University of Pennsylvania
UPenn Biostatistics Working Papers

Year 2006 *Paper* 4

Use of Unbiased Estimating Equations to Estimate Correlation in Generalized Estimating Equation Analysis of Longitudinal Trials

Wenguang Sun* Justine Shults† Mary Leonard‡

*

†University of Pennsylvania, jshults@cceb.med.upenn.edu

‡University of Pennsylvania, leonard@email.chop.edu

Use of Unbiased Estimating Equations to Estimate Correlation in Generalized Estimating Equation Analysis of Longitudinal Trials

Wenguang Sun, Justine Shults, and Mary Leonard

Abstract

In a recent publication, Wang and Carey (Journal of the American Statistical Association, 99, pp. 845-853, 2004) presented a new approach for estimation of the correlation parameters in the framework of generalized estimating equations (GEE). They considered correlated continuous, binary and count data with a generalized Markov correlation structure that includes the first-order autoregressive AR(1) and Markov structures as special cases. They made detailed comparisons with pseudo-likelihood (PL) and the first stage of quasi-least squares (QLS), a two-stage approach in the framework of generalized estimating equations (GEE). In this note we extend their comparisons for the second (bias corrected) stage of QLS. We comment on their earlier findings, which were overwhelmingly in favor of the Wang-Carey (WC) approach relative to stage one of QLS. We prove that WC and QLS are identical for equally spaced data with an AR(1) structure. Furthermore, we demonstrate via simulations that neither QLS, PL or WC is uniformly superior for unequally spaced data with a Markov structure. We give general recommendations regarding the relative merits of each approach for analysis of unbalanced and unequally spaced longitudinal data and demonstrate their application in an analysis of a longitudinal study of obesity following renal transplantation in children.

Use of unbiased estimating equations to estimate correlation in generalized estimating equation analysis of longitudinal trials

Wenguang Sun[1]

Justine Shults*[2]

Mary Leonard[3]

Department of Biostatistics and Epidemiology[1,2,3], **Center for Clinical Epidemiology and Biostatistics**[1,2,3], and **Department of Pediatrics**[3], **University of Pennsylvania School of Medicine, Philadelphia, PA 19034, U.S.A.**

SUMMARY.

In a recent publication, Wang and Carey (*Journal of the American Statistical Association*, **99**, pp. 845-853, 2004) presented a new approach for estimation of the correlation parameters in the framework of generalized estimating equations (GEE). They considered correlated continuous, binary, and count data with a generalized Markov correlation structure that includes the first-order autoregressive AR(1) and Markov structures as special cases. They made detailed comparisons with pseudo-likelihood (PL) and the first stage of quasi-least squares (QLS), a two-stage approach in the framework of generalized estimating equations (GEE). In this note we extend their comparisons

[0] **Corresponding author's email address:* jshults@cceb.upenn.edu

Key words: Generalized Estimating Equations; Longitudinal Data; Pseudo-Likelihood; Quasi-Least Squares; Unbiased Estimating Equations

for the second (bias corrected) stage of QLS. We comment on their earlier findings, which were overwhelmingly in favor of the Wang-Carey (WC) approach relative to stage one of QLS. We prove that WC and QLS are identical for equally spaced data with an AR(1) structure. Furthermore, we demonstrate via simulations that neither QLS, PL, or WC is uniformly superior for unequally spaced data with a Markov structure. We give general recommendations regarding the relative merits of each approach for analysis of unbalanced and unequally spaced longitudinal data and demonstrate their application in an analysis of a longitudinal study of obesity following renal transplantation in children.

1. Introduction

In longitudinal analyses with generalized estimating equations (GEE) (Liang and Zeger, 1986), interest often focuses on the dependence of the outcome variable on covariates, while the correlation among repeated measurements per subject is termed a "nuisance" that is of secondary interest. However, careful estimation of the association among the repeated measurements per subject in longitudinal trials can be helpful, e.g. in avoiding the breakdown in iterative procedures such as GEE that occurs when the estimated correlation matrices are not positive definite (Shults and Chaganty, 1998); improving efficiency in estimation of the regression parameters β by proper choice of covariance structure (Sutradhar and Das, 2000 and Wang and Carey, 2003); and in potentially enhancing scientific understanding. For example, if a longitudinal educational intervention is successful, this might result in more positive responses over time in the intervention subjects, which could be reflected in higher intra-subject associations due to the greater similarity in

their responses. Assessment of the intra-subject correlation could therefore provide additional insight into the effectiveness of the intervention.

In a recent manuscript, Wang and Carey (2004) present a new approach for estimation of the correlation parameters, that we refer to as Wang-Carey (WC) estimation. They made comparisons with the well-established method of pseudo-likelihood (PL) (Carroll and Ruppert, 1988; Davidian and Giltinan, 1995) and the newer approach of quasi-least squares (QLS) that is based on generalized estimating equations (GEE) (Liang and Zeger, 1986). They demonstrated that QLS is inferior to PL and WC estimation both in terms of efficiency and bias. However, QLS is a two-stage procedure and they only made comparisons with the first stage (Chaganty, 1997; Shults and Chaganty, 1998) that was available when they began their research. In this note, we extend their comparisons for stage two (bias corrected) QLS (Chaganty and Shults, 1999). We also present comparisons for a wider range of the correlation parameter α, e.g., they only considered $\alpha \in \{0.05, 0.25\}$ for binary data. In addition, we consider a higher degree of variability in the temporal spacing of measurements. Our goal is to provide information that should prove useful to the statistician who is seeking an appropriate method for estimation of the correlation parameters in the framework of GEE. Our note is organized as follows: We give a brief description of each approach (Section 2); compare the methods for a Markov correlation structure (Section 3); apply the methods in an analysis of a renal study in children (Section 4); and make some concluding remarks (Section 5).

2. Three Approaches for Estimation of the Correlation Parameter

2.1 *Notation and Overview*

We assume the usual set-up for longitudinal analyses with GEE. Measurements $Y_i = (y_{i1}, \cdots, y_{in_i})'$ and associated covariates $x'_{ij} = (x_{ij1}, \cdots, x_{ijp})$ are collected on subject i at times $T_i = (t_{i1}, \cdots, t_{in_i})'$, for $i = 1, \cdots, m$. When $n_i = n \ \forall \ i$ and $|t_{ij} - t_{ik}| = \gamma \ \forall \ i, \ j, \ k$, we refer to the data as balanced and equally spaced, respectively. The expected value and variance of measurement y_{ij} on subject i can be expressed as $E(y_{ij}) = g^{-1}(x'_{ij}\beta) = u_{ij}$ and $Var(y_{ij}) = \phi h(u_{ij})$, respectively, where ϕ is a known or unknown scale parameter. Observations on different subjects are independent. Within subjects, they are correlated, with a pattern of association described by the working correlation structure for observations on subject i, $Corr(Y_i) = W_i(\alpha)$, that depends on correlation parameter α. We assume that the working structure is correctly specified. The covariance matrix of Y_i is then given by $Cov(Y_i) = \phi {A_i}^{(1/2)} W_i(\alpha) {A_i}^{(1/2)}$, where $A_i = diag(h(u_{i1}), \ldots, h(u_{in_i}))$.

WC, PL, and QLS can be considered approaches in the framework of GEE because they alternate until convergence between (i) updating their estimate of β by solving the GEE estimating equation (Liang and Zeger, 1986) for β at the current estimate of α and (ii) updating their estimate of α with a consistent estimate $\hat{\alpha}$. However, while GEE typically uses moment estimates for α, WC and QLS estimate α by solving an unbiased estimating equation evaluated at the current estimate of β. We provide each method's estimating equation for α here.

2.2 *Pseudo-Likelihood Estimating Equation for* α

PL obtains an updated estimate $\hat{\alpha}_{PL}$ by solving the following estimating equation for α:

$$\frac{\partial}{\partial\alpha}\left\{\sum_{i=1}^{m} Z_i'(\beta)\left\{W_i^{-1}(\alpha)\right\}Z_i(\beta)\right\} - \phi\sum_{i=1}^{m} trace\left\{\frac{\partial W_i^{-1}(\alpha)}{\partial\alpha}W_i(\alpha)\right\} = 0, \quad (1)$$

where $U_i = E(Y_i)$ and $Z_i(\beta) = A_i^{-1/2}(Y_i - U_i)$ is the vector of Pearson residuals on subject i. We refer to the first term on the left-hand side of (1) as the *derivative* (with respect to α) of the *generalized error sum of squares*, or D_G. The second term is easily shown to equal the expectation $E(D_G)$ of D_G, so that this etimating equation is unbiased for α. Because (1) involves ϕ, PL requires updating the estimate of this parameter within each iteration. We applied the moment estimate $\hat{\phi} = (1/N)\sum_{i=1}^{m} Z_i(\hat{\beta})'Z_i(\hat{\beta})$ that is very similar to that suggested in Liang and Zeger (1986), where $N = \sum_{i=1}^{m} n_i$.

2.3 *Quasi-Least Squares Estimating Equation for* α

QLS is a two stage procedure that obtains a solution to unbiased estimating equation (1) by equating each of its two terms with zero. Stage one (Chaganty, 1997 for $n_i = n$; Shults and Chaganty, 1998 for $n_i \neq n$) obtains the estimate $\hat{\alpha}_{QONE}$ as the solution to the following equation for α:

$$D_G = \frac{\partial}{\partial\alpha}\left\{\sum_{i=1}^{m} Z_i'(\beta)\left\{W_i^{-1}(\alpha)\right\}Z_i(\beta)\right\} = 0. \quad (2)$$

Because (2) is biased the stage two estimate $\hat{\alpha}_{QLS}$ is obtained as the solution to the following estimating equation for α:

$$\sum_{i=1}^{m} trace\left\{\frac{\partial W_i^{-1}(\delta)}{\partial\delta}W_i(\alpha)\right\}\Bigg|_{\delta=\hat{\alpha}_{QONE}} = 0. \quad (3)$$

2.4 *Wang-Carey Estimating Equation for α*

Wang and Carey (2004) worked with the Cholesky decomposition $W_i^{-1}(\alpha) = B_i^\tau J_i B_i$, where B_i is an upper (or lower) triangular matrix and J_i is a diagonal matrix. They expressed the derivative of the generalized error of sum squares, D_G, as the sum of two terms. They proved that one of the terms will always have expectation zero and they equated this term with zero, to obtain the following unbiased estimating equation for α:

$$\sum_{i=1}^{m} Z_i^\tau (\partial B_i/\partial\alpha)^\tau J_i B_i Z_i = 0. \tag{4}$$

Because B_i can be taken as an upper or a lower triangular matrix, the estimating function in (4) can also be derived as the sum of two estimating functions; we implement this in Section 3.

2.5 *Relationship Between the Three Approaches*

As shown in the previous sections, QLS, WC, and PL can be viewed as methods that work with the function D_G to obtain bias-corrected estimating functions for construction of unbiased estimating equations for α. PL removes the bias by subtracting the expectation of D_G from itself and setting $D_G - E(D_G) = 0$. QLS first sets $D_G = 0$ and solves for α to obtain the stage one estimate. It then equates the expectation of D_G with 0 and solves for α to obtain the stage two estimate. QLS is therefore also an unbiased estimating procedure for estimating α by combining two stages. WC proceeds by first decomposing D_G as the sum of two terms, removing the biased term from D_G and then using the unbiased term as the estimating function for α. One important difference between these approaches is that PL requires estimation of ϕ within each iteration, while WC and QLS do not.

3. Comparison of WC, QLS, and PL for the Markov Correlation Structure

3.1 *Set-up and Results of Simulations*

We will consider the Markov structure, for which $\text{Corr}(y_{ij}, y_{ik}) = \alpha^{|t_{ij}-t_{ik}|}$. This structure is useful for longitudinal studies because we often expect that measurements on a subject will be more similar (and thus more highly correlated) if they are measured more closely in time. It also includes the AR(1) structure as a special case, when $T_i = (1, 2, \ldots, n_i)$ for each subject i, i.e. when the measurements are equally spaced in time.

In Appendix A we prove that WC and QLS are identical for the AR(1) structure. However, as demonstrated in our simulations, they are not identical for the Markov structure. We therefore compared WC, QLS, and PL for this structure, in simulation studies that are similar to those in Wang and Carey (2004) but that consider a wider range of values of α. As in Wang and Carey (2004) we focus on estimation of α, because the methods differ with respect to this parameter and because improved estimation of α should result in improved estimation of β. In Appendix B we describe the approaches used to simulate correlated normal, Poisson (with and without overdispersion), and binary data. This includes a proof that the approach for Poisson data described in Wang and Carey (2004), modified to also include a "burn-in" period, will asymptotically yield data with an AR(1) structure and overdispersion. Table 1 displays the mean square error (MSE $= 1/500 \sum_{r=1}^{500} (\hat{\alpha}_r - \alpha_r)^2$) and bias (BIAS $= 1/500 \sum_{r=1}^{500} (\hat{\alpha}_r - \alpha_r)$) based on 500 simulation runs for stage one of QLS (QONE), QLS, WC, and PL with β treated as known. We conducted simulations for several different sample sizes

(m), number of measurements per subject (n) (prior to randomly dropping measurements), and true values of α. For brevity, a subset of the results are shown in Table 1; the complete tables (with results similar to those displayed here) are available on request.

[Table 1 about here.]

To compare the MSE between two approaches for a greater range of values of α, we also ran additional simulations. Figures 1, 2, and 3 display the ratios $MSE(WC)/MSE(QLS)$, $MSE(PL)/MSE(QLS)$, and $MSE(WC)/MSE(PL)$ versus α, respectively, for normal, binary, and Poisson (with overdispersion) data. For these graphs, $m = 30$; $n = 5$; and α increases in value from 0.1 to 0.9 by 0.02 in each step.

[Figure 1 about here.]

[Figure 2 about here.]

[Figure 3 about here.]

The simulation results suggest the following: *QLS is clearly superior to QONE* in terms of reduced MSE and bias (Table 1), especially for larger values of α; we comment on this in Section 3.2. *For comparison of QLS versus WC* (Figure 1), QLS outperforms WC for smaller values of α, but tends to have larger MSE for higher values. However, aside from normal data with smaller α, the two approaches are similar. The bias is similar (Table 1), but tends to be slightly smaller for WC. *For comparison of QLS versus PL* (Figure 2), QLS has greater MSE for normal data for most values of α. It is

similar to PL for Poisson and binary data for smaller values of α ($\alpha < 0.5$ for binary, $\alpha < 0.4$ for Poisson), but has greater MSE for higher values. The greatest loss in efficiency for use of QLS versus PL occurs for Poisson data with larger values of α. The bias (Table 1) is similar for both approaches, but slightly smaller for PL, overall. *For comparison of PL versus WC* (Figure 3) for normal data, PL outperforms WC for most values of α, with the greatest difference occuring for smaller values of α. The two methods were very similar for binary data with $\alpha < 0.70$; for $\alpha > 0.7$ WC outperformed PL. They were also similar for Poisson data with $\alpha < 0.40$; for $\alpha > 0.40$, PL outperformed WC. With regard to bias (Table 1) WC and PL were very similar. Overall, results for Poisson data without overdispersion (Table 1) were very similar to those for Poisson data with overdispersion (Figures 1-3), although the range of values for α was more restricted. When we increased the sample size (both in number of subjects and number of visits per subject) or reduced the degree of imbalance in the data (so that the true structure became closer to an AR(1) structure), differences between the methods became less pronounced and the graphs (Figures 1-3) became "smoother". (In our simulations the temporal spacing and number of measurements per subject varied greatly between simulation runs, especially for normal data, which resulted in greater variability in the MSE and a jagged appearance for the figures.)

3.2 *Recommendations for Analysis of Unbalanced and Unequally Spaced Longitudinal Data*

Our simulation studies suggested the following. For the analyst who is choosing between QLS, WC, or PL for estimation of α in the framework of GEE, we suggest application of PL if there is any suspicion of overdispersion

in the data, i.e. if $\phi \neq 1$. In this situation (and as demonstrated in our simulations) we might expect PL to outperform the other approaches because it incorporates estimation of ϕ into each iteration, while QLS and WC ignore this parameter. Also, if the data do not deviate seriously from multivariate normality, then we would also recommend application of PL because PL is very similar to the maximum likelihood (ML) approach for normal data. We also suggest application of PL for highly correlated Poisson data (with or without overdispersion). However, for highly correlated binary data, WC is the superior approach. For binary or Poisson data with very small correlations, QLS is slightly preferable to WC and PL (which are similar). In general, if one method has been selected as most appropriate for estimation of α, it would be beneficial to implement the other approaches as well, to assess the sensitivity of the results to the method of estimation.

3.3 *Comments on Earlier Findings*

Wang and Carey (2004) concluded that stage one of QLS is not appropriate because WC is more reliable in terms of bias and MSE. However, our simulations demonstrated that stage two of QLS offers a great improvement over stage one (QONE) in terms of reduced MSE for higher values of α. Here we give an intuitive explanation for this reduction. By studying the decomposition formula for the MSE of an estimator

$$\text{MSE}(\hat{\alpha}) = [\text{Bias}(\hat{\alpha})]^2 + \text{Var}(\hat{\alpha}), \tag{5}$$

we see that the MSE of $\hat{\alpha}_{QONE}$ can be reduced by use of $\hat{\alpha}_{QLS}$ in two ways. First, $\hat{\alpha}_{QONE}$ is biased, while $\hat{\alpha}_{QLS}$ is consistent. As a result, the first term in the MSE is reduced by using $\hat{\alpha}_{QLS}$ instead of $\hat{\alpha}_{QONE}$. Second, the variance of $\hat{\alpha}_{QLS}$ might be reduced by the function f that relates the stage one and

stage two estimate. For example, when the working correlation structure is AR(1), the function f is given by

$$\hat{\alpha}_{QLS} = f(\hat{\alpha}_{QONE}) = 2\hat{\alpha}_{QONE}/(1 + \hat{\alpha}^2_{QONE}). \quad (6)$$

According to the Delta method,

$$\begin{aligned} \text{Var}(\hat{\alpha}_{QLS}) &= (\frac{\partial f}{\partial \alpha})\text{Var}(\hat{\alpha}_{QONE})(\frac{\partial f}{\partial \alpha}), \text{ where} \\ \frac{\partial f}{\partial \alpha} &= 2(1-\alpha^2)/(1+\alpha^2). \end{aligned}$$

The second term of the MSE is therefore also reduced when $|\alpha| > 0.58$; e.g. when $\alpha = 0.75$, $\text{Var}(\hat{\alpha}_{QLS}) \approx 0.31\text{Var}(\hat{\alpha}_{QONE})$. (However, we note that although stage two offered a substantial improvement over stage one, WC still had smaller MSE than QLS for larger α, for unequally spaced data with a Markov structure.)

4. Analysis of Obesity Following Renal Transplant in Children

In a study conducted at the Children Hospital of Philadelphia, body mass index (BMI) and related variables were measured on 100 children following a kidney transplant. Between 2 and 11 measurements (mean = 5.9) were taken on each patient, who had from 0.25 to 8 years of follow-up (mean = 0.25 years). The primary goal of this analysis was to describe the change in likelihood of obesity (OBESE = 1 if BMI z-score exceeds the 95^{th} percentile for a subject's age and height; is 0 otherwise) following transplant and to assess potential correlates of obesity that included age in years at transplant (AgeTrans), baseline measure of BMI z-score (BaseBMIZ), and African-American ethnicity (AAEthnicity = 1 for African-Americans; is 0 otherwise). The full analysis of this study that will consider additional covariates will appear elsewhere.

To relate the likelihood of obesity with covariates, we fit the regression model $E(Obese) = g^{-1}(\beta_0+\beta_1\text{Time}+\beta_2\text{Time}^2+\beta_3\text{BaseBMIZ}+\beta_4\text{AAEthnicity}+\beta_5\text{AgeTrans})$ with a logistic link function $g^{-1}(\cdot)$. To account for the highly unequal spacing of the data (mean time-lag = 0.59 years, range = (0.017, 3)), we implemented a Markov working correlation structure to describe the pattern of association among the repeated measurements on each subject.

Table 2 displays the results of the analysis for WC, QLS, and PL, with standard errors obtained using the robust sandwich variance estimator (Liang-Zeger,1986) of the variance-covariance matrix of $\hat{\beta}$. The estimates were similar for the three approaches, although there were some interesting differences. After transplant children are given massive doses of steroids, which usually results in marked weight gain. As time progresses, fewer steroids are given and as a result, it is anticipated that the child's weight will return to normal (or at least to a weight that is closer to their pre-transplant weight). This pattern was displayed in this study, e.g. a plot of BMI versus time (not shown) revealed an initial increase in weight during the first months post-transplant, followed by a slow and steady decline in weight. This resulted in an overall frequency of obesity that initially increased following transplant, and then decreased over time. This is also reflected in the signs of the coefficients for Time and Time^2, which were positive and negative, respectively, for all approaches. However, the coefficient for Time^2 only differed significantly from zero for WC. Ethnicity was not significantly associated with obesity for any approach, but the regression coefficient (which was negative for all three approaches) was closer to zero for the WC approach. A smaller value might be anticipated based on prior knowledge that an increased likelihood

of obesity, as opposed to the decrease suggested by negative coefficients, is typically associated with African-American ethnicity. All three approaches, as expected, identified baseline BMI z-score as a significant correlate of current obesity status; all approaches failed to identify a significant association with age at transplantation. The estimated correlations were relatively high for each approach, but largest for WC ($\hat{\alpha}_{QLS}$=0.5456; $\hat{\alpha}_{WC}$=0.7729; and $\hat{\alpha}_{PL}$=0.5868).

In summary, this analysis suggests that subjects have an increased likelihood of obesity following kidney transplantation and that African-American ethnicity and age at transplant are not associated with increased likelihood of obesity. However, only WC identified a significant subsequent leveling off (or decline) in the likelihood of obesity following the initial weight gain, that was expected based on our knowledge of steroids and also graphical displays of these data. This suggests that WC was more sensitive to the apparent time course of weight gain in this study, although the results for all three approaches were similar.

[Table 2 about here.]

5. Discussion

In this note we compared three approaches that provide estimate α by solving unbiased estimating equations in the framework of GEE. We proved that two of the methods (WC and QLS) are identical for an AR(1) structure and we gave recommendations based on simulations regarding selection of an approach for analysis of unbalanced and unequally spaced discrete or continuous longitudinal data. We believe that further simulation studies are needed to definitively declare the superiority of one particular approach; e.g., it would

be of interest to compare the methods with regard to estimation of β for different study designs. It is also important to note that ease of application is an important consideration. For example, WC is based on the factorization $B_i^\tau J_i B_i$ of the inverse of working correlation structure. This decomposition required non-trivial calculations for the Markov structure (Wang and Carey, 2004), which has a relatively simple tri-diagonal structure for its inverse. For other structures, it might not be straightforward to implement their approach. In addition, for complex structures, PL, which requires directly solving the estimating equation (1), might be more difficult to implement than QLS, whose estimating equations do not involve the scalar parameter ϕ. As a result, QLS might have the broadest range of application for correlation structures that have not yet been applied in the framework of GEE. Implementing new correlation structures and developing approaches for choosing between them will be of continued interest to these authors. Other potential areas of interest include comparison of the methods under model misspecification and with other approaches that are available for analysis of correlated discrete and continuous data.

ACKNOWLEDGEMENTS

This work was supported by a grant from the National Institutes of Health (R01-CA096885). We are grateful to Dr. Vincent Carey and Dr. You-Gon Wang for useful discussions.

References

Carroll, R.J. and Ruppert, D. (1998). *Transformations and Weighting in Regression*, Chapman and Hall, New York

Chaganty, N. R. (1997). An alternative approach to the analysis of longitudinal data via generalized estimating equations, *Journal of Statistical Planning and Inference* **63,** 39-54.

Chaganty, N.R. and Shults, J. (1999). On eliminating the asymptotic bias in the quasi-least squares estimate of the correlation parameter, *Journal of Statistical Planning and Inference* **76,** 127-144.

Davidian, M. and Giltinan, D.M. (1995). *Nonlinear Models for Repeated Measures Data*, Chapman and Hall, New York

Emrich, L.J. and Piedmonte, M.R.. (1991) A method for generating high-dimensional multivariate binary variables, *American Statistician* **45,** 302-304.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73,** 13-22.

Prentice, R.L. (1988). Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44,** 1033-1048.

Shults, J. and Chaganty, N.R. (1998). Analysis of serially correlated data using quasi-least squares, *Biometrics* **54,** 1622-1630.

Sim, C.H. (1993). Generation of Poisson and gamma random vectors with given marginals and covariance matrix, *Journal of Statistical Computation and Simulation* **47,** 1-10.

Sutradhar, B.C. and Das, K. (2000). On the accuracy of efficiency of estimating equation approach, *Biometrics* **56,** 622-625.

Wang, Y.G. and Carey, V.J. (2003). Working correlation misspecification,

estimation and covariate design: implications for generalized estimating equation performance, *Biometrika* **90,** 29-41.

Wang, Y.G. and Carey, V.J. (2004). Unbiased estimating equations from working correlation models for irregularly timed repeated measures, *Journal of the American Statistical Association* **99,** 845-852.

Appendix A

Proof that WC and QLS are Identical for the AR(1) Structure

We consider unbalanced data, so that the dimension of all matrices in this Appendix is $n_i \times n_i$. The AR(1) structure $R_i(\alpha) = (\alpha^{|j-k|})$ has inverse

$$R_i^{-1}(\alpha) = \frac{1}{1+\alpha^2}\{I_{n_i} + \alpha^2 C_{2i} - \alpha C_{1i}\},$$

where C_{1i}, and C_{2i}s are matrices that can be expressed as

$$C_{1i} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix},$$

and $C_{2i} = \text{diag}\{0, 1, 1, \cdots, 1, 0\}$.

For QLS and known β, the stage one estimating equation (2) can be expressed as:

$$a_m^* \alpha^2 - 2b_m^* \alpha + a_m^* = 0, \tag{A.1}$$

where $a_m^* = \sum_i Z_i' C_{1i} Z_i$ and $b_m^* = \sum_i Z_i'(C_{2i}+I_{n_i})Z_i$. The stage two estimate (Chaganty and Shults, 1999) is then given by $\hat{\alpha}_{QLS} = 2\hat{\alpha}_{QONE}/(1+\hat{\alpha}_{QONE}^2)$ which can be expressed in terms of a_m^* and b_m^* as $\hat{\alpha}_{QLS} = a_m^*/b_m^*$.

For WC, as shown in Wang and Carey (2004), we can express the Cholesky decomposition of the inverse of R_i as:

$$W_i^{-1}(\alpha) = B_i' J_i B_i = B_i J_i^* B_i',$$

where

$$B_i = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ -\alpha & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\alpha & 1 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & -\alpha & 1 \end{pmatrix},$$

$J_i(\alpha) = \text{diag}\{1, \frac{1}{1-\alpha^2}, \cdots, \frac{1}{1-\alpha^2}\}$, and $J_i^*(\alpha) = \text{diag}\{\frac{1}{1-\alpha^2}, \cdots, \frac{1}{1-\alpha^2}, 1\}$.

The WC estimating equation (4) can then be expressed as:

$$\sum_{i=1}^{m} Z_i^\tau (\partial B_i/\partial\alpha)^\tau J_i B_i Z_i + \sum_{i=1}^{m} Z_i^\tau (\partial B_i/\partial\alpha) J_i^* B_i^\tau Z_i = 0, \qquad \text{(A.2)}$$

which simplifies to

$$\frac{1}{1-\alpha^2} \sum_i Z_i'(-C_{1i} + \alpha(I_{n_i} + C_{2i}))Z_i = 0. \qquad \text{(A.3)}$$

The solution $\hat{\alpha}_{WC}$ to (A.3) is given by $\hat{\alpha}_{QLS} = a_m^*/b_m^*$, which is identical to $\hat{\alpha}_{QLS}$. This completes the proof.

Appendix B

Approaches Used for Simulating Correlated Normal, Poisson, and Binary Data

To simulate data with a Markov structure, we directly simulated the normal data. For Poisson and binary data, we first simulated data with an AR(1) structure and then randomly dropped 50 percent of the measurements, in order to introduce imbalance and yield data with a Markov structure. Each approach is (briefly) described below.

Multivariate Normal Data: We simulated unequally spaced data for each subject by first simulating their timings from a uniform $(k-1, k)$ distribution, for $k = 1, 2, \ldots, n$. We then used the Matlab function `'mvnrnd'` to simulate n observations from a normal distribution with a Markov correlation structure with the specified timings and means $u_{ij} = 1 + I(j > n/2)$. To introduce imbalance, we randomly dropped 20 percent of observations, after simulating the complete data set for m subjects.

Correlated Binary Data:

We used the alorgithm of Emrich and Piedmonte (1991) to simulate correlated binary outcomes with an AR(1) correlation structure. This involved using bisection to solve equation (2.1) in Emrich and Piedmonete, with $\delta_{jk} = \alpha^{|j-k|}$. For the marginal probabilities chosen by Wang and Carey (2004) in their Table 5, the correlation parameters are bounded by a short interval (see Prentice, 1988). To allow for comparison over a broader interval, we assumed that all p_js in (2.1) were 0.5; α could then take value in the interval $(-1, 1)$.

Correlated Poisson Data with Over-Dispersion:

We implemented the following approach described in Wang and Carey (2004), but modified to include a "burn-in" period: First simulate an observation o_1 from a Poisson distribution with parameter λ_1, $Poi(\lambda_1)$. Next, for $s = 2, \ldots, N+n$, simulate observations o_s from $Poi(\lambda_s + \alpha\frac{\lambda_s}{\lambda_{s-1}}(y_{s-1} - \lambda_{s-1}))$, where N is a relatively large number, e.g. 50 or 100. The last n observations in this series will represent measurements $y_{i1}, \ldots, y_{in}$ for subject i. To ensure the correlation structure of measurements is AR(1), the λ_is need to be chosen appropriately; we chose all λ_is equal to 5, as suggested by Wang and Carey

(2005, personal correspondence).

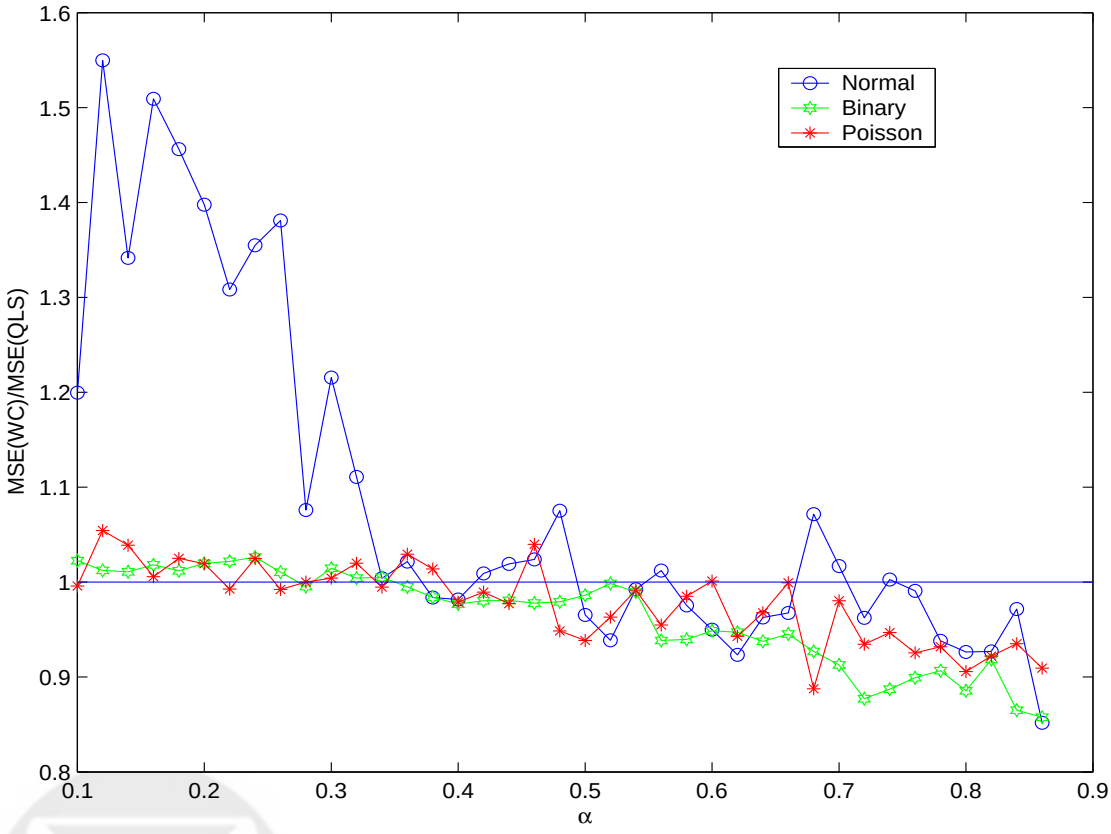
To prove that this approach yields a correlation structure that is asymptotically AR(1) (as $N \to \infty$) for identical λs, we first notice that $\mathrm{Var}(o_s) = \lambda_s + \alpha^2(\frac{\lambda_s}{\lambda_{s-1}})^2\mathrm{Var}(o_{s-1}) = \lambda_s + \alpha^2\mathrm{Var}(o_{s-1})$. So if we start from $s = 1$, we have: $Var(o_1) = \lambda_1$, $Var(o_2) = \lambda_2 + \alpha^2\lambda_1 = \lambda_1(1 + \alpha^2)$, $\cdots\cdots$, and finally, $Var(o_n) = \lambda_1\frac{1-\alpha^{2n}}{1-\alpha^2}$. Also, we have: $Cov(o_n, o_{n+k}) = Cov(o_n, E(o_{n+k}|o_n)) = \alpha^k\lambda_1\frac{1-\alpha^{2n}}{1-\alpha^2}$, so that $lim_{n\to\infty}Corr(o_n, o_{n+k}) = lim_{n\to\infty}\frac{\alpha^k\lambda_1\frac{1-\alpha^{2n}}{1-\alpha^2}}{\sqrt{\lambda_1\frac{1-\alpha^{2n}}{1-\alpha^2}}\sqrt{\lambda_1\frac{1-\alpha^{2n+2k}}{1-\alpha^2}}}$ $= \alpha^k$. The data therefore do have an AR(1) structure asymptotically. Also, after a certain number of steps, the variance tends to be stable and varies around a constant, which is asymptotically given by $Var(o_s) \simeq \frac{\lambda_1}{1-\alpha^2} > \lambda_1 =$ M$ean(o_s)$. The simulated Poisson data are therefore over-dispersed.

Correlated Poisson Data without Over-Dispersion:

We applied Sim's (1993) method that can be used to simulate data with a fixed covariance matrix and positive entries. Because this method has constraints that need to be satisfied to avoid failure in the data generating procedure, we considered α in the range of 0.1-0.5 (Table 1).

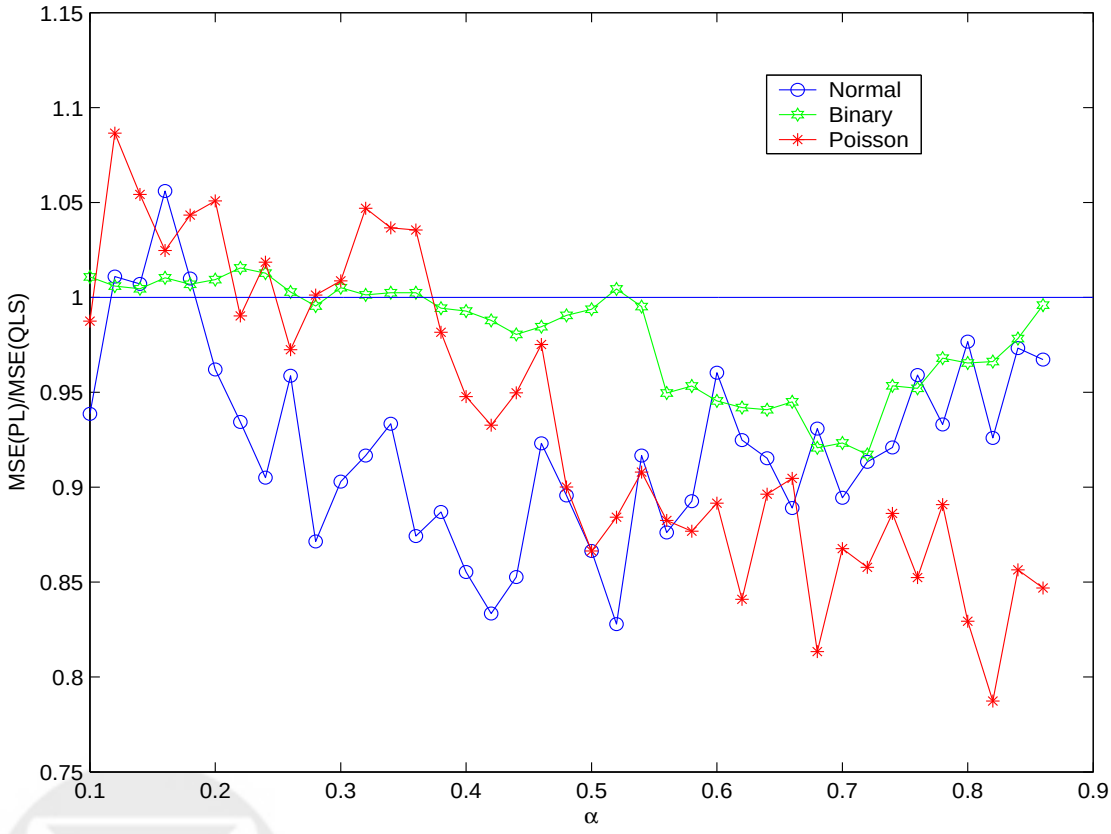
[h]

Figure 1. Small sample efficiency comparisons: $MSE(WC)/MSE(QLS)$ vs. α



[h]

Figure 2. Small sample efficiency comparisons: $MSE(PL)/MSE(QLS)$ vs. α



[h]

Figure 3. Small sample efficiency comparisons: $MSE(WC)/MSE(PL)$ vs. α

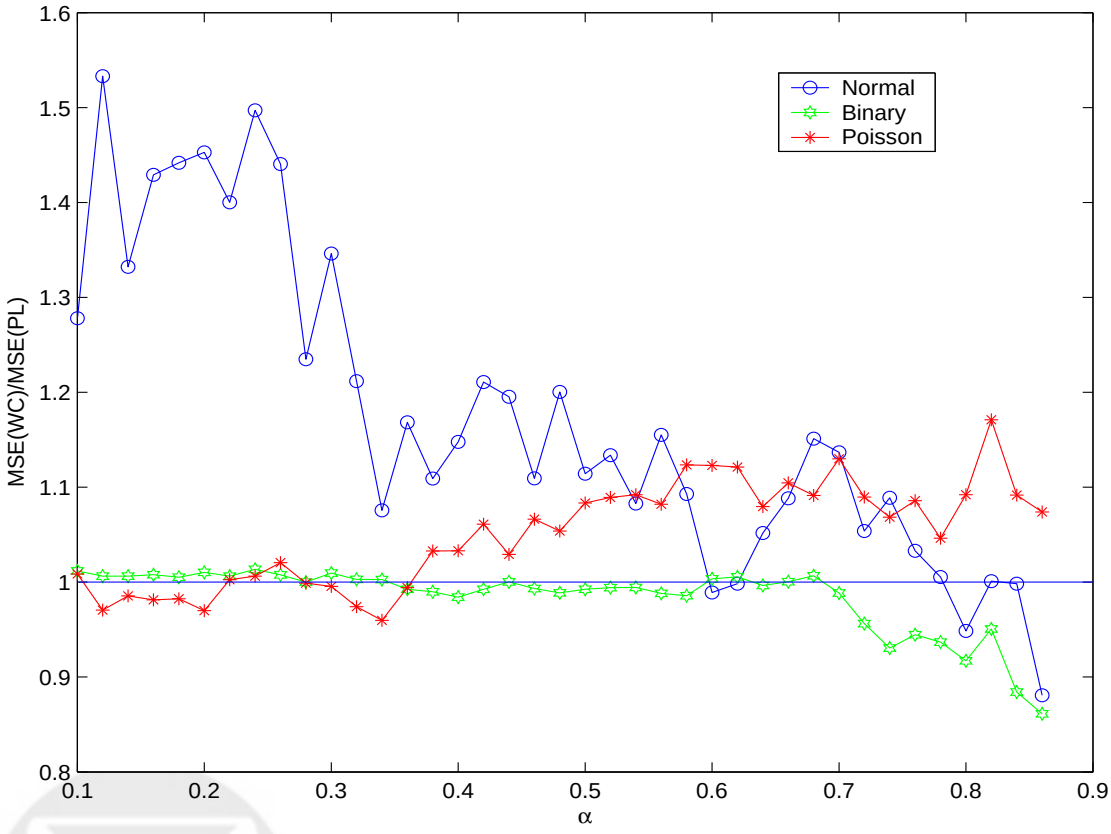


Table 1

Simulation Results

Design			Bias				MSE			
m	n	α	QONE	QLS	WC	PL	QONE	QLS	WC	PL
					Correlated Normal Data					
30	5	0.15	-0.0709	0.0272	0.0150	0.0247	0.0076	0.0055	0.0074	0.0056
30	5	0.30	-0.1816	-0.0122	-0.0108	-0.0099	0.0392	0.0106	0.0119	0.0097
30	5	0.50	-0.2710	-0.0210	-0.0121	-0.0182	0.0777	0.0094	0.0090	0.0085
30	5	0.75	-0.3129	-0.0164	-0.0121	-0.0159	0.1012	0.0036	0.0037	0.0034
					Correlated Binary Data					
40	6	0.15	-0.0777	-0.0107	-0.0110	-0.0109	0.0125	0.0234	0.0238	0.0236
40	6	0.30	-0.1468	-0.0122	-0.0120	-0.0115	0.0274	0.0182	0.0178	0.0177
40	6	0.50	-0.2261	-0.0171	-0.0166	-0.0156	0.0573	0.0137	0.0126	0.0130
40	6	0.75	-0.2769	-0.0193	-0.0193	-0.0189	0.0817	0.0057	0.0052	0.0055
					Correlated Poisson Data with Over-dispersion					
40	6	0.15	-0.0800	-0.0145	-0.0143	-0.0137	0.0117	0.0194	0.0200	0.0204
40	6	0.30	-0.1517	-0.0204	-0.0191	-0.0190	0.0285	0.0178	0.0179	0.0177
40	6	0.50	-0.2418	-0.0405	-0.0379	-0.0354	0.0645	0.0156	0.0146	0.0134
40	6	0.75	-0.3018	-0.0462	-0.0409	-0.0426	0.0963	0.0081	0.0077	0.0068
					Correlated Poisson Data without Over-dispersion					
30	5	0.15	-0.0772	-0.0081	-0.0078	-0.0067	0.0105	0.0166	0.0167	0.0170
30	5	0.30	-0.1531	-0.0195	-0.0180	-0.0176	0.0284	0.0168	0.0169	0.0168
30	5	0.50	-0.2368	-0.0242	-0.0239	-0.0220	0.0612	0.0128	0.0126	0.0118

[h]

Table 2

Analysis of renal data

Covariate	Quasi-Least Squares Estimate	p-value	Wang-Carey Estimate	p-value	Pseudo-Likelihood Estimate	p-value
Constant	-1.3943 (0.6140)	0.0116	-1.5119 (0.6134)	0.0069	-1.4181 (0.6142)	0.0105
Time	0.1586 (0.0924)	0.0431	0.1983 (0.0963)	0.0197	0.1651 (0.0924)	0.0371
Time^2	-0.0103 (0.0081)	0.1006	-0.0132 (0.0078)	0.0465	-0.0107 (0.0080)	0.0903
BaseBMIZ	1.0442 (0.2570)	0.0000	1.0957 (0.2542)	0.0000	1.0482 (0.2565)	0.0000
AAEthnicity	-0.2313 (0.4946)	0.3201	-0.0905 (0.4600)	0.4220	-0.2141 (0.4903)	0.3311
AgeTrans	-0.0025 (0.0428)	0.4771	-0.0075 (0.0425)	0.4300	-0.0027 (0.0428)	0.4744
Correlation	$\hat{\alpha}_{QLS}$=0.5456		$\hat{\alpha}_{WC}$=0.7729		$\hat{\alpha}_{PL}$=0.5868	