STATISTICS IN TRANSITION new series, March 2019
Vol. 20, No. 1, pp. 153–169, DOI 10.21307/stattrans-2019-009

AN ALTERNATIVE MATRIX TRANSFORMATION TO THE F TEST STATISTIC FOR CLUSTERED DATA

Sukanya Intarapak[1], **Thidaporn Supapakorn** [2]

ABSTRACT

For the regression analysis of clustered data, the error of cluster data violates the independence assumption. Consequently, the test statistic based on the ordinary least square method leads to incorrect inferences. To overcome this issue, the transformation is required to apply to the observations. In this paper we propose an alternative matrix transformation that adjusts the intra-cluster correlation with Householder matrix and apply it to the F test statistic based on generalized least squares procedures for the regression coefficients hypothesis. By Monte Carlo simulations of the balanced and unbalanced data, it is found that the F test statistic based on generalized least squares procedures with Adjusted Householder transformation performs well in terms of the type I error rate and power of the test.

Key words: adjusted Householder, clustered data, F test statistic, generalized least squares, intra-cluster correlation.

1. Introduction

Clustered data arise in many situations such as health research (multiple patients within a hospital) (see Miall and Oldham (1955) and Ng et al. (2004)), education study (multiple students within a school) (see McCulloch and Shayle (2001)) and biological science (multiple children within a family) (see Agarwal et al. (2005)). Clustered data are characterized as data that can be classified into a number of distinct groups or clusters (see Galbraith et al. (2010)). Any two responses from different clusters are independent, but pairs of responses within clusters are correlated, and the correlation is the same for all pairs of individuals from the same cluster, which is called the intra-cluster correlation (see Eldridge et al. (2009)). In general, the regression technique assumes that the errors in observations are independent, identically and normally distributed. This assumption will not be always held for clustered data. Battese et al. (1988) proposed a regression method for analysing clustered data, which is called the nested error regression model.

The nested error regression model is expressed as

$$y_{ij} = \mathbf{x}_{ij}\beta + u_i + e_{ij}, \quad i = 1,\dots,c; j = 1,\dots,n_i, \tag{1}$$

[1]Department of Mathematics, Faculty of Science, Srinakharinwirot University, Bangkok, Thailand. E-mail: sukanyain@g.swu.ac.th. ORCID ID: https://orcid.org/0000-0002-6770-9503.

[2]Corresponding Author, Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand. E-mail: fscitdps@ku.ac.th. RCID ID: https://orcid.org/0000-0003-0019-9884.

where y_{ij} is the observed response for the jth sample unit in the ith cluster, $\mathbf{x}_{ij} = (x_{ij0}, x_{ij1}, \ldots, x_{ij,k-1})$ is the $n \times k$ matrix of explanatory variables and x_{ij0} is the $n \times 1$ column vector where entries are all 1, $\beta = (\beta_0, \beta_1, \ldots, \beta_{k-1})'$ is the k vector of regression coefficients and n_i is the number of sample units observed in the ith cluster ($\sum_{i=1}^{c} n_i = n$). The random effect u_i and random error e_{ij} are assumed to be independent of each other and distributed as $N(0, \sigma_u^2)$ and $N(0, \sigma_e^2)$, respectively.

The model (1) can be written as

$$y = \mathbf{X}\beta + \varepsilon, \tag{2}$$

where $y = (y_1, \ldots, y_c)'$ with $y_i = (y_{i1}, \ldots, y_{in_i})$, $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_c)'$ with $\mathbf{X}_i = (X_{i10}, \ldots, X_{in_i,k-1})$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_c)'$ with $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{in_i})$. Further, $\varepsilon_{ij} = u_i + e_{ij}, \varepsilon \sim N(0, \sigma^2\mathbf{V}), \sigma^2 = \sigma_u^2 + \sigma_e^2, \mathbf{V}$ has block-diagonal variance-covariance matrix with $\mathbf{V}_i = (1-\rho)\mathbf{I}_{n_i} + \rho\mathbf{J}_{n_i}$ for the ith cluster where $\rho = \sigma_u^2/\sigma^2$ is the intra-cluster correlation, $\mathbf{I}_{n_i}$ is the $n_i \times n_i$ identity matrix and $\mathbf{J}_{n_i}$ is the $n_i \times n_i$ matrix consisting of all 1s.

For testing the hypothesis about regression coefficients in the nested error regression model, formerly, Wu et al. (1988), Rao et al. (1993) and Lahiri and Li (2009) showed that the F test statistic based on ordinary least squares procedures leads to highly inflated type I error rate. Wu et al. (1988) proposed a modification of the F test statistic with known intra-cluster correlations, which is much better than the F test statistic based on ordinary least squares procedures by type I error rate. Rao et al. (1993) presented the F test statistic based on generalized least squares procedures with the Fuller–Battese transformation (see Galbraith et al. (2010)) in order to make observations independent and then applied the F test statistic to the observations under valid assumption. The F test statistic with the Fuller–Battese transformation performs similar to the modification of the F test statistic in controlling the type I error rate. Furthermore, the power of F test statistic with the Fuller–Battese transformation increases as the intra-cluster correlation increases, whereas the power of the modification of F test statistic decreases. The power of the F test statistic based on ordinary least squares procedures is not comparable because of type I error rate inflation.

Recently, Lahiri and Li (2009) suggested the transformation for the F test statistic based on generalized least squares procedures that is part of the Helmert matrix (see Lancaster (1965)), unlike the Fuller–Battese transformation. Like previous work, the F test statistic with part of Helmert matrix performs as well as in controlling the type I error rate, but the power of the test is not considered.

In this paper, we propose an alternative transformation for the generalized least squares procedures by applying Householder matrix (see Householder (1958)). In Section 2 we review several F test statistics for testing linear hypothesis regarding the regression coefficients under the nested error regression model. Monte Carlo study concerning the type I error rate and the power of the F test statistic is conducted in Section 3 as well as the real application. The results of the simulation are presented in Section 4.

2. The F test statistics

2.1. Prior F test statistics

Under model (1), suppose that the hypothesis of interest is $H_0 : \mathbf{C}\beta = \mathbf{q}$, where $\mathbf{C}$ is a known $m \times k$ matrix of rank $m(< k)$, and $\mathbf{q}$ is a known $m \times 1$ constant vector. The F-statistic based on ordinary least squares procedures is

$$F_{OLS} = \frac{(\mathbf{C}\widehat{\beta} - \mathbf{q})'\{\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'\}^{-1}(\mathbf{C}\widehat{\beta} - \mathbf{q})/m}{(\mathbf{y} - \mathbf{X}\widehat{\beta})'(\mathbf{y} - \mathbf{X}\widehat{\beta})/(n-k)}$$

where $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

F_{OLS} leads to highly inflated type I error rate when the intra-cluster correlation increases. When the intra-cluster correlation is known, Wu et al. (1988) proposed a modification of the F test statistic by multiplying the numerator and the denominator of F_{OLS} by a chi-squared distribution with $n-k$ and m degrees of freedom, respectively.

The modification of the F test statistic is

$$F_{WU} = F_{OLS} \times \frac{\{n - \mathrm{tr}(\mathbf{PV})\}/(n-k)}{\mathrm{tr}(\mathbf{P}_C\mathbf{V})/m},$$

where $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \mathbf{P}_C = \mathbf{X}_C(\mathbf{X}'_C\mathbf{X}_C)^{-1}\mathbf{X}'_C, \mathbf{X}_C = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'$ and tr is the trace operator.

Afterwards Rao et al. (1993) presented the F test statistic based on generalized least squares procedures with the Fuller–Battese transformation under model (1).

Define $\mathbf{T}_i = \mathbf{I}_{n_i} - n_i^{-1}(1 - [\{1-\rho\}/\{1+(n_i-1)\rho\}]^{1/2})\mathbf{J}_{n_i}, \mathbf{y}_i^* = \mathbf{T}_i\mathbf{y}_i, \mathbf{X}_i^* = \mathbf{T}_i\mathbf{X}_i$, and $\varepsilon_i^* = \mathbf{T}_i\varepsilon_i$. Then the transformed model can be written as $\mathbf{y}^* = \mathbf{X}^*\beta + \varepsilon^*$, where $\varepsilon \sim \mathbf{N}(0, \sigma_e^2\mathbf{I}_n)$ and $\sigma_e^2 = \sigma^2(1-\rho)$. Thus, the F test statistic based on generalized least squares procedures with the Fuller–Battese transformation is

$$F_{RAO} = \frac{(\mathbf{C}\beta^* - \mathbf{q})'(\mathbf{X}_C^{*\prime}\mathbf{X}_C^*)^{-1}(\mathbf{C}\beta^* - \mathbf{q})/m}{(\mathbf{y}^* - \mathbf{X}^*\beta^*)'(\mathbf{y}^* - \mathbf{X}^*\beta^*)/(n-k)},$$

where $\beta^* = (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{y}^*$ and $\mathbf{X}_C^* = \mathbf{X}^*(\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{C}'$.

Recently, unlike the Fuller–Battese transformation, Lahiri and Li (2009) proposed the transformation for the F test statistic based on generalized least squares procedures, which is part of the Helmert matrix. Generally, the Helmert matrix is orthogonal (see Farhadian and Asadian (2017)), but Lahiri and Li (2009) used the Helmert matrix by ignoring the first row. Thus, the part of the Helmert matrix is not orthogonal.

Let $\mathbf{G}_i$ be an $(n_i - 1) \times n_i$ matrix which is part of the Helmert matrix by ignoring the first row, i.e. $1'_{n_i}/\sqrt{n_i}$. Multiplying both sides of the model (2) by $\mathbf{G}_i$ then the transformed model is written as $\mathbf{y}^\star = \mathbf{X}^\star\beta + \varepsilon^\star$, where $\varepsilon^\star \sim \mathbf{N}(0, \sigma_e^2\mathbf{I}_{n-c})$ and $\sigma_e^2 = \sigma^2(1-\rho)$. The F test statistic based on generalized least squares procedures with part of the Helmert transformation is

$$F_{LAH} = \frac{(\mathbf{C}\beta^\star - \mathbf{q})'(\mathbf{X}_C^{\star\prime}\mathbf{X}_C^\star)^-(\mathbf{C}\beta^\star - \mathbf{q})/m}{(\mathbf{y}^\star - \mathbf{X}^\star\beta^\star)'(\mathbf{y}^\star - \mathbf{X}^\star\beta^\star)/(n-c-k)},$$

where $\beta^\star = (\mathbf{X}^{\star\prime}\mathbf{X}^\star)^{-1}\mathbf{X}^{\star\prime}\mathbf{y}^\star$ and $\mathbf{X}_C^\star = \mathbf{X}^\star(\mathbf{X}^{\star\prime}\mathbf{X}^\star)^{-1}\mathbf{C}'$.

2.2. F test statistic with an alternative transformation

For the F test statistic based on generalized least squares procedures, the transformation matrix is the necessary part to make the observations independent. Unlike the previous transformations, we propose an alternative transformation that adjusts the Householder matrix.

The Householder matrix (see Appendix) is taken into account because of its orthogonal, symmetry and idempotent properties, which are necessary for the transformation matrix. The Householder matrix for the ith cluster, denoted by $\mathbf{H}_i$, can be written in a simple form as

$$\mathbf{H}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & \frac{-1}{\sqrt{n_i-1}} & \frac{-1}{\sqrt{n_i-1}} & \cdots & \frac{-1}{\sqrt{n_i-1}} \\ 0 & \frac{-1}{\sqrt{n_i-1}} & \frac{(n_i-1)(n_i-3)+\sqrt{n_i-1}}{(n_i-1)(n_i-2)} & \cdots & \frac{-1}{(n_i-1)+\sqrt{n_i-1}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \frac{-1}{\sqrt{n_i-1}} & \frac{-1}{(n_i-1)+\sqrt{n_i-1}} & \cdots & \frac{(n_i-1)(n_i-3)+\sqrt{n_i-1}}{(n_i-1)(n_i-2)} \end{bmatrix}.$$

Even if $\mathbf{H}_i$ is orthogonal, the error term of the transformed model is still fallacious, that is the error term is not independent. Therefore, $\mathbf{H}_i$ is required to be adjusted. Let $\mathbf{D}_i$ be an $n_i \times n_i$ matrix, which is defined as

$$\mathbf{D}_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\rho\sqrt{n_i-1} & \sqrt{1+(n_i-2)\rho-(n_i-1)\rho^2} & 0 & \cdots & 0 \\ 0 & 0 & \sqrt{1-\rho} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sqrt{1-\rho} \end{bmatrix}$$

such that $(\mathbf{H}_i\mathbf{D}_i)(\mathbf{H}_i\mathbf{D}_i)' = \mathbf{V}_i$. Now, we have the alternative matrix transformation, $\mathbf{P}_i$, which is the inverse of matrix $(\mathbf{H}_i\mathbf{D}_i)$, also, $\mathbf{y}_i^{AH} = \mathbf{P}_i\mathbf{y}_i, \mathbf{X}_i^{AH} = \mathbf{P}_i\mathbf{X}_i$ and $\varepsilon_i^{AH} = \mathbf{P}_i\varepsilon_i$.

Then, the transformed model can be written as $\mathbf{y}^{AH} = \mathbf{X}^{AH}\beta + \varepsilon^{AH}$, where

$$\mathbf{y}^{AH} = (\mathbf{y}_1^{AH}, \ldots, \mathbf{y}_c^{AH})' \quad \text{with} \quad \mathbf{y}_i^{AH} = (y_{i1}^{AH}, \ldots, y_{in_i}^{AH}),$$

$$\mathbf{X}^{AH} = (\mathbf{X}_1^{AH}, \ldots, \mathbf{X}_c^{AH})' \quad \text{with} \quad \mathbf{X}_i^{AH} = (X_{i10}^{AH}, \ldots, X_{in_i,k-1}^{AH})$$

$$\text{and} \quad \varepsilon^{AH} = (\varepsilon_1^{AH}, \ldots, \varepsilon_c^{AH})' \quad \text{with} \quad \varepsilon_i^{AH} = (\varepsilon_{i1}^{AH}, \ldots, \varepsilon_{in_i}^{AH}).$$

Currently, the assumption of the error is valid, i.e. $\text{var}(\varepsilon_i^{AH}) = \text{var}(\mathbf{P}_i\varepsilon_i) = \text{var}\{(\mathbf{H}_i\mathbf{D}_i)^{-1}\varepsilon_i\} =$

$\sigma^2(\mathbf{H}_i\mathbf{D}_i)^{-1}\mathbf{V}_i\,(\mathbf{H}_i\mathbf{D}_i)^{-1'}=\sigma^2\mathbf{I}_{n_i}$, and $\mathrm{cov}(\varepsilon_i^{AH},\varepsilon_j^{AH})=0$ for $i\neq j$, that is $\mathrm{cov}(y_i^{AH},y_j^{AH})=0$ for $i\neq j$.

Ultimately, the F test statistic based on generalized least squares procedures with Adjusted Householder transformation is

$$F_{AH}=\frac{(\mathbf{C}\beta^{AH}-\mathbf{q})'(\mathbf{X}_C^{AH'}\mathbf{X}_C^{AH})^{-1}(\mathbf{C}\beta^{AH}-\mathbf{q})/m}{(\mathbf{y}^{AH}-\mathbf{X}^{AH}\beta^{AH})'(\mathbf{y}^{AH}-\mathbf{X}^{AH}\beta^{AH})/(n-k)},$$

where $\hat{\beta}^{AH}=(\mathbf{X}^{AH'}\mathbf{X}^{AH})^{-1}\mathbf{X}^{AH'}\mathbf{y}^{AH}$ and $\mathbf{X}_C^{AH}=\mathbf{X}^{AH}(\mathbf{X}^{AH'}\mathbf{X}^{AH})^{-1}\mathbf{C}'$.

3. Simulation study

3.1. A Monte Carlo simulation

In this section the data sets are randomly generated to illustrate how various methods of statistical inference perform for analysing the clustered data. Following Wu et al. (1988), Rao et al. (1993) and Lahiri and Li (2009), the nested error regression model with two covariates (i.e. x_1 and x_2) is considered:

$$y_{ij}=\beta_0+\beta_1x_{ij1}+\beta_2x_{ij2}+u_i+e_{ij},\quad i=1,\ldots,c;j=1,\ldots,n_i. \tag{3}$$

The data sets of (x_{ij1},x_{ij2}) are generated from the bivariate normal distribution with additional random effects components to allow for the intra-cluster correlations ρ_{x_1} and ρ_{x_2} on both x_1 and x_2, respectively:

$$x_{ij1}=\mu_{x_1}+u_{x_1i}+e_{x_1ij},\quad x_{ij2}=\mu_{x_2}+u_{x_2i}+e_{x_2ij},$$

where $u_{x_1i}\sim N(0,\sigma_{ux_1}^2)$, $e_{x_1i}\sim N(0,\sigma_{ex_1}^2)$, $u_{x_2i}\sim N(0,\sigma_{ux_2}^2)$, $e_{x_2i}\sim N(0,\sigma_{ex_2}^2)$, $\rho_{x_1}=\sigma_{ux_1}^2/\sigma_{x_1}^2$, $\rho_{x_2}=\sigma_{ux_2}^2/\sigma_{x_2}^2$, $\sigma_{x_1}^2=\sigma_{ux_1}^2+\sigma_{ex_1}^2$, $\sigma_{x_2}^2=\sigma_{ux_2}^2+\sigma_{ex_2}^2$, $\sigma^2=\sigma_u^2+\sigma_e^2$, $\sigma_u^2=\sigma_{ux_1}^2+\sigma_{ux_2}^2$, $\sigma_e^2=\sigma_{ex_1}^2+\sigma_{ex_2}^2$. u_{x_1},u_{x_2} and e_{ij} are independent. Moreover, u_{x_1i} and u_{x_2i} are correlated with covariance $\sigma_{ux_1x_2}$, and e_{x_1i} and e_{x_2i} are correlated with covariance $\sigma_{ex_1x_2}$.

Let $\rho_{x_1x_2}=\sigma_{ux_1x_2}/\sigma_{x_1}\sigma_{x_2}$ and $\mathrm{corr}(x_1,x_2)=\sigma_{x_1x_2}/\sigma_{x_1}\sigma_{x_2}$, where $\sigma_{x_1x_2}=\sigma_{ux_1x_2}+\sigma_{ex_1x_2}$ and $\mathrm{corr}(x_1,x_2)$ denote the correlation between x_{ij1} and x_{ij2}. For the nested error regression model with two covariates, the parameters are set accordingly to the previous researchers (see Wu et al. (1988), Rao et al. (1993) and Lahiri and Li (2009)). Then, without loss of generality, $\sigma_{x_1}^2=\sigma_{x_2}^2=20$, $\rho_{x_1}=0.1,\rho_{x_2}=0.5$, $\rho_{x_1x_2}=0,\mathrm{corr}(x_1,x_2)=-0.33$, $\mu_{x_1}=100,\mu_{x_2}=200,\beta_0=10,\beta_1=\beta_2=0$ and $\sigma^2=10$.

Given (x_{ij1},x_{ij2}), y_{ij} is generated by model (3) with five different values for intra-cluster correlation (ρ = 0, 0.05, 0.1, 0.3 and 0.5) and five different numbers of clusters (c = 3, 4, 5, 10 and 15) for the balanced data. When the data is unbalanced, there are three clusters (c = 3) and the sets of the sample size are varied. The simulated data (y_{ij},x_{ij1},x_{ij2}) are repeated 10,000 replications for all conditions and the F test statistics are computed for each replication to obtain the type I error rate and the power of the test.

3.2. Type I error rate

Type I error rate is obtained by the proportion of times that the p-value for the F test statistic is smaller than the nominal level. The measurements are the binary variables corresponding to the rejection regions of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$. We then test the null hypothesis of no effect of the regression coefficients at 5% and 10% nominal levels and the confidence interval of the type I error rate ($\hat{\alpha}$) is calculated from $\hat{\alpha} + Z_{\alpha/2}\sqrt{\hat{\alpha}(1-\hat{\alpha})/10000}$ (see Lahiri and Li (2009)). If the nominal level is 5% and 10%, the type I error rate should not exceed 5.43% and 10.49%, respectively.

In Table 1, for the balanced data with the sample sizes (n) of 20, 30, 100 and 150, the results show that F_{AH}, F_{LAH}, F_{RAO} and F_{WU} perform as well as in controlling the type I error rate. Under the F test statistic based on generalized least squares procedures, F_{AH} performs as well as F_{RAO} in terms of the type I error rate. The F test statistic with the Fuller–Battese transformation and the F test statistic with the Adjusted Householder transformation are slightly different but they come up with the same hypothesis testing conclusion, consequently, the type I error rate of F_{RAO} is disregarded. Furthermore, the type I error rates of F_{WU} exceed the limit for large intra-cluster correlation (ρ = 0.5) and F_{OLS} leads to highly inflated the type I error rate for almost situations.

When the sample sizes of each cluster are unbalanced, the results show that F_{AH}, F_{LAH} and F_{WU} perform well in controlling the type I error rate for a small sample size (n=15) and small intra-cluster correlations ($\rho \leq$0.1) as shown in Table 2. Under the unbalanced data and large sample sizes (n=30 and 90), F_{AH}, F_{LAH} and F_{WU} maintain the nominal level for small intra-cluster correlations ($\rho \leq$0.1) when the sample sizes of each cluster are slightly different, such as (n_1, n_2, n_3)=(9,10,11), (29,30,31). While the sample sizes of each cluster are widely varied, such as (n_1, n_2, n_3)=(5,10,15), (3,3,24), (10,20,60), (3,3,84), all F test statistics lead to highly inflated type I error rate for all intra-cluster correlations, except for ρ = 0.

3.3. Power of the test

Power of the test is obtained by the proportion of times that rejects the null hypothesis when the alternative hypothesis is true at the nominal level. Table 3 reports the power of the F test statistic of the null hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the specified alternative at nominal 5% and 10% levels for the balanced data. For large intra-cluster correlations, the power of F_{AH} gains over the others. For very small intra-cluster correlations ($\leq$0.05), the power of F_{AH} performs as well as F_{WU}, on the contrary, the power of F_{LAH} is the lowest as shown in Figures 1 - 4. For example, when $c_i \times n_i = 10 \times 10$ and ρ =0.05, the power of F_{LAH} is approximately 53% compared to the power of F_{AH}, which is 74%. Note that a slight decrease of the power of F_{AH} occurs when ρ increases from 0 to 0.1.

For a small sample size, the powers of F_{AH} and F_{WU} are similar when $\rho \leq 0.1$ as shown clearly in Figure 1 and 2, whereas the power of F_{AH} is higher than that of F_{WU} when $\rho \geq 0.3$. For a large sample size, Figure 3 and 4 confirm that the power of F_{AH}

is higher than that of F_{WU} and the powers of F_{AH} and F_{LAH} increase as ρ increases, while the power of F_{WU} decreases as ρ increases, theoretically corresponding to the power of test established by Rao and Wang (1995). The illustration is shown in Figure 4(b). That is, when the nominal level is 10% and the alternative hypothesis is $H_1 : \beta_1 = \beta_2 = 0.2$, the power of F_{AH} increases from 87.07% to 90.42%, but the power of F_{WU} decreases from 87.07% to 54.14% as ρ increases for $c_i \times n_i = 10 \times 10$.

Similar to the balanced data, when the sample sizes of each cluster are unbalanced, the powers of F_{AH} and F_{WU} are higher than the power of F_{LAH} for almost all situations as shown in Table 4.

Table 1. Type I error rates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ at nominal 5% and 10% levels for the balanced data

		Nominal level 5%				Nominal level 10%			
$c \times n_i$	ρ	F_{OLS}	F_{WU}	F_{LAH}	F_{AH}	F_{OLS}	F_{WU}	F_{LAH}	F_{AH}
4×5	0	5.05	5.05	4.28	5.05	10.15	10.15	8.73	10.15
	0.05	5.57*	4.83	3.69	4.88	11.23*	10.11	8.23	9.95
	0.1	6.29*	4.73	3.55	4.69	12.14*	9.75	7.70	9.72
	0.3	9.84*	4.76	4.19	4.83	16.54*	9.44	8.37	9.77
	0.5	14.74*	4.82	3.88	5.03	23.22*	9.52	8.43	9.89
3×10	0	5.20	5.20	4.42	5.20	9.84	9.84	9.24	9.84
	0.05	6.44*	4.85	4.29	4.97	12.24*	10.10	9.17	9.86
	0.1	8.15*	4.74	4.29	4.99	14.59*	9.99	9.17	9.91
	0.3	15.30*	4.37	4.29	4.96	23.04*	9.21	9.17	10.04
	0.5	22.74*	4.04	4.29	5.11	31.83*	8.72	9.17	9.97
5×20	0	4.88	4.88	4.68	4.88	9.97	9.97	9.46	9.97
	0.05	8.99*	4.75	4.82	4.73	15.71*	9.80	9.86	9.94
	0.1	13.38*	5.13	5.05	4.86	21.22*	10.00	9.86	10.06
	0.3	28.99*	5.02	4.75	4.68	38.41*	9.47	9.45	9.54
	0.5	43.21*	5.72*	5.04	5.38	51.83*	10.89*	10.20	10.27
10×10	0	5.18	5.18	4.76	5.18	10.42	10.42	10.15	10.42
	0.05	7.28*	5.43	4.53	5.12	12.74*	9.90	9.32	9.97
	0.1	9.14*	4.76	4.92	4.82	16.06*	9.67	9.93	9.86
	0.3	19.11*	5.11	4.54	4.85	27.73*	10.18	9.40	9.62
	0.5	28.47*	5.49*	4.99	5.21	37.31*	10.11	9.86	9.86
15×10	0	5.01	5.01	4.87	5.01	9.93	9.93	9.78	9.93
	0.05	7.09*	4.91	5.17	4.94	13.17*	9.93	10.07	9.79
	0.1	10.17*	5.26	5.20	5.40	16.69*	10.29	9.93	10.49
	0.3	19.15*	5.04	4.71	4.85	27.92*	9.99	9.35	9.59
	0.5	28.05*	5.50*	5.20	5.28	37.03*	10.10	10.18	10.01

*indicates that the type I error rate exceeded the limit

Table 2. Type I error rates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ at nominal 5% and 10% levels for the unbalanced data

		Nominal level 5%				Nominal level 10%			
$c \times n_i$	ρ	F_{OLS}	F_{WU}	F_{LAH}	F_{AH}	F_{OLS}	F_{WU}	F_{LAH}	F_{AH}
4,5,6	0.0	5.41	5.41	3.73	5.41	10.33	10.33	7.76	10.33
	0.05	5.50*	5.09	3.60	5.19	10.84*	10.26	7.65	10.32
	0.1	5.74*	5.00	3.27	4.91	11.10*	10.03	7.52	10.30
	0.3	8.85*	6.05*	4.47	6.90*	15.75*	11.73*	9.09	13.14*
	0.5	13.05*	7.15*	5.11	8.95*	21.18*	12.98*	10.21	15.71*
3,3,9	0.0	4.76	4.76	3.43	4.76	10.04	10.04	7.54	10.04
	0.05	5.46*	5.23	3.71	5.28	10.53*	10.19	7.93	10.13
	0.1	6.62*	6.13*	4.23	6.19*	11.87*	11.27*	8.65	11.66*
	0.3	8.85*	7.35*	5.76*	8.76*	15.75*	13.26*	11.13*	15.27*
	0.5	12.90*	9.53*	7.48*	12.09*	20.90*	16.33*	13.82*	19.90*
9,10,11	0.0	4.64	4.64	4.36	4.64	9.77	9.77	8.87	9.77
	0.05	5.90*	4.79	4.05	4.97	11.58*	9.70	9.13	9.63
	0.1	8.53*	5.79*	4.38	5.88*	14.82*	11.19*	9.49	10.93*
	0.3	15.52*	5.81*	5.32	6.89*	24.06*	11.53*	10.29	12.97*
	0.5	23.11*	6.05*	5.80*	7.79*	32.43*	11.88*	11.18*	14.09*
5,10,15	0.0	4.83	4.83	4.25	4.83	9.60	9.60	8.47	9.60
	0.05	6.32*	5.68*	5.22	6.01*	11.87*	11.11*	10.16	11.36*
	0.1	7.81*	6.54*	5.88*	7.03*	14.29*	12.42*	10.98*	13.17*
	0.3	16.03*	10.98*	10.16*	12.50*	23.78*	17.84*	16.76*	20.06*
	0.5	23.94*	13.89*	14.91*	17.97*	33.20*	22.08*	22.66*	26.56*
3,3,24	0.0	4.69	4.69	4.02	4.69	9.60	9.60	8.59	9.60
	0.05	6.19*	5.82*	5.08	5.85*	11.79*	11.55*	10.07	11.39*
	0.1	7.76*	7.19*	6.37*	7.38*	14.60*	13.77*	12.36*	13.99*
	0.3	15.66*	12.99*	11.96*	14.40*	24.21*	20.94*	19.41*	22.28*
	0.5	23.63*	17.91*	17.59*	20.93*	32.39*	26.32*	26.12*	29.77*
29,30,31	0.0	4.86	4.86	4.87	4.86	10.19	10.19	10.07	10.19
	0.05	10.22*	5.12	5.06	5.34	17.59*	10.37	9.97	10.49
	0.1	15.97*	6.02*	5.04	5.85*	24.07*	11.13*	10.29	11.27*
	0.3	32.75*	6.06*	5.68*	6.53*	41.29*	10.71*	10.80*	12.14*
	0.5	45.53*	4.89	6.15*	7.11*	53.74*	9.93	11.60*	12.68*
10,20,60	0.0	5.19	5.19	4.88	5.19	10.12	10.12	9.60	10.12
	0.05	10.50*	7.50*	7.95*	8.03*	17.17*	13.40*	13.90*	13.91*
	0.1	15.18*	9.15*	10.97*	11.04*	22.81*	15.27*	18.26*	18.30*
	0.3	32.84*	14.70*	23.37*	23.80*	41.13*	22.05*	31.27*	31.56*
	0.5	46.12*	17.70*	34.42*	35.03*	54.39*	24.33*	42.84*	43.77*
3,3,84	0.0	5.16	5.16	5.04	5.16	10.23	10.23	10.01	10.26
	0.05	10.80*	10.65*	10.00*	10.71*	17.95*	17.78*	16.86*	17.67*
	0.1	15.88*	15.61*	14.47*	15.57*	23.78*	23.32*	22.29*	23.51*
	0.3	33.25*	32.08*	30.92*	31.96*	41.91*	40.65*	39.75*	40.85*
	0.5	45.95*	43.53*	43.56*	44.61*	54.28*	52.10*	51.78*	52.87*

*indicates that the type I error rate exceeded the limit

Table 3. Power estimates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ versus specified alternatives at nominal 5% and 10% levels for the balanced data

				Nominal level 5%			Nominal level 10%		
$c \times n_i$	β_1	β_2	ρ	F_{WU}	F_{LAH}	F_{AH}	F_{WU}	F_{LAH}	F_{AH}
4×5	0.1	0.1	0	8.18	5.41	8.18	14.63	10.50	14.63
			0.05	7.90	5.00	7.91	14.88	10.52	14.78
			0.1	7.61	4.89	7.39	13.93	10.02	13.88
			0.3	7.40	5.95	7.78	13.67	11.34	14.50
			0.5	7.25	6.41	8.63	13.43	12.20	15.44
	0.2	0.2	0	18.30	9.28	18.30	28.02	16.51	28.02
			0.05	18.19	9.39	17.91	28.21	17.27	28.04
			0.1	16.93	9.83	16.80	26.63	17.73	27.16
			0.3	15.38	12.02	17.29	25.32	20.24	27.62
			0.5	15.20	15.18	21.13	24.13	25.37	32.14
3×10	0.1	0.1	0	9.38	6.80	9.38	17.02	12.93	17.02
			0.05	9.55	7.37	9.48	16.50	14.02	16.52
			0.1	9.34	7.09	9.11	15.79	13.46	16.25
			0.3	8.17	8.07	9.53	14.71	14.52	17.03
			0.5	-	10.02	11.31	-	17.26	18.75
	0.2	0.2	0	26.02	15.71	26.02	37.86	25.40	37.86
			0.05	24.55	16.92	24.19	36.20	27.36	36.01
			0.1	23.21	16.94	23.93	34.16	27.42	35.42
			0.3	19.32	21.36	25.76	29.40	32.44	37.35
			0.5	-	29.28	32.19	-	41.92	45.22
5×20	0.1	0.1	0	25.94	16.45	25.94	37.39	25.86	37.39
			0.05	20.84	16.40	21.48	31.47	26.33	32.06
			0.1	18.13	17.29	20.97	28.02	27.34	31.13
			0.3	14.13	22.11	23.83	21.23	32.29	34.65
			0.5	-	28.62	30.15	-	40.66	41.95
	0.2	0.2	0	76.63	53.93	76.63	84.63	66.49	84.63
			0.05	65.53	55.20	68.96	75.89	67.79	79.17
			0.1	58.24	58.35	67.68	69.56	70.21	77.99
			0.3	39.62	69.27	72.96	50.81	79.21	82.22
			0.5	-	83.74	85.32	-	90.69	91.72
10×10	0.1	0.1	0	26.73	15.21	26.73	38.15	24.72	38.15
			0.05	23.88	16.43	23.91	34.81	25.35	35.13
			0.1	22.65	17.44	23.71	33.89	27.58	35.35
			0.3	17.86	21.33	25.23	26.34	31.74	36.14
			0.5	-	26.76	29.42	22.16	38.90	41.54
	0.2	0.2	0	79.46	51.02	79.46	87.07	63.60	87.07
			0.05	73.96	52.92	74.30	82.52	65.02	83.35
			0.1	69.77	56.65	73.14	80.17	68.67	82.64
			0.3	53.22	66.54	74.73	64.90	77.44	83.65
			0.5	-	80.28	84.00	54.14	87.68	90.42

-indicates that power of the test cannot be compared

Table 3. Power estimates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ versus specified alternatives at nominal 5% and 10% levels for the balanced data (cont.)

				Nominal level 5%			Nominal level 10%		
$c \times n_i$	β_1	β_2	ρ	F_{WU}	F_{LAH}	F_{AH}	F_{WU}	F_{LAH}	F_{AH}
15×10	0.1	0.1	0	39.85	21.93	39.85	52.53	33.09	52.53
			0.05	35.06	22.85	35.22	47.41	33.09	47.74
			0.1	31.40	23.75	33.93	43.96	35.84	46.54
			0.3	23.01	29.53	34.22	32.92	41.73	47.22
			0.5	-	39.55	43.25	28.74	52.81	56.39
	0.2	0.2	0	93.48	69.79	93.48	96.56	80.00	96.56
			0.05	90.04	72.15	90.59	94.54	82.09	95.01
			0.1	86.16	74.57	89.01	91.89	83.93	93.76
			0.3	70.50	85.17	90.52	79.78	91.41	94.76
			0.5	-	94.34	96.12	70.21	97.13	98.12

-indicates that power of the test cannot be compared

Table 4. Power estimates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ versus specified alternatives at nominal 5% and 10% levels for the unbalanced data

				Nominal level 5%			Nominal level 10%		
n_1, n_2, n_3	β_1	β_2	ρ	F_{WU}	F_{LAH}	F_{AH}	F_{WU}	F_{LAH}	F_{AH}
4,5,6	0.1	0.1	0	7.11	4.49	7.11	13.59	7.38	13.59
			0.05	6.56	4.11	6.51	12.56	7.01	12.64
			0.1	7.21	4.52	7.10	13.90	7.96	14.02
			0.3	-	5.59	-	-	9.96	-
			0.5	-	7.21	-	-	13.35	-
	0.2	0.2	0	13.27	9.22	13.27	22.12	14.38	22.12
			0.05	12.42	8.77	12.50	21.59	14.16	21.73
			0.1	13.70	9.73	13.78	22.98	15.23	23.28
			0.3	-	11.07	-	-	18.28	-
			0.5	-	13.73	-	-	22.43	-
3,3,9[a]	0.1	0.1	0	6.63	4.33	6.63	13.53	7.82	13.53
			0.05	6.83	4.59	6.82	14.08	8.53	13.90
			0.1	-	5.55	-	-	9.93	-
	0.2	0.2	0	13.13	9.26	13.13	22.40	15.60	22.40
			0.05	15.81	12.76	16.00	35.21	27.59	35.43
			0.1	-	11.04	-	-	17.28	-
9,10,11[b]	0.1	0.1	0	9.47	6.93	9.47	26.05	17.06	26.05
			0.05	9.06	6.58	8.91	23.75	16.79	23.87
			0.1	-	7.95	-	-	18.73	-
			0.3	-	9.57	-	-	23.76	-
	0.2	0.2	0	17.08	13.21	17.08	37.99	26.94	37.99
			0.05	17.03	13.73	16.97	36.61	28.11	36.86
			0.1	-	14.19	-	-	29.18	-
			0.3	-	16.11	-	-	34.60	-

-indicates that power of the test cannot be compared

For $H_1 : \beta_1 = \beta_2 = 0.1$ and $H_1 : \beta_1 = \beta_2 = 0.2$,

[a] when ρ = 0.3 and 0.5, all F-statistics cannot control the type I error rate,

[b] when ρ = 0.5, all F-statistics cannot control the type I error rate

Table 4. Power estimates (%) of the test $H_0 : \beta_1 = \beta_2 = 0$ versus specified alternatives at nominal 5% and 10% levels for the unbalanced data (cont.)

				Nominal level 5%			Nominal level 10%		
n_1, n_2, n_3	β_1	β_2	ρ	F_{WU}	F_{LAH}	F_{AH}	F_{WU}	F_{LAH}	F_{AH}
5,10,15[c]	0.1	0.1	0	9.38	7.73	9.38	25.64	19.10	25.64
			0.05	-	9.20	-	-	21.60	-
	0.2	0.2	0	16.58	13.93	16.58	37.61	30.03	37.61
			0.05	-	16.14	-	-	32.97	-
3,3,24[c]	0.1	0.1	0	9.35	7.92	9.35	25.36	20.95	25.36
			0.05	-	8.95	-	-	22.23	-
	0.2	0.2	0	16.30	14.33	16.30	37.15	31.68	37.15
			0.05	-	15.76	-	-	33.09	-
29,30,31	0.1	0.1	0	22.06	15.51	22.06	66.77	50.09	66.77
			0.05	18.79	16.52	19.84	56.97	52.79	61.31
			0.1	-	16.58	-	-	54.96	-
			0.3	-	-	-	-	-	-
			0.5	10.33	-	-	28.31	-	-
	0.2	0.2	0	32.76	24.81	32.76	77.10	63.12	77.10
			0.05	28.29	25.82	30.38	68.79	65.08	72.92
			0.1	-	26.16	-	-	67.57	-
			0.3	-	-	-	-	-	-
			0.5	17.41	-	-	38.58	-	-
10,20,60[d]	0.1	0.1	0	22.26	18.53	22.26	68.03	58.74	68.03
	0.2	0.2	0	33.35	28.68	33.35	77.79	70.11	77.79
3,3,84[d]	0.1	0.1	0	22.00	21.31	22.00	66.90	64.93	66.91
	0.2	0.2	0	32.80	31.34	32.79	77.35	75.29	77.35

-indicates that power of the test cannot be compared

For $H_1 : \beta_1 = \beta_2 = 0.1$ and $H_1 : \beta_1 = \beta_2 = 0.2$,

[c] when ρ = 0.1, 0.3 and 0.5, all F-statistics cannot control the type I error rate,

[d] when ρ = 0.05, 0.1, 0.3 and 0.5, all F-statistics cannot control the type I error rate

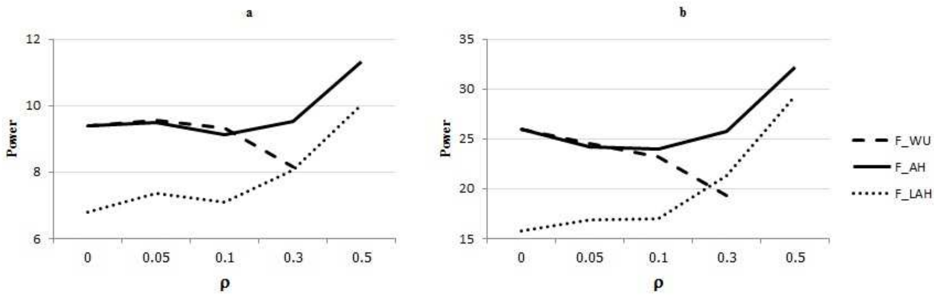

Figure 1: Power estimates (%) of F test statistic versus intra-cluster correlation at nominal 5% level and $c \times n_i = 3 \times 10$ corresponding to the alternatives hypothesis (a) $H_1 : \beta_1 = \beta_2 = 0.1$ and (b) $H_1 : \beta_1 = \beta_2 = 0.2$

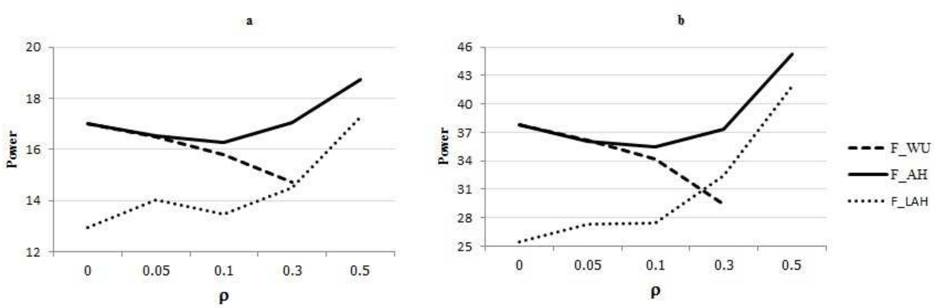

Figure 2: Power estimates (%) of F test statistic versus intra-cluster correlation at nominal 10% level and $c \times n_i = 3 \times 10$ corresponding to the alternatives hypothesis (a) $H_1 : \beta_1 = \beta_2 = 0.1$ and (b) $H_1 : \beta_1 = \beta_2 = 0.2$

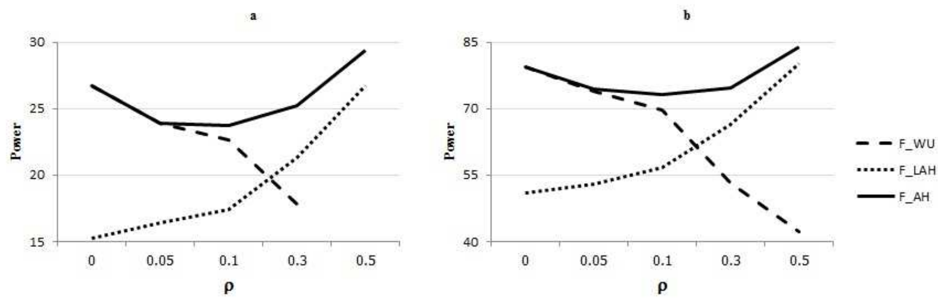

Figure 3: Power estimates (%) of F test statistic versus intra-cluster correlation at nominal 5% level and $c \times n_i = 10 \times 10$ corresponding to the alternatives hypothesis (a) $H_1 : \beta_1 = \beta_2 = 0.1$ and (b) $H_1 : \beta_1 = \beta_2 = 0.2$

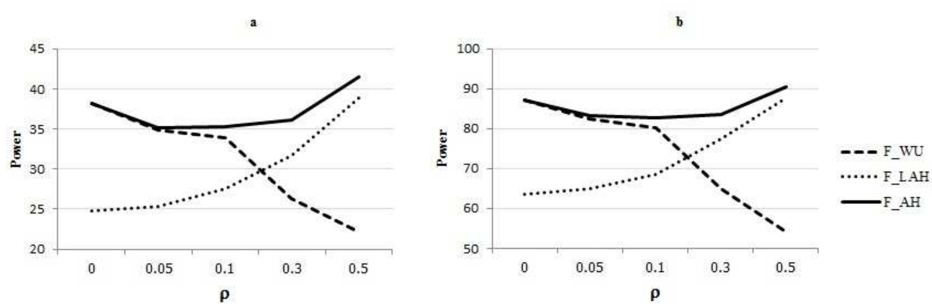

Figure 4: Power estimates (%) of F test statistic versus intra-cluster correlation at nominal 10% level and $c \times n_i = 10 \times 10$ corresponding to the alternatives hypothesis (a) $H_1 : \beta_1 = \beta_2 = 0.1$ and (b) $H_1 : \beta_1 = \beta_2 = 0.2$

3.4. An application

In this section, we consider the data from Smith (1980). This data set covers the values of pattern intensity on soles of 14 families chosen from Polish family data (see Table 5). The families consist of siblings, together with their mothers and fathers. Here, $\mathbf{y}$ is the 49×1 vector of values of pattern intensity on soles of feet of siblings and $\mathbf{X}$ is the 49×3 matrix of ones in the first column and values of pattern intensity on soles of feet of mother and father. In the real data, the intra-cluster correlation is usually unknown and it must be estimated for the F test statistic. The Srivastava estimator of the intra-cluster correlation, 0.4922, (see Srivastava and Katapa (1986)) is applied in this section. In order to test the regression coefficients for the nested error regression model, the p-value of F_{AH}, F_{WU}, F_{LAH} and F_{OLS} are less than 0.05, then we reject the null hypothesis (H_0) at the significance level of 5%. This indicates that at least one regression coefficient is significant to the model. Note that the intra-cluster correlation estimator, the important characteristic of the clustered data, is not used to compute F_{OLS} and F_{LAH}. In addition, the errors in observations of F_{OLS} do not correspond to the assumption of regression analysis even the power of the test is quite high. Therefore, the F_{AH} and F_{WU} using Srivastava estimator of the intra-cluster correlation are suggested and the power of F_{AH} is higher than F_{WU} for applying to this application.

Table 5. Values of pattern intensity on soles of feet in 14 families

Family no.	Mother	Father	Siblings
1	2	3	2,2
2	2	3	2,3
3	2	3	2,2,2
4	2	4	2,2,2,2,2
5	6	7	6,6
6	4	3	4,3,3
7	4	3	2,2,3,6,3,5,4
8	3	7	2,4,7,4,4,7,8
9	5	5	5,6
10	5	4	4,5,4
11	5	6	5,3,4,4
12	2	4	2,4
13	6	3	4,3,3,3
14	2	3	2,2,2

Table 6. The F test statistics, p-values and powers for data set in Table 5

Method	F	p-value	power
F_{OLS}	21.2343	0.0000003	0.9999997
F_{WU}	4.7388	0.0135	0.9865
F_{LAH}	7.4865	0.0021	0.9979
F_{AH}	6.4756	0.0033	0.9967

4. Conclusion

For clustered data analysis with compound symmetry correlation structure of known intra-cluster correlation, the proposed transformation by Adjusted Householder matrix can be used to adjust the correlation of the error term and then allowed to be applied to the F test statistic based on generalized least squares procedures. The simulation study shows that the F test statistic with Adjusted Householder transformation performs as well as the other methods for the balanced and unbalanced data, except for the F test statistic based on standard ordinary least squares procedures, in controlling the type I error rate regarding regression coefficients hypothesis testing for small and large sample sizes. The power of the F test statistic with Adjusted Householder transformation is always higher than that with part of the Helmert transformation for the balanced and unbalanced data. Also, the power of the F test statistic with Adjusted Householder (F_{AH}) and part of the Helmert (F_{LAH}) transformations are the increasing functions of the intra-cluster correlation whereas the power of the modification of the F test statistic (F_{WU}) is the decreasing function.

REFERENCES

AQARWAL, G. G., AWASTHI, S., WALTER, S. D., (2005). Intra-class Correlation Estimates for Assessment of Vitamin A Intake in Children. Journal of Health, Population, and Nutrition, 23 (1), pp. 66–73.

BATTESE, G. E., HARTER, R. M., FULLER, W. A., (1988). An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. Journal of the American Statistical Association, 83 (401), pp. 28–36.

ELDRIDGE, S. M., UKOUMUNNE, O. C., CARLIN, J. B., (2009). The Intra-Cluster Correlation Coefficient in Cluster Randomized Trials: A Review of Definitions. International Statistical Review, 77 (3), pp. 378–394.

FARHADIAN, R., ASADIAN, N., (2017). On the Helmert Matrix and Application in Stochastic Processes. International Journal of Mathematics and Computer Science, 12 (2), pp. 107–115.

GALBRAITH, S., DANIEL, J. A., VISSEL, B., (2010). A Study of Clustered Data and Approaches to Its Analysis. The Journal of Neuroscience, 30 (32), pp. 10601–10608.

HOUSEHOLDER, A. S., (1958). Unitary Triangularization of a Nonsymmetric Matrix. Journal of the ACM, 5, pp. 339–342.

LAHIRI, P., LI, Y., (2009). A New Alternative to the Standard F Test for Clustered Data. Journal of Statistical Planning and Inference, 139, pp. 3430–3441.

LANCASTER, H. O., (1965). The Helmert Matrices. The American Mathematical Monthly, 72, pp. 4–12.

MCCULLOCH, C. E., SHAYLE, R. S., (2001). Generalized, linear, and mixed models, New York: John Wiley and Sons.

MIALL, W. E., OLDHAM, P. D., (1955). A Study of Arterial Blood Pressure and Its Inheritance in a Sample of the General Population. Clinical Science, 14 (3), pp. 459–488.

NG, S. K., MCLACHLAN, G. J., YAU, K. K. W., LEE, A. H., (2004). Modeling the Distribution of Ischaemic Stroke-specific Survival Time using an EM-based Mixture Approach with Random Effects Adjustment. Statistics in Medicine, 23, pp. 2729–2744.

RAO, J. N. K., SUTRADHAR, B. C., YUE, K., (1993). Generalized Least Squares F Test in Regression Analysis with Two-Stage Cluster Samples. Journal of the American Statistical Association, 88 (424), pp. 1388–1391.

RAO, J. N. K., WANG, S. G., (1995). On the Power of F Tests under Regression Models with Nested Error Structure. Journal of Multivariate Analysis, 53, pp. 237–246.

SMITH, C. A. B., (1980). Estimating Genetic Correlations. Ann. Human Genetics, 43, pp. 265–284.

SRIVASTAVA, M. S., KATAPA, R. S., (1986). Comparison of Estimators of Interclass and Intraclass Correlations from Familial Data. The Canadian Journal of Statistics, 14 (1), pp. 29–42.

WU, C. F. J., HOLT, D., HOLMES, D. J., (1988). The Effect of Two-Stage Sampling on the F Statistic. Journal of the American Statistical Association, 83 (401), pp. 150–159.

APPENDIX

Householder matrix

For any $n \times n$ symmetric matrix, in this paper we consider the variance-covariance matrix $\mathbf{V} = [v_{ij}]; i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, n$, which is the compound symmetry correlation structure. The corresponding Householder matrix, $\mathbf{H}$, is a symmetry and orthogonal matrix in the form

$$\mathbf{H} = \mathbf{I}_n - 2\mathbf{w}\mathbf{w}'.$$

Let $\mathbf{w} = (w_1, \ldots, w_n)'$ be a column vector which is a unit vector of Euclidean norm where

$$w_1 = 0, w_2 = \frac{v_{21} - \alpha}{2\gamma} \quad \text{and} \quad w_l = \frac{v_{l1}}{2\gamma}; l = 3, \ldots, n.$$

α and γ are determined by

$\alpha = -\mathrm{sgn}(v_{21})\sqrt{\Sigma_{i=2}^{n} v_{i1}^2}$ where $\mathrm{sgn}(v_{21}) = \begin{cases} -1 & \text{for} \quad v_{21} < 0 \\ +1 & \text{for} \quad v_{21} > 0 \end{cases}$

and $\gamma = \sqrt{\frac{1}{2}(\alpha^2 - v_{21}\alpha)}$.