Reviewer's report

Title: Estimation Methods with Ordered Covariate Subject to Measurement Error and Missingness in Semi-Ecological Design

Version: 1 Date: 10 May 2012

Reviewer: Lianne Sheppard

Reviewer's report:

Summary:
This paper proposes new methods to handle limited (i.e. incomplete or missing) mis-measured exposure data with natural application to occupational epidemiological studies where there is good insight that exposures to be used in the epidemiological study analysis vary strongly by job class. They wrestle with the challenge of improving the health effect estimate when the group mean exposures are poorly estimated. The authors assume that there is prior understanding about the ordering with respect to exposure of groups (i.e. by job class) and use this information to improve inference by relying on isotonic regression methods; this is particularly applicable to the case where there are limited exposure samples on individuals in the epidemiological study. The authors assume exposure is randomly distributed according to a mixed model with group means and subject-specific variances. They further assume that exposure measurements adhere to a classical measurement error model (which provides an additional source of independent normally distributed errors). They assess the impact of various estimation approaches on inference about a health effect parameter in linear and logistic disease models where the exposure effect is linear. They compare health effect estimates and their uncertainties in simulations and in an example dataset based on four approaches to estimation in the presence of missing and mis-measured exposure data: naïve, Monte Carlo EM (CEM), group-based strategy (GBS) and their new constrained group-based strategy (CGBS). They develop a constrained EM estimation approach incorporating isotonic regression for both the setting of measurement error only (i.e. the mis-measured exposure is available on all study subjects) and measurement error with missing covariates (i.e. where some/many subjects do not have any exposure measurements; their group (e.g. job class) is known so that quantification of exposure based on group is possible given data from other members of their group).

Overall I found this paper to be useful and informative. I believe it can be improved through revision. My major comments are:

• The model for the exposure needs to be revisited and fully aligned throughout the paper. There is implicit assumption of prior knowledge of the ordering of the group means that could be stated more clearly. The model for the group information isn’t handled consistently in the model parameterization. Referring to the model introduced on page 4 and the description of the simulation study on
page 9, it is unclear whether the data-generating mechanism for the exposure mean µg is completely random (with some unspecified mechanism to determine the ordering; see p. 4) or fixed (as implied on p. 9).

• The use of the term “covariate” needs to be clarified. As used in the section heading on p. 7, it seems to imply both W and X. In other places it implies only X (p. 2) or the reference is unclear (p. 3). It may be useful to use the word “exposure” in place of “covariate”, particularly since the so-called attributes (e.g. job class) can be considered covariates in the model for exposure.

• The methods section covers GBS, CGBS and CEM but the flow does not help the reader easily understand the distinctions between these and their linkages with the simulation study. In the methods section some readers might believe CEM is the approach to estimation for CGBS, but it distinct in the simulation study. It appears that GBS and CGBS condition on predicted group mean exposure in the regression analysis, with some correction for the specification bias (due to using the group mean instead of the individual exposure) in the logistic regression model. I suggest the authors distinguish the overall model formulation (given in the beginning of the methods section) from the estimation approaches and add some summarizing statements (e.g. at the end of each subsection on estimation approaches) to help readers with the flow.

• The discussion should incorporate insights gained into measurement error from work applied to air pollution epidemiology. For instance see work by Szpiro et al (2011a) and discussed further in Szpiro et al (2011b). Szpiro et al discuss their results in the context of Berkson-like and classical-like components to the measurement error. (The terminology “-like” is introduced because the errors are not purely Berkson or classical.) I would appreciate if the authors could discuss how those concepts apply to their work.

Discretionary revisions

• I suggest commenting on the effect of the isotonic regression. As shown in Table 1, it appears to collapse groups that do not adhere to the a priori order. How does this impact the Berkson- and classical-like components of the measurement error?

• The authors could provide additional insight into the results of their simulation and example by comparing the relative variability in the group means, subject-specific exposure variation, and measurement error variance.

• The example dataset has complexity that isn’t taken advantage of in the application. In particular many workers have repeat exposure measures, there is most likely variability between factories that isn’t acknowledged, and the analysis subset only uses workers with at least one exposure measurement. It would be helpful to add some comments acknowledging the impact of these features and possible extensions to this methods work to accommodate them.

Minor comments:

• P 5: Add a reference to PAVA.

• P. 7: I believe the authors mean that both X* and W* are all missing.
• P 11: Clarify the exposure data and ordering in the example dataset. Is the ordering in Figure 1 based on prior knowledge about the order of exposures across groups?

• P 11: It appears that the approach to estimating the standard error of betahat1 is different in the models with and without missing exposure data. Clarify if this is correct and convey insight as to why this is preferable to reporting the same quantities for analyses with and without missing exposures. It is surprising to note in Table 2 that for two of the three estimation approaches the standard errors of betahat1 for the full set of exposure data are larger than any of the cases with missing exposure data.

• The order of the results in the tables should be consistent across tables. In particular, CEM is the most advanced estimation approach and should appear in the final set of columns.

Responses to questions:

1. Is the question posed by the authors well defined?
   Yes

2. Are the methods appropriate and well described?
   The methods are appropriate and for the most part they are well described although I give suggestions above for better sign-posting of this section. I think the reference to the “covariate” is not well described and there is implicit understanding about the prior knowledge about the groups (e.g. job class) that is not incorporated into the paper. Additional attention to the isotonic regression approach will also be helpful to include.

3. Are the data sound?
   This is mostly a methods paper so consideration of the soundness of the data does not dominate. The application to carbon black exposure among workers in the carbon black manufacturing industry is appropriate but the example could be more thoroughly described in general. For the analysis in the paper there has been some simplification of the example that is appropriate for a methods paper, but it should be more clearly acknowledged that the simplification of the dataset and analysis to align the application with the methods implicitly makes the methods less generally applicable. While this is often the case for methods development work, the authors should include comments on this topic in their discussion.

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?
   Yes

5. Are the discussion and conclusions well balanced and adequately supported by the data?
   • I think the discussion should address where the knowledge about the exposure ordering comes; this topic is not addressed at all in the paper. For instance, on page 5 the authors state “if information on order of group means is correct”. This
should be taken up in the discussion.

• The authors could also put their work into better context with features of measurement error based on exposure modeling such as the ideas developed in recent work on air pollution exposure prediction for epidemiological inference. (see e.g. Szpiro et al 2011a,b)

6. Are limitations of the work clearly stated?
Greater attention to this topic is needed.

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?
Generally yes. Areas for improvement: The authors discuss the semi-ecological design briefly at the beginning of the paper but don’t mention any of the ecological/semi-ecological study literature. In particular I think they should acknowledge the specification bias property in nonlinear disease models (e.g. logistic) in conjunction with their presentation on the logistic model approximation. In addition there is new literature on measurement error that has emerged in the field of air pollution epidemiology that these authors may not be aware of; it should be cited and its insights incorporated into this paper.

8. Do the title and abstract accurately convey what has been found?
Yes although the abstract doesn’t really reflect the methodological focus of the paper.

9. Is the writing acceptable?
Generally yes although the paper can be improved by editing for grammar and clarity.

References

Level of interest: An article of importance in its field

Quality of written English: Needs some language corrections before being published

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:
I declare that I have no competing interests