Comment: Probabilistic index models

Olivier Thas, Jan De Neve, Lieven Clement and Jean-Pierre Ottoy

Wicher Bergsma, London School of Economics and Political Science
Marcel Croon, Tilburg University
Jacques A. Hagenaars, Tilburg University
Andries van der Ark, Tilburg University

We would like to point out the relation to Bergsma, Croon and Hagenaars (2009), where PIMs were introduced under the name of *Bradley-Terry type models*, and full maximum likelihood for fitting and testing with categorical variables was used. Below we also point out possible interpretational problems with certain PIMs, and how to avoid these.

We begin by giving a justification for the use of the probabilistic index. Consider a set of ordinal random variables $\{Y_i, i \in \mathcal{I}\}$ (not necessarily iid). Being ordinal, the Y_i are only meaningful comparatively, i.e., an individual Y_i has no meaning. However, a set of meaningful sufficient statistics is

$$\{\text{sign}(Y_i - Y_j) | i \neq j\}.$$

This suggests the use of

$$L_{ij} = E\,\text{sign}(Y_i - Y_j) = P(Y_i > Y_j) - P(Y_i < Y_j),$$

which is related to the probabilistic index via

$$\text{PI}_{ij} = (1 - L_{ij})/2.$$

In the notation of Thas et al, write $Y_i = (Y|\mathbf{X} = i)$ so that

$$\text{PI}_{ij} = P(Y < Y^*|\mathbf{X} = i, \mathbf{X}^* = j).$$

We see that models based on the L_{ij} or the PI_{ij} are truly ordinal, in contrast to, e.g., McCullagh's logistic models and normal threshold models, which assume ordinal data are realizations of some underlying interval level variable.

It might be tempting to interpret $L_{ij} > 0$ as "$Y_i > Y_j$". However, a problem is that it is possible that

$$L_{ij} > 0, L_{jk} > 0, \text{ and } L_{ik} > 0,$$

so $Y_i > Y_j$, $Y_j > Y_k$, and $Y_k > Y_i$, i.e., the inequality relation is intransitive. For PIM (31) in Thas et al, an intransitive solution arises if $\beta_1 = \beta_2 = \beta_4 = 0$, $\beta_3 > 0$, $\text{SES}_i = \text{SES}_j = \text{SES}_k$, and $\text{MI}_i > \text{MI}_j > \text{MI}_k > \text{MI}_i$.

Ideally, we would like to be able to interpret L_{ij} as a difference in location of Y_i and Y_j. However, this is not possible in general, since we may have

$$L_{ij} + L_{jk} \neq L_{ik}.$$

However, if the following Bradley-Terry type model holds,

$$L_{ij} = \lambda_i - \lambda_j \tag{1}$$

then

$$L_{ij} + L_{jk} = L_{ik},$$

and the λs can be interpreted as ordinal location parameters for the Ys. A regression model for the ordinal locations λ_i can then be formulated as

$$\lambda_i = \mathbf{X}_i^T \boldsymbol{\beta}. \tag{2}$$

More generally than (1), for a link g, we can consider

$$g(L_{ij}) = \lambda_i - \lambda_j. \tag{3}$$

Subsitution of (2) into (3) yields

$$g(L_{ij}) = (\mathbf{X}_i - \mathbf{X}_j)^T \boldsymbol{\beta},$$

which is a subclass of the PIMs considered by Thas et al. Note that, assuming (3) holds, our formulation (2) is easy to interpret and falls within the classical regression framework.

Bergsma, Croon and Hagenaars (2009) considered a very broad class of models, which includes PIMs, and derived multinomial likelihood equations. These equations apply to PIMs for the case that the response variable is categorical. However, the Lagrangian algorithm described there (and implemented in Bergsma and Van der Ark, 2009) appears to suffer from numerical problems when covariates are continuous. We wonder how a full likelihood method could be implemented for the continuous case.

Bergsma, Croon and Hagenaars (2009). *Marginal models for dependent, clustered and longitudinal categorical data.* Springer NY.
Bergsma and Van der Ark (2009). *CMM: Categorical marginal models.* R package