Assessment of Agreement Between Two Methods with Replicated Observations

Anuradha Roy
Department of Management Science and Statistics
The University of Texas at San Antonio
San Antonio, Texas 78249, USA
Assessment of Agreement Between Two Methods with Replicated Observations

Anuradha Roy
Department of Management Science and Statistics
The University of Texas at San Antonio
San Antonio, Texas 78249, USA

Abstract

We study the problem of assessing the agreement between two methods with any number of replicated observations using linear mixed effects (LME) model in a doubly multivariate set-up. This method can also be used in the case of unbalanced designs when number of replications for each patient is unequal, as well as when the number of replications for each patient by respective methods is unequal. This method can easily incorporate any covariate, especially categorical to substantiate its effect on the method assessment. The model is implemented using MIXED procedure of SAS. We demonstrate our method with three real data sets.

Keywords: Assessment of Agreement, Kronecker Product, Maximum Likelihood Estimates, Mixed Effects Model, Replicated Observations, PROC MIXED.

JEL Code: M110

1 Introduction

It is often required to compare a new measurement technique with an established one of measuring some quantity such as carbon dioxide production, blood pressure, body fat or even child’s weight. The simple and relatively inexpensive methods for gathering quantitative data in comparison to the expensive gold standard one are always appraised. It is often needed to see whether they agree so that both of them can be used interchangeably. The question to be answered in this paper is, “Do the two methods of measurement agree statistically?” so that one can switch them, if needed. The problem has been discussed by many authors (Bland and Altman (1983, 1986, 1990, 1999); Lee, Koh and Ong, 1989; Lin 1989; St. Laurent, 1998, Bartko, 1994; Argall et.al. 2003; Choudhary and Nagaraja, 2005). A common feature of all these approaches is that they used various factors, such as a systematic bias, a difference in variabilities and a low correlation that cause disagreement. Choudhary and Nagaraja (2005) combined all the above factors into
a single measure using the intersection-union principle. However, all these authors except Bland and Altman (1986, 1999) used only a single measurement on each subject for each method. Bland and Altman pointed out correctly that a single measurement on each subject is not be able to judge which method is more precise; lack of preciseness can certainly interfere with the comparison of two methods. They also mentioned in their paper that for more than two replicated measurements the calculations become very complicated; but, strongly recommended the simultaneous estimation of repeatability and agreement by collecting replicated data. Repeatability is very important to the study of method comparison because repeatabilities of the methods of measurements limit the amount of agreement and the best way to check repeatability is to take replicated measurements on a series of subjects. As mentioned in Bland and Altman (1986) repeatability plays a significant role in method comparison study. If one method has poor repeatability in the sense of considerable variation, the agreement between the two methods is bound to be poor. Even if the old method is the more variable one, a new method which is perfect will not agree with it. By replicates Bland and Altman meant two or more measurements on the same individual taken in identical conditions. In general these mean that the measurements are taken in quick succession. We can assume that these replicated measurements are equicorrelated and we must take this equicorrelated structure of the replicates into account while assessing the agreement between the two methods. Bland and Altman (1999) calculated the repeatability coefficient for each method, regrettably they did not test the agreement between them formally. They also calculated the bias, again unfortunately they did not test its statistical significance. These two authors explored the agreement between two measurement methods by asking the question “Do the two methods of measurement agree sufficiently closely?”. And, they answered this question by estimating two limits of agreement. But, this idea of limits of agreement is too limiting. We try to solve this problem by fitting linear mixed effects (LME) model, instead of straightforward graphical techniques and tedious statistical calculations, the computations of which becomes very complicated for more than two replicated measurements (Bland and Altman, 1999).

It is worth noting that specifically in this article we propose a method to appraise the agreement between the established method and a new method, with any number of replicated observations using LME model in a doubly multivariate set-up, by properly testing the bias as well as the agreement between the repeatability coefficients of the two methods. We deem that the proposed LME model, which can handle any number of replicated measurements very easily, can serve as a surrogate, or as a substitute, to Bland and Altman’s (1999) technique of method comparison studies. By doubly multivariate set-up we mean the information in each patient is multivariate in two ways, one in the
number of methods and the other one in the number of replicated measurements. We approached the problem by using the maximum likelihood estimation where the replicate observations are linked over time. We can easily extend the method to situation where the replicate measurements are not linked. To the best of the author’s knowledge this is the first time that the hypotheses testing on the bias and the repeatability coefficient between two methods are accomplished in a formal way, with any number of replicated measurements. The model is very easy to implement using PROC MIXED of SAS and the results are straightforward too. Thus, obviates the tedious and complex statistical calculations. Since PROC MIXED can handle missing values, our method can be applied when number of replications for each patient is unequal, as well as when the number of replications for each patient by respective methods is unequal. Moreover our method can handle any number of replicated observations very easily.

Correlation coefficient-type approaches are used by many authors to study the agreement between two analytical methods. Correlation coefficient-type approaches based on a bivariate normal distribution of the data are also given in (Lin, 1989; St. Laurent, 1998; Bartko, 1994). Recently Argall et. al. (2003) used Pearson correlation coefficient to compare the two methods of weight estimation and described that the correlation coefficient 0.82 is good in comparing the two methods. Bland and Altman (1983) mentioned that since correlation cannot cope with replicated data, few studies are there involving replications. Nevertheless, there are few contemporary studies in the literature that deal with correlation coefficient with repeated measurements. Lam, Webb and O’Donnell (1999) estimated the correlation coefficient between two variables with repeated observations on each variable. Then Hamlett et al. (2003, 2004) and lately Roy (2006) estimated it by using LME model. Roy modeled the true overall correlation coefficient between the two variables by calculating it in two parts; the partial correlation coefficient (without the subject effect) between the two variables, and then added the subject effect to it. In this article we use this overall correlation coefficient along with the bias and the repeatability coefficients to compare the agreement between two methods. We maintain the value 0.82, like Argall et.al. (2003), as the edge of the overall correlation coefficient while comparing two methods, but one can always change it according to one’s requirement. We propose the following three conditions, using the three factors as mentioned previously, to verify whether two methods for measuring a quantitative variable can be considered interchangeable.

1. No significant bias, i.e. the difference between the two mean readings is not “statistically significant”.
2. High overall correlation coefficient.
3. The agreement between the two methods by testing their \textit{repeatability coefficients} (defined later).

Testing of means is normal with the mixed effects model. The output of \texttt{PROC MIXED} always gives the bias, its $t$-value, its $p$-value, and its confidence interval. It also gives the overall correlation coefficient between the two methods. Nevertheless, it is not straightforward to check the agreement of the repeatability coefficients between the two methods. We will accomplish it by the indirect use of \texttt{PROC MIXED} in two steps. We will use likelihood ratio test

$$
\Lambda = \frac{\max_{H_0} L}{\max_{H_1} L},
$$

\begin{align*}
H_o : & \text{ the two methods have the agreement with the repeatability coefficients} \\
\text{vs. } H_1 : & \text{ the two methods lack the agreement with the repeatability coefficients.}
\end{align*}

(1)

It is well known that, $L = -2 \ln \Lambda$ is approximately distributed as $\chi^2_\nu$ under $H_o$ for large sample size and under normality assumption. The degrees of freedom $\nu$ is equal to the number of parameters estimated under $H_1$ minus the number estimated under $H_o$.

\section{Linear Mixed Effects Model}

Let $p$ be the maximum number of replications for each patient or subject. For two methods we have then $2p$ maximum number of observations for each subject. We arrange these $2p$ observations by a $2p \times 1$ dimensional vector $y$ by stacking the 2 responses of the 2 methods at the first replication, then stacking 2 responses at the second replication and so on. We assume that $y$ follows a multivariate normal distribution with mean vector $\mu$ and with a positive definite variance covariance matrix $\Omega$. The $2 \times 2$ block diagonal matrix in $\Omega$ gives the covariance matrix between the 2 methods. Let $y_i$ represent the response vector for the $i$th subject, $i = 1, 2, \ldots, N$. As mentioned in the introduction the number of replicated measurements for each patient may not be equal. Suppose, for the $i$th subject each method is measured over $m_i$ times. So for subject $i$, $y_i$ is $n_i \times 1$-dimensional, $1 \leq n_i \leq 2p$, where $n_i = 2m_i$.

Let $y_{it} = (e_{it}, n_{it})'$ be a $2 \times 1$ vector of measurements on the $i$th patient at the $t$th replicate, $i = 1, 2, \ldots, N; t = 1, 2, \ldots, p$. The quantity $e$ represents the established method and $n$ the new method. Thus, $y_i = (y'_i, y'_i, \ldots, y'_i)'$. 

4
Consider a LME model as described by Laird and Ware (1982)

\[ y_i = X_i\beta + Z_i b_i + e_i, \]

\[ b_i \sim N_m(0, D), \]

\[ e_i \sim N_{n_i}(0, R_i), \]

where \( b_1, b_2, \ldots, b_N, e_1, e_2, \ldots, e_N \) are independent, and \( y_1, y_2, \ldots, y_N \) are also all independent. \( X_i \) and \( Z_i \) are \( n_i \times l \) and \( n_i \times m \) dimensional design matrices of known covariates, \( \beta \) is a \( l \)-dimensional vector containing the fixed effects, \( b_i \) is a \( m \)-dimensional vector containing the random effects, and \( e_i \) is a \( n_i \)-dimensional vector of residual components. The variance-covariance matrix \( D \) is a general \( (m \times m) \)-dimensional matrix and \( R_i \) is a \((n_i \times n_i)\)-dimensional covariance matrix which depends on \( i \) only through its dimension \( n_i \). If a patient has the maximum number of repeated measures i.e., \( n_i = 2p \), then the number of unknown parameters to be estimated in the unstructured variance-covariance matrix \( R_i \) is \( 2p(2p + 1)/2 \), otherwise the number of unknown parameters in \( R_i \) is \( n_i(n_i + 1)/2 \).

The marginal density function of \( y_i \sim N_{n_i}(X_i\beta, Z_iDZ_i' + R_i) \), where \( R_i \) represents the partial variance covariance matrix corresponding to the \( i \)th individual. The \( 2 \times 2 \) block diagonal of this gives the partial variance covariance matrix of the 2 methods. We assume \( R_i = \text{dim}_{n_i}(V \otimes \Sigma) \), where \( V \) and \( \Sigma \) respectively are \( p \times p \) and \( 2 \times 2 \) dimensional positive definite matrices and \( \otimes \) represents the Kronecker product structure. The notation \( \text{dim}_{n_i}(V \otimes \Sigma) \), represents a \( n_i \times n_i \) dimensional submatrix obtained from a \( 2p \times 2p \) dimensional matrix \( (V \otimes \Sigma) \), by appropriately keeping the columns and rows corresponding to the \( n_i \) dimensional response vector \( y_i \). The matrix \( \Sigma = \begin{bmatrix} \sigma_e^2 & \sigma_{en} \\ \sigma_{en} & \sigma_n^2 \end{bmatrix} \), represents the partial variance covariance matrix of the established method and a new method for any replicates; where \( \sigma_e^2 \) and \( \sigma_n^2 \) are the partial variances of the established method and a new method respectively and \( \sigma_{en} \) is the partial covariance between the two methods. It is assumed that \( \Sigma \) is same for all replications. The correlation matrix \( V \) of the replicated measurements on a given method is assumed to be the same for both the methods (p. 279, Timm and Mieczkowski, 1997; p. 401, Timm, 2002). Since compound symmetry (CS) correlation structure assumes equal correlation among all the measurements, we assume that the correlation matrix \( V \) of the replicated measurements has CS correlation structure. We further improve the model by incorporating the subject effect. The number of random effects and the form of \( Z_i \) can be chosen to fit the observed \((n_i \times n_i)\) dimensional overall variance-covariance matrix for the \( i \)th individual as

\[
\text{Cov}(y_i) = \Omega_i = Z_i D Z_i' + \text{dim}_{n_i} \left( V \otimes \begin{bmatrix} \sigma_e^2 & \sigma_{en} \\ \sigma_{en} & \sigma_n^2 \end{bmatrix} \right).
\]
Thus, the covariance matrix have the same structure for each subject, except that of the dimension. The $2 \times 2$ block diagonals in the estimated residual overall variance-covariance matrix $\Omega_i$ gives the overall variance-covariance matrix of the 2 methods.

3 PROC MIXED of SAS

We use PROC MIXED of SAS to get the maximum likelihood estimates of $\beta, D, R_i$ and $\Omega_i$. RANDOM and REPEATED statements specify the structure of the covariance matrices $D$ and $R_i$. The advantage of PROC MIXED is that it can handle the separable covariance structure of the variance covariance matrix $R_i = \text{dim}_i (V \otimes \Sigma)$, and it can calculate a $n_i \times n_i$ dimensional submatrix $R_i$, from a $2p \times 2p$ dimensional matrix $V \otimes \Sigma$, and eventually calculates $n_i \times n_i$ dimensional $\Omega_i$. At present, PROC MIXED can only have option $\Sigma$ as unstructured and $V$ as unstructured, AR(1) or CS structure. METHOD=ML specifies PROC MIXED to calculate the maximum likelihood estimates of the parameters. REML is the default method of SAS; which offers non-biased REML estimates of the covariance parameters. CLASS statement specifies the categorical variables. DDFM=KR specifies the Kenward-Roger (1997) correction for computing the denominator degrees of freedom for the fixed effects. Kenward-Roger correction is suggested whenever one has replicated or repeated measures data as well as for missing data. Options $V$ and VCORR in the RANDOM statement prints the estimate of $\Omega$ variance covariance matrix and the corresponding $\Omega$ correlation matrix for the first subject. The $2 \times 2$ block diagonal in the $\Omega$ correlation matrix gives the overall correlation matrix between the two methods. When the correlation matrix $V$ on the repeated measures has CS structure and $\Sigma$ is unstructured, we can either use TYPE= UN @ CS along with SUBJECT=PATIENT option or use TYPE= UN along with SUBJECT=REPlicate(PATIENT) option in the REPEATED statement. We will use the second option in this article. The only disadvantage with this is that it does not give the whole $n_i \times n_i$ dimensional $R_i$ matrix, but only the $2 \times 2$ block diagonal matrix $\Sigma$. We only need this information to calculate the repeatability coefficients for the two methods. Options $R$ and RCORR in the REPEATED statement prints the estimate of $R$ variance covariance matrix and the corresponding $R$ correlation matrix for the first subject. One can get the $\Omega$ variance covariance matrix and the corresponding $\Omega$ correlation matrix for all patients by specifying $V=1$ to $N$, and VCORR=1 to N in the RANDOM statement. For detail information one must see SAS/STAT User’s Guide (Version 9, 2004). Since PROC MIXED can handle covariates, our model can easily see its effect, especially categorical to substantiate its effect on the method assessment.
4 Repeatability Coefficient and Related Hypothesis Testing

Following Bland and Altman (1999) we name $1.96\sqrt{\frac{2}{\sigma_e}}$ as the repeatability coefficient of the established method, where $\sigma_e^2$ is the partial variance of the established method as defined earlier. Similarly, the repeatability coefficient of the new method. For 95% of subjects two replicated measurements by the same method will be within this repeatability coefficient.

As mentioned in the introduction to test the agreement between the two methods it is crucial to test the equality of their repeatability coefficients. We will accomplish this simply by testing the following hypothesis:

$$H_0 : \sigma_e^2 = \sigma_n^2 \text{ vs. } H_1 : \sigma_e^2 \neq \sigma_n^2.$$  

We apply the likelihood ratio test for this hypothesis testing. To compute the test statistic $-2 \ln \Lambda$, where

$$-2 \ln \Lambda = \left[ -2 \ln \max_{H_0} L \right] - \left[ -2 \ln \max_{H_1} L \right],$$

the likelihood function under both null hypothesis and alternating hypothesis must be maximized separately. We do this by setting the option METHOD=ML in PROC MIXED statement. The options TYPE=UN and TYPE=CS along with SUBJECT = REPLICATE(PATIENT) in the REPEATED statement are used to calculate the “-2 Log Likelihood” for the covariance structure under $H_1$ and $H_0$ respectively. PROC MIXED calculates this under the heading of goodness of fit statistics. Since $\Sigma$ is 2 x 2 dimensional, one can also use TYPE=AR(1) or TOEP along with SUBJECT=REPLICATE(PATIENT) in the REPEATED statement to calculate the “-2 Log Likelihood” for the covariance structure under $H_0$. The above test statistic $-2 \ln \Lambda$ under $H_0$ follows a chi-square distribution with degrees of freedom $\nu$, where $\nu$ is computed as

$$\nu = \text{LRT df (under } H_1) - \text{LRT df (under } H_0).$$

5 Some Examples

We demonstrate the proposed method by considering three real data sets. All the data sets are taken from different papers of Bland and Altman (1986, 1999). The first and the second data sets are of smaller in sizes, whereas the third one is larger. The first data set has unbalanced replications while the second and the third data sets have balanced replications.
Example 1. (Cardiac Data): This data set is taken from Bland and Altman (1986). This data set (Table 1) has measurements of cardiac output by two methods, radionuclide ventriculography (RV) and impedance cardiography (IC), on 12 patients. The number of repeated observations differs by patient. Such data may occur if patients are measured at regular intervals during surgery.

Table 1  Repeated measurements of Cardiac output by two methods RV and IC for 12 patients.

<table>
<thead>
<tr>
<th>Patient #</th>
<th>RV</th>
<th>IC</th>
<th>Patient #</th>
<th>RV</th>
<th>IC</th>
<th>Patient #</th>
<th>RV</th>
<th>IC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7.83</td>
<td>6.57</td>
<td>5</td>
<td>3.13</td>
<td>3.03</td>
<td>9</td>
<td>4.48</td>
<td>3.17</td>
</tr>
<tr>
<td>1</td>
<td>7.42</td>
<td>5.62</td>
<td>5</td>
<td>2.98</td>
<td>2.86</td>
<td>9</td>
<td>4.92</td>
<td>3.12</td>
</tr>
<tr>
<td>1</td>
<td>7.89</td>
<td>6.90</td>
<td>5</td>
<td>2.85</td>
<td>2.77</td>
<td>9</td>
<td>3.97</td>
<td>2.96</td>
</tr>
<tr>
<td>1</td>
<td>7.12</td>
<td>6.57</td>
<td>5</td>
<td>3.17</td>
<td>2.46</td>
<td>10</td>
<td>4.22</td>
<td>4.35</td>
</tr>
<tr>
<td>1</td>
<td>7.88</td>
<td>6.35</td>
<td>5</td>
<td>3.09</td>
<td>2.32</td>
<td>10</td>
<td>4.65</td>
<td>4.62</td>
</tr>
<tr>
<td>2</td>
<td>6.16</td>
<td>4.06</td>
<td>5</td>
<td>3.12</td>
<td>2.43</td>
<td>10</td>
<td>4.74</td>
<td>3.16</td>
</tr>
<tr>
<td>2</td>
<td>7.26</td>
<td>4.29</td>
<td>6</td>
<td>5.92</td>
<td>5.90</td>
<td>10</td>
<td>4.44</td>
<td>3.53</td>
</tr>
<tr>
<td>2</td>
<td>6.71</td>
<td>4.26</td>
<td>6</td>
<td>6.42</td>
<td>5.81</td>
<td>10</td>
<td>4.50</td>
<td>3.53</td>
</tr>
<tr>
<td>2</td>
<td>6.54</td>
<td>4.09</td>
<td>6</td>
<td>5.92</td>
<td>5.70</td>
<td>11</td>
<td>6.78</td>
<td>7.20</td>
</tr>
<tr>
<td>3</td>
<td>4.75</td>
<td>4.71</td>
<td>6</td>
<td>6.27</td>
<td>5.76</td>
<td>11</td>
<td>6.07</td>
<td>6.09</td>
</tr>
<tr>
<td>3</td>
<td>5.24</td>
<td>5.50</td>
<td>7</td>
<td>7.13</td>
<td>5.09</td>
<td>11</td>
<td>6.52</td>
<td>7.00</td>
</tr>
<tr>
<td>3</td>
<td>4.86</td>
<td>5.08</td>
<td>7</td>
<td>6.62</td>
<td>4.63</td>
<td>11</td>
<td>6.42</td>
<td>7.10</td>
</tr>
<tr>
<td>3</td>
<td>4.78</td>
<td>5.02</td>
<td>7</td>
<td>6.58</td>
<td>4.61</td>
<td>11</td>
<td>6.41</td>
<td>7.40</td>
</tr>
<tr>
<td>3</td>
<td>6.05</td>
<td>6.01</td>
<td>7</td>
<td>6.93</td>
<td>5.09</td>
<td>11</td>
<td>5.76</td>
<td>6.80</td>
</tr>
<tr>
<td>3</td>
<td>5.42</td>
<td>5.67</td>
<td>8</td>
<td>4.54</td>
<td>4.72</td>
<td>12</td>
<td>5.06</td>
<td>4.50</td>
</tr>
<tr>
<td>4</td>
<td>4.21</td>
<td>4.14</td>
<td>8</td>
<td>4.81</td>
<td>4.61</td>
<td>12</td>
<td>4.72</td>
<td>4.20</td>
</tr>
<tr>
<td>4</td>
<td>3.61</td>
<td>4.20</td>
<td>8</td>
<td>5.11</td>
<td>4.36</td>
<td>12</td>
<td>4.90</td>
<td>3.80</td>
</tr>
<tr>
<td>4</td>
<td>3.72</td>
<td>4.61</td>
<td>8</td>
<td>5.29</td>
<td>4.20</td>
<td>12</td>
<td>4.80</td>
<td>3.80</td>
</tr>
<tr>
<td>4</td>
<td>3.87</td>
<td>4.68</td>
<td>8</td>
<td>5.39</td>
<td>4.36</td>
<td>12</td>
<td>4.90</td>
<td>4.20</td>
</tr>
<tr>
<td>4</td>
<td>3.92</td>
<td>5.04</td>
<td>8</td>
<td>5.57</td>
<td>4.20</td>
<td>12</td>
<td>5.10</td>
<td>4.50</td>
</tr>
</tbody>
</table>

Table 2  Regression results for the variables RV and IC with CS correlation structure on $V$.

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>DF</th>
<th>t-value</th>
<th>Pr &gt;</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>4.68</td>
<td>0.35</td>
<td>12</td>
<td>13.34</td>
<td>&lt;0.001</td>
<td>3.92</td>
<td>5.45</td>
</tr>
<tr>
<td>RV</td>
<td>0.70</td>
<td>0.26</td>
<td>11.9</td>
<td>2.67</td>
<td>0.020</td>
<td>0.12</td>
<td>1.28</td>
</tr>
<tr>
<td>IC</td>
<td>0</td>
<td>0.26</td>
<td>11.9</td>
<td>2.67</td>
<td>0.020</td>
<td>0.12</td>
<td>1.28</td>
</tr>
</tbody>
</table>

The regression results from the output of PROC MIXED is given in Table 2. Since the data set has an imbalanced number of observations per patient, the overall variance-covariance matrix $\Omega$ for each patient will have different dimensions. Like, the patient 1 will have $10 \times 10$ dimensional variance-covariance matrix, as there are 5 repetitions for patient 1.
For patient 2, it will be $8 \times 8$ dimensional, as patient 2 has 4 repetitions. The $2 \times 2$ block diagonals Block $\hat{\Omega}$ in the estimated residual variance-covariance matrix $\hat{\Omega}$, gives the overall variance-covariance matrix between the two methods RV and IC.

We see that the bias between the two methods RV and IC is statistically significant with $p$-value = 0.0204. The estimate of the partial residual variance-covariance matrix $\hat{\Sigma}$ at a single time point is as follows

$$
\hat{\Sigma} = \begin{bmatrix}
0.1072 & 0.0372 \\
(\text{<0.0001}) & (0.0429) \\
0.0372 & 0.1379 \\
(0.0429) & (\text{<0.0001})
\end{bmatrix}.
$$

The second row in italics gives the $p$-values to the corresponding entities. The estimates of the variances are exactly the same as obtained by Bland and Altman (1999). However, Bland and Altman did not calculate the covariance between the two methods. The estimate of the corresponding partial correlation matrix is

$$
\text{Corr} \hat{\Sigma} = \begin{bmatrix}
1.000 & 0.3056 \\
0.3056 & 1.000
\end{bmatrix}.
$$

Simple calculation depicts that the partial correlation between the two methods at a single time point is 0.3056, a poor one. The repeatability coefficient for RV method is 0.9075 and the repeatability coefficient for IC method is 1.0293. Thus, the repeatability of the method IC is little more than (13%) that of the RV method. The test statistic $-2 \ln \Lambda = (173.9) - (173.1) = 0.8$, where 173.9 and 173.1 are the values of “-2 Log Likelihood” reported by SAS for the two models under $H_0$ and $H_1$ respectively. This test statistic under $H_0$ follows a chi-square distribution with degrees of freedom $\nu$, where $\nu$ is computed as $\nu = 5 - 4 = 1$. The corresponding $p$-value = 0.3711. Therefore, the repeatabilities of RV and IC are not statistically significant. The $2 \times 2$ block diagonals Block $\hat{\Omega}$ in the estimated overall residual variance-covariance matrix $\hat{\Omega}$, $i = 1, \ldots, 12$, gives the residual overall variance-covariance matrix between the RV and the IC methods

$$
\text{Block} \hat{\Omega} = \begin{bmatrix}
1.7396 & 1.1799 \\
1.1799 & 1.5877
\end{bmatrix}.
$$

The overall correlation coefficient between the two methods is 0.7100; implying that the two methods are not highly correlated.

We see that there exists statistically significant bias between the two methods. Also the methods do not have a high correlation coefficient even though the repeatability coefficients of RV and IC are statistically insignificant; So the methods do not have total agreement. Therefore, on the basis of three conditions stated in the introduction we as a
statistician do not recommend to switch the two methods.

**Example 2. (Peak Expiratory Flow Rate Data):** This data set (Bland and Altman, 1986) compares the two methods of measuring peak expiratory flow rate (PEFR). The sample was collected from a wide range of PEFR, but was not from any defined population. Two measurements (Table 3) were made with a Wright peak flow meter (X) and two with a mini Wright peak flow meter (Y), in random order. All measurements were taken using the same two instruments. The regression results from the output of *PROC MIXED* is given in Table 4.

**Table 3** PEFR measured with Wright peak flow and mini Wright peak flow meter

<table>
<thead>
<tr>
<th>Subject</th>
<th>First PEFR (1/min)</th>
<th>Second PEFR (1/min)</th>
<th>First PEFR (1/min)</th>
<th>Second PEFR (1/min)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>494</td>
<td>490</td>
<td>512</td>
<td>525</td>
</tr>
<tr>
<td>2</td>
<td>395</td>
<td>397</td>
<td>430</td>
<td>415</td>
</tr>
<tr>
<td>3</td>
<td>516</td>
<td>512</td>
<td>520</td>
<td>508</td>
</tr>
<tr>
<td>4</td>
<td>434</td>
<td>401</td>
<td>428</td>
<td>444</td>
</tr>
<tr>
<td>5</td>
<td>476</td>
<td>470</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>6</td>
<td>557</td>
<td>611</td>
<td>600</td>
<td>625</td>
</tr>
<tr>
<td>7</td>
<td>413</td>
<td>415</td>
<td>364</td>
<td>460</td>
</tr>
<tr>
<td>8</td>
<td>442</td>
<td>431</td>
<td>380</td>
<td>390</td>
</tr>
<tr>
<td>9</td>
<td>650</td>
<td>638</td>
<td>658</td>
<td>642</td>
</tr>
<tr>
<td>10</td>
<td>433</td>
<td>429</td>
<td>445</td>
<td>432</td>
</tr>
<tr>
<td>11</td>
<td>417</td>
<td>420</td>
<td>432</td>
<td>420</td>
</tr>
<tr>
<td>12</td>
<td>656</td>
<td>633</td>
<td>626</td>
<td>605</td>
</tr>
<tr>
<td>13</td>
<td>267</td>
<td>275</td>
<td>260</td>
<td>227</td>
</tr>
<tr>
<td>14</td>
<td>478</td>
<td>492</td>
<td>477</td>
<td>467</td>
</tr>
<tr>
<td>15</td>
<td>178</td>
<td>165</td>
<td>259</td>
<td>268</td>
</tr>
<tr>
<td>16</td>
<td>423</td>
<td>372</td>
<td>350</td>
<td>370</td>
</tr>
<tr>
<td>17</td>
<td>427</td>
<td>421</td>
<td>451</td>
<td>443</td>
</tr>
</tbody>
</table>

**Table 4** Regression results for the PEFR Measurements with Wright peak flow meter and Mini Wright peak flow meter

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>DF</th>
<th>t-value</th>
<th>Pr &gt;</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>453.91</td>
<td>26.1862</td>
<td>17</td>
<td>17.33</td>
<td>&lt;0.0001</td>
<td>398.66</td>
<td>509.16</td>
</tr>
<tr>
<td>Wright peak flow meter</td>
<td>-6.0294</td>
<td>7.8127</td>
<td>17</td>
<td>-0.77</td>
<td>0.4509</td>
<td>-22.5128</td>
<td>10.4540</td>
</tr>
<tr>
<td>Mini Wright peak flow meter</td>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
We see that there is a non-significant \((p-\text{values} = 0.4509)\) bias \(-6.0294\, \text{min}^{-1}\) between the two methods. The estimate of the residual partial variance-covariance matrix \(\Sigma\) for any single replication is given by

\[
\hat{\Sigma} = \begin{bmatrix}
234.29 & 2.0000 \\
(0.0018) & (0.9784) \\
2.0000 & 396.44 \\
(0.9784) & (0.0018)
\end{bmatrix}.
\]

Therefore the partial correlation coefficient between the two meters is 0.0066. As before the quantities in the parentheses gives the \(p-\)values of the corresponding entries. The coefficient of repeatability for the larger Wright peak flow meter is \(42.4275\, \text{min}^{-1}\), and the coefficient of repeatability for the mini Wright peak flow meter is \(55.1899\, \text{min}^{-1}\). Therefore repeatability of the Mini Wright peak flow meter is 30\% more than the repeatability of the larger Wright peak flow meter.

To test the hypothesis of the equality of repeatabilities of these peak flow meters, we calculate the value of the test statistic \(-2\ln \Lambda = (689.4) - (688.2) = 1.2\), where 689.4 and 688.2 are the values of “-2 Log Likelihood” reported by SAS for the two models under \(H_0\) and \(H_1\) respectively. The above test statistic under \(H_0\), follows a chi-square distribution with degrees of freedom \(\nu\), where \(\nu\) is computed as \(\nu = 5 - 4 = 1\). The corresponding \(p-\)value = 0.2733. Therefore, the repeatabilities of the two flow meters are statistically insignificant.

The \(2 \times 2\) block diagonals Block \(\hat{\Omega}\) in the estimated residual overall variance-covariance matrix \(\Omega\) is as follows

\[
\text{Block} \hat{\Omega} = \begin{bmatrix}
13105 & 11805 \\
11805 & 11855
\end{bmatrix}.
\]

The overall correlation coefficient between the two methods is 0.9471. Therefore the two flow meters do not have significant bias, and they have high correlation and the repeatabilities of the two flow meters are statistically insignificant.

Therefore on the basis of three conditions stated in the introduction our statistical recommendation is that one can use the two meters interchangeably.

**Example 3. (Systolic Blood Pressure Data):** This data set is also taken from Bland and Altman (1999). Simultaneous measurements of systolic blood pressure were made by each of the two experienced observers (denoted J and R) using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted by S). Three sets of readings were made in quick succession on 85 subjects. We want to examine whether either of the two observer can be replaced by the semi-automatic blood pressure monitor. To see this we
first analyze the data by taking the observer J and the machine S, and then by taking the observer R and the machine S. The regression results from the output of `PROC MIXED` are given in Tables 5 and 6 respectively.

**Table 5** Regression results for the observer J and an automatic blood pressure machine S

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>DF</th>
<th>t-value</th>
<th>Pr &gt;</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>143.03</td>
<td>3.4283</td>
<td>85</td>
<td>41.72</td>
<td>&lt;0.0001</td>
<td>136.21</td>
<td>149.84</td>
</tr>
<tr>
<td>J</td>
<td>-15.6196</td>
<td>2.0416</td>
<td>85</td>
<td>-7.65</td>
<td>&lt;0.0001</td>
<td>-19.6788</td>
<td>-11.5605</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>.</td>
<td>85</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

**Table 6** Regression results for the observer R and an automatic blood pressure machine S

<table>
<thead>
<tr>
<th>Effect</th>
<th>Estimate</th>
<th>SE</th>
<th>DF</th>
<th>t-value</th>
<th>Pr &gt;</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>143.03</td>
<td>3.4283</td>
<td>85</td>
<td>41.72</td>
<td>&lt;0.0001</td>
<td>136.21</td>
<td>149.84</td>
</tr>
<tr>
<td>R</td>
<td>-15.7059</td>
<td>2.0263</td>
<td>85</td>
<td>-7.75</td>
<td>&lt;0.0001</td>
<td>-19.7348</td>
<td>-11.6770</td>
</tr>
<tr>
<td>S</td>
<td>0</td>
<td>.</td>
<td>85</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

We see that the mean difference $-15.6196$ mmHg between the observer J and the automatic blood pressure machine S is statistically significant with $p$--value <0.0001.

The estimate of the residual partial variance-covariance matrix $\Sigma$ for any single replication is given by

$$\Sigma = \begin{bmatrix} 37.4078 & 16.0627 \\ \left( <0.0001 \right) & \left( 0.0003 \right) \\ 16.0627 & 83.1412 \\ \left( 0.0003 \right) & \left( <0.0001 \right) \end{bmatrix}.$$ 

The estimates of the variances are exactly the same as obtained by Bland and Altman (1999). All the entries in this matrix are statistically significant. The corresponding $p$--values are given in the parentheses. The coefficient of repeatabilities for the observers J and S are 16.9532 mmHg and 25.2743 mmHg respectively. Therefore repeatability of the machine S is 49% more than the repeatability of the observer J. To test the equality of these two repeatabilities we calculate the test statistic $-2 \ln \Lambda = (4090.1) - (4061.5) = 28.6$, where 4090.1 and 4061.5 are the values of “-2 Log Likelihood” reported by SAS for the two models under $H_0$ and $H_1$ respectively. As before the test statistic under $H_0$ follows a chi-square distribution with 1 degree of freedom. The corresponding $p$--value $= 8.8982E-8$, a negligible quantity. Therefore, the repeatabilities of the observer J and the machine S are statistically significant. Simple calculation depicts that the partial correlation coefficient between the observer J and the machine S is 0.2880. The $2 \times 2$ block diagonals $\mathbf{Block} \hat{\Omega}$
in the estimated residual overall variance-covariance matrix $\Omega$ gives the overall variance-covariance matrix between the two observers and the automatic machine.

$$\text{Block} \hat{\Omega} = \begin{bmatrix} 961.39 & 801.31 \\ 801.31 & 1054.44 \end{bmatrix}.$$  

Therefore, the overall correlation coefficient between the observer J and the machine S is 0.7959.

From Table 6 we see that the bias $-15.7059 \text{mmHg}$ between the observer R and the machine S is statistically significant with $p$-value < 0.0001. The residual $\Sigma$ variance-covariance matrix is as follows

$$\Sigma = \begin{bmatrix} 37.9804 & 17.3333 \\ (0.0001) & (0.0001) \end{bmatrix},$$

$$\begin{bmatrix} 17.1412 & 83.1412 \\ (0.0003) & (0.0001) \end{bmatrix}.$$  

Here also the estimates of the variances are exactly the same as obtained by Bland and Altman (1999). But as before they did not calculate the covariance between the observer R and the machine S. The partial correlation coefficient between the observer R and the machine S is 0.3085. The coefficient of repeatabilities for the observer R and the machine S are 17.0825 mmHg and 25.2743 mmHg respectively. Therefore the repeatability of the machine S is 48% more than the repeatability of the observer R. As before to test the equality of these two repeatabilities we calculate the test statistic $-2 \ln \Lambda = (4087.5) - (4059.6) = 27.9$, where 4087.5 and 4059.6 are the values of “-2 Log Likelihood” reported by SAS for the two models under $H_o$ and $H_1$ respectively. This test statistic under $H_o$ follows a chi-square distribution with 1 degree of freedom. The corresponding $p$-value = 1.2775 $E - 7$, a negligible quantity. Therefore, the repeatabilities of the observer R and the machine S are statistically significant. The $2 \times 2$ block diagonals Block $\hat{\Omega}$ in the estimated residual overall variance-covariance matrix $\Omega$ gives the overall variance-covariance matrix between the observer R and the automatic machine.

$$\text{Block} \hat{\Omega} = \begin{bmatrix} 944.11 & 795.95 \\ 795.95 & 1054.44 \end{bmatrix}.$$  

Therefore the overall correlation coefficient between the observer J and the machine S is 0.7977.

So the biases between each of the two observers J and R, and the machine S are statistically significant. Also, they do not exhibit high correlation coefficient ($\geq 0.82$). Furthermore, the repeatabilities of each of the observer and the machine S are statistically significant. Therefore, on the basis of three conditions stated in the introduction we do not recommend to substitute any of the observer with the machine.
6 Conclusions

In this article we present a new method using the LME model to assess the agreement between a new method and an established method with any number of replicated observations. The topic is of practical relevance in many practical fields, especially in medical and biomedical sciences. The method is easily understandable by either a statistician or a non-statistician, and is very easy to implement using \textit{PROC MIXED} of SAS. The interpretation of the results is also straightforward. A few lines of computer program can be used by any person with little bit of programming expertise. The power of the likelihood ratio test mentioned in this paper may depend on specific sample size and specific number of replicated observations. One needs to do some simulation study for this. We will report it in a future correspondence.

Acknowledgements

The author would like to acknowledge the generous support for the summer grant from the College of Business at the University of Texas at San Antonio.

References


