

REGRESSION ANALYSIS USING SCRAMBLED RESPONSES

SARJINDER SINGH¹, A.H. JOARDER²
AND MAXWELL L. KING³

*The Australian Bureau of Statistics, University of Sydney
and Monash University*

Summary

This paper investigates the general linear regression model $Y = X\beta + e$ assuming the dependent variable is observed as a scrambled response using Eichhorn & Hayre's (1983) approach to collecting sensitive personal information. The estimates of the parameters in the model remain unbiased, but the variances of the estimates increase due to scrambling. The Wald test of the null hypothesis $H_0: \beta = \beta_0$, against the alternative hypothesis $H_a: \beta \neq \beta_0$, is also investigated. Parameter estimates obtained from scrambled responses are compared to those from conventional or direct-question surveys, using simulation. The coverage by nominal 95% confidence intervals is also reported.

Key words: Regression analysis; scrambled responses; sensitive issues; randomized response technique; Wald test.

1. Introduction

The randomized response technique is used for reducing problems caused by false responses when potentially sensitive questions are present in surveys of human populations. Direct questioning of the respondents about sensitive issues often results in either refusal to respond or falsification of the answers. This introduces a non-sampling error that can bias sample-based estimates. The problem arises when a survey asks personal questions on a sensitive issue, e.g. the illegal use of drugs, level of income, incidents of incest or acts of domestic violence. The most serious difficulty in studying these types of problems is the lack of reliable data on their incidence. Social stigma and fear of reprisals sometimes cause respondents to lie when they are approached with the conventional or direct-question survey method.

Warner (1965) was the first to suggest an ingenious method of collecting information on sensitive characteristics. His idea has spawned a vast literature

Received December 1995; revised June 1996; accepted June 1996.

¹Methodology Division, The Australian Bureau of Statistics, Belconnen, ACT 2616.

²Dept of Mathematical Sciences, The University of Sydney, NSW 2006.

³Dept of Econometrics, Monash University, Clayton, Vic 3168.

Acknowledgements. The authors thank an Associate Editor and a referee for constructive suggestions that have improved this paper. This research was conducted when all three authors were members of the Dept of Econometrics at Monash University.

which has been reviewed by Fox & Tracy (1986), Chaudhuri & Mukherjee (1987, 1988) and Sheers (1992). Some recent modifications to the model have been suggested by Franklin (1989), Kuk (1990), Mangat (1994), Mangat & Singh (1990), Singh & Singh (1993), Singh (1994), Singh *et al.* (1994) and Bansal *et al.* (1994). Maddala (1983) suggested combining randomized response survey data with other personal information to estimate logit models predicting immoral, unpopular, or unlawful behaviour. To our knowledge, Kerkvliet (1994) was the first to implement Maddala's technique using real data. Maddala's (1983) model deals with qualitative characteristics only. Eichhorn & Hayre (1983) introduced an interesting scrambled randomized response method to deal with qualitative sensitive characteristics. Their method can be summarized as follows: respondents scramble their responses by multiplying them by a random scrambling variable, and then reveal only the scrambled results to the interviewer. Mahajan *et al.* (1994) extended the Eichhorn & Hayre (1983) model to the stratified random sampling scheme and developed the theory for finding the optimum strata boundaries.

This paper has two aims. The first is to develop new theory for scrambled responses obtained by adopting Eichhorn & Hayre's (1983) approach to the general linear regression model. The second aim is to show, through simulation, that the proposed approach provides fruitful results in the application of regression analysis to the scrambled responses.

2. Theory

We assume the variable of interest Y_i is related to k non-stochastic regressors via the classical linear regression model

$$Y = X\beta + e, \quad (2.1)$$

where Y is an n -vector of values Y_i , X is an $n \times k$ matrix of regressors, β is a k -vector of unknown regression coefficients and e is an n -vector of independent normally distributed error terms with variance σ^2 . In our case, Y_i is a sensitive variable whose observations have to be obtained by survey methods. Because some respondents are unlikely to reply truthfully to questions about behaviour which is immoral, unpopular or unlawful, Eichhorn & Hayre's (1983) scrambled response approach is applied as follows:

each respondent is requested to report the product $Y_i S_i$ where S_i is the value of the scrambling variable drawn by the i th respondent. The privacy of the respondent is protected by the fact that S_i is not known to the interviewer, although its distribution and, in particular, its mean $E(S_i) = \theta$ and variance $\text{var}(S_i) = \gamma^2$ are known. The scrambling device may be a deck of cards, spinner etc., following some suitable distribution, e.g. Normal, Weibull or any discrete distribution. The $Y_i S_i$ value obtained from the i th respondent can be standardized as $Z_i = Y_i S_i / \theta$ after collection.

Our interest is in estimating and testing β using the model

$$Z = X\beta + \eta, \quad (2.2)$$

where Z is the n -vector of Z_i values and η is an n -vector of errors whose distribution is unknown.

Theorem 1. $E(\eta) = 0$ and the OLS estimator $\hat{\beta}_* = S^{-1}X^tZ = (X^tX)^{-1}X^tZ$ of β in (2.2), where $S = X^tX$, is unbiased.

This theorem follows immediately from the fact that

$$E(Z) = E_M E_R(Z) = E_M(Y) = X\beta,$$

where E_M and E_R denote, respectively, the expected value with respect to model (2.1) and the distribution of the randomization device providing the S_i values.

To find the variance of the OLS estimator $\hat{\beta}_*$, we need the following lemma.

Lemma 1. The variance-covariance matrix V_R of Z over the distribution of S_i is given by

$$V_R(Z) = C_\gamma^2 \text{diag}(Y_1^2, \dots, Y_n^2) \quad (2.3)$$

where $C_\gamma = \gamma/\theta$ is the coefficient of variation of the scrambling device.

Theorem 2. The variance-covariance matrix $V(\hat{\beta}_*)$ of the estimator $\hat{\beta}_*$ equals

$$\sigma^2(1 + C_\gamma^2)S^{-1} + C_\gamma^2 S^{-1} X^t \text{diag} \left[\left(\sum_{j=1}^k X_{1j} \beta_j \right)^2, \dots, \left(\sum_{j=1}^k X_{nj} \beta_j \right)^2 \right] X S^{-1}. \quad (2.4)$$

Proof. Let V_M denote the variance-covariance matrix over the model (2.1). Then $V(\hat{\beta}_*)$ equals

$$\begin{aligned} & E_M(V_R(\hat{\beta}_*)) + V_M E_R(\hat{\beta}_*) \\ &= E_M(V_R[S^{-1}X^tZ]) + V_M E_R[S^{-1}X^tZ] \\ &= E_M[S^{-1}X^t C_\gamma^2 \text{diag}(Y_1^2, \dots, Y_n^2) X S^{-1}] + V_M[S^{-1}X^t Y] \\ &= \sigma^2[S^{-1} + C_\gamma^2 S^{-1}] + C_\gamma^2 S^{-1} X^t \text{diag} \left[\left(\sum_{j=1}^k X_{1j} \beta_j \right)^2, \dots, \left(\sum_{j=1}^k X_{nj} \beta_j \right)^2 \right] X S^{-1}. \end{aligned}$$

Hence the theorem.

Theorem 3. An estimator of σ^2 is given by

$$\hat{\sigma}_*^2 = \frac{\hat{\eta}^t \hat{\eta} - C_\gamma^2 \sum_{i=1}^n \left(\sum_{j=1}^k \hat{\beta}_{*j} X_{i,j} \right)^2 M_{ii}}{n - k + C_\gamma^2 \sum_{i=1}^n M_{ii}}, \quad (2.5)$$

where $\hat{\eta} = Z - X\hat{\beta}_*$ is the OLS residual vector from (2.2) and M_{ii} denotes the i th diagonal element of the matrix $M = I - XS^{-1}X^t$.

Proof. We have

$$\begin{aligned} E(\hat{\eta}^t \hat{\eta}) &= E_M E_R [(Z - X\hat{\beta}_*)^t (Z - X\hat{\beta}_*)] \\ &= E_M E_R [Z^t (I - XS^{-1}X^t) Z] \\ &= \left(n - k + C_\gamma^2 \sum_{i=1}^n M_{ii} \right) \sigma^2 + C_\gamma^2 \sum_{i=1}^n \left(\sum_{j=1}^k \beta_j X_{i,j} \right)^2 M_{ii}. \end{aligned} \quad (2.6)$$

By the method of moments, (2.6) with unknown β_j values replaced by estimates gives (2.5). Hence the theorem.

Theorem 4. The Wald test statistic to test the null hypothesis $H_0: \beta = \beta_0$ against the alternative hypothesis $H_a: \beta \neq \beta_0$ for the scrambled response model is given by

$$(\hat{\beta}_* - \beta_0)^t [\hat{V}(\hat{\beta}_*)]^{-1} (\hat{\beta}_* - \beta_0) \sim_d \chi^2(p) \quad \text{under } H_0,$$

assuming that $\hat{\beta}_* \sim_d N(\beta, V(\hat{\beta}_*))$ and where $\hat{V}(\hat{\beta}_*)$, an estimator of $V(\hat{\beta})$ is

$$\hat{\sigma}_*^2 (1 + C_\gamma^2) S^{-1} + C_\gamma^2 S^{-1} X^t \text{diag} \left[\left(\sum_{j=1}^k X_{1j} \hat{\beta}_{*j} \right)^2, \dots, \left(\sum_{j=1}^k X_{n,j} \hat{\beta}_{*j} \right)^2 \right] X S^{-1}.$$

3. Empirical Study of the Properties of the Proposed Model

An empirical simulation study was conducted to check the properties of the proposed procedure. The study also assessed estimates and test outcomes calculated from the proposed scrambling procedure, in comparison with those that might be obtained from direct questioning with different simulated levels of lying. The model used throughout the experiment was

$$Y_i = 20 + 0.5X_i + 0.9W_i + E_i \quad (i = 1, \dots, n), \quad (3.1)$$

where the regressors X_i and W_i and the errors E_i are generated as independent standard normal variates using the RNOR subroutine from the NSWCLIB library of mathematics subroutines. We used the following method to generate Y_i^* , the

reported response from respondent i when directly questioned about his/her Y_i value. Another uniform random variable L_i ($i = 1, \dots, n$), was generated to select randomly true and false responses under the assumption that all respondents have an equal probability of reporting untruthfully. If the value of L_i is less than p , where p is a fixed proportion, then the i th respondent reports $Y_i^* = rY_i$; otherwise he/she reports $Y_i^* = Y_i$, where $r \leq 1$ is the fixed under-reporting percentage represented as a proportion. Then the model

$$Y_i^* = a + bX_i + cW_i + E_i^* \quad (i = 1, \dots, n), \quad (3.2)$$

was fitted by OLS (ordinary least squares) and the properties of the estimates observed over 2000 iterations for $r = 0.6, 0.8, 0.9, 1.0$ and $p = 0.25, 0.50$. Here $r = 1.0$ represents the true model, i.e. the model in which all the respondents report truthfully. The situation in which respondents with higher values of Y_i report untruthfully was also simulated. In this case, respondents with the highest 10% and highest 25% of Y_i values in the sample report $Y_i^* = rY_i$ while the remainder report $Y_i^* = Y_i$.

In generating scrambled responses for Y_i , we used the uniform distribution over different sets of known positive integers because the mean and variance of these distributions can be easily calculated. We used the subroutine URGET from the same library of subroutines to generate random integers S_i with equal probability over any set of integers. If S_i is any integer in the set $\{991, \dots, 999\}$ then $C_\gamma = 0.26\%$, or in the set $\{981, \dots, 999\}$ then $C_\gamma = 0.55\%$, or in the set $\{800, \dots, 999\}$ then $C_\gamma = 6.42\%$, or in the set $\{700, \dots, 999\}$ then $C_\gamma = 10.2\%$ or in the set $\{600, \dots, 999\}$ then $C_\gamma = 14.4\%$. The simulation assumes 100% cooperation from the respondents when scrambling is used so that $Z_i = Y_i S_i / \theta$. Then the model

$$Z_i = a + bX_i + cW_i + \eta_i \quad (i = 1, \dots, n), \quad (3.3)$$

was fitted by OLS and the properties of the estimates observed over 2000 iterations. For each situation and coefficient, we computed the bias, mean squared error and standard deviation of the 2000 estimates for all models. The coverage of the nominal 95% confidence interval (CCI) was also calculated. Here CCI is the proportion of times the actual parameter β_j falls in the interval defined as $\hat{\beta}_j \pm 1.96\sqrt{\hat{V}(\hat{\beta}_j)}$ in 2000 iterations, where $\hat{\beta}_j$ is the particular estimator being studied.

The results obtained are presented in Tables 1–3 for sample size $n = 50, 100$ and 800. The values of the estimates hardly differ between $n = 500$ and $n = 800$. Table 1 shows that as the percentage of randomly selected lying increases (i.e. r decreases), the value of the MSE (mean square error) of estimates for the direct-question model also increases. The situation is worse when the respondents with higher values of Y_i lie (Table 2) because then the value of CCI reduces to zero

TABLE 1

Average estimates of the parameters, with SD and MSE of estimates, estimated size α , coefficient of determination R^2 and power $P(w)$ of Wald test over 2000 iterations in the direct-question model at 4 levels of randomly selected untruthful answering

	$r = 1.00$ (true model)			$r = 0.90$			$r = 0.80$			$r = 0.60$		
	a	b	c	a	b	c	a	b	c	a	b	c
$n = 50, p = 0.25$												
est.	19.985	0.481	0.886	19.461	0.465	0.864	18.937	0.451	0.842	17.890	0.420	0.798
SD	0.143	0.144	0.139	0.184	0.187	0.183	0.283	0.291	0.284	0.521	0.534	0.525
CCI	0.943	0.948	0.948	0.182	0.942	0.945	0.023	0.941	0.947	0.002	0.940	0.949
MSE	0.021	0.021	0.020	0.323	0.036	0.035	1.208	0.087	0.085	4.722	0.292	0.287
bias	-0.015	-0.019	-0.014	-0.539	-0.035	-0.036	-1.063	-0.049	-0.058	-2.110	-0.080	-0.102
$n = 50, p = 0.50$												
est.				18.985	0.455	0.843	17.984	0.430	0.799	15.983	0.379	0.713
SD				0.192	0.202	0.191	0.309	0.325	0.308	0.582	0.608	0.509
CCI				0.000	0.939	0.937	0.000	0.936	0.935	0.000	0.934	0.936
MSE				1.067	0.043	0.040	4.157	0.111	0.105	16.473	0.384	0.371
bias				-1.015	-0.045	-0.057	-2.016	-0.070	-0.101	-4.017	-0.121	-0.187
$n = 100, p = 0.25$												
est.	19.993	0.494	0.893	19.492	0.477	0.869	18.992	0.460	0.845	17.991	0.426	0.796
SD	0.102	0.103	0.097	0.133	0.116	0.128	0.200	0.204	0.194	0.362	0.366	0.352
CCI	0.941	0.942	0.954	0.031	0.948	0.948	0.001	0.945	0.943	0.000	0.947	0.941
MSE	0.011	0.011	0.010	0.275	0.019	0.017	1.056	0.043	0.041	4.167	0.139	0.134
bias	-0.007	-0.006	-0.007	-0.508	-0.023	-0.031	-1.008	-0.040	-0.055	-2.009	-0.074	-0.104
$n = 100, p = 0.50$												
est.				18.993	0.464	0.845	17.994	0.435	0.796	15.994	0.374	0.698
SD				0.143	0.141	0.138	0.226	0.223	0.221	0.417	0.413	0.411
CCI				0.000	0.951	0.932	0.000	0.952	0.925	0.000	0.949	0.924
MSE				1.033	0.021	0.022	4.076	0.539	0.060	16.222	0.186	0.210
bias				-1.007	-0.036	-0.055	-2.006	-0.065	-0.104	-4.006	-0.126	-0.202
$n = 800, p = 0.25$												
est.	19.998	0.499	0.899	19.499	0.486	0.875	19.001	0.474	0.851	18.005	0.448	0.804
SD	0.035	0.035	0.035	0.047	0.046	0.046	0.072	0.069	0.069	0.130	0.125	0.127
CCI	0.952	0.949	0.951	0.000	0.933	0.912	0.000	0.933	0.889	0.000	0.939	0.881
MSE	0.001	0.001	0.001	0.252	0.002	0.003	1.001	0.005	0.007	3.994	0.018	0.025
bias	-0.002	-0.001	-0.001	-0.501	-0.014	-0.025	-0.999	-0.026	-0.049	-1.995	-0.052	-0.096
$n = 800, p = 0.50$												
est.				18.996	0.474	0.853	17.996	0.449	0.807	15.994	0.398	0.716
SD				0.049	0.049	0.049	0.079	0.079	0.081	0.147	0.148	0.151
CCI				0.000	0.919	0.857	0.000	0.911	0.809	0.000	0.909	0.785
MSE				1.009	0.003	0.005	4.022	0.009	0.015	16.061	0.322	0.057
bias				-1.004	-0.026	-0.047	-2.004	-0.051	-0.093	-4.006	-0.102	-0.184
where												
p	R^2	α	$P(w)$	R^2	α	$P(w)$	R^2	α	$P(w)$	R^2	α	$P(w)$
$n = 50$												
0.25	0.519	0.057	0.922	0.377	0.056	0.694	0.213	0.057	0.321	0.095	0.055	0.116
0.50				0.346	0.054	0.595	0.174	0.052	0.226	0.064	0.055	0.085
$n = 100$												
0.25	0.519	0.058	0.999	0.377	0.061	0.962	0.205	0.058	0.62	0.077	0.057	0.33
0.50				0.338	0.056	0.915	0.159	0.054	0.44	0.054	0.054	0.11
$n = 800$												
0.25	0.514	0.055	1.000	0.372	0.057	1.000	0.197	0.055	1.00	0.064	0.054	0.922
0.50				0.334	0.054	1.000	0.152	0.057	0.999	0.023	0.053	0.649

TABLE 2

Average estimates of the parameters, with SD and MSE, estimated size α , coefficient of determination R^2 and power $P(w)$ of Wald test over 2000 iterations in the direct-question model with 3 levels of untruthful answering at a higher value of study variable Y_i

	$r = 0.90$			$r = 0.80$			$r = 0.60$		
	a	b	c	a	b	c	a	b	c
$n = 50, p = 0.10$									
est.	19.767	0.346	0.642	19.548	0.212	0.398	19.110	-0.055	-0.091
SD	0.143	0.134	0.134	0.154	0.175	0.165	0.201	0.321	0.284
CCI	0.556	0.767	0.489	0.229	0.606	0.172	0.104	0.599	0.113
MSE	0.057	0.041	0.084	0.228	0.113	0.280	0.832	0.412	1.063
bias	-0.233	-0.154	-0.258	-0.452	-0.288	-0.502	-0.890	-0.555	-0.991
$n = 50, p = 0.25$									
est.	19.431	0.232	0.431	18.876	-0.016	-0.024	17.766	-0.513	-0.933
SD	0.148	0.119	0.126	0.184	0.185	0.175	0.297	0.391	0.341
CCI	0.014	0.382	0.037	0.001	0.229	0.003	0	0.295	0.006
MSE	0.346	0.086	0.236	1.297	0.301	0.884	5.076	1.178	3.479
bias	-0.569	-0.268	-0.469	-1.124	-0.516	-0.924	-2.234	-1.013	-1.833
$n = 100, p = 0.10$									
est.	19.771	0.357	0.647	19.549	0.219	0.399	19.104	-0.056	-0.948
SD	0.105	0.095	0.092	0.116	0.121	0.113	0.152	0.217	0.185
CCI	0.311	0.652	0.200	0.039	0.362	0.011	0.002	0.305	0.003
MSE	0.063	0.029	0.073	0.216	0.093	0.263	0.824	0.375	3.449
bias	-0.229	-0.143	-0.253	-0.451	-0.281	-0.501	-0.896	-0.556	-1.848
$n = 100, p = 0.25$									
est.	19.453	0.248	0.450	18.912	0.002	0.007	17.831	-0.490	-0.879
SD	0.112	0.086	0.088	0.139	0.127	0.121	0.218	0.262	0.233
CCI	0.000	0.162	0.002	0.000	0.034	0.000	0.000	0.055	0.000
MSE	0.311	0.071	0.210	1.201	0.263	0.812	4.749	1.049	3.222
bias	-0.547	-0.252	-0.450	-1.088	-0.498	-0.893	-2.169	-0.990	-1.779
$n = 800, p = 0.10$									
est.	19.773	0.361	0.649	19.549	0.223	0.401	19.101	-0.053	-0.097
SD	0.036	0.033	0.034	0.040	0.043	0.041	0.054	0.076	0.068
CCI	0.000	0.014	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MSE	0.052	0.020	0.064	0.204	0.078	0.251	0.811	0.313	1.000
bias	-0.227	-0.139	-0.251	-0.451	-0.277	-0.499	-0.899	-0.553	-0.987
$n = 800, p = 0.25$									
est.	19.453	0.254	0.458	18.909	0.009	0.017	17.821	-0.480	-0.864
SD	0.039	0.029	0.032	0.048	0.043	0.043	0.076	0.090	0.083
CCI	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
MSE	0.299	0.061	0.196	1.191	0.242	0.781	4.751	0.968	3.119
bias	-0.547	-0.246	-0.442	-1.091	-0.491	-0.883	-2.179	-0.980	-1.764
where									
p	R^2	α	$P(w)$	R^2	α	$P(w)$	R^2	α	$P(w)$
$n = 50$									
0.10	0.408	0.050	0.478	0.153	0.050	0.111	0.037	0.053	0.097
0.25	0.287	0.064	0.206	0.036	0.059	0.110	0.154	0.059	0.059
$n = 100$									
0.10	0.407	0.058	0.818	0.142	0.065	0.286	0.019	0.056	0.190
0.25	0.294	0.055	0.391	0.018	0.061	0.251	0.132	0.060	0.153
$n = 800$									
0.10	0.402	0.049	0.999	0.132	0.052	0.930	0.004	0.046	0.732
0.25	0.294	0.045	0.998	0.002	0.042	0.895	0.119	0.041	0.653

TABLE 3
Average estimates of the parameters, with SD and MSE, estimated size α , and power $P(w)$ of Wald test
over 2000 iterations in the proposed scrambled randomized response model
at 5 values for the coefficient of variation C_γ of the scrambling variable

n	$C_\gamma = 0.26\%$			$C_\gamma = 0.55\%$			$C_\gamma = 6.42\%$			$C_\gamma = 10.2\%$			$C_\gamma = 14.4\%$		
	a	b	c	a	b	c	a	b	c	a	b	c	a	b	c
50	est.	19.985	0.480	0.887	19.985	0.480	0.889	19.982	0.479	0.917	19.980	0.479	0.935	19.977	0.479
	SD	0.143	0.144	0.139	0.143	0.145	0.140	0.226	0.237	0.226	0.316	0.331	0.317	0.427	0.446
	CCI	0.957	0.946	0.951	0.982	0.943	0.953	1.000	0.950	0.962	1.000	0.952	0.962	1.000	0.956
	MSE	0.021	0.021	0.020	0.021	0.021	0.020	0.051	0.057	0.052	0.101	0.110	0.102	0.183	0.199
	bias	-0.015	-0.020	-0.013	-0.015	-0.020	-0.011	-0.018	-0.021	0.017	-0.020	-0.021	0.035	-0.013	-0.021
100	est.	19.994	0.495	0.894	19.994	0.495	0.895	19.993	0.493	0.909	19.992	0.491	0.918	19.991	0.489
	SD	0.102	0.103	0.097	0.103	0.104	0.098	0.168	0.167	0.163	0.235	0.232	0.229	0.316	0.312
	CCI	0.965	0.942	0.956	0.994	0.941	0.956	1.000	0.953	0.966	1.000	0.951	0.972	1.000	0.958
	MSE	0.011	0.011	0.010	0.012	0.012	0.010	0.028	0.028	0.027	0.055	0.054	0.053	0.100	0.098
	bias	-0.006	-0.005	-0.006	-0.006	-0.005	-0.005	-0.007	-0.007	0.009	-0.008	-0.009	0.018	-0.009	-0.011
800	est.	19.998	0.499	0.899	19.998	0.499	0.899	19.997	0.501	0.902	19.997	0.501	0.904	19.997	0.502
	SD	0.036	0.035	0.035	0.036	0.035	0.036	0.058	0.057	0.058	0.081	0.079	0.080	0.109	0.108
	CCI	0.999	0.948	0.949	1.000	0.948	0.957	1.000	0.978	0.993	1.000	0.984	0.996	1.000	0.984
	MSE	0.001	0.001	0.001	0.001	0.001	0.001	0.003	0.003	0.003	0.006	0.006	0.006	0.012	0.012
	bias	-0.002	-0.001	-0.001	-0.002	-0.001	-0.001	-0.003	0.001	0.002	-0.003	0.001	0.004	-0.003	0.002
where															
n	α	$P(w)$		α	$P(w)$		α	$P(w)$		α	$P(w)$		α	$P(w)$	
50	0.066	0.923		0.066	0.932		0.064	0.586		0.062	0.311		0.049	0.169	
100	0.061	0.999		0.064	0.999		0.056	0.868		0.052	0.556		0.050	0.298	
800	0.057	1.000		0.056	1.000		0.050	1.000		0.041	1.000		0.033	0.991	

in most of the cases. Also the increase in bias in Table 2 is more rapid than in Table 1. In addition, the estimation biases are a problem only for the intercept term in Table 1 while being a problem for all coefficient estimates in Table 2. It appears that random under-reporting does not strongly bias estimates of slope parameters, whereas selective under-reporting can result in biased estimates of all parameters. The value of CCI obtained from the scrambled response model is greater than 0.95 in almost all the cases and significantly so in the case of a , the intercept term (Table 3). In addition, the bias is smaller than for the direct questioning approach. Table 3 also shows that the MSEs of estimates for the scrambled response model are an increasing function of C_γ . If $C_\gamma \leq 0.55\%$, then the MSE values of the estimates are the same as for the true model. But here respondent cooperation is likely to be least. It seems logical that as C_γ increases, the respondents' cooperation would typically increase. In contrast, the efficiency of the estimates decreases.

Table 3 shows that if C_γ increases to 14.4%, the increase in MSE of the estimates of a , b and c for the scrambled response model is 0.021 to 0.183, 0.021 to 0.199 and 0.020 to 0.187, respectively, for a sample of size 50. Thus the respective relative efficiency of the estimators,

$$RE = \frac{MSE_{r=1}}{MSE_{C_\gamma > 0}} \times 100, \quad (3.4)$$

is given by 11.475%, 10.553% and 10.969% for the estimators of a , b and c respectively when $n = 50$ and $C_\gamma = 14.4\%$. If $n = 50$ and $C_\gamma = 6.42\%$ then the respective values of RE are 41.176%, 36.842% and 38.462%. The same trend is observed for the other large sample cases, though the relative gain is less. Thus the value of RE decreases very rapidly for given n if the value of C_γ increases from 6.42% to 14.4%. Clearly a very high value of C_γ is not recommended.

Tables 1 and 2 show that if the respondents report less than the actual value, then we typically under-estimate the actual parameters. The bias in the estimates increases due to the level of untruthful response. If we define the percentage relative efficiency of the estimators under different levels of lying as

$$RE_L = \frac{MSE_{r=1}}{MSE_{r < 1}} \times 100, \quad (3.5)$$

then one can see easily from Table 1 that RE_L is a drastically decreasing function of $(1 - r)$, namely the level of lying. For example, if $n = 50$ and $r = 0.60$, then the RE_L values for the estimators of a , b and c take the extremely low values of 0.445%, 7.192% and 6.968%, respectively. The same trend is observed in Table 2. On the other hand, Table 3 shows that if the sample size is large, then the estimates remain close to the actual values of the parameters in the case of scrambled responses. For example, if $n = 800$ and $C_\gamma = 6.42\%$ then the average estimates of a , b and c are 19.997, 0.501 and 0.902 respectively, with respective

standard deviations of 0.058, 0.057 and 0.058. Thus these average estimates are close to the actual values of the parameters in model (3.1) as predicted by Theorem 1.

In the simulation study the null hypothesis, $H_0: b = 0, c = 0$, is tested against the alternative hypothesis $H_a: b \neq 0, c \neq 0$. The probability of a Type I error, α , and selected powers of the Wald test, $P(w)$, are estimated — in the case of the α values, by resimulating (3.1) with $b = 0$ and $c = 0$. In almost all cases, the resultant estimated α values are not significantly different from 0.05. The value of $P(w)$ decreases as $(1 - r)$ and/or p increase in Tables 1 and 2 for the given n . Also we see $P(w)$ is a decreasing function of C_γ in Table 3.

On the basis of our empirical study, we conclude that scrambled responses are a quite useful device to study sensitive issues in society. Scrambling does have a cost in terms of efficiency but if it can be used to gain truthful responses, we are less likely to end up with biased estimates and misleading inferences. Our study has assumed that respondents cooperate fully when being questioned using the scrambling method. We acknowledge that this is open to doubt and note it is an issue worthy of further examination.

References

- BANSAL, M.L., SINGH, S. & SINGH, R. (1994). Multi-character survey using randomized response technique. *Comm. Statist. A — Theory Methods* **23**, 1705–1715.
- CHAUDHURI, A. & MUKHERJEE, R. (1987). Randomized response techniques: a review. *Statist. Neerlandica* **41**, 27–44.
- & — (1988). *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.
- EICHHORN, B.H. & HAYRE, L.S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *J. Statist. Plann. Inference* **7**, 307–316.
- FOX, J. & TRACY, P. (1986). *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills: Sage Publication.
- FRANKLIN, L.A. (1989). A comparison of estimators for randomized response sampling with continuous distributions from a dichotomous population. *Comm. Statist. A — Theory Methods* **18**, 489–505.
- KERKVLIT, J. (1994). Estimating a logit model with randomized data: the case of cocaine use. *Austral. J. Statist.* **36**, 9–20.
- KUK, A.Y.C. (1990). Asking sensitive questions indirectly. *Biometrika* **77**, 436–438.
- MADDALA, G. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
- MAHAJAN, P.K., GUPTA, J.P. & SINGH, R. (1994). Determination of optimum strata boundaries for scrambled randomized response. *Statistica*, **54**, 375–381.
- MANGAT, N.S. (1994). An improved randomized response strategy. *J. Roy. Statist. Soc. Ser. B* **56**, 93–95.
- & SINGH, R. (1990). An alternative randomized response procedure. *Biometrika* **77**, 439–442.
- SHEERS, N. (1992). A review of randomized response technique. *Measurement and Evaluation in Counselling and Development* **25**, 27–41.
- SINGH, S. (1994). Unrelated question randomized response sampling using continuous distributions. *J. Indian Soc. Argil. Statist.* **46**, 349–361.
- , MANGAT, N.S. & SINGH, R. (1994). On estimation of mean/total of stigmatized quantitative variables. *Statistica* **54**, 383–386.

- & SINGH, R. (1993). Generalised Franklin's model for randomized response sampling. *Comm. Statist. A — Theory Methods* **22**, 741-755.
- WARNER, S.L. (1965). Randomized response: a survey technique for eliminating evasive answer bias. *J. Amer. Statist. Assoc.* **60**, 63-69.