involve

Variations of the Greenberg unrelated question binary model

David P. Suarez and Sat Gupta

msp

Variations of the Greenberg unrelated question binary model

David P. Suarez and Sat Gupta

(Communicated by Kenneth S. Berenhaut)

We explore different variations of the Greenberg unrelated question RRT model for a binary response. In one of the variations, we allow multiple independent responses from each respondent. In another variation, we use inverse sampling. It turns out that both of these variations produce more efficient models, a fact validated by both theoretical comparisons as well as extensive computer simulations.

1. Introduction

Social desirability response bias (SDB) is a major concern in surveys involving sensitive topics. One method that could help circumvent SDB is the randomized response technique, introduced originally by Warner [1965] and then generalized by other researchers such as Greenberg et al. [1969; 1971], Warner [1971], Klein and Spady [1993], Gupta et al. [2002; 2013].

RRT models have been used extensively in field surveys. Abernathy et al. [1970] used RRT models to obtain estimates of induced abortion rates in urban North Carolina. From the open survey, it was noticed that female respondents would have hesitated to respond truthfully to the sensitive question of induced abortions. Striegel et al. [2006] used indirect questioning techniques to measure the prevalence of doping among elite athletes. In order to study the effect of higher education in favourable attitudes towards foreigners in Germany, Ostapczuck et al. [2009] used two survey methods: direct questioning and RRT. The results obtained by these two survey methods demonstrated great variation. Based upon the respondents who used RRT, the results obtained showed a sharp decline in the estimates for the proportion of xenophiles among both the less educated and highly educated. Gill et al. [2013] conducted a survey which used an RRT model to estimate the risky sexual behaviors among students at the University of North Carolina at Greensboro. The binary question of interest was "Have you been told by a healthcare professional

that you have a sexually transmitted disease?", whereas the quantitative question of interest was "How many sexual partners have you had in the last 12 months?" The survey was conducted using three methods: RRT method, direct face-to-face interviewing and anonymous check-box survey method. It was observed that the optional unrelated question RRT method's estimates were closer to the check-box survey method's estimates, and the lowest point estimate was obtained by face-to-face interview method, which is expected as it provided the lowest anonymity. More recently, Chhabra et al. [2016] used these models to estimate the prevalence of sexual abuse of female college students by either a friend or an acquaintance.

In this paper, we discuss some variations of the Greenberg et al. (1969) unrelated question RRT model. In one of the variations, we allow a respondent to provide multiple independent responses. In another variation, we use the inverse sampling technique.

2. Proposed models

2.1. *Using multiple independent responses in the Greenberg model.* Let us first recall the Greenberg et al. (1969) unrelated question RRT model, which we will henceforth refer to as the Greenberg model. Let π_x be the unknown prevalence of a sensitive attribute X in the population and π_y be the known prevalence of a nonsensitive attribute Y. A randomization device offers respondents a choice between two questions, a sensitive question and an unrelated question with respective probabilities p and $1 - p$. Let p_y be the probability of a "yes" response. Then

$$p_y = \pi_x p + \pi_y (1 - p), \tag{1}$$

which leads to the estimator

$$\hat{\pi}_G = \frac{\hat{p}_y - \pi_y(1 - p)}{p}, \tag{2}$$

where $\hat{p}_y$ is the sample proportion of "yes" responses.

The mean of the estimator in (2) is given by

$$E(\hat{\pi}_G) = \pi_x,$$

which signifies that $\hat{\pi}_G$ is an unbiased estimator of π_x.

The variance of the estimator in (2) is given by

$$\mathrm{Var}(\hat{\pi}_G) = \frac{p_y(1 - p_y)}{np^2}. \tag{3}$$

Now suppose we allow m independent responses from each respondent in a sample of size n. Let T_i be the number of "yes" responses provided by the i-th

respondent. Then

$$T_i \sim \text{Binomial}(m, p_y) \quad \text{and} \quad E(T_i) = mp_y.$$

If $\bar{T} = \left(\sum T_i\right)/n$, then we know that $E(\bar{T}) = mp_y$.
Estimating mp_y by $\bar{T}$, the estimator for π_x in (2) can be refined to

$$\hat{\pi}_{GM} = \frac{\bar{T}/m - (1-p)\pi_y}{p}. \tag{4}$$

Note that

$$E(\hat{\pi}_{GM}) = \frac{E(\bar{T})/m - (1-p)\pi_y}{p} = \pi_x. \tag{5}$$

The variance of the estimator $\hat{\pi}_{GM}$ is given by

$$\text{Var}(\hat{\pi}_{GM}) = \frac{1}{m^2 p^2} \text{Var}(\bar{T}) = \frac{1}{m^2 p^2} \frac{mp_y(1-p_y)}{n} = \frac{p_y(1-p_y)}{nmp^2}. \tag{6}$$

2.2. Inverse sampling: waiting for the first "yes" response. Let each respondent continue to use the Greenberg model repeatedly until a "yes" response is recorded. Let S_i be the total number of trials needed by the i-th respondent to get to the first "yes" response. Then,

$$S_i \sim \text{Geometric}(p_y),$$

with $E(S_i) = 1/p_y$ and $\text{Var}(S_i) = (1 - p_y)/p_y^2$, where p_y is defined in (1).

Also, let there be a sample of n respondents and $\bar{S}$ be the sample mean of the S_i's. Then $1/p_y$ can be estimated by $\bar{S}$ leading to $\hat{p}_y = 1/\bar{S}$ as an estimator of p_y. Using first-order Taylor's approximation of $1/S$, we can write

$$\frac{1}{\bar{S}} \approx \frac{1}{E(S)} + (\bar{S} - E(S))\left(\frac{-1}{(E(S))^2}\right), \tag{7}$$

where $E(S) = 1/p_y$.

With this approximation,

$$E\left(\frac{1}{\bar{S}}\right) \approx \frac{1}{E(S)} = p_y. \tag{8}$$

Then, using $1/\bar{S}$ as an estimator of p_y, the estimator in (2) becomes

$$\hat{\pi}_{GI} = \frac{1/\bar{S} - (1-p)\pi_y}{p}. \tag{9}$$

Note that

$$E(\hat{\pi}_{GI}) = \frac{E(1/\bar{S}) - (1-p)\pi_y}{p} \approx \frac{p_y - (1-p)\pi_y}{p} = \pi_x$$

since $E(1/\bar{S}) \approx p_y$, as argued in (8).

Thus, we see that $\hat{\pi}_{GI}$ is an unbiased estimator of π_x, up to first order of approximation.

From (9),

$$\text{Var}(\hat{\pi}_{GI}) = \frac{1}{p^2} \text{Var}\left(\frac{1}{\bar{S}}\right). \tag{10}$$

But

$$\text{Var}\left(\frac{1}{\bar{S}}\right) \approx \text{Var}\left(\frac{1}{E(\bar{S})} + (\bar{S} - E(\bar{S}))\left(-\frac{1}{(E(\bar{S}))^2}\right)\right) = \text{Var}(-\bar{S}p_y^2)$$

$$= p_y^4 \text{Var}(\bar{S}) = p_y^4 \frac{\text{Var}(S)}{n} = p_y^4 \left(\frac{1 - p_y}{np_y^2}\right) = \frac{p_y^2(1 - p_y)}{n}. \tag{11}$$

Thus, we have

$$\text{Var}(\hat{\pi}_{GI}) \approx \frac{1}{p^2}\left(\frac{p_y^2(1 - p_y)}{n}\right) = \frac{p_y^2(1 - p_y)}{np^2}. \tag{12}$$

2.3. Inverse sampling: waiting for k "yes" responses. Let S_i be the total number of trials needed to reach the k-th "yes" response. Then, we see that $S_i \sim$ Negative Binomial(p_y, k) with

$$E(S_i) = \frac{k}{p_y} \tag{13}$$

and

$$\text{Var}(S_i) = \frac{k(1 - p_y)}{p_y^2}. \tag{14}$$

Also, let there be a sample of n respondents and $\bar{S}$ be the sample mean of the n responses. Then

$$E(\bar{S}) = E(S_i) = \frac{k}{p_y}. \tag{15}$$

Therefore, k/p_y can be estimated by $\bar{S}$ and $\hat{p}_y = k/\bar{S}$ can be used as an estimator of p_y. Using first-order Taylor's approximation,

$$\frac{1}{\bar{S}} = \frac{1}{E(S)} + (\bar{S} - E(S))\left(-\frac{1}{(E(S))^2}\right), \tag{16}$$

where $E(S) = k/p_y$.

Thus, our estimator for π_x in (2) becomes

$$\hat{\pi}_{GI_k} = \frac{k/\bar{S} - \pi_y(1 - p)}{p}. \tag{17}$$

From (17), we get the following for the mean of $\hat{\pi}_x$:

$$E(\hat{\pi}_{GI_k}) = \frac{kE(1/\bar{S}) - \pi_y(1-p)}{p}$$

$$\approx \frac{k(p_y/k) - \pi_y(1-p)}{p} = \frac{p_y - \pi_y(1-p)}{p} = \pi_x. \qquad (18)$$

Thus, $\hat{\pi}_{GI_k}$ is an unbiased estimator of π_x, up to first order of approximation.

From (17), we also get,

$$\text{Var}(\hat{\pi}_{GI_k}) = \frac{k^2}{p^2} \text{Var}\left(\frac{1}{\bar{S}}\right). \qquad (19)$$

But,

$$\text{Var}\left(\frac{1}{\bar{S}}\right) \approx \text{Var}\left(\frac{1}{E(\bar{S})} + (\bar{S} - E(\bar{S}))\left(-\frac{1}{(E(\bar{S}))^2}\right)\right) = \frac{p_y^4}{k^4} \text{Var}(\bar{S})$$

$$\approx \frac{p_y^4}{k^4}\left(\frac{k(1-p_y)/p_y^2}{n}\right) = \frac{p_y^4}{k^4}\left(\frac{k(1-p_y)}{p_y^2 n}\right) = \frac{p_y^2(1-p_y)}{k^3 n}. \qquad (20)$$

Thus, we have

$$\text{Var}(\hat{\pi}_{GI_k}) \approx \frac{k^2}{p^2}\left(\frac{p_y^2(1-p_y)}{k^3 n}\right) = \frac{p_y^2(1-p_y)}{knp^2}. \qquad (21)$$

3. Efficiency comparisons

In this section, we compare the efficiencies of the following Greenberg estimators:

$\hat{\pi}_G$ = standard estimator,

$\hat{\pi}_{GM}$ = estimator using m independent responses,

$\hat{\pi}_{GI}$ = estimator using inverse sampling, waiting for the first "yes" response,

$\hat{\pi}_{GI_k}$ = estimator using inverse sampling, waiting for the k-th "yes" response.

Since

$$\text{Var}(\hat{\pi}_{GM}) = \frac{p_y(1-p_y)}{nmp^2} = \frac{1}{m}\left(\frac{p_y(1-p_y)}{np^2}\right) = \frac{1}{m}\text{Var}(\hat{\pi}_G), \qquad (22)$$

we have $\text{Var}(\hat{\pi}_{GM}) < \text{Var}(\hat{\pi}_G)$ for $m > 1$. Thus, the Greenberg multiple response model is more efficient than the single response model.

Since

$$\text{Var}(\hat{\pi}_{GI}) = \frac{p_y^2(1-p_y)}{np^2} = p_y\left(\frac{p_y(1-p_y)}{np^2}\right) = p_y \text{Var}(\hat{\pi}_G), \qquad (23)$$

we have $\mathrm{Var}(\hat{\pi}_{GI}) < \mathrm{Var}(\hat{\pi}_G)$ for $p_y < 1$. Thus, the inverse sampling model is more efficient than the Greenberg model.

Since

$$\mathrm{Var}(\hat{\pi}_{GI}) = \frac{p_y^2(1-p_y)}{np^2} = p_y\left(\frac{p_y(1-p_y)}{np^2}\right) = mp_y\left(\frac{p_y(1-p_y)}{nmp^2}\right)$$
$$= mp_y\,\mathrm{Var}(\hat{\pi}_{GM}), \qquad (24)$$

we have $\mathrm{Var}(\hat{\pi}_{GI}) < \mathrm{Var}(\hat{\pi}_{GM})$ for $mp_y < 1$. Thus, the inverse sampling model is more efficient than the Greenberg multiple response model when $m < 1/p_y$.

Since

$$\mathrm{Var}(\hat{\pi}_{GI_k}) = \frac{p_y^2(1-p_y)}{nkp^2} = \frac{1}{k}\left(\frac{p_y^2(1-p_y)}{np^2}\right) = \frac{1}{k}\,\mathrm{Var}(\hat{\pi}_{GI}), \qquad (25)$$

we have $\mathrm{Var}(\hat{\pi}_{GI_k}) < \mathrm{Var}(\hat{\pi}_{GI})$ for $k > 1$. Thus, the inverse sampling model that waits for k "yes" responses is more efficient than the inverse sampling model that waits for the first "yes" response.

We can summarize the above observations as follows:

$$\mathrm{Var}(\hat{\pi}_{GI_k}) < \begin{cases} \mathrm{Var}(\hat{\pi}_{GI}) & \text{if } k > 1, \\ \mathrm{Var}(\hat{\pi}_{GM}) & \text{if } mp_y < 1, \\ \mathrm{Var}(\hat{\pi}_G) & \text{if } m > 1. \end{cases} \qquad (26)$$

4. Simulation results

All of the preceding theoretical formulas were tested empirically through computer simulations. Table 1 below presents simulation results that were obtained using SAS for a total of 10000 simulations with a sample size of 500, $\pi_x = 0.30$, $\pi_y = 0.7$ and $p = 0.85$. Note that the simulation results support the formulas for the means and variances of various estimators, even when first-order approximation is used.

5. Conclusion

Based on Table 1, we can see that the regular Greenberg model has higher variance (theoretical and empirical) than the modified Greenberg model with multiple responses as well as the models based on inverse sampling. Hence, the proposed variants of the Greenberg model are more efficient; although greater effort is needed in using these newer models. Given that the gain in efficiency with newer models is quite substantial, the newer models are worth trying. However, in practice, we need to keep m and k small, such as $m \leq 3$ and $k \leq 3$.

	$\hat{\pi}_G$	$\widehat{\text{Var}}(\hat{\pi}_G)$	$\text{Var}(\hat{\pi}_G)$
	0.3001781	0.000638371	0.000637785
m	$\hat{\pi}_{GM}$	$\widehat{\text{Var}}(\hat{\pi}_{GM})$	$\text{Var}(\hat{\pi}_{GM})$
1	0.3001781	0.000638371	0.000637785
2	0.3002513	0.000318622	0.000318893
3	0.3000282	0.000214628	0.000212595
4	0.3000342	0.000156454	0.000159446
5	0.3000586	0.000126368	0.000127557
	$\hat{\pi}_{GI}$	$\widehat{\text{Var}}(\hat{\pi}_{GI})$	$\text{Var}(\hat{\pi}_{GI})$
	0.3004394	0.000229236	0.000229603
k	$\hat{\pi}_{GI_k}$	$\widehat{\text{Var}}(\hat{\pi}_{GI_k})$	$\text{Var}(\hat{\pi}_{GI_k})$
1	0.3004394	0.000229236	0.000229603
2	0.3002714	0.000112837	0.000114801
3	0.3000640	0.000076382	0.000076534
4	0.3000161	0.000058789	0.000057401
5	0.3002176	0.000046028	0.000045921

Table 1. Estimators of π_x with corresponding empirical and theoretical variances.

References

[Abernathy et al. 1970] J. R. Abernathy, B. G. Greenberg, and D. G. Horvitz, "Estimates of induced abortion in urban North Carolina", *Demography* **7**:1 (1970), 19–29.

[Chhabra et al. 2016] A. Chhabra, B. K. Dass, and S. Gupta, "Estimating prevalence of sexual abuse by an acquaintance with an optional unrelated question RRT model", *North Carolina J. Math. Stat.* **2** (2016), 1–9.

[Gill et al. 2013] T. S. Gill, A. Tuck, S. Gupta, M. Crowe, and J. Figueroa, "A field test of optional unrelated question randomized response models: estimates of risky sexual behaviors", pp. 135–146 in *Topics from the 8th annual UNCG regional mathematics and statistics conference*, edited by J. Rychtář et al., Springer Proceedings in Mathematics and Statistics **64**, Springer, New York, 2013.

[Greenberg et al. 1969] B. G. Greenberg, A.-L. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz, "The unrelated question randomized response model: theoretical framework", *J. Amer. Statist. Assoc.* **64**:326 (1969), 520–539. MR

[Greenberg et al. 1971] B. G. Greenberg, R. R. Kuebler, J. R. Abernathy, and D. G. Horvitz, "Application of the randomized response technique in obtaining quantitative data", *J. Amer. Statist. Assoc.* **66**:334 (1971), 243–250.

[Gupta et al. 2002] S. Gupta, B. Gupta, and S. Singh, "Estimation of sensitivity level of personal interview survey questions", *J. Statist. Plann. Inference* **100**:2 (2002), 239–247. MR Zbl

[Gupta et al. 2013] S. Gupta, A. Tuck, T. S. Gill, and M. Crowe, "Optional unrelated-question randomized response models", *Involve* **6**:4 (2013), 483–492. MR Zbl

[Klein and Spady 1993] R. W. Klein and R. H. Spady, "An efficient semiparametric estimator for binary response models", *Econometrica* **61**:2 (1993), 387–421. MR Zbl

[Ostapczuk et al. 2009] M. Ostapczuk, J. Musch, and M. Moshagen, "A randomized-response investigation of the education effect in attitudes towards foreigners", *Eur. J. Soc. Psychol.* **39**:6 (2009), 920–931.

[Striegel et al. 2006] H. Striegel, P. Simon, J. Hansel, A. M. Niess, and R. Ulrich, "Doping and drug use in elite sports: an analysis using the randomized response technique", *Med. Sci. Sports Exerc.* **38**:5 (2006), art. id. S247.

[Warner 1965] S. L. Warner, "Randomized response: a survey technique for eliminating evasive answer bias", *J. Amer. Statist. Assoc.* **60**:309 (1965), 63–69. Zbl

[Warner 1971] S. L. Warner, "The linear randomized response model", *J. Amer. Statist. Assoc.* **66**:336 (1971), 884–888.

d_suarez@uncg.edu   *Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC, United States*

sngupta@uncg.edu   *Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC, United States*

involve

msp.org/involve

involve