A strictly positive estimator of intra-cluster correlation for the one-way random effects model

Siegfried Gabler Matthias Ganninger Partha Lahiri

May 12, 2011

1 Introduction

Estimates of intra-cluster correlations are routinely produced for designing and analyses of large cross-national sample surveys like the European Social Survey (see Ganninger 2010:19). Kish (1962) defined intra-cluster correlation as the ratio of the between cluster variance to the total variance. Thus, according to Kish's definition, intra-cluster correlation is a strictly positive parameter. It is well-known (Wang et al. 1991) that standard variance component methods, such as the ANOVA method, can frequently produce negative estimates, especially when the true intra-cluster correlation and the number of clusters are small, a situation that can arise in practice (Killip et al. 2004). A standard solution to this problem is to truncate the intra-cluster correlation estimate to 0. (Campbell et al. 2005).

In this paper, we present a new estimator of the intra-cluster correlation for the one-way random effects model and prove that it is strictly positive. We compare the proposed estimator with the ANOVA estimator using a Monte Carlo simulation study.

2 The one-way random effects model

Let y_{ij} denote the value of the study variable for the jth unit of the ith cluster ($i = 1, \ldots, m;\ j = 1, \ldots, n_i$). We consider the following one-way random effects model:

$$y_{ij} = \mu + v_i + \varepsilon_{ij}, \tag{1}$$

where μ is a fixed unknown parameter; v_i and ε_{ij} are independent random errors with $v_i \sim N\left(0, \sigma_v^2\right)$ and $\varepsilon_{ij} \sim N\left(0, \sigma_\varepsilon^2\right)$, $i = 1, \ldots, m; j = 1, \ldots, n_i$. Under this model, the expected value and the covariances of the study variable are given by

$$\begin{aligned} E\left(y_{ij}\right) &= \mu \\ Cov\left(y_{ij}, y_{i'j'}\right) &= \begin{cases} \sigma_v^2 + \sigma_\varepsilon^2 & i = i', j = j' \\ \sigma_v^2 & i = i', j \neq j' \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

respectively. Thus, under the model,

$$\begin{aligned} E\left(\bar{y}_i\right) &= \mu \\ Var\left(\bar{y}_i\right) &= \frac{\sigma_\varepsilon^2}{n_i} + \sigma_v^2 = \sigma_\varepsilon^2 \left(\frac{1}{n_i} + \lambda\right) \\ \lambda &= \frac{\sigma_v^2}{\sigma_\varepsilon^2} \quad . \end{aligned}$$

The intra-cluster correlation coefficient is defined as

$$\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\varepsilon^2} = \frac{\lambda}{1 + \lambda} > 0 \quad . \tag{2}$$

The problem is to find an estimator $\hat{\lambda}$ such that the resulting estimator $\hat{\rho}$ fulfils the constraint $\hat{\rho} > 0$.

Define $MSW = \frac{1}{n_T - m}\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$ and $MSB = \frac{1}{m-1}\sum_{i=1}^{m} n_i(\bar{y}_i - \bar{y})^2$. Following Ghosh and Lahiri (1987), we estimate λ by

$$\hat{\lambda}_{ANOVA} = b\frac{MSB - MSW}{MSW} = b(F-1) \quad ,$$

where $b = \frac{m-1}{\sum_{i=1}^{m} n_i\left(1 - \frac{n_i}{\sum n_j}\right)}$ and $F = \frac{MSB}{MSW}$. The well-known ANOVA estimator of ρ is given by

$$\hat{\rho}_{ANOVA} = \frac{\hat{\lambda}_{ANOVA}}{1 + \hat{\lambda}_{ANOVA}} \quad . \tag{3}$$

Note that $\hat{\rho}_{ANOVA}$ is negative that if $MSW > MSB$ in which case $\hat{\rho}_{ANOVA}$ is usually truncated to zero.

3 A strictly positive estimator of ρ

First consider the case when both μ and σ_ε^2 are known. We define an adjusted likelihood function of λ as $L^c_{\text{adj}}(\lambda|\mathbf{y}) = \lambda^{\frac{c}{2}} L(\lambda|\mathbf{y})$ for any arbitrary but fixed $0 < c < m$. For the special case of $c = 2$, see Li and Lahiri (2010).

The maximization of $L^c_{\text{adj}}(\lambda|\mathbf{y})$ is equivalent to that of $\frac{c}{2}\log(\lambda) + \log[L(\lambda|\mathbf{y})]$. The adjusted maximum likelihood estimator of λ is obtained as a solution of

$$\frac{dl\left(\mu, \sigma_\varepsilon^2, \lambda|\mathbf{y}\right)}{d\lambda} + \frac{c}{2\lambda} = 0,$$

which is equivalent to

$$\frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{m}\frac{(\bar{y}_i - \mu)^2}{\left(\frac{1}{n_i} + \lambda\right)^2} - \sum_{i=1}^{m}\left(\frac{1}{n_i} + \lambda\right)^{-1} + \frac{c}{\lambda} = 0 \quad . \tag{4}$$

3.1 The balanced case

In this section, we assume $n_i = n; b_i = \frac{1}{n_i} = \frac{1}{n} \quad \forall i$ so that $b = \frac{1}{n}$. Defining $a_i = \frac{(\bar{y}_i - \mu)^2}{\sigma_\varepsilon^2}$ and $a = \frac{1}{m}\sum_{i=1}^{m} a_i$, we can write (4) as

$$f_c(\lambda) = \frac{ma}{(\lambda + b)^2} - \frac{m}{(\lambda + b)} + \frac{c}{\lambda} \tag{5}$$

Theorem 1. *For $m > 2$ there exists exactly one positive λ with $f_c(\lambda) = 0$ which has the solution*

$$\tilde{\lambda}_c = \frac{ma - (m-2c)b + \sqrt{(ma - (m-2c)b)^2 + 4c(m-c)b^2}}{2(m-c)} \tag{6}$$

Remark 1. *The second solution of the quadratic equation obtained from equation (5) is obviously negative for $0 < c < m$.*

The proof follows from

Lemma 1. *The inequality $\tilde{\lambda}_c > 0$ can be sharpened by*

$$\tilde{\lambda}_c > \frac{c}{m-c}(a+b) > \frac{c}{m-c}b \quad .$$

Remark 2. *For $c = 0$ the maximum likelihood solution is*

$$\tilde{\lambda}_0 = \begin{cases} 0 & \text{for } a < b \\ a - b & \text{for } a \geq b \end{cases} \quad .$$

For arbitrary positive a, b and c with $0 < c < m$ we define

$$\lambda_c = \tilde{\lambda}_c - \frac{c}{m-c}(a+b) = \frac{(m-2c)a - mb + \sqrt{(ma - (m-2c)b)^2 + 4c(m-c)b^2}}{2(m-c)}. \tag{7}$$

Lemma 2. *For arbitrary positive* a, b

$$\lambda_c = \frac{(m-2c)\,a - mb + \sqrt{(ma-(m-2c)\,b)^2 + 4c\,(m-c)\,b^2}}{2\,(m-c)}$$

is a monotone increasing function of c *in* $(0, m)$.

Remark 3. *From Lemma 2 it follows*

$$\frac{a-b+|a-b|}{2} = \lambda_0 < \lambda_c < \frac{a^2}{(a+b)} = lim_{c\to m}\lambda_c \tag{8}$$

Lemma 3. *For arbitrary positive* a, b

$$\lambda_c = \frac{(m-2c)\,a - mb + \sqrt{(ma-(m-2c)\,b)^2 + 4c\,(m-c)\,b^2}}{2\,(m-c)}$$

is a concave function of c *in* $(0, m)$.

Lemma 2 and Lemma 3 show that λ_c is a monotone increasing and concave function of c where $c \in (0, m)$.

Now consider the case when μ and σ_ε^2 are unknown. We propose to estimate them by their usual unbiased estimators:

$$\begin{aligned} \hat{\mu} &= \bar{y} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij} = \frac{1}{m}\sum_{i=1}^{m}\bar{y}_i \\ \hat{\sigma}_\varepsilon^2 &= \frac{1}{m\,(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2 = MSW \qquad . \end{aligned}$$

Furthermore, we define $MSB = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2$ and

$$\hat{a} = \frac{m-1}{mn}\frac{MSB}{MSW} \qquad .$$

Then, from

$$E\left(\hat{\sigma}_\varepsilon^2\right) = E\,(MSW) = \sigma_\varepsilon^2$$

and

$$E\left(\frac{1}{n}MSB\right) = \frac{\sigma_\varepsilon^2}{n} + \sigma_v^2 = \sigma_\varepsilon^2(\lambda + b)$$

it follows that

$$E\,(\hat{a}) \approx \frac{m-1}{m}(\lambda + b)$$

If we now substitute $\frac{m}{m-1}\hat{a}$ for a in equation (6) we get as an estimator of λ

$$\hat{\lambda}_c = \frac{(m-2c)\,\frac{m}{m-1}\hat{a} - mb + \sqrt{\left(m\frac{m}{m-1}\hat{a} - (m-2c)\,b\right)^2 + 4c\,(m-c)\,b^2}}{2(m-c)} \qquad . \tag{9}$$

This estimator has the property $\hat{\lambda}_c > 0$ and as a consequence of inequality (8) has expected value

$$\lambda \approx E\left(\frac{m}{m-1}\hat{a} - b\right) < E\left(\hat{\lambda}_c\right) < E\left(\frac{m}{m-1}\hat{a}\right) \approx \lambda + b \qquad .$$

3.2 The unbalanced case

For the unbalanced case, the (adjusted) likelihood method can have local maxima even for $c = 0$, which corresponds to the usual maximum likelihood function. Using a theorem of Descartes (Anderson et al. 1998), it can be shown (see Gabler et al. 2011) that the adjusted maximum likelihood function is quasi-concave in $\mathbb{R}_+$ when $c = m - 1$. Therefore, there is a unique positive maximum in $\mathbb{R}_+$.

An extension of the balanced case is obtained by substituting in (9) $\hat{a}$ by $\hat{a} = \frac{m-1}{m} b \frac{MSB}{MSW}$. Thus, the estimator of λ is given by

$$\hat{\lambda}_c = \frac{(m-2c)\, b \frac{MSB}{MSW} - mb + \sqrt{\left(mb\frac{MSB}{MSW} - (m-2c)\, b\right)^2 + 4c\,(m-c)\, b^2}}{2(m-c)}. \tag{10}$$

Note that $\hat{\lambda}_c > 0$ and the inequality (8), we obtain

$$\lambda \approx E\left(b\frac{MSB}{MSW} - b\right) < E\left(\hat{\lambda}_c\right) < E\left(b\frac{MSB}{MSW}\right) \approx \lambda + b.$$

Using the fact $c\,(m-c) > 0$ and comparing the new estimator $\hat{\lambda}_c$ with the ANOVA estimator

$$\hat{\lambda}_{ANOVA} = b\left(\frac{MSB}{MSW} - 1\right) \tag{11}$$

we get $\hat{\lambda}_c > \hat{\lambda}_{ANOVA}$.

4 Simulation Study

We compare the proposed estimator with the ANOVA estimator using a Monte Carlo simulation study. We consider a universe of $M = 1000$ clusters, each of size $N_i = N = 500$. A set of study variables $\mathbf{y}$ is generated using model (1) with different combinations of $(\sigma_\varepsilon^2, \sigma_v^2)$. Thus, twelve different $\mathbf{y}$ vectors are generated with different values of rho. From these populations, repeated two-stage samples are drawn with different sampling fractions at the first and second stages. We choose three different values for $m = \{30, 250, 625\}$. For a fair comparison, the sample sizes for different sample design settings are kept approximately the same. In particular, we choose $n_i = n = \{85, 10, 4\}$, respectively, such that $n \cdot m \approx 2500$.

Figures 1 and 2 display an overview of the distribution of the 1000 estimates under twelve different scenarios described above for large ($n = 85$) and small ($n = 4$) cluster sizes. In Figure 1, ANOVA estimates produces some negative values for small values of ρ_{M}, i.e. in scenarios displayed in the two lower rows of the panel plot. Our estimates remain positive all the time and in general improve on the ANOVA estimates. This effect is more pronounced when small cluster sizes are considered as can be seen in Figure 2. Here, the ANOVA estimator can yield negative values even for moderate ρ_{M} as can be seen in the panels to the right of the upper row of the plot.

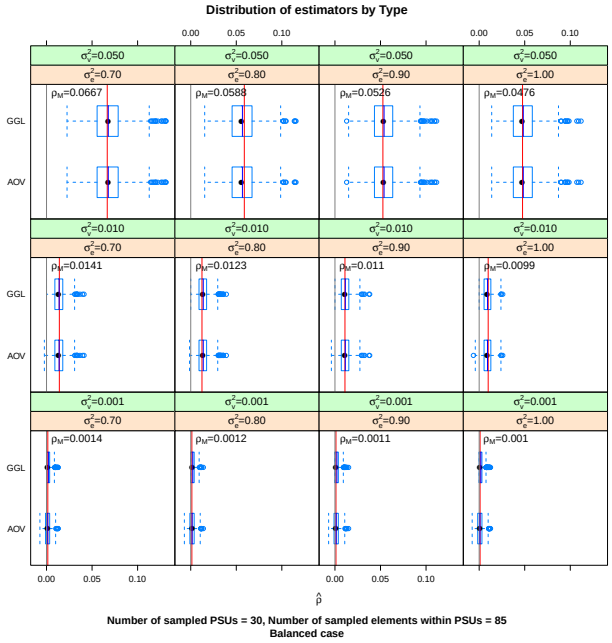

Figure 1: Distribution of 1000 estimates under different scenarios for continuous study variable - large cluster sizes (balanced)

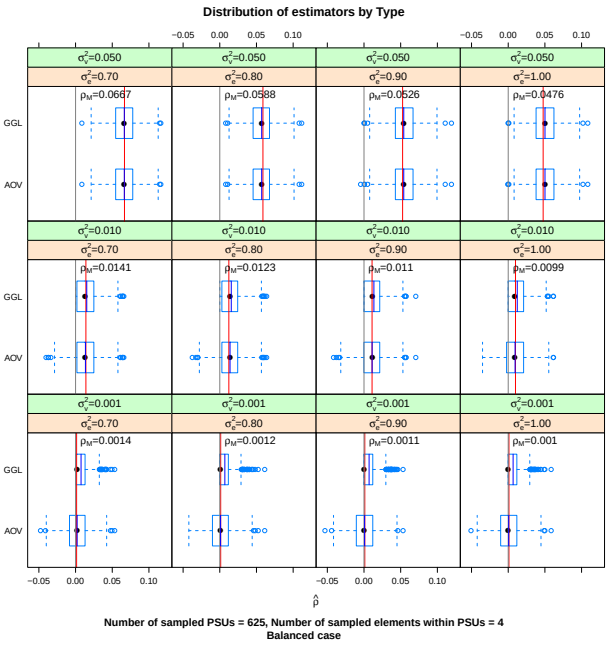

Figure 2: Distribution of 1000 estimates under different scenarios for continuous study variable - small cluster sizes (balanced)

5 Summary

For small values of ρ, our simulation studies show that the choice of c has an influence on the point estimates. Overall, the proposed estimator of the intra-cluster correlation is appealing in a wide range of scenarios. If c is sufficiently small, the proposed estimator is better than the usual ANOVA estimator in terms of the mean square error criterion.

References

Anderson, B., Jackson, J., and Sitharam, M. (1998). Descartes' rule of signs revisited. *American Mathematics Monthly*, 105:447–451.

Campbell, M. K., Fayers, P. M., and Grimshaw, J. M. (2005). Determinants of the intracluster correlation coefficient in cluster randomized trials: the case of implementation research. *Clinical Trials*, 2:99–107.

Ganninger, M. (2010). *Design Effects: Model-based versus Design-based Approach*. Number 3 in GESIS–Series. GESIS, Bonn.

Ghosh, M. and Lahiri, P. (1987). Robust empirical bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82:1153–1162.

Killip, S., Mahfoud, Z., and Pearce, K. (2004). What is an intracluster correlation coefficient? crucial concepts for primary care researchers. *Annals of Family Medicine*, 2(3):204–208.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57:92–115.

Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101:882–892.

Wang, C. S., Yandell, B. S., and Rutledge, J. J. (1991). Bias of maximum likelihood estimator of intraclass correlation. *Theoretical and Applied Genetics*, 82:421–424.