

Roy uses an LME model approach to provide a set of formal tests for method comparison studies.

Four candidate models are fitted to the data.

These models are similar to one another, but for the imposition of equality constraints.

These tests are the pairwise comparison of candidate models, one formulated without constraints, the other with a constraint.

Carstensen also uses a LME model for examining MCS with replicates.

Roy's model uses fixed effects $\beta_0 + \beta_1$ and $\beta_0 + \beta_1$ to specify the mean of all observations by methods 1 and 2 respectively.

Carstensen allocates a fixed, but unknown, mean for each individual. [Grubbs(1948) model.]

His interest lies in calculating the LoA as opposed to Formalized testing.

Roy adheres to Random Effect ideas in ANOVA

Roy treats items as a sample from a population.

Allocation of fixed effects and random effects are very different in each model

Carstensen's interest lies in the difference between the population from which they were drawn.

Carstensen's model is a mixed effects ANOVA.

$$Y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \quad c_{mi} \sim \tau_{\Downarrow}^{\epsilon}, \quad e_{mir} \sim \sigma_{\Downarrow}^{\epsilon},$$

This model includes a method by item iteration term.

Carstensen presents two models. One for the case where the replicates, and a second for when they are linked.

Carstensen's model does not take into account either between-item or within-item covariance between methods.

Importantly, estimates required to calculate the limits of agreement are not extractable, and therefore the calculation must be done by hand.

In cases where there is negligible covariance between methods, the limits of agreement computed using Roy's model accord with those computed using Carstensen's model. In cases where some degree of covariance is present between the two methods, the limits of agreement computed using models will differ.

In the presented example, it is shown that Roy's LoAs are lower than those of Carstensen.

Carstensen makes some interesting remarks in this regard.

The only slightly non-standard (meaning "not often used") feature is the differing residual variances between methods.

0.1 Remarks on the Multivariate Normal Distribution

The multivariate normal distribution of a k -dimensional random vector $X = [X_1, X_2, \dots, X_k]$ can be written in the following notation:

$$X \sim \mathcal{N}(\mu, \Sigma),$$

or to make it explicitly known that X is k -dimensional,

$$X \sim \mathcal{N}_k(\mu, \Sigma).$$

with k -dimensional mean vector

$$\mu = [\mathbb{E}[X_1], \mathbb{E}[X_2], \dots, \mathbb{E}[X_k]]$$

and $k \times k$ covariance matrix

$$\Sigma = [\text{Cov}[X_i, X_j]], \quad i = 1, 2, \dots, k; \quad j = 1, 2, \dots, k$$

1. Univariate Normal Distribution

$$X \sim \mathcal{N}(\mu, \sigma^2),$$

2. Bivariate Normal Distribution

(a)

$$X \sim \mathcal{N}_2(\mu, \Sigma),$$

(b)

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix}.$$

Diligence is required when considering the models. Carstensen specifies his models in terms of the univariate normal distribution. Roy's model is specified using the bivariate normal distribution. This gives rise to a key difference between the two models, in that a bivariate model accounts for covariance between the variables of interest.

1 Introduction

Roy (2009) uses an approach based on linear mixed effects (LME) models for the purpose of comparing the agreement between two methods of measurement, where replicate measurements on items, typically individuals, by both methods are available. She provides three tests of hypothesis appropriate for evaluating the agreement between the two methods of measurement under this sampling scheme. These tests consider null hypotheses that assume: absence of inter-method bias; equality of between-subject variabilities of the two methods; equality of within-subject variabilities of the two methods. By inter-method bias we mean that a systematic difference exists between observations recorded by the two methods. Differences in between-subject variabilities of the two methods arise when one method is yielding average response levels for individuals than are more variable than the average response levels for the same sample of individuals taken by the other method. Differences in within-subject variabilities of the two methods arise when one method is yielding responses for an individual than are more variable than the responses for this same individual taken by the other method. The two methods of measurement can be considered to agree, and subsequently can be used interchangeably, if all three null hypotheses are true.

Let y_{mir} be the r th replicate measurement on the i th item by the m th method, where $m = 1, 2$, $i = 1, \dots, N$, and $r = 1, \dots, n_i$. When the design is balanced and there is no ambiguity we can set $n_i = n$. The LME model can be written

$$y_{mir} = \beta_0 + \beta_m + b_{mi} + \epsilon_{mir}. \quad (1)$$

Here β_0 and β_m are fixed-effect terms representing, respectively, a model intercept and an overall effect for method m . The b_{1i} and b_{2i} terms represent random effect parameters corresponding to the two methods, having $E(b_{mi}) = 0$ with $\text{Var}(b_{mi}) = g_m^2$ and $\text{Cov}(b_{mi}, b_{m'i}) = g_{12}$. The random error term for each response is denoted ϵ_{mir} having $E(\epsilon_{mir}) = 0$, $\text{Var}(\epsilon_{mir}) = \sigma_m^2$, $\text{Cov}(b_{mir}, b_{m'ir}) = \sigma_{12}$, $\text{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$ and $\text{Cov}(\epsilon_{mir}, \epsilon_{m'ir'}) = 0$. When two methods of measurement are in agreement, there is no

significant differences between β_1 and β_2 , g_1^2 and g_2^2 , and σ_1^2 and σ_2^2 .

2 Roy's Hypotheses Tests

In order to express Roy's LME model in matrix notation we gather all $2n_i$ observations specific to item i into a single vector $\mathbf{y}_i = (y_{1i1}, y_{2i1}, y_{1i2}, \dots, y_{mir}, \dots, y_{1in_i}, y_{2in_i})'$. The LME model can be written

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$ is a vector of fixed effects, and \mathbf{X}_i is a corresponding $2n_i \times 3$ design matrix for the fixed effects. The random effects are expressed in the vector $\mathbf{b} = (b_1, b_2)'$, with \mathbf{Z}_i the corresponding $2n_i \times 2$ design matrix. The vector $\boldsymbol{\epsilon}_i$ is a $2n_i \times 1$ vector of residual terms.

It is assumed that $\mathbf{b}_i \sim N(0, \mathbf{G})$, $\boldsymbol{\epsilon}_i$ is a matrix of random errors distributed as $N(0, \mathbf{R}_i)$ and that the random effects and residuals are independent of each other.

\mathbf{G} is the variance covariance matrix for the random effects \mathbf{b} . i.e. between-item sources of variation. The between-item variance covariance matrix \mathbf{G} is constructed as follows:

$$\text{Var} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} = \mathbf{G} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix}$$

It is important to note that no special assumptions about the structure of \mathbf{G} are made. An example of such an assumption would be that \mathbf{G} is the product of a scalar value and the identity matrix.

\mathbf{R}_i is the variance covariance matrix for the residuals, i.e. the within-item sources of variation between both methods. Computational analysis of linear mixed effects models allow for the explicit analysis of both \mathbf{G} and \mathbf{R}_i .

Hamlett et al. (2004) shows that \mathbf{R}_i can be expressed as $\mathbf{R}_i = \mathbf{I}_{n_i} \otimes \boldsymbol{\Sigma}$. The partial within-item variance?covariance matrix of two methods at any replicate is denoted $\boldsymbol{\Sigma}$, where σ_1^2 and σ_2^2 are the within-subject variances of the respective methods, and σ_{12} is the within-item covariance between the two methods. It is assumed that the within-item variance?covariance matrix $\boldsymbol{\Sigma}$ is the same for all replications. Again it is

important to note that no special assumptions are made about the structure of the matrix.

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \quad (2)$$

For expository purposes consider the case where each item provides three replicates by each method. Then in matrix notation the model has the structure

$$\mathbf{y}_i = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_{1i} \\ b_{2i} \end{pmatrix} + \begin{pmatrix} \epsilon_{1i1} \\ \epsilon_{2i1} \\ \epsilon_{1i2} \\ \epsilon_{2i2} \\ \epsilon_{1i3} \\ \epsilon_{2i3} \end{pmatrix}, \quad (3)$$

where

$$\mathbf{G} =$$

and

$$\mathbf{R}_i =$$

It is assumed that $\mathbf{b}_i \sim N(0, \mathbf{G})$, ϵ_i is a matrix of random errors distributed as $N(0, \mathbf{R}_i)$ and that the random effects and residuals are independent of each other. Assumptions made on the structures of \mathbf{G} and \mathbf{R}_i will be discussed in due course.

The overall variability between the two methods is the sum of between-item variability \mathbf{G} and within-item variability $\mathbf{\Sigma}$. Roy (2009) denotes the overall variability as Block - $\mathbf{\Omega}_i$. The overall variation for methods 1 and 2 are given by

$$\begin{pmatrix} \omega_1^2 & \omega_{12} \\ \omega_{12} & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g_1^2 & g_{12} \\ g_{12} & g_2^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

The computation of the limits of agreement require that the variance of the difference of measurements. This variance is easily computable from the estimate of the Block - $\mathbf{\Omega}_i$ matrix. Lack of agreement can arise if there is a disagreement in overall variabilities. This may be due to the disagreement in either between-item variabilities or within-item variabilities, or both. Roy (2009) allows for a formal test of each.

2.1 Hypothesis Testing

The formulation presented above usefully facilitates a series of significance tests that advise as to how well the two methods agree. These tests are as follows:

- A formal test for the equality of between-item variances,
- A formal test for the equality of within-item variances,
- A formal test for the equality of overall variances.

These tests are complemented by the ability to consider the inter-method bias and the overall correlation coefficient. Two methods can be considered to be in agreement if criteria based upon these methodologies are met. Additionally Roy makes reference to the overall correlation coefficient of the two methods, which is determinable from variance estimates.

3 Carstensen's Limits of agreement

Carstensen et al. (2008) presents a methodology to compute the limits of agreement based on LME models. Importantly, Carstensen's underlying model differs from Roy's model in some key respects, and therefore a prior discussion of Carstensen's model is required.

3.1 Carstensen's Model

Carstensen (2004) presents a model to describe the relationship between a value of measurement and its real value. The non-replicate case is considered first, as it is the context of the Bland Altman plots. This model assumes that inter-method bias is the only difference between the two methods.

A measurement y_{mi} by method m on individual i is formulated as follows;

$$y_{mi} = \alpha_m + \mu_i + e_{mi} \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2) \quad (4)$$

The differences are expressed as $d_i = y_{1i} - y_{2i}$. For the replicate case, an interaction term c is added to the model, with an associated variance component. All the random effects are assumed independent, and that all replicate measurements are assumed to be exchangeable within each method.

$$y_{mir} = \alpha_m + \mu_i + c_{mi} + e_{mir}, \quad e_{mi} \sim \mathcal{N}(0, \sigma_m^2), \quad c_{mi} \sim \mathcal{N}(0, \tau_m^2). \quad (5)$$

Of particular importance is terms of the model, a true value for item i (μ_i). The fixed effect of Roy's model comprise of an intercept term and fixed effect terms for both methods, with no reference to the true value of any individual item. A distinction can be made between the two models: Roy's model is a standard LME model, whereas Carstensen's model is a more complex additive model.

3.2 Assumptions on Variability

Aside from the fixed effects, another important difference is that Carstensen's model requires that particular assumptions be applied, specifically that the off-diagonal elements of the between-item and within-item variability matrices are zero. By extension the overall variability off diagonal elements are also zero.

Also, implementation requires that the between-item variances are estimated as the same value: $g_1^2 = g_2^2 = g^2$. Necessarily Carstensen's method does not allow for a formal test of the between-item variability.

$$\begin{pmatrix} \omega_2^1 & 0 \\ 0 & \omega_2^2 \end{pmatrix} = \begin{pmatrix} g^2 & 0 \\ 0 & g^2 \end{pmatrix} + \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$$

In cases where the off-diagonal terms in the overall variability matrix are close to zero, the limits of agreement due to Carstensen et al. (2008) are very similar to the limits of agreement that follow from the general model.

4 Note 1: Coefficient of Repeatability

The coefficient of repeatability is a measure of how well a measurement method agrees with itself over replicate measurements (Bland and Altman, 1999). Once the within-item variability is known, the computation of the coefficients of repeatability for both methods is straightforward.

5 Note 2: Model terms

It is important to note the following characteristics of this model.

- Let the number of replicate measurements on each item i for both methods be n_i , hence $2 \times n_i$ responses. However, it is assumed that there may be a different number of replicates made for different items. Let the maximum number of replicates be p . An item will have up to $2p$ measurements, i.e. $\max(n_i) = 2p$.
- Later on \mathbf{X}_i will be reduced to a 2×1 matrix, to allow estimation of terms. This is due to a shortage of rank. The fixed effects vector can be modified accordingly.
- \mathbf{Z}_i is the $2n_i \times 2$ model matrix for the random effects for measurement methods on item i .
- \mathbf{b}_i is the 2×1 vector of random-effect coefficients on item i , one for each method.
- $\boldsymbol{\epsilon}$ is the $2n_i \times 1$ vector of residuals for measurements on item i .
- \mathbf{G} is the 2×2 covariance matrix for the random effects.
- \mathbf{R}_i is the $2n_i \times 2n_i$ covariance matrix for the residuals on item i .
- The expected value is given as $E(\mathbf{y}_i) = \mathbf{X}_i\boldsymbol{\beta}$. (Hamlett et al., 2004)
- The variance of the response vector is given by $\text{Var}(\mathbf{y}_i) = \mathbf{Z}_i\mathbf{G}\mathbf{Z}_i' + \mathbf{R}_i$ (Hamlett et al., 2004).

References

- Bland, J. and D. Altman (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research* 8(2), 135–160.
- Carstensen, B. (2004). Comparing and predicting between several methods of measurement. *Biostatistics* 5(3), 399–413.
- Carstensen, B., J. Simpson, and L. C. Gurrin (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics* 4(1).
- Hamlett, A., L. Ryan, and R. Wolfinger (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures. *Proceedings of the Statistics and Data Analysis Section, SAS Users Group International* 198-229, 1–7.
- Roy, A. (2009). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics* 19, 150–173.