ELSEVIER
Volume 51, issue 12
15 August 2007
ISSN 0167-9473
COMPUTATIONAL STATISTICS & DATA ANALYSIS
Incorporating Statistical Software Newsletter
The official journal of
iasc
The International Association for Statistical Computing
A Section of The International Statistical Institute
Available online at
ScienceDirect
www.sciencedirect.com

Available online at www.sciencedirect.com
ScienceDirect

ELSEVIER

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 right stack, and when the answer is *no*, he names the color of the card from the left stack. In this case we have $c = \frac{1}{5}$ and $d = \frac{3}{5}$.

3. The RR binary logistic regression model as a GLM

3.1. The random component

A GLM is described by a random component, a systematic component and a link function. The standard logistic regression model is a GLM, see, e.g., Agresti (2002) or Aitkin et al. (2005). Regarding the random component, the binary response variable Y with values $\{0, 1\}$ is a Bernoulli trial. In the RR logistic regression, the misclassified version Y^* of a binary Y is also a Bernoulli trial given that the latent Y is a Bernoulli trial. It follows that

$$Y^* \sim \text{Bernoulli}(c + d\mathbb{P}(Y = 1)).$$

Note that $0 < c + d\mathbb{P}(Y = 1|\boldsymbol{x}_i) < 1$ holds since $0 < c$, $0 < d$, and $c + d = \mathbb{P}(Y^* = 1|Y = 1) < 1$ in an RR design.

3.2. The systematic component

If data are given by $(y_i, \boldsymbol{x}_i)$, $i \in \{1, \ldots, n\}$, the systematic component of the standard logistic regression is

$$\mu_i = \mathbb{E}[Y_i|\boldsymbol{x}_i] = \mathbb{P}(Y_i = 1|\boldsymbol{x}_i) \quad \text{and} \quad g(\mu_i) = \log \frac{\mu_i}{1 - \mu_i} = \boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_i,$$

where $\boldsymbol{\beta}$ is the parameter vector, and the link function g is a monotonic differentiable function on $(0, 1)$. In the RR logistic regression model we have

$$\mu_i^* = \mathbb{E}[Y_i^*|\boldsymbol{x}_i] = c + d\mathbb{P}(Y_i = 1|\boldsymbol{x}_i) = c + d\,\frac{\exp(\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_i)}. \tag{2}$$

It follows that the link function g^* of the RR logistic regression model is given by

$$g^*(\mu_i^*) = \log \frac{\mu_i^* - c}{c + d - \mu_i^*} = \boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_i.$$

The function g^* is a monotonic differentiable function on $(0, 1)$ since

$$c < \mu_i^* = c + d\mathbb{P}(Y_i = 1|\boldsymbol{x}_i) < c + d,$$

which holds since $\mathbb{P}(Y_i = 1|\boldsymbol{x}_i) \in (0, 1)$. Notice that the inverse of g^* is given by (2). Fig. 1 depicts this inverse for the standard GLM, where $c = 0$ and $d = 1$, for the GLM with the Kuk design, where $c = \frac{1}{5}$ and $d = \frac{3}{5}$, and for the GLM with the forced response design, where $c = \frac{1}{6}$ and $d = \frac{3}{4}$. The graph shows what is already implicit in (1), i.e., given c and d, $\mu^* \in (c, c + d)$. If there is no misclassification, we have of course $(c, c + d) \equiv (0, 1)$. The graph also shows that the forced response design discriminates better between values of $\eta = \boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}$ than the Kuk design with $c = \frac{1}{5}$ and $d = \frac{3}{5}$. In the application section, we compare the efficiency of the RR designs numerically.

3.3. Maximum likelihood estimation

To estimate parameter vector $\boldsymbol{\beta}$ in the RR logistic regression model, the likelihood is maximized. In the case of RR, the likelihood is given by

$$\begin{aligned} L(\boldsymbol{\beta}|(y_i^*, \boldsymbol{x}_i), i \in \{1, \ldots, n\}, c, d) &\propto \prod_{y_i^*=0} 1 - \mathbb{P}(Y_i^* = 1|\boldsymbol{x}_i) \prod_{y_i^*=1} \mathbb{P}(Y_i^* = 1|\boldsymbol{x}_i) \\ &= \prod_{y_i^*=0} 1 - \left(c + d\,\frac{\exp(\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_i)}\right) \prod_{y_i^*=1} c + d\,\frac{\exp(\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_i)}{1 + \exp(\boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}_i)}. \end{aligned}$$

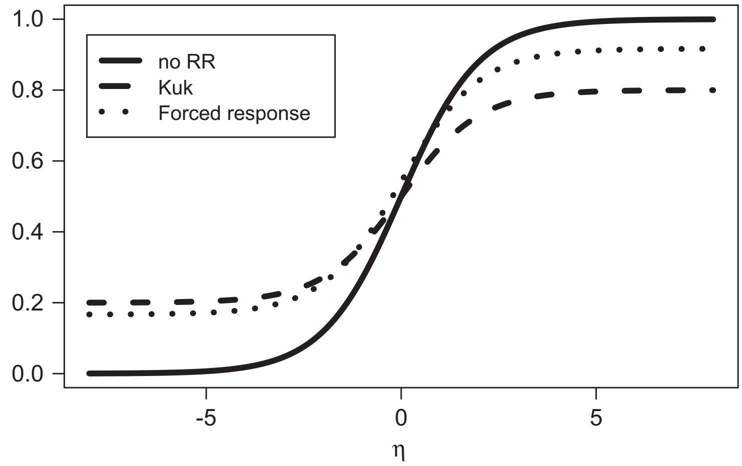

Fig. 1. The inverse of the link functions, given $\eta = \boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}$, for the standard GLM, the RR GLM with the Kuk design, and the RR GLM with the forced response design.

This is the general formulation of the likelihood in the RR setting. Maximization is possible by general Newton–Raphson methods, but as is shown in the Appendix, we can also adapt standard GLM software to take into account the RR design by defining a specific link function. The Appendix shows how this can be done in R and in GLIM. It is also possible to use the Fisher scoring algorithm directly, as presented in the next section.

4. The RR multivariate logistic regression model

4.1. The random component

For the RR multivariate logistic regression model, we describe the RR design per variable Y by using a matrix $\boldsymbol{P}_Y = (p_{jk})$, where $p_{jk} = \mathbb{P}(Y^* = j|Y = k)$. If the variable has K categories, matrix $\boldsymbol{P}_Y$ is a $K \times K$ non-singular matrix and the misclassification design is given by $\boldsymbol{\pi}^* = \boldsymbol{P}\boldsymbol{\pi}$, where $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_K^*)^{\mathrm{t}}$ is the distribution of the observed variable Y^* and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)^{\mathrm{t}}$ is the distribution of the latent variable Y. The case of a binary Y has already been illustrated by the forced response design (1). For this design it follows that

$$\boldsymbol{P}_Y = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} = \begin{pmatrix} 10/12 & 1/12 \\ 2/12 & 11/12 \end{pmatrix}.$$

Note that $c = p_{10}$ and $d = p_{11} - p_{10}$. The advantage of the matrix notation becomes apparent when we assess the multivariate RR design. As an example, if the misclassification of Y_1 is described by $\boldsymbol{P}_{Y_1}$ and the misclassification of Y_2 is described by $\boldsymbol{P}_{Y_2}$, the misclassification of the Cartesian product $\boldsymbol{Y} = (Y_1, Y_2)$ is described by $\boldsymbol{P}_{\boldsymbol{Y}} = \boldsymbol{P}_{Y_1} \otimes \boldsymbol{P}_{Y_2}$, where $\otimes$ denotes the Kronecker product, see, e.g., Chaudhuri and Mukerjee (1988).

For the general situation, assume that Y is multinomially distributed with parameter vector $\boldsymbol{\pi}$. Let the RR design be given by the non-singular matrix $\boldsymbol{P}$. From the fact that $\boldsymbol{P}$ is non-singular and $\boldsymbol{P}^{\mathrm{t}}$ is a transition matrix, it follows that Y^* is multinomially distributed with parameter vector $\boldsymbol{\pi}^* = \boldsymbol{P}\boldsymbol{\pi}$.

4.2. The systematic component

The multivariate logistic regression model is described by Glonek and McCullagh (1995). The general form of the link in the multivariate model between the linear predictor and the expected response vector is

$$\boldsymbol{\eta} = \boldsymbol{C}^{\mathrm{t}} \log(\boldsymbol{L}\boldsymbol{\pi}), \tag{3}$$

where $\boldsymbol{C}$ is the contrast matrix and $\boldsymbol{L}$ is the marginal indicator. We illustrate the model for the bivariate case with the binary responses Y_1 and Y_2. Let $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})^{\mathrm{t}}$, where $\pi_{kl} = \mathbb{P}(Y_1 = k, Y_2 = l)$, for $k, l \in \{0, 1\}$.

Given $\boldsymbol{\eta} = (0, \eta_{Y_1}, \eta_{Y_2}, \eta_{Y_1Y_2})^{\mathrm{t}}$, the link functions are given by

$$\eta_{Y_1} = \log \frac{\pi_{1+}}{1-\pi_{1+}}, \quad \eta_{Y_2} = \log \frac{\pi_{+1}}{1-\pi_{+1}} \quad \text{and} \quad \eta_{Y_1Y_2} = \log \frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}}, \tag{4}$$

where the plus subscript denotes summation, for example, $\pi_{1+} = \pi_{10} + \pi_{11}$. The regression equations are given by

$$\eta_{Y_1} = \boldsymbol{\beta}_{Y_1}^{\mathrm{t}} \boldsymbol{x}_{Y_1}, \quad \eta_{Y_2} = \boldsymbol{\beta}_{Y_2}^{\mathrm{t}} \boldsymbol{x}_{Y_2} \quad \text{and} \quad \eta_{Y_1Y_2} = \boldsymbol{\beta}_{Y_1Y_2}^{\mathrm{t}} \boldsymbol{x}_{Y_1Y_2}. \tag{5}$$

This model is a marginal model—it implies univariate logistic models for both Y_1 and Y_2 marginally (Glonek and McCullagh, 1995). The odds ratio is used to model the dependence between Y_1 and Y_2. The contrast matrix and the marginal indicator for this model are given by

$$\boldsymbol{C}^{\mathrm{t}} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{L}^{\mathrm{t}} = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

The first element of $\boldsymbol{\eta}$ represents the null contrast $\log(\pi_{++}) = 0$. For an observation i, the regression is given by $\boldsymbol{\eta}_i = \boldsymbol{X}_i \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_{Y_1}^{\mathrm{t}}, \boldsymbol{\beta}_{Y_2}^{\mathrm{t}}, \boldsymbol{\beta}_{Y_1Y_2}^{\mathrm{t}})^{\mathrm{t}}$ and

$$\boldsymbol{X}_i = \begin{pmatrix} 0 & 0 & 0 \\ \boldsymbol{x}_{Y_1}^{\mathrm{t}} & 0 & 0 \\ 0 & \boldsymbol{x}_{Y_2}^{\mathrm{t}} & 0 \\ 0 & 0 & \boldsymbol{x}_{Y_1Y_2}^{\mathrm{t}} \end{pmatrix}.$$

Given an RR design, the link between the linear predictor and the expected response vector is

$$\boldsymbol{\eta} = \boldsymbol{C}^{\mathrm{t}} \log(\boldsymbol{L}\boldsymbol{\pi}) = \boldsymbol{C}^{\mathrm{t}} \log(\boldsymbol{L}\boldsymbol{P}^{-1}\boldsymbol{\pi}^*). \tag{6}$$

4.3. *Maximum likelihood estimation*

To estimate the RR multivariate logistic regression model, we adapt the method that is presented in Glonek and McCullagh (1995), who do not consider the RR case. Assume that data are given by independent observations $(\boldsymbol{y}_i^*, \boldsymbol{X}_i)$, $i = 1, \ldots, n$. Let the RR design be described by $\boldsymbol{P}$, be it a univariate or a multivariate situation. The log likelihood is

$$l(\boldsymbol{\beta} \,|(\boldsymbol{y}_i^*, \boldsymbol{X}_i), i \in \{1, \ldots, n\}, \boldsymbol{P}) = \sum_{i=1}^{n} (\boldsymbol{y}_i^*)^{\mathrm{t}} \log \boldsymbol{\pi}^* = \sum_{i=1}^{n} (\boldsymbol{y}_i^*)^{\mathrm{t}} \log \boldsymbol{P}\boldsymbol{\pi}. \tag{7}$$

The maximization of the log likelihood consists of two parts that are iterated. Given a starting value of $\boldsymbol{\beta}$, the first part uses a Newton–Raphson iteration to obtain $\boldsymbol{\pi}$ given $\boldsymbol{\eta}$. This part does not need an adaption for the RR design since the relation between the linear predictor and the latent distribution does not change. The second part is a Fisher scoring algorithm that takes the RR into account and updates the estimation of $\boldsymbol{\beta}$.

Part one: To invert Eq. (3), work with $\boldsymbol{v} = \log(\boldsymbol{\pi})$. The Newton–Raphson iterations are described by

(a) The initial approximation is $\boldsymbol{v}_0$.
(b) Then set

$$\boldsymbol{v}_n = \boldsymbol{v}_{n-1} - \{\boldsymbol{C}^{\mathrm{t}} (\mathrm{diag}(\boldsymbol{L} \exp(\boldsymbol{v}_{n-1})))^{-1} \boldsymbol{L}\}^{-1} \{\boldsymbol{C}^{\mathrm{t}} \log(\boldsymbol{L} \exp(\boldsymbol{v}_{n-1})) - \boldsymbol{\eta}\}.$$

For details see Glonek and McCullagh (1995, Section 3).

Part two: The adaptation of the scoring algorithm is straightforward given the above. The derivative $\partial\boldsymbol{\pi}/\partial\boldsymbol{\beta}$ is given by $(\boldsymbol{C}^{\mathrm{t}} \,\mathrm{diag}(\boldsymbol{L}\boldsymbol{\pi})^{-1}\boldsymbol{L})^{-1}\boldsymbol{X}$. The score vector and the Fisher information matrix are given by

$$s(\boldsymbol{\beta}|\boldsymbol{y}_i^*) = \left(\boldsymbol{P} \frac{\partial \boldsymbol{\pi}}{\partial \boldsymbol{\beta}}\right)^{\mathrm{t}} \mathrm{diag}(\boldsymbol{P}\boldsymbol{\pi})^{-1} \boldsymbol{y}_i^*,$$

and

$$\mathscr{I}(\boldsymbol{\beta}|\boldsymbol{y}_i^*) = \mathbb{E}_{Y_i^*}[s(\boldsymbol{\beta};\boldsymbol{y}_i)s^{\mathrm{t}}(\boldsymbol{\beta};\boldsymbol{y}_i)] = \left(\boldsymbol{P}\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\beta}}\right)^{\mathrm{t}} \mathrm{diag}(\boldsymbol{P}\boldsymbol{\pi})^{-1}\left(\boldsymbol{P}\frac{\partial\boldsymbol{\pi}}{\partial\boldsymbol{\beta}}\right). \tag{8}$$

If the cross-classification of the latent response variables contains one or more sample zeros, we propose to smooth the data in the maximization procedure. That is, after part one, add a small probability mass to each entry of $\widehat{\boldsymbol{\pi}}$, normalize $\widehat{\boldsymbol{\pi}}$ such that the entries sum up to one, and then go to part two. Sample zeros in the cross-classification of RR data are discussed in Van den Hout and Van der Heijden (2004).

5. Applications

This section illustrates the foregoing by analyzing data taken from an RR survey concerning unemployment benefit in the Netherlands in 2004. This analysis is limited in the sense that we utilize a limited number of variables from the original survey and only consider main-effect models.

The sample size is $n = 753$. For $i \in \{1, \ldots, n\}$, let y_{1i}^* denote the observed answer to the RR question whether the number of job applications of respondent i is below the required minimum ($yes = 1, no = 0$). Likewise let y_{2i}^* denote the observed answer to the question whether respondent i did any voluntary work without reporting this. In addition, x_{1i} is sex ($male = 1, female = 0$), x_{2i} is age in years and x_{3i} denotes whether the income of respondent i constitutes the larger part of the income of the household ($yes = 1, no = 0$).

The RR design is given by $p_{00} = 0.813$ and $p_{11} = 0.933$ and it is a slightly adapted form of the forced response design.

5.1. Univariate regression

The univariate logistic regression analyses are presented in Table 1, where AIC stands for the Akaike Information Criterion. The results in the table—bar the AICs—consist of the output of the `glm()` procedure in R if this procedure is adapted for RR as described in the Appendix.

The first estimated model and the reductions in the deviance show that covariate x_2 is important when modelling the question whether respondents apply frequently enough for new jobs. The second model shows that covariates x_1 and x_2 are associated with the question whether respondents did any unreported voluntary work. Looking at the parameter estimates we note that the estimates of β_3 are pretty close in the two models. However, variable x_3 does not induce a significant reduction in deviance after controlling for x_1 and x_2. We shall use these results in the multivariate modelling in the next section.

Table 1
Univariate logistic regression analyses of the unemployment data, where the RR design is taken into account

Full model				Analysis of deviance (terms added sequentially)			
$\mathrm{logit}(\mathbb{P}(Y=1\|\boldsymbol{x})) = \boldsymbol{\beta}^{\mathrm{t}}\boldsymbol{x}$					Reduction in deviance	Deviance	AIC
	Estimate		SE				
Y_1	β_0	−0.048	0.556	(Intercept)		952.95	954.95
	β_1	−0.330	0.369	x_1	0.74	952.21	956.21
	β_2	−0.034	0.014	x_2	4.65	947.56	953.56
	β_3	0.534	0.390	x_3	1.93	945.63	953.63
Y_2	β_0	−3.515	0.717	(Intercept)		979.26	981.26
	β_1	0.168	0.378	x_1	5.81	973.44	977.44
	β_2	0.043	0.015	x_2	12.33	961.12	967.12
	β_3	0.310	0.370	x_3	0.49	960.62	968.62

Variable Y_1 is job application and Y_2 is voluntary work. Sample size is 753 and β_0 is the intercept.

Table 2
Standard errors for RR designs w.r.t. univariate logistic regression analysis of the unemployment data in Table 1

Design	Standard errors of $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, and $\widehat{\beta}_4$			
Current RR design with $c = 0.187$ and $d = 0.746$	0.556	0.369	0.014	0.390
Forced response with $c = \frac{1}{6}$ and $d = \frac{3}{4}$	0.545	0.362	0.014	0.383
Kuk with $c = \frac{1}{5}$ and $d = \frac{3}{5}$	0.684	0.453	0.018	0.480
Kuk with $c = \frac{1}{10}$ and $d = \frac{4}{5}$	0.485	0.321	0.012	0.340
No RR, i.e., $c = 0$ and $d = 1$	0.355	0.234	0.009	0.246

The response variable is Y_1 and we assume $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.

Table 3
Deviances of bivariate logistic regression models taken the RR design into account

Model		# parameters	Deviance	AIC
1	Intercept(1,2,3)	3	1928.033	1934.033
2	Intercept(1,2,3), $x_2(1, 2)$	5	1905.071	1915.071
3	Intercept(1,2,3), $x_1(2)$, $x_2(1, 2)$	6	1903.411	1915.411
4	Intercept(1,2,3), $x_2(1, 2)$, $x_3(2)$	6	1903.098	1915.098
5	Intercept(1,2,3), $x_2(1, 2)$, $x_3(1, 2)$, $\beta_{Y_13} = \beta_{Y_23}$	6	1901.638	1913.638
6	Intercept(1,2,3), $x_2(1, 2)$, $x_3(1)$	6	1903.799	1915.799
7	Intercept(1,2,3), $x_2(1, 2)$, $x_3(1, 2)$	7	1901.611	1915.611
8	Intercept(1,2,3), $x_1(1, 2)$, $x_2(1, 2)$, $x_3(1, 2)$	9	1900.889	1918.889

Sample size is 753. Notation: $x_j(k, \ldots)$ means the covariate x_j is included in the kth regression equation, where the order of the equations is given by (5).

To apply the Fisher scoring algorithm of Section 4 to the univariate case, choose

$$\boldsymbol{C}^{\mathrm{t}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{L}^{\mathrm{t}} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

As a reviewer of the paper suggested, the information matrix (8) can be used to investigate the RR design with respect to the extra variance induced by the misclassification. The matrix does not depend on observed values of the response variable—its link with the RR design is $\boldsymbol{P}$. Given a value of $\boldsymbol{\beta}$, we can compare the variance for various choices of c and d. For job application (Y_1) and corresponding $\widehat{\boldsymbol{\beta}}$, standard errors for five designs are presented in Table 2. Regarding the extra variance, the design in our application is in between the two Kuk options and similar to the original forced response design. The closer (c, d) to $(0, 1)$ the better in terms of standard errors, but the closer to $(0, 1)$, the greater the lack of privacy protection and, hence, the higher the probability that respondents do not follow the instructions of the RR design, see Böckenholt and van der Heijden (in press). A discussion of the efficiency with respect to the Warner model and the unrelated-question model can be found in Scheers and Dayton (1988).

5.2. *Bivariate regression*

Continuing with the application in the previous section, we now turn to the multivariate models. Table 3 presents the results of the analysis. As an illustration, Model 8 is defined by (3), (4) and the regression equations $\boldsymbol{\eta}_i = \mathbf{X}_i \boldsymbol{\beta}$ which are given by $\boldsymbol{\beta} = (\beta_{Y_10}, \beta_{Y_11}, \beta_{Y_12}, \beta_{Y_13}, \beta_{Y_20}, \beta_{Y_21}, \beta_{Y_22}, \beta_{Y_23}, \beta_{Y_1Y_20})^{\mathrm{t}}$ and

$$\mathbf{X}_i = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & x_{Y_11,i} & x_{Y_12,i} & x_{Y_13,i} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_{Y_21,i} & x_{Y_22,i} & x_{Y_23,i} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

So the odds ratio is estimated by using an intercept only. Models 1–7 are defined by restrictions on the parameters, see Table 3. The order of the models in the table reflects the findings in the previous section. Model 3 versus Model 2 shows that adding x_1 does not seem necessary. With respect to adding x_3, the best model according to the AIC is Model 5 in which the regression coefficient of x_3 is restricted to be the same in the marginal models for Y_1 and Y_2, and $\beta_{Y_1 1} = \beta_{Y_2 1} = 0$. Coefficient estimates for this model are given by

$$\begin{aligned}
&\widehat{\beta}_{Y_1 0} = -0.03\ (0.54), \quad \widehat{\beta}_{Y_2 0} = -3.31\ (0.68),\\
&\widehat{\beta}_{Y_1 2} = -0.03\ (0.01), \quad \widehat{\beta}_{Y_2 2} = 0.04\ (0.013),\\
&\widehat{\beta}_{Y_1 3} = \widehat{\beta}_{Y_2 3} = 0.44\ (0.33) \quad \text{and} \quad \widehat{\beta}_{Y_1 Y_2 0} = 1.30\ (0.54),
\end{aligned}$$

where the estimated standard errors are within the brackets. The standard errors are estimated by evaluating the information matrix at the optimum. The Wald 95% confidence interval of $\exp(\beta_{Y_1 3}) = \exp(\beta_{Y_2 3})$ is (0.81; 2.99) which suggest that including x_3 in the model is not necessary. If we use the deviance to select a model, we would opt for model 2.

The interpretation of the parameters of the marginal models for Y_1 and Y_2 is the same as in the univariate logistic regression model. In the multivariate model, we can easily investigate whether an effect of a covariate is the same in the marginal models. This is illustrated by Model 5 which states that the effect of whether or not the income of the respondent constitutes the larger part of the income of the household is the same for the prevalence of regulatory non-compliance regarding job applications and the prevalence of non-compliance regarding voluntary work. More specific, Model 5 states that a change from $x_3 = 0$ to 1, means that the odds on non-compliance regarding both Y_1 and Y_2 changes multiplicatively by $\exp(\widehat{\beta_{Y_1 3}}) = \exp(\widehat{\beta_{Y_2 3}}) = \exp(0.44) = 1.56$. So the odds on non-compliance is higher when a person's income constitutes the larger part of the income of the household.

In addition, the multivariate model describes the association between the two response variables Y_1 and Y_2 by estimating the odds ratio. In Model 5, the ratio of the odds in the two rows of the cross-classification of Y_1 and Y_2 is estimated to be $\exp(1.30) = 3.67$ with Wald 95% confidence interval (1.28; 10.49). Thus, interpreting this model, respondents appeared to be nearly 4 times as likely to ignore rules concerning job applications as to do voluntary work without reporting them.

6. Discussion

This paper presents methods to estimate univariate and multivariate logistic regression models where the response variable is subject to RR. The framework is quite general in the sense that the methods can deal with various RR designs and that the RR regression models have the same flexibility as the standard regression models. The application section shows that once the methods are implemented, analyzing RR data and interpreting the results resemble the analyses and interpretations in the situations without RR. The Fisher information matrix can be used to compare the variance induced by various RR designs.

To take RR into account, we extended the multivariate logistic regression model as discussed in Glonek and McCullagh (1995). Both the extension and the standard model are marginal models. Bergsma and Rudas (2002) introduce a general definition of marginal models that includes the multivariate logistic regression model. They also consider aspects of marginal models that we did not discuss, e.g., the existence of a joint distribution with certain restrictions on some of its marginals.

An important assumption in the paper is that respondents comply in the sense that they follow the instructions of the RR design. This assumption will not always be justified. For instance, it might be that some respondents do not trust the privacy protection offered by the RR design and give a socially desirable answer anyway. The general idea, however, is that RR performs relatively well, see Lensvelt-Mulders et al. (2005). Future RR surveys may profit from research into cheating with respect to the RR design, see Böckenholt and van der Heijden (in press) and the references therein.

Acknowledgment

Most of the work of the first author was done at the Department of Methodology and Statistics, Utrecht University, the Netherlands.

Appendix A. Using GLM software to estimate the univariate model

Using R (R Development Core Team, 2005). It is possible to extend `glm()` with a user-specified link function that takes the RR design into account. The general idea how to define your own link function was explained by B. Ripley (R mailing List, 2001).

1. Make a new family object `my.binomial<-edit(binomial)` by changing

```
  if(any(linktemp= =c("logit","probit","cloglog","log")))
  stats<-make.link(linktemp)
```

into

```
  if(any(linktemp= =c("logit","probit","cloglog","log","RRlogit")))
  stats<-my.make.link(linktemp)
```

2. Make a new link function `my.make.link<-edit(make.link)` by extending `make.link` with a user-specified link `RRlogit`. This new link is the same as the `logit` link except for `linkfun`, `linkinv`, and `mu.eta`, which should be adapted such that they describe the new link function as given in Section 3.2

3. The function call is `glm( ...., family =my.binomial(link="RRlogit"))`

Using GLIM (Francis et al., 1993). For the forced response RR design we have $c = 0.16667$ and $d = 0.75$. Assume that $n = 1000$ is the sample size. For the RR univariate logistic regression model with `ystar` and `x1` and `x2`:

```
$sl 1000 $data ystar x1 x2 $din 'RRdata.dat' $err b n
$calc n=1 $yvar ystar $num c d $ass c=0.16667: d=0.75
$macro ownlink
$calc emineta=%exp(-%lp)
$c Calculate mu as in eqn (2)
$calc %fv=c+d/(1+emineta)
$c Calculate the derivative of eta w.r.t. the mu from eqn (2)
$calc %dr=(1+emineta)**2/(d*emineta)
$endmac
$c Fit a model without RR to get starting values
$fit x1+x2 $di e
$c Fit univariate RR model
link o ownlink $fit . $di e
```

References

Agresti, A., 2002. Categorical Data Analysis. second ed. Wiley, New York.

Aitkin, M., Francis, B., Hinde, J., 2005. Statistical Modelling in GLIM4. second ed. Oxford University Press, Oxford.

Bergsma, W.P., Rudas, T., 2002. Marginal models for categorical data. Ann. Statist. 30, 140–159.

Böckenholt, U., van der Heijden, P.G.M. Item randomized-response models for measuring noncompliance: risk-return perceptions, social influences, and self-protective responses. Psychometrika, in press.

Boruch, R.F., 1971. Assuring confidentiality of responses in social research: a note on strategies. Amer. Sociologist 6, 308–311.

Chaudhuri, A., Mukerjee, R., 1988. Randomized Response: Theory and Techniques. Marcel Dekker, New York.

Chen, T.T., 1989. A review of methods for misclassified categorical data in epidemiology. Statist. Med. 8, 1095–1106.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models. second ed. Springer, New York.

Francis, B., Green, M., Payne, C., 1993. The GLIM System, The Release Manual. Clarendon Press, Oxford.

Glonek, G.F.V., McCullagh, P., 1995. Multivariate logistic models. J. Roy. Statist. Soc. B 57, 533–546.

Greenberg, B.G., Abul-Ela, A., Simmons, W.R., Horvitz, D.G., 1969. The unrelated question randomized response model: theoretical framework. J. Amer. Statist. Assoc. 64, 520–539.

Kuk, A.Y.C., 1990. Asking sensitive questions indirectly. Biometrika 77, 436–438.

Lensvelt-Mulders, G.J.L.M., Hox, J.J., Van der Heijden, P.G.M., Maas, C.J.M., 2005. Meta-analysis of randomized response research: thirty-five years of validation. Soc. Methods Res. 33, 319–348.

Lensvelt-Mulders, G.J.L.M., Van der Heijden, P.G.M., Laudy, O., 2006. A validation of a computer-assisted randomized response survey to estimate the prevalence of fraud in social security. J. Roy. Statist. Soc. A 169, 305–318.

Maddala, G.S., 1983. Limited Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge.

Magder, L.S., Hughes, J.P., 1997. Logistic regression when the outcome is measured with uncertainty. J. Amer. Statist. Assoc. 146, 195–203.

R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL ⟨http://www.R-project.org⟩.

Scheers, J., Dayton, C.M., 1988. Covariate randomized response models. J. Amer. Statist. Assoc. 83, 969–974.

Van der Heijden, P.G.M., Van Gils, G., 1996. Some logistic regression models for randomized response data. In: Forcina, A., Marchetti, G.M., Hatzinger, R., Falmacci, G. (Eds.), Proceedings of the 11th International Workshop on Statistical Modelling.

Van den Hout, A., Van der Heijden, P.G.M., 2004. The analysis of multivariate misclassified data with special attention to randomized response data. Soc. Methods Res. 32, 310–336.

Warner, S.L., 1965. Randomized response: a survey technique for eliminating answer bias. J. Amer. Statist. Assoc. 60, 63–69.