Inference of Adaptive methods for Multi-Stage skew-t Simulated Data

Loai M. A. Al-Zou'bi,
Amer I. Al-Omari,
Ahmad M. Al-Khazalah,
Al al-Bayt University, Department of Mathematics, Mafraq , Jordan
Raed A. Alzghool
Al-Balqa Applied University, Department of Mathematics, Salt, Jordan

doi: 10.19044/esj.2017.v13n24p448 URL:http://dx.doi.org/10.19044/esj.2017.v13n24p448

Abstract

Multilevel models can be used to account for clustering in data from multi-stage surveys. In some cases, the intra-cluster correlation may be close to zero, so that it may seem reasonable to ignore clustering and fit a single level model. This article proposes several adaptive strategies for allowing for clustering in regression analysis of multi-stage survey data. The approach is based on testing whether the cluster-level variance component is zero. If this hypothesis is retained, then variance estimates are calculated ignoring clustering; otherwise, clustering is reflected in variance estimation. A simple simulation study is used to evaluate the various procedures.

Keywords: Adaptive estimation, variance components, cluster sampling, multi-level models, Huber-White variance estimator, skew-t distribution, balanced data.

1 Introduction

1.1 Cluster and Multistage Sampling

The basic idea of sampling is to make inference about the population of interest. The use of probability methods to minimize decision in the choice of survey units will provide good designs in sampling methods (Cochran, 1977). We have many sampling methods; one of them is the two-stage sampling. It is used in many surveys of social, health, economics and demographic topics (Kish, 1965). In this type of sampling the population is divided into groups called primary sampling units (PSUs), and then a simple random sample from each PSU is selected. If we select all elements within each considered PSU then two-stage sampling is called cluster sampling (Cochran, 1977). One of the main advantages of cluster sampling is that it reduces the cost; actually it is the most economical form of sampling designs, (Bennet et al., 1991). It, almost always reduces the listing and travel cost. However, the sampling variance calculated using cluster sampling is greater than the sampling variance calculated using simple random sampling. For example, PSUs might be schools and units might be students in schools, or PSUs might be households and units might be people, or PSUs might be geographic areas and units might be households (see for example Snijders and Bosker, 1999; Cochran, 1977; Kish, 1965).

Two-stage sampling is typically used because

- There is no sampling frame of final units, but a frame of PSUs is available.
- Cost efficiency; for example it is much cheaper to draw a two-stage sample of 100 persons from 10 households than draw a simple random sample of 100 persons, as those persons might be dispersed over 100 households (Snijders and Bosker, 1999).
- Within-group correlations may be of interest in their own right. For example, the correlation between values for students in the same school might be of interest.

The intraclass correlation (ICC), ρ, measures the similarity within PSUs for a particular variable (Killip and Pearce, 2004). Therefore, it is given by $\rho = \frac{\sigma_b^2}{\sigma_b^2+\sigma_e^2}$ (Kish, 1965; Commenges and Jacqmin, 1994).

In practice the ICC is often quite small. It is zero if units within PSUs are no more homogenous than units over all PSUs. It is 1, as well if units from the same PSU have equal values. It is close to zero for many variables (e.g. age and sex), and small for others (for example, $\rho = 0.03$ to 0.05) (Nations, 2005). It may take a negative value, but in practice it is generally positive. If each PSU in the population contains M units, the smallest possible value of ρ is $-1/(M-1)$. This occurs when the population is finite with high heterogeneity within PSUs, and zero variance between PSU means (Hansen et al., 1953, p.260, show this for repeated probability sampling from a fixed finite population).

The ICC can be high for some variables (for example, for clinical studies if PSUs are villages (PSUs) and elements are persons in these villages who have or not have access to clinics in these villages). ρ is usually less than 0.1 when PSUs are geographic areas and final units are households in these areas (Verma et al., 1980). When PSUs are households and final units are people in households it is usually between 0 and 0.2 (Clark and Steel, 2002).

Multilevel models are generalization of linear regression models. Assume a dependent variable of interest, y_{ij}, and a vector of covariates for unit j in primary sampling unit (PSU) i, x_{ij}. Goldstein (2003) defined the two-level linear mixed model (LMM) by

$$y_{ij} = \boldsymbol{\beta}' x_{ij} + b_i + e_{ij}, \quad i = 1,2,\dots,c, \quad j = 1,2,\dots,m_i \tag{1}$$

where c denotes the number of PSUs in the sample, m_i denotes the number of observations selected in PSU i, $\boldsymbol{\beta}$ is the vector of unknown regression coefficients, b_i is a PSU specific random effect with variance σ_b^2, we assume that b_i has a skew-t distribution with parameter ν and λ, $b_i \sim skt(\nu, \lambda)$.

In Al-Zou'bi et al. (2010), the methods were based on fitting a linear mixed model. Data were assumed to be normally distributed. These methods were applied in Abushrida et al. (2014) and Al-Zoubi (2015) on data generated from skewed normal and exponential distributions, respectively. In this article, the plan is to see if these methods still working for balanced two-stage skew-t data rather than normal data.

The skewed distributions have been grown widely in applications and research in the last three decades, from the idea of skewed normal distribution introduced by Azzalini (1985). The popularity of skewed distributions is due to many reasons:

- Flexibility of modelling skewed data
- Easily extended of their symmetric counterparts;
- have a number of properties of standard symmetric distributions; and
- have well-mannered multivariate versions.

Azzalini and Capitanio (2003) introduced another family of skewed distributions, called the skew-t for which the symmetric base distribution is a heavy-tailed Student's t distribution. This distribution is a flexible model, which is able to deal with skewness and kurtosis in the data. It affords an alternative to the implementation of robust procedures, when normality assumption is not reached. In fact the skew-t likelihood function can be used as robust likelihood in an adaptive estimation procedure. Jones (2003) proposed a family of distributions which includes the the symmetric t-distribution as special cases. This family includes extensions of the t-distribution as well as a non-zero skewness. If $a > 0$ and $b > 0$ be the parameters then the pdf of the random variable X which follows this new distribution is given by:

$$f(x) = f(x; a, b) = C_{a,b}^{-1} \left[1 + \frac{x}{(a+b+x^2)^{\frac{1}{2}}}\right]^{a+\frac{1}{2}} 1 - \left[\frac{x}{(a+b+x^2)^{\frac{1}{2}}}\right]^{b+\frac{1}{2}} \tag{2}$$

where $C_{a,b} = 2^{a+b-1} B(a,b)(a+b)^{\frac{1}{2}}$, $B(.,.)$ denotes the beta function. f reduces to the t-distribution with $2a$ degrees of freedom when $a = b$. When $a < b$, f is negatively skewed and it is positively skewed when $a > b$. Also,

$f(x;a,b) = f(-x;a,b)$. Charalambides et al. (2001) derived the mean and the variance of the skew t random variable X; respectively as:

$$E(X) = \mu = \frac{(a+b)^{\frac{1}{2}}(a-b)\Gamma\left(a-\frac{1}{2}\right)\Gamma\left(a\mathrm{b}-\frac{1}{2}\right)}{\Gamma(a)\Gamma(b)};$$

$$Var(X) = \frac{a+b}{4}\left[\frac{(a-b)^2+a+b-2}{(a-1)(b-1)} - \left\{\frac{(a-b)\Gamma\left(a-\frac{1}{2}\right)\Gamma\left(a\mathrm{b}-\frac{1}{2}\right)}{\Gamma(a)\Gamma(b)}\right\}^{\frac{1}{2}}\right]. \quad (3)$$

where $a+b=\nu$, ν is the degrees of freedom and $a-b=\lambda$, $\lambda>0$ is the shape parameter. The skew-t distribution reduces to the standard Student's t distribution when $a=b$ (Zhang et al., 2013). As well as the standard Student's t distribution, the skew t includes the skew normal distribution when $\nu \to \infty$ and the normal distribution when $a=b$ and $\nu \to \infty$.

2 Fitting the linear mixed model

2.1 The model

Model (1) can be written in a general form as:

$$Y = X\boldsymbol{\beta} + \mathbf{b} + e, \quad (4)$$

where X is the design matrix with dimension $n \times p$ and it is assumed to be of rank p and $Y = (y'_1, \dots, y'_c)'$ be the complete set of $n = \sum_{i=1}^{c} m_i$ observations in the c PSUs, where $y_i = (y_{i1}, \dots, y_{im_i})'$ is the observed vector for the i^{th} PSU, b is a vector $n \times 1$ vector of random coefficients. The variance of Y is defined to be V, where V is a block diagonal matrix, $V = diag(V_i, i=1,\dots,c)$, and

$$V_i \quad = \quad \sigma_b^2 J_{m_i} + \sigma_e^2 I_{m_i}, \quad (5)$$

where J_{m_i} is an $m_i \times m_i$ matrix with all entries are ones, and I_{m_i} is the $m_i \times m_i$ identity matrix. $\boldsymbol{\beta}$ describes patterns of change in the mean response over time in the population of interest.

If we set x_{ij} to 1 $\forall(i,j)$ then model (1) will reduce to the intercept-only model. This model includes just a grand mean parameter, it is defined as

$$y_{ij} = \beta + b_i + e_{ij}, \quad i = 1,2,\dots,c, \quad j = 1,2,\dots,m_i, \quad (6)$$

where c denotes number of the sample PSUs, m_i denotes the number of units selected in PSU i, $b_i \overset{iid}{\sim} skt(\nu,\lambda)$ represents the i^{th} individual's deviation from the population mean intercept after the effect of covariates have been accounted for with variance σ_b^2, and e_{ij} is assumed to be $skt(\infty, 0)$ with variance σ_e^2. The parameters σ_b^2 and σ_e^2 are the between- and within-PSUs variance components. Observations for different units from the same PSU are correlated. It is assumed that b_i is uncorrelated with e_{ij}, and that b_i and $b_{i'}$ for $i \neq i'$ are uncorrelated also (Rao, 1997).

2.2 Likelihood Theory Estimation of $\mathrm{var}(\widehat{\boldsymbol{\beta}})$

The variances of the estimated regression coefficients and their estimators will be discussed in this section. The estimated variance of the REML estimate of the regression coefficients, $\widehat{\boldsymbol{\beta}}_R$, is given by

$$\left.\begin{aligned} \widehat{var}(\hat{\beta}_R) &= (X'V^{-1}X)^{-1} \\ &= \left(\sum_{i=1}^{c} x'_i V_i^{-1} x_i\right)^{-1} \end{aligned}\right\} \quad (7)$$

This will simplify in the balanced data case, $m_i = m$, to

$$\widehat{var}(\hat{\beta}_R) = \frac{1}{c}\left[\hat{\sigma}_b^2 + \frac{\widehat{\sigma_e^2}}{m}\right] \quad (8)$$

Hence the $(1-\alpha)100\%$ confidence interval for β could be obtained as

$$(1-\alpha)100\%CI = \hat{\beta}_R \pm t_{\left(df,1-\frac{\alpha}{2}\right)}\sqrt{\widehat{var}\left(\hat{\beta}_R\right).} \quad (9)$$

There is no clear form for the degrees of freedom in (9) for mixed models. This paper will use Faes et al. (2009) approach which relies on the effective sample size (ne) as degrees of freedom for mixed models, with $\widehat{ne} = \frac{n}{\widehat{deff(\beta)}}$. Other approaches were suggested by Satterthwaite (1941) and Kenward and Roger (1997) for defining the degrees of freedom. Faes et al. (2009) approach is preferred over other approaches as it extends naturally to non-normal models.

2.3 Huber-White Estimator of $var(\widehat{\boldsymbol{\beta}})$

The generalized estimating equation (GEE) (Liang and Zeger, 1986) is a general approach for modeling correlated data (Burton et al., 1998). It is an alternative to the ML and REML approaches for modeling longitudinal (Diggle et al., 1994). GEE is used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes (Liang and Zeger, 1986; Hardin and Hilbe, 2003). This approach can use either ordinary least squares (OLS) or generalized least squares (GLS) to linear modeling of clustered data. GEEs have consistent and asymptotically normal solutions even with misspecification of the correlation structure (Hedeker and Gibbons, 2006).

The OLS estimator for $\boldsymbol{\beta}$ is defined by

$$\widehat{\boldsymbol{\beta}}_{ols} = (X'X)^{-1}X'Y. \quad (10)$$

The OLS estimator of $\boldsymbol{\beta}$ is unbiased (Scott and Holt, 1982) when the same PSU observations are correlated with common intraclass correlation ρ but the observations from different PSUs are uncorrelated. Therefore

$$var(\widehat{\boldsymbol{\beta}}_{ols}) = (X'X)^{-1}X'VX(X'X)^{-1}. \qquad (11)$$

As V is not known, in general, and it can be estimated by V, therefore Equation (11) becomes:

$$\widehat{var}(\widehat{\boldsymbol{\beta}}_{ols}) = (X'X)^{-1}X'\hat{V}X(X'X)^{-1}. \qquad (12)$$

In Equation (7), $\widehat{var}(\widehat{\boldsymbol{\beta}})$ was estimated by substituting REML estimates of σ_b^2 and σ_e^2 into V_i. An alternative estimator of V_i is $V_i^{Hub} = e_i e_{i'}$, where $e_i = y_i - x'_i\widehat{\boldsymbol{\beta}}$. V_i^{Hub} is approximately unbiased for V_i even if (3) does not apply. Note that

$$\begin{aligned} var(\widehat{\boldsymbol{\beta}}) &= var((\textstyle\sum_{i=1}^{c} x'_i V_i^{-1} x_i)^{-1}(\sum_{i=1}^{c} x'_i V_i^{-1} y_i)) \\ &\approx (\textstyle\sum_{i=1}^{c} x'_i V_i^{-1} x_i)^{-1}(\sum_{i=1}^{c} x'_i V_i^{-1} V_i V_i^{-1} x_i) \qquad (13) \\ &\quad (\textstyle\sum_{i=1}^{c} x'_i V_i^{-1} x_i)^{-1}. \end{aligned}$$

Huber-White variance estimators are valid even if the variance-covariance model is substantially incorrect since they give unbiased estimators of these parameters as long as there are no covariances between observations from units in different groups.

The estimator $\widehat{var}(\widehat{\boldsymbol{\beta}})$ in (7) is approximately unbiased provided that the variance model (3) is correct. Otherwise, $\widehat{var}(\widehat{\boldsymbol{\beta}})$ will be biased and inference will be incorrect. The robust variance estimate approach described by Liang and Zeger (1986) is an alternative to ML or REML estimates of $var(\widehat{\boldsymbol{\beta}})$ in the context of modeling longitudinal data using GEEs. This approach can be applied to the analysis of data collected using PSUs, where observations within PSUs might be correlated and the observations in different PSUs are independent.

This approach can be referred to as robust or Huber-White variance estimation (Huber, 1967; White, 1982). It will be used as an alternative approach to estimating $var(\widehat{\boldsymbol{\beta}})$. The method yields asymptotically consistent covariance matrix estimates even if the variances and covariances assumed in model (1) are incorrect. It is still necessary to assume that observations from different PSUs are independent. A robust estimator of $var(\widehat{\boldsymbol{\beta}})$ can be constructed by substituting the robust estimator V_i^{Hub} in (13) (Liang and Zeger, 1986).

In the balanced data case and the intercept only model (x_{ij}=1), this variance becomes

$$\widehat{var}(\hat{\beta}) = \frac{1}{c(c-1)}\sum_{i=1}^{c} (\bar{y}_{i.} - \bar{y}_{..})^2 \qquad (14)$$

Exact confidence intervals can then be calculated with degrees of freedom equal to c-1 (MacKinnon and White, 1985).

2.4 Restricted Likelihood Ratio Test (RLRT)

A better option is to use REML estimators to derive the likelihood ratio test (LRT) statistic for testing $H_0: \sigma_b^2 = 0$.

The problem of testing $H_0: \sigma_b^2 = 0$ using the likelihood ratio test is discussed by (MacKinnon and White, 1985). using ML estimators for the variance components. Self and Liang (1987) allowed the true parameter values to be on the boundary of the parameter space, and showed that the large sample distribution of the likelihood ratio test is a mixture of χ^2 distributions under nonstandard conditions assuming that response variables are iid. This assumption does not generally hold in linear mixed models, at least under the alternative hypothesis.

The restricted log-likelihood function is given by West et al. (2007, p.28) as

$$\begin{aligned} \ell_R \quad &= \quad -\tfrac{1}{2}[(n-1)log(2\pi) + log|V| + log|X'V^{-1}X| \\ &\qquad +Y'V^{-1}\{I - X(X'V^{-1}X)^{-1}X'\}V^{-1}Y], \end{aligned} \qquad (15)$$

where $V = diag(V_i)$ and V_i are given by (3). Maximizing (15) with respect to σ_b^2 and σ_e^2 gives the REML estimates of these parameters.

From (15), the restricted likelihood ratio test is given by

$$\begin{aligned} \Lambda &= -2\ \ log(RLRT) \\ &= 2\overset{MAX}{H_A}\ \ \ell_R(\boldsymbol{\beta}, \sigma_b^2, \sigma_e^2) - 2\overset{MAX}{H_0}\ \ \ell_R(\boldsymbol{\beta}, \sigma_b^2, \sigma_e^2). \end{aligned} \qquad (16)$$

In the intercept-only model case Visscher (2006) defined the REML-based likelihood ratio test (RLRT) as

$$\Delta = -2\log(LRT) = \begin{cases} (n-1)\ log\left(\frac{n-c}{n-1} + \frac{c-1}{n-1}\ F\right) - (c-1)\ log(F) & if \quad F > 1 \\ 0 & if \quad F \leq 1. \end{cases} \qquad (17)$$

where $F = \frac{MSB}{MSE}$, MSB is the mean square between PSUs and MSE is the within PSUs mean square error.

The large sample distribution of the likelihood ratio Λ is a 50:50 mixture of χ^2 distribution with 0 and 1 degrees of freedom as the parameter values fall on the boundary of the parameter space (Self and Liang, 1987).

3 Adaptive strategies

Two adaptive strategies are employed in this paper both of them relying on testing $H_0: \sigma_b^2 = 0$. The first uses $var(\hat{\beta})$ defined in Equation (7) if H_0 is rejected and the standard LM with independent errors if H_0 is not rejected. This strategy is explained in Figure 1, where $\widehat{var}_{\mathrm{LM}}(\hat{\beta})$ is the estimator of $var_{\mathrm{LM}}(\hat{\beta})$ using the LM strategy, $\widehat{var}_{\mathrm{LMM}}(\hat{\beta})$ is the estimator of $var_{\mathrm{LMM}}(\hat{\beta})$ using the LMM strategy and $\widehat{var}_{\mathrm{ADM}}(\hat{\beta})$ is the adaptive estimator.

Figure 1: Flowchart explaining the adaptive procedure using the estimated variance extracted from the LMM

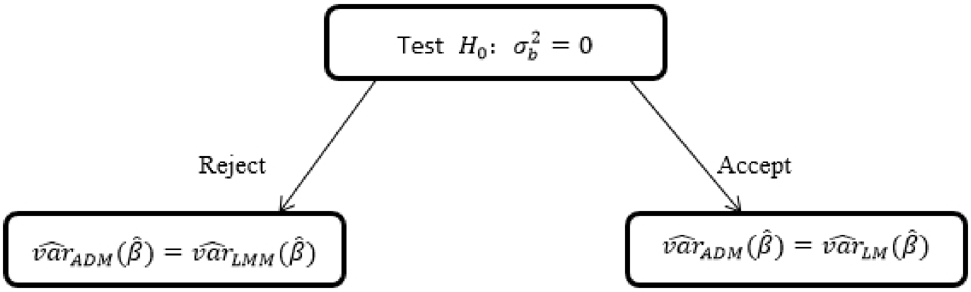
Test H_0: $\sigma_b^2 = 0$
Reject
Accept
$\widehat{var}_{ADM}(\hat{\beta}) = \widehat{var}_{LMM}(\hat{\beta})$
$\widehat{var}_{ADM}(\hat{\beta}) = \widehat{var}_{LM}(\hat{\beta})$

Figure 2: Flowchart explaining the adaptive procedure using Huber-White estimator

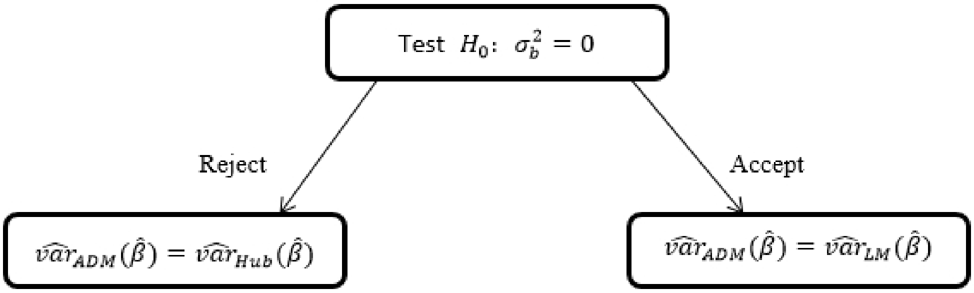
Test H_0: $\sigma_b^2 = 0$
Reject
Accept
$\widehat{var}_{ADM}(\hat{\beta}) = \widehat{var}_{Hub}(\hat{\beta})$
$\widehat{var}_{ADM}(\hat{\beta}) = \widehat{var}_{LM}(\hat{\beta})$

Figure 2 explains the second adaptive strategy which is similar to the first, except that it uses the Huber-White variance estimator ($\widehat{var}_{\mathrm{Hub}}(\hat{\beta})$ if H_0 is rejected.

The advantage of the adaptive strategy is that we use the simple linear model to derive variance estimators, unless there is strong evidence that $H_0: \sigma_b^2 > 0$. This has benefit of simplifying the model and may also give tighter confidence intervals. However, it is not clear whether the adaptive approaches will give valid confidence intervals for β, because the confidence intervals assume non-adaptive procedures.

A simulation study was performed to compare the adaptive and non-adaptive methods for estimating $var(\hat{\beta})$ using PSUs with equal sample sizes. Data were generated from model (6), with equal PSU sizes, $m_i = m$, assuming b_i has a skew-t distribution. A range of values of ρ was used: 0, 0.025 and 0.1. The number of PSUs, c, and the number of observations in each PSU, m, were varied over a range of values of 2, 5, 10 and 25. In each case 1000 samples were generated. The hypothesis $H_0\colon \sigma_b^2 = 0$ was tested as described in Subsection 2.4 using the RLRT defined by Equation (17)

Tables 1 - 3 show the balanced data case results for values of ρ = 0, 0.025 and 0.1 using the ADM, ADH, LMM and Huber strategies of estimation. The ratio of the mean estimated variance to the estimated variance of $\hat{\beta}$, $E(\widehat{var}(\hat{\beta}))/var(\hat{\beta})$ was shown in these tables. A significance level of $\alpha = 0.1$ is used for testing $\sigma_b^2 = 0$. It is assumed that $\beta = 0$ in all tables. The non-coverage rates of 90% confidence intervals for β are shown also in the tables. They show the average lengths of these confidence intervals as well as the proportion of samples where $H_0\colon \sigma_b^2 = 0$ was rejected.

The ADM and ADH variance estimators were approximately biased for $\rho = 0$, when there were 5 sample PSUs or less for all values of m as well as when $c = 10$ with $m = 5$ and 10. On the other hand for $\rho = 0.025$, the variance estimators were unbiased for all values of c and m except for $c = 2$ with $m \leq 5$. For $\rho = 0.1$, the variance estimators almost always were unbiased.

The LMM variance estimators were approximately biased for all all values of c and m, whe $\rho = 0$. For $\rho = 0.025$, these estimators were biased for $c \leq 5$ with all values of m and for $c = 10$ with $m \leq 5$. The variance estimators were also biased when $\rho = 0.1$ for $c = 2$ with $m \leq 5$ and for $c = 5$ with $m = 2$. Otherwise, they were biased.

The LMM variance estimators were almost biased for $\rho = 0$ when there were 10 or less sample PSUs, and when there were 25 sample PSUs with 5 or 10 observations each. For $\rho = 0.025$, these estimators were biased for $c \leq 5$ for all values of m and when $c = 10$ with $m \leq 5$. For $\rho = 0.1$, the LMM variance estimators were almost unbiased except for $c = 2$ with $m \leq 5$ and for $c = 5$ with $m = 2$.

The Huber-White variance estimators were, in general, unbiased regardless the values of ρ, c and m.

Non-coverage rates for β were pretty close to the nominal rate (10%) when $\rho = 0$ for all methods. These rates were far from the nominal rate for $\rho = 0.025$ with about 7-61% using TADM, TADH and TLMM methods. When $\rho = 0.1$, these rates were 7-56% higher than nominal rate for $c \geq 5$ and for $c = 2$ with $m \leq 5$ and much higher when $c = 2$ with $m > 5$. Huber non-coverage rates were pretty close to the nominal rate in all cases.

The adaptive confidence intervals were shorter than their corresponding non-adaptive confidence interval when $\rho = 0$ for $c \leq 5$. Otherwise, they were approximately equal. For $\rho \neq 0$ the adaptive confidence intervals were shorter.

5 Conclusion

i. The performance of adaptive confidence intervals is poor in extreme designs with 2 sample PSUs with 2 and 5 observations each and with 5 sample PSUs with 2 observations each. In designs with 5 PSUs or less none-coverage rates are higher than desirable non-coverage (10%) for all values of $\rho \neq 0$. Therefore, clustering must be allowed for in variance estimates, even if it is not statistically significant.

ii. In comparing the Linear Mixed Model with the adaptive version (ADM), we find that:

- ADM approach has close to nominal non-coverage when $\rho = 0$.
- The ADM confidence intervals are narrower (5-25%) than the LMM for $c \leq 10$.

iii. In comparing the robust Huber-White approach with the adaptive version (ADH), we find that:

- The Huber approach has close to nominal non-coverage in all cases. So does the ADH approach, except for the designs mentioned in i.
- The Huber method gives wide confidence intervals when $c \leq 10$ with order of 10-75%. The reason for that is the degrees of freedom for Huber method is equal to (c-1), while the ADH method degrees of freedom are equal to (n-1) if the PSU-level variance component is not significant and (c-1) if it is significant.

Table 1: Variance ratios, average length and non-coverage of the 90% confidence intervals for β, and power of testing $H_0\colon \sigma_b^2 = 0$ using RLRT in the balanced data case with ρ=0.

PSUs	Obs	$E(\widehat{var}(\hat{\beta}))/var(\hat{\beta})$				Non-Coverage of CI for β (%)				Pr(Rej H_0) (%)	Confidence Interval Length			
C	m	VADM	VADH	VLMM	VHub	TADM	TADH	TLMM	THub	RLRT	CADM	CADH	CLMM	CHub
2	2	1.251	1.251	1.49	1.068	8.9	8.9	11.8	10.5	9.9	4.915	2.908	5.491	5.047
	5	1.169	1.169	1.373	0.944	10.1	10	10.3	8.4	5.4	1.272	1.538	1.349	3.172
	10	1.185	1.185	1.432	0.975	9.4	9.4	9.4	9.6	4.7	0.852	1.011	0.937	2.26
	25	1.208	1.208	1.454	1.008	10.3	10.3	11.5	8	4.3	0.537	0.638	0.59	1.443
5	2	0.934	0.934	1.029	0.846	10.2	10.3	10.1	10.5	8.6	1.159	1.164	1.176	1.228
	5	1.126	1.126	1.268	1.068	9	8.9	9.5	9.1	8.6	0.723	0.73	0.75	0.828
	10	1.118	1.118	1.241	1.012	9.2	9.2	10.1	10.5	7.8	0.505	0.51	0.52	0.573
	25	1.109	1.11	1.237	0.979	9.2	9.2	9.3	10.1	6.5	0.316	0.318	0.326	0.355
10	2	1.126	1.126	1.206	1.07	8.5	8.2	8.5	8.9	11.1	0.786	0.787	0.794	0.797
	5	1.024	1.024	1.08	0.958	9.2	9.1	9.6	9.5	9.3	0.494	0.495	0.494	0.509
	10	1.141	1.141	1.213	1.043	7.3	7.3	9	10.2	7.4	0.345	0.345	0.345	0.353

	25	1.008	1.008	1.065	0.922	10.3	10.4	11.1	11.8	7.5	0.217	0.218	0.217	0.224
25	2	1.063	1.063	1.111	1.032	9.3	9.3	9.8	10.6	10	0.48	0.48	0.481	0.48
	5	1.06	1.06	1.068	1.013	9.5	9.5	9.7	9.6	9.6	0.305	0.304	0.3	0.304
	10	1.012	1.012	1.022	0.968	10	10	10.9	10.6	8.5	0.214	0.214	0.212	0.214
	25	1.096	1.096	1.111	1.043	8.6	8.6	8.9	8.5	8.4	0.135	0.135	0.133	0.135

Table 2: Variance ratios, average length and non-coverage of the 90% confidence intervals for β, and power of testing $H_0: \sigma_b^2 = 0$ using RLRT in the balanced data case with ρ=0.025.

PSUs	Obs	$E(\widehat{var}(\hat{\beta}))/var(\hat{\beta})$				Non-Coverage of CI for β (%)				Pr(Rej H_0) (%)	Confidence Interval Length			
c	m	VADM	VADH	VLMM	VHub	TADM	TADH	TLMM	THub	RLRT	CADM	CADH	CLMM	CHub
2	2	1.118	1.118	1.325	0.965	10.4	10.4	13.1	11.2	8.9	4.673	2.874	5.27	5.089
	5	1.12	1.12	1.334	0.958	11.6	11	13.6	9.8	7.8	1.301	1.697	1.402	3.362
	10	1.054	1.054	1.324	0.993	13.8	13.8	12.6	9.9	6.6	0.89	1.145	1.024	2.526
	25	1.063	1.063	1.302	1.082	16.1	16.1	15.3	9.4	13	0.679	0.992	0.785	1.908
5	2	1.087	1.087	1.206	1.011	9.2	9	10	10.1	10.3	1.21	1.218	1.234	1.296
	5	1.072	1.072	1.185	1.024	10.7	10.7	10.2	8.9	11.6	0.751	0.76	0.775	0.856
	10	1.018	1.018	1.143	1.004	11.8	11.6	11.2	9.6	13.3	0.539	0.546	0.565	0.632
	15	1.074	1.074	1.207	1.086	12.4	12.2	12	8.4	18.9	0.466	0.473	0.491	0.548
10	2	1.022	1.022	1.101	0.988	8.6	8.5	8.9	9.7	10.3	0.797	0.8	0.807	0.818
	5	1.042	1.042	1.117	1.029	10.4	10.3	9.8	10.3	13.5	0.509	0.511	0.519	0.538
	10	0.93	0.93	1.005	0.961	12.3	12.2	11.6	10.4	20.1	0.374	0.374	0.383	0.405
	25	0.959	0.959	1.025	1.016	13.4	13.3	12.3	9.2	41.5	0.272	0.273	0.282	0.295
25	2	0.997	0.997	1.044	0.977	10.3	10.5	10.9	10.8	11.7	0.486	0.487	0.485	0.489
	5	1.021	1.021	1.039	1.032	9.9	9.9	9.9	9.3	16.9	0.315	0.314	0.314	0.322
	10	0.922	0.922	0.946	0.965	10.6	10.2	10.2	9.1	29.9	0.233	0.231	0.234	0.242
	25	0.939	0.939	0.953	0.972	12.8	12.6	12.8	10.1	65.8	0.17	0.17	0.172	0.175

Table 3: Variance ratios, average length and non-coverage of the 90% confidence intervals for β, and power of testing $H_0: \sigma_b^2 = 0$ using RLRT in the balanced data case with ρ=0.1.

PSUs	Obs	$E(\widehat{var}(\hat{\beta}))/var(\hat{\beta})$				Non-Coverage of CI for β (%)				Pr(Rej H_0) (%)	Confidence Interval Length			
c	m	VADM	VADH	VLMM	VHub	TADM	TADH	TLMM	THub	RLRT	CADM	CADH	CLMM	CHub
2	2	1.008	1.008	1.204	0.925	11.1	11.1	13.7	10.8	11.3	5.562	3.132	6.336	5.464
	5	1.023	1.023	1.224	0.987	15.4	14.8	15.1	11.1	12.1	1.523	2.138	1.668	3.982
	10	0.923	0.923	1.07	0.929	21.9	21.7	18.5	10.9	17.8	1.183	1.964	1.324	3.302
	25	0.851	0.851	0.941	0.887	29.1	29	22.2	8.4	32	1.158	2.024	1.276	2.867
5	2	1.012	1.012	1.128	1.002	11.2	11.1	11.2	10	17.1	1.291	1.305	1.344	1.422
	5	0.936	0.936	1.048	0.975	13.4	13.2	12.6	10.4	22.4	0.858	0.875	0.907	1.01
	10	0.976	0.977	1.065	1.034	17.2	16.8	13.9	10.2	40.2	0.721	0.738	0.762	0.837
	25	0.989	0.989	1.02	1.012	15.6	14.8	13.5	10	68.3	0.666	0.675	0.684	0.714
10	2	0.992	0.992	1.072	1	11	10.7	10.7	10.3	20	0.858	0.86	0.879	0.896
	5	0.838	0.838	0.899	0.882	14.7	14.9	13.7	12.5	33.6	0.588	0.587	0.605	0.631
	10	0.876	0.876	0.917	0.915	15	14.8	14	11.4	57.5	0.489	0.49	0.503	0.518
	25	0.971	0.971	0.98	0.981	10.9	10.3	9.8	8.7	89.6	0.446	0.443	0.45	0.45

25	2	0.993	0.994	1.049	1.009	9.9	9.6	10.3	9.9	20.6	0.517	0.516	0.522	0.527
	5	0.9	0.9	0.913	0.938	13.5	13.5	13	11.8	54.1	0.369	0.367	0.371	0.38
	10	1.003	1.003	1.007	1.013	9.3	9.2	9.6	8.7	88.4	0.316	0.313	0.317	0.316
	25	0.939	0.939	0.953	0.972	12.800	12.600	12.800	10.100	65.800	0.170	0.170	0.172	0.175

References:

1. Abushrida, R. M., Al-Zoubi, L. M., and Al-Omari, A. I. (2014). Robustness of adaptive methods for balanced non-normal data: Skewed normal data as an example. Revista Investigatio´n Operacional, 35(2):157-172.
2. Al-Zoubi, L. M. (2015). Adaptive inference for multi-stage unbalanced exponential survey data based on a simulation from an intercept-only model. Electronic Journal of Applied Statistical Analysis, 8(2):136-153.
3. Al-Zou'bi, L. M., Clark, R. G., and Steel, D. G. (2010). Adaptive inference for multi-stage survey data.
4. Communications in Statistics - Simulation and Computation, 39(7):1334-1350.
5. Azzalini, A. (1985). Aclass of distributions which includes the normal ones. Scandinavian Journal of Statistics, 12(2):171-178.
6. Azzalini, A. and Capitanio, A. (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t distribution. Journal of Royal Statistical Society, Series B, 65:367-389.
7. Bennet, S., Woods, T., Liyanage, W. M., and Smith, D. L. (1991). A simplified general method for cluster-sample surveys of health in developing countries. World Health Statistics Quarterly, 44(3):98-106.
8. Burton, P., Gurin, L., and Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. Statistics in Medicine, 17:1261-1291.
9. Charalambides, C., Koutras, M., and Balakrishnan, N. (2001). Probability and Statistical Models with Applications. CHAPMAN & HALL/CRC.
10. Clark, R. G. and Steel, D. G. (2002). The effect of using household as a sampling unit. International Statistical Review, 70(2):289-314.
11. Cochran, W. G. (1977). Sampling Techniques. John Wiley and Sons, Inc.
12. Commenges, D. and Jacqmin, H. (1994). The intraclass correlation coefficient: Distribution-free definition and test. Biometrics, 50(2):517 - 526.

13. Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). Analysis of Longitudinal Data. Clarendon Press, New York.
14. Faes, C., Molenberghs, H., Aerts, M., Verbeke, G., and Kenward, M. G. (2009). The effective sample size and an alternative small-sample degrees-of-freedom method. The American Statistician, 63(4):389-399.
15. Goldstein, H. (2003). Multilevel Statistical Models. Kendall's Library of Statistics 3. Arnold, London, third edition.
16. Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953). Sample Survey Methods and Theory, volume 1, 2. John Wiley and Sons, Inc., New York.
17. Hardin, J. and Hilbe, J. (2003). Generalized Estimating Equation. London, Chapman and Hall/CRC.
18. Hedeker, D. and Gibbons, R. D. (2006). Longitudinal Data Analysis. Wiley Interscience.
19. Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. Proceedings of the Fifth Berekley Symposium on Mathematical Statistics and Probability, University of California, Berekley, 11:221-233.
20. Jones, M.C., F. M. (2003). A skew extension of the t-distribution, with applications. Journal of the Royal Statistical Society: Series B (Statistical Methodology, 65(1):159-174.
21. Kenward, M. G. and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. Biometrics, 53(3):983-997.
22. Killip, S. amd Mahfoud, Z. and Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. Anaals of Family Medicine, 2(3):204-208.
23. Kish, L. (1965). Survey Sampling. John Wiley and Sons, Inc.
24. Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1):13-22.
25. MacKinnon, J. G. and White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics, 29:305-325.
26. Nations, U. (2005). Household Sample Surveys in Developing and Transition Countries. Number 96 in F. Department of Economic and Social A airs, Statistics Division, Studies and Methods.
27. Rao, P. S. R. S. (1997). Variance Components Estimation, Mixed Models, Methodologies and Applications. Chapman and Hall.
28. Satterthwaite, F. E. (1941). Synthesis of variance. Psychometrika, 6:309-316.

29. Scott, A. J. and Holt, D. (1982). The effect of two-stage sampling on ordinary least squares methods. Journal of the American Statistical Association, 77(380):848-854.
30. Self, S. G. and Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions Journal of the American Statistical Association, 82(398):605-610.
31. Snijders, T. A. B. and Bosker, R. J. (1999). Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. SAGE publications.
32. Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample design and sampling errors for the world fertility survey. Journal of the Royal Statistical Society, Series A (General), 143(4):431-473.
33. Visscher, P. M. (2006). A note on the asymptotic distribution of likelihood ratio tests to test variance components. Twin Research and Human Genetics, 9(4):490-495.
34. West, B. T., Welch, K. B., and Galecki, A. T. (2007). Linear Mixed Model: A Practical Guide Using Statistical Software. Chapman and Hall/CRC, Boca Raton, Florida.
35. White, H. (1982). Maximum likelihood estimation of misspecified models. Econometrica, 50(1):1-25.
36. Zhang, Y., Liao, X., and Li, T. (2013). Mills ratio of skew-t distribution and their applications. Proceedings 59th ISI World Statistics Congress, 25-30 August 2013, Hong Kong (Session CPS110).