Estimation of bias-corrected intraclass correlation coefficient for unbalanced clustered studies with continuous outcomes

Guogen Shan

Department of Epidemiology and Biostatistics, School of Public Health, University of Nevada Las Vegas, Las Vegas, Nevada, USA

ABSTRACT
Intraclass correlation coefficient for data in a clustered study is traditionally estimated from a one-way random-effects model. This model assumes normality for the random cluster effect and the residual effect. When the normality assumption is questionable, we find that the estimated correlation could be much below the nominal level when data are highly skewed or data have low kurtosis. We develop a bias-corrected estimator based on the approach by Thomas and Hultquist for a study with unbalanced cluster sizes. For multivariate normal data or non-normal data with moderate skewness, we compare the performance of the new bias-corrected estimator with two existing estimators with regards to accuracy and precision. When correlation is small, the existing ANOVA estimator works well. When correlation is medium to large, the proposed new estimator has the correlation close to the nominal level, and its mean squared error is smaller than others.

ARTICLE HISTORY
Received 3 April 2020
Accepted 12 August 2020

KEYWORDS
ANOVA; Bias corrected; Clustered study; Inter-rater study; Intraclass correlation coefficient; Skewness and kurtosis

1. Introduction
Clustered data are found in many studies: such as repeated-measure studies, healthcare studies, inter-rater reliability studies, and household surveys. Dependence of outcomes from the same cluster is known as the intraclass correlation coefficient (ICC), which is traditionally estimated from a one-way random-effects model (Demetrashvili, Wit, and van den Heuvel 2016; Donner 1986; Lai 2020; Shan and Ma 2014; Ukoumunne 2002). This estimator $\rho$ can be expressed as a function of the sum of squares from the analysis of variance (ANOVA) table. This estimator is referred to as the ANOVA estimator which can be applied to studies with balanced or unbalanced cluster sizes. Thomas and Hultquist (1978) developed an asymptotic estimator for $\rho$ when a study has unbalanced cluster sizes (named as the TH estimator), and they showed that the TH estimator works very well. These two estimators were developed under the normality assumption. When this assumption is not met, a non-parametric measure proposed by Rothery (1979) may be considered. However, this rank-based estimator tends to be over-estimated in many configurations.

In one of our ongoing projects to study the impact of cognitive impairment on patients with Parkinson’s Disease (PD), skewness of some biomarkers could be very small (e.g. $-4$), which is considered to be highly skewed. A large number of existing studies showed that the ANOVA estimator works not too bad for skewed data when kurtosis is very large (e.g. over 5)
When kurtosis is large, data have a huge spike and outcomes are close to each other. It is not clear whether the estimated $q$ is close to the nominal level or not when kurtosis is small or even negative. For this reason, we conduct extensive simulation studies to investigate the robustness of $q$ estimators to non-normality with a wide range of skewness and kurtosis. These findings provide guidance on the $q$ estimation for other researchers.

Recently, Atenafu et al. (2012) developed a bias-corrected estimator for $q$ by using the second-order Taylor series expansion for a balanced study. Following their approach, we develop a bias-corrected estimator for a study with unbalanced cluster sizes. In a balanced study, the ratio of sum of square among groups and its variance follows a chi-squared distribution (Donner and Wells 1986). This important property is generally not satisfied in an unbalanced study. For this reason, Thomas and Hultquist (1978) proposed an asymptotic version of that ratio: the variance of cluster means is used to approximate the sum of square among groups. Similarly, we use the second-order Taylor expansion of the log transformation to derive the bias of the TH estimator. We then compare the new bias-corrected estimator with the existing ANOVA estimator and the TH estimator with regards to accuracy and precision.

We organize this article as follows. In Sec. 2, we first introduce the two commonly used ICC estimators, and then derive the biased-corrected estimator. In Sec. 3, we conduct simulation studies to compare the three estimators with regards to accuracy and precision under a wide range of configurations. An example from the Parkinson’s Progression Markers Initiative (PPMI) is used to illustrate the application of the new estimator and the existing estimators. Lastly, we provide some comments and suggestions in Sec. 4.

2. Methods

Suppose a clustered study has $k$ clusters, and the $i$-th cluster has $n_i$ subjects. When $n_1 = n_2 = \cdots = n_k$, it becomes a balanced study. Atenafu et al. (2012) developed a bias-corrected estimator for $q$ in a balanced study from a one-way random-effects model. We extend their approach to a study with unbalanced cluster sizes which often occur in practice (e.g. different length of follow up, or unequal numbers of raters for each participant). The one-way random-effects model is expressed as

$$Y_{ij} = \mu + a_i + e_{ij}$$

where $i = 1, 2, ..., k, j = 1, 2, ..., n_i$. $Y_{ij}$ is the outcome for the $j$-th subject in the $i$-th cluster. The random cluster effect $a_i$ is assumed to follow $N(0, \sigma_a^2)$, and the residual effect $e_{ij}$ follows $N(0, \sigma_e^2)$. These two effects are assumed to be independent from each other. It is obvious that this model is under normal assumptions of the random effect and the residual effect. The parameter of interest $\rho$ is defined as

$$\rho = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2}. \quad (1)$$

The ICC can be estimated by using the sums of squares from the ANOVA table

$$\hat{\rho}_{\text{ANOVA}} = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{\sigma}_e^2} = \frac{\text{SSA}}{k-1} - \frac{\text{SSE}}{N-k} \frac{\text{SSA}}{N-k} + (m-1) \frac{\text{SSE}}{N-k}, \quad (2)$$

where $m = \frac{1}{k-1} \left[ N - \sum_{i=1}^{k} \frac{n_i^2}{N} \right]$ is the harmonic mean of the cluster sizes, $\text{SSA} = \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y}_-)^2$ is the sum of square among clusters, $\text{SSE} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$ is the sum of square within clusters, $\bar{Y}_i$ is the mean of outcomes from the $i$-th cluster, and $\bar{Y}_-$ is the grand mean of all outcomes.
The ICC in Equation (1) can be reorganized as
\[ \rho = \frac{\sigma_a^2 / \sigma_e^2}{1 + \sigma_a^2 / \sigma_e^2} = \frac{F}{1 + F}, \]
where \( F = \sigma_a^2 / \sigma_e^2 \). It is well known that \( SSA / (m \sigma_a^2 + \sigma_e^2) \) follows a chi-squared distribution in an unbalanced study if and only if \( \sigma_a^2 = 0 \) (Donner and Wells 1986; Thomas and Hultquist 1978). Thus, the estimator for \( F \) in a balanced study by Atenafu et al. (2012) can be not directly extended to an unbalanced study.

For an unbalanced study, Thomas and Hultquist (1978) developed a new ratio test by using the sample variance of the cluster means \( S^2_c \) and \( SSE \), where
\[ S^2_c = \frac{1}{k - 1} \left[ \sum_{i=1}^{k} Y_i^2 - \frac{1}{k} \left( \sum_{i=1}^{k} Y_i \right)^2 \right]. \]
They showed that \( m(k - 1)S^2_c / (m \sigma_a^2 + \sigma_e^2) \) asymptotically follows a chi-squared distribution with the degree freedom of \( (k - 1) \). Let \( SSA = m(k - 1)S^2_c \). It should be noted that \( SSA \) is SSA in a balanced study. We then have the following theorem for an asymptotically unbiased estimator for \( F \) in a study with unbalanced cluster sizes.

**Theorem 2.1.** An asymptotically unbiased estimator for \( F \) is
\[ \hat{F} = \frac{(N - k - 2) \overline{SSA} \overline{SSE} - (k - 1)}{m(k - 1)}. \]

**Proof.** From the results by Thomas and Hultquist (1978), we have
\[ \frac{SSA}{m \sigma_a^2 + \sigma_e^2} \overset{asy}{\sim} \chi^2(k - 1), \]
and
\[ \frac{SSE}{\sigma_e^2} \sim \chi^2(N - k). \]
Then, an asymptotically unbiased estimator for \( F \) is presented as
\[ \hat{F} = \frac{(N - k - 2) \overline{SSA} \overline{SSE} - (k - 1)}{m(k - 1)}. \]
By using \( E(\overline{SSA}) = (k - 1)(m \sigma_a^2 + \sigma_e^2) \) and \( E(1/SSE) = 1/[\sigma_e^2(N - k - 2)] \), it is easy to show that \( \hat{F} \) is an unbiased estimator of \( F: E(\hat{F}) = F \).

This estimator was originally proposed by Thomas and Hultquist (1978) who extended the results from a balanced study by Searle (1971). When \( \hat{F} \) is derived from the aforementioned theorem, an estimator for \( \rho \) is given as
\[ \hat{\rho}_{TH} = \frac{\hat{F}}{1 + \hat{F}}. \]
A log transformation is applied on \( \rho \) to compute the bias: \( \log \rho = \log F - \log (F + 1) \). Then, the second-order Taylor expansion of the log function can be used as an approximation of the log function:
\[ \log (\hat{F} + d) \approx \log (F + d) + \frac{1}{F + d} (\hat{F} - F) - \frac{1}{2(F + d)^2} (\hat{F} - F)^2, \]
where \( d = 0 \), and 1. It follow that
\[
\log\left( \frac{F}{b_F} \right) / C_0 \log\left( \frac{F + 1}{b_F + 1} \right) / C_2 - \log\left( \frac{F}{b_F} \right) / C_0 \log\left( \frac{F + 1}{b_F + 1} \right) / C_2.
\]
Thus, the estimated bias of \( \hat{\rho}_{TH} \) is
\[
\frac{1}{2} \left[ \frac{1}{F^2} - \frac{1}{(F + 1)^2} \right] \hat{\text{Var}}(\hat{F}), \quad \text{where} \quad \hat{\text{Var}}(\hat{F}) = \frac{N-k-2}{m^2(k-1)} \times \left[ \frac{k+1}{N-k-4} - \frac{k-1}{N-k-2} \right] (m\hat{F} + 1)^2.
\]
Therefore the bias-corrected estimator for \( \rho \) is
\[
\hat{\rho}_{THbc} = \hat{\rho}_{TH} \times \exp \left[ \frac{1}{2} \left( \frac{1}{F^2} - \frac{1}{(F + 1)^2} \right) \hat{\text{Var}}(\hat{F}) \right].
\] (4)
Since the bias is always positive, the bias-corrected estimator \( \hat{\rho}_{THbc} \) is always larger than \( \hat{\rho}_{TH} \). When the estimated \( F \) is small, the quantity of \( \frac{1}{2} \left( \frac{1}{F^2} - \frac{1}{(F + 1)^2} \right) \) in the bias-corrected estimator would be too large. For this reason, the second order Taylor expansion of \( \log(1 - \rho) \) is used to compute the bias, and the bias-corrected estimator is
\[
\hat{\rho}_{THbc} = 1 - (1 - \hat{\rho}_{TH}) \times \exp \left[ - \frac{1}{2(F + 1)^2} \hat{\text{Var}}(\hat{F}) \right].
\] (5)
We use 0.1 as the threshold for \( \hat{F} \) on whether the estimator in Equation (4) or the one in Equation (5) is used as the bias-corrected estimator.

3. Results

We compare the performance of the three ICC estimators based on the ANOVA method, the TH method, and the bias-corrected TH method (\( \hat{\rho}_{ANOVA}, \hat{\rho}_{TH}, \) and \( \hat{\rho}_{THbc} \)) with regards to accuracy and precision when data are from multivariate normal distributions and non-normal distributions. The clustered data are simulated by using the R function \textit{mvnormnorm} from the \textit{semTools}

Figure 1. Average \( \rho \) value and the mean squared error (MSE) for the multivariate normal data when \( k = 40 \) and \( n = 30 \) (top), and the multivariate non-normal data when \( k = 40, n = 100, \) skewness = 0.5, and kurtosis = 0.8 (bottom).
package, with the mean vector of zero (R codes are presented in the Supplementary material). The algorithm for simulating multivariate normal and non-normal data in the R function was developed by Vale and Maurelli (1983). Skewness and kurtosis can be specified in this R function for non-normal data. When skewness is below $-0.5$ or above $0.5$, such data are considered as highly skewed. We first investigate $q$ estimators with data normally distributed or moderately skewed: skewness $=-0.5, 0,$ and $0.5$. The considered values of kurtosis are: $-0.8, 0,$ and $0.8$. When both skewness and kurtosis are zero, data are normally distributed. The pre-specified ICC values are set as: $0.1, 0.3, 0.5, 0.7,$ and $0.9$. The variance $r^2 = r^2_a + r^2_e$ is assumed to be $1$. Suppose $n$ is the maximum of cluster sizes: $n = \max\{n_i : i = 1, 2, \ldots, k\}$. The size of each cluster is uniformly distributed between $1$ and $n$. For each given configuration $(k, n, \sigma^2, \text{skewness, kurtosis})$, we generate $B = 10,000$ data sets where each data set has $k$ clusters.

Figure 1 shows the average of $\rho$ estimators with the mean squared error (MSE) for the multivariate normal data (top) when $k = 40$ and $n = 30$, and the multivariate non-normal data (bottom) when $k = 40, n = 100$, skewness $= 0.5$, and kurtosis $= 0.8$. The MSE is defined as

$$MSE = \frac{1}{B} \sum_{b=1}^{B} (\hat{\rho}_b - \rho)^2,$$

where $\hat{\rho}_b$ is the estimated $\rho$ from the $b$-th simulation and $\rho$ is the pre-specified ICC. It can be seen from the figure that the three methods have similar average correlations being very close to the nominal level. For the multivariate normal case, the new $\hat{\rho}_{TH_{bc}}$ is slightly above the nominal level while the other two methods are slightly below the nominal level. When $\rho$ is small (e.g. $0.1$ or $0.3$), the traditionally used estimator $\hat{\rho}_{ANOVA}$ has the smallest MSE, and $\hat{\rho}_{TH_{bc}}$ is the best estimator when $\rho$ is medium or large. Similar results are observed when data are non-normal.

We present the relationship between the estimated ICC and variance ($\sigma^2$) from $0.3$ to $1,000$ in Figure 2 for multivariate normal data when the pre-specified $\rho$ is $0.1$ (top) and $0.7$ (bottom). The other parameters are: $k = 30$ and $n = 40$. The average of the estimated $\rho$ values for each method is
Figure 3. Average of the ANOVA estimator as a function of skewness and kurtosis when $\rho = 0.1$ (top), 0.5 (middle), and 0.9 (bottom).
consistent as $\sigma^2$ changes. The MSE also seems independent of $\sigma^2$. These plots show a detailed comparison between the three ICC estimators, having the results similar to the findings in Figure 1. When $\rho = 0.1$, the ANOVA method has the estimate close to the nominal level, and the new bias-corrected estimator becomes the best when $\rho = 0.7$.

In the existing literature for non-normal data, skewness is often small or skewness is above 1 but kurtosis is large. We conduct simulation studies with skewness from $-2$ to 2, and kurtosis from $-2$ to 3 ($-2$, $0$, $0.5$, and $3$). Figure 3 shows the average of the $\hat{p}_{ANOMA}$ when $\rho = 0.1$ (top), $0.5$ (middle), and $0.9$ (bottom). The results from three plots are similar to each other. It can be seen that the average of $\hat{p}_{ANOMA}$ is close to the nominal level when skewness between $-1$ and 1 and kurtosis is above 0.5. When kurtosis is small or even negative, the average of $\hat{p}_{ANOMA}$ could be much below the nominal level, especially in the cases of kurtosis $= -2$. When kurtosis is near 0 and skewness is between 0.5 and 1 (or $-1$ to $-0.5$), the average of $\hat{p}_{ANOMA}$ is about 10% below the nominal level. This simulation shows that the larger kurtosis is, the higher the estimated ICC is.

3.1. Example

The PPMI study is an international study having 33 clinical sites with the goal to develop new treatments for PD by discovering new and promising biomarkers. The test battery of Hopkins Verbal Learning Test-Revised (HVLT-R) is one of the tests in the PPMI study to measure verbal learning and memory. For PD patients in the PPMI study, the average HVLT-R test score is 26.1, and the range of visits is from 1 to 6. The estimated ICC based on the three considered methods are: $\hat{p}_{ANOMA} = 0.170$, $\hat{p}_{TH} = 0.236$, and $\hat{p}_{THbc} = 0.244$. The TH estimator and the new estimator are higher than the traditionally used ANOVA estimator. The average of skewness values across the visits is $-0.487$, which is considered as moderately skewed. Kurtosis is close to zero as 0.024.

The Montreal Cognitive Assessment (MoCA) is another measure that is designed as a screening instrument for cognitive impairment, with the score range between 0 and 30. A high score is considered to be not impaired. In the PPMI study, the average score of MoCA for PD is 27.5, which is similar to results from other reported studies (Bayram et al. 2020; Cummings and Fulkerson 2018; Shan et al. 2018). The ICC is estimated as: $\hat{p}_{ANOMA} = 0.508$, $\hat{p}_{TH} = 0.423$, and $\hat{p}_{THbc} = 0.428$. Skewness has the range from $-0.22$ to $-4.58$. The skewness seems a decreasing function of visit. Meanwhile, the estimated kurtosis is an increasing function of visit and it is as high as 32 at the last visit.

4. Conclusion

In this article, we develop an asymptotic bias-corrected estimator for the ratio of $\sigma_a^2$ and $\sigma_e^2$ in the ICC calculation. We show that this new estimator performs better than the existing TH estimator and the traditionally used ANOVA estimator with regards to accuracy and precision for medium to large ICC when data are not highly skewed (Shan 2015, 2018; Shan et al. 2018; Shan and Zhang 2019). The bias in the $F$ estimate is computed from a second-order Taylor expansion of a log transformation. We explore the third-order Taylor expansion, and the results are similar to the proposed estimator (Shan 2020a, 2020b; Shan and Chen 2018; Shan et al. 2016). The impact of the estimated $\rho$ from the third moment is relatively small as compared to that from the second moment.

5. Discussion

We conduct extensive simulation studies to show that ICC could be significantly underestimated when data are highly skewed or kurtosis is near zero or negative (Shan et al. 2010, 2020).
Kurtosis seems to have a positive relationship with the ICC. For this reason, simulations from existing literature showed good accuracy of ICC methods as kurtosis is often large. Based on this finding, we would recommend researchers to use the ICC methods with caution when kurtosis is small or negative and skewness is below −1 or above 1. In such cases, non-parametric methods for ICC may be considered. The rank based measure by Rothery (1979) does not have estimates close to the nominal level. Thus, it would be interesting to develop new non-parametric measures to improve the accuracy (Shan 2014; Shan et al. 2018; Shan, Young, and Kang 2014; Zhang and Guogen 2020). Alternatively, the approaches to calculate Pearson correlation when normality is not met, could be possibly extended for clustered data (Pearson 1947).

In addition to ICC estimate, confidence interval of ICC is also another important research topic. Hartung and Knapp (2000) proposed a Wald type confidence interval for ICC in the one-way random effects model. Liu, Zhao, and Li (2016) utilized the confidence interval distribution approach by combining asymptotic confidence distributions in the one-way random effects model. Recently, Al–Sarraj, von Brömssen, and Forkman (2019) proposed a new generalized prediction interval based on the approximate methods of Satterthwaite and Kenward and Roger for random effects in linear random–effects models, and it was shown to have the coverage probabilities being closer to the nominal level than the traditional intervals based on the restricted maximum likelihood method.

Methods for constructing confidence intervals for variance component ratios in general unbalanced mixed models are developed. The methods are based on inverting the distribution of the signed root of the log-likelihood ratio statistic constructed from either the restricted maximum likelihood or the full likelihood (Jiang, Cao, and Shan 2020). As this distribution is intractable, the inversion is rather based on using a saddlepoint approximation to its distribution. Apart from Wald’s exact method, the resulting intervals are unrivaled in terms of achieving accuracy in overall coverage, underage, and overage. Issues related to the proper “reference set” with which to judge the coverage as well as issues connected to variance ratios being nonnegative with lower bound 0 are addressed. Applications include an unbalanced nested design and an unbalanced crossed design.

It can be seen from the PPMI example, the estimated skewness and kurtosis are not small for the MoCA measure. One possible reason is that such measures always have a boundary, and a certain population (e.g. PD without MCI) has the outcome close to the upper boundary due to learning effect. Thus, data are likely to be skewed. In addition, as disease status progresses in the disease groups, outcomes are skewed to the worse direction. Meanwhile, kurtosis could change as time goes. Generally, we assume a constant skewness and kurtosis in the ICC calculation which can be violated in practice. When this occurs, homogeneity of skewness and kurtosis should be tested.

Acknowledgments

The authors are very grateful to Editor, Associate Editor, and two reviewers for their insightful comments that help improve the manuscript significantly. Data used in the preparation of this article were obtained from the Parkinson’s Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI - a public-private partnership - is funded by the Michael J. Fox Foundation for Parkinson’s Research and funding partners, including Abbvie, Allergan, Avid, Biogen, Biolegend, Bristol-Myers Squibb, JenaLi, GE Healthcare, Genetnetch, GlaxoSmithKline, Lilly, Lundbeck, Merck, Mesoscale Discovery, Pfizer, Piramal, Prevail, Roche, SanofiGenzyme, Servier, Takeda, Teva and UCB.

ORCID

Guogen Shan http://orcid.org/0000-0001-8690-6599
Shan, G., A. Vexler, G. E. Wilding, and A. D. Hutson. 2010. Simple and exact empirical likelihood ratio tests for normality based on moment relations. *Communications in Statistics - Simulation and Computation* 40 (1): 129–46. doi:10.1080/03610918.2010.532896.

Shan, G., G. E. Wilding, A. D. Hutson, and S. Gerstenberger. 2016. Optimal adaptive two-stage designs for early phase II clinical trials. *Statistics in Medicine* 35 (8):1257–66. doi:10.1002/sim.6794.

Shan, G., D. Young, and L. Kang. 2014. A new powerful nonparametric rank test for ordered alternative problem. *PLoS ONE*, 9 (11):e112924. doi:10.1371/journal.pone.0112924.

Shan, G., and H. Zhang. 2019. Two-stage optimal designs with survival endpoint when the follow-up time is restricted. *BMC Medical Research Methodology* 19 (1):74. doi:10.1186/s12874-019-0696-x.

Thomas, J. D., and R. A. Hultquist. 1978. Interval estimation for the unbalanced case of the one-way random effects model. *The Annals of Statistics* 6 (3):582–87. doi:10.1214/aos/1176344202.

Ukoumunne, O. C. 2002. A comparison of confidence interval methods for the intraclass correlation coefficient in cluster randomized trials. *Statistics in Medicine* 21 (24):3757–74. doi:10.1002/sim.1330.

Vale, C. D., and V. A. Maurelli. 1983. Simulating multivariate nonnormal distributions. *Psychometrika* 48 (3): 465–71. doi:10.1007/BF02293687.

Zhang, H., and S. Guogen. 2020. Letter to the Editor: A novel confidence interval for a single proportion in the presence of clustered binary outcome data (SMMR, 2019). *Statistical Methods in Medical Research* 29 (2): 636–37.