

POWER AND SAMPLE SIZE CALCULATIONS FOR GENERALIZED ESTIMATING EQUATIONS VIA LOCAL ASYMPTOTICS

Zhigang Li and Ian W. McKeague

Dartmouth College and Columbia University

Abstract: We consider the problem of calculating power and sample size for tests based on generalized estimating equations (GEE), that arise in studies involving clustered or correlated data (e.g., longitudinal studies and sibling studies). Previous approaches approximate the power of such tests using the asymptotic behavior of the test statistics under *fixed* alternatives. We develop a more accurate approach in which the asymptotic behavior is studied under a sequence of *local* alternatives that converge to the null hypothesis at root- m rate, where m is the number of clusters. Based on this approach, explicit sample size formulae are derived for Wald and quasi-score test statistics in a variety of GEE settings. Simulation results show that in the important special case of logistic regression with exchangeable correlation structure, previous approaches can inflate the projected sample size (to obtain nominal 90% power using the Wald statistic) by over 10%, whereas the proposed approach provides an accuracy of around 2%.

Key words and phrases: Clustered and correlated data, GEE, local alternatives, longitudinal data analysis, marginal models.

1. Introduction

Power and sample size formulae play an important role in the design of experimental and observational studies. There is an extensive literature on this topic, especially for hypothesis tests based on the method of generalized estimating equations (GEE), as introduced by Liang and Zeger (1986) for handling correlated longitudinal or clustered data. In this setting, however, all previous sample size formulae have been derived using the asymptotic behavior of test statistics under *fixed* alternatives. In particular, fixed alternatives were used by Liu and Liang (1997) (henceforth LL) to derive sample size formulae for quasi-score statistics, and by Shih (1997) (henceforth Shih) for Wald test statistics, see also Rochon (1998), Pan (2001), Liu, Shih, and Gehan (2002), Jung and Ahn (2003, 2005) and Kim, Williamson, and Lyles (2005).

In this article, a more accurate approach to power and sample size calculations in the GEE setting is developed using local asymptotic theory

(Le Cam (1960)). That is, rather than directly calculating the power of a test of $H_0: \psi = \psi_0$ vs. a fixed alternative $H_1: \psi = \psi_1$, where ψ is a p -vector representing the parameters of interest, we first calculate the asymptotic power of the test under a sequence of *local* alternatives $H_{1m}: \psi = \psi_0 + h/\sqrt{m}$, where m is the sample size (the number of clusters) and h is a fixed p -vector (the local parameter). The local asymptotic approach is considered to be standard in settings that do not involve clusters of correlated data (van der Vaart (1998); Lehmann and Romano (2005)), but it has not previously been attempted in the GEE setting as far as we know. For use of the approach in a survival analysis setting, see Lin, Yao, and Ying (1999).

The justification for our approach is provided by a result showing that in general GEE settings the Wald and quasi-score test statistics are asymptotically noncentral chi-squared under A sequence of local alternatives. This result also provides a suitable small sample approximation to the asymptotic power function involving a weighted average of the gradient of the estimating equation. Under marginal models, this approximation can be expressed directly in terms of the distribution of the covariates. This leads to explicit sample size formulae by inverting the small sample approximation to the asymptotic power function at the local parameter value $h = \sqrt{m}(\psi_1 - \psi_0)$, and solving for m in the usual way.

Our approach has several advantages over previous approaches. For the quasi-score test, previous methods of power calculation depend on knowing the limiting value of the nuisance parameter estimator under H_1 ; such estimators are generally inconsistent under fixed alternatives, cf. Self and Mauritsen (1988). This can involve the additional burden of having to numerically find the root of a nonlinear equation; our approach, on the other hand, does not require this extra step because the nuisance parameter estimator is *consistent* under the local alternatives. For the Wald test, our approach has better accuracy. Detailed comparisons of our results with those of LL and Shih are made in the setting of marginal models with exchangeable correlation structure under various sampling designs.

The paper is organized as follows. Preliminary material about GEE is provided in Section 2. Our main results are stated and discussed in Section 3. In Section 4, we compare the sample size formulae provided by our approach with those of LL and Shih. The results of a simulation study comparing the accuracy of the various formulae are reported in Section 5. An application involving exposure to arsenic in drinking water is given in Section 6. Concluding remarks appear in Section 7.

2. Preliminaries

In this section we provide the background material we need on the GEE method and marginal models (see Diggle et al. (2002); Fitzmaurice et al. (2009)). For convenience, we consider longitudinal data as a special type of clustered data in which “cluster” can refer to (repeated measures on) a single subject, or a group of subjects.

2.1. Generalized estimating equations and marginal models

Let m be the number of clusters and n_i the number of units in the i th cluster, $i = 1, \dots, m$. Let \mathbf{y}_{ij} denote the outcome, \mathbf{x}_{ij} the p -vector of covariates of interest, \mathbf{z}_{ij} the q -vector of confounding covariates, and μ_{ij} the conditional mean for the j th unit in the i th cluster. The GEE approach of Liang and Zeger (1986) produces consistent parameter estimates given that the model for the marginal means μ_{ij} is correctly specified, regardless of misspecification of the intracluster correlation matrix. The marginal model assumes

$$g(\mu_{ij}) = \mathbf{z}_{ij}^T \boldsymbol{\kappa} + \mathbf{x}_{ij}^T \boldsymbol{\psi}, \quad i = 1, 2, \dots, m, \quad j = 1, 2, \dots, n_i, \quad (2.1)$$

where g is a known link function, $\boldsymbol{\psi}$ contains the parameters of interest, and $\boldsymbol{\kappa}$ contains nuisance parameters, including the intercept. Let ϕ denote the univariate dispersion parameter of the model, an s -dimensional vector $\boldsymbol{\alpha}$ denote the correlation parameters. The full vector of parameters is denoted $\boldsymbol{\beta} = (\phi, \boldsymbol{\alpha}^T, \boldsymbol{\kappa}^T, \boldsymbol{\psi}^T)^T$.

Let $\boldsymbol{\theta} = (\boldsymbol{\kappa}^T, \boldsymbol{\psi}^T)^T$. The GEE approach provides a consistent estimator of $\boldsymbol{\theta}$ by solving the following estimating equation, for given values of $\boldsymbol{\alpha}$ and ϕ , regardless of misspecification of the intracluster correlation matrix:

$$\sum_{i=1}^m U_i(\boldsymbol{\theta}, \boldsymbol{\alpha}, \phi) = 0, \quad (2.2)$$

where $U_i(\boldsymbol{\theta}, \boldsymbol{\alpha}, \phi) = D_i^T V_i^{-1} S_i$, $D_i = \partial \mu_i / \partial \boldsymbol{\theta}$, $S_i = \mathbf{y}_i - \boldsymbol{\mu}_i$, and $V_i = \Delta_i^{1/2} R_i(\boldsymbol{\alpha}) \Delta_i^{1/2}$ is the working covariance matrix. Here $\Delta_i = \text{diag}[\text{Var}(\mathbf{y}_{i1}), \dots, \text{Var}(\mathbf{y}_{in_i})]$ and $R_i(\boldsymbol{\alpha})$ is the working correlation matrix (all such quantities being conditional on the covariates and cluster sizes). The parameters $\boldsymbol{\alpha}$ and ϕ are usually unknown. Further estimating equations can be introduced to estimate them; see Fitzmaurice et al. (2009, Chap. 3) for detailed discussion. This results in a combined estimating equation for the full set of parameters $\boldsymbol{\beta}$. Numerical methods to solve for $\boldsymbol{\beta}$ are widely implemented in statistical packages.

2.2. General setting

The dimension n_i of \mathbf{y}_i is assumed to be uniformly bounded. We are interested in estimating $\boldsymbol{\beta} \in \mathbb{B} \subset \mathbb{R}^k$, a k -vector of unknown parameters indexing the

conditional marginal means and variance-covariance matrices of the \mathbf{y}_i , where \mathbb{B} is compact. Let Ψ_i be a Borel function from $\mathbb{R}^{n_i} \times \mathbb{B}$ to \mathbb{R}^k , $i = 1, \dots, m$, and

$$s_m(b) = \sum_{i=1}^m \Psi_i(\mathbf{y}_i, b), b \in \mathbb{B}, \quad (2.3)$$

where the estimating function Ψ_i also depends on covariates (suppressed in the notation; all expectations and variances are taken to be conditional on the covariates and the dimension n_i). Suppose that $E_\beta\{\Psi_i(\mathbf{y}_i, \beta)\} = 0$ for all i , where the subscript β indicates the true value of the parameter. The GEE estimator, $\hat{\beta} \in \mathbb{B}$, satisfies $s_m(\hat{\beta}) = 0$. Under mild conditions, see Shao (2003, Sec. 5.4) (henceforth Shao),

$$\sqrt{m}(\hat{\beta} - \beta) \xrightarrow{d} N_k(0, \Sigma_\beta) \quad (2.4)$$

as $m \rightarrow \infty$, where $\Sigma_\beta = \lim_{m \rightarrow \infty} m\{M_m(\beta)\}^{-1}[\sum_{i=1}^m \text{Var}_\beta\{\Psi_i(\mathbf{y}_i, \beta)\}]\{M_m(\beta)\}^{-1}$, $M_m(\beta) = -E_\beta\{\nabla_\beta s_m(\beta)\}$ and $\nabla_\beta s_m(\beta) = \partial s_m(\beta)/\partial \beta$. The limiting covariance matrix Σ_β can be consistently estimated by replacing β with $\hat{\beta}$ and $\text{Var}_\beta\{\Psi_i(\mathbf{y}_i, \beta)\}$ with $\Psi_i(\mathbf{y}_i, \hat{\beta})\{\Psi_i(\mathbf{y}_i, \hat{\beta})\}^T$; the resulting estimator is denoted by $\hat{\Sigma}$.

3. Power and Sample Size Calculation Method

In this section we first develop the local asymptotic behavior of the GEE estimators and the Wald and quasi-score statistics under the general setting in Section 2.2. Based on that, we propose a power and sample size calculation procedure for the GEE under marginal models. Throughout we restrict attention to the testing of a simple null hypothesis of the form $H_0: \psi = \psi_0$ vs. the alternative $H_1: \psi \neq \psi_0$, where ψ is the vector consisting of the last p components of β . Let λ be the vector of the first $k - p$ components of β , so $\beta = (\lambda^T, \psi^T)^T$, and for marginal models $\lambda = (\phi, \alpha^T, \kappa^T)^T$.

The Wald statistic is given by

$$W_m = m(\hat{\psi} - \psi_0)^T (B\hat{\Sigma}B^T)^{-1}(\hat{\psi} - \psi_0),$$

where $B = (0_{p \times (k-p)}, I_p)$, $0_{p \times (k-p)}$ is the $p \times (k - p)$ zero matrix and I_p is the $p \times p$ identity matrix. Here and elsewhere, expressions involving B are simply used to extract the relevant submatrix or subvector. The quasi-score statistic or generalized score statistic (Rotnitzky and Jewell (1990, p.488) or Boos (1992, p.329)) is given by

$$T_m = \{Bs_m(\tilde{\beta})\}^T V_T^{-1}\{Bs_m(\tilde{\beta})\},$$

where

$$V_T = \{BM_m(\tilde{\beta})^{-1}B^T\}^{-1} \\ \times \left[BM_m(\tilde{\beta})^{-1} \left\{ \sum_{i=1}^m \Psi_i(\mathbf{y}_i, \tilde{\beta}) \Psi_i(\mathbf{y}_i, \tilde{\beta})^T \right\} M_m(\tilde{\beta})^{-1} B^T \right] \\ \times \{BM_m(\tilde{\beta})^{-1}B^T\}^{-1},$$

$\tilde{\beta} = (\tilde{\lambda}^T, \psi_0^T)^T$ and $\tilde{\lambda}$, constructed in the Appendix, is a consistent estimator of λ under H_0 . Both W_m and T_m converge in distribution under H_0 to χ_p^2 , and H_0 is rejected at significance level ζ if the test statistic is greater than $\chi_{p,1-\zeta}^2$, where $\chi_{p,1-\zeta}^2$ is the 100(1 - ζ)th percentile of χ_p^2 .

We illustrate our approach to obtaining power and sample size for the quasi-score test; the procedure is the same for the Wald test and the power and sample size formulae for both tests coincide. The power function of the quasi-score test is $\psi \mapsto \pi_m(\psi) = P(T_m \geq \chi_{p,1-\zeta}^2 | \psi \neq \psi_0)$, the probability of rejecting the null hypothesis given that the true value of the parameter ψ belongs to the alternative. To obtain power or sample size under a specific value of ψ , say $\psi = \psi_A$, we propose to study the power based on a sequence of alternatives $H_{1m} : \psi = \psi_{1m} = \psi_0 + h/\sqrt{m}$ that converge at root- m rate to the null hypothesis; here h is an arbitrarily fixed p -vector. We show that W_m is asymptotically noncentral chi-squared under H_{1m} and approximate the power $\pi_m(\psi_{1m})$ under $\psi = \psi_{1m}$ based on this result. The power under the fixed alternative $\psi = \psi_A$ given a sample size m is then obtained by specifying h so that $\psi_{1m} = \psi_A$, in which case $\pi_m(\psi_{1m})$ becomes identically $\pi_m(\psi_A)$. This approach calculates power and sample sizes under the fixed value ψ_A although it is based on local asymptotic theory since ψ_A is on the path along which ψ_{1m} converges to ψ_0 as $m \rightarrow \infty$. The justification for this procedure depends on finding the asymptotic behavior of $\hat{\psi}$ under H_{1m} , and we do this by extending the approach in Shao to a sequence of contiguous alternatives.

3.1. Local asymptotics

The results in this section are developed under the general setting described in Section 2.2. Various regularity conditions are needed, as provided in the Appendix. Let P_m denote the joint distribution of $\{\mathbf{y}_i\}_{i=1,2,\dots,m}$ under H_{1m} conditional on the covariates. Let $\beta_0 = (\lambda_0^T, \psi_0^T)^T$ and $\beta_m = (\lambda_0^T, \psi_{1m}^T)^T$, where λ_0 is the true value of λ . Our first result establishes the asymptotic behavior of $\hat{\psi}$ under P_m as $m \rightarrow \infty$; a sketch of the proof is given in the Appendix.

Theorem 1. *Under regularity conditions (C1)–(C9) in the Appendix, $\sqrt{m}(\hat{\psi} - \psi_0)$ converges to $N_p(h, B\Sigma_{\beta_0}B^T)$ in distribution under P_m .*

Remark 1. In the special case that the \mathbf{y}_i are i.i.d. and $\hat{\beta}$ is the MLE of β , the limiting covariance matrix Σ_{β_0} is the inverse of the Fisher information matrix, and our result for the local asymptotic distribution of $\sqrt{m}(\hat{\beta} - \beta_0)$ reduces to the classical result obtained by applying Le Cam's third lemma; see, e.g., Chapter 7 of van der Vaart (1998).

Remark 2. Theorem 5.14 in Shao can be considered as a special case of the above theorem with $h = 0$, where the asymptotic behavior of the GEE estimator $\hat{\psi}$ is provided under a fixed value of ψ , see (2.4). The limiting covariance matrix remains the same under the local alternatives, namely the submatrix of Σ_{β_0} corresponding to the components in ψ_0 . The major extra challenge to show the above result, beyond the theorem in Shao, is that the distributions of the \mathbf{y}_i are allowed to vary.

We now state our result giving the asymptotic behavior of T_m and W_m under H_{1m} ; a sketch of the proof is given in the Appendix. The result is the basis of our approach to obtain power and sample size.

Theorem 2. *Under regularity conditions (C1)–(C9) in the Appendix, the test statistics T_m and W_m converge under P_m in distribution to a noncentral chi-squared random variable $\chi_p^2(\nu)$ with non-centrality parameter $\nu = h^T(B\Sigma_{\beta_0}B^T)^{-1}h$.*

3.2. Power and sample size calculation procedure

In this section, we propose a power and sample size calculation procedure for GEE under marginal models. To carry out the calculations at the design stage of a study, we first need a good approximation to finite sample distributions of T_m and W_m . Theorem 2 provides such an approximation, but we propose to replace the non-centrality parameter ν by what we consider to be a better approximation for a given sample size m :

$$\nu_m = m\xi_{m\psi}^T \Sigma_{m\psi}^{-1} \xi_{m\psi}, \quad (3.1)$$

where

$$\xi_{m\psi} = B\{M_m(\beta_0)\}^{-1}E_{\beta_m}\{s_m(\beta_0)\}, \quad (3.2)$$

$$\Sigma_{m\psi} = mB\{M_m(\beta_0)\}^{-1}\text{Var}_{\beta_m}\{s_m(\beta_0)\}\{M_m(\beta_0)\}^{-1}B^T. \quad (3.3)$$

From the proof of Theorem 1 provided in the supplementary material, we have $\sqrt{m}\xi_{m\psi}$, $\Sigma_{m\psi}$, and ν_m converge to h , $B\Sigma_{\beta_0}B^T$, and ν respectively.

Note that expressions (3.1)–(3.3) are conditional on the covariates and cluster sizes. If the design is non-random and prespecified, then these expressions could be used directly. However, in general, these expressions would not be available

and would need to be estimated from pilot data. For the rest of this section we restrict attention to the case that the covariates and cluster sizes $(\mathbf{z}_i, \mathbf{x}_i, n_i)$, $i = 1, \dots, m$, are i.i.d., where $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{in_i})$ and $\mathbf{z}_i = (\mathbf{z}_{i1}, \mathbf{z}_{i2}, \dots, \mathbf{z}_{in_i})$. Suppose it is of interest to achieve nominal power $1 - \eta$ at significance level ζ under $\psi = \psi_A$. As mentioned, we replace h in the expression of ν_m by $\sqrt{m}(\psi_A - \psi_0)$. Ideally, for the purpose of sample size calculation, we would integrate M_m , $E_{\beta_m}\{s_m(\beta_0)\}$ and $\text{Var}_{\beta_m}\{s_m(\beta_0)\}$ over the distribution of $(\mathbf{z}_1, \mathbf{x}_1, n_1)$. After those steps (details are provided in Section 2 of the supplementary material), under the marginal model described in Section 2.1, the non-centrality parameter ν_m can be approximated by $\tilde{\nu}_m = m\tilde{\xi}_\psi^T \tilde{\Sigma}_\psi^{-1} \tilde{\xi}_\psi$, where

$$\tilde{\xi}_\psi = \bar{B}\{E(D_1^T V_1^{-1} D_1)\}^{-1} E[D_1^T V_1^{-1} \{\mu_1(\theta_A) - \mu_1(\theta_0)\}], \quad (3.4)$$

$\mu_1(\theta)$ is the conditional mean vector of \mathbf{y}_1 given $(\mathbf{z}_1, \mathbf{x}_1, n_1)$ and the value of θ , and

$$\begin{aligned} \tilde{\Sigma}_\psi &= \bar{B}\{E(D_1^T V_1^{-1} D_1)\}^{-1} \\ &\times [E\{D_1^T V_1^{-1} \text{Var}_{\beta_A}(\mathbf{y}_1 | \mathbf{z}_1, \mathbf{x}_1, n_1) V_1^{-1} D_1\}] \{E(D_1^T V_1^{-1} D_1)\}^{-1} \bar{B}^T. \end{aligned} \quad (3.5)$$

Here D_1 and V_1 are evaluated under H_0 , the expectations are taken with respect to the pre-specified distribution of $(\mathbf{z}_1, \mathbf{x}_1, n_1)$, $\theta_A = (\kappa_0^T, \psi_A^T)^T$, $\beta_A = (\lambda_0^T, \psi_A^T)^T$, $\bar{B} = (0_{p \times q}, I_p)$, and $\lambda_0 = (\phi_0, \alpha_0^T, \kappa_0^T)^T$ is the true value of the nuisance parameter. The power at $\psi = \psi_A$ is given by

$$\pi_m(\psi_A) = P \left\{ \chi_p^2 \left(m\tilde{\xi}_\psi^T \tilde{\Sigma}_\psi^{-1} \tilde{\xi}_\psi \right) \geq \chi_{p,1-\zeta}^2 \right\}.$$

The final step is to solve an equation for sample size m for achieving $1 - \eta$ nominal power at significance level ζ :

$$\pi_m(\psi_A) = 1 - \eta, \quad (3.6)$$

where $\tilde{\nu}$ satisfies $P \left\{ \chi_p^2(\tilde{\nu}) \geq \chi_{p,1-\zeta}^2 \right\} = 1 - \eta$, and then the sample size is given by

$$m = \frac{\tilde{\nu}}{\tilde{\xi}_\psi^T \tilde{\Sigma}_\psi^{-1} \tilde{\xi}_\psi}. \quad (3.7)$$

Remark 3. The vector h is chosen to be $\sqrt{m}(\psi_A - \psi_0)$, so the fixed value ψ_A is on the path along which H_{1m} converges to H_0 . The method calculates power and sample sizes under a fixed value ψ_A of ψ although it is based on local asymptotic theory since ψ_A is on the path along which ψ_{1m} converges to ψ_0 as $m \rightarrow \infty$. We will see that our proposed method works better than previous approaches in various commonly seen cases with fixed alternative values, as discussed in a simulation study later.

Remark 4. An alternative approach is to calculate the power $\pi_m(\psi_{1m})$ with ν_m replaced by ν (cf., van der Vaart (1998); Lehmann and Romano (2005)). To calculate ν , this approach uses the limiting variance $B\Sigma_{\beta_0}B^T$ that does not depend on the alternative value ψ_A . However, the variance of $\sqrt{m}(\hat{\psi} - \psi_0)$ may depend on the alternative value ψ_A for a given sample size m under some models, e.g., logistic regression models, where variance of the outcome is a function of its mean.

In summary, our sample size calculation proceeds as follows,

1. Choose type I error rate ζ and power $1 - \eta$.
2. Specify the joint distribution of covariates and cluster size.
3. Give ψ_0 , ψ_A and λ_0 and specify the working correlation matrix (based on pilot data if available). Calculate D_1 and V_1 in (2.2) under $\beta = \beta_0$.
4. Calculate $\text{Var}_{\beta_A}(\mathbf{y}_1|\mathbf{z}_1, \mathbf{x}_1, n_1)$ using the choice of the working correlation matrix (in Step 3) in place of the true correlation matrix.
5. Based on the results in Steps 3 and 4, find, using numerical integration methods if necessary, the expectations within (3.5) under the joint distribution of the covariates and cluster size given in Step 2.
6. Calculate $\tilde{\xi}_\psi$, $\tilde{\Sigma}_\psi$, $\tilde{\nu}$ and the sample size m according to (3.4), (3.5), (3.6), and (3.7).

The joint distribution in Step 2 and values of the nuisance parameters in Step 3 may need to be estimated from pilot data. One possibility in Step 2 would simply be to use the empirical distribution of the pilot data; alternatively, a parametric model could be fit to the pilot data (e.g., the normal distribution in the arsenic study discussed in Section 6), in which case numerical integration might be needed in Step 5. Of course, if the true correlation matrix is known, it could be used in Step 4, but typically such information is not available at the design stage (cf., Liu and Liang (1997); Shih (1997); Liu, Shih, and Gehan (2002); Jung and Ahn (2005)). As mentioned by LL, sensitivity analysis should also be carried out to assess how the sample size varies according to changes of the specified working correlation matrix. Sample size re-estimation (SSR) is often carried out when interim data are available for updating design parameters (e.g., Shih (2001); Friede and Kieser (2006)). SSR for studies with correlated data have been actively discussed (e.g., Shih and Gould (1995); Lake et al. (2002); Zucker and Denne (2002); Yin and Shen (2005)).

4. Comparison with Previous Approaches

In this section, we compare our approach with LL's and Shih's for several important examples of marginal models. The working correlation matrix is assumed to be the true correlation matrix. Background and discussion of LL's and Shih's approaches are provided in Section 3 and Section 4 of the supplementary material.

LL restricted attention to the case of discrete covariates, but for continuous covariates their approach requires an initial (ad hoc) discretization. As pointed out by Shieh (2000), there is no consensus in how the discretization should be done. For LL's approach, it is also necessary to derive the limiting value of the nuisance parameter estimator under the alternative hypothesis. Compared to Shih's approach, our approach has much better accuracy under various circumstances.

4.1. Continuous outcomes and cluster level covariates

Consider a study with common cluster size n , continuous outcomes and covariates: $x_{ij} \equiv x_i$, $z_{ij} \equiv 1$ (intercept), $j = 1, \dots, n$. For simplicity, suppose that the working correlation matrix $R_i \equiv R$ coincides with the true correlation matrix. The standard linear regression version of (2.1) is $\mu_{ij} = \kappa + x_i\psi$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$, and there is a constant conditional marginal variance $\text{Var}(y_{ij}|x_i) = \sigma^2$. The null hypothesis of interest is $H_0: \psi = \psi_0$. To obtain a desired power $1 - \eta$ at significance level ζ for detecting $\psi = \psi_A$, the sample size from our approach (3.7) is

$$m = \frac{(z_{1-\zeta/2} + z_{1-\eta})^2 (\mathbf{1}_n^T R^{-1} \mathbf{1}_n)^{-1} \sigma^2}{(\psi_A - \psi_0)^2 \text{Var}(x_1)}, \quad (4.1)$$

where $\mathbf{1}_n$ is the n -vector of 1's. Derivation of this formula is provided in Section 8 of the supplementary material. When the correlation has an exchangeable correlation structure with $\rho_{ij} = \rho$, the numerator in (4.1) is $(z_{1-\zeta/2} + z_{1-\eta})^2 \{1 + (n-1)\rho\} \sigma^2$. Note that the presence of the intracluster correlation has effectively increased the noise variance from σ^2 to $\{1 + (n-1)\rho\} \sigma^2$. The inflation factor $1 + (n-1)\rho$ is known as the *design effect* (Scott and Holt (1982)), and provides a direct measure of how the sample size (needed to achieve a fixed nominal power) increases as a function of the intracluster correlation. The formula (4.1) agrees with Shih's. In this example, LL's approach requires (a) x_1 has a discrete distribution, or (b) an ad hoc discretization is made to its continuous distribution if x_1 is a continuous variable; in the former case, LL's result agrees with (4.1).

4.2. Binary outcomes and cluster level covariates

This example uses the same assumptions as 4.1 above except that we use the logistic regression version of (2.1):

$$\text{logit}(\mu_{ij}) = \kappa + x_i\psi, \quad i = 1, 2, \dots, m; \quad j = 1, 2, \dots, n. \quad (4.2)$$

The null hypothesis of interest is $H_0: \psi = \psi_0$. To obtain a desired power $1 - \eta$ for detecting $\psi = \psi_A$, the sample size from (3.7) is

$$m = (z_{1-\zeta/2} + z_{1-\eta})^2 (\mathbf{1}_n^T R^{-1} \mathbf{1}_n)^{-1} \\ \times \frac{\{E(v_{1x})(E(xv_{0x}))^2 + E(x^2v_{1x})(E(v_{0x}))^2 - 2E(xv_{1x})E(xv_{0x})E(v_{0x}))\}}{[E(v_{0x})E(xp_{1x}) - E(v_{0x})E(xp_{0x}) - E(xv_{0x})\{E(p_{1x}) - E(p_{0x})\}]^2},$$

where $p_{0x} = \text{expit}(\kappa_0 + x\psi_0)$, $p_{1x} = \text{expit}(\kappa_0 + x\psi_A)$, $v_{0x} = p_{0x}(1 - p_{0x})$ and $v_{1x} = p_{1x}(1 - p_{1x})$. Here p_{0x} and p_{1x} are the risks of disease at exposure level x under $\psi = \psi_0$ and $\psi = \psi_A$, respectively, $\text{expit}(\cdot)$ is the inverse of the logit function, and the expectations are taken under the distribution of x_1 . Derivation of this formula is provided in Section 9 of the supplementary material. Shih's approach gives

$$m = \frac{(z_{1-\zeta/2} + z_{1-\eta})^2 E(v_{1x})(\mathbf{1}_n^T R^{-1} \mathbf{1}_n)^{-1}}{(\psi_A - \psi_0)^2 [E(v_{1x})E(x^2v_{1x}) - \{E(xv_{1x})\}^2]}.$$

Again, LL's approach requires (a) x_1 has a discrete distribution, or (b) an ad hoc discretization is made to its continuous distribution; in both cases, LL's approach requires a solution to a nonlinear equation for κ^* in terms of a discrete distribution $P(x_1 = u_l) = \omega_l, l = 1, \dots, L$:

$$\sum_{l=1}^L \omega_l \{\text{expit}(\kappa_0 + u_l\psi_A) - \text{expit}(\kappa^* + u_l\psi_0)\} = 0.$$

However, if the null hypothesis of interest is $\psi_0 = 0$, this equation has an explicit solution and their result agrees with ours in case (a). In this example, there is a one-dimensional nuisance parameter and the equation is readily solved, but in general it would be more challenging to reach a solution. Moreover, the solution of this nonlinear equation is sensitive to the choice of discretization.

4.3. Sibling studies with binary outcomes and unit level binary covariates

Consider a cohort study of m pairs of siblings in which one sibling is exposed and the other unexposed, and the outcome is binary (disease or non-disease). The covariates are $\mathbf{x}_i = (x_{i1}, x_{i2})^T = (1, 0)^T$, where 1 and 0 designate exposed

and unexposed, respectively, and $\mathbf{z}_i = (z_{i1}, z_{i2})^T$ with $z_{i1} = z_{i2} = 1$ (intercept) for $i = 1, \dots, m$. Let ρ denote the common correlation between siblings. Logistic regression is used, and the cluster size $n = 2$. Let p_0 and p_1 denote the risk of the disease in unexposed and exposed subjects, respectively. Suppose the null hypothesis of interest is $H_0: \psi = \psi_0$. To obtain a desired power $1 - \eta$ for detecting $\psi = \psi_A$, the sample size m from (3.7) is

$$m = \frac{(z_{1-\zeta/2} + z_{1-\eta})^2 (v_0^2 v_1 + v_0 \tilde{v}_0^2 - 2\rho v_0 \tilde{v}_0 \sqrt{v_0 v_1})}{v_0^2 (p_1 - \tilde{p}_0)^2},$$

where $\tilde{p}_0 = \text{expit}(\kappa_0 + \psi_0)$, $\tilde{v}_0 = \tilde{p}_0(1 - \tilde{p}_0)$, $p_0 = \text{expit}(\kappa_0)$, $p_1 = \text{expit}(\kappa_0 + \psi_A)$, $v_0 = p_0(1 - p_0)$, and $v_1 = p_1(1 - p_1)$. Derivation of this formula is provided in Section 10 of the supplementary material. Shih's method leads to

$$m = \frac{(z_{1-\zeta/2} + z_{1-\eta})^2 (v_0 + v_1 - 2\rho\sqrt{v_0 v_1})}{(\psi_A - \psi_0)^2 v_0 v_1}.$$

LL's approach requires solving an equation for κ^* :

$$\left(1 - \frac{\rho p_1^*}{c_0 p_0^*}\right) (p_0 - p_0^*) + \left(1 - \frac{\rho c_0 p_0^*}{p_1^*}\right) (p_1 - p_1^*) = 0, \quad (4.3)$$

where $c_0 = \exp(\psi_0/2)$, $p_0^* = \text{expit}(\kappa^*)$, and $p_1^* = \text{expit}(\kappa^* + \psi_0)$. Again, if the null hypothesis of interest is $\psi_0 = 0$, (4.3) has the explicit solution $\kappa^* = \text{logit}[(p_0 + p_1)/2]$, and their result agrees with ours.

5. Simulation Study

In this section we report the results of a simulation study for some of the examples in the previous section in which the sample size formulae from previous approaches disagree with that from our approach. We generated 10,000 replicated samples according to the marginal model in each setting. For each sample, we estimated the parameters and implemented the tests. The empirical power is given by the proportion of samples with test statistic value exceeding $\chi_{p,1-\zeta}^2$. The software SAS v9.2 was used to generate the data sets and calculate the test statistics. Accuracy of the sample size formula is then determined by how close the empirical power is to the nominal power. We set the type I error rate to be $\zeta = 0.05$ and the nominal power at 90%. The Monte Carlo standard errors of the empirical powers reported in the simulation study are about 0.3%. In each example, we also provide the sample size that generates empirical power closest to the nominal power 90% in the simulation study and call it the "actual" sample size.

Table 1. Simulation results comparing our approach with Shih’s approach in the two-sample comparison problem with a binary outcome and a cluster level binary exposure. Sample sizes are for achieving a nominal power 90% to detect different relative risks (p_1/p_0) at 0.05 significance level for different correlations ρ . The reference risk p_0 is fixed at 0.1 and $\psi_0 = 0$. Empirical powers are based on the Wald test. The “actual” are the sample sizes with empirical powers closest to the nominal power 90%.

RR	ρ	Our approach		Shih’s approach		“Actual”	
		m	Power (%)	m	Power (%)	m	Power (%)
2.5	0.20	156	89.82	172	92.43	158	89.96
	0.50	195	89.47	215	92.28	196	90.03
	0.80	234	89.40	258	92.68	238	90.06
3.0	0.20	95	89.32	110	93.65	96	90.11
	0.50	119	89.52	138	93.52	122	90.49
	0.80	142	89.45	165	93.25	144	89.86
3.5	0.20	65	89.11	79	93.85	66	89.87
	0.50	81	88.20	99	94.28	84	89.62
	0.80	97	88.76	118	94.45	100	89.98

5.1. Simulation results for binary outcomes and cluster level binary exposures

Consider a special case of Example 4.2 involving a two-group comparison study with common cluster size 2, binary outcomes (disease or non-disease) and covariates $x_{ij} \equiv x_i = 1$ (exposed) or 0 (unexposed), and $z_{ij} \equiv 1$ (intercept), $j = 1, 2$. Suppose half of the clusters are unexposed and the other half exposed. We set the reference risk of disease in the unexposed group to be $p_0 = 0.1$, and set $\psi_0 = 0$. The simulation results are presented in Table 1. The empirical powers of the sample sizes from Shih’s approach are substantially larger than the nominal power 90%, whereas from our approach they are mostly within one or two Monte Carlo standard errors of the nominal power. The sample sizes from our approach almost coincide with the “actual”. Shih’s approach overestimates the sample sizes by over 8% for the first three cases, and over 10% for the last six cases. In particular, for the last three cases the sample sizes are overestimated by about 18%. On the other hand, our approach is accurate to within about 1–3% for the first six cases and 1–4% for the last three cases. The reason for the overestimation of sample size based on Shih’s method is unclear, but perhaps it is due to the use of the limiting variance of $\sqrt{m}(\hat{\psi} - \psi_A)$ in the calculation of the non-centrality parameter, as discussed in Remark 4 of Section 3.

Table 2. Simulation results comparing our approach with Shih’s approach in the sibling study with a binary outcome and a binary exposure. Sample sizes are for achieving a nominal power 90% to detect different relative risks (p_1/p_0) at 0.05 significance level for different correlations ρ . The reference risk p_0 is fixed at 0.1 and $\psi_0 = 0$. Empirical powers are based on the Wald test.

RR	ρ	Our approach		Shih’s approach		“Actual”	
		m	Power (%)	m	Power (%)	m	Power (%)
2.0	0.10	238	90.34	251	91.72	236	90.02
	0.15	225	90.31	238	91.85	220	90.03
	0.20	213	90.93	225	91.95	209	89.96
2.5	0.10	118	90.16	130	92.91	118	90.16
	0.15	112	90.69	124	93.18	109	90.08
	0.20	106	90.70	117	93.37	102	89.96
3.0	0.10	72	90.05	84	93.67	72	90.05
	0.15	68	89.75	79	93.90	68	89.75
	0.20	65	90.36	75	93.88	64	90.00

5.2. Simulation results for sibling studies with binary outcomes and binary exposures

First consider the sibling study (Example 4.3) with $p_0 = 0.1$ and $\psi_0 = 0$, see Table 2. Shih’s approach still overestimates the sample size. The sample sizes from our approach are all accurate to within 1–4% compared to the actual, whereas Shih’s approach overestimates the sample sizes by 6–18%. Table 3 compares our approach with LL when ψ_0 changes to 0.5 (LL’s approach agrees with ours when $\psi_0 = 0$). The empirical powers of the sample sizes from our approach are mostly within one Monte Carlo standard error of the nominal power, whereas those from LL’s approach fall outside such limits.

6. Application to an Arsenic Study

Exposure to arsenic through drinking water represents a major threat to human health (Karagas (2010)). Recently, a cohort study has been initiated by the Children’s Environmental Health and Disease Prevention Center and Superfund Basic Research Program at Dartmouth College to investigate the impact of arsenic exposure on the health of pregnant women and children in New England. A primary aim of the study is to assess the association between arsenic exposure during pregnancy and infant growth. Suppose the binary outcome ‘short stature’ is to be recorded at 0.5, 1, 1.5, and 2 years of age, giving the cluster size of $n = 4$ in model (4.2). The proportion of ‘short’ children (<5 years of age) in New Hampshire is estimated to be 6.2% according to the Pediatric Nutrition Surveillance (2009), and we use this information to specify the intercept $\kappa_0 = \log\{0.062/(1 - 0.062)\} = -2.717$ corresponding to $x_i = 0$ in (4.2).

Table 3. Simulation results comparing our approach with LL’s approach in the sibling study with a binary outcome and a binary exposure. Sample sizes are for achieving a nominal power 90% to detect different effect sizes relative risks (p_1/p_0) at 0.05 significance level for different correlations ρ . Risk p_0 is fixed at 0.1 and $\psi_0 = 0.5$. Empirical powers are based on the quasi-score test.

RR	ρ	Our approach		LL’s approach		“Actual”	
		m	Power (%)	m	Power(%)	m	Power (%)
2.5	0.10	395	90.04	388	89.47	395	90.04
	0.15	373	89.83	366	89.59	374	90.10
	0.20	351	89.82	345	89.11	351	89.82
3.0	0.10	180	90.16	176	89.53	179	89.99
	0.15	170	90.22	166	89.48	169	90.06
	0.20	160	90.55	157	89.81	158	89.95
3.5	0.10	104	89.58	102	89.57	105	89.87
	0.15	99	90.11	96	88.80	99	90.11
	0.20	93	90.19	91	89.70	92	89.88
4.0	0.10	68	89.12	66	88.11	70	89.89
	0.15	65	90.12	63	89.23	65	90.12
	0.20	61	90.37	59	89.50	61	90.37

Data indicate that mean log-arsenic concentration ($\mu\text{g/L}$) in the water supply in New Hampshire is -0.942 , so in (4.2) we set $x_i = \log\text{-exposure} - (-0.942)$. Pilot data on the study population (Gilbert-Diamond et al. (2012)) show that log-arsenic exposure is (approximately) distributed as $N(-0.04, 4)$, so we specify $x_i \sim N(0.902, 4)$.

First we consider an AR(1) structure, $\rho_{ij} = \rho^{|i-j|}$. Pilot data on children’s heights in this cohort were not provided, so it is not possible to estimate ρ and we consider a range of values: $\rho = 0.2, 0.5$ and 0.8 . For each ρ , we calculate the sample size needed to detect an odds ratio of 1.5 ($\psi_A = 0.406$) due to a unit increase in log-arsenic exposure (to achieve 90% power at significance level 0.05). Our approach gives sample sizes of 70, 105, and 157, and Shih’s gives 69, 103, and 154 for the values of ρ , respectively. Using LL’s approach with a discretization of 150 equispaced values within 3 standard deviations of the mean log-arsenic exposure, the sample sizes are 73, 109, and 164, respectively. Under an exchangeable structure with $\rho_{ij} = \rho$, for $\rho = 0.2, 0.5$ and 0.8 , the sample sizes are 84, 131, and 178 from our approach, 82, 128, and 174 from Shih’s, and 88, 137, and 185 from LL’s, respectively. In this example, we used Monte Carlo integration in Step 5. SAS code is provided in the supplementary material.

7. Discussion

We have developed a method to calculate sample size for studies with correlated data. Under the framework of marginal models, our approach gives the same sample size formulae for the Wald and quasi-score tests. We study power under a sequence of alternatives that converge at root- m rate to the null hypothesis, and show that the statistics converge in distribution to a noncentral chi-squared. Then we link the sequence of alternatives to a fixed alternative by choosing a specific value of the local parameter vector h so that ψ_1 is on the path along which H_{1m} converges to H_0 .

As shown in simulation study, our approach provides considerable improvements in terms of accuracy. Hanfelt and Liang (1995) developed an approximate likelihood ratio test for GEE, that could be used as the basis for power and sample size calculations. We conjecture that our approach could be extended to the approximate likelihood ratio test and would lead to the same conclusion since the likelihood ratio test is equivalent to the other two tests for independent data.

We used correlation to measure the association between binary outcomes. An alternative to the correlation as a measure of association between pairs of binary responses is the odds ratio, which has a more straightforward interpretation. To estimate the odds ratio as a measure of association, a second set of estimating equations (Lipsitz, Laird, and Harrington (1991); Carey, Zeger, and Diggle (1993)) can be used. The power and sample size calculations approach developed in the present paper can readily be adapted to these settings.

A referee raised the question of whether our approach could be extended to ordinal or nominal outcomes. Lipsitz, Kim, and Zhao (1994) developed a GEE for such outcomes; This could be handled under our local asymptotic approach, but further work would be needed to develop a specific procedure. Another issue raised by a referee concerns the handling of missing data. We refer the interested reader to Robins, Rotnitzky, and Zhao (1995) who developed an inverse-weighted GEE method for missing data. Again, further work would be needed to develop a specific procedure.

Monte Carlo simulations are sometimes used to obtain power and sample size, especially for complicated designs where explicit sample size formulae are hard to derive. However, simulations have some serious disadvantages from a practical perspective: they are time-consuming, computational expertise is needed and, in the case of small sample sizes, the results can be highly sensitive to distributional assumptions. On the other hand, our formulae only require plugging a few numbers into a calculator (or at most some routine numerical integrations), are more appealing to practitioners, and distributional assumptions play a limited role. With an explicit sample size formula, we are able to calculate the minimal sample size to find the most efficient design strategy in the planning stages of

a study. Even with complicated designs, where explicit formulae are not available, explicit sample size formulae are still useful for providing initial sample size estimates that can later be refined using simulations.

Acknowledgements

The authors are grateful for many helpful comments from an associate editor and the referees. The research of Dr. Ian McKeague was supported in part by NIH Grant R01GM095722. We thank Dr. Margaret Karagas and the Children's Environmental Health and Disease Prevention Center at Dartmouth College for providing preliminary data in the arsenic study example (supported by P20 ES018175 and P42 ES007373 from the NIEHS and NIH; and RD-83459901-0 from the USEPA).

Appendix

Regularity conditions

- (C1) The parameter space \mathbb{B} is compact and β_0 belongs to its interior.
 (C2) There exists $\delta > 1$ such that

$$c_0 \equiv \sup_{m \geq 1} \max_{i=1, \dots, m} E_{\beta_m} |h_i(\mathbf{y}_i)|^{1+\delta} < \infty \text{ and } c_1 \equiv \sup_{m \geq 1} \max_{i=1, \dots, m} E_{\beta_m} |\mathbf{y}_i|^\delta < \infty,$$

where $h_i(\mathbf{y}_i) \equiv \sup_{b \in \mathbb{B}} |\Psi_i(\mathbf{y}_i, b)|$.

- (C3) For any bounded sequence $\{y_i\}$ with $y_i \in \mathbb{R}^{n_i}$, the functions $b \mapsto \Psi_i(y_i, b)$ are equicontinuous and uniformly bounded on \mathbb{B} .
 (C4) $\sup_{i \geq 1} |E_{b_1} \{\Psi_i(\mathbf{y}_i, b)\} - E_{b_1^*} \{\Psi_i(\mathbf{y}_i, b^*)\}| \lesssim |b_1 - b_1^*| + |b - b^*|$ for all $b_1, b_1^*, b, b^* \in \mathbb{B}$, where “ \lesssim ” means “smaller than up to a constant”.
 (C5) For all $\epsilon > 0$, $\inf_{m \geq 1, |b - \beta_0| > \epsilon} |f_{m,0}(b)| > 0 = |f_{m,0}(\beta_0)|$, where $f_{m,0}(b) = E_{\beta_0} \{m^{-1} s_m(b)\}$.
 (C6) $\varphi_i(y, b) = \nabla_b \Psi_i(y, b)$ exists for all $y \in \mathbb{R}^{n_i}, b \in \mathbb{B}$, and its j th row (denoted by $\varphi_{ij}(y, b)$ for later use) satisfies conditions (C1)–(C4) in place of $\Psi_i(y, b)$, for each $j = 1, \dots, k$.
 (C7) There exists a neighborhood N of β_0 such that $\sup_{i \geq 1, j=1, \dots, k} |V_{ij}(b) - V_{ij}(\beta_0)| \lesssim |b - \beta_0|$ for all $b \in N$, where $V_{ij}(b)$ is the j th column of $\text{Var}_b(\Psi_i(\mathbf{y}_i, \beta_0))$.
 (C8) The elements of $(1/m)M_m(\beta_0)$ and $(1/m) \sum_{i=1}^m \text{Var}_{\beta_0}(\Psi_i(\mathbf{y}_i, \beta_0))$ converge to finite limits, where $M_m(\beta) = -E_{\beta_0} \{\nabla_{\beta} s_m(\beta)\}$.
 (C9) $M(\beta_0) = \lim_{m \rightarrow \infty} \{m^{-1} M_m(\beta_0)\}$ is non-singular.

Remark 5. Condition (C2) is slightly stronger than the corresponding condition in Shao, namely that $\sup_i E|h_i(\mathbf{y}_i)|^{1+\delta} < \infty$ and $\sup_i E|\mathbf{y}_i|^\delta < \infty$ for some $\delta > 0$; we need the stronger condition because in our case the distribution of \mathbf{y}_i is changing with sample size. The consistency result established in the first part of the proof of Theorem 1 still holds if we replace $\delta > 1$ by $\delta > 0$ in (C2), but we need $\delta > 1$ for the asymptotic normality result. Condition (C3) is also a little stronger because we make the additional assumption that the functions are uniformly bounded on \mathbb{B} . Similar to the existence of a *well-separated* point of maximum in Theorem 5.7 in van der Vaart (1998), (C5) requires that $f_m(b)$ has a unique zero β_0 , and only values close to β_0 yield a value of $f_m(b)$ close to zero. (C4) and (C7) are Lipschitz conditions.

Remark 6. Under marginal models, $(1/m)M_m(\beta_0)$ and $(1/m)\sum_{i=1}^m \text{var}_{\beta_0}(\Psi_i(\mathbf{y}_i, \beta_0))$ in (C8) reduce to $(1/m)\sum_{i=1}^m D_i^T V_i^{-1} D_i$ and

$$\frac{1}{m} \sum_{i=1}^m D_i^T V_i^{-1} \text{Cov}(\mathbf{y}_i | \mathbf{z}_i, \mathbf{x}_i) V_i^{-1} D_i,$$

respectively; These are important components of the covariance matrix of the GEE estimator developed in Liang and Zeger (1986). Also, the matrix $M(\beta_0)$ in (C9) is the inverse of the asymptotic covariance of $\sqrt{m}(\hat{\beta} - \beta_0)$ if the working correlation is correctly specified.

Sketch of proof of Theorem 1

The detailed proof is provided in Section 6 of the supplementary material. Under (C1)–(C5), it can be shown that $\hat{\beta} \xrightarrow{P_m} \beta_0$ and $\hat{\psi} \xrightarrow{P_m} \psi_0$ by adapting the proof of Theorem 5.7 of van der Vaart (1998).

By using the extra conditions (C6)–(C9), we can further show that $\sqrt{m}(\hat{\beta} - \beta_0)$ converges in distribution under P_m to $N(h^*, \Sigma_{\beta_0})$, where $h^* = (\mathbf{0}_{k-p}^T, h^T)^T$ and $\mathbf{0}_{k-p}$ is the $(k-p)$ -dimensional zero vector. The proof is completed by noticing that $\hat{\psi} = B\hat{\beta}$.

Sketch of proof of Theorem 2

The detailed proof is provided in Section 7 of the supplementary material.

Asymptotic distribution of W_m

The first step is to show that $\hat{\Sigma}$ converges in probability under H_{1m} to Σ_{β_0} . Thus $B\hat{\Sigma}B^T$ converges in probability under H_{1m} to $B\Sigma_{\beta_0}B^T$. From Theorem 1, $\sqrt{m}(\hat{\psi} - \psi_0)$ converges in distribution under H_{1m} to $N(h, B\Sigma_{\beta_0}B^T)$. Therefore, by Slutsky's Lemma and the Continuous Mapping Theorem, W_m converges in distribution under H_{1m} to noncentral χ_p^2 with non-centrality parameter $\nu = h^T(B\Sigma_{\beta_0}B^T)^{-1}h$.

Asymptotic distribution of T_m

An estimate $\tilde{\lambda}$ of the nuisance parameter vector λ under H_0 is needed to calculate the quasi-score statistic. For this purpose it suffices to use the first $k - p$ estimating equations, so $\tilde{\lambda}$ can be taken as a solution of $Cs_m(\lambda, \psi_0) = 0$, where $C = (I_{(k-p)}, 0_{(k-p) \times p})$. Recall that $\tilde{\lambda}$ and ψ_0 combine to form $\tilde{\beta}$. Write the quasi-score statistic as

$$T_m = \left\{ m^{-1/2} B s_m(\tilde{\beta}) \right\}^T (m^{-1} V_T)^{-1} \left\{ m^{-1/2} B s_m(\tilde{\beta}) \right\}.$$

It can be shown that T_m is asymptotically equivalent to W_m , concluding the proof.

Contents of supplementary material

Detailed proofs, derivations of the sample size formulae, outlines of LL's and Shih's approaches, and SAS code for calculating the sample sizes in the arsenic example.

References

- Boos, D. D. (1992). On generalized score tests. *Amer. Statist.* **46**, 327-333.
- Carey, V., Zeger, S. L. and Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika* **82**, 517-526.
- Diggle, P. J., Heagerty, P., Liang, K.-Y. and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. 2nd edition. Oxford University Press.
- Fitzmaurice, G. M., Davidian, M., Verbeke, G. and Molenberghs, G. (2009). *Longitudinal Data Analysis: Handbooks of Modern Statistical Methods*. Chapman and Hall/CRC, Boca Raton.
- Friede, T. and Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical J.* **48**, 537-555.
- Gilbert-Diamond, D., Cottingham, K. L., Gruber, J. F., Punshonb, T., Sayarath, V., Gandolfi, A. J., Baker, E. R., Jackson, B., Folt, C. L. and Karagas, M. R. (2012). Rice consumption contributes to arsenic exposure in U.S. women. *The Proceedings of the National Academy of Sciences of the United States of America*, In press.
- Hanfelt, J. J. and Liang, K.-Y. (1995). Approximate likelihood ratios for general estimating functions. *Biometrika* **82**, 461-477.
- Jung, S. H. and Ahn, C. (2003). Sample size estimation for GEE method for comparing slopes in repeated measurements data. *Statist. Medicine* **22**, 1305-1315.
- Jung, S. H. and Ahn, C. (2005). Sample size for a two-group comparison of repeated binary measurements using GEE. *Statist. Medicine* **24**, 2583-2596.
- Karagas, M. R. (2010). Arsenic-related mortality in Bangladesh. *The Lancet* **376**, 213-214.
- Kim, H.-Y., Williamson, J. M. and Lyles, C. M. (2005). Sample-size calculations for studies with correlated ordinal outcomes. *Statist. Medicine* **21**, 1337-1350.
- Lake, S., Kammann, E., Klar, N. and Betensky, R. (2002). Sample size re-estimation in cluster randomization trials. *Statist. Medicine* **24**, 2977-2987.
- Le Cam, L. (1960). Locally asymptotically normal families of distributions. *University of California Publications in Statistics* **3**, 37-98.

- Lehmann, E. and Romano, J. (2005). *Testing Statistical Hypotheses*. 3rd edition. Springer, New York.
- Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13-22.
- Lin, D. Y., Yao, Q. and Ying, Z. (1999). A general theory on stochastic curtailment for censored survival data. *J. Amer. Statist. Assoc.* **94**, 510-521.
- Lipsitz, S. R., Kim, K. and Zhao, L. (1994). Analysis of repeated categorical data using generalized estimating equations. *Statist. Medicine* **13**, 1149-1163.
- Lipsitz, S. R., Laird, N. M. and Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association. *Biometrika* **78**, 153-160.
- Liu, A., Shih, W. J. and Gehan, E. (2002). Sample size and power determination for clustered repeated measurements. *Statist. Medicine* **21**, 1787-1801.
- Liu, G. and Liang, K.-Y. (1997). Sample size calculations for studies with correlated observations. *Biometrics* **53**, 937-947.
- Pan, W. (2001). Sample size and power calculations with correlated binary data. *Controlled Clinical Trials* **22**, 211-227.
- Polhamus, B., Dalenius, K., Mackintosh, H., Smith, B. and Grummer-Strawn, L., Pediatric Nutrition Surveillance (2009). Report. Atlanta: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2011.
- Robins, J. M., Rotnitzky, A. and Zhao, L. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc.* **90**, 106-121.
- Rochon, J. (1998). Application of GEE procedures for sample size calculations in repeated measures experiments. *Statist. Medicine* **17**, 1643-1658.
- Rotnitzky, A. and Jewell, N. P. (1990). Hypothesis testing of regression parameters in semi-parametric generalized linear models for cluster correlated data. *Biometrika* **77**, 485-497.
- Scott, A. J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *J. Amer. Statist. Assoc.* **77**, 848-854.
- Self, S. G. and Mauritsen, R. H. (1988). Power/sample size calculations for generalized linear models. *Biometrics* **44**, 79-86.
- Shao, J. (2003). *Mathematical Statistics*. Springer-Verlag, New York.
- Shieh, G. (2000). On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics* **56**, 1192-1196.
- Shih, W. J. (1997). Sample size and power calculations for periodontal and other studies with clustered samples using the method of generalized estimating equations. *Biometrical J.* **39**, 899-908.
- Shih, W. J. (2001). Sample size re-estimation—journey for a decade. *Statist. Medicine* **20**, 515-518.
- Shih, W. J. and Gould, A. L. (1995). Re-evaluating design specifications of longitudinal clinical trials without unblinding when the key response is rate of change. *SStatist. Medicine* **14**, 2239-2248.
- van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge, U.K..
- von Bahr, B. and Esseen, C. (1965). Inequalities for the r th absolute moment of a sum of random variables, $1 \leq r \leq 2$. *Ann. Math. Statist.* **36**, 299-303.

Yin, G. and Shen, Y. (2005). Adaptive design and estimation in randomized clinical trials with correlated observations. *Biometrics* **61**, 362-369.

Zucker, D. M. and Denne, J. (2002). Sample-size redetermination for repeated measures studies. *Biometrics* **58**, 548-559.

Dartmouth Medical School, Hanover, NH 03755, USA.

E-mail: Zhigang.Li@dartmouth.edu

Department of Biostatistics, Columbia University, 722 West 168th Street, New York, NY 10032, USA.

E-mail: im2131@columbia.edu

(Received April 2011; accepted December 2011)