A multivariate logistic regression model for randomized response data

Ardo van den Hout[1], Peter G.M. van der Heijden[2] and Robert Gilchrist[3]

[1] MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, CB2 2SR, UK.
[2] Department of Methodology and Statistics, Faculty of Social Sciences, Utrecht University, P.O. box 80140, 3508 TC Utrecht, the Netherlands
[3] STORM Research Center, London Metropolitan University, 2-16 Eden Grove, London N7 8EA, UK.

Abstract: A multivariate logistic regression model is discussed given that the dependent variables are subject to randomized response. Randomized response is an interview technique that can be used when sensitive questions have to be asked and respondents are reluctant to answer directly. The multivariate model is an adaption of a model without randomized response variables and a Fisher scoring algorithm is used to maximize the likelihood. Randomized response data taken from a study into social benefit fraud are used to illustrate the approach.

Keywords: Multivariate logistic regression; Misclassification; Randomized response; 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.
In recent years, RR techniques have been investigated and applied in the Netherlands. Van Gils, Van der Heijden, Laudy, and Ross (2003) report about rule transgression with respect to social benefits and Elffers, Van der Heijden, and Hezemans (2003) used RR to study rule transgression regarding two Dutch instrumental laws. RR is also applied outside the Netherlands, see, e.g., Fisher, Kupferman, and Lesser (1992), and Lara, Strickler, Olavarrieta, and Ellertson (2004). A meta-analysis by Lensvelt-Mulder, Hox, Van der Heijden, and Maas (2005) shows that RR yields more valid prevalence estimates than other methods for sensitive questions.

In the case the sensitive question is a *yes*/*no* question, the univariate logistic regression model can be used to investigate how the sensitive behavior is associated with covariates such as, e.g., gender and age. Given the misclassification induced by the RR design, the standard logistic regression model has to be adapted to take into account the extra perturbation. For the Warner RR design, the univariate RR logistic regression model was first described by Maddala (1983). Further research in this area is presented by Van der Heijden and Van Gils (1996). In case all the covariates are discrete, loglinear models can be applied, see Chen (1989) and Van den Hout and Van der Heijden (2004).
Often in RR surveys there are several sensitive questions and we would like to be able to investigate - within one model - the associations between a number of RR variables and covariates. The present paper introduces a multivariate logistic regression model for dependent variables subject to RR. The model makes it possible to assess several RR variables and a set of covariates jointly. There are various ways to define a multivariate logistic regression model, see 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. The estimation of this RR multivariate model is an extension of the Fisher scoring algorithm as presented by Glonek and McCullagh (1995).
The application at the end of the paper illustrates how the method can be used by analyzing RR data taken from a study into social benefit fraud.

2 The randomized response design

The forced response design (Boruch 1971) is an illustrative example of an RR design. Assume that the sensitive question asks for a *yes* or a *no*. 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 implies a latent *yes*.
Let Y be the latent binary 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

$$\begin{aligned} I\!P(Y^*=1) &= I\!P(Y^*=1|Y=0)I\!P(Y=0) + I\!P(Y^*=1|Y=1)I\!P(Y=1) \\ &= 2/12 I\!P(Y=0) + 11/12 I\!P(Y=1) \end{aligned}$$

Note that probabilities $I\!P(Y^*=j|Y=k)$ are fixed for $j,k \in \{0,1\}$ by the known distribution of the sum of the two dice. RR variables can be seen as

misclassified variables, where conditional misclassification probabilities are given by $I\!P(Y^* = j|Y = k)$.
The general RR design for variable Y can be described by using a matrix $\mathbf{P}_Y = (p_{jk})$, where $p_{jk} = I\!P(Y^* = j|Y = k)$. If the variable has 2 categories, the misclassification design is given by $\pi^* = \mathbf{P}_Y\pi$, where $\pi^* = (\pi_0^*, \pi_1^*)^t$ is the distribution of the observed variable Y^* and $\pi = (\pi_0, \pi_1)^t$ is the distribution of the latent variable Y. 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 $\mathbf{P}_{Y_1}$ and the misclassification of Y_2 is described by $\mathbf{P}_{Y_2}$, the misclassification of the Cartesian product $Y = (Y_1, Y_2)$ is described by $\mathbf{P}_Y = \mathbf{P}_{Y_1} \otimes \mathbf{P}_{Y_2}$, where $\otimes$ denotes the Kronecker product.
In RR surveys there is always the question whether respondents comply with the design. An example of non-compliance would be a respondent that answers *no* whereas the design forces the answer to be *yes*. We shall not address this problem here although its importance is obvious. Instead we refer to Cruyff, Van den Hout, Van der Heijden, and Böckenholt (2006) - a paper in these proceedings - for further discussion and references.

3 Multivariate logistic regression

In the multivariate logistic regression model, the link between the linear predictor η and the expected response vector π is

$$\eta = \mathbf{C}^t \log(\mathbf{L}\pi), \tag{1}$$

where $\mathbf{C}$ is the contrast matrix and $\mathbf{L}$ is the marginal indicator (Glonek and McCullagh 1995) .
We illustrate the model for the bivariate case with the binary responses Y_1 and Y_2. Let $\pi = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})^t$, where $\pi_{kl} = I\!P(Y_1 = k, Y_2 = l)$, for $k, l \in \{0, 1\}$. Given $\eta = (0, \eta_{Y_1}, \eta_{Y_2}, \eta_{Y_1Y_2})^t$, the link functions are given by

$$\eta_{Y_1} = \log\frac{\pi_{1+}}{1-\pi_{1+}}, \qquad \eta_{Y_2} = \log\frac{\pi_{+1}}{1-\pi_{+1}}, \qquad \eta_{Y_1Y_2} = \log\frac{\pi_{00}\pi_{11}}{\pi_{01}\pi_{10}}, \tag{2}$$

where the plus subscript denotes summation over the index. The regression equations are given by

$$\eta_{Y_1} = \beta_{Y_1}^t x_{Y_1}, \qquad \eta_{Y_2} = \beta_{Y_2}^t x_{Y_2}, \qquad \eta_{Y_1Y_2} = \beta_{Y_1Y_2}^t x_{Y_1Y_2}. \tag{3}$$

This model is a marginal model - it implies univariate logistic models for both Y_1 and Y_2 marginally. 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

$$\mathbf{C}^t = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 & 1 \end{pmatrix}$$

and

$$\mathbf{L}^t = \begin{pmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

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

$$\mathbf{X}_i = \begin{pmatrix} 0 & 0 & 0 \\ x_{Y_1,i}^t & 0 & 0 \\ 0 & x_{Y_2,i}^t & 0 \\ 0 & 0 & x_{Y_1Y_2,i}^t \end{pmatrix}.$$

To estimate the RR multivariate logistic regression model, assume that data are given by independent observations $(y_i^*, \mathbf{X}_i)$, $i = 1, ..., n$, where y_i^* is a draw from a multinomial distribution with parameter π^*. Let the RR design be described by $\mathbf{P}$. The kernel of the loglikelihood is

$$l(\beta|\text{ data}, \mathbf{P}) = \sum_{i=1}^{n} (y_i^*)^t \log \pi^* = \sum_{i=1}^{n} (y_i^*)^t \log \mathbf{P}\pi.$$

Note that $\mathbf{P}$ will be the Kronecker product of the matrices that describe the RR design per RR question. For example, if there are two binary RR questions, $\mathbf{P}$ is a 4×4 matrix, and $y_i^* \in \{(1,0,0,0)^t, (0,1,0,0)^t, (0,0,1,0)^t, (0,0,0,1)^t\}$, for $i = 1, ..., n$.

The maximization of the loglikelihood consists of two parts that are iterated. Given a starting value of β, the first part uses a Newton-Raphson iteration to obtain π given η. 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 β.

Part one. To invert equation (1), work with $v = \log(\pi)$. The Newton-Rahpson iterations are described by.

(a) Initial approximation is v_0.

(b) Next

$$v_n = v_{n-1} - \{\mathbf{C}^t(\text{diag}(\mathbf{L}\exp v_{n-1}))^{-1}\mathbf{L}\}^{-1}\{\mathbf{C}^t \log(\mathbf{L}\exp v_{n-1}) - \eta\}.$$

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

Part two. The adaptation of the scoring algorithm is straigthforward given the above. The derivative $\partial\pi/\partial\beta$ is given by $\left(\mathbf{C}^t\text{diag}(\mathbf{L}\pi)^{-1}\mathbf{L}\right)^{-1}\mathbf{X}$. The score statistic and the Fisher information matrix can be derived from the loglikelihood and are given by

$$\begin{aligned} s(\beta|y_i^*) &= \left(\mathbf{P}\frac{\partial\pi}{\partial\beta}\right)^t \text{diag}(\mathbf{P}\pi)^{-1} y_i^* \\ \mathcal{I}(\beta|y_i^*) &= I\!E_{Y_i^*}[s(\beta; y_i^*)s^t(\beta; y_i^*)] = \left(\mathbf{P}\frac{\partial\pi}{\partial\beta}\right)^t \text{diag}(\mathbf{P}\pi)^{-1}\left(\mathbf{P}\frac{\partial\pi}{\partial\beta}\right). \end{aligned}$$

The information matrix is derived by using the fact that Y_i^* is multinomially distributed with parameter vector $\pi^* = \mathbf{P}\pi$. That is, $I\!E_{Y_i^*}[Y_{ij}^*] = \pi_j^* = I\!E_{Y_i^*}[(Y_{ij}^*)^2]$ and $I\!E_{Y_i^*}[Y_{ij}^* Y_{ik}^*] = 0$ for $j \neq k$, see, e.g., Agresti (2002, Section 14.1).
Note that the above model does not make sense if the cross classification of the latent dependent variables contains one or more structural zeroes. For instance, in a 2×2 table with one structural zero, the estimated odds ratio is infinity. In the case of sample zeroes, we propose to smooth the data in the maximization procedure. That is, after part one add small probability mass to each entry of $\widehat{\pi}$, normalize $\widehat{\pi}$ such that the entries sum up to one, and next go to part two.

4 Application

We illustrate the above method by analyzing data taken from an RR survey concerning unemployment benefit in the Netherlands in 2004. Let y_1^* denote the observed answer to the question whether the respondent applied frequently enough for new jobs (*yes* = 1, *no* = 0). Likewise let y_2^* denote the observed answer to the question whether the respondent did any voluntary work without reporting this. In addition, x_1 is sex (*male* = 1, *female* = 0), x_2 is age in years, and x_3 denotes whether the income of the respondent constitutes the larger part of the income of the household (*yes* = 1, *no* = 0). The sample size is $n = 753$.
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.
Table 1 presents the results of the analysis. Model 6 is defined by (1) and (2) and the regression equations $\eta_i = \mathbf{X}_i\beta$ are given by $\beta = (\beta_{Y_1 0}, \beta_{Y_1 1}, \beta_{Y_1 2}, \beta_{Y_1 3}, \beta_{Y_2 0}, \beta_{Y_2 1}, \beta_{Y_2 2}, \beta_{Y_2 3}, \beta_{Y_1 Y_2 0})^t$ and

$$\mathbf{X}_i = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & x_{Y_1 1,i} & x_{Y_1 2,i} & x_{Y_1 3,i} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & x_{Y_2 1,i} & x_{Y_2 2,i} & x_{Y_2 3,i} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

So the odd ratio is estimated by using one parameter, i.e., an intercept. Models 1 up to 5 are defined by restrictions on the parameters, see Table 1. Both the analysis of deviance and the Akaike Information Criterion (AIC) favor model 4 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

$$\begin{aligned} \widehat{\beta}_{Y_1 0} &= -0.028\ (0.544) & \widehat{\beta}_{Y_2 0} &= -3.309\ (0.683) \\ \widehat{\beta}_{Y_1 2} &= -0.037\ (0.014) & \widehat{\beta}_{Y_2 2} &= 0.039\ (0.013) \\ \widehat{\beta}_{Y_1 3} &= \widehat{\beta}_{Y_2 3} = 0.444\ (0.333) & \widehat{\beta}_{Y_1 Y_2 0} &= 1.300\ (0.536), \end{aligned}$$

where the estimated standard errors are within the brackets. The standard errors are estimated by evaluating the information matrix in the optimum.

TABLE 1. Deviances of bivariate logistic regression models taken the RR design into account. Sample size is 753. Notation: $x_j(k,..)$ means the covariate x_j is included in the k^{th} regression equation, where the order of the equations is given by (3).

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) $x_2(1,2)$, $x_3(1)$	6	1903.799	1915.799
4	Intercept(1,2,3) $x_2(1,2)$, $x_3(1,2)$ and $\beta_{Y_1 3} = \beta_{Y_2 3}$	6	1901.638	1913.638
5	Intercept(1,2,3) $x_2(1,2)$, $x_3(1,2)$	7	1901.611	1915.611
6	Intercept(1,2,3) $x_1(1,2)$, $x_2(1,2)$, $x_3(1,2)$	9	1900.889	1918.889

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. One of the advantages of the multivariate model, however, is that we can easily investigate whether an effect of a covariate is the same in the marginal models. For the data at hand, this is illustrated by the chosen model 4 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 fraud regarding applying for jobs and the prevalence of fraud regarding voluntary work. More specific, model 4 states that a change from $x_3 = 0$ to $x_3 = 1$, means that the odds on fraud both on Y_1 and Y_2 changes multiplicatively by $\exp(\widehat{\beta}_{Y_1 3}) = \exp(\widehat{\beta}_{Y_2 3}) = \exp(0.444) = 1.559$. So the odds on fraud is higher when a person's income constitutes the larger part of the income of the household.

References

Agresti, A. (2002). *Categorical data analysis, second edition.* New York: Wiley.

Boruch, R.F. (1971). Assuring confidentiality of responses in social research: a note on strategies. *The American Sociologist*, **6**, 308-311.

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

Chen, T. T. (1989). A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine*, **8**, 1095-1106.

Cruyff, M.J.L.F., Van den Hout, A., Van der Heijden, P.G.M., Böckenholt, U. (2006). A Log-Linear Randomized-Response Model to Account for Cheating. In: *Proceedings of the 21st international workshop on statistical modelling.*

Elffers, E., Van der Heijden, P.G.M., and Hezemans, M. (2003). Explaining regulatory non-compliance: A survey study of rule transgression for two Dutch instrumental laws, applying the randomized response method. *Journal of Quantitative Criminology*, **19**, 409-439.

Fahrmeier, L., and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models. Second Ed.* Berlin: Springer-Verlag.

Fisher, M., Kupferman, L.B., and Lesser, M. (1992). Substance use in school-based clinic population: Use of the randomized response technique to estimate prevalence. *Journal of Adolescent Health*, **13**, 281-285.

Glonek, G.F.V., and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Association, Series B*, **57**, 533-546.

Lara, D., Strickler, J., Olavarrieta, C.D., and Ellertson, C. (2004). Measuring induced abortion in Mexico: A comparison of four methodologies. *Sociological Methods and Research*, **32**, 529-558.

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. *Sociological Methods and Research*, **33**, 319-348.

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

Van den Hout, A., and Van der Heijden, P.G.M. (2004). The analysis of multivariate misclassified data with special attention to randomized response data. *Sociological Methods and Research*, **32**, 310-336.

Van der Heijden, P.G.M, and Van Gils, G. (1996). Some logistic regression models for randomized response data. In: *Proceedings of the 11th international workshop on statistical modelling.*

Van Gils, G., Van der Heijden, P.G.M., Laudy, O., Ross, R (2003). Regeloovertreding in de sociale zekerheid. Den Haag: Ministerie van Social Zaken en Werkgelegenheid. (In Dutch)

Warner, S.L. (1965). Randomized response: A survey technique for eliminating answer bias. *Journal of the American Statistical Association*, **60**, 63-69.