Available online at www.sciencedirect.com
SCIENCE DIRECT®
Computational Statistics & Data Analysis 50 (2006) 3311–3323
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Estimating the linear regression model with categorical covariates subject to randomized response

Ardo van den Hout[a,*], Peter Kooiman[b]

[a] *Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. box 80140, 3508 TC Utrecht, The Netherlands*
[b] *Department Labour Market and Welfare State, CPB Netherlands Bureau for Economic Policy Analysis, P.O. Box 80510, 2508 GM Den Haag, The Netherlands*

Received 22 July 2004; received in revised form 16 June 2005; accepted 17 June 2005
Available online 19 July 2005

Abstract

The maximum likelihood estimation of the iid normal linear regression model where some of the covariates are subject to randomized response is discussed. Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly. RR variables are described as misclassified categorical variables where conditional misclassification probabilities are known. The likelihood of the linear regression model with RR covariates is derived and a fast and straightforward EM algorithm is developed to obtain maximum likelihood estimates. The basis of the algorithm consists of elementary weighted least-squares steps. A simulation example demonstrates the feasibility of the method.

Keywords: EM algorithm; Linear regression model; Misclassification; Post randomization method; Randomized response; Statistical disclosure control

1. Introduction

Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly (Warner, 1965;

* Corresponding author. Tel.: +30 2539237; fax: +30 2535797.
E-mail address: a.vandenhout@fss.uu.nl (A. van den Hout).

0167-9473/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2005.06.008

Chaudhuri and Mukerjee, 1988). Examples of sensitive questions are questions about alcohol consumption, sexual behavior or fraud. RR variables can be seen as misclassified categorical variables where conditional misclassification probabilities are known. The misclassification protects the privacy of the individual respondent.

This paper applies the ideas in Spiegelman et al. (2000) to iid normal linear regression models where some of the covariates are subject to RR. Spiegelman et al. (2000) discuss the logistic regression model with misclassified covariates and estimate the misclassification using main study/validation study designs. The misclassification model of an RR design, however, is different since conditional misclassification probabilities are known to the analyst of an RR data set. This paper specifies the misclassification model of RR and shows how the misclassification can be taken into account in the maximum likelihood estimation of the linear regression model. Furthermore, as an alternative to Newton–Raphson maximization of the likelihood function an EM algorithm (Dempster et al., 1977) is presented.

There is quite some literature about RR and the adjustment for the misclassification in the analysis, see, e.g., probability estimation in Chaudhuri and Mukerjee (1988), Bourke and Moran (1988), Moors (1981) and Migon and Tachibana (1997), the logistic regression model with a RR dependent variable in Maddala (1983), and loglinear models in Chen (1989) and Van den Hout and Van der Heijden (2004). RR variables as covariates, however, have not been dealt with. The possibility to include RR variables in regression models enlarges the possible application of RR. As an example, consider the situation where one variable depends on a second variable that models sexual behavior. If respondents are reluctant to answer about their behavior directly, RR can be used. In that case, a standard regression model is incorrect since it does not take into account the misclassification due to the use of RR.

A second field that may benefit from the discussion in this paper is statistical disclosure control. There is a similarity between RR designs and the post randomization method (PRAM) as a method for disclosure control of data matrices, see Gouweleeuw et al. (1998). Disclosure control aims at safeguarding the identity of respondents after data have been collected, see, e.g., Bethlehem et al. (1990). If privacy is sufficiently protected, data producers, such as national statistical institutes, can safely pass on data to a third party. The idea of PRAM is to misclassify some of the categorical variables in the original data matrix and to release the perturbed data together with information about the misclassification mechanism. In this way, PRAM introduces uncertainty in the data, i.e., the user of the data cannot be sure whether the individual information in the matrix is original or perturbed due to PRAM. Since the variables that are perturbed are typically covariates such as, e.g., Gender, Ethnic Group, Region, it is important to know how to adjust regression models in order to take into account the misclassification. PRAM can be seen as a specific form of RR and the idea to use RR in this way goes back to the founder of RR, see Warner (1971). Similarities and differences between PRAM and RR are discussed in Van den Hout and Van der Heijden (2002). Domingo-Ferrer and Torra (2001) compare PRAM with other methods for disclosure control and Willenborg and De Waal (2001, Chapter 5) discuss the derivation of misclassification probabilities by means of linear programming.

The outline of the paper is as follows. Section 2 introduces the RR model. Section 3 discusses the linear regression model with RR covariates. In Section 4, an EM algorithm is presented that maximizes the likelihood formulated in Section 3. Section 5 discusses