Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 162 (2005) 387–402
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/dsw

An algorithm based on discrete response regression models suitable to correct the bias of non-response in surveys with several capture tries

Carmen Anido [a], Carlos Rivero [b], Teófilo Valdés [b,*]

[a] *Departamento de Análisis Económico: Economía, Cuantitativa, Facultad de CC Económicas y Empresariales, Universidad Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain*
[b] *Departamento de Estadística e Investigación, Operativa, Facultad de CC Matemáticas, Universidad Complutense de Madrid, Plaza de las Ciencias 3, 28040 Madrid, Spain*

Received 22 February 2002; accepted 2 June 2003
Available online 3 December 2003

Abstract

The use of survey plans, which contemplate several tries or call-backs when endeavouring to capture individual data, may supply unarguable information in certain sampling situations with non-ignorable non-response. This paper presents an algorithm whose final aim is the estimation of the individual non-response probabilities from a general perspective of discrete response regression models, which includes the well known probit and logit models. It will be assumed that the respondents supply all the variables of interest when they are captured. Nevertheless, the call-backs continue, even after previous captures, for a small number of tries, r, which has been fixed beforehand only for estimating purposes. The different retries or call-backs are supposed to be carried out with different capture intensities. As mentioned above, the response probabilities, which may vary from one individual to another, are sought by discrete response regression models, whose parameters are estimated from conditioned likelihoods evaluated on the respondents only. The algorithm, quick and easy to implement, may be used even when the capture indicator matrix has been partially recorded. Finally, the practical performance of the proposed procedure is tested and evaluated from empirical simulations whose results are undoubtedly encouraging.

Keywords: Multivariate statistics; Estimation algorithms; Discrete response models; Non-ignorable non-response; Conditional likelihood

1. Antecedents and preliminaries

This paper has as precedents the works of Politz and Simmons (1949), Drew and Fuller (1980), Särndal (1980), Huggins (1989) and Alho (1990) where several approaches to non-response under certain sampling

* Corresponding author. Tel.: +34-91-3944422; fax: +34-91-3944606.
E-mail address: teofilo_valdes@mat.ucm.es (T. Valdés).

0377-2217/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2003.06.031

plans with call-backs (or repeated retries in the capture of data) being present. In addition to the former direct references, several papers within the statistical bibliography on missing data have dealt with different situations of non-ignorable non-response in sample plans, e.g. the papers included in the three volumes of the National Academic of Sciences on *Panel on Incomplete Data* (Madow et al., 1983; Madow and Olkin, 1983; and finally, Madow et al., 1983) which provided, at the time, a cunning vision on the topic. Later on, Glynn et al. (1986) and Little and Rubin (1987, ch. 12) also presented different situations of non-ignorable missing data in surveys.

No distributional hypotheses on the variables under study need to be established in this work. We have only assumed the existence of r retries of information capture, which are maintained with all the individuals included in the sample no matter whether their information had or had not been previously captured. In the former case, the call-backs are not intended to supply data that has previously been obtained. On the contrary, we seek to be provided with some additional information about the underlying probabilities of capture. Only under this criterion the call-backs are not superfluous and the available information seems to be much more useful than that which could be obtained if the tries were only kept on as far as the first capture of the individual data. Usually, in captures by telephone or mail services, the additional cost derived from the maintenance of the retries is minimal, thus the described situation appears in fact to be quite viable. Additionally, it will be assumed throughout this paper that the available information about the capture results is only partially known, although it does include, as a minimum, all the marginal values of the complete capture indicator matrix. The non-response probabilities will be estimated by general discrete response models defined on the survey variables. Thus, some well known models, such as logit and probit, may be considered as particular cases to be included within the field of interest of this paper, needing only to adjust the iterative estimating procedure proposed here. The parameters involved in the selected discrete response model will be estimated by means of the conditional likelihoods evaluated on the respondents only. A similar scheme of estimation has been used, for instance, in Sanathanan (1972), Huggins (1989) and Alho (1990). Finally, the estimated probabilities will be used to treat the non-ignorable non-response by means of Horvitz–Thompson type estimates which, as it is well known, simply weigh each observation with the inverse of the overall response probabilities.

We have organised this paper as follows. In Section 2, we describe the missing data mechanism, the motor estimates and the conditioned likelihoods, which underlie our final estimates. Section 3 is completely devoted to justifying the iterative estimating algorithm proposed here. We start by analysing the case in which the capture indicator matrix is completely known and, later on, we readjust the steps so as to consider the situation in which the former matrix is only partially known. In Section 4, we incorporate several simulation studies whose results empirically show the performance of the iterative estimating process in three particular discrete models (to wit, logit, probit and with double exponential distribution function). Finally, Section 5 briefly includes some final observations and comments.

2. Non-response process and estimating background

The individuals selected in the sample will be identified by the set of indices $I = \{1, \ldots, i, \ldots, n\}$. Let us assume next that the sample design allows us to establish the probability π_i that each individual i of the population has to be included in the sample. The data vector of the individual i will be denoted by $x_i = (x_{i1}, \ldots, x_{ip})^{\mathrm{T}}$. Thus, in absence of non-response, the $n \times p$ data matrix is

$$X = (x_1^{\mathrm{T}}, \ldots, x_n^{\mathrm{T}})^{\mathrm{T}}$$

and the unbiased Horvitz–Thompson estimates of the population mean vector $\mu = (\mu_1, \ldots, \mu_p)$ and the mean cross-products matrix $\Lambda = N^{-1}\Sigma_i^N x_i x_i^{\mathrm{T}}$ are

$$\overline{X} = N^{-1}X^{\mathrm{T}}\Pi^{-1}1_n, \quad \widehat{\Lambda} = N^{-1}X^{\mathrm{T}}\Pi^{-1}X,$$