Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2008/12

Robust Random Effect Models: a diagnostic approach based on the Forward Search

Bruno Bertaccini,
Roberta Varriale

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Robust Random Effects Models: a diagnostic approach based on the Forward Search

B. Bertaccini[*,a], R. Varriale[b]

[a]Department of Statistics "G.Parenti", University of Florence, Viale Morgagni 59, 50134 Firenze, Italy

[b]Department of Methodology and Statistics, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands

---

Abstract

This paper presents a robust procedure for the detection of atypical observations and for the analysis of their effect on model inference in random effects models. Given that the observations can be outlying at different levels of the analysis, we focus on the evaluation of the effect of both first and second level outliers and, in particular, on their effect on the higher level variance which is statistically evaluated with the Likelihood-Ratio Test. A cut-off point separating the outliers from the other observations is identified through a graphical analysis of the information collected at each step of the Forward Search procedure; the Robust Forward *LRT* is the value of the classical *LRT* statistic at the cut-off point. Through Montecarlo simulations we are able to claim the clear superiority of our proposal since the probability of the type I error computed with the proposed method is much lower than the one computed with the classical approach when data are contaminated.

Key words: Random effects model, Forward Search, Outliers, Robustness

*Corresponding author.

Email address: bruno.bertaccini@ds.unifi.it (B. Bertaccini)

1. Introduction

Outliers in a dataset are observations which appear to be inconsistent with the rest of the data [5], [12], [3] and can influence the statistical analysis of such a dataset leading to invalid conclusions. Outliers can be contaminants (arising from other distributions) or can be atypical observations (extreme values) generated from the assumed model [2]. They can often be masked and should always be examined to see if they follow particular patterns, come from recording errors, or could be explained adequately by alternative models. Starting from the work of Bertaccini and Varriale [4], the purpose of this work is to implement the Forward Search method proposed by Atkinson and Riani [1] in the random effects models, in order to detect and investigate the effect of outliers on model inference. Since random effects models belong to the wider class of multilevel models we will refer interchangeably to the random effects and multilevel models notations [11].

While there is an extensive literature on the detection and treatment of single and multiple outliers for ordinary regression, these topics have been little explored in the area of multilevel models. As an example, in a one-way random effects model, Wellmann and Gather [13] distinguish three types of outliers in terms of their position relative to the main part of the observations and suggest some simple rules for their identification; in particular, the authors propose some robust procedure using median based statistics for the estimation of the model parameters and the prediction of the random effects. In the context of multilevel models, Langford and Lewis [6] propose several techniques for data exploration for outliers at first level of the analysis, such as the use of deviance reduction, measures based on residuals, leverage values and hierarchical cluster analysis and Shi and Chen [9] provide some approximate formulas for outliers detection in estimating both fixed and random parameters under the mean-shift outlier model and propose a test for multiple outliers at all levels of the analysis.

In this work, we are able to detect, at the same time, both the outliers at the first and at the second level, proposing a diagnostic method based on the Forward Search approach. The basic idea of this approach is to fit the hypothised model to an increasing subset of units, where the order of entrance of observations into the subset is based on their closeness to the previously fitted model. During the search, parameter estimates, residuals and other informative statistics are collected and these information are analysed in order to identify a cut-off point separating the outliers from the other observations. At the moment, there are no rules that allow the automatic identification of this point, so we advocate the use of a graphical approach. The robustness of the method does not derive from the choice of a particular estimator with a high breakdown point, but from the progressive inclusion of units into a subset which, in the first steps, is intended to be outlier free. Our procedure can detect the presence of more than one outlier; of course, the membership of almost all the outliers to the same group (factor level) suggests the presence of an outlying unit at the higher level of the analysis.

After a brief review of the classical linear random effects model, in Section 2 we illustrate the problems of the classical Likelihood-Ratio (LR) Test in the presence of atypical observations. In Section 3 we present the proposed forward search algorithm, describing the advantages of the proposed approach in identifying outliers through the analysis of two case studies. In Section 4 we illustrate the results of a Montecarlo study that allows the evaluation of the performance of the Robust Forward LR Test. Section 5 draws some conclusions.

2. The random effects model

The simplest random effects model is a two level linear model without covariates, also known as a random effects ANOVA. Forward search for fixed effects ANOVA has been proposed by the authors Bertaccini and Varriale [4]; in the following, we will extend this work to the random effects framework.

Let y_{ij} be the observed outcome variable of individual i ($i = 1, 2, \ldots, n_j$) within group, or factor level, j ($j = 1, 2, \ldots, J$) where J is the total number of groups and $N = \sum_{j=1}^{J} n_j$ is the total number of individuals. The simplest linear model in this framework is expressed by:

$$y_{ij} = \mu + u_j + e_{ij} = \mu + \xi_{ij} \tag{1}$$

where μ is the grand mean outcome in the population, u_j is the group effect associated with unit j and e_{ij} is the residual error at the lower level of the analysis. This model can be interpreted as a fixed or random effects model, depending on the assumptions about the nature of u_j. When u_j are interpreted as the effects attributable to a finite set of levels of a factor that occur in the data, we have a fixed effect model. On the contrary, when u_j are the effects attributable to a infinite set of levels of a factor of which only a random sample are deemed to occur in the data, we have a random effects model [7]. In this approach, each observed response y_{ij} differs from the grand mean μ by a total residual ξ_{ij} given by the sum of two random error components, u_j and e_{ij}, representing, respectively, the residual error at the higher and lower level of the analysis.

Usual assumptions for the random effects model are:

$$\begin{aligned} e_{ij} &\sim N(0,\sigma^2) \ \ \forall i,j \\ u_j &\sim N(0,\tau^2) \ \ \forall j \\ cov(e_{ij},e_{i'j'}) &= 0 \ \ \forall i \neq i' \ and \ j \neq j' \\ cov(u_j,u_{j'}) &= 0 \ \ \forall j \neq j' \\ cov(e_{ij},u_j) &= 0 \ \ \forall i,j \end{aligned}$$

and, as a consequence:

$$var(y_{ij}) = var(u_j) + var(e_{ij}) = \tau^2 + \sigma^2. \tag{2}$$

Thus, τ^2 and σ^2, expressing respectively the variability between and within groups, are components of the total variance of y_{ij}; for this reason, the model is also known as variance components or components of variance model. Furthermore, τ^2 expresses also the intra-class covariance, namely the covariance between every pair of observation in the same group expressed by:

$$cov(y_{ij},y_{i'j}) = \tau^2 \ \ \forall i \neq i'. \tag{3}$$

If model (1) holds, the expected values of the observed within-group variance, $S^2_{within} = \frac{1}{N-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij}-\bar{y}_{.j})^2$ with $\bar{y}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$, is exactly equal to the population within-group variance: $E(S^2_{within}) = \sigma^2$.

The expected value of the observed between-group variance, $S^2_{between} = \frac{1}{J-1}\sum_{j=1}^{J} n_j(\bar{y}_{.j}-\bar{y}_{..})^2$ with $\bar{y}_{..} = \frac{1}{N}\sum_{j=1}^{J} n_j\bar{y}_{.j}$, is a bit more complicated because the first-level residuals e_{ij} also contribute, although to a minor extent, to $S^2_{between}$. ANOVA estimators are obtained equating observed and expected values; they have the advantage that they can be represented by explicit formulae. For equal group sizes $n = n_j$ we have that $E(S^2_{between}) = n\tau^2 + \sigma^2$; in this case the ANOVA estimators are:

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{N-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_{.j})^2 \\ \hat{\tau}^2 &= (S^2_{between} - \hat{\sigma}^2)/n.\end{aligned}$$

For the general case formulation of the ANOVA estimators see Searle and Casella and McCulloch [7] in section 3.6. Other much used estimators are those produced by the maximum likelihood (ML) and residual maximum likelihood (REML) methods. For balanced data, ANOVA estimators are the same as the REML estimators; for unbalanced data ML and REML estimators are slightly more efficient than the ANOVA ones.

In many applications of hierarchical analysis, one common research question is whether the variability of the random effects at the group level u_j is significatively equal to 0, namely:

$$H_0 : \tau^2 = 0. \tag{4}$$

The most used procedure to test this hypothesis is the Likelihood-Ratio Test. In random effects models, when there is only one variance being set to zero in the reduced model, the asymptotic distribution of the LRT statistic is a $50:50$ mixture of a χ^2_k and χ^2_{k+1} distributions, where k is the number of the other restricted parameters in the reduced model that are unaffected by boundary conditions [8]. A rule of thumb in the applied research is to divide by two the asymptotic p-value of the Chi-squared LRT statistic distribution

[10]. Due to the presence of outliers in the data, the value of the LRT statistic can erroneously suggest to reject the null hypothesis H_0 even when there is no second level residual variability. As an example, consider the two balanced datasets represented in Fig. 1, with $n_{ij} = 10$ first level units in each group j and the total number of groups, J, equal to 25. While the bulk of the data has been generated by the same model in both cases:

$$y_{ij} = \mu + e_{ij} \tag{5}$$

with $\mu = 0$ and $e_{ij} \sim N(0, 1)$, the outliers have very different features: in the first case there are more than one first level outliers, while in the second case there is only one outlier at the second level of the analysis. In particular, the 8 outliers in the first case have been generated from a Uniform $U(10, 11)$ distribution, while in the second case, the first level units belonging to the outlier group have been generated by the $N(0 + \gamma, 1)$ distribution where γ is an observation from the $U(4, 5)$ distribution. In both cases, the LRT statistic for testing H_0 has 1 degree of freedom and its value - respectively 4.8132 and 94.4937 with halved p-value of 0.0141 and < 0.0001 falls in the rejection region due to the contamination. Obviously, in these datasets outliers are so different from the bulk of the data that they are easily identifiable by any approach; these were done only to introduce the problem more clearly.

The LR Test can often lead to erroneous conclusions also in the presence of "lighter" contamination. Let us consider some datasets with an increasing number of balanced groups ($J = 15, 20, 25, 30$) and an increasing number of observations for each group ($n_{ij} = 10, 15, 20$). While $(1 - \epsilon)N$ observations are generated by a Standard Normal distribution and are randomly assigned to all groups, the $(1 - \epsilon)N$ outliers are generated by a Normal $N(2, 1)$ dis-

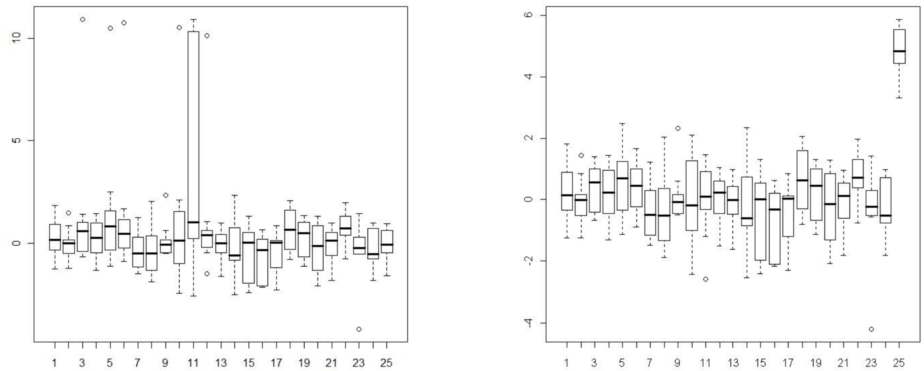

Figure 1: Boxplot showing the compositions of the described datasets.

tribution and are randomly assigned to the first half of the total number of groups. Table 1 shows the relative frequencies $rf_{(J,n_{ij},\epsilon)}$ over 10000 simulations in which the LRT statistic falls in the rejection area at the nominal significance level of $\alpha = 0.05$.

For example, for $J = 20, n_{ij} = 15$ and $\epsilon = 0.08$ the classical LRT rejects the null hypothesis (4) 1889 times giving a "real" α value of 0.1889. Obviously, the larger the ϵ is and the stronger the effect of the contamination on the LRT is.

3. Forward Search

The Forward Search is a statistical methodology proposed by Atkinson and Riani [1] useful both to detect and investigate observations that differ from the bulk of the data and to analyse their effect on the estimation of parameters and on model inference. The basic idea of this "forward" procedure is to fit the hypothised model to an increasing subset of units until all

Table 1: Classical LR Test: approximation of the true type I error probability with contaminated data

		ϵ					
J	n_{ij}	0.05	0.06	0.07	0.08	0.09	0.1
15	10	0.0578	0.0655	0.0747	0.0960	0.1153	0.1513
	15	0.0607	0.0767	0.1088	0.1385	0.1754	0.2183
	20	0.0781	0.1053	0.1436	0.1931	0.2546	0.3355
20	10	0.0662	0.0814	0.1005	0.1287	0.1680	0.2104
	15	0.0795	0.1108	0.1441	0.1889	0.2601	0.3381
	20	0.1038	0.1477	0.2067	0.2837	0.3816	0.4809
25	10	0.0678	0.0938	0.1334	0.1668	0.2077	0.2772
	15	0.0959	0.1261	0.1895	0.2620	0.3523	0.4627
	20	0.1277	0.1872	0.2657	0.3708	0.4896	0.6119
30	10	0.0781	0.1019	0.1313	0.1767	0.2265	0.2902
	15	0.0966	0.1433	0.2051	0.2704	0.3557	0.4656
	20	0.1278	0.1947	0.2815	0.3823	0.5098	0.6380

data are fitted. In particular, the entrance order of the observations into the subset is based on their closeness to the fitted model that is expressed by the residuals.

The forward search algorithm consists of three steps: the first concerns the choice of an initial subset, the second refers to the way in which the forward search progresses and the third relates to the monitoring of the statistics during the search. In this work, the methodology is adapted to the peculiarity of the random effects model taking into account the presence of groups in the data structure. In particular, focusing on the inferential issue expressed in equation (4), we propose a procedure to obtain a Robust Forward LR Test (LRT_F), individuating a cut-off point of all the classical LRT values computed during the search, cut-off point that divides the group of outliers from the other observations. Programming codes for R and S-Plus, developed by the authors, are available on demand.

3.1. Step 1: choice of the initial subset

The first step of the forward pocedure consists in the choice of an initial subset of observations supposed to be outliers free, S^*. Many robust methods were proposed to sort the data into a clean and a potentially contaminated part and the forward search is not sensitive to the method used to select the initial subset, provided unmasked outliers are not included at the start [1]. In the random effects framework, our proposal is to include in the initial subset of observations the y_{ij} which satisfy:

$$min|y_{ij} - med_j| \quad \forall j = 1, 2, \ldots, J \tag{6}$$

where med_j is the group j sample median. We impose that every group has to be represented by at least two observations; in this way, every group contributes to the estimation of the within random effects and the initial subset dimension, $m = \sum_k^J m_k$, is at least $2 * J$.

3.2. Step 2: adding observations during the search

At each step, the forward search algorithm adds to the subset the observations closer to the previously fitted model. Formally, given the subset $S^{(m)}$ of dimension $m = \sum_t^J m_t$, where the m_ts are the number of observations in group j at step m, the forward search moves to $S^{(m+1)}$ in the following way: after the random effects model is fitted to the $S^{(m)}$ subset, all the n_{ij} observations are ordered inside each group according to their squared total residuals $\hat{\xi}_{ij}^2 = (y_{ij} - \hat{y}_{ij,S^{(m)}})^2$. Since $\hat{y}_{ij,S^{(m)}} = \hat{\mu}_{S^{(m)}}$, the total residuals express the closeness of each unit to the grand mean estimate, making possible the detection of both first and second level outliers. For each group j we choose the first m_j ordered observations and add the one with the smallest squared residual among the remaining. The random effects model is now fitted to $S^{(m+1)}$ and the procedure ends when all the N observations are entered into the model. In moving from $S^{(m)}$ to $S^{(m+1)}$, while most of the time just one new unit joins the previous subset, it may also happen that two or more new units enter $S^{(m+1)}$ as one or more leave, given that all the groups have to be always represented in the subset with at least 2 observations.

The procedure allows the choice between different parameters' estimators; available estimators are ANOVA, ML and REML (default is ML).

3.3. Step 3: monitoring the search

At each stage of the search, it is possible to collect information on parameter estimates, residuals and other relevant statistics, to guide the researcher in the outliers detection.

In order to illustrate the application and the advantages of the forward search approach we show the methodology using the two datasets described in Fig. 1. In both cases, even if they represent very different situations characterised by the presence of, respectively, some first level outliers and just one second level outlier, the LRT computed with the classical approach "erroneously" falls in the rejection area of the null hypothesis expressed in equation (4). In the following, we separately describe the analysis of these two datasets carried out with the forward search approach. Figures 1 to 5 refer to the first dataset, while the others refer to the second one; all figures represent the steps of the forward search on the x-axis.

Fig. 2 shows how the observations join the subset $S^{(m)}$ during the search. The last observations joining $S^{(m)}$ belong to different second level units (right panel of Fig. 2), precisely to the groups 3, 6, 10, 11 and 12, and are represented by the bold lines that lie under the other lines at the end of the search; this suggests the possible presence of outliers in these groups.

Fig. 3 shows the N absolute total residuals $\hat{\xi}_{ij}$ computed at each step of the forward search. Throughout the search, all the residuals are very small except those related to the last 8 entered observations. These units can be considered outliers in any fitted subset and even when they are included in the algorithm in the last steps of the search their residuals decrease only slightly. Furthermore, Fig. 3 clearly highlights the sensitivity of the forward search

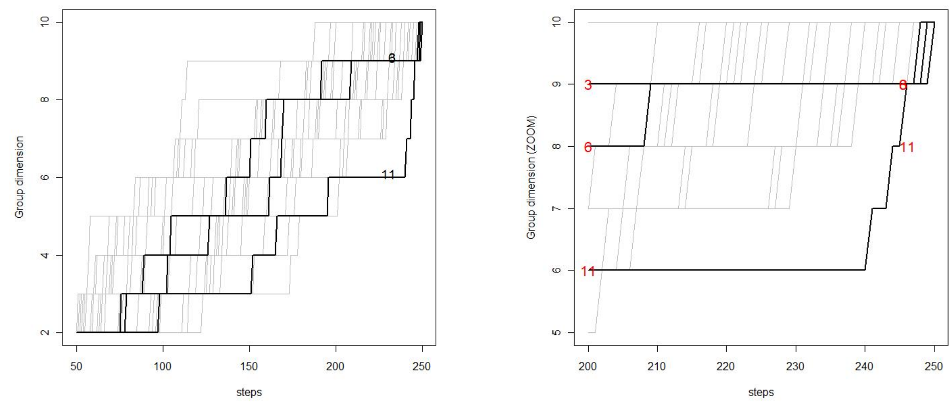

Figure 2: Forward plots of the groups dimensions: during the search (a) and zoom of the last 50 steps (b).

that also recognises the presence of an additional anomalous observation generated randomly from the Standard Normal distribution; this observation belongs to the group 23 and join the search at step 242 just before the other eight outlier observations.

Other two plots useful for detecting the presence of outliers are shown in Fig. 4, representing the values of the within ($\hat{\sigma}^2$) and between ($\hat{\tau}^2$) variance estimated during the search. In the graphs, broken lines represent, at each step of the search, the 5% and 95% quantiles of the empirical variance distribution, obtained from a Montecarlo simulation of 10000 samples free from contamination. Initially, $\hat{\tau}^2$ is quite high due to the effect of the instability in the composition of the fitting subsets $S^{(m)}$. Subsequently, both estimates become closer to their real values: the inclusion of each further unit causes a light increase in the value of the estimate of σ^2, which tends to 1 as the number of units included in the subset tends to N, while the estimated value

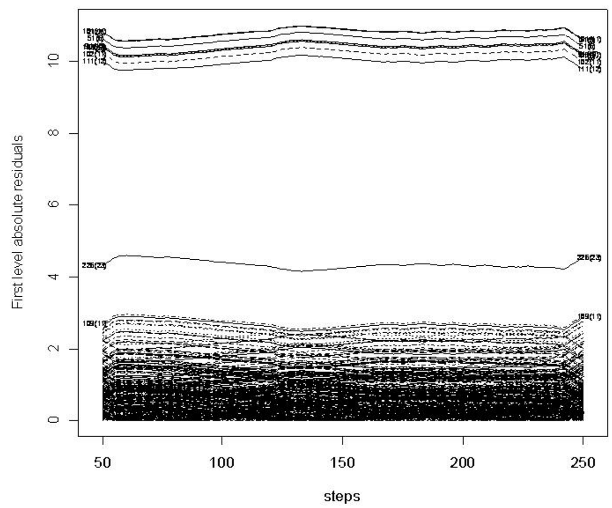

Figure 3: Forward plot of the estimated absolute residuals.

of τ^2 is always quite near to 0. At the end of the search, both estimates increase sharply due to the effect of the outliers. Finally, Fig. 5 represents the halved p-value obtained, at each step of the search, from the LRT for the null hypothesis: $H_0 : \tau^2 = 0$. During almost all the search the p-value is very high, but in the last steps it moves, erroneously, to the rejection area; at the last step, halved $p - value = 0.0141$, as indicated in Section 2.

The second example is characterized by the presence of one second level outlier. In this case, as shown in Figure 6, the observations joining $S^{(m)}$ during the last steps of the search belong to the same second level unit, 25, suggesting the presence of an anomalous group of observations.

Fig. 7 shows the total residuals computed during the search, representing, as mentioned above, the distance between the observed responses and the estimated grand mean $\hat{\mu}$. In particular, the plot is characterized by the

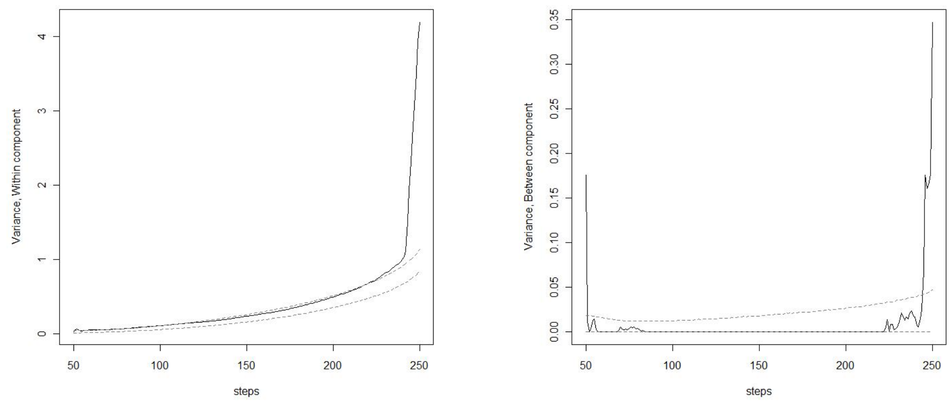

Figure 4: Forward plots of the Within (a) and Between (b) components of the error variance. The dotted lines represent the 5% and 95% quantile of the empirical distributions obtained from a simulation study with clean data.

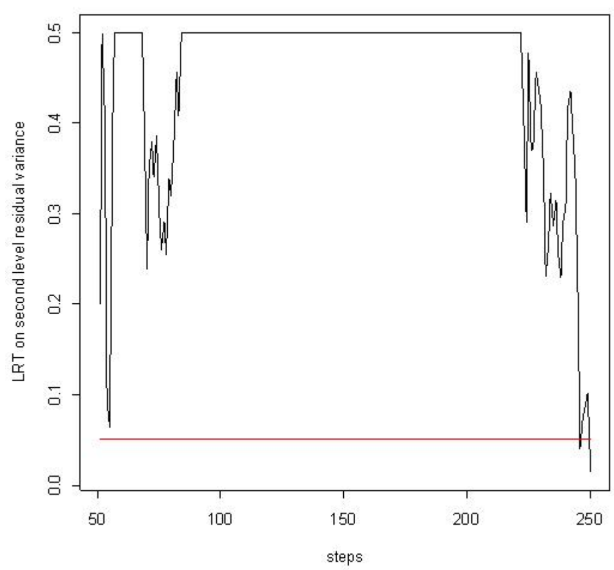

Figure 5: Forward plot of the Likelihood-Ratio Test. The horizontal line represents the chosen halved α value.

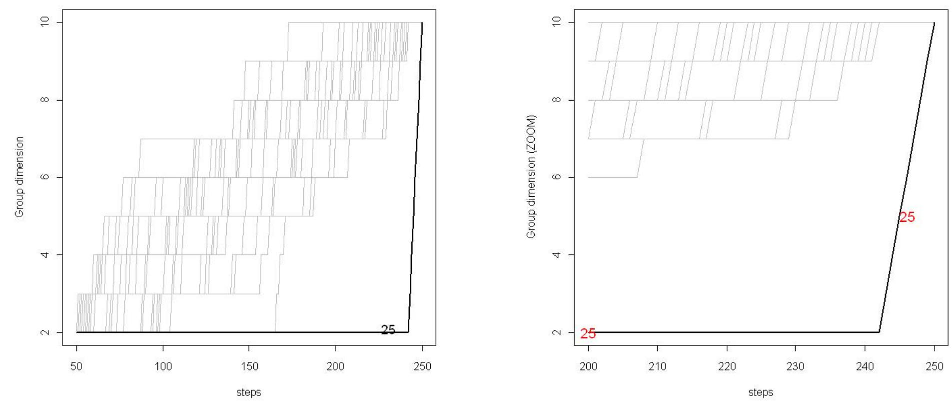

Figure 6: Forward plots of the groups dimensions: during the search (a) and zoom of the last 50 steps (b).

presence of two opposite patterns of lines: while the values of the residuals of the bulk of the data increase sharply at the beginning of the search, then they decrease smoothly and they start again to increase at the end of the search, the values of the residuals of the outlier observations follow a specular pattern (first decrease, then increase and, at the end, decrease again). This feature is due to the fact that at least two observations belonging to the outlier group are in the initial subset S^*. For this reason, the estimated grand mean is relatively high in the first steps of the search; then it starts to decrease as the number of clean observations joining $S^{(m)}$ increases and it increases again at the end of the search when all the other outliers join $S^{(m)}$.

Fig. 8 represents the values $\hat{\sigma}^2$ and $\hat{\tau}^2$ estimated during the search. Obviously, also the shape of these graphs is strongly influenced by the search rules. As said before, the two outliers of group 25 present in the initial subset will increase the grand mean value and, since the ordering of the entrance

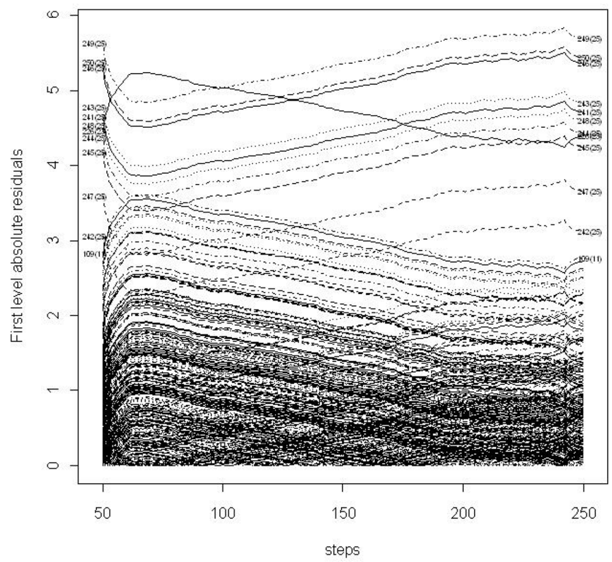

Figure 7: Forward plot of the estimated absolute residuals.

of observations in $S^{(m)}$ is decided according to their closeness to the grand mean estimate, the first few observations entering the search are those positioned in the highest part of each group and far away from their group mean; therefore, the whithin variance is overestimated and falls outside the 95% confidence interval computed with clean data. After all the observation of the non outling groups are included in $S^{(m)}$, the Within variance estimate decreases slightly entering the bands.

The behaviour of $\hat{\tau}^2$ is less regular: indeed, after being very high for the first half of the search, it moves toward its real value as more non outlying observations join $S^{(m)}$ contributing to move $\hat{\mu}$ toward its real value. Anyway, at the end of the search, it obviously increases due to the effect of the outliers.

Finally, Fig. 9 represents a very interesting behaviour of the the halved

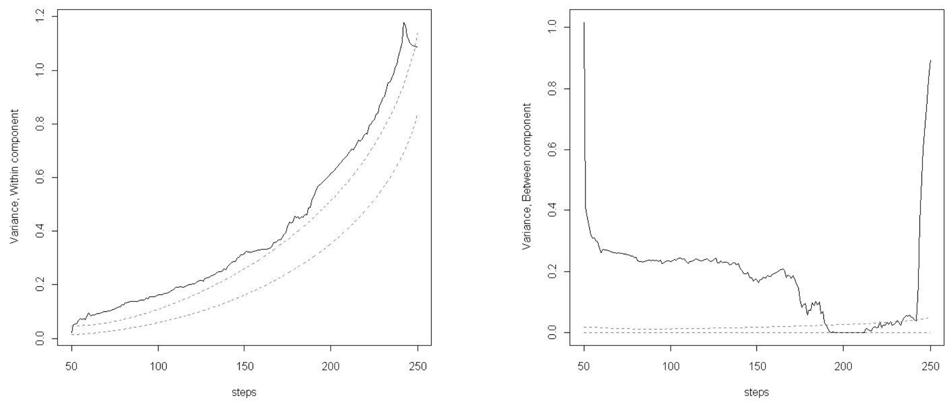

Figure 8: Forward plots of the Within (a) and Between (b) components of the error variance. The dotted lines represent the 5% and 95% quantile of the empirical distributions obtained from a simulation study with clean data.

p-value obtained with the LRT. Contrary to the first example, during the search the p-value is always very low since the units belonging to the outlier group that are in $S^{(m)}$ lead to the wrong conclusion of the presence of second level variability. Then, the LRT correctly increases as the number of non outlying units entering the subset $S^{(m)}$ increases, and it obviously sharply decreases when the units of the outlier group finally enter the search.

4. Forward Likelihood-Ratio Test and its evaluation

The Forward Likelihood-Ratio Test can be defined as a collection of the values of the classical LR Test statistic computed at each step of the search (see Figs. 5 and 9); to obtain a Robust Forward LR Test we need to identify a cut-off point of the progress procedure that best divides the group of observations that differ to the bulk of the data from the others. As showed, the

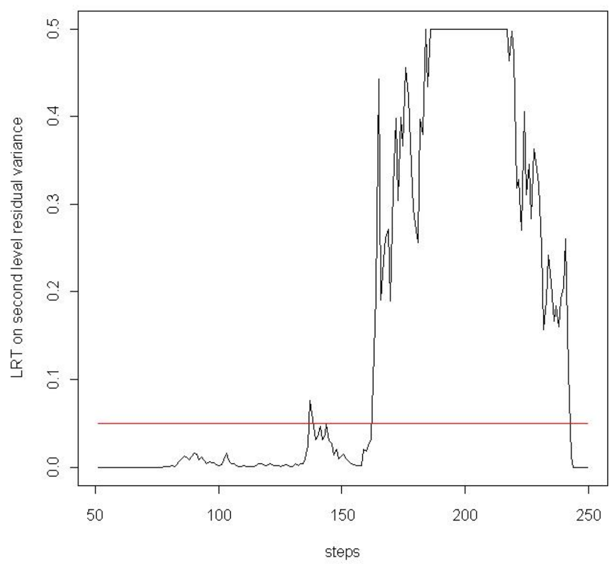

Figure 9: Forward plot of the Likelihood-Ratio test. The horizontal line represents the chosen halved α value.

search of the cut-off point cannot be "automatic" but is completely based on graphical analysis and ad hoc considerations, and is strictly connected to the context of the observed phenomenon. Since is clearly impossible to carry out a simulation to evaluate the goodness of the test without an "automatic" cut-off point, we refer to the same simulations illustrated in the Section 2 to show how the proposed robust test behaves. In particular, because of the complexity of the model and the computational burden of the simulations, we evaluate only some cases presented in Table 1.

The datasets used for the analysis are composed by J balanced groups of n_{ij} observations coming from a Standard Normal distribution, with $J = 20, 25, 30$ and $n_{ij} = 10, 15, 20$. The cut-off point is located at $\epsilon \cdot N$ steps before the end of the search, with $\epsilon = 0.05, 0.08, 0.10$. Table 2 shows frequencies over 10000 simulations in which Robust Forward LR Test falls in the rejection area at the nominal significance level of $\alpha = 0.05$; these values can be compared with the results obtained with the classical approach (Table 1). For example, for $J = 20, n_{ij} = 15$ and $\epsilon = 0.08$ the Robust Forward LR Test produces an evidence versus the alternative hypothesis in 519 samples while with the classical test the frequency is of 1889 over 10000 replicates. With the proposed method, the probability of accepting H_1 when H_0 is true is always lower than the same probability showed in Table 1. However, since we did not analyse every sample with a graphical approach as the forward search suggests and we choice the cut-off point with an automatic procedure, there are small variations from the nominal significance level $\alpha = 0.05$.

Table 2: Approximation of the true type I error probability of the Robust Forward LR Test with contamination in the data

J	n_{ij}	0.05	0.08	0.1
20	10	0.0394	0.0451	0.0535
20	15	0.0363	0.0519	0.0635
20	20	0.0426	0.0637	0.0757
25	10	0.0409	0.0503	0.0645
25	15	0.0411	0.0636	0.0774
25	20	0.0445	0.0694	0.0910
30	10	0.0405	0.0528	0.0625
30	15	0.0454	0.0619	0.0744
30	20	0.0491	0.0681	0.0935

5. Concluding remarks

In a dataset, outliers can affect the inferetial results on a statistical model and can lead to erroneous conclusions. For this reason, researchers should always be aware of the presence of outliers in the analysed dataset. Our work concerns the effect of outliers in the random effects modeling framework. In particular, we implemented the Forward Search method for random effects models, in order to individuate the outliers and to analyse their effect on the estimation of parameters and on model inference. The proposed methodology takes into account the presence of groups in the data structure and can identify the presence of both first and second level outliers. The basic idea of the Forward Search approach is to fit the hypothesised model to an increasing subset of units, where the order of entrance of observations into the subset is based on their closeness to the fitted model. At every step of the forward search we compute parameters estimates, residuals and some other informative statistics in order to identify, with a graphical approach, a cut-off point separating the outliers from the other observations. In particular, the value of the classical Likelihood-Ratio Test evaluated at the cut-off point is the Robust Forward Likelihood-Ratio Test value. The results of a Montecarlo simulation study show the clear superiority of our proposal since the probability of the type I error computed with the proposed method is much lower than the one computed with the classical approach when data are contaminated.

References

[1] Atkinson, A.C., Riani, M., 2000. *Robust Diagnostic Regression Analysis*, Springer, NewYork.

[2] Barnett, V., 1988. Outlier and order statistics. *Commun. Statist. Theory Methods* **17**, 2109–2118.

[3] Barnett, V., Lewis, T., 1993. *Outliers in Statistical Data*, third ed. Wiley, NewYork.

[4] Bertaccini, B., Varriale, R., 2007. Robust ANalysis Of VAriance: An approach based on the Forward Search. *Computational Statistics & Data Analysis* **51**, 5172–5183.

[5] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A., 1986. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, NewYork.

[6] Langford, I.H., Lewis, T., 1998. Outliers in multilevel data. *Journal of the Royal Statistical Society A* **161**, 121–160.

[7] Searle, S.R., Casella, G., McCulloch, C.E., 1992. *Variance components*, John Wiley & Sons, Inc., New York etc.

[8] Self, S.G., Liang, K.Y., 1987. Asymptotic properties of maximum likelihood estimators and like lihood ratio tests under nonstandard conditions. *JASA* **82**, 605–610.

[9] Shi, L., Chen, G., 2008. Detection of outliers in multilevel models. *Journal of statistical planning and inference* **138**, 3189–3199.

[10] Skrondal, A., Rabe-Hesketh, S., 2004. *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equation Models*, Chapman & Hall/CRC, Boca Raton, FL.

[11] Snijders, T.A.B., Bosker, R.J., 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publishers, London etc.

[12] Staudte, R.G., Sheather, S.J., 1990. *Robust Estimation and Testing.* Wiley, NewYork.

[13] Wellmann, J., Gather, U., 1990. Identification of outliers in a one-way random effects model *Statistical papers*. Springer, Berlin/Heidelberg.