Bayesian BA - Philip J Schluter

Bayesian Bland Altman Approaches A multivariate hierarchical Bayesian approach to measuring agreement in repeated measurement method comparison studies

http://www.biomedcentral.com/1471-2288/9/6

Background

Assessing agreement in method comparison studies depends on two fundamentally important components; validity (the between method agreement) and reproducibility (the within method agreement).

The Bland-Altman limits of agreement technique is one of the favoured approaches in medical literature for assessing between method validity. However, few researchers have adopted this approach for the assessment of both validity and reproducibility.

This may be partly due to a lack of a flexible, easily implemented and readily available statistical machinery to analyse repeated measurement method comparison data.

Methods

Adopting the Bland-Altman framework, but using Bayesian methods, we present this statistical machinery. Two multivariate hierarchical Bayesian models are advocated, one which assumes that the underlying values for subjects remain static (exchangeable replicates) and one which assumes that the underlying values can change between repeated measurements (non-exchangeable replicates).

Results

We illustrate the salient advantages of these models using two separate datasets that have been previously analysed and presented; (i) assuming static underlying values analysed using both multivariate hierarchical Bayesian models, (ii) assuming each subject's underlying value is continually changing quantity and analysed using the non-exchangeable replicate multivariate hierarchical Bayesian model.

Conclusion These easily implemented models allow for full parameter uncertainty, simultaneous method comparison, handle unbalanced or missing data, and provide estimates and credible regions for all the parameters of interest. Computer code for the analyses in also presented, provided in the freely available and currently cost free software package WinBUGS. ¡hr¿

Bayesian Approach

A multivariate hierarchical Bayesian approach to measuring agreement in repeated measurement method comparison studies PJ Schluter - BMC medical research methodology, 2009 - biomedcentral.com

- Assessing agreement in method comparison studies depends on two fundamentally important components; validity (the between method agreement)

and reproducibility (the within method agreement).

- The Bland-Altman limits of agreement technique is one of the f

1 Escaramis

1.1 Background

In an agreement assay, it is of interest to evaluate the degree of agreement between the different methods (devices, instruments or observers) used to measure the same characteristic. We propose in this study a technical simplification for inference about the total deviation index (TDI) estimate to assess agreement between two devices of normally-distributed measurements and describe its utility to evaluate inter- and intra-rater agreement if more than one reading per subject is available for each device.

1.2 Methods

We propose to estimate the TDI by constructing a probability interval of the difference in paired measurements between devices, and thereafter, we derive a tolerance interval (TI) procedure as a natural way to make inferences about probability limit estimates. We also describe how the proposed method can be used to compute bounds of the coverage probability.

1.3 Results

The approach is illustrated in a real case example where the agreement between two instruments, a handle mercury sphygmomanometer device and an OMRON 711 automatic device, is assessed in a sample of 384 subjects where measures of systolic blood pressure were taken twice by each device. A simulation study procedure is implemented to evaluate and compare the accuracy of the approach to two already established methods, showing that the TI approximation produces accurate empirical confidence levels which are reasonably close to the nominal confidence level.

1.4 Conclusions

The method proposed is straightforward since the TDI estimate is derived directly from a probability interval of a normally-distributed variable in its original scale, without further transformations. Thereafter, a natural way of making inferences about this estimate is to derive the appropriate TI. Constructions of TI based on normal populations are implemented in most standard statistical packages, thus making it simpler for any practitioner to implement our proposal to assess agreement.

Lin defined the TDI as the boundary, κ_P which capyures a large proportion p of paired based differences from two devices or observers within the boundary.

The value of κ_P that yeilds $P(|D| < \kappa_p) = p$ where D is the paired-difference variate.

$$\kappa_P = F^{-1}(p) = \sigma_D \sqrt{\chi^2(p, 1, \mu_D^2/\sigma_d^2)}$$

$$\kappa_P = Z_{\frac{1+p}{2}} \|\varepsilon\|$$

Tolerance Interval around the TDI estimate

$$\hat{\kappa_p} = \hat{\mu}_D = Z_{p_i} \sigma_d$$

Coverage Probability is another user friendly measure of agrre,ment which is related to the computation of the TDI.

2 Influence analysis

Likelihood based estimation methods, such as ML and REML, are sensitive to unusual observations. Influence diagnostics are formal techniques that assess the influence of observations on parameter estimates for β and θ. A common technique is to refit the model with an observation or group of observations omitted.

west examines a group of methods that examine various aspects of influence diagnostics for LME models. For overall influence, the most common approaches are the 'likelihood distance' and the 'restricted likelihood distance'.

2.1 Cook's 1986 paper on Local Influence

Cook 1986 introduced methods for local influence assessment. These methods provide a powerful tool for examining perturbations in the assumption of a model, particularly the effects of local perturbations of parameters of observations.

The local-influence approach to influence assessment is quitedifferent from the case deletion approach, comparisons are of interest.

2.2 Overall Influence

An overall influence statistic measures the change in the objective function being minimized. For example, in OLS regression, the residual sums of squares serves that purpose. In linear mixed models fit by maximum likelihood (ML) or restricted maximum likelihood (REML), an overall influence measure is the likelihood distance [Cook and Weisberg].

2.3 Influence

schab examines the use and implementation of influence measures in LME models.

Influence is understood to be the ability of a single or multiple data points, through their presences or absence in the data, to alter important aspects of the analysis, yield qualitatively different inferences, or violate assumptions of the statistical model (*schabenberger*).

Outliers are the most noteworthy data points in an analysis, and an objective of influence analysis is how influential they are, and the manner in which they are influential.

schab describes a simple procedure for quantifying influence. Firstly a model should be fitted to the data, and estimates of the parameters should be obtained. The second step is that either single of multiple data points, specifically outliers, should be omitted from the analysis, with the original parameter estimates being updated.

This is known as '*leave one out leave k out*' analysis. The final step of the procedure is comparing the sets of estimates computed from the entire and

reduced data sets to determine whether the absence of observations changed the analysis.

schabenberger notes that it is not always possible to derive influence statistics necessary for comparing full- and reduced-data parameter estimates.

2.4 Influence

Broadly defined, "*influence* is understood as the ability of a single or multiple data points, through their presence or absence in the data, to alter important aspects of the analysis, yield qualitatively different inferences, or violate assumptions of the statistical model.

The goal of influence analysis is not primarily to mark data points for deletion so that a better model fit can be achieved for the reduced data, although this might be a result of influence analysis. The goal is rather to determine which cases are influential and the manner in which they are important to the analysis. Outliers, for example, may be the most noteworthy data points in an analysis. They can point to a model breakdown and lead to development of a better model.

In recent years, mixed models have become invaluable tools in the analysis of experimental and observational data. In these models, more than one term can be subject to random variation. Mixed model technology enables you to analyze complex experimental data with hierarchical random processes, temporal, longitudinal, and spatial data, to name just a few important applications.

schab remarks that the concept of critiquing the model-data agreement applies in mixed models in the same way as in linear fixed-effects models. In fact, because of the more complex model structure, you can argue that model and data diagnostics are even more important. For example, you are not only concerned with capturing the important variables in the model. You are also concerned with "distributing them correctly between the fixed and random components of the model. The mixed model structure presents unique and interesting challenges that prompt us to reexamine the traditional ideas of influence and residual analysis.

3 Hawkins : Diagnostics for conformity of paired quantitative measurements

- Matched pairs data arise in many contexts in case-control clinical trials, for example, and from cross-over designs. They also arise in experiments to verify the equivalence of quantitative assays. This latter use (which is the main focus of this paper) raises difficulties not always seen in other matched pairs applications.

- Since the designs deliberately vary the analyte levels over a wide range, issues of variance dependent on mean, calibrations of differing slopes, and

curvature all need to be added to the usual model assumptions such as normality.

- Violations in any of these assumptions invalidate the conventional matched pairs analysis.
- A graphical method, due to Bland and Altman, of looking at the relationship between the average and the difference of the members of the pairs is shown to correspond to a formal testable regression model.
- Using standard regression diagnostics, one may detect and diagnose departures from the model assumptions and remedy them for example using variable transformations. Examples of different common scenarios and possible approaches to handling them are shown.

Roy Test

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 Variabili
> testW   = anova(Ref.Fit,NMW.fit)                         # Within-Subject Variabilities
> testO     = anova(Ref.Fit,NMO.fit)                         # Overall Variabilities
```

4 Profile Function with "lmer"

The profile() function for lmer models is now available in the latest version of lme4, to be installed by typing:

install.packages("lme4",repos="http://r-forge.r-project.org")

also

The mle function from the stats4 package is a wrapper of optim, which makes it quite easy to produce profile likelihood computations.

See help("profile,mle-method", package = "stats4") for more information.

http://people.upei.ca/hstryhn/stryhn208.pdf

The profile likelihood (or likelihood or likelihood ratio) methid is applicable to all likelihood based statstical analysis and is generally less sensitive to the difficulties encountered by walkd-Tyoe CIs.

5 Quiroz Burdick

Assessmet of individual agreements with repeated measurements based on Generalized Confidence intervals.

Bootstrap confidence intervals. Coverage probability (CP) Equivalence Studies Individual agreements Generalized Confidence intervals (GCI) Total deviation index (TDI) Variance components

Proposing an equivalence test for assessing individual agreement based on TDI and CP. The bounds used in the tests are constucted using a bootstrap approach and generalized confidence intervals (GCI).

Equivalence testing is an approach commonly used to determine the acceptability of a new method against a reference method.

Both the TDI and CP are attractive criteria as they are easy to interpret.

Bootstrap approach was later applied to mixed models with repeated measurements by Choudhary (2007)

T for test measurement, R for reference measurement

$\otimes$ is the Kroneckor Product operator.

$$\Sigma_{MS} = \begin{bmatrix} \sigma^2_{TS} & 0 \\ 0 & \sigma^2_{RS} \end{bmatrix}$$