Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 6060–6069

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

The logistic regression model with response variables subject to randomized response

Ardo van den Hout[a,*], Peter G.M. van der Heijden[b], Robert Gilchrist[c]

[a] *MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 0SR, UK*
[b] *Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands*
[c] *STORM Research Centre, London Metropolitan University, Holloway Road, London N7 8DB, UK*

Received 12 June 2006; received in revised form 4 December 2006; accepted 5 December 2006
Available online 2 January 2007

Abstract

The univariate and multivariate logistic regression model is discussed where response variables are subject to randomized response (RR). 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 may be described as misclassified categorical variables where conditional misclassification probabilities are known. The univariate model is revisited and is presented as a generalized linear model. Standard software can be easily adjusted to take into account the RR design. The multivariate model does not appear to have been considered elsewhere in an RR setting; it is shown how a Fisher scoring algorithm can be used to take the RR aspect into account. The approach is illustrated by analyzing RR data taken from a study in regulatory non-compliance regarding unemployment benefit.

Keywords: Multivariate and univariate logistic regression; Misclassification; Randomized response; Regulatory non-compliance; Sensitive questions

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; 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 fixed by design. The misclassification protects the privacy of the individual respondent. A meta-analysis by Lensvelt-Mulders et al. (2005) shows that RR yields more valid prevalence estimates than other methods for sensitive questions.

This paper discusses the univariate and the multivariate logistic regression model when the response variables are subject to RR. The discussion consists of two parts. First, we consider the univariate logistic regression model for binary RR response variables and present this model as a generalized linear model (GLM). By presenting the model as a GLM, two aspects of the model will become clear that have not been noticed in the literature. (i) The model inherits useful properties from the standard GLM; for example, properties of parameter estimates. (ii) The model can

[*] Corresponding author.
E-mail address: ardo.vandenhout@mrc-bsu.cam.ac.uk (A. van den Hout).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.12.002

be assessed by adjusting standard software for GLMs. The paper shows how adjustments of routines in R and GLIM are possible, making the assessment of logistic regression models for RR response variables reliable and fast.

The second part of the discussion in this paper is the presentation of a multivariate logistic regression model for RR response variables. As far as we know this model has not been considered elsewhere. The model makes it possible to investigate the relation between several RR response variables and a set of covariates jointly. There are various ways to define a multivariate logistic regression model without RR (Fahrmeir and Tutz, 2001, Section 3.5). The present paper extends the multivariate logistic regression model as presented by Glonek and McCullagh (1995) and shows how the model can be adapted to take the RR design into account.

We briefly review the literature on univariate logistic regression for RR response variables. Maddala (1983) was the first to present the likelihood of the model with respect to the RR design by Warner (1965). Scheers and Dayton (1988) discuss the model with respect to both the Warner model and the unrelated-question model (Greenberg et al., 1969). Van der Heijden and Van Gils (1996) present the model where the response variable is subject to either the RR design by Boruch (1971) or the RR design by Kuk (1990). A recent application of the univariate model is presented in Lensvelt-Mulders et al. (2006). Chen (1989) describes the link between RR and misclassification in the context of log-linear models. Magder and Hughes (1997) discuss the logistic regression model where the response variable is subject to misclassification comparable to the perturbation induced by the RR design. As far as we know, the encompassing GLM framework has not yet been discussed.

The outline of the paper is as follows. Section 2 describes the RR design. Section 3 present the logistic regression model given a binary RR response variable. Section 4 discusses the multivariate logistic regression model for RR response variables. In Section 5, applications are discussed using RR data from a Dutch study in regulatory non-compliance regarding unemployment benefit. Section 6 concludes the paper.

2. The RR design

This section starts with the forced response design (Boruch, 1971) as an example of an RR design and shows how the design can be seen as a misclassification design.

Assume that the sensitive question asks for a *yes* or a *no*. The forced response design is as follows. After the sensitive question is asked, the respondent throws two dice and keeps the outcome hidden from the interviewer. If the outcome is 2, 3 or 4, the respondent answers *yes*. If the outcome is 5, 6, 7, 8, 9 or 10, the respondent answers according to the truth. If the outcome is 11 or 12, the respondent answers *no*. This design protects the privacy since an observed *yes* does not necessarily imply a latent *yes*. In other words, the interviewer does not know whether a respondent answers *yes* because he or she answers the question truthfully—it might also be the case that the respondent answers *yes* because he or she is forced to do so by the design.

Let Y be the latent binary RR variable that models the sensitive item, Y^* the binary variable that models the observed answer, and *yes* $\equiv 1$ and *no* $\equiv 0$. The RR design of the forced response design is given by

$$\begin{aligned}\mathbb{P}(Y^*=1) &= \mathbb{P}(Y^*=1|Y=0)\mathbb{P}(Y=0)+\mathbb{P}(Y^*=1|Y=1)\mathbb{P}(Y=1)\\ &= 1/6+3/4\mathbb{P}(Y=1).\end{aligned} \tag{1}$$

Note that probabilities $\mathbb{P}(Y^*{=}j|Y{=}k)$ are fixed for $j, k \in \{0, 1\}$ by the known distribution of the sum of the two dice. The first equation of (1) shows that RR variables can be seen as misclassified variables, where conditional misclassification probabilities are given by $\mathbb{P}(Y^*=j|Y=k)$. We assume that the respondent complies with the instructions provided to him in the interview setting. Given this assumption, the distribution of the misclassification error is under control by the researcher.

If we write $\mathbb{P}(Y^*=1)=c+d\mathbb{P}(Y=1)$, other RR designs can be described in the same way (compare Böckenholt and van der Heijden, in press). In the original Warner (1965) design every person in a population belongs to either group A or group B. With probability $p \in (1/2, 1)$ the question is "Do you belong to group A?" and with probability $1-p$ the question is "Do you belong to group B?" In this case we have $c=1-p$ and $d=2p-1$.

In the Kuk (1990) design, a *yes*-or-*no* question is asked using stack of cards. As an example, let two stacks of cards contain both black and red cards. In the right stack the proportion of red cards is $\frac{8}{10}$ and in the left stack $\frac{2}{10}$. The respondent is asked to draw one card from each stack and to keep the color of the cards hidden from the interviewer. Next, the question is asked. Instead of answering the question directly, the respondent names the color of the card he took from the related stack, i.e., when the answer is *yes*, the respondent names the color of the card he took from the