Computational Statistics and Data Analysis 103 (2016) 124–138

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A covariate nonrandomized response model for multicategorical sensitive variables☆

Heiko Groenitz *

Philipps-University Marburg, School of Business and Economics, Working group Statistics, Universitätsstraße 25, 35032 Marburg, Germany

ARTICLE INFO

Article history:
Received 28 November 2014
Received in revised form 21 April 2016
Accepted 22 April 2016
Available online 7 May 2016

Keywords:
Answer refusal
EM algorithm
Fisher scoring
Generalized linear model
Multivariate logistic regression
Untruthful answers

ABSTRACT

The diagonal method (DM) is an innovative technique to obtain trustworthy survey data on an arbitrary categorical sensitive characteristic Y^* (e.g., income classes, number of tax evasions). The estimation of the unconditional distribution of Y^* from DM data has already been shown. Now, a covariate extension of the DM, that is, methods to investigate the dependence of Y^* on nonsensitive covariates, is sought. For instance, the dependence of income on gender and profession may be under study. The covariate extensions of privacy-protecting survey designs are broadened by the covariate DM, especially because existing methods focus on binary Y^*. LR-DM estimation and stratum-wise estimation are described, where the former is based on a logistic regression model, leads to a generalized linear model, and requires computer-intensive methods. The existence of a certain regression estimate is investigated. Moreover, the connection between efficiency of the LR-DM estimation and the degree of privacy protection is studied and appropriate model parameters of the DM are searched. This problem of finding suitable model parameters is rarely addressed for privacy-protecting survey methods for multicategorical Y^*. Finally, the LR-DM estimation is compared with the stratum-wise estimation. MATLAB programs that conduct the presented estimations are provided as supplemental material (see Appendix E).

1. Introduction

Let us assume a survey containing a sensitive characteristic such as financial situation, evaded taxes, insurance or social fraud, cheating in examinations, political preferences, or discrimination against women in the job is to be conducted. For instance, the characteristic income is often relevant in social surveys on living conditions (e.g., the German General Social Survey and the Socio-Economic Panel) and in market research whereas the question of whether a person is employed but not registered is of interest in studies on the loss through moonlighting. Then, however, direct questions like "How much do you earn?" or "Have you ever employed illicit workers?" are not recommended, because they induce answer refusal (i.e., missing values) or untruthful answers. Consequently, a serious bias will occur if the distribution of a sensitive variable is estimated based on the responses to direct questions.

This problem motivates the application of survey designs that protect the respondents' privacy. The first privacy-protecting survey method was the randomized response (RR) model by Warner (1965). Today comprehensive literature

☆ Supplemental material for this article is available online (see Appendix E).

* Tel.: +49 6421 2823742; fax: +49 6421 2823713.
E-mail address: groenitz@staff.uni-marburg.de.

http://dx.doi.org/10.1016/j.csda.2016.04.007
0167-9473/

on RR methods exists (for an overview, see, for example, Chaudhuri, 2011). In RR procedures, the respondents must conduct a random experiment with the help of a randomization device. An alternative to RR designs is the item count technique (ICT). Nearly the complete literature on the ICT concentrates on binary sensitive questions (e.g., the question on illicit workers as mentioned above). For such binary questions, the basic principle of the ICT is that the respondents receive a list with some nonsensitive questions and one sensitive question and reveal only the number of questions that must be answered with a yes. For more details on the ICT, see, e.g., Blair and Imai (2012) or Groenitz (2014c) and the references therein. Another alternative to RR protocols is the class of nonrandomized response (NRR) methods, which require a scrambled answer depending on the sensitive attribute and a nonsensitive auxiliary characteristic (e.g., "When is your birthday?"). The NRR approach does not need a random experiment conducted by the respondents. The absence of such a procedure causes a reduction in survey complexity.

The literature on NRR models has been growing in recent years (see, for example, Tian and Tang, 2014). When multicategorical sensitive characteristics are intended to be studied, the nonrandomized diagonal method (DM) by Groenitz (2014a) can be used. The DM is applicable to both variables with at least one nonsensitive category (e.g., the variable "number of illicit workers" with categories 0, 1–2, 3–5, more than 5) and variables which are sensitive as a whole (e.g., income divided into classes, political preferences). The DM facilitates interviewees' cooperation and has a simple procedure. The estimation of the distribution of a sensitive variable from DM data is derived in Groenitz (2014a,b). Additionally, the exploitation of prior information by Bayes methods is explained in Groenitz (2015). The aim of this current paper is to derive a covariate extension of the DM. This means that we present methods that enable the analysis of the dependence of the sensitive attribute on nonsensitive covariates. In other words, we develop procedures to investigate the influence of nonsensitive explanatory variables on the sensitive variable. For example, the dependence of income on age and profession or the reasons (clarity of laws, inspections, punishments) that promote or oppose the employment of illicit workers might be under study.

This covariate extension of the DM is worthwhile, especially because the existing literature on covariate extensions of privacy-protecting survey designs focuses on binary sensitive variables, as the following overview shows. Regarding covariate extensions of RR models, the first contribution can be found in the book of Maddala (1983, pp. 54–56), who analyzes the relation between nonsensitive exogenous variables and a binary sensitive variable where data on the sensitive attribute are gathered by the RR technique according to Boruch (1971). The paper by Scheers and Dayton (1988) extends the RR model by Warner (1965) and the unrelated question model (see Greenberg et al., 1969) with covariates. It also contains a real-data study on the influence of the grade point average on academic cheating behavior. The work by van der Heijden and van Gils (1996) presents a covariate version of the RR method by Kuk (1990). Van den Hout et al. (2007) deal with the analysis of the influence of covariates on multiple binary sensitive characteristics and their association based on RR data. They also present a real-data example regarding social benefit fraud, more precisely, the illegal receipt of unemployment benefit in the Netherlands. Further real-life investigations on determinants of social fraud based on RR surveys are available in van der Heijden et al. (2000) and in Lensvelt-Mulders et al. (2006). Ostapczuk et al. (2009) investigate the education effect in attitudes towards foreigners where a certain RR technique is involved in the survey. Interesting recent papers on the analysis of the influence of explanatory variables on sensitive attributes where an ICT is applied are, for example, Imai (2011) and Kuha and Jackson (2014). The first addresses racism, the second deals with the purchase of stolen goods. In the field of NRR designs, Jann et al. (2012) present a covariate extension of the crosswise model by Yu et al. (2008) and study factors affecting plagiarism.

This article continues with a review of the DM in Section 2. In Section 3, we consider a multicategorical sensitive $Y^* \in \{1, \ldots, k\}$ and nonsensitive characteristics $X_1^*, \ldots, X_p^*$ where we assume that the DM is applied to elicit information about Y^*. The aim of Section 3 is to develop methods to investigate the influence of $X^* = (X_1^*, \ldots, X_p^*)$ on Y^*. For this, we present the LR-DM estimation and a stratum-wise estimation. LR-DM inference is based on a logistic regression model (LRM), leads to a certain generalized linear model, and requires computer-intensive methods (e.g., a special Fisher scoring algorithm). In Section 4, important features of the derived methods are shown. First, we discuss unlike in other papers on covariate extensions of privacy-protecting survey methods the existence of a maximum likelihood estimate by considering the convergence behavior of a Fisher scoring algorithm and an expectation maximization (EM) algorithm (Section 4.1). Subsequently, in Section 4.2, we present a simulation-based method to analyze the important relation between efficiency of the estimation based on a LRM and the degree of privacy protection. Here, we give practical steps on how to choose the specifications of the DM. Such statements can rarely be found in the existing literature on privacy-protecting survey methods for multicategorical sensitive variables. Finally, in Section 4.3, we compare the efficiency of the estimation based on a LRM with the efficiency of the stratum-wise estimation.

To support other researchers and practitioners, we provide self-created MATLAB programs that compute estimates via Fisher scoring and the EM algorithm as supplemental material (see Appendix E).

2. Diagonal method

Groenitz (2014a) proposes a nonrandomized response technique for multichotomous sensitive variables, namely the diagonal method. This procedure enables the estimation of the distribution of a sensitive characteristic Y^* with codomain $\{1, \ldots, k\}$ by the frequencies of certain nonrandomized answers A^*, which depend on an auxiliary variable $W^* \in \{1, \ldots, k\}$. The auxiliary variable is assumed to be nonsensitive and independent from Y^* with a known distribution. Moreover, we assume that the interviewer does not know the respondents' values for W^*. For instance, W^* could describe the birthday of