THE UNIVERSITY OF TEXAS AT SAN ANTONIO, COLLEGE OF BUSINESS

Working Paper SERIES

February 13, 2007

WP # 0008MSS-253-2007

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

Copyright ©2006 by the UTSA College of Busines. All rights reserved. This document can be downloaded without charge for educational purposes from the UTSA College of Business Working Paper Series (business.utsa.edu/wp) without explicit permission, provided that full credit, including © notice, is given to the source. The views expressed are those of the individual author(s) and do not necessarily reflect official positions of UTSA, the College of Business, or any individual department.

UTSA COLLEGE OF BUSINESS KNOWLEDGE FOR A NEW WORLD

ONE UTSA CIRCLE
SAN ANTONIO, TEXAS 78249-0631
210 458-4317 | BUSINESS.UTSA.EDU

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 *repeatability coefficients* (defined later).

Testing of means is normal with the mixed effects model. The output of *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 *PROC MIXED* in two steps. We will use likelihood ratio test

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

to test the null hypothesis:

$$\begin{aligned} & H_o : \text{the two methods have the agreement with the repeatability coefficients} \\ \text{vs.} \quad & H_1 : \text{the two methods lack the agreement with the repeatability coefficients.} \end{aligned} \tag{1}$$

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

2 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 $\boldsymbol{y}$ 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 $\boldsymbol{y}$ follows a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and with a positive definite variance covariance matrix $\boldsymbol{\Omega}$. The 2×2 block diagonal matrix in $\boldsymbol{\Omega}$ gives the covariance matrix between the 2 methods. Let $\boldsymbol{y}_i$ represent the response vector for the ith 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 ith subject each method is measured over m_i times. So for subject i, $\boldsymbol{y}_i$ is $n_i \times 1$-dimensional, $1 \leq n_i \leq 2p$, where $n_i = 2m_i$.

Let $\boldsymbol{y}_{it} = (e_{it}, n_{it})'$ be a 2×1 vector of measurements on the ith patient at the tth replicate, $i = 1, 2, \ldots, N; t = 1, 2, \ldots, p$. The quantity e represents the established method and n the new method. Thus, $\boldsymbol{y}_i = (\boldsymbol{y}'_{i1}, \boldsymbol{y}'_{i2}, \ldots, \boldsymbol{y}'_{ip})'$.

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

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

$$\boldsymbol{b}_i \sim N_m(\boldsymbol{0}, \boldsymbol{D}),$$

$$\boldsymbol{\epsilon}_i \sim N_{n_i}(\boldsymbol{0}, \boldsymbol{R}_i),$$

where $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_N, \boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, \ldots, \boldsymbol{\epsilon}_N$ are independent, and $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N$ are also all independent. $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$ are $n_i \times l$ and $n_i \times m$ dimensional design matrices of known covariates, $\boldsymbol{\beta}$ is a l-dimensional vector containing the fixed effects, $\boldsymbol{b}_i$ is a m-dimensional vector containing the random effects, and $\boldsymbol{\epsilon}_i$ is a n_i-dimensional vector of residual components. The variance-covariance matrix $\boldsymbol{D}$ is a general $(m \times m)$-dimensional matrix and $\boldsymbol{R}_i$ 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 $\boldsymbol{R}_i$ is $2p(2p+1)/2$, otherwise the number of unknown parameters in $\boldsymbol{R}_i$ is $n_i(n_i+1)/2$.

The marginal density function of $\boldsymbol{y}_i \sim N_{n_i}(\boldsymbol{X_i\beta}, \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{R}_i)$, where $\boldsymbol{R}_i$ represents the partial variance covariance matrix corresponding to the ith individual. The 2×2 block diagonal of this gives the partial variance covariance matrix of the 2 methods. We assume $\boldsymbol{R}_i = \dim_{n_i}(\boldsymbol{V} \otimes \boldsymbol{\Sigma})$, where $\boldsymbol{V}$ and $\boldsymbol{\Sigma}$ respectively are $p \times p$ and 2×2 dimensional positive definite matrices and $\otimes$ represents the Kronecker product structure. The notation $\dim_{n_i}(\boldsymbol{V} \otimes \boldsymbol{\Sigma})$, represents a $n_i \times n_i$ dimensional submatrix obtained from a $2p \times 2p$ dimensional matrix $(\boldsymbol{V} \otimes \boldsymbol{\Sigma})$, by appropriately keeping the columns and rows corresponding to the n_i dimensional response vector $\boldsymbol{y}_i$. The matrix $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_e^2 & \sigma_{en} \\ \sigma_{en} & \sigma_n^2 \end{bmatrix}$, represents the partial variance covariance matrix of the established method and a new method for any replicates; where σ_e^2 and σ_n^2 are the partial variances of the established method and a new method respectively and σ_{en} is the partial covariance between the two methods. It is assumed that $\boldsymbol{\Sigma}$ is same for all replications. The correlation matrix $\boldsymbol{V}$ 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 $\boldsymbol{V}$ 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 $\boldsymbol{Z}_i$ can be chosen to fit the observed $(n_i \times n_i)$ dimensional overall variance-covariance matrix for the ith individual as

$$\mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{\Omega}_i = \boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{Z}_i' + \dim_{n_i}\left(\boldsymbol{V} \otimes \begin{bmatrix} \sigma_e^2 & \sigma_{en} \\ \sigma_{en} & \sigma_n^2 \end{bmatrix}\right).$$

Thus, the covariance matrix have the same structure for each subject, except that of the dimension. The 2×2 block diagonals in the estimated residual overall variance-covariance matrix $\boldsymbol{\Omega}_i$ 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 $\boldsymbol{\beta}, \boldsymbol{D}$, $\boldsymbol{R}_i$ and $\boldsymbol{\Omega}_i$. RANDOM and REPEATED statements specify the structure of the covariance matrices $\boldsymbol{D}$ and $\boldsymbol{R}_i$. The advantage of *PROC MIXED* is that it can handle the separable covariance structure of the variance covariance matrix $\boldsymbol{R}_i = \dim_{n_i}(\boldsymbol{V} \otimes \boldsymbol{\Sigma})$, and it can calculate a $n_i \times n_i$ dimensional submatrix $\boldsymbol{R}_i$, from a $2p \times 2p$ dimensional matrix $\boldsymbol{V} \otimes \boldsymbol{\Sigma}$, and eventually calculates $n_i \times n_i$ dimensional $\boldsymbol{\Omega}_i$. At present, *PROC MIXED* can only have option $\boldsymbol{\Sigma}$ as unstructured and $\boldsymbol{V}$ 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 $\boldsymbol{\Omega}$ variance covariance matrix and the corresponding $\boldsymbol{\Omega}$ correlation matrix for the first subject. The 2×2 block diagonal in the $\boldsymbol{\Omega}$ correlation matrix gives the overall correlation matrix between the two methods. When the correlation matrix $\boldsymbol{V}$ on the repeated measures has CS structure and $\boldsymbol{\Sigma}$ 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 $\boldsymbol{R}_i$ matrix, but only the 2×2 block diagonal matrix $\boldsymbol{\Sigma}$. 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 $\boldsymbol{R}$ variance covariance matrix and the corresponding $\boldsymbol{R}$ correlation matrix for the first subject. One can get the $\boldsymbol{\Omega}$ variance covariance matrix and the corresponding $\boldsymbol{\Omega}$ 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 $1.96\sqrt{2}\,\sigma_e$ as the *repeatability coefficient* of the established method, where σ_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:

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

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

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

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_o respectively. *PROC MIXED* calculates this under the heading of goodness of fit statistics. Since $\mathbf{\Sigma}$ is 2×2 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_o. The above test statistic $-2\ln\Lambda$ under H_o follows a chi-square distribution with degrees of freedom ν, where ν is computed as

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

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.

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

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

Effect	Estimate	SE	DF	t-value	Pr > |t|	Lower	Upper
Intercept	4.6836	0.3510	12	13.34	<0.0001	3.9189	5.4484
RV	0.7040	0.2634	11.9	2.67	0.0204	0.1298	1.2782
IC	0	.	.	.	.	.	.

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 $\boldsymbol{\Omega}_i$ for each patient will have different dimensions. Like, the patient 1 will have 10×10 dimensional variance-covariance matrix, as there are 5 repetitions for patient 1.

For patient 2, it will be 8×8 dimensional, as patient 2 has 4 repetitions. The 2×2 block diagonals Block $\hat{\mathbf{\Omega}}$ in the estimated residual variance-covariance matrix $\mathbf{\Omega}_i$ 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 $\mathbf{\Sigma}$ at a single time point is as follows

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

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

$$\text{Corr}\,\hat{\mathbf{\Sigma}} = \begin{bmatrix} 1.000 & 0.3056 \\ 0.3056 & 1.0000 \end{bmatrix}.$$

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_o and H_1 respectively. This test statistic under H_o follows a chi-square distribution with degrees of freedom ν, where ν 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 2×2 block diagonals Block $\hat{\mathbf{\Omega}}$ in the estimated overall residual variance-covariance matrix $\mathbf{\Omega_i}$, $i = 1, \ldots 12$, gives the residual overall variance-covariance matrix between the RV and the IC methods

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

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

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

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

Effect	Estimate	SE	DF	t-value	Pr > |t|	Lower	Upper
Intercept	453.91	26.1862	17	17.33	<0.0001	398.66	509.16
Wright peak flow meter	-6.0294	7.8127	17	-0.77	0.4509	-22.5128	10.4540
Mini Wright peak flow meter	0	.	.	.	.	.	.

We see that there is a non-significant ($p-$values $= 0.4509$) bias $-6.0294\,\text{min}^{-1}$ between the two methods. The estimate of the residual partial variance-covariance matrix $\boldsymbol{\Sigma}$ for any single replication is given by

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} 234.29 & 2.0000 \\ \textit{(0.0018)} & \textit{(0.9784)} \\ 2.0000 & 396.44 \\ \textit{(0.9784)} & \textit{(0.0018)} \end{bmatrix}.$$

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_o and H_1 respectively. The above test statistic under H_o, follows a chi-square distribution with degrees of freedom ν, where ν 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 2×2 block diagonals Block $\hat{\boldsymbol{\Omega}}$ in the estimated residual overall variance-covariance matrix $\boldsymbol{\Omega}$ is as follows

$$\text{Block}\,\hat{\boldsymbol{\Omega}} = \begin{bmatrix} 13105 & 11805 \\ 11805 & 11855 \end{bmatrix}.$$

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

Effect	Estimate	SE	DF	t-value	Pr > |t|	Lower	Upper
Intercept	143.03	3.4283	85	41.72	<0.0001	136.21	149.84
J	-15.6196	2.0416	85	-7.65	<0.0001	-19.6788	-11.5605
S	0	.	.	.	.	.	.

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

Effect	Estimate	SE	DF	t-value	Pr > |t|	Lower	Upper
Intercept	143.03	3.4283	85	41.72	<0.0001	136.21	149.84
R	-15.7059	2.0263	85	-7.75	<0.0001	-19.7348	-11.6770
S	0	.	.	.	.	.	.

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 $\mathbf{\Sigma}$ for any single replication is given by

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

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_o and H_1 respectively. As before the test statistic under H_o 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 2×2 block diagonals Block $\hat{\mathbf{\Omega}}$

in the estimated residual overall variance-covariance matrix $\mathbf{\Omega}$ gives the overall variance-covariance matrix between the two observers and the automatic machine.

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

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 $\mathbf{\Sigma}$ variance-covariance matrix is as follows

$$\hat{\mathbf{\Sigma}} = \begin{bmatrix} 37.9804 & 17.3333 \\ \textit{(<0.0001)} & \textit{(0.0001)} \\ 17.1412 & 83.1412 \\ \textit{(0.0003)} & \textit{(<0.0001)} \end{bmatrix}.$$

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 $17.0825\,\text{mmHg}$ and $25.2743\,\text{mmHg}$ 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 $= 1.2775E-7$, a negligible quantity. Therefore, the repeatabilities of the observer R and the machine S are statistically significant. The 2×2 block diagonals $\text{Block}\,\hat{\mathbf{\Omega}}$ in the estimated residual overall variance-covariance matrix $\mathbf{\Omega}$ gives the overall variance-covariance matrix between the observer R and the automatic machine.

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

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 (≥ 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 *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

[1] Argall, J. A. W., Wright, N., Mackway-Jones, K. and Jackson R. (2003). A comparison of two commonly used methods of weight estimation, *Archives of Disease in Childhood*, 88, 789-790

[2] Bartko, J. J. (1994). General methodology II. Measures of agreement: A single procedure. *Stat. Med.*, 13, 737745.

[3] Bland J. M. and Altman D. G. (1983). Measurement in Medicine: the Analysis of Method Comparison Studies, *Statistician*, 32, 307-17.

[4] Bland J. M. and Altman D. G. (1986). Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement, *The Lancet*, 8, 307-310.

[5] Bland J. M. and Altman D. G. (1990). A Note on the Use of the Intraclass Correlation Coefficient in the Evaluation of Agreement Between Two Methods of Measurement, *Comput. Biol. Med.* 20(5), 337-340.

[6] Bland J. M. and Altman D. G. (1999). Measuring Agreement in Method Comparison Studies, *Statistical Methods in Medical Research*, 8, 135-160.

[7] Choudhary P. K. and Nagaraja H. N. (2005). Assessment of Agreement Using Intersection-Union Principle. *Biometrical Journal* 47(5), 674-681.

[8] Hamlett A., Ryan L., Serrano-Trespalacios P., Wolfinger R. (2003). Mixed models for assessing correlation in the presence of replication, Journal of the Air & Waste Management Association 53, 442-450.

[9] Hamlett A., Ryan L., Wolfinge R. (2004). On the use of PROC MIXED to estimate correlation in the presence of repeated measures. SAS Users Group International, Proceedings of the Statistics and Data Analysis Section, Paper 198-29; 1-7.

[10] Kenward M.G., J.H. Roger, (1997). Small sample inference for fixed effects from restricted maximum likelihood, Biometrics 53, 983-997.

[11] Lam M., Webb K.A., O'Donnell D.E., (1999). Correlation between two variables in repeated measures, American Statistical Association, Proceedings of the Biometric Section 213-218.

[12] Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal Data. *Biometrics* 38, 963-974.

[13] St. Laurent R.T. (1998). Evaluating agreement with a gold standard in method comparison studies, Biometrics 54, 537-545.

[14] Lee, J., Koh, D. and Ong, C. N. (1989). Statistical Evaluation of Agreement between two Methods for Measuring a Quantitative Variable. *Comput. Biol. Med.* 19(1), 61-70.

[15] Lin, L. K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 45, 255-268.

[16] Roy A. (2006). Estimating Correlation Coefficient between Two Variables with Repeated Observations using Mixed Effects Model, Biometrical Journal 48, 286-301.

[17] SAS Institute Inc. (2004). SAS/STAT User's Guide Version 9, SAS Institute Inc., Cary, NC.

[18] Timm N.H. (2002). Applied Multivariate Analysis, New York: Springer- Verlag.

[19] Timm N.H., Mieczkowski T.A. (1997). Univariate & Multivariate General Linear Models: Theory and Applications using SAS Software, Cary, NC: SAS Institute Inc.