Journal of Sciences, Islamic Republic of Iran 35(3): 221 - 231 (2024) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Modeling Some Repeated Randomized Responses

M. Tarhani, M. R. Zadkarami*, S. M. R. Alavi

Department of Statistics, Faculty of Mathematical Sciences and Computer, Shahid Chamran University of Ahvaz, Ahvaz, Islamic Republic of Iran

Received: 25 September 2024 / Revised: 2 January 2025 / Accepted: 1 February 2025

Abstract

Some social surveys address sensitive topics for which respondents do not report reliable responses. Randomized response techniques (RRTs) are employed to increase privacy levels and provide honest answers. However, estimates obtained from this method tend to exhibit increased variances. Repeating randomized responses for each individual increases the sample size, and the mean of observations for each individual reduces the variance of the parameter's estimator, bringing them closer to reality. In this study, considering continuous additive repeated randomized responses (RRRs), we apply the averaged RR of each individual using the linear regression model for sensitive variable mean. Data on the income of family heads were collected from students, and each respondent was asked to randomize their responses five times. The maximum likelihood estimators of parameters are obtained by two methods. In the first method, the response variable is the first reported observation, and in the second method, we considered the averaged RR for each individual. The results emphasize that the estimators from the second method are closer to reality and have lower variance.

Keywords: Randomized Response; Repeated Randomized Response; Linear Regression Model; Continuous Sensitive Variable; Repeated Individual Observations.

Introduction

In many social sampling surveys, some questions may be sensitive to respondents, leading to insecurity in providing honest answers. A sensitive variable has a high level of social privacy or pertains to individuals' private lives. For example, research related to addiction, bribery, specific political views, socially undesirable behaviors, or income. The RR technique is a sampling process in which respondents are more willing and confident in providing honest answers to questions.

As the pioneer paper, the RR technique for sensitive binary questions was introduced in 1965 (1). It included answering a sensitive question or its complement using a Bernoulli trial (tossing a coin). Considering this trick, the sensitive answer remains hidden from the researcher, preserving the respondent's privacy. Afterward, many methods were proposed to examine sensitive qualitative data, including the unrelated response method or Simon's method (2). Many authors extended this method (3-5). Another method is the forced response technique introduced in 1971 (6). An estimate of the sensitive proportion through the maximum likelihood method was obtained using the proposed RRR (7). The RRR technique increases privacy protection and provides a truthful answer by reporting different responses by an individual. The logistic regression parameters for RR data gathered using Warner's method were estimated (8,

* Corresponding Author: Tel: 06133369509; Email: zadkarami@yahoo.co.uk, zadkarami_m@scu.ac.ir

9). Subsequently, many researchers estimated the sensitive proportion and regression parameters using various RRTs for univariate or multivariate logistic regression models (10-16). The additive RR method was used when the sensitive attribute is a discrete quantitative variable, and the mean of the item-sum technique was estimated (17-18).

For continuous sensitive variables, the mean estimate is obtained using the systematic random sampling design in the presence of a non-sensitive auxiliary variable (19). Additionally, methods for estimating the mean of sensitive variables in the presence of measurement error have been developed (20-21), and variance estimators for sensitive variables using RRT have also been proposed (22). Quantitative RRTs were investigated to enhance respondent trust (23- 24). The effect of the initial non-response on the regression estimator in panel surveys was reduced using RRT (25). By assuming truthful responses about domain membership, non-sensitive quantitative variables were estimated for specific sensitive domains (26). RRs can shorten the length of certain confidence intervals with a conditional coverage guarantee (27).

Modeling for continuous RRs is a less explored topic. In many cases, the sensitive variable is continuous. For example, income, tax evasion, expenses for election campaigns, drug or alcohol consumption during a week, student grade point averages, and financial or ethical corruption. The unrelated question design was employed in 1971 (28) to estimate the mean of the quantitative sensitive variable. The sensitive response was added to a random number of the scramble variable (a variable with known-finite mean and variance) (29). The multiplicative method was introduced by multiplying the sensitive variable by the scramble variable (30). Additive and multiplicative approaches, the optional and mixture RR methods, increase reliability and reduce bias in the reported responses (31, 32). Regression-cum-ratio estimator estimates the sensitive variable mean (33).The authors apply several estimation methods. The regression parameters using forced RRT and the EM algorithm in a Poisson distribution were estimated (34). Multiplicative RR regression parameters were estimated using the least squares method (35). The regression coefficients were estimated using the maximum likelihood method for the multiplicative design when the scramble variable was distributed as uniform (36). Later, a multiplicative RR design was applied as the dependent variable, and the regression parameters were estimated using the least squares method (37). The regression parameters for the model introduced in (39) were estimated (38). The model parameters were estimated for a generalized linear mixed effects model employing the forced-response technique (40). There are some reasons for the limited research on modeling based on RRs, including the complexity of the model and the limited packages in commonly used software. Furthermore, changing the method of randomizing responses also affects the modeling, making it more complex (40).

A privacy criterion was introduced (41). The larger this criterion, the more confidential the RRT becomes, and respondents are expected to be more willing to participate in the study. A measure for comparing quantitative RR methods based on the variance-to-privacy was proposed (42). The smaller the value, the greater the privacy for the RRT. In this paper, we use this criterion to evaluate the privacy of quantitative RRTs.

The main focus of this article is to study models for continuous RRRs. Using RRs gets the parameter estimators closer to reality and improves efficiency; however, it increases the variance of the estimators. We consider the RRR model for the mean individual observations, which can remedy the variance growth by increasing the number of responses for each respondent. It is worth mentioning that the scramble variable with a known mean should be chosen so that the true sensitive value cannot be discerned from the participant's reported value. Otherwise, they may lack confidence in providing honest answers.

We studied repeated additive RR responses from 512 students in 2018. The information included the number of family members, education, occupation, age of the family head, and the monthly income in millions of the family head. The monthly income of the family head was added to an existing random number of the scramble variable, and the result was reported, and this process was repeated five times. Regression parameter estimators were obtained using the first response of each respondent and the average of each respondent's responses, which was reported.

The remainder of the paper is structured as follows. In the second section, the parameters for the additive, multiplicative, mixture, and optional techniques are estimated when considering the normal sensitive and scramble variable(s). The third section explains the real data application. In the fourth section, simulations are performed to evaluate the parameter estimates. Their privacy is compared using thecriteria above. The discussion is in the last section.

1. Randomized Response Techniques (RRTs)

Let $Y \sim N(\mu, \sigma^2)$, and $S \sim N(\mu_s, \sigma_s^2)$ denote the sensitive variable and the scramble variable (μ_s and σ_s^2 known), respectively. We consider two cases. In the first case, to reduce response bias and enhance privacy, each respondent should add their response with a random value of the scramble variable, report only the result, and

then repeat the procedure m times. Let RR variable denote byZ, then for individual i, the j-th RR is given by: $z_{ij} = y_i + s_{ij} \quad i = 1, .., n\,, \quad j = 1, \dots, m,$

where y_i and s_{ij} denote the true value of the sensitive variable and a random value of the scramble variable for the i-th individual in j-th repeat of RR. If T denotes the predictor variable for the sensitive variable, the unbiased prediction for i-th individual in j-th RR is as follows: $t_{ij} = z_{ij} - \mu_s, \quad i = 1, .., n\,, \quad j = 1, \dots, m.$

Then, the estimation of the sensitive variable mean and its variance are calculated as follows:

$$\hat{\mu}_t = \bar{z} - \mu_s, \qquad V(\hat{\mu}) = \frac{\sigma^2 + \sigma_s^2}{n}$$

In the second case, considering the average of m RRs for each individual as observation, the predictor for i-th individual is $\bar{t}_i = \bar{z}_i - \mu_s, i = 1, .., n$, where $\overline{T} \sim N(\mu, \sigma^2 + \frac{\sigma_s^2}{m})$, and $\hat{\mu} = \bar{\bar{z}} - \mu_s\,, V(\hat{\mu}) = (\frac{\sigma^2}{n} + \frac{\sigma_s^2}{nm})$.

The matrix form of the model for the sensitive variable y is as follows:

$$y = X\beta + \varepsilon \qquad (1)$$

where X denotes the matrix of explanatory variables. Suppose the error term ε has zero mean and variance$v = \sigma_\varepsilon^2 I$, where I denote the identity matrix. Since its true value is not observable, the RR variable is used. Therefore, the model that uses the averaged RRs is as follows:

$$\bar{z} = g(X\beta) + \xi \qquad (2)$$

$g(X\beta) = X\beta + \mu_s\,, \quad \xi = \varepsilon + \delta_s\,,$

where $\delta_s = (\bar{s} - \mu_s) \sim N(0, \tau)$ is the error of selecting the scramble value, and $\tau = \frac{\sigma_s^2}{m} I$ is the covariance matrix of the vector $\bar{s}$. Assuming independence of S and Y, ξ is distributed as $N(0, \psi)$ where $\psi = \Sigma + \tau$.

The model parameters are estimated using the maximum likelihood method. The log-likelihood function is given by

$$l_Z(\beta, \sigma_\varepsilon^2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\ln|\psi| - \frac{1}{2}(\bar{z} - X\beta - \mu_s)'\psi^{-1}(\bar{z} - X\beta - \mu_s),$$

and the maximum likelihood estimators (MLEs) of the unknown parameters are

$$\hat{\beta} = (X'X)^{-1}X'(\bar{z} - \mu_s),$$

$$\hat{\sigma}^2 = \frac{1}{n}(\bar{z} - X\hat{\beta} - \mu_s)'(\bar{z} - X\hat{\beta} - \mu_s) - {\sigma_s^2}/{m}.$$

The distribution of the regression coefficient estimators is as follows:

$$\hat{\beta} \sim N\left(\beta, \left(\sigma^2 + \frac{\sigma_S^2}{m}\right)(X'X)^{-1}\right).$$

Then, the use of RRTs can increase the variance of parameter estimates.

1.1. Additive-Scrambled RR Technique

Suppose respondents report RR variable $Z = aY + bS$ instead of the sensitive value Y where a and b are known constant values, and S denotes a random value from the independent scramble variable $S \sim N(\mu_S, \sigma_S^2)$. Then, $Z \sim N(a\mu + b\mu_S, a^2\sigma^2 + b^2\sigma_S^2)$ and, the j-th reported RR variable of individual i is $z_{ij} = ay_i + bs_{ij}\,, \quad i = 1, \dots, n\,, \quad j = 1, \dots, m.$

The predictor variable based on one observation, T, is given by $T = \frac{aY + bS - b\mu_S}{a} = \frac{Z - b\mu_S}{a} \sim N(\mu, (a^2\sigma^2 + b^2\sigma_S^2)/a^2)$, and the unbiased predictor of Y using m repetitions is $\overline{T} = \frac{\overline{Z} - b\mu_S}{a} \sim N(\mu, \sigma^2 + \frac{b^2}{ma^2}\sigma_S^2))$.

The log-likelihood function for the predictor variable is given by

$$l(\mu, \sigma^2) = -\frac{1}{2\left\{n\ln 2\pi + \ln\left(\sigma^2 + \frac{b^2}{ma^2}\sigma_S^2\right) + \frac{1}{\left(\sigma^2 + \frac{b^2}{ma^2}\sigma_S^2\right)}(\overline{T} - \mu)'(\overline{T} - \mu)\right\}}.$$

The MLEs of the model parameters are $\hat{\mu} = \bar{\bar{t}}$ and $\hat{\sigma}^2 = \frac{(\overline{T} - \hat{\mu})'(\overline{T} - \hat{\mu})}{n} - \frac{b^2}{a^2}\sigma_S^2$ and, the variance of the estimator of $\hat{\mu}$ is $V(\hat{\mu}) = V(\bar{\bar{t}}) = \frac{V(\overline{Z})}{na^2} = \frac{\sigma^2}{n} + \frac{b^2\sigma_S^2}{nma^2}$.

Let the model error for the sensitive variable distribute as $\varepsilon \sim N(0, \sigma^2 I)$. Due to the lack of the latent variable, its predictor variable, $\bar{t}$, is used. The model is given by

$$\bar{t} = X\beta + \varepsilon^*, \qquad (3)$$

in which $\varepsilon^* \sim N(0, \sigma_{\varepsilon^*}^2 I)$, where $\sigma_{\varepsilon^*}^2 = (a^2\,\sigma^2 + \frac{b^2\sigma_S^2}{m})/a^2$.

The log-likelihood function for the RR model is

$$l(\beta, \sigma^2) = -1/2\left\{n\ln 2\pi + \ln\left(\sigma^2 + \frac{b^2}{ma^2}\sigma_S^2\right) + \frac{1}{\left(\sigma^2 + \frac{b^2}{ma^2}\sigma_S^2\right)}(\overline{T} - X\beta)'(\overline{T} - X\beta)\right\}.$$

The MLEs of the proposed model parameters are as follows:

$$\hat{\beta} = (X'X)^{-1}X'\overline{T}, \text{ and } \hat{\sigma}^2 = \frac{1}{n}\left(\overline{T} - X\hat{\beta}\right)'\left(\overline{T} - X\hat{\beta}\right) - \frac{b^2\sigma_S^2}{ma^2}.$$

The distribution of the regression coefficients estimators is given by:

$$\hat{\beta} \sim N\left(\beta, \left(\sigma^2 + \frac{b^2\sigma_S^2}{ma^2}\right)(X'X)^{-1}\right).$$

1.2. Additive-Scrambled-Scrambled Technique

Suppose the respondents multiply their sensitive answer by a known constant value a, and randomly selecte two independent values, S_1 and S_2,from known scramble variables $N(\mu_{S_1}, \sigma_{S_1}^2)$ and $N(\mu_{S_2}, \sigma_{S_2}^2)$, respectively and it is reported the RR variable $Z = aY + bS_1 + cS_2$ to the researcher for two known constants b and c. Then, the reported variable Z is distributed as $Z \sim N(a\mu + b\mu_{S_1} + c\mu_{S_2}, a^2\sigma^2 + b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2)$. The j-th reported value of Z for individual i is:

$$z_{ij} = ay_i + bs_{1ij} + cs_{2ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, m.$$

The unbiased predictor variable for a single RR, T, is as follows:

$$T = \frac{aY + bS_1 + cS_2 - b\mu_{S_1} - c\mu_{S_2}}{a} = \frac{Z - b\mu_{S_1} - c\mu_{S_2}}{a} \sim N\left(\mu, \sigma^2 + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{a^2}\right),$$

where the averaged RR for each individual is:

$$\bar{T} = \frac{\bar{Z} - b\mu_{S_1} - c\mu_{S_2}}{a} \sim N\left(\mu, \sigma^2 + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}\right).$$

The log-likelihood function is given by

$$l(\mu, \sigma^2) = -\frac{1}{2}\left\{ nln2\pi + nln\left(\sigma^2 + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}\right) + \frac{1}{\left(\sigma^2 + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}\right)}(\bar{T} - \mu)'(\bar{T} - \mu)\right\},$$

The MLE's of parameters are given by $\hat{\mu} = \bar{\bar{t}}$, and $\hat{\sigma}^2 = \frac{(\bar{T} - \hat{\mu})'(\bar{T} - \hat{\mu})}{n} - \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}$ where, The variance of $\hat{\mu}$ is $V(\hat{\mu}) = V(\bar{\bar{t}}) = \frac{V(\bar{Z})}{na^2} = \frac{\sigma^2}{n} + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{nma^2}$.

Consider the sensitive variable, which is defined using equation (1). Consequently, the variance of the model error is $\sigma_{\varepsilon^*}^2 = \sigma^2 + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}$ for the predictor variable in equation (3), where $\varepsilon^* \sim N(0, \sigma_{\varepsilon^*}^2 I)$.

The log-likelihood function for estimated MLE's of model parameters is as follows:

$$l(\beta, \sigma^2) = -1/2\left\{ nln2\pi + nln\left(\sigma^2 + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}\right) + \frac{1}{\left(\sigma^2 + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}\right)}(\bar{T} - X\beta)'I(\bar{T} - X\beta)\right\},$$

Then, the parameters MLE's are $\hat{\beta} = (X'X)^{-1}X'\bar{T}$ and $\hat{\sigma}^2 = \frac{1}{n}(\bar{T} - X\hat{\beta})'(\bar{T} - X\hat{\beta}) - \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}$.

The distribution of the regression coefficients estimators is $\hat{\beta} \sim N\left(\beta, \left(\sigma^2 + \frac{b^2\sigma_{S_1}^2 + c^2\sigma_{S_2}^2}{ma^2}\right)(X'X)^{-1}\right)$.

1.3. Optional RR Technique

In the additive-optional RRT, respondents either report the sensitive value or add it with a random value of the scramble variable. Let $Y \sim N(\mu, \sigma^2)$, and G be the sensitive variable and a Bernoulli random variable with probability p, respectively, then the reported variable is $Z = YG + (Y + S)(1 - G)$. It shows that the sensitivity level of the variable Y is $(1 - p)$. The j-th reported value of Z for individual i is as follows:

$$z_{ij} = \begin{cases} y_i, & with\ probability\ p \\ y_i + s_{ij}, & with\ probability\ 1 - p \end{cases}, \quad i = 1, \dots, n, \quad j = 1, \dots, m$$

Then, the additive-optional RR variable has the following mixture density function

$$f_Z(Z) = p\varphi\left(\frac{y - \mu}{\sigma}\right) + (1 - p)\varphi\left(\frac{y + s - (\mu + \mu_s)}{\sqrt{\sigma^2 + \sigma_s^2}}\right). \quad (4)$$

From density (4), we have the following equation: $E_p(E_R([(Z|G)])) = E_p(pY + (1 - p)(Y + S))$.

The mean and variance of Z are $\mu_z = \mu + (1 - p)\mu_S$ and, $\sigma_z^2 = (p - p^2){\mu_s}^2 + (1 - p)\sigma_s^2 + \sigma^2$ respectively and, the MLE's of the parameters are $\hat{\mu} = \bar{z} - (1 - p)\mu_S$ and $\hat{\sigma}^2 = \left(\frac{(Z - \hat{\mu}_z)'(Z - \hat{\mu}_z)}{n} - (1 - p)\sigma_s^2 - (p - p^2){\mu_s}^2\right)$ where, the variance of $\hat{\mu}$ is:

$$V(\hat{\mu}) = \frac{(p - p^2)\mu_S^2 + (1 - p)\sigma_S^2 + \hat{\sigma}^2}{n}.V(\hat{\mu}).$$

The unbiased predictor variable, T, and the averaged RR, $\bar{T}$, have variances $\sigma_t^2 = \sigma^2 + (1 - p)\sigma_s^2$ and $\sigma_{\bar{z}}^2 = \sigma_{\bar{t}}^2 = \sigma^2 + (1 - p)\,\sigma_s^2/m + (p - p^2)\mu_S^2/m$, respectively where, $T = Z - (1 - G)\mu_S$ and, $\bar{T} = \bar{Z} -$

$(1-p)\mu_S$. Based on $\bar{T}$, the log-likelihood function for the predictor variable is:

$$l(\mu,\sigma^2) \propto \sum ln\left(\frac{1}{\sigma_{\bar{z}}} exp\left\{-\frac{1}{2}\left(\frac{\bar{z}_i-\mu_z}{\sigma_{\bar{z}}}\right)^2\right\}\right)$$
$$= \sum ln\left(\frac{1}{\sigma_{\bar{t}}} exp\left\{-\frac{1}{2}\left(\frac{\bar{t}_i-\mu}{\sigma_{\bar{t}}}\right)^2\right\}\right).$$

Furthermore, the MLE's of the parameters are:

$$\hat{\mu} = \bar{\bar{z}} - (1-p)\mu_S\,,$$
$$\hat{\sigma}^2 = \left(\frac{(\bar{t}-\hat{\mu})'(\bar{t}-\hat{\mu})}{n} - (1-p)\,\sigma_s^2/m - (p-p^2)\mu_S^2/m\right) \quad (5)$$

Consider regression equation (1) for the sensitive variable, the log-likelihood function is rewritten as

$$l(\mu,\sigma^2) = -1/2\left\{nln2\pi + nln(\sigma_{\bar{z}}^2) + \frac{1}{\sigma_{\bar{z}}^2}(\bar{Z}-(X\beta+(1-p)\mu_S))'(\bar{Z} - (X\beta+(1-p)\mu_S))\right\}$$
$$= -1/2\left\{nln2\pi + nln\left(\sigma_{\bar{t}}^2\right) + \frac{1}{\sigma_{\bar{t}}^2}(\bar{t}-X\beta)'(\bar{t}-X\beta)\right\}$$

Then, the MLE's of parameters are as follows:

$$\hat{\beta} = (X'X)^{-1}X'(\bar{Z}-(1-p)\mu_S\,),$$
$$\hat{\sigma}^2 = \frac{1}{n}\left(\bar{Z}-X\hat{\beta}-(1-p)\mu_S\right)'\left(\bar{Z}-X\hat{\beta}-(1-p)\mu_S\right) - (1-p)\,\sigma_s^2/m - \frac{(p-p^2){\mu_s}^2}{m}.$$

Therefore, the distribution of the regression coefficients estimators is given by:

$$\hat{\beta}\sim N\left(\beta,\left((1-p)\,\sigma_s^2/m + (p-p^2){\mu_s}^2/m + \sigma^2\right)(X'X)^{-1}\right).$$

1.4. Productive RR Technique

Assume that respondents multiply their sensitive value Y by a known value from the scramble variable S. Then, the RR variable is $Z = YS$, and the j-th answer for individual i is given by:

$$z_{ij} = y_i s_{ij}\,, \qquad i = 1,\dots,n\,, \qquad j = 1,\dots,m.$$

Then, the unbiased predictor variable $T = \frac{Z}{\mu_s}$ is defined for a single RR. The mean and variance estimators of the sensitive variable are $\hat{\mu} = \bar{t}$ and $\hat{\sigma}^2 = \frac{\left(\frac{(t-\hat{\mu})'(t-\hat{\mu})}{n} - \frac{\mu^2\sigma_s^2}{{\mu_s}^2}\right)}{\left(1+\frac{\sigma_s^2}{{\mu_s}^2}\right)}$ respectively when, the variance of $\hat{\mu}$ is $(\hat{\mu}) = \sigma^2 + (\frac{\mu^2+\sigma^2}{n})\frac{\sigma_s^2}{\mu_s^2}$.

The averaged RR for each individual is $\bar{T} = \frac{\bar{Z}}{\mu_s}$ which is an unbiased predictor variable with the mean and variance μ and, $\sigma^2 + (\frac{\mu^2+\sigma^2}{m})\frac{\sigma_s^2}{\mu_s^2}$, respectively. Then, the estimators of the mean and variance are $\hat{\mu} = \bar{\bar{t}}$, and $\hat{\sigma}^2 = \frac{\left(\frac{(t-\hat{\mu})'(t-\hat{\mu})}{n} - \frac{\mu^2\sigma_s^2}{m{\mu_s}^2}\right)}{\left(1+\frac{\sigma_s^2}{m{\mu_s}^2}\right)}$, respectively where, the variance of $\hat{\mu}$ is $V(\hat{\mu}) = \frac{\sigma^2}{n} + (\frac{\mu^2+\sigma^2}{mn})\frac{\sigma_s^2}{\mu_s^2}$ and the privacy level is calculated as $P_L = (\mu^2+\sigma^2)\left(\frac{\sigma_s^2}{m} + (\mu_s-1)^2\right)$.

The log-likelihood equation using the predictor variable $\bar{T}$ is given by

$$l(\mu,\sigma^2) = log\left(\int_{-\infty}^{\infty} \frac{1}{2\pi|\bar{s}|\sqrt{\frac{\sigma^2\sigma_s^2}{m}}} exp\left(-\frac{m}{2}\left(\frac{\bar{s}-\mu_s}{\sigma_s}\right)^2 - \frac{1}{2\sigma^2}\left(\frac{\bar{z}}{\mu_s}-\mu\right)^2\right) ds\right).$$

Numerical methods estimate these parameters since the likelihood equation does not lead to a closed-form solution

Considering the regression equation (1) for sensitive variable, the log-likelihood function using equation (3), is as follows:

$$l(\beta,\sigma^2) = log\left(\int_{-\infty}^{\infty} \frac{1}{2\pi|\bar{s}|\sqrt{\frac{\sigma^2\sigma_s^2}{m}}} exp\left(-\frac{m}{2}\left(\frac{\bar{s}-\mu_s}{\sigma_s}\right)^2 - \frac{1}{2\sigma^2}(\bar{T}-X\beta)^2\right) ds\right).$$

where $= \sigma_{\varepsilon^*}^2 + (\frac{\mu^2+\sigma^2}{m})\frac{\sigma_s^2}{\mu_s^2}$. The likelihood equations do not have closed-form solutions, so numerical methods are used.

2. Application

In this section, the real data is applied to investigate the proposed RR method. In the data collection, fifty random values from the normal scramble variable $S\sim N(3.6, 0.5^2)$ are selected and recorded in fifty cards. The income of the family head is one of the sensitive questions in social sciences studies. In a questionnaire, 512 bachelor's students at Shahid Chamran University were asked to report the number of family members, education level, occupation, and age of the head of the family. they also summed up the monthly income (of millions) of the family head with one of the given random

scramble values and repeated this process five times. The students randomly selected one card from the deck of 50 cards and without anyone noticing, added the income of their family head to the number on the card and returned the card to the deck. The cards were then shuffled to maintain privacy, and only the sum of two values was reported. We repeated the process five times and reported the results for each repetition. The j-th RR for an i-th individual was as follows:

$Z_{ij} = Y_i + S_{ij} \quad i = 1, \dots, 512, \quad j = 1, \dots, 5.$

Considering the RR Model (case one), the MLEs of the mean and variance of the family head income were obtained as $\hat{\mu}_Y = 3.50662$ and $\hat{\sigma}^2 = 1.97$, respectively. The explanatory variables include the number of family members, the age of the family head, the level of education (coded as a binary variable: 1 for university attendance and 0 for non-attendance), and the occupation of the family head. Occupation is treated as a nominal variable with five categories: "others" (used as the reference level), "self-employed," "doctor," "engineer," and "retired or deceased".

The results summarized in Table 1 indicate that the number of family members and the age of the family head were not statistically significant. Considering "others" as the reference level for the occupation, levels of "doctor" and "engineer" had a significant impact on income compared to employees. The results also showed that having university attendance compared to non-attendance led to a significant increase in income.

Table 1 also shows that the family head jobs "doctor" and "engineer" had a significantly increasing effect on family head income compared to "others". However, "self-employed" and "retired" did not significantly affect family head income compared with "others". Our findings indicated that the variance of the sensitive variable was estimated at 1.83.

For RR model 1 (case two), the estimated mean and variance of the family head income were $\hat{\mu}_Y = 3.579$ and, $\hat{\sigma}^2 = 1.932$, respectively. The estimated parameters and their significant levels are presented in Table 2 where the estimated variance of the sensitive variable is 2.09.

The results of Tables 1 and 2, are consistent with previous ones; however, the standard error of estimates decreased (Table 2).

3. Simulation Study

For the models presented in Section 2, simulation and comparison were conducted using privacy criteria. Let $\beta_0 = 5$ and $\beta_1 = 2$, and the covariate X and model error ε were generated from normal $N(1,4)$ and $N(0,1)$, respectively. Therefore, the sensitive variable had a normal distribution of $N(5 + 2x, 1)$. On the other hand, the distribution of the scramble variables must be such that their mean falls within the parameter space of the

Table 1. Estimated Parameters of the RR model (case one)

Parameter		Coefficient	SE	p-value
Intercept		2.46	0.65	< .001
Age		0.0091	0.01	0.414
Education	non-attendance	-	-	-
	university attendance	1.13	0.2	< .001
Family number		-0.04	0.06	0.447
Occupation of	others	--	--	--
the family head	self-employed	-0.16	0.21	0.449
	Doctor	2.6	0.48	< .001
	Engineer	2.37	0.37	< .001
	Retired	0.1	0.35	0.765

Table 2. Estimated Parameters for averaged RR (case two).

Parameter		Coefficient	SE	95% CI	t -value	p-value
Intercept		2.16	0.62	[1.18, 3.75]	3.51	< .000
Age		0.017	0.01	[-0.01, 0.03]	1.6	0.11
Education	Non-attendance	-	-	-	-	-
	University attendance	1.146	0.19	[0.73, 1.53]	5.96	< .001
Family number		-0.042	0.055	[-0.16, 0.07]	-0.76	0.45
Occupation of the	Others	--	--	--	--	--
family head	self-employed	-0.2	0.2	[-0.57, 0.25]	-1	0.317
	Doctor	2.77	0.46	[1.65, 3.55]	6.08	< .001
	Engineer	2.28	0.35	[1.64, 3.11	6.45	< .001
	Retired	0.076	0.33	[-0.57, 0.78]	0.23	0.815

Table 3. MSE and bias of parameter Estimations for additive-scrambled RR.

	n	σ_Y	β_1	β_0
Est.	100	0.944	2	4.995
Bias		-0.056	0?	-0.005
MSE		0.046	0.019	0.104
Est.	50	0.893	1.993	5.021
Bias		-0.107	-0.007	0.021
MSE		0.11	0.04	0.219
Est.	20	0.817	1.993	5.005
Bias		-0.183	-0.007	0.005
MSE		0.177	0.07	0.378

Table 4. MSE and bias of parameter Estimations for averaged-additive-scrambled RR.

	n	σ_Y	β_1	β_0
Est.	100	0.984	2.001	5
Bias		-0.016	0.0006	0.0004
MSE		0.0094	0.0095	0.049
Est.	50	0.96	2.001	4.998
Bias		-0.04	0.0008	-0.002
MSE		0.019	0.019	0.105
Est.	20	0.9	2.006	4.98
Bias		-0.101	0.006	-0.021
MSE		0.052	0.054	0.279

sensitive variable. The parameters were estimated using the maximum likelihood method. The simulation was repeated K times, and the results included the average parameter estimatesand the bias and mean squared error (MSE) of these estimates.

The simulations are as follows:

3.1 We considered the additive model with one scramble variable. This variable was sampled from normal $N(6,4)$. We consider $a = 3$ and $b = 2$, so the RR variable has a normal distribution of $Z \sim N(27 + 6x, 25)$. For $m = 5$ times repeat of RR for each individual, the averaged RR variable has a normal distribution of $\bar{Z} \sim N(27 + 6x, 12.2)$.

Tables 3 and 4 present the simulation results for $k = 2000$ repetitions for both RR and averaged RR models, respectively.

3.2. We considered the additive-scrambled-scrambled RR model with two scramble variables. The scramble data were generated from a normal distribution of $S_1 \sim N(6,4)$ and$S_2 \sim N(8,16)$. Setting $a = 3, b = 2$ and, $c = 2$, the RR variable $Z = aY + bS_1 + cS_2$ had a normal distribution of $N(43 + 6x, 89)$. The mean of m = 5 times the repeat of RR for each individual had a normal distribution of $\bar{Z} \sim N(27 + 6x, 25)$. Simulation results for both cases are provided in Tables 5 and 6, respectively.

3.3. Given a normal distribution $N(6,4)$, and a sensitivity level of 0.6, we used an optional RRT model. So, the probability of answering the sensitive variable was $p = 0.4$. The regression model is as follows:
$y_i = 5 + 2x_i + \varepsilon_i$, $\quad i = 1, \dots, n$, $\varepsilon \sim N(0,1)$.

Table 5. MSE and bias of parameter Estimations for additive-scrambled-scrambled RR.

	n	σ_Y	β_1	β_0
Est.	100	0.793	1.997	5.012
Bias		-0.207	-0.003	0.012
MSE		0.398	0.0675	0.364
Est.	50	0.731	2.003	4.986
Bias		-0.27	0.003	-0.014
MSE		0.55	0.138	0.752
Est.	20	0.707	2.016	4.938
Bias		-0.293	0.016	-0.061
MSE		0.662	0.254	1.32

Table 6. MSE and bias of parameter Estimations for averaged-additive-scrambled-scrambled RR.

	n	σ_Y	β_1	β_0
Est.	100	0.953	2	5.001
Bias		-0.047	0?	0.001
MSE		0.043	0.019	0.103
Est.	50	0.893	2.008	4.975
Bias		-0.106	0.0078	-0.025
MSE		0.102	0.04	0.211
Est.	20	0.743	1.991	5.037
Bias		-0.257	-0.009	-0.037
MSE		0.243	0.114	0.607

The optional RR variable had a mean of $8.6 + 2X$ and variance of 12.04. The mean of RRs for $m = 5$ observations per individual had the same mean and a variance of 10.12. Parameter estimatesand their MSE and biases are provided for k =2000 repetitions in Tables 7 and 8.

3.4. Finaly, simulation results were provided for the multiplicative RR. The scramble variable data were sampled from $N(6,4)$, so the mean and variance of the multiplicative RR variable were $\mu_z = 30 + 12x$ and$\sigma_z^2 = 16(5 + 2x)^2 + 52$, respectively.

The mean of the multiplicative RRs for $m = 5$ observations per individual had the same mean and variance $\sigma_z^2 = 3.2(5 + 2x)^2 + 39.2$. The simulation results are provided for k=2000 repetitions in Tables 9 and 10.

According to the simulation results, the maximum likelihood estimates were very close to the true values with high accuracy. Moreover, as the number of simulated data, n, increases, the accuracy of estimates improves, whereas the variance and bias of the estimates decrease.

3.5 Privacy criteria

The privacy criterion, P_L, (the privacy level), is the mean squared difference between the RR, Z, and the true response Y or $P_L = E(Z - Y)^2$. The measure $\delta = \frac{V(\hat{\mu})}{P_L}$ was proposed for comparing quantitative RR methods (42). The privacy evaluation criteria for single and repeated observations of techniques are presented in Table 11. For each n, the techniques in terms of the P_L criterion are sorted as follows:

The technique with two scramble variables is the best, and after that, the techniques are sorted as follows: The technique with one scramble variable, the multiplicative technique, and the optional technique. However, considering the δ criterion, the multiplicative technique came first, followed by the techniques with one and two

Table 7. MSE and bias of parameter Estimates for optional RR.

	n	σ_Y	β_1	β_0
Est.	100	0.791	2.005	5
Bias		-0.209	0.005	-0.001
MSE		0.318	0.079	0.444
Est.	50	0.686	2.001	5.006
Bias		-0.314	0.001	0.006
MSE		0.422	0.17	0.911
Est.	20	0.615	1.984	5.029
Bias		-0.385	-0.016	0.029
MSE		0.521	0.301	1.61

Table 8. MSE and bias of parameter Estimates for averaged-optional RR.

	n	σ_Y	β_1	β_0
Est.	100	0.924	2.001	4.998
Bias		-0.076	0.001	-0.002
MSE		0.066	0.022	0.116
Est.	50	0.849	2.005	4.991
Bias		-0.15	0.005	-0.008
MSE		0.138	0.048	0.246
Est.	20	0.717	2.007	4.985
Bias		-0.283	0.007	-0.015
MSE		0.276	0.129	0.69

Table 9. MSE and bias of parameter Estimates for multiplicative RR.

	n	σ_Y	β_1	β_0
Est.	100	3.246	2.002	4.998
Bias		2.246	0.002	-0.002
MSE		0.078	0.084	0.27
Est.	50	3.18	2.003	4.988
Bias		2.18	0.003	-0.012
MSE		0.145	0.166	0.57
Est.	20	3.039	1.993	4.983
Bias		2.038	-0.007	-0.017
MSE		0.352	0.5	1.73

Table 10. MSE and bias of parameter Estimates for averaged-multiplicative RR.

	n	σ_Y	β_1	β_0
Est.	100	1.61	1.999	5.008
Bias		0.613	-0.001	0.008
MSE		0.017	0.02	0.08
Est.	50	1.695	1.992	5.014
Bias		0.695	-0.008	0.014
MSE		0.034	0.048	0.18
Est.	20	1.67	1.997	5.01
Bias		0.67	-0.003	0.01
MSE		0.088	0.124	0.511

Table 11. Privacy criteria of the RR techniques.

	n	Privacy evaluation criteria	Results from one observation		Results for the mean of $m = 5$ observations	
			Mean	Var.	Mean	**Var.**
$Z = aY + bS,$	20	P_L	943.1	5209.87	928	**5719.02**
		δ	0.0016	3×10^{-8}	-0.0011	$\mathbf{55 \times 10^{-7}}$
	50	P_L	944.36	3193.87	932.19	**2280.68**
		δ	0.00098	54×10^{-9}	-0.00045	$\mathbf{18 \times 10^{-9}}$
	100	P_L	943.58	1597.88	930.127	**1155.3**
		δ	0.00049	69×10^{-10}	-0.00012	$\mathbf{2 \times 10^{-9}}$
$Z = aY + bS_1 + cS_2,$	20	P_L	2214.71	31578.69	2165.62	**17363.07**
		δ	0.0018	14×10^{-8}	0.0011	$\mathbf{15 \times 10^{-8}}$
	50	P_L	2219.29	17797.91	2159.43	**7626.24**
		δ	0.0011	29×10^{-9}	0.00045	$\mathbf{11 \times 10^{-9}}$
	100	P_L	2225.44	9425.88	2158.821	**3702.596**
		δ	0.00054	42×10^{-9}	0.00023	$\mathbf{13 \times 10^{-10}}$
$Z = YG + (Y + S)(1 - G),$	20	P_L	23.93	24.81	15.17	**6.13**
		δ	0.111	0.00016	0.04	**0.0001**
	50	P_L	24.05	14.77	15.14	**2.5**
		δ	0.066	37×10^{-6}	0.017	$\mathbf{10 \times 10^{-6}}$
	100	P_L	24.1	7.61	15.16	**1.21**
		δ	0.033	64×10^{-6}	0.009	$\mathbf{14 \times 10^{-7}}$
$Z = YS$	20	P_L	88.32	122.42	2280.34	**126752**
		δ	0.0004	39×10^{-10}	0.0001	$\mathbf{46 \times 10^{-11}}$
	50	P_L	87.75	47.41	2266.81	**47221.5**
		δ	0.00015	29×10^{-11}	0.00004	$\mathbf{29 \times 10^{-12}}$
	100	P_L	87.88	23.23	2271.15	**22501.78**
		δ	0.00008	37×10^{-12}	0.00002	$\mathbf{38 \times 10^{-13}}$

scramble variables, and the optional technique was the last one.

For techniques with averaged RRs, the best-performing technique in terms of the P_L criterion was the multiplicative technique.The technique with two scramble variables was the second one, followed by the technique with a single scramble variable.The last was the optional technique. The behavior of the δ criterion

for this case is consistent with the single-observation case.

When comparing privacy criteria between single-observation models and models with averaged RRs, the P_L criterion significantly increased for the multiplicative model with averaged observations. Models with one and two scramble variables showed a slight reduction in P_L, while the optional model had nearly a halving of P_L. The δ criterion favors single-observation responses across all techniques, emphasizing the preference for models with averaged RRs.

Results and Discussion

In social surveys, when studying a sensitive variable, respondents may refuse to answer questions or provide socially desirable responses. The RR techniques help mitigate this issue. The RRR technique is one approach that increases privacy levels while moderating the increase in estimates variance. When studying continuous RR data, collecting multiple observations from each increases the sample size and improves parameter estimates. Averaging the observations for each respondent helps achieve more precise estimations. Linear models are applied for the mean of observations. The findings of this study demonstrate that the averaged RRs for each individual in various RR techniques yield more accurate estimations and reduce their variance.

In the study of the family head income, modeling the RR techniques are evaluated with demographic variables, including the number of family members and age, education level, and occupation of the family head. The results of the averaged RR model indicate that the number of family members and the age of the family head are not statistically significant. Levels of “doctor” and “engineer” of occupation variable, have a significant impact on income compared to the reference category, “others”. The results also show that having a university education may lead to a significant increase in income. This finding provides a valuable avenue for further investigations in this field.

References

1. Warner SL. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association. 1965;60(309):63-9.
2. Abul-Ela A-LA. Randomized Response Models for Sample Surveys on Human Population: University of North Carolina, Chapel Hill; 1966.
3. G. Horvitz BVS, Walt R. Simmons. The unrelated question randomized response modeld. Research Triangle Institute and -National Center for Health Statistics;1967.
4. Greenberg BG, Abul-Ela A-LA, Simmons WR, Horvitz DG. The unrelated question randomized response model: Theoretical framework. Journal of the American Statistical Association. 1969;64(326):520-39.
5. Moors J. Optimization of the unrelated question randomized response model. Journal of the American Statistical Association. 1971;66(335):627-9.
6. Boruch RF. Assuring Confidentiality of Responses in Social Research: A Note on Strategies. The American Sociologist. 1971;6(4):308-11.
7. Alavi SMR, Tajodini M. Maximum likelihood estimation of sensitive proportion using repeated randomized response techniques. Journal of Applied Statistics. 2016;43(3):563-71.
8. Maddala GS. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge: Cambridge University Press; 1983.
9. Scheers NJ, Dayton CM. Improved estimation of academic cheating behavior using the randomized response technique. Research in Higher Education. 1987;26(1):61-9.
10. Cruyff MJLF, Böckenholt U, van der Heijden PGM, Frank LE. Chapter 18 - A Review of Regression Procedures for Randomized Response Data, Including Univariate and Multivariate Logistic Regression, the Proportional Odds Model and Item Response Model, and Self-Protective Responses. In: Chaudhuri A, Christofides TC, Rao CR, editors. Handbook of Statistics. 34: Elsevier; 2016. p. 287-315.
11. Blair G, Imai K, Zhou Y-Y. Design and Analysis of the Randomized Response Technique. Journal of the American Statistical Association. 2015;110(511):1304-19.
12. Chang CH, Cruyff M, Giam X. Examining conservation compliance with randomized response technique analyses. Conserv Biol. 2018;32(6):1448-56.
13. Chang P-C, Pho K-H, Lee S-M, Li C-S. Estimation of parameters of logistic regression for two-stage randomized response technique. Computational Statistics. 2021;36(3):2111-33.
14. Hsieh S-H, Perri PF. A Logistic Regression Extension for the Randomized Response Simple and Crossed Models: Theoretical Results and Empirical Evidence. Sociological Methods & Research. 2022;51(3):1244-81.
15. Halim A, Arshad IA, Alomair AM, Alomair MA. Estimation of hidden logits using several randomized response techniques. Symmetry. 2023;15(9):1636.
16. Sayed KH, Cruyff MJ, van der Heijden PG. The analysis of randomized response “ever” and “last year” questions: A non-saturated Multinomial model. Behavior Research Methods. 2024;56(3):1335-48.
17. Chaudhuri A, Christofides TC, Chaudhuri A, Christofides TC. Randomized response techniques to capture qualitative features. Indirect questioning in sample surveys. 2013:29-94.
18. Trappmann M, Krumpal I, Kirchner A, Jann B. Item sum: a new technique for asking quantitative sensitive questions. Journal of Survey Statistics and Methodology. 2014;2(1):58-77.
19. Qureshi MN, Balqees S, Hanif M. Mean estimation of scrambled responses using systematic sampling. 2022.
20. Khalil S, Zhang Q, Gupta S. Mean estimation of sensitive variables under measurement errors using optional RRT

models. Communications in Statistics-Simulation and Computation. 2021;50(5):1417-26.
21. Tiwari KK, Bhougal S, Kumar S, Rather KUI. Using randomized response to estimate the population mean of a sensitive variable under the influence of measurement error. Journal of Statistical Theory and Practice. 2022;16(2):28.
22. Gupta S, Qureshi MN, Khalil S. Variance estimation using randomized response technique. REVSTAT-Statistical Journal. 2020;18(2):165–76–-76.
23. Gupta S, Zhang J, Khalil S, Sapra P. Mitigating lack of trust in quantitative randomized response technique models. Communications in Statistics-Simulation and Computation. 2024;53(6):2624-32.
24. Azeem M, Ijaz M, Hussain S, Salahuddin N, Salam A. A novel randomized scrambling technique for mean estimation of a finite population. Heliyon. 2024;10(11).
25. Khan M. A randomized response technique for reducing the effect of initial non-response of the regression estimator in panel surveys. 2021.
26. Ahmed S, Shabbir J. On use of randomized response technique for estimating sensitive subpopulation total. Communications in Statistics-Theory and Methods. 2023;52(5):1417-30.
27. Kivaranovic D, Leeb H. A (tight) upper bound for the length of confidence intervals with conditional coverage. arXiv preprint arXiv:200712448. 2024.
28. Greenberg BG, Kuebler RR, Abernathy JR, Horvitz DG. Application of the Randomized Response Technique in Obtaining Quantitative Data. Journal of the American Statistical Association. 1971;66(334):243-50.
29. Warner SL. The Linear Randomized Response Model. Journal of the American Statistical Association. 1971;66(336):884-8.
30. Eichhorn BH, Hayre LS. Scrambled randomized response methods for obtaining sensitive quantitative data. Journal of Statistical Planning and Inference. 1983;7(4):307-16.
31. Gupta S, Gupta B, Singh S. Estimation of sensitivity level of personal interview survey questions. Journal of Statistical Planning and Inference. 2002;100(2):239-47.
32. Gupta S, Kalucha G, Shabbir J. A regression estimator for finite population mean of a sensitive variable using an optional randomized response model. Communications in Statistics - Simulation and Computation. 2017;46(3):2393-405.
33. Singh N, Vishwakarma G, Kumar N, Singh PH. Estimation of Mean of Sensitive Variable Using Multiplicative Scramble Variable Under Measurement Error. Journal of Statistical Theory and Practice. 2022;16:46.
34. Cao M, Breidt FJ, Solomon JN, Conteh A, Gavin MC. Understanding the drivers of sensitive behavior using Poisson regression from quantitative randomized response technique data. PloS one. 2018;13(9):e0204433.
35. Singh S, Tracy DS. Ridge regression using scrambled responses. Metron-International Journal of Statistics. 1999:147-57.
36. Strachan R, King M, Singh S. Theory and Methods: Likelihood-based Estimation of the Regression Model with Scrambled Responses. Australian & New Zealand Journal of Statistics. 1998;40(3):279-90.
37. Singh S, Joarder A, King M. Regression analysis using scrambled responses. Australian Journal of Statistics. 2008;38:201-11.
3. Rueda MD, Cobo B, Arcos A. Regression Models in Complex Survey Sampling for Sensitive Quantitative Variables. Mathematics [Internet]. 2021; 9(6).
39. Arcos A, Rueda MdM, Singh S. A generalized approach to randomised response for quantitative variables. Quality & Quantity. 2015;49(3):1239-56.
40. Fox J-P, Veen D, Klotzke K. Generalized Linear Mixed Models for Randomized
Responses. Methodology. 2018;15(1):1-18.
41. Yan Z, Wang J, Lai J. An Efficiency and Protection Degree-Based Comparison Among the Quantitative Randomized Response Strategies. Communications in Statistics - Theory and Methods. 2008;38(3):400-8.
42. Gupta S, Mehta S, Shabbir J, Khalil S. A unified measure of respondent privacy and model efficiency in quantitative RRT models. Journal of Statistical Theory and Practice. 2018;12:506-11.
43. Anthony YCK. Asking Sensitive Questions Indirectly. Biometrika. 1990;77(2):436-8.
44. Arellano-Valle R, Bolfarine H, Lachos V. Skew-normal Linear Mixed Models. 2004;3.
45. Arellano-Valle RB, Genton MG. On fundamental skew distributions. Journal of Multivariate Analysis. 2005;96(1):93-116.