Reviewer's report

Title: Grid multi-class response logistic models

Version: 2
Date: 23 August 2014

Reviewer: Ben Van Calster

Reviewer's report:

I find the topic of grid model fitting interesting, and it is of value to extend this to multiclass outcomes. However, in my opinion the paper should be much improved (better focus, better case study, more elaborate discussion of merits and limitations):

Major compulsory

1. The discussion is extremely short. It is awkward that the conclusion section is as long as the discussion. The discussion should definitely discuss practical aspects of grid model fitting: this paper deals with one single model fitting step, which is often not the case in practice. For example, issues often are variable selection, selection of transformations. In addition, Step 1 and Step 2 (cf p4) should be repeated until convergence, which means that local (site specific) investigators have to obtain gradient/Hessian matrices multiple times for one single model. This also causes practical concerns. What happens if the outcome variable cannot be shared (needed for performance evaluation)?

2. The case studies are uninteresting: there is no context/background given (what is the purpose of the study), the data are randomly split up to obtain multicenter data, and the samples are very small. They do not add to the paper in my view, but a real multicenter study of sufficient size definitely would.

3. I have serious concerns with two statements made on top of p7: (1) 'under the correct underlying model assumption ordinal logistic model fitting outperforms multinomial logistic model fitting when the sample size and number of covariates are the same'; (2) 'it is not appropriate to fit an ordinal logistic model for data generated from a multinomial logistic model ... however it is not too bad to fit a multinomial logistic model for data generated from an ordinal logistic model'.

Regarding (1), under the correct model assumption you are dealing with two completely different situations (one situation has data that follow an ordinal model, another has data that follow a multinomial model) so it does not make any sense to compare performance – it’s like comparing apples with pears.

Regarding (2), this is obvious. These statements do not add to the paper, and are in fact not dealing with the main scope of this paper.

4. P5 'to check the added value we need to compare the grid AUC score with the AUC score for each data subset': I do not think such a comparison can show the added value of grid computing. In the reported case studies, data is randomly split into subsets such that every subset has the same underlying population
(which is typically not the case in multicenter studies). Hence the lower subset AUCs, or the invalid estimates (cf p8), mainly show that the subsets are far smaller than the grid model's sample size.

Minor essential

5. On p1 the term 'distributed' data is used. The authors should be more clear what they mean with this; I assume something like multicenter data?

6. Section 'grid area under the ROC curve': the reported rationale of grid models seems to be the rationale for multicenter studies; the rationale of grid models seems to be related to privacy and confidentiality concerns? In any case, such statements belong in the Background/Introduction section.

7. Introduction of the logistic models (p2-3): there is a lot of technical information that is not needed in this paper and that in my opinion distracts from the main message. I would reserve technicalities for the section on grid multiclass models. That said, the IIA assumption for multinomial logistic regression is interesting but requires a description that is understandable from a medical point of view. I mainly find explanations of this assumption from the field of choice theory (where outcome categories can be present or absent on purpose), but this represents a very different application area of multiclass modeling: in medical applications alternatives are not present or absent at will.

8. P6. Please check whether references to the simulations studies 1 to 4 is always correct. E.g. the text says that the IIA test was checked for studies 1 and 4, and the proportional odds test for studies 3 and 4; I would say that IIA was checked in 2 and 3, and p.o. in 2 and 3? Later on p6 the term log odds log is used, I do not understand.

9. P7 'p=0.366, hence the assumption for the ordinal logistic model was valid'. This conclusion is too strong because power may well be far too low.

10. P8, SYMPT2 is used twice for different variables.

11. P8, the outcome ME seems ordinal instead of nominal: never, within a year, over a year. Why do the authors claim it is nominal?

12. P8 'both separate models produced invalid estimates': please provide the reason for this. Is it perhaps related to the distribution of ME (small numbers for some categories)?

13. P8, discussion, 'hence the covariate values cannot be protected either when there is only one covariate in the model': please provide more information. Is this the case because covariate values can be derived from Y and the model output?

Discretionary

14. P1, 'deidentified': do the authors mean 'coded' or 'anonymous'?
15. Background, 5th line: ‘phrase’ should be phase?

16. Background, 'check multiclass logistic model fitting': do you refer to the HL test or do you refer to checks for influential observations?

17. On p5 the authors seem to introduce an ad hoc ordinal AUC measure. This and other AUC measures for ordinal outcomes are described in Van Calster et al (Biom J, 2012).

18. Last sentence of the background section (p2) seems obsolete.

19. P7 'none of these models passed the HL test': what could be the reason? I guess it might be related to nonlinear effects of the continuous variable age, or to omitted interaction terms. (Of course sample size is too low to adequately check for this.)

20. P8, discussion: 'Cox proportional hazards model'.

21. P10, 'the HL test might be too strong and can produce too many false significant test results': several problems with the HL test have been published already.

**Level of interest:** An article whose findings are important to those with closely related research interests

**Quality of written English:** Acceptable

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

**Declaration of competing interests:**

I declare that I have no competing interests