Author's response to reviews

Title: Estimating equations for biomarker based exposure estimation under non-steady-state conditions

Authors:

Scott M Bartell (sbartell@uci.edu)
Wes O Johnson (wjohnson@uci.edu)

Version: 2 Date: 28 March 2011

Author's response to reviews: see over
Response to reviewer comments for MS 6186704284631047: “Estimating equations for biomarker based exposure estimation under non-steady-state conditions”

The authors would like to thank the reviewers for their careful reviews and detailed suggestions. We feel the revised manuscript is much improved.

Reviewer #1

*What would be the exposure estimates if a steady-state model was adopted?*

We have added text on p. 14 addressing this question.

*Major Compulsory Revisions:*

1. *It is suggested that an example with real data be provided with mean exposure and variance estimates by the proposed method. For example, the technique may be applied to estimate maternal mercury exposures based on blood mercury concentration data for women of childbearing age from the National Health and Nutrition Survey database, and to compare the results with those of the existing methods. In this way, one can then assess the contributions of the proposed method for exposure assessment more easily.*

The use of NHANES data is an interesting idea, but poses some challenges due to the complex survey sampling design (a four stage cluster sample with post-stratification weighting to a US Census standard population). Pfeffermann (International Statistical Review 1993, 61:317-337) reviews general approaches for complex survey data including weighted estimating functions and other methods that might work in this setting, but this would require significant modifications beyond the scope of the present manuscript.

2. *(Algorithm, 5th paragraph) The authors reparameterize the gamma distribution parameters and using and . However, the notations and are defined as vectors of parameters of the covariance matrix and the mean vector previously. Different notations must be used for distinction. Also, why the subscripts i of and disappeared when reparameterize using and ? Supposedly these two new parameters should represent population exposure mean and variance. A clarification of this is needed. Following this step, it is stated that the vectors EY and D are identical. However, the former is a vector and the latter is a matrix. The dimensions of the two do not match.*

Although the notation was missing from our copy of the reviewer comments, we believe we understand his comment. We intentionally choose \( \alpha \) and \( \beta \) to match the notation used earlier for the unknown mean and covariance parameters, because the scalars \( \alpha \) and \( \beta \) from the reparameterized zero-inflated gamma distribution are the parameters being estimated with the estimating equations. Because scalar \( \beta \) is the only unknown mean parameter in this parameterization, \( p=1 \) and \( D \) is indeed a
vector. We have revised the manuscript to make this clearer by introducing the reparameterization immediately after equation 3 on page 7, and with additional clarifying text on page 10. We have also eliminated the subscripts on the parameters \(a, \lambda, \alpha, \) and \(\beta\) throughout the manuscript in order to emphasize that these parameters are assumed not to vary across individuals.

3. (Algorithm, 5th & 6th paragraphs) While complicated mathematical derivations of the formulas of the matrices \(D, V, D^*,\) and \(V^*\) may not be necessary for an applied paper, it is difficult to justify the correctness of the expressions. Perhaps some details of these may be provided as an appendix or attached materials for readers to replicate the results.

Intermediate calculations for expectations and variances for a single day's exposure have been added to p. 7. The calculations for \(E(Y)\) and \(\text{Var}(Y)\) are now described in enough detail for replication on p. 11, along with an additional citation.

4. The authors present the simulation results only through Figures 1 to 4. To be able to assess the scales of bias and variations, tabulated results are desired in addition to diagrams. Also, the simulation results only show the effects of different sample sizes. What would be the effects with different scenarios of \(a\) and \(t\)? It is also of interest to know the corresponding 95% CIs of the estimates.

The results shown in Figures 1-2 are now tabulated in Tables 1-2, respectively. We have also added Tables 3-4 showing the actual coverage rates for the nominal 95% CIs assuming approximate normality.

Changing \(t\) does not substantially affect the results except for very small values of \(t\)—a case where the steady-state model would never be applied. Although many assumptions and parameters could be varied for simulations, we focused on the exposure frequency and sample size because of their strong effects on the results.

Minor Essential Revisions:
1. It is suggested that the authors merge the Algorithm and the first paragraph of Testing and Implementation subsections of the Results section with the Methods section, together with an example dataset for illustration. The results section may also be expanded with the data analysis and simulations of different scenarios of parameter settings and a table to summarize the results numerically. The first two sentences of the Results paragraph in the Abstract also appear to belong to the Methods paragraph.

Following the journal guidelines for methodology articles, we presented the estimating equations approach and our specific method in the Results section. We have entirely rewritten the abstract to comply with the 75-word limit.

2. There are several references in the text do not appear in the references list. These include Degroot (1989), Steenland and Greenland (2004), and Prentice and Zhao
There is a typo error in Sherlock et al. (1982) and year of Wedderburn is missing. Also, the publication year for Hardin and Hilbe should be corrected to 2003 (last line of the Background section).

We have added the missing references and corrected the Harden and Hilbe reference. The Steenland and Greenland manuscript is now omitted, as it is actually an example of post-hoc correction for an unmeasured confounder.

3. **Testing and Implementation** The unit of the mean exposure is not given.

Units have now been added for all parameters on p. 12.

4. A brief description of the underlying single compartment biokinetic (PK) model and the conditions as well the solutions for steady-state and non-steady state might be helpful.

We have added this description to p. 5.

**Reviewer #2**

**Major Compulsory Revisions**

1) General: The authors present an interesting approach to estimating the distribution of toxicant exposure using biomarker data within the framework of a non-steady state model. However, there are limitations which suggest substantially more work is needed before the approach can be recommended outside of settings where exposure distribution is rather homogeneous across subjects. The authors should emphasize, and in some instances spell-out in greater detail, these limitations.

We have added more explicit discussion of the limitations to p. 15.

2) Equation 3: x in the numerator of equation (3) should be I_{ij}

We have corrected the equation as suggested.

3) Equation (3): The authors should emphasize that unique a_i and #_i are (presumably) only estimable if they are modeled using fixed covariate effects (e.g. job type), something not explicitly considered in this paper.

We have decided to omit the i subscript from these parameters altogether in order to avoid confusion, as per our response to major comment #2 from reviewer #1.

4) Equation (3): The authors should state that they are assuming daily exposures are independent within-person, and comment on limitations of this assumption.
We have added text to p. 7 describing this assumption and several scenarios that would cause correlations.

5) Line 229: DeGroot (1989) should be added to the reference list.

We have corrected the reference list.

6) Line 244 (last paragraph under Algorithm): Given that estimating \# is one of the main goals of the paper (being a parameter of the exposure distribution), requiring an external estimate of \# seems the most important issue. If an external estimate of \# is being recommended, then it should be explicitly mentioned as a limitation in the context of the paper as a whole.

We have added text to p. 15 emphasizing this limitation.

7) After Line 276 (4'th paragraph of Testing and Implementation): In the simulation studies the authors have assumed that the distribution of exposure (given that exposure takes place) over 1000 exposure measurement intervals are homogeneous, both within and between subjects. They have also assumed that the residual error variance is small, known, and homogeneous across subjects, and that the frequency of exposure is known for the entire period of 1000 exposure intervals for each subject. Given these rather strong restrictions, the authors should comment on how they might expect their model to perform in general scenarios where exposure levels vary (e.g. based on preliminary simulation results), or the residual model error is not small, or exposure frequency is measured with error. If there’s no basis for speculation, then explicitly mention that performance in more general settings is unknown should be made.

We have added text to p. 15 acknowledging that more work should be done to evaluate the robustness of both the steady-state and non-steady-state methods when various assumptions are violated.

8) Line 294 (first paragraph of Discussion): The statement “It may be possible to extend the estimating equations to handle multiple biomarkers per individual” conflicts with line 288, where the authors recommend multiple biomarker values be sampled per person. If it is possible then state as such; otherwise indicate that this needs to be explored.

It is indeed possible; we have now stated that more clearly in the discussion.

9) Discussion: It’s not obvious how biokinetic parameters would be incorporated into the process as random effects (at least not easily, given what the score-equations might look like). Likewise, it is not clear how to incorporate information “...about the nature of exposure related behaviors.” Presumably the latter gets at making the exposure distribution much more general (e.g. incorporating covariates, random
exposure effects, unobserved exposure frequency, etc.). The authors should explain how this might actually be done, or emphasize it as future research.

We have removed this paragraph. Random effects for the biokinetic parameters is a very complex issue for which we have done only preliminary work in the frequentist setting. The statement regarding conditional means and variances is simply an idea for future work, and as such is best omitted from the present manuscript.

10) Conclusions: A need for caution and future research appears warranted, because the assessment done in this manuscript – although very interesting – seem based on strong assumptions about the exposure process and known exposure frequency over potentially long periods of time. Likewise, emphasize the importance of correctly specifying the biomarker-exposure model, validating exposure distributional assumptions, the need to have external estimates of #2#, etc.

We strongly agree and have added appropriate text to p. 15.

Minor Essential Revisions 1) Line 30: “...each biomarker measurement as a weighted linear...”

The abstract has been rewritten to comply with the 75-word limit, and no longer contains this phrase.

2) Line 80: The examples considered in the paper seek to estimate an exposure distribution that is assumed to be common across all subjects (given that exposure takes place). Hence the line ‘estimating individual and population exposures’ is somewhat confusing. Suggest being a little more specific w/r what is being estimated and how (e.g. without using any actual exposure data).

We have modified this section to be more explicit about the target and means of estimation.

3) Equation (3): it would be helpful to the reader to emphasize that throughout the rest of the paper, the authors will assume that the conditional exposure distribution given exposure takes place (#i=1) is the same for each subject, or that a_i=a and #_i=# for all i..

We have decided to omit the i subscript from these parameters altogether in order to avoid confusion, as per our response to major comment #2 from reviewer #1.

4) Line 178 (end of Results): it would be helpful to readers unfamiliar with estimating equation methods to emphasize why there’s no need to have exposure measurements to estimate the mean exposure level.

We have added some text along these lines to the beginning of the Discussion section.
5) Line 228 (6th paragraph of Algorithm): might mention that using this parameterization, # and # are the logged mean and logged variance, respectively, of the (conditional) exposure distribution.

We have added this text as suggested.

6) Line 229: Since D is not generally a vector, it might be better to say “..., EY and D are identical vectors...” rather than “the vectors EY and D are identical,...”

We have reworded this section as described in our response to major comment #2 from reviewer #1.

7) Lines 259-260 (first paragraph of Testing and Implementation): introducing still more notation gets confusing. Perhaps use exp(#) and exp(#) rather than μ and #g2.

We acknowledge the potential confusion regarding multiple parameterizations, but believe the results are more easily interpreted with the familiar mu and sigma^2 notation describing the mean and variance of the non-negative exposure magnitudes. We believe the reorganization and additional text explains this more clearly than in the previous draft.

8) Line 291 (end of Testing and Implementation): To be consistent with paragraph starting with line 242, perhaps say # and #2# here (recognizing that #2g is a function of #).

We have corrected the notation in this line to α, as we believe the reviewer suggested.

8) Line 295 (second line of Discussion): “...parameters to be randomly drawn...”

This sentence has been deleted.

9) Figure 2: Adding a reference line at ‘5’ would help the reader.

We have added appropriate reference lines to Figure 1 and Figure 2. We have also replaced all four figures with color versions that are easier to read.

Discretionary Revisions:
1) Some of the development of the estimating equations might be better left to an appendix

Because this paper is submitted in the methodology category and the main result is the estimating equations, we have left their development in the main body of the manuscript.
2) Line 177 (end of Results): It might not be too difficult to express the examples in terms of likelihoods, and maximize using ML methods, if the gamma exposures are independent within subject. In that case \( Y_{it} = Z_{it} + X_{it} \), where \( Z_{it} \) is a scaled (#it) gamma random variable and \( X_{it} \) is normal random variable. Comparing the performance of ML to the estimating equation methods could be interesting.

We have explored ML approaches to some extent and believe the convolution of weighted zero-inflated gammas to be more complicated than suggested by the reviewer (i.e., not resulting in a scaled gamma distribution for \( Z_{it} \)). Augmented data approaches, saddlepoint approximations, or other methods may be reasonable here, though. We may explore these other approaches in the future, but they are beyond the scope of the present manuscript.