THE LOGISTIC GENERALIZED REGRESSION ESTIMATOR WITH RANDOMIZED RESPONSE SAMPLING WITHOUT REPLACEMENT IN FINITE POPULATIONS: A UNIFYING APPROACH

Víctor H. Soberanis Cruz[1], Víctor Miranda-Soberanis
Universidad de Quintana Roo, Colonia del Bosque, Chetumal, Quintana Roo, México, CP77000

ABSTRACT
The randomized response technique (RM) introduced by Warner (1965) was designed to avoid non-answers to questions about sensitive issues and protect the privacy of the interviewee. In this paper, a model assisted survey sampling approach is used to propose an estimator of the total of individuals with some sensitive characteristic; i.e., we use an auxiliar variable (Fuller and Park, *et al*., 2006) in a logistic regression model to improve the estimator. We also propose a model (the G model) that unifies several other RM approaches under the finite population sampling scheme and the π −estimators (Särndal, *et al*., 1992; Cassel, *et al*., 1976) framework. We also propose an estimator for the variance of the estimator.

MSC: 62D05

KEY WORDS: randomized responses, logistic regression, π -estimator.

RESUMEN
La técnica de respuestas Aleatorizadas (RM) introducida por Warner (1965) fue diseñada para evitar la no-respuesta en preguntas sensitivas y para proteger la privacidad del entrevistado. En este artículo, el enfoque de muestreo asistido por un modelo es usado para proponer un estimador para el total de individuos con una característica sensitiva; i.e., disponemos de una variable auxiliar (Fuller and Park, *et al.*, 2006) en un modelo de regresión logística para mejorar la estimación. Proponemos también un modelo, el modelo G, que recoge a varios otros mecanismos aleatorios bajo un esquema de muestreo sin reemplazo de poblaciones finitas y en el marco de la teoría de los estimadores π (Särndal, *et al*., 1992; Cassel, *et al*., 1977). También proponemos un estimador para la varianza de nuestro estimador.

1. INTRODUCTION

In survey studies, interest is frequently focused on issues that are sensitive or confidential for the interviewees, such as use of drugs, tax evasion, sexual preferences, honesty in exams, opinions on authorities, etc. For this reason, some interviewees refuse to respond (no response phenomenon) to the questions designed to obtain information on a sensitive aspect, or they give a false answer. In either case the estimations are biased.

The random response technique, introduced by Warner (1965), proposes a solution to protect the privacy of the interviewee consisting of using a random mechanism (RM) by which one of two questions is selected: Do you belong to the group with characteristic A? or Do you belong to the group that does not have characteristic A?, in which A is the sensitive characteristic of interest. The interviewee will answer yes or no, and the interviewer, thus protecting her or his privacy.

The RM technique has encouraged a series of approaches, among which the following models are outstanding: (a) the W model (Warner, 1965), (b) the U model with an innocuous unrelated question W (Greenberg *et al*., 1969), (c) the C model with an innocuous unrelated question W correlated with the

1 1 Corresponding autor vsobera@uqroo.mx

sensitive variable Y; (d) the H model (Horvitz *et al*., 1976), (e) the D model (Devore, 1977), and (f) the M model (Mangat and Singh, 1990). Each of these is described below.

The U model (Greenberg *et al*., 1969) is a randomized response approach with unrelated questions. As the W model, it has a random mechanism that selects one of two questions, but while one question refers to a sensitive subject (Do you belong to the group with characteristic A?), the second question has nothing to do with the sensitive subject, but is about some other innocuous aspect W; that is, it does not affect the interviewee's sensitivity. For example, if the first question is "Do you evade taxes?", the second question could be "Do you like the movies?" The W and U models were compared within the framework of infinite populations (Moors, 1971), and the outcome revealed that the U model was more efficient than the W model. The C model that we introduce is like the U model except that the innocuous aspect W is correlated with the sensitive characteristic Y.

Horvitz *et al*. (1976) proposed the H model, which allows for greater protection of the interviewee's anonymity, without the use of the complementary question. Each element of the sample responds randomly to one of three propositions: (1) the sensitive question, (2) an instruction that says "yes", and (3) an instruction that says "no", to be chosen with probabilities of P_1, P_2 and P_3 respectively, with $P_1 + P_2 + P_3 = 1$

In the M model, the random mechanism provides *n* independent responses with two random components. The D model is analogous to U, with a basic difference: belonging to the innocuous group W is established with probability one.

Chaudhuri and Mukerjee (1988) present a good review of the pioneer work in randomized responses. Other studies are those of Lakshmi and Raghavarao (1992), Mangat*et al*. (1993), Chua and Tsui (2000), Padmawar and Vijayan (2000), and Chaudhuri (2001). A Bayesian approach to the Warner model can be seen in Winkler and Franklin (1979) and Bar-Lev *et al*. (2003).

This paper proposes a logistic regression estimator whose auxiliary variable $\underline{x}$ is innocuous and is correlated with the sensitive variable Y, but does not affect the individual's sensitivity maintaining the privacy of the interviewee. In this way we get a better estimation in terms of bias and variance, under finite populations without replacement sampling setting. Also, it is proposed that these schemes be unified into a G randomized response model, such that the W, U, C, H, D, and M models are peculiar cases. Estimators for the variances of the different models are proposed, and dispersion of the estimator is studied by simulation.

2. FRAMEWORK

A finite population $U = \{1, 2, \dots, N\}$ is considered. It is assumed that the size of population *N* is known. Sample size is denoted by *n*, which is not necessarily fixed. Let *y* be the dichotomized variable that refers to the individual's belonging to the group with the sensitive characteristic of interest, with y_k the value of *y* for the k^{th}element of the population. Thus, $y_k$is unknown but not random with $y_k = 1$ if the k^{th} individual has the sensitive characteristic A, and $y_k = 0$ otherwise. What is to be estimated is $t_A = \sum_U y_k$, the total number of individuals of the population with the sensitive characteristic A.

3. SAMPLING PROCEDURE

The sampling procedure is as follows:
Step 1 (sample selection). A sample size of size *n* is extracted in accordance with the sampling design *p*(*s*) with positive probabilities of inclusion π_k and π_{kl} where

$$\pi_k = \Pr(k \epsilon S) = \sum_{\{k \epsilon s\}} p(s) \quad y \quad \pi_{kl} = \Pr(k \& l \epsilon S) = \sum_{\{k \& l \epsilon s\}} p(s)$$

For each element *k* in sample *S*, $I_k = 1$ if $k \epsilon S$; otherwise, $I_k = 0$. Note that $I_k(S)$ is a function of the random variable S. Also,

$$Cov(I_k, I_l) = \Delta_{kl} = \pi_{kl} - \pi_k \pi_l, k \neq l;\ \Delta_{kk} = \pi_k(1 - \pi_k);\ \breve{\Delta}_{kl} = {}^{\Delta_{kl}}/_{\pi_{kl}}$$

Step 2 (information gathering, we follow Cassel, *et al*., 1976). The interviews of individuals of the sample are conducted in accordance with the RM defined for the randomized response model used. For each $k \in S$, RM induces a random variable Z_k, so that the linear combination $\hat{Z}_k = aZ_k + b_k$ is an unbiased estimation of y_k, where a and b_k are known constants that do not depend on RM; therefore, $E_{RM}(\hat{Z}_k) = y_k, V_{RM}(\hat{Z}_k) = a^2 V_{RM}(Z_k), E_{RM}(b_k) = b_k, k \epsilon S$, where $V_{RM}(\hat{Z}_k)$ represents the variance of $\hat{Z}_k$ computed from the randomized response technique (RM), i.e., the variance obtained for the W Model (assisted for) is $V_W(\hat{Z}_k)$, and so on. The same interpretation follows for $E_{RM}(\hat{Z}_k)$. Thus

$$y_k = E_{RM}(\hat{Z}_k) = E_{RM}(aZ_k + b_k) = aE_{RM}(Z_k) + b_k,$$

$$E(\hat{Z}_k) = E_\xi E_{RM}(\hat{Z}_k) = E_\xi(y_k) = \mu_k,$$

$$\lambda_k \equiv E(Z_k) = E_\xi E_{RM}(Z_k) = E_\xi \frac{(y_k - b_k)}{a}$$

$$= \frac{1}{a}(\mu_k - b_k)$$

where the operator $E_\xi()$ is computed in relation to some superpopulation model ξ, and

$$\mu_k = E_\xi(Y_k) = \Pr\left(Y_k = 1 \middle| \underline{x}; \underline{\beta}\right) = \frac{1}{1 + \exp(-\underline{x}_k^t \underline{\beta})}, \forall k \in U$$

so

$$\mu_k = a\lambda_k + b_k.$$

4. LOGISTIC REGRESSION MODEL ESTIMATOR

The Generalized Logistic Regression Estimator $\hat{t}_{AG,LGREG}$ for t_A that we present is an extension of the estimator by Lehtonen and Veijanen (1998). We assume that $\underline{y} = (y_1, \ldots, y_k, \ldots, y_N)^t$ is a realization of the random vector $\underline{Y} = (Y_1, \ldots, Y_k, \ldots, Y_N)^t$whose components are independent random variables with distribution given by

$$\Pr\{Y_k = 1 | \underline{x}_k; \underline{\beta}\} = \frac{\exp\{\underline{x}_k^t \underline{\beta}\}}{1 + \exp\{\underline{x}_k^t \underline{\beta}\}}, k = 1, 2, \ldots, N$$

This superpopulation model will be referred as ξ.
Now

$$t_A = \sum_U y_k = \sum_U \mu_k + \sum_U (y_k - \mu_k)$$

which allows introducing the Logistic Generalized Regression Estimator (LGREG) by Lehtonen and Veijanen (1998), which is given by:

$$\hat{t}_{A,LGREG} = \sum_U \hat{\mu}_k + \sum_s \frac{(\hat{Z}_k - \hat{\mu}_k)}{\pi_k}$$

where

$$\hat{\mu}_k = \mu\left(\underline{x}_k^t \hat{\underline{\beta}}\right) = \frac{1}{1 + \exp\{-\underline{x}_k^t \hat{\underline{\beta}}\}}$$

and $\hat{\underline{\beta}}$ is the Maximum Likelihood Estimator of $\underline{\beta}$ in the model ξ, which is obtained in the usual manner in the superpopulation model setup:

$$L\left(\underline{\beta} \middle| \underline{z}\right) = \prod_U \Pr\{Z_k = z_k\}$$

$$= \prod_U \lambda_k^{z_k}(1-\lambda_k)^{1-z_k} I_{\{0.1\}}(z_k)$$

This function is the complete finite population likelihood, i.e., as if $(z_k, \underline{x}_k)$ were observed for all $k \epsilon U$, like in a census. Then

$$l\left(\underline{\beta}\middle|\underline{z}\right) = \log L\left(\underline{\beta}\middle|\underline{z}\right)$$
$$= \sum_U [z_k \ln \lambda_k + (1-z_k)\ln(1-\lambda_k)]$$

where $\lambda_k = E(Z_k)$.

5. MAXIMUM LIKELIHOOD ESTIMATOR OF $\underline{\beta}$

The equation $\frac{\partial}{\partial \underline{\beta}} l\left(\underline{\beta}\middle|\underline{z}\right) = \underline{0}$ defines the parameter $\underline{\beta}$, then we estimate it through the $\pi - estimator$ of $l\left(\underline{\beta}\middle|\underline{z}\right)$:

$$\hat{l}_\pi\left(\underline{\beta}\middle|\underline{z}\right) = \sum_s \pi_k^{-1}[z_k \ln \lambda_k + (1-z_k)\ln(1-\lambda_k)]$$

In general, we maximize the likelihood function numerically by appropriate numerical methods such as Newton-Raphson algorithm. Now

$$\frac{\partial \lambda_k}{\partial \beta_j} = \frac{1}{a}\frac{\partial \mu_k}{\partial \beta_j} = \frac{1}{a}\mu_k(1-\mu_k)x_{kj}, j = 0,1,2,\dots,p,$$

So

$$\frac{\partial \hat{l}_\pi}{\partial \underline{\beta}} = \underline{0} \Leftrightarrow \frac{1}{a}\sum\nolimits_s \pi_k^{-1}[(z_k-\lambda_k)\frac{\mu_k(1-\mu_k)}{\lambda_k(1-\lambda_k)}]x_{kj} = 0, j = 0,1,2,\dots,p$$
$$\Leftrightarrow \sum\nolimits_s \pi_k^{-1}[(z_k-\lambda_k)\frac{\mu_k(1-\mu_k)}{\lambda_k(1-\lambda_k)}]x_{kj} = 0, j = 0,1,2,\dots,p$$

$$\Leftrightarrow \sum\nolimits_s \pi_k^{-1}\alpha_k x_{kj} = 0, j = 0,1,2,\dots,p$$
$$\Leftrightarrow (\frac{\alpha_1}{\pi_1}, \frac{\alpha_2}{\pi_2}, \dots, \frac{\alpha_{n_s}}{\pi_{n_s}}) \begin{pmatrix} 1 & x_{11} x_{12} \dots x_{1p} \\ 1 & x_{21} x_{22} \dots x_{2p} \\ & \vdots \\ 1 & x_{n_s 1} & x_{n_s 2} \dots x_{n_s p} \end{pmatrix}$$

$$\Leftrightarrow \underline{\alpha}_{S,\pi}^t X_S = \underline{0}^t \Leftrightarrow X_S^t \underline{\alpha}_{S,\pi} = \underline{0},$$

where $\alpha_k = (z_k - \lambda_k)\frac{\mu_k(1-\mu_k)}{\lambda_k(1-\lambda_k)}$ and $\underline{\alpha}_{S,\pi} = (\frac{\alpha_k}{\pi_k})_{k \epsilon S}$. Solving $\frac{\partial \hat{l}_\pi}{\partial \underline{\beta}} = \underline{0} \Leftrightarrow X_S^t \underline{\alpha}_{S,\pi} = \underline{0}$, we obtain $\hat{\underline{\beta}}$. Once we get $\hat{\underline{\beta}}$, the estimator that we propose for the total of individuals with the sensitive characteristic is given by:

$$\hat{t}_{AG,LGREG} = \sum_U \hat{\mu}_k + a\sum_S \frac{z_k - \hat{\lambda}_k}{\pi_k}$$

$$= \sum_U \hat{\mu}_k + \sum_S \frac{\hat{z}_k - \hat{\mu}_k}{\pi_k} \quad (1)$$

where $\hat{\lambda}_k = \frac{\hat{\mu}_k - b_k}{a}, \hat{Z}_k = aZ_k + b_k, k \epsilon S, and\ \hat{\mu}_k = 1\Big/(1+\exp\left(-\underline{x}^t\hat{\underline{\beta}}\right))$ $\forall k \in U$. It is interesting to note that the estimator in (1) is analogous to the Minimum Dispersion Estimator proposed by Gutierrez and Breidt (2009).

6. ESTIMATOR OF THE VARIANCE ESTIMATOR

Following the theory of π-estimator (Ës̈arndal, *et al*, 1992) and Lehtonen and Veijanen(1998), we propose

$$\hat{V}\left(\hat{t}_{AG,LGREG}\right) = a^2 \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{z_k-\hat{\lambda}_k}{\pi_k}\right)\left(\frac{z_l-\hat{\lambda}_l}{\pi_l}\right) \quad (2)$$

$$= \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{\hat{Z}_k-\hat{\mu}_k}{\pi_k}\right)\left(\frac{\hat{Z}_l-\hat{\mu}_l}{\pi_l}\right)$$

as an estimator for $V\left(\hat{t}_{AG,LGREG}\right)$. The goodness of this estimator is analyzed via simulation.

7. THE UNIFIED APPROACH

7.1 WARNER'S MODEL: W MODEL

$$Z_k = \begin{cases} y_k & \text{with probability P;} \\ 1-y_k & \text{with probability } 1-\text{P} \end{cases}$$

for $k \in S$, and

$$\hat{Z}_k = \frac{Z_k-(1-P)}{2P-1}$$

$$= \frac{1}{2P-1}Z_k - \frac{1-P}{2P-1}$$

$$= aZ_k + b$$

7.2 *H MODEL*

$$Z_k = \begin{cases} y_k & with\ probability\ P_1; \\ 1 & with\ probability\ P_2; \\ 0 & with\ probability\ 1-P_1-P_2 \end{cases}$$

$$then\ \ E_{RC}(Z_k) = y_kP_1 + P_2 \quad and \quad \hat{Z}_k = \frac{Z_k-P_2}{P_1} = \frac{1}{P_1}Z_k - \frac{P_2}{P_1} = aZ_k + b \quad \text{i.e.} \quad \hat{Z}_k = aZ_k + b$$

7.3 *U AND C MODELS*

$$Z_k = \begin{cases} y_k & with\ probability\ P; \\ w_k & with\ probability\ 1-P \end{cases}$$

$$then \quad E_{RC}(Z_k) = y_kP + w_k(1-P) \quad and \quad \hat{Z}_k = \frac{Z_k-(1-P)w_k}{P}$$

$$= \frac{1}{P}Z_k - \frac{1-P}{P}w_k$$

$$= aZ_k + b_k$$

7.4 *DEVORE'S MODEL: D MODEL*

$$Z_k = \begin{cases} y_k & with\ probability\ P; \\ 1 & with\ probability\ 1-P \end{cases}$$

$$then \quad E_{RC}(Z_k) = y_k P + (1-P) \quad \text{and} \quad \hat{Z}_k = \frac{Z_k - (1-P)}{P}$$

$$= \frac{1}{P} Z_k - \frac{1-P}{P}$$

$$= aZ_k + b$$

7.5 *M MODEL*

$$Z_k = \begin{cases} y_k & with\ probability\ T; \\ Py_k + (1-P)(1-y_k) & with\ probability\ 1-T \end{cases}$$

$$then \quad E_{RC}(Z_k) = y_k T + (1-T)[Py_k + (1-P)(1-y_k)]$$

$$= [T + (1-T)(2P-1)]y_k + (1-T)(1-P)$$

and $$\hat{Z}_k = \frac{Z_k - (1-T)(1-P)}{[T+(1-T)(2P-1)]}$$

$$= \frac{1}{[T + (1-T)(2P-1)]} Z_k - \frac{(1-T)(1-P)}{[T + (1-T)(2P-1)]}$$

$$= aZ_k + b$$

The following table shows the values of a and b_k for the RMs considered in this work.

Table 1. Values of a and b_k for the different randomized response techniques.

	a	b_k
W	$\frac{1}{2P-1}$	$-\frac{1-P}{2P-1}$
H	$\frac{1}{P_1}$	$-\frac{P_2}{P_1}$
U	$\frac{1}{P}$	$-\frac{(1-P)w_k}{P}$
C	$\frac{1}{P}$	$-\frac{(1-P)w_k}{P}$
D	$\frac{1}{P}$	$-\frac{(1-P)}{P}$
M	$\frac{1}{T + (1-T)(2P-1)}$	$-\frac{(1-T)(1-P)}{T + (1-T)(2P-1)}$

8. SIMPLE RANDOM SAMPLING

$p(s) = \Pr(S = s) = \binom{N}{n}^{-1} \forall s \in \mathscr{A}$,

$$\pi_k = \frac{n}{N} = f;\ \pi_{kl} = \begin{Bmatrix} \frac{n(n-1)}{N(N-1)} & ,k \neq l \\ \pi_k & ,k = l \end{Bmatrix}, \Delta_{kl} = \begin{Bmatrix} \frac{f(f-1)}{N-1} & ,k \neq l \\ f(f-1) & ,k = l \end{Bmatrix}, \frac{\Delta_{kl}}{\pi_{kl}\pi_k\pi_l}$$

$$= \begin{Bmatrix} -\frac{1-f}{(n-1)f^2} & ,k \neq l \\ \frac{1-f}{f^2}, & k = l \end{Bmatrix}$$

$$D_{nxn} = D = \left[\frac{\Delta_{kl}}{\pi_{kl}\pi_k\pi_l}\right]_{nxn} = \frac{1-f}{f^2}\begin{pmatrix} 1 & -\frac{1}{n-1} & -\frac{1}{n-1} & -\frac{1}{n-1} & -\frac{1}{n-1} \\ -\frac{1}{n-1} & 1 & -\frac{1}{n-1} & -\frac{1}{n-1} & -\frac{1}{n-1} \\ & & \vdots & & \\ & -\frac{1}{n-1} & -\frac{1}{n-1}\cdots & & 1 \end{pmatrix}$$

and

$$\hat{V}(\hat{t}_{AG,LGREG}) = \sum\sum_S \frac{\Delta_{kl}}{\pi_{kl}}\left(\frac{\hat{Z}_k - \hat{\mu}_k}{\pi_k}\right)\left(\frac{\hat{Z}_l - \hat{\mu}_l}{\pi_l}\right) = \left(\underline{\hat{Z}}_S - \underline{\hat{\mu}}_S\right)^t D\left(\underline{\hat{Z}}_S - \underline{\hat{\mu}}_S\right)$$

9. SIMULATION

We simulated a finite population with N=700, $t_A = 477$, n=140 and did compare models W and H. The following results were obtained:

Table 2. Total population and variances estimated for models W and H with =700, $t_A = 477$ and n=140

	$\hat{t}_A$	$\sqrt{Var(\hat{t}_A)}$	$\sqrt{\hat{V}ar(\hat{t}_A)}$
W	475.94	74.35	64.004
H	478.34	36.03	34.04

The following table shows that high positive correlate W_1 variable in the random mechanism with Y in model C produces a very important reduction in the variance estimator. For each row we simulate M=700 times and fixed cor.yW1

The table 3 shows that high positive correlations between the W_1 variable in the random mechanism and Y in model C produces a very important reduction in the variance of the estimator.

In this work a model assisted (Model C) survey sampling approach by using an auxiliary variable was used to propose an estimator of the total of individuals with some sensitive characteristics. Simulations were carried out under this framework.

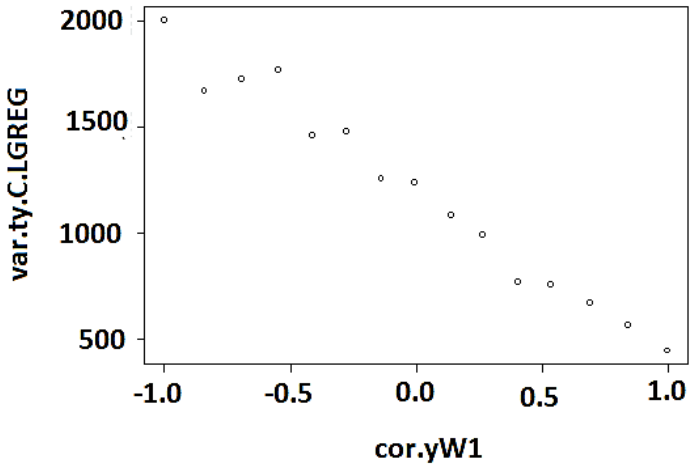
Results of simulations for Model C were compared with those for Models W and H (Table 2 and Table3). Our results suggest that model C is more efficient than these traditional randomized response techniques also based on a model assisted approach, by producing a significant reduction in the variance of the estimator. This relationship follows under positive correlation between W_1 and Y, as well as under the finite population sampling scheme and the π −estimators.

Table 3. Total population estimated and standard deviation of the estimator, for different fixed correlations between W_1 variable and Y in model C (M = 700).

MODEL C

	Mean.$\hat{t}_A$	cor.yW1	$\sqrt{Var(\hat{t}_A)}$	$\sqrt{\hat{V}ar(\hat{t}_A)}$
1	476.66	-1.00	44.77	40.11
2	478.43	-0.84	40.88	38.96
3	479.01	-0.69	41.50	37.94
4	476.82	-0.54	42.03	36.78
5	479.34	-0.41	38.16	35.73
6	476.90	-0.27	38.43	34.60
7	479.41	-0.13	35.40	33.19
8	478.04	-0.00	35.15	31.99
9	478.00	0.13	32.88	30.73
10	477.23	0.26	31.47	29.27
11	477.05	0.40	27.73	27.74
12	477.44	0.53	27.42	26.23
13	477.95	0.69	25.87	24.66
14	477.39	0.84	23.78	22.58
15	477.56	0.99	21.04	20.85

Figure 1. Estimated variances (horizontal axis) against correlations for W_1 variable in the random mechanism, and Y in model C .



Other approaches for RM´s may be considered in order to compare with our results, as well as for emphasize that the proportion of reduction of the variance for the estimator is a very important advantage of our approach.

RECEIVED DECEMBER 2010
REVISED JULY, 2011

REFERENCES

[1] BAR-LEV, S. K. BOBOVICH, E. and . BOUKAI, B. (2003): A common conjugate prior structure for several randomized response models. **Test**, 12 , 101-113 .

[2] CASSEL, C.M, SÄRNDAL, C.E, and WRETMAN, J.H. (1976): Some results on generalized difference estimation and generalized regression estimation for finite populations. **Biometrika**, 63, 3, 615-20.

[3] CHAUDHURI, A. and MUKERJEE, R. (1988): **Randomized Response: Theory and Techniques.** Marcel Dekker, New York.

[4] CHAUDHURI, A. (2001): Using randomized response from a complex survey to estimate a sensitive proportion in a dichotomous finite population. **Jour. Stat. Plan. Inf**. 94:37–42.

[5] CHUA, T. C., and A. K. TSUI. (2000). Procuring honest responses indirectly. **Journal of Statistical Planning and Inference**. 90: 107–116.

[6] DEVORE, J. L. 81977): A note on the randomized response technique. **Communications in StatisticsTheory and Methods**. 6: 1525–1529.

[7] FULLER, W. A. and PARK, M. (2006): Generalized Regression Estimators. **Encyclopedia of Envirometrics.**

[8] GREENBERG B.G., ABUL-ELA, ABDEL-LATIF, A., SIMMONS, W.R., and HORVITZ, D.C . (1969):The unrelated question RR model: theoretical framework. **Journal of the American Statistical Association**, 64, 520-539.

[9] GUTIÉRREZ, H.A, and BREIDT, F.J. (2009): Estimation of the Population Total using Generalized Difference Estimator and Wilcoxon Ranks. **Revista Colombiana de Estadística**, 32, 123-143.

[10] HORVITZ,D.C., GREENBERG, B. G., and ABERNATHY, J. R.(1976): RR: a data gathering device for sensitive questions. **Internat.Statist.Rev**. **44**,181-196.

[11] LAKSHMI, D. V., and D. RAGHAVARAO. (1992): A test for detecting untruthful answering in randomized response procedures. **Journal of Statistical Planning and Inference**. 31: 387–390.

[12] LEHTONEN, R., and VEIJANEN, A., (1998): Logistic Generalized Regression Estimators. **Survey Methodology**, 24, 51-55.

[13] MANGAT, N. S., SINGH, R. (1990): An alternative randomized response procedure, **Biometrika**, 77, 439-442.

[14] MOORS, J.J.A. (1971):. Optimization of the unrelated question RR model.**Journal of the American Statistical Association,66**, 627-629.

[15] PADMAWAR, V. R., and K. VIJAYAN (2000): Randomized response revisited. **Journal of Statistical Planning and Inference.** 90: 293– 304.

[16] SÄRNDAL, C.E., SWENSSON, B., and WRETMAN, J. (1992): **Model Assisted Survey Sampling**. New York: Springer Verlag.

[17] WINKLER,R. and FRANKLIN,L. (1979): Warner's randomized response model: A Bayesian approach. **J. Amer. Statist. Assoc**, 74 207-214.

[18] WARNER, S. L. (1965): Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias, **Journal of the American Statistical Association**, 60, 63-69.

APPENDIX

```
N<-700
Iset<-1:N
x<-sample(14:70,N,replace=T)
B0<--3
B1<-0.1
mu<-function(B0,B1,x) (exp(B0+B1*x)/(1+exp(B0+B1*x)))
p<-mu(B0,B1,x)
plot(x,p)
y<-rep(2,N)
u1<-runif(N)
for (k in 1:N){if (u1[k]<=p[k]) (y[k]<-1) else
                (y[k]<-0)
        }
A<-sum(y)
A

#MU HATS
#Bhat Matrix of estimated betas via Newton-Raphson
mu.hat.U<-matrix(rep(0,N*M),nrow=N)
for (j in 1:M){
            for (i in 1:N) (mu.hat.U[i,j]<-mu(Bhat[j,1],Bhat[j,2],x[i]))
        }
mu.hat.s<-matrix(rep(0,n*M),nrow=n)
for (j in 1:M) (mu.hat.s[,j]<-mu.hat.U[s[,j],j])

#W&MAS&LGREG
Zs<-matrix(rep(2,n*M),nrow=n)
U<-matrix(rep(0,n*M),nrow=n)
tY.W.LGREG<-rep(0,M)
P<-0.70
a<-1/(2*P-1)
b<--(1-P)/(2*P-1)
for (j in 1:M){
        U[,j]<-runif(n)
        for (k in 1:n){
                if (U[k,j]<P) (Zs[k,j]<-y[s[k,j]])
                if (U[k,j]>P)(Zs[k,j]<-1-y[s[k,j]])
                }
            Zs[,j]
            tY.W.LGREG[j]<-sum(mu.hat.U[,j])+(1/f)*(sum(a*Zs[,j]+b)-sum(mu.hat.s[,j]))
        }
hist(tY.W.LGREG)
mean.tY.W.LGREG<-mean(tY.W.LGREG)
var.tY.W.LGREG<-var(tY.W.LGREG)
D<-matrix(rep(-1/(n-1),n*n),nrow=n)
diagD<-rep(1,n)
diag(D)<-diagD
D<-((1-f)/(f*f))*D
var.hat.tY.W.LGREG<-rep(0,M)
for (j in 1:M){var.hat.tY.W.LGREG[j]<-(t((a*Zs[,j]+b)-mu.hat.s[,j])%*%D%*%((a*Zs[,j]+b)-mu.hat.s[,j]))}
mean.var.hat.tY.W.LGREG<-mean(var.hat.tY.W.LGREG)
cWarner<-c(mean.tY.W.LGREG,var.tY.W.LGREG,mean.var.hat.tY.W.LGREG)
cWarner



#H&MAS&LGREG
Zs<-matrix(rep(2,n*M),nrow=n)
U<-matrix(rep(0,n*M),nrow=n)
tY.H.LGREG<-rep(0,M)
P1<-0.70
P2<-0.15
for (j in 1:M){
        U[,j]<-runif(n)
        for (k in 1:n){
                if (U[k,j]<P1) (Zs[k,j]<-y[s[k,j]])
                if  ((P1<=U[k,j])&(U[k,j]<P1+P2))(Zs[k,j]<-1)
```

```
                        if (U[k,j]>P1+P2)(Zs[k,j]<-0)
                        }
                Zs[,j]
                tY.H.LGREG[j]<-sum(mu.hat.U[,j])+(1/P1)*(1/f)*sum(Zs[,j]-(P1*mu.hat.s[,j]+P2))
        }
hist(tY.H.LGREG)
mean.tY.H.LGREG<-mean(tY.H.LGREG)
var.tY.H.LGREG<-var(tY.H.LGREG)
a<-1/P1
b<--P2/P1
D<-matrix(rep(-1/(n-1),n*n),nrow=n)
diagD<-rep(1,n)
diag(D)<-diagD
D<-((1-f)/(f*f))*D
var.hat.tY.H.LGREG<-rep(0,M)
for (j in 1:M){var.hat.tY.H.LGREG[j]<-(t((a*Zs[,j]+b)-mu.hat.s[,j])%*%D%*%((a*Zs[,j]+b)-mu.hat.s[,j]))}
mean.var.hat.tY.H.LGREG<-mean(var.hat.tY.H.LGREG)
cH<-c(mean.tY.H.LGREG,var.tY.H.LGREG,mean.var.hat.tY.H.LGREG)
cH


#W is a matrix whose columns are non-related questions with differents correlation with Y.
N<-700
y1<-y
M<-700
W<-matrix(rep(0,M*N),nrow=N)
for (k in 1:M)(y1[k]<-1-y1[k])&(W[,k]<-y1)
cor.yW<-cor(y,W)

#C0&MAS&LGREG
Zs<-matrix(rep(2,n*M),nrow=n)
U<-matrix(rep(0,n*M),nrow=n)
tY.C0.LGREG<-rep(0,M)
P<-0.70
a<-1/P
b<--(1-P)/P
B0<-b*W[,200]
for (j in 1:M){
        U[,j]<-runif(n)
        for (k in 1:n){
                        if (U[k,j]<P) (Zs[k,j]<-y[s[k,j]])
                        if (U[k,j]>P)(Zs[k,j]<-W[s[k,j],200])
                        }
                Zs[,j]
                tY.C0.LGREG[j]<-t(rep(1,N))%*%mu.hat.U[,j]+(1/f)*t(rep(1,n))%*%(a*Zs[,j]+B0[s[,j]]-mu.hat.s[,j])
        }
hist(tY.C0.LGREG)
mean.tY.C0.LGREG<-mean(tY.C0.LGREG)
var.tY.C0.LGREG<-var(tY.C0.LGREG)
D<-matrix(rep(-1/(n-1),n*n),nrow=n)
diagD<-rep(1,n)
diag(D)<-diagD
D<-((1-f)/(f*f))*D
var.hat.tY.C0.LGREG<-rep(0,M)
for (j in 1:M){var.hat.tY.C0.LGREG[j]<-t(a*Zs[,j]+B0[s[,j]]-mu.hat.s[,j])%*%D%*%(a*Zs[,j]+B0[s[,j]]-mu.hat.s[,j])}
mean.var.hat.tY.C0.LGREG<-mean(var.hat.tY.C0.LGREG)
cC0<-c(mean.tY.C0.LGREG,var.tY.C0.LGREG,mean.var.hat.tY.C0.LGREG)
cC0


#C&MAS&LGREG
x<-seq(0,700,by=50)
x[1]<-1
x<-x[sort.list(-x)]
x
W1<-W[,x]
cor.yW1<-cor(y,W1)
cor.yW1
lx<-length(x)
tY.C.LGREG<-matrix(rep(0,lx*M),nrow=M)
```

```
P<-0.70
a<-1/P
b<--(1-P)/P
B<-b*W1#B<-b*W[,x]
Zs<-matrix(rep(2,n*M),nrow=n)
U<-matrix(rep(0,n*M),nrow=n)
for (j1 in 1:lx){
                  for (j in 1:M){
                        U[,j]<-runif(n)
                              for (k in 1:n){
                                          if (U[k,j]<P) (Zs[k,j]<-y[s[k,j]])
                                          if (U[k,j]>P)(Zs[k,j]<-W[s[k,j],x[j1]])
                                          }
               Zs[,j]
               tY.C.LGREG[j,j1]<-t(rep(1,N))%*%mu.hat.U[,j]+(1/f)*t(rep(1,n))%*%(a*Zs[,j]+B[s[,j],j1]-mu.hat.s[,j])
                        }
            }
mean.tY.C.LGREG<-apply(tY.C.LGREG,2,mean)
var.tY.C.LGREG<-apply(tY.C.LGREG,2,var)
var.hat.tY.C.LGREG<-matrix(rep(0,lx*M),nrow=M)
for (j1 in 1:lx){
            for (j in 1:M){var.hat.tY.C.LGREG[j,j1]<-t(a*Zs[,j]+B[s[,j],j1]-mu.hat.s[,j])%*%D%*%(a*Zs[,j]+B[s[,j],j1]-mu.hat.s[,j])
                     }
            }
var.hat.tY.C.LGREG<-apply(var.hat.tY.C.LGREG,2,mean)
df2<-data.frame(mean.tY.C.LGREG,cor.yW1,var.tY.C.LGREG,var.hat.tY.C.LGREG)
sd.tY.C.LGREG<-sqrt(var.tY.C.LGREG)
sd.hat.tY.C.LGREG<-sqrt(var.hat.tY.C.LGREG)
df3<-data.frame(mean.tY.C.LGREG,cor.yW1,sd.tY.C.LGREG,sd.hat.tY.C.LGREG)
plot(cor.yW1,var.tY.C.LGREG)
```