Contents

1 Lorelia

Method comparison studies are performed in order to prove equivalence between two measurement methods or instruments. The identification of outliers is an important

part of data analysis as outliers can indicate serious errors in the measurement process. Common outlier tests proposed in the literature require a homogeneous sample distribution and homoscedastic random error variances. However, datasets in method comparison studies usually do not meet these assumptions. To overcome this problem, different data transformation methods are proposed in the literature. However, they will only be applicable if the random errors can be described by simple additive or multiplicative models. In this work, a new outlier test based on robust linear regression is proposed which provides a general solution to the above problem. The LORELIA (LOcal RELIAbility) residual test is based on a local, robust residual variance estimator, given as a weighted sum of the observed residuals. Outlier limits are estimated from the actual data situation without making assumptions on the underlying error variance model. The performance of the new test is demonstrated in examples and simulations.

Method comparison studies are performed in order to prove equivalence between two measurement methods or instruments. The identification of outliers is an important part of data analysis as outliers can indicate serious errors in the measurement process. Common outlier tests proposed in the literature require a homogeneous sample distribution and homoscedastic random error variances. However, datasets in method comparison studies usually do not meet these assumptions. To overcome this problem, different data transformation methods are proposed in the literature. However, they will only be applicable if the random errors can be described by simple additive or multiplicative models. In this work, a new outlier test based on robust linear regression is proposed which provides a general solution to the above problem. The LORELIA (LOcal RELIAbility) residual test is based on a local, robust residual variance estimator, given as a weighted sum of the observed residuals. Outlier limits are estimated from the actual data situation without making assumptions on the underlying error variance model. The performance of the new test is demonstrated in examples and simulations.

2 Note on Roy's paper

1. Basic model:

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i}, \qquad i = 1, \ldots, n$$

$$\boldsymbol{Z_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma^2 I})$$

Assumptions are made about homoskedasticity.

2. General model:

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}, \qquad i = 1, \ldots, n$$
$$\boldsymbol{Z_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \quad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma^2 \Lambda})$$

Assumptions about homoskedasticity are relaxed (Pinheiro and Bates, 1994, pg.202).

3. $\sigma^2\boldsymbol{\Lambda}$ is the general form for the VC structure for residuals.

4. The response vector $\boldsymbol{y}_i$ comprises the observations of the subject, as measured by two methods, taking three measurements each. Hence a 6×1 random vector corresponding to the ith subject.

$$\boldsymbol{y}_i = (y_i^{j1}, y_i^{Jj2}, y_i^{j3}, y_i^{s1}, y_i^{s2}, y_i^{s3})\prime \tag{1}$$

5. The number of replicates is p. A subject will have up to $2p$ measurements, for the two instrument case, i.e. $Max(n_i) = 2p$. (Let k denote number of instruments, which is assumed to be 2 unless stated otherwise.) For the blood pressure data $p = 3$.

6. $\boldsymbol{\Psi}$ refers to the between-subject sources of variation. $\boldsymbol{R}_i$ refers to the within-subject source of variation between two methods. LME models allow for the explicit analysis of each.

7. $\boldsymbol{\Psi}$ is the variance covariance structure for the random effects.

   There is three alternative structures for $\boldsymbol{\Psi}$, the diagonal form, the identity form and the general form.

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_1^2 & 0 \\ 0 & \psi_2^2 \end{pmatrix} \quad \text{or} \quad \boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \quad \text{or} \quad \boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$$

8. $\boldsymbol{\epsilon}_i$ is a n_i-dimensional vector comprised of residual components. For the blood pressure data $n_i = 85$.

9. $\boldsymbol{\beta}$ is the solutions of the means of the two methods. In the LME output, the bias ad corresponding t-value and p-values are presented. This is relevant to Roy's first test.

10. $\boldsymbol{b}_i$ is a $m-$dimensional vector comprised of the random effects.

$$\boldsymbol{b}_i = \begin{pmatrix} b_{1i} \\ b_{21} \end{pmatrix} \tag{2}$$

11. $\boldsymbol{\Psi}$ is the variance-covariance matrix of the random effects , with 2×2 dimensions.

$$\boldsymbol{\Psi} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \tag{3}$$

12. $\boldsymbol{\Sigma}$ represents the partial VC matrix of the established matrix and the new method for any replicates.

$$\Sigma = \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \tag{4}$$

- σ_e^2 - partial variance of the established method.
- σ_n^2 - partial variance of the new method.
- σ_{en} - partial covariance between both methods.

13. $\boldsymbol{V}$ represents the correlation matrix of the replicated measurements on a given method. $\boldsymbol{\Sigma}$ is the within-subject VC matrix.

14. $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are positive definite matrices. The dimensions of $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ are $3 \times 3 (= p \times p)$ and $2 \times 2 (= k \times k)$.

15. It is assumed that $\boldsymbol{V}$ is the same for both methods and $\boldsymbol{\Sigma}$ is the same for all replications.

16. $\boldsymbol{V} \bigotimes \boldsymbol{\Sigma}$ creates a $6 \times 6 (= kp \times kp)$ matrix. $\boldsymbol{R}_i$ is a sub-matrix of this.

17. The variance covariance structure $\boldsymbol{R}_i$ has a separable covariance structure.

18. The overall variability Block$\boldsymbol{\Omega_i}$ is the sum of the between-subject variability $\boldsymbol{\Psi}$ and the within subject variability $\boldsymbol{\Sigma}$

$$\begin{pmatrix} \omega_e^2 & \omega^{en} \\ \omega_{en} & \omega_n^2 \end{pmatrix} = \begin{pmatrix} \psi_e^2 & \psi^{en} \\ \psi_{en} & \psi_n^2 \end{pmatrix} + \begin{pmatrix} \sigma_e^2 & \sigma^{en} \\ \sigma_{en} & \sigma_n^2 \end{pmatrix} \tag{5}$$

19. No special form of the random effects VC matrix $\boldsymbol{\Psi}$ is assumed. Form cans be specified. The `pdMat class is used by the 'nlme' package to specify patterned VC matrices.`

    - `pdDiag - Assumes random effects are independent, with different variance.`
    - `pdIdent - Assumes random effects are independent, with same variance.`
    - `pdSymm - General symmetric positive definite matrix.`
    - `pdCompSymm`

3 Lambda Structure

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Lambda}) \tag{6}$$

1. A simple assumption is to assumes that residuals are independent and homoscedastic, i.e. $\boldsymbol{Lambda = I}$.

2. For the Bland Altman blood pressure data,$\boldsymbol{\Lambda}$ has kronecker product structure and has dimensions 6×6.

3.1 Variance-Covariance Structures

3.1.1 Independence

As though analyzed using between subjects analysis.

$$\begin{pmatrix} \psi^2 & 0 & 0 \\ 0 & \psi^2 & 0 \\ 0 & 0 & \psi^2 \end{pmatrix}$$

3.1.2 Compound Symmetry

Assumes that the variance-covariance structure has a single variance (represented by ψ^2) for all 3 of the time points and a single covariance (represented by ψ_{ij}) for each of the pairs of trials.

$$\begin{pmatrix} \psi^2 & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi^2 & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi^2 \end{pmatrix}$$

3.1.3 Unstructured

Assumes that each variance and covariance is unique. Each trial has its own variance (e.g. s12 is the variance of trial 1) and each pair of trials has its own covariance (e.g. s21 is the covariance of trial 1 and trial2). This structure is illustrated by the half matrix below.

3.1.4 Autoregressive

Another common covariance structure which is frequently observed in repeated measures data is an autoregressive structure, which recognizes that observations which are more proximate are more correlated than measures that are more distant.

4 Basic Models Fits

Further to **?**, several simple LME models are constructed for the blood pressure data. This data set is the subject of a method comparison study in Bland and Altman (1999).

4.1 Implementing the Mixed Models Fits

They are implemented using the following `R` code, utilising the 'nlme' package. An analysis of variance is used to compare the model fits.

The `R` script:

```
fit1 = lme( BP ~ method, data = dat, random = ~1 | subject )
fit2 = update(fit1, random = ~1 | subject/method )
fit3 = update(fit1, random = ~method - 1 | subject )
#analysis of variance
anova(fit1,fit2,fit3)
```

1. Simplest workable model, allows differences between methods and incorporates a random intercept for each subject. For subject 1 we have

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{Z}_i = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{b}_i = b$$

   where $\mathrm{E}(b) = 0$ and $\mathrm{var}(b) = \psi$.

2.

$$\boldsymbol{Z}_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \quad \boldsymbol{b}_i = \begin{pmatrix} b_1 & 0 \\ 0 & b_2 \end{pmatrix}$$

where $\mathrm{E}(b_i) = 0$ and $\mathrm{var}(\boldsymbol{b}) = \boldsymbol{\Psi}$.

The variance of error terms is a 6×6 matrix.

4.2 Laird Ware Formulation

$$\boldsymbol{y_i} = \boldsymbol{X_i\beta} + \boldsymbol{Z_i b_i} + \boldsymbol{\epsilon_i}, \qquad i = 1, \ldots, 85$$

$$\boldsymbol{Z_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma^2\Lambda})$$

4.3 Model Fit 1

This is a simple model with no interactions. There is a fixed effect for each method and a random effect for each subject.

$$y_{ijk} = \beta_j + b_i + \epsilon_{ijk}, \qquad i = 1, \ldots, 2, j = 1, \ldots, 85, k = 1, \ldots, 3$$

$$b_i \sim \mathcal{N}(0, \sigma_b^2), \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

```
Linear mixed-effects model fit by REML
  Data: dat
  Log-restricted-likelihood: -2155.853
  Fixed: BP ~ method
(Intercept)     methodS
  127.40784    15.61961

Random effects:
 Formula: ~1 | subject
        (Intercept) Residual
StdDev:    29.39085 12.44454

Number of Observations: 510
Number of Groups: 85
```

4.4 Model Fit 2

This is a simple model, this time with an interaction effect. There is a fixed effect for each method. This model has random effects at two levels b_i for the subject, and another, b_{ij}, for the respective method within each subject.

$$y_{ijk} = \beta_j + b_i + b_{ij} + \epsilon_{ijk}, \qquad i = 1, \ldots, 2, j = 1, \ldots, 85, k = 1, \ldots, 3$$

$$b_i \sim \mathcal{N}(0, \sigma_1^2), \qquad b_{ij} \sim \mathcal{N}(0, \sigma_2^2), \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

In this model, the random interaction terms all have the same variance σ_2^2. These terms are assumed to be independent of each other, even within the same subject.

```
Linear mixed-effects model fit by REML
  Data: dat
  Log-restricted-likelihood: -2047.714
  Fixed: BP ~ method
(Intercept)     methodS
  127.40784    15.61961

Random effects:
 Formula: ~1 | subject
        (Intercept)
StdDev:    28.28452

 Formula: ~1 | method %in% subject
        (Intercept) Residual
StdDev:    12.61562 7.763666

Number of Observations: 510
Number of Groups:
            subject method %in% subject
                 85                 170
```

4.5 Model Fit 3

This model is a more general model, compared to 'model fit 2'. This model treats the random interactions for each subject as a vector and allows the variance-covariance matrix for that vector to be estimated from the set of all positive-definite matrices. $\boldsymbol{y_i}$ is the entire response vector for the ith subject. $\boldsymbol{X_i}$ and $\boldsymbol{Z_i}$ are the fixed- and random-effects design matrices respectively.

$$\boldsymbol{y_i} = \boldsymbol{X_i}\boldsymbol{\beta} + \boldsymbol{Z_i}\boldsymbol{b_i} + \boldsymbol{\epsilon_i}, \qquad i = 1, \ldots, 85$$

$$\boldsymbol{Z_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Psi}), \qquad \boldsymbol{\epsilon_i} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\sigma^2\Lambda})$$

For the first subject the response vector, $\boldsymbol{y_1}$, is:

observation	BP	subject	method	replicate
1	100.00	1	J	1
86	106.00	1	J	2
171	107.00	1	J	3
511	122.00	1	S	1
596	128.00	1	S	2
681	124.00	1	S	3

The fixed effects design matrix $\boldsymbol{X_i}$ is given by:

(Intercept)	method S
1	0
1	0
1	0
1	1
1	1
1	1

The random effects design matrix $\boldsymbol{Z_i}$ is given by:

method J	method S
1	0
1	0
1	0
0	1
0	1
0	1

The following output was obtained.

```
Linear mixed-effects model fit by REML
  Data: dat
  Log-restricted-likelihood: -2047.582
  Fixed: BP ~ method
(Intercept)     methodS
  127.40784    15.61961

Random effects:
 Formula: ~method - 1 | subject
 Structure: General positive-definite, Log-Cholesky parametrization
         StdDev    Corr
methodJ  30.455093 methdJ
methodS  31.477237 0.835
Residual  7.763666

Number of Observations: 510
Number of Groups: 85
```

5 Simplifying GLS (K Hayes)

5.1 Introduction

Hayes and Haslett (1998) present an approach to the problem of **general least squares** estimation of the general linear model in terms of constrained optimization, which is in turn solved via Lagrange multipliers. The crux of the proposed approach is that one system of equations is sufficiently versatile, and provides for

- the estimation of new observations,
- estimation of fixed parameters in regression
- estimation of fixed and random effects in mixed models,
- the diagnostics associated with conditional and marginal residuals
- and of subset deletion.

5.2 Overview

Hayes and Haslett (1998) have demonstrated how the problem of best linear unbiased estimation can be posed in terms of Lagrange multipliers. Both BLUE and BLUP can be treated as distinct estimation problems from the following equation.

$$\begin{pmatrix} V & X \\ X^t & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}_z \\ \boldsymbol{\gamma}_z \end{pmatrix} = \begin{pmatrix} \text{cov}(Y, Z) \\ A^t \end{pmatrix} \tag{7}$$

Hence BLUE and BLUP can be considered as the estimation of two different variables from Y. This equation has a natural role in the derivation of *leave- k-out* residuals and diagnostic measures, and replaces the traditional approach of using a variety of clumsy updating formulas. Note that this approach may be used to determine the impact of deletion on any quantity computed from Y.

General Linear model

6 General Linear model

Mixed Effects Models are seen as especially robust in the analysis of unbalanced data when compared to similar analyses done under the General Linear Model framework

(Pinheiro and Bates, 2000).

A Mixed Effects Model is an extension of the General Linear Model that can specify additional random effects terms

6.1 Equivalence of LME model

Henderson's mixed model equations are presented on page 147 of Youngjo et al. Youngjo et al demonstrate that this formulation is equivalent to an augmented general linear model.

Youngjo et al show that the linear mixed effects model can be shown to be the augmented classical linear model involving fixed effects parameters only.

7 Augmented GLMs

7.1 Augmented linear model

The subscript M is a label referring to the mean model.

$$\begin{pmatrix} Y \\ \psi_M \end{pmatrix} = \begin{pmatrix} X & Z \\ 0 & I \end{pmatrix} \begin{pmatrix} \beta \\ \nu \end{pmatrix} + e^* \tag{8}$$

The error term e^* is normal with mean zero. The variance matrix of the error term is given by

$$\Sigma_a = \begin{pmatrix} \Sigma & 0 \\ 0 & D \end{pmatrix}. \tag{9}$$

$$X = \begin{pmatrix} T & Z \\ 0 & I \end{pmatrix} \delta = \begin{pmatrix} \beta \\ \nu \end{pmatrix} \tag{10}$$

$$y_a = T\delta + e^* \tag{11}$$

Weighted least squares equation

References

Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research 8*(2), 135–160.

Pinheiro, J. and D. Bates (1994). *Mixed Effects Models in S and S plus* (2nd ed.). Reading, Massachusetts: Springer.