A randomized response technique for reducing the effect of initial non-response of the regression estimator in panel surveys

Mursala Khan[†]

[†]*Department of Mathematics and Statistics, Riphah International University, I-14, Islamabad, Pakistan. .*

Email: mursalakhan04@gmail.com, mursala.khan@riphah.edu.pk

Abstract

The results of Sample surveys play a vital role in decision making. One of the main issues being faced by survey statisticians during the collection of survey data is the problem of non-response which may affect survey cost and accuracy of estimates. The problem of non-response becomes more severe if the survey contains sensitive questions like related to family planning methods, use of drugs. To diminish the non-response rate arising in the case of direct questioning (DQ) technique, Warner (1965) proposed an indirect survey technique known as the randomized response (RR) technique. He addressed this problem for a cross-sectional data. This method is a well-known procedure that produces more valid responses on sensitive questions in surveys. The method avoids the direct link between respondent's response and the sensitive question through the help of a randomization device. Thereby protecting respondent's privacy which in turn greatly increases survey response rate. However, due to the complex nature of panel estimator, the work is missing the in the context of RR technique. To cover this gap, we propose a linear regression model in the context of panel surveys/longitudinal studies under the application of the RR technique. We solve all these issues through simulation study.

Keywords: Panel surveys, non-response, randomized response, linear model, fade-away effect.

1. Introduction

Whenever we are talking about surveys or survey sampling, the first word which ultimately comes to our minds is the word “non-response”. Non-response not only condenses case numbers of the data but may lead to substantial bias in the estimates (descriptive and causal parameters). This holds for both cross-sectional surveys as well as panel surveys. However, in panel surveys, the case numbers or sample size is further reduced by panel attrition which happens after the initial wave the so-called successive drop-out of respondents in later panel waves. The cumulative dropping of initial non-response and attrition are increasing with the length of the panel. In general, if attrition is selective for the response then it may worsen the size of initial non-response bias in later panel waves. On the contrary, if the attrition is non-selective for response or if the dropout rate is low compared to initial non-response, then under this aspect an initial non-response bias may reduce in later panel waves. It was the interesting findings of Sisto (2003) and Rendtel (2013) that in panel surveys sampled from registers the estimates obtained from the initially intended sample persons and the respondent sample (only respondents) converged to the true solution in subsequent panel waves. They did it for income quintiles and poverty states by selecting a sub-sample from the European Community Household Panel (ECHP) and poverty rates from the European Union Statistics of Income and Living Conditions (EU-SILC). Rendtel (2013) used the term “fade-away” for the initial non-response. The idea of the fade-away of bias goes back to the earlier study of Fitzgerald et al. (1998) who find out that without any correction for the initial response or for attrition the distortion effect between the estimates from the Panel Study of income dynamics (PSID) and counts from the current US population survey diminished in subsequent panel waves.

Motivated from the empirical studies of Sisto (2003) and Rendtel (2013) where information about the non-respondents is taken from registers Alho et al. (2017) extends the fade-away phe-

nomenon in the context of the Markov chains model. In their influential article, the main tool is the contraction theorem which states that under some regularity conditions the distributional differences between two chains fade-away over time. Alho and Rendtel (2020) conducted a similar study in the context of regression analysis. To explain the fade-away of initial non-response bias they proposed a linear panel model of general length ($t = \text{time} = \text{wave}$) in the framework of regression analysis having a single covariate. In their proposed model the covariate and the error term are decomposed into two parts: the permanent parts and the time-varying parts. They concluded that a large initial bias of the OLS estimates due to non-response fades-away in the case of low dynamic components of the covariate and error terms. They proposed an approximation formula for the bias of the OLS estimates of the true slope coefficient under the not missing at random (NMAR) mechanism at the start of a panel survey. However, the approximations formula becomes very intractable in later panel waves of a panel survey.

To avoid such difficulty arising in the case of approximation formula, Khan (2020) extend the results to a longer panel wave of a panel survey using a simulation study. Using the linear model of Alho and Rendtel (2020) for the response, Khan (2020) also concluded that if the size of the time-invariant and dynamic components of the covariate and error terms is small then large initial bias of the OLS estimates fades-away in later panel waves. In the application part, Khan used income and satisfaction data from SOEP which are also completely in line with the theoretical results of Alho. Surveys where interviews are conducted directly are relatively easy than the indirect survey techniques in terms of cheaper survey cost and faster way of obtaining information. However, it may cause a substantial non-response rate, which distorts the distribution of interest, e.g., from normal to non-normal, and also have the potential to bias statistical analysis. One of the main sources of substantial non-response in direct surveys is the sensitivity of survey questions. Where respondents are reluctant to answer certain survey questions due to privacy concerns. Questions such as abortion, the use of the drug, tax evasion, black money, divorce, and plagiarism, or any other sensitive issues, etc., are considered sensitive. If a survey questionnaire contains such type of sensitive questions then applying direct survey methods will cause substantial bias in estimates due to non-response.

To cover this gap in the case of cross-sectional data, Warner (1965) proposed an alternative indirect survey technique known as the randomized response (RR) technique. This method reduces the extent of large non-response arising in DQ technique by protecting the respondent's privacy when answering a certain sensitive question. This method, therefore, allows interviewers to obtain more accurate information on sensitive issues while guaranteeing the privacy of respondents. In this method the respondents use a randomization device, such as selecting a card, spinning a spinner, or tossing a coin whose result is hidden from the interviewer. Thus by introducing random noise in their responses through a randomization device, the RR method conceals respondent responses and protects their privacy. As privacy is of great concern for survey respondents, in direct response surveys their privacy is nowhere protected resulting in a large part of missing data due to non-response. It is also possible that they may provide false answers which may either under/over estimates true parameters. On the other hand, when they are certain that their privacy will be protected through randomization, they are intended to answer more truthfully. This also includes non-respondents who refuse or reluctant to answer a certain sensitive question in the DQ technique thus increasing survey response rate. As the main purpose of RR technique is to increase the survey response rate, once it is increased through randomization it yields more reliable survey estimates on sensitive issues than the estimates provided by direct survey methods. The basic idea of Warner RR model is to estimate the proportion of respondents in a dichotomous population whose truthful response to a certain sensitive question would be "yes", without exposing the respondent's privacy to the interviewer and therefore avoiding social stigma or fear of reprisals. To increase survey response rate Warner uses a randomization device, one such device could be e.g., a deck of cards with each card having one of the following two statements:

- I belong to group A (group that contain sensitive questions),
- I do not belong to group A (non-sensitive group).

Each survey participants performs a random trial consisting of sensitive questions to whom they pretended to respond e.g., (i) I never pay the income tax (i.e., the respondents belong to group A);

(ii) I pay the income tax (i.e., the respondents doesn't belongs to group A). The above statements have the probabilities P and $1 - P$, in the deck of cards. These relative frequencies are also known as the design probabilities (DP) $0 \leq P \leq 1$. P is the probability of success that the randomizing device provides instructions to the respondent to responds to the real question. The purpose is to estimate π_A ($0 < \pi < 1$) the population proportion of respondents belong to the sensitive group A. This is also known as the true probability of selecting the first true answer to the real question. Under this aspect, a sample of size n is selected from the population of size N through simple random sampling with or without replacement schemes. The member of the selected sample is asked to select a card at random from a pack of ordinary playing 52 cards. Due to the sensitivity of the survey question or due to privacy concerns the respondent is advised not to show the cards to the interviewer. Suppose out of n survey respondents, n_1 respondents are those who say "Yes" answer to a sensitive question. It is assumed that the distribution of n_1 responses follows binomial distribution having parameters n and $P_y = P\pi_A + (1 - P)(1 - \pi_A)$ i.e., $n_1 \sim N(n, P_y)$.

Warner obtained an estimator (maximum likelihood which is unbiased) for π_A which exists for $P \neq 0.5$ and is defined by

$$\hat{\pi}_A(w) = \frac{n_1/n - (1 - P)}{2P - 1}. \quad (1)$$

with variance

$$\begin{aligned} V(\hat{\pi}_A(w)) &= \frac{\pi_A(1 - \pi_A)}{n} + \frac{P(1 - P)}{n(2P - 1)^2} \\ &= V(\hat{\pi}_A(D)) + \frac{P(1 - P)}{n(2P - 1)^2}. \end{aligned} \quad (2)$$

Where the notation D stands for the direct response (DR) survey and $V(\hat{\pi}_A(D)) = \frac{\pi_A(1 - \pi_A)}{n}$ is the variance of the estimator in a DR survey.

From the above discussion, we concluded that estimates obtained through the Warner RR model estimator are unbiased and of course more consistent with the true population parameters than estimates produced by DR surveys. Nevertheless, in the indirect response surveys, one has to pay more cost in terms of variance. This is because RR survey techniques are most cost-effective

and put more burden on the respondents in protecting the privacy of the respondents. This can be seen from variance equation (2) that for $P = 0$ the variance of the estimator reduces to the variance of the DR survey i.e., $V(\hat{\pi}_A(W)) = \text{Var}(\hat{\pi}_A(D))$. For all other values of P ($0 < P < 1$) the variance of the Warner model estimator will always be larger than the variance of the estimator provided by the DR survey.

After the pioneering work of Warner on the RR model, numerous alternative techniques have been incorporated in the standard RR models to estimate the mean responses of respondents bearing a sensitive group. For more details introduction see e.g., Greenberg et al. (1969), Moors (1971), Boruch (1972), Frankiln (1989), Kuk (1990), Arnab (1996) Chaudhuri (2001), Fenton et al. (2001), Gjstvang and Singh (2006), Van den Hout et al. (2007), in the survey literature on RR models. These RR techniques or procedures are incorporated to estimate the mean responses of respondents belong to the sensitive group of a qualitative random variable (Bernoulli or Multinomial). However, in the survey where the outcome variable is continuous and where the objective is to estimate the average responses then these standard RR models are not the appropriate survey techniques. To overcome this gap Eichhorn and Hayre (1983) proposed a new RR model estimator for the estimation of the population mean of the sensitive quantitative response variable. Such type of models are very beneficial in studies where the response variable is extremely sensitive in its nature. For example, studies where the focus is on addressing issues e.g., i) Not only whether or not a particular woman had gone an abortion, but in addition, how many number of abortions she underwent; ii) Not only whether or not an individual had used condom, but also how many times he/she used condoms; iii) Not only whether a medical student used illicit drugs, but also the number of occasions in which he/she taken drugs; vi) Not only if an individual received black money, but in addition how many time he/she received such money, more such type of example can be seen in Bar-Lev et al. (2004) and references therein.

According to their procedure, respondents are asked about their value of the sensitive quantitative response variable. They are also informed that if they feel that the underlying survey question is sensitive then respond with coded value comprising of their true response value. The coded response (scrambled response) is obtained by multiplying the true response value by some

random number which can be drawn from a given distribution such as from normal or exponential distribution. Due to the randomization of response, the interviewer is unaware or unable to track the random number which is used by the respondent. However, he/she knows the underlying parent distribution to which the random numbers are generated for coding responses.

To understand this procedure let the Y be a sensitive random variable whose outcome we want to estimate. Further, let Z be another random variable that is used for the random numbers that is used for the coding/scrambling response. It is also assumed that both these variables are independently distributed having coded response $Y^* = YZ$ provided by the respondent to a sensitive question. The expectations of these variable are denoted by $\mu_y = E(Y)$, and $\mu_z = E(Z)$, and their variances by $\sigma_y^2 = V(Y)$, and $\sigma_z^2 = V(Z)$, respectively. The coefficient of variations of these variables are denoted by C_y and C_z . Under this aspect the coded response variable $Y^* = YZ$ has an expectation $E(Y^*) = \mu_y\mu_z$ and variance $V(Y^*) = \sigma_y^2\mu_z^2 + \mu_y^2(1 + C_y^2)\sigma_z^2$, respectively. Then using the coded responses of size n survey respondents Eichhorn and Hayre (1983) proposed an unbiased estimator for the estimation of population mean μ_y of the un-coded response variable Y , is given by:

$$\hat{\mu}_{y(EH)} = \frac{\hat{\mu}_{y^*}}{\mu_z}, \quad (3)$$

where $\hat{\mu}_{y^*} = \frac{1}{n} \sum_{i=1}^n Y_i^*$ is the sample mean of coded responses of size n . The estimator $\hat{\mu}_{y(EH)}$ is unbiased and its variance is given as follows:

$$V(\hat{\mu}_{y(EH)}) = \frac{1}{n} [\sigma_y^2 + \mu_y^2(1 + C_y^2)C_z^2]. \quad (4)$$

Which has larger variance than the variance of the estimator for population mean in simple random sampling with open or DR survey, i.e., $V(\bar{y}) = \sigma_y^2/n$. More detailed discussion can be seen in Eichhorn and Hayre (1983).

Bar-Lev et al. (2004) proposed an estimation procedure for the estimation of the population mean of the sensitive variable, which is a generalization of Eichhorn and Hayre (1983) RR procedure. This RR procedure has several positive features which it exploits for the probability

of response i.e., it incorporates two mechanisms of two different models (qualitative response model, quantitative response model). Firstly, it uses the randomizing parameter which is used in the Warner (1965) model the so-called “design parameter”. Secondly, it uses the coding mechanism of the Eichhorn and Hayre (1983) RR model. Due to the implementation of these positive RR features for randomization of responses this procedure, therefore, produces quite promising results, which are more efficient than Eichhorn and Hayre (1983) RR procedure in terms of lower variance. Let us consider the previously used variables Y , Z , and a design parameter P ($0 < P < 1$). For the probability of response, Bar-Lev et al. (2004) assume that with probability P the respondents respond to the true answer of the response variable Y , whereas with probability $1 - P$ he/she responds with coded or scrambling variable YZ . Under this respect the layout of the individual response to a sensitive question can be written as:

$$Y^* = \begin{cases} Y, & \text{with probability } P, \\ YZ, & \text{with probability } (1 - P). \end{cases} \quad (5)$$

The coded response variable Y^* has expectation $E(Y^*) = \mu_y(P + (1 - P)\mu_z)$ and variance $V(Y^*) = \mu_y^2[(1 + C_y^2)\{P + (1 - P)\mu_z^2(1 + C_z^2)\} - \{P + (1 - P)\mu_z\}^2]$. For $P = 0$, this procedure reduces to that of Eichhorn and Hayre (1983) RR procedure.

Then according to Bar-Lev et al. (2004) an unbiased estimator $\hat{\mu}_y$ for μ_y and its variance are obtained as follows:

$$\hat{\mu}_{y(BBB)} = \frac{\hat{\mu}_y^*}{P + (1 - P)\mu_z}, \quad (6)$$

with a sampling variance of $\hat{\mu}_{y(BBB)}$ is given by:

$$V(\hat{\mu}_{y(BBB)}) = \frac{1}{n}[\sigma_y^2 + \mu_y^2(1 + C_y^2)C_{zp}^*], \quad (7)$$

where, $C_{zp}^* = \frac{P + (1 - P)E(Z^2)}{P + (1 - P)\mu_z^2} - 1$. More details about C_{zp}^* can be seen in Bar-Lev et al. (2004) and some latest references on RR techniques.

2. The fade-away of bias in regression analysis

Alho and Rendtel (2020) extended the fade-away phenomenon into time series regression analysis. They proposed a linear model of the panel over time t (=wave), which takes the form:

$$Y_{i,t} = a_t + b_t X_{i,t} + e_{i,t}, \quad \text{for } t = 1, 2, 3, \dots, T. \quad (8)$$

It is evident from the model that the relationship between the response variable $Y_{i,t}$ and the covariate or predictor variable $X_{i,t}$ varies over time t . a_t is the intercept, and b_t coefficient. The disturbance terms $e_{i,t}$ satisfy the basic assumptions of the classical linear regression model having means $E(e_{i,t}) = 0$ and constant variances $Var(e_{i,t}) = \sigma^2$. The covariate $X_{i,t}$ consists two components: the fixed permanent component M_i and a time-varying components $Z_{i,t}$ which is expressed in the following form:

$$X_{i,t} = M_i + Z_{i,t},$$

These two components of the covariate $X_{i,t}$ are assumed to be independent from each other and having their expectations $E(M_i) = E(Z_{i,t}) = 0$, and variance $Var(M_i) = \kappa$ and $Var(Z_{i,t}) = 1 - \kappa$, where $0 \leq \kappa \leq 1$.

In addition, the disturbance terms $e_{i,t}$ of the model has a variance component structure which is decomposed into two components: the permanent component V_i and the transient components $U_{i,t}$ which has the following structure:

$$e_{i,t} = V_i + U_{i,t},$$

Following a similar argument, these components of the error term is independent from each other with expectations $E(V_i) = E(U_{i,t}) = 0$ and variances $Var(V_i) = \gamma\sigma^2$ and $Var(U_{i,t}) = (1 - \gamma)\sigma^2$, where $0 \leq \gamma \leq 1$.

The covariate $X_{i,t}$ is assumed to be independent with the model disturbance terms $e_{i,t}$, where it

also mean that M , Z , V , and U are independent from each other. As the covariate and the error terms of the model have a time varying components, it is therefore, assumed that they follow an auto-regressive model of order one i.e., AR(1) which is defined as follows:

$$Z_{i,2} = \rho Z_{i,1} + \varepsilon_{i,2}, \quad (9)$$

and

$$U_{i,2} = \phi U_{i,1} + \xi_{i,2}. \quad (10)$$

where ρ and ϕ are the coefficient of auto-covariance of the time-varying parts of the covariates and the residual factors, and where $\varepsilon_{i,2}$ and $\xi_{i,2}$ are the model “noise terms” which satisfy all the assumptions of the classical linear regression model i.e., $E(\varepsilon_{i,2}) = E(\xi_{i,2}) = 0$ with variances $Var(\varepsilon_{i,2}) = (1 - \kappa)(1 - \rho^2)$ and $Var(\xi_{i,2}) = (1 - \gamma)(1 - \phi^2)\sigma^2$.

Regarding the participation in the survey the proposal is that, at the beginning of the survey (wave 1=time 1), an individual decides whether he/she would participate in the survey or not. Let $R_{i,1} = 1$ be the binary response indicator of the i person who is willing to participate in the survey initially, while for $R_{i,1} = 0$ we have a non-response or stated otherwise. Then using the linear model of Alho and Rendtel (2020), the response probability of a person i is defined by

$$P(R_{i,1} = 1 | Y_{i,1}) = \alpha + \beta Y_{i,1}, \quad (11)$$

where $0 < \alpha < 1$ and $0 < \beta < 1$ are the non-response parameters. The values of these parameters are be selected in a such a manner such that the resulting value from the response model (11) falls within the interval $[0, 1]$. It is also worthwhile to mention that the corresponding response model depends on $Y_{i,1}$, which means that the missingness is not random (NMAR), which states that the distribution of the variable of interest $Y_{i,1}$ at the initial wave 1 is highly selective for the

response. Then the marginal probability of response in initial wave is then approximated by

$$\begin{aligned} P(R_{i,1} = 1) &= E[P(R_{i,1} = 1|Y_{i,1})] \\ &= \alpha + \beta a_1. \end{aligned} \tag{12}$$

where $a_1 = E(Y_{i,1})$. The following approximation results are due to Alho and Rendtel (2020).

The OLS estimates of b_t of those who initially participate in the survey is approximated by:

$$\begin{aligned} p \lim_{n \rightarrow \infty} (\hat{b}_{t,com}) &\approx Cov(X_t, Y_t | R_t = 1) / Var(X_t | R_t = 1) \\ &\approx b_t \left(1 - \frac{\beta^2 \sigma^2 (\kappa + \rho^t (1 - \kappa)) (\gamma + \phi^t (1 - \gamma))}{(\alpha + \beta a_1)^2 - \beta^2 b_t^2 (\kappa + \rho^t (1 - \kappa))^2} \right), \end{aligned}$$

with a non-response which is stated by

$$Bias(\hat{b}_{t,com}) = \frac{\beta^2 \sigma^2 (\kappa + \rho^t (1 - \kappa)) (\gamma + \phi^t (1 - \gamma))}{(\alpha + \beta a_1)^2 - \beta^2 b_t^2 (\kappa + \rho^t (1 - \kappa))^2}. \tag{13}$$

Conclusion: Note that the under the above condition the bias at time t will be always smaller/lower than the bias at time $t - 1$ i.e., $Bias(\hat{b}_{t,com}) < Bias(\hat{b}_{(t-1),com})$, always holds true for all values of the time-invariant and dynamic components of the covariate and error terms in the interval $0 < \kappa < \rho < 1$ and $0 < \gamma < \phi < 1$.

3. Model for response using RR technique

In this section, we introduce a linear model for the response in the application using RR technique. Non-response is present in almost every non-mandatory survey, however, its effects are varying in different surveys depending on the nature of the surveys and the type of the questionnaire. If the survey questions are non-sensitive then almost everyone in the survey is willing to respond. However, if the question of the surveys is sensitive then a huge part of the data is missing due to non-response. To reduce the effect of non-response in cross-sectional data, Warner

(1965) proposed RR technique which improves survey response rate through randomization device. To reduce the effect of initial non-response, Alho and Rendtel (2020) proposed a time series model. Alho, don't use the randomization device to control the effect of non-response. To cover this gap, we propose a linear model for the response in the context of panel survey when the outcome variable is subjected to RR. The proposed model has the advantage that it exploits both the randomizing mechanism used in Warner (1965) randomized response model for qualitative data and the coding scheme used in Bar-Lev et al. (2004) for quantitative data. By letting know to individuals that their privacy will be fully protected through the RR technique, suppose those who agree to participate in the survey by answering sensitive survey question(s) in the initial and following subsequent waves of the panel.

Let $Y_{i,t}$ be the quantitative response variable of individual i at time t and let S_i be the corresponding random variable representing the random number that is used for the coding mechanism. It is also assumed that the response variable $Y_{i,t}$ is uncorrelated with the random variable S_i . Also, let $0 \leq P \leq 1$ be a DP of the RR model during the data collection which is controlled by the experimenter and is used for randomizing the respondent response. Under this aspect, if the respondent thinks that the underlying survey question is non-sensitive he/she responds with the true value of the response variable $Y_{i,t}$ with probability P . On the other hand, if the underlying survey question is sensitive he/she responds with the coded value of the response variable say the response $Y_{i,t}S_i$ with probability $1 - P$. Small value of P indicate that the underlying survey question is sensitive which means that most of the survey respondents will respond with the coded variable i.e., $Y_{i,t}S_i$. Let us suppose that $R_{i,t}$ be the binary response indicator of a person i which is equal to one $R_{i,t} = 1$ if the person agrees to respond at the start of the panel, and $R_{i,t} = 0$ otherwise. Then following the linear RR approach of Bar-Lev et al. (2004), the response of individual i at the initial and following panel waves can be represented by the following equation:

$$R_{i,t} = \begin{cases} Y_{i,t}, & \text{with probability } P, \\ Y_{i,t}S_i, & \text{with probability } (1 - P). \end{cases} \quad (14)$$

In this pattern, the probability of response of the individual i to a sensitive question can be approximately described by the following linear model:

$$P(R_{i,t} = 1 | Y_{i,t} \text{ or } Y_{i,t}S_i) = \begin{cases} 0, & \text{if } PY_{i,t} + (1-P)Y_{i,t}S_i < 0, \\ PY_{i,t} + (1-P)Y_{i,t}S_i, & \text{if } 0 \leq PY_{i,t} + (1-P)Y_{i,t}S_i \leq 1, \\ 1, & \text{if } PY_{i,t} + (1-P)Y_{i,t}S_i > 1, \end{cases} \quad (15)$$

In the context of panel surveys, Alho and Rendtel (2020) proposed a bias approximation formula for the cross-sectional OLS regression estimates in the presence of non-ignorable non-response at the start of the panel. Later on, Khan 2020 conducted a simulation study to verify the approximate results of Alho and Rendtel (2020) under different model stability of the covariate and error terms. Motivated by the work of Alho and Rendtel (2020), and Khan (2020) we extend the idea of fade-away of bias of the OLS regression estimates in the context of panel surveys when the outcome variable is subjected to RR.

To understand this idea let us take two finite samples selected from the same target population of sufficient size. Also, let us assume that the samples follow a Markov chain model and change their state according to the same transition probabilities law. Further, suppose that the first sample consists of all persons who participate in the surveys, we denote this sample as Full-Sample. This means that the Full-Sample consists of both respondents and non-respondents who are in the target population. As the Full-Sample includes both the survey respondents and non-respondents, there is virtually no bias in the estimated model parameter under this sample. The second sample then consists only of respondents of the Full-Sample or it contains only the respondents who participated in the survey initially. We denote this sample by Resp-Sample (respondent sample). At this point, we further assume that the distribution of the variable of interest in the Resp-Sample is highly selective or skewed for the response at the start of the panel. For the initial non-response of a person i , we suppose that it depends on the $Y_{i,1}$ or $Y_{i,1}S_i$. This assumption is automatically satisfied from the linear response model given in equation (11). Therefore, if the probability of response holds true then under this case the initial distribution under the Resp-

Sample is somewhat far away from the distribution of the Full-Sample. Under this respect, if there is no further drop out of observations in following panel waves, then according to the bias approximation formula of Alho and Rendtel (2020) in Section 2 and the simulation study of Khan (2020) the distorting effects of initial non-response on the distribution of $Y_{i,t}$ under the Resp-Sample has to faded-away over time $t = 2, 3, 4, \dots, T$ by reaching to the steady-state distribution of the Markov chain. Alternatively, if the model for response does not depend on either $Y_{i,1}$ and/or $Y_{i,t}S_i$ then in this case there will be no bias in the estimates of the Resp-Sample or non-response is missing at random (MAR). When there is no bias in the Resp-Sample then in this case the distribution of the Resp and the Full samples will be the same. This assumption holds true for small values of the DP (for details see figures).

Therefore, we assume that the initial non-response is not NMAR or highly selective for the estimation of population parameters. And therefore the initial distribution of the Resp-Sample at the start of the panel is somewhat away from the distribution of the Full-Sample. Under this respect, if there is no further drop out after wave 1, then according to the bias approximation formula proposed by Alho and Rendtel (2020) in Section 2 the distorting effects of initial non-response on the distribution of $Y_{i,t}$ in the Resp-Sample has to fade-away over time $t = 2, 3, 4, \dots, T$.

4. Monte Carlo simulation study

In the previous section, we theoretically discussed the fade-away of initial non-response of the OLS regression estimator in the context of panel survey using the RR technique. To verify the theoretical results of the fade-away effect we conducted a Monte Carlo simulation study. To generate data from the proposed model we choose different starting values of the parameters, where $a_t = 1$ for the intercept and $b_t = 1$ for the slope coefficient. We generated a synthetic population of size 1,000 observations from the model over 100 Monte Carlo replications. For the data generation, we assume that the underlying distributions of the covariate $X_t = M + Z_t$ and the error term $e_t = V + U_t$ are normally distributed. Under the assumption of uncorrelated variance components M and Z_t of the covariate X_t , having expectations $\mu_m = \mu_{z_t} = 0$, and variances $\sigma_m^2 = \kappa$

and $\sigma_{z_t}^2 = (1 - \kappa)$, respectively.

$$M \sim N(0, \sigma_m^2) \quad \text{and} \quad Z_t \sim N(0, \sigma_{z_t}^2),$$

Following a similar pattern, the residual term e_t is decomposed into two uncorrelated components known as: the permanent component V and the transient component U_t having their expectations $\mu_v = \mu_{u_t} = 0$, and variances $\sigma_v^2 = \gamma\sigma^2$ and $\sigma_{u_t}^2 = (1 - \gamma)\sigma^2$, respectively.

$$V \sim N(0, \sigma_v^2) \quad \text{and} \quad U_t \sim N(0, \sigma_{u_t}^2),$$

Further, we assume that the white noise error terms ε_t and ξ_t of the AR(1) models are independently and identically normally distributed random variables having expectations zero's, and variances $\sigma_\varepsilon^2 = (1 - \kappa)(1 - \rho^2)$ and $\sigma_\xi^2 = (1 - \gamma)(1 - \phi^2)\sigma^2$, respectively.

$$\varepsilon_t \sim N(0, \sigma_\varepsilon^2) \quad \text{and} \quad \xi_t \sim N(0, \sigma_\xi^2),$$

In addition, we assume that the random variable S which is used for the coding mechanism also, follow normal distribution having expectation $\mu_s = 0$ and variance $\sigma_s^2 = 0.20$, i.e., $S \sim N(0, \sigma_s^2 = 0.20)$.

As the fade-away hypothesis strongly depends on the size of the permanent and transient components of the covariate and error terms and the size of the DP (P). We, therefore, check its turnover for various stabilities of model covariates. Then two cases arise. In the first case, we assign equal weights to permanent and transient components of the covariate and the error term. This case is known as an equal stability scenario. While there are some intermixed cases where we assign different weights to permanent and transient components of the covariate and the error term. This case is known as an unequal stability scenario. All these cases are thoroughly discussed in Section 2. In this section we are only considering the equal case stability scenarios denoted by Scenario A – C and is given by:

- Scenario A: Low stability $\kappa = \gamma = \rho = \phi = 0.10$,
- Scenario B: Medium stability $\kappa = \gamma = \rho = \phi = 0.50$,

- Scenario C: High stability $\kappa = \gamma = \rho = \phi = 0.90$,

In each scenario, we allow the design parameter P to vary in the interval $0 \leq \sigma^2 \leq 1$, having values 0.1, 0.2, 0.3, 0.5, 0.8, and 1.0, respectively. The estimate of the model slope coefficient b_r is then obtained in each replication r , ($r = 1, 2, 3, \dots, 100$), of the Monte Carlo simulation study. where r stands for the number of Monte Carlo replications. Let $\hat{b}_{tr}$ be the OLS regression estimator of b_{tr} in r^{th} replication at time point t . Their average overall replications is denoted by $\hat{b}_t = \text{mean}(\hat{b}_{tr}) = \frac{1}{100} \sum_{r=1}^{100} \hat{b}_{tr}$. Where b_t is the true slope coefficient at time t . Finally, we compute the properties of the estimator of $\hat{b}_t$ at time t in terms of relative bias (RB) and the mean squared error (MSE). To make things clear we multiply the RB and MSE by 100, e.g., $\text{RB}(\hat{b}_t) = 100 \times (\text{mean}(\hat{b}_{tr}) - b_t)/b_t$ and $\text{MSE}(\hat{b}_t) = 100 \times E(\text{mean}(\hat{b}_{tr}) - b_t)^2$, respectively.

5. Results discussion

This paper presents the OLS estimation method to estimate the slope coefficients of the linear regression model where the response variable is sensitive or is subjected to RR. The bias of the OLS estimates in the application of the RR technique is then compared with the bias of estimates of the estimates using DQ technique. The general idea of the study is that when the outcome variable for the response then RR techniques provide more accurate results than the DQ survey techniques. Therefore, in this article, we use the RR survey technique for obtaining survey information on sensitive issues. This technique improves response through the randomization device. In general, the RR technique performs well as compared to any direct surveys see e.g., Lensvelt-Mulders et al. (2005). This is also verified from the simulation results of our simulation study which demonstrate that the proposed RR technique provides accurate estimates. Further, to check the size of the initial bias and its fade-away in follow-up panel waves we restrict our analysis up to 10 panel waves of the survey. Usually, ten waves of the panel are sufficient to check the behavior of the estimator and its fade-away effect.

Surveys, where a huge part of the data is missing due to non-response, may produce bias statistical analysis that not only reduces sample size but may produce bias in estimates. There are

many reasons for the high non-response rate (we discuss it earlier). However, some such reasons for non-response are questions on sensitive issues such as the use of drugs, abortion, watching porn, black money, divorce, etc. Such questions are considered sensitive questions. If such types of questions are included in the survey questionnaire then survey respondents are reluctant or afraid to answer. As a result survey response rate is low or in other words the non-response rate is high. High non-response may not only reduce case numbers of data but it has the power to make the analysis biased.

To improve response rate Warner (1965) proposed an alternative survey technique known as the randomized response (RR) technique which improves survey response rate by protecting respondent privacy through randomization devices. Therefore, to check the size of the estimated slope coefficients and their fade-away over time we run a simulation study for various values of the DP (denoted by P) of the Warner model. As the values of the DP varies in the interval $0 \leq P \leq 1$. For $P = 1$ the Warner RR technique is equivalent to the DQ technique. For $P = 1$, it also means that respondents are asked to respond to sensitive questions directly without using any randomization device which protects their privacy. So we expect a large non-response rate on a large value of P say $0.5 \leq P \leq 1$, while for small or low values of P say $0.1 \leq P \leq 0.5$ we expect a low non-response rate. Low values of P usually less than $P \leq 0.5$ indicate respondent privacy is well protected. In other words, it means that small values of $P \leq 0.5$ indicate that respondents respond to sensitive questions in coded values.

As the Full-Sample consists of both the respondents and the non-respondents, therefore there is no bias in estimates under this sample. This holds true for all simulation scenarios (Scenario A-C). We denote the bias of the estimates under the Full-Sample by “Bias Full” highlighted by the color “green”. The results of our simulation study are summarized in Figure 1 to Figure 3. On Y -axis we have the bias (relative bias in percent) of the estimate, while on the X -axis we have the wave of the panel. The colored’s circle points in the graph show the relative bias of the estimates obtained from our regression model in a specific wave of the panel. The graphical results demonstrate that for large values of P the bias of the estimates is substantially larger than the bias of the estimates for small values of P . Regarding the fade-away effect, it strongly depends

on the stability of the model parameters $(\kappa, \gamma, \rho, \phi)$. For example in Figure 1 (Scenario-A) the bias of the estimate in wave 1 for $P = 1$ is 37% (color: blue), while for $P = 0.8, 0.5, 0.3, 0.2, 0.1$ the bias of the estimate is 37% (color: orange), 35% (color: red), 25% (color: brown), 14% (color: pink), and 2% (color: black), respectively. Here it can be seen that for all values of P the bias of the estimate is very small at $P = 0.1$ this is since the respondents have the higher chance to scrambled their responses through the RR technique.

Regarding the fade-away effect, we see that small/large initial non-response bias of the OLS estimates for various values of P vanishes in follow-up panel waves. Moreover, the bias of the estimated slope coefficients in wave 2 and later panel waves converge to zero. This happens due to low stability (Stability-A: $\kappa = \gamma = \rho = \phi = 0.10$) of the model covariate and the error term. Thus the speed of the fade-away effect is high in the case where the size of the variance components of the covariate and/or the error term is less correlated or independent over different cross-sections or over time. We then check the size of the fade-away effect in Scenario-B (Medium stability: $\kappa = \gamma = \rho = \phi = 0.50$). We demonstrate the results under this scenario in Figure 2. Here the bias of the estimates in the initial wave is the same as the bias of the estimates in Scenario-A. However, the size of the biases after the initial wave is larger than the bias of estimates in Scenario-A. Therefore, large subsequent biases of the estimates dampen the speed of the fade-away effect. This is clearly shown in Figure 2. Nevertheless, the initial bias of the estimates and its selection effects on the estimates reduces geometrically in follow-up panel waves. However, after some waves, the effect of non-response on the estimated slope coefficients remains permanent. This indicates the fixed permanent component of the error term which is related to non-response.

Finally, Scenario C (High stability: $\kappa = \gamma = \rho = \phi = 0.90$) is the more extreme scenario where the AR covariate and AR error term of the model is highly correlated with its previous time values of the model. Also as have been discussed in the previous discussion that the fade-away of bias depends on the size of the fixed permanent components of the covariate and the error term. If the size of permanent components is large then their distribution remains permanent and hence the effect of initial non-response doesn't reduce in subsequent panel waves. Thus no fade-away is present for Scenario-C. This can also be checked from the bias approximation formula

of Alho and Rendtel (2020), and the simulation study of Khan (2020). Now, the situation is completely different for the same Scenario-C under the RR technique. E.g., by plotting the results under this scenario which are illustrated in Figure 3. For the said scenario the estimated slope coefficients of the simulation study swing into the steady-state distribution of the Markov chain. The reason for such a change is clear with the RR procedure respondent are more confident about their privacy to a certain sensitive question. Thus more respondents participate in the survey with more true answers, as a result, non-response bias fades-away over in later panel waves. Overall, our results are completely in line with the theoretical results of Alho and Rendtel (2020), the simulation results of Khan (2020), the theoretical results of the Warner (1965) model, and the coding mechanism in Bar-Lev et al. (2004).

Concerning the privacy protection of the respondent, we see that the design parameter plays a central role in the privacy protection of the respondent to a certain sensitive survey question. And as can be seen from the probability of response given in equation 1, a small value of P means that the privacy of respondents to a certain sensitive question is well protected. Thus for $P = 0.1$ almost 90% of the survey respondents protect their privacy through RR technique. Thus if survey contains sensitive question then the use of the RR technique provide consistent parameter estimates but it does not guarantee for the unbiasedness as a bias of 2% is always present in the “Resp-Samples” estimates.

Figure 1: The fade-away effect of the initial bias of the OLS slope coefficient of $b_t = 1$ in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for various stability of the design parameter P .

Data generated from the model through Monte Carlo simulation study. Artificial population of size $N = 1,000$. Number of Monte Carlo replications $R = 100$. Solid lines with colored's circle's points: Bias of the OLS estimate in wave t . Source: Own computations.

Figure 2: The fade-away effect of the initial bias of the OLS slope coefficient of $b_t = 1$ in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for various stabilities of the design parameter P .

Data generated from the model through Monte Carlo simulation study. Artificial population of size $N = 1,000$. Number of Monte Carlo replications $R = 100$. Solid lines with colored's circle's points: Bias of the OLS estimate in wave t . Source: Own computations.

Figure 3: The fade-away effect of the initial bias of the OLS slope coefficient of $b_t = 1$ in Scenario C ($\kappa = \gamma = \rho = \phi = 0.90$) for various stabilities of the design parameter P .

Data generated from the model through Monte Carlo simulation study. Artificial population of size $N = 1,000$. Number of Monte Carlo replications $R = 100$. Solid lines with colored's circle's points: Bias of the OLS estimate in wave t . Source: Own computations.

We also compared the mean squared error (MSE) of the OLS estimate of b_t over 100 Monte Carlo replications for various values of the DP in Scenario A-C. The visualization of the MSE's results exhibits different patterns for various values of DP in these stability scenarios which are visualized in Figure 4 to Figure 6. In the figures, the MSE of the estimator is displayed on the Y-axis which is plotted against various panel waves on the X-axis. For some coefficients, the MSE's estimates are very small so we multiply all the MSE estimates by 100. To distinguish between Full-Sample and Resp-Sample estimates we assign them different names and different colors. First, consider the MSE's of the estimator under the "Full-Sample". We denote the MSE estimates under this sample by the letter "MSE Full" highlighted with the color "green". Regarding the MSE estimates of the Full-Sample, previously we have shown that the bias of the estimates under this sample is very small nearly close to zero. This is true for all Scenario A-C. Having a negligible bias, plus a variance component which, therefore, leads to the smallest MSE estimates. The MSE property of the estimator under the "Full-Sample" holds true for Scenario A-C. Concerning the fade-away, effect we see that there is no fade-away present in the estimates. The reason for this constant behavior is that the estimated slope coefficients have very small MSE values, close to zero.

However, for the "Resp-Sample" the analysis is affected by non-response up to some extent depending on the size of the DP and the size of the stability parameters ($\kappa, \gamma, \rho, \phi$). In the figure we see that in all scenarios the estimates of the "Resp-Sample" is less affected by bias for low values of P say for $P = 0.10$ (10%), while is it greatly affected for all other values of P . Therefore, the MSE of the estimates are very small for $P = 0.10$, and it is substantial for all other values $P > 0.10$. Again this is true for all Scenarios A-C. Regarding the fade-away effect, it is present in all the figures, i.e., over time it has the tendency of downward MSE's. However, the speed of the fade-away is fast in Figure 4 (Scenario-A: $\kappa = \gamma = \rho = \phi = 0.10$). This is because the regression models are less correlated (10%) over different cross-sections and over different time periods. For Scenario-B (Figure 5) the speed of the fade-away is strong due to medium correlation over different cross-sections and over different time periods. While the speed factor of the fade-away effect in Scenario-C (Figure 6) is quite lower than the speed factor in Scenario A-B (Figure 4 and

Figure 5). The reason for this change is that the regressions in this scenario are not varying so much over time and thus it is the main reason for the slow fade-away of MSE's. As expected from the theoretical results of Alho, with low stability of covariates and error term the OLS estimates have the smallest bias which in turn lead to the smallest MSE value. Therefore, the efficiency of the OLS estimator is highest in Scenario-A. While the efficiency of the estimator is reduced considerably in Scenario B-C, when the size of permanent and time-varying parts of the covariate and error terms are medium or high (see Figure 5 and Figure 6).

Overall we concluded that the speed of the fade-away effect of the bias/MSE strongly depends on: i) the size of the permanent/temporal components of the covariates and the model noise error term, ii) It also strongly depends on the DP if the value of DP is small then we expect substantial reduction not only for initial waves estimates but also for the follow-up panel waves, iii) It also depends on the fast turn-over covariates (measured and unmeasured) of the model if the covariates of the model are time-invariant or change slowly over time then we would not expect any fade-away for the bias or for the MSE estimates.

Figure 4: The fade-away effect of the initial MSE of the OLS slope coefficient of $b_t = 1$ in Scenario A ($\kappa = \gamma = \rho = \phi = 0.10$) for various stabilities of the design parameter P .

Data generated from the model through Monte Carlo simulation study. Artificial population of size $N = 1,000$. Number of Monte Carlo replications $R = 100$. Solid lines with colored's circle's points: MSE of the OLS estimate in wave t . Source: Own computations.

Figure 5: The fade-away effect of the initial MSE of the OLS slope coefficient of $b_t = 1$ in Scenario B ($\kappa = \gamma = \rho = \phi = 0.50$) for various stabilities of the design parameter P .

Data generated from the model through Monte Carlo simulation study. Artificial population of size $N = 1,000$. Number of Monte Carlo replications $R = 100$. Solid lines with colored's circle's points: MSE of the OLS estimate in wave t . Source: Own computations.

Figure 6: The fade-away effect of the initial MSE of the OLS slope coefficient of $b_t = 1$ in Scenario C ($\kappa = \gamma = \rho = \phi = 0.90$) for various stabilities of the design parameter P .

Data generated from the model through Monte Carlo simulation study. Artificial population of size $N = 1,000$. Number of Monte Carlo replications $R = 100$. Solid lines with colored's circle's points: MSE of the OLS estimate in wave t . Source: Own computations.

Compliance with ethical standards

Conflict of interest I declare that I has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals

performed by any of the authors.

Bibliography

- Alho, J. and Rendtel, U. (2020). Demographics of selection bias in follow-up studies: When does initial bias fade-away? Unpublished manuscript.
- Alho, U., Müller, G., Pflieger, V., and Rendtel, U. (2017). The fade-away of an initial bias in longitudinal surveys, Discussion paper Economics FB Wirtschaftswissenschaft FUB 2017/25 . Online available at <http://hdl.handle.net/10419/168347>.
- Arnab, R. (1996). Randomized response trials: a unified approach for qualitative data. *Communications in Statistics - Theory and Methods*, 25(6):1173–1183.
- Bar-Lev, S. K., Bobovitch, E., and Boukai, B. (2004). A note on randomized response models for quantitative data. *Metrika*, 60:255–260.
- Boruch, R. F. (1972). Relations among statistical methods for assuring confidentiality of social research data. *Social Science Research*, 1(4):403–414.
- Chaudhuri, A. (2001). Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. *Journal of Statistical Planning and Inference*, 94:37–42.
- Eichhorn, B. H. and Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, 7:307–316.
- Fenton, K. A., Johnson, A. M., McManus, S., and Erens, B. (2001). Measuring sexual behaviour: methodological challenges in survey research. *Sexually Transmitted Infections*, 77(2):84–92.
- Fitzgerald, J., Gottschalk, P., and Moffitt, R. (1998). An analysis of sample attrition in panel data, The Michigan Panel Study of Income Dynamics. *The Journal of Human Resources*, 33(2):251–299.

- Franklin, L. A. (1989). Randomized response sampling from dichotomous population with continuous randomization. *Survey Methodology*, 15(2):225–235.
- Gjestvang, C. R. and Singh, S. (2006). A new randomized response model. *Journal of the Royal Statistical Society: Series B*, 68(3):523–530.
- Greenberg, B. G., Abul-El, A. A., Simmons, W. R., and Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64:520–539.
- Khan, M. (2020). *The fade-away phenomenon of initial non-response bias in panel surveys*. PhD thesis, Institute of Statistics Freie Universität Berlin.
- Kuk, A. Y. C. (1990). Asking sensitive questions directly. *Biometrika*, 77(2):436–438.
- Lensvelt-Mulders, G. J. L. M., Hox, J. J., Van der Heijden, P. G. M., and Maas, C. J. M. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods and Research*, 33:319–348.
- Moors, J. J. A. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association*, 66:627–629.
- Rendtel, U. (2013). The fade-away effect of initial non-response in panel surveys, Empirical results for EU-SILC. Euro-stat methodologies and working papers (ISSN 1977-0375) edition 2013, doi:10.2785/21863.
- Sisto, J. (2003). Attrition effects on the design-based estimates of disposable household income. CHINTEX working paper No. 9 03/2003, URL: www.destatis.de/chintex.
- Van den Hout, A., van der Heijden, P. G. M., and Gilchrist, R. (2007). The logistic regression model with response variables subject to randomized response. *Computational Statistics and Data Analysis*, 51(12):6060–6069.
- Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69.