ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Small sample inference for probabilistic index models

G. Amorim [a,*], O. Thas [a,b], K. Vermeulen [a], S. Vansteelandt [c,d], J. De Neve [e]

[a] Department of Mathematical Modelling Statistics and Bioinformatics, Ghent University, Belgium
[b] National Institute for Applied Statistics Research Australia (NIASRA), School of Mathematics and Applied Statistics, University of Wollongong, Australia
[c] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
[d] Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine, United Kingdom
[e] Department of Data Analysis, Ghent University, Belgium

ARTICLE INFO

Article history:
Received 22 March 2017
Received in revised form 8 November 2017
Accepted 11 November 2017
Available online xxxx

Keywords:
Bootstrap
Empirical likelihood
Rank estimation

ABSTRACT

Probabilistic index models may be used to generate classical and new rank tests, with the additional advantage of supplementing them with interpretable effect size measures. The popularity of rank tests for small sample inference makes probabilistic index models also natural candidates for small sample studies. However, at present, inference for such models relies on asymptotic theory that can deliver poor approximations of the sampling distribution if the sample size is rather small. A bias-reduced version of the bootstrap and adjusted jackknife empirical likelihood are explored. It is shown that their application leads to drastic improvements in small sample inference for probabilistic index models, justifying the use of such models for reliable and informative statistical inference in small sample studies.

1. Introduction

Probabilistic index models were introduced by Thas et al. (2012) as a class of semiparametric models that can be used to complement rank tests with interpretable effect size measures, without the need to assume a location-shift model. Bergsma et al. (2009) described similar models, which they referred to as *Bradley–Terry type models*. For more details on the connection between probabilistic index models and their Bradley–Terry type models, we refer to Bergsma et al. (2012). Probabilistic index models can also be used to generate and extend many of the well-known rank tests such as, for example, the Wilcoxon–Mann–Whitney test, Kruskal–Wallis and Friedman tests (De Neve and Thas, 2015).

A probabilistic index model parameterizes the conditional probabilistic index

$$\text{pr}\left(Y \preccurlyeq Y^* \mid \boldsymbol{X}, \boldsymbol{X}^*\right) = \text{pr}\left(Y < Y^* \mid \boldsymbol{X}, \boldsymbol{X}^*\right) + 0.5\text{pr}\left(Y = Y^* \mid \boldsymbol{X}, \boldsymbol{X}^*\right),$$

in which Y and Y^* are outcomes associated with the covariates $\boldsymbol{X}$ and $\boldsymbol{X}^*$, respectively, with $(Y, \boldsymbol{X})$ and $(Y^*, \boldsymbol{X}^*)$ identically and independently distributed random vectors. The covariate may be vector-valued. The probabilistic index model is defined as

$$\text{pr}\left(Y \preccurlyeq Y^* \mid \boldsymbol{X}, \boldsymbol{X}^*\right) = m(\boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}) \quad (\boldsymbol{X}, \boldsymbol{X}^*) \in \mathcal{X}, \tag{1}$$

where $\boldsymbol{\beta}$ is a p-dimensional parameter vector and $m(\boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\beta})$ is a known function that satisfies $0 \leq m(\boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\beta}) \leq 1$, $m(\boldsymbol{X}, \boldsymbol{X}; \boldsymbol{\beta}) = 0.5$ and $m(\boldsymbol{X}, \boldsymbol{X}^*; \boldsymbol{\beta}) + m(\boldsymbol{X}^*, \boldsymbol{X}; \boldsymbol{\beta}) = 1$ for all $(\boldsymbol{X}, \boldsymbol{X}^*) \in \mathcal{X}$. The vector $\boldsymbol{Z}$ depends on the regressors,

* Correspondence to: Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium.
E-mail address: ggca@outlook.com (G. Amorim).

https://doi.org/10.1016/j.csda.2017.11.005
0167-9473/

e.g. $\boldsymbol{Z} = \boldsymbol{X}^* - \boldsymbol{X}$. The model thus takes the form of a generalized linear model by relating $\boldsymbol{Z}^T\boldsymbol{\beta}$ to the conditional probabilistic index through a known link function $g(\cdot)$. The set $\mathcal{X}$ is the subset of covariate pairs $(\boldsymbol{X}, \boldsymbol{X}^*)$ for which the probabilistic index model is defined.

Estimation of $\boldsymbol{\beta}$ in (1) was discussed in detail in Thas et al. (2012). Given a random sample of identically and independently distributed $(Y_i, \boldsymbol{X}_i)$, $i = 1, \ldots, n$, the outcomes are first transformed to so-called pseudo-observations I_{ij} defined as 1 if $Y_i < Y_j$, 1/2 if $Y_i = Y_j$ and 0 otherwise $(i, j = 1, \ldots, n)$. For each pair $(\boldsymbol{X}_i, \boldsymbol{X}_j)$ the vector $\boldsymbol{Z}_{ij}$ is constructed. An estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, is then obtained by solving $\boldsymbol{U}_n(\boldsymbol{\beta}) = \boldsymbol{0}$, with

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n}\boldsymbol{U}_{ij}(\boldsymbol{\beta}) = \frac{1}{|\mathcal{I}_n|}\sum_{(i,j)\in\mathcal{I}_n}\boldsymbol{A}\left(\boldsymbol{Z}_{ij};\boldsymbol{\beta}\right)\left\{I_{ij} - g^{-1}\left(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta}\right)\right\}, \tag{2}$$

where $\boldsymbol{A}\left(\boldsymbol{Z}_{ij};\boldsymbol{\beta}\right)$ is a p-dimensional vector function of the regressors $\boldsymbol{Z}_{ij}$ and the parameter vector $\boldsymbol{\beta}$, and the sum is limited to all pairs (i, j) for which $(\boldsymbol{X}_i, \boldsymbol{X}_j)$ are in $\mathcal{X}$ (the set of such index pairs is denoted by $\mathcal{I}_n$ and has $|\mathcal{I}_n|$ elements). Under mild regularity conditions, the solution $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normal (Thas et al., 2012; De Neve, 2013), with covariance matrix that can be consistently estimated by a sandwich estimator, which is given by

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \left\{\sum_{(i,j)\in\mathcal{I}_n}\frac{\partial\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}^T}\right\}^{-1}\left\{\sum_{(i,j)\in\mathcal{I}_n}\sum_{(k,l)\in\mathcal{I}_n}\phi_{ijkl}\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}})\boldsymbol{U}_{kl}^T(\hat{\boldsymbol{\beta}})\right\}\left\{\sum_{(i,j)\in\mathcal{I}_n}\frac{\partial\boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}})}{\partial\boldsymbol{\beta}^T}\right\}^{-1},$$

where ϕ_{ijkl} is an indicator variable taking value 1 if the pseudo-outcomes I_{ij} and I_{kl} share an index, and 0 otherwise.

Simulation studies in Thas et al. (2012) and De Neve (2013) confirm the asymptotic distribution theory. However, even with only two regressors, their results also indicate that $\hat{\boldsymbol{\beta}}$, the sandwich estimator and the coverages of asymptotic Wald-based confidence intervals are only reliable for sample sizes of 50 or more. This is an important limitation, particularly because the above methods provide natural extensions of rank tests by, for example, allowing inference for treatment effects while controlling for covariates (De Neve and Thas, 2015; Vermeulen et al., 2015). Only for the special case of a Wilcoxon-rank sum test in randomized experiments, a covariate adjustment, based on a probabilistic index model, has been proposed for which permutation p-values are available (Vermeulen et al., 2015).

To overcome the aforementioned limitations, we explore methods that are designed to give better small sample results. Resampling techniques, such as the bootstrap and jackknife, are often used as alternative approaches to increase accuracy in many statistical applications (Basu, 2001). However, they sometimes require strong computational power. For instance, direct application of the traditional non-parametric bootstrap to probabilistic index models requires solving B times (number of bootstrap samples) the nonlinear estimating equation $\boldsymbol{U}_n(\boldsymbol{\beta}) = \boldsymbol{0}$ for $\boldsymbol{\beta}$. Even solving the equation only once may already be computationally demanding because fitting a probabilistic index model requires modelling the pseudo-observations, resulting in an inflated number of estimating functions. We solve this issue by applying the bootstrapping U-statistics method of Jiang and Kalbfleisch (2012). This method simplifies the computational demands by resampling properly studentized terms from an asymptotic approximation of the estimating function that is a U-statistic of degree 1 or 2. It hence avoids the need to repeatedly solve nonlinear equations and our simulation results show that the resulting coverages are often close to the nominal values.

In addition to bootstrap, we also use methods based on empirical likelihood to improve small sample inference for probabilistic index models. The empirical likelihood method (Owen, 1988, 1990) maximizes a non-parametric likelihood subject to restrictions given by the estimating equations. The ratio of the maximized empirical likelihood over the maximized nonparametric likelihood, which corresponds to an unconstrained model, is known as the empirical likelihood ratio statistic for which a Wilks' theorem needs to be proven. Just as for the parametric likelihood ratio statistic, this Wilks' theorem gives the asymptotic distribution under the hypothesis that the constrained model holds true. Confidence intervals of the parameters are subsequently found by inverting the empirical likelihood ratio test. For U-statistics, the model restrictions are non-linear, leading to a computationally expensive estimation process. The jackknife empirical likelihood method (Jing et al., 2009) reduces this computational cost by rewriting the U-statistic as a sum of asymptotically linear independent terms, making the constraints linear. Its use for probabilistic index models was also suggested by Zhou (2012). However, this method, as any empirical likelihood method, requires that the constraints always have a solution, which is not necessarily true. Chen et al. (2008) proposed to adjust the empirical likelihood by including an artificial "pseudo-observation" so that a solution can always be obtained and Zhao et al. (2015) later adapted this approach to the jackknife empirical likelihood setting. In this paper we further adapt the adjusted jackknife empirical likelihood to the probabilistic index model setting and evaluate its performance in simulation studies. This method performs generally well, but may be strongly affected by finite-sample bias. To alleviate this problem, we propose a bias-reduced adjusted jackknife empirical likelihood approach that shows good empirical results for samples as small as 20 and with coverages close to the nominal values.

2. Bias-reduced estimator for probabilistic index models

Consider the estimating function (2) with

$$\boldsymbol{A}\left(\boldsymbol{Z}_{ij};\boldsymbol{\beta}\right) = \frac{\partial}{\partial\boldsymbol{\beta}}m\left(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta}\right)V^{-1}\left\{m\left(\boldsymbol{Z}_{ij}^T\boldsymbol{\beta}\right)\right\}, \tag{3}$$

Please cite this article in press as: Amorim G., et al., Small sample inference for probabilistic index models. Computational Statistics and Data Analysis (2018), https://doi.org/10.1016/j.csda.2017.11.005.