phi}$ (equation 9) depends on the parameters as well. To compute the log determinant of the asymptotic covariance matrix of the estimate of $\boldsymbol{\phi}$, we need the values of the parameters, which are not known exactly. We could plug in the point estimates of the parameters from previous experiments to obtain a local optimal design. This method is often not appropriate for the following reasons. When the estimates of the parameters are different from the true parameters, the achieved local design could be quite different from the true optimal design. In medical application, the found optimal design will be used to estimate the parameters for many different patients. Different patients may have different parameter values. Local optimal design could be good for some patients, but bad for

the other patients. Since the optimal design will be used for the estimation for all patients, the optimal design should be able to accommodate a range of parameter values. Instead of using a local optimal design, we minimize the expected value of log determinant of the asymptotic covariance matrix of the estimate of $\boldsymbol{\phi}$ with respect to a prior distribution of the parameters. The resulted optimal design is referred to as a Bayesian optimal design, which doesn't require exact values of the parameters. The Bayesian optimal design minimizes the following objective function with respect to the control variables vector $\boldsymbol{\delta}$

$$\int \log \left| A\left(\boldsymbol{\delta};\boldsymbol{\theta}\right) I\left(\boldsymbol{\delta};\boldsymbol{\theta}\right)^{-1} A'\left(\boldsymbol{\delta};\boldsymbol{\theta}\right) \right| f\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta}$$

where $f\left(\boldsymbol{\theta}\right)$ is the prior density function of the parameters $\boldsymbol{\theta}$. Prior distribution could be summarized from historical data. One example can be found in Clyde et al. (1996).

Optimal designs found using this objective function may require usage of high concentration of expensive reagents. In order to take cost into consideration, we further extend the objective function by adding a term to penalize designs with high cost. Such an objective function can be expressed as:

$$\int \log \left| A\left(\boldsymbol{\delta};\boldsymbol{\theta}\right) I\left(\boldsymbol{\delta};\boldsymbol{\theta}\right)^{-1} A'\left(\boldsymbol{\delta};\boldsymbol{\theta}\right) \right| f\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} + w\left(\boldsymbol{\delta}\right)$$

where $w(.)$ is a monotone function of the cost of the design. The optimal design is the design which minimizes the objective function.

More specifically the optimal design can be found by minimizing the following objective function:

$$\int \log |A(\boldsymbol{\delta};\boldsymbol{\theta})\, I(\boldsymbol{\delta};\boldsymbol{\theta})^{-1}\, A'(\boldsymbol{\delta};\boldsymbol{\theta})|\, f(\boldsymbol{\theta})\, d\boldsymbol{\theta} + c(\boldsymbol{\delta})\, \omega$$

where $c(\delta)$ is the total cost of the experiment design; ω is a nonnegative weight chosen by the researchers to represent the relative importance of saving money compared to increasing estimation precision.

2.4 An Example

Development of this project is motivated by an application in pharmacodynamic study.

2.4.1 PKPD Models

Pharmacodynamics is the study of the biochemical and physiological effects of drugs on bodies (human or animal) or on microorganisms or parasites within or on the bodies and the mechanisms of drug action and the relationship between drug concentrations and effects (Lees et al., 2004).

Pharmacokinetics is a branch of pharmacology dedicated to the determination of the fate of substances administered externally to a living organism.

Pharmacodynamics is often summarized as the study of what a drug does to the body, whereas pharmacokinetics is the study of what the body does to a drug. Pharmacodynamics is sometimes abbreviated as "PD". Pharmacokinetics is sometimes abbreviated as "PK". Pharmacokinetics is often studied in conjunction with pharmacodynamics. In this case they are often jointly referred to as "PKPD". Statistical models for PKPD processes are referred to as PKPD models in literature.

2.4.2 Gaddum/Schild Model

Gaddum/Schild model is a widely used PD model. It has been used in a number of cases to model responses to combined agonist and antagonist stimuli (Williams et al. 1988; Swartz et al. 1992; Lazareno and Birdsall, 1993; Maguire and Davenport, 1995; Maguire et al. 1997; Motulsky and Christopoulos, 2003). An agonist is a drug that binds to a receptor and causes a response. An antagonist is a drug that does not provoke a response itself, but blocks agonist-mediated response. The relationship is described by the following equation:

$$Y = f([A],[B]) + \varepsilon^* = Min + \frac{Max - Min}{1 + \left(\frac{10^{pEC50}\left(1 + \left(\frac{[B]}{10^{-pA2}}\right)^S\right)}{[A]} \right)^{Hill}} + \varepsilon^*$$

where Y is a continuous response, $[A]$ and $[B]$ are two quantitative predictors. $[A]$ is the agonist concentration and $[B]$ is the antagonist concentration. There are 6 parameters in this model: Min, Max, pEC50, Hill, S, and pA2. Min is the baseline response; Max is the maximum response; pEC50 is the negative logarithm of the EC50; Hill is the Hill slope; S is the Schild slope; pA2 is the negative logarithm of the antagonist concentration which doubles the EC50; and ε^* is a random error with mean 0. For simple competitive antagonist, Schild slope equals to 1.

EC50 is the agonist concentration that provokes a response halfway between baseline and maximum response (Figure 2).

The maximum response y_0 for given concentration of agonist is obtained when there is no antagonist present. IC50 is defined as the concentration of the antagonist that reduces the response to $(y_0 + Min)/2$ (Figure 3). More generally ICx is defined to be the antagonist concentration, which brings the response to the value $Min + x\%\,(y_0 - Min)$.

Using the definition of ICx, it can be shown that ICx is a function of the parameters and the given agonist concentration $[A]_g$. In practice, $[A]_g$ is

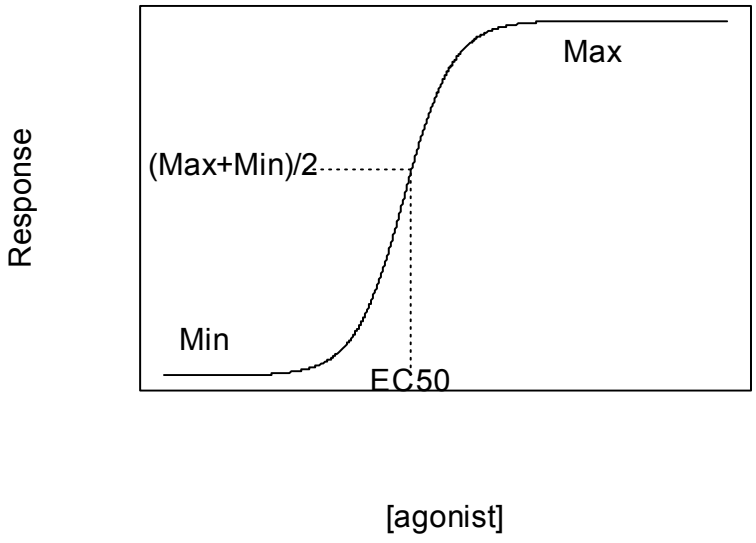


Figure 2: EC50

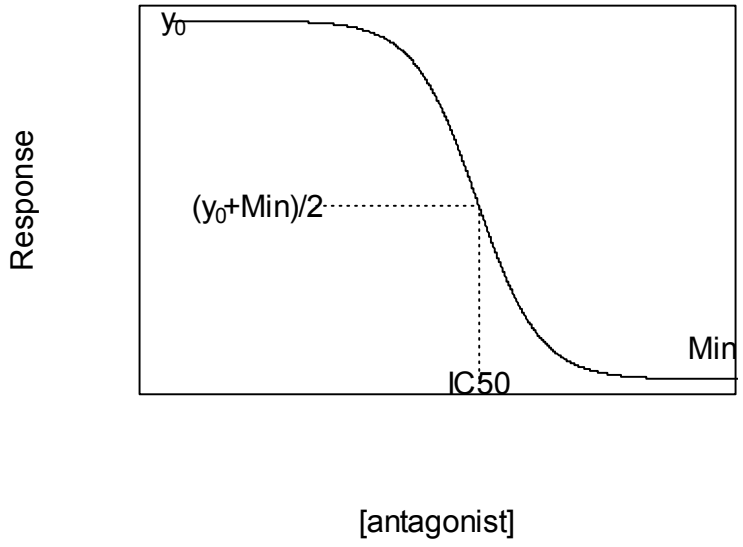


Figure 3: IC50

provided by the researcher based on the goal of the research.

$$ICx = \left(\left(\frac{1}{x\%} - 1 + \frac{1}{x\%} \left(\frac{10^{pEC50}}{[A]_g} \right)^{Hill} \right)^{1/Hill} \frac{[A]_g}{10^{pEC50}} - 1 \right)^{1/S} - 10^{-pA2}$$

In drug discovery, researchers are mainly interested in estimating both EC50 and ICx. Notice that they are nonlinear combinations of the parameters in Gaddum/Schild model.

For illustrative purpose, we show a contour plot of the square root transformed response surface as a function of the log concentration of the agonist and antagonist in figure 4.

Optimal Design for the Estimation of pEC50 and log(IC50) We considered the optimal design for Gaddum/Schild model. The purpose of the experiment is to precisely estimate pEC50 and log(IC50) simultaneously in a cost efficient way.

In practice since the solution should be easily prepared, researchers usually prepare solutions with 0 concentration and serial dilutions with same dilution factors. These concentrations may be replicated several times. Using more complicated design schemes may increase measurement error because more complex manipulation is required.

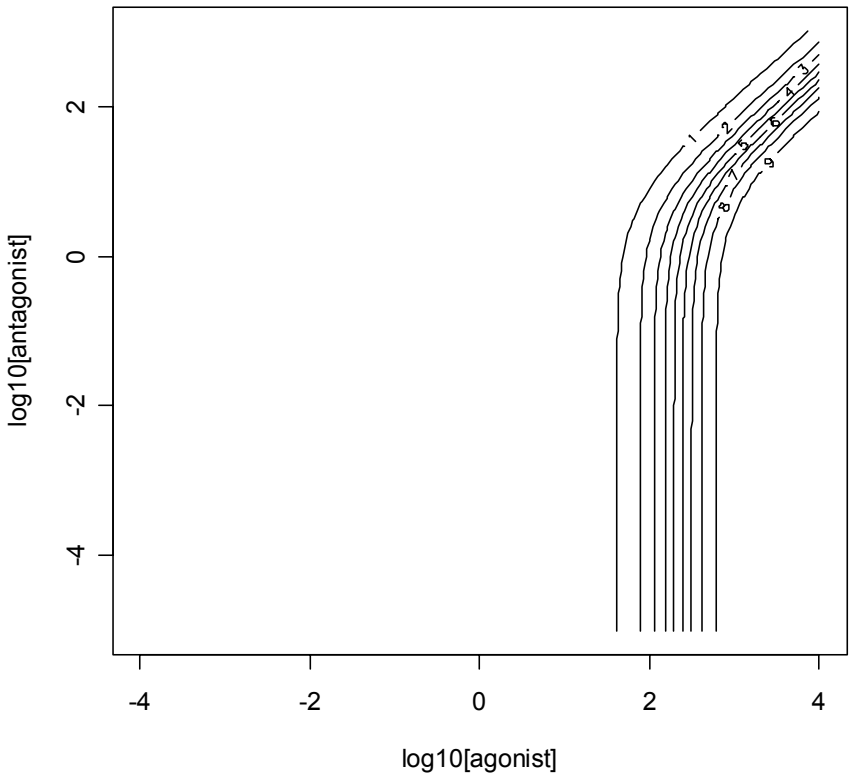


Figure 4: Example Contour Plot for Square Root Transformed Gaddum/Schild Model

Table 20: Illustration of Dilution Series Design

	[AA] →									
	0	3.9	7.8	15.6	31.3	62.5	125	250	500	1000
[BB] 0	X	X	X	X	X	X	X	X	X	X
↓ 30	X	X	X	X	X	X	X	X	X	X
300	X	X	X	X	X	X	X	X	X	X
3000	X	X	X	X	X	X	X	X	X	X

Let's consider a hypothetical experiment. In this experiment, AA binds to its receptor and triggers the release of YY. BB binds to the receptor of AA and blocks its binding with AA. AA is the agonist in this experiment. BB is the antagonist in this experiment. The relationship between the concentration of YY and the concentration of AA, BB is well modeled by Gaddum/Schild model.

Suppose previous experiments show that square root transformation on both sides of Gaddum/Schild model leads to normal and stable random error. That is

$$Z = \sqrt{Y} = \sqrt{Min + \frac{Max - Min}{1 + \left(\frac{10^{pEC50}\left(1 + \left(\frac{[B]}{10^{-pA2}}\right)^{S}\right)}{[A]} \right)^{Hill}}} + \varepsilon$$

where $\varepsilon \sim$Normal$(0, \sigma^2)$. In this example, we assume $\sigma^2 = 0.163$.

In table 20, we illustrate one possible zero concentration plus serial dilution experimental design.

In this design, the maximum agonist concentration (max_a) is 1000; the maximum antagonist concentration (max_b) is 3000. The dilution factor for agonist (d_a) is 0.5. The dilution factor for antagonist (d_b) is 0.1. The number of dilutions for agonist (nd_a) is 8. The number of dilutions for antagonist (nd_b) is 2. The total number of distinct design points is 40. Each distinct design point is replicated twice. Here max_a, max_b, d_a, d_b, nd_a, and nd_b are design variables controlled by the researchers.

The prior distribution of the parameters can be extracted from historical data. We assume the parameters have the following uniform distribution and all the parameters are independent of each other.

Min$\sim$Uniform(0.00319, 2.629); Max$\sim$Uniform(21.54, 218.11);

pEC50$\sim$Uniform(1.91, 2.98); Hill$\sim$Uniform(1.48, 2.73);

pA2$\sim$Uniform(-2.24, 0.39); S$\sim$Uniform(0.32, 1.5);

We also use the following setting in the example. The unit price for the agonist is 1. The unit price for the antagonist is 10. The researchers think reducing the average log generalized variance (logvar0) 1 unit is worth spending 1000 unit money. Every design point is replicated twice. We use the following objective function to find the cost efficient optimal design for

precisely estimating both pEC50 and log(IC50).

$$\int\int\int \log|A\left(\max.a, \max.b, d.a, d.b, nd.a, nd.b; Min, Max, pEC50, Hill, pA2, S\right)$$

$$I\left(\max.a, \max.b, d.a, d.b, nd.a, nd.b; Min, Max, pEC50, Hill, pA2, S\right)^{-1}$$

$$A'\left(\max.a, \max.b, d.a, d.b, nd.a, nd.b; Min, Max, pEC50, Hill, pA2, S\right)|$$

$$f(Min, Max, pEC50, Hill, pA2, S) dMin\ dMax\ dpEC50\ dHill\ dpA2\ dS +$$

$$c\left(\max.a, \max.b, d.a, d.b, nd.a, nd.b\right)\omega$$

where $c\left(\max.a, \max.b, d.a, d.b, nd.a, nd.b\right) =$

$$\max.a(1 - d.a^{(nd.a+1)})/\left(1 - d.a\right)\left(nd.a + 2\right) +$$

$$\max.b\left(1 - d.b^{(nd.b+1)}\right)/\left(1 - d.b\right)\left(nd.b + 2\right)10;$$

A is the derivative matrix of pEC50 and log(IC50) with respect to the parameters Min, Max, pEC50, Hill, pA2, and S; I is the information matrix of the parameters; f(.) is the prior density function.

With the total number of distinct design points to be 40, there are 4 possible combinations of nd.a and nd.b. They are (nd_a=2, nd_b=8), (nd_a=8, nd_b=2), (nd_a=3, nd_b=6), and (nd_a=6, nd_b=3). Table 21 lists the optimal designs for each of these combinations with $w = 0$. These optimal designs don't take cost into consideration.

The favorite optimal design without taking cost into consideration has

Table 21: Optimal Designs without Taking Cost into Consideration

nd_a	nd_b	max_a	max_b	d_a	d_b	logvar0
8	2	4836	60	0.60	0.14	-11.23
2	8	3700	1026	0.19	0.24	-10.91
6	3	6974	1881	0.49	0.07	-11.29
3	6	3828	72	0.30	0.23	-11.00

nd.a: number of dilution for agonist;
nd.b: number of dilution for antagonist;
max_ a: maximum agonist concentration;
max_ b: maximum antagonist concentration;
d_ a: dilution factor for agonist;
d_ b: dilution factor for antagonist;
logvar0: average log generalized variance for pEC50 and log(IC50) wtih respect to prior distribution of parameters.

the number of dilution for agonist equals to 6 and number of dilution for antagonist equals to 3.

Table 22 lists the cost efficient Bayesian optimal designs with $w = 0.001$.

After taking cost into consideration, the favorite optimal design has the number of dilution for agonist equals to 8 and the number of dilution for antagonist equals to 2. This is because the antagonist is more expensive. Putting more points on the antagonist side will increase the cost and therefore make the design less favored.

Table 23 compares the regular and cost efficient optimal design holding the number of dilution for agonist equals to 6 and number of dilution for

Table 22: Cost Efficient Bayesian Optimal Designs

nd_a	nd_b	max_a	max_b	d_a	d_b	logvar0	cost	logvar
8	2	350	5	0.39	0.21	-3.99	5843	1.84
2	8	207	8.2	0.41	0.24	-3.07	7427	4.35
6	3	315	6	0.37	0.20	-3.84	6193	2.35
3	6	238	7.4	0.39	0.21	-3.33	7034	3.7

nd.a: number of dilution for agonist;
nd.b: number of dilution for antagonist;
max_ a: maximum agonist concentration;
max_ b: maximum antagonist concentration;
d_ a: dilution factor for agonist;
d_ b: dilution factor for antagonist;
logvar0: average log generalized variance for pEC50 and log(IC50) wtih respect to prior distribution of parameters.
cost: total cost of the experiment.
logvar is the value of the objective function.

antagonist equals to 3.

The cost of the cost efficient optimal design is much smaller compared to the regular optimal design, while the estimation precision of the regular optimal design is much higher than the cost efficient optimal design. While using cost efficient optimal design, we sacrifice estimation precision to cut down cost. The maximum concentration of agonist is reduced 20 times in the cost efficient optimal design compared to the regular optimal design. In contrast the maximum concentration of the antagonist is reduced 200 times. This is because the antagonist is more expensive, reducing the antagonist

Table 23: Optimal Designs with Uniform Priors

design	max_a	max_b	d_a	d_b	logvar0	cost
cost efficient	315	6	0.37	0.20	-3.84	6193
regular	6974	1881	0.49	0.07	-11.29	459422

nd.a: number of dilution for agonist;
nd.b: number of dilution for antagonist;
max_a: maximum agonist concentration;
max_b: maximum antagonist concentration;
d_a: dilution factor for agonist;
d_b: dilution factor for antagonist;
logvar0: average log generalized variance for pEC50 and log(IC50) wtih respect to prior distribution of parameters.
cost: total cost of the experiment.
logvar is the value of the objective function.

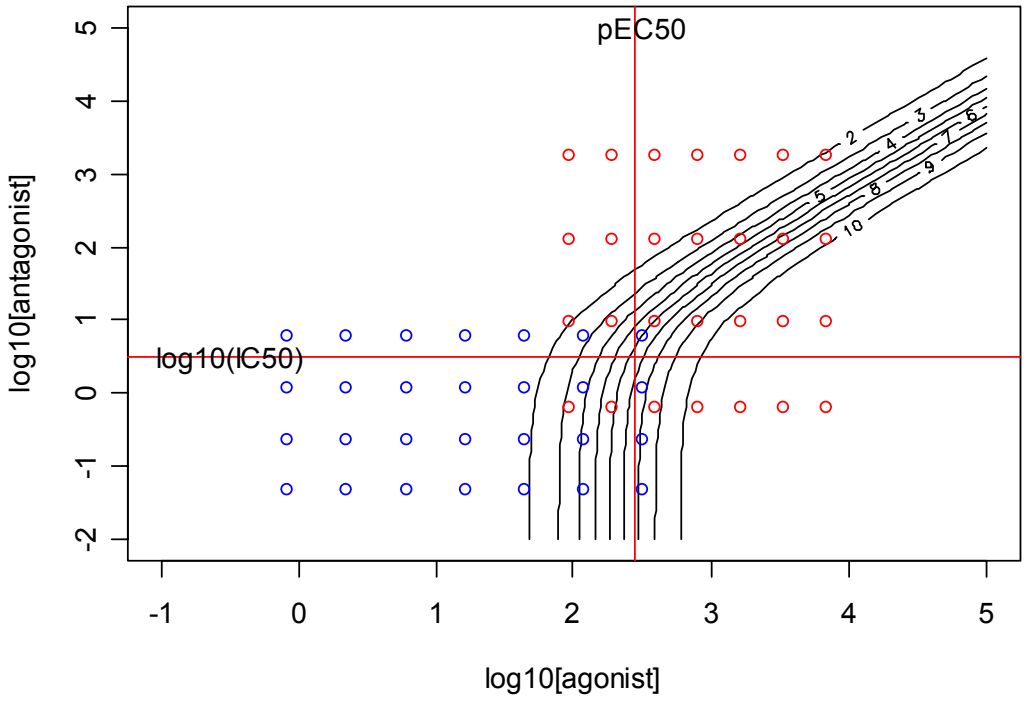
concentration is more efficient way to cut cost.

Figure 5 compares the cost efficient and regular Bayesian optimal design holding the number of dilution for agonist equals to 6 and number of dilution for antagonist equals to 3. Notice that the cost efficient Bayesian optimal design moves points to lower agonist and antagonist concentration to cut the cost.

2.5 Implementation

Bayesian optimal design has existed for a while. But very few Bayesian optimal designs have been used in real scientific research (Chaloner, 1995). One

Figure 5: Comparison Between Regular and Cost Efficient Bayesian Optimal Design Holding the Number of Dilution of agonist equals to 6, and Number of Dilution for Antagonist Equals to 3. Red Dots Represent the Regular Optimal Design. Blue Dots Represent the Cost Efficient Optimal Design.



of the reasons is lack of appropriate software. To help researchers implement the proposed method, an R package ODK has been developed.

Building package ODK involves many cutting-edge techniques such as numerical computation of multiple-dimension derivatives, multiple-dimension integrals, multiple-dimension optimization, adaptive quadratures, symbolic computation, passing unknown number of parameters. Multiple-dimension derivatives can be found symbolically using R function deriv in stat package (Griewank and Corliss, 1991). In package ODK, multiple-dimension integral is computed using adaptive quadrature. Basically adaptive quadrature is a process in which the integral of a function is approximated using quadrature rules on adaptively refined subintervals. Adaptive quadrature is effective for "bad behaved" integrand, for which traditional methods fail (Rice, 1975). In package ODK, adaptive quadrature is implemented using R function adapt in adapt package. R has a very flexible data type called list. A list can accomodate different types of data with different length. In package ODK, passing unknown number of parameters is achieved by using lists as arguments for defined functions. The proposed method requires multiple-dimension optimization, which is implemented using R function optim in stat package.

Function "OD" implements the proposed cost efficient Bayesian optimal

design. It requires the following input.

- Expressions for the nonlinear model, nonlinear combinations of parameters to be precisely estimated, and the derivatives of these expressions with respect to the parameters in the model.
- A prior density function of the parameters.
- A function which generates the design points using design variables as input.
- Unit costs for all predictors.
- A weight (ω) value used in the objective function.
- Initial guess of the values of the control variables which optimize the objective function.
- Lower bounds and upper bounds of the control variables space.
- Choice of optimization algorithem.
- Variance of the random error in the nonlinear model after normalization.
- Arguments for the argument functions provided.

Symbolic mathematical expression can be formed using R function as.formula. The generated output includes the values of the control variables, and the value of the objective function for the optimal design.

See appendix for example R codes about how to use the functions in the pacakge to find a cost efficient Bayesian optimal design.

Since there are many arguments for the OD function, some interface will be built in the future to prompt the user to provide all the information required for the computation.

With a computer with a Pentium(R) 4, 2.8 GHz CPU and a 2.79 GHz, 1.99 GB RAM, it takes about 100 minutes to find the cost efficient optimal design for a case similar to the previous example. The algorithm is computationally expensive at this point. Considering the fast development of computation technology, this limitation won't be an issue in the future.

The proposed method requires inversing the Fisher Information matrix. Because of precision limit for different computation system, computationally singular matrix may be encountered. If this happens, a computer with higher computation precision can solve the problem.

The package contains a function called "logdm.cost.int". This function can compute the value of the objective function for any design that a re-

searcher can come up with. Considering the complexity of reality, this function could be very helpful. The function provides a way for the researchers to select the best design from several candidate designs.

The package has two functions related to dilution series design scheme. The function "dilution" generates design points for a dilution series design scheme. The function "design.points" adds deisgn points from a dilution series design scheme to an existing plot.

2.6 Conclusion

In the proposed work, a cost efficient Bayesian optimal design method is proposed for precisely estimating nonlinear (or linear) combinations of parameters in a nonlinear model with quantitative predictors. The proposed method balances the information for multiple parameters of interest. It balances the cost of the design and the estimation precision. It accommodates a range of parameter values. An R package has been developed to implement the proposed method.

2.7 References

1. Atkinson, A. C., and Donev, A. N. (1992), *Optimum Experimental Designs*, Oxford, U.K.: Oxford University Press.

2. Bates, D., and Watts, D. (2007), *Nonlinear Regression Analysis and Its Application*, Wiley, Hoboken, NJ.

3. Box, M. J. (1971), "An Experimental Design Criterion for Precise Parameter Estimation of a Subset of the Parameters in a Nonlinear Model," *Biometrika*, 58, 149-153.

4. Box, G. E. P., and Draper, N. R. (1987), *Empirical Model Building and Response Surfaces*, Wiley, New York, NY.

5. Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995), "A limited memory algorithm for bound constrained optimization,". *SIAM J. Scientific Computing*, 16, 1190–1208.

6. Carroll, R. J., and Ruppert, D. (1982), "Robust Estimation in Heteroscedastic Linear Models, The Annals of Statistics," 10, 429-441. —–(1988), *Transformation and Weighting in Regression*, London: Chapman and Hall.

7. Casella G., and Berger. R. L. (2002), *Statistical Inference*, 2nd ed. Duxbury Press, Pacific Grove, CA.

8. Chaloner K and Verdinlli I (1995), "Bayesian Experimental Design: A Review," *Statistical Science*, 3, 273-304.

9. Clyde, M., Muller, P., Parmigiani, G., (1996), "Inference and design strategies for a hierarchical logistic regression model." In: Berry, D., Stangel, D. (Eds.), *Bayesian Biostatistics*, Marcel Dekker (New York).

10. Davidian, M., and Carroll, R. J. (1987), "Variance Function Estimation," *Journal of the American Statistical Association*, 82, 1079-1091.

11. Davidian, M., Carroll, R. J., and Smith, W. (1988), "Variance Functions and the Minimum Detectable Concentrations in Assays," *Biometricka*, 75, 549-556.

12. Griewank, A. and Corliss, G. F. (1991) "Automatic Differentiation of Algorithms: Theory, Implementation, and Application," *SIAM* proceedings, Philadelphia.

13. Jennrich, R. I. (1969), "Asymptotic Properties of Nonlinear Least Squares Estimators," *The Annals of Mathematical Statistics*, 40: 633-643.

14. Lazareno, S., and Birdsall, NJM. (1993), "Estimation of competitive antagonist affinity from functional inhibition curves using the Gaddum, Schild and Cheng-Prusoff equations," *British Journal of Pharmacology*, 109 (4), 1110-1119.

15. Lees, P., Cunningham, F. M., Elliott, J. (2004), "Principles of pharmacodynamics and their applications in veterinary pharmacology," *Journal of Veterinary Pharmacology and Therapeutics,* 27, 397–414.

16. Lindley, D. V. (1956), "On the Measure of Information Provided by an Experiment," *Annal of Statistics*, 27: 986-1005.

17. Maguire, J. J., and Davenport, A. P. (1995), "ET(A) Receptor-mediated Constrictor Responses to Endothelin Peptides in Human Blood-vessels in-vitro," *British Journal of Phamacology*, 115: 191-197.

18. Maguire, J. J., Kuc, R. E., and Davenport, A. P. (1997), "Affinity and Selectivity of PD156707, a Novel Nonpeptide Endothelin Antagonist, for Human ET(A) and ET(B) receptors," *Journal of Pharmacology and Experimental Therapeutics*, 280: 1102-1108.

19. Matthew, J. N. S. and Allcock, G. C. (2004), "Optimal Designs for

Michaelis-Menton Kinetics Studies," *Statistics in Medicine*, 23: 477-491.

20. Motulsky, H., and Christopoulos, A. (2003), *Fitting Models to Biological Data Using Linear and Nonlinear Regression*, 256-295, GraphPad Software, Inc., San Diego, CA.

21. O'Brien, T. E. (2005), "Designing for Parameter Subsets in Gaussian Nonlinear Regression Models," *Journal of Data Science*, 3: 179-197.

22. O'Brien, T. E. and Funk, G. M. (2003), "A Gentle Introduction to Optimal Designs for Regression Models," *The American Statistician*, 57(4), 265-267.

23. Raab, G. M. (1981), "Estimation of a Variance Function With Application in Immunoassay," *Applied Statistics*, 30, 32-40.

24. Rice, J. R. (1975), "A Metalgorithem for Adaptive Quadrature," *Journal of the Association of Computing Machinary*, 22, 61-82.

25. Rocke, D. M. and Jones, G. (1997), "Optimal Design for ELISA and Other Forms of Immunoassay," *Technometrics*, 39, 162-170.

26. Seber, G. A. and Wild, C. J. (1989), *Nonlinear Regression*, Wiley, Hoboken, NJ.

27. Sibson, R. (1974), "D_A-optimality and duality," In *Progress in Statistics*, vol II (Edited by Gani, J., Sakadi, K., and Vincze, I.), 9th European Meeting of Statistics, Budapest, 1972, Amsterdam, North-Holland.

28. Silvey, S. D. (1980), *Optimum Design*, Chapman & Hall, London.

29. Swartz, K. J., Koroshetz, W. J., Rees, A. H., and Huettner, J. E. (1992), "Competitive antagonism of Glutamate Receptor Channels by Substituted Benzazepines in Cultured Cortical-neurons," *Molecular Pharmacology*, 41(6), 1130-1141.

30. Verdinelli, I. (1990), *Bayesian Statistics*, Oxford University, Oxford.

31. Verdinelli, I. and Kadane, J. B. (1992), "Bayesian Designs for Maximizing Information and Outcome," *Journal of the American Statistics Association*, 87: 510-515.

32. Williams, T. L., Smith, DAS, Burton, N. R., and Stone, T. W. (1988), "Amino-acid Pharmacology in Neocortical Slices-evidence for Biomole-

cular Actions Form an Extension of the Hill and Gaddum-Schild equations," *British Journal of Pharmacology*, 95(3), 805-810.

3 Appendix

3.1 R Codes for Implementing Cost Efficient Bayesian Optimal Design and Ancillary Functions

```
#The following functions help to find a cost-efficient Bayesian optimal
    #design for any combinations of parameters for any model with stable
    #normal random error and quantitative predictors with respect to some
    #design variables;
    #Reference: Tang and Yang (2008)
    library(faraway)
    library(adapt)
    dilution <- function(xp, d.args) {
    nd.a <- d.args[[1]]; nd.b <- d.args[[2]]; zero <- d.args[[3]]; km <- d.args[[4]];
    #Checking input arguments for vaildity;
    if(!is.vector(xp, mode="numeric")) stop("control variables must be in a
numeric vector!")
    max.a <- 10^xp[1];max.b <- 10^xp[2];dilution.a <- ilogit(xp[3]);dilution.b
<- ilogit(xp[4]);
    #design points
```

```
#a is the concentrations for agonist;

#b is the concentrations for antagonist;

a <- dilution.a^seq(from=0, to=nd.a)*max.a;

if (zero==TRUE) {a <- c(0, a); nd.a2 <- nd.a+2;}

else {nd.a2 <- nd.a+1};

b <- dilution.b^seq(from=0, to=nd.b)*max.b;

if (zero==TRUE) {b <- c(0, b); nd.b2 <- nd.b+2;}

else {nd.b2 <- nd.b+1};

Data <- matrix(0, nd.a2*nd.b2, 2)

for (i in 1:nd.a2) {

for (j in 1:nd.b2) {

Data[(i-1)*nd.b2+j,] <- c(a[i], b[j])

}

}

Data.all <- matrix(0, km*nd.a2*nd.b2, 2)

for (k in 1:km) {

Data.all[(nd.a2*nd.b2*(k-1)+1):(nd.a2*nd.b2*k),] <- Data

}

Data.all
```

```
}
cost <- function(design.points, pr) {
d <- length(pr)
if(d!=dim(design.points)[2]) stop("predictor dimension and weight dimension do not match!")
cost <- drop(apply(design.points, 2, sum)%*%pr)
}
#logdm.cost computes log|A'I-1A|+cost*weight
logdm.cost <- function(theta, x, design, sigma, input, funs, pr, weight, d.args, jpdf, ...) {
#check validity of arguments
if (!is.vector(theta, mode="numeric")) stop("input parameters must be a numeric vector!")
if (!is.list(input)) stop("input for parameters of interest must be a list with lists as elements!")
#generate design points and replace 0 by a very small value
design.points <- design(x, d.args)
design.points[design.points==0] <- 1e-30
#information matrix
```

```
d <- length(theta);
I <- matrix(0, d, d);
for (i in 1:dim(design.points)[1]) {
point <- design.points[i,]
point.list <- vector("list", dim(design.points)[2])
for (j in 1:dim(design.points)[2]) {
point.list[[j]] <- point[j]
}
u <- attr(do.call(funs[[1]], c(theta, point.list)), "gradient")
V <- (t(u)%*%u)/sigma
I <- I+V
}
#delta method variance
n <- length(input);
A <- matrix(0, n, d)
for (i in 1:n) {
A[i,] <- attr(do.call(funs[[i+1]], c(theta, input[[i]])), "gradient")
}
AIA <- A%*%solve(I, tol=1e-30)%*%t(A)
```

```
    res <- (log(det(AIA))+cost(design.points, pr)*weight)*jpdf(theta, ...)
    }
    logdm.cost.int <- function(x, int.low, int.up, design, sigma, input, funs,
pr, weight, d.args, jpdf, ...) {
    if (!is.vector(int.low, mode="numeric") | !is.vector(int.up, mode="numeric"))
stop("lower and upper
    bound of parameters must be numeric vectors!")
    if (length(int.low)!=length(int.up)) stop("lower and upper bound of pa-
rameters should have same dimension!")
    ndim <- length(int.low)
    if (ndim==1) {integrate(logdm.cost, lower=int.low, upper=int.up, x=x,
    design=design, sigma=sigma, input=input, funs=funs, pr=pr, weight=weight,
d.args=d.args, jpdf=jpdf, ...)$value }
    else { adapt(ndim=ndim, lower=int.low, upper=int.up, funct=logdm.cost,
x=x,
    design=design, sigma=sigma, input=input, funs=funs, pr=pr, weight=weight,
d.args=d.args, jpdf=jpdf, ...)$value }
    }
    OD <- function(initial, method, int.low, int.up, design, sigma, input,
```

```
funs, pr, weight, d.args, jpdf, ...) {
    od <- optim(par=initial, fn=logdm.cost.int, method=method, int.low=int.low,
int.up=int.up,
    design=design, sigma=sigma, input=input, funs=funs, pr=pr, weight=weight,
d.args=d.args, jpdf=jpdf, ...)
    design.points <- design(od$par, d.args);
    design.cost <- cost(design.points, pr);
    logvar0 <- logdm.cost.int(x=od$par, int.low=int.low, int.up=int.up, de-
sign=design, sigma=sigma,
    input=input, funs=funs, pr=pr, weight=0, d.args=d.args, jpdf=jpdf, ...);
    logvar <- od$value;
    c(control=od$par, logvar=logvar, convergence=od$convergence, counts=od$counts,
cost=design.cost,
    logvar0=logvar0, pr=pr, weight=weight);
    }

    design.points <- function(max.a, max.b, d.a, d.b, nd.a, nd.b, color) {
    a <- max.a*d.a^seq(0, nd.a)
    b <- max.b*d.b^seq(0, nd.b)
    for (i in 1:length(a)){
```

```
for (j in 1:length(b)){
points(log10(a[i]),log10(b[j]), col=color)
}
}
}
```

3.2 R Codes for Generating Clustered Binary Data Using Qaqish's Method

```
################################
#The following functions generate clustered binary data
#using Qaqish (2003);
#Specifically the data is generated with marginal means
#and correlations within clusters;
#The function corbin generates clustered binary data
#based on logit of the response, correlation parameters,
#and the correlation sturcture (exchangable or ar1);
#Variables
#ly: logit vector of response;
```

```
#alpha: correlation parameter in exchangalbe or ar1
#structure;
#p: marginal probability vector of success of response;
#v: marginal variance vector of response;
#b: linear coeffecients vector in the conditional linear
#family;
#l: conditional probability; The conditional probability
#could be bigger than 1 and generates NA values;
############################
library(faraway);
library(bindata);
logloginv <- function(x) {
1-exp(-exp(x))
}
corbin.ex <- function(ly, alpha, ilink) {
p <- ilink(ly)
v <- p*(1-p)
n <- length(ly)
y <- rbinom(1, 1, p[1])
```

```
for (i in 2:n) {
b <- numeric(0)
for (j in 1:(i-1)){
b[j] <- alpha/(1+(i-2)*alpha)*(v[i]/v[j])^(1/2)
}
l <- p[i]+b%*%(y-p[1:(i-1)])
y <- c(y, rbinom(1, 1, l))
}
}
#test.y <- corbin.ex(c(1,2, 3), 0.3)
#print(test.y)
#simdata <- numeric(3)
#
#for (i in 1:1000) {
# y <- t(corbin.ex(c(1,1,1), 0.6))
# if (sum(is.nan(y)==0)) {
# simdata <- rbind(simdata, y)
# }
#}
```

```
#simdata <- simdata[-1,]
#print(cor(simdata))
corbin.ar1 <- function(ly, alpha, ilink) {
p <- ilink(ly)
v <- p*(1-p)
n <- length(ly)
y <- rbinom(1, 1, p[1])
for (i in 2:n) {
l <- p[i]+alpha*(y[i-1]-p[i-1])*(v[i]/v[i-1])^(1/2)
y <- c(y, rbinom(1, 1, l))
}
}
#print(corbin.ar1(c(1,2,3), 0.3))
#simdata <- numeric(3)
#
#for (i in 1:1000) {
# y <- t(corbin.ar1(c(1,1,1), 0.6))
# if (sum(is.nan(y)==0)) {
# simdata <- rbind(simdata, y)
```

```
    # }
    #}
    #simdata <- simdata[-1,]
    #print(cor(simdata))
    corbin <- function(ly, alpha, cor.type, ilink) {
    if (cor.type=="ex") corbin.ex(ly, alpha, ilink)
    else if (cor.type=="ar1") corbin.ar1(ly, alpha, ilink)
    else stop("Error: correlation type doesn't exist")
    }
    #print(corbin(c(1,2,3), 0.3, "ar1", ilink=logloginv))
    ################################
    #The following functions generate clustered binary data using qaqish
method;
    #simulation replicates
    #r
    #dimeansion
    #nXd;
    #correlation parameter
    #alpha
```

```
    #correlation structure
    #cor.type
    #linear predictor coefficeints
    #b
    sim.x <- function(rand.gen, args, powers=rep(1, length(x.type))) {
    dx <- length(rand.gen)
    x <- numeric(dx)
    for (i in 1:dx) {
    x[i] <- do.call(rand.gen[[i]], args[[i]])^powers[i]
    }
    }
    #args <- list(list(n=1, min=-1, max=1), list(n=1, min=-1, max=1))
    #print(sim.x(c(runif, runif), args))
    sim.yp <- function(d, b, rand.gen1, rand.gen2, args1, args2, powers1,
powers2) {
    x1 <- sim.x(rand.gen1, args1, powers1)
    yp <- numeric(d)
    for (i in 1:d) {
    x2 <- sim.x(rand.gen2, args2, powers2)
```

```
    x <- c(1, x1, x2)
    yp[i] <- ilogit(b%*%x)
    }
    yp
    }
    #print(sim.yp(d=2, b=c(0, 0.8, 0.8), rand.gen1=c(runif), rand.gen2=c(runif),
    # args1=list(list(n=1, min=-1, max=1)), args2=list(list(n=1, min=-1,
max=1)),
    # powers1=c(1, 1), powers2=c(1, 1)))
    simdata <- function(r, n, d, b, rand.gen, x.type, alpha, cor.type, ilink,
args, powers=rep(1, length(x.type)),
    addition="none") {
    dx <- length(b)-1
    dx0 <- length(rand.gen)
    simdata <- matrix(0, r*n*d, dx+4)
    for (i in 1:r) {
    count <- 0
    while (count<n) {
    X <- matrix(0, d, dx+1)
```

```
X[,1] <- 1
for (j in 1:dx0) {
if (x.type[j]=="C") {
X[,j+1] <- do.call(rand.gen[[j]], args[[j]])^powers[j]
}
else {
for (k in 1:d) {
X[k,j+1] <- do.call(rand.gen[[j]], args[[j]])^powers[j]
}
}
}
if (addition=="int") {
X[,4] <- X[,2]*X[,3]
}
if (addition=="qua") {
X[,4] <- X[,3]^2
}
ly <- X%*%b
X <- X[,-1]
```

```
y <- corbin(ly, alpha, cor.type, ilink)
Y <- cbind(X, y, 1:d, rep(count+1, d), rep(i, d))
if (sum(is.nan(y))==0) {
count <- count + 1
simdata[((i-1)*n*d+(count-1)*d+1):((i-1)*n*d+count*d),] <- Y
}
}
}
simdata <- data.frame(simdata)
names(simdata) <- c(paste("x", 1:dx, sep=""), "y", "t", "id", "r")
simdata
}
#args <- list(list(n=1, min=-1, max=1), list(n=1, size=1, prob=0.5))
#powers <- c(2, 1); addition <- "int";
#testsim <- simdata(2, 5, 2, c(0, 0.8, 0.8, 0.8), c(runif, rbinom), c("C",
"T"), 0.4, "ex", ilink=logloginv, args, powers, addition)
#print(testsim)
##########################################
#Function models generate simulation data based on different models;
```

```
models <- function(r, n, d, b, rand.gen, x.type, alpha, cor.type, ilink,
outpath, fn, ...) {
start.time <- proc.time()
model <- simdata(r, n, d, b, rand.gen, x.type, alpha, cor.type, ilink, ...)
end.time <- proc.time()
lapse <- end.time - start.time
print(lapse[3])
file <- paste(outpath, fn, sep="")
write.csv(model, file, row.names=F)
}
```

3.3 SAS Macro for Implementing Proposed LOF Methods for Logsitic GEE Models

```
/*****************************************************
The following macro implements the lack of fit method for logistic
GEE model using Tang and Yang (2008);
dataset: the data set contains the data;
x: the covariate variables in the model;
y: the response variable in the mode;
```

```
id: the variable indicating an individual or a cluster;
corrtype: the correlation type used in the fitted model;
class: class variables in the model (including covariates,
     subject variable and within subject variable);
time: variable indicating observations within the same cluster;
NOTES:
If the full model with different intercepts and different
slopes cause non-positive-definite Hessian, a full model
with different intercepts only or different slopes only are
fitted;
This macro uses another macro exist;
*****************************************************/
%macro lofty(dataset=, x=, y=, id=, corrtype=, class=, time=);
     %let p=0;
     %do %while(%scan(&x,&p+1)^=);
     %let p=%eval(&p+1);
     %let x&p=%scan(&x, &p);
     %end;                                  *partition of covariates;
```

```
ods listing close;
proc genmod data=&dataset desc;
      class &class;
      model &y=&x / dist=bin link=logit;
      repeated subject=&id / type=&corrtype within=&time;
      output out=linout predicted=p;
run;
ods listing;                              *fitted model;
proc means data=linout noprint;
      var p;
      output out=linout2 median=m;
run;
proc sql noprint;
      select m into :m from linout2;
quit;
data groups;
      set linout;
      if p<&m then g=1;
      else g=2;
```

```
        run;                                    *grouping data based on median
fitted probabilities;
        ods listing close;
        proc genmod data=groups desc;
            class &class g / param=effect;
            model &y=&x g
            %do j=1 %to &p;
                g*&&x&j
            %end;
            / dist=bin link=logit;
            repeated subject=&id /type=&corrtype within=&time;
            contrast 'TY' g 1,
            %do k=1 %to &p;
                g*&&x&k 1
            %end;
            / e;
            ods output contrasts=gof;
        run;
        ods listing;
```

```
*partitioned model with different intercepts and different slopes for
different groups;
%exist(gof, exist);
%if &exist=0 %then %do;
    ods listing close;
    proc genmod data=groups desc;
        class &class g / param=effect;
        model &y=&x g / dist=bin link=logit;
        repeated subject=&id /type=&corrtype within=&time;
        contrast 'TY' g 1 / e;
        ods output contrasts=gof;
    run;
    ods listing;
    *partitioned model with different intercepts for different groups;
    %if &exist=0 %then %do;
        ods listing close;
        proc genmod data=groups desc;
            class &class g / param=effect;
            model &y=&x
```

```
                %do j=1 %to &p;
                    g*&&&x&j
                %end;
                / dist=bin link=logit;
                repeated subject=&id /type=&corrtype within=&time;
                contrast 'TY'
                %do k=1 %to &p;
                    g*&&&x&k 1
                %end;
                / e;
                ods output contrasts=gof;
            run;
            ods listing;
            *partitioned model with different slopes for different groups;
        %end;
    %end;
%mend;

/************************************************
The following macro check the existence of a variable named "dsn".
```

```
If the data set exists then set a global variable "varn" to 1,
otherwise "varn" is set to be 0;
**********************************************/
%macro exist(dsn, varn);
    %global &varn;
    %if &dsn ne %then %do;
        data _null_;
            set &dsn;
        run;
    %end;
    %if &syserr=0 %then %let &varn=1;
    %else %let &varn=0;
%mend exist;
*%exist(gof3, exist3);
```

3.4 Example R Code of Using ODK Package to Find a Cost Efficient Bayesian Optimal Deisgn

```
#The following functions help to find a cost-efficient Bayesian optimal
  #design for any combinations of parameters for any model with stable
```

```
    #normal random error and quantitative predictors with respect to some
    #quantitative predictors;
    #The expression of square root transformed Gaddum/Schild model
    sqrtGaddum <- as.formula(y~sqrt(theta1+(theta2-theta1)/
        (1+(10^theta3*(1+(b/10^(-theta5))^theta6)/a)^theta4)))
    parms <- c("theta1", "theta2", "theta3", "theta4", "theta5", "theta6")
    arg1 <- c(parms, "a", "b")
    #The expression of the derivative of square root transformed Gaddum/Schild
model
    sqrtGaddum.D <- deriv(sqrtGaddum, parms, arg1)
    #The expression of pEC50
    pEC50 <- as.formula(y~theta3)
    arg2 <- parms
    #The expression of the derivative of pEC50
    pEC50.D <- deriv(pEC50, parms, arg2)
    #input for pEC50.D
    input2 <- list();
    #The expression of logIC
    logIC <- as.formula(y~log(((1/pct-1+1/pct*(10^theta3/a)^theta4)^(1/theta4)*a/10^theta3-
```

```
1)^(1/theta6)*10^(-theta5)))
    arg3 <- c(parms, "a", "pct")
    #The expression of the derivative of logIC
    logIC.D <- deriv(logIC, parms, arg3)
    #The input for logIC.D
    input3 <- list(500, 0.5)
    input <- list(input2, input3);
    funs <- c("sqrtGaddum.D", "pEC50.D", "logIC.D")
    xp <- c(2.5,1, -0.5,-2)
    #arguments for the function dilution
    d.args <- list(8, 2, TRUE, 2)
    sigma <- 0.163
    low <- c(0.00319, 21.54, 1.91, 1.48, -2.24, 0.32);
    up <- c(2.629, 218.11, 2.98, 2.73, 0.39, 1.5);
    jpdf <- function(theta, lb, ub) {
    1/prod(ub-lb)
    }
    initial <- c(2, 2, -1, -1)
    od <- OD(initial, method="L-BFGS-B", int.low=low, int.up=up, de-
```

```
sign=dilution,
    sigma=sigma, input=input, funs=funs, pr=c(1, 10), weight=0.001,
        d.args=d.args, jpdf=jpdf, lb=low, ub=up)
    print(od)
```