Computational Statistics and Data Analysis 104 (2016) 110–129

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Semi-parametric copula sample selection models for count responses

Giampiero Marra, Karol Wyszynski [*]

Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK

ARTICLE INFO

Article history:
Received 11 December 2015
Received in revised form 4 May 2016
Accepted 13 June 2016
Available online 5 July 2016

Keywords:
Non-random sample selection
Copula
Penalized regression spline
Count response

ABSTRACT

In observational studies, a response of interest (as well as some individual level characteristics) may be observed for a non-randomly selected sample of the population. In this situation, standard models such as linear and probit regressions will yield biased and inconsistent parameter estimates. Selection models can address this issue and mainly consist of two regressions: a binary selection equation which determines whether the statistical units will enter the sample, and an outcome equation which models the response. While sample selection models for continuous and binary outcomes have been widely studied in the literature, the case of count response has not received as much attention. Sample selection models for count data which allow for the use of potentially any discrete distribution, non-Gaussian dependencies between the selection and outcome equations, and flexible covariate effects are introduced. The estimation algorithm is based on the penalized likelihood estimation framework. The method is illustrated in simulation and using data from a United States Veterans Administration Survey.

1. Introduction

Non-random sample selection arises when individuals select themselves into (out of) the sample based on features that are observed and unobserved. In this case, statistical analysis based on commonly known models such as linear and probit regressions will yield biased and inconsistent parameter estimates. One way of addressing this issue is to employ sample selection models.

The motivating example of this work stems from data collected through the 2001 United States Veterans Administration Survey (USVA, 2001). Here, the interest is in estimating the impact of certain observed patients' characteristics on number of visits in Veterans Administration (VA) and non-VA medical facilities and the predicted average number of visits (Lahiri and Xing, 2004). This study is challenging because of the likely presence of relevant unobserved factors (e.g., attitude towards health related risks); neglecting the difference in the unobserved attributes of the individuals who used the facilities and those who did not use them may have an adverse effect on parameter estimation (Lahiri and Xing, 2004; Trivedi and Zimmer, 2007, p. 76). In this case, an appropriate approach such as the sample selection methodology is required to obtain consistent estimates.

Sample selection models, also known as Heckman-type models, were first introduced by Gronau (1974), Lewis (1974) and Heckman (1976), and discussed more thoroughly in Heckman (1979). They typically consist of a selection equation and an outcome equation. The former models whether an observation will be missing on the response (e.g., decision to use

[*] Corresponding author.
E-mail addresses: giampiero.marra@ucl.ac.uk (G. Marra), k.m.wyszynski@gmail.com (K. Wyszynski).

http://dx.doi.org/10.1016/j.csda.2016.06.003

the facilities) and is usually achieved using a probit regression. The latter models the response of interest (e.g., number of hospital visits) and the type of regression employed typically depends on the nature of the response. The two equations are allowed to be associated which will be crucial when non-random sample selection is based on unobservables. The literature on models tackling selection bias is vast and without claim of exhaustiveness we mention below some interesting variants. Chib et al. (2009) and Wiesenfarth and Kneib (2010) introduced Bayesian sample selection models which allow researchers to flexibly estimate covariate effects, whereas a frequentist counterpart was introduced by Marra and Radice (2013a). Li (2011) considered the situation in which there is more than one selection mechanism, and Omori and Miyawaki (2010) extended selection models to allow threshold values to depend on individuals' characteristics. These models have also been compared to principal stratification in the context of causal inference with non-ignorable missingness (Mealli and Pacini, 2008). Liu et al. (2012) employed sample selection models based on a three equation system, whereas (Marra and Radice, 2013b) focused on binary outcomes. Greene (1997), Terza (1998) and Miranda and Rabe-Hesketh (2006) discussed the case of count responses; the approaches of these authors have the main drawbacks that (typically computationally expensive) quadrature or simulation methods have to be employed to obtain certain key quantities required for model fitting, data-driven semi-parametric effects are not allowed for and, generally speaking, the unconditional distribution of the response of the outcome equation is unknown. Taking a different view to the problem of non-observable response, it is possible to assign a zero value to the outcome whenever an observation on it was not generated (e.g., the individual did not use the facilities) and assume that such a value is "genuine". In this case, two-part and zero inflated models may, for example, be employed (see Humphreys, 2013; Lambert, 1992 and references therein). The former consists of fitting two regressions, one for modeling the occurrence of zeros and the other for modeling non-zero values. The latter uses a mixture of Bernoulli and discrete distributions which accounts for excess of zeros. These approaches essentially aim at modeling zero and non-zero values, all of which observed, instead of dealing with missing observations on the response which is when selection models are most useful. Therefore, the key question for the researcher is to determine why zero values are present in the response and then choose an appropriate methodology.

Selection approaches that rely on the commonly used bivariate normality assumption are often criticized since if this fails to hold then the resulting estimator will not yield consistent estimates (e.g., Pigini, 2012; Smith, 2003). The literature offers some alternatives to the assumption of Gaussianity, including non/semi-parametric and copula approaches. The former tend to be computer-intensive and typically do not allow for much flexibility in the model specification (compare: Pigini, 2012). Furthermore, convergence problems may arise when fitting models with several types of covariate effects (Wojtyś et al., 2016). The copula approach is more feasible as it uses maximum likelihood techniques. It also allows for simultaneous estimation of all model parameters which may lead to important efficiency gains (Smith, 2003). However, it may deliver estimators that are not consistent when the distributional assumption is not correct. Nevertheless, copulae allow for a piece-wise model specification and for many modeling options; for instance, it is possible to use any two marginal distributions when the copula linking them is Joe or Clayton. This is advantageous as the user can assess the sensitivity of results to different modeling assumptions. Genius and Strazzera (2008) pointed out that the copula approach allows for direct estimation of the dependence between the two equations, while non/semi-parametric methods do not. Hasebe and Vijverberg (2012) established a sample selection model based on copulae and the Generalized Tukey Lambda (GTL) marginal distribution. The authors argue that GTL is an appealing choice as it allows for skewness and thin and heavy tailed response behavior. Marchenko and Genton (2012) developed the selection-t model in which the errors are modeled using a bivariate t-distribution. Finally, it is worth mentioning the works of Cameron and Trivedi (2005), Cameron and Trivedi (2013) and Cameron et al. (2004) who exploited copulae for modeling count data in various contexts including non-random selectivity.

The practical relevance of the selection approach is supported by the number of hits received by Google Scholar when typing "sample selection model" (over 1200 hits since 2014, the majority of which relate to applied articles and reports). This paper contributes to the literature by introducing a flexible copula-based sample selection modeling approach for count data. The proposed method allows for the use of several (copula) dependence structures and (potentially) any discrete outcome margin (as long as the probability mass function (pmf) and cumulative distribution function (cdf) are known). Covariate effects are flexibly determined by the data using, for instance, penalized thin plate regression splines or P-splines commonly known in the context of Generalized Additive Models (e.g., Wood, 2006). The proposed framework is termed as semi-parametric (following the typical convention adopted in the statistical modeling literature when covariate effects are estimated flexibly) and is not affected by the aforementioned drawbacks of available selection approaches for count data. Previous works on selection models have considered separately the use of copulae, semi-parametric covariate effects and discrete distributions. This paper brings together these strands of research which required a considerable methodological effort and some careful structuring when implementing the proposed class of models. Moreover, our approach allows one to model distribution specific parameters as functions of semi-parametric effects as advocated by Stasinopoulos and Rigby (2005) in the context of univariate generalized additive models. To the best of our knowledge, this development has never been considered in the context of the models introduced in this paper.

The remainder of this paper is organized as follows. Section 2 introduces the models and likelihood. Section 3 discusses parameter estimation which is based on penalized maximum likelihood, whereas Section 4 briefly mentions how to construct confidence intervals and carry out model selection. Section 5 presents some simulation results, and Section 6 illustrates the framework using USVA data. Concluding remarks are given in Section 7. All developments are implemented in the `SemiParSampleSel` package (Marra et al., 2016) for the `R` environment (R Development Core Team, 2016).