In :

Antonio Forcina, Giovanni M. Marchetti,
Reinhold Hatzinger, Gianfranco Galmacci (Eds.)

Statistical Modelling

Proceedings of the
11th International Workshop on Statistical Modelling
Orvieto, Italy, 15–19 July, 1996

Some logistic regression models for randomized response data

Peter G.M. van der Heijden
Ger van Gils

ABSTRACT: We introduce these alternative randomized response procedures and discuss how the logistic regression model can be adjusted in order to deal with randomized response data. We also will illustrate these models on an example dealing with social security fraud in the Netherlands.

KEYWORDS: Logistic Regression, Randomized Responses

1 Introduction

Some social topics are difficult to investigate with standard surveys to due their sensitivity. Examples are abortion, drug use, and homosexuality. Standard survey techniques usually fail when applied in these areas because many respondents do not answer questions honestly.

The randomized response technique aims to circumvent the sensitivity problem by offering individuals respondents safety. The basic idea is that an answer of an individual respondent to a sensitive question does not lead to 100 percent certainty about his/her behavior due to the use of a randomization procedure in the question. Many forms exist. The form originally proposed by Warner (1965) is this: consider that we are interested in the use of hard drugs. A respondent gets two versions of the same question:

A. I have used hard drugs in the past ten years.

B. I have never used hard drugs in the past ten years.

A randomized device determines which question a respondent is supposed to answer. The outcome of this device is known to the respondent, but unknown to the interviewer. The interviewer only hears the respondent say 'yes' or 'no', but does not know to which question this answer refers.

Let the probability of the randomized device to select form A be equal to p_A. This probability is specified by the design of the study. Let the probability of having used hard drugs be Π. Then the probability of a 'yes' answer is $P(yes) = p_A\,\Pi + (1-p_A)\,(1-\Pi)$. If p_A is unequal to .5, an estimate for Π can be obtained from the sample proportion of P(yes). Thus individual respondents are safe, whereas it is possible to obtain the information a researcher is interested in. Of course, there is no such thing as a free lunch, and the price paid here

is that the standard error for Π is larger than a comparably standard error for an estimate obtained from standard survey techniques. However, the latter estimate will probably be biased due to systematic underreporting.

The original proposal of Warner (1965) was not used very often, but it stimulated much research. In particular many alternative forms of randomized response were proposed. The estimates for Π obtained with these alternatives are more efficient than the original proposal.

In this paper we study two of these alternative forms. The question dealt with in this paper is how to relate covariates to the answers obtained with the randomized response procedure. Earlier work in this area has been done by Maddala (1983) and by Scheers and Dayton (1988). Maddala deals with this question in the context of the so-called forced response model, and Scheers and Dayton deal with this question in the context of the model originally proposed by Warner, and the so-called unrelated-question design that we will not discuss here. We will use an approach similar to the one taken by them to deal with the question in the context of a randomized response procedure recently proposed by Kuk (1990).

In section 2 we will introduce these alternative randomized response procedures. In section 3 we will discuss how the logistic regression model can be adjusted in order to deal with randomized response data. In section 4 we will illustrate these models on an example dealing with social security fraud in the Netherlands.

2 Two alternative randomized response procedures

In this section we discuss two alternatives for the randomized response procedure proposed by Warner (1965). For an overview of alternatives we refer to Warner Fox and Tracy (1986) and Chaudhuri and Mukerjee (1988).

A first alternative is the so-called forced response procedure (see, for example, Fox and Tracy, 1986, p.24; Maddela, 1983). In this alternative the respondent is offered the sensitive question, and a randomized device determines whether the respondent is supposed to answer 'yes', 'no', or answer honestly to the sensitive question.

Let the probability to answer forced 'yes' be P_1, let the probability to answer forced 'no' be P_2, and let the probability to answer honestly be $P_3 = 1-P_1-P_2$. Let Π be the probability to answer 'yes' to the sensitive question. Then the probability of a 'yes' answer is $P(yes) = P_1 + P_3\Pi$, and $P(no) = P_2 + P_3(1-\Pi)$. The sample estimate of P(yes) can then be used to obtain an estimate of Π. See for more details Fox and Tracy (1986) and Chaudhuri and Mukerjee (1988, p.16-17).

A second alternative is proposed by Kuk (1990). The idea behind this proposal was that it was sometimes difficult for respondents to answer 'yes', even though the design forced them to do this. An example is a question about homosexuality among soldiers in the US Army, where it was secretly checked whether soldiers who had a forced 'yes' really answered 'yes'. This did not turn

out to be the case.

In Kuk's procedure the interviewer has two stacks of cards in front of him/her. In the left stack the proportion of red cards is P_1, and in the right stack it is P_2. The respondent is supposed to draw one card from each stack, and to name the color of the card from the left stack if he should answer 'yes' to the sensitive question, and the color of the card from the right stack if he should answer 'no' to the sensitive question.

Thus P(red) = $P_1\Pi + P_2(1\text{-}\Pi)$, and P(black) = $(1\text{-}P_1)\Pi + (1\text{-}P_2)(1\text{-}\Pi)$. So in this alternative the interviewer only hears 'black' or 'red', and for the whole group he can estimate Π from the proportion of respondents having answered 'red' and 'black'. For more details about this procedure we refer to Kuk (1990).

3 Adjusted logistic regression models

If a direct procedure was used to obtain answers on a senstive question, and the sensitive question was considered to be a response variable, then the logistic regression model would be a natural candidate to relate covariates to these response variables. In this section we shortly introduce the logistic regression model for direct questions, and then show how this model is to be adjusted to deal with the two types of randomized response data discussed in section 2.

Let $\Pi_{j|i}$ be the probability to give answer $j(j = 1, 2)$ as a function of covariate vector i. In this context assume $j = 1$ means 'no' and $j = 2$ means 'yes'. Let the k'th covariate value in covariate vector indexed by i be denoted by x_{ik}, where x_{ik} is either a continuous variable or a dummy-coded variable. Let β_k be the regression parameter determining the relation between covariate k and response j. Then the logistic regression model is defined as

$$\Pi_{j|i} = \frac{1}{1 + exp \sum_k x_{ik}\beta_k} \text{if } j = 1 \tag{1}$$

$$\Pi_{j|i} = \frac{exp \sum_k x_{ik}\beta_k}{1 + exp \sum_k x_{ik}\beta_k} \text{if } j = 2 \tag{2}$$

Let n_{ij} be the number of responses j for the observations with covariate vector i. Then the log likelihood for the model is

$$\log L = \sum_i \sum_j n_{ij} \log \Pi_{j|i} = \sum_i n_{i1} \log \Pi_{1|i} + \sum_i n_{i2} \log \Pi_{2|i} \tag{3}$$

This loglikelihood can be maximized over the parameters β_k. The first and second derivatives, given in the appendix, can be used to estimate model (1)-(2) with, for example, the Newton-Raphson algorithm.

We will now adjust the logistic regression model to the situation where we deal with forced response data. For the forced response model this is worked out by Maddala (1983, p.54-56; see Scheers and Dayton, 1988, for a similar

proposals for different randomized response procedures). Assume that $\Pi_{j|i}$ is the probability to fall into answer category j of the sensitive question ($j = 1$ for 'no', and $j = 2$ for 'yes'), given a covariate vector indexed by i. This probability is then supposed to follow a logit model. Let P_1 be the probability of forced 'yes', P_2 the probability of forced 'no', and P_3 the probability of honest response. In a direct question situation $P_1 = 0$, $P_2 = 0$, $P_3 = 1$. The probabilities for the sensitive question are denoted by $\Pi_{j|i}$, where $j = 1$ is 'no' and $j = 2$ is 'yes'. This leads to $P_i(yes) = P_1 + P_3\ \Pi_{2|i}$, and $P_i(no) = P_2 + P_3\ \Pi_{1|i}$.

Thus the loglikelihood for the randomized response data becomes

$$\log L = \sum_i n_{i1} \log(P_2 + P_3 \Pi_{1|i}) + \sum_i n_{i2} \log(P_1 + P_3 \Pi_{2|i}). \qquad (4)$$

and this loglikelihood is to be maximized over the parameters of the logit model for $\Pi_{j|i}$. (see (1)-(2); first and second partial derivatives can be found in the appendix).

The second randomized response procedure for which we will adjust the logistic regression model is Kuk's (1990) procedure (see section (2)). In this procedure the respondent has two stacks of cards in front of him/her. In the left stack the proportion of red cards is P_1, and in the right stack it is P_2. For ease of notation we define $P_3 = 1 - P_1$ and $P_4 = 1 - P_2$. This leads to $P_i(red) = P_1 \Pi_{1|i} + P_2\ \Pi_{2|i}$ and $P_i(black) = P_3\ \Pi_{1|i} + P_4 \Pi_{2|i}$. Let n_{i1} be the number of responses 'red' for covariate vector indexed by i, and n_{i2} the number of 'black' responses.

Thus the loglikelihood becomes

$$\log L = \sum_i n_{i1} \log(P_2 + P_3 \Pi_{1|i}) + \sum_i n_{i2} \log(P_1 + P_3 \Pi_{2|i}). \qquad (5)$$

and again this loglikelihood is to be optimized over the parameters of the logit model for $\Pi_{j|i}$ (see (1)-(2); first and second partial derivatives can be found in the appendix).

4 Example

Both procedures are illustrated using data from a pilot study on social security fraud. For part of the sample we know from the social security office that the respondents are caught for income fraud, however, these respondents do not know that we know this. Thus we can investigate the validity of their answers obtained with the randomized response method, and the regression models can be used to investigate which covariates influence this validity.

The research was presented to the respondents as an investigation into "making ends meet in case of unemployment benefit". After about half an hour the sensitive questions were introduced. The sensitive question we study here was "did you ever fail to report a part of your income to the Social Service, while you were supposed to do this by law? (This can be income from labour,

extra earnings, gifts, alimony and the like)." For more details we refer to van Gils and van der Heijden (1996).

For the forced response procedure we let the respondent throw two dice, and when the sum equals 2, 3 or 4, he/she is supposed to answer 'yes' to our sensitive question, when it is 11 or 12, he/she is supposed to answer 'no', and when the sum is between 5 and 10 he is supposed to answer honestly. For Kuk's procedure in the left stack of cards the proportion of red cards is .8, and in the right stack of cards the proportion of red cards is .2. In case of fraud the respondent is supposed to name the colour of the card from the left stack, and in case of no fraud, from the right stack. The results of the randomized response procedures are compared with two other procedures, namely one in which respondents are asked sensitive questions directly, and another procedure where respondents are allowed to type in answers privately on a portable computer brought along by the interviewer (we call this method 'self completion').

Since we know that all respondents committed fraud, the question answered by the design is whether the way in which the question is asked (i.e. by randomized response or not) influences the probability that respondents admit fraud, in other words, answer honestly. Table 1 summarizes the analyses. In panel 1 model fit is reported for each of the four conditions. In model 1 only a constant is used in the logistic regression model, and the estimated probability of honest answers is given in panel 3. It turns out that the randomized response conditions yield significantly more honest answers than self completion and direct answers, but the estimated probabilities of .43 and .49 are a bit disappointing. Due to time restrictions in preparing this workshop paper we now limit ourselves to a study of the relation of three covariates with the probability to answer honestly. The variables are sex of respondent (male = 1, female = 2), age (in years) and whether the respondent is born in the Netherlands or not (born in NL = 1, not born in NL = 2). Panel 1 shows for models 2, 3 and 4 the model fitting results for each covariate separately, and model 5 shows the results for the three covariates jointly. Panel 2 shows the parameter estimates for panel 5. Strictly speaking (i.e. comparing in panel 1 differences in chi-square between model 1 versus 2, 3 and 4 with 3.8) not many variables are significant: for forced response age is significant; for Kuk's procedure and self completion none of the variables is significant; for direct questions born in the Netherlands is significant.

However, the sample size is rather small for all of the conditions (especially for the randomized response conditions), and also, the parameter estimates for model 5 (see panel 2) show that all conditions show that the variables have effects into the same directions: for males answer more honestly, younger people answer more honestly, and people born in the Netherlands answer more honestly. In panel 3 the estimated proportions for combinations of sex, age (for two arbitrary ages, namely 30 and 50), and born in the Netherlands are shown. Born in the Netherlands is especially strong in direct questions, and randomized response procedures seem to be helpful in eliciting more honest answers from respondents not born in the Netherlands. Males are always more

Panel 1: model fitting, reported is -2 times likelihood ratio

	Foced response n=96	Kuk n=105	Self completion n=47	Direct n=99
1. Constant	133.04	145.55	45.91	111.89
2. 1 + sex	130.96	144.24	42.56	109.01
3. 1+ age	129.34	145.06	43.80	111.24
4. 1+ born in NL	133.01	145.47	45.89	100.71
5. 1+ sex, age, NL	127.01	143.64	40.52	98.27

Panel 2: Parameter estimates (with s.e. in parentheses) for model 5

	Foced response n=96		Kuk n=105		Self completion n=47		Direct n=99	
sex (m=1, f=2)	-.86	(.62)	-.79	(.73)	-1.47	(.69)	-.76	(.52)
age(in years)	-.06	(.03)	-.02	(.03)	-.06	(.04)	-.02	(.02)
born in NL (1=y, 2=n)	-.43	(.68)	-.28	(.69)	-.23	(.75)	-1.67	(.60)
costant	-.3.63	(2.13)	2.22	(1.95)	-3.07	(2.18)	2.85	(1.40)

Panel 3: Estimated proportions of honest answers

	Foced response n=96	Kuk n=105	Self completion n=47	Direct n=99
Model 1:	0.43	0.49	0.19	0.25
Model 5 (using panel 2):				
Men NL 30	.65	.63	.38	.47
50	.38	.54	.15	.39
not NL 30	.55	.57	.32	.14
50	.29	.47	.12	.10
Women NL 30	.45	.44	.12	.30
50	.21	.34	.03	.23
not NL 30	.35	.37	.10	.07
50	.15	.28	.03	.05

TABLE 1. Model fitting, parameter estimates (with standard errors in parentheses), and estimated proportions of honest answers under four conditions: a Forced response procedure, Kuk's randomized response procedure, Self completion and Direct questions

inclined to admit that they have committed fraud, but is especially the case in the self-completion condition.

5 Conclusion

The randomized response procedures employed show an important increase in the proportion of honest answers when the sensitive topic is a specific king of social security fraud. However, the proportions of honest answers still leave much to be desired.

The study reported here is a pilot study for the application of randomized response techniques for a nation-wide survey among people getting social security money. Since Kuk's procedure is leading to the highest number of honest answers, this procedure will be used in this nation-wide survey. Results for Kuk's procedure from this pilot study will be used to obtain correction factors for proportions of fraud obtained in the national survey. For example, the pilot study shows that male respondents more easily admit social security fraud than female responents . Therefore, in order to obtain realistic nation-wide fraud estimates, we will correct the proportion of males admitting fraud in the nation-wide survey with a smaller factor than the proportion of females. At this moment we are studying how to obtain such correction factors.

Appendix

The logistic regression model was defined in (1) and (2) in the text. In the derivatives shown below it was useful to note that

$$\frac{\partial \Pi_{1|i}}{\partial \beta_m} = -X_{im}\Pi_{1|i}\Pi_{2|i} \tag{6}$$

$$\frac{\partial \Pi_{2|i}}{\partial \beta_m} = X_{im}\Pi_{1|i}\Pi_{2|i} \tag{7}$$

The first and second derivatives for the likelihood of the standard logistic regression model defined in (3) were

$$\frac{\partial \log L}{\partial \beta_m} = -\sum_i n_{i1} X_{im}\Pi_{2|i} + \sum_i n_{i2} X_{im}\Pi_{1|i} . \tag{8}$$

$$\frac{\partial^2 L}{\partial \beta_m \partial \beta_n} = -\sum_i n_{i+}\Pi_{1|i}\Pi_{2|i}X_{im}X_{in} \tag{9}$$

The first and second derivatives for the likelihood of the forced response logistic regression model defined in (4) were

$$\frac{\partial \log L}{\partial \beta_m} = \sum_i P_3 X_{im}\Pi_{1|i}\Pi_{2|i}\left(\frac{n_{i2}}{P_i(yes)} - \frac{n_{i1}}{P_i(no)}\right) \tag{10}$$

$$\frac{\partial^2 LogL}{\partial\beta_m \partial\beta_n} = \sum_i P_3 X_{im} X_{in} \Pi_{1|i} \Pi_{2|i}$$
$$\left\{ (\Pi_{1|i} - \Pi_{2|i}) \left(\frac{n_{i2}}{P_i(yes)} - \frac{n_{i1}}{P_i(no)} \right) - P_3 \Pi_{1|i} \Pi_{2|i} \left(\frac{n_{i2}}{P_i(yes)^2} + \frac{n_{i1}}{P_i(no)^2} \right) \right\} \quad (11)$$

And the first and second derivatives for the likelihood of the Kuk's logistic regression model defined in (5) were

$$\frac{\partial LogL}{\partial\beta_m} = \sum_i X_{im} \Pi_{1|i} \Pi_{2|i} \left(\frac{n_{i1}(P_4 - P_3)}{P_1(black)} + \frac{n_{i2}(P_2 - P_1)}{P_i(red)} \right) \quad (12)$$

$$\frac{\partial^2 LogL}{\partial\beta_m \partial\beta_n} = \sum_i X_{im} X_{in} \Pi_{1|i} \Pi_{2|i} (\Pi_{1|i} - \Pi_{2|i})$$
$$\left(\frac{n_{i1}(P_4 - P_3)}{P_i(black)} + \frac{n_{i2}(P_2 - P_1)}{P_i(red)} \right) - \Pi_{1|i} \Pi_{2|i} \left(\frac{n_{i1}(P_4 - P_3)^2}{(P_i(black))^2} + \frac{n_{i2}(P_2 - P_1)^2}{(P_i(red))^2} \right) \quad (13)$$

References

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

Fox, J.A. and Tracy, P.E. (1986). *Randomized response. A method for sensitive surveys.* Beverly Hills: Sage, Quantitative Applications in the Social Sciences.

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

Maddala, G.S. (1983). *Limited dependent and qualitative variables in econometrics.* New York: Cambridge University Press.

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

van Gils, G. and van der Heijden, P. G. M. (1996). *Reporting social security fraudin Surveys.* Utrecht: University of Utrecht, Faculty of Social Sciences.

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

The det
correlati

Geert Verb
Emmanuel
Larry J. Bi

ABSTRACT:
nary least squ
lation structu
is extended to
intercepts, an
timore Longit

KEYWORDS
data.

1 Introdu

In medical scien
parameter whic
Such longitudin
ual changes ov
likely to influer
often be analys
However, such
observational st
are collected ca
and withdraw
number of time
vations may be

One possibl
e_i, with correl
rors are assum
is completed b
many covarianc
from the gene
Zeger (1994)).
distributed wit