Published in final edited form as:
J Stat Comput Simul. 2016 ; 86(10): 1891–1900. doi:10.1080/00949655.2015.1089873.

A covariance correction that accounts for correlation estimation to improve finite-sample inference with generalized estimating equations: A study on its applicability with structured correlation matrices

Philip M. Westgate
Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, KY, 40536, U.S.A., Phone: (859) 218-2082, philip.westgate@uky.edu

Abstract

When generalized estimating equations (GEE) incorporate an unstructured working correlation matrix, the variances of regression parameter estimates can inflate due to the estimation of the correlation parameters. In previous work, an approximation for this inflation that results in a corrected version of the sandwich formula for the covariance matrix of regression parameter estimates was derived. Use of this correction for correlation structure selection also reduces the over-selection of the unstructured working correlation matrix. In this manuscript, we conduct a simulation study to demonstrate that an increase in variances of regression parameter estimates can occur when GEE incorporates structured working correlation matrices as well. Correspondingly, we show the ability of the corrected version of the sandwich formula to improve the validity of inference and correlation structure selection. We also study the relative influences of two popular corrections to a different source of bias in the empirical sandwich covariance estimator.

Keywords

bias correction; correlation selection; efficiency; empirical covariance matrix; generalized estimating equations

1. Introduction

Generalized estimating equations (GEE) [1] are commonly utilized for the analysis of correlated data when a marginal model is desired. When GEE incorporates an unstructured working correlation matrix, it has been shown that the covariance matrix of the regression parameter estimates may inflate due to the need to estimate nuisance correlation parameters [2]. We note that although relatively unknown, to our knowledge, in the GEE literature before the work of Westgate [2], this type of small-sample variance inflation is very well known when data arise from a multivariate normal distribution and a linear mixed model is

Supplemental material
One supplemental file includes additional simulation results. Specifically, results for a true AR-1 structure and normal outcomes are presented, as well as results from settings in which outcomes are binary. The other supplemental file includes an R function that implements GEE and outputs results based on the methods presented in this manuscript.

used [3]. In fact, the Kenward and Roger method [4, 5], which accounts for this inflation, has enjoyed great popularity when the working covariance structure is assumed to be correctly specified such that model-based standard error estimates can be utilized.

With respect to GEE, Westgate [2] derived an approximation for this inflation when utilizing an unstructured working correlation matrix, resulting in a corrected version of the well-known sandwich formula for the covariance matrix of the regression parameter estimates. Furthermore, Westgate [6] showed that, in order to improve regression parameter estimation via correlation structure selection, this covariance correction can be used to penalize the estimation of the multiple nuisance correlation parameters within the unstructured matrix in order to reduce its over-selection. In this manuscript, we conduct a simulation study to demonstrate that even when GEE incorporates structured correlation matrices, the variances of the regression parameter estimates can still inflate. Therefore, we also apply and study the use of the covariance inflation correction when GEE incorporates structured correlation matrices. Specifically, we show that use of this correction can improve inference when using structured working correlation matrices, and that this correction should be utilized as a penalty by correlation selection criteria for all structures under consideration. Furthermore, unrelated to the covariance inflation correction, a correction for the bias in the meat of the empirical sandwich covariance matrix estimator is needed in small-sample settings. Therefore, we study the relative influences of two such corrections, proposed by Kauermann and Carroll [7] and Mancl and DeRouen [8], that have found popularity.

Section 2 introduces notation and discusses GEE, the covariance inflation correction, estimation of the sandwich covariance matrix, and correlation structure selection. Our simulation study is presented in Section 3. Finally, concluding remarks are given in Section 4.

2. Notation, GEE, Covariance Correction and Estimation, and Correlation Selection

Assume we have data from N independent clusters. The observed outcome vector for the ith cluster is denoted by $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]^T$, which has a marginal mean given by $E(\boldsymbol{Y}_i) = \boldsymbol{\mu}_i$ that is linked to covariates via a function, f, such that $f(\mu_{ij})=\mathbf{x}_{ij}^T\beta$ for $\mathbf{x}_{ij} = [1, x_{1ij}, \ldots, x_{(p-1)ij}]^T$ and $\beta = [\beta_0, \beta_1, \ldots, \beta_{p-1}]^T$. The corresponding working covariance matrix for $\boldsymbol{Y}_i$ is given by $\boldsymbol{V}_i=\boldsymbol{A}_i^{1/2}\boldsymbol{R}_i(\alpha)\boldsymbol{A}_i^{1/2}$, $i = 1, \ldots, N$. Here, $\boldsymbol{A}_i = diag[\phi\nu(\mu_{i1}), \ldots, \phi\nu(\mu_{in_i})]$ is a diagonal matrix of working marginal variances, ϕ is an assumed common dispersion parameter, ν is a known function, and $\boldsymbol{R}_i(\alpha)$ is a working correlation matrix with 1 along the diagonal and one or more parameters given by α.

Let $\boldsymbol{D}_i = \partial\boldsymbol{\mu}_i/\partial\beta^T$, and denote a consistent working estimate for β by $\tilde{\beta}$. To obtain the final estimate of the regression parameters, $\hat{\beta}$, using GEE [1], we iteratively solve

$$\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{R}_i^{-1}\left(\hat{\alpha}(\tilde{\beta})\right)\boldsymbol{A}_i^{-1/2}(\boldsymbol{Y}_i-\boldsymbol{\mu}_i)=\mathbf{0}, \quad (1)$$

for which $\tilde{\beta} = \hat{\beta}$ at the end of the iterative procedure. The well-known sandwich formula for the covariance matrix of $\hat{\beta}$ is given by $Cov(\hat{\beta}) \approx$

$$\boldsymbol{\Sigma}=\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\mathrm{Cov}(\boldsymbol{Y}_i)\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}. \quad (2)$$

The sandwich formula of Equation (2) assumes correlation parameters are known, although in practice they must be estimated. Additionally, $\hat{\alpha}(\beta)$ must be replaced with $\hat{\alpha}(\tilde{\beta})$ in Equation (1). As a result, because $\boldsymbol{R}_i^{-1}\left(\hat{\alpha}(\tilde{\beta})\right)$ varies about $\boldsymbol{R}_i^{-1}\left(\hat{\alpha}(\beta)\right)$, the estimation variability of GEE can increase, thus inflating $Cov(\hat{\beta})$ [2]. Specifically, Westgate [2] showed that

$$\mathrm{Cov}(\hat{\beta}) \approx (\boldsymbol{I}_p+\boldsymbol{G})\boldsymbol{\Sigma}(\boldsymbol{I}_p+\boldsymbol{G})^T \quad (3)$$

after accounting for covariance inflation via a Taylor series expansion. Here, $\boldsymbol{I}_p$ is a $p \times p$ identity matrix, and $\boldsymbol{G} = (\boldsymbol{G}_0, \boldsymbol{G}_1, \ldots, \boldsymbol{G}_{p-1})$,

$$\boldsymbol{G}_r=-\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1/2}\boldsymbol{R}_i^{-1}\frac{\partial\boldsymbol{R}_i\left(\hat{\alpha}(\beta)\right)}{\partial\beta_r}\boldsymbol{R}_i^{-1}\boldsymbol{A}_i^{-1/2}\left(\boldsymbol{Y}_i-\boldsymbol{\mu}_i(\beta)\right).$$

We note that although Westgate [2, 6] only applied the covariance inflation correction in Equation (3) when GEE incorporates an unstructured working correlation matrix, this correction is not restricted and can be applied with GEE regardless of the type of working correlation structure, as will be demonstrated in our simulation study.

In practice, unknown parameters must be estimated within the formula for $Cov(\hat{\beta})$. Therefore, an arbitrary estimator can be denoted by $(\boldsymbol{I}_p + \hat{\boldsymbol{G}})\hat{\Sigma}(\boldsymbol{I}_p + \hat{\boldsymbol{G}})^T$. Within $\boldsymbol{G}$, unknown parameters can be estimated using $\hat{\beta}$, resulting in $\hat{\boldsymbol{G}}$. However, Σ can be estimated in different manners. If we assume the working covariance structure is correctly specified, then $Cov(\boldsymbol{Y}_i)$ in Equation (2) can be replaced with $\boldsymbol{V}_i$, $i = 1, \ldots, N$, resulting in the model-based estimator $\hat{\Sigma}_{MB}=\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{V}_i^{-1}\boldsymbol{D}_i\right)^{-1}$. However, if the working structure is misspecified, $\hat{\Sigma}_{MB}$ will be biased. Therefore, a common form for $\hat{\Sigma}$ is the Liang and Zeger [1] empirical estimator, $\hat{\Sigma}_{LZ}$, that replaces $Cov(\boldsymbol{Y}_i)$ in Equation (2) with $(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)^T$, $i = 1, \ldots, N$. This estimator is routinely used with GEE because it generally is a consistent estimate for $Cov(\hat{\beta})$ that does not require the working and true covariance structures to be equivalent [1]. We further note that in small-sample settings, $\hat{\Sigma}_{LZ}$ can be biased for Σ because $(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i)$, $i = 1, \ldots, N$, tends to be too small [8]. Therefore, multiple corrections have been proposed to reduce this bias, such as the popular corrections proposed by Kauermann and Carroll [7] and Mancl and DeRouen [8]. As these two corrections can yield notably different standard error estimates in small-sample settings, in our simulation study we will assess the performances of both corrections in conjunction with the covariance inflation correction.

Accurate modeling of the working correlation structure has the potential to improve estimation efficiency [1, 9]. Therefore, multiple criteria have been proposed to select a working structure, many of which are summarized in and studied by Westgate [6]. For instance, the 'correlation information criterion' (CIC) and 'trace of the empirical covariance matrix' (TECM) criterion have been shown to work well [6, 10]. When incorporating the covariance inflation correction, as proposed by Westgate [6] to penalize, or account for, the estimation of nuisance correlation parameters, the CIC selects the working structure that gives the smallest value for $\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_I^{-1}(\boldsymbol{I}_p+\hat{\boldsymbol{G}})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I}_p+\hat{\boldsymbol{G}})^T\right)$, where $\hat{\boldsymbol{\Sigma}}_I=\left(\sum_{i=1}^{N}\boldsymbol{D}_i^T\boldsymbol{A}_i^{-1}\boldsymbol{D}_i\right)^{-1}$, and the TECM chooses the structure that yields the smallest value for $\mathrm{tr}\left((\boldsymbol{I}_p+\hat{\boldsymbol{G}})\hat{\boldsymbol{\Sigma}}(\boldsymbol{I}_p+\hat{\boldsymbol{G}})^T\right)$. We note that $\hat{\Sigma}$ must be $\hat{\Sigma}_{LZ}$ or a bias-corrected version of this estimator. Furthermore, Westgate [6] only applied the covariance inflation correction for the unstructured correlation matrix, whereas in our simulation study we show that it should be applied with all working structures that are under consideration for selection. For instance, structures such as independence, exchangeable, AR-1, and less parsimonous Toeplitz forms do not all have the same number of correlation parameters, and therefore each will have a different degree of covariance inflation that must be taken into account.

3. Simulation Study

3.1. Study Description

We now conduct a simulation study to show that variances of regression parameter estimates can inflate when GEE incorporates well-known structured working correlation matrices. Furthermore, we demonstrate the corresponding use, and study the necessity and utility, of the covariance inflation correction in Equation (3). Specifically, we study standard error (SE) estimation and the validity of inference via empirical covarage probabillities (CPs) of 95% confidence intervals (CIs). We further study correlation selection accuracy via the ability of correlation selection criteria to choose the true structure. As our focus in this manuscript is on structured working correlation matrices, we do not present results from an unstructured working matrix. Furthermore, because we focus on small-sample settings, we study the use of the covariance inflation correction in conjunction with the Kauermann and Carroll [7] and Mancl and DeRouen [8] corrections in order to assess the impact these latter two corrections for the bias in $\hat{\Sigma}_{LZ}$ has on the necessity and utility of the covariance inflation correction.

Multivariate normal data were generated from

$$Y_{ij}=\beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+\varepsilon_{ij};\quad j=1,\dots,n,$$

where $\beta = [0, 0.3, 0.3]^T$ and $Var(\varepsilon_{ij}) = 1$, $j = 1, \dots, n$; $i = 1, \dots, N$, and correlated binary outcomes were generated from the marginal model given by

$$\mathrm{logit}(\mu_{ij})=\beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij},\quad j=1,\dots,n,$$

where $\beta = [0, 0.1, 0.1]^T$. In both models, x_{1ij} and x_{2ij}, $j = 1, \ldots, n$, were independently generated from *Uniform*(0, 1). Models are similar to the ones used in Hin et al. [11], Hin and Wang [10], and Westgate [6].

Simulations were conducted in R version 2.13.1 [12]. Each setting was examined via 1,000 replications. Normal outcomes were generated using `rmvnorm` of the `mvtnorm` package [13, 14], whereas binary outcomes were generated using `rmvbin` of the `bindata` package [15]. When correlated outcomes are binary, additional constraints are required on the correlation parameters [16, 17]. Therefore, to avoid problems with data generation, and to enhance the stability of working correlation matrices, we utilized $\alpha_{exch} = \alpha_{AR-1} = 0.2$ in these settings.

In Tables 1 and 2, we focus on results for normal outcomes that only correspond to $\hat{\beta}_1$, as results for $\hat{\beta}_2$ are similar and the intercept is not of interest. Furthermore, we present two sets of results based on the use of either $\hat{\alpha}(\beta)$ (no resulting covariance inflation) or $\hat{\alpha}(\tilde{\beta})$ (results in covariance inflation, as will realistically be the case in practice) within Equation (1), denoted by "Theoretical Analyses" and "Realistic Analyses", respectively. SE estimates corresponding to the theoretical and realistic analyses are obtained from $\hat{\Sigma}$ and $(\boldsymbol{I}_p+\hat{\boldsymbol{G}})\hat{\Sigma}(\boldsymbol{I}_p+\hat{\boldsymbol{G}})^T$, respectively, and are denoted by SE_T and SE_R. We note that $\hat{\Sigma}$ is $\hat{\Sigma}_{LZ}$ with either the Kauermann and Carroll [7] or Mancl and DeRouen [8] correction. For each type of analysis, we present empirical standard deviations (ESDs) of $\hat{\beta}_1$, empirical means of SE estimates and corresponding 95% confidence interval (CI) empirical coverage probabilities (CPs). Variance inflation does not occur with the Theoretical Analyses, in which case SE_R is not applicable and is therefore not presented. As in Westgate [2], CIs use critical values based on a t-distribution with $N - p$ degrees of freedom. The working correlation structures for which results are presented are exchangeable, AR-1, and Toeplitz. In Table 1, we present results from when the Kauermann and Carroll [7] correction is used, whereas in Table 2 the Mancl and DeRouen [8] correction is utilized. Results are from settings in which the true structure is exchangeable with $\alpha_{exch} = 0.5$. Results for other true structures do not provide additional insight. We therefore include corresponding results from a true AR-1 structure with $\alpha_{AR-1} = 0.5$ in Supplementary Material. For the same reason, results from settings in which correlated outcomes were binary are also included in Supplementary Material.

In Tables 3 and 4, we present results from the use of unpenalized, based on $\hat{\Sigma}$ and thus unrealistically assuming correlation parameters are known, and penalized, based on $(\boldsymbol{I}_p + \hat{\boldsymbol{G}})\hat{\Sigma}(\boldsymbol{I}_p + \hat{\boldsymbol{G}})^T$, versions of the CIC and TECM to select either independence, exchangeable, AR-1, or Toeplitz working correlation matrices for normal outcomes. Corresponding results for binary outcomes are given in Supplementary Material. For each version of each criterion, we present the number of times each structure was selected, with the goal of selecting the true structure as often as possible. For each setting, the true structure is either exchangeable or AR-1 with true parameter value of 0.5, and $n = 4$. In Table 3, we present results based on the incorporation of the Kauermann and Carroll [7] correction, whereas in Table 4 the Mancl and DeRouen [8] correction is utilized.

3.2. Results

Numerical instability was encountered with the working Toeplitz structure when $N = 10$. In such instances, we do not present ESDs or empirical means of SE estimates because they were highly influenced. However, because stable results were observed for the analyses of most simulated datasets, empirical CPs are still presented. Furthermore, we include an additional setting with $N = 25$, as this is a small-sample setting in which results were stable.

For theoretical analyses, the ESD of $\hat{\beta}_1$ and corresponding empirical mean of SE_T were typically close in value when utilizing the Kauermann and Carroll [7] correction (Table 1). Furthermore, corresponding CPs were often relatively close to the nominal 0.95 value. Specifically, to determine if empirical CPs are acceptably close to 0.95, we note that empirical CPs between 0.936 and 0.964 have corresponding 95% CIs that cover 0.95. In short, these results suggest that if β did not have to be estimated for use in $\hat{\alpha}$ within Equation (1), then SE estimates obtained from $\hat{\Sigma}$ and utilizing the Kauermann and Carroll [7] correction would result in valid inference. Alternatively, the Mancl and DeRouen [8] correction (Table 2) sometimes yielded positive bias in SE estimates, particularly for small N, and thus resulted in over-coverage of CIs in such settings.

ESDs of $\hat{\beta}_1$ from the realistic analyses were greater than the corresponding ESDs from the theoretical analyses, demonstrating that variance inflation does occur when GEE incorporates structured working correlation matrices due to the need for replacing β with $\tilde{\beta}$ inside $\hat{\alpha}$ within Equation (1). However, empirical means for SE_T were approximately the same for both theoretical and realistic analyses in most settings. Therefore, when utilizing the Kauermann and Carroll [7] correction, SE_T was negatively biased, or smaller than the ESD, for the realistic analyses in some settings. This bias was also inherently observed via the degree of undercoverage by the CI that is constructed with SE_T in the realistic analyses. In contrast, use of the inflation correction worked very well at approximating the magnitude of variance inflation, and therefore improved inference overall when used in conjunction with the Kauermann and Carroll [7] correction. Specifically, utilizing SE_R notably reduced bias relative to SE_T, and typically resulted in near-nominal empirical CPs (Table 1). Alternatively, when utilizing the Mancl and DeRouen [8] correction (Table 2), use of the covariance inflation correction was often not needed due to the Mancl and DeRouen [8] correction resulting in a positively biased estimate for Σ. However, use of both corrections did perform best when incorporating a Toeplitz working structure.

The magnitude of covariance inflation that occurred, and therefore the need for the inflation correction, depended on n, N, and the number of estimated correlation parameters. For instance, a more notable variance inflation occurred when $n = 2$ (in which case the three working correlation structures are equivalent), particularly for $N = 10$, relative to the use of exchangeable or AR-1 when $n = 4$ because fewer empirical correlations were used to estimate the single correlation parameter. Furthermore, the magnitude of inflation increased as N decreased or the dimension of $\hat{\alpha}$ increased. For AR-1 and exchangeable, only one correlation parameter was estimated. Therefore, in settings in which $n = 4$, the covariance inflation was very small, especially when $N = 50$. Due to this result, SE_T and SE_R were similar, on average, in these settings. More notable inflations occurred with the Toeplitz structure due to the need to estimate three correlation parameters when $n = 4$. We further note that the number of parameters this structure estimates increases with n. Therefore, the need for the covariance inflation correction will be more apparent for larger values of n when utilizing this structure.

Although it is ideal to select the true structure, either exchangeable or AR-1, in Tables 3 and 4, unpenalized versions of the TECM and CIC using $\hat{\Sigma}$ selected Toeplitz more often. However, penalized versions using $(\boldsymbol{I}_p + \hat{\boldsymbol{G}})\hat{\Sigma}(\boldsymbol{I}_p + \hat{\boldsymbol{G}})^T$ appropriately took into account the degree of covariance inflation that occurs with each of these structures, and therefore the corresponding penalized versions of these criteria correctly chose the true, simpler structure much more frequently, greatly reducing the number of times Toeplitz was selected. We note that the selection accuracy of the penalized criteria enhanced as N increased, because $(\boldsymbol{I}_p + \hat{\boldsymbol{G}})\hat{\Sigma}(\boldsymbol{I}_p + \hat{\boldsymbol{G}})^T$ is estimated more precisely. Another interesting result is that selection frequencies were similar whether using the Kauermann and Carroll [7] or Mancl and DeRouen [8] correction. In short, the TECM and CIC were not notably influenced by this type of correction for $\hat{\Sigma}_{LZ}$, whereas use of the covariance inflation correction with all working structures under consideration greatly improved the performances of the TECM and CIC.

4. Concluding Remarks

With GEE, correlation parameters are estimated, therefore potentially inflating the covariance matrix of the regression parameter estimates. Westgate [2] derived an approximation for this inflation when utilizing an unstructured working correlation matrix, and Westgate [6] proposed the use of this approximation to penalize the estimation of the unstructured matrix's parameters. In this manuscript, we showed that the resulting corrected version of the well-known sandwich covariance formula and the use of this correction as a correlation selection penalty are also applicable when GEE incorporates structured working correlation matrices. In our study, use of the corrected formula improved standard error estimation, and thus the validity of inference, when the Kauermann and Carroll [7] correction was used. Alternatively, when the Mancl and DeRouen [8] correction was used, the covariance inflation correction appeared to be useful for attaining valid inference only when GEE incorporated a working Toeplitz structure, as the Mancl and DeRouen [8] correction often over-corrected for the bias in the Liang and Zeger [1] empirical sandwich estimator. Furthermore, irrelevent of which correction is applied to the Liang and Zeger [1] empirical sandwich estimator, use of the covariance inflation correction as a penalty greatly improved correlation structure selection accuracy.

Simulation results showed that, even for small N, the inflation of the variances of regression parameter estimates can be negligible for the AR-1 and exchangeable structures. Therefore, it is no surprise that, to our knowledge, this variance inflation has gone relatively unnoticed in practice with these working structures that require the estimation of only one nuisance parameter. This also implies that the inflation correction is often not needed to penalize these structures when compared against the working independence structure, which is a comparison that is routinely demonstrated in the GEE correlation selection literature. However, the need for the covariance inflation correction can be apparent for structures that require multiple nuisance parameters to be estimated. Another situation in which multiple correlation parameters may be estimated is when different trial arms, for instance, are allowed to have different exchangeable or AR-1 parameter values, in which case the covariance inflation correction can be useful. Furthermore, the need for the covariance inflation correction increases as the number of independent clusters decreases.

An alternative approach to GEE is the quadratic inference function (QIF) method [18]. Theoretically, the QIF approach is equally or more efficient than GEE. However, finite-sample covariance inflation of the regression parameter estimates must be taken into account, as is done in Westgate [19, 20]. Westgate [20] proposed a method that utilizes the TECM to select both a working correlation structure and one of these two methods, analogous to the approach we used in this manuscript. Therefore, our study results imply that the covariance inflation correction can also be used with GEE incorporating structured working correlation matrices in the context of Westgate [20].

An R function that implements GEE and outputs results based on the methods presented in this manuscript can be found in Supplementary Material or obtained by contacting the author.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

I would like to thank Mr. Woodrow W. Burchett for his input with respect to this manuscript and his assistance with the R function.

Funding

This publication was supported by the National Center for Research Resources and the National Center for Advancing Translational Sciences, National Institutes of Health, through Grant UL1TR000117. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

1. Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. Biometrika. 1986; 73:13–22.
2. Westgate PM. A bias correction for covariance estimators to improve inference with generalized estimating equations that use an unstructured correlation matrix. Statistics in Medicine. 2013; 32:2850–2858. [PubMed: 23255154]
3. Kackar AN, Harville DA. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. Journal of the American Statistical Association. 1984; 79:853–862.
4. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. 1997; 53:983–997. [PubMed: 9333350]
5. Kenward MG, Roger JH. An improved approximation to the precision of fixed effects from restricted maximum likelihood. Computational Statistics and Data Analysis. 2009; 53:2583–2595.
6. Westgate PM. Improving the correlation structure selection approach for generalized estimating equations and balanced longitudinal data. Statistics in Medicine. 2014; 33:2222–2237. [PubMed: 24504841]
7. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. Journal of the American Statistical Association. 2001; 96:1387–1396.
8. Mancl LA, DeRouen TA. A covariance estimator for gee with improved small-sample properties. Biometrics. 2001; 57:126–134. [PubMed: 11252587]
9. Wang YG, Carey V. Working correlation structure misspecification, estimation and covariate design: implications for generalised estimating equations performance. Biometrika. 2003; 90:29–41.
10. Hin LY, Wang YG. Working-correlation-structure identification in generalized estimating equations. Statistics in Medicine. 2009; 28:642–658. [PubMed: 19065625]

11. Hin LY, Carey VJ, Wang YG. Criteria for working-correlation-structure selection in gee. The American Statistician. 2007; 61:360–364.
12. R Development Core Team. R. Vienna, Austria: R Foundation for Statistical Computing; 2011. A language and environment for statistical computing. ISBN 3-900051-07-0; Available from: http://www.R-project.org/
13. Genz A, Bretz F, Miwa T, Mi X, Leisch F, Scheipl F, Hothorn T. mvtnorm: Multivariate normal and t distributions. 2013 r package version 0.9-9995; Available from: http://CRAN.R-project.org/package=mvtnorm.
14. Genz, A.; Bretz, F. Lecture Notes in Statistics. Heidelberg: Springer-Verlage; 2009. Computation of multivariate normal and t probabilities.
15. Leisch F, Weingessel A, Hornik K. bindata: Generation of artificial binary data. 2011 r package version 0.9-18; Available from: http://CRAN.R-project.org/package=bindata.
16. Shults J, Sun W, Tu X, Kim H, Amsterdam J, Hilbe JM, Ten-Have T. A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. Statistics in Medicine. 2009; 28:2338–2355. [PubMed: 19472307]
17. Prentice RL. Correlated binary regression with covariates specific to each binary observation. Biometrics. 1988; 44:1033–1048. [PubMed: 3233244]
18. Qu A, Lindsay BG, Li B. Improving generalised estimating equations using quadratic inference functions. Biometrika. 2000; 87:823–836.
19. Westgate PM. A bias-corrected covariance estimate for improved inference with quadratic inference functions. Statistics in Medicine. 2012; 31:4003–4022. [PubMed: 22807168]
20. Westgate PM. Criterion for the simultaneous selection of a working correlation structure and either generalized estimating equations or the quadratic inference function approach. Biometrical Journal. 2014; 56:461–476. [PubMed: 24431030]

Table 1

Empirical standard deviations (ESD) (in bold, underneath $\hat{\beta}_1$), empirical mean standard error (SE) estimates (underneath $\hat{\beta}_1$) and corresponding empirical 95% confidence interval coverage probabilities (CP) for both Theoretical and Realistic GEE Analyses. The Kauermann and Carroll [7] correction is utilized.

				Theoretical Analyses		Realistic Analyses	
Working Correlation	***N***	***n***	**SE Estimator**	$\hat{\beta}_1$	**CP**	$\hat{\beta}_1$	**CP**
			ESD	**0.713**		**0.892**	
	10	2	SE_T	0.675	0.940	0.704	0.901
			SE_R			0.870	0.942
Any							
			ESD	**0.305**		**0.309**	
	50	2	SE_T	0.300	0.944	0.300	0.942
			SE_R			0.307	0.947
			ESD	**0.434**		**0.441**	
	10	4	SE_T	0.425	0.952	0.425	0.948
			SE_R			0.438	0.954
Exchangeable							
			ESD	**0.191**		**0.192**	
	50	4	SE_T	0.193	0.953	0.193	0.953
			SE_R			0.194	0.955
			ESD	**0.459**		**0.472**	
	10	4	SE_T	0.444	0.951	0.443	0.946
			SE_R			0.466	0.952
AR-1							
			ESD	**0.201**		**0.202**	
	50	4	SE_T	0.206	0.955	0.206	0.953
			SE_R			0.207	0.955
			ESD	—		—	

Working Correlation	N	n	SE Estimator	Theoretical Analyses $\hat{\beta}_1$	Theoretical Analyses CP	Realistic Analyses $\hat{\beta}_1$	Realistic Analyses CP
	10	4	SE_T	—	0.946	—	0.906
			SE_R			—	0.940
			ESD	**0.287**		**0.303**	
Toeplitz	25	4	SE_T	0.274	0.945	0.275	0.925
			SE_R			0.299	0.942
			ESD	**0.190**		**0.196**	
	50	4	SE_T	0.192	0.951	0.192	0.943
			SE_R			0.198	0.953

N - number of independent subjects; n - number of repeated measurements per subject

SE_T; theoretical SE estimate obtained from $\hat{\Sigma}$ that assumes correlation parameters are known

SE_R, realistic SE estimate obtained from 9 $(\boldsymbol{I}_p + \hat{G})\hat{\Sigma}(\boldsymbol{I}_p + \hat{G})^T$

GEE-generalized estimating equations

Theoretical Analyses use $\hat{\alpha}(\beta)$ within Equation (1)

Realistic Analyses use $\hat{\alpha}(\tilde{\beta})$ within Equation (1)

Table 2

Empirical standard deviations (ESD) (in bold, underneath $\hat{\beta}_1$), empirical mean standard error (SE) estimates (underneath $\hat{\beta}_1$) and corresponding empirical 95% confidence interval coverage probabilities (CP) for both Theoretical and Realistic GEE Analyses. The Mancl and DeRouen [8] correction is utilized.

				Theoretical Analyses		Realistic Analyses	
Working Correlation	***N***	***n***	**SE Estimator**	$\hat{\beta}_1$	**CP**	$\hat{\beta}_1$	**CP**
Any	10	2	ESD	**0.713**		**0.892**	
			SE_T	0.796	0.963	0.885	0.938
			SE_R			1.838*	0.969
	50	2	ESD	**0.305**		**0.309**	
			SE_T	0.308	0.952	0.308	0.947
			SE_R			0.315	0.953
Exchangeable	10	4	ESD	**0.434**		**0.441**	
			SE_T	0.465	0.973	0.465	0.970
			SE_R			0.482	0.975
	50	4	ESD	**0.191**		**0.192**	
			SE_T	0.196	0.959	0.196	0.959
			SE_R			0.197	0.960
AR-1	10	4	ESD	**0.459**		**0.472**	
			SE_T	0.491	0.966	0.490	0.961
			SE_R			0.518	0.968
	50	4	ESD	**0.201**		**0.202**	
			SE_T	0.210	0.959	0.209	0.958
			SE_R			0.211	0.961
			ESD	—		—	

Working Correlation	N	n	SE Estimator	Theoretical Analyses $\hat{\beta}_1$	Theoretical Analyses CP	Realistic Analyses $\hat{\beta}_1$	Realistic Analyses CP
	10	4	SE_T	—	0.966	—	0.940
			SE_R			—	0.957
			ESD	**0.287**		**0.303**	
Toeplitz	25	4	SE_T	0.289	0.955	0.311	0.934
			SE_R			0.333	0.950
			ESD	**0.190**		**0.196**	
	50	4	SE_T	0.195	0.956	0.195	0.950
			SE_R			0.201	0.959

N - number of independent subjects; n - number of repeated measurements per subject

SE_T; theoretical SE estimate obtained from $\hat{\Sigma}$ that assumes correlation parameters are known

SE_R, realistic SE estimate obtained from 10 $(I_p + \hat{G})\hat{\Sigma}(I_p + \hat{G})^T$

GEE-generalized estimating equations

Theoretical Analyses use $\hat{\alpha}(\beta)$ within Equation (1)

Realistic Analyses use $\hat{\alpha}(\tilde{\beta})$ within Equation (1)

*Empirical mean influenced by outlying estimates

Table 3

Empirical frequencies of selecting each working correlation structure out of 1,000 replications from the use of the given correlation selection criterion. The Kauermann and Carroll [7] correction is utilized.

True Correlation	N	Criterion	Selection Frequencies: Ind	Exch	AR-1	Toeplitz
Exchangeable	25	$TECM_T$	18	376	113	493
		$TECM_R$	33	651	150	166
		CIC_T	16	285	95	604
		CIC_R	28	690	134	148
Exchangeable	50	$TECM_T$	0	393	62	545
		$TECM_R$	1	699	91	201
		CIC_T	0	277	52	671
		CIC_R	1	719	78	202
AR-1	25	$TECM_T$	15	97	315	573
		$TECM_R$	33	167	594	206
		CIC_T	12	65	277	646
		CIC_R	36	158	625	181
AR-1	50	$TECM_T$	1	44	331	624
		$TECM_R$	2	83	669	246
		CIC_T	0	21	256	723
		CIC_R	1	81	675	243

N - number of independent subjects

Ind - Independence; Exch - Exchangeable

TECM - 'trace of the empirical covariance matrix' criterion

CIC - 'correlation information criterion'

T - No penalty: The criterion is based on $\hat{\Sigma}$ that theoretically assumes correlation parameters are known

R - Penalty: The criterion is based on $(\boldsymbol{I}_p + \hat{G})\hat{\Sigma}(\boldsymbol{I}_p + \hat{G})^T$ that realistically accounts for, or penalizes, covariance inflation due to correlation parameter estimation

Table 4

Empirical frequencies of selecting each working correlation structure out of 1,000 replications from the use of the given correlation selection criterion. The Mancl and DeRouen [8] correction is utilized.

True Correlation	N	Criterion	Selection Frequencies: Ind	Exch	AR-1	Toeplitz
Exchangeable	25	$TECM_T$	16	387	109	488
		$TECM_R$	31	669	141	159
		CIC_T	15	299	90	596
		CIC_R	24	705	126	145
Exchangeable	50	$TECM_T$	0	405	57	538
		$TECM_R$	1	706	90	203
		CIC_T	0	285	48	667
		CIC_R	1	730	74	195
AR-1	25	$TECM_T$	17	115	328	540
		$TECM_R$	33	184	599	184
		CIC_T	14	74	287	625
		CIC_R	34	168	628	170
AR-1	50	$TECM_T$	1	48	343	608
		$TECM_R$	2	86	670	242
		CIC_T	0	24	274	702
		CIC_R	1	86	678	235

N - number of independent subjects

Ind - Independence; Exch - Exchangeable

TECM - 'trace of the empirical covariance matrix' criterion

CIC - 'correlation information criterion'

T - No penalty: The criterion is based on $\hat{\Sigma}$ that theoretically assumes correlation parameters are known

R - Penalty: The criterion is based on $(\boldsymbol{I}_p + \hat{\boldsymbol{G}})\hat{\Sigma}(\boldsymbol{I}_p + \hat{\boldsymbol{G}})^T$ that realistically accounts for, or penalizes, covariance inflation due to correlation parameter estimation