Estimation Of Binary Logistic Regression Parameters In Sensitive Surveys For A Two-Stage Randomized Response Technique

Neelam[1], Syed Muhammad Asim[2], Soofia Iftikhar[3], Balqis Khalil[4], Safia Murad[5], Mehwish Dalail[6], Said Farooq Shah[7]

[1,2,7]Department of Statistics, University of Peshawar, Pakistan
[3]Department of Statistics, Shaheed Benazir Bhutto Women University Peshawar, Pakistan
[4]Department of Statistics, Islamia College Peshawar, Pakistan
[5]Sarhad Institute of Health Sciences, Sarhad University of Science and Information Technology, Pakistan
[6]Sarhad University of Science and Information Technology, Pakistan
Email address in order: [1]neelam_arbab@uop.edu.pk, [2]smasim@uop.edu.pk,
[3]soofia.iftikhar@sbbwu.edu.pk, [4]bilqeeskhalil@gmail.com,
[5]safia.murad55@gmail.com, [6]dmehwish310@gmail.com, [7]farooqshahqau@uop.edu.pk

Abstract

When collecting data on sensitive topics such as abortion, harassment, tax evasion, and income, respondents often provide untruthful answers, leading to biased results. To address this issue, Warner introduced the Randomized Response Technique (RRT) to obtain sensitive information while ensuring respondent privacy. Building on this, Narjis and Shabbir proposed a two-stage RRT to estimate the prevalence of a sensitive attribute. This study extends their two-stage RRT by incorporating covariates through logistic regression to analyze their effect on the sensitive characteristic. Parameters of the logistic regression model are estimated under simple random sampling, and the performance of maximum likelihood estimators is evaluated through simulation.

Keywords Randomized response technique, Sensitive survey, Logistic regression, Maximum likelihood estimation, Sensitive characteristic

Introduction

In various fields, especially in the social and behavioral sciences, the primary goals of research are often connected to understanding and measuring human attitudes and behaviors. Self-reported data play a crucial role in these studies; knowing the rates, proportions, correlations, and causes and effects of specific behaviors or attitudes allows researchers to uncover underlying theories and, when appropriate, suggest practical solutions. This data is typically gathered through a structured process known as a survey. When only a subset of the population is examined systematically, it is referred to as a sample survey, whereas a census involves collecting information from every individual within the entire population.

When examining population characteristics, sample surveys are typically conducted using common methods such as questionnaires, observations, or interviews. While surveys can yield a wealth of information from respondents on various topics, they can become particularly challenging when addressing stigmatized behaviors. In other words, collecting data on sensitive subjects can be quite difficult and almost impossible, if the attribute or variable is highly sensitive. Nevertheless, there are solutions available to address this issue. In 1965, Warner [1] for the very first time coined a method or procedure for dealing with sensitive attributes. Sensitive attributes refer to the collection of data or information related to stigmatized or undesirable variables. For instance, gathering information on topics such as tax evasion, abortions, smuggling, cheating, fraud, or any other undesirable variables can negatively affect how individuals are

perceived by others. Collecting such data is quite challenging; therefore, Warner [1] developed the Randomized Response Technique (RRT) to gather information on sensitive topics. Direct questioning about these issues often fails to yield accurate data, but employing this technique can significantly enhance the collection of sensitive information. Randomization is a specific approach in which questions about sensitive attributes are posed in a way that ensures respondents' confidentiality, making it impossible for the interviewer to know their answers. This sense of confidentiality and anonymity can encourage respondents to feel more secure in providing honest answers through a randomization process.

The randomization procedure, Warner [1] has adopted, is to first divide the overall population into two different parts, the sensitive group and the non-sensitive group. Then, using a randomization procedure and some pre-determined probabilities/proportions of the selection of questions, data is gathered from the respondents. Respondents are required to randomly choose between two questions and then answer the selected question with a simple yes or no. The confidentiality of their responses is maintained, as the interviewer is unaware of which question the respondent answered. Consequently, RRTs enhances confidentiality, leading to higher response rates since respondents feel more at ease when providing answers. This approach significantly increases the likelihood of obtaining unbiased responses and maximizing participation.

Later on, a lot of work has been done on RRTs in the literature by different scholars. The first modification was done by Greenberg et al. [2] in 1969 followed by another in 1972 in which Greenberg et al. [3] extended the work to quantitative RRTs which are very suitable for those surveys that are conducted by mail. Chow and Liu [4], Goodstadt and Gruson [5], Eichhron [6], Scheers and Dyton [7], Chaudhry and Mukerjee[8], Mangat and Singh [9], Kuk [10], Mangat et al. [11], Bar lev et al. [12], Haung [13], Kim et al. [14], Saha [15], Singh and Greewal [16], Blair et al. [17], Lee et al. [18], Shah et al. [19], Hsieh et al. [20], Narjis and Shabbir [21] and many more worked on RRT, which is very useful as RRTs not only helps in estimation of unknown population parameters of sensitive variables but also helps in estimating and ensuring honest responses. Given that data privacy is a significant concern today and can lead to biased results, it cannot be overlooked. This is the primary reason RRTs are utilized across various fields where sensitive data is involved, whether it pertains to simple sensitive questions, quality control, paired comparisons, or regression parameters. Shah et al. [22] applied RRTs in the field of quality control by introducing masking to ensure privacy and reduce data falsification. They employed Shewhart-Control charts for the masked data, using Average Run Length (ARL) as the performance measure in their study. The findings indicated that the BL control charts were effective under RRT designs for detecting smaller shifts, while the GB charts under the unrelated question model performed better for larger shifts.

The RRTs was also extended to social desirability bias in paired comparison experiments by Shah et al. [23]. Bayesian analysis was used for the estimation purpose in their research article. As a result, RRTs have found applications in nearly every sensitive field. However, until 1983, a number of works were done on RRTs but all RRTs were only designed to estimate a certain proportion/mean of people of certain specific sensitive characteristics. None of the researchers had presented the idea for the analysis of determinants of these factors. In 1983, Maddala [24] was the first to introduce the concept of measuring the effect of covariates on randomized response variables. That is, to measure the effect of exogenous variables on the true proportion π of the sensitive variable.

In 1988, Sheers and Dyton [7] proposed a covariate extension to Warner's [1] and Greenberg's [2] unrelated RRT by utilizing a logistic regression model to explore the relationship between the sensitive population proportion and its covariates. Hsieh-et-al [25, 26] developed a semi-parametric method to estimate the parameters of logistic regression in case of missing covariates in RRTs.

In 2021, Narjis and Shabbir [21] presented a two-stage RRT for estimating the proportion of a sensitive characteristic. In their RRT, they utilize two randomization devices, R_1 and R_2. In the first stage, the choice

is given to the respondents either to answer directly to the sensitive question or utilize the second randomization device, where, the sensitive question is asked using a specific randomization tool. The model was compared with Warner [1], Mangat-et-al [11], Bhagava and Singh [27], and Gupta-et-al [28] and found more efficient than all of these models.

In the proposed research article, a covariate extension of the two-stage RRT of Narjis and Shabbir [21] is considered in order to find the effect of covariates on the sensitive characteristics in logistic regression model. A lot of RRTs have been proposed over time yet a gap persists in the literature. An effective method is introduced to estimate the population proportion with sensitive characteristics through the estimates of logistic regression parameter.

Review of RR Techniques

The RRT of Warner [1] and some others are discussed briefly.

Warner RRT:

Warner [1] proposed RRT for the first time to deal with the survey of sensitive questions. He is the pioneer of the RRTs. In his survey, he designed the questions such that both the questions are the complement of each other. The respondents are assured that the interviewer does not know which question they are answering, thereby ensuring confidentiality.

The questions are like:

Do you have the sensitive characteristic?, or

Do you not have the sensitive characteristics?

with probability of selection P and 1-P respectively, using a specific randomization tool. Let π represent the population proportion of people belonging to the sensitive group, then the unbiased estimate of population proportion is given in equation (1) as,

$$\hat{\pi}_w = \frac{\hat{\theta} - (1-P)}{2P-1} \tag{1}$$

where $\hat{\theta} = \frac{n_1}{n}$ the proportion of yes responses in the sample.

with variance given in equation (2),

$$Var(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{P(1-P)}{n(2P-1)^2} \tag{2}$$

Narjis and Shabbir RRT

Narjis and Shabbir [21] proposed a two-stage RRT for estimating the proportion of people who have the sensitive characteristic. They made use of two randomization devices R_1 and R_2. In the first stage, R_1 the first randomization device consists of the following two statements.

- Do you have the sensitive characteristic *A*?
- Proceed to the second randomization device with the probability *F* and 1-*F* respectively.

That is, the choice is given to the respondents at the first stage to either answer directly to the sensitive question or go to the second randomization device where the sensitive question is asked indirectly using a randomization tool. The second randomization device consists of the following questions.

1. Do you have the sensitive characteristic *A*?

2. Do you not possess the sensitive characteristic *A*?

3. Draw another card

with the probability of selection of questions 1, 2, and 3 as P_1, P_2 *and* P_3 respectively.

where, $\sum_{i=1}^{3} P_i = 1$.

If the third card is selected, then the individual is suggested to re-follow the procedure without substituting the card. In the repetition process, if the third card is selected another time, then the individual is asked to reveal his/her true status.

Now, the probability of getting a yes answer is given by

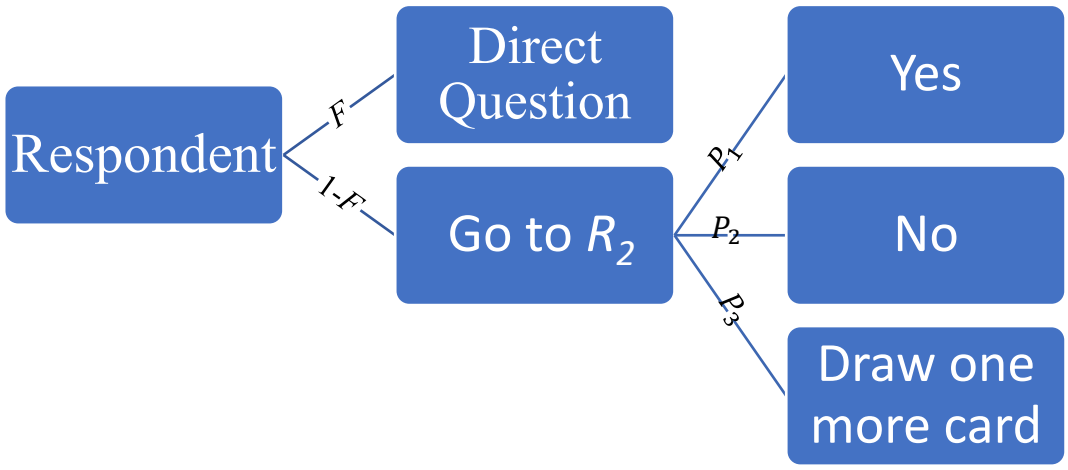
$$\theta_N = F\pi + (1-F)\left[P_1\pi + P_2(1-\pi)\left(1 + P_3\frac{k}{(k-1)}\right) + P_3^2\frac{k}{(k-1)}\pi \right] \tag{3}$$

where in equation (3), k is the total number of cards in the second randomization device.

Now the unbiased estimate of π , the proportion of sensitive group is given in equation (4)

$$\hat{\pi}_N = \frac{\hat{\theta}_N - (1-F)P_2\left(1 + P_3\frac{k}{(k-1)}\right)}{\left[F + (1-F)\left\{(P_1 - P_2) + P_3\frac{k}{(k-1)}(P_1 - P_2 + P_3)\right\}\right]} \tag{4}$$

with variance given in equation (5)



$$Var(\hat{\pi}_N)=\frac{\pi(1-\pi)}{n}+\frac{\pi(1-F)P_3\left(1-\frac{k}{(k-1)}\right)}{n\left[F+(1-F)\left\{(P_1-P_2)+P_3\frac{k}{(k-1)}(P_1-P_2+P_3)\right\}\right]}+\frac{\left[(1-F)P_2\left(1+P_3\frac{k}{(k-1)}\right)\right]\left[1-(1-F)P_2\left(1+P_3\frac{k}{(k-1)}\right)\right]}{n\left[F+(1-F)\left\{(P_1-P_2)+P_3\frac{k}{(k-1)}(P_1-P_2+P_3)\right\}\right]^2} \quad (5)$$

Fig 1: Probabilities under two-stage RR design of Nargis and Shabbir (2021)

Maddala approach to estimating the Logistic Regression Model in Sensitive Surveys:

Maddala [22] in 1983 proposed a method of employing the logistic model to the sensitive data to explore the relationship between the auxiliary variable and randomized response survey data using Warner's [1] randomized response data. For instance, in the case of a sensitive question such as, "Do you cheat on others?", responses are obtained using a RRT, alongside various independent or exogenous variables X, including age, gender, education, religion, parental income, parental education, geographical area, and personal income, among others. It is needed to find the effect of each of these explanatory variables on cheating.

Now, there are two cases. First, when there is no randomization. In this case, simply a logistic model is used, as the response variable is binary in nature, so logistic regression will be used.

For example, let

y = 1 If the person cheats on others, and
y = 0 Otherwise

Thus, the MLE for estimating the β is given below,

$$L=\prod_{y_i=1}P(y_i=1)\prod_{y_i=0}P(y_i=0) \quad (6)$$

where,

$$P\left(y_i = 1\right) = \frac{e^{\beta' x}}{1 + e^{\beta' x}}$$

so, equation (6) becomes

$$L = \prod_{y_i=1} \frac{e^{\beta' x}}{1 + e^{\beta' x}} \prod_{y_i=0} \frac{1}{1 + e^{\beta' x}} \tag{7}$$

However, in the presence of RRT, it will be a bit different, like a randomization procedure will be involved for getting responses from the respondents. For more understanding, let there be different colors of balls in a box and let the randomization procedure be as follows.

Let the box contain black, yellow, and brown color of balls. The respondent has to select a ball from the box and answer accordingly.

- Suppose if a black ball is selected the respondent has to answer one.
- If a yellow ball is selected, he/she has to answer zero.
- However, if a brown ball is selected then the respondent has to answer the question,

"Do you cheat on others?"

He/she has to answer 1 if he cheats and 0 otherwise. Let the selection of black, yellow, and brown balls have predetermined probabilities with P_1, P_2 *and* $1 - P_1 - P_2$ respectively. Thus,

$$P\left(y_i = 1\right) = P\left(yes\ response\right) = P_1 + \left(1 - P_1 - P_2\right)\pi \tag{8}$$

and,

$$P\left(y_i = 0\right) = P\left(no\ response\right) = P_2 + \left(1 - P_1 - P_2\right)\left(1 - \pi\right) \tag{9}$$

where,

$$\pi = \frac{e^{\beta' x}}{1 + e^{\beta' x}} \tag{10}$$

Using equation (8), (9) and (10), equation (6) becomes

$$L = \prod_{y_i=1} \left(P_1 + \left(1 - P_1 - P_2\right) \frac{e^{\beta' x}}{1 + e^{\beta' x}} \right) \prod_{y_i=0} \left(P_2 + \left(1 - P_1 - P_2\right) \frac{1}{1 + e^{\beta' x}} \right) \tag{11}$$

(The probability of cheating others has an "i" subscript because the model assumes the probability of cheating others, is dependent on characteristic *X*, so different individuals will have different probabilities of cheating others).

Suppose,

$$R_1 = \frac{P_1}{\left(1 - P_1 - P_2\right)}$$

$$R_2 = \frac{P_2}{\left(1 - P_1 - P_2\right)}$$

$$W_i = R_1 + \left(1 + R_1\right)\mathrm{e}^{\beta' x}$$

$$Z_i = 1 + R_2 + R_2\mathrm{e}^{\beta' x}$$

Then, equation (11) becomes,

$$\log L = constant + \sum\left[y \log W_i + \left(1 - y\right)\log Z_i - \log\left(1 + \mathrm{e}^{\beta' x}\right)\right] \tag{12}$$

where,

$$\frac{\partial \log L}{\partial \beta} = 0$$

Differentiating equation (12) w.r.t to *β, we get*

$$\sum\left[\frac{y\left(1 + R_1 + R_2\right)\mathrm{e}^{\beta' x}}{W_i Z_i} - \frac{\mathrm{e}^{\beta' x}}{Z_i\left(1 + \mathrm{e}^{\beta' x}\right)}\right] X_i = 0 \tag{13}$$

and,

$$\frac{\partial^2 \log L}{\partial\beta\partial\beta'} = \sum\left[\frac{y\left(1 + R_1 + R_2\right)\mathrm{e}^{\beta' x}}{W_i Z_i}\left(\frac{1 + R_2}{Z_i} + \frac{R_i}{W_i} - 1\right) - \frac{\mathrm{e}^{\beta' x}}{1 + \mathrm{e}^{\beta' x}}\frac{1}{Z_i}\left(\frac{1 + R_2}{Z_i} - \frac{\mathrm{e}^{\beta' x}}{1 + \mathrm{e}^{\beta' x}}\right)\right] X_i X_i^{'} \tag{14}$$

By using equation (**13)** and **(14),** we can get the estimates of β using Newton Raphson Method.

Thus, through this method, one can estimate the effect of exogenous variable *X* on the true probability π for the occurrence of a sensitive variable.

Method

This research article aims to explore the relationship between the covariates and randomized response variables using the logistic regression model. A covariate extension of a two-stage RRT of Nargis and Shabbir [21] is presented. It is aimed to estimate the parameters of logistic regression for the covariates which affect the sensitive attribute considering the simple random sampling with replacement only and simulation work is performed to explore the sample performances of maximum likelihood estimators of parameters of logistic regression.

Proposed Work

Model and Estimation

Let Y_1 represents the response of the respondent in first stage to a direct question where

- $Y_1 = 1$ if the respondent answer yes and
- $Y_1 = 0$ if the respondent answer no

Let Y_2 represent the respondents answer in the second stage to the sensitive question asked indirectly using RRT. $Y_2 = 1$, if the respondent answer yes $Y_2 = 0$ if the respondent response is no to the sensitive question asked indirectly using RRT in the second stage. Let we model the probability of an observed 'Yes' response Y that accounts for both (i) direct 'Yes' responses from the first stage device Y_1 and (ii) 'Yes' responses

obtained through the second stage RRT Y_2. Let X represents the vector of covariates. The logistic regression model for the sensitive attribute is represented by $\theta(X;\beta)$ given by

$$\theta(X;\beta) = G(\beta^T \chi) = \frac{1}{1+e^{-\beta^T \chi}} \tag{15}$$

where $\chi = (1, X^T)^T$. Let $h(\beta)$ represents the probability of yes response, then,

$$P(Y=1|X) = h(\beta) = M\theta(X;\beta) + (1-M)\left[P_1\theta(X;\beta) + P_2(1-\theta(X;\beta)) + p_3\left(\frac{k}{k-1}\right)P_1\theta(X;\beta) + P_3\left(\frac{k}{k-1}\right)P_2(1-\theta(X;\beta)) + P_3\left(\frac{k}{k-1}\right)P_3\theta(X;\beta)\right] \tag{16}$$

or

$$h(\beta) = M\theta(X;\beta) + (1-M)\left[\left\{P_1\theta(X;\beta) + P_2\left(1-\theta(X;\beta)\right)\right\}\left\{1+P_3\left(\frac{k}{k-1}\right)\right\} + P_3^2\left(\frac{k}{k-1}\right)\theta(X;\beta)\right] \tag{17}$$

The probability of yes response is represented by $h(\beta)$ and probability of no response is represented by $1-h(\beta)$. Since Y represents the yes response, so, $1-Y$ represents the no response. Let Y has a binomial distribution i.e, $(Y/X) \sim binom[1, h(\beta), 1-h(\beta)]$. Let $\{(Y_i, X_i) : i = 1,2,3,\ldots\ldots,n\}$ is a random sample.

The likelihood function is $L(\beta) = \prod_{i=1}^{n}\left(h(\beta)\right)^{Y_i}\left(1-h(\beta)\right)^{1-Y_i}$ and

$$\log L(\beta) = \sum_{i=1}^{n}\left[Y_i \log h_i(\beta) + (1-Y_i)\log\left(1-h_i(\beta)\right)\right]$$

Thus the score function is given by

$$U_n(\beta) = \frac{\partial}{\partial\beta}\log L(\beta)$$

$$= \sum_{i=1}^{n}\left[\frac{Y_i}{h_i(\beta)}\frac{\partial}{\partial\beta}h_i(\beta) - \frac{1-Y_i}{1-h_i(\beta)}\left(\frac{\partial}{\partial\beta}h_i(\beta)\right)\right]$$

$$= \sum_{i=1}^{n}\Psi_i(\beta) \tag{18}$$

where

$$\Psi_i(\beta) = \frac{Y_i}{h_i(\beta)}\frac{\partial}{\partial\beta}h_i(\beta) - \frac{1-Y_i}{1-h_i(\beta)}\left(\frac{\partial}{\partial\beta}h_i(\beta)\right) \tag{19}$$

for

$$\frac{\partial}{\partial\beta}\,h_i(\beta)=\frac{\partial}{\partial\beta}\left[M\theta(X;\beta)+(1-M)\left[\left\{P_1\theta(X;\beta)+P_2\left(1-\theta(X;\beta)\right)\right\}\left\{1+P_3\left(\frac{k}{k-1}\right)\right\}+P_3^2\left(\frac{k}{k-1}\right)\theta(X;\beta)\right]\right]$$

$$=M\theta^1(X;\beta)\left(0,\chi^t\right)^t+(1-M)\left[\left\{P_1\theta^1(X;\beta)\left(0,\chi^t\right)^t-P_2\theta^1(X;\beta)\left(0,\chi^t\right)^t\right\}\left(1+P_3\left(\frac{k}{k-1}\right)\right)+\left\{P_3^2\left(\frac{k}{k-1}\right)\theta^1(X;\beta)\left(0,\chi^t\right)^t\right\}\right]$$

$$=\theta^1(X;\beta)\left(0,\chi^t\right)^t\left[M+(1-M)\left\{\left(P_1-P_2\right)\left(1+P_3\left(\frac{k}{k-1}\right)\right)+\left(P_3^2\left(\frac{k}{k-1}\right)\right)\right\}\right] \tag{20}$$

(

with $\theta^1(X;\beta)=\theta(X;\beta)\left(1-\theta(X;\beta)\right)$. To derive easily the large sample properties of the ML estimators of β, we simplify the expression of $\psi_i(\beta)$ in (19) in terms of $h_i(\beta)$, Y_i, and the first derivatives of $h_i(\beta)$ with respect to β in Lemma 1.

Lemma 1:

$\psi_i(\beta)$ in (19) can be rewritten as

$$\Psi_i(\beta)=\left(\frac{\partial}{\partial\beta}h(\beta)\right)V_i^{-1}(\beta)\left[Y_i-h(\beta)\right],i=1,2,...,n \tag{21}$$

Where $h(\beta)$ is the predicted probability for $Y_i=1$.

$$V_i(\beta)=Var(Y_i)=h(\beta)(1-h(\beta)) \tag{22}$$

The proof of Lemma 1 is provided in the appendix. The large-sample properties of the ML estimators are given in the next section.

Large sample Properties

Consider the estimating function $U_n(\beta)$ as follows.

$$U_n(\beta)=\sum_{i=1}^{n}\psi_i(\beta)$$

Where $\psi_i(\beta)$ is given in (19). To investigate the large sample properties of $\hat{\beta}$, the following regularity conditions are required.

1. The expected value of the second derivative (Fisher information) is finite and positive in a neighborhood of the true parameter β_0. This ensures that the score function provides sufficient curvature to identify the parameter:

$$E\left[\left(\frac{\partial \psi_i(\beta)}{\partial \beta}\right)^2\right] > 0,$$

where $\psi_i(\beta)$ is the score function.

2. The score function $\psi_i(\beta)$ has finite variance, ensuring stability in estimation:

$$E\left[\psi_i(\beta)^2\right] < \infty.$$

3. The first and second derivatives of the log-likelihood function with respect to β exist almost surely and are uniformly bounded in a neighborhood of the true β_0:

$$\left|\frac{\partial l_i(\beta)}{\partial \beta}\right| < M_1 \text{ and } \left|\frac{\partial^2 l_i(\beta)}{\partial \beta^2}\right| < M_2,$$

where M_1 and M_2 are finite constants.

The asymptotic properties of $\hat{\beta}$ are stated in Theorem 1.

Theorem 1 Under the regularity conditions (1), (2), and (3),

I. $\hat{\beta}_n \xrightarrow{p} \beta$ as $n \to \infty$.

II. $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N\left(0, I^{-1}(\beta_0)\right)$, where $I(\beta_0) = E\left[\left(\frac{\partial}{\partial \beta} h(\beta)\right)^t V_i^{-1}(\beta)\left(\frac{\partial}{\partial \beta} h(\beta)\right)\right]$

where $I(\beta_0)$ is the Fisher information matrix.

The proof of Theorem 1 is provided in the appendix.

Simulation Study

A simulation experiment is carried out to study the finite sample performance of the ML estimation method. For each configuration of these experiments, r=2000 replications were performed with sample sizes n=1000, 2000, 3000. For each Ml estimator, bias, asymptotic standard error (ASE), standard deviation (SD) and coverage probability (CP) of a 95% confidence interval (CI) is conducted.

Under the proposed covariate extension of the two-stage RR design of Nargis and Shabbir [21], a predictor X is considered. We estimated the logistic regression parameters and the proportion with the sensitive characteristic through the estimates of the logistic regression parameters. The data of X are generated from the standard normal distribution. We generated the data of Y given X from Binomial distribution with probabilities 1 and $h(\beta)$, i.e, Binom(1, $h(\beta)$), where

$$h(\beta) = M\theta(X;\beta) + (1-M)\left[\left\{P_1\theta(X;\beta) + P_2\left(1-\theta(X;\beta)\right)\right\}\left\{1 + P_3\left(\frac{k}{k-1}\right)\right\} + P_3^2\left(\frac{k}{k-1}\right)\theta(X;\beta)\right]$$

$$\theta(X;\beta) = G\left(\beta^T \chi\right) = \frac{1}{1+e^{-\beta_0-\beta_1 X}} = \frac{1}{1+\mathrm{e}^{-\beta^T \chi}}$$

For the j^{th} simulated data set, $j = 1,2,....,r = 2000$, let X_{ij} be the simulated covariate vector of respondent i, $\hat{\beta}_j = (\hat{\beta}_{0j}, \hat{\beta}_{1j})$ the estimate of β, $\theta_j = \theta(X_j, \beta)$, $\hat{\theta}_j = \theta(X_j, \hat{\beta}_j)$

We set $p_1 = 0.7$ and considered the four settings for β: $\beta = (0,-0.1)$, $\beta = (0,0.1)$, $\beta = (0.1,0.1)$, $\beta = (-0.1,-0.1)$, $\beta = (0,0.05)$, $\beta = (0,-0.05)$, $\beta = (-0.2,0.1)$, $\beta = (0.2,-0.1)$, $\beta = (-0.1,0.2)$, $\beta = (0.1,-0.2)$. Y can be considered to follow a binomial distribution with a successive probability of $h(\beta)$.

Table 1 Summary of estimates of $\beta = (0,-0.1)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	0.0062	0.006196924	0.0633984	0.06287181	0.9505
β_1	-0.0860	0.01394103	0.06359208	0.06281593	0.945
n=2000					
β_0	0.00514019	0.00514019	0.04479392	0.04507761	0.9495
β_1	-0.08354492	0.01645508	0.04488357	0.04485077	0.9355
n=3000					
β_0	0.006423619	0.006423619	0.03656455	0.03693449	0.942
β_1	-0.08311984	0.01688016	0.03663034	0.03661939	0.926

Table 2 Summary of estimates of $\beta = (0,0.1)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	0.00650486	0.00650486	0.06339787	0.06212589	0.9465
β_1	0.08619352	-0.01380648	0.06359188	0.06279554	0.9425
n=2000					
β_0	0.007602258	0.007602258	0.04479398	0.04491366	0.9495
β_1	0.08347569	-0.01652431	0.04488363	0.04495784	0.9325
n=3000					
β_0	0.005246614	0.005246614	0.03656431	0.03667307	0.945
β_1	0.08298424	-0.01701576	0.03662995	0.0366581	0.9285

Table 3 Summary of estimates of $\beta = (0.1,0.1)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	0.09055997	-0.009440028	0.06346204	0.06214025	0.955
β_1	0.08538253	-0.01461747	0.06365438	0.0632904	0.941
n=2000					

β_0	0.08984205	-0.01015795	0.04484004	0.04633794	0.936
β_1	0.08453439	-0.01546561	0.04493007	0.04385336	0.9425
n=3000					
β_0	0.08886761	-0.01113239	0.03659996	0.03669592	0.9445
β_1	0.0826594	-0.0173406	0.03666463	0.03635292	0.924

Table 4 Summary of estimates of $\beta = (-0.1, -0.1)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	-0.07726889	0.02273111	0.06344416	0.06122341	0.9435
β_1	-0.08592028	0.01407972	0.06363822	0.06331324	0.9445
n=2000					
β_0	-0.07733629	0.02266371	0.04482778	0.04584458	0.916
β_1	-0.08415364	0.01584636	0.0449174	0.04402524	0.9405
n=3000					
β_0	-0.07603468	0.02396532	0.03659042	0.03717841	0.893
β_1	-0.08287949	0.01712051	0.03665512	0.03578425	0.931

Table 5 Summary of estimates of $\beta = (0, 0.05)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	0.00614044	0.00614044	0.06335653	0.06355361	0.948
β_1	0.04414354	- 0.005856459	0.06346514	0.06294338	0.9445
n=2000					
β_0	0.007603877	0.007603877	0.04476478	0.04441177	0.9535
β_1	0.04206578	-0.007934224	0.04479648	0.04480193	0.952
n=3000					
β_0	0.00557232	0.00557232	0.03654094	0.03711075	0.945
β_1	0.04161611	-0.008383894	0.03655964	0.03623139	0.951

Table 6 Summary of estimates of $\beta = (0, -0.05)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	0.005738907	0.005738907	0.06335589	0.06314004	0.949
β_1	-0.04434691	0.005653086	0.06346417	0.06261629	0.9485
n=2000					
β_0	0.005293243	0.005293243	0.04476506	0.04516969	0.9455

β_1	-0.04216278	0.007837223	0.04479683	0.04478767	0.954
n=3000					
β_0	0.005868165	0.005868165	0.0365409	0.03695879	0.949
β_1	-0.04153061	0.008469392	0.03655953	0.03627812	0.9475

Table 7 Summary of estimates of $\beta = (-0.2, 0.1)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	-0.1616657	0.03833434	0.06360504	0.06327265	0.9025
β_1	0.08347589	-0.01652411	0.06379305	0.06497568	0.9385
n=2000					
β_0	-0.1601983	0.03980168	0.04493661	0.04501556	0.8565
β_1	0.08215801	-0.01784199	0.04502232	0.04524885	0.931
n=3000					
β_0	-0.159669	0.04033103	0.03668148	0.03675896	0.805
β_1	0.08359344	-0.01640656	0.03674703	0.03697009	0.919

Table 8 Summary of estimates of $\beta = (0.2, -0.1)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	0.174273	-0.02572701	0.06363908	0.06317541	0.93
β_1	-0.08390506	0.01609494	0.06382773	0.06478818	0.934
n=2000					
β_0	0.1734011	-0.02659887	0.04496149	0.04536354	0.908
β_1	-0.08173175	0.01826825	0.04504685	0.04591719	0.9295
n=3000					
β_0	0.1722073	-0.0277927	0.03670108	0.03649937	0.885
β_1	-0.08442803	0.01557197	0.03676762	0.03685898	0.923

Table 9 Summary of estimates of $\beta = (-0.1, 0.2)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	-0.07568171	0.02431829	0.06361391	0.06463696	0.9285
β_1	0.1697707	-0.03022929	0.06414161	0.06427212	0.926
n=2000					
β_0	-0.07623936	0.02376064	0.04494132	0.04532624	0.9095
β_1	0.1663412	-0.03365878	0.04525952	0.04588558	0.8725

n=3000					
β_0	-0.07579136	0.02420864	0.03668503	0.03741202	0.895
β_1	0.1664519	-0.0335481	0.03693773	0.03752776	0.8405

Table 10 Summary of estimates of $\beta = (0.1, -0.2)$

	Mean	Bias	ASE	SD	CP
n=1000					
β_0	0.08856722	-0.01143278	0.06363118	0.06454501	0.943
β_1	-0.1698753	0.03012473	0.06415907	0.06435114	0.9255
n=2000					
β_0	0.08929859	-0.01070141	0.04495404	0.04517871	0.9415
β_1	-0.1666228	0.03337723	0.04527298	0.04597417	0.8735
n=3000					
β_0	0.08920893	-0.01079107	0.03669484	0.03739025	0.94
β_1	-0.166177	0.03382304	0.03694638	0.03752709	0.833

Simulation results

Summarized in Tables 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10 are the estimates of β_0 and β_1 for different parameter settings, including $\beta = (0, -0.1)$, $\beta = (0, 0.1)$, $\beta = (0.1, 0.1)$, $\beta = (-0.1, -0.1)$, $\beta = (0, 0.05)$, $\beta = (0, -0.05)$, $\beta = (-0.2, 0.1)$, $\beta = (0.2, -0.1)$, $\beta = (-0.1, 0.2)$ and $\beta = (0.1, -0.2)$. The results are presented for sample sizes n=1000,2000,3000, evaluating key metrics such as mean, bias, asymptotic standard error (ASE), standard deviation (SD), and empirical coverage probability (CP).

For $\beta = (0, -0.1)$ the bias of β_0 and β_1 remained relatively small across all sample sizes, with ASE and SD values converging as n increased. The empirical coverage probability remained close to 95%, indicating good performance of the estimator. Similar trends were observed for $\beta = (0, 0.1)$, where biases were slightly lower, and coverage probabilities were consistently above 92%.

However, as sample size increased, a notable trend emerged, the coverage probability of β_1 tended to decrease. This suggests that while larger sample sizes reduce the bias and variance of the estimates, the standard error estimates may slightly underestimate the true variability, leading to narrower confidence intervals and a drop in coverage probability. This effect was particularly noticeable for more extreme parameter settings, such as $\beta = (0.2, -0.1)$ where the empirical CP of β_1 dropped below 93% for n=3000.

In summary, for moderate values of β_0 and β_1, the estimates performed well, with bias decreasing and ASE values aligning closely with SD. The empirical coverage probabilities were generally near the nominal 95%, except for larger n where β_1 showed slight under coverage. These results highlight that while increasing sample size improves estimation precision, adjustments in confidence interval calculations may be necessary to maintain accurate coverage probabilities, especially for β_1.

Conclusion:

We have developed a covariate extension of the two-stage RRT proposed by Narjis and Shabbir [21], utilizing logistic regression to examine the effect of covariates on a sensitive characteristic. The relationship between covariates and randomized response variables was explored through the logistic regression model, and an effective method was introduced to estimate the population proportion of individuals with the sensitive characteristic using logistic regression parameters. Additionally, we derived and proved the large-sample properties of the maximum likelihood (ML) estimators for these parameters.

Simulation studies were conducted to assess the finite-sample performance of the ML estimators. The means, biases, asymptotic standard errors (ASE), standard deviations (SD), and coverage probabilities of the estimates were analyzed for sample sizes n=1000,2000, and 3000 with 2000 replications. Results indicated a decreasing pattern in ASE and SD as sample size increased. For moderate values of β_0 and β_1, the estimators performed well, with bias reduction and ASE values closely aligning with SD. The empirical coverage probabilities were generally near the nominal 95%, except for larger n, where β_1 exhibited slight under coverage. These findings suggest that while increasing sample size improves estimation precision, adjustments in confidence interval calculations may be necessary to maintain accurate coverage probabilities, particularly for β_1.

The proposed methodology can also be applied to other randomized response designs. Future work will focus on developing estimation methods for logistic regression parameters when covariates are missing in data collected through the two-stage RRT of Narjis and Shabbir [21].

Appendix:

Proof of Lemma 1

$\psi_i(\beta)$ in (19) can be written as

$$\Psi_i(\beta) = \frac{Y_i}{h(\beta)}\left(\frac{\partial}{\partial\beta}h(\beta)\right) - \frac{1-Y_i}{1-h(\beta)}\left(\frac{\partial}{\partial\beta}h(\beta)\right)$$

$$= \left(\frac{\partial}{\partial\beta}h(\beta)\right)\left[\frac{Y_i}{h(\beta)} - \frac{1-Y_i}{1-h(\beta)}\right]$$

$$= \left(\frac{\partial}{\partial\beta}h(\beta)\right)\left[\frac{Y_i - h(\beta)}{h(\beta)\left(1-h(\beta)\right)}\right]$$

where,

a. $h(\beta) = \theta\left(\beta, X_i\right)$ is the predicted probability for $Y_i = 1$

b. $\frac{\partial}{\partial\beta}h(\beta)$ is the derivative of θ w.r.t β

c. $Y_i \in (0,1)$

Also, $V_i = h(\beta)\left(1-h(\beta)\right)$, so,

$$\Psi_i(\beta)=\left(\frac{\partial}{\partial\beta}h(\beta)\right)V_i^{-1}(\beta)\left[Y_i-h(\beta)\right], i=1,2,...,n$$

Hence, the score function $U_n(\beta)$ can be written as

$$U_n(\beta)=\sum_{i=1}^{n}\psi_i(\beta)=\Psi_i(\beta)=\left(\frac{\partial}{\partial\beta}h(\beta)\right)V_i^{-1}(\beta)\left[Y_i-h(\beta)\right] \tag{23}$$

Proof of Theorem 1

(a) Proof of Consistency of $\hat{\beta}$

Because from conditions 1 and 2 and the inverse function theorem of Foutz (1977), $U_n(\beta)=0$, has a unique solution, the ML estimator $\hat{\beta}$ is a consistent estimator of β.

(b) Proof of asymptotic normality of $\sqrt{n}\left(\hat{\beta}-\beta\right)$

The score function is defined as :

$$U_n(\beta)=\frac{\partial}{\partial\beta}\ln(\beta)=\frac{\partial}{\partial\beta}\log_n(\beta)$$

where,

$$\ln(\beta)=\sum_{i=1}^{n}\log_i(\beta)$$

At the MLE $\hat{\beta}$, the score function satisfies $U_n(\hat{\beta})=0$, using a first order Tylor expansion, expand $U_n(\beta)$ around β_0, that is,

$$U_n(\hat{\beta})=U_n(\beta)+\frac{\partial}{\partial\beta}U_n(\beta)\Big|_{\beta=\beta_0}(\hat{\beta}-\beta_0)+R_n \tag{24}$$

Where, R_n is the remainder term which vanishes when $n\to\infty$.

Since, $U_n(\hat{\beta})=0$, the above simplifies to

$$0=U_n(\beta_0)+\frac{\partial}{\partial\beta}U_n(\beta)\Big|_{\beta=\beta_0}(\hat{\beta}-\beta_0)+R_n \tag{25}$$

As we know that, $U_n(\beta_0)=\sum_{i=1}^{n}\psi_i(\beta_0)$, where $\psi_i(\beta_0)=\frac{\partial}{\partial\beta}\log_i(\beta)$.

Thus, by central limit theorem, as $n\to\infty$,

$$U_n(\beta_0)\xrightarrow{d}N(0,I(\beta_0))$$

Where, $I(\beta_0) = -E\left(\frac{\partial^2}{\partial\beta^2}\log_i(\beta)\right)$.

We define the Hessian matrix (the second derivative of the log likelihood)

$$H_n(\beta) = \frac{\partial^2}{\partial\beta^2}\log_n(\beta)$$

At β_0, the normalized Hessian converges to the Fisher information matrix, that is,

$$\frac{1}{n}H_n(\beta_0) \xrightarrow{P} I(\beta_0) \text{ as } n \to \infty$$

Thus, from the Tylor expansion given in (25), and the consistency of $\hat{\beta}$, the remainder term R_n vanishes asymptotically. Thus,

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) = -\left(\frac{1}{n}H_n(\beta_0)\right)^{-1}\frac{1}{\sqrt{n}}U_n(\beta_0) \tag{26}$$

Using the results:

a. $\frac{1}{n}H_n(\beta_0) \xrightarrow{P} I(\beta_0)$

b. $\frac{1}{\sqrt{n}}U_n(\beta_0) \xrightarrow{d} N\left(0, I\left(\beta_0\right)\right)$

Thus, it is concluded that

$$\sqrt{n}\left(\hat{\beta} - \beta_0\right) \xrightarrow{d} N\left(0, I^{-1}\left(\beta_0\right)\right)$$

Proved

References

[1] Warner, S. L. Randomized response: A survey technique for eliminating evasive answer bias. Jour Amer Stat Assoc 1965, 60(309), 63-69.
[2] Greenberg, B. G.; Abul-Ela, A. L. A.; Simmons, W. R.; Horvitz, D. G. The unrelated question randomized response model: Theoretical framework. Jour of the Amer Stat Assoc 1969, 64(326), 520-539
[3] Greenberg, B. G.; Kuebler Jr; R. R.; Abernathy, J. R.; Horvitz, D. G. Application of the randomized response technique in obtaining quantitative data. Jourl of the Amer Stat Assoc 1971, 66(334), 243-250
[4] Chow, L. P.; Liu, P. T. A new randomized response technique: the multiple answer model. Dep of Pop Dyn, John Hop Uni, Balt, Md, 1973
[5] Goodstadt, M. S.; Gruson, V. The randomized response technique: A test on drug use. Jour of the Amer Stat Assoc, 1975, 70(352), 814-818.
[6] Eichhorn, B. H.; Hayre, L. S. Scrambled randomized response methods for obtaining sensitive quantitative data. Jour Stat Plan infe, 1983, 7(4), 307-316.
[7] Scheers, N. J.; Dayton, C. M. Improved estimation of academic cheating behavior using the randomized response technique. Res in High Edu, 1987, 26(1), 61-69.
[8] Chaudhuri, A; Mukerjee,R. Randomized response: theory and techniques; Indian Statistical Institute: Calcutta, 1988
[9] Mangat, N. S.; Singh, R. An alternative randomized response procedure, Biometrika, 1990, 77, 439-442.

[10] Kuk, A. Y. Asking sensitive questions indirectly. Biometrika, 1990, 77, 436-438
[11] Mangat, N. S.; Singh, S.; Singh, R. On use of a modified randomization device in Warner model. Journal of the Indian Soc in Stat and Oper Res, 1995, 16, 65-69.
[12] Bar-Lev, S. K.; Bobovitch, E.; Boukai, B. A note on randomized response models for quantitative data. Metrika, 2004, 60(3), 255-260
[13] Huang, K. C. (2004). A survey technique for estimating the proportion and sensitivity in a dichotomous finite population. Statistica Neerlandica, 58(1), 75-82.
[14] Kim, J. M.; Tebbs, J. M.; An, S.W. Extensions of Mangat's randomized-response model. Jour of Stat Plan and Inf, 2006, 136(4), 1554 -1567
[15] Saha, A. A simple randomized response technique in complex surveys. Metron, 2007, 65(1), 59-66
[16] Singh, S., Grewal, I. S. Geometric distribution as a randomization device: implemented to the Kuk's model. Inter Jour of Cont Math Sci 2013, 8(5), 243-248.
[17] Blair, G.; Imai, K.; Zhou, Y. Y. Design and analysis of the randomized response technique. Jour Amer Stat Assoc 2015, 110(511), 1304-1319.
[18] Lee, S. M., Peng, T. C., Tapsoba, J. D. D., Hsieh, S. H. Improved estimation methods for unrelated question randomized response techniques. Comm in Stat-Theo and Meth 2017, 46(16), 8101-8112.
[19] Kim, J. M.; Tebbs, J. M.; An, S.W. Extensions of Mangat's randomized-response model. Jour of Stat Plan and Inf, 2006, 136(4), 1554 -1567
[20] Shah, S. F.; Hussain, Z.; Cheema, S. A. Combining answers to direct and indirect questions: An implementation of Kuk's randomized response model. Com in Stat-Theo and Meth 2019, 1-17.
[21] Narjis, G., Shabbir, J. An improved two-stage randomized response model for estimating the proportion of sensitive attribute. Socio Meth & Res 2023, 52(1), 335-355.
[22] Shah, S. F., Hussain, Z., Riaz, M., Cheema, S. A. Shewhart-Type Charts for Masked Data: A Strategy for Handling the Privacy Issue. Math Prob in Eng, 2020, (1), 5104753.
[23] Shah, S. F., Cheema, S. A., Hussain, Z., Shah, E. A. Masking data: a solution to social desirability bias in paired comparison experiments. Comm in Stat-Simu and Comp 2022, 51(6), 3149-3167.
[24] Maddala GS. Limited-dependent and qualitative variables in econometrics. Camb Uni Press, Camb 1983
[25] Hsieh SH, Lee SM, Shen PS. Semiparametric analysis of randomized response data with missing covariates in logistic regression. Comput Stat Data Anal 2009, 53, 2673–2692
[26] Hsieh SH, Lee SM, Shen PS. Logistic regression analysis of randomized response data with missing covariates. J Stat Plan Inference 2010,140, 927–940
[27] Bhargava, M. and R. Singh. A Modified Randomized Device for Warner's Model. Stat 2000, 60(2), 315-21.
[28] Gupta, S., J. Shabbir, R. Iembo. Modifications to Warner's Model Using Blank Cards. Ameri Jour of Math and Manag Sci 2006, 26(1),185-96.