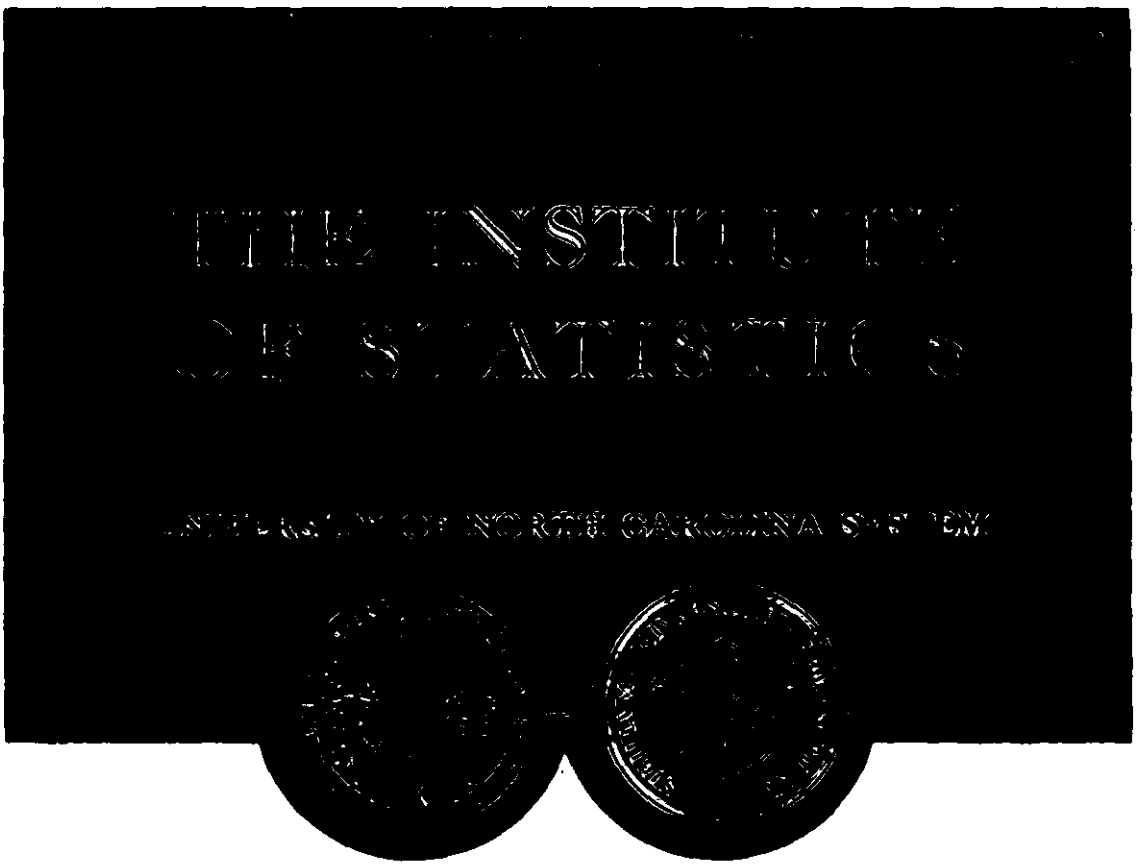


GEE1a1b Brief Manual

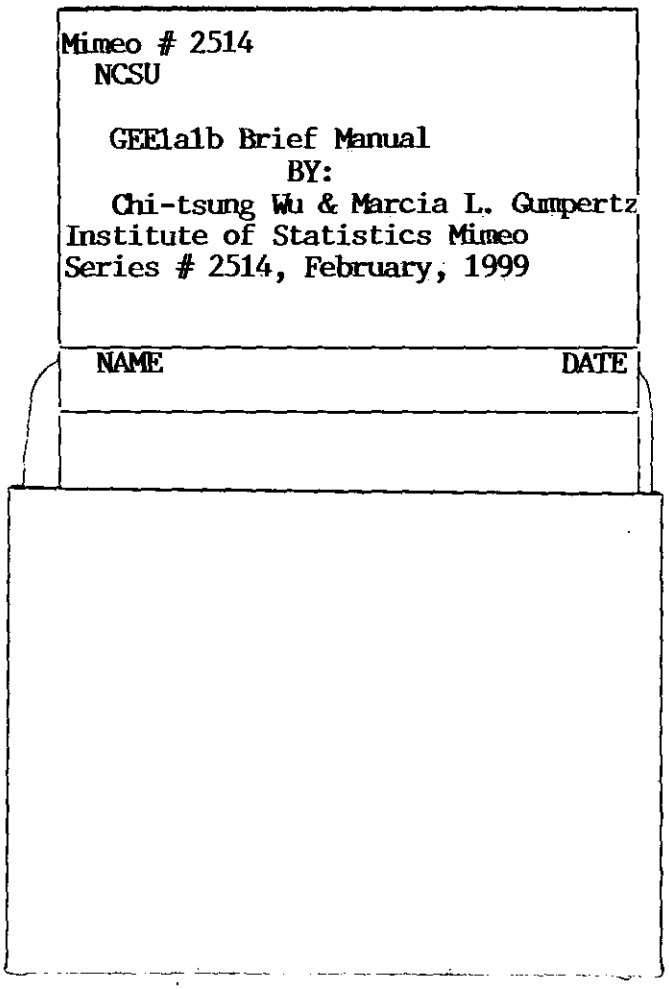
by

Chi-tsung Wu and Marcia L. Gumpertz

Institute of Statistics Mimeo Series Number 2514

February, 1999

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina


Mimeo # 2514
NCSU
GEE1a1b Brief Manual
BY:
Chi-tsung Wu & Marcia L. Gumpertz
Institute of Statistics Mimeo
Series # 2514, February, 1999
NAME
DATE

GEE1a1b Brief Manual *

Chi-tsung Wu and Marcia L. Gumpertz
Department of Statistics,
North Carolina State University,
Raleigh, NC 27695

Contents

*This work was funded by the USDA Forest Service Southern Research Station under Agreement No SRS 33-CA-97-115. Thanks to Dr. Thomas Holmes of the Southeastern Forest Experiment Station, RTP, NC for sponsoring this project.

1 Introduction

The GEE1a1b module fits marginal logistic regression models to spatially correlated binary (0 or 1) data where there are several independent realizations of a spatial process. The objective is to describe the relationship between the probability of response and some explanatory variables, while using the information about spatial autocorrelations. One example of this type of data comes from a neuroimaging study in which one CT (computer tomography) scan was taken per patient to examine the characteristics and spatial pattern of stroke-induced lesions in the brain (Albert & McShane 1995). The lesion frequency is thought to be a function of age and sex of the patient and of spatial location on the brain scan. A second example comes from a manufacturing study in which specific sites on a wafer were checked for defects. The aim in this study was to reduce defect frequency by modeling the relationship between manufacturing conditions, spatial locations on a wafer and defect occurrences (Taam 1995).

The model is

$$E(Y_{ij}) = p_{ij},$$

the probability that $Y = 1$ for subject i at location j, where

$$\log(\frac{p_{ij}}{1 - p_{ij}}) = \boldsymbol{x_{ij}}\boldsymbol{\beta}.$$

The explanatory variables may include both variables whose values change within subjects (within-subject covariates) and variables such as age whose values do not change within a subject (between-subject covariates). In this documentation we use the term "subject" to indicate an independent realization. In the neuroimaging example the subject is the patient and in the manufacturing example the subject is the wafer. The within-subject covariates may include spatial variables such as functions of location coordinates.

The observations are assumed to be spatially correlated with an exponential correlation structure,

$$\mathrm{Corr}(Y_{ij}, Y_{ik}) = e^{d_{ij}/a}.$$

In the exponential structure the parameter a is related to the range of spatial autocorrelation, which is the distance beyond which pairs of points have little correlation. The parameter a=range/3. The spatial locations (sites) must

be the same on all subjects and missing data are not allowed. The spatial correlation may be a nuisance or may be of interest itself.

The GEE1a1b modules calls a set of SAS IML modules to fit the marginal logistic regression model. Generalized estimating equations (Zeger & Liang 1986) are used to estimate both the regression coefficients and the spatial autocorrelation parameter. The user has the choice of two slightly different GEE algorithms, called GEE1a and GEE1b in this program. On average, the two methods give similar estimates of $\boldsymbol{\beta}$. If the spatial autocorrelation is large, GEE1b estimates of $\boldsymbol{\beta}$ tend to be slightly less variable than GEE1a estimates. In both cases the robust standard errors of $\hat{\boldsymbol{\beta}}$ tend to underestimate the variability of $\hat{\boldsymbol{\beta}}$. GEE1a tends to give better estimates of the autocorrelation parameter α if the correlation is small, and GEE1b gives better estimates of α if the correlation is large, however the differences between the two estimators is not large. GEE1a tends to run faster than GEE1b.

2 How to Install and Run the GEE1a1b Module

2.1 Installing the GEE1a1b module

1. Create a subdirectory to hold the GEE modules. Henceforth, we will refer to this subdirectory as "your GEE subdirectory". Store the files gee1a1b.source, and gee1a1b.example in this subdirectory. Change to this directory before going any further.

2. Invoke SAS.

3. Include the file gee1a1b.source in the SAS program editor window by typing on the command line

```
command===> include gee1a1b.source
```

4. Change the libname statement (the first line) of the gee1a1b.source program to reflect the name of your GEE subdirectory. In the file gee1a1b.source the subdirectory is called " /cwu/GEEIML" and the libname statement reads

```
libname modules '~/cwu/GEEIML';
```

Replace ' /cwu/GEEIML' with the correct pathname for your GEE subdirectory.

5. Submit the program. This compiles all of the GEE modules and stores them in a SAS catalog called gee.sct01 in your GEE subdirectory.

2.2 Running the program

1. Change directories to your GEE subdirectory.
2. Create a SAS dataset that contains variables for subject id, site id, x coordinate, y coordinate, the response variable and the explanatory variables. Here a "subject" refers to a separate realization, such as a patient, a wafer, or a year. The sites are the nodes or locations where measurements were taken on each subject. The SAME SET OF SITES must be measured on all subjects. Missing values are not allowed.
3. Sort your SAS dataset by subject and site.
4. Use the following SAS commands to load the GEE modules. In the USE command, fill in the blank (...) with your SAS dataset name.

```
proc iml symsize=250000;
   reset storage=modules.gee;
   load module=_all_;
   use ...;
```

5. Run the GEE1a1b module.

3 Syntax

The GEEa1ab module takes several arguments. The syntax is

```
run gee1a1b(subject, siteid, xcoord, ycoord, response,
            xb, xw, distmeas, alpha, beta, estmeth);
```

Example syntax:

```
run gee1a1b('kk','site','locx','locy','z','',{'locx','locy'},'euclid',
            1,{0,0,0},'gee1b');
```

3.1 Arguments to the GEE1a1b module.

The arguments must be entered in the order listed, with commas as delimiters. Most of the arguments are literal strings or vectors containing variable names. All literal strings (which includes variable names) must be entered in quotes.

subject The name of the subject variable in quotes.

siteid The name of the site variable in quotes.

xcoord The name of the variable containing X-coordinates or longitude values, in quotes.

ycoord The name of the variable containing Y-coordinates or latitude values, in quotes.

response The name of the response variable in quotes.

xb A vector of between-subject covariate names.

xw A vector of within-subject covariate names.

distmeas Specify either Euclidean distance ('euclid') or great circle distance in land miles ('gc').

alpha Give a starting value (initial guess) for α.

beta Give a vector of starting values for β.

estmethod Specify either estimation method 'gee1a' or 'gee1b'.

4 Example

Simulated data: The response variable, Z, is measured on a 10x10 grid on each of 10 realizations (subjects). The variable KK gives the realization number. The nodes of the grid, stored in the variable SITE, are labeled from 1 to 100 and the coordinates are stored in the variables LOCX and LOCY. In this example the odds that Z=1 vary smoothly in the horizontal and vertical directions, so we regress logit(Z) on the variables LOCX and LOCY. The first 20 lines of the SAS dataset "sim" are shown here.

OBS	KK	LOCX	LOCY	Z	SITE
1	1	1	1	1	1
2	1	1	2	0	2
3	1	1	3	1	3
4	1	1	4	1	4
5	1	1	5	0	5
6	1	1	6	0	6
7	1	1	7	0	7
8	1	1	8	1	8
9	1	1	9	1	9
10	1	1	10	1	10
11	1	2	1	0	11
12	1	2	2	0	12
13	1	2	3	0	13
14	1	2	4	1	14
15	1	2	5	0	15
16	1	2	6	1	16
17	1	2	7	1	17
18	1	2	8	1	18
19	1	2	9	1	19
20	1	2	10	0	20

4.1 Initial values

You must provide initial values for β and α. In this example, the initial values for β are obtained by running PROC LOGISTIC. The initial value for

α is obtained by plotting the semivariogram of the Pearson residuals from PROC LOGISTIC. On this graph the range of spatial correlation appears to be about 6, which corresponds to a range parameter of $\alpha = 2$ (α=range/3 in the exponential model). SAS code such as the following can be used to obtain separate semivariograms for each realization, compute the average semivariogram value for each lag distance, and plot the average semivariogram against lag distance.

```
proc logistic data=sim descending;
   model z=locx locy;
   output out=outl reschi=pearson;
proc sort data=outl;
   by kk;
proc variogram data=outl outvar=outvario;
   by kk;
   var pearson;
   coordinates xcoord=locx ycoord=locy;
   compute lagdistance=1 maxlags=10;
proc means data=outvario noprint;
   class distance;
   var variog;
   output out=vario mean=variog;
proc gplot data=vario;
   plot variog*distance;
   run;
```

4.2 SAS code for GEE1b

The SAS code to fit the complete model using the GEE1b method follows. Note that there are no between-subject covariates and the within-subject covariates in this example are just the same as the spatial coordinates.

```
libname modules '~/cwu/GEEIML';

proc sort data=sim;
 by kk site;
proc iml symsize=250000;
   reset storage=modules.gee;
```

```
load module=_all_;
use sim;
run gee1a1b('kk','site','locx','locy','z','' ,{'locx','locy'},'euclid',
          2,{.006,-.02,-.03},'gee1b');
```

4.3 Output

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                                      13:49 Wednesday, July 29, 1998   1

               Marginal Logistic Regression Model Information


                                                K
                   number of subjects          10


                                                N
                    number of sites           100

      number (including intercept) and names of regressor variables


                           P    XB     XW
                           3           locx
                                       locy

                  distance measure and estimation method


                       DISTMEAS    ESTMETH
                       euclid      gee1b

    Minimum, 25th Percentile, Median, 75th Percentile, Maximum Distance

                  1 3.1622777 5.0990195 7.0710678 12.727922
```

initial values for beta and alpha

BETA	ALPHA
0.006	2
-0.02	
-0.03	

Iteration 0. Ordinary logistic regression

Ordinary logistic regression estimates

and robust and model-based standard errors

BETA	ROBUST	MODEL
0.006003	0.2524636	0.1839879
-0.02228	0.0317494	0.0222646
-0.028709	0.0206968	0.0222707

Iteration history of score function values

IT	UBETA			UALPHA
1	-0.000192	-0.001078	-0.000595	-35.46425
2	-0.000042	-0.000287	-0.000183	-3.172115
3	-9.607E-7	-9.005E-6	-6.596E-6	-0.20973
4	-3.097E-8	-4.127E-7	-3.336E-7	-0.0123
5	-1.699E-9	-2.387E-8	-1.951E-8	-0.000729

Iteration history of parameter estimates

IT	BETAFLO			ALPHAFLO
1	0.0150214	-0.022359	-0.02948	0.3951705
2	0.0113684	-0.02233	-0.029167	0.3343285
3	0.0109257	-0.022326	-0.029129	0.3259752

```
        4    0.0108936 -0.022326 -0.029126    0.3253563
        5    0.0108916 -0.022326 -0.029126    0.3253188

Final parameter estimates, robust, and model-based standard errors


     BETA       ROBUST        MODEL
0.0108916    0.2537026     0.200812
-0.022326    0.0326801    0.0242613
-0.029126    0.0208013    0.0242685

Robust and model-based estimates of Var(betahat)


   ROBUST                            MODEL
 0.064365 -0.005903 -0.003774    0.0403254 -0.003231 -0.003228
-0.005903  0.001068 0.0000498    -0.003231 0.0005886  1.575E-6
-0.003774 0.0000498 0.0004327    -0.003228  1.575E-6  0.000589


                                  ALPHA
Final estimate of alpha    0.3253188
```

5 Statistical Method

The GEE1a1b module estimates the parameters of the mean and correlation models for the marginal logistic regression model with spatially correlated observations. It uses two variants of the GEE1 method proposed by Prentice (1988) for correlated binary data in longitudinal studies. The generalized

estimating equations proposed by Prentice (1988) are

$$\begin{aligned} \boldsymbol{U}_\beta &= \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^t \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \\ \boldsymbol{U}_{1a} &= \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{v}_i}{\partial \boldsymbol{\alpha}}\right)^t \boldsymbol{W}_i^{-1}(\boldsymbol{z}_i - \boldsymbol{v}_i) = \mathbf{0}, \end{aligned}$$

where $\boldsymbol{V}_i = \mathrm{Var}(\boldsymbol{Y}_i)$, $\boldsymbol{W}_i$ is a working covariance matrix of $\boldsymbol{z}_i$, and $\boldsymbol{z}_i$ is defined by

$$\boldsymbol{z}_i = \begin{pmatrix} (Y_{i1} - \mu_{i1})(Y_{i2} - \mu_{i2}) \\ (Y_{i1} - \mu_{i1})(Y_{i3} - \mu_{i3}) \\ \vdots \\ (Y_{i,n_i-1} - \mu_{i,n_i-1})(Y_{i,n_i} - \mu_{i,n_i}) \end{pmatrix}.$$

The parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated by iteratively solving the estimating equations $\boldsymbol{U}_\beta$ and $\boldsymbol{U}_{1a}$. We call this method GEE1a.

The second method, called GEE1b, is based on modeling the spatial association using a semivariogram rather than a covariogram function. The variogram is a measure of spatial dependence, defined as follows. Let

$$\mathrm{Var}(Y_i(\boldsymbol{s}_j) - Y_i(\boldsymbol{s}_k)) = 2\gamma(\boldsymbol{s}_j - \boldsymbol{s}_k), \text{for all } \boldsymbol{s}_j, \boldsymbol{s}_k \in \mathcal{R}^d.$$

The quantity $\gamma(\cdot)$ is called a semivariogram.

The semivariogram is based on cross-products of differences. We define a new vector called $\boldsymbol{z}^\star$ corresponding to $\boldsymbol{z}$ in $\boldsymbol{U}_{1a}$ for subject i as

$$\boldsymbol{z}_i^\star = \begin{pmatrix} \frac{[(Y_{i1}-\mu_{i1})-(Y_{i2}-\mu_{i2})]^2}{2} \\ \frac{[(Y_{i1}-\mu_{i1})-(Y_{i3}-\mu_{i3})]^2}{2} \\ \vdots \\ \frac{[(Y_{i,n_i-1}-\mu_{i,n_i-1})-(Y_{i,n_i}-\mu_{i,n_i})]^2}{2} \end{pmatrix} = \begin{pmatrix} z^\star_{i12} \\ z^\star_{i13} \\ \vdots \\ z^\star_{i,n_i-1,n_i} \end{pmatrix}.$$

Estimating the variogram can have some advantages over the covariance function: (1) estimation of the variogram is more stable than estimation of the covariogram in the presence of trend contamination; and (2) taking differences tends to make the $\boldsymbol{z}^\star$s less highly correlated with each other than the $\boldsymbol{z}$ values, potentially improving estimation of α.

Since for each pair of different locations $\boldsymbol{s}_j$ and $\boldsymbol{s}_k$

$$\mathrm{E}[(Y_{ij} - \mu_{ij}) - (Y_{ik} - \mu_{ik})]^2 = \mathrm{Var}(Y_{ij} - Y_{ik}) = 2\gamma_{ijk},$$

we define an objective function for estimating $\boldsymbol{\alpha}$ as follows

$$\boldsymbol{U}_{1b} = \sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\gamma}_i}{\partial \boldsymbol{\alpha}}\right)^t \boldsymbol{W}_i^{\star -1} (\boldsymbol{z}_i^\star - \boldsymbol{\gamma}_i) = \mathbf{0}$$

where $\boldsymbol{W}_i^\star$ is a working variance matrix for $\boldsymbol{z}_i^\star$ and $\boldsymbol{\gamma}_i = \mathrm{E}(\boldsymbol{z}_i^\star)$. One possible type of working variance matrix is given by

$$\boldsymbol{W}_i^\star = \begin{pmatrix} \mathrm{Var}(z_{i12}^\star) & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \mathrm{Var}(z_{i,n_i-1,n_i}^\star) \end{pmatrix}.$$

The GEE1b method estimates $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ by iteratively solving the equations $\boldsymbol{U}_\beta = \mathbf{0}$ and $\boldsymbol{U}_{1b} = \mathbf{0}$.

In general, the GEE equations do not have a closed form solution, so an iterative algorithm is followed. The updating equations for GEE1a are shown here; the equations for GEE1b are similar. For GEE1a $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are updated at the $s+1^{st}$ iteration as

$$\boldsymbol{\beta}_{s+1} = \boldsymbol{\beta}_s + \left(\sum_{i=1}^{K} \boldsymbol{D_i}^t \boldsymbol{V_k}^{-1} \boldsymbol{D_i}\right)^{-1} \sum_{i=1}^{K} \boldsymbol{D_i}^t \boldsymbol{V_i}^{-1} (\boldsymbol{Y_i} - \boldsymbol{p_i}),$$

and

$$\boldsymbol{\alpha}_{s+1} = \boldsymbol{\alpha}_s + \left(\sum_{i=1}^{K} \boldsymbol{E_i}^t \boldsymbol{W_k}^{-1} \boldsymbol{E_i}\right)^{-1} \sum_{i=1}^{K} \boldsymbol{E_i}^t \boldsymbol{W_i}^{-1} (\boldsymbol{z_i} - \boldsymbol{v_i}),$$

where $\boldsymbol{D_i} = \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}}$ and $\boldsymbol{E_i} = \frac{\partial \boldsymbol{v_i}}{\partial \boldsymbol{\alpha}}$. The model-based and robust estimates of the variances of the parameter estimates are
Model-based

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \left(\sum_{i=1}^{K} \widehat{\boldsymbol{D}_i}^t \widehat{\boldsymbol{V}_i}^{-1} \widehat{\boldsymbol{D}_i}\right)^{-1}$$

Robust

$$\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}}) = \left(\textstyle\sum_{i=1}^{K} \widehat{\boldsymbol{D}_i}^t \widehat{\boldsymbol{V}_i}^{-1} \widehat{\boldsymbol{D}_i}\right)^{-1} \left(\textstyle\sum_{i=1}^{K} \widehat{\boldsymbol{D}_i}^t \widehat{\boldsymbol{V}_i}^{-1} (\boldsymbol{Y_i} - \widehat{\boldsymbol{p}_i})(\boldsymbol{Y_i} - \widehat{\boldsymbol{p}_i})^t \widehat{\boldsymbol{V}_i}^{-1} \widehat{\boldsymbol{D}_i}\right) \left(\textstyle\sum_{i=1}^{K} \widehat{\boldsymbol{D}_i}^t \widehat{\boldsymbol{V}_i}^{-1} \widehat{\boldsymbol{D}_i}\right)^-$$

6 References

Albert, P. S., and McShane, L. M. (1995), "A Generalized Estimating Equations Approach for Spatially Correlated Binary Data:Applications to the Analysis of Neuroimaging Data," *Biometrics,* 51, 627-638.

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

Taam, W. (1997), "A quasi-likelihood approach to model Bernoulli data with spatial dependence," *Communications in Statistics - Simulation and Computation,*26:591-603.

Zeger, S. L., and Liang, K.-Y. (1986), "Longitudinal Data Analysis for Discrete and Continuous Outcomes," *Biometrics,* 42, 121-130.

A Modules contained in the file gee1a1b.source

Table 1 lists the SAS IML modules called by GEE1a1b, along with the function of each module.

Name	Brief Description
geelind	Estimate $\boldsymbol{\beta}$ using ordinary logistic regression.
seBetaId	Compute robust and model-based $s(\widehat{\boldsymbol{\beta}})$, ordinary logistic regression
gee1	Update $\widehat{\boldsymbol{\beta}}$ solving $\boldsymbol{U}_{\boldsymbol{\beta}}$
gee2	Update $\hat{\alpha}$ solving $\boldsymbol{U}_{1a}$
gee4	Update $\hat{\alpha}$ solving $\boldsymbol{U}_{1b}$
seBeta	Estimate variance for $\widehat{\boldsymbol{\beta}}$
U1	Calculate the value of $\boldsymbol{U}_{\beta}$
U2	Calculate the value of $\boldsymbol{U}_{1a}$
U4	Calculate the value of $\boldsymbol{U}_{1b}$
infoM1	Calculate $\sum \boldsymbol{D}^t \boldsymbol{V}^{-1} \boldsymbol{D}$
infoM2	Calculate $\sum \boldsymbol{E}^t \boldsymbol{W}^{-1} \boldsymbol{E}$
infoM4	Calculate $\sum \boldsymbol{E}^{\star^t} \boldsymbol{W}^{\star^{-1}} \boldsymbol{E}^{\star}$
vop	gives locations of lower triangular elements of a square matrix
diagvec	gives locations of diagonal elements of a square matrix
Hmat	Compute Euclidean distances among sites
gcdist	Compute great circle distance among sites
gee1a	Calls gee1 and gee2 modules iteratively to convergence
gee1b	Calls gee1 and gee4 modules iteratively to convergence
gee1a1b	Calls geelind, then gee1a or gee1b, outputs results

Table 1: Brief description of all modules in the file modules.source