Contemporary Clinical Trials 36 (2013) 244–253

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Covariate-adjusted confidence interval for the intraclass correlation coefficient

Mohamed M. Shoukri [a,c,*], Allan Donner [b], Abdelmoneim El-Dali [d]

[a] National Biotechnology Center, KFSHRC, Saudi Arabia
[b] Schulich School of Medicine and Dentistry, Canada
[c] Al-Faisal University College of Medicine, Saudi Arabia
[d] Department of Biostatistics, KFSHRC, Saudi Arabia

ARTICLE INFO

Article history:
Received 2 April 2013
Received in revised form 4 July 2013
Accepted 7 July 2013
Available online 16 July 2013

Keywords:
Multi-level models
Intra-class correlation
Generalized Estimating Equations
Percentile bootstrap confidence intervals
Monte-Carlo simulations

ABSTRACT

A crucial step in designing a new study is to estimate the required sample size. For a design involving cluster sampling, the appropriate sample size depends on the so-called design effect, which is a function of the average cluster size and the intracluster correlation coefficient (ICC). It is well-known that under the framework of hierarchical and generalized linear models, a reduction in residual error may be achieved by including risk factors as covariates. In this paper we show that the covariate design, indicating whether the covariates are measured at the cluster level or at the within-cluster subject level affects the estimation of the ICC, and hence the design effect. Therefore, the distinction between these two types of covariates should be made at the design stage. In this paper we use the nested-bootstrap method to assess the accuracy of the estimated ICC for continuous and binary response variables under different covariate structures. The codes of two SAS macros are made available by the authors for interested readers to facilitate the construction of confidence intervals for the ICC. Moreover, using Monte Carlo simulations we evaluate the relative efficiency of the estimators and evaluate the accuracy of the coverage probabilities of a 95% confidence interval on the population ICC. The methodology is illustrated using a published data set of blood pressure measurements taken on family members.

1. Introduction

Estimation of the intraclass correlation coefficient (ICC) is relevant to many applications in survey sampling, genetic epidemiology, reliability studies and other fields. In genetic epidemiology it is used as a measure of familial aggregation, i.e. as a measure of similarity of responses among siblings who belong to the same family [1,2], while in interobserver agreement studies, it is used as a measure of reliability [3]. In cluster randomized trials and observational studies that involve aggregates of individuals as sampling units; the ICC measures the degree of similarity among individuals belonging to the same cluster and must be taken into account in both the estimation of sample size and the statistical analysis.

The ICC may be defined as the ratio of the between cluster variance divided by the total variance (the sum of between cluster variance and within cluster variance). When the trait of interest is measured on quantitative scale (e.g. blood pressures, body mass index) the ICC may be estimated using standard expressions for variance components under the assumption of multivariate normality. The most common model used for this purpose is one-way random effects analysis of variance (ANOVA) [4,5]. When the trait is measured on a binary scale, the ANOVA model may be used as well to find a point estimator for the ICC.

* Corresponding author at: National Biotechnology Center, KFSHRC, Saudi Arabia. Tel.: +966 509491454.
E-mail address: shoukri@kfshrc.edu.sa (M.M. Shoukri).

1551-7144
http://dx.doi.org/10.1016/j.cct.2013.07.003

It may be desirable for the purpose of increasing precision to extend the one-way random effects model to include one or more covariates. In this case the selected covariate structure would be expected to affect the estimated ICC and its standard error. In particular we discuss three scenarios: when the covariate is measured at the cluster level, measured at the individual level and when measured at both levels of hierarchy.

In summary, the main objective of this paper is to derive a covariate adjusted variance components estimator for the ICC with corresponding standard error under the three proposed covariate structures and under the assumption of multivariate normality. For the case in which the response variable is measured on a binary scale, we use the Generalized Estimating Equations to find a working correlation estimate, accounting for the measured covariates. We construct confidence limits for the ICC using the non-parametric "Accelerated Bias-corrected percentile" bootstrap known as BCa interval [6,7]. The asymptotic relative efficiency of the ICC estimators corrected for the effect of measured covariates will be assessed relative to the estimator obtained when covariate effects are not accounted for, using Monte Carlo simulations. Moreover, we use simulations to evaluate the coverage probabilities of the 95% confidence intervals on the population parameter. We illustrate the methodology presented in this paper on published arterial blood pressure data collected from nuclear families.

The paper is structured as follows: In Section 2 we introduce the normal linear mixed model and the ANOVA estimator of the ICC. Covariate adjusted ICC estimators and their large sample standard errors are obtained using the delta method, with comparisons made with the standard errors obtained using bootstrap. In Section 3 we discuss the case of a binary outcome, and introduce a BCa confidence interval for the ICC. Section 4 presents an example using a published data set of arterial blood pressures taken on nuclear families.

In Section 5, we design a Monte Carlo study to evaluate the asymptotic efficiency of the proposed estimators and evaluate the adequacy of the constructed confidence intervals on the population values of the ICC. Two SAS macros for nested-bootstrap cluster re-sampling that may be used to facilitate the construction of confidence intervals for ICC in the continuous and binary case are available from the first author.

2. Effect of the covariate structure

In what follows we investigate four statistical models. The first, we call the baseline or the unconditional mean model [8]. The second includes one covariate measured at the cluster level; while the third includes a covariate measured at the individual level, and the fourth includes both types of covariates.

2.1. No covariates (baseline models)

The most commonly used model for estimating the ICC is the one-way random effects model given by:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij} \tag{1}$$

where μ is the grand mean of all measurements in the population, τ_i reflects the effect of cluster i, and ϵ_{ij} is the error term ($i = 1, 2, \ldots$ k; $j = 1, 2, \ldots n_i$). It is assumed that the cluster effects $\{\tau_i\}$ are normally and identically distributed with mean 0 and variance $\sigma_\tau{}^2$, the errors $\{\epsilon_{ij}\}$ are normally and identically distributed with mean 0 and variance $\sigma_\epsilon{}^2$, and the $\{\tau_i\}$ and $\{\epsilon_{ij}\}$ are independent. For this model the ICC, which may be interpreted as the correlation ρ between any two members of a cluster, may be defined as

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2}. \tag{2}$$

It is seen by definition that the ICC is defined as non-negative in this model, a plausible assumption for the application of interest here. We also note that the variance components $\sigma_\tau{}^2$ and $\sigma_\epsilon{}^2$ can be estimated from the one-way ANOVA mean squares [9–11] given in expectation by

$$\mathrm{E(MSB)} = \sigma^2 + n_0\sigma_\tau^2, \tag{3}$$

where, $n_0 = \frac{1}{k-1}\left[N - \sum_{i=1}^{k} n_i^2/N\right]$, and $N = \sum_{i=1}^{k} n_i$.

$$\mathrm{E(MSW)} = \sigma_\epsilon^2.$$

The ANOVA estimator of the population intraclass correlation is thus given by:

$$\widehat{\rho}_0 = \frac{\mathrm{MSB} - \mathrm{MSW}}{\mathrm{MSB} + (n_0 - 1)\mathrm{MSW}} \tag{4}$$

where MSB and MSW are, obtained from the usual ANOVA table, with corresponding sums of squares

$$\mathrm{SSB} = \sum_{i=1}^{k} n_i(\overline{y}_i - \overline{y})^2$$
$$\mathrm{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2.$$

Using the delta method, and to the first order of approximation, the variance of $\widehat{\rho}_0$ [5] is given by:

$$var(\widehat{\rho}_0) = \frac{2(1-\rho)^2(1 + (n_0 - 1)\rho)^2}{n_0^2(k-1)\left(1 - \frac{k}{N}\right)}. \tag{5}$$

Note that when $n_i = n, i = 1, 2, \ldots k$, Eq. (5) reduces to

$$var(\widehat{\rho}_0) = \frac{2(1-\rho)^2(1 + (n-1)\rho)^2}{n(n-1)(k-1)}. \tag{6}$$

This equation differs from the variance expression given in [12,13] by a factor $\left(1 - \frac{1}{k}\right)$, which for large number of clusters is 1. Note also that when $n_i = 1$ (as in twin studies), $var(\widehat{\rho}_0) = k^{-1}(1-\rho^2)^2$.

An approximate (1 − α) 100% confidence interval on ρ may then be constructed as:

$$\widehat{\rho}_0 \pm z_{1-\alpha/2}\sqrt{var(\widehat{\rho}_0)}. \tag{7}$$

Extensive simulations to evaluate the coverage probabilities of the above interval showed [14] that this approximation is adequate over a wide range of the parameter combinations (ρ, k, n). For the different estimators of ICC that will be