An Empirical Assessment of the Performance of Variance Components Tests under Contaminated Error Distributions with Application to Random-Intercept Regression Models

Yahia S. El-Horbaty
Department of Mathematics, Insurance and Applied Statistics, Helwan University, Egypt

Abstract

Testing zero variance components is a common practice under random-intercept models. Various tests exist to check the need for random effects in such models. Although many of those tests have correct Type-I error rates even when the error components are not normally distributed, an empirical assessment of the performance of these tests when the distribution is contaminated in the form of possessing heavy tails, heavy skewness, or contains outliers does not exist. This article investigates the performance of four recently proposed variance components tests under such violations using extensive simulation studies. Results indicate that the simulation-based test based on the likelihood ratio test statistic is much preferred to the other tests unless the response space suffers from the presence of outliers. Under the latter case, none of the competing tests revealed satisfactory performance.

Keywords Heavy-Skewed Distribution, Outliers, Variance Components, Likelihood Ratio Test.

1. Introduction

The random-intercept model is a famous two-level model that can be used in various applications (Goldstein, 2011). Assume that the data is collected from m main groups, and nested within those groups are n_i observations per group where $i = 1, \dots, m$. The model can be represented as

$$[1] \qquad y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij},$$

where y_{ij} denotes the j^{th} response ($j = 1, \dots, n_i$) in the i^{th} group, $\boldsymbol{x}_{ij}^T$ is a p-vector of explanatory variables, $\boldsymbol{\beta}$ is a p-vector of fixed effects, u_i denotes a common random effect in the i^{th} group, and e_{ij} denotes the residual error. Both u_i and e_{ij} are independently assumed to follow a distribution with mean zero and constant variance, say σ_u^2 and σ_e^2, respectively.

Generally, testing zero variance components gained fame over the last two decades. Although the null hypothesis of this test simply equates $\sigma_u^2 = 0$, the test remains challenging as the asymptotic distribution of the likelihood ratio test statistic is not tractable. Crainiceanu and Ruppert (2004) proposed obtaining the finite sample distribution of the likelihood ratio test (LRT) using a simulation-based approach. Fitzmaurice *et al.* (2007) proposed a permutation test for evaluating the performance of the LRT under the random-intercept model. A wide range of simulation comparisons of various tests have been considered in Scheipl *et al.* (2008). Samuh *et al.* (2012) proposed a permutation test for approximating the distribution of the analysis-of-variance *F* statistic. The authors showed that the *F* statistic can detect the departures from the null hypothesis when the zero variance components are evidently positive. Drikvandi *et al.* (2013) proposed a permutation test using a simple test statistic that depends on directly estimating σ_u^2 under the alternative hypothesis $\sigma_u^2 > 0$. Most of those tests used simulation-based tests to assess the performance of the proposed test statistics therein. Importantly, estimation methods such as the maximum likelihood, least squares, or variance least squares (VLS) Amemiya (1977) are commonly employed in calculating the corresponding test statistics.

In this article, we focus on assessing the empirical size and power of those existing tests when u_i and e_{ij} have contaminated distributions such as the contaminated normal distribution and heavy-tail distribution. We also assess the performance of the tests in the presence of outliers in the response space. This is performed, as shown in Section 3, via simulation experiments to recommend the most reliable test under each scenario.

The rest of this paper is organized as follows. Section 2 provides a detailed description of the tests that will be covered under the simulation experiments where the distributions of the error components are contaminated. The results of our empirical investigation are presented in Section 3. An application to a real dataset involving the presence of potential outliers and other violations is considered in Section 4. Section 5 concludes the performance of the competing tests and provides recommendations for the most appropriate test according to the scenarios that have been investigated. It also provides some directions for future work.

2. Variance Components Tests

The hypothesis under consideration is given by

[2] $$H_0 \colon \sigma_u^2 = 0 \text{ versus } H_1 \colon \sigma_u^2 > 0.$$

Note that, under the null hypothesis, model [1] reduced to the traditional multiple regression model with independent observations. Thus, estimation methods such as least squares or maximum likelihood are commonly employed before the test statistics are calculated under each test. Differently, Drikvandi *et al.* (2013) employs the VLS method in calculating the test statistic used in their proposed test. Next, we present each test that will undergo our simulation experiments in some details.

2.1 *ANOVA-Type F-test*

The analysis of variance (ANOVA) famous F test statistic is defined under model [1] as follows

[3] $$F_{AOV} = \sum_{i=1}^{m} n_i \left(\bar{\hat{\zeta}}_{i.} - \bar{\hat{\zeta}}_{..}\right)^2 / \sum_{i=1}^{m} \sum_{j=1}^{n_i} \left(\hat{\zeta}_{ij} - \bar{\hat{\zeta}}_{i.}\right)^2,$$

where $\bar{\hat{\zeta}}_{i.} = n_i^{-1} \sum_{j=1}^{n_i} \hat{\zeta}_{ij}$, $\bar{\hat{\zeta}}_{..} = m^{-1} \sum_{i=1}^{m} \bar{\hat{\zeta}}_{i.}$, $\hat{\zeta}_{ij} = y_{ij} - \boldsymbol{x}_{ij}^T \widehat{\boldsymbol{\beta}}_{ols}$, and $\widehat{\boldsymbol{\beta}}_{ols}$ denotes the least squares estimate of $\boldsymbol{\beta}$. Note that the degrees of freedom in [3] are ignored since the test statistic is not approximated by the F-distribution any further. The algorithm for obtaining the empirical p-values of the test statistic in [3] are presented in the sequel. Note that under the null hypothesis the indices $j = 1, \dots, n_i$ over all groups are exchangeable and thus permutable. Thus, Samuh *et al.* (2012) proposed a permutation test for testing the hypothesis in [2] using the F_{AOV} test statistic in [3].

2.2 *A Direct Test Based on VLS Estimation of* σ_u^2

In order to test the null hypothesis in [2], Drikvandi *et al.* (2013) used the following test statistic

[4] $$T_{Drik} = m^{-1} \hat{\sigma}_u^2 \sum_{i=1}^{m} n_i,$$

where $\hat{\sigma}_u^2$ is obtained using the VLS method as follows. Rewrite model [1] in a compact form as

[5] $$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$$

where $\boldsymbol{Y}$ is a vector of $n = \sum_{i=1}^{m} n_i$ observations, $\boldsymbol{\beta}$ is a vector of p unknown fixed effects, $\boldsymbol{u}$ is a vector of m unobservable random effects, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known $n \times p$ and $n \times m$ matrices for the fixed effects and the random effects respectively, and $\boldsymbol{e}$ is a vector of n unobservable residual errors. Note that $\boldsymbol{Z} = diag(\boldsymbol{Z}_{n_1}, \dots, \boldsymbol{Z}_{n_m})$ and $\boldsymbol{Z}_{n_i} = \boldsymbol{1}_{n_i}$. Further, $E(\boldsymbol{u}) = \boldsymbol{0}$, $var(\boldsymbol{u}) = \sigma_u^2 \boldsymbol{I}_m$, $E(\boldsymbol{e}) = \boldsymbol{0}$, $var(\boldsymbol{e}) = \sigma_e^2 \boldsymbol{I}_n$, and $cov(\boldsymbol{u}, \boldsymbol{e}) = \boldsymbol{0}$.

The VLS method can be employed as follows. By first defining that $\boldsymbol{w} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and $\hat{\boldsymbol{e}}_i = \boldsymbol{Y}_i - \boldsymbol{X}_i \widehat{\boldsymbol{\beta}}_{ols}$ where $\widehat{\boldsymbol{\beta}}_{ols} = \boldsymbol{wX}^T\boldsymbol{Y}$, an unbiased VLS estimator of $\hat{\sigma}_u^2$ under model [5] can be explicitly derived from the following equation

$$\widehat{\boldsymbol{U}}_{VLS} = \frac{1}{q}\left(\{q\boldsymbol{H}^{-1} + \boldsymbol{H}^{-1}\boldsymbol{cc}^T\boldsymbol{H}^{-1}\} \sum_{i=1}^{m}(\boldsymbol{1}_i^T \hat{\boldsymbol{e}}_i \otimes \boldsymbol{1}_i^T \hat{\boldsymbol{e}}_i) - \boldsymbol{H}^{-1}\boldsymbol{c} \sum_{i=1}^{m} \hat{\boldsymbol{e}}_i{}^T \hat{\boldsymbol{e}}_i\right)$$

where

$$\boldsymbol{c} = vec\left(\sum_{i=1}^{m}\{\boldsymbol{Z}_i^T\boldsymbol{Z}_i - \boldsymbol{Z}_i^T\boldsymbol{X}_i\boldsymbol{w}\boldsymbol{X}_i^T\boldsymbol{Z}_i\}\right),\ q = \sum_{i=1}^{m} n_i - m - \boldsymbol{c}^T\boldsymbol{H}^{-1}\boldsymbol{c},\ \text{and}$$
$$\boldsymbol{H} = \sum_{i=1}^{m}(\boldsymbol{Z}_i^T\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i^T\boldsymbol{Z}_i - \boldsymbol{Z}_i^T\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i^T\boldsymbol{X}_i\boldsymbol{w}\boldsymbol{X}_i^T\boldsymbol{Z}_i - \boldsymbol{Z}_i^T\boldsymbol{X}_i\boldsymbol{w}\boldsymbol{X}_i^T\boldsymbol{Z}_i \otimes \boldsymbol{Z}_i^T\boldsymbol{Z}_i)$$
$$+\{\sum_{i=1}^{m} \boldsymbol{Z}_i^T\boldsymbol{X}_i\boldsymbol{w} \otimes \boldsymbol{Z}_i^T\boldsymbol{X}_i\boldsymbol{w}\}\{\sum_{i=1}^{m}\boldsymbol{X}_i^T\boldsymbol{Z}_i \otimes \boldsymbol{X}_i^T\boldsymbol{Z}_i\}$$

The value of $\hat{\sigma}_u^2$ is then extracted from $\widehat{\boldsymbol{U}}_{VLS} = vec(var[\boldsymbol{u}])$. For further details, see Demidenko (2004).

2.3 *Simulation-based Distributions for the LRT Statistic*

Under the compact model in [5], suppose that the loglikelihood functions that can be maximized under null and the alternative hypothesis in [2] are denoted by l_0^{ML} and l_1^{ML}. The asymptotic distribution of the LRT statistic $T_{LRT} = -2[l_0^{ML} - l_1^{ML}]$ does not have the familiar chi-square density, but can be approximated (as $m \to \infty$) as the mixture $0.5\chi^2_{(0)} + 0.5\chi^2_{(1)}$ or the mixture $0.65\chi^2_{(0)} + 0.35\chi^2_{(1)}$ (Stram and Lee, 1994; Self and Liang, 1987; Fitzmaurice et al., 2007) where $\chi^2_{(0)}$ denotes a point mass at zero and $\chi^2_{(1)}$ is a chi-square distribution with one degree of freedom.

Using the eigen-decomposition of the LRT statistic, Crainiceanu and Ruppert (2004) provide a simulation-based algorithm to obtain the finite sample distribution of the test statistic $T_{CR} = -2[l_0^{ML} - l_1^{ML}]$ under a linear mixed models framework. Zhang et al. (2016) extended the algorithm to test multiple variance components in the class of linear mixed models. Fitzmaurice et al. (2007) proposed a permutation test that provides a one-sided p-value for the LRT statistic $PLRT = -2[l_0^{ML} - l_1^{ML}]$.

3. Simulation Study

In our simulation experiments, we consider four tests whose test statistics are mentioned in Section 2. Namely, the test statistics under consideration are F_{AOV} (Samuh *et al.*, 2012), T_{Drik} (Drikvandi *et al.*, 2013), T_{CR} (Crainiceanu and Ruppert, 2004), and $PLRT$ (Fitzmaurice et al., 2007). We distinguish our investigation by evaluating the size and power of these tests when the distribution of the error components is contaminated. The choice of a reliable test in such case comes at the cost of maintaining a correct size while proving a high power since the violations are known to influence the power of those tests. Indeed, this is not surprising as the violations are expected to increase the variability of the estimates of the fixed effects and the variance components even when the null hypothesis is true.

Using the following Monte Carlo (MC) algorithm, the permutation p-values can be calculated where needed:

1) Compute the test statistic (for each of F_{AOV}, T_{Drik}, or $PLRT$) using the original sample data $(y_{ij}, \boldsymbol{x}_{ij})$ for $i = \mathbf{1}, \dots m$ and denote the test statistic by TS^{obs}.
2) Randomly permute the cluster indices holding fixed the number of observations within cluster. Then, recalculate the test statistic as in each particular test for the new permutation sample.
3) Repeat the process in the previous step B times, giving B new values of the test statistic, denoted by $TS^{(g)}$, $g = 1, \dots, B$. We choose $B=2000$.
4) Compute the empirical p-value as the proportion of permutation samples with $TS^{(g)}$ greater than or equal to TS^{obs}.
5) Given the nominal level, say α, reject the null hypothesis if α exceeds the empirical p-value.

3.1 *Simulation Setup*

In the simulations, the model for y_{ij} given the random effects u_i is

[6] $$y_{ij} = \eta + u_i + e_{ij}$$

where $j = 1, \dots., n_i$, $i = 1, \dots, m$, $m = 30, 40$ clusters, $n_i = 3, 10$ observations within a cluster and $\eta = 2$. Without loss of the generality of the simulation results, model [6] assumes that $\boldsymbol{\beta} = \mathbf{0}$ in model [5]. Let the intra-cluster correlation (ICC) take the values 0.1, 0.2, and 0.3 where $ICC = \rho = \sigma_u^2/(\sigma_u^2 + \sigma_e^2)$. The value of $\rho = 0$ is used to examine the empirical size (Type I error) of the tests under consideration.

Evaluating the empirical power of the tests (i.e. $\rho > 0$) is considered as long as we detect that the competing tests possess the correct size under the null hypothesis $H_0: \sigma_u^2 = 0$. Power considerations are investigated under the following violation schemes. We assume that the residual error term e_{ij} follows: (1) a symmetric contaminated normal distributed and (2) a skewed contaminated normal distribution. We further consider the situation where the random error components follow a normal distribution involving outliers. Note that the value of σ_u^2 is chosen such that the ICC takes its aforementioned values. The value of σ_e^2 is mentioned under each scheme separately. The detailed setup under each of these schemes of contamination is explained next.

3.1.1 *Symmetric Contaminated Normally Distribution*

A symmetric contaminated normal distribution is a mixture of two normal distributions with mixing probabilities $(1 - \delta)$ and δ where $0 < \delta < 1$. For any random variable, say ε, that follows a normal distribution with density function $f(\varepsilon;\ \mu, \sigma)$ where μ and σ denote, respectively, the mean and the standard deviation of the distribution, the contaminated normal density can be expressed as $f(\varepsilon) = (1 - \delta)f(\varepsilon;\ \mu, \sigma) + \delta\, f(\varepsilon;\ \mu, \lambda\sigma)$ where $\lambda > 1$ is a parameter that determines the standard deviation of the wider component. In the simulations, we consider $\delta = 20\%, 30\%$ as levels of contamination in the distribution of the residual errors e_{ij}, $\lambda = 5$, $\mu = 0$ and $\sigma_e^2 = 1$. The random effects u_i are assumed to follow a normal distribution with zero mean and variance σ_u^2. Table 1 summarizes the outcomes under this scheme.

3.1.2 *Skewed Contaminated Normally Distribution*

Here we investigate the performance of the tests when e_{ij} is generated from a normal distribution that is contaminated, as defined in Section 3.1.1, with a skewness parameter equal to 5. The level of contamination is set at $\delta = 20\%, 30\%$, where $\lambda = 5$, $\mu = 0$ and $\sigma_\epsilon^2 = 1$. The random effects u_i are assumed to follow a normal distribution with zero mean and variance σ_u^2. The results under this scheme are tabulated in Table 2.

3.1.3 *Outliers*

Assuming that under the null hypothesis $e_{ij} \sim N(\mu = 0, \sigma_e^2 = 0.5)$, this scheme replaces 5% of each of e_{ij} (over all observations) by random variables from $N(5, 15^2)$. This is replicated also at a 10% replacement of e_{ij} by random variables from $N(5, 15^2)$. Indeed, maximum likelihood, least squares, and the VLS estimation methods are known to be inefficient under the presence of outliers. Table 3 emphasizes this fact by displaying the Type I error rates that are achieved by each of the competing tests.

Table 1 Empirical size and power (as percentage) of tests when the residual errors are generated from symmetric contaminated normal distribution*

m	n_i	ICC	Contamination (20%)				Contamination (30%)			
			F_{AOV}	PLRT	T_{CR}	T_{Drik}	F_{AOV}	PLRT	T_{CR}	T_{Drik}
30	3	0.0	2.30	7.00	4.40	4.60	2.70	6.90	6.00	5.80
		0.1	7.20	4.00	5.20	5.60	5.10	4.20	5.00	4.80
		0.2	4.60	4.00	7.40	5.00	4.20	4.40	6.90	4.80
		0.3	4.80	5.00	9.80	9.60	5.10	4.90	8.70	7.50
	10	0.0	3.40	3.00	6.20	4.60	3.20	3.00	6.80	6.20
		0.1	4.65	4.00	8.80	7.45	4.35	4.20	7.90	6.25
		0.2	6.40	6.00	20.4	10.0	5.80	5.80	17.2	8.90
		0.3	7.85	8.00	36.0	19.8	7.15	7.80	33.0	16.4
40	3	0.0	4.00	3.50	6.20	6.00	3.80	3.50	6.70	6.50
		0.1	5.00	5.50	6.00	5.80	5.00	5.30	5.40	5.00
		0.2	6.00	5.50	7.60	6.40	5.60	5.30	7.00	6.20
		0.3	7.60	7.00	10.2	8.45	7.00	6.80	9.50	7.95
	10	0.0	6.00	5.50	4.60	3.00	6.70	5.90	5.90	3.60
		0.1	6.45	7.00	13.2	9.40	6.00	6.50	11.3	8.20
		0.2	8.20	7.50	25.2	10.8	7.50	7.10	22.8	9.20
		0.3	9.85	10.5	43.6	27.6	9.05	9.50	35.6	24.6

*Nominal Level is set to be 5%

From Table 1 and Table 2, the following notes can be taken. Under the null hypothesis of zero variance components, (i.e. when ICC=0) the four competing tests tend to possess empirical sizes that are somehow still close to the nominal level when the contamination level is 20%. However, as that level increases to 30%, the empirical sizes tend to deviate farther from the nominal level, indicating worse control of the Type-I error rate. It seems that the contamination forms in Section 3.1.1 and Section 3.1.2 have an effect on the size of these tests. We justify the high variability in the empirical sizes as the critical values are determined by the produced permutation test. The instability in the empirical shape of the distribution can be justified due to the high variability in the estimation of the unknown parameters that is fed into the calculation of the test statistic. Then, we proceed in Table 1 and Table 2 by reporting the empirical power of the tests (i.e. ICC > 0). We conclude that both the test statistics F_{AOV} and PLRT have close power values where both tend to be less powerful compared to the power produced by the statistic T_{Drik}. Interestingly, the test statistic T_{CR} remains the champion under all schemes and this suggests its reliability and resistance in producing higher power values even when the error distribution is contaminated.

Table 2 Empirical size and power (as percentage) of tests when the residual errors are generated from skewed contaminated distribution

m	n_i	ICC	Contamination (20%)				Contamination (30%)			
			F_{AOV}	PLRT	T_{CR}	T_{Drik}	F_{AOV}	PLRT	T_{CR}	T_{Drik}
30	3	0.0	3.80	5.00	7.80	8.00	3.50	4.00	7.90	8.40
		0.1	8.60	8.50	6.60	5.60	7.30	7.40	6.10	4.90
		0.2	9.60	10.5	9.80	7.20	8.90	9.95	9.20	6.70
		0.3	13.6	14.5	15.4	11.8	12.8	13.9	14.9	11.1
	10	0.0	6.00	6.50	5.40	4.20	7.20	6.80	6.40	3.80
		0.1	17.0	16.0	24.2	17.8	16.1	15.4	22.0	15.9
		0.2	22.2	24.0	30.2	28.4	21.2	23.4	28.4	26.9
		0.3	50.2	53.0	54.6	53.0	48.8	52.0	52.8	51.0
40	3	0.0	4.70	5.00	4.60	4.60	5.80	6.20	3.40	3.40
		0.1	6.85	7.00	7.80	8.20	5.55	6.20	6.30	7.60
		0.2	7.00	7.50	11.0	9.00	5.90	6.40	10.0	8.15
		0.3	12.4	12.0	16.2	14.2	11.0	11.5	14.9	12.2
	10	0.0	6.00	6.50	4.20	5.00	6.80	6.95	3.65	4.10
		0.1	8.85	13.5	16.2	14.0	7.25	11.5	15.0	12.8
		0.2	23.8	30.5	35.2	33.8	22.4	29.7	33.9	31.9
		0.3	57.6	61.0	65.2	62.0	54.6	60.0	63.8	60.8

Table 3 Empirical size (as percentage) of tests when data corrupted by outliers (ICC = 0)

	5% outliers				10% outliers			
(m, n_k)	F_{AOV}	PLRT	T_{CR}	T_{Drik}	F_{AOV}	PLRT	T_{CR}	T_{Drik}
(30, 3)	1.30	1.70	1.90	1.40	1.55	1.85	1.70	1.35
(30, 10)	0.99	1.35	1.25	0.92	1.59	1.45	1.35	1.52
(30, 20)	1.80	1.75	1.95	1.52	1.30	1.55	1.55	1.72
(40, 3)	1.10	1.67	1.52	0.86	1.20	1.77	1.35	1.95
(40, 10)	0.95	0.81	0.55	1.40	0.85	0.85	0.95	1.50
(40, 20)	2.05	1.81	2.25	2.00	2.15	1.85	2.05	2.40

From Table 3, it is obvious that all the four tests have been severely influenced by the presence of outliers due to the inefficiencies in the estimation methods that are used in the calculations of the four tests statistics. To sum up, the simulation experiments in this section have revealed the possibility of using any of the four tests to check the need for random effects (i.e. test zero variance components) while favoring the T_{CR} as most powerful. This conclusion holds as long as the response variable does not suffer from the presence of outliers in the response space. Indeed, this opens door for further the research.

4. Application to Real Data

In this section, we consider the rat pup dataset (Pinheiro and Bates, 2000) that comes from a study in which 30 female rats were randomly assigned to receive one of three doses (high, low, or control) of an experimental compound. Under random intercept modelling framework, the study was originally designed to compare the birth weights of pups from litters born to female rats that received the high and low dose treatments to the birth weights of pups from litters that received the control treatment. The data consists of 27 litters, which are randomly assigned to a specific level of treatment, and 322 rat pups are nested within these litters. The study has an unbalanced design, since the numbers of pups per litter are unequal, where the smallest litter has a size of 2 pups and the largest litter has a size of 18 pups. In addition, the numbers of litters per treatment are also unequal, such that 10 litters were assigned to the control level of treatment, 7 to the high level of treatment and 10 litters were assigned to the low level of treatment.

A summary for weights by treatment and sex is shown in Table 4. We note that the experimental treatments, high and low, appear to have a negative effect on mean birth weight. That is the sample means of the birth weights for the pups born in litters that received high and low treatments are lower than the mean of the birth weights for those born in litters that received the control dose. Besides, the sample mean birth weights of male pups are higher than those of females within all levels of treatment.

Table 4 Summary statistics for weight by treatment and sex

Treatment	Sex	Number of Observation	Mean	Standard deviation	Minimum	Maximum
High	Female	32	5.85	0.600	4.48	7.68
	Male	33	5.919	0.691	5.01	7.70
Low	Female	65	5.838	0.45	4.75	7.73
	Male	61	6.025	0.380	5.25	7.13
Control	Female	54	6.116	0.685	3.68	7.57
	Male	77	6.471	0.754	4.57	8.33

Figure 1 describes the litter effect on the rat pup birth weights using 27 box plots such that the first 10 belong to control level followed by 7 box plots that belong to a high level and the last 10 belong to the low level of treatment. It is obvious that the medians of the 27 box plots are not same where the largest medians appear in litters 8, 17 and 27 and the smallest medians are in litters 1, 11, 12 and 18. Potential outliers are also recognized in Figure 1 since some pups appear to have either lower or higher weights than the other pups that belong to the same litter. Former analyses of this dataset (e.g., Pinheiro and Bates, 2000) focused on using the conventional likelihood-based methods to infer about the effect of the different treatment levels on the birth weight. Nevertheless, these methods did not figure out the potential effect of outliers (see Figure 1) on the efficiency of the estimates and the consequent inference under the random intercept modelling framework. In the rest of this section, we highlight the gains from using the robust rank-based estimation method in terms of estimating both the fixed effects and the variance components with higher efficiency compared to the likelihood-based estimates.

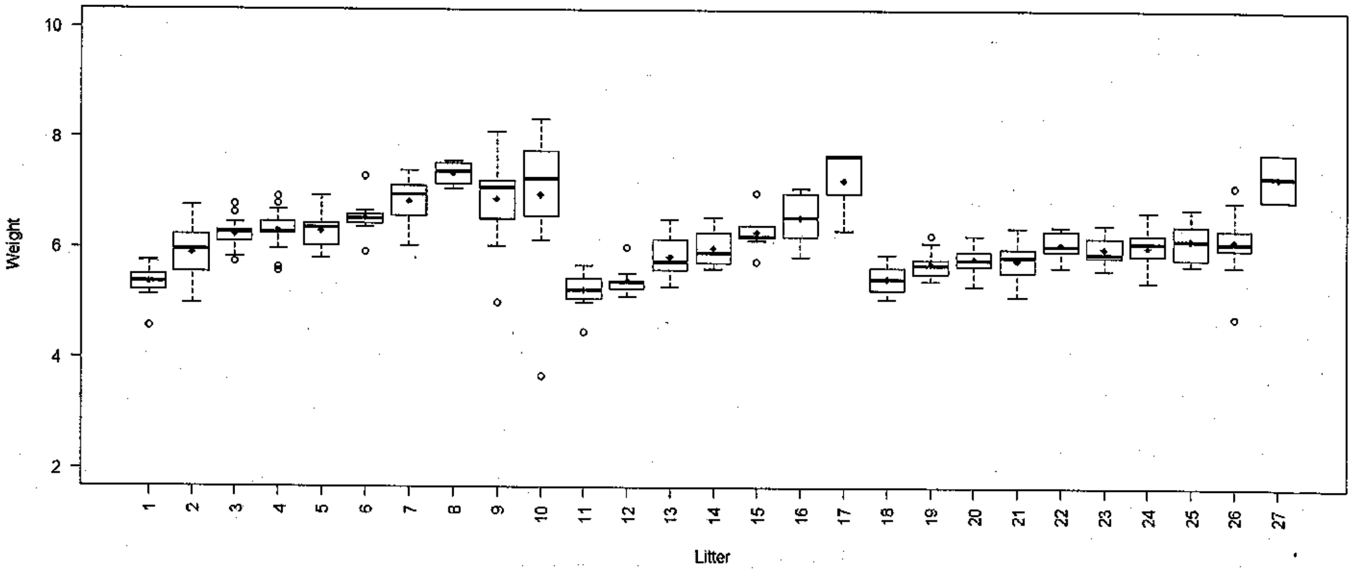

Figure 1 Box plots for rat pup birth weights by litter

Figure 1 indicates a potential varying litter effect on the distribution of the values of the rat pup birth weights in each litter. Considering this effect to be random, the individual birth weight observation ($WEIGHT_{kj}$) of the j^{th} rat pup within the k^{th} litter can be modeled using the following two-level random intercept regression model:

$$WEIGHT_{ij} = \beta_0 + \beta_1\, TREAT1_k + \beta_2\, TREAT2_k + \beta_3\, SEX_{kj} + \beta_4\, LITSIZE_k \\ + \beta_5\, TREAT1_k\, SEX_{kj} + \beta_6\, TREAT2_k\, SEX_{kj} + b_k + \epsilon_{kj};$$

$$k = 1, \ldots, 27\,, j = 1, \ldots, n_k$$

where n_k refers to the litter size that ranges between 2 and 18 pups per litter, $WEIGHT_{ij}$ is the response variable, $TREAT1_k$ and $TREAT2_k$ denote level-2 indicator variables for receiving the high and low levels of treatment, respectively, SEX_{kj} is a level-1 indicator variable for female rat pup and, $LITSIZE_k$ refers to the size of litter k, where $k = 1, \ldots, 27$. The random litter effect, b_k, is assumed to have normal distribution with mean zero and constant variance σ^{2}_{litter} and the residual error term, ϵ_{kj} is also assumed to have a normal distribution with mean zero and constant variance $\sigma^{2}_{residuals}$ (Pinheiro and Bates, 2000).

Testing the need of random effect is conducted to decide whether the random effects that associated with the intercepts for each litter can be omitted from the above model. Based on the original rat pup data set, the four test statistics were calculated where all of them yielded *p-values* that are less than the 5% nominal level which allows the random effect b_k (where $k = 1, \dots, 27$) interpretation and that the random litter effects should be retained in this model.

5. Conclusions

In this article, we conducted empirical investigations of the performance of four most commonly used tests for variance components in the literature under random-intercept models. Our comparison criteria are distinguished in the sense that we explored the size and power of the tests when the error components distribution is contaminated. Our simulation studies revealed that the simulation-based test using the LRT statistic (T_{CR}) is much powerful compared to the remaining tests. Nevertheless, the results pointed out the unfortunate poor performance, in terms of the empirical test size, of all competing tests when outliers are present in the response space. This opens the door for future research where outlier resistant test statistics can solve the latter problem.

References

Amemiya, T. (1977). A note on a heteroscedastic model. *Journal of Econometrics* 6, 365–370.

Crainiceanu, C., and Ruppert, D., (2004). Likelihood ratio tests in linear mixed models with one variance component. *Journal of Royal Statistical Society B* 66, 165-185.

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. New York: Wiley

Drikvandi, R., Verbeke, G., Khodadadi, A. and Nia, V.P. (2013). Testing multiple variance components in linear mixed-effects models. *Biostatistics* 14, 144-159.

Fitzmaurice, G.M., Lipsitz, S.R., and Ibrahim, J.G. (2007). A note on permutation tests for variance components in multilevel generalized linear mixed models. *Biometrics* 63, 942-946.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). John Wiley & Sons

Pinheiro, J. and Bates, D. (2000). *Mixed-effects models in S and S-PLUS*. Springer Science & Business Media.

Samuh, M.H., Grilli, L., Rampichini, C., Salmaso, L. and Lunardon, N. (2012). The Use of Permutation Tests for Variance Components in Linear Mixed Models. *Communications in Statistics - Theory and Methods* 41: 3020-3029.

Scheipl, F., Greven, S., and Kuchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics and Data Analysis* 52, 3283-3299.

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.

Stram, D.O., and Lee, J.W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics* 1171-1177.

Zhang, Y., Staicu, A.M., and Maity, A. (2016). Testing for additivity in non-parametric regression. *Canadian Journal of Statistics* 44(4), 445-462.