A ZIP model accounting for response bias in randomized response.

Maarten Cruyff[1], Ulf Böckenholt[2] and Peter van der Heijden[1]

[1] Utrecht University, Netherlands
[2] McGill University, Canada

Abstract: Randomized response (RR) is an interview technique that ensures confidentiality when sensitive questions are asked. In RR the answer to a sensitive question is partly determined by a randomizing device. Respondent may nevertheless feel uncomfortable with the RR design and give the least sensitive response, irrespective of the outcome of the randomizing device. This kind of self-protective response behavior yields biased prevalence estimates. In this paper we present the zero-inflated Poisson model to account for self-protective response bias in the presence of an RR variable with multiple, ordered response categories. An example from a Dutch social security survey is provided.

Keywords: randomized response; self-protective response bias; zero-inflated Poisson.

1 Introduction

In many surveys sensitive questions need to asked. Examples are questions about alcohol abuse or sexuality. Randomized response (RR) is an interview method especially designed for sensitive questions. In RR the answer depends partly on the true status of the respondent and partly on the outcome of a randomizing device. The outcome is known only to the respondent so that confidentiality is guaranteed.
It is however unrealistic to assume that all respondents comply with the RR design. Studies by Edgell, Himmelfarb and Duchan (1982), and Van der Heijden, Van Gils, Bouts and Hox (2000) show that respondents may exhibit self-protective response behavior by giving the least sensitive answer, regardless the outcome of the randomizing device. Recently models have been proposed that account for self-protective response bias in the presence of multiple binary RR variables. These models use the fact that self-protective response behavior results in an excess of observation of the response profile denying the sensitive characteristic on all questions. As an example, Böckenholt and Van der Heijden (2007) propose IRT models for RR profile data that includes a parameter accounting for the excess of observations of the response profile consisting of only *no* responses.

This paper proposes the zero-inflated Poisson (ZIP) model (Lambert, 1992) to account for self-protective response bias in the presence of a single RR variable with multiple ordered response categories. As an example we present the analysis of a variable from a Dutch social security survey. If this variable is analyzed with the general multinomial RR model we observe an unsatisfactory fit, with too many observations in the nonsensitive response category and too few observations in the more sensitive response categories. This pattern of residuals clearly suggests the presence of self-protective response behavior. If we assume that the categories of the sensitive characteristic follow a Poisson distribution, then we can model the excess of observations in the zero count as zero-inflation due to self-protective response behavior.
The paper is structured as follows. In section 2 we present the question from the Dutch survey on social security fraud. Section 3 presents the general multinomial RR models, and derives the ZIP model for a RR variable with multiple ordered response categories under the assumptions of a Poisson distribution and the presence of self-protective response behavior. In Section 4 we compare the main results of the RR models presented in section 3. Section 5 concludes.

2 The social security survey

In 2004 the Dutch Department of Social Affairs conducts a nationwide survey to assess the level of regulatory noncompliance with the Social Security Law. In this survey 2.580 beneficiaries of social security benefits are asked the following question:

A On average, how much money a month have you earned in the past 12 months in addition to your social security benefits by working off the books ?

The question has six response categories, ranging from 0 euros to 250 euros or more. For this question according the forced-response (FR) design (Boruch, 1971) is used. In this particular application, the respondent tosses two dice and answers the question truthfully if the sum of the two dice equals $5, 6, 7, 8, 9$ or 10. If the sum of the two dice is equal to $2, 3, 4, 11$ or 12, the respondent tosses a single die, and answers the question by naming the number of eyes on the die. The observed response frequencies n_j^* are
The following eleven variables are included in this study as covariates. Beneficiaries of 3 different social security acts are distinguished by the dummy variables *AIA* (Assistance Insurance Act) and *DIA* (Disability Insurance Act), with reference category the Unemployment Insurance Act. The background variables are *gender*, *age* and (level of) *education*. The variables *costs compliance*, *benefits noncompliance*, *norm conformity* and *informal control* assess motives for noncompliance with the rules. The two variables

TABLE 1. Response frequencies.

j	1	2	3	4	5	6
in EUR	0	1-50	51-75	76-100	101-250	250^+
n_j^*	2014	245	108	74	72	67

trust and *understanding* assess the respondent's attitude towards the RR design.

3 The RR ZIP model

Let random variable U denote the true status on the sensitive characteristic and let U^* denote the response to the sensitive question, for $u^*, u \in \{1, \ldots, 6\}$. If we define π_{u^*} as the probability that U^* takes on the value u^*, and π_u as the probability that U takes on value u, and we assume a multinomial distribution for U, we obtain the general multinomial RR model

$$\pi_{u^*} = \sum_{u=1}^{6} p_{u^*|u} \pi_u, \tag{1}$$

where $p_{u^*|u}$ denotes the conditional misclassification probability of observing response u^* given the true status u (Chaudhuri and Mukerjee, 1988). The conditional misclassification probabilities can be derived from the distribution of the dice. In the given FR design

$$p_{u^*|u} = \begin{cases} 19/24 & \text{if} \quad u^* = u \\ 1/24 & \text{if} \quad u^* \neq u. \end{cases}$$

We can also interpret the response categories to the sensitive question as counts of units of income. We can then define the variable $Y = U - 1$, so that for $Y = 0$ denote the event that "0 units of income" are earned, and $Y = 5$ that "5 units of income or more". If, for $y, y^* \in \{0, 1, \ldots, 5\}$, we assume a censored Poisson distribution for Y we obtain the Poisson RR model

$$\pi_{y^*}^{RR} = \begin{cases} \sum_{y=0}^{4} p_{y^*|y} \pi_y & \text{if } y < 5 \\ p_{y^*|5} \left(1 - \sum_{y=0}^{4} p_{y^*|y} \pi_y\right) & \text{if } y = 5, \end{cases} \tag{2}$$

with $p_{y^*|y}$ equal to $p_{u^*|u}$, and $\pi_{y^*}^{RR}$ denoting the probability of observing the event $Y^* = y^*$ under the RR scheme.

We now introduce the zero-inflated Poisson model (Lambert, 1992) with the zero-inflation parameter θ to account for the effect of self-protective response behavior. Since self-protective response behavior results in an excess of zeros on the observed variable Y^*, we obtain the ZIP RR model

$$\pi_{y^*} \quad = \quad (1-\theta)\pi_{y^*}^{RR} + I_{(y^*=0)}\theta, \tag{3}$$

where π_{y^*} denotes the probability of observing the event $Y^* = y^*$, and $I_{(y^*=0)}$ is an indicator variable taking on the value 1 if $Y^* = 0$, and 0 otherwise.
The covariate vectors $\mathbf{x_i}$ explaining the RR generated count distribution and $\mathbf{z_i}$ accounting for the self-protective response bias can be included in model (3) by respectively

$$\lambda_i = \exp(\mathbf{x_i}\beta) \text{ and } \theta_\mathbf{i} = \frac{\exp(\mathbf{z_i}\gamma)}{\mathbf{1} + \exp(\mathbf{z_i}\gamma)}, \tag{4}$$

for $i \in \{1, 2, \ldots, n\}$.

4 Main results

Table 2 compares the fit of the models presented in the previous section. Only the final model (4) includes covariates. The two variables *trust* and *understanding* compose the covariate vector $\mathbf{z}$ and are used to explain the zero-inflation. The remaining nine variables compose the covariate vector x and explain differences in the individual Poisson parameters.

TABLE 2. Model fitting

RR model	AIC	X^2
Model (1): multinomial model	4425	43.4
Model (2): Poisson model	4422	46.8
Model (3): ZIP model	4376	5.3
Model (4): ZIP model with covariates	4279	0.7

In terms of the Akaike Information Criterion (AIK) Model (2) assuming a Poisson distribution for the units of money earned by working off the books fits slightly better than the multinomial model (1). The X^2 statistics indicate however that the fit of both models is far from satisfactory. Substantial reductions in the AIK are accomplished by the inclusion of a zero-inflation parameter in Model (3), and the inclusion of the 9 β and 2 γ parameters in Model (4). In terms of X^2 both ZIP models seem to exhibit a satisfactory fit.

Model (4) yields an estimate of the zero-inflation parameter of $\widehat{\theta} = 0.40$, and an estimated probability vector for the units of money earned by working off the books of $\widehat{\pi}_y = (.792, .163, .034, .008, .002, .001)$, with Pearson residuals vector $(.1, -.6, .3, -.1, .5, .0)$. As a comparison, consider the corresponding vectors $\widehat{\pi}_y = (.935, .065, 0, 0, 0, 0)$ and $(2.2, .8, .0, -3.3, -3.4, -3.9)$ of the multinomial RR model (1). The vector with the Pearson residuals of this model shows a systematic pattern which is in line with our assumption about self-protective response behavior: in the observed data the zero count is overrepresented and the higher counts are underrepresented.

TABLE 3. Parameter estimates β and γ of Model (4)

covariate	$\widehat{\beta}$	se	covariate	$\widehat{\gamma}$	se
constant	-2.08	0.45	constant	-0.74	0.32
AIA	-0.23	0.26	*trust*	-0.33	0.42
DIA	0.04	0.11	*understanding*	-0.75	0.44
gender	0.15	0.19			
age	0.02	0.02			
education	0.48	0.24			
costs compliance	1.03	0.33			
benefits noncompliance	-2.30	0.34			
norm conformity	0.32	0.43			
informal control	1.25	0.41			

Table 3 shows the estimates for the β and γ parameters. For the β parameters the estimates of *cost of compliance*, *benefits of noncompliance*, *education* and *informal control* are significant, indicating that high costs of compliance, high benefits of noncompliance, higher education and less informal control are associated with higher earnings from working off the books. Judging by the standard errors the estimates of the γ parameters do not seem to be significant. However, the removal of the covariate *understanding* from the model results in a significant likelihood-ratio ($LR = 7.0$, df=1, $p < .001$). This results suggests that the zero-inflation in the responses to the sensitive question is negatively related to the extent to which the respondent understands when to answer *yes* and when *no*.

4.1 Conclusions

In this paper we present a ZIP model accounting for self-protective response bias in RR and we illustrate the model with an example from a Dutch social security survey. Obviously, the RR variable in our example is a pseudo count variable and therefore the Poisson assumption is questionable. However, we feel that results of the ZIP model are more realistic than those of the models

that assume a multinomial distribution and ignore self-protective response bias.

Table 2 clearly shows that the introduction of the zero-inflation parameter θ substantially improves the fit. Since θ was introduced into the Poisson model to capture the effect of self-protective responses, it seems that the estimate $\widehat{\theta} = 0.40$ indicates the proportion of respondents who exhibit self-protective response behavior. However, this interpretation disregards the possibility of true zero-inflation, for example due to respondents who are unable to perform labor or who are fundamentally opposed to working off the books. We therefore interpret the parameter θ in our model as a reflection of multiple sources of zero-inflation, of which self-protective response behavior is only one source. As a consequence, we take the estimate of 40% as overestimate of self-protective bias.

Lastly, an interesting result is the observed relation between the covariate *understanding* and self-protective response behavior. This result suggests that improving the instructions to the respondents may lead to more valid responses to the sensitive question, while improving the respondent's trust in the RR method does not seem to have a significant effect. It would be interesting to see if this result can be reproduced in other studies.

References

Böckenholt, U., and P.G.M. Van der Heijden (2007). 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, *The American Sociologist* **6**, 308-311.

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

Edgell, S.E., S. Himmelfarb and K.L. Duncan (1982). Validity of forced response in a randomized response model. *Sociological Methods and Research* **11**, 89-110.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**, 1-14.

Van der Heijden, P.G.M., G. Van Gils, J. Bouts and J. Hox (2000). A comparison of Randomized Response, Computer-Assisted Self-Interview and Face-To-Face Direct-Questioning. *Sociological Methods and Research* **28**, 505-537.