J. R. Statist. Soc. B (2008)
70, *Part* 1, *pp.* 21–30

The analysis of randomized response sum score variables

Maarten J. L. F. Cruyff,

Utrecht University, The Netherlands

Ardo van den Hout

Medical Research Council Biostatistics Unit, Cambridge, UK

and Peter G. M. van der Heijden

Utrecht University, The Netherlands

[Received August 2006. Final revision July 2007]

Summary. Randomized response (RR) is an interview technique that ensures confidentiality when questions are sensitive. In RR the answer to a sensitive question depends to a certain extent on a probability mechanism. As a result the observed data are partially misclassified, and the true status of the respondent is obscured. RR data are commonly analysed in a univariate way, with models that relate the observed responses to the prevalence of the sensitive characteristic, and with the more recent logistic regression models that relate the sensitive characteristic to a set of covariates. In an RR design with multiple sensitive questions, interest is usually not confined to the univariate prevalence and regression parameter estimates. Additional multivariate information may be obtained from an RR sum score variable, assessing the sum of sensitive characteristics that are associated with the respondent. However, the construction of an RR sum score variable is by no means straightforward, which might explain why sum scores have not yet been used within the context of RR. We present two models for RR sum score variables: the RR sum score model that relates the observed sum scores to the true sum scores and the RR proportional odds model that relates the true sum scores to covariates. The models are applied to RR data from a Dutch survey on non-compliance with social security regulations.

Keywords: Proportional odds model; Randomized response; Regulatory non-compliance; Sum score variable

1. Introduction

In surveys and questionnaires, questions are sometimes regarded as sensitive or embarrassing. Especially if personal characteristics like the respondent's use of drugs, alcohol consumption or sexual behaviour are assessed, the questions may be perceived as an invasion of privacy, and respondents will be reluctant to give a direct answer. Randomized response (RR) is an interview technique that was designed to protect the privacy of the respondent. In RR, the answer to a sensitive question depends partly on the respondent's true status and partly on the outcome of a randomizing device. The RR technique was originally introduced by Warner (1965). In the Warner design the respondent is given two complementary sensitive questions, e.g. 'I have used drugs' and 'I have never used drugs', and the outcome of a randomizing device determines

Address for correspondence: Maarten J. L. F. Cruyff, Department of Methodology and Statistics, Utrecht University, PO Box 80.140, 3508 TC Utrecht, The Netherlands.
E-mail: m.cruyff@uu.nl

 1369–7412/08/70021

which of the two questions the respondent must answer. So, a respondent who has never used drugs answers *false* if the former question must be answered, and *true* if the latter question must be answered. Since the outcome of the randomizing device is not known to the interviewer, the true status of the respondent remains uncertain, and confidentiality is ensured.

Usually the main objective of the RR design is to obtain a prevalence estimate of the sensitive characteristic, and this estimate can be obtained with a model that relates the observed response to the true status of the respondent. In the Warner design, the model $\pi^* = \theta\pi + (1-\theta)(1-\pi)$ describes the probability π^* of observing a true response as a function of the prevalence π of drug use, and the probability θ that the statement 'I have used drugs' is selected. Since θ is determined by the design and the sample proportion of true responses is an estimate of π^*, the prevalence of the sensitive characteristic π can be estimated. Similar models have been presented for other RR designs such as the unrelated question design (Horvitz *et al.*, 1967), the forced response design (Boruch, 1971) and the Kuk design (Kuk, 1990).

In addition to the prevalence, the determinants of the sensitive characteristic are of interest. Maddala (1983) and Scheers and Dayton (1988) presented logistic regression models that can be used to analyse the dependence of an RR variable on a set of covariates. Recently, Elffers *et al.* (2003) have applied these models to RR data to study the motives for regulatory non-compliance with two Dutch instrumental laws.

In many RR applications, more than one sensitive question is asked. A meta-analysis of prevalence estimation in RR research (Lensvelt-Mulders *et al.*, 2005) reveals that, in 39 RR surveys, a total of 264 sensitive questions are asked, or an average of approximately seven questions in each survey. In a design with multiple RR variables, interest is usually not confined to the univariate prevalence and regression parameter estimates of the separate sensitive characteristics. Böckenholt and van der Heijden (2007) and Fox (2005) introduced item response theory models for RR profiles. In these models the person parameter is based on multiple assessments of the sensitive characteristic and individual differences are explained by covariates. van den Hout *et al.* (2007) present a multivariate logistic regression model describing the associations between multiple binary RR variables and a set of covariates.

An alternative approach to analyse multivariate RR data is to construct a sum score variable denoting the individual sum of sensitive characteristics. In this approach interest is primarily in the distribution of the number of sensitive characteristics and the dependence of the number of sensitive characteristics on covariates. Examples of sum score variables in the context of RR are variables assessing the number of different drugs that the respondent has used, the number of different criminal activities that the respondent has engaged in or the number of potentially traumatic events that the respondent has experienced. To the best of our knowledge, sum score variables have not yet been used in the context of RR.

Since the observed data are partially misclassified, the construction of an RR sum score variable is not straightforward. This paper demonstrates how to construct an RR sum score variable and presents two models for analysing RR sum score variables. The RR sum score model relates the sum of affirmative responses to the sum of the sensitive characteristics and is used to estimate the probability distribution of the sum of sensitive characteristics. The RR proportional odds model is an adjusted version of the proportional odds model that was presented by McCullagh (1980) and describes the dependence of the sum of the sensitive characteristics on a set of covariates. As an example, the models are applied to RR data from a Dutch survey assessing regulatory non-compliance with social security legislation.

Section 2 describes the social security survey data and the forced response design that was used in this survey. The first part of Section 3 presents the RR sum score model and the second

part the RR proportional odds model. The example is presented in Section 4. Section 5 discusses boundary solutions and presents an example. Section 6 gives the conclusions.

2. Social security survey 2002

Employees in the Netherlands are insured under the Social Security Law. The Disability Insurance Act insures them against a loss of income due to a complete or partial inability to work. To be eligible for financial benefits, one has to comply with various rules and regulations. In 2002 the Dutch Department of Social Affairs conducted a nationwide survey to evaluate the level of non-compliance with the rules and regulations in the Disability Insurance Act (for more details see Lensvelt-Mulders *et al.* (2006) and van Gils *et al.* (2003)). A sample of 1760 recipients were asked two questions about their health status (questions 1 and 2) and two questions about receiving income from work in addition to the disability benefit (questions 3 and 4).

Question 1. 'At a Social Services check-up, have you ever acted as if you were sicker or less able to work than you actually were?'

Question 2. 'For periods of any length at all, do you ever feel stronger and healthier and able to work more hours without informing the Department of Social Services?'

Question 3. 'Have you done any small jobs for or via friends or acquaintances in the past year, or paid jobs of any size without reporting it to the Department of Social Services? (This only pertains to monetary payments.)'

Question 4. 'Have you worked off the books in the past year in addition to your disability benefit?'

Owing to the sensitive nature of the questions, the forced response design (Boruch, 1971) was applied. In the forced response design the respondent tosses two dice and is instructed to answer *yes* to the question if the sum of the two dice is 2, 3 or 4, and *no* if the sum of the two dice is 11 or 12, irrespective of the respondent's true status. If the sum of the two dice is 5, 6, 7, 8, 9 or 10, the respondent must answer truthfully. The outcome of the dice is known only to the respondent.

Misclassification occurs if respondents are forced to give an answer that is in disagreement with their true status. The probabilities of a forced yes and a forced no response follow from the probability distribution of the sum of two dice; it can be easily verified that $\mathbb{P}(\text{forced yes}) = 1/6$, and $\mathbb{P}(\text{forced no}) = 1/12$. (The programmer inadvertently programmed the virtual dice so that $\mathbb{P}(\text{forced yes}) = 0.1868$ and $\mathbb{P}(\text{forced no}) = 0.0671$.) Given that the respondent's true answer is no, the probability of misclassification $\mathbb{P}(\text{observed yes}|\text{true no}) = \mathbb{P}(\text{forced yes})$, and similarly, given a true yes response, the probability of misclassification $\mathbb{P}(\text{observed no}|\text{true yes}) = \mathbb{P}(\text{forced no})$. Since, irrespective of the true response, the probability of misclassification is non-zero, confidentiality is assured.

Let the variables Y_1^*–Y_4^* denote the answers to questions 1–4, with $y_1^*, \ldots, y_4^* \in \{0 \equiv \text{no}, 1 \equiv \text{yes}\}$. The frequencies of the observed response profiles $0000, 0001, \ldots, 1111$, with the score on the last variable changing first, are given by the vector $\mathbf{n}^* = (694, 117, 188, 81, 179, 43, 65, 41, 117, 41, 37, 26, 62, 14, 27, 28)$. The set of covariates consists of the variables gender, age, last job contract, education, degree of disability and time unemployed. Gender, age, job contract and degree of disability are binary variables with respective reference categories male, younger than 45 years, other (*versus* regular job) and less than 80%. The categories of education are low, middle and high. Time unemployed is a continuous variable that denotes the logarithm of the number of years (plus 1) that have passed since the respondent was last employed.

3. The models

In this section, we present the two models. The RR sum score model relates the sum of the observed yes responses to the number of rule violations. The RR proportional odds model relates the number of rule violations to the covariates.

3.1. The randomized response sum score model

In an RR design with M sensitive questions, let variable Y_m denote the true response to the mth question, for $m \in \{1, \ldots, M\}$ and $y_m \in \{0 \equiv \text{no}, 1 \equiv \text{yes}\}$. The RR sum score variable denoting the number of true yes responses is defined by

$$Z = \sum_{m=1}^{M} Y_m. \tag{1}$$

Analogously, let the sum score variable $Z^* = \Sigma_{m=1}^{M} Y_m^*$ denote the number of observed yes responses. The probability of observing sum score s on variable Z^*, for $s \in \{0, \ldots, M\}$, is given by the RR sum score model

$$\pi_s^* = \sum_{t=0}^{M} q_{s|t}\pi_t, \tag{2}$$

where $\pi_s^* = \mathbb{P}(Z^* = s)$, $\pi_t = \mathbb{P}(Z = t)$ and $q_{s|t} = \mathbb{P}(Z^* = s|Z = t)$.

Lemma 1. Denote the misclassification probabilities of the variables Y_m by $p_{i|j} = \mathbb{P}(Y_m^* = i|Y_m = j)$, for $i, j \in \{0, 1\}$, and let $p_{i|j}$ be the same for all $m \in \{1, \ldots, M\}$. The misclassification probabilities of Z are given by

$$q_{s|t} = \sum_{j=0, 0 \leqslant s+j-t \leqslant M-t}^{t} \binom{t}{j} \binom{M-t}{s+j-t} p_{1|1}^{t-j} p_{0|1}^{j} p_{1|0}^{s+j-t} p_{0|0}^{M-s-j}. \tag{3}$$

The index j in equation (3) denotes the number of positions where $Y_m^* = 0$ among the t positions m where $Y_m = 1$, and the index $s + j - t$ denotes the number of positions where $Y_m^* = 1$ among the $M - t$ positions m where $Y_m = 0$. Equation (3) follows from the fact that the pairs (Y_m^*, Y_m) are independent and identically distributed for all $m \in \{1, \ldots, M\}$, and the order of 1s and 0s in the response profile $(Y_1, \ldots, Y_M)$ is not relevant for the result. (We thank a referee for contributing to the final formulation of lemma 1.)

3.1.1. Estimation

The RR sum score model is most easily estimated with the method of moments (MM). The MM estimator is most conveniently presented by using matrix notation,

$$\hat{\boldsymbol{\pi}} = \mathbf{Q}^{-1}\hat{\boldsymbol{\pi}}^*, \tag{4}$$

where $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_M)'$, $\boldsymbol{\pi}^* = (\pi_0^*, \ldots, \pi_M^*)'$ and π_s^* is estimated by n_s^*/n, with n_s^* denoting the frequency of the observed sum score s on variable Z^*. The matrix $\mathbf{Q}$ is an $(M+1)\times(M+1)$ transition matrix with entries $(s+1, t+1)$ given by the conditional misclassification probabilities $q_{s|t}$, for $s, t \in \{0, \ldots, M\}$. The MM solution always fits the data but can result in probability estimates that are outside the boundaries of the parameter space defined by (0,1).

The maximum likelihood (ML) estimates of the RR sum score model are obtained by maximizing the kernel of the observed data log-likelihood

$$\ln\{l(\boldsymbol{\pi}|n_0^*,\ldots,n_M^*)\}=\sum_{s=0}^{M} n_s^* \ln\left(\sum_{t=0}^{M} q_{s|t}\pi_t\right), \tag{5}$$

for $\pi_t \in (0,1)$. Kuha and Skinner (1997) have provided EM algorithms. van den Hout and van der Heijden (2002) showed that, if the MM estimates are in the interior of the parameter space, the ML solution is identical to the MM solution. Otherwise, one or more ML estimates will be on the boundary.

3.2. The randomized response proportional odds model

We now present the model for the regression of an RR sum score variable on a set of covariates. Assume that the sum scores are on an ordinal scale and let $\mathbb{P}(Z=t|\mathbf{x})$ denote the probability that the sum score variable Z takes the value t given the covariate vector $\mathbf{x}$. Define $\gamma_t=\mathbb{P}(Z\leqslant t|\mathbf{x})$. Then the proportional odds model (McCullagh, 1980) states that

$$\gamma_t=\frac{\exp(\alpha_t-\mathbf{x}'\boldsymbol{\beta})}{1+\exp(\alpha_t-\mathbf{x}'\boldsymbol{\beta})}, \tag{6}$$

where the threshold parameters α_t can be thought of as the values on a latent trait variable that mark the transition from $Z=t-1$ to $Z=t$. The threshold parameters satisfy the condition

$$-\infty<\alpha_0\leqslant\alpha_1\leqslant\ldots\leqslant\alpha_M\equiv\infty. \tag{7}$$

For $M=1$, the order of the threshold parameters is $-\infty<\alpha_0\leqslant\alpha_1\equiv\infty$, and expression (6) reduces to the binary logistic regression model (with a negative sign for $\boldsymbol{\beta}$).

A property of the proportional odds model is that the logarithm of the cumulative odds

$$\ln\left\{\frac{\mathbb{P}(Z\leqslant t|\mathbf{x}_0)/\mathbb{P}(Z>t)|\mathbf{x}_0)}{\mathbb{P}(Z\leqslant t|\mathbf{x}_1)/\mathbb{P}(Z>t)|\mathbf{x}_1)}\right\}=(\mathbf{x}_1-\mathbf{x}_0)'\boldsymbol{\beta} \tag{8}$$

is proportional to the distance between $\mathbf{x}_0$ and $\mathbf{x}_1$, and does not depend on t. McCullagh (1980) called this property the proportional odds assumption.

In the RR design, Z is not directly observed. Therefore, the cumulative probabilities $\mathbb{P}(Z\leqslant t|\mathbf{x})$ are modelled through the observed variable Z^*, with the relationship between Z^* and Z given by the RR sum score model. The RR proportional odds model is given by

$$\gamma_s^*=\sum_{j=0}^{s}\sum_{t=0}^{M} q_{j|t}(\gamma_t-\gamma_{t-1}), \tag{9}$$

where $\gamma_s^*=\mathbb{P}(Z^*\leqslant s|\mathbf{x})$.

3.2.1. Estimation

The ML estimator of model (9) is obtained by maximization of the kernel of the observed data log-likelihood, which is given by

$$\ln\{l(\boldsymbol{\beta},\boldsymbol{\alpha}|z_i^*,\ldots,z_n^*,\mathbf{x}_i,\ldots,\mathbf{x}_n)\}=\sum_{i=1}^{n}\ln\left\{\sum_{t=0}^{M} q_{z_i^*|t}(\gamma_t-\gamma_{t-1})\right\}, \tag{10}$$

where $\gamma_{-1}=0$ and $\gamma_M=1$. To identify the model, we use the convention $\alpha_0=0$. For the maximization of expression (10) standard optimization routines can be used. To estimate the models in the social security survey examples we use the quasi-Newton optimization routine `QNewtonmt` of the statistical package GAUSS. The gradients and Hessian matrix are computed numerically by using the Broyden–Fletcher–Goldfarb–Shanno method. For solutions in the interior of the parameter space standard asymptotic theory applies with respect to the normal distribution of

the estimators, and we report the asymptotic standard errors that are derived from the estimated Hessian matrix. In case of a boundary solution the normality assumption is no longer valid, and we report 95% bootstrap confidence intervals that are derived from 500 non-parametric bootstrap samples by using the percentile method.

4. The example

In this section, we analyse the sum score variable $Z = \Sigma_{m=1}^{3} Y_m$, denoting the number of yes responses to questions 1–3 of the social security survey, with the RR sum score model and the RR proportional odds model. The frequencies of the sum scores 0, 1, 2 and 3 that are observed in the sample are given by the vector $\mathbf{n}^* = (811, 649, 245, 55)$.

The respective MM sum score probability estimates of the RR sum score model are $\hat{\boldsymbol{\pi}} = (0.850, 0.075, 0.058, 0.017)$. Since the MM estimates are all in the interior of the parameter space, the ML solution is identical. The log-likelihood of the ML solution is -1949.54. The same probability estimates and log-likelihood can also be obtained with the RR proportional odds null model, i.e. the model without any covariates except the intercept. The parameter estimates of the null model are $\hat{\beta}_0 = -1.74$, $\hat{\alpha}_1 = 0.77$ and $\hat{\alpha}_2 = 2.32$, and the sum score probabilities are found by plugging these estimates into $\hat{\gamma}_t$ defined in equation (6), and using the expression $\hat{\pi}_t = \hat{\gamma}_t - \hat{\gamma}_{t-1}$.

Table 1 presents the parameter estimates of the RR proportional odds model with all six covariates. The log-likelihood of this model is -1937.84, yielding a likelihood ratio test statistic of 23.4 with 6 degrees of freedom in relation to the null model. The parameter estimates of the covariates gender, age, last job contract and education are significant. To interpret these results, we use the property of the proportional odds model that, for all t, the odds of non-compliance with more than t rules change with a factor $\exp(-\beta_j)$ for each unit increase in covariate j, holding all other covariates constant. The parameter estimate for gender indicates that for men the odds of non-compliance are about 2.3 times those for women. Similarly, the odds of non-compliance for people above the age of 45 years and for people who had a regular job contract are about 1.8 times that for younger people and people who had a different kind of job contract respectively. Finally, the odds of non-compliance decrease with a factor 0.73 for each increase in the level of education.

To test whether the proportional odds assumption holds for this model, we performed a likelihood ratio test with respect to the RR unconstrained partial proportional odds model (Peterson and Harrell, 1990), that is given by

Table 1. Parameter estimates of the RR proportional odds model

Parameter	*Estimate (standard error)*	*t-value*
α_1	0.99 (0.31)	3.10
α_2	2.46 (0.38)	6.46
Intercept	−0.85 (0.46)	−1.84
Gender	−0.81 (0.26)	−3.14
Education	0.32 (0.16)	2.05
Age	−0.57 (0.28)	−2.23
Time unemployed	0.13 (0.16)	0.80
Last job contract	−0.57 (0.29)	−1.99
Degree of disability	−0.26 (0.25)	−1.05

$$\text{logit}(\gamma_t) = \alpha_t - \mathbf{x}'\boldsymbol{\beta} - \mathbf{w}'\boldsymbol{\eta}_t, \tag{11}$$

where the $k \times 1$ vector $\mathbf{w}$ contains a subset of the values in $\mathbf{x}$, and $\boldsymbol{\eta}_t$ is a $k \times 1$ vector with regression parameters, for $t \in \{1, \ldots, M-1\}$. If $\boldsymbol{\eta}_t = \mathbf{0}$ for all $t \in \{1, \ldots, M-1\}$, the RR unconstrained partial proportional odds model reduces to the RR proportional odds model. The likelihood ratio simultaneously tests the null hypothesis that for all covariates in $\mathbf{w}$ the cumulative odds ratios do not depend on t. For the model with all six covariates included in $\mathbf{w}$ and the parameter vector $\boldsymbol{\eta}_t$ specified for $t \in \{1, 2\}$, the likelihood ratio statistic of 8.2 with 12 degrees of freedom ($p = 0.77$) indicates that the proportional odds assumption need not be rejected. Note that the likelihood ratio statistic at the same time implies that the proportional odds assumption holds for the four significant covariates in Table 1. By setting the contribution to the likelihood ratio statistic LR of the two non-significant covariates to 0 we obtain LR = 8.2, 8 degrees of freedom, and $p = 0.41$.

5. Boundary solutions

Fitting the RR proportional odds null model to the observed frequency vector $\mathbf{n}^* = (694, 601, 329, 108, 28)$ of $Z^* = \Sigma_{m=1}^{4} Y_m^*$ denoting the number of yes responses to the four questions 1–4 yields the solution $\hat{\beta}_0 = -1.31$, $\hat{\alpha}_1 = -0.46$, $\hat{\alpha}_2 = 1.98$ and $\hat{\alpha}_3 = 2.22$. Note that this solution does not satisfy condition (7), since

$$\hat{\alpha}_1 < \alpha_0 \equiv 0 < \hat{\alpha}_2 < \hat{\alpha}_3.$$

The vector $\hat{\boldsymbol{\pi}} = (0.906, -0.065, 0.134, 0.013, 0.012)'$ that is implied by this solution coincides with the MM solution of the RR sum score model. Obviously, this is not a valid solution since $\hat{\pi}_1$ is outside the parameter space.

To force the threshold parameter estimates to satisfy condition (7) we use the parameterization

$$\alpha_t = \alpha_0 + \sum_{j=1}^{t} \exp(\dot{\alpha}_j), \tag{12}$$

and we maximize log-likelihood (10) for $\dot{\alpha}_j$ and $\boldsymbol{\beta}$, with α_0 constrained to 0. This parameterization yields the solution $\hat{\dot{\alpha}}_1 = -10.92$, $\hat{\dot{\alpha}}_2 = 0.46$ and $\hat{\dot{\alpha}}_3 = -0.02$ (corresponding to $\hat{\alpha}_1 = 0.00$, $\hat{\alpha}_2 = 1.58$ and $\hat{\alpha}_3 = 2.56$), and $\hat{\beta}_0 = -1.88$. The vector $\hat{\boldsymbol{\pi}} = (0.867, 0.000, 0.102, 0.019, 0.012)'$ that is implied by this solution is valid and coincides with the ML estimates of the RR sum score model.

Table 2 presents the parameter estimates of the full RR proportional odds model by using parameterization (12). Since we have a boundary solution with the estimate of $\dot{\alpha}_1$ tending to $-\infty$, we report the 95% bootstrap confidence intervals. The confidence intervals of the threshold parameters α_t are obtained after applying equation (12) to the bootstrap estimates of the parameters $\dot{\alpha}_j$. The log-likelihood of the model is -2251.87, yielding a likelihood ratio test statistic of 19.9 with 6 degrees of freedom in comparison with the corresponding null model. The parameter estimates for the covariates gender and last job contract show significance.

Since the RR logistic regression model is a special case of the RR proportional odds model, it is informative to compare the results of both models for respectively the binary variables Y_1–Y_4 and the sum score variable Z. Table 3 presents the regression parameter estimates of the RR logistic model specified as in expression (6), i.e. with a negative sign for the vector $\boldsymbol{\beta}$. The probability estimates $\hat{\pi}_1$ are obtained by fitting separate RR sum score models for each Y-variable. The solution of the RR logistic regression model with dependent variable Y_1^* is unstable with large parameter estimates and standard errors. The instability of this model is most likely due to

Table 2. Parameter estimates and 95% bootstrap confidence intervals CI_{boot} of the full RR proportional odds model with parameterization $\dot{\alpha}$

Parameter	*Estimate*	*95% CI_{boot}*
α_1	0.00	(0.00, 0.31)
α_2	2.01	(1.12, 3.02)
α_3	2.53	(1.98, 3.84)
Intercept	−1.02	(−2.01, −0.25)
Gender	−0.76	(−1.26, −0.26)
Education	0.21	(−0.06, 0.46)
Age	−0.42	(−0.86, 0.05)
Time unemployed	0.13	(−0.10, 0.38)
Last job contract	−0.60	(−1.14, −0.09)
Degree of disability	−0.25	(−0.71, 0.29)

Table 3. Parameter estimates (with standard errors in parentheses) of the RR logistic regression model for variables Y_1–Y_4

Parameter	Y_1	Y_2	Y_3	Y_4
$\hat{\pi}_1$	0.018	0.099	0.125	0.047
Intercept	−5.36 (5.68)	−1.42 (0.57)	−1.38 (0.47)	−1.93 (0.83)
Gender	2.53 (5.38)	−0.94 (0.34)	−0.83 (0.30)	−0.46 (0.59)
Education	1.43 (1.42)	0.58 (0.22)	0.13 (0.16)	−0.28 (0.35)
Age	−7.44 (30.8)	−0.77 (0.33)	−0.14 (0.30)	0.10 (0.51)
Time unemployed	−1.36 (1.01)	0.10 (0.18)	0.08 (0.16)	−0.03 (0.14)
Last job contract	−0.75 (1.64)	−0.55 (0.37)	−0.59 (0.34)	−1.15 (0.62)
Degree of disability	0.07 (0.28)	−0.46 (0.31)	−0.13 (0.32)	0.37 (0.69)

the fact that $\hat{\pi}_1$ is close to 0, so that little information is available to estimate the parameters. In the model with Y_2 the covariates age, education and gender are significant, and the last is also significant in the model with Y_3. The model with Y_4 shows no significant results. In comparison, the RR proportional odds models also show significant results for the covariates age, education and gender, but in addition reveal a significant relationship between regulatory non-compliance and the covariate last job contract. This shows that both models may provide different insights into the relationship between the dependent variables and the covariates; covariates that are significantly related to the sum scores of multiple sensitive characteristics may not be significantly related to any of the separate sensitive characteristics.

6. Conclusions

This paper discusses the construction and analysis of RR sum score variables that are composed of multiple binary RR variables measuring a range of sensitive characteristics. The paper introduces the RR sum score model that can be used to obtain the probability distribution of the sum scores of the sensitive characteristics, and the RR proportional odds model that can be used to analyse the dependence of the sum score probabilities of the sensitive characteristics on a set of covariates. Special attention is devoted to various estimation methods and to bound-

ary solutions that are characterized by sum score probability estimates on the boundary of the parameter space. Both of the models are applied to two sets of sum score data from a social security survey, and the analysis of one data set illustrates a boundary solution.

The analysis of a sum score variable provides additional information about the distribution of the sensitive characteristics that are under study. For example, the distribution and determinants of the sum score probabilities of regulatory non-compliance may contain valuable information for law enforcers and policy makers. Moreover, the analysis of sum score data may reveal associations that remain undetected if the data are analysed in a univariate way. In the examples, the RR proportional odds model detected an association between regulatory non-compliance and the last job contract, an association that was not found in the RR logistic model. These differences result from the fact that each model addresses different questions. Therefore the choice of a model should ultimately be based on the research question; the RR logistic regression model is appropriate if interest is in the predictors of a single sensitive characteristic, and the RR proportional odds model is appropriate if interest is in the predictors of the sum score distribution of multiple sensitive characteristics.

The second example shows that the RR proportional odds model can successfully handle boundary solutions. However, this does not necessarily mean that the model is correctly specified. In this respect, the validity of the model depends on how the boundary solution came about. One explanation for the occurrence of boundary solutions is chance. For example, if the prevalence of the sensitive characteristic is 0 or close to 0, a boundary solution is obtained if the proportion of respondents who throw 2, 3 or 4 with the two dice is less than 1/6. Obviously, this type of chance result does not invalidate the model. Another explanation for a boundary solution is that respondents protect their privacy by answering no when according to the outcome of the dice they should have answered yes. Böckenholt and van der Heijden (2007) propose a Rasch model with an extra parameter to account for the effects of self-protective response bias on the response profiles of multiple RR variables. The results of this study suggest that self-protective responses significantly affect the prevalence estimates. In the case of RR sum score data, self-protective responses would lead to a systematic overestimation of the zero sum score probability. If self-protective responses occur, the RR sum score model and the RR proportional odds model are both misspecified, and additional research is needed to account for this kind of response bias.

To conclude we mention that the RR proportional odds model can be extended to weighted sum scores, where Z and Z^* are weighted sums of respectively Y_m and Y_m^*, with the weights given by w_m, $m \in \{1, \ldots, M\}$. By analogy with the sum score variables, the conditional misclassification probabilities for the weighted sum score variables can be found as a function of the misclassification probabilities for the binary variables Y_m and Y_m^*, since these are not affected by the weights.

References

Böckenholt, U. and van der Heijden, P. G. M. (2007) Item randomized-response models for measuring noncompliance: risk-return perceptions, social influences, and self-protective responses. *Psychometrika*, **72**, 245–262.

Boruch, R. F. (1971) Assuring confidentiality of responses in social research: a note on strategies. *Am. Sociol.*, **6**, 308–311.

Elffers, H., van der Heijden, P. G. M. and Hezemans, M. (2003) Explaining regulatory noncompliance: a survey study of rule transgression for two Dutch instrumental laws, applying the randomized-response method. *J. Quant. Crimin.*, **4**, 409–439.

Fox, J. P. (2005) Randomized item response theory models. *J. Educ. Behav. Statist.*, **30**, 1–24.

van Gils, G., van der Heijden, P. G. M., Laudy, O. and Ross, R. (2003) *Regelovertreding in de Sociale Zekerheid*. The Hague: Ministry of Social Affairs and Employment.

Horvitz, D. G., Shah, B. V. and Simmons, W. R. (1967) The unrelated question randomized response model. *Proc. Socl Statist. Sec. Am. Statist. Ass.*, 65–72.

van den Hout, A. and van der Heijden, P. G. M. (2002) Randomized response, statistical disclosure control and misclassification: a review. *Int. Statist. Rev.*, **70**, 269–288.

van den Hout, A., van der Heijden, P. G. M. and Gilchrist, R. (2007) The logistic regression model with response variables subject to randomized response. *Computnl Statist. Data Anal.*, **51**, 6060–6069.

Kuha, J. and Skinner, C. (1997) Categorical data analysis and misclassification error. In *Survey Measurement and Process Quality* (eds L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz and D. Trewin). New York: Wiley.

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

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

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M. and Maas, C. (2005) Meta-analysis of randomized response research, thirty-five years of validation. *Sociol. Meth. Res.*, **33**, 319–348.

Maddala, G. (1983) *Limited Dependent and Quantitative Variables in Econometrics*. New York: Cambridge University Press.

McCullagh, P. (1980) Regression models for ordinal data (with discussion). *J. R. Statist. Soc.* B, **42**, 109–142.

Peterson, B. and Harrell, Jr, F. E. (1990) Partial proportional odds models for ordinal response variables. *Appl. Statist.*, **39**, 205–217.

Scheers, N. J. and Dayton, C. M. (1988) Covariate randomized response models. *J. Am. Statist. Ass.*, **83**, 969–974.

Warner, S. L. (1965) Randomized response: a survey technique for eliminating answer bias. *J. Am. Statist. Ass.*, **60**, 63–69.