Categorical Data Analysis

PENALIZED GEE (PGEE)

with small or sparse longitudinal binary data

김 민 수　김 종 영　김 학 빈　진 영 화

2019 12 23

Contents

1 When PGEE is needed

2 PGEE

3 Simulation Study

4 Empirical Result1 - confidence interval

5 Empirical Result2 - real data analysis

When PGEE is needed

correlated binary data

- **Repeated observations** on a subject are typically correlated

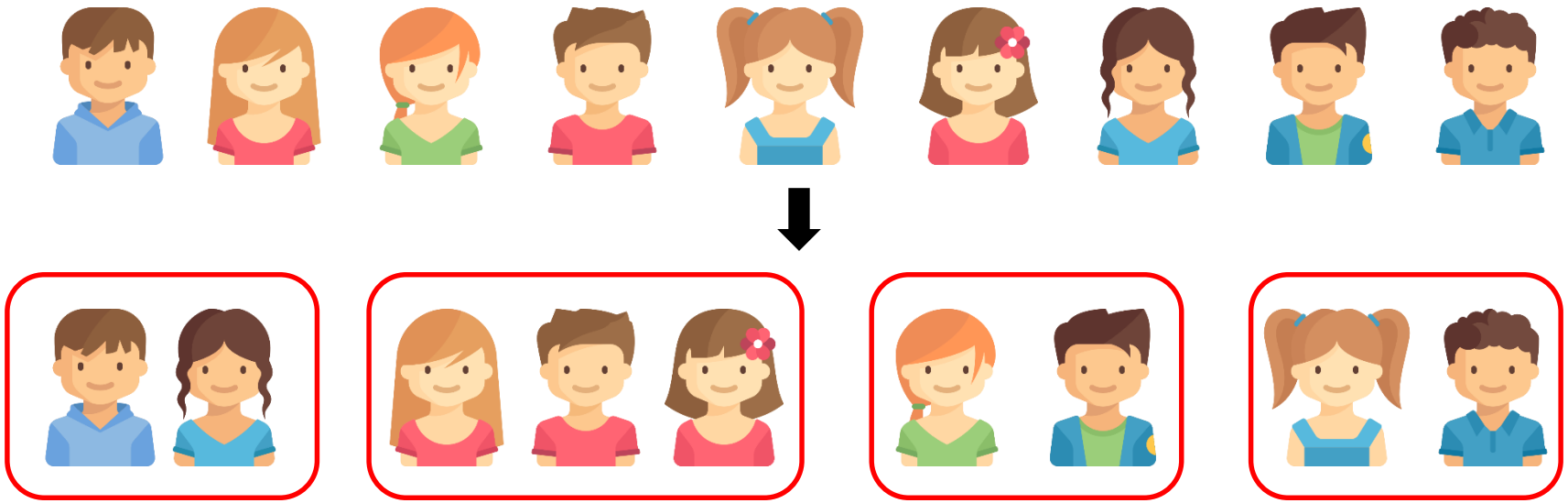
Diagnosis Severity	Treatment	Response at Three Times							
		NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
Mild	Standard	16	13	9	3	14	4	15	6
Mild	New drug	31	0	6	0	22	2	9	0
Severe	Standard	2	2	8	9	9	15	27	28
Severe	New drug	7	2	5	2	31	5	32	6

(N = Normal, A = Abnormal)

id	Diagnosis Severity	Treatment	Response
1	Mild	Standard	N
1	Mild	Standard	N
1	Mild	Standard	N
2	Mild	Standard	N
2	Mild	Standard	N
2	Mild	Standard	N
⋮	Mild	Standard	N
16	Mild	Standard	N
16	Mild	Standard	N
16	Mild	Standard	N

correlated binary data

- Response variable is observed for **matched sets** of subjects



Analyses that ignore the correlation can estimate model parameters well, but the **standard error estimators can be badly biased**

Generalized estimating equation (GEE)

the GEE method is a multivariate type of quasi-likelihood method

Maximum likelihood method

- assume a particular type of probability distribution for **Y**

Quasi-likelihood method

- assumes only a relationship between **μ** and **Var(Y)**
- usual variance formula but multiplies it by a **constant** that is itself estimated using the data

indepentent	$\mathrm{Var}(Y) = n\pi(1-\pi)$
	φ
Quasi-likelihood	$\varphi n\pi(1-\pi)$

Generalized estimating equation (GEE)

- multivariate generalizations of the equations

Rather than assuming a particular type of distribution for $(Y_1,...,Y_T)$, this method only links each **marginal mean to a linear predictor** and provides a guess for the **variance–covariance structure** of $(Y_1,...,Y_T)$

valid standard errors result from an adjustment the GEE method makes using the empirical dependence the actual data exhibit

Problems of separation in GEE

small correlated binary data / even in large data with rare outcome and/or high intra-cluster correlation

I. Complete Separation

		Y	
		1	0
X	exposed	20	0
	unexposed	0	20

II. Quasi-Complete Separation

		Y	
		1	0
X	exposed	20	0
	unexposed	8	12

III. Near-to-Quasi-Complete separation

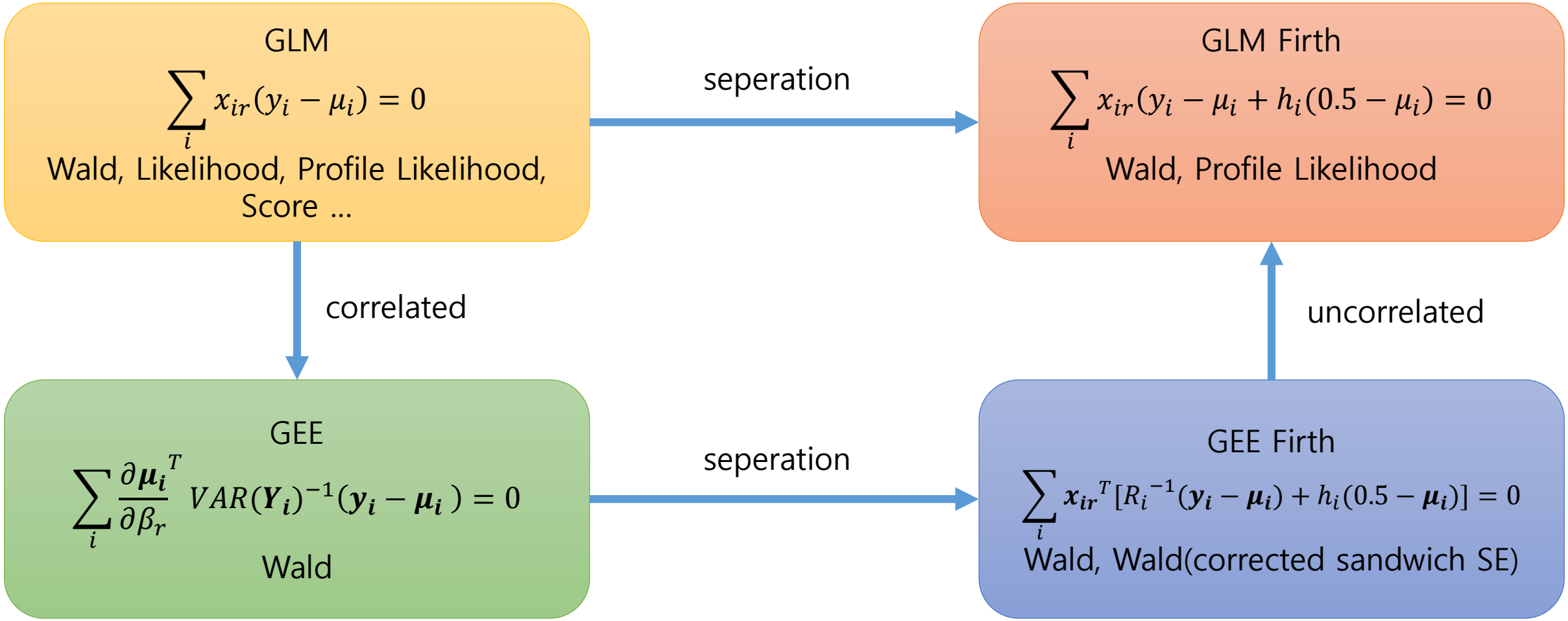
		Y	
		1	0
X	exposed	18	2
	unexposed	8	12

cell count to less than 15% of total observations (N × n)

frequent **convergence failure**, and even if it converges, it provides biased or often **infinite estimate** of at least one regression coefficients as well as its standard error, which are unrealistic and not interpretable and provide misleading inference

PGEE

OUTLINE


GLM
$\sum_i x_{ir}(y_i - \mu_i) = 0$
Wald, Likelihood, Profile Likelihood, Score …
seperation
GLM Firth
$\sum_i x_{ir}(y_i - \mu_i + h_i(0.5 - \mu_i) = 0$
Wald, Profile Likelihood
correlated
uncorrelated
GEE
$\sum_i \frac{\partial \boldsymbol{\mu_i}^T}{\partial \beta_r} VAR(\boldsymbol{Y_i})^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) = 0$
Wald
seperation
GEE Firth
$\sum_i \boldsymbol{x_{ir}}^T [R_i^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i}) + h_i(0.5 - \boldsymbol{\mu_i})] = 0$
Wald, Wald(corrected sandwich SE)

Firth's modified score equation

- Firth's modified scored equation for removing the first order (O(1/n)) bias in the MLE of the rth regression parameter of GLM takes the following form :

$$U_r{}^*(\boldsymbol{\beta}) = U_r(\boldsymbol{\beta}) + a_r(\boldsymbol{\beta}) = 0 \quad \text{for r=1,....,p}$$

where $\boldsymbol{\beta}=(\beta_1, \ldots., \beta_p)^T$, $U_r(\boldsymbol{\beta}) = \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_r}$, $\mathrm{I}(\boldsymbol{\beta}) = \frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}$, $a_r(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_r}(0.5\log|\mathrm{I}(\boldsymbol{\beta})|)$.

- At this time, the modified log likelihood is

$$l^*(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + 0.5\log(|\mathrm{I}(\boldsymbol{\beta})|)$$

and the modified likelihood is

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta})\ |\mathrm{I}(\boldsymbol{\beta})|^{0.5}$$

Firth's modified score equation for logistic regression

-logistic regression

$$\Pr(y_i = 1 \,|\, \boldsymbol{x}_i, \boldsymbol{\beta}) = \mu_i = \left[1 + \exp\left(-\sum_{r=1}^{p} \boldsymbol{x}_{ir}\beta_r\right)\right]^{-1}, \quad (i = 1, \ldots, n).$$

-The observed information matrix : I($\boldsymbol{\beta}$) = $X^T WX$, $W_{ii} = \mu_i(1\text{-}\mu_i)$, X is design matrix.

-Then, the score equation for the rth coefficient is

$$U_r(\boldsymbol{\beta}) = \sum_i x_{ir}(y_i - \mu_i) = 0 \qquad \text{for r=1,....,p}$$

Firth's modified score equation for logistic regression

-and $a_r(\boldsymbol{\beta}) = = \frac{\partial}{\partial \beta_r}$ (0.5log| I($\boldsymbol{\beta}$) |) = $\sum_i [h_i(0.5-\mu_i)\ ]x_{ir}$, $h_i = H_{ii}$, $H = W^{0.5}X(X^TWX)^{-1}X\ W^{0.5}$

=> This use the formula $\frac{\partial}{\partial \beta_r}$ (0.5log| I($\boldsymbol{\beta}$) |) = 0.5*trace(I($\boldsymbol{\beta}$)$^{-1}$ $\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_r}$)

-Then, Firth's modified scored equation for logistic regression is

$$U_r^*(\boldsymbol{\beta}) = \sum_i x_{ir}(y_i - \mu_i + h_i(0.5 - \mu_i) = 0 \qquad \text{for r=1,....,p}$$

-Then, the solution of $U_r^*(\boldsymbol{\beta})$ =0 is FL estimate. And

$$U_r^*(\beta) = \sum_{i=1}^{n} \left[\{(y_i - \mu_i)(1 + h_i/2) + (1 - y_i - \mu_i)h_i/2\}x_{ir}\right]$$

$$= \sum_{i=1}^{n} \left[y_i - \mu_i + h_i\,(1/2 - \mu_i)\right] x_{ir} = 0, \quad (r = 1, \ldots, p),$$

Application of Firth's method for solving separation in GEE

-GEE(Generalized Estimating Equation)
Let $\boldsymbol{y_i} = (y_{i1}, \ldots, y_{ini})^T$ (i=1,...,N) and n_i * p covariance matrix X_i.
Then the marginal expectation $E(\boldsymbol{Y_i} \mid X_i) = \boldsymbol{\mu_i} = (\mu_{i1}, \ldots, \mu_{i1})^T$ is modeled by $g(\boldsymbol{\mu_i}) = X_i^T \boldsymbol{\beta}$.

And let $V_i = W_i^{0.5} R_i(\boldsymbol{\alpha}) W_i^{0.5}$ (This is working covariance matrix.)
where W_i=diag($v(\mu_{ij})$) is a known variance function and $R_i(\boldsymbol{\alpha})$ is the corresponding working correlation matrix, which may depend on some parameters $\boldsymbol{\alpha}$ and is generally unknown.

Under the assumption of $R_i(\boldsymbol{\alpha})$, regression coeffficient $\boldsymbol{\beta}$ can be estimated by solving the following equations (called GEE) :

$$U(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} D_i^T V_i^{-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i) = \mathbf{0} \qquad \text{where } D_i = \partial \mu_i / \partial \boldsymbol{\beta}.$$

=>Under the assumption of independent working correlation, GEE is equivalent to GLM score equation.

Application of Firth's method for solving separation in GEE

-For the rth regression coefficient, the PGEE(Penalized Generalized Estimating Equation) is given by

$$U_r^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = U_r(\boldsymbol{\beta}, \boldsymbol{\alpha}) + A_r(\boldsymbol{\beta}, \boldsymbol{\alpha}), \quad (r = 1, \ldots, p), \tag{4}$$

with the penalty $A_r(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{1}{2}\text{trace}\left[I(\boldsymbol{\beta}, \boldsymbol{\alpha})^{-1}\{\partial I(\boldsymbol{\beta}, \boldsymbol{\alpha})/\partial\beta_r\}\right].$

- $U_r(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_i \frac{\partial \boldsymbol{\mu_i}^T}{\partial \beta_r} VAR(\boldsymbol{Y_i})^{-1}(\boldsymbol{y_i} - \boldsymbol{\mu_i})$ (3)

- And $A_r(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is….

Application of Firth's method for solving separation in GEE

$$\frac{\partial U(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N}\left[D_i^T V_i^{-1}\left(-\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right) + (\boldsymbol{y}_i - \boldsymbol{\mu}_i)\left\{D_i^T \frac{\partial V_i^{-1}}{\partial \boldsymbol{\beta}} + V_i^{-1}\frac{\partial D_i^T}{\partial \boldsymbol{\beta}}\right\}\right]. \quad (5)$$

For the logit link $g(\boldsymbol{\mu_i}) = \log(\boldsymbol{\mu_i}/(1- \boldsymbol{\mu_i})) = {X_i}^T \boldsymbol{\beta}$, the derivative $D_i = \frac{\partial \boldsymbol{\mu_i}}{\partial \boldsymbol{\beta}} = W_i\ X_i$.
(W_i=diag($v(\mu_{ij})$) , X_i is ith design matrix)
Using this, the expression for information matrix can be derived from Equation (5) as

$$I(\boldsymbol{\beta},\boldsymbol{\alpha}) = -E\left[\frac{\partial U(\boldsymbol{\beta},\boldsymbol{\alpha})}{\partial \boldsymbol{\beta}}\right] = \sum_{i=1}^{N} D_i^T V_i^{-1} D_i = \sum_{i=1}^{N} (W_i X_i)^T V_i^{-1} W_i X_i, \quad (6)$$

Application of Firth's method for solving separation in GEE

Now, with $V_i = W_i^{1/2} R_i(\boldsymbol{\alpha}) W_i^{1/2}$ and changing the notation $R_i(\boldsymbol{\alpha})$ to R_i, Equation (6) is reduced to

$$I(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} X_i^T W_i^{1/2} R_i^{-1} W_i^{1/2} X_i. \tag{7}$$

Using the chain rule and the product rule in matrix calculus, the derivative $\partial I(\boldsymbol{\beta}, \boldsymbol{\alpha}) / \partial \beta_r$ for the rth coefficient can be expressed as

$$\frac{\partial I(\boldsymbol{\beta}, \boldsymbol{\alpha})}{\partial \beta_r} = 2 \sum_{i=1}^{N} X_i^T W_i^{1/2} R_i^{-1} W_i^{1/2} X_i Q_i, \tag{8}$$

where, $Q_i = \text{diag}\{x_{ir1}(1/2 - \mu_{i1}), \ldots, x_{irn}(1/2 - \mu_{in})\}$.

Application of Firth's method for solving separation in GEE

Finally, putting Equations (3), (7), and (8) into Equation (4), the modified GEE (PGEE) for rth coefficient can be expressed as

$$U_r^*(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{N} \left[R_i^{-1}(\boldsymbol{y}_i - \boldsymbol{\mu}_i) + \boldsymbol{h}_i \left(\mathbf{1/2} - \boldsymbol{\mu}_i\right)\right]\boldsymbol{x}_{ir} = 0, \quad (r = 1, \ldots, p), \tag{9}$$

=>Under the assumption of independent working correlation, PGEE is equivalent to FL score equation.

$$U_r{}^*(\boldsymbol{\beta}) = \sum_i (y_i - \mu_i + h_i(0.5 - \mu_i)x_{ir} = 0 \qquad \text{for r=1,....,p}$$

PGEE – ESTIMATE OF $\boldsymbol{\beta}$

The PGEE estimates of $\boldsymbol{\beta}$ can be estimated using the following standard iterative process by Liang and Zeger except for the first step of taking the initial value of $\boldsymbol{\beta}$.

[Algorithm]

1. Choose an initial estimate $\boldsymbol{\beta}^{(0)}$ of $\boldsymbol{\beta}$ as the FL estimate obtained using Firth's procedure for standard logistic regression considering independent working correlation.

2. Given $\boldsymbol{\beta}^*$ ($\boldsymbol{\beta}^* = \boldsymbol{\beta}^{(0)}$ at the first iteration), calculate moment estimate $\boldsymbol{\alpha}^*$ of $\boldsymbol{\alpha}$ of the working correlation matrix $R(\boldsymbol{\alpha})$.

For example, for exchangeable working correlation

$$\alpha^* = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i(n_i-1)}\sum_{j\neq k}^{n_i} e_{ij}^* e_{ik}^*, \text{ where } e_{ij}^* = \frac{y_{ij}-\mu_{ij}^*}{\sqrt{v(\mu_{ij}^*)}}.$$

why?

PGEE – ESTIMATE OF $\boldsymbol{\beta}$

Similarly, For AR(1),

$$\alpha^* = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{n_i - 1}\sum_{j\le n_i-1}^{n_i} e_{ij}^* e_{i,j+1}^*.$$

3. Given the working correlation matrix $R(\boldsymbol{\alpha}^*)$ obtained in step 2, the current estimate $\boldsymbol{\beta}^{*t}$ is updated according to the PGEE using Newton-Raphson method as

$$\boldsymbol{\beta}^{*(t+1)} = \boldsymbol{\beta}^{*t} + [I(\boldsymbol{\beta}, \boldsymbol{\alpha})]^{-1}\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{*t}} U^*(\boldsymbol{\beta}, \boldsymbol{\alpha})\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}^{*t}}.$$

4. Iterate steps 2 and 3 until a desired convergence achieved.

PGEE – ESTIMATE OF Var-Cov($\boldsymbol{\beta}^*$)

-The Var-Cov($\boldsymbol{\beta}^*$) can be consistently estimated by substituting the PGEE estimtates ($\boldsymbol{\beta}^*$) into the so-called sandwich estimator of Liang and Zeger.

-Therefore, the sandwich covariance estimator for the PGEE estimator can be expressed as

$$\hat{V}(\boldsymbol{\beta}^*) = \{I_0(\boldsymbol{\beta}^*)\}^{-1} I_1(\boldsymbol{\beta}^*) \{I_0(\boldsymbol{\beta}^*)\}^{-1},$$

where

$$I_0(\boldsymbol{\beta}^*) = \sum_{i=1}^{N} D_i^{*T} V_i^{*-1} D_i^*,$$

$$I_1(\boldsymbol{\beta}^*) = \frac{n^* - 1}{n^* - p} \frac{N}{N-1} \sum_{i=1}^{N} (\boldsymbol{d}_i - \bar{\boldsymbol{d}})(\boldsymbol{d}_i - \bar{\boldsymbol{d}})^T,$$

$$\boldsymbol{d}_i = D_i^{*T} V_i^{*-1} (\boldsymbol{y}_i - \boldsymbol{\mu}_i^*),$$

$$\bar{\boldsymbol{d}} = N^{-1} \sum_{i=1}^{N} \boldsymbol{d}_i, \text{ and } n^* = \sum_{i=1}^{N} n_i.$$

PGEE – ESTIMATE OF Var-Cov($\boldsymbol{\beta}^*$)

-As the problem of separation is likely to occur in small sample case for which we proposed penalized estimator of $\boldsymbol{\beta}$, the sandwich estimator $\hat{V}(\boldsymbol{\beta}^*)$ may possess some bias in such situation, like those for the standard GEE estimator.

-Several bias corrections to Liang and Zeger's sandwich estimator have been proposed.

-The proposal by Morel et al has some additional benefits that the correction not only reduces type 1 error rate but also guarantees that the estimated variance covariance matrix is positive definite. And the adjustment can be applied to the whole "sandwich" rather than to the individual residuals of $I_1(\boldsymbol{\beta}^*)$.

-In addition, the correction term vanishes as N increases, which is similar in nature to the Firth's penalty we used for estimating $\boldsymbol{\beta}$ coefficient.

PGEE – ESTIMATE OF Var-Cov($\boldsymbol{\beta}^*$)

-Using the similar notations of Morel et al, the bias corrected estimator can be defined as

$$\tilde{V}(\boldsymbol{\beta}^*) = \hat{V}(\boldsymbol{\beta}^*) + \hat{\delta}\hat{\kappa}\{I_0(\boldsymbol{\beta}^*)\}^{-1},$$

where

$$\hat{V}(\boldsymbol{\beta}^*) = \{I_0(\boldsymbol{\beta}^*)\}^{-1} I_1(\boldsymbol{\beta}^*)\{I_0(\boldsymbol{\beta}^*)\}^{-1},$$

$$\hat{\kappa} = \max\left[1, \operatorname{trac}\left\{I_0(\boldsymbol{\beta}^*)^{-1} I_1(\boldsymbol{\beta}^*)\right\}/p\right],$$

$$\hat{\delta} = \min\{0.5, p/(n-p)\}.$$

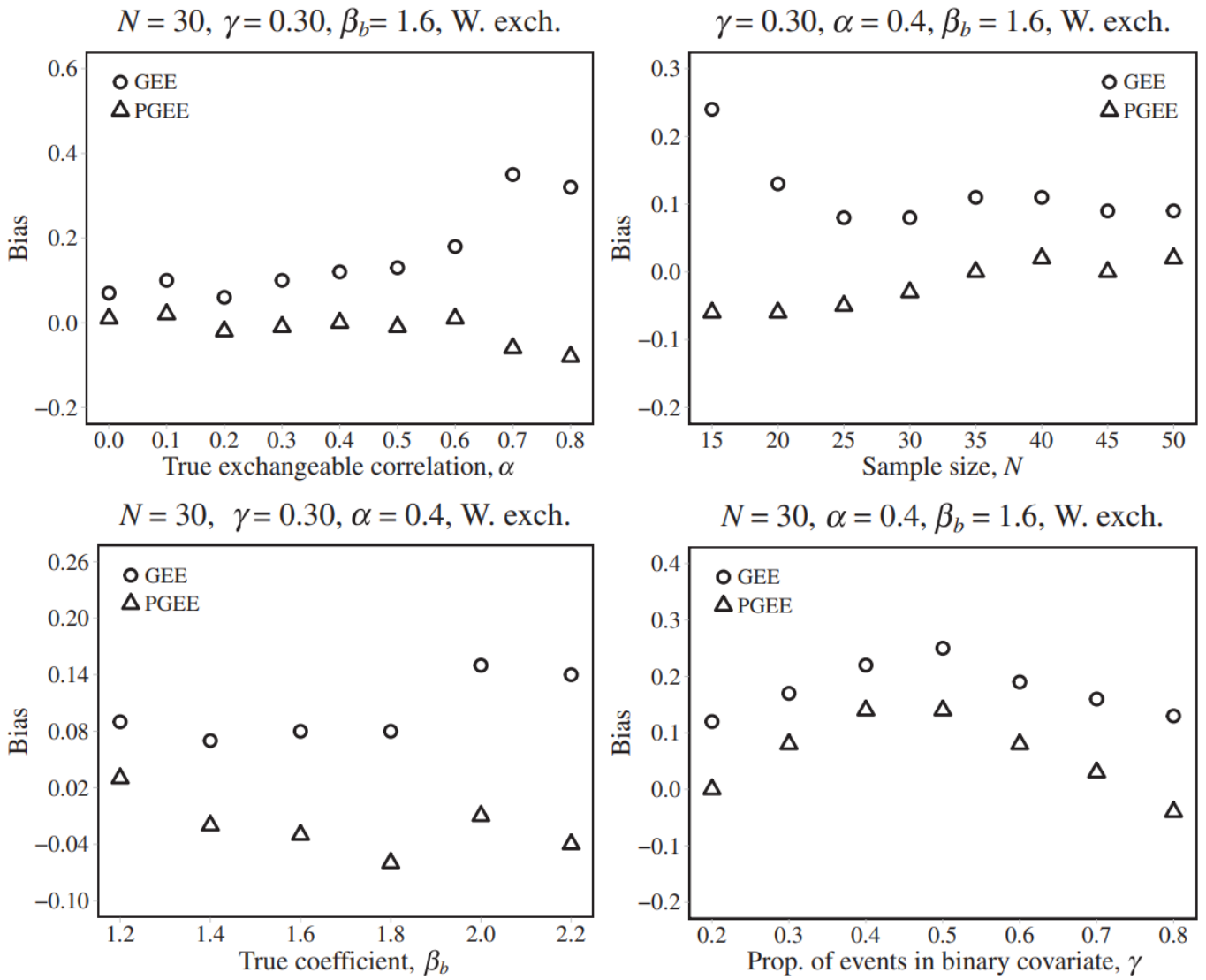
Simulation Study

$$\mu_{ij} = \Pr\left(y_{ij=1} \middle| x_i\right) = \left[1 + \exp\left\{-\left(\beta_0 + \beta_b x_{ij1} + \beta_c x_{ij2}\right)\right\}\right]^{-1}$$

- β_0 = coefficient for intercept
- β_b = coefficient for binary covariate
- β_c = coefficient for continuous covariate
- N = num of cluster
- n = num of observation in each cluster
- α = intra-cluster correlation
- γ = proportion of events in binary covariate

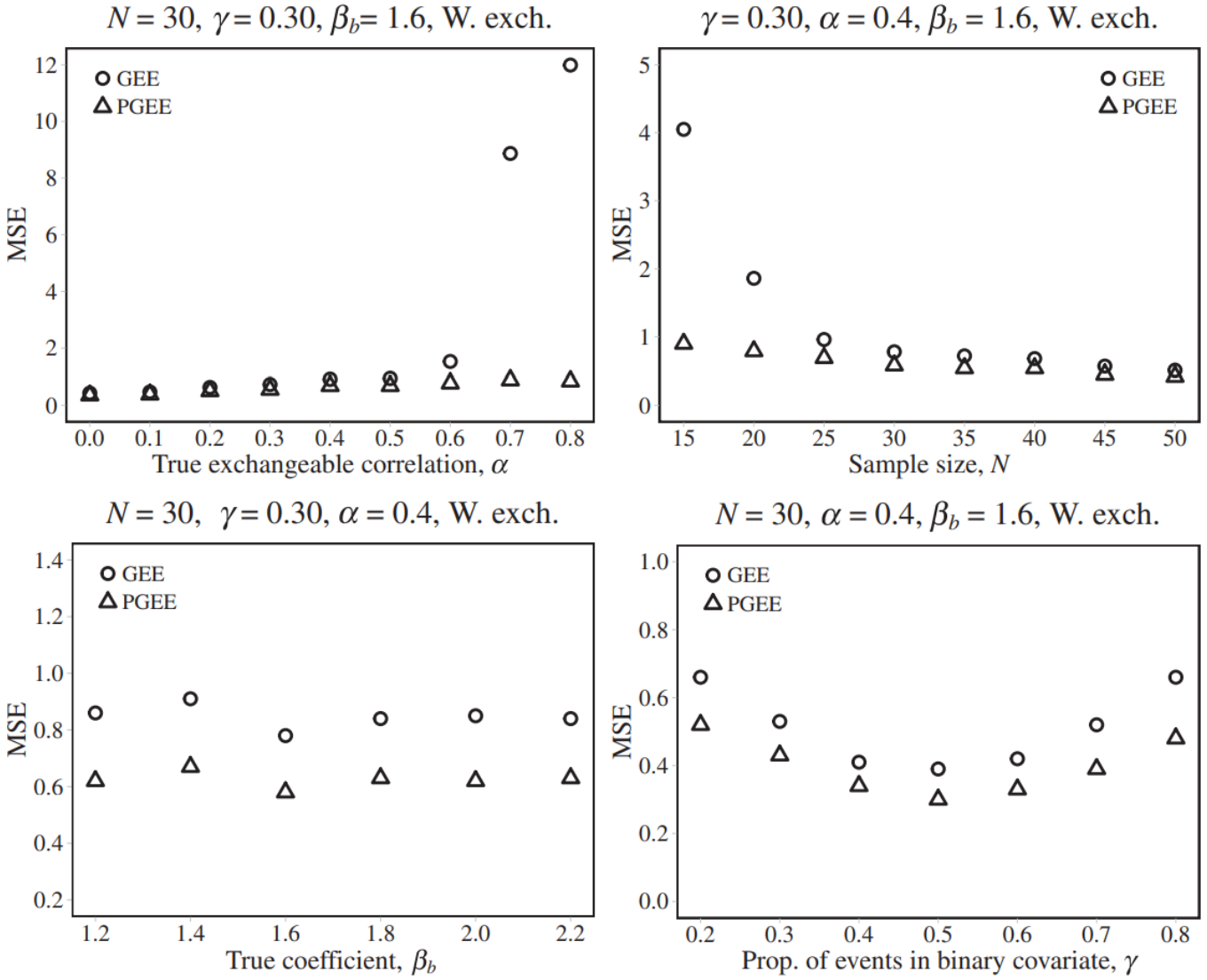
	id	yij	intercept	X1i	obstime
1	1	0	1	0	0.2
2	1	0	1	0	0.4
3	1	0	1	0	0.6
4	1	1	1	0	0.8
5	2	0	1	0	0.2
6	2	1	1	0	0.4
7	2	0	1	0	0.6
8	2	0	1	0	0.8
9	3	0	1	0	0.2
10	3	0	1	0	0.4
11	3	0	1	0	0.6
12	3	0	1	0	0.8
13	4	0	1	1	0.2
14	4	1	1	1	0.4
15	4	0	1	1	0.6
16	4	0	1	1	0.8
17	5	0	1	0	0.2
18	5	0	1	0	0.4
19	5	0	1	0	0.6
20	5	0	1	0	0.8
21	6	0	1	1	0.2
22	6	0	1	1	0.4
23	6	0	1	1	0.6
24	6	0	1	1	0.8
25	7	0	1	0	0.2
26	7	0	1	0	0.4
27	7	0	1	0	0.6
28	7	0	1	0	0.8

- Bias of GEE and PGEE estimates of β_b



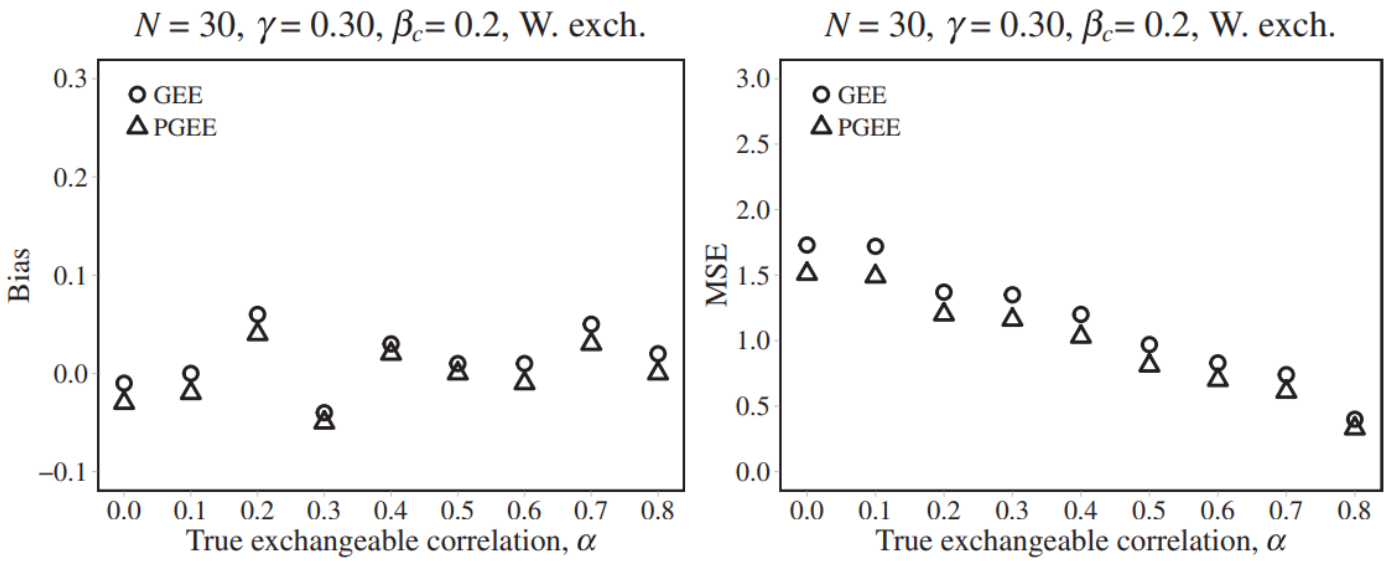
- β_0 = coefficient for intercept
- β_b = coefficient for binary covariate
- β_c = coefficient for continuous covariate
- N = num of cluster
- n = num of observation in each cluster
- α = intra-cluster correlation
- γ = proportion of events in binary covariate

- MSE of GEE and PGEE estimates of β_b



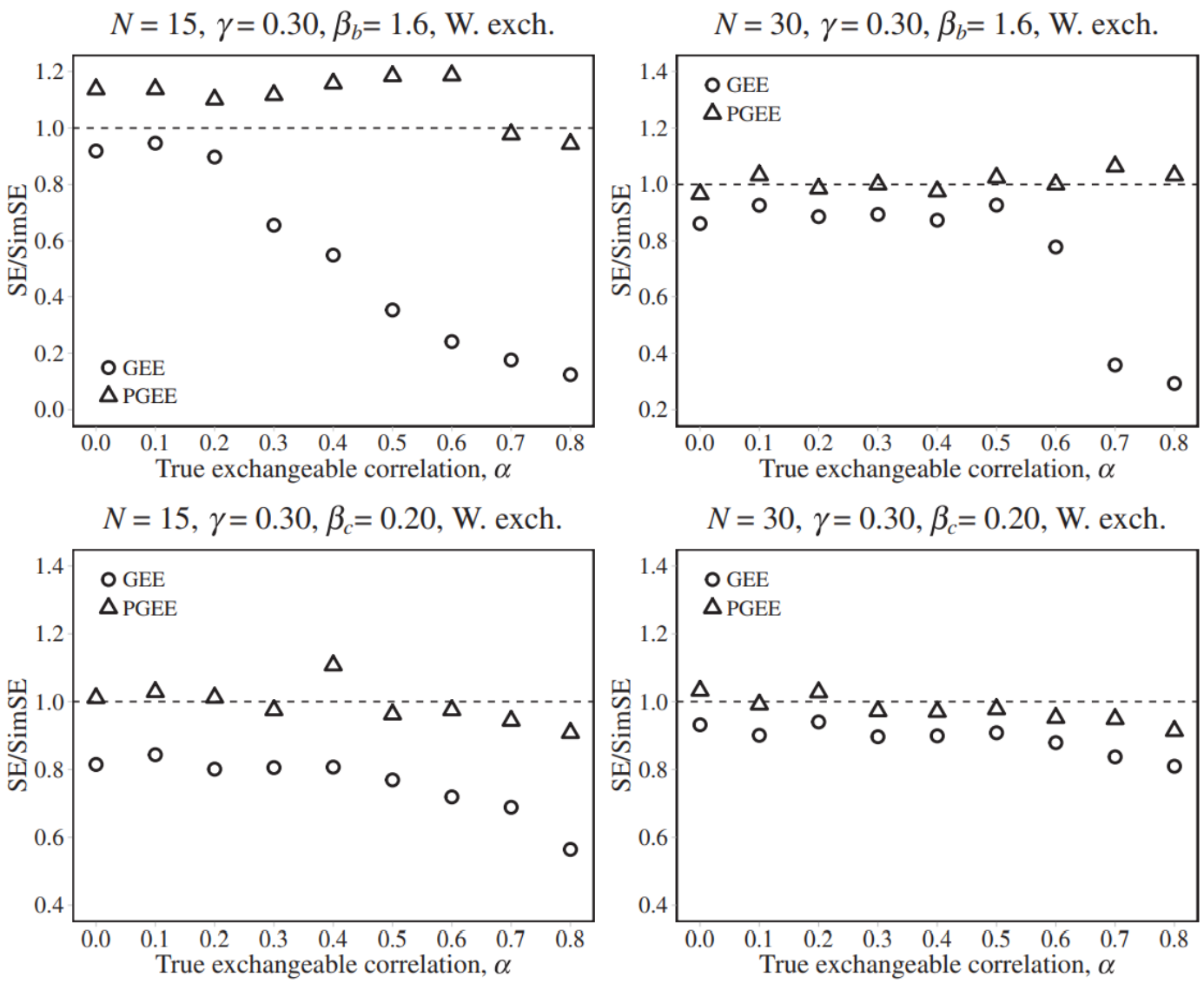
- β_0 = coefficient for intercept
- β_b = coefficient for binary covariate
- β_c = coefficient for continuous covariate
- N = num of cluster
- n = num of observation in each cluster
- α = intra-cluster correlation
- γ = proportion of events in binary covariate

- MSE of GEE and PGEE estimates of β_c



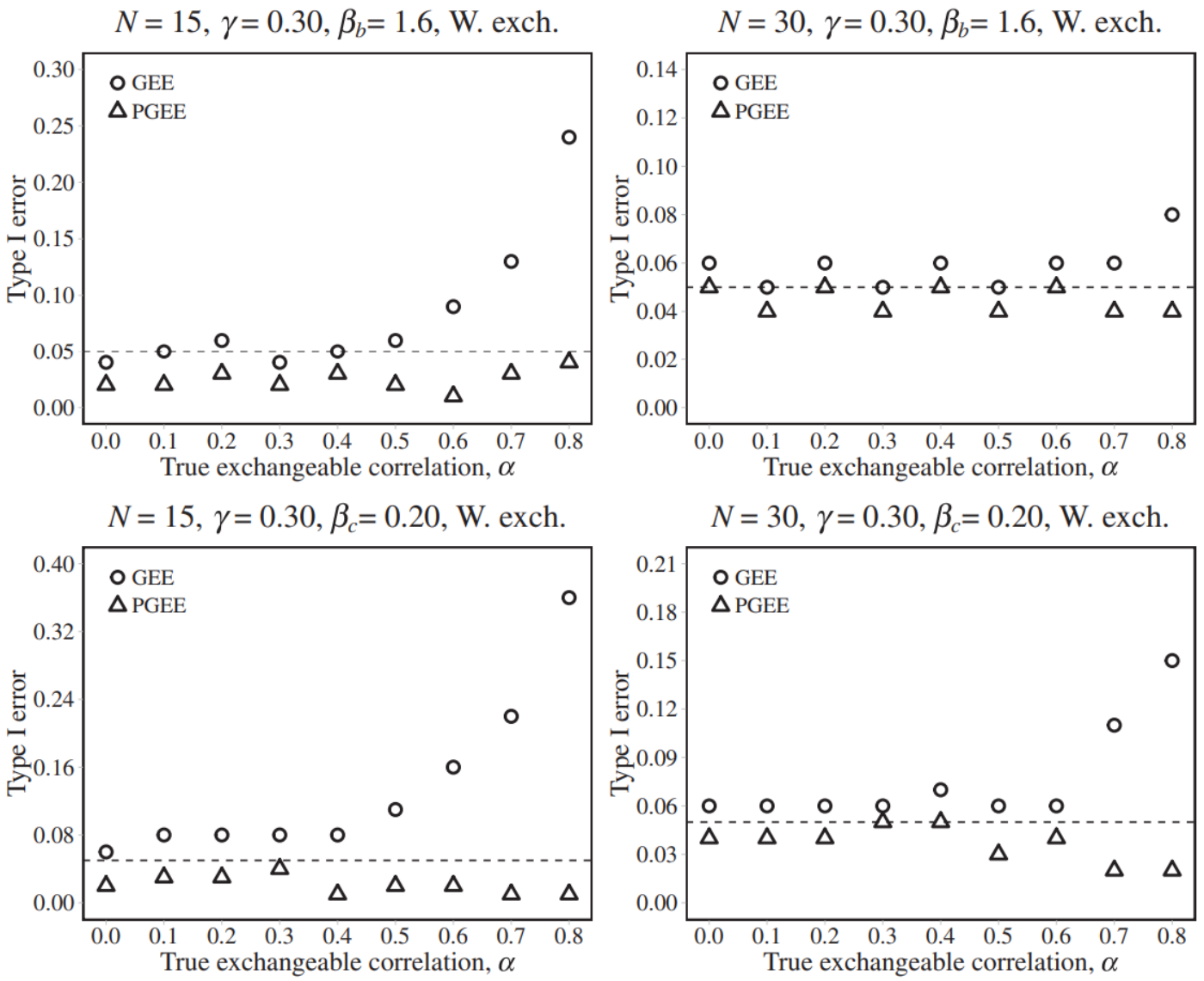
- β_0 = coefficient for intercept
- β_b = coefficient for binary covariate
- β_c = coefficient for continuous covariate
- N = num of cluster
- n = num of observation in each cluster
- α = intra-cluster correlation
- γ = proportion of events in binary covariate

- Relative standard error of GEE and PGEE estimates of β_c , β_b



- β_0 = coefficient for intercept
- β_b = coefficient for binary covariate
- β_c = coefficient for continuous covariate
- N = num of cluster
- n = num of observation in each cluster
- α = intra-cluster correlation
- γ = proportion of events in binary covariate

- Type I error rate associated with GEE and PGEE
- $H_0 : \beta_b = 1.6$ and $H_0 : \beta_c = 0.2$
- 5% level of significance



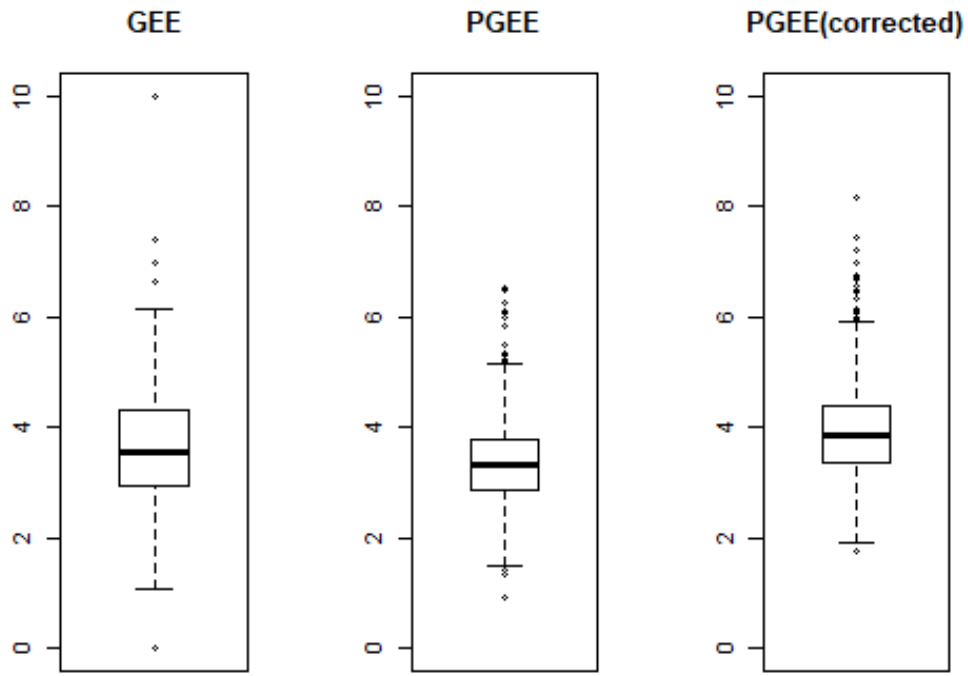
- β_0 = coefficient for intercept
- β_b = coefficient for binary covariate
- β_c = coefficient for continuous covariate
- N = num of cluster
- n = num of observation in each cluster
- α = intra-cluster correlation
- γ = proportion of events in binary covariate

Empirical Result1 – confidence interval

4

Empirical result 1 - confidence interval

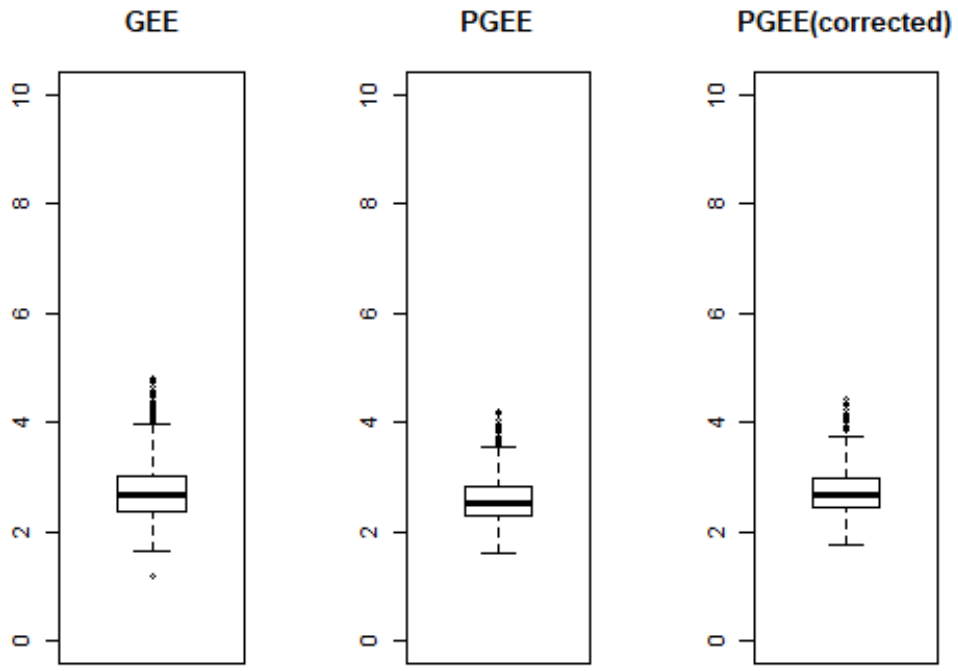
- Simulation1 Result ($N = 15, \gamma = 0.3,\ \beta_0 = -2.5,\ \beta_b = 1.6,\ \beta_c = 0.2$)



N = 15	α(rho) = 0.0	α(rho) = 0.2	α(rho) = 0.4	α(rho) = 0.6
GEE	900	845	779	699
PGEE	906	866	802	737
PGEE (corrected)	940	901	878	846

Boxplot of the length of confidence interval(β_b)
- α(rho) = 0.2 case

- Simulation2 Result ($\mathrm{N} = 30, \gamma = 0.3,\ \beta_0 = -2.5,\ \beta_b = 1.6,\ \beta_c = 0.2$)



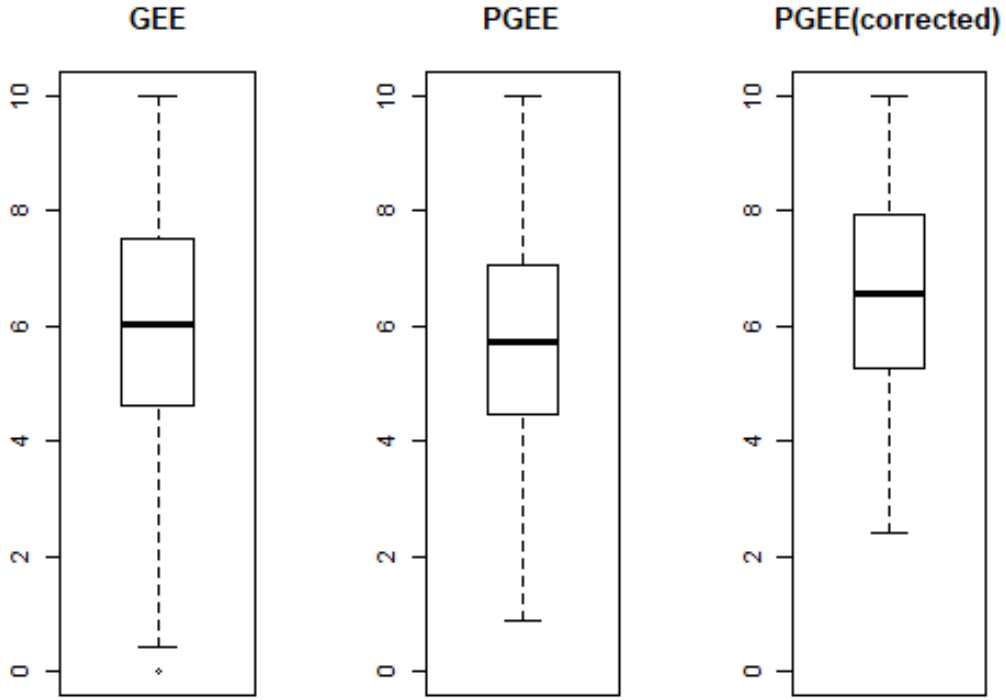
N = 30	α(rho) = 0.0	α(rho) = 0.2	α(rho) = 0.4	α(rho) = 0.6
GEE	947	942	936	903
PGEE	952	944	932	901
PGEE (corrected)	965	957	945	913

Boxplot of the length of confidence interval(β_b)
- α(rho) = 0.2 case

4

Empirical result 1 - confidence interval

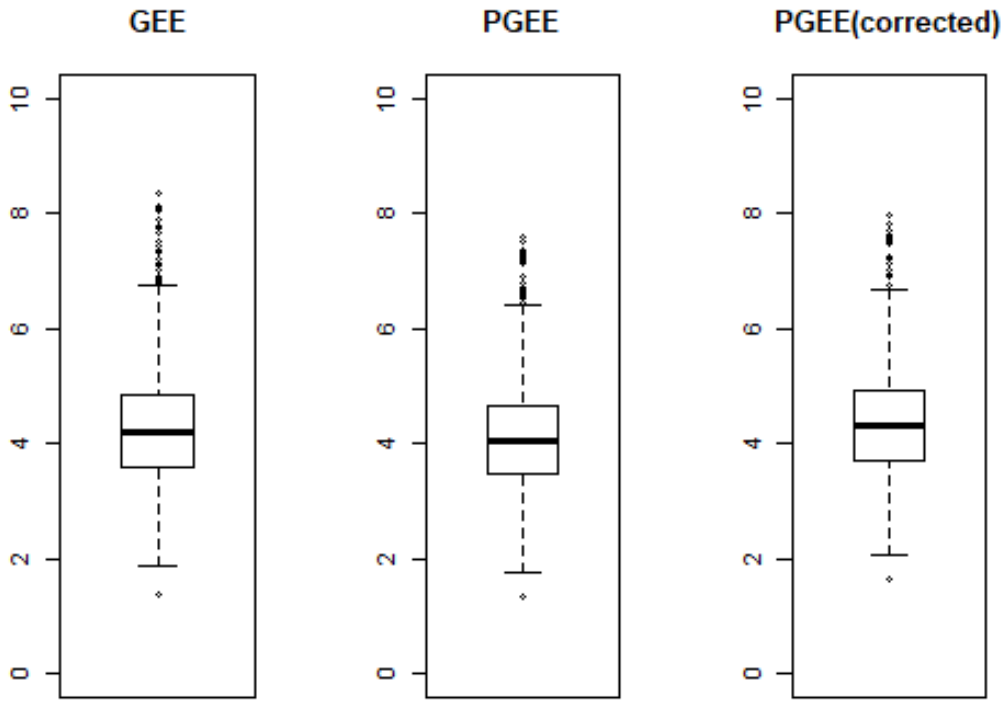
- Simulation3 Result ($N = 15, \tau = 0.3, \ \beta_0 = -2.5, \ \beta_b = 1.6, \ \beta_c = 0.2$)



N = 15	α(rho) = 0.0	α(rho) = 0.2	α(rho) = 0.4	α(rho) = 0.6
GEE	904	893	865	844
PGEE	915	896	856	840
PGEE (corrected)	947	945	938	955

Boxplot of the length of confidence interval(β_c)
- α(rho) = 0.2 case

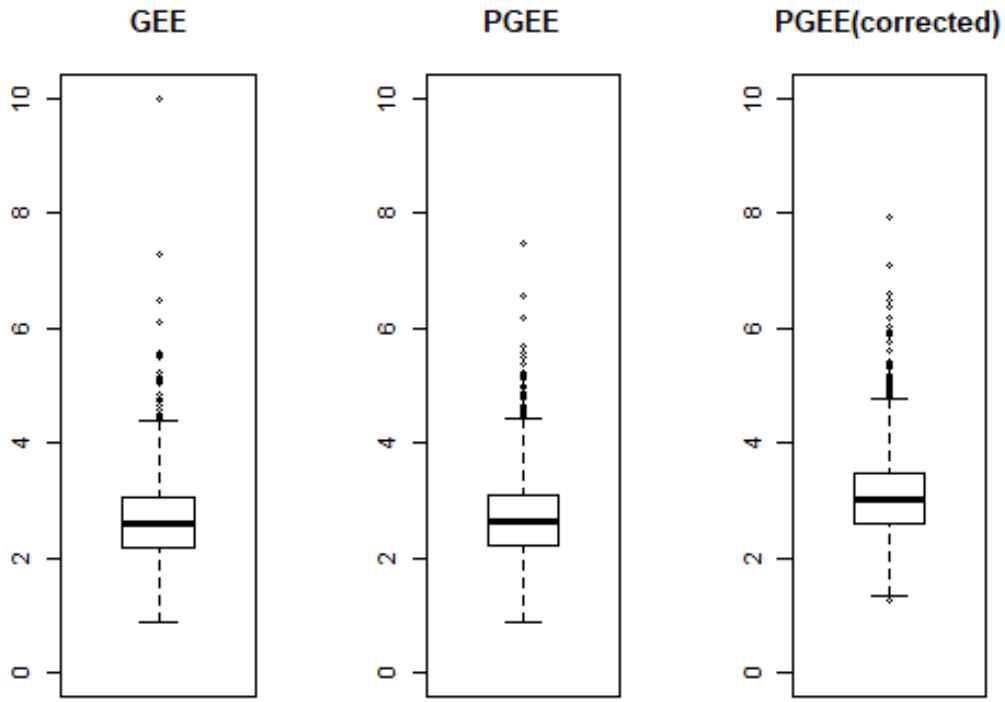
- Simulation4 Result ($N = 15, \tau = 0.3, \ \beta_0 = -2.5, \ \beta_b = 1.6, \ \beta_c = 0.2$)



N = 30	α(rho) = 0.0	α(rho) = 0.2	α(rho) = 0.4	α(rho) = 0.6
GEE	950	929	923	929
PGEE	954	927	915	911
PGEE (corrected)	961	944	937	954

Boxplot of the length of confidence interval(β_c)
- α(rho) = 0.2 case

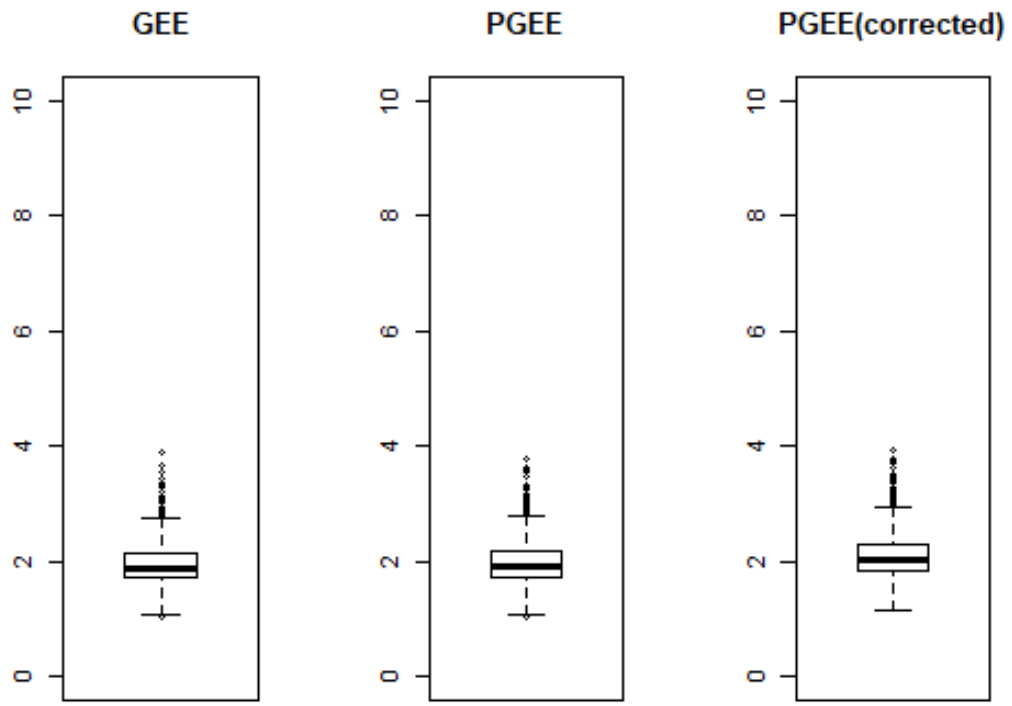
- Simulation5 Result ($N = 15, \gamma = 0.3,\ \beta_0 = -2.5,\ \beta_b = 0.16,\ \beta_c = 4$)



N = 15	α(rho) = 0.0	α(rho) = 0.1
GEE	878	890
PGEE	911	914
PGEE (corrected)	958	961

Boxplot of the length of confidence interval(β_b)
- α(rho) = 0.1 case

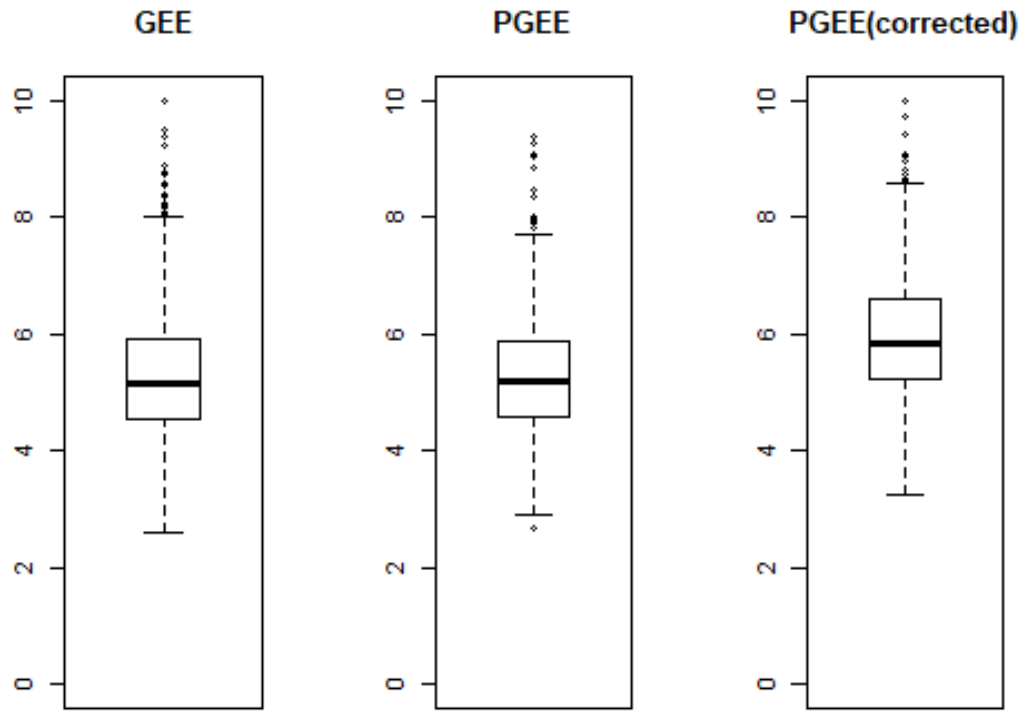
- Simulation6 Result ($\mathrm{N} = 30, \gamma = 0.3,\ \beta_0 = -2.5,\ \beta_b = 0.16,\ \beta_c = 4$)



N = 30	α(rho) = 0.0	α(rho) = 0.1
GEE	924	928
PGEE	932	940
PGEE (corrected)	948	957

Boxplot of the length of confidence interval(β_b)
- α(rho) = 0.1 case

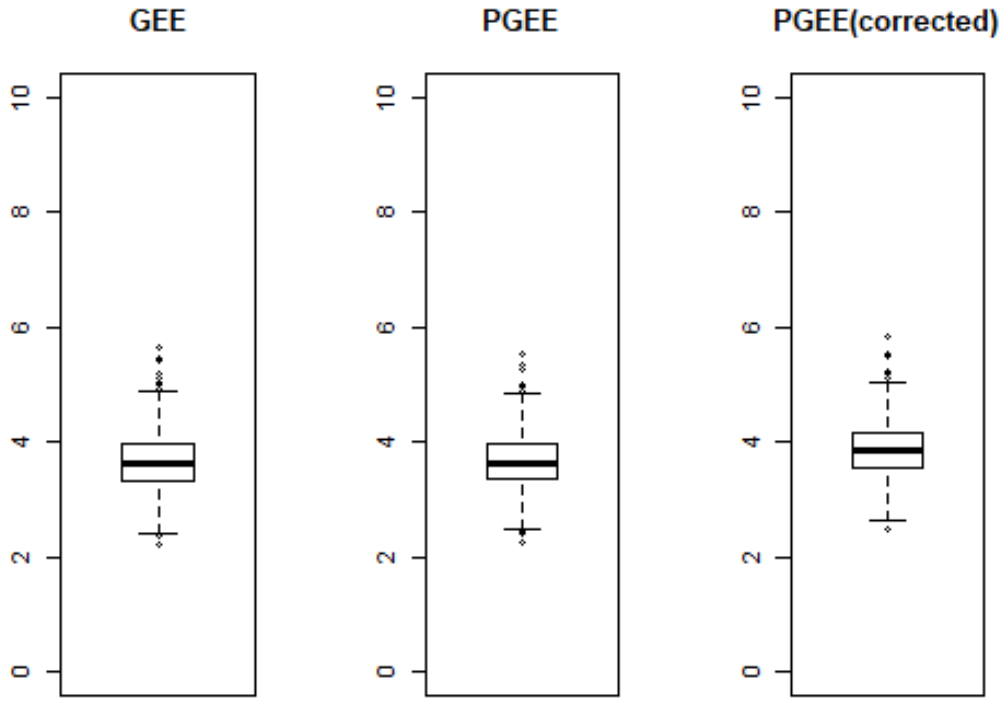
- Simulation7 Result ($N = 15, \tau = 0.3, \ \beta_0 = -2.5, \ \beta_b = 0.16, \ \beta_c = 4$)



N = 15	α(rho) = 0.0	α(rho) = 0.1
GEE	928	929
PGEE	944	943
PGEE (corrected)	968	969

Boxplot of the length of confidence interval(β_c)
- α(rho) = 0.1 case

- Simulation8 Result ($N = 30, \gamma = 0.3, \ \beta_0 = -2.5, \ \beta_b = 0.16, \ \beta_c = 4$)



N = 30	α(rho) = 0.0	α(rho) = 0.1
GEE	949	943
PGEE	956	945
PGEE (corrected)	965	959

Boxplot of the length of confidence interval(β_c)
- α(rho) = 0.1 case

Empirical Result2 - real data analysis

Dataset – Clinical trial of patients with respiratory illness

```
> head(df, 15)
   id hospital treatment time0 time1 time2 time3 time4
1   1        1         P     0     0     0     0     0
2   2        1         P     0     0     0     0     0
3   3        1         A     1     1     1     1     1
4   4        1         P     1     1     1     1     0
5   5        1         P     1     1     1     1     1
6   6        1         A     0     0     0     0     0
7   7        1         P     0     1     0     1     1
8   8        1         A     0     0     0     0     0
9   9        1         A     1     1     1     1     1
10 10        1         P     1     0     1     1     0
```

- id - patient identification
- hospital – clinic
- treatment – A(Active), P(Placebo) / 1=A, 0=P
- time - (time 0~4) / 1=Yes, 0=No (Respiratory illness)
- N = 111 (56 patients from clinic 1 and 55 from clinic 2)

The dataset is publicly available and used by Fitzmaurice et al31 in the book titled "Applied Longitudinal Analysis."

Scenarios

TABLE 4 Contingency tables between dichotomous covariates (treatment) and binary response (respiratory illness status)

I. Quasi-complete Sep. ($N=15$)				II. Near-to-Sep ($N=15$)			III. Near-to-Sep. ($N=15$)			IV. No Sep. ($N=111$)		
		Respiratory illness			Respiratory illness			Respiratory illness			Respiratory illness	
		yes	no		yes	no		yes	no		yes	no
Treatment	Active	25	0	Active	17	2	Active	16	10	Active	171	99
	Placebo	26	24	Placebo	28	28	Placebo	19	30	Placebo	127	158

Each of the scenarios I, II, and III is based on a subset of patients from full dataset and scenario IV is based on full data.

- Scenario Ⅰ
  - Quasi-complete-Separation (the responses and nonresponses are nearly separated by the predictor)
- Scenario Ⅱ
  - Near-to-Separation (nonzero cell but with few observations)
- Scenario Ⅲ
  - Near-to-Separation (nonzero cell but with few observations, more overlap than Scenario Ⅱ)
- Scenario Ⅳ
  - No-Separation (full data, N=111)

- Ⅰ, Ⅱ, Ⅲ - taken a random subsample of patients of size N = 15 from the clinic 2
- Near-to-Separation ($75 \times 0.15 = 11.25$)

The dataset is publicly available and used by Fitzmaurice et al31 in the book titled "Applied Longitudinal Analysis."

Result

TABLE 5 The summary of the fitted PGEE and GEE on the Clinical Trial of Patients with Respiratory Illness data. The value in the parenthesis is the estimated Std. error using the PGEE(M)

Estimate		I. Quasi-Complete Sep. ($N = 15$)		II. Near-to-Sep. ($N = 15$)		III. Near-to-Sep. ($N = 15$)		IV. No Sep. ($N = 111$)	
		PGEE	GEE	PGEE	GEE	PGEE	GEE	PGEE	GEE
Coefficients	Intercept	−0.16	−0.16	−0.56	−0.57	−0.38	−0.39	−0.39	−0.40
	Treatment	3.85	44.5	1.78	1.95	0.85	0.94	0.74	0.76
	Time	0.08	0.08	0.36	0.37	0.18	0.19	0.06	0.06
	$\hat{\alpha}$	0.11	0.16	0.20	0.24	0.35	0.35	0.46	0.46
Std. error	Intercept	0.42	0.40	0.51	0.50	0.66	0.65	0.25	0.25
	Treatment	0.62	0.58	0.70	0.73	0.77	0.76	0.29	0.29
	Time	0.16	0.15	0.21	0.20	0.14	0.13	0.05	0.05

Abbreviations: GEE, generalized estimating equation; PGEE, penalized GEE.

- The GEE estimate of the regression coefficient of the dichotomous covariate (treatment status) that caused quasi-complete separation is reported to very large, which is not interpretable.

Result

TABLE 5 The summary of the fitted PGEE and GEE on the Clinical Trial of Patients with Respiratory Illness data. The value in the parenthesis is the estimated Std. error using the PGEE(M)

Estimate		I. Quasi-Complete Sep. ($N = 15$)		II. Near-to-Sep. ($N = 15$)		III. Near-to-Sep. ($N = 15$)		IV. No Sep. ($N = 111$)	
		PGEE	GEE	PGEE	GEE	PGEE	GEE	PGEE	GEE
Coefficients	Intercept	−0.16	−0.16	−0.56	−0.57	−0.38	−0.39	−0.39	−0.40
	Treatment	3.85	44.5	1.78	1.95	0.85	0.94	0.74	0.76
	Time	0.08	0.08	0.36	0.37	0.18	0.19	0.06	0.06
	$\hat{\alpha}$	0.11	0.16	0.20	0.24	0.35	0.35	0.46	0.46
Std. error	Intercept	0.42	0.40	0.51	0.50	0.66	0.65	0.25	0.25
	Treatment	0.62	0.58	0.70	0.73	0.77	0.76	0.29	0.29
	Time	0.16	0.15	0.21	0.20	0.14	0.13	0.05	0.05

Abbreviations: GEE, generalized estimating equation; PGEE, penalized GEE.

- The GEE estimate becomes small with the increasing amount of overlapping in near-to-quasi-complete separation.

Result

TABLE 5 The summary of the fitted PGEE and GEE on the Clinical Trial of Patients with Respiratory Illness data. The value in the parenthesis is the estimated Std. error using the PGEE(M)

Estimate		I. Quasi-Complete Sep. ($N = 15$)		II. Near-to-Sep. ($N = 15$)		III. Near-to-Sep. ($N = 15$)		IV. No Sep. ($N = 111$)	
		PGEE	GEE	PGEE	GEE	PGEE	GEE	PGEE	GEE
Coefficients	Intercept	−0.16	−0.16	−0.56	−0.57	−0.38	−0.39	−0.39	−0.40
	Treatment	3.85	44.5	1.78	1.95	0.85	0.94	0.74	0.76
	Time	0.08	0.08	0.36	0.37	0.18	0.19	0.06	0.06
	$\hat{\alpha}$	0.11	0.16	0.20	0.24	0.35	0.35	0.46	0.46
Std. error	Intercept	0.42	0.40	0.51	0.50	0.66	0.65	0.25	0.25
	Treatment	0.62	0.58	0.70	0.73	0.77	0.76	0.29	0.29
	Time	0.16	0.15	0.21	0.20	0.14	0.13	0.05	0.05

Abbreviations: GEE, generalized estimating equation; PGEE, penalized GEE.

- In contrast, the PGEE provided comparatively very smaller estimate of the corresponding regression coefficient for each of the scenarios.

Result

TABLE 5 The summary of the fitted PGEE and GEE on the Clinical Trial of Patients with Respiratory Illness data. The value in the parenthesis is the estimated Std. error using the PGEE(M)

Estimate		I. Quasi-Complete Sep. ($N = 15$)		II. Near-to-Sep. ($N = 15$)		III. Near-to-Sep. ($N = 15$)		IV. No Sep. ($N = 111$)	
		PGEE	GEE	PGEE	GEE	PGEE	GEE	PGEE	GEE
Coefficients	Intercept	−0.16	−0.16	−0.56	−0.57	−0.38	−0.39	−0.39	−0.40
	Treatment	3.85	44.5	1.78	1.95	0.85	0.94	0.74	0.76
	Time	0.08	0.08	0.36	0.37	0.18	0.19	0.06	0.06
	$\hat{\alpha}$	0.11	0.16	0.20	0.24	0.35	0.35	0.46	0.46
Std. error	Intercept	0.42	0.40	0.51	0.50	0.66	0.65	0.25	0.25
	Treatment	0.62	0.58	0.70	0.73	0.77	0.76	0.29	0.29
	Time	0.16	0.15	0.21	0.20	0.14	0.13	0.05	0.05

Abbreviations: GEE, generalized estimating equation; PGEE, penalized GEE.

- Both the GEE and PGEE provided almost similar results in case of full dataset where there is no separation.

Result

TABLE 5 The summary of the fitted PGEE and GEE on the Clinical Trial of Patients with Respiratory Illness data. The value in the parenthesis is the estimated Std. error using the PGEE(M)

Estimate		I. Quasi-Complete Sep. ($N = 15$)		II. Near-to-Sep. ($N = 15$)		III. Near-to-Sep. ($N = 15$)		IV. No Sep. ($N = 111$)	
		PGEE	GEE	PGEE	GEE	PGEE	GEE	PGEE	GEE
Coefficients	Intercept	−0.16	−0.16	−0.56	−0.57	−0.38	−0.39	−0.39	−0.40
	Treatment	3.85	44.5	1.78	1.95	0.85	0.94	0.74	0.76
	Time	0.08	0.08	0.36	0.37	0.18	0.19	0.06	0.06
	$\hat{\alpha}$	0.11	0.16	0.20	0.24	0.35	0.35	0.46	0.46
Std. error	Intercept	0.42	0.40	0.51	0.50	0.66	0.65	0.25	0.25
	Treatment	0.62	0.58	0.70	0.73	0.77	0.76	0.29	0.29
	Time	0.16	0.15	0.21	0.20	0.14	0.13	0.05	0.05

Abbreviations: GEE, generalized estimating equation; PGEE, penalized GEE.

- Both methods showed comparable results for the coefficient associated with "time," which did not make separation.

Result

TABLE 5 The summary of the fitted PGEE and GEE on the Clinical Trial of Patients with Respiratory Illness data. The value in the parenthesis is the estimated Std. error using the PGEE(M)

Estimate		I. Quasi-Complete Sep. ($N = 15$)		II. Near-to-Sep. ($N = 15$)		III. Near-to-Sep. ($N = 15$)		IV. No Sep. ($N = 111$)	
		PGEE	GEE	PGEE	GEE	PGEE	GEE	PGEE	GEE
Coefficients	Intercept	−0.16	−0.16	−0.56	−0.57	−0.38	−0.39	−0.39	−0.40
	Treatment	3.85	44.5	1.78	1.95	0.85	0.94	0.74	0.76
	Time	0.08	0.08	0.36	0.37	0.18	0.19	0.06	0.06
	$\hat{\alpha}$	0.11	0.16	0.20	0.24	0.35	0.35	0.46	0.46
Std. error	Intercept	0.42	0.40	0.51	0.50	0.66	0.65	0.25	0.25
	Treatment	0.62	0.58	0.70	0.73	0.77	0.76	0.29	0.29
	Time	0.16	0.15	0.21	0.20	0.14	0.13	0.05	0.05

Abbreviations: GEE, generalized estimating equation; PGEE, penalized GEE.

- The estimated standard error for the PGEE estimate was comparatively larger than those with GEE for all scenarios related to separation.

Result

TABLE 5 The summary of the fitted PGEE and GEE on the Clinical Trial of Patients with Respiratory Illness data. The value in the parenthesis is the estimated Std. error using the PGEE(M)

Estimate		I. Quasi-Complete Sep. ($N = 15$)		II. Near-to-Sep. ($N = 15$)		III. Near-to-Sep. ($N = 15$)		IV. No Sep. ($N = 111$)	
		PGEE	GEE	PGEE	GEE	PGEE	GEE	PGEE	GEE
Coefficients	Intercept	−0.16	−0.16	−0.56	−0.57	−0.38	−0.39	−0.39	−0.40
	Treatment	3.85	44.5	1.78	1.95	0.85	0.94	0.74	0.76
	Time	0.08	0.08	0.36	0.37	0.18	0.19	0.06	0.06
	$\hat{\alpha}$	0.11	0.16	0.20	0.24	0.35	0.35	0.46	0.46
Std. error	Intercept	0.42	0.40	0.51	0.50	0.66	0.65	0.25	0.25
	Treatment	0.62	0.58	0.70	0.73	0.77	0.76	0.29	0.29
	Time	0.16	0.15	0.21	0.20	0.14	0.13	0.05	0.05

Abbreviations: GEE, generalized estimating equation; PGEE, penalized GEE.

- However both showed comparable results when there was no separation.

Reference

[1] A.AGRESTI. AN INTRODUCTION TO CATEGORICAL DATA ANALYSIS
[2] M.MONDOL, M.RAHMAN. Bias-reduced and separation-proof GEE with small or sparse longitudinal binary data

감 사 합 니 다