Identification of Outliers in a One-Way Random Effects Model

Jürgen Wellmann[⋆,†] and Ursula Gather[†]

[⋆] University of Münster, Institute of Epidemiology and Social Medicine, D-48129 Münster, Germany
[†] Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

Keywords mixed linear model; variance components; random effects; robust statistics; median; median absolute deviation

Abstract

We distinguish between three types of outliers in a one-way random effects model. These are formally described in terms of their position relative to the main part of the observations. We propose simple rules for identifying such outliers and give an example which involves median-based statistics.

1 Introduction

A one-way random effects model assumes for continuous random variables Y_{ij} that

$$Y_{ij} = \mu + U_i + E_{ij}, \qquad i = 1, \ldots, \ell, j = 1, \ldots, n_i. \tag{1}$$

In applications Y_{ij} may represent the jth measurement taken in the ith laboratory taking part in an interlaboratory testing procedure to investigate the quantity of a certain ingredient in some given substance. The measurements deviate from the fixed (unknown) quantity μ by $U_i + E_{ij}$, where U_i is a normally distributed random effect ('laboratory effect') with mean 0 and variance $\sigma_U^2 \geq 0$, i.e. $U_i \sim N(0, \sigma_U^2), i = 1, \ldots, \ell$. The variables E_{ij} are $N(0, \sigma_E^2)$ distributed random variables with $\sigma_E^2 > 0$. They represent the individual measurement errors, $i = 1, \ldots, \ell, j = 1, \ldots, n_i$.

The parameters σ_U^2 and σ_E^2 are called variance components. Note that model (1) implies that the random vectors $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ are independent and follow a multivariate normal distribution with mean $\mu \mathbf{1}_{n_i}$ and covariance matrix $\sigma_U^2 \mathbf{J}_{n_i} + \sigma_E^2 \mathbf{I}_{n_i}$ (cf. Searle (1987)), where $\mathbf{1}_{n_i}$ denotes a vector of ones of length n_i, $\mathbf{I}_{n_i}$ denotes the identity matrix and $\mathbf{J}_{n_i}$ the matrix of ones of order $(n_i \times n_i), i = 1, \ldots, \ell$.

Model (1) is invariant under linear transformations

$$y \mapsto ay + b, \qquad a \neq 0. \tag{2}$$

That is to say, a model of the form (1) is still valid when all data are transformed as in (2).

Figure 1 gives some results from an intercomparison of radon detectors, described in Kreienbrock et al. (1999). The scatterplot shows 25 measurements of α-energy, emitted by radioactive radon gas, which were taken under identical conditions. Each detector supplies one measurement after preparation in a laboratory. Five laboratories took part in this investigation, each with five detectors.

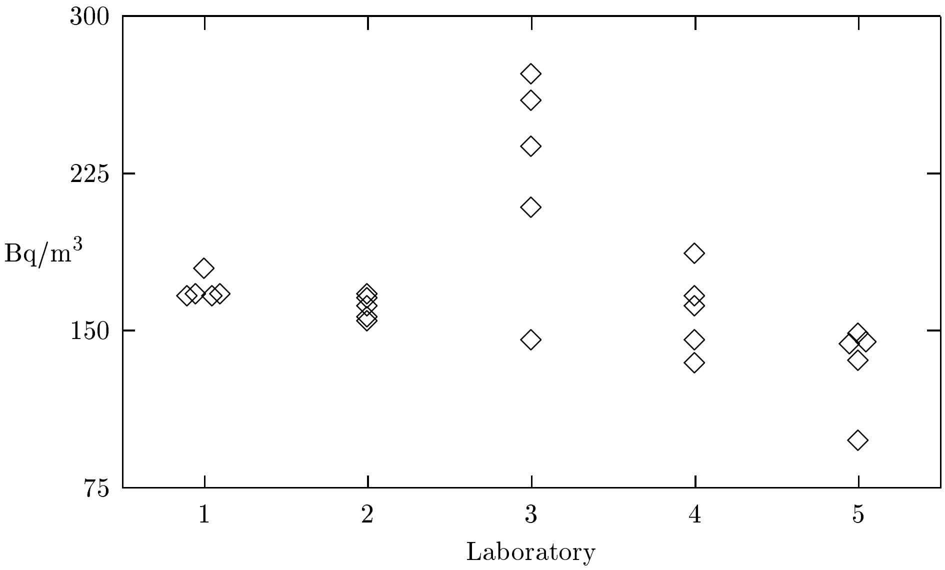

Figure 1: Radon measurements from an interlaboratory test

One would expect that the measurements lie close together because the same quantity was measured and the laboratories used the same standardized analytical technique. Furthermore the variation within each laboratory should be the same. Therefore model (1) should be appropriate for these data.

But in fact one can observe some types of 'outliers'. Barnett and Lewis (1994) "... define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". Three types of outliers can be distinguished in random effects models. These are seen in figure 1 for the following data.

1. There is one observation in each of the laboratories 3 and 5 which is remarkably small, compared to the other observations of the same laboratory.

2. Except for the lower outlier laboratory 3 generally supplies larger measurements.

3. Laboratories 1 and 3 differ from the others with respect to the variation of the data. There is very little variation in laboratory 1, whereas the data in laboratory 3 show higher variation, even when the lower outlier is neglected.

If model (1) is satisfied, these outliers are not likely to occur because of the light tails of the normal distribution and the assumption of homoscedasticity. Therefore model (1) is considered to describe the ideal situation without outliers.

Our aim is to set up formal rules that identify these outliers. At first we will give the term outlier a more precise meaning. We will then consider one example of robust estimators and predictors, based on medians. Robust statistics for the one-way random effects model are extensively discussed by e.g. Stahel and Welsh (1992), Wellmann (1994), and Wellmann (2000). They are of interest in their own. However, they are only used here to construct rules for the identification of outliers. We will then suggest a general form for such rules and provide details for a specific example which involves the median-based estimators discussed before. This method is illustrated using the data from the introductory example (fig. 1).

2 Outliers in a one-way random effects model

2.1 Outlier regions

For univariate data Davies and Gather (1993) have defined so called outlier regions. These are tail regions of the target distribution. For a normal $N(\mu, \sigma^2)$ distribution with mean μ and variance $\sigma^2 > 0$, the *δ-outlier region* with respect to $N(\mu, \sigma^2)$ is

$$\mathrm{out}_L(\delta, \mu, \sigma) = \left\{x : |x - \mu| > z_{1-\delta/2}\sigma\right\}, \tag{3}$$

where $\delta \in (0, 1)$ is some given number and z_q is the q-quantile of the normal distribution. The outlier region is chosen to be symmetric about μ because of the symmetry of the normal distribution. We note that a random variable X from $N(\mu, \sigma^2)$ will be located in $\mathrm{out}_L(\delta, \mu, \sigma)$ with probability δ,

$$\Pr\left(X \in \mathrm{out}_L(\delta, \mu, \sigma)\right) = \delta. \tag{4}$$

A real number x is called *δ-outlier with respect to $N(\mu, \sigma^2)$* if $x \in \mathrm{out}_L(\delta, \mu, \sigma)$ (Davies, Gather (1993)). Here we will consider three types of outlier regions in order to describe the above mentioned types of outliers in a one-way random effects model.

The region $\mathrm{out}_L(\delta, 0, \sigma_E)$ corresponds to outliers in the E's, i.e. outliers within the classes. We call a real number y a *location-δ-outlier within the ith class* if it is an observation of Y_{ij} and the corresponding unobservable random variable E_{ij} is realized in $\mathrm{out}_L(\delta, 0, \sigma_E)$. More conveniently we may call any real number y a location-δ-outlier within the ith class if it belongs to

$$\mathrm{OUT}(\delta, \mu, \sigma_E, U_i) = \left\{y : |y - \mu - U_i| > z_{1-\delta/2}\sigma_E\right\} \qquad i = 1, \ldots, \ell. \tag{5}$$

This outlier region depends on a random effect and is therefore a random set. But it could as well have been formulated with the unobservable realizations of the random effects.

Globally larger or smaller observations in some class correspond to an outlier in the U's, which is described by $\mathrm{out}_L(\delta, 0, \sigma_U)$. When a random effect $U_i, i = 1, \ldots \ell$, is observed in $\mathrm{out}_L(\delta, 0, \sigma_U)$ we call the corresponding class a *location-δ-outlier within the random effects.*

Extremely large or small variation within one class may be reflected in corresponding values of an estimator of scale s. The statistic s is called a scale estimator if it is location invariant and scale equivariant. That is to say, if $\boldsymbol{y}$ is the vector of observations, then

$$s(a\boldsymbol{y} + b\mathbf{1}) = |a| s(\boldsymbol{y}) \geq 0 \tag{6}$$

for any scalar constants $a \neq 0$ and b, cf. Lax (1985). We define an outlier region corresponding to an estimator of scale s to be a set of vectors $\boldsymbol{y}$ which lead to $s(\boldsymbol{y})$ sufficiently far away from σ_E and choose this region to be symmetric about $\ln(\sigma_E)$ on a logarithmic scale as

$$\mathrm{out}_S(\delta, \sigma_E; s) = \{\boldsymbol{y} \in \mathbb{R}^m : |\ln(s(\boldsymbol{y})) - \ln(\sigma_E)| > c_\delta\}. \tag{7}$$

The constant c_δ is chosen to satisfy

$$\Pr\left(\boldsymbol{Y} \in \mathrm{out}_S(\delta, \sigma_E; s)\right) = \delta \tag{8}$$

for a multivariate normal random vector $\boldsymbol{Y}$ with a covariance matrix of the form $\sigma_U^2 \mathbf{J}_m + \sigma_E^2 \mathbf{I}_m$ as in model (1). We will apply this concept to observations of the vectors $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, $i = 1, \ldots, \ell$, but we call any vector $\boldsymbol{y}$ in $\mathrm{out}_S(\delta, \sigma_E; s)$ a *scale-δ-outlier with respect to s*.

The scale estimator s should be resistant against outliers in order to reduce confusion between high variation and low variation plus single outliers within the class.

The outlier region out_S is invariant under the transformation (2) in the sense that a transformed vector of observations $a\boldsymbol{y}_i + b\mathbf{1}_{n_i}$ is a scale-outlier if and only if $\boldsymbol{y}_i$ is a scale-outlier in the original dataset. Analogous results hold true for the other outlier regions, which involve the unobservable random variables U_i and E_{ij}, where we adopt the following interpretation of invariance. Motivated by equation (1) we think of the transformed observable random variables $aY_{ij} + b$ as the sum of the new 'true value' $a\mu + b$, class effects aU_i and measurement errors aE_{ij}, $i = 1, \ldots, \ell, j = 1, \ldots, n_i$. This convention is consistent with the assumption that the unobservable random variables have zero means.

Two further types of outliers could be considered. The outlier region

$$\mathrm{out}_\mu = \left\{ y : |y - \mu| > z_{1-\delta/2} \sqrt{\sigma_U^2 + \sigma_E^2} \right\}$$

corresponds to observations far away from μ. The region

$$\mathrm{out}_{L+S} = \left\{ \boldsymbol{y} \in \mathbb{R}^{n_i} : \frac{\sum_{j=1}^{n_i} (y_j - \bar{y}_\bullet)^2}{\sigma_E^2} + \frac{n_i (\bar{y}_\bullet - \mu)^2}{\sigma_E^2 + n_i \sigma_U^2} < \xi_{1-\delta} \right\},$$

where $\xi_{1-\delta}$ denotes the $(1-\delta)$-quantile of the central χ^2-distribution with n_i degrees of freedom and $\bar{y}_\bullet = \sum_{j=1}^{n_i} y_i/m$, considers observations of $\boldsymbol{Y}_i, i = 1, \ldots, \ell$, as outliers which lie outside the smallest ellipsoid with probability mass $1-\delta$ under the ideal model (1). Thus location and spread of the classes are considered simultaneously.

Further aspects of outliers and robustness in the one-way random effects model are discussed in Davies (1991). We do not discuss these approaches any further.

2.2 Model assumptions and outliers

The element of surprise which the outliers provoke depends on what one expects to observe, or in other words, on the ideal model that one assumes before the data are available. A model similar to our ideal model (1), which could also be appropriate for the data in our introductory example, is the fixed effects model

$$\begin{aligned} Y_{ij} = \mu + \theta_i + E_{ij}, \qquad & i = 1, \ldots, \ell, j = 1, \ldots, n_i, \\ & \sum_{i=1}^{\ell} \theta_i = 0, E_{ij} \sim N(0, \sigma_i^2), \sigma_i > 0. \end{aligned} \tag{9}$$

The special case of this model with $\sigma_1 = \cdots = \sigma_\ell = \sigma_E$ is even more similar to (1).

Both models assume a normal distribution for the data. The normal distribution implies that the data are crowded together because of the light tails of this distribution. A Cauchy distribution, for example, will generate aberrant values much more easily.

Furthermore, outliers in the $E's$ can be considered in the fixed effects model as well as in the random effects model.

But the fixed effects model per se gives no reason to identify location-outliers in the class effects, since these are arbitrary parameters in this model. The random effects model, on the other hand, states that the class effects stem from a common source and therefore should not differ too much.

However, the fixed-effects model allows a test for the hypothesis $H_0 : \theta_1 = \cdots = \theta_\ell = 0$ or a multiple testing procedure to compare individual class effects. But this hypothesis seems to be more restrictive (though not directly comparable) than the assumption of model (1) about the class effects, where some variation is allowed. A less restrictive hypothesis on the θs in model (9) could be formulated, but then one has to decide how much variation in the class effects should be allowed. This is also true for the random effects model, but there this decision is assisted by the assumption of the distribution of the random class effects.

Searching for unusual variation of the data within the classes can be based on a heteroscedastic model like (9), or the analogous model with random class effects, by a test for the hypotheses $\sigma_1 = \cdots = \sigma_\ell$. Again, less restrictive hypotheses can be formulated for the scale parameters or one can take the viewpoint of outlier identification as described above and search for unusual realizations of the random variables $s(\boldsymbol{Y}_i), i = 1, \ldots, \ell$, under a homoscedastic model. Note however that our approach may be of limited use in a situation where a common scale parameter for all classes is not reasonable.

In more structured models further aspects of outliers may occur; see Terbeck and Davies (1999) for a discussion of the two-way analysis of variance.

2.3 Standardization of outlier regions

Following the ideas of Davies and Gather (1993) we use outlier regions with $\delta = \alpha_n$ depending on a prespecified $\alpha \in (0,1)$ and $n = \sum_{i=1}^{\ell} n_i$ such that

$$\Pr\left(\exists\, i,j : Y_{ij} \in \mathrm{OUT}(\alpha_n, \mu, \sigma_E, U_i)\right) = \alpha \tag{10}$$

under model (1). This can be achieved by taking $\alpha_n = 1 - (1-\alpha)^{1/n}$. We use in the same way $\alpha_\ell = 1 - (1-\alpha)^{1/\ell}$ in order to get

$$\Pr\left(\exists\, i : U_i \in \mathrm{out}_L(\alpha_\ell, 0, \sigma_U)\right) = \alpha, \tag{11}$$

or

$$\Pr\left(\exists\, i : s(\boldsymbol{Y}_i) \in \mathrm{out}_S(\alpha_\ell, \sigma_E; s)\right) = \alpha, \tag{12}$$

respectively, under this model. Thus the identification of outliers in the one-way random effects model, as it is considered in this paper, aims at the identification of

1. location-α_n-outliers within the classes, that means observations y_{ij} in $\mathrm{OUT}(\alpha_n, \mu, \sigma_E, U_i)$,

2. location-α_ℓ-outliers within the random effects, i.e. observations of the U_i in $\mathrm{out}_L(\alpha_\ell, 0, \sigma_U)$,

3. scale-α_ℓ-outliers, i.e. observations of the vectors $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$ which lie in $\mathrm{out}_S(\alpha_\ell, \sigma_E; s)$, $\quad i = 1, \ldots, \ell$, for a robust scale estimator s. Here the median absolute deviation (MAD) is used (see below).

3 Median-based estimators and predictors

The above mentioned outlier regions depend on the unknown parameters μ, σ_U^2, and σ_E^2 as well as on the unobservable random effects $U_1, \ldots, U_\ell$. In order to identify outliers we need estimates of these parameters and predictors of the random effects. Experience with univariate data shows that robust procedures are preferable and especially median based statistics are a good choice to avoid a masking effect in outlier detection rules (Davies, Gather (1993)).

Given data y_{ij}, let

$$\text{med}_i = \text{median}\left(y_{i1}, \ldots, y_{in_i}\right) = \frac{1}{2}\left(y_{i,([(n_i+1)/2])} + y_{i,([n_i/2]+1)}\right)$$

be the median of the observations in the ith class, where $y_{i,(1)} \leq \ldots \leq y_{i,(n_i)}$ are the ordered observations $y_{i1}, \ldots, y_{in_i}$ in the ith class, and $[\bullet]$ denotes rounding off to the nearest integer. Let

$$\text{mad}_i = \text{median}\left(|y_{i1} - \text{med}_i|, \ldots, |y_{in_i} - \text{med}_i|\right)$$

denote the median absolute deviation within class i. The normalized median absolute deviation

$$s_i = e(n_i) \cdot \text{mad}_i, \qquad i = 1, \ldots, \ell,$$

is used to describe the variation within the classes. The factor $e(m)$ has been found by Croux and Rousseeuw (1992) and ensures that the median absolute deviation of m stochastically independent normally distributed random variables is an approximately unbiased estimator for the underlying standard deviation. This normalizing factor equals $e(m) = 1.4826\ b(m)$. The factor 1.4826 is the reciprocal of the MAD of the $N(0,1)$-distribution and is multiplied by $b(m)$, where

$m =$	2	3	4	5	6	7	8	9
$b(m) =$	1.196	1.495	1.363	1.206	1.200	1.140	1.129	1.107

and $b(m) = m/(m - 0.8)$ for $m > 9$. We use

$$\hat{\mu} = \text{median}\left(\text{med}_1, \ldots, \text{med}_\ell\right), \tag{13}$$

$$\widehat{\sigma_U^2} = e_U \text{median}\left((\text{med}_1 - \hat{\mu})^2, \ldots, (\text{med}_\ell - \hat{\mu})^2\right) \tag{14}$$

$$\widehat{\sigma_E^2} = e_E \text{median}\left(s_1^2, \ldots, s_\ell^2\right) \tag{15}$$

as estimators of μ, σ_U^2, and σ_E^2 respectively. The factor e_E in (15) is given by

$$e_E = 0.9797 + 1.1188\,\frac{\ell - 3.5592}{n}$$

and achieves that the estimator $\widehat{\sigma_E^2}$ becomes approximately unbiased under model (1).

The estimator $\widehat{\sigma_U^2}$ is constructed following the form of the estimator of Hartung (1981) for σ_U^2 in the balanced case, i.e. when $n_1 = \cdots = n_\ell = m$, say. This estimator is proportional to the sum of squares of the class averages. It is always non-negative, but biased, it's mean equals $\sigma_U^2 + \sigma_E^2/m$. This is equal to the variance of the average of the observations from one class. Our estimator is proportional to the MAD of the class medians. The variance of the ith class-median med_i is

$$\mathrm{Var}(U_i + \mathrm{median}(E_{i1}, \ldots, E_{in_\ell})) = \sigma_U^2 + v(n_i)\sigma_E^2, \quad i = 1, \ldots, \ell, \tag{16}$$

where $v(m)$ denotes the variance of the median of m independent $N(0,\ 1)$ random variables. Cadwell (1952) gives approximations for $v(m)$. By means of simulations we found the factor

$$e_U = \frac{\ell}{\ell + 1.56}\, e(\ell)^2$$

to achieve that the mean of (14) is approximately equal to this variance in the balanced case, i.e. $\sigma_U^2 + v(m)\sigma_E^2$.

The above reasoning suggests that

$$\widetilde{\sigma_U^2} = \widehat{\sigma_U^2} - v(m)\widehat{\sigma_E^2} \tag{17}$$

is unbiased for σ_U^2. Note that $\widetilde{\sigma_U^2}$ can become negative. A simple remedy of this defect is to replace negative values of this estimator by zero,

$$\overline{\sigma_U^2} = \max\left\{\widetilde{\sigma_U^2}, 0\right\}, \tag{18}$$

but this will again introduce a bias, cf. Verdooren (1980). Problems of this kind are well known in classical estimation theory of variance components, see e.g. LaMotte (1973).

The factors e_U and e_E approximate the reciprocals of simulated means of the uncorrected estimators under model (1). They have been simulated in the balanced case for several values of ℓ and m. We use them in the unbalanced case as well.

The estimators defined above don't take account of the number of observations per class. Therefore it should be possible to improve these estimators, especially in the unbalanced situation, by considering appropriate weights for each class. We do not follow up this issue further.

Predictors for the random effects U_i are constructed similar to the non-robust 'best linear unbiased predictors', cf. Searle (1987),

$$\widehat{u_i} = \frac{n_i \widehat{\sigma_U^2}}{\widehat{\sigma_E^2} + n_i \widehat{\sigma_U^2}} \left(\mathrm{med}_i - \hat{\mu}\right), \quad i = 1, \ldots, \ell. \tag{19}$$

4 Identification of outliers

Let $\hat{\mu}, \widehat{\sigma_U^2}$ and $\widehat{\sigma_E^2}$ be estimators of μ, σ_U^2 and σ_E^2 respectively and let $\widehat{u_i}$ be a predictor of the unobservable random effect $U_i, i = 1, \ldots, \ell$. We also need robust estimates of scale $s_i = s(\boldsymbol{y}_i), i = 1, \ldots, \ell$. We restrict our attention to nonnegative estimators of the variance component σ_U^2. This excludes for example the estimator (17), but admits its truncated version (18). Common estimators for σ_E^2 are nonnegative with probability one. Note however that the MAD becomes zero when half of the data are identical.

The identification rules proposed below need estimators $\widehat{\sigma_U}$ and $\widehat{\sigma_E}$ for σ_U and σ_E. When only estimators for the variance components are available, we simply take their square roots.

There are numerous suggestions in the literature on how to identify outliers in univariate data (Barnett and Lewis (1994), Hawkins (1980)). Appropriate modifications of these procedures can be applied to the $\widehat{u_i}$ to find location-outliers within the random effects, to the s_i to identify scale-outliers (especially procedures for non-negative data) or to the $(y_{ij} - \widehat{u_i})$ to identify location-outliers within the classes, $i = 1, \ldots, \ell, j = 1, \ldots, n_i$.

Our definition of the task of identifying outliers in one-way random effects models reads: Find all points in the outlier regions defined in section 2.1. Therefore we define empirical versions of these regions. These are given by procedures which identify outliers when appropriately defined residuals exceed a critical value. These values may depend on estimators $\widehat{\gamma}$ of the unknown ratio $\gamma = \sigma_U / \sigma_E$.

1. Identify y_{ij} to be a location-α_n-outlier within the ith class, if

$$|y_{ij} - \hat{\mu} - \widehat{u_i}| > c_E(\alpha, \widehat{\gamma}) \widehat{\sigma_E}. \tag{20}$$

2. The ith random effect is identified as location-α_ℓ-outlier within the random effects, if

$$\widehat{\sigma_U} > 0 \quad \text{and} \quad |\widehat{u}_i| > c_U(\alpha, \widehat{\gamma})\widehat{\sigma_U}. \tag{21}$$

When $\widehat{\sigma_U} = 0$, which may occur with positive probability for some estimators, e.g. the truncated estimator (18), we take this as a hint that there are in fact no random effects (cf. Searle (1971), p. 407). In this case one would not look for outlying random effects either.

3. The ith class is identified as scale-α_ℓ-outlier, if

$$|\ln(s_i) - \ln(\widehat{\sigma_E})| > c_S(\alpha, \widehat{\gamma}), \qquad i = 1, \ldots, \ell. \tag{22}$$

These identification rules should be invariant under linear transformations of the data. This is fulfilled whenever the location estimator is location and scale equivariant and the scale estimators as well as the predictors are location invariant and scale equivariant, which is commonly requested for such statistics.

The functions c_U, c_E, and c_S should be chosen to achieve that under the assumptions of model (1), where outliers are not likely to occur, there is only a small probability, $\alpha \in (0,1)$ say, to detect any outlier,

$$\Pr\left(\exists\, i,j : \frac{|Y_{ij} - \hat{\mu} - \widehat{u}_i|}{\widehat{\sigma_E}} > c_E(\alpha, \widehat{\gamma})\right) = \alpha, \tag{23}$$

$$\Pr\left(\widehat{\sigma_U} > 0 \quad \text{and} \quad \exists\, i : \frac{|\widehat{u}_i|}{\widehat{\sigma_U}} > c_U(\alpha, \widehat{\gamma})\right) = \alpha, \tag{24}$$

$$\Pr\left(\exists\, i : |\ln(s_i) - \ln(\widehat{\sigma_E})| > c_S(\alpha, \widehat{\gamma})\right) = \alpha, \qquad i = 1, \ldots, \ell. \tag{25}$$

These normalizing constraints are close to (10)–(12) and thus allow indeed to interpret these procedures as rules for the identification of α_n- or α_ℓ-outliers, respectively.

Of course such rules should detect as many true α_n-, α_ℓ-outliers as possible, i.e. they should maximize some criterion like the expected number of detected α-outliers. For this purpose we look for procedures which avoid the so called masking and the swamping effect. This means that an identification rule is misled by the outliers themselves and detects too few or too many outliers, cf. Davies and Gather (1993).

5 An identification rule based on medians

We investigate identification rules as outlined in section 4, using the estimators and predictors of section 3. Approximations to the functions c_E, c_U, and c_S ((20)–(22)) are found by simulating critical values for different values of γ that satisfy conditions (23)–(25) with $\widehat{\gamma}$ replaced by γ.

The simulated values for $c_S(\alpha, \gamma)$ seem to be constant in γ. This is also true for the critical values for location-outliers in the random effetcs, when the identification rule (21) is simplified to the rule which identifies the ith class as α_ℓ-location-outlier if

$$\widehat{\sigma_U} > 0 \quad \text{and} \quad |\text{med}_i - \hat{\mu}| > c_U(\alpha)\widehat{\sigma_U}, \qquad i = 1, \ldots, \ell. \tag{26}$$

But the c_E depent on γ. Nonlinear functions were fitted to the simulated values for c_E, c_U, and c_S, yielding

$$\begin{aligned} c_E(\alpha,\gamma) \approx \widetilde{c_E}(\alpha,\gamma) &= z_{1-\alpha_n/2} + \frac{\ell + \theta_{E1}}{n} + \theta_{E2}\left(\frac{\gamma}{\gamma + \theta_{E3}}\right)^2 && (27) \\ c_U(\alpha,\gamma) \approx \widetilde{c_U}(\alpha,\gamma) &= z_{1-\alpha_\ell/2} + \theta_{U1}\left(\ell - \theta_{U2}\right)^{\theta_{U3}} && (28) \\ c_S(\alpha,\gamma) \approx \widetilde{c_S}(\alpha,\gamma) &= \theta_{S1} + \frac{\theta_{S2}\ell}{(\ell + \theta_{S3})(m + \theta_{S4})}. && (29) \end{aligned}$$

Here m equals n/ℓ, rounded to the nearest integer. The quantiles $z_{1-\alpha_n/2}$ and $z_{1-\alpha_\ell/2}$ would be the correct critical values if the model parameters were known. Note that the fitted functions converge to these values or the constant θ_{S1}, respectively, when n and γ in (27), ℓ in (28) with $\theta_{U3} < 0$, and m and/or ℓ in (29) are growing. The θ's depend on α and on whether ℓ or m is even or odd. They are tabulated in the appendix.

In practice γ is unknown and is therefore replaced by the estimator $\widehat{\gamma} = \widehat{\sigma_U}/\widehat{\sigma_E}$ in (27). Simulations confirm that conditions (10–12) are satisfied in general when proceeding as above. However, the identification rule for outliers within the classes seems to be somewhat conservative or liberal in some situations, depeding on the combination of ℓ and m. This may be partly due to the fact that the critical values depend on an estimate of γ.

6 An Example

Table 1 lists the data of the introductory example, along with some auxiliary statistics which help to calculate the statistics introduced above. The seventh

Table 1: Results from radon intercomparison

Lab.	measurements					statistics			
i	in Becquerel/m^3					med_i	$\lvert \mathrm{med}_i - \hat{\mu} \rvert$	mad_i	s_i
1	166	166	167	167	179	167	6	1	1.788
2	156	161	167	154	165	161	0	5	8.940
3	$\underline{145}$[1]	237	259	208	272	$\underline{237}$[1]	76	29	51.852
4	186	161	166	134	145	161	0	16	28.608
5	148	144	143	135	$\underline{97}$[1]	143	18	5	8.940
					some medians	$\hat{\mu} = 161$	6		8.940
estimates of variance components							$\widehat{\sigma_U^2} = 87.723$		$\widehat{\sigma_E^2} = 83.456$

[1] outliers identified at $\alpha = 0.05$

column lists the within laboratory medians med_i, their median is used as estimator of the location parameter μ, i.e. $\hat{\mu} = 161$.

Columns 8 and 10 show the absolute residuals $|\mathrm{med}_i - \hat{\mu}|$ and the scale estimates $s_i = e(5)\mathrm{mad}_i$, which are needed to identify outliers in the random effects and scale-outliers, respectively. Here $e(5) = 1.4826 \cdot 1.206 = 1.788$. Squaring the medians of these quantities and multiplying them with e_U and e_E, resp., gives the estimates of the variance components. Here, $e_U = 5(5 + 1.56)(1.4826 \cdot 1.206)^2 = 2.4367$ and $e_E = 0.9797 + 1.1188(5 - 3.5592)/25 = 1.0442$. Note that taking medians and squaring of positive numbers is interchangeable.

For the identification of outliers critical values according to (27)–(29) are calculated. At first, a estimate of γ is obtained as $\sqrt{87.723/83.456} = 1.0252$. Using the parameters for $\alpha = 0.05$ from the appendix gives $\widetilde{c_E} = 4.561, \widetilde{c_U} = 5.066, \widetilde{c_S} = 1.834$. The lowest observations from laboratories 3 and 5 differ more than $\widetilde{c_E}\widehat{\sigma_E} = 4.561 \cdot 9.135 = 41.665$ from the corresponding median and are therefore identified as within class outliers. The median value of lab. 3 is more than $\widetilde{c_U}\widehat{\sigma_U} = 5.066 \cdot 9.366 = 47.448$ greater than the overall median $\hat{\mu} = 161$, therefore the ith laboratory is identified as outlier in the random effects. No scale-outlier is identified, since none of the $\ln(s_i)$ deviates more than $\widetilde{c_S} = 1.834$ from $\ln(\widehat{\sigma_E}) = 2.212$.

7 Discussion

Figure 1 exemplifies several patterns of departure from the assumptions of model (1). These patterns are described to a great extent by our definition of outliers. The procedures of sections 3, 4, and 5 translate these concepts into easily applicable statistical methods.

In our example, the procedure identifies the most obvious outliers from laboratories 3 and 5. However, no scale-outlier is identified, despite the great variability among the scale estimates within the laboratories.

It can be expected that the methods can be improved by using more efficient and robust estimators and predictors and by adapting stepwise procedures for the detection of outliers in univariate data.

Acknowledgement

Research was supported by Deutsche Forschungsgemeinschaft (DFG), Sonderforschungsbereich 475, and DFG Grants DA 237/1-2 and GA 338/2-2. The authors are grateful to Prof. Laurie Davies for stimulating discussions.

References

Barnett, V., and T. Lewis. *Outliers in Statistical Data.* John Wiley & Sons, New York, third edition, 1994.

Cadwell, J. H. The distribution of quantiles of small samples. *Biometrika*, 39:207–211, 1952.

Croux, C., and P. J. Rousseeuw. Time-efficient algorithms for two highly robust estimators of scale. In Y. Dodge and J. Whittaker, editors, *Computational Statistics*, volume 1, pages 411–428. Physika-Verlag, Heidelberg, 1992.

Davies, P. L. A stochastic model for interlaboratory tests. *Computational Statistics & Data Analysis*, 12:201–209, 1991.

Davies, P. L. and U. Gather. The identification of multiple outliers. Invited paper with discussion and reply. *Journal of the American Statistical Association, Theory and Methods*, 88:782–792, 1993.

Hartung, J. Nonnegative minimum biased invariant estimation in variance component models. *The Annals of Statistics*, 9:278–292, 1981.

Hawkins, D. M. *Identification of Outliers.* Chapman and Hall, London, 1980.

Kreienbrock, L., A. Poffijn, M. Tirmache, M. Feider, A. Kies, and S. C. Darby. Intercomparisons of passive radon-detectors under field conditions in epidemiological studies. *Health Physics*, 76(5):558–563, 1999.

LaMotte, L. R. On non-negative quadratic unbiased estimation of variance components. *Journal of the American Statistical Association*, 68:728–730, 1973.

Lax, D. A. Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association, Theory and Methods*, 80:736–741, 1985.

Searle, S. R. *Linear Models.* John Wiley & Sons, New York, 1971.

Searle, S. R. *Linear Models for Unbalanced Data.* John Wiley & Sons, New York, 1987.

Stahel, W. A., and A. Welsh. Robust estimation of variance components. Technical report, ETH Zürich, Switzerland, 1992.

Terbeck, W., and P. L. Davies. Interactions and outliers in the two-way analysis of variance. *Annals of Statistics*, 26:1279–1305, 1998.

Verdooren, L. R. On estimation of variance components. *Statistica Neerlandica*, pages 83–106, 1980.

Wellmann, J. *Robuste statistische Verfahren und Ausreißeridentifikation beim Modell der Einfachklassifikation mit zufälligen Effekten.* PhD thesis, Department of Statistics, University of Dortmund, 1994.

Wellmann, J. Robustness of an S-Estimator in the One-Way Random Effects Model. *Biometrical Journal*, 42(2):215–221, 2000.

A Approximation of critical values

The results for the critical values c_E, c_U, and c_S are based on 10,000 simulations for balanced designs with $\ell, m = 3(1)12$, $15(3)30$, $\gamma = 0$, $1/10$, $1/4$, $1/3$, $1/2$, $2/3$, $3/4$, 1, 2, 4, 7, 10, $\alpha = 0.01$, 0.025, 0.05, 0.075, 0.1, 0.125, 0.15, and 0.2. The averages and maxima of the absolute relative deviation of the simulated critical values $\hat{c}$ from the fitted curve $\tilde{c}$, defined as $(|\hat{c} - \tilde{c}|)/\hat{c}$, are tabulated too.

Table 2: Approximation for $c_E(\alpha, \gamma)$

α	ℓ	m	θ_{E1}	θ_{E2}	θ_{E3}	abs. rel. dev. average	abs. rel. dev. max
0.01	even	even	43.2945	0.2168	0.0983	0.013	0.050
		odd	98.7835	0.0000	0.4675	0.063	0.272
	odd	even	66.7645	1.6639	5.2218	0.050	0.433
		odd	149.7464	0.0000	0.0000	0.134	0.584
0.025	even	even	28.7092	0.1976	0.0811	0.013	0.058
		odd	67.3591	0.0841	1.3156	0.046	0.208
	odd	even	43.9021	0.4244	1.3694	0.030	0.269
		odd	90.4922	2.3905	499517.3	0.072	0.376
0.05	even	even	20.0756	0.1701	0.0826	0.013	0.062
		odd	46.8162	0.1144	0.3953	0.035	0.156
	odd	even	29.2568	0.2408	0.4617	0.023	0.165
		odd	59.8869	0.3051	6.7851	0.046	0.259
0.075	even	even	15.1429	0.1546	0.0861	0.014	0.063
		odd	35.7368	0.1304	0.2639	0.029	0.138
	odd	even	22.4798	0.1939	0.3233	0.021	0.123
		odd	45.2405	0.1852	1.9774	0.035	0.208
0.1	even	even	11.8195	0.1424	0.0923	0.015	0.070
		odd	28.7363	0.1355	0.2151	0.025	0.126
	odd	even	18.0670	0.1693	0.2694	0.020	0.104
		odd	36.2594	0.1606	0.9856	0.029	0.163
0.125	even	even	9.4091	0.1316	0.1006	0.016	0.074
		odd	23.9408	0.1339	0.2078	0.022	0.117
	odd	even	14.8461	0.1538	0.2436	0.020	0.096
		odd	29.9873	0.1542	0.6556	0.025	0.141
0.15	even	even	7.5064	0.1224	0.1103	0.016	0.076
		odd	20.3330	0.1300	0.2105	0.020	0.108
	odd	even	12.3927	0.1397	0.2351	0.021	0.096
		odd	25.3297	0.1469	0.5351	0.021	0.124
0.2	even	even	4.5492	0.1062	0.1262	0.017	0.082
		odd	14.5664	0.1227	0.2069	0.016	0.094
	odd	even	8.7006	0.1194	0.2333	0.021	0.091
		odd	18.6299	0.1342	0.4301	0.017	0.097

Table 3: Approximation for $c_U(\alpha, \gamma)$

α	ℓ	θ_{U1}	θ_{U2}	θ_{U3}	abs. rel. dev. average	max
0.01	even	22.1191	1.7531	-0.8739	0.019	0.143
	odd	38.0453	2.5403	-1.0935	0.038	0.238
0.025	even	14.7482	0.9919	-0.8284	0.013	0.079
	odd	19.1942	2.4546	-0.9435	0.023	0.176
0.05	even	11.5456	-0.2643	-0.8212	0.010	0.068
	odd	10.9473	2.3549	-0.8344	0.016	0.144
0.075	even	11.5518	-1.9997	-0.8608	0.009	0.051
	odd	7.6956	2.2635	-0.7732	0.012	0.094
0.1	even	13.4682	-4.3361	-0.9282	0.008	0.053
	odd	5.8465	2.1881	-0.7278	0.010	0.099
0.125	even	21.6681	-8.2832	-1.0666	0.007	0.049
	odd	4.6777	2.1073	-0.6938	0.010	0.103
0.15	even	44.9717	-13.7962	-1.2527	0.007	0.038
	odd	3.8730	2.0134	-0.6676	0.009	0.082
0.2	even	45.0000	-19.4365	-1.2523	0.008	0.044
	odd	2.8178	1.7854	-0.6282	0.008	0.075

The model for c_S was actually

$$c_S(\alpha,\gamma) = \theta_{S1} + x_0 \left(\frac{\theta_{S2,0}\ell}{(\ell+\theta_{S3,0})(m+\theta_{S4,0})} \right) + x_1 \left(\frac{\theta_{S2,1}\ell}{(\ell+\theta_{S3,1})(m+\theta_{S4,1})} \right),$$

with $x_0 = 1$ for m even, $x_0 = 0$ otherwise and $x_1 = 1 - x_0$. This results in a unique parameter θ_{S1} for both m odd and even and therefore a unique asymptotic critical value for growing m. As a consequence the absolute relative deviation can not be distinguished for odd and even m.

Table 4: Approximation for $c_S(\alpha,\gamma)$

						abs. rel. dev.	
α	m	θ_{S1}	θ_{S2}	θ_{S3}	θ_{S4}	average	max
0.01	even	0.4477	15.8189	2.5456	0.1735	0.021	0.088
	odd		15.3729	2.3991	-0.9559		
0.025	even	0.4125	14.4887	3.1908	0.2629	0.020	0.077
	odd		14.0051	3.0530	-0.9224		
0.05	even	0.3832	13.5444	3.9216	0.3426	0.020	0.095
	odd		12.9677	3.7308	-0.8865		
0.075	even	0.3663	12.9185	4.4797	0.3627	0.019	0.096
	odd		12.3486	4.2342	-0.8680		
0.1	even	0.3532	12.4912	4.9454	0.3879	0.019	0.105
	odd		11.9512	4.7216	-0.8481		
0.125	even	0.3426	12.2155	5.4184	0.4213	0.019	0.112
	odd		11.6554	5.1678	-0.8325		
0.15	even	0.3336	11.9883	5.8711	0.4410	0.018	0.115
	odd		11.4529	5.6309	-0.8162		
0.2	even	0.3185	11.6439	6.6663	0.5010	0.019	0.129
	odd		11.0477	6.4034	-0.7937		