Reviewer’s report

**Title:** Tweedie Distributions for Fitting Semicontinuous Health Care Utilization Cost Data

**Version:** 0  **Date:** 18 Aug 2017

**Reviewer:** Seamus Kent

**Reviewer’s report:**

I agree with the premise of the author that most studies that have compared statistical models for cost data focused only on positive realisations of costs, and have tended to somewhat blithely recommend the use of two-part models when confronted with zero-cost observations. However, zero-cost observations are very common, if not the norm, in healthcare cost data. I think this paper can offer a valuable addition to the literature, comparing two-part and single-distribution models on data with zero-cost observations. I do, however, have a number of concerns, particularly regarding the methods of model comparison, which I feel need to be addressed if this study is to make a useful contribution to the literature.

*References to previous studies*

This paper cites a number of previous studies that have compared statistical models for cost data, and noted that all had focused on only positive cost realisations. I think is by and large a fair characterisation of the literature. And I agree that it is important to compare models in the probably more common scenario of some zero-cost observations. The author has not, however, referenced a major paper in this literature that has addressed this question (Buntin and Zaslavsky, 2004). In particular, it compared single-equation and two-part models in Medical Expenditure Panel Survey data with 9% zero costs. This paper does not negate the contribution of the submitted paper, but may complement it. If the author is not aware of this paper, they should read it, and incorporate the lessons into their paper, in both the design (e.g. comparator models), and interpretation.

*Choice of comparison models*

In the paper by Buntin and Zaslavsky (2004) discussed above, the single-equation Poisson model was found to perform well despite the presence of zero-cost observations. I wonder whether this should be a candidate model in the present study.

More generally, the way in which candidate models were selected is not clear. I don't think, for instance, standard specification tests were performed to identify variance and link functions for the GLM model.
Perhaps the reason this wasn't done for the simulation study (but not the RAND data application) was because data was simulated from gamma distributions. Why were other distributions not considered?

*Assessment of model performance*

Models are compared in terms of AIC and root mean square error (RMSE). I wonder why the authors did not also consider mean error (ME) and mean absolute error (MAE), as is standard in this literature. These metrics provide useful information about bias and individual level predictive accuracy, both important qualities in modelling cost data, and would complement RMSE.

In the simulation study, rather than presenting absolute values of AIC and RMSE for the different models in the different scenarios, a ranking method is used. I am concerned that this ranking method gives the impression of greater differences between models than is, in reality, the case. I suspect that the RMSE values for different candidate models (perhaps with the exception of the Tobit) are very similar. This is certainly the case for the RAND data application, where values differ only to the first decimal place (except for the Tobit). I also think these differences are over interpreted on page 7, line 19 of the Discussion: "on the RAND HIE data, the tweedie model shows slightly better predictive accuracy". It's not clear these differences are at all important.

Relatedly, in much of the recent leading literature in this field, metrics of fit are compared not just at the aggregate or overall level, but across the distribution of data, defined by, for example, deciles of a linear predictor. I recognise the challenges of presenting such data, with other parameters varying, but this would give a richer pitcher of model fit.

Proportions of zero-cost observations range from 0.05 to 0.7. This choice is not explained. In many cost data sets, particularly of inpatient costs, the proportion of zero-costs can be larger.

The author states that they performed a "5-fold cross validation". No more description is given, but further description is required. The limitations of such validation should also be noted in the Discussion.

In the simulation study, all models were fitted with N = 5000. Even if this choice can be justified, it has not been in the paper. I understand the challenges of presenting results across multiple dimensions but it is important to understand whether the conclusions about the relative performance of the candidate models would be different with larger sample sizes.

Using RAND HIE data, models are compared in terms of marginal effects for the variables in the model. On page 6, lines 24-33, the author notes that "care must be taken when comparing the marginal effects of single distribution and two-part models" because the second part of the two-part models, model only positive costs. If marginally effects for only the second part of the two-part models is presented, then the results are simply incomparable and the utility of this exercise
is unclear to me. I wonder why marginal effects were not calculated combining the first and second parts of the two-part models. This would make the results comparable.

*Abstract & Introduction*

I think the introduction could be made somewhat clearer. For instance in the first sentence of the Introduction (pg1, line 43), the author writes: "...the non-negative response variable is usually right-skewed with a positive probability of zero outcome for non-users". It would make much more sense to turn this around and first state that there are zeros, and then focus on the positive realisations. In addition the construction "positive probability of zero outcome for non-users" is false. A similar construction appears elsewhere, for instance, in the abstract: "...model the probability of zero outcome for non-users of healthcare utilisation...". You are not modelling probability of zero costs for non-users but rather than probability of zero costs, i.e. of being a non-user.

*Other*

Add axis titles to figure 2. Also misidentified as figure 1 in text (page 6, line 46).

In describing the characteristics of the RAND data on page 5 lines 43-46, it may be useful to provide further details, including the standard deviation and skewness, and for both overall costs and only positive realisations of cost.

*References*


Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

Yes

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No
Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I am able to assess the statistics

Quality of written English
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interests.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.
I agree to the open peer review policy of the journal