Multivariate Analysis of Randomized Response Data

Maarten Cruyff

Multivariate Analysis of Randomized Response Data

Multivariate analyse van randomized response data

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. J.C. Stoof,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op maandag 23 juni 2008, des ochtends te 10.30 uur

door

Maarten Jan Leo Frans Cruyff

geboren op 22 februari 1961, te Den Haag.

Promotoren: Prof. dr. P.G.M. van der Heijden
Prof. dr. U. Böckenholt
Co-promotor: Dr. A. van den Hout

Acknowledgements

This thesis has been written under supervision Peter van der Heijden, Ardo van den Hout and Ulf Böckenholt. I thank Peter for giving me the freedom to pursue my own ideas, while at the same preventing that I got carried away by them. I am indebted to Ardo providing a solid foundation for this thesis with his previous work on randomized response, and for his assistance on mathematical and computational issues. I thank Ulf for the very pleasant and inspiring conversations about statistical modeling of randomized response data; the majority of the models presented in this thesis are based on his ideas.

The research for this thesis was done at the Department of Methodology and Statistics of the Faculty of Social Sciences at Utrecht University. I would like to thank my colleagues for providing me with a friendly and inspiring place to work. My special thanks go out to Laurence Frank, with whom I had the pleasure to frequently exchange ideas on randomized response modeling.

Last but not least, I like to thank Margreet, Floor and Raf for putting up with me during my occasional periods of absent-mindedness, and for showing me that there are more interesting things in this world than randomized response.

Contents

Chapter 1

Introduction

Randomized response is an interview technique that exists for more than forty years. In recent years the focus has slowly shifted from a univariate to a multivariate approach to the analysis of randomized response data. This thesis presents four models for the multivariate analysis of randomized response data. This chapter introduces the randomized response design, presents a brief overview of existing multivariate models, and concludes with an outline for the subsequent chapters.

1.1 Randomized Response

An important topic within the social sciences is the study of the attitudes, opinions and behavior. Surveys and questionnaires are often used to gather information about the way people feel, think or act with respect to such issues as climate change, politics, family values, and so on. This approach works well as long as the respondents answer the questions honestly, but this is usually not the case if the questions concern a sensitive issue, like for example sexual behavior, drug and alcohol consumption or criminal activities. In response to such questions many respondents give the evasive answer (i.e. the answer that denies the sensitive characteristic). It was for this kind of sensitive questions that the randomized response method was developed.

Randomized response is an interview technique in which the answer to the question partly depends on the outcome of a randomizer, like a pair of dice or a deck of cards. In the original randomized response design, that was developed in 1965 by Warner (Warner, 1965), the respondent is presented

with two complementary statements, for example "I use drugs" and "I do not use drugs". The respondent then operates a randomizing device that has potential outcomes A and B. The respondent answers the first statement in case the outcome is A and the second statement in case the outcome is B. Since only the respondent knows the outcome of the randomizer, the interviewer does not know which of the two statements was answered by the respondent, and confidentiality is guaranteed.

There are many variations on the Warner design. Well known examples are the Kuk design (Kuk, 1990), and the forced response design (Boruch, 1971). In the forced response design the respondent is presented with a single question, for example "Do you use drugs?", and then tosses two dice. The respondent is forced to answer *yes* if the outcome of the two dice is $2, 3$ or 4, and *no* if the outcome is 11 or 12. If the outcome is $5, 6, 7, 8, 9$ or 10, the respondent has to answer truthfully.

Due to the randomization, the answer given by the respondent may not coincide with the true behavior of the respondent. Consider a person who uses drugs. In the forced response design this person answers the question "Do you use drugs?" with *yes* if the sum of the dice is $2, 3, 4$ (forced *yes*) or if the sum of the dice is $5, 6, 7, 8, 9, 10$ (truthful answer). So the probability of answering *yes* given the use of drugs is

$$\begin{aligned} I\!P(yes \mid \text{drugs}) &= I\!P(\text{forced } yes) + I\!P(\text{truthful}) \\ &= I\!P_{dice}(2,3,4) + I\!P_{dice}(5,6,7,8,9,10) \\ &= 1/6 + 3/4 \\ &= 11/12. \end{aligned} \tag{1.1}$$

Similarly, the probability that someone who does not use drugs answers *yes* to the question is equal to the probability of a forced *yes* response, and is equal to

$$I\!P(yes \mid \text{no drugs}) = 1/6. \tag{1.2}$$

The probability of a *yes* response thus depends on the unknown probability that someone uses drugs or not and on the known conditional probabilities in (1.1) and (1.2) according to

$$\begin{aligned} I\!P(yes) &= I\!P(yes \mid \text{drugs})I\!P(\text{drugs}) + I\!P(yes \mid \text{no drugs})I\!P(\text{no drugs}) \\ &= 11/12 I\!P(\text{drugs}) + 1/6[1 - I\!P(\text{drugs})]. \end{aligned} \tag{1.3}$$

Since $\mathbb{P}(yes)$ can be estimated from the proportion of observed *yes* responses in the sample, expression (1.3) can be used to estimate the prevalence of drug use in the population. For other randomized response designs similar expressions can be formulated.

Several studies show that data collected with the randomized response design are more valid than data collected with direct questioning designs. In an experimental study van der Heijden *et al.* (2000) compared randomized response techniques to direct questioning and computer-assisted self-interview with respondents who were known to have committed social security fraud. The randomized response conditions yielded the highest prevalence estimates of fraud. A meta-analysis of randomized response studies (Lensvelt-Mulders *et al.*, 2005) also shows that randomized response results in higher prevalence estimates of the sensitive behavior than other methods, and that this effect becomes stronger as the sensitivity of the questions increases.

The randomized response technique has been used in the context of such divers topics as abortion, sexual behavior, drugs, alcohol, criminal offences, ethnical issues, charity, cheating on exams and environmental issues (see Lensvelt-Mulders *et al.*, 2005). In the Netherlands the randomized response method has been used extensively by the Dutch administration to assess law compliance. This research included nationwide surveys about rule compliance with respect to taxi licences, mineral adminstration by farmers, storing of food products by cafeteria, contamination of surface waters by industrial companies, application for individual rent subsidies and social welfare rules and regulations.

1.2 Existing Multivariate Models

The analysis of randomized response data has traditionally focussed on prevalence estimate of the sensitive behavior in question. There are however other interesting research questions that can only be retrieved using a multivariate approach. This section provides an overview of recent examples.

Researchers are usually not only interested in the prevalence of a sensitive behavior itself, but also in the associations between different sensitive behaviors. Randomized response surveys often include multiple sensitive questions that assess different kind of sensitive behavior. Associations patterns between these behaviors can be studied with the *log-linear model*. Chen (1989) adapted the log-linear model to accommodate randomized variables, and a

recent application by van den Hout and van der Heijden (2004) studies associations between noncompliance with various social security regulations.

Another important research question concerns the relationship between sensitive behavior and personal characteristics, such as gender, age, education, and so on. This a question can be answered with a *logistic regression analysis*. Maddala (1983) and Scheers and Dayton (1988) adapted the logistic regression model to randomized response, and recently Elffers, van der Heijden and Hezemans (2003) used the logistic regression model to study the motives for regulatory noncompliance with two Dutch laws.

Item Response Theory (IRT) models are used to study profiles of different sensitive behaviors. The main assumptions underlying this model are that each respondent is characterized by a score on a latent trait variable that explain the observed behavior profile, and that these latent trait scores are in turn explained by personal characteristics. Recent applications in the randomized response context are Böckenholt and van der Heijden (2004, 2007) and Fox (2005).

Although the randomized response method is designed to eliminate evasive response behavior, it highly unlikely that all respondents comply with the instructions. Studies by Edgell, Himmelfarb and Duncan (1982), van der Heijden *et al.* (2000) and Boeije and Lensvelt-Mulders (2002) reveal that respondents have a tendency to protect their own privacy and to give the least incriminating response, regardless of their own status or the outcome of the randomizer. It is obvious that such responses constitute a serious threat to the validity of the data. Böckenholt and van der Heijden (2004, 2007) propose IRT models with an extra parameter that allows for this self-protective response behavior.

1.3 The Models

Chapters 2 to 5 introduce four models for the multivariate analysis of randomized response data and present examples. The models are applied to randomized response data from the the social security surveys that were conducted by the Dutch Department of Social Affaires in the years 2000, 2002, 2004 and 2006. The chapters are based on papers that written for publication in international journals.

Chapter 2 discusses a log-linear model for randomized response data. The distinctive feature of this model is the inclusion of a parameter that

accounts for self-protective response behavior. The model is used to obtain prevalence estimates of and study the associations patterns between multiple sensitive behaviors. An important assumption of the model is that there no highest-order interaction present in the data. Special attention is given to the robustness of the model against violations of this assumption.

Chapter 3 introduces the proportional odds model for randomized response sum score variables. The dependent variable of this regression model denotes the sum score of *yes* responses to multiple binary questions about violations of the social security regulations. The model is used to study the relationship between the number of rule violations and personal characteristics of the social security beneficiaries.

Chapter 4 presents the zero-inflated Poisson model for randomized response sum score variables. As in the previous the dependent variable denotes the sum score of *yes* responses to multiple binary questions about rule violations, but in this model the dependent variable is assumed to follow a Poisson distribution. The model also allows for self-protective response through the inclusion of a zero-inflation parameter. The Poisson and zero-inflation parameters are modeled as a function of covariates. Special attention is given to the tenability of the Poisson assumption with respect the number of rule violations.

Chapter 5 introduces the doubly zero-inflated Poisson model. This model applies to randomized response questions with multiple response categories that denote counts or pseudo counts of a sensitive characteristic (in the presented example the response categories denote amounts of illegally earned money by social security beneficiaries). The model includes two zero-inflation parameters that respectively allow for self-protective response behavior and for persons who are incapable of generating any kind of illegal income. The Poisson and the two zero-inflation parameters are modeled as a function of covariates.

Chapter 2

The Log-linear Model

2.1 Introduction

Since most people are reluctant to answer questions about sensitive topics like the use of drugs or alcohol, sexuality or anti-social behavior, sensitive characteristics are often underreported in surveys and questionnaires. Randomized Response (RR) is an interview technique that is especially designed to eliminate evasive response bias (Warner 1965, Chaudhuri and Mukerjee, 1988). In the RR design, the answer is to a certain extent determined by the outcome of a randomizing device, e.g. a pair of dice or the draw of card. Since the outcome is only known to the respondent, confidentiality is guaranteed. A meta-analysis shows that RR yields more valid prevalence estimates than direct-questioning designs (Lensvelt-Mulders, Hox, Van der Heijden, and Maas 2005).

Although the respondents' privacy is protected, RR does not completely eliminate evasive response bias. Several studies show that some respondents do not always give the affirmative answer when this is required by the RR design. In line with Böckenholt and Van der Heijden (2004), we refer to this answer strategy as self-protection (SP). Edgell, Himmelfarb and Duchan (1982) show the presence of SP in an experimental study, where the outcomes of the randomizing device were fixed in advance. To a question about having experiences with homosexuality, 25% of the respondents who had to

[1]Published as Cruyff, M.J.L.F., van den Hout, A., van der Heijden, P.G.M. and Böckenholt, U. (2007). Log-Linear Randomized-Response Models Taking Self-Protective Response Behavior into Account, *Sociological Methods and Research* **26**, 266-282.

answer *yes* by design gave an SP *no* response. In another study, Van der Heijden, Van Gils, Bouts and Hox (2000) apply different interview techniques to subjects identified as having committed social welfare fraud. Although the RR condition elicited more admission of fraud than direct questioning or computer-assisted self-interviews, a substantial percentage of the subjects still deny having committed fraud. In a study by Boeije and Lensvelt-Mulders (2002), most of the respondents who participated in a computer-assisted RR survey found it difficult to give a false *yes* response and some of them admitted that they had answered *no*.

Some studies have recently focussed on the detection and estimation of SP in RR designs. Clark and Desharnais (1998), who use the term cheating to denote SP, propose to split the sample in two groups and assign different randomization probabilities to each group. They show that significant cheating can be detected if cheating behavior and randomization probability are assumed to be independent. A multivariate approach is taken by Böckenholt and Van der Heijden (2004), who assume an underlying non-compliance scale for a set of RR variables and estimate SP using an item-response model.

In this paper we present a log-linear modeling approach to account for SP in an RR design. This SP model is derived from the log-linear randomized-response (LLRR) model (Chen 1989) by the introduction of an SP parameter. The three main results of the SP model are: (1) an estimate of the probability of SP; (2) log-linear parameter estimates describing the associations between RR variables and; (3) prevalence estimates of the sensitive behavior corrected for SP. The model is illustrated with two examples from the 2000 Social Welfare Survey conducted in the Netherlands (Van Gils, Van der Heijden, Rosebeek, 2001; see also Lensvelt-Mulders, Van der Heijden, Laudy, and Van Gils, 2006).

In the remainder of this paper we present the questions and the RR design used in the Social Welfare Survey. We introduce the general RR model and shows that identification problems arise when a SP parameter is included. We present the SP model as an extension of the LLRR model. We present two examples from the Social Welfare Survey and investigates the robustness of the parameter estimates against violations of model assumptions. We close with our conclusions.

2.2 The Social Welfare Survey

Employees in the Netherlands are insured under various social welfare acts against the loss of income due to redundancy, disability or sickness. Social benefit recipients have to comply with the rules and regulations of these acts. Non-compliance with the rules is considered fraud and can have serious repercussions. In 2000, 2002 and 2004, the Dutch Department of Social Affaires has conducted a nationwide survey to monitor the degree of non-compliance with respect to these rules.

We present two examples from the 2000 Social Welfare Survey. The sample consists of $1,308$ persons who receive benefits within the framework of the Disability Benefit Act (DBA). The DBA offers financial benefits to employees who, due to sickness or an accident, have been unable to work for a period longer than one year. The amount depends on the degree of disablement, with a maximum of 70% of the last earned wage. To be eligible for benefits, beneficiaries are required to report all additional income from work and improvements in their health status. A detailed description of the sampling procedures used in the Social Welfare Survey can be found in Lensvelt-Mulders et al. (2006).

The examples consist of one set of three work-related questions and one set of four health-related questions. The work-related questions are:

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

2 Have you ever in the past 12 months had a job or worked for an employment agency in addition to your disability benefit without informing the Department of Social Services?

3 Have you worked off the books in the past 12 months in addition to your disability benefit?

Let the variables A^*, B^* and C^* denote the answers to these questions, for $a^*, b^*, c^* \in \{1 \equiv yes, 2 \equiv no\}$. The observed-response profiles frequencies $111, 112, \ldots, 222$ are given by $\boldsymbol{n}^* = (66, 67, 68, 169, 52, 95, 123, 668)^t$.

The health-related question are:

4 Has a doctor or specialist ever told you that the symptoms your disability classification is based upon have decreased without you informing the Department of Social Services of this change?

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

6 Have you yourself ever noticed an improvement in the symptoms causing your disability, for example in your present job, in volunteer work or the chores you do at home, without informing the Department of Social Services of this change?

7 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 of this change?

Let the variables D^*, E^*, F^*, G^* analogously denote the answers to the questions 4 to 7. The observed-response profile frequencies $1111, 1112, \ldots, 2222$ are given by $\boldsymbol{n}^* = (43, 22, 10, 34, 20, 31, 40, 93, 30, 29, 40, 91, 60, 86, 146, 533)^t$.

The questions were all answered according to the Kuk design (Kuk, 1990; Van der Heijden et al, 2000). In this RR design the respondent is given two decks with red and black playing cards. One deck contains 80% red cards and 20% black cards and is called the *yes* deck. The other deck contains 80% black cards and 20% red cards and is called the *no* deck. For each sensitive question, the respondent draws one card from both decks and answers the question by naming the color of the card from the deck corresponding to the true answer. So if the true answer is *yes*, the respondent names the color of the card from the *yes* deck, and if the true answer is *no*, the respondent names the color of the card from the *no* deck.

2.3 The General Randomized-Response Design

Consider a multivariate RR design with K dichotomous sensitive questions. The true responses are denoted by the random variables $A, B, \ldots$ and the random variable X denotes the $D = 2^K$ true-responses profiles $A = a, B = b, \ldots$. Analogously, define the variables $A^*, B^*, \ldots$ for the observed responses and X^* for the observed-response profiles. Let $\boldsymbol{P}_K$ be a $D \times D$ dimensional

transition matrix, with elements (i,j) given by the conditional misclassification probabilities $p_{ij} = I\!P(X^* = i|X = j)$, for $i, j \in \{1, \ldots, D\}$. For the univariate Kuk design, the transition matrix is given by

$$\boldsymbol{P}_K = \boldsymbol{P}_1 = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 8/10 & 2/10 \\ 2/10 & 8/10 \end{pmatrix}. \tag{2.1}$$

In a multivariate design, the transition matrix $\boldsymbol{P}_K$ is found by taking the Kronecker product of the univariate transition matrices. For $K = 3$, the multivariate transition matrix $\boldsymbol{P}_3$ is found by taking the Kronecker product $\boldsymbol{P}_1 \otimes \boldsymbol{P}_1 \otimes \boldsymbol{P}_1$, where

$$\boldsymbol{P}_1 \otimes \boldsymbol{P}_1 = \begin{pmatrix} p_{11}\boldsymbol{P}_1 & p_{12}\boldsymbol{P}_1 \\ p_{21}\boldsymbol{P}_1 & p_{22}\boldsymbol{P}_1 \end{pmatrix},$$

is a 4×4 transition matrix.

The general RR model is given by

$$\boldsymbol{\pi}^* = \boldsymbol{P}_K \boldsymbol{\pi}, \tag{2.2}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_D)^t$ is a vector denoting the true-response profile probabilities and $\boldsymbol{\pi}^* = (\pi_1^*, \ldots, \pi_D^*)^t$ is a vector denoting the observed-response profile probabilities. Model (2.2) is estimated by maximization of the kernel of the loglikelihood

$$\begin{aligned} \ln \ell(\boldsymbol{\pi}|\boldsymbol{n}^*, \boldsymbol{P}_K) &= \sum_{i=1}^{D} n_i^* \ln \pi_i^* \\ &= \sum_{i=1}^{D} n_i^* \ln \left(\sum_{j=1}^{D} p_{ij}\pi_j \right), \end{aligned} \tag{2.3}$$

for $\pi_1, \ldots, \pi_D \in (0, 1)$.

2.4 Boundary Solutions, SP and Identification

The general RR model sometimes exhibits a lack of fit. In case the general RR model lacks fit a boundary solution is obtained, which is characterized by probability estimates on the boundary of the parameter space (Van der

Hout and Van der Heijden, 2002). A lack of fit is a somewhat unexpected result because the general RR model is a saturated model in the sense that the number of independent parameters equals the number of independent observed-response frequencies. There are two potential reasons for boundary solutions to occur.

Boundary solutions occur if a relative observed-response frequency is below (or above) chance level, with chance level defined as the probability of observing a *yes* response given a true *yes* response probability of zero (or equivalently as the probability of observing a *no* response given a true *no* response probability of one). We illustrate this for one variable by writing out the probability of π_1^* in (2.1) and (2.2), with subscript $1 \equiv yes$. Since in the univariate case $\pi_2 = 1 - \pi_1$, it follows that

$$\begin{aligned} \pi_1^* &= p_{11}\pi_1 + p_{12}\pi_2 \\ &= 0.2 + 0.6\ \pi_1, \end{aligned}$$

Solving this equation for π_1 yields the moment estimator

$$\widehat{\pi}_1 = \frac{\widehat{\pi}_1^* - 0.2}{0.6}, \tag{2.4}$$

with $\widehat{\pi}_1^*$ estimated by the relative observed-response frequency n_1^*/n. If π_1^* is smaller than the chance level of 0.2, a negative moment estimate of π_1 is obtained. It follows that in this case π_2^* is greater than the change level of 0.8, and the moment estimate $\widehat{\pi}_2 > 1$. Since the probability estimates obtained by maximizing loglikelihood (2.3) are constrained to be the interval $(0, 1)$, the model will not exhibit a perfect fit.

One potential reason for boundary solutions is RR sampling variation. By this we mean the sampling fluctuation in the frequency of red cards, given the true-response frequencies. If the number of red cards drawn in the sample is less than expected on the basis of the randomization probabilities, the percentage of observed *yes* responses might fall below chance level, especially when the frequency of the true *yes* responses is near zero. The other potential reason for a boundary solution is SP, which has a similar effect on the frequency of the observed *yes* responses as RR sampling variation. If respondents answer *no* when the answer required by the randomizing device is *yes*, the percentage of the observed *yes* responses may also be below the chance level.

In the univariate setting, the effects of SP and RR sampling variation on the observed-response frequencies are confounded. The effect of sample

proportions of red cards larger than the corresponding conditional misclassification probabilities p_{11} or p_{12} described in (2.1) cancel out the effect of SP, whereas smaller sample proportions reinforce the effect of SP. In a multivariate setting, the situation is more complicated because the effect of RR sampling variation on the sample proportion of red cards is different for each variable.

In this paper, we define SP respondents as persons who answer *no* to every question, regardless of their true status or the outcome of the randomizing device. Given this definition, we account for SP by introducing an SP parameter θ in the general RR model, such that

$$\boldsymbol{\pi}^* \;=\; (1-\theta)\boldsymbol{P}_K\boldsymbol{\pi} + \theta\mathbf{v}, \tag{2.5}$$

where θ denotes the probability of SP, $\mathbf{v}$ is the D-dimensional vector $(0,\ldots,0,1)^t$. Notice that model (2.5) implies that SP can only result in the observed-response profile consisting of only *no* responses, and that all true-response profiles are equally likely to be subject to SP. The model can also be rewritten as

$$\boldsymbol{\pi}^* = \boldsymbol{Q}_K\boldsymbol{\pi} \tag{2.6}$$

where the transition matrix $\boldsymbol{Q}_K$ has elements

$$q_{ij} = \begin{cases} (1-\theta)p_{ij} & \text{for } i \neq D, j \in \{1,\ldots,D\} \\ \\ (1-\theta)p_{ij} + \theta & \text{for } i = D, j \in \{1,\ldots,D\} \end{cases} \tag{2.7}$$

Model (2.6) is not identified. We illustrate this with the work-related questions of the Social Welfare Survey. We estimated the true-response probabilities by fitting models to the respective observed-response (profile) frequencies $\boldsymbol{n}^* = (309, 999)$ of variable C^*, $\boldsymbol{n}^* = (118, 162, 191, 873)$ of the variables B^* and C^*, and $\boldsymbol{n}^* = (66, 67, 68, 169, 52, 95, 123, 668)$ of the variables A^*, B^* and C^*. The models were estimated by maximizing the kernel of the loglikelihood

$$\ln \ell(\boldsymbol{\pi}|\boldsymbol{n}^*, \boldsymbol{P}_K, \theta) = \sum_{i=1}^{D} n_i^* \ln\left(\sum_{j=1}^{D} q_{ij}\pi_j\right). \tag{2.8}$$

for fixed values of θ in the interval $(0, 1)$. Figure 2.1 shows the likelihood-ratio statistic L^2 of the models as a function of the value of θ.

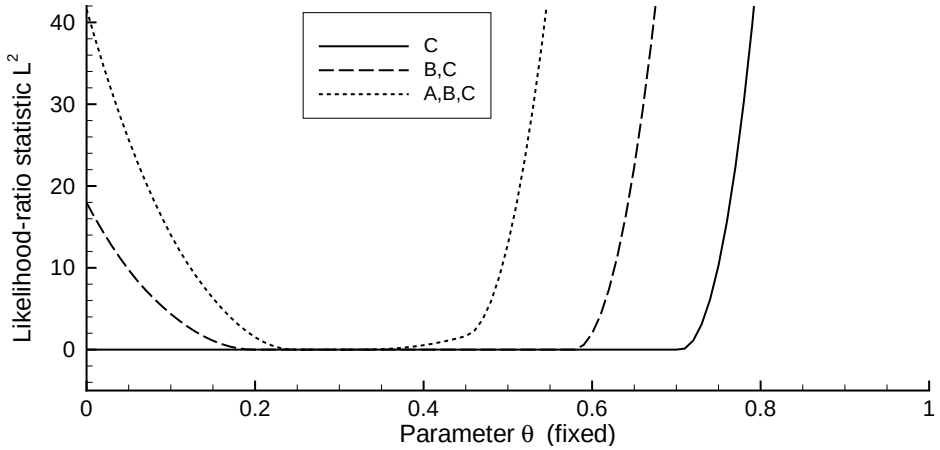

Figure 2.1: The likelihood-ratio statistic when fitting the general RR model to uni- and multivariate data with fixed values of θ.

In case of variable C (solid line), there is a serious identification problem, since the model exhibits a perfect fit for all $\theta \in (0, 0.72)$. The interval of θ for which the model fits perfectly is reduced to $(0.2, 0.6)$ when the variable B is added to the model (dashed line). If the model is estimated for all three variables A, B and C simultaneously (dotted line), the interval of θ for which a perfect fit is obtained is further reduced to $(0.25, 0.32)$. In the next section we show the identification problem can be overcome by using a log-linear model.

2.5 The LLRR and SP Models

The log-linear randomized-response (LLRR) model is presented by Chen (1989) in the context of misclassification of categorical data and is further developed by Van den Hout and Van der Heijden (2004). In this section we briefly review the theory of this model and then introduce the SP model.

Consider the true-response variables A, B and C, with the true-response profiles abc, for $a, b, c \in \{1, 2\}$. For $j \in \{1, \ldots, D\}$, let π_j denote the probabilities of the respective true-response profiles $111, 112, \ldots, 222$. Then the

saturated LLRR model $[ABC]$ is given by

$$\pi_j = \exp\left(\lambda_0 + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{bc}^{BC} + \lambda_{abc}^{ABC}\right), \tag{2.9}$$

where the λ terms are constrained to sum to zero over any subscript. The kernel of the loglikelihood of the LLRR model

$$\ln \ell(\boldsymbol{\lambda}|\boldsymbol{n}^*, \boldsymbol{P}_K) = \sum_{i=1}^{D} n_i^* \ln\left(\sum_{j=1}^{D} p_{ij}\pi_j\right), \tag{2.10}$$

is identical to the kernel of the loglikelihood (2.3) of the general RR model, except that loglikelihood (2.10) is maximized as a function of the log-linear parameters. Constrained LLRR models are formulated by setting log-linear parameters in (2.9) to zero or by imposing equality constraints. For a more detailed discussion of the LLRR model we refer to Chen (1989) and Van den Hout and Van der Heijden (2004).

The LLRR model can be adapted to accommodate SP by replacing the elements p_{ij} of transition matrix $\boldsymbol{P}_K$ in the loglikelihood function (2.10) by the elements q_{ij} of transition matrix $\boldsymbol{Q}_K$ defined in (2.7). Since the matrix $\boldsymbol{Q}$ contains the SP parameter θ, this results in an overparametrized model. We solve this problem by constraining the highest-order interaction parameter of the log-linear model to zero. In a design with K variables, constraining the K-factor interaction parameter preserves the hierarchical structure of the model. The saturated SP model is the model $\theta, [AB, AC, BC]$, that is given by

$$\pi_j = \exp\left(\lambda_0 + \lambda_a^A + \lambda_b^B + \lambda_c^C + \lambda_{ab}^{AB} + \lambda_{ac}^{AC} + \lambda_{bc}^{BC}\right), \tag{2.11}$$

and where the term saturated is used in the sense that the number of free parameters in the model equals the number of independent observed-response frequencies. As with the LLRR model, constrained SP models are formulated by imposing restrictions on the log-linear parameters. The kernel of the loglikelihood of the SP model is given by

$$\ln \ell(\boldsymbol{\lambda}, \theta|\boldsymbol{n}^*, \boldsymbol{P}_K) = \sum_{i=1}^{D} n_i^* \ln\left(\sum_{j=1}^{D} q_{ij}\pi_j\right). \tag{2.12}$$

The SP model is estimated by maximizing loglikelihood (2.12) as a function of the SP parameter θ in (2.7) and the log-linear parameters in (2.11). The estimation can be performed with standard optimization routines. A code written for the statistical programme Gauss can be found on the website www.randomizedresponse.nl.

2.6 Examples

Table 2.1 presents the model selection results for the Work and the Health example. The table reports the likelihood-ratio statistics L^2 obtained from fitting various LLRR models and SP models by maximization the respective loglikelihoods (2.10) and (2.12). The table also presents the estimates of θ for the SP models.

Table 2.1: Model Selection and Cheating Parameter Estimates

Data	Model	$\widehat{\theta}$	L^2	df
Work	W0: $[ABC]$	-	41.6	0
	W1: $[AB, AC, BC]$	-	42.6	1
	W2: $\theta, [AB, AC, BC]$	.25 (.04)	.0	0
	W3: $\theta, [AB, AC, BC]$ [a]	.26 (.04)	5.2	2
	W4: $\theta, [AB, BC]$ [b]	.25 (.04)	.2	2
	W5: $\theta, [A, B, C]$	.36 (.02)	18.4	3
Health	H0: $[DEFG]$	-	37.3	0
	H1: $[DEF, DEG, DFG, EFG]$	-	38.9	1
	H2: $\theta, [DEF, DEG, DFG, EFG]$	.15 (.03)	7.1	0
	H3: $\theta, [DE, DF, DG, EF, EG, FG]$	.13 (.05)	7.1	4
	H4: $\theta, [DE, DF, DG, EF, EG, FG]$ [a]	.13 (.05)	36.2	9
	H5: $\theta, [DE, EF, FG]$ [c]	.15 (.03)	8.4	8
	H6: $\theta, [D, E, F, G]$	.27 (.02)	82.9	10

a. equality constraints on all interaction parameters

b. equality constraints $\lambda_{ab}^{AB} = \lambda_{bc}^{BC}$

c. equality constraints $\lambda_{de}^{DE} = \lambda_{ef}^{EF}$

The models W0, H0, W1 and H1 are LLRR models. The saturated LLRR models W0 and H0 both fit poorly. In the models W1 and H1, the highest-order interaction parameters λ_{abc}^{ABC} and λ_{defg}^{DEFG} are constrained to zero. The slight deterioration in fit suggests that no substantial K-factor interaction is present in the data when SP is not taken into account.

The results are reported of four SP models (W2 to W5) of the Work example. Model W2 fits perfectly, with an estimated SP probability of 0.25. The most parsimonious model is W4, with the parameters λ_{ab}^{AB} and λ_{bc}^{BC} constrained to be equal. The deterioration of fit for the models W3 with equality constraints on all interaction parameters and W5 with only main

effects illustrate that no further restrictions on the parameters are feasible.

For the Health example, the results are shown of five SP models (H2 to H6). Elimination of all 3-factor interaction parameters in model H3 does not affect the fit. Model H5, with equality of the interaction parameters λ_{de}^{DE} and λ_{ef}^{EF}, is the most parsimonious model. In this model the estimated probability of SP is 0.15. Model H4 and model H6 illustrate that the fit deteriorates if more constraints are imposed.

Table 2.2: Estimated 2-way Interactions

Model	Interaction	Odds Ratio	$\widehat{\lambda}$
W4	$AB = BC$	26.5	.82 (.22)
H5	$DE = EF$	181.3	1.30 (.30)
	FG	29.2	.84 (.23)

Table 2.2 reports the estimated odds ratios and interaction parameters. The results for the Work example suggest that the status on variable A (small jobs for friends) is positively associated to the status on variable B (job or employment agency). The odds of having the same status on both variables are estimated to be 26 to 1. As indicated by the equality constraint $AB = BC$, the positive association between the status on the variables B and C (working off the books) is roughly equally strong. Furthermore, giving the status on variable B, there is no evidence for a significant association between the variables A and C. Similar association patterns are found in the Health example. The estimated odds ratios of 181 imply a high probability of the same status on the variables E (pretending to be sick at the check-up) and D (withholding doctor's information about symptom improvements) and on the variables E and variable F (not reporting symptom improvements noticed by respondent himself). The results also show a positive, although somewhat less strong association between the status on variable F and G (not reporting feeling stronger and more able to work).

The estimated true-response profile probabilities $\pi_1, \ldots, \pi_D$ are shown in Table 2.3. The large odds ratio estimates in Table 2.2 turn out to be caused by probability estimates that are close to their boundary values. In the Work example, the response profile nyn has an estimated probability smaller than 0.01. In the Health example more than half of the response profiles have an

Table 2.3: True-Response Probability Estimates

W4		$B=y$		$B=n$	
		$C=y$	$C=n$	$C=y$	$C=n$
	$A=y$	.092	.027	.017	.160
	$A=n$	.019	.006	.065	.615
H5		$F=y$		$F=n$	
		$G=y$	$G=n$	$G=y$	$G=n$
$D=y$	$E=y$	.068	.014	.001	.005
	$E=n$	.001	.000	.002	.013
$D=n$	$E=y$	.017	.004	.000	.001
	$E=n$	.038	.008	.120	.708

estimated probability smaller than 0.01.

Table 2.4 reports the univariate non-compliance estimates with corresponding confidence intervals, obtained with the parametric bootstrap method. In comparing the results of the LLRR and SP models, the correction for SP has a substantial effect on the estimated non-compliance probabilities.

Table 2.4: Estimated Non-Compliance Probabilities and 95% Confidence Intervals

Model	A	B	C	
W0	.14 (.10,.18)	.09 (.06,.13)	.09 (.07,.13)	
W4	.30 (.23,.38)	.14 (.10,.20)	.19 (.13,.27)	
Health	D	E	F	G
H0	.07 (.05,.11)	.08 (.06,.12)	.11 (.09,.15)	.16 (.12,.20)
H5	.10 (.07,.17)	.11 (.08,.17)	.15 (.11,.21)	.25 (.20,.32)

2.7 Robustness Against Model Violations

In this section we evaluate the robustness of the SP model against model violations. First, we examine the robustness of the SP parameter and the

univariate prevalence estimates against violations of the assumption that the K-factor interaction is zero. Second, we investigate the extent to which the SP parameter captures the effects of RR sampling variation. Lastly, we generate the sampling distribution of the likelihood-ratio statistic for the models W4 and H5 and infer the critical value.

We evaluate the robustness of the SP parameter and univariate prevalence estimates of the models W4 and H5 against non-zero K-factor interaction λ_k^K by fitting the SP models $\theta, [AB, AC, BC]$ and $\theta, [DEF, DEG, DFG, EFG]$ to three manipulated data sets $\boldsymbol{n}^*_{(\lambda_k^K)}$, for $K \in \{3, 4\}$. The data sets are computed for different log-linear parameter vectors $\boldsymbol{\lambda}$, that consist of the log-linear parameter estimates $\widehat{\boldsymbol{\lambda}}$ of the models W4 and H5, extended with the K-factor interaction parameter λ_k^K, for $\lambda_k^K \in \{-1, 0, 1\}$. The data sets are computed using the equations $\ln(\boldsymbol{\pi}_{(\lambda_k^K)}) = \boldsymbol{M}\boldsymbol{\lambda}$ and $\boldsymbol{n}^*_{(\lambda_k^K)} = n\boldsymbol{Q}_K\boldsymbol{\pi}_{(\lambda_k^K)}$, with $n = 1,308$. In the latter equation, the transition matrix $\boldsymbol{Q}_K$ is based on the estimated values $\widehat{\theta} = .249$ for model W4 and $\widehat{\theta} = .146$ for model H5. Since the expectation of $\boldsymbol{Q}_K$ is used to compute the observed-response frequencies, the data are not affected by RR sampling variation.

Table 2.5: Bias in estimated parameters of the saturated SP model as a function of ignored K-factor interaction

		$\lambda_k^K = 0$	$\lambda_k^K = -1$		$\lambda_k^K = 1$	
Model	Parameter	True=Est.	True	Est.	True	Est.
W4	θ	.249	.249	.300	.249	.235
	π_1 (A)	.295	.113	.145	.614	.597
	π_1 (B)	.142	.081	.110	.250	.239
	π_1 (C)	.192	.073	.102	.400	.386
H5	θ	.146	.146	.142	.146	.149
	π_1 (D)	.104	.135	.133	.094	.095
	π_1 (E)	.110	.151	.149	.096	.097
	π_1 (F)	.150	.189	.187	.137	.138
	π_1 (G)	.249	.539	.535	.151	.153

The results are shown in Table 2.5. The "True" columns refer to the parameter values used to construct the data, and the columns labeled "Est." refer to the estimates of the saturated SP models. The upper panel of Table 2.5 shows that in the event of negative K-factor interaction, the SP parameter

and univariate non-compliance probabilities are overestimated. The effects are reversed if the K factor interaction is positive. In the lower panel, the effects of the K-factor interaction are opposite to those in the upper panel in both conditions and for all parameters. In comparing the true values and the estimates, the results show that, given the absence of RR sampling variation, the SP model is unbiased when the K-factor interaction is zero, and that otherwise the bias in the SP parameter and univariate probability estimates is relatively small.

We perform a parametric bootstrap to examine the bias in the SP parameter estimate resulting from RR sampling variation. We draw two sets of $1,000$ random samples from the fitted vectors $\widehat{\boldsymbol{n}}^*$ of the models W4 and H5, and fit the SP models $\theta, [AB, AC, BC]$ and $\theta, [DEF, DEG, DFG, EFG]$ to the respective bootstrap samples. We subtract the fitted values $\widehat{\theta} = .249$ of model W4 and $\widehat{\theta} = .146$ of model H5 from the respective SP parameter averages in the bootstrap. Table 2.6 shows that the SP parameters are overestimated by .003 for model W4 and by .008 for model H5. These results suggest that the SP parameter estimate is not substantially affected by the effects of RR sampling variation.

Table 2.6: Parametric bootstrap of models W4 and H5

Bootstrap Model	Fitted Model	Bias in $\widehat{\theta}$	$L^2_{95\%}$
W4	$\theta, [AB, AC, BC]$	.003	1.4
H5	$\theta, [DEF, DEG, DFG, EFG]$	.008	11.1

Lastly, the parametric bootstraps are used to generate the distribution of the likelihood-ratio statistic for the models W4 and H5. We find an average value of 0.3 for the samples based on model W4 and of 4.4 for the samples based on model H5. The fact that these averages do not equal zero shows that the SP parameter cannot entirely account for the lack of fit resulting from RR sampling variation. It also shows that even though the SP model is correctly specified, it may not always fit perfectly. To find the rejection area of the saturated SP models we determined the 95^{th} percentile value $L^2_{95\%}$ of the likelihood-ratio statistic in the parametric bootstrap. These are shown for model W4 and H5 in the last column of Table 2.6. The likelihood-ratio

statistic of 7.1 of the saturated SP model ($H2$) in Table 2.1 does not exceed the critical value of 11.1 obtained in the bootstrap. The result suggests that lack of fit is attributable to RR sampling variation, and that therefore the model need not be rejected.

2.8 Conclusions

The SP model is a useful tool to analyze RR data that are potentially affected by self-protective response bias. The two applications presented in this paper show that the SP model fits significantly better than models that do not take SP into account. The SP model is unbiased if the assumption of zero K-factor interaction is fulfilled and RR sampling variation is absent. Given that RR sampling variation is present, the SP parameter and univariate prevalence estimates are slightly positively biased. If K-factor interaction is present in the data, the bias in the SP parameter and univariate prevalence estimates are relatively small. Furthermore, in real data the highest-order interaction parameter is usually not significant, unless the sample size is large relative to the number of variables. The costs of a priori setting this parameter to zero thus seem to be low.

In this paper we restrict ourselves to the assumption that SP always results in the observed-response profile with only *no* responses, regardless of the outcome of the randomizing device or the true status of the respondent. This assumption implies that SP is independent of the true-response profile. Therefore the prevalence estimates of the model are unbiased if SP and non-compliance are independent. However, if SP correlates positively with non-compliance, the prevalence of non-compliance is underestimated. Similarly, the SP model will overestimates the prevalence if SP correlates negatively with non-compliance. Different assumptions about SP are possible, for example that the probability of SP depends on the true-response profile or on person characteristics. However, the new identifiability problems that arise when SP is assumed to depend the true-response profile are beyond the scope of this paper. An interesting question is to what extent SP depends on person characteristics. For example, if SP is due to a lack of trust of the RR design, improved instructions might reduce the probability of SP. The development of regression models in which the SP parameter is defined as a function of covariates is an interesting topic for future research.

If the number of variables in the RR design is large and the variables are

strongly associated, the response profile data can rapidly become sparse. In this case it would be interesting to compare the SP model to an approach proposed by Gilula and Haberman (1991), that combines log-linear modeling and a summarization of the true-response profile data, that is obtained after correcting the observed-response profile for RR. The methodology of Gilula and Haberman seems especially suited when the number of variables is large and SP is absent. However, it is less obvious how their methodology can be applied when SP responses are present and the probability of observing an SP response has to be estimated from the data.

The SP model is estimated by maximizing the loglikelihood function. It would also be interesting to model SP within a Bayesian framework. An advantage of the Bayesian approach is the possibility of using an informative prior for the SP parameter. In this way, knowledge of the prevalence of SP from other RR research can be taken into account. Within the Bayesian framework it is also be possible to use fully specified distributions of the SP parameter in a sensitivity analysis. If the distribution of the SP parameter is specified, there is no identification problem. By choosing different distributions, one can study the effect of these distributions on the estimated log-linear parameters and the univariate prevalence estimates of the sensitive characteristics.

Chapter 3

The Proportional Odds Model

3.1 Introduction

In surveys and questionnaires, questions are sometimes regarded as sensitive or embarrassing. Especially if personal characteristics like the respondent's drug use, alcohol consumption or sexual behavior 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 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, for example "I have used drugs" and "I have never used drugs", and the outcome of a randomizing device determines which of the two questions the respondent has to answer. So, a respondent who has never used drugs answers $false$ if the former question has to be answered, and *true* if the latter question has to 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

[1]Published as Cruyff, M.J.L.F., van den Hout, A., and van der Heijden, P.G.M. (2008). The analysis of randomized response sum score variables, *Journal of the Royal Statistical Society, Series B*, **70**, 21-30.

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) present logistic regression models that can be used to analyze 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 noncompliance 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) introduce IRT models for randomized-response 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.* (2006) present a multivariate logistic regression model describing the associations between multiple binary RR variables and a set of covariates.

An alternative approach to analyze 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 the respondent has used, the number of different criminal activities the respondent has engaged in, or the number of potentially traumatic events 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 analyzing 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 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 noncompliance with the Social Security legislation.

Section 3.2 describes the Social Security Survey data and the forced-response design used in this survey. The first part of Section 3.3 presents the RR sum score model and the second part the RR proportional odds model. The example is presented in Section 3.4. Section 3.5 discusses boundary solutions and presents an example. Section 3.6 gives the conclusions.

3.2 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 a number of rules and regulations. In 2002 the Dutch Department of Social Affairs conducted a nationwide survey to evaluate the level of noncompliance 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 $1,760$ recipients were asked two questions about their health status (Q1 and Q2) and two questions about receiving income from work in addition to the disability benefit (Q3 and Q4):

Q1 At a Social Services check-up, have you ever acted as if you were sicker or less able to work than you actually were?

Q2 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?

Q3 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.)

Q4 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 (FR) 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 has to answer truthfully. The outcome of the dice is only known 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 $I\!P(\text{forced } yes) = 1/6$, and $I\!P(\text{forced } no) = 1/12$. (The programmer inadvertently programmed the virtual dice so that $I\!P(\text{forced } yes) = 0.1868$ and $I\!P(\text{forced } no) = 0.0671$). Given that the respondent's true answer is *no*, the probability of misclassification $I\!P(\text{observed } yes|\text{true } no) = I\!P(\text{forced } yes)$, and similarly, given a true *yes* response the probability of misclassification $I\!P(\text{observed } no|\text{true } yes) = I\!P(\text{forced } no)$. Since irrespective of the true response, the probability of misclassification is non-zero, confidentiality is assured.

Let the variables Y_1^* to Y_4^* denote the answers to the questions 1 to 4, with $y_1^*, \ldots, y_4^* \in \{0 \equiv no, 1 \equiv yes\}$. 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 $\boldsymbol{n}^* = (694, 117, 188, 81, 179, 43, 65, 41, 117, 41, 37, 26, 62, 14, 27, 28)$. 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, 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.

3.3 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.

3.3.1 The RR sum score model

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

$$Z = \sum_{m=1}^{M} Y_m. \tag{3.1}$$

Analogously, let the sum score variable $Z^* = \sum_{m=1}^{M} Y_m^*$ 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

$$\pi_s^* = \sum_{t=0}^{M} q_{s|t}\pi_t, \tag{3.2}$$

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

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

$$q_{s|t} = \sum_{j=0,\ 0\le s+j-t\le M-t}^{t} \binom{t}{j}\binom{M-t}{s+j-t} p_{1|1}^{t-j} p_{0|1}^{j} p_{1|0}^{s+j-t} p_{0|0}^{M-s-j}. \tag{3.3}$$

The index j in (3.3) 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 $Y_m = 0$. Lemma (3.3) 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 ones and zeros 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.)

Estimation The RR sum score model is most easily estimated with the

method of moments (MM). The MM estimator is most conveniently presented using matrix notation,

$$\widehat{\boldsymbol{\pi}} = \boldsymbol{Q}^{-1}\widehat{\boldsymbol{\pi}}^{*}, \tag{3.4}$$

where $\boldsymbol{\pi} = (\pi_0, \ldots, \pi_M)', \boldsymbol{\pi}^* = (\pi_0^*, \ldots, \pi_M^*)'$ and π_s^* estimated by n_s^*/n, with n_s^* denoting the frequency of the observed sum score s on variable Z^*. The matrix $\boldsymbol{Q}$ 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 outside the boundaries of 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

$$\ln \ell(\boldsymbol{\pi}|n_0^*, \ldots, n_M^*), = \sum_{s=0}^{M} n_s^* \ln \left(\sum_{t=0}^{M} q_{s|t}\pi_t \right), \tag{3.5}$$

for $\pi_t \in (0, 1)$. Kuha and Skinner (1997) provide EM algorithms. van den Hout and van der Heijden (2002) show 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.

3.3.2 The RR 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 $I\!P(Z = t|\mathbf{x})$ denote the probability that the sum score variable Z takes on the value t given the covariate vector $\mathbf{x}$. Define $\gamma_t = I\!P(Z \leq t|\mathbf{x})$. Then the proportional odds model (McCullagh, 1980) states that

$$\gamma_t = \frac{\exp(\alpha_t - \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\alpha_t - \mathbf{x}'\boldsymbol{\beta})}, \tag{3.6}$$

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

$$-\infty < \alpha_0 \leq \alpha_1 \leq \ldots \leq \alpha_M \equiv \infty. \tag{3.7}$$

Note that for $M = 1$, the order of the threshold parameters is $-\infty < \alpha_0 \leq \alpha_1 \equiv \infty$, and expression (3.6) reduces to the binary logistic regression model (with a negative sign for $\boldsymbol{\beta}$).

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

$$\ln\left[\frac{I\!P(Z \leq t|\mathbf{x}_0)/I\!P(Z > t)|\mathbf{x}_0)}{I\!P(Z \leq t|\mathbf{x}_1)/I\!P(Z > t)|\mathbf{x}_1)}\right] = (\mathbf{x}_1 - \mathbf{x}_0)'\boldsymbol{\beta} \tag{3.8}$$

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 $I\!P(Z \leq t|\mathbf{x})$ are modeled through the observed variable Z^*, with the relation between Z^* and Z given by the RR sum score model. The RR proportional odds model is given by

$$\gamma_s^* = \sum_{j=0}^{s}\sum_{t=0}^{M} q_{j|t}(\gamma_t - \gamma_{t-1}), \tag{3.9}$$

where $\gamma_s^* = I\!P(Z^* \leq s|\mathbf{x})$.

Estimation The maximum likelihood estimator (MLE) of model (3.9) is obtained by maximization of the kernel of the observed data log likelihood, given by

$$\ln \ell(\boldsymbol{\beta}, \boldsymbol{\alpha}|z_i^*, \ldots, z_n^*, \mathbf{x}_i, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \ln\left(\sum_{t=0}^{M} q_{z_i^*|t}(\gamma_t - \gamma_{t-1})\right), \tag{3.10}$$

where $\gamma_{-1} = 0$ and $\gamma_M = 1$. To identify the model, we use the convention $\alpha_0 = 0$. For the maximization of (3.10) 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 using the Broyden-Flechter-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 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 derived from 500 nonparametric bootstrap samples using the percentile method.

Table 3.1: Parameter estimates of the RR proportional odds model.

Parameters	*Estimates (se)*	*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

3.4 The Example

In this section, we analyze the sum score variable $Z = \sum_{m=1}^{3} Y_m$, denoting the number of *yes* responses to the questions Q_1, Q_2 and Q_3 of the Social Security Survey, with the RR sum score model and the RR proportional odds model. The frequencies of the sum scores $0, 1, 2, 3$ observed in the sample are given by the vector $\boldsymbol{n}^* = (811, 649, 245, 55)$.

The respective MM sum score probability estimates of the RR sum score model are $\widehat{\boldsymbol{\pi}} = (0.850, 0.075, 0.058, 0.017)$. Since the MM estimates are all in the interior of the parameter space, the ML solution is identical. The log likelihood of 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 $\widehat{\beta}_0 = -1.74$, $\widehat{\alpha}_1 = 0.77$ and $\widehat{\alpha}_2 = 2.32$, and the sum score probabilities are found by plugging these estimates into $\widehat{\gamma}_t$ defined in (3.6), and using expression $\widehat{\pi}_t = \widehat{\gamma}_t - \widehat{\gamma}_{t-1}$.

Table 3.1 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 noncompliance 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 noncompliance are about 2.3 times those of women. Similarly, the odds of noncompliance for people above the age of 45 and for people who had a regular job contract are about 1.8 times that of younger people and people who had a different kind of job contract, respectively. Finally, the odds of noncompliance 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

$$\text{logit}(\gamma_t) = \alpha_t - \mathbf{x}'\boldsymbol{\beta} - \mathbf{w}'\boldsymbol{\eta}_t. \tag{3.11}$$

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 (LR) statistic of 8.2 with 12 degrees of freedom ($p = 0.77$) indicates that the proportional odds assumption need not be rejected. Notice that the LR statistic at the same time implies that the proportional odds assumption holds for the four significant covariates in Table 3.1. By setting the contribution to the LR statistic of the two nonsignificant covariates to zero, we obtain $LR = 8.2, df = 8, p = 0.41$.

3.5 Boundary solutions

Fitting the RR proportional odds null-model to the observed frequency vector $\boldsymbol{n}^* = (694, 601, 329, 108, 28)$ of $Z^* = \sum_{m=1}^{4} Y_m^*$ denoting the number of *yes* responses to the four questions Q_1 to Q_4, yields the solution $\widehat{\beta}_0 = -1.31$,

$\widehat{\alpha}_1 = -0.46$, $\widehat{\alpha}_2 = 1.98$ and $\widehat{\alpha}_3 = 2.22$. Note that this solution does not satisfy condition (3.7), since

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

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

To force the threshold parameter estimates to satisfy condition (3.7) we use the parametrization

$$\alpha_t = \alpha_0 + \sum_{j=1}^{t} \exp(\dot{\alpha}_j), \tag{3.12}$$

and maximize log likelihood (3.10) for $\dot{\alpha}_j$ and $\boldsymbol{\beta}$, with α_0 constrained to zero. This parametrization yields the solution $\widehat{\dot{\alpha}}_1 = -10.92$, $\widehat{\dot{\alpha}}_2 = 0.46$, and $\widehat{\dot{\alpha}}_3 = -0.02$ (corresponding to $\widehat{\alpha}_1 = 0.00$, $\widehat{\alpha}_2 = 1.58$, and $\widehat{\alpha}_3 = 2.56$), and $\widehat{\beta}_0 = -1.88$. The vector $\widehat{\boldsymbol{\pi}} = (0.867, 0.000, 0.102, 0.019, 0.012)'$ implied by this solution is valid and coincides with the ML estimates of the RR sum score model.

Table 3.2 presents the parameter estimates of the full RR proportional odds model using parametrization (3.12). 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 (3.12) 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 to 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 to Y_4 and the sum score variable Z. Table 3.3 presents the regression parameter estimates of the RR logistic model specified as in expression (3.6), i.e. with a negative sign for the vector $\boldsymbol{\beta}$. The probability estimates $\widehat{\pi}_1$ 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

Table 3.2: Parameter estimates and 95% bootstrap confidence intervals (CI_{boot}) of the full RR proportional odds model with parametrization $\dot{\alpha}$

Parameters	*Estimates*	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)

due to the fact that $\widehat{\pi}_1$ is close to zero, 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 latter 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 relation between regulatory noncompliance and the covariate last job contract. This shows that both models may provide different insights in the relation 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.

3.6 Conclusions

This paper discusses the construction and analysis of RR sum score variables 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

Table 3.3: Parameter estimates (standard errors) of the RR logistic regression model for variables Y_1 to Y_4.

Parameters	Y_1	Y_2	Y_3	Y_4
$\widehat{\pi}_1$	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)

sensitive characteristics, and the RR proportional odds model that can be used to analyze 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 boundary solutions 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 distribution of the sensitive characteristics under study. For example, the distribution and determinants of the sum score probabilities of regulatory noncompliance may contain valuable information for law enforcers and policymakers. Moreover, the analysis of sum score data may reveal associations that remain undetected if the data are analyzed in a univariate way. In the examples, the RR proportional odds model detected an association between regulatory noncompliance and the last job contract, an association that was not found in the RR logistic model. These differences result from the fact that both models address 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 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 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 zero or close to zero, 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 weights given by $w_m, m \in \{1, \ldots, M\}$. In analogy to 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 and Y^*, since these are not affected by the weights.

Chapter 4

The Zero-inflated Poisson Model

4.1 Introduction

In 2004 the Dutch Department of Social Affairs conducted a nationwide survey to assess the level of compliance with the Unemployment Insurance Act. Under this act employees who have lost their income due to unemployment are entitled to financial benefits, provided that they comply with the rules and regulations stipulated in the act. The participants in the survey were asked if they had ever violated against the regulations in the year preceding the survey. Since the disclosure of a rule violation may have serious financial consequences for the respondent, the randomized response design was used.

The randomized response method was first introduced in 1965 by Warner as an interview technique that protects the respondents' privacy Warner (1965). In Warner's design the respondent is presented with two complementary statements, for example "I am a marihuana user" and "I am not a marihuana user". The respondent then operates a randomizing device, like a pair of dice or a deck of cards, and the outcome of this device determines which of the two statements the respondents has to answer. Since only the respondent knows the outcome of the randomizing device, confidentiality is guaranteed.

[1]Published as Cruyff, M.J.L.F. Böckenholt, U., van den Hout, A., and van der Heijden, P.G.M. (2008). Zero-Inflated Poisson Regression Models for Randomized Response Sum Score Data, *Annals of Applied Statistics*, **2**, 316-331.

A meta-analysis of randomized response studies shows that the randomized response design generally yields higher and more valid prevalence estimates of the sensitive characteristic than direct-questioning designs Lensvelt *et al.* (2006). However, a number of studies suggest that respondents do not always follow the instructions of the randomized response design. In an experimental randomized response design (Edgell *et al.*, 1982) with the outcomes of the randomizing device fixed in advance, about 25% of the respondents answers *no* to a question about having had homosexual experiences, while according to the design these respondents should have answered *yes*. In another experimental study (van der Heijden *et al.*, 2000) all respondents were known to have offended against social security regulations . Although the randomized response condition yielded higher estimates than the direct question design, the prevalence estimate of offenders obtained with randomized response was only about 50%. Another study involved an interview of participants in an randomized response survey (Boeije and Lensvelt-Mulders, 2002). Many of the participants indicated that they had found it difficult to falsely incriminate themselves when they were forced to do so by the outcome of the dice. Some of them admitted that in this situation they had given the non-incriminating answer instead.

A recent topic of investigation in the field of randomized response is the estimation of evasive response bias. Clark and Desharnais (1998) show that the presence of evasive responses can be detected in an randomized response design with two groups that each use a randomizing device with different outcome probabilities. Kim and Warde (2005) present a multinomial randomized response model taking evasive response bias into account in designs with a sensitive question with multiple response categories that increase in sensitivity. The term self-protection (SP) was introduced in by Böckenholt and van der Heijden (2007, 2008) to describe the responses by respondents who consistently give the evasive answer, without taking the outcome of the randomizing device into account. According to this definition the SP response profile consists of non-incriminating (i.e. *no*) responses only. The authors use models from item response theory to obtain prevalence estimates of the sensitive characteristics corrected for SP. The SP assumption is also used in log-linear randomized response models that study the association patterns between the sensitive characteristics and obtain prevalence estimates corrected for SP (Cruyff *et al.*, 2007).

The definition of SP implies that the probability of an evasive response does not explicitly depend on the sensitivity of the question or on the true

status of the respondent. Although it is possible to formulate more complex assumptions with respect to the generation of evasive response bias, SP seems to provide an adequate description of the process. A study by Böckenholt, Barlas and van der Heijden (2007) modeling evasive response behavior in randomized response as a function of both the sensitivity of the question and the true status of the respondent found no compelling evidence for the superiority of these models in relation to the corresponding SP models.

In this paper we introduce a regression model that allows for SP in randomized response sum score data. The model assumes a Poisson distribution for the true sum score variable assessing the individual number of sensitive characteristics. The model further assumes that the observed sum score variable denoting the number of incriminating responses is partly generated by the randomized response design, and partly by SP. Since SP by definition results in an observed sum score of zero, the distribution of the observed sum score variable is zero-inflated with respect to the Poisson randomized response distribution of the true sum score variable. The model allows for predictors that explain individual differences in the Poisson parameters as well predictors that explain individual differences in the probability of SP. Since the distribution of the observed sum score variable is a mixture of a Poisson randomized response distribution and observed zero-inflation, the model is called the zero-inflated Poisson randomized response regression model.

The model is applied to randomized response data from a social security survey conducted in the Netherlands in 2004. Section 4.2 describes the data. Section 4.3 derives the zero-inflated Poisson regression model based as an extension of existing randomized response models for multinomial and sum score data. The section also includes a description of a maximum likelihood (ML) estimation procedure and an evaluation of the validity of the Poisson assumption with respect to the true sum score variable. The results for the social security data are presented in Section 4.4. Section 4.5 discusses some assumptions and interpretations of the model.

4.2 The Data

In 2004 the Department of Social Affairs in the Netherlands conducted a nationwide survey to assess the level of noncompliance with the Social Security Law (compare Lensvelt *et al.*, 2006). The survey includes 870 participants who receive financial benefits under the Unemployment Insurance Act (UIA).

Persons who have become (partially) unemployed are eligible for benefits. A beneficiary receives about 70% of the last earned wages, and the duration of the benefits depend on the length of the persons' employment history. Beneficiaries are required to report all activities that generate income in addition to their benefits or that might conflict with reintegration into the labor market. The failure to report such an activity may be sanctioned.

The social security survey includes the following five questions assessing noncompliance with UIA regulations:

1 Have you in the past 12 months ever had a job or worked for an employment agency in addition to your benefit without informing the Department of Social Services?

2 Have you in the past 12 months ever refused to accept a suitable job, or have you ever deliberately made sure you were not hired even though you had a chance of getting the job?

3 Have you in the past 12 months ever deliberately put in an insufficient number of job applications for a sustained period of time?

4 Have you in the past 12 months attended any day courses without informing the Department of Social Services?

5 Have you in the past 12 months had any income in addition to your benefit, for example from alimony, a scholarship, subletting, other benefits, gifts, interest and so forth, without informing the Department of Social Services?

Due to the sensitive nature of the questions the randomized response method is used. The respondents answer the questions with the use of a computer according to the forced response design (Boruch, 1971). Before answering the question the respondent throws two virtual dice, and is instructed to answer *yes* if the sum of the dice is $2, 3$ or 4, and to answer *no* if the sum of the dice is 11 or 12. If the sum of the dice is $5, 6, 7, 8, 9$ or 10, the respondent has to answer the question truthfully. The misclassification probabilities, that are conditional on the true status of the respondent, can be derived from the probability distribution of the sum of two dice. Given regulatory noncompliance, the probability of a *yes* response is $11/12$ and that of *no* response $1/12$. Given regulatory compliance, the probability of a *yes* response is $1/6$ and the probability of a *no* response $5/6$. In the actual social security survey however, the programmer inadvertently programmed the virtual dice so that the probability of a *yes* response given regulatory noncompliance was 0.9329,

and that of a *yes* response given regulatory compliance .18678. The number of observed *yes* responses to the five questions are respectively 122, 195, 168, 207 and 274. Counting the total number of *yes* responses for each respondent on the five questions yields the frequencies $n_0 = 288$, $n_1 = 295$, $n_2 = 207$, $n_3 = 68$, $n_4 = 7$ and $n_5 = 5$ (with the subscript denoting the number of observed *yes* responses).

The social security survey includes two kinds of predictors we like to explore, one concerning demographic variables and the other concerning variables related to the forced response design. The demographic variables *gender*, *age*, *year unemployment*, *education* and *knowledge rules* are used as predictors of regulatory noncompliance. The variables *gender* and *age* are dummy-coded with "male" ($n = 483$) and "older than 26" ($n = 832$) as respective reference categories. The variable *year unemployment* is a dummy variable denoting the last year of being employed, with the year 2004 as reference category ($n = 257$). The variable *education* (mean $= 2.25, \text{sd} = .67$) measures increasing levels of eduction. The variable *knowledge rules* (mean $= 3.8, \text{sd} = .90$) denotes on a 5-point scale of the respondents' general knowledge of the social security regulations. The two variables *trust* and *understanding* are related to the forced response design and are used as predictors of SP. The variable *trust* (mean $= 3.5, \text{sd} = .92$) is constructed as the average score on four 5-point scale variables (Cronbach's Alpha = .87) assessing different aspects of the respondents' beliefs in the confidentiality and privacy protection of the forced response design. A high score on this variable corresponds to a high degree of trust. The variable *understanding* (mean $= 4.2, \text{sd} = .85$) assesses on a 5-point scale to what extent the respondent feels that he understood when to answer *yes* and when to answer *no* to an forced response question. High scores correspond to a good understanding of the forced response design.

Figure 4.1 depicts the associations between the observed sum scores and the predictors. At this point we would like to emphasize that the plots should not be interpreted as depicting associations between the predictors and the true sum scores (i.e. the number of rule violations), since the observed sum scores are not corrected for the misclassification due to randomized response, nor for SP. The plots at the top of the figure show the observed sum score proportion conditional on the categories within the dummy variables *gender*, *age* and *year unemployment*. The profiles of males and females look similar. The plot for *age* shows that the proportion of zeros for the younger respondents (about 15%) is about half that of the older respondents. The

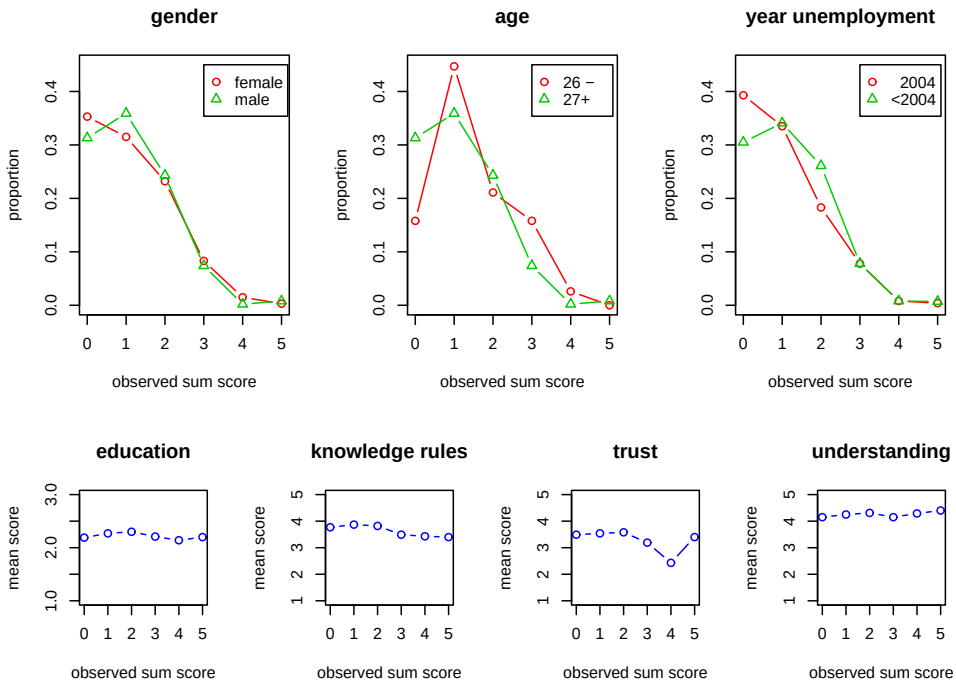

Figure 4.1: Observed sum score proportion given the categories within the dummy variables *gender*, *age* and *year unemployment* (upper plots), and mean predictor score within observed sum score for the continuous variables *education*, *knowledge rules*, *trust* and *understanding* (lower plots).

younger respondents also have a relatively high proportion of ones (about 45%). The profiles within *year unemployment* are again relatively similar, although persons who became unemployed in 2004 have a higher percentage of zero response (about 40%) compared to the respondents who became unemployed before 2004 (30%). The four plots at the bottom show the mean predictor scores within the observed sum scores for the respective continuous variables *education*, *knowledge rules*, *trust* and *understanding*. The predictor means do not show a any clear linear associations with the observed sum score, although for the variable *knowledge rules* the means seem to slightly decrease with increasing sum scores. The effect of sum score is most pronounced on the means of the predictor *trust*, but the association pattern is erratic.

4.3 The Model

4.3.1 The Multinomial Randomized Response Model

Consider an randomized response design with M sensitive questions, each assessing the presence or absence of a sensitive characteristic. Let the random variable Y_m^* denote the observed response to the m^{th} question, with $y_m^* \in \{0 \equiv no, 1 \equiv yes\}$ and $m \in \{1, \ldots, M\}$. Similarly, let Y_m denote the true status with respect to the sensitive characteristic, with $y_m \in \{0 \equiv \text{absent}, 1 \equiv \text{present}\}$. The binomial randomized response model for the binary variable Y_m^* is given by

$$I\!P(Y_m^* = y_m^*) = \sum_{y_m=0}^{1} p_{y_m^*|y_m} \pi_{y_m}, \tag{4.1}$$

where $\pi_{y_m} = I\!P(Y_m = y_m)$ and $p_{y_m^*|y_m} = I\!P(Y_m^* = y_m^*|Y_m = y_m)$ are the conditional misclassification probabilities that can be derived from the probability distribution of the randomizing device. For a more detailed discussion of this model we refer to Chaudhuri and Mukerjee (1988).

Next consider the true sum score of the M sensitive characteristics, denoted by the variable

$$S = \sum_{m=1}^{M} Y_m. \tag{4.2}$$

If S follows a multinomial distribution with parameters $\pi_0, \ldots, \pi_M$, then the multinomial randomized response model

$$I\!P(S^* = s^*) = \sum_{s=0}^{M} q_{s^*|s} \pi_s, \tag{4.3}$$

applies, where S^* denotes the number of observed *yes* responses on the M sensitive questions and $q_{s^*|s} = I\!P(S^* = s^*|S = s)$, for $s, s^* \in \{0, \ldots, M\}$.

The misclassification probabilities $q_{s^*|s}$, that exist if and only if the $p_{y_m^*|y_m}$ are the same for all m, can be derived as the multinomial probabilities

$$q_{s|t} = \sum_{j=0}^{t} \binom{t}{j} \binom{M-t}{s+j-t} p_{1|1}^{t-j} p_{0|1}^{j} p_{1|0}^{s+j-t} p_{0|0}^{M-s-j}, \tag{4.4}$$

for $s, t \in \{0, 1, \ldots, M\}$ and $t \leq s + j \leq M$ (Cruyff, van den Hout and van der Heijden, 2008).

As an illustration, consider the forced response design of the social security survey with two binary variables Y_1 and Y_2. Application of (4.4) for $M = 2$ yields the misclassification probabilities $q_{s^*|s}$

$$\begin{pmatrix} q_{0|0} & q_{0|1} & q_{0|2} \\ q_{1|0} & q_{1|1} & q_{1|2} \\ q_{2|0} & q_{2|1} & q_{2|2} \end{pmatrix} = \begin{pmatrix} p_{0|0}^2 & p_{0|0}p_{0|1} & p_{0|1}^2 \\ 2p_{0|0}p_{1|0} & p_{1|0}p_{0|1} + p_{0|0}p_{1|1} & 2p_{0|1}p_{1|1} \\ p_{1|0}^2 & p_{1|0}p_{1|1} & p_{1|1}^2 \end{pmatrix}.$$

4.3.2 The Poisson Randomized Response Model

Assume that the true sum score S is generated by a Poisson process with parameter λ. Since realizations of S are limited to the maximum value of M, the Poisson distribution of S is truncated at the right (Cameron and Trivedi, 1998), so that

$$I\!P(S = s|s \leq M) = \frac{\pi_s}{\sum_{s=0}^{M} \pi_s}, \tag{4.5}$$

with

$$\pi_s = \frac{\exp(-\lambda)\lambda^s}{s!}. \tag{4.6}$$

Substitution of the multinomial probabilities π_s in model (4.3) for the expression at right-hand side of (4.5), with π_s defined as in (4.6), yields the (right-truncated) Poisson randomized response model

$$I\!P(S^* = s^*|s^*, s \leq M) = \sum_{s=0}^{M} q_{s^*|s} \frac{\pi_s}{\sum_{s=0}^{M} \pi_s}. \tag{4.7}$$

4.3.3 The Zero-Inflated Randomized Response Regression Model

Count data are often characterized by an excess of zeros relative to a Poisson distribution. To account for the excess of zeros Lambert (1992) introduced the zero-inflated Poisson (ZIP) model

$$I\!P(S = s) \;=\; (1 - \theta)\pi_s + I\theta, \tag{4.8}$$

with $S \in \{0, 1, 2, \ldots\}$, π_s defined as in (4.6), and I an indicator variable taking on value 1 if $S = 0$, and 0 otherwise. The parameter θ denotes the probability of an excess zero in the observed counts, i.e. a zero count that is not generated by the Poisson process.

Now suppose that in the context of randomized response the true sum score variable S is generated by a Poisson process. In the absence of SP the observed sum score variable S^* is entirely generated by the Poisson randomized response process. In the presence of SP however, S^* is generated partly by the Poisson randomized response process and partly by SP. Let parameter θ^* denote the probability that the observed sum score is generated by SP, and let $1-\theta^*$ denote the probability that the observed sum score is generated by Poisson randomized response process. The distribution of S^* is then given by

$$I\!P(S^* = s^* | s^*, s \leq M) = (1 - \theta^*) \sum_{s=0}^{M} q_{s^*|s} \frac{\pi_s}{\sum_{s=0}^{M} \pi_s} + I^* \theta^*, \tag{4.9}$$

where I^* is an indicator variable taking on the value 1 if $S^* = 0$, and 0 otherwise.

Both parameters λ and θ^* in (4.9) can be modeled as a function of predictors. Let variable S_i denote the true sum of sensitive characteristics of individual i, for $i \in \{1, \ldots, n\}$, and let $\mathbf{x}_i = (x_{i0}, \ldots, x_{ik})'$ and $\mathbf{z}_i = (z_{i0}, \ldots, z_{il})'$ be vectors that may or may not contain the same predictors. Let the Poisson parameter of individual i depend on $\mathbf{x}_i$ according to

$$\lambda_i = \exp(\mathbf{x}_i' \boldsymbol{\beta}), \tag{4.10}$$

and let the probability of zero-inflation depend on $\mathbf{z}_i$ according to

$$\theta_i^* = \frac{\exp(\mathbf{z}_i' \boldsymbol{\gamma})}{1 + \exp(\mathbf{z}_i' \boldsymbol{\gamma})}, \tag{4.11}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)'$ and $\boldsymbol{\gamma} = (\gamma_0, \ldots, \gamma_l)'$ are parameter vectors. The ZIP randomized response regression model is given by

$$I\!P(S_i^* = s_i^* | s_i^*, s_i \leq M, \mathbf{x}_i, \mathbf{z}_i) = (1 - \theta_i^*) \sum_{s_i=0}^{M} q_{s_i^*|s_i} \frac{\pi_{s_i}}{\sum_{s_i=0}^{M} \pi_{s_i}} + I^* \theta_i^*, \tag{4.12}$$

where $\pi_{s_i} = \exp(-\lambda_i) \lambda_i^{s_i} / s_i!$.

4.3.4 Estimation

The ZIP randomized response regression model (4.12) (as well as the other models presented in this section) can be estimated by maximizing the kernel of the observed-data log-likelihood

$$\ln \ell^*(\boldsymbol{\beta}, \boldsymbol{\gamma} | \mathbf{s}^*, \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{n} \ln \left((1 - \theta_i^*) \sum_{s_i=0}^{M} q_{s_i^* | s_i} \frac{\pi_{s_i}}{\sum_{s_i=0}^{M} \pi_{s_i}} + I^* \theta_i^* \right) \quad (4.13)$$

with respect to the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. We have written code for the quasi Newton-Raphson procedure `QNewtonmt` of the statistical software programm GAUSS to estimate the model parameters. The procedure uses the BFGS method with numerically computed gradients and Hessian matrix, and standard errors are obtained from the inverse of the estimated Hessian. Convergence is generally fast, but due to machine imprecision problems may be encountered with the inversion of the Hessian. The use of slightly different starting values usually solves this problem. The observed-data likelihood is convex and unimodal when evaluated as a function of the parameters θ and λ. Figure 4.2 shows the shape likelihood function for the ZIP randomized response model 4.9 given the social security data. This model does not include any predictors for the parameters λ and θ, and the likelihood function is evaluated for the SP parameter $\theta \in (0, ..25)$ and the Poisson parameter $\lambda \in (.25, .75)$.

4.3.5 The Poisson Assumption

It is a well known statistical result that for $M \gg 1$, $\pi \ll 1$ and $M\pi \approx 1$, the distribution of the sum of M i.i.d. Bernoulli variables with success probability π is approximated by a Poisson distribution with parameter $\lambda = M\pi$. The Poisson randomized response models presented in this paper are based on the assumption that the five randomized response variables Y_m are Bernoulli variables and that the sum of these variables follows a Poisson distribution with parameter $\lambda = \sum_{m=1}^{M} \pi_{1_m}$, where π_{1_m} denotes the success probability of variable Y_m (i.e. the prevalence of the sensitive characteristic). In this section we evaluate the validity of this assumption, given that in our example M is relatively small and that the success probabilities π_{1_m} are not identical for different m.

To evaluate the adequacy of the Poisson assumption, we first derive the exact distribution of the sum of five independent Bernoulli variables with

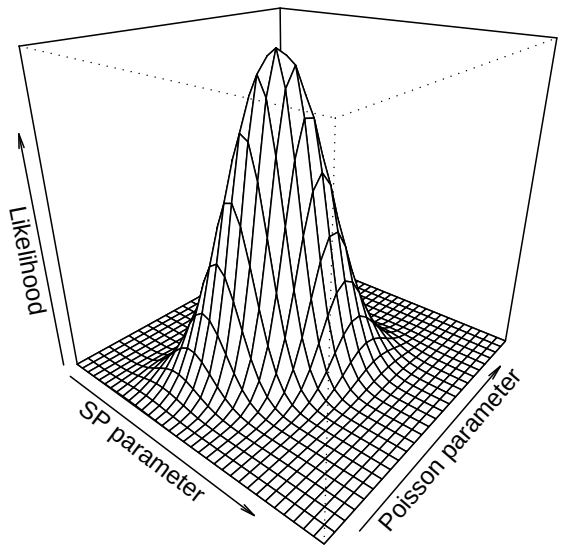

Figure 4.2: Likelihood function for the ZIP randomized response model evaluated for the SP parameter $\theta^* \in (0, .25)$ and the Poisson parameter $\lambda \in (.25, .75)$.

success probabilities equal to the prevalence estimates $\widehat{\pi}_{1_m}$ of the five variables Y_m of the social security survey. The prevalence estimates obtained with the multinomial randomized response model (4.1) are $\widehat{\pi}_{1_1} < .001$, $\widehat{\pi}_{1_2} = .050$, $\widehat{\pi}_{1_3} = .009$, $\widehat{\pi}_{1_4} = .069$ and $\widehat{\pi}_{1_1} = .172$, with $\widehat{\pi}_{1_1}$ set equal to .001. We then approximate this distribution by a Poisson distribution with $\lambda = \sum_{m=1}^{5} \widehat{\pi}_{1_m} = .301$. The two distributions are shown in the table below.

	0	1	2	3	4	5
Exact distribution	.7250	.2498	.0244	.0008	.0000	.0000
Poisson approximation	.7401	.2228	.0335	.0033	.0002	.0000

The Poisson approximation assigns more mass to the zero count and to counts larger than 2, and thus overestimates the true variance. It underestimates the probability of count 1 by .0270, which corresponds to a relative difference of approximately 11%. In view of the fact that the absolute deviations in probability of the remaining counts are smaller, the Poisson approximation seems satisfactory for all practical purposes.

4.4 Analysis of the Social Security Data

Table 4.1 presents fit indices for the multinomial randomized response model ($\mathcal{M}$), the Poisson randomized response model ($\mathcal{P}$), the ZIP randomized response null-model ($\mathcal{Z}_0$), the model $\mathcal{Z}_\beta$ including the five demographic predictors of the Poisson parameter, and the full model $\mathcal{Z}_{\gamma,\beta}$ with the additional two predictors of SP. The table reports the loglikelihood, the Akaike Information Criterion (AIC) given by $2k - 2\ln\ell^*$, the Bayesian Information Criterion (BIC) given by $k\ln n - 2\ln\ell^*$, with $\ln\ell^*$ the maximized loglikelihood and k the number of independently estimated parameters. For the models without predictors we present the Pearson chi-square statistic X^2 with $df = M - k$, where M denotes the number of independently observed sum score frequencies. The last column of Table 4.1 presents the SP probability estimates $\widehat{\theta}^* = (\sum_{i=1}^n I_{(S_i^*=0)})^{-1} \sum_{i=1}^n I_{(S_i^*=0)}\widehat{\theta}_i$ for the three ZIP models.

Table 4.1: Loglikelihoods, AIC's, BIC's and Pearson X^2 statistics, and SP estimates $\widehat{\theta}^*$.

	Model	loglik.	AIC	BIC	k	X^2	df	$\widehat{\theta}^*$
$\mathcal{M}$	Multinomial	-1170.8	2351.6	2375.4	5	6.1	0	-
$\mathcal{P}$	Poisson	-1183.3	2368.6	2373.4	1	56.0	5	-
$\mathcal{Z}_0$	ZIP (null)	-1173.2	2350.5	2360.0	2	19.6	4	.126
$\mathcal{Z}_\beta$	ZIP (incl. β)	-1167.0	2348.1	2381.5	7	-	-	.124
$\mathcal{Z}_{\gamma,\beta}$	ZIP (full)	-1165.0	2348.0	2391.0	9	-	-	.121

Although model $\mathcal{M}$ is saturated, the fitted response frequencies $\widehat{n}_0 = 272.0$, $\widehat{n}_1 = 319.1$, $\widehat{n}_2 = 195.3$, $\widehat{n}_3 = 66.7$, $\widehat{n}_4 = 12.6$, $\widehat{n}_5 = 4.3$ do not equal the corresponding observed response frequencies. The fact that X^2 is non-

zero with zero degrees of freedom indicates that one or more of the estimates are on the boundary of the parameter space (van den Hout and van der Heijden, 2002). The expected distribution of the true sum score variable S is

$$\widehat{\boldsymbol{\pi}}(\mathcal{M}) = (.878, .000, .116, .000, .000, .006),$$

with (near) zero-probability estimates for one, three and four rule violations. An interesting result is that the probability estimate of .6% for five rule violations is inconsistent with the fact the smallest univariate prevalence estimate of regulatory noncompliance is only .1% (see Section 4.3.5).

Model $\mathcal{P}$ clearly does not fit well, indicating that for our application the Poisson assumption does not hold. SP is introduced in model $\mathcal{Z}_0$ with an estimated probability of 12.6%. This model fits substantially better and is the best model in terms of BIC. The Pearson chi-square of 19.6 with 4 degrees of freedom however indicates lack of fit. The fitted frequencies $\widehat{n}_0 = 287.2$, $\widehat{n}_1 = 298.9$, $\widehat{n}_2 = 199.5$, $\widehat{n}_3 = 70.1$, $\widehat{n}_4 = 13.3$ and $\widehat{n}_5 = 1.1$ show that the lack of fit is primarily due to the underestimation of n_5, this cell contributes about 80% (14.6) to the total X^2 value. In terms of AIC the models $\mathcal{Z}_\beta$ and $\mathcal{Z}_{\gamma,\beta}$ fit best. Both models estimate the SP probability a little above 12%. The best model is $\mathcal{Z}_{\gamma,\beta}$, with the marginal distribution of the fitted values of S_i given by

$$\widehat{\boldsymbol{\pi}}(\mathcal{Z}_{\gamma,\beta}) = (.657, .267, .063, .011, .002, .000).$$

The AIC and BIC disagree with respect to model choice. Since randomized response requires much larger samples than direct question designs and the BIC punishes for sample size, we feel that the BIC might be too conservative. Therefore we prefer the model with the lowest AIC, which is $\mathcal{Z}_{\gamma,\beta}$. This choice is further motivated by the fact that the AIC decreases to 2343.4 when the four nonsignificant regression parameters in this model (see Table 4.2) are set to zero. In this case the BIC becomes 2367.2, so that according to this criterion $\mathcal{Z}_0$ remains the preferred model. The disagreement between the two criteria indicates that the evidence for the effects of the predictors in model $\mathcal{Z}_{\gamma,\beta}$ is not strong.

Table 4.2 presents the parameter estimates of the predictors in model $\mathcal{Z}_{\gamma,\beta}$. The upper part of the table shows the results for the predictors in the vector $\mathbf{x}$. The last column reports the effect size $\exp(\widehat{\beta})$, expressing the relative change in the Poisson parameter for a unit change in the predictor. (For continuous variables the standardized effect size can be computed by raising the reported effect size to the power of the standard deviation of the

predictor.) The variables *year unemployed* and *knowledge rules* are significant predictors of the Poisson parameter. Regulatory noncompliance increases after the first year of unemployment; the estimated number of rule violations for a person unemployed longer than 1 year is 1.78 times that of a person unemployed less than 1 year. Better knowledge of the rules is associated with lower levels of regulatory noncompliance; the standardized effect size of .78 denotes the factor change in the Poisson parameter for each standard deviation increase in the score on *knowledge rules* (sd = .90).

Table 4.2: Parameter estimates, standard errors (se), t-values, and effect sizes for model $\mathcal{Z}_{\gamma,\beta}$.

predictors in $\mathbf{x}$	$\widehat{\beta}$	(se)	t-val.	$\exp(\widehat{\beta})$
constant	-.13	(.38)	-.32	-
gender (female)	.21	(.22)	.95	1.23
age (< 26)	.50	(.36)	1.39	1.65
education	.19	(.18)	1.07	1.21
year unemployment (< 2004)	.58	(.29)	1.97	1.78
knowledge rules	-.27	(.12)	-2.34	.76
predictors in $\mathbf{z}$	$\widehat{\gamma}$	(se)	t-val.	$\exp(\widehat{\gamma})$
constant	-.64	(1.04)	-.61	-
trust	.14	(.33)	.43	1.15
understanding	-.46	(.23)	-1.99	.63

The lower part of Table 4.2 reports the parameter estimates for the predictors in the vector $\mathbf{z}$. The last column reports the effect size $\exp(\widehat{\gamma})$, which expresses the relative change in the odds of SP for a unit change in the predictor. The parameter estimate for the variable *understanding* is significant. Better understanding of the forced response method results in less self-protective responses; the standardized effect indicates that the odds of SP decrease by approximately two-third (.67) for each standard deviation increase in the score of *understanding* (sd = .85).

In order to assess the fit of model $\mathcal{Z}_{\gamma,\beta}$ more closely, we evaluate the correspondence between the observed and fitted frequencies within the response categories of each the predictor variables. Figure 4.3 plots the Pearson residuals $(n_{s^*x_{jk}} - \widehat{n}_{s^*x_{jk}})/\sqrt{\widehat{n}_{s^*x_{jk}}}$, with $n_{s^*x_{jk}}$ denoting the observed frequency

of persons with sum score s^* and score k on predictor j, and $\widehat{n}_{s^*x_{jk}}$ denoting the corresponding fitted frequency. Because of the low frequencies of the observed sum scores 4 ($n = 7$) and 5 ($n = 5$), these two categories have been collapsed into the single sum score category 4/5.

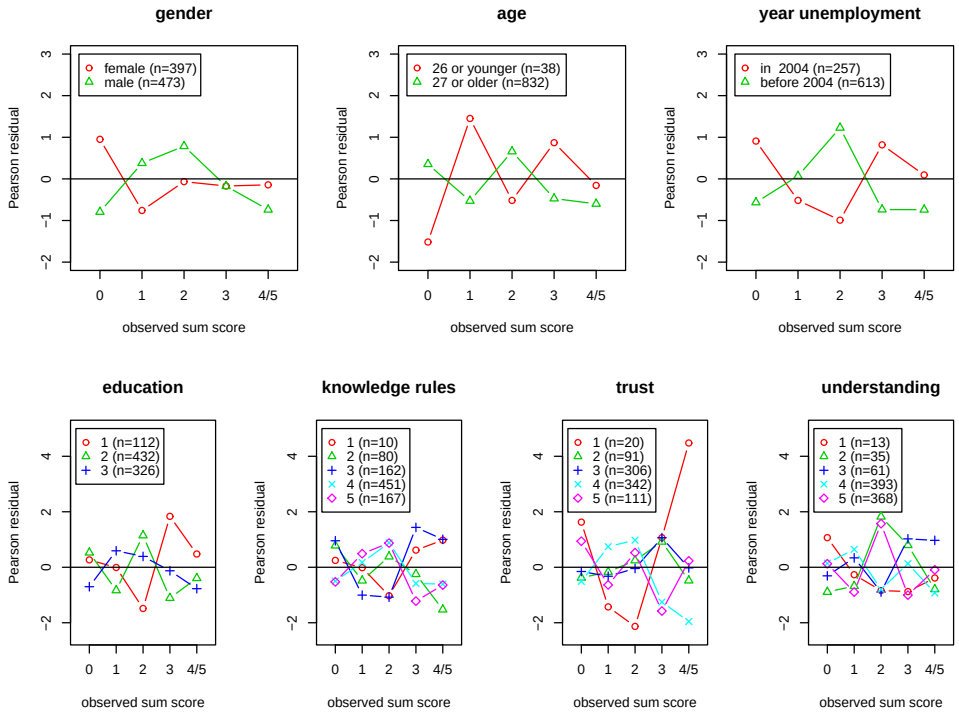

Figure 4.3: Pearson residuals given predictor score and observed sum score (with sum scores 4 and 5 collapsed into single category).

The upper three plots of Figure 4.3 do not show systematic patterns or outliers for the dummy variables. Only the category of respondents younger than 27 with sum score zero shows a moderately large (negative) residual, indicating that this group is slightly overestimated. The remaining four plots in the lower part of Figure 4.3 show large residuals for the predictor *trust*. The plot shows underestimation of respondents who have no trust in the randomized response design and who have an observed sum score of either zero or four or five, with an exceptionally large residual in the latter category ($n = 3$). Since the combination of no trust in the randomized response design and *yes* responses to (almost) all sensitive questions is somewhat counterintu-

itive, it suggests the presence of a response mechanism opposite to that of SP; it may be the case that are (a few) respondents who do not trust randomized response and who therefore (almost) always answer *yes*, irrespective of the outcome of dice. Although this is only a tentative explanation for the large residual, it would be interesting to see whether a similar response mechanism could also detected in other randomized response applications.

The results presented in this section show evidence for the presence of SP in the data; the models with the SP parameter fit better than the other models. The degree of freedom needed to estimate the SP parameter is gained by the Poisson assumption for the sum scores of regulatory noncompliance. In this application the Poisson models fit the data well, except for the underestimation of the five cases with an observed sum score 5. Further research is needed to explore the nature of this misfit.

4.5 Discussion

In this paper we introduce a zero-inflated Poisson regression model for the analysis of randomized response sum score data. The central assumption underlying the model is that the true sum score variable follows a Poisson distribution, and that the presence of SP results in a zero-inflated distribution of the observed sum score variable. We present an example with an randomized response sum score variable assessing noncompliance with social security regulations. The ZIP randomized response model is used to find (1) the probability distribution of regulatory noncompliance, (2) the probability of SP, (3) significant predictors of regulatory noncompliance and (4) significant predictors of SP.

From a substantive point of view, the ZIP randomized response regression model yields some interesting results. For officials at the Department of Social Affaires the negative effect of rule knowledge on regulatory noncompliance, suggesting that noncompliance is to a certain extent due to ignorance, may be of assistance in the formulation of new policies. The negative association between understanding of the forced response design and the probability of SP is especially interesting to social scientists interested in randomized response method. This result, that coincides with the conclusions of a study of psychological aspects of randomized response (Landsheer, van der Heijden and van Gils, 1999), suggests that adjustments in the instructions that would help the respondents to better understand the forced response design may

reduce response bias and thereby enhance the validity of the responses.

The central assumption of the model that the true sum score variable is generated by a Poisson process is questionable since (1) the number of Bernoulli (i.e. binary randomized response) variables making up the randomized response sum score variable is limited and (2) the success probabilities (i.e. the prevalence of the sensitive characteristics) are not identical. Based on the univariate prevalence estimates, we demonstrate that in our example the Poisson approximation is satisfactory. An evaluation study with manipulation of the numbers of Bernoulli variables and of success probabilities (not reported here) shows that the quality of Poisson approximation is most affected if one (or more) of the success probabilities becomes larger. This finding is corroborated by the more formal result of Serfling (1978) that the absolute deviations between a series of Bernoulli variables with different success probabilities and its Poisson approximation increase as a function of the squared success probabilities. Although it is difficult to give exact figures, we feel that the Poisson assumption is justified as long as the prevalence estimates of the binary randomized response variables do not exceed .25. Since randomized response deals with sensitive characteristics that are rare by definition, the risk of this happening should be small.

Chapter 5

The Doubly Zero-Inflated Poisson Model

5.1 Introduction

In the Netherlands there has been a growing political interest in regulatory compliance. The reason is that, in modern society, there are many rules and regulations and it is not always clear how well these are followed by the public. In politics the general feeling is that if rules and laws are imposed to the public, it is also important to investigate the level of compliance. This feeling has been intensified by two disasters that took place in the Netherlands in 2000. A fireworks explosion near the center of a middle sized city killing twenty-three inhabitants happened because the fireworks were not stored according to the rules, and a fire in a discotheque on new years eve in which fourteen teenagers lost their lives, occurred because fire regulations had not been followed.

To monitor the compliance levels with instrumental legislation the Dutch administration regularly conducts nationwide surveys among target populations. Recently example include surveys to assess compliance with regulations on the storage of food products by cafeteria, individual rents subsidy applications, dumping of industrial waste water in open waters, working schedules for taxi drivers and social welfare benefits. A drawback of interviewing people about rule violations is that this is a particularly sensitive

[1]Cruyff, M.J.L.F., Böckenholt, U. and van der Heijden. A Doubly Zero-Inflated Poisson Regression Model for Randomized Response Count Data, submitted

topic, so that respondents may not be willing to reveal their true behavior. In all the surveys the questions about regulatory noncompliance were therefore posed in the randomized response format.

The randomized response technique was introduced by Warner in 1965 as an interview technique to elicit sensitive information while protecting the respondents' privacy (Warner, 1965). In randomized response the respondent operates a randomizer, like a pair of dice or a deck of cards, and answers the question based on the outcome of the randomizer. In the original design by Warner the outcome of the randomizer determines whether the respondent answers the sensitive statement, for example "I use illegal drugs", or the complementary statement "I do not use illegal drugs". Since the interviewer does not know the outcome of the randomizer, confidentiality is ensured. Since Warner alternative randomized response designs have been developed that make use of other question and response formats. A well known example is the unrelated question design (Greenberg, Abel-Ela, Simmons and Horvitz, 1969), in which the outcome of the randomizer determines whether the respondent has to answer the sensitive question or an unrelated, neutral question. Another example is the forced response design (Boruch, 1971). In this design the respondent answers a single sensitive question, and the outcome of the randomizer determines whether the respondent has to give a truthful answer or a forced 'yes' or a forced 'no' response.

The vast majority of randomized response designs use a binary question format with the answer categories denoting the presence or absence of a sensitive characteristic, while only a very limited number of designs allow for questions with multiple answer categories (compare Lensvelt-Mulders, Hox, van der Heijden and Maas, 2005). An early example is the multi-proportions randomized response design by Abul-Ela, Greenberg and Horvitz (1967). In this design the answer categories denote membership of mutually exclusive groups that are assumed to follow a multinomial distribution. An example that allows for quantitative responses is the unrelated question design with by Greenberg, Kuebler, Abernathy and Horvitz (1971). The authors present two applications in which the responses denote the number times a woman has had an abortions and the income in dollars per head of a household. In both applications nonparametric estimates of the mean and variance of these sensitive characteristic are obtained. As a final example, Liu and Chow (1976) propose a design similar to the forced response design in which the response categories consist of a limited number of integers that correspond to sensitive quantities that are assumed to follow a multinomial distribution.

As an example of such sensitive quantities the authors mention the number of abortions.

In this paper we introduce a model for a randomized response variable denoting censored counts of a sensitive event. Such a censored sensitive event count could for instance be the number of abortions, with all counts above a certain number taken together in a single category. The model assumes that the sensitive event counts are generated by a zero-inflated Poisson process. This assumption implies that the population consists of two latent groups; a group that is incapable of experiencing the sensitive event (for example infertile women in case of abortions) and therefore has a zero count with probability 1, and another group for which the number of sensitive events follows a Poisson distribution. Since sensitive events are usually rare events, the Poisson assumption is justified by *law of rare events* that states that if events have a low probability of occurrence and the number of trials is large, the event counts will approximately be Poisson distributed.

Experimental studies by Edgell, Himmelfarb and Duchan (1982) and van der Heijden, van Gils, Bouts and Hox (2000) demonstrate that randomized response does not completely eliminate response bias; the results suggest that in response to questions about homosexuality and social welfare fraud part of the respondents protects their own privacy by giving the nonsensitive response, irrespective of the outcome of the randomizer. In the model for sensitive event counts we therefore distinguish between two kinds of response processes; randomized responses are generated by respondents who answer in keeping with the randomized response design, and *self-protective zeros* are generated by respondents who do not follow the randomized response and who answer 'zero' regardless of the outcome of the randomizer. This assumption toward of self-protective response behavior is more strict than the one offered by Kim and Ward (2005), who propose a multinomial randomized response model in which respondents who do not comply with the design are allowed to give a response that is below their true count, but this need not be the 'zero' response. The presence of self-protective zeros in the present model results in a zero-inflated response distribution in comparison to the distribution that would have been obtained if all respondent had followed the randomized response design.

The model we present has three parameters; (1) a Poisson parameter that specifies the sensitive event count distribution for persons who are capable of experiencing the sensitive event, (2) a it Poisson zero-inflation parameter that allows for persons who are incapable of experiencing the sensitive event and

(3) a it self-protective zero-inflation parameter that allows for respondents who do not follow the randomized response. Each of the three parameters can be written as a function of predictors. Given the presence of the Poisson parameter and the two zero-inflation parameters, we refer to the model as the doubly zero-inflated Poisson regression model.

We present two examples from a social security survey conducted in 2004 by the Department of Social Affaires in the Netherlands. In this survey 2.580 respondents are asked about the amount of illegal income they obtained in the past year in addition to their social security benefit with small jobs for friends or acquaintances and with working off the books. Several background variables are selected to serve as predictors of the model parameters.

The paper is structured as follows. Section 5.2 presents the data and the randomized response design used in the social security survey. Section 5.3 presents the model, and includes a simulation study to assess the identifiability of the three model parameters. The examples are presented in Section 5.4. Section 5.5 concludes.

5.2 The Data

In 2004 the Dutch Department of Social Affairs conducted a nationwide survey to assess the level of noncompliance with social security regulations. The sample consisted of 2.580 respondents receiving financial benefits under of three separate social security insurance acts. The Unemployment Act provides benefits to people who have recently become unemployed, the Disability Act provides benefits to people who have become (partially) disabled to perform labor, and the Assistance Act provides income to starters on the labor market and to long-term unemployed.

To be eligible for financial benefits one has to comply with various rules and regulations. Some of these rules are related to income that is obtained in addition to the social security benefit. In general it is allowed to keep a small percentage of additional income, provided that the income is reported to the social welfare agency and that the money is legally obtained. The respondents were asked the following two questions:

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 doing small jobs for friends or acquaintances (without reporting this to the social welfare agency)?

Question B 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?

Both questions have the same six response categories coded 1 to 6, respectively denoting '0 euros', '1 to 50 euros', '51 to 75 euros', 76 to 100 euros', '101 to 250 euros' and '251 euros or more'. The questions were answered according to the forced response design (Boruch, 1971), which was adapted to allow for multiple response categories. In this design the respondent is instructed to toss two dice and to answer 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 has to toss another die and has to answer the question by giving the number of eyes on that die. The following response frequencies were observed,

	1	2	3	4	5	6
jobs	1882	304	110	104	91	98
work	2014	245	108	74	72	67

with *jobs* referring to the income implied by question A, and *work* referring to the income implied by question B. We recoded the original response categories into censored counts of 'income units', with each response categories denoting a successive count (with '0 euros' corresponding to the count 'zero' and '250 euros or more' corresponding the censored count '5 or more'). The Poisson assumption for these pseudo count variables is again motivated by *law of rare events*. Since employment generating illegal income is usually rare, irregular and short-term, the probability of a 'success' is small, and a large number of 'trials' will be needed to generate a substantial amount of income. Under these conditions the pseudo counts of 'income units' will roughly follow a Poisson distribution.

The survey includes 11 variables that we use as predictors of the model parameters. We include age, gender, sector of last employment and the perceived cost and benefits of regulatory (non)compliance as predictors of the Poisson parameter specifying the the income count distribution. *Age* is a dummy-coded variable with reference category persons younger than 30 years ($n = 421$) and two dummies "31 to 50" ($n = 1201$) and "50 or older" ($n = 958$). *Gender* is a dummy-coded predictor denoting "males" ($n = 1069$), with "females" ($n = 1511$) as reference category. The variable *job sector* denotes the sector the respondent was last employed in and has dummies "food

& drinks" denoting the hotel, restaurant and bar sector ($n = 126$), and "construction" denoting the building and construction sector ($n = 79$). These two sectors are included for their reputation for offering ample opportunities to work off the books. The variables *cost compliance* and *benefits noncompliance* respectively assess on a 5-point scale the perceived effort to comply with the social security rules ($m = -0.24, sd = 0.53$), and the perceived benefits of not complying with the rules ($m = -0.09, sd = 0.57$).

The variables *insurance act* and *law conformity* are included in the model as predictors of the Poisson zero-inflation parameter. The dummy "UA" denotes beneficiaries of the Unemployment Act ($n = 870$), and the dummy "AA" beneficiaries of the Assistance Act ($n = 880$). The reference category Disability Act, consisting of persons who are (partly) unable to perform labor ($n = 830$), is expected to contribute more to the Poisson zero-inflation than the other two groups. The variable *law conformity* assesses the respondents' belief that you always have to comply with law. The variable is measured on a 5-point scale ($m = 0.14, sd = 0.43$), and is included because law-abiding persons are expected to have a higher probability to abstain from any activities that yield illegal income.

The variables *education*, *trust* and *understanding* are included in the model as predictors of the self-protective zero-inflation parameter. The variable *trust* ($m = 0.25, sd = 0.46$) is computed as the average score on four 5-point scales that assess different aspects of the respondents' believe that the forced response design offers confidentiality (Cronbach's $\alpha = .83$). *Understanding* assesses on a 5-point scale ($m = 0.61, sd = .43$) to what extent the respondent has understood how to answer the randomized response questions (for a more thorough discussion of these variables see Landsheer, van der Heijden and van Gils, 1999). Since *understanding* may be subject to social desirability (people generally do not like to admit that they did not understand something), the variable *education* was also included as a proxy of understanding the instructions. The variable denotes increasing levels of education ($m = 0.17, sd = .70$).

5.3 The Model

This section introduces the model. The first part relates the randomized response design to the Poisson distribution of the sensitive event counts, the second part introduces the Poisson zero-inflation parameter and the third

part the self-protective zero-inflation parameter. Estimation of the model by the maximum likelihood method is discussed in the fourth part. The section ends with a small simulation study that addresses potential identification problems with respect to the model parameters.

5.3.1 The Poisson Model

Consider a sensitive question with $K+1$ response categories ranging from 0 to K denoting censored counts of a sensitive event, with response category K denoting 'K sensitive events or more'. Let the random variable Y_i denote the number of sensitive events for individual i, for $i \in \{1, 2, \ldots, n\}$ and $y_i \in \{0, 1, \ldots, K\}$. If the counts are generated by a Poisson process with Poisson parameter λ_i depending on the covariate vector $\mathbf{x}_i$, then the count probabilities $\pi_{y_i}^P = I\!P(Y_i = y_i|\mathbf{x}_i)$ are given by the censored Poisson regression model (Cameron and Trivedi, 1998),

$$\pi_{y_i}^P = \begin{cases} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} & \text{for } y_i < K \\ \\ 1 - \sum_{y_i=0}^{K-1} \frac{\exp(-\lambda_i)\lambda_i^{y_i}}{y_i!} & \text{for } y_i = K \end{cases}, \tag{5.1}$$

where

$$\ln(\lambda_i) = \mathbf{x}_i\boldsymbol{\beta}, \tag{5.2}$$

and $\boldsymbol{\beta}$ is a parameter vector.

Let random variable Y_i^*, for $y_i^* \in \{0, 1, \ldots, K\}$, denote the observed response to the randomized response question. In the forced response design that was used the social security survey with K denoting '5 events or more', the conditional misclassification probabilities of observing response $Y_i^* = y_i^*$ given that the sensitive event count $Y_i = y_i$ follow from the probability distribution of the sum of the dice. Correct classification occurs with probability $p = 3/4$, which is the probability that the sum of two dice is $4, 5, 6, 7, 8, 9$ or 10. Hence the probability q that response y_i^* is a forced response equals $1/24$. Since one of the six possible forced responses coincides with the true sensitive event count, the conditional misclassification probabilities are given by

$$p_{y_i^*|y_i} = \begin{cases} p + q = 19/24 & \text{if} \quad y_i^* = y_i \\ \\ q = 1/24 & \text{if} \quad y_i^* \neq y_i \end{cases}. \tag{5.3}$$

Note that although the conditional misclassification probabilities do not depend on any individual characteristics, the subscript i in $p_{y_i^*|y_i}$ is maintained to properly define the summation $\sum_{y_i=0}^{K} p_{y_i^*|y_i}$ used in the subsequent models.

If the sensitive event counts were assumed to follow a multinomial distribution, with parameters $\pi_{y_i}^{\mathcal{M}} = (\pi_{0i}, \ldots, \pi_{Ki})$, we would obtain the multinomial randomized response model

$$\pi_{y_i^*} = \sum_{y_i=0}^{K} p_{y_i^*|y_i} \pi_{y_i}^{\mathcal{M}}. \tag{5.4}$$

For an excellent overview of this model we refer to Chaudhuri and Mukerjee (1988). The *Poisson randomized response model* (Poisson model) is derived from this model by substituting of $\pi_{y_i}^{\mathcal{M}}$ for $\pi_{y_i}^{P}$, giving

$$\pi_{y_i^*} = \sum_{y_i=0}^{K} p_{y_i^*|y_i} \pi_{y_i}^{P}. \tag{5.5}$$

where $\pi_{y_i}^{P}$ defined as in (5.1).

5.3.2 Poisson Zero-Inflation

The zero-inflated Poisson model is a mixture of a Poisson distribution and a degenerate distribution with all the mass in the zero count (Lambert, 1992). Let ϕ_i denote the probability that a person is incapable to experience the sensitive event, then the zero-inflated Poisson regression model is given by

$$\pi_{y_i}^{ZIP} = (1 - \phi_i)\pi_i^{P} + I_{(y_i=0)}\phi_i, \tag{5.6}$$

where $I_{(y_i=0)}$ being an indicator variable that takes value 1 if $y_i = 0$, and 0 otherwise. In model (5.6) the parameter ϕ_i depends on the covariate vector $\mathbf{z}_i$ according to the logistic function

$$\ln\left(\frac{\phi_i}{1-\phi_i}\right) = \mathbf{z}_i\boldsymbol{\gamma}, \tag{5.7}$$

where $\boldsymbol{\gamma}$ is a parameter vector.

In the randomized response setting with Y_i following a censored Poisson distribution, the distribution of Y_i^* is given by the it zero-inflated Poisson randomized response model (ZIP model)

$$\pi_{y_i^*} = \sum_{y_i=0}^{K} p_{y_i^*|y_i} \pi_{y_i}^{ZIP}, \tag{5.8}$$

where $\pi_{y_i}^{ZIP}$ is defined as in (5.6).

5.3.3 Self-Protective Zero-Inflation

We now turn attention to the self-protective zero responses. Define θ_i as the probability that individual i gives a self-protective zero response, with the log odds of θ_i depending on the covariate vector $\mathbf{u}_i$,

$$\ln\left(\frac{\theta_i}{1-\theta_i}\right) = \mathbf{u_i}\boldsymbol{\psi}, \tag{5.9}$$

where $\boldsymbol{\psi}$ is a parameter vector.

In a randomized response context where the sensitive event counts are generated by a Poisson process and in which all individuals have a nonzero probability of a positive sensitive event count, we obtain the *self-protective randomized response model* (SP model)

$$\pi_{y_i^*} = (1-\theta_i)\sum_{y_i=0}^{K} p_{y_i^*|y_i}\pi_{y_i}^{P} + I_{(y_i^*=0)}\theta_i, \tag{5.10}$$

where $I_{(y_i^*=0)}$ is an indicator variable that takes value 1 if the observed response $y_i^* = 0$, and 0 otherwise, and $\pi_{y_i}^{P}$ defined as in (5.1).

In a randomized response context where the sensitive event counts follow a zero-inflated Poisson distribution, the observed response distribution is given by the *self-protective zero-inflated Poisson model* (SP ZIP model)

$$\pi_{y_i^*} = (1-\theta_i)\sum_{y_i=0}^{K} p_{y_i^*|y_i}\pi_{y_i}^{ZIP} + I_{(y_i^*=0)}\theta_i, \tag{5.11}$$

with $\pi_{y_i}^{ZIP}$ defined as in (5.6).

5.3.4 Estimation

The models presented in this section can be estimated by maximizing the kernel of the observed-data log-likelihood. For the SP ZIP model defined in (5.11) the log-likelihood function

$$\ln \ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\psi}|\mathbf{y}^*, \mathbf{X}, \mathbf{Z}, \mathbf{U}) = \sum_{i=1}^{n} \ln(\pi_{y_i^*}), \tag{5.12}$$

with $\pi_{y_i^*}$ defined in (5.11), is maximized with respect to the parameter vectors $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{\psi}$. We have written code for the quasi Newton-Raphson procedure `QNewtonmt` of the statistical software programm GAUSS to obtain the parameter estimates. The procedure uses the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method with numerically computed gradients and Hessian matrix. In order to minimize the condition number of the Hessian matrix (and thus avoiding numerical problems with respect to the inversion of the Hessian), the predictors in the model should all be on approximately the same scale. In the examples the continuous predictors *costs compliance*, *social control*, *law conformity*, *education*, *trust* and *understanding*, that were originally measured on a scale ranging from 1 to 5, were all linearly transformed to a scale ranging from -1 to 1.

5.3.5 Parameter Identification

In this section we perform a simulation study to assess potential problems with the identifiability of the parameters λ_i, ϕ_i and θ_i of the SP ZIP model (5.11). Consider a censored count variable with values from '0' to '5 or more', and homogenous parameters λ, ϕ and θ. Suppose the responses are generated under the conditions:

Poisson	$\lambda = 0.4$	$\phi = 0$	$\theta = 0$
ZIP	$\lambda = 0.4$	$\phi = 0.4$	$\theta = 0$
SP	$\lambda = 0.4$	$\phi = 0$	$\theta = 0.4$
SP/ZIP	$\lambda = 0.4$	$\phi = 0.4$	$\theta = 0.4$.

Figure 5.1 depicts the probability distribution of the observed responses for the forced response design under the different data generating conditions. The figure shows that the response probabilities $0, 1$ and 2 are different in all conditions, but probabilities of counts above 2 are equal in the conditions without self-protective zeros (Poisson and ZIP) and in the conditions with self-protective zeros (SP and SP/ZIP). The high zero response probabilities in the conditions with a nonzero parameter θ show that this parameter has a large effect on the zero-inflation of the observed responses.

To investigate potential identification problems, we fitted the ZIP model (5.8), the SP model (5.10) and the SP ZIP model (5.11) to the observed response distributions under the ZIP, SP and SP/ZIP conditions (Figure 5.1) for a sample size $n = 1,000$. The parameter estimates are shown in Table 5.1. The three models all yield consistent estimates if correctly specified. The ZIP model underestimates the λ if the data generation process is SP,

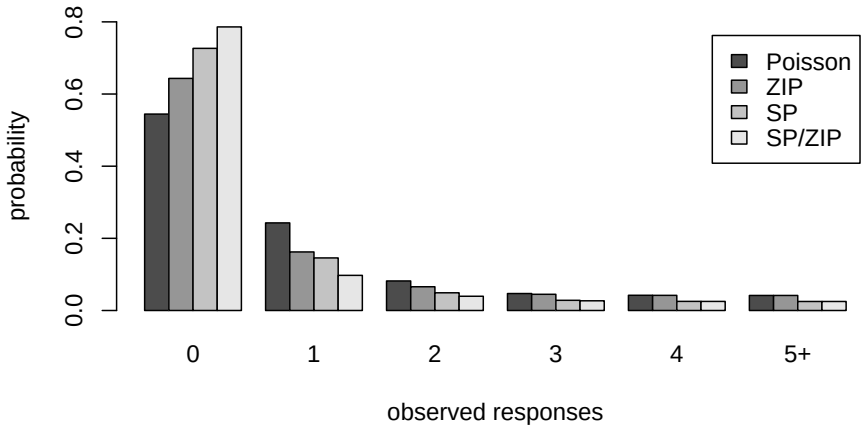

Figure 5.1: Observed response distributions given that the data are generated under the Poisson, ZIP, SP, or SP/ZIP condition.

and severely underestimates both λ and ϕ if the data is generated under the SP/ZIP condition. The SP model underestimates λ if the data are from a ZIP or a ZIP/SP distribution, and in the latter case θ is also slightly underestimated. The SP ZIP model yields consistent estimates in all conditions, and is therefore unbiased even in cases that the model is misspecified.

5.4 Social Security Survey Applications

Table 5.2 compares the fit statistics for the multinomial model (no predictors included), the SP ZIP null-model (no predictors included) and the full SP ZIP model (all predictors included) with the respective dependent variables *jobs* and *work*. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are included to assess the relative fit of the model, and the Pearson X^2 statistic to assess absolute model fit. The latter statistic is not presented for the full SP ZIP model, since the inclusion of predictors prohibits determination of the appropriate number of degrees of freedom.

The fit statistics clearly indicate that the multinomial randomized re-

Table 5.1: Parameter estimates of the ZIP, SP and SP ZIP randomized response models for data generated under the ZIP, SP and SP/ZIP conditions.

		Condition		
Fitted model		ZIP	SP	SP/ZIP
ZIP	$\widehat{\lambda}$	0.400	0.147	0.070
	$\widehat{\phi}$	0.400	0.000	0.003
SP	$\widehat{\lambda}$	0.214	0.400	0.193
	$\widehat{\theta}$	0.000	0.400	0.370
SP ZIP	λ	0.402	0.400	0.400
	$\widehat{\phi}$	0.399	0.001	0.400
	$\widehat{\theta}$	0.003	0.400	0.400

Table 5.2: Fit statistics and number of independent parameters (k) of the multinomial model, the SP ZIP null-model and the full SP ZIP model (including all predictors) with dependent variables *jobs* and *work*.

	model	loglike	k	AIC	BIC	X^2 (df)
jobs	Multinomial	-2532.0	5	5074.0	5103.3	6.6 (0)
	SP ZIP (null)	-2529.7	3	5065.5	5083.0	1.2 (2)
	SP ZIP (full)	-2447.1	17	4928.3	5027.8	-
work	Multinomial	-2207.6	5	4425.2	4454.5	43.4 (0)
	SP ZIP (null)	-2183.6	3	4373.1	4390.7	0.2 (2)
	SP ZIP (full)	-2116.9	17	4267.7	4367.2	-

sponse model is inadequate. Although this is a saturated model, the fitted response frequencies diverge from the observed response frequencies, resulting in a nonzero X^2. The Pearson residuals presented below

counts	0	1	2	3	4	5^+
jobs	0.7	0.3	0.2	-0.4	-1.6	-1.8
work	2.2	0.8	0.0	-3.3	-3.4	-3.9

show that in both models the counts $0, 1$ and 2 are underestimated, while the counts above 2 are overestimated. Apparently there is a significant underreporting of the higher counts, which provides strong evidence for the presence of self-protective zeros. The inclusion of the zero-inflation parameters in the SP ZIP null-models results in a decrease in the AIC and BIC, and the values of the X^2 statistics attain nonsignificant values. The inclusion of the predic-

tors in the full SP ZIP model further reduces the AIC and BIC, making this the preferred model.

Table 5.3 presents the parameters estimates, standard errors and t-values for the predictors in the full SP ZIP models with *jobs* and *work* as dependent variables. Also included are three effect size estimates. For predictors of the Poisson parameter the effect size $\Delta\lambda = \exp(\delta\beta)$ indicates relative change in the expected income count for a change δ in the predictor, holding the other predictors constant. The effect sizes $\Delta\phi/(1-\phi) = \exp(\delta\gamma)$ and $\Delta\theta/(1-\theta) = \exp(\delta\psi)$ respectively denote the relative change in the odds of Poisson zero-inflation and the relative change in the odds of a self-protective zero for a change δ in the predictor, holding the other predictors constant. All effect sizes are computed for $\delta = 1$ in case of the dummy-coded predictors, and for $\delta = s$ in case of the continuous predictors, with s denoting the standard deviation of the predictor.

Table 5.3: Parameter estimates, standard errors (se), t-values and relative change in λ and the odds of ϕ and θ per unit change in the categorical predictors or per standard deviation change in the continuous predictors for the SP ZIP model with dependent variables *jobs* (left) and *work* (right).

	jobs			*work*		
covariate	$\widehat{\beta}$ (se)	t	$\Delta\lambda$	$\widehat{\beta}$ (se)	t	$\Delta\lambda$
gender (male)	0.37 (.20)	1.8	1.45	0.14 (.23)	0.6	1.15
age (31-50)	-0.14 (.18)	-0.7	0.87	-0.33 (.24)	-1.4	0.72
age (51 or older)	0.01 (.14)	0.0	1.00	-0.38 (.27)	-1.4	0.69
job sector (food & drinks)	0.31 (.24)	1.3	1.36	0.38 (.28)	1.3	1.46
job sector (construction)	-0.41 (.38)	-1.1	0.66	1.19 (.34)	3.5	3.29
costs compliance	0.47 (.17)	3.8	1.29	0.31 (.20)	1.5	1.18
benefits noncompliance	1.23 (.15)	8.2	2.00	1.04 (.18)	5.7	1.81
social control	-0.65 (.17)	-3.8	0.75	-0.70 (.21)	-3.4	0.73
	$\widehat{\gamma}$ (se)	t	$\Delta\frac{\phi}{1-\phi}$	$\widehat{\gamma}$ (se)	t	$\Delta\frac{\phi}{1-\phi}$
Insurance Act (UA)	-0.52 (.45)	-1.2	0.59	-0.18 (.63)	-0.3	0.84
Insurance Act (AA)	-1.97 (.92)	-2.1	0.14	-2.02 (1.3)	-1.6	0.13
law conformity	1.51 (.66)	2.3	1.92	2.05 (1.3)	1.6	2.46
	$\widehat{\psi}$ (se)	t	$\Delta\frac{\theta}{1-\theta}$	$\widehat{\psi}$ (se)	t	$\Delta\frac{\theta}{1-\theta}$
education	-0.34 (0.16)	-2.2	0.71	-0.42 (.14)	-3.0	0.75
trust	-0.74 (0.28)	-2.7	0.72	-0.19 (.25)	-0.8	0.92
understanding	-0.22 (0.43)	-0.5	0.91	-0.32 (.20)	-1.6	0.87

The upper part of the table shows the parameters estimates for the predictors of the Poisson parameter, representing the expected 'unit of income' count. In the model with dependent variable *jobs* the significant predictors of the income count are *cost compliance*, *benefits noncompliance* and *social control*. Higher perceived costs of compliance are associated with a higher income count, which increases with 29% for each standard deviation increase in the score on *cost compliance*. Income is strongly related to perceived benefits of noncompliance, the income count doubles for a standard deviation change in the score on *benefits noncompliance*. More social control is associated with less income, the expected income count decreases with 25% if *social control* increases a standard deviation. In the model with dependent variable *work*, the effects of the predictors *benefits noncompliance* and *social control* have a similar interpretation in the model with dependent variable *jobs*, while in the latter *costs noncompliance* is nonsignificant. Additionally, the significant parameter for the 'construction' of the variable *job sector* indicates that persons who were last employed in the construction sector have an expected income count that is more than three times higher than that of persons who were last employed in a different sector.

The middle part of Table 5.3 shows the results for the predictors of Poisson zero-inflation. The variable *Insurance Act* is a significant predictor of the odds of an inflated zero income count in the model with *jobs*. The odds of being unable to work for friends are almost eight times higher for Disability Act beneficiaries than for Assistance Act beneficiaries. The model with *work* shows comparable estimates for *Insurance Act*, but the Wald tests nor the likelihood-ratio test for the joint effect of the dummies UA and AA ($LR = 5.2, df = 2, p = .08$) show significance.

The predictors of self-protective response behavior are shown in the bottom part of Table 5.3. The estimates for *education* are significant and have a similar interpretation in both models; the odds of a self-protective zero count decrease with 25 to 29% for each standard deviation increase in the respondents' educational level. Trust in the privacy protection of the forced response design is negatively associated with self-protective response behavior in the *jobs* model, with the odds of a self-protective zero decreasing by 28% for each standard deviation increase in *trust*. There is no evidence for an effect of *trust* in the *work* model.

Table 5.4 presents the estimated joint distribution of self-protective zeros, the zero-inflated income counts and the Poisson distributed income counts under the SP ZIP model. The estimate $\widehat{\theta}$ denoting the probability of a self-

Table 5.4: Joint probability distribution of self-protective zeros ($\widehat{\theta}$), zero-inflated illegal income counts ($\widehat{\pi}_0^{ZIP}$) and Poisson distributed illegal income counts ($\widehat{\pi}_0^P$ to $\widehat{\pi}_{5+}^P$) under the SP ZIP model.

	$\widehat{\theta}$	$\widehat{\pi}_0^{ZIP}$	$\widehat{\pi}_0^P$	$\widehat{\pi}_1^P$	$\widehat{\pi}_2^P$	$\widehat{\pi}_3^P$	$\widehat{\pi}_4^P$	$\widehat{\pi}_{5+}^p$
jobs	.289	.231	.324	.108	.032	.010	.004	.002
work	.401	.167	.314	.088	.021	.006	.002	.001

protective zero count is obtained as the sample average of the parameter estimates $\widehat{\lambda}_i$. The estimates of about 30% for *jobs* and about 40% for *work* suggest that self-protective response behavior is more likely to occur when the sensitivity of the question increases (working off the books is generally considered a more serious offence than not reporting income from jobs for friends and relatives). The probability estimate $\widehat{\pi}_0^{ZIP}$ of a zero-inflated Poisson count is computed as $\widehat{\phi}(1-\widehat{\theta})$, with $\widehat{\phi}$ being the sample average of the parameter estimates $\widehat{\phi}_i$. This probability is about 23% for *jobs* and 17% for *work*. These results suggest there are slightly more persons who never do do jobs for relatives than there are persons who never work off the books. The estimates $\widehat{\pi}_j^P$, for $j \in \{0, 1, \ldots, 5\}$, denoting the probabilities of the Poisson distributed income counts are computed with the Poisson parameter $\widehat{\lambda}$ equal to the sample average of the individual parameter estimates λ_i. For *jobs* the estimates are slightly higher than for *work*, indicating that more money is earned with doing small jobs for relatives than with working off the books.

Using the results in Table 5.4 we obtain the conditional probabilities of regulatory compliance given compliance with the randomized response design as $(.231 + .324)/(1 - .289) = .781$ for *jobs* and $(.167 + .314)/(1 - .401) = .803$ for *work*. If self-protective response behavior does not depend on regulatory noncompliance, then the estimates that 29.9% earns illegal income with *jobs* and 19.7% with *work* are unbiased estimates for the entire population. If however self-protective response behavior is positively related to regulatory noncompliance, then these are percentages are underestimates of the true levels of regulatory noncompliance.

5.5 Discussion

The SP ZIP model presented in this paper allows for zero-inflation on the level of the sensitive event counts and on the level of the observed responses, thus stratifying the sample into self-protective respondents, respondents who always comply with the rules and respondents who are potentially noncomplying. Although the model was originally meant for true count data, the examples from the social security survey show that the model can also be successfully applied to pseudo count data. It is probably true that imposing a Poisson distribution onto the pseudo counts may have induced bias in the prevalence estimates of regulatory noncompliance, but we feel that this problem is by far outweighed by the correction for self-protective response bias.

The model fits the data significantly better than the existing multinomial randomized responses, and provides interesting information to policy makers and social scientists. To policy makers the estimates with respect to the zero-inflated Poisson distribution are particularly interesting. The Poisson parameter estimates provide insight in the distribution of the amounts of money that are illegally earned in addition to social security benefit, with the zero-inflation parameter provide additional insight in the proportion of the population that is not at risk of earning illegal income. The corresponding predictors allow for a risk assessment of of groups with particular characteristics, and are therefore a useful tool for the formulation of more effective and efficient law enforcement strategies. To social scientists the estimates related to the self-protective zero-inflation parameter are most interesting. Since violations with respect to *jobs* are generally regarded as less serious than violations with respect to *work*, the estimates of this parameter suggest that self-protective response behavior is positively related to the sensitivity of the questions. The predictors of self-protective response behavior are useful to evaluate the validity of the randomized response design. The fact that in both examples education level is negatively related to self-protective response behavior suggests that the instructions with respect to the randomization procedure may not have been clear to respondents with a low level of education.

An interesting property of the doubly zero-inflated Poisson model is that the prevalence of self-protective zeros can be estimated on the basis of a single sensitive question. Böckenholt and van der Heijden (2007), Cruyff, van den Hout, van der Heijden and Böckenholt (2008), and Cruyff, Böckenholt,

van den Hout and van der Heijden (2008) investigate the presence of self-protective response bias in randomized response designs with multiple binary sensitive questions. In these studies it is assumed that part of the respondents exhibit *self-protective 'no'-saying*, which is defined as consistently giving the nonsensitive response ('no') to all the questions without taking the outcome of the randomizer into account, while the other respondents all consistently follow the randomized response design. The prevalence estimates of self-protective 'no' response profiles in these studies range between 10 and 25% and are substantially lower than the prevalence estimates of self-protective zeros of 30% and 40% in the present study. An explanation for this result is that the definition of self-protective 'no'-saying is too strict; self-protective 'no'-sayers may not be consistent over all questions, and may occasionally give a 'yes' response in keeping with the outcome of the randomizer. An alternative explanation is that in the present study the two-stage procedure with the dice is more complicated than in designs with binary questions, where only one toss of the dice is required. Further research on this topic may provide insights in the effect of the randomizing procedure on the prevalence of self-protective response behavior.

References

Abul-Ela, A-L. A., Greenberg, G. B., and Horvitz, D. G. (1967), A Multi-Proportions Randomized Response Model. *Journal of the American Statistical Association*, 62, 990-1008.

Böckenholt, U., and van der Heijden, P.G.M. (2004). Measuring Noncompliance in Insurance Benefit Regulations with Randomized Response Methods for Multiple Items, pp 106-110, in *Proceedings of the 19th International Workshop on Statistical Modelling* edited by A. Biggeri, E. Dreassi, C. Lagazio and M. Marchi. Florence, Italy.

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.

Böckenholt, U., Barlas, S., and van der Heijden, P.G.M. (2008). Do randomized-response designs eliminate response biases? An empirical study of non-compliance behavior. *Journal of Applied Econometrics*, in press.

Cameron, A. C. and Trivedi, P. K. (1998) *Regression Analysis of Count Data*. Econometric Monographs, 30, Cambridge University Press, Cambridge.

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

Boeije, H. and Lensvelt-Mulders, G.J.L.M. (2002). Honest by Chance: A Qualitative Interview Study to Clarify Respondents' (Non)-Compliance with Computer-Assisted Randomized Response. *Bulletin de Methodologie Sociologique*, **75**, 24-39.

Chaudhuri, A., and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*, Statistics: Textbooks and Monographs, 85, Marcel

Dekker Inc., New York.

Chen, T.T. (1989). A Review Of Methods For Misclassified Categorical Data In Epidemiology. *Statistics in Medicine*, **8**, 1095-1106.

Clark, S.J. and Desharnais, R.A. (1998). Honest Answers to Embarrassing Questions: Detecting Cheating in the Randomized Response Model. *Psychological Methods*, **3**, 160-168.

Cruyff, M.J.L.F., Böckenholt, U., van den Hout, A., and van der Heijden P. G. M., (2008), Zero-Inflated Poisson Regression Models for Randomized Response Sum Score Data, *Annals of Applied Statistics*, in press.

Cruyff, M.J.L.F., van den Hout, A., van der Heijden, P. G. M. and, Böckenholt, U. (2007), Log-linear Randomized Response Models Taking Self-Protective Response Behavior into Account. *Sociological Methods & Research*, **36**, 266-282.

Edgell, S.E., Himmelfarb, S and Duncan, K.L. (1982). Validity of Forced Response in a Randomized Response Model. *Sociological Methods and Research*, **11**, 89-110.

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. *Journal of Quantitative Criminology*, **4**, 409-439.

Fox, J.P. (2005). Randomized item response theory models. Journal of Educational and Behavioral Statistics, **30**, 1-24.

Gilula, Z. and Haberman, S.J. (2001). Analysis of Categorical Response Profiles by Informative Summaries. *Sociological Methodology*, **31**, 129-187.

Greenberg, B.G., Abel-Ela, A-L.,A., Simmons, W. R., and Horvitz, D.G. (1969). The Unrelated Question Randomized Response Model: Theoretical framework. *Journal of the American Statistical Association*, **64**, 520-539.

Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., and Horvitz, D.G. (1971), "Application of the Randomized Response Technique in Obtaining Quantitative Data", *Journal of the American Statistical Association*, **66**, 243-250.

Horvitz, D.G., Shah, B.V., and Simmons, W.R. (1967). The unrelated question randomized response model. Proceedings in the Social Statistics Section, American Statistical Association, 65-72.

Kim, J., and Warde, W.D. (2005). Some New Results on the Multinomial Randomized Response Model. *Communications in Statistics: Theory and Methods*, **34**, 847-856.

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

Kuk, Anthony Y.C. 1990. Asking Sensitive Questions Indirectly. *Biometrika*, **77**, 436-438.

Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, **34**, 1-14.

Landsheer, J.A., van der Heijden, P.G.M., and van Gils, G. (1999). Trust and Understanding, Two Psychological Aspects of Randomized Response. *Quality and Quantity*, **33**, 1-12.

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

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. *Journal of the Royal Statistical Association, Series A: Statistics in Society*, **169**, 305-318.

Liu, P.T., and Chow, L.P. (1976). A New Discrete Quantitative Randomized Response Model. *Journal of the American Statistical Association*, **71**, 72-73.

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

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statististical Society: Series B*, **42**, 109-142.

Peterson, B. and Harrell, F.E. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, **39**, 205-217.

Scheers, N.J. and Dayton, C.M. (1988). Covariate randomized response models. *J. American Statistical Association*, **83**, 969-974.

Serfling, R.J. (1978). Some elementary results on Poisson approximation in a sequence of Bernoulli trials. *SIAM Review*, **20**, 567-579.

van Gils, G., van der Heijden, P.G.M. and Rosebeek, A. (2001). *Randomized Response: Onderzoek naar Regelovertreding*, NIPO (in Dutch).

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

van den Hout, A., and van der Heijden, P.G.M. (2002). Randomized Response, Statistical Disclosure Control and Misclassification: A Review. *International Statistical Review*, **70**, 269-288.

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

van den Hout, A. and van der Heijden, P.G.M. (2004). The Analysis of Multivariate Misclassified Data with Special Attention to Randomized Response. *Sociological Methods and Research*, **32**, 384-410.

van den Hout, A., van der Heijden, P.G.M., and Gilchrist, R. (2006). The logistic regression model with response variables subject to randomized response. *Computational Statistics and Data Analysis*, **50**, 6060-6069.

Warner, Stanley L. (1965). Randomized response: a Survey Technique for Eliminating Answer Bias. *Journal of the American Statistical Association*, **60**, 63-69.

Samenvatting

Dit proefschrift behandelt de multivariate analyse van randomized response data. Randomized response werd in 1965 geïntoduceerd als een interviewmethode om ontwijkend antwoordgedrag op gevoelige vragen te elimineren (Warner, 1965). Het beantwoorden van de vragen gebeurd op basis van een kansmechanisme, waardoor de privacy van de respondent wordt beschermd. Uit onderzoek is gebleken dat de randomized response methode inderdaad tot eerlijkere antwoorden leidt dan directe vragen (zie van der Heijden *et al.*, 2000 en Lensvelt-Mulders *et al.*, 2005). De randomized response methode is in de laatste 10 jaar veelvuldig door de Nederlandse overheid gebruikt om regelnaleving te meten.

Een bekende randomized response methode is forced response (Boruch, 1971). Deze methode werkt als volgt. De respondent gooit twee dobbelstenen en krijgt dan een vraag voorgelegd, waarop het antwoord *ja* het gevoelige gedrag bevestigd, en het antwoord *nee* het gevoelige gedrag ontkent. De respondent beantwoord nu de vraag op basis van het aantal ogen dat hij met zijn worp heeft gegooid. Is het aantal ogen $2, 3$ of 4, dan is de respondent verplicht *ja* te antwoorden, en als het aantal ogen 11 of 12 is dan is de respondent verplicht *nee* te anwoorden. Voor de overige uitkomsten van de dobbelstenen beantwoordt de respondent de vraag naar waarheid. Aangezien allen de respondent op de hoogte is van de uitmomst van de dobbelstenen, is de vertrouwelijkheid van de gegeven informatie gewaarborgd.

Aanvankelijk werd de randomized response methode toegepast om een univariate prevalentieschatting van het gevoelige gedrag te verkrijgen. De laatste jaren is er echter een tendens om ook andere variabelen in de analyse van randomized response data te betrekken. Hierdoor kunnen ook andere interessante onderzoeksvragen worden beantwoord, zoals bijvoorbeeld: "Met welke persoonskenmerken hangt het gevoelige gedrag samen?", of: "In welke mate hangen verschillende soorten gevoelig gedrag samen?". Een tweede

ontwikkeling is dat men zich er van bewust is geworden dat randomized response niet tot algehele eliminatie van ontwijkend antwoordgedrag leidt. Inmiddels is gebleken is dat met behulp van multivariate analysetechnieken een prevalentieschatting van zogenaamd zelfbeschermend antwoordgedrag kan worden verkregen (het ontkennend antwoorden ongeacht de uitkomst van de randomisatieprocedure) (Böckenholt and van der Heijden, 2004 en 2007). Na correctie voor de zelfbeschermende antwoorden leidt tot zuiverdere prevalentieschattingen van het gevoelige gedrag.

In deze thesis worden vier statistische modellen voor de multivariate analyse van randomized response data geïntroduceerd. De modellen worden alle toegepast op data uit de *sociale zekerheidsmonitor*, die in de jaren 2000, 2002, 2004 en 2006 in Nederland is gehouden, en welke tot doel heeft regelnaleving van door uitkeringsgerechtigden vast te stellen. Het eerste model wordt behandeld in hoofdstuk 2 betreft een *log-lineair model* dat de samenhang tussen verschillende soorten regelovertreding analyseert, waarbij gecorrigeerd wordt voor zelfbeschermend antwoordgedrag. Het model is gebaseerd op eerder werk van Chen (1989) en van den Hout en van der Heijden (2004). Hoofdstuk 3 behandeld het *proportional odds model* (McGullagh, 1980), waarin gekeken de kansverdeling van de som van een aantal randomized response variabelen met betrekking tot regelovertreding wordt verklaard uit diverse persoonskenmerken. Het *zero-inflation Poisson model* in hoofdstuk 4 schat de prevalentie van de som van een aantal regelovertredingen aan de hand van een Poissonverdeling, en schat tegelijkertijd de prevalentie van zelfbeschermend antwoordgedrag. Hoofdstuk 5 behandelt het *doubly zero-inflated Poisson model*. Dit model wordt toegepast op vragen met meer dan twee andwoordcategoriën, waarbij de categoriën tellingen zijn een sensitieve gebeurtenis, zoals bijvoorbeeld het aantal keer dat een bepaalde regel is overtreden. Prevalentieschattingen van zelfbeschermend antwoordgedrag en van personen die nooit de regels overtreden zijn in dit model inbegrepen.

Curriculum Vitae

Maarten Cruyff was born on February 22 1961, in The Hague, the Netherlands. After completing secondary school at the Erasmus College in Zoetermeer (1980), he studied Languages and Cultures of Latin America at Leiden University. After graduation in 1992 he worked for a consultancy bureau on environmental issues, which from 1998 on he combined with a study psychology at Leiden University. In 2002 he graduated (cum laude) in methodology and statistics, and in that same year he started working for the Department Methodology and Statistics of Utrecht University. At present he is preparing an application for a VENI-scholarship on capture-recapture problems.