# Contents

# Chapter 1

# Appendices

## 1.1 Implementation in R

To implement an LME model in `R`, the `nlme` package is used. This package is loaded into the `R` environment using the library command, (i.e. `library(nlme)`). The `lme` command is used to fit LME models. The first two arguments to the `lme` function specify the fixed effect component of the model, and the data set to which the model is to be fitted. The first candidate model ('MCS1') fits an LME model on the data set 'dat'. The variable 'method' is assigned as the fixed effect, with the response variable 'BP' (i.e. blood pressure).

The third argument contain the random effects component of the formulation, describing the random effects, and their grouping structure. The `nlme` package provides a set of positive-definite matrices , the `pdMat` class, that can be used to specify a structure for the between-subject variance-covariance matrix for the random effects. For Roy's methodology, we will use the `pdSymm` and `pdCompSymm` to specify a symmetric structure and a compound symmetry structure respectively. A full discussion of these structures can be found in Pinheiro and Bates (1994, pg. 158).

Similarly a variety of structures for the with-subject variance-covariance matrix can be implemented using `nlme`. To implement a particular matrix structure, one must

specify both a variance function and correlation structure accordingly. Variance functions are used to model the variance structure of the within-subject errors. `varIdent` is a variance function object used to allow different variances according to the levels of a classification factor in the data. A compound symmetry structure is implemented using the `corCompSymm` class, while the symmetric form is specified by `corSymm` class. Finally, the estimation methods is specified as "ML" or "REML". The first of Roy's candidate model can be implemented using the following code;

```
MCS1 = lme(BP ~ method-1, data = dat,
random =  list(subject=pdSymm(~ method-1)),
weights=varIdent(form=~1|method),
correlation = corSymm(form=~1 | subject/obs), method="ML")
```

For the blood pressure data used in Roy (2009b), all four candidate models are implemented by slight variations of this piece of code, specifying either `pdSymm` or `pdCompSymm` in the second line, and either `corSymm` or `corCompSymm` in the fourth line. For example, the second candidate model 'MCS2' is implemented with the same code as MCS1, except for the term `pdCompSymm` in the second line, rather than `pdSymm`.

```
MCS2 = lme(BP ~ method-1, data = dat,
random = list(subject=pdCompSymm(~ method-1)),
weights = varIdent(form=~1|method),
correlation = corSymm(form=~1 | subject/obs), method="ML")
```

Using this `R` implementation for other data sets requires that the data set is structured appropriately (i.e. each case of observation records the index, response, method

and replicate). Once formatted properly, implementation is simply a case of re-writing the first line of code, and computing the four candidate models accordingly.

To perform a likelihood ratio test for two candidate models, simply use the `anova` command with the names of the candidate models as arguments. The following piece of code implement the first of Roy's variability tests.

```
> anova(MCS1,MCS2)
Model df    AIC    BIC  logLik   Test L.Ratio p-value
MCS1     1  8 4077.5 4111.3 -2030.7
MCS2     2  7 4075.6 4105.3 -2030.8 1 vs 2 0.15291  0.6958
>
```

The fixed effects estimates are the same for all four candidate models. The inter-method bias can be easily determined by inspecting a summary of any model. The summary presents estimates for all of the important parameters, but not the complete variance-covariance matrices (although some simple `R` functions can be written to overcome this). The variance estimates for the random effects for MCS2 is presented below.

```
Random effects:
Formula: ~method - 1 | subject
Structure: Compound Symmetry
StdDev Corr
methodJ  30.765
methodS  30.765 0.829
Residual  6.115
```

Similarly, for computing the limits of agreement the standard deviation of the differences is not explicitly given. Again, A simple R function can be written to calculate the limits of agreement directly.

## 1.2 Demonstration of Roy's testing

Roy provides three case studies, using data sets well known in method comparison studies, to demonstrate how the methodology should be used. The first two examples used are from the 'blood pressure' data set introduced by Bland and Altman (1999). The data set is a tabulation of simultaneous measurements of systolic blood pressure were made by each of two experienced observers (denoted 'J' and 'R') using a sphygmomanometer and by a semi-automatic blood pressure monitor (denoted 'S'). Three sets of readings were made in quick succession. Roy compares the 'J' and 'S' methods in the first of her examples.

The inter-method bias between the two method is found to be 15.62 , with a $t-$value of $-7.64$, with a $p-$value of less than 0.0001. Consequently there is a significant inter-method bias present between methods $J$ and $S$, and the first of the Roy's three agreement criteria is unfulfilled.

Next, the first variability test is carried out, yielding maximum likelihood estimates of the between-subject variance covariance matrix, for both the null model, in compound symmetry (CS) form, and the alternative model in symmetric (symm) form. These matrices are determined to be as follows;

$$\hat{\boldsymbol{D}}_{CS} = \begin{pmatrix} 946.50 & 784.32 \\ 784.32 & 946.50 \end{pmatrix}, \qquad \hat{\boldsymbol{D}}_{Symm} = \begin{pmatrix} 923.98 & 785.24 \\ 785.24 & 971.30 \end{pmatrix}.$$

A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the null model is $-2030.7$, and for the alternative model $-2030.8$. The test statistic, presented with greater precision than the log-likelihoods, is 0.1592. The $p-$value is 0.6958. Consequently we fail to reject the null model, and by extension,

conclude that the hypothesis that methods $J$ and $S$ have the same between-subject variability. Thus the second of the criteria is fulfilled.

The second variability test determines maximum likelihood estimates of the within-subject variance covariance matrix, for both the null model, in CS form, and the alternative model in symmetric form.

$$\hat{\mathbf{\Lambda}}_{\boldsymbol{CS}} = \begin{pmatrix} 60.27 & 16.06 \\ 16.06 & 60.27 \end{pmatrix}, \qquad \hat{\mathbf{\Lambda}}_{Symm} = \begin{pmatrix} 37.40 & 16.06 \\ 16.06 & 83.14 \end{pmatrix}.$$

Again, A likelihood ratio test is perform to compare both candidate models. The log-likelihood of the alternative model model is $-2045.0$. As before, the null model has a log-likelihood of $-2030.7$. The test statistic is computed as 28.617, again presented with greater precision. The $p-$value is less than 0.0001. In this case we reject the null hypothesis of equal within-subject variability. Consequently the third of Roy's criteria is unfulfilled. The coefficient of repeatability for methods $J$ and $S$ are found to be 16.95 mmHg and 25.28 mmHg respectively.

The last of the three variability tests is carried out to compare the overall variabilities of both methods. With the null model the MLE of the within-subject variance covariance matrix is given below. The overall variabilities for the null and alternative models, respectively, are determined to be as follows;

$$\hat{\mathbf{\Sigma}}_{CS} = \begin{pmatrix} 1007.92 & 801.65 \\ 801.65 & 1007.92 \end{pmatrix}, \qquad \hat{\mathbf{\Sigma}}_{Symm} = \begin{pmatrix} 961.38 & 801.40 \\ 801.40 & 1054.43 \end{pmatrix},$$

The log-likelihood of the alternative model model is $-2045.2$, and again, the null model has a log-likelihood of $-2030.7$. The test statistic is 28.884, and the $p-$value is less than 0.0001. The null hypothesis, that both methods have equal overall variability, is rejected. Further to the second variability test, it is known that this difference is specifically due to the difference of within-subject variabilities.

Lastly, Roy considers the overall correlation coefficient. The diagonal blocks $\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii}$ of the correlation matrix indicate an overall coefficient of 0.7959. This is less than the threshold of 0.82 that Roy recommends.

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii} = \begin{pmatrix} 1 & 0.7959 \\ 0.7959 & 1 \end{pmatrix}$$

The off-diagonal blocks of the overall correlation matrix $\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii'}$ present the correlation coefficients further to Hamlett et al. (2004).

$$\hat{\boldsymbol{r}}_{\boldsymbol{\Omega}ii'} = \begin{pmatrix} 0.9611 & 0.7799 \\ 0.7799 & 0.9212 \end{pmatrix}.$$

The overall conclusion of the procedure is that method $J$ and $S$ are not in agreement, specifically due to the within-subject variability, and the inter-method bias. The repeatability coefficients are substantially different, with the coefficient for method $S$ being 49% larger than for method $J$. Additionally the overall correlation coefficient did not exceed the recommended threshold of 0.82.

## 1.3 Worked Eamples:LikelihoodRatio Tests

Conventionally LME models can be tested using Likelihood Ratio Tests, wherein a reference model is compared to a nested model.

```
> Ref.Fit = lme(y ~ meth-1, data = dat,   #Symm , Symm#
+     random = list(item=pdSymm(~ meth-1)),
+     weights=varIdent(form=~1|meth),
+     correlation = corSymm(form=~1 | item/repl),
+     method="ML")
```

Roy(2009) presents two nested models that specify the condition of equality as required, with a third nested model for an additional test. There three formulations share the same structure, and can be specified by making slight alterations of the code for the Reference Model. Nested Model (Between-Item Variability)

```
> NMB.fit  = lme(y ~ meth-1, data = dat,   #CS , Symm#
+     random = list(item=pdCompSymm(~ meth-1)),
+     correlation = corSymm(form=~1 | item/repl),
+     method="ML")
```

```
Nested Model (Within ?item Variability)
> NMW.fit = lme(y ~ meth-1, data = dat,   #Symm , CS#
+     random = list(item=pdSymm(~ meth-1)),
+     weights=varIdent(form=~1|meth),
+     correlation = corCompSymm(form=~1 | item/repl),
+     method="ML")
```

Nested Model (Overall Variability) Additionally there is a third nested model, that can be used to test overall variability, substantively a a joint test for between-item and within-item variability. The motivation for including such a test in the suite is not clear, although it does circumvent the need for multiple comparison procedures in certain circumstances, hence providing a simplified procedure for non-statisticians.

```
> NMO.fit = lme(y ~ meth-1, data = dat,   #CS , CS#
+     random = list(item=pdCompSymm(~ meth-1)),
+     correlation = corCompSymm(form=~1 | item/repl),
+     method="ML")
```

ANOVAs for Original Fits The likelihood Ratio test is very simple to implement in R. All that is required it to specify the reference model and the relevant nested

mode as arguments to the command anova(). The figure below displays the three tests described by Roy (2009).

```
> testB    = anova(Ref.Fit,NMB.fit)                          # Between-Subject Var
> testW   = anova(Ref.Fit,NMW.fit)                         # Within-Subject Variabi
> testO     = anova(Ref.Fit,NMO.fit)                        # Overall Variabilitie
```

## 1.4 Fat Data Examples: LoAs for Carstensen's data

Carstensen et al. (2008) presents a data set 'fat', which is a comparison of measurements of subcutaneous fat by two observers at the Steno Diabetes Center, Copenhagen. Measurements are in millimeters (mm). Each person is measured three times by each observer. The observations are considered to be 'true' replicates.

A linear mixed effects model is formulated, and implementation through several software packages is demonstrated. All of the necessary terms are presented in the computer output. The limits of agreement are therefore,

$$0.0449 \pm 1.96 \times \sqrt{2 \times 0.0596^2 + 0.0772^2 + 0.0724^2} = (-0.220, 0.309). \tag{1.1}$$

Carstensen et al. (2008) describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm 2\sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}.$$

For the 'Fat' data set, the inter-method bias is shown to be 0.045. The limits of agreement are $(-0.23, 0.32)$

For Carstensen's 'fat' data, the limits of agreement computed using Roy's method are consistent with the estimates given by Carstensen et al. (2008); $0.044884 \pm 1.96 \times 0.1373979 = (-0.224, 0.314)$.

## 1.5 HBA1C Diabetes Example

Carstensen et al. (2008) describes the sampling method when discussing of a motivating example

Diabetes patients attending an outpatient clinic in Denmark have their $HbA_{1c}$ levels routinely measured at every visit. Venous and Capillary blood samples were obtained from all patients appearing at the clinic over two days. Samples were measured on four consecutive days on each machines, hence there are five analysis days.

Carstensen et al. (2008) notes that every machine was calibrated every day to the manufacturers guidelines. Measurements are classified by method, individual and replicate. In this case the replicates are clearly not exchangeable, neither within patients nor simulataneously for all patients.

## 1.6 Oximetry Data

Carstensen et al. (2008) introduces a second data set; the oximetry study. This study done at the Royal Childrens Hospital in Melbourne to assess the agreement between co-oximetry and pulse oximetry in small babies.

In most cases, measurements were taken by both method at three different times. In some cases there are either one or two pairs of measurements, hence the data is unbalanced. Carstensen et al. (2008) describes many of the children as being very sick, and with very low oxygen saturations levels. Therefore it must be assumed that a biological change can occur in interim periods, and measurements are not true replicates.

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods.

Limits of agreement are determined using Roy's methodology, without adding any additional terms, are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$. Roy's methodology assumes that replicates are linked. However, following Carstensen's example, an addition interaction term is added to the implementation of Roy's model to assess the effect, the limits of agreement estimates do not change. However there is a conspicuous difference in within-subject matrices of Roy's model and the modified model (denoted 1 and 2 respectively);

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the 'Oximetry' data set. When the interaction term is omitted,

the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

Limits of agreement are determined using Roy's methodology, without adding any additional terms, are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$. Roy's methodology assumes that replicates are linked. However, following Carstensen's example, an addition interaction term is added to the implementation of Roy's model to assess the effect, the limits of agreement estimates do not change. However there is a conspicuous difference in within-subject matrices of Roy's model and the modified model (denoted 1 and 2 respectively);
(The variance of the additional random effect in model 2 is 3.01.)

Akaike (1974) introduces the Akaike information criterion ($AIC$), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit.Two candidate models can said to be equally good if there is a difference of less than 2 in their AIC values.

The Akaike information criterion (AIC) for both models are $AIC_1 = 2304.226$ and $AIC_2 = 2306.226$ , indicating little difference in models. The AIC values for the Carstensen 'unlinked' and 'linked' models are 1994.66 and 1955.48 respectively, indicating an improvement by adding the interaction term.

The $\hat{\mathbf{\Lambda}}$ matrices are informative as to the difference between Carstensen's unlinked and linked models. For the oximetry data, the covariance terms (given above as 11.67 and 2.6 respectively ) are of similar magnitudes to the variance terms. Conversely for the 'fat' data the covariance term $(-0.00032)$ is negligible. When the interaction term is added to the model, the covariance term remains negligible. (For the 'fat' data, the difference in AIC values is also approximately 2).

To conclude, Carstensen's models provided a rigorous way to determine limits of agreement, but don't provide for the computation of $\hat{\boldsymbol{D}}$ and $\hat{\mathbf{\Lambda}}$. Therefore the test's proposed by Roy (2009b) can not be implemented. Conversely, accurate limits of

agreement as determined by Carstensen's model may also be found using Roy's method. Addition of the interaction term erodes the capability of Roy's methodology to compare candidate models, and therefore shall not be adopted.

Finally, to complement the blood pressure (i.e.'J vs S') method comparison from the previous section (i.e.'J vs S'), the limits of agreement are $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

Carstensen et al. (2008) demonstrate the necessity of accounting for linked replicated by comparing the limits of agreement from the 'oximetry' data set using a model with the additional term, and one without. When the interaction is accounted for the limits of agreement are (-9.62,14.56). When the interaction is not accounted for, the limits of agreement are (-11.88,16.83). It is shown that the failure to include this additional term results in an over-estimation of the standard deviations of differences.

Carstensen et al. (2008) demonstrates the use of the interaction term when computing the limits of agreement for the 'Oximetry' data set. When the interaction term is omitted, the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

## 1.7 Linked replicates

Carstensen et al. (2008) proposes the addition of an random effects term to their model when the replicates are linked. This term is used to describe the 'item by replicate' interaction, which is independent of the methods. This interaction is a source of variability independent of the methods. Therefore failure to account for it will result in variability being wrongly attributed to the methods.

Carstensen et al. (2008) introduces a second data set; the oximetry study. This study done at the Royal Childrens Hospital in Melbourne to assess the agreement between co-oximetry and pulse oximetry in small babies.

In most cases, measurements were taken by both method at three different times. In

some cases there are either one or two pairs of measurements, hence the data is unbalanced. Carstensen et al. (2008) describes many of the children as being very sick, and with very low oxygen saturations levels. Therefore it must be assumed that a biological change can occur in interim periods, and measurements are not true replicates.

Carstensen et al. (2008) demonstrate the necessity of accounting for linked replicated by comparing the limits of agreement from the 'oximetry' data set using a model with the additional term, and one without. When the interaction is accounted for the limits of agreement are (-9.62,14.56). When the interaction is not accounted for, the limits of agreement are (-11.88,16.83). It is shown that the failure to include this additional term results in an over-estimation of the standard deviations of differences.

Limits of agreement are determined using Roy's methodology, without adding any additional terms, are found to be consistent with the 'interaction' model; $(-9.562, 14.504)$. Roy's methodology assumes that replicates are linked. However, following Carstensen's example, an addition interaction term is added to the implementation of Roy's model to assess the effect, the limits of agreement estimates do not change. However there is a conspicuous difference in within-subject matrices of Roy's model and the modified model (denoted 1 and 2 respectively);

$$\hat{\mathbf{\Lambda}}_1 = \begin{pmatrix} 16.61 & 11.67 \\ 11.67 & 27.65 \end{pmatrix} \qquad \hat{\mathbf{\Lambda}}_2 = \begin{pmatrix} 7.55 & 2.60 \\ 2.60 & 18.59 \end{pmatrix}. \tag{1.2}$$

(The variance of the additional random effect in model 2 is 3.01.)

Akaike (1974) introduces the Akaike information criterion ($AIC$), a model selection tool based on the likelihood function. Given a data set, candidate models are ranked according to their AIC values, with the model having the lowest AIC being considered the best fit.Two candidate models can said to be equally good if there is a difference of less than 2 in their AIC values.

The Akaike information criterion (AIC) for both models are $AIC_1 = 2304.226$ and $AIC_2 = 2306.226$ , indicating little difference in models. The AIC values for the Carstensen 'unlinked' and 'linked' models are 1994.66 and 1955.48 respectively, indicating an improvement by adding the interaction term.

The $\hat{\mathbf{\Lambda}}$ matrices are informative as to the difference between Carstensen's unlinked and linked models. For the oximetry data, the covariance terms (given above as 11.67 and 2.6 respectively ) are of similar magnitudes to the variance terms. Conversely for the 'fat' data the covariance term ($-0.00032$) is negligible. When the interaction term is added to the model, the covariance term remains negligible. (For the 'fat' data, the difference in AIC values is also approximately 2).

To conclude, Carstensen's models provided a rigorous way to determine limits of agreement, but don't provide for the computation of $\hat{\boldsymbol{D}}$ and $\hat{\mathbf{\Lambda}}$. Therefore the test's proposed by Roy (2009a) can not be implemented. Conversely, accurate limits of agreement as determined by Carstensen's model may also be found using Roy's method. Addition of the interaction term erodes the capability of Roy's methodology to compare candidate models, and therefore shall not be adopted.

Finally, to complement the blood pressure (i.e.'J vs S') method comparison from the previous section (i.e.'J vs S'), the limits of agreement are $15.62 \pm 1.96 \times 20.33 = (-24.22, 55.46)$.)

## 1.8 Alternative agreement indices

As an alternative to limits of agreement, Lin et al. (2002) proposes the use of the mean square deviation is assessing agreement. The mean square deviation is defined as the expectation of the squared differences of two readings . The MSD is usually used for the case of two measurement methods $X$ and $Y$ , each making one measurement for the same subject, and is given by

$$MSDxy = E[(x-y)^2] = (\mu_x - \mu_y)^2 + (\sigma_x - \sigma_y)^2 + 2\sigma_x\sigma_y(1-\rho_{xy}).$$

Barnhart et al. (2007) advises the use of a predetermined upper limit for the MSD value, $MSD_{ul}$, to define satisfactory agreement. However, a satisfactory upper limit may not be properly determinable, thus creating a drawback to this methodology.

Barnhart et al. (2007) proposes both the use of the square root of the MSD or the expected absolute difference (EAD) as an alternative agreement indices. Both of these indices can be interpreted intuitively, being denominated in the same units of measurements as the original measurements. Also they can be compare to the maximum acceptable absolute difference between two methods of measurement $d_0$.

$$EAD = E(|x-y|) = \frac{\sum |x_i - y_i|}{n}$$

The EAD can be used to supplement the inter-method bias in an initial comparison study, as the EAD is informative as a measure of dispersion, is easy to calculate and requires no distributional assumptions.

Barnhart et al. (2007) remarks that a comparison of EAD and MSD , using simulation studies, would be interesting, while further adding that 'It will be of interest to investigate the benefits of these possible new unscaled agreement indices'. For the Grubbs' 'F vs C' and 'F vs T' comparisons, the inter-method bias, difference variances, limits of agreement and EADs are shown in Table 1.5. The corresponding Bland-Altman plots for 'F vs C' and 'F vs T' comparisons were depicted previously on Figure 1.3. While the inter-method bias for the 'F vs T' comparison is smaller, the

EAD penalizes the comparison for having a greater variance of differences. Hence the EAD values for both comparisons are much closer.

| | F vs C | F vs T |
|---|---|---|
| Inter-method bias | -0.61 | 0.12 3 |
| Difference variances | 0.06 | 0.22 |
| Limits of agreement | (-1.08, -0.13) | (-0.81,1.04) |
| EAD | 0.61 | 0.35 |

Table 1.8.1: Agreement indices for Grubbs' data comparisons.

Further to Lin (2000) and Lin et al. (2002), individual agreement between two measurement methods may be assessed using the the coverage probability (CP) criteria or the total deviation index (TDI). If $d_0$ is predetermined as the maximum acceptable absolute difference between two methods of measurement, the probability that the absolute difference of two measures being less than $d_0$ can be computed. This is known as the coverage probability (CP).

$$CP = P(|x_i - y_i| \leq d_0) \tag{1.3}$$

If $\pi_0$ is set as the predetermined coverage probability, the boundary under which the proportion of absolute differences is $\pi_0$ may be determined. This boundary is known as the 'total deviation index' (TDI). Hence the TDI is the $100\pi_0$ percentile of the absolute difference of paired observations.

## 1.9 Limits of agreement for Carstensen's data

**?** describes the calculation of the limits of agreement (with the inter-method bias implicit) for both data sets, based on his formulation;

$$\hat{\alpha}_1 - \hat{\alpha}_2 \pm 2\sqrt{2\hat{\tau}^2 + \hat{\sigma}_1^2 + \hat{\sigma}_2^2}.$$

For the 'Fat' data set, the inter-method bias is shown to be 0.045. The limits of agreement are $(-0.23, 0.32)$

Carstensen demonstrates the use of the interaction term when computing the limits of agreement for the 'Oximetry' data set. When the interaction term is omitted, the limits of agreement are $(-9.97, 14.81)$. Carstensen advises the inclusion of the interaction term for linked replicates, and hence the limits of agreement are recomputed as $(-12.18, 17.12)$.

### 1.9.1 RV-IV

For the the RV-IC comparison, $\hat{D}$ is given by

$$\hat{D} = \begin{bmatrix} 1.6323 & 1.1427 \\ 1.1427 & 1.4498 \end{bmatrix} \tag{1.4}$$

The estimate for the within-subject variance covariance matrix is given by

$$\hat{\Sigma} = \begin{bmatrix} 0.1072 & 0.0372 \\ 0.0372 & 0.1379 \end{bmatrix} \tag{1.5}$$

The estimated overall variance covariance matrix for the the 'RV vs IC' comparison is given by

$$Block\Omega_i = \begin{bmatrix} 1.7396 & 1.1799 \\ 1.1799 & 1.5877 \end{bmatrix}. \tag{1.6}$$

The power of the likelihood ratio test may depends on specific sample size and the specific number of replications, and the author proposes simulation studies to examine this further.

## 1.10 LME - Pankaj Choudhury

Consistent with the conventions of mixed models, (**?**) formulates the measurement $y_{ij}$from method $i$ on individual $j$ as follows;

$$y_{ij} = P_{ij}\theta + W_{ij}v_i + X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{1.7}$$

The design matrix $P_{ij}$ , with its associated column vector $\theta$, specifies the fixed effects common to both methods. The fixed effect specific to the $j$th method is articulated by the design matrix $W_{ij}$ and its column vector $v_i$. The random effects common to both methods is specified in the design matrix $X_{ij}$, with vector $b_j$ whereas the random effects specific to the $i$th subject by the $j$th method is expressed by $Z_{ij}$, and vector $u_j$. Noticeably this notation is not consistent with that described previously. The design matrices are specified so as to includes a fixed intercept for each method, and a random intercept for each individual. Additional assumptions must also be specified;

$$v_{ij} \sim N(0, \Sigma), \tag{1.8}$$

These vectors are assumed to be independent for different $i$s, and are also mutually independent. All Covariance matrices are positive definite. In the above model effects can be classed as those common to both methods, and those that vary with method. When considering differences, the effects common to both effectively cancel each other out. The differences of each pair of measurements can be specified as following;

$$d_{ij} = X_{ij}b_j + Z_{ij}u_j + \epsilon_{ij}, (j = 1, 2, i = 1, 2....n) \tag{1.9}$$

This formulation has seperate distributional assumption from the model stated previously.

This agreement covariate $x$ is the key step in how this methodology assesses agreement.

## 1.11 The Hat Matrix

The projection matrix $H$ (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values $Y$, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^T X)^{-1} X^T \tag{1.10}$$

$H$ describes the influence each observed value has on each fitted value. The diagonal elements of the $H$ are the 'leverages', which describe the influence each observed value has on the fitted value for that same observation. The residuals ($R$) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{1.11}$$

The variances of $Y$ and $R$ can be expressed as:

$$\begin{aligned} \text{var}(Y) &= H\sigma^2 \\ \text{var}(R) &= (I - H)\sigma^2 \end{aligned} \tag{1.12}$$

Updating techniques allow an economic approach to recalculating the projection matrix, $H$, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

## 1.12 Cross Validation

Cross validation techniques for linear regression employ the use 'leave one out' recalculations. In such procedures the regression coefficients are estimated for $n - 1$ covariates, with the $Q^{th}$ observation omitted.

Let $\hat{\beta}$ denote the least square estimate of $\beta$ based upon the full set of observations, and let $\hat{\beta}^{-Q}$ denoted the estimate with the $Q^{th}$ case excluded.

In leave-one-out cross validation, each observation is omitted in turn, and a regression model is fitted on the rest of the data. Cross validation is used to estimate the

generalization error of a given model. alternatively it can be used for model selection by determining the candidate model that has the smallest generalization error.

Evidently leave-one-out cross validation has similarities with 'jackknifing', a well known statistical technique. However cross validation is used to estimate generalization error, whereas the jackknife technique is used to estimate bias.

### 1.12.1 Cross Validation: Updating standard deviation

The variance of a data set can be calculated using the following formula.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i^2) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} \tag{1.13}$$

While using bivariate data, the notation $Sxx$ and $Syy$ shall apply to the variance of $x$ and of $y$ respectively. The covariance term $Sxy$ is given by

$$Sxy = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{n-1} \tag{1.14}$$

Let the observation $j$ be omitted from the data set. The estimates for the variance identities can be updating using minor adjustments to the full sample estimates. Where $(j)$ denotes that the $j$th has been omitted, these identities are

$$Sxx^{(j)} = \frac{\sum_{i=1}^{n}(x_i^2) - (x_j)^2 - \frac{((\sum_{i=1}^{n} x_i) - x_j)^2}{n-1}}{n-2} \tag{1.15}$$

$$Syy^{(j)} = \frac{\sum_{i=1}^{n}(y_i^2) - (y_j)^2 - \frac{((\sum_{i=1}^{n} y_i) - y_j)^2}{n-1}}{n-2} \tag{1.16}$$

$$Sxy^{(j)} = \frac{\sum_{i=1}^{n}(x_i y_i) - (y_j x_j) - \frac{((\sum_{i=1}^{n} x_i) - x_j)(\sum_{i=1}^{n} y_i) - y_k)}{n-1}}{n-2} \tag{1.17}$$

The updated estimate for the slope is therefore

$$\hat{\beta}_1^{(j)} = \frac{Sxy^{(j)}}{Sxx^{(j)}} \tag{1.18}$$

It is necessary to determine the mean for $x$ and $y$ of the remaining $n-1$ terms

$$\bar{x}^{(j)} = \frac{(\sum_{i=1}^{n} x_i) - (x_j)}{n-1}, \tag{1.19}$$

$$\bar{y}^{(j)} = \frac{(\sum_{i=1}^{n} y_i) - (y_j)}{n-1}. \tag{1.20}$$

The updated intercept estimate is therefore

$$\hat{\beta}_0^{(j)} = \bar{y}^{(j)} - \hat{\beta}_1^{(j)} \bar{x}^{(j)}. \tag{1.21}$$

## 1.13 Updating Estimates

### 1.13.1 Updating of Regression Estimates

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row. In time series problems, there will be scientific interest in the changing relationship between variables. In cases where there a single row is to be added or deleted, the procedure used is equivalent to a geometric rotation of a plane.

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row.

### 1.13.2 Updating Standard deviation

A simple, but useful, example of updating is the updating of the standard deviation when an observation is omitted, as practised in statistical process control analyzes. From first principles, the variance of a data set can be calculated using the following formula.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i^2) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} \tag{1.22}$$

While using bivariate data, the notation $Sxx$ and $Syy$ shall apply hither to the variance of $x$ and of $y$ respectively. The covariance term $Sxy$ is given by

$$Sxy = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{n-1}. \tag{1.23}$$

### 1.13.3 Updating of Regression Estimates

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row. In time series problems, there will be scientific interest in the changing relationship between variables. In cases where there a single row is to be added or deleted, the procedure used is equivalent to a geometric rotation of a plane.

Consider a $p \times p$ matrix $X$, from which a row $x_i^T$ is to be added or deleted. **?** sets $A = X^T X$, $a = -x_i^T$ and $b = x_i^T$, and writes the above equation as

$$(X^T X \pm x_i x_i^T)^{-1} = \quad (X^T X)^{-1} \mp \quad \frac{(X^T X)^{-1}(x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \tag{1.24}$$

### 1.13.4 Updating Regression Estimates

Let the observation $j$ be omitted from the data set. The estimates for the variance identities can be updating using minor adjustments to the full sample estimates. Where $(j)$ denotes that the $j$th has been omitted, these identities are

$$Sxx^{(j)} = \frac{\sum_{i=1}^n (x_i^2) - (x_j)^2 - \frac{((\sum_{i=1}^n x_i) - x_j)^2}{n-1}}{n-2} \tag{1.25}$$

$$Syy^{(j)} = \frac{\sum_{i=1}^n (y_i^2) - (y_j)^2 - \frac{((\sum_{i=1}^n y_i) - y_j)^2}{n-1}}{n-2} \tag{1.26}$$

$$Sxy^{(j)} = \frac{\sum_{i=1}^n (x_i y_i) - (y_j x_j) - \frac{((\sum_{i=1}^n x_i) - x_j)(\sum_{i=1}^n y_i) - y_k)}{n-1}}{n-2} \tag{1.27}$$

The updated estimate for the slope is therefore

$$\hat{\beta}_1^{(j)} = \frac{Sxy^{(j)}}{Sxx^{(j)}} \tag{1.28}$$

It is necessary to determine the mean for $x$ and $y$ of the remaining $n-1$ terms

$$\bar{x}^{(j)} = \frac{(\sum_{i=1}^n x_i) - (x_j)}{n-1}, \tag{1.29}$$

$$\bar{y}^{(j)} = \frac{(\sum_{i=1}^n y_i) - (y_j)}{n-1}. \tag{1.30}$$

The updated intercept estimate is therefore

$$\hat{\beta}_0^{(j)} = \bar{y}^{(j)} - \hat{\beta}_1^{(j)} \bar{x}^{(j)}. \tag{1.31}$$

### 1.13.5 Inference on intercept and slope

$$\hat{\beta}_1 \pm t_{(\alpha,n-2)}\sqrt{\frac{S^2}{(n-1)S_x^2}} \tag{1.32}$$

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \tag{1.33}$$

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_0)} \tag{1.34}$$

**Inference on correlation coefficient**

This test of the slope is coincidentally the equivalent of a test of the correlation of the $n$ observations of $X$ and $Y$.

$$\begin{aligned} H_0 &: \rho_{XY} = 0 \\ H_A &: \rho_{XY} \neq 0 \end{aligned} \tag{1.35}$$

## 1.14 influence.ME

*influence.ME* allows you to compute measures of influential data for mixed effects models generated by lme4.

*influence.ME* provides a collection of tools for detecting influential cases in generalized mixed effects models. It analyses models that were estimated using lme4. The basic rationale behind identifying influential data is that when iteratively single units are omitted from the data, models based on these data should not produce substantially different estimates.

To standardize the assessment of how influential a (single group of) observation(s) is, several measures of influence are common practice, such as DFBETAS and Cook's Distance. In addition, we provide a measure of percentage change of the fixed point estimates and a simple procedure to detect changing levels of significance.

### 1.14.1 Influence() command

`influence()` is the workhorse function of the influence.ME package. Based on a priorly estimated mixed effects regression model (estimated using lme4), the `influence()` function iteratively modifies the mixed effects model to neutralize the effect a grouped set of data has on the parameters, and which returns returns the fixed parameters of these iteratively modified models. These are used to compute measures of influential data.

### 1.14.2 Classical Model

The classical model is based on measurements $y_{mi}$ by method $m = 1, 2$ on item $i = 1, 2 \ldots$

$$y_{mi} + \alpha_m + \mu_i + e_{mi}$$

$$e_{mi} \sim \backslash(0, \sigma_m^2)$$

Even though the separate variances can not be identified, their sum can be estimated by the empirical variance of the differences.

Like wise the separate $\alpha$ can not be estimated, only theiir difference can be estimated as $\bar{D}$

## 1.15 Lai Shiao

Lai and Shiao (2005) advocates the use of LME models to study method comparison problems. The authors analyse a data set typical of method comparison studies using SAS software, with particular use of the *'Proc Mixed'* package. The stated goal of this study is to determine which factor from a specified group of factors is the key contributor to the difference in the two methods.

The study relates to oxygen saturation, the most investigated variable in clinical nursing studies (Lai and Shiao, 2005). The two method compared are functional saturation (SO2, percent functional oxy-hemoglobin) and fractional saturation (HbO2, percent fractional oxy-hemoglobin), which is considered to be the 'gold standard' method of measurement.

Lai and Shiao (2005) establishes an LME model for analysing the differences $D_{ijtl}$, where $D_{ijtl}$ is the differences of the measurements (i.e = $SO2_{ijtl}$ - $HbO2_{ijtl}$) for the ith donor at the $j$th level of foetal haemoglobin percent (Fhbperct) and the $t$th repeated measurement by the $l$th practitioner of the experiment.

(Carstensen (2004) also advocates the use of LME models in comparing methods, but with a different emphasis.) Lai and Shiao (2005) use mixed models to determine the factors that affect the difference of two methods of measurement using the conventional formulation of linear mixed effects models.

If the parameter **b**, and the variance components are not significantly different from zero, the conclusion that there is no inter-method bias can be drawn. If the fixed effects component contains only the intercept, and a simple correlation coefficient is used, then the estimate of the intercept in the model is the inter-method bias. Conversely the estimates for the fixed effects factors can advise the respective influences each factor has on the differences. It is possible to pre-specify different correlation structures of the variance components **G** and **R**.

Oxygen saturation is one of the most frequently measured variables in clinical nursing studies. 'Fractional saturation' ($HbO_2$) is considered to be the gold standard method of measurement, with 'functional saturation' ($SO_2$) being an alternative method. The method of examining the causes of differences between these two methods is applied to a clinical study conducted by **?**. This experiment was conducted by 8 lab practitioners on blood samples, with varying levels of haemoglobin, from two donors. The samples have been in storage for varying periods ( described by the variable 'Bloodage') and are categorized according to haemoglobin percentages(i.e 0%,20%,40%,60%,80%,100%). There are 625 observations in all.

Lai and Shiao (2005) fits two models on this data, with the lab technicians and the replicate measurements as the random effects in both models. The first model uses haemoglobin level as a fixed effects component. For the second model, blood age is added as a second fixed factor.

**Single fixed effect**

The first model fitted by Lai and Shiao (2005) takes the blood level as the sole fixed effect to be analyzed. The following coefficient estimates are estimated by 'Proc Mixed';

$$\begin{aligned} &\text{fixed effects}: 2.5056 - 0.0263\text{Fhbperct}_{ijtl} \\ &(\text{p-values}: = 0.0054, < 0.0001, < 0.0001) \end{aligned} \tag{1.36}$$

$$\begin{aligned} \text{random effects}: u(\sigma^2 = 3.1826) + e_{ijtl}(\sigma_e^2 = 0.1525, \rho = 0.6978) \\ (\text{p-values}: = 0.8113, < 0.0001, < 0.0001) \end{aligned}$$

With the intercept estimate being both non-zero and statistically significant ($p = 0.0054$), this models supports the presence inter-method bias is 2.5% in favour of $SO_2$. Also, the negative value of the haemoglobin level coefficient indicate that differences will decrease by 0.0263% for every percentage increase in the haemoglobin .

In the random effects estimates, the variance due to the practitioners is 3.1826, indicating that there is a significant variation due to technicians ($p = 0.0311$) affecting the differences. The variance for the estimates is given as 0.1525, ($p < 0.0001$).

**Two fixed effects**

Blood age is added as a second fixed factor to the model, whereupon new estimates are calculated;

$$\begin{aligned} \text{fixed effects}: \; & -0.2866 + 0.1072\text{Bloodage}_{ijtl} - 0.0264\text{Fhbperct}_{ijtl} \\ & (\text{p-values}: \; = 0.8113, < 0.0001, < 0.0001) \\ \\ \text{random effects}: \; & u(\sigma^2 = 10.2346) + e_{ijtl}(\sigma_e^2 = 0.0920, \rho = 0.5577) \\ & (\text{p-values}: \; = 0.0446, < 0.0001, < 0.0001) \end{aligned} \tag{1.37}$$

With this extra fixed effect added to the model, the intercept term is no longer statistically significant. Therefore, with the presence of the second fixed factor, the model is no longer supporting the presence of inter-method bias. Furthermore, the second coefficient indicates that the blood age of the observation has a significant bearing on the size of the difference between both methods ($p < 0.0001$). Longer storage times for blood will lead to higher levels of particular blood factors such as MetHb and HbCO (due to the breakdown and oxidisation of the haemoglobin). Increased levels of MetHb and HbCO are concluded to be the cause of the differences. The coefficient for the haemoglobin level doesn't differ greatly from the single fixed factor model, and has a much smaller effect on the differences. The random effects estimates also indicate significant variation for the various technicians; 10.2346 with $p = 0.0446$.

Lai and Shiao (2005) demonstrates how that linear mixed effects models can be used to provide greater insight into the cause of the differences. Naturally the addition of further factors to the model provides for more insight into the behavior of the data.

## 1.16 Sherman Morrison Woodbury Formula

The 'Sherman Morrison Woodbury' Formula is a well known result in linear algebra;

$$(A + a^T B)^{-1} \quad = \quad A^{-1} - A^{-1}a^T(I - bA^{-1}a^T)^{-1}bA^{-1} \tag{1.38}$$

This result is highly useful for analyzing regression diagnostics, and for matrices inverses in general. Consider a $p \times p$ matrix $X$, from which a row $x_i^T$ is to be added or deleted. **?** sets $A = X^T X$, $a = -x_i^T$ and $b = x_i^T$, and writes the above equation as

$$(X^T X \pm x_i x_i^T)^{-1} = \quad (X^T X)^{-1} \mp \quad \frac{(X^T X)^{-1}(x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \tag{1.39}$$

The projection matrix $H$ (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values $Y$, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^T X)^{-1} X^T \tag{1.40}$$

$H$ describes the influence each observed value has on each fitted value. The diagonal elements of the $H$ are the 'leverages', which describe the influence each observed value has on the fitted value for that same observation. The residuals ($R$) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{1.41}$$

The variances of $Y$ and $R$ can be expressed as:

$$\begin{aligned} \text{var}(Y) &= H\sigma^2 \\ \text{var}(R) &= (I - H)\sigma^2 \end{aligned} \tag{1.42}$$

Updating techniques allow an economic approach to recalculating the projection matrix, $H$, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

### 1.16.1 Hat Values for MCS regression

With A as the averages and D as the casewise differences.

```
fit = lm(D~A)
```

$$H = A\left(A^\top A\right)^{-1} A^\top,$$

# Bibliography


Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Barnhart, H., M. Haber, and L. Lin (2007). An overview of assessing agreement with continuous measurements. *Journal of Biopharmaceutical Statistics 17*, 529–569.

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.

Lai, D. and S.-Y. P. K. Shiao (2005). Comparing two clinical measurements: a linear mixed model approach. *Journal of Applied Statistics 32*(8), 855–860.

Lin, L. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in medicine 97*, 255–270.

Lin, L., A. Hedayat, B. Sinha, and M. Yang (2002). Statistical methods in assessing agreement: Models issues and tools. *Journal of American Statistical Association 97*, 257–270.

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

Roy, A. (2009a). An application of linear mixed effects model to assess the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.

Roy, A. (2009b). An application of the linear mixed effects model to ass the agreement between two methods with replicated observations. *Journal of Biopharmaceutical Statistics 19*, 150–173.