Replies to referee reports on the manuscript JSS-861

*Ulrich Halekoh and Søren Højsgaard : A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models – the R Package pbkrtest*

General replies

Thanks for very thorough reviews.

We apologize for the rather long time it took for the revision. The main reason is, that we became aware that the development of the lme4 package is now continued on GitHub with important structural changed in relation to the CRAN housed version which must be considered static. Because of the structural changes we had to revise parts of our program.

Please not that we made two additions to the paper:

1. In section 6 we show how to do a parametric bootstrap with a sequential determination of the resample size.
2. Section 7 shows how to test a linear hypothesis by using a restriction matrix and obtaining a test on several contrasts (this addendum was suggested by both reviewers)

Reviewer 1:

Comments from the reviewer are in lines starting with “\*”;

Our replies are marked in yellow and given in lines starting with “>>”

In general the paper and the pbkrtest package are clearly written and an excellent addition to the R ecosystem.

>> Thank you.

Most of my comments are saved as notes on the PDF. Here are some more general points.

>> Thank you for your very thorough work. All the comments have been taken into account.

STATISTICAL:

\* What bias or other artifacts might be induced by discarding PB samples on which lmer fails to converge?

>> The short answer is that one obtains artificially large p-values. For this reason we have based the current version of the manuscript on the version of lme4 which is on GibHub because it has better numerical properties: The algorithm(s) converge; also on simulated data. We hope GitHub-versions of lme4 will soon appear on CRAN.

\* It surprises me that the PB results are \_less\_ accurate than the KR results. I would have expected that the PB results would be more or less exactly right, for sufficiently large sample sizes, but in general (according to Table 1) KR is closer to the nominal value.

>> We see no compelling reason that PB should give exactly the right coverage even for large number of bootstrap samples.  
  
 \* would it be possible to extend the usage of PB to compute confidence intervals and tests of specific contrasts, or is this too broad a scope for the package?

>> We find that this would be outside the scope of the package. However, pbkrtest provides a function vcovAdj() which returns the adjusted covariance matrix for K&R and a bootstrap version of the covariance matrix based on the PB approach. The result from these functions can be used as input to, for example, the glht() function in the multcomp package and the esticon() function in the doBy package ot the lsemans() function of the lsmeans package. We have added examples (see section Testing a hypothesis Lbeta=Lbeta\_H) in the paper illustrating how to do so.

COSMETIC/PACKAGING:

\* It would be nice to provide an original reference for the 'beets' data set if possible.

>> We were not able to identify a printed source, the data were used by one of our collegue of us could recollect that the data must have been collected about 30 years ago in a field trial.

\* I dislike the default format of "PBmodcomp"; giving all of the different corrections at once seems rather "SAS"-ish and would seem to be an invitation to cherry-picking. I would prefer a "method" argument that selected the method(s) to choose -- it could have an "all" option, I suppose, to restore the current behaviour -- but its default would preferably be set to a sensible default (maybe c("PBkd","PBtest")?)

>> We have changed the output such that only the results from the asymptotic chi^2-distribution and the PBtest values appear in the output. In addition we have added a summary method that will provide all the tests.

\* is there any point in trying to connect the bootstrap results with the format and capabilities of the boot package? (Although I will say (a) I find the format of "boot" objects slightly clunky and (b) the

adjusted confidence intervals in the boot package don't work very easily with parametric bootstrap results (see http://tinyurl.com/bl6le26)

>> We do not see immediate advantages of connecting our work to the boot package, partly because the programming in pbkrtest is already made (and it is simple), but more importantly because a simple extension of PBkrtest allows to let the number of simulations be determined sequentially (in the spirit of the work on sequential Monte Carlo p-values of Besag and Clifford). ).

\* you could make the output of PBmodcomp slightly prettier by using printCoefmat(), and particularly using na.print="" or na.print="." to eliminate the NA values of df and ddf where they aren't relevant.

>> Good suggestion, thanks. This has been implemented.

\* would it make sense to replace calls to snow functions with calls to functions from the parallel package, which is (a) included as a core package (now) and (b) allows both snow-like and multicore functionality?

>> We have made the package use the parallel package instead of the snow package.

\* does it make sense to report p values as <(1/N+1), rather than 0? More generally, I have sometimes been told (don't have a reference right now, sorry) that it is most appropriate

to include the observed data as part of the ensemble for calculating p-values, i.e. a PB sample of size N with p values >= the observed statistic would be given a p-value of (1+p)/(1+N)

>> We stick to reporting 0.000.. in the output table, but we agree to add the additional information suggested above. We also agree to employ the suggested definition of the bootstrap p-values (as given in Davison and Hinkley).

\* A different and potentially more powerful way to plot the reference distributions is with nominal vs observed p-values (see attached plot)

>> Very good idea. This has been implemented, and there is now a plot method for the objects returning from PBmodcomp().

\* Could we have the simulation code in an appendix?

>> Yes, has been added.

\* I think it might be nice to convert Table 1 to a figure, for example (the colours here are pretty but superfluous -- it would be easy to do a black & white version)

>> While we appreciate the suggestion, we find that the table conveys the information clearly enough

Reviewer 2:

1. The article presents its subject very clearly and is well written. The presentation is well developed with good illustrative data examples that are used to motivate and explain the methodology. The detailed matrix examples using this data are very effective.

>> Thank you.

2. The package seems to have problems with stability. I used it a month ago and it worked properly in the sense that it gave results consistent with those in the article. To prepare a final version of this report, I re-installed the package today (August 15, 2012), Version: 0.3-1, Built: R 2.15.1; ; 2012-08-13 18:52:18 UTC; windows, and the examples applying 'KRmodcomp' to the 'beets' data set, gave results very different

from those in the article, e.g.

>> Package is under continuous development.We use in the article throughout the GitHub version of lme4 which proved to be numerically more stable than the CRAN version.

3. I commend the authors for using the 'parallel' package for parallel generation of bootstrap samples. Note that the current help file for 'PBmodcomp' was written for the 'snow' package. The example showing a call to 'makeSOCKcluster' does not work with the 'parallel' package.

>> The current version of the package uses the parallel package.

4. Since the package extends the 'lme4' package it might be a good idea to clarify the scope of the models to which pbkrtest applies. It seems to be intended only for linear models with a gaussian response. I have been able to invoke 'PBmodcomp' on a binomial model without producing an error message although I don't think that the simulated null values were drawn from a binomial. It would be desirable to make this

more clear both in the article and in error or warning messages in the package.

>> The scope has been clarified and call with a ‘glmer’ fit yield now an error message.

5. The last sentence of the second paragraph following equation (1) on p. 2 reads “Notice that the structural forms of the random components of the two models are mes[sic] identical”. I presume that 'mes' is a typographical error but what I find interesting here is that there appears to be no need for the random structures to be identical. For K&R, it is only the random structure of the large model that matters. The role of the smaller model is merely to provide the fixed effects model subspace from which the matrix can be obtained. In contrast, for the PB test, it is only the random structure of the smaller model that matters.

>> The reviewer is correct in mentioning this. The discussion has been extended with a discussion of this.

6. There are very few typographical errors and relatively few questions concerning usage and grammar. A professional copy editor would correct

these much better than I could. On request I would be happy to forward those I believe I have found. The only 'error' I have found that could be

missed by a copy editor follows Appendix A.1 equation (22):

“elements of the lower triangular ”

should read:

“elements of the lower triangle of ”

>> Another reviewer has already provided detailed info about these errors, but thanks for the offer.

7. The following are further comments that are merely suggestions for future consideration.

General Linear Hypotheses

Although KRmodcomp computes a linear hypothesis matrix to compute a Wald test, it is not possible to specify the hypothesis in the form:

or, more generally,

i.e. using the 'restriction matrix' directly without requiring the user to specify and fit the null model.

This would be a very easy and, in my opinion a desirable, enhancement.

It would also be easy to allow for matrices that are not of full rank by performing the overall test by transforming to an equivalent semiorthogonal

hypothesis matrix whose rows span the same space as that spanned by the rows of . The packages has all the tools are in place to do

this easily. The default could be simply testing the hypotheses with , i.e. producing standard regression output with K&R estimates for

denominator degrees of freedom for each single-degree-of-freedom effect in the model. This would be highly welcome by many users – perhaps

too welcome!

The output could produce tests for the hypotheses represented by each row of , each with the appropriate degrees of freedom, as well as the joint hypothesis that all row hypotheses are satisfied.

This output would be very useful in estimating, for example, a full set of pairwise differences among levels of a categorical factor.

>> These facilities do exist in the package, and we have added an example showing how use these facilities. Additionally we point to the possibility to calculate lsmeans with bias corrected standard errors using the lsmeans packages of Russell Leth, who uses the pbkrtest package.

Algorithms

>> We appreciate that the reviewer can support our approach to the implementation.

An additional note is that, as long as is only obtained as a restriction matrix from two models, strong collinearity in the rows of – as might occur

if were supplied directly by the user – is not a concern. The issue becomes more salient if can be supplied as an argument.

>> This has been illustrated in an example, cfr. reply above.