US-China Education Review, ISSN 1548-6613
January 2011, Vol. 8, No. 1, 25-32

Logistic Regression Analysis of Employment Behavior Data Using Randomized Response Technique

Velo Suthar, Habshah Bt. Midi
University Putra Malaysia, Selangor, Malaysia

Naeem A. Qureshi
Sindh Agriculture University, Sindh, Pakistan

Direct survey techniques deal with collecting information on sensitive issues data, such as induced abortion, drug addiction, and so on. RR (randomized response) techniques are available for many interviewees, who do not feel comfortable to disclose their personal data due to privacy risks. RR techniques are used in the estimation of the number of people having a sensitive attribute say A. When the research is conducted on the disgraceful or ignominious characteristics of persons like rash driving, tax elusion, induced abortion, testing HIV (human immunodeficiency virus) positive etc., RR techniques are used to make sure that the estimates obtained are efficient and unbiased. During these types of surveys, privacy of the respondent is also managed. Among others, the conflict between efficiency and protection of privacy was also discussed by Nayak in 1994. In RR-related techniques, the SRS (simple random sampling) is statistically used in the sample selection. In this paper, RR procedure is used that allows us to estimate the population proportion in addition to the probability of providing a truthful answer. This study also quantifies a method for the estimation of the model having one variable (univariate) while studying logistic regression, where the dependent variables are subject to RR. In addition, an efficiency comparison is carried out to investigate the performance of the proposed technique. It is also assumed that during the study, the respondents will respond keeping in view the instructions of the RR design. The general idea about findings of current study, though, is so as to perform RR techniques comparatively fine.

Keywords: RR variable, logistic regression, AIC (Akaike information criterion), privacy risk, RR design, maximum likelihood, survey

Introduction

In the direct survey techniques of collecting information on sensitive issues data, such as induced abortion, drug addiction, and so on, many interviewees do not feel comfortable to disclose their personal data due to privacy risks. They sometimes refuse to provide information or might decide to give false data. Warner (1965) introduced an ingenious technique known as RR (randomized response) technique for estimating π_x, the proportion of population possessing certain stigmatized character x (say) by protecting the privacy of respondents, and preventing the unacceptable rate of non-response. Warner's (1965) technique was modified by Horvitz, Shah and Simmons (1967), Greenberg, Abul-Ela, Simmons and Horvitz (1969), Raghavrao (1978), Kim (1978), Franklin (1989), Mangat and R. Singh (1990), Kuk (1990), Tracy and Mangat (1995), Mangat, R.

Velo Suthar, Ph.D. candidate, Institute for Mathematical Research, University Putra Malaysia.
Habshah Bt. Midi, Institute for Mathematical Research, University Putra Malaysia.
Naeem A. Qureshi, Department of Statistics, Sindh Agriculture University.

Singh and S. Singh (1997), S. Singh (1999; 2003), Mahmood, S. Singh and Horn (1998) among other researchers for improving greater cooperation and efficiency. The present study mainly focuses on the development of the univariate logistic regression model when the dependent variable is expressed in terms of the RR. Similarly, in the present study, an attempt was made to asses the qualities, such as reliability and less time consumption of univariate logistic regression model, when the response variable is binary in nature.

Experimental analysis addicted to employment behavior is desired to appreciate the level and the source of the problem, but existing research is susceptible for the reason that it is difficult to obtain sensitive information about employment behavior from individuals. Many researchers, such as Lanke (1976), Leysieffer and Warner (1976), and Anderson (1976; 1977), and others have described how to maintain the conflict between efficiency and protection of privacy of respondents. The inability to directly observe individual employment benefit behavior is among the various limitations which researchers are facing in investigating employment behavior. It is usually observed that in the description of the employment behavior, individuals' self- reports/surveys provide the basis for the empirical evidence. Because of the receptive nature of the topic, surveys of employment behavior are much more complicated as compared to other type of surveys. Generally, employment behavior is perceived to be an illegal and socially undesirable behavior.

It is the nature of the people that they are reluctant to admit while availed the illegal employment benefit. The main reasons due to which the respondents either lie about their employment behavior (response bias) or to refuse to take part in the surveys and avoid answering sensitive questions (non-response bias) are the threat of penalties, court hearings and stigmatization. Getting reliable estimates of the employment benefits are difficult, because the response and non-response biases in the survey affect the validity and the generalized ability of the results. The investigators are in front of the troubles of how to give confidence to participants to respond, and then to provide truthful response in surveys (Nayak, 1994; Leysieffer & Warner, 1976). A suggested solution is the RR technique that was first developed by Warner (1965). There are two major advantages for which the RR technique was designed: (1) It reduces both response bias and non-response bias in surveys which ask sensitive questions; and (2) It uses probability theory to protect the privacy of an individual's response and this theory has been used successfully in several sensitive research areas, such as abortion, drugs and assault.

Numerous studies were carried out on RR technique for the last 4 decades and two and half decades on application of logistic regression for RR variables. The available literature on logistic regression for randomized response variables was briefly reviewed. Warner (1965) is the pioneer of RR technique that eliminates the response bias and protects the privacy of respondents in sensitive survey issues. Among several researchers, Horvitz, et al. (1967) modified a Warner's (1965) RR technique. Maddala (1983) was the first to present the likelihood of the model with respect to the RR design by Warner (1965). Scheers and Dayton (1988) discussed the model with respect to both the Warner model and the unrelated-question model (Greenberg, et al., 1969). Van der Heijden and Van Gils (1996) presented the model where the response variable was subject to either the RR design by Boruch (1971) or the RR design by Kuk (1990). A recent application of the univariate model was presented by Lensvelt-Mulders, Van der Heijden and Laudy in 2006. The link between RR and misclassification in the context of log linear models was described by CHEN in 1989. Magder and Hughes (1997) discussed the logistic regression model where the response variable was subject to misclassification comparable to the perturbation induced by the RR design. The question of the investigation of employment behavior had been raised several times from the employment behavior research community by using RR technique, but a review of literature showed that little work had been done. Motivated by the need to gather

more reliable and meaningful data on employment behaviors and to improve the research methods, a present survey was conducted to achieve the following objectives:

(1) To estimate the proportion and type of employment behaviors of individuals in two institutes of UPM (University Putra Malaysia);

(2) To assess the effectiveness of the RR technique in reducing response and non-response biases in surveys asking sensitive questions;

(3) To investigate the relationship between employment behaviors and key covariates variables.

Methodology

The Warner Model

The RR technique was introduced by Warner in 1965. This technique is used to protect the privacy of the respondents in the situation when the sensitive questions are being asked to them. This technique advocates as a useful tool in eliciting sensitive information, and respondents are reluctant to answer directly. The present technique works logically, i.e., the respondents are given two logically opposite questions and are instructed to answer one or the other depending on the outcome of a randomizing device. For example, $i \in \{1, \ldots, n\}$, let y^*_i denote the observed answer to the RR question, whether the respondent ever used a sick leave when he/she was not really sick (Yes $\equiv 1$, No $\equiv 0$). This could be done by asking the respondent to throw a dice, and the outcome of the dice determines which question they answer.

FRD (Forced Response Design)

It is an example of RR design and is used to answer the question that how design can be seen as a misclassification design. When the sensitive question is asked, the respondent throws the dice and keeps the outcome hidden from the interviewer.

If the outcome is 1 or 2, the respondent answers “Yes”.

If the outcome is 6, the respondent answers “No”.

If the outcome is 3, 4 or 5, the respondent answers according to the truth.

This design protects the privacy, since an observed “Yes” does not necessarily imply a latent “Yes”. When the respondent answers “Yes” or “No”, the researcher does not know whether the respondent is answering. Thus, the privacy of the respondent is protected. The use of probability theory allows the researcher to estimate the proportion of affirmative responses to answer “Yes” and the associated sampling variance using the following equations:

Suppose that “Y” represents the binary RR variable that models the sensitive item, “Y^*” the binary variable that models the observed answer, Yes=1 and No=0.

The RR design of the forced response design is given by:

$$P(Y^*=1)=P(Y^*=1|Y=0)P(Y=0)+P(Y^*=1|Y=1)P(Y=1)=1/3+1/2\ P(Y=1) \quad (1)$$

$$\lambda = p\pi + (1-p)(1-\pi) \quad (2)$$

Thus,

$$\pi = (\lambda + p - 1)/(2p-1) \quad (p \neq 0) \quad (3)$$

and

$$\mathrm{Var}(\pi)=[\pi(1-\pi)/n]+[p(1-p)/n(2p-1)^2] \quad (4)$$

where: π =the estimated proportion of "Yes" responses;

$\hat{\lambda}$ =the observed proportion of "Yes" responses;

p=the probability of answering Yes=1;

n=the sample size.

Note that probabilities $P(Y^*=j|Y=k)$ are fixed for $j, k \in \{0,1\}$ by the known distribution sum of the two dice. The conditional misclassification probabilities are given by $P(Y^*=j|Y=k)$

$$\pi_1^* = p_{00}\pi_1 + p_{10}\pi_0$$

In Matrix algebra:

$$\pi_0^* = p_{01}\pi_1 + p_{11}\pi_0$$

$$\begin{pmatrix}\pi_1^*\\ \pi_0^*\end{pmatrix} = \begin{pmatrix}p_{00} & p_{01}\\ p_{10} & p_{11}\end{pmatrix} = \begin{pmatrix}4/6 & 1/6\\ 2/6 & 5/6\end{pmatrix}\begin{pmatrix}\pi_1\\ \pi_0\end{pmatrix}$$

$$\Downarrow$$

$$\pi^* = P_W \pi$$

The Warner design is a MC (misclassification) design, where the MC is desired by a transition matrix (Kuha & Skinner, 1997).

The potentiality of the RR technique regarding the reduction of both response and non-response bias resulting from sensitive survey questions was claimed by Warner in 1965. But this design has a major limitation that it increases the variance of the estimator due to the introduction of the randomizing procedure into the design. Because of this exaggerated variance, Warner stressed the importance of using the RR technique only for sensitive issues, so as to offset the increased variance of the estimate with the lower mean square error produced by more truthful reporting.

Hypotheses Development

In this section, a questionnaire survey was reported which was conducted in 2008 at two institutes of UPM. The survey was concerned about the individual employment behavior using RR techniques.

Survey instrument was designed by using the RR technique to ask sensitive question. The assessment of RR technique about the effective reduction in the response and non-response biases in the surveys related to sensitive questions was done by testing the following two hypotheses:

$H_0$1: The proportion of "Yes" response rate will be higher than "No" response rate for individuals receiving the RR survey technique.

$H_0$2: The proportion of individuals admitting to use the sick leave, when respondent is not really sick will be higher as compared to those individuals admitting not to use sick leave, when respondent is not really sick.

Use of RR technique reduced the response bias was tested in the above hypotheses. It is based on the assumption that a higher proportion of respondents admitting to use the sick leave, that indicates more truthful reporting. Five more hypotheses ($H_0$3, $H_0$4, $H_0$5, $H_0$6 and $H_0$7) were also developed for the investigation of the relationship between employment behavior and 5 covariates (x_{1i}=age in years, x_{2i}=gender (male=1, female=0), x_{3i}=place of grow up (village=0, city=1), x_{4i}=current job position in the institute and x_{5i}=job experience in years) were used in the model.

Protection of Privacy

Chaudhuri, Tasos and Amitava (2008) reported that numerous studies confined to SRSWR (simple random sampling with replacement) of the respondents from a population (Lanke, 1976; Leysieffer & Warner, 1976; Anderson, 1976, 1977; Nayak, 1994; Warner, 1965; Kuk, 1990; Greenberg, et al., 1969; Horvitz, et al., 1967; Mangat & Singh, 1990; Christofides, 2003). Under SRSWR, let P(Y=1)=θ=Y/N=P(A) be the probability that a respondent chosen from "U" at random bears the sensitive attribute "A". Denoting by "Yes" and "No" as the response announcing that the "card type" or the outcome of a random device "matches" or "mis-matches" respectively the respondent's attribute "A" or "A^c", it was possible for Nayak (1994) to generalize the works of his predecessors as follow:

Letting P(Yes|A)=a and P(No|A^c)=b, Nayak (1994) noted that,

$$P(A|\text{yes}) = \frac{\theta a}{\theta a + (1-\theta)(1-b)}, \quad P(A|\text{no}) = \frac{\theta(1-a)}{\theta(1-a) + (1-\theta)b},$$

On applying Bayes' theorem, since P(A)=θ and P(Ac)=1−θ.

Departures of P(A|Yes) from θ and P(A^c|No) from 1−θ could be treated as measures of revelation of secrecy. Treating R as a response "Yes" or "No",

$$P(A|R) = \frac{\theta P(R|A)}{\theta P(R|A) + (1-\theta)P(R|A^c)}$$

and

$$P(A^c|R) = \frac{(1-\theta)P(R|A^c)}{(1-\theta)P(R|A^c) + \theta P(R|A)}$$

could be respectively regarded as "revealing probabilities" in announcing R about a person's response concerning A or A^c.

If P(A|R)>θ, R is jeopardizing with respect to A and if P(A^c|R)>(1−θ), then R is jeopardizing with respect to A^c. Combining these two,

$$J(R) = \frac{P(A|R)/\theta}{P(A^c|R)/(1-\theta)}$$

is treated as a "measure of jeopardy" inherent in a response R concerning A or A^c.

The higher its value is, the more its deviation from "unity".

The previous discussion was applied to the case of selecting the respondents with simple random sampling with replacement. To cover sampling of respondents from "U" by arbitrary probabilities we proceeded in the following way as partially introduced by Chaudhuri and Saha (2004).

Suppose that $L_i(0<L_i<1)$ is the probability that y_i takes on the value "1" for the unit labeled i and that $L_i(R)$ denotes the "conditional probability" that the "ith" respondent has the stigmatizing characteristic given that his/her randomized response is R. Then

$$J_i(R) = \frac{L_i(R)/L_i}{[1-L_i(R)]/(1-L_i)}, i \in U$$

will be defined as the "response-specific" and "jeopardy measure" for the RR obtained as "R" from the respondent "i" on adopting a specific RR technique. This measure depends on the specific response of the participant. However, since a measure of jeopardy quantifies the risk of revealing his/her status (i.e., whether he/she belongs to the stigmatizing group) which a person undertakes by agreeing to use the randomization device, it should be made known to the participants before they agree to participate in the survey, i.e., before any response is available. It is therefore justified to use a measure which is not response-specific but rather

could be regarded as a technical characteristic of the device.

Research Design

The RR technique was designed and used to ask the sensitive questions about employment behavior, the dice/instrument offered respondents the protection of anonymity. Use of randomized procedure had increased the protection of the RR technique's respondents. For the selection of the most suitable RR design, few decisions were made. The encoding technique for dependent variable is presented in Table 1, the original value "Yes" is coded with internal value 1, and the original value "No" is coded with internal value 0.

Table 1

Dependent Variable Encoding

Original value	Internal value
No	0
Yes	1

Randomizing Procedure

The success of RR technique entirely depends upon the randomization process. By the application of probability theory, respondents' replies of sensitive questions were being protected. Even the researchers themselves are unknown about the answers of the respondents. The most common randomizing device used in RR survey questionnaire by respondents is random number dice supplied by a researcher. Probability of answering the sensitive question, "p", is important parameter that has an impact on the variance of the estimate. There exists an inverse relationship between level of probability and the protection offered to the respondents, i.e., the smaller the level of "p" (i.e., the fewer respondents who are instructed by the randomizing device to respond to the sensitive questions), the greater the protection offered to the respondents. However, this also means that the sampling variance of the estimator will increase.

Due to the above-mentioned reasons, the investigators working on RR methodology make use of the "p" large enough to avoid the sampling error when the samples are very small (Lanke, 1975). Soecken and Macready (1982) recommended that "p" be chosen between 0.7 and 0.85 to obtain sufficient efficiency in the design and still protect the privacy of respondents. We struck a balance between respondent jeopardy and estimation efficiency in the study.

Survey Procedure

Since it is required by the randomized response technique that the selected samples must be the large for effective data analysis, a well defined questionnaire was prepared for the survey. The present survey was conducted on the small scale in order to obtain a representative sample and as well as pilot study of UPM. The people working in two different institutes of the UPM were selected as the target population. Total of 48 respondents were selected for the present survey. The larger sample for the RR instrument was meant to compensate for the inflated sampling variances caused by the randomizing procedure. The survey was started by asking the screening questions as an instrument. Prior to the survey instrument, the researchers explained to the respondents the purpose and the procedure of the study and encouraged them to participate in the study.

Application

The demonstration of the present survey is done by analyzing the data regarding employment behavior to avail sick leave at two institutes of UPM in 2008. A sample size of n=48 was used in this survey. For example, $i \in \{1, \ldots, n\}$, let y_i denote the observed answer to the RR question, whether the respondent ever used a sick leave when he/she was not really sick (Yes=1, No=0). One of the limitation of the analysis is that for the consideration of the main effect model from the original survey, only limited number of variables were selected and analyzed. In addition, x_{1i} is age in years (quantitative), x_{2i} is gender (male=1, female=0), x_{3i} is place of grown up (village=0, city=1), x_{4i} is current job position in the institute and x_{5i} is job experience in years (quantitative). The RR design is given by p_{00}=0.813 and p_{11}=0.933 and it is a slightly adapted form of the FRD (forced response design).

Classification of Cases

One method of assessing the success of a model is to evaluate its ability to predict correctly the outcome category for cases for which outcome is known. If a respondent has identified as he/she availed a sick leave, for instance, it can be seen if a respondent is correctly classified as sick on the basis of other predictor variables. Table 2 shows the classification of respondents of 42 individuals. It is obvious from the Table 2 that 17 (80% age of the total cases) respondents were in support and they were correctly predicted by the model. Likewise, 16 (77.3% age of the total cases) individuals are in underneath employment behavior and the model correctly predicted them. On an overall basis, the proposed model could correctly classified about 78.6% age of the total cases.

Table 2

Classification (a)

Observed		Predicted		
		Sick leave		Percentage correct (%)
		No	Yes	
Sick leave	No	16	4	80.0
	Yes	5	17	77.3
Overall percentage (%)				78.6

Note. (a) The cut value is 0.50.

The proposed model is comparatively better than that of Lawson, et al. (2004) who developed the logistic regression model and correctly predicted about 51% of total cases while the logistic model of the present study correctly classified about 78.6% of cases. At the same time, the proposed model is slightly less efficient than Suthar, Pasha and Lohano (2007), they expanded the logistic model and correctly predicted 81.5%. Unlike diagonal entries, the off-diagonal entries indicate the incorrectly classified individuals by the model. There were 4 individuals not avail the sick leave, but they were incorrectly predicted "Yes" by the model. Five cases were incorrectly classified as avail the sick by the model while in real situation they were recorded to be "No". On an overall basis, there were 21.4% of the cases that were incorrectly classified by the proposed model.

Univariate Regression

The analysis of univariate logistic regression is presented in Table 3. The AICs' (Akaike information criterion) value consists of the output of the generalized linear model procedure in "R-code" statistical software, whereby, this procedure is adapted for RR variable (Ardo, Peter, & Robert, 2007). The variables X_3 and X_5 are significant while other variables are not significant at 5% level of confidence. This variable suggests, covariate

X_3 (place where respondent was grown up) and X_5 (job experience in years) are important in modeling the question of whether respondents ever availed the sick leave when he/she was not really sick. The covariate X_1 (age) is significant at 10% when respondent change the level from 0 to 1 ("No=0" to "Yes=1").

Table 3

Analysis of Univariate Logistic Regression of the Unemployment Data

Full model: Logit ($P(Y=1\|X)=\beta$ t xi				
Coefficients: (R-code)				
Variables	Estimate	Std. error	z value	Pr(>|z|)
(Intercepte)	8.7650	5.5900	1.568	0.1169
X_1	-0.3316	0.1903	-1.743	0.0814*
X_2	-1.2664	1.1894	-1.065	0.2870
X_3	3.1587	1.5562	2.030	0.0424**
X_4	-1.0337	0.7043	-1.468	0.1422
X_5	0.4548	0.2225	2.044	0.0410**

Notes. Null deviance: 58.129; Residual deviance: 46.230; AIC: 58.23.

Conclusion

The current research was attempted to estimate univariate logistic regression model where the randomized response variables are used as dependent variables. The framework is quite general in the sense that the methods can deal with various RR designs and that the RR regression models have the same flexibility as the standard regression models (Ardo, et al., 2007). The application section shows that once the methods are implemented, analyzing RR data and interpreting the results resemble the analyses and interpretations in the situations without RR. The general idea, however, is that RR performs relatively well (Lensvelt-Mulders, Heijden, & Laudy, 2006). Future RR surveys may profit from research into cheating with respect to the RR design (Böckenholt & Heijden, 2007). An important assumption in the paper is that respondents comply in the sense that they follow the instructions of the RR design. This assumption will not always be justified. For instance, it might be that some respondents do not trust the privacy protection offered by the RR design and give a socially desirable answer anyway.

References

Aitkin, M., Francis, B., & Hinde, J. (2005). *Statistical modeling in GLIM4* (2nd ed.). Oxford: Oxford University Press.

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

Chaudhary, A., & Saha, A. (2004). Extending, sitter's mirror-match, bootstrap to cover, Rao-Hartley-Cochran: Sampling in two-stages with simulated illustration. *The Indian Journal of Statistics, 66*(4), 791- 802.

Kim, J. I. (1978). *Randomized response technique for surveying human populations* (Doctoral dissertation, Temple University, 1978).

Kuha, J. T., & Skinner, C. J. (1997) Categorical data analysis and misclassification. In L. Lyberg et al (eds.), *Survey measurement and process quality*. New York: Wiley, 633-670.

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

Mahmood, M., Singh, S., & Horn, S. (1998). On the confidentiality guaranteed under randomized response sampling: A comparison with several new techniques. *Biom. J., 40*(2), 237-242.

Scheers, J., & Dayton, C. M. (1988). Covariate randomized response models. *J. Amer. Statist. Assoc., 83*, 969-974.

Singh, S. (2003). *Advanced sampling theory with applications: How Micheal "selected" Amy*. The Netherlands: Kluwer Academic Publishers.

Soeken, K. L., & Macready, G. B. (1982). Respondents' perceived protection when using randomized response. *Psychological Bulletin, 92*(September), 487-489.

Van der Heijden, P. G. M., & Van Gils, G. (1996). Some logistic regression models for randomized response data. In A. Forcina, G. M. Marchetti, R. Hatzinger, & G. Falmacci (Eds.), Proceedings of *the 11th International Workshop on Statistical Modeling*.