mathematics

MDPI

Article

Estimation of Non-Linear Parameters with Data Collected Using Respondent-Driven Sampling

Ismael Sánchez-Borrego [1], María del Mar Rueda [1,*] and Héctor Mullo [2]

[1] Department of Statistics and Operations Research, University of Granada, 18071 Granada, Spain; ismasb@ugr.es

[2] Facultad de Ciencias, Escuela Superior Politécnica de Chimborazo (ESPOCH), 060155 Riobamba, Ecuador; hmullo@espoch.edu.ec

* Correspondence: mrueda@ugr.es

Received: 23 July 2020; Accepted: 5 August 2020; Published: 7 August 2020

Abstract: Respondent-driven sampling (RDS) is a snowball-type sampling method used to survey hidden populations, that is, those that lack a sampling frame. In this work, we consider the problem of regression modeling and association for continuous RDS data. We propose a new sample weight method for estimating non-linear parameters such as the covariance and the correlation coefficient. We also estimate the variances of the proposed estimators. As an illustration, we performed a simulation study and an application to an ethnic example. The proposed estimators are consistent and asymptotically unbiased. We discuss the applicability of the method as well as future research.

Keywords: Respondent-driven sampling; regression; network dependence

1. Introduction

Respondent-driven sampling (RDS) is a refined form of snowball sampling for collecting data from hidden populations that lack a sampling frame. Therefore, these populations are difficult to reach and cannot be dealt with using traditional sampling techniques. RDS was first introduced by Heckathorn [1] and was developed afterwards by Salganik and Heckathorn [2] and Volz and Heckathorn [3]. Some recent papers in parameter and variance estimation are given by [4–6]. Some popular examples of RDS are HIV at-risk people, the LGBTI community and injection drug users [7–9]. Other examples are given in [10,11].

A long-standing problem in non-probabilistic sampling is regression modeling for RDS data. Several authors have considered this issue from different perspectives. For instance, there have been a surge of studies in machine learning and statistical framework to solve similar problems. Wong et al. [12] studies the problem of biased standard errors of non-linear transport models. Imani et al. [13,14] use an approximation of a distribution using a Markov chain Monte carlo (MCMC) algorithm and an approximate MCMC implementation, respectively.

Model-fitting should incorporate sample weights as well as information about correlation between sample units [3]. Avery et al. [15] review some available methods in an RDS framework dealing with this problem in a simulation study with binary response. One popular method for addressing the problem of the correlation structure between recruits and recruiter in a network is clustering, that is, transforming RDS data into clustered data [16–18]. Clustering is connected with the homophily in the population, that is the tendency to associate with those with similar characteristics. Becket et al. [16] use clusters to consider correlation for estimating diabetes prevalence and discrete covariates in an empirical study. Nevertheless, using clustering can be problematic as sometimes clusters, if they exist, may be difficult to identify. Hubbart et al. [19] point some of the problems involved with using

generalized estimating equations (GEE) when the number of clusters is small and Rao et al. [20] stress some of the problems involved with not adjusting properly to clusters.

Regression methods for analyzing data have not been fully validated [16]. Therefore, as no clear approach for regression modeling in RDS is available, some authors use standard statistical methods without adjusting for RDS data [21,22]. On the other hand, some authors use weighted regression for estimating prevalence of characteristics of interest in real-life examples [23–26]. Most of them use individual weights calculated (typically with the respondent driven sampling analytical tool (RDSAT)) and export them to standard statistical software to apply the weighted method. Methods incorporating sample weights will tend to improve their performance when homophily is small in the population, as they typically do not account for the potential dependence of the units. While this might be an issue in populations with high homophily, there is no clear reliable regression method in RDS accounting for clustering, that can be extensively used in applications. Adjusting for clusters requires knowledge of the population and if it is not well performed in practice, or even if clusters do not actually exist in the population, might result in biased estimates [20]. Our method addresses the problem of regression modeling and association between continuous variables by proposing a new sample weight estimation method for continuous data. The focus of our work was to propose a method for estimating non-linear parameters such as the covariance and the correlation coefficient. We derived expressions for the estimators that make use of the RDS estimators admitting continuous data and showed that they share properties with them, such as being consistent and asymptotically unbiased. A diagram is given in Figure 1. We also estimated the variances of the proposed methods. Our method may fill a gap as no such an approach has achieved in an RDS framework: most studies incorporate the weights using standard statistical software and unlike our proposal, they are focused on prevalence estimation.

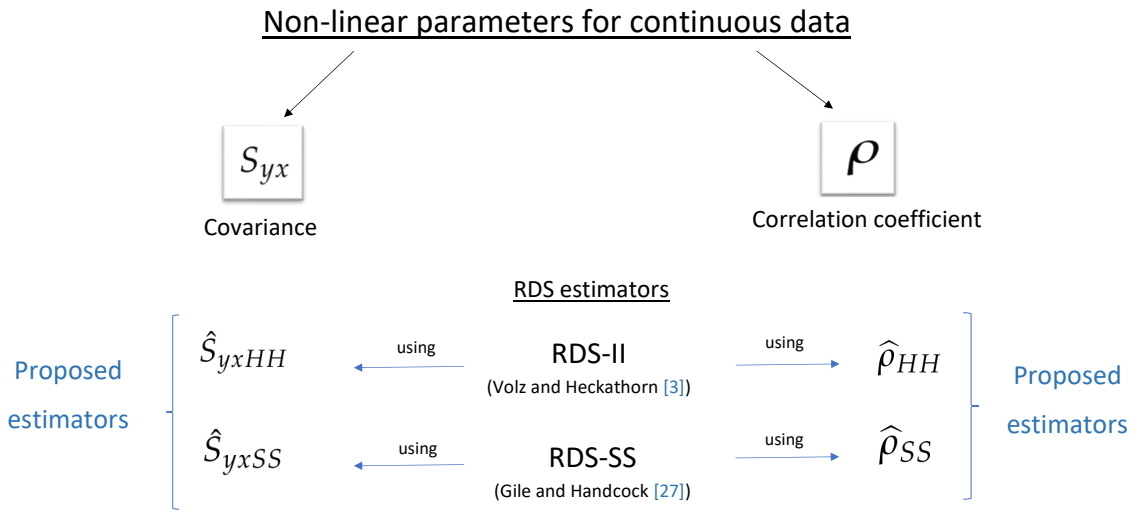

Figure 1. Schematic representation of the proposed method.

We begin in the next section by introducing respondent-driven sampling. In Section 3, we propose estimators for the population covariance and for the correlation coefficient. Estimators of the variances of the proposed estimators are considered in Section 4. A simulation study was performed to illustrate their performance, described in Section 5. An application to study the living conditions of Indigenous, Montubios and Afro-Ecuadorian young people are presented in Section 6. Finally, Section 7 presents concluding remarks.

2. Background

The main idea behind the estimation in RDS [3] is to treat this sampling as a random walk on an undirected network. It is well known from Markov chain theory that the stationary (equilibrium) probability of a node is then proportional to its degree.

We assume the target population consists of N people (nodes) with labels $1, ..., N$. We assume the target population is connected by a network of mutual relations with $N \times N$ adjacency matrix $\mathbf{Z}$. That is $z_{ij} = z_{ji} = 1$ if i and j are connected and 0 otherwise. We define the nodal degree of a the person i, $\delta_i = \sum_j z_{ij}$, as the number of network ties or alters of node i.

An small initial sample s is selected from the population members accessible to researchers that are called the seeds and comprise wave 0 of the sample. Each member of wave v is given a number of uniquely identified coupons to distribute among their alters. Coupon recipients returning their coupons to the study center are subsequently enrolled in the study. The wave number of a respondent is one more than that of their recruiter. This procedure is repeated until the desired sample size, n, is attained.

Let the N-vector y represent a variable of interest. If y have binary response and groups A and B, the more usual estimators in RDS are the RDS-I ratio estimator, the RDS-II estimator [3] and the Gile and Hanckock [27] version for sampling with replacement.

The RDS-I estimator for estimating proportions with y binary response and groups A and B is defined as

$$\hat{p}_A = \hat{C}_{BA}\hat{D}_B / (\hat{C}_{BA}\hat{D}_B + \hat{C}_{AB}\hat{D}_A), \tag{1}$$

with $\hat{C}_{AB} = \frac{r_{AB}}{r_{AB}+r_{AA}}$, r_{AB} is the number of people of $A's$ recruiting $B's$ in the sample, r_{AA} the number of people of $A's$ recruiting $A's$ in the sample, $\hat{C}_{BA} = \frac{r_{BA}}{r_{BA}+r_{BB}}$, r_{BA} is the number of people of $B's$ recruiting $A's$ in the sample, n_A and n_B the number of sample units belong to groups A and B respectively, and $\hat{D}_A$ and $\hat{D}_B$ are the average degree of people in groups A and B, respectively.

The RDS-SS [27] estimator for estimating proportions:

$$\hat{p}_A = \sum_{k \in s_A} (\hat{\pi}(\delta_k)^{-1}) / \sum_{k \in s} \hat{\pi}(\delta_k)^{-1}, \tag{2}$$

with $\hat{\pi}(\delta_k)$ the estimated population distribution of degrees through successive sampling.

The RDS-II estimator of the mean $\overline{Y}$ allows continuous variables and takes the form of the Hajek estimator as follows:

$$\hat{\overline{Y}} = \sum_{k \in s} (\delta_k^{-1} y_k) / \sum_{k \in s} \delta_k^{-1}, \tag{3}$$

with δ_k the degree reported by respondent k.

3. Estimation of Some Non-Linear Parameters

The widespread use of regression based on sample survey data requires a careful assessment of the use of standard techniques. It is clear that usual estimators of parameters involved in regression are not valid in the case of RDS scheme. In this section, we develop some estimators for population variances, covariances and the correlation coefficient.

3.1. Estimation of the Variance and the Covariance

We define the population covariance as:

$$S_{yx} = \frac{1}{N-1} \sum_U (y_k - \overline{Y})(x_k - \overline{X}).$$

We can write this parameter as:

$$S_{yx} = \frac{1}{N-1} T_{yx} - \frac{1}{N(N-1)} T_y T_x = \theta = f(\theta_1, \theta_2, \theta_3),$$

being $T_{yx} = \theta_1 = \sum_U y_k x_k$, $T_y = \theta_2 = \sum_U y_k$ and $T_x = \theta_3 = \sum_U x_k$.

Similarly, the finite population variances are defined as

$$S_y^2 = \frac{1}{N-1} T_{yy} - \frac{1}{N(N-1)} T_y^2,$$

and

$$S_x^2 = \frac{1}{N-1} T_{xx} - \frac{1}{N(N-1)} T_x^2.$$

Let us construct estimators for these parameters assuming that y_k and x_k are observed for the units of the RDS sample s.

If there exists $\hat{\theta}_1$, $\hat{\theta}_2$ and $\hat{\theta}_3$ consistent estimators of θ_1, θ_2 and θ_3, a consistent estimator of S_{yx} will be

$$\hat{S}_{yx} = \frac{1}{N-1} \hat{\theta}_1 - \frac{1}{N(N-1)} \hat{\theta}_2 \hat{\theta}_3 = \hat{\theta}. \tag{4}$$

We can estimate these totals with the RDS-II estimator:
$\hat{T}_{yHH} = \frac{N\hat{\delta}_v}{n} \sum_s y_k \delta_k^{-1}$, $\hat{T}_{yyHH} = \frac{N\hat{\delta}_v}{n} \sum_s y_k^2 \delta_k^{-1}$, and $\hat{T}_{yxHH} = \frac{N\hat{\delta}_v}{n} \sum_s y_k x_k \delta_k^{-1}$, being $\hat{\delta}_v = \frac{n}{\sum_U \delta_k^{-1}}$ the average degree.

Then, the estimator of the covariance is

$$\hat{S}_{yxHH} = \frac{1}{N-1} \hat{T}_{yxHH} - \frac{1}{N(N-1)} \hat{T}_{yHH} \hat{T}_{xHH}. \tag{5}$$

If N is large, $\hat{T}_{yHH}$ can be written in a more straightforward way that does not depends on N:

$$\hat{S}_{yxHH} = \frac{\hat{\delta}_v}{n} \sum_s y_k x_k \delta_k^{-1} - \frac{\hat{\delta}_v^2}{n^2} \sum_s y_k \delta_k^{-1} \sum_s x_k \delta_k^{-1}. \tag{6}$$

Using the idea of the RDS-SS estimator, we propose to estimate the totals as:
$\hat{T}_{ySS} = \sum_s y_k \hat{\pi}(\delta_k)^{-1}$, $\hat{T}_{yxSS} = \sum_s y_k x_k \hat{\pi}(\delta_k)^{-1}$ and $\hat{T}_{yySS} = \sum_s y_k^2 \hat{\pi}(\delta_k)^{-1}$, being $\hat{\pi}(\delta_k)$ the estimated population distribution of degrees through successive sampling.

Then, the estimator of the covariance is

$$\hat{S}_{yxSS} = \frac{1}{N-1} \hat{T}_{yxSS} - \frac{1}{N(N-1)} \hat{T}_{ySS} \hat{T}_{xSS}.$$

If N is unknown, a consistent estimator for S_{yx} is:

$$\hat{S}_{yxSS} = \frac{1}{\hat{N}-1} \hat{T}_{yxSS} - \frac{1}{\hat{N}(\hat{N}-1)} \hat{T}_{ySS} \hat{T}_{xSS}, \tag{7}$$

with $\hat{N} = \sum_s \hat{\pi}(\delta_j)^{-1}$.

RDS-SS and RDS-II estimators of a total are asymptotically unbiased, thus the proposed estimators will be asymptotically unbiased.

3.2. Estimation of the Correlation Coefficient

In this section, we consider the estimation of the correlation coefficient between two variables, say y and x, defined by

$$\rho = S_{yx} / S_y S_x.$$

Two estimators for this parameter can be obtained by using RDS-II and RDS-SS estimators which are previously defined:

$$\widehat{\rho}_{HH} = \frac{\hat{S}_{yxHH}}{\hat{S}_{yHH} \hat{S}_{xHH}}, \tag{8}$$

and

$$\widehat{\rho}_{SS} = \frac{\hat{S}_{yxSS}}{\hat{S}_{ySS}\hat{S}_{xSS}}, \tag{9}$$

being $\hat{S}_{yHH} = \frac{1}{N-1}\hat{T}_{yyHH} - \frac{1}{N(N-1)}\hat{T}^2_{yHH}$, $\hat{S}_{xHH} = \frac{1}{N-1}\hat{T}_{xxHH} - \frac{1}{N(N-1)}\hat{T}^2_{xHH}$, $\hat{S}_{ySS} = \frac{1}{N-1}\hat{T}_{yySS} - \frac{1}{N(N-1)}\hat{T}^2_{ySS}$ and $\hat{S}_{xSS} = \frac{1}{N-1}\hat{T}_{xxSS} - \frac{1}{N(N-1)}\hat{T}^2_{xSS}$.

4. Estimation of the Variances

We consider the variance estimation of the covariance of $\widehat{S}_{yx}$

Using a Taylor linearization, we write

$$\widehat{\theta} \simeq \widehat{\theta_o} = \theta + \sum_1^3 w_j(\widehat{\theta}_j - \theta_j),$$

with

$$w_j = \frac{\partial f(\widehat{\theta_1}(s), ..., \widehat{\theta_3}(s))}{\partial \widehat{\theta}_j}|_{\theta_1,...,\theta_3}.$$

$$V(\widehat{\theta_o}) = V(\sum w_j\widehat{\theta}_j) = \sum w_j^2 V(\widehat{\theta}_j) + \sum w_i w_j cov(\widehat{\theta}_i, \widehat{\theta}_j),$$

and

$$\widehat{V}(\widehat{\theta}) \simeq \widehat{V}(\widehat{\theta_o}) = \sum \hat{w}_j^2 \widehat{V}(\widehat{\theta}_j) + \sum \hat{w}_i \hat{w}_j \widehat{cov}(\widehat{\theta}_i, \widehat{\theta}_j),$$

being $w_1 = \frac{1}{N-1}$, $w_2 = -\frac{\theta_3}{N(N-1)}$, $w_3 = -\frac{\theta_2}{N(N-1)}$.

They are estimated by

$$\hat{w}_1 = w_1, \hat{w}_2 = -\frac{\hat{T}_x}{N(N-1)}, \hat{w}_3 = -\frac{\hat{T}_y}{N(N-1)}.$$

Note: A more straightforward computational expression can be derived from formulae 5.5.10 in Särndal et al. [28]

We estimate the variances and covariances of the above-mentioned totals for the RDS-II estimator as

$$\hat{V}(\hat{T}_{yxHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k x_k \delta_k^{-1} - \hat{T}_{yxHH})^2,$$

$$\hat{V}(\hat{T}_{yHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k \delta_k^{-1} - \hat{T}_{yHH})^2,$$

$$\hat{V}(\hat{T}_{xHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v x_k \delta_k^{-1} - \hat{T}_{xHH})^2,$$

$$\widehat{cov}(\hat{T}_{yxHH}, \hat{T}_{xHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k x_k \delta_k^{-1} - \hat{T}_{yxHH})(N\hat{\delta}_v x_k \delta_k^{-1} - \hat{T}_{xHH}),$$

$$\widehat{cov}(\hat{T}_{yxHH}, \hat{T}_{yHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k x_k \delta_k^{-1} - \hat{T}_{yxHH})(N\hat{\delta}_v y_k \delta_k^{-1} - \hat{T}_{yHH}),$$

and

$$\widehat{cov}(\hat{T}_{yHH}, \hat{T}_{xHH}) = \frac{1}{(n-1)n}\sum_s (N\hat{\delta}_v y_k \delta_k^{-1} - \hat{T}_{yHH})(N\hat{\delta}_v x_k \delta_k^{-1} - \hat{T}_{xHH}).$$

The proposed RDS-II estimator is only analogous to the Hansen and Hurvitz estimator [29], but as data are correlated in an RDS framework, the above-mentioned estimators can perform poorly. Even though Volz and Heckathorn [3] derived a variance estimator that accounts the MCMC structure of the sample for categorical variables, we can not use this variance estimator in this context.

We estimate now the variances and covariances of the totals for the RDS-SS estimator by using the Deville and Särndal [30] method for estimating the variance of the Horvitz–Thompson estimator. The variances are estimated as

$$\widehat{V}(\widehat{T}_{yxSS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l x_l/\hat{\pi}(\delta_l))^2,$$

$$\widehat{V}(\widehat{T}_{ySS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l/\hat{\pi}(\delta_l))^2,$$

and

$$\widehat{V}(\widehat{T}_{xSS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l x_l/\hat{\pi}(\delta_l))^2.$$

The covariances are estimated as

$$\widehat{cov}(\widehat{T}_{yxSS}, \widehat{T}_{ySS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l x_l/\hat{\pi}(\delta_l))(\frac{y_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l/\hat{\pi}(\delta_l)),$$

$$\widehat{cov}(\widehat{T}_{yxSS}, \widehat{T}_{xSS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l x_l/\hat{\pi}(\delta_l))(\frac{x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l x_l/\hat{\pi}(\delta_l)),$$

and

$$\widehat{cov}(\widehat{T}_{ySS}, \widehat{T}_{xSS}) = \frac{1}{1-\sum_{k\in s} a_k^2} \sum_{k\in s}(1-\hat{\pi}(\delta_k))(\frac{y_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l y_l/\hat{\pi}(\delta_l))(\frac{x_k}{\hat{\pi}(\delta_k)} - \sum_{l\in s} a_l x_l/\hat{\pi}(\delta_l)),$$

where $a_k = (1-\hat{\pi}(\delta_k))/\sum_{l\in s}(1-\hat{\pi}(\delta_l))$.

As the correlation coefficient estimators are ratio estimators, the estimators of their variances can be easily obtain by using Taylor linearization (see e.g., [31]).

5. Simulation Experiments

In this section, a limited simulation study was carried out to illustrate the performance of the proposed estimators under different scenarios. The main factor of interest was the estimation of the population covariance and the correlation between continuous covariates. We used our own code written in `R` to compute the proposed estimators. Programming details and code are available from the authors.

The simulated population size was $N = 10000$. A NxN network connection indicator matrix C was randomly generated, with c_{ij} either 0 or 1, a connection indicator between node i and j, for $i, j = 1, \ldots, N$. Resulting c_{ij} will determine degree, as $\sum_{i\in U, i\neq j} c_{ij} = \delta_j$. Ten seeds were selected at random from the network with probability proportional to their degree, with three maximal coupons issued for each participant.

The values of the variable of interest y were generated from a normal distribution $y_j \sim N(5000, 500)$, for $j = 1, \ldots, 5000$. Three auxiliary variables were then generated from the values of y, which were: $x_1 = (y - e_1)/0.5$ with $e_1 \sim N(500, 500)$, $x_2 = (y - e_2)/0.5$ with $e_2 \sim N(500, 700)$ and $x_3 = (y - e_3)/0.5$, where $e_3 \sim N(500, 300)$. The resulting correlation coefficients were $\rho = 0.7007$ for x_1, $\rho = 0.571$ for x_2 and $\rho = 0.8579$ for x_3, respectively. The simulations were also performed for other different covariates and therefore different values of ρ, but the results were qualitatively similar and hence are not reported here. Sample size was $n = 500$ and samples were selected using simple random sampling without replacement, just like RDS is usually conducted in practice.

For each regression model, we computed the two proposed estimators of the population covariance S_{yx} and the correlation coefficient ρ. We investigated the percent relative bias

$$rb\% = E_{MC}(\widehat{\theta} - \theta)/\theta * 100,$$

and the percent relative mean squared error

$$rmse\% = E_{MC}[(\widehat{\theta} - \theta)^2]/\theta^2 * 100,$$

for each estimator $\widehat{S}_{yx}$ and $\widehat{\rho}$. Simulation results were based on $B = 1000$ samples and E_{MC} denotes the average of the Monte Carlo replications.

The estimators of the covariance are approximately unbiased, as relative biases are around 1% for all scenarios considered, with even lower biases for the correlation coefficient estimates, with all of them less than 1%, as shown in Tables 1 and 2. Small relative efficiency values for estimating the parameters with quite similar results obtained with both estimators, indicating that they are effective in estimating these non-linear parameters.

Table 1. Percent relative bias ($rb\%$) and the relative mean squared error ($rmse\%$) for estimating S_{yx} with estimators $\widehat{S}_{yx,RDS-II}$ and $\widehat{S}_{yx,RDS-SS}$ in the three scenarios. RDS—respondent-driven sampling.

Estimators	$\widehat{S}_{yx,RDS-II}$		$\widehat{S}_{yx,RDS-SS}$	
	rb%	***rmse%***	***rb%***	***rmse%***
Scenario 1	1.4953	0.8158	1.5055	0.8065
Scenario 2	1.7857	1.0906	1.7897	1.0782
Scenario 3	1.2745	0.6516	1.2924	0.6447

Table 2. Percent relative bias ($rb\%$) and the relative mean squared error ($rmse\%$) for estimating the correlation coefficient ρ with estimators $\widehat{\rho}_{RDS-II}$ and $\widehat{\rho}_{RDS-SS}$ in the three scenarios.

Estimators	$\widehat{\rho}_{RDS-II}$		$\widehat{\rho}_{RDS-SS}$	
	rb%	***rmse%***	***rb%***	***rmse%***
Scenario 1	0.4738	0.1262	0.4786	0.1245
Scenario 2	0.8656	0.3347	0.8628	0.3304
Scenario 3	0.1889	0.0244	0.2003	0.0242

6. Application to a Real Survey

In this section, the proposed estimators were applied to a real survey involving discrimination and the under-representation of young Indigenous, Montubios and Afro-Ecuadorian people in Ecuador. The RDS methodology was applied to a population of young (18 to 29 years old) Indigenous, Montubios and Afro-Ecuadorian people living in the city of Riobamba (Ecuador). They have historically been suffering from exclusion and under-representation and therefore, this group lacks a reliable sampling frame [32–35]. A total of 814 people were recruited in six waves and questioned on their social and economic background and living conditions using a dual system of incentives to motivate recruitment. The reported income of the household is the variable of interest and the age of the respondent is the covariate. This is unpublished data that is intended for publication in a manuscript that is in preparation [36].

Good overall performance of the two proposed estimators for the covariance and the correlation coefficient, with a bias approximately around 5% and similar small values of the relative mean squared error *rmse*, as shown in Tables 3 and 4.

Table 3. Percent relative bias (*rb*%) and the relative mean squared error (*rmse*%) for estimating S_{yx} with estimators $\widehat{S}_{yx,RDS-II}$ and $\widehat{S}_{yx,RDS-SS}$ for the ethnic example.

$\widehat{S}_{yx,RDS-II}$		$\widehat{S}_{yx,RDS-SS}$	
***rb*%**	***rmse*%**	***rb*%**	***rmse*%**
−5.671021	0.8551362	−7.147357	0.5216048

Table 4. Percent relative bias (*rb*%) and the relative mean squared error (*rmse*%) for estimating the correlation coefficient ρ with estimators $\widehat{\rho}_{RDS-II}$ and $\widehat{\rho}_{RDS-SS}$ for the ethnic example.

$\widehat{\rho}_{RDS-II}$		$\widehat{\rho}_{RDS-SS}$	
***rb*%**	***rmse*%**	***rb*%**	***rmse*%**
6.1371	0.4460	4.9065	0.2407

7. Discussion

RDS were used extensively to study the prevalence of a disease. As more RDS practitioners are incorporating this methodology to their toolbox, model-fitting in an RDS framework has become an important issue of interest. We proposed a new sample weight estimation method for continuous data. Our approach is most appropriate for situations in which homophily is small. While we consider this is a novel approach for continuous RDS data, accounting for clustering remains an open question. It is possible to extend this methodology to adjusting to clusters, as part of future research.

As an illustration of the applicability of the proposed method, we performed a simulation study and an application to an ethnic example. Nevertheless, the focus of our work has been to propose a method for estimating non-linear parameters with new sample weights. We derived expressions of the variances and showed that the proposed estimators have desirable properties. Our simulation study does not show significant differences in terms of bias or root mean square error between the two proposed estimators. Furthermore, the calculation complexity of the two estimators is similar. There is therefore no objective reason to prefer one over the other.

Taken together, the results about the dependence between continuous variables presented in this paper add to the growing literature on respondent-driven sampling, allowing researchers to obtain better information about key hidden populations.

Author Contributions: The authors contributed equally to this work in conceptualization, methodology, software and original draft preparation. All authors have read and agree to the published version of the manuscript.

Funding: The work was supported by the Ministerio de Economia, Industria y Competitividad, Spain, under Grant MTM2015-63609-R.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RDS	Respondent-Driven Sampling
MCMC	Markov Chain Monte Carlo

References

1. Heckathorn, D. Respondent-driven sampling: A new approach to the study of hidden populations. *Soc. Probl.* **1997**, *44*, 174–199. [CrossRef]
2. Salganik, M.; Heckathorn, D. Sampling and estimation in hidden populations using respondent-driven sampling. *Sociol. Methodol.* **2004**, *34*, 193–240. [CrossRef]

3. Volz, E.; Heckathorn, D. Probability based estimation theory for respondent driven sampling. *J. Off. Stat.* **2008**, *14*, 79–97.
4. Spiller, M.W.; Gile, K.J.; Handcock, M.S.; Mar, C.M.; Wejnert, C. Evaluating Variance Estimators for Respondent-Driven Sampling. *J. Surv. Stat. Methodol.* **2018**, *6*, 23–45. [CrossRef] [PubMed]
5. Beaudry, I.S.; Gile, K.J. Correcting for differential recruitment in respondent-driven sampling data using ego-network information. *Electron. J. Stat.* **2020**, *14*, 2678–2713. [CrossRef]
6. Rocha, L.E.; Thorson, A.E.; Lambiotte, R.; Liljeros, F. Respondent-driven sampling bias induced by community structure and response rates in social networks. *J. R. Stat. Soc. Ser. A Stat. Soc.* **2017**, *180*, 99–118. [CrossRef]
7. Kan, M.; Garfinkel, D.; Samoylova, O.; Gray, R.; Little, K. Social network methods for HIV case-finding among people who inject drugs in Tajikistan. *J. Int. AIDS Soc.* **2018**, *21*, 57–64. [CrossRef]
8. Sypsa, V.; Psichogiou, M.; Paraskevis, D. Rapid decline in HIV incidence among persons who inject drugs during a fast-track combination prevention program after an HIV outbreak in Athens. *J. Infect. Dis.* **2017**, *215*, 1496–1505.
9. Card, K.G.; Lachowsky, N.J.; Cui, Z. Exploring the role of sex-seeking apps and websites in the social and sexual lives of gay, bisexual and other men who have sex with men: a cross-sectional study. *Sex Health* **2017**, *14*, 229–237. [CrossRef]
10. Bernard, J.; Daňková, H.; Vašát, P. Ties, sites and irregularities: pitfalls and benefits in using respondent-driven sampling for surveying a homeless population. *Int. J. Soc. Res. Methodol.* **2018**, *21*, 603–618. [CrossRef]
11. Hipp, L.; Kohler, U.; Leumann, S. How to Implement Respondent-Driven Sampling in Practice: Insights from Surveying 24-Hour Migrant Home Care Workers. *Surv. Methods Insights Field* **2019**. [CrossRef]
12. Wong, W.; Wang, S.; Liu, H. Bootstrap standard error estimations of non-linear transport models based on linearly projected data. *Transp. A* **2018**, *15*, 1–35.
13. Imani, M.; Ghoreishi, S.F.; Braga-Neto, U. Bayesian Control of Large MDPs with Unknown Dynamics in Data-Poor Environments. *Adv. Neural Dyn.* **2018**, 8146–8156.
14. Imani, M.; Dougherty, E.R.; Braga-Neto, U. Boolean Kalman Filter and Smoother under Model Uncertainty. *Automatica* **2020**. [CrossRef]
15. Avery, L.; Rotondi, N.; McKnight, C.; Firestone, M.; Smylie, J.; Rotondie, M. Unweighted regression models perform better than weighted regression techniques for respondent-driven sampling data: results from a simulation study. *BMC Med. Res. Methodol.* **2019**, *19*, 202. [CrossRef]
16. Beckett, M.; Firestone, M.A.; McKnight, C.D. A cross-sectional analysis of the relationship between diabetes and health access barriers in an urban First Nations population in Canada. *BMJ Open* **2017**, *8*, e018272. [CrossRef]
17. da Silva Lima, F.S.; Merchán-Hamann, E.; Urdaneta, M. Fatores associados à violência contra mulheres profissionais do sexo de dez cidades brasileiras. *CAD Saude Publica* **2017**, *33*, 1–15.
18. Selvaraj, B.; Boopathi, K.; Paranjape, R.; Mehendale, S. A single weighting approach to analyze respondent-driven sampling data. *Indian J. Med. Res.* **2016**, *144*, 447–459.
19. Hubbart, A.E.; Ahern, J.; Fleischer, N.L.; Van der Laan, M.; Lippman, S.A.; Jewell, T.B.; Satariano, W.A. To GEE or not to GEE. *Epidemiology* **2010**, *21*, 467–474. [CrossRef]
20. Rao, S.; LaRoque, R.; Jentes, E. Comparison of methods for clustered data analysis in a non-ideal situation: results from an evaluation of predictors of yellow fever vaccine refusal in the global TravEpiNet (GTEN) consortium. *Int. J. Stat. Med. Res.* **2014**, *3*, 215–223. [CrossRef]
21. Lyons, C.E.; Grosso A.; Drame, F.M.; Physical and sexual violence affecting female sex workers in Abidjan, Côte d'Ivoire: prevalence, and the relationship with the work environment, HIV, and access to health services. *J. Acquir. Immune Defic. Syndr.* **2017**, *75*, 9–17. [CrossRef] [PubMed]
22. Schwartz, S.; Papworth, E.; Thiam-Niangoin, M. An urgent need for integration of family planning services into HIV care. *J. Acquir. Immune Defic. Syndr.* **2015**, *68*, 91–98. [CrossRef] [PubMed]
23. de Matos, M.A.; da Silva França, D.D.; dos Santos Carneiro, M.A. Viral hepatitis in female sex workers using the respondent-driven sampling. *Rev. Saude Publica* **2017**, *51*, 1–11. [CrossRef]
24. Scheim, A.; Bauer, G.; Coleman, T. Sociodemographic differences by survey mode in a respondent-driven sampling study of transgender people in Ontario, Canada. *LGBT Health* **2016**, *3*, 391–395. [CrossRef] [PubMed]

25. Pan, X.; Wu, M.; Ma, Q. High prevalence of HIV among men who have sex with men in Zhejiang, China: A respondent-driven sampling survey. *BMJ Open* **2015**, *5*, 1–7. [CrossRef] [PubMed]
26. Maragh-Bass, A.C.; Powell, C.; Park, J. Sociodemographic and access related correlates of health-care utilization among African American injection drug users: the BESURE study. *J. Ethn. Subst. Abuse* **2017**, *16*, 344–362. [CrossRef]
27. Gile, K.; Handcock, M. Respondent-driven sampling: An assessment of current methodology. *Sociol. Methodol.* **2010**, *40*, 285–327. [CrossRef]
28. Särndal C.E.; Swensson B.; Wretman J. *Model Assisted Survey Sampling*; Springer: Berlin/Heidelberg, Germany, 1992.
29. Hansen, M.H.; Hurvitz, W.N. On the theory of sampling from finite populations. *Ann. Math. Stat.* **1943**, *14*, 333–362. [CrossRef]
30. Deville J.C.; Särndal C.E. Calibration estimators in survey sampling. *J. Am. Stat. Assoc.* **1992**, *87*, 376–382. [CrossRef]
31. Wolter, K. *Introduction to Variance Estimation*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2007.
32. Larrea, C.; Torres, F.; López, N.; Rueda, M. *Pueblos indíGenas, Desarrollo Humano y Discriminación en el Ecuador*; Abya Yala: Quito, Ecuador, 2007.
33. Chisaguano, S. La Población Indígena del Ecuador (Análisis de Estadísticas Socio-Demográficas). INEC. Available online: https://www.acnur.org/fileadmin/Documentos/Publicaciones/2009/7015.pdf (accessed on 5 May 2020).
34. Araki, H. Movimientos éTnicos y Multiculturalismo en el Ecuador: Pueblos indíGenas, Afrodescendientes y Montubios. Master's Thesis, University of Kanagawa, Kanagawa, Japan, 2012. Available online: http://klibredb.lib.kanagawa-u.ac.jp/dspace/handle/10487/12164 (accessed on 5 May 2020).
35. Uquillas, J.; Carrasco, T.; Rees, M. *Exclusión Social y Estrategias de vida de los indíGenas Urbanos en Perú*; Banco Mundial: Quito, Ecuador, 2003.
36. Mullo, H.S.; Sánchez-Borrego, I.; Pasadas, S. Respondent-driven sampling for surveying ethnic minority in Ecuador. Manuscript in preparation.

© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).