African Journal of Pure & Applied Sciences

ISSN: 2710-0758
DOI 10.33886/ajpas.v3i1.255

Journal homepage: http://journal.ku.ac.ke/index.php/AJPAS

Research Article Open Access

A Bayesian Generalized Regression Approach for estimating parameters in the Item Sum Technique of sensitive surveys

Alakija T. O.[1*], Adeleke I. A.[2], Adamu M. O.[3], Adekeye K. S.[4]

[1*]Department of Statistics, Yaba College of Technology, Yaba, Lagos, Nigeria.
[2]Department of Actuarial Sciences, University of Lagos, Lagos, Nigeria.
[3,4]Department of Mathematics, University of Lagos, Lagos, Nigeria.

[1*]Corresponding Author's Email: lawaltope2003@yahoo.com

Received: 28th September 2021 / Accepted: 1st March 2022 / Published online: 1st June 2022

How to Cite:
Alakija, T., Adeleke, I., Adamu, M., & Adekeye, K. (2022). A Bayesian Generalized Regression Approach for estimating parameters in the Item Sum Technique of sensitive surveys. *African Journal of Pure and Applied Sciences*, *3*(1), 155-168.

ABSTRACT

Most respondents find it difficult to respond to sensitive questions in surveys especially providing information that require one to disclose behaviors that are possibly against social norms. This study involves solving the problem of non-response or falsification of information in a sample survey as a result of sensitive questions using the Item Sum Technique (IST). The IST is a recent indirect method for collecting data for continuous sensitive characteristics and it is a quantitative variant of the Item Count Technique (ICT). The calibration method has previously been used to estimate the IST, but the method is design-based and only useful for sufficiently large domains. There are situations where the domain sample size is not adequate to meet the needs of the precision of the estimates, hence the small-area estimation which is model-based may be needed. The Bayesian method is a model-based estimation method that tackles the problem of small area sample size. Results from simulation indicate the efficiency of the Bayesian Generalized Regression (BGREG) method for both small and large domain samples as compared with the Generalized Regression (GREG) and calibration method using three different criteria, AIC, BIC and MSE. A small sample survey on the use of substance abuse and internet gambling among university students from Nigeria was further used to show that the BGREG outperformed the GREG and Calibration methods using the AIC, BIC and MSE criterion.

Keywords: Bayesian Generalized Regression, Calibration method, Item Sum Technique, Large domain, Indirect method.

INTRODUCTION

The need for a better technique to handle the issue of sensitive questions in a survey has led to the development of various indirect questioning methods. The first ingenious indirect questioning method called the Randomized Response Technique (RRT) was introduced by Warner (1965). The RRT process deals with a simple principle of the question-and-answer method using a randomized device. The respondent obscures the meaning of an answer communicated to the interviewer using the randomized device which may be coins, dice, spinner, etc. The outcome of the device is hidden from the interviewer. The RRT designs create a probabilistic relationship between the observed answers and the unknown true scores. This is essential in extracting the truth from respondents and ensuring that their privacy is protected. The advantage of this approach is that it lessens the motivation of respondents to deliberately give a wrong report of their attitudes and it also allows for better cooperation from respondents. These methods, however, seem to be limited by various constraints which range from cost constraint to deficiencies in methodology, the RRT has also not improved nonresponse in a survey.

The ICT is another indirect questioning strategy that deals with qualitative sensitive variables. The ICT can also be called the list experiment (Miller, 1984), unmatched count technique (Dalton *et al.*, 1994), and the unmatched block design, or block total response (Raghavarao and Federer, 1979). The ICT was proposed by Miller (1984) which provides much secrecy and confidentiality on sensitive attributes. For the ICT, there are numbers of

statements related to the respondent's behavior called the non-sensitive items, and one statement of behavior related to the sensitive attribute which is called the sensitive item (Guo-Liang and Man-Lai, 2014). Potential limitation of the ICT is that information regarding the question of interest can only be gotten from the treatment group, while the control items included in the survey design only complicates the design and adds uncertainty its estimation. Moreover, the ICT use count data, which are limited in application (Kuha and Jackson, 2014).

The IST is recent and a variant of the ICT as it measures quantitative responses. It has been used by some researchers (Chaudhuri and Christofides, 2013; Zawar *et al.*, 2017; Perri *et al.*, 2017; Rueda *et al.,* 2017; Wolter and Herold 2018; Alakija *et al.*, 2019; Priyanka and Trisandhya, 2019; Trisandhya *et al.*, 2020). The IST has many advantages among which are; relatively low cognitive effort are required from respondents; a randomizing device will not be needed; it is easy to implement in both interviewer- and self-administered interviews (Trappmann *et al.*, 2014). The survey respondents for the IST are randomly divided into two independent groups (subsamples). Members of the first group are given a list containing non-sensitive characteristics (g_i for i = 1, 2,…, n) and one related to sensitive characteristics, while respondents from the second subsample are presented with a list containing only the non-sensitive items. These lists are quantitative in nature. Respondents in the two groups are then requested to report the total score applicable to them, without reporting the individual scores on each of the items. An unbiased estimator of the population means of a sensitive item, say, y from the IST data can be estimated by the mean difference of answers between the two random subsamples.

Methodologies are evolving for large sample domains, most recently is the use of calibration method of estimation by (Rueda *et al.*, 2017); and recently GREG method of estimation by Alakija *et al.* (2019). The calibration framework was introduced by Deville and Sarndal (1992) and Deville *et al.* (1993) to estimate population characteristics, while the GREG method was introduced by Alakija *et al.* (2019) to model sensitive response in the IST. The variance of the calibration method of estimation may be unacceptably large for certain domains, and moderately high for the GREG methods. Situations may however arise, such that, the size of the domain sample is too large to meet demands concerning the precision of the estimates, hence, there could be a need for a methodological advancement for small-area (model-based) estimation. Calibration is used in surveys to adjust for the weights assigned to members of a given sample to satisfy some assumed constraints. The calibration method was extended by introducing a GREG method of estimation for comparison of models and determine the reliability of the method. There have been no contributions regarding the IST with two sensitive variables, multiple innocuous variables, and a single auxiliary variable using the BGREG estimator to the best of our knowledge.

However, in this paper the GREG method of estimating sensitive response variable is extended to BREG, and the results compared with GREG and calibration methods, especially for small area domains. The BGREG is however the Bayesian approach to the generalized regression model as defined in this work.

The remaining parts of this paper is organized as the methodology section where the BGREG estimator for the IST was introduced. Next section is the empirical study which contains the simulation of both small and large samples for comparison of methods and to determine the efficiency of the BGREG. The results are also presented under this section for small sample area estimation using real-life data. Finally, the concluding remarks given in the conclusion section.

METHODOLOGY

Let S be a sensitive variable, g be a non-sensitive variable, z be the total score of the answers from the non-sensitive and sensitive questions from participants in the first subsample, and w be the reported total score of the non-sensitive questions reported by respondents in the second subsample. An unbiased estimator of the population means of the sensitive variable is then given by:

$$\bar{y} = \bar{z} - \bar{w} \tag{1}$$

such that $\bar{z}$ is the sample mean of the first subsample and $\bar{w}$ is the sample mean of the second group. The variance of (1) is given by

$$V(y) = \frac{\sigma_s^2}{n_1} + \frac{n\sum_{i=1}^{g}\sigma_{g_i}^2}{n_1 n_2} \tag{2}$$

where σ_s^2 is the population variance of the sensitive chracteristic and $\sigma_{g_i}^2$ is the population variance of the i^{th} non-sensitive variable.

According to Rueda *et al.* (2017), the mean of the sensitive characteristic for the IST using the Calibration type estimator can also be given by

$$\hat{\bar{y}}_c = \hat{\bar{z}}_c - \hat{\bar{w}}_c \tag{3}$$

Where

$\hat{\bar{z}}_c = \frac{1}{N}\sum_{i\in s_j} w_{i1} z_i = \hat{\bar{z}}_{HT} + \left(\bar{X} - \hat{\bar{X}}_c\right)^t \hat{B}_{s_1}$ is the calibration estimator of $\hat{\bar{Z}}_c$ obtained based on the LL sample s_1, with

$\hat{B}_{s_1} = F_{s_1}^{-1}\sum d_i q_i x_i z_i$ and

$$\hat{\bar{w}}_c = {}^{1}\!/_{N} \sum_{i\in S_L} \varphi_{i\in S_L} w_i \tag{4}$$

where φ is the weights based on the first and second subsamples.

The design-based or model-assisted-based are not appropriate in the small area estimations. This is as a result of small or even zero area sample sizes. The use of indirect estimation methods is therefore pertinent such that, information is taken from related areas by determining models based on survey data and the available auxiliary information (Sarndal, 2007). The calibration method has previously been investigated and it seems to perform well for large sample sizes (Rueda *et al.*, 2017). The BGREG method will subsequently be compared with the calibration method for small and large sample size estimation.

BGREG *Estimator in the IST*

The Generalized Regression (GREG) approach defines a linear regression model of the matrix form

$$y = X\beta + e, \tag{5}$$

where y is an $N'1$ matrix of dependent variables, X is $n'p$ matrix of p independent variables, b is $p'1$ matrix of unknown parameters, and e an n´1 matrix of assumed uncorrelated model errors.

The GREG estimator of the regression coefficient for the IST is then given as

$$\hat{\beta} = \left[\sum\sum_{j\in S} \frac{x_j x_j^{'}}{\sigma_2^2 \pi_{ij}}\right]^{-1} \sum_{j\in S} \frac{x_j (z_j - w_j)}{\sigma_1^2 \pi_j} \tag{6}$$

where $y_j = z_j - w_j$, π_j and π_{ij} are the first order inclusion probability, and the second-order inclusion probability defined as $\pi_j = \Pr(j \in S) = \sum_{S:j\in S} P(S)$ and $\pi_{ij} = \Pr(i, j \in S) = \sum_{S:i,j\in S} P(S)$ respectively; $\hat{\beta}$ is the design weighted estimator of β, x_j is the auxiliary variable for element j, σ_2^2 is the variance among clusters for all cluster i and σ_1^2 is the variance within clusters for all element j.

Lemma 1. The sample variance, s_2^2 is an unbiased estimator of σ_2^2.

Proof: To show the unbiased estimator of s_i^2.

Let

$$E\left(s_2^2\right) = E_1\left[E_2\left(s_2^2\right)\right]$$

$$= E_1\left[E_2\left[\frac{\sum_{i\in S}\sum_{j\in s_i}\left[\left(z_{ij} - w_{ij}\right) - \left(\bar{z}_i - \bar{w}_i\right)\right]^2}{(m-1)n}\right]\right]; i = 1,2,\ldots,n; j = 1,2,\ldots,m_i$$

$$= E_1\left[\frac{1}{n}\sum_{i\in S} E_2\left[\frac{\sum_{j\in s_i}\left[\left(z_{ij} - w_{ij}\right) - \left(\bar{z}_i - \bar{w}_i\right)\right]^2}{(m-1)}\right]\right]$$

$$= E_1\left[\frac{1}{n}\sum_{i\in S}\sigma_2^2\right] = \frac{1}{N}\sum_{i=1}^{N}\sigma_2^2$$

$$E\left(s_2^2\right) = \sigma_2^2$$

Lemma 2. The variance s_1^2 is a biased estimator σ_1^2 and it is an overestimate of σ_2^2.

Proof: To show the biasedness of s_1^2.
Let

$$E_2\left(\bar{y}_i^2\right) = E_2\left(\bar{z}_i - \bar{w}_i\right)^2 = \left(\bar{Z}_i - \bar{W}_i\right)^2 + \left(1 - f_2\right)\frac{\sigma_2^2}{M}$$

and also from the IST equations for within a cluster and among clusters, we have

$$E\left(\hat{\bar{Y}}^2\right) = \left[E_2\left(\hat{\bar{Z}} - \hat{\bar{W}}\right)\right]^2 + V_2\left(\hat{\bar{Z}} - \hat{\bar{W}}_2\right)$$

$$= \left[\sum_{i\in S}\frac{\bar{z}_i - \bar{w}_i}{n}\right]^2 + \frac{1 - f_2}{n^2}\sum_{i\in S}\frac{\sigma_i^2}{m}$$

Let

$$\bar{Y}_n = \sum_{i\in S}\frac{\bar{Z}_i - \bar{W}_i}{n}$$

Then

$$(n-1)E_2\left(S_1^2\right)=E_2\left[\sum_{i\in S}\left(\bar{Z}_i-\bar{W}_i\right)^2-n\left(\hat{\bar{Z}}_i-\hat{\bar{W}}_i\right)^2\right]$$

$$=\sum_{i\in S}\left(\bar{Z}_i-\bar{W}_i\right)^2+\frac{1-f_2}{m}\sum_{i\in S}\sigma_i^2-n\left(\bar{Z}_i-\bar{W}\right)^2-\frac{1-f_2}{nm}\sum_{i\in S}\sigma_i^2$$

$$=\sum_{i\in S}\left[\left(\bar{Z}_i-\bar{W}_i\right)^2-\left(\bar{Z}_n-\bar{W}_n\right)\right]^2+\frac{(1-n)(1-f_2)}{nm}\sum_{i\in S}\sigma_i^2$$

Multiply

both sides by $\frac{1-f_1}{n(n-1)}$ to derive

$$\frac{1-f_1}{n}E_2\left(S_1^2\right)=\left(\frac{1-f_1}{n}\right)\sum_{i\in S}\frac{\left[\left(Z_i-W_i\right)^2-\left(Z_n-W_n\right)\right]^2}{n-1}+\frac{(1-f_1)(1-f_2)}{n}\sum_{i\in S}\sigma_i^2$$

Applying E_1 to both sides of the equation gives

$$E\left(S_1^2\right)=\left(\frac{1-f_1}{n}\right)\sigma_1^2+\frac{(1-f_1)(1-f_2)}{nm}\sigma_2^2$$

And it is very easy to see that the expectation of S_1^2 will over estimate σ_1^2 by $\left(\frac{1-f_1}{m}\right)\sigma_2^2$ since $(1-f_1)(1-f_2)+f_1(1-f_2)=(1-f_2)$ hence, the proof is complete.

For Bayesian methods, priors can take any form. This is because the researcher's information before obtaining any data is usually reflected in the prior. It is however common to select groups of priors whose computation and interpretation are uncomplicated, such as natural conjugate priors. A conjugate prior distribution when combined with the likelihood, gives a posterior that is in a similar class of distributions as the prior. The prior information can therefore be interpreted in the same way as the likelihood function information. Hence, the prior can be construed as gotten from a fictitious data set using a similar procedure that brought about the actual data. For the GREG model, the prior distribution is elicited. The form of the likelihood function in the BGREG model suggests that the natural conjugate prior will be a Normal-Gamma distribution.

The BGREG model is a modification of the GREG model for regression estimation and the following theorem characterizes the variance of the precision given the sensitive questions.

Theorem 1. Given the model $y_i=\beta_1+\beta_2x_{i1}+\mathsf{L}\ +\beta_rx_{ir}+e_i$. If the model is estimated with BGREG estimator and has a prior Normal-Gamma distribution with parameters β, V, S^{-1} and v then, the variance of h given m is given by

$$Var(h\mid m)=\frac{2\bar{S}^{-1}}{\bar{v}}$$

Where

$\underline{Var}(h\mid m)$ is the variance of h given m, and $h=\sigma^{-2}$.

$v=n-1$

$m=y_0$ such that $y_0=(y_1,y_2,\ldots y_n)'$

Proof. Suppose we have data in a quantitative value, Z_i total score applicable to the sensitive and non-sensitive character for the i^{th} population unit under consideration and Y total score applicable to the sensitive character, for $i=1,\mathsf{K}\ ,N$. Since $Z=Y_i+W_i$

where

W_i- is the variable density representing the total score for the Λ non-sensitive questions

Y_i- is the variable density which represents the total score for the S sensitive questions

$Z-$ is the total score applicable to the non-sensitive & sensitive question.

Let X denote the auxiliary variable that can supply information about Y, then, the regression model is given as

$$y_i=\beta_1+\beta_2x_{i1}+\mathsf{L}\ +\beta_rx_{ir}+e_i \tag{7}$$

where $Y_i=Z_i-W_i$ but X is auxiliary information independent of variables Z and W.

As such

$$y_0=\begin{bmatrix}y_1\\y_2\\\mathsf{L}\\y_N\end{bmatrix};\quad \varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\mathsf{L}\\\varepsilon_N\end{bmatrix};\quad \beta=\begin{bmatrix}\beta_1\\\beta_2\\\mathsf{L}\\\beta_N\end{bmatrix}$$

and $N\times r$ matrix

$$X=\begin{bmatrix}1&x_{12}&\mathsf{L}&x_{1r}\\1&x_{22}&\mathsf{L}&x_{xr}\\\mathsf{M}&\mathsf{M}&\mathsf{O}&\mathsf{L}\\1&x_{N2}&\mathsf{L}&x_{Nr}\end{bmatrix}$$

Therefore, we write

$$y_0=X\beta+\varepsilon$$

The likelihood function is therefore normal with the assumptions that :

• All elements or unit of X are either fixed or are independent of ε with a probability density of $p(x/\phi)$ if they are random variables, where ϕ is a vector of parameters excluding the β and h.

$$\boxed{Note \;\; h = \frac{1}{\sigma^2}} \leftarrow \text{precision}$$

• Error term (ε) follows a multivariate Normal distribution with mean O_N and variance covariance matrix $\sigma^2 l_N$. That is

$$\varepsilon : \;\; N(O_N, h^{-1} l_N)$$

In matrix representation, the variance covariance matrix is given as

$$var(\varepsilon) = \begin{bmatrix} var(\varepsilon_1) & cov(\varepsilon_1, \varepsilon_2) & \mathsf{L} & cov(\varepsilon_1, \varepsilon_N) \\ cov(\varepsilon_1, \varepsilon_2) & var(\varepsilon_2) & \mathsf{L} & cov(\varepsilon_2, \varepsilon_N) \\ \mathsf{L} & \mathsf{L} & \mathsf{O} & \mathsf{L} \\ cov(\varepsilon_1, \varepsilon_N) & \mathsf{L} & \mathsf{L} & \mathsf{L} \end{bmatrix}$$

$$= \begin{bmatrix} h^{-1} & 0 & \mathsf{L} & 0 \\ 0 & h^{-1} & \mathsf{L} & 0 \\ \mathsf{L} & \mathsf{L} & \mathsf{O} & \mathsf{L} \\ 0 & 0 & \mathsf{L} & h^{-1} \end{bmatrix}$$

Thus: $var(\varepsilon_{ii}) = h^{-1}$ and $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$

Given by the first assumption of X, we can consider the conditional probability of Y_i on X_i and treat such as the likelihood function. i.e. $P(Y \mid X, \beta, h)$ as the likelihood function. But since X is independent and non-random, we drop X and have $P(y \mid \beta, h)$.

$$P(y_o \mid \beta, h) = \frac{h^{N/2}}{(2\pi)^{N/2}} \left\{ \exp\left[-\frac{h}{2} (y_o - X\beta)'(y_o - X\beta) \right] \right\} \quad (8)$$

recall that $y_i = z_i - w_i$, then;

$$P(y_o \mid \beta, h) = \frac{h^{N/2}}{(2\pi)^{N/2}} \left\{ \exp\left[-\frac{h}{2} (y_o - (Z - w)\beta)'(y_o - (Z - w)\beta) \right] \right\} \quad (9)$$

To prevent notational conflict, we replace the y_o as m

$$P(m \mid \beta, h) = \frac{h^{N/2}}{(2\pi)^{N/2}} \left\{ \exp\left[-\frac{h}{2} (m - (Z - w)\beta)'(m - (Z - w)\beta) \right] \right\} \quad (10)$$

From the likelihood pdf of equ (7), the estimate of the parameters given as

$$\hat{\beta} = (X'X)^{-1} X'm \qquad \boxed{v = N - r}$$

where $y_i = z_i - w_i$

$$\hat{\beta} = \left((Z_i - w_i)'(Z_i - w_i) \right)^{-1} (Z - w)'m \quad (11)$$

and

$$S^2 = \frac{(m - X\hat{\beta})'(m - X\hat{\beta})}{N - r} \quad (12)$$

Following the normal formular; we have that,

$$P[x \mid \mu, \sigma^2] = \frac{1}{(2\pi)^{N/2}} \sigma^{-N} \left\{ \exp\left[-\frac{\sigma^{-2}}{2} (x - \mu)^2 \right] \right\}$$

Separating the parameter, we have:

$$= \frac{1}{(2\pi)^{N/2}} \left[\left\{ \sigma^{-1} \exp\left[-\frac{\sigma^{-2}}{2} (x - \mu)^2 \right] \right\} \left\{ \sigma^{-N+1} \exp\left[-\frac{\sigma^2}{2} \right] \right\} \right]$$

Now from equ (7), separating the parameters, we have:

$$P(m \mid \beta, h) = \frac{1}{(2\pi)^{N/2}} \left\{ h^{1/2} \exp\left[-\frac{h}{2} (\beta - \hat{\beta})'(Z - w)'(Z - w)(\beta - \hat{\beta}) \right] \right\} \left\{ h^{v/2} \exp\left[-\frac{hv}{2S^{-2}} \right] \right\} \quad (13)$$

Inversely,

$$P(m \mid \beta, h) = \frac{1}{(X\pi)^{N/2}} \left\{ h^{1/2} \exp\left[-\frac{h}{2} (\beta - \hat{\beta})' X'X (\beta - \hat{\beta}) \right] \right\} \left\{ h^{v/2} \exp\left[-\frac{hv}{2S^{-2}} \right] \right\} \quad (14)$$

From equ (10) & (11); if we elicit a prior for β conditional on h, form:

$\beta \mid h : \;\; N(\underline{\beta}; h^{-1} V)$

and a prior for h of the form

$h : \;\; G(\underline{S}^{-2}, \underline{v})$

That is, the prior follows a Normal-Gamma Distribution with parameters: $\underline{\beta}, \underline{V}, \underline{S}^{-1}$ & $\underline{v}$:

Notationally:

$\beta, h : \;\; NG(\underline{\beta}, \underline{V}, \underline{S}^{-1}, \underline{v})$

where

$\underline{\beta}$ is now a r-vector containing the prior means for the r-regression coefficients, $\beta_1, \mathsf{K}, \beta_r$

$\underline{V}$ is now a $r \times r$ positive definite prior Variance Covariance matrix.

Thus, the notation for the prior density is

$$P(\beta, h) = fNG(\beta, h \mid \underline{\beta}, \underline{V}, \underline{S}^{-1}, \underline{v})$$

Therefore, the posterior is proportional to the product of the likelihood function and the prior; such that:

$$P(\beta, h \mid m) \alpha P(m \mid \beta, h) \cdot P(\beta, h)$$

we have,

$$\beta, h \mid m: \quad NG(\overline{\beta}, \overline{V}, \overline{S}^{-1}, \overline{v})$$

Where,

$$\overline{V} = (\underline{V}^{-1} + X'X)^{-1}$$

$$\overline{\beta} = \overline{V}(\underline{v}^{-1}\underline{\beta} + X'X\hat{\beta})$$

$$\overline{v} = \underline{v} + N$$

and $\overline{S}^{-2}$ is implicitly defined as :

$$\overline{v}s^2 = \underline{v}\underline{s}^2 + (\hat{\beta} - \underline{\beta})'[\underline{V} + (X'X)^{-1}]^{-1}(\hat{\beta} - \underline{\beta})$$

The $\beta, h \mid m$ is the joint posterior distribution. The marginal posterior for β when h is integrated out is given as

$$\beta \mid m: \quad t(\overline{\beta}, \overline{S}^2\overline{V}, \overline{v})$$

it the follows that from the t-distribution,

$$E(\beta \mid m) = \overline{\beta} \quad \mathsf{L} \quad Unbiasdness$$

and

$$Var(\beta \mid m) = \frac{\overline{v}S^2}{\overline{v} - 2}\overline{V}$$

Similarly, the properties of the Normal-Gamma distribution implies that:

$h \mid m: \quad G(\overline{S}^{-1}, \overline{v})$ hence, $E(h \mid m) = \overline{S}^{-2}$ and we have the variance as $Var(h \mid m) = \frac{2\overline{S}^{-1}}{\overline{v}}$.

The likelihood is normal and if it is not normal it can be normalized by a simple transformation formula. The posterior is also normal-gamma. In summary, the prior is normal-gamma, with a normal likelihood and the posterior is also normal-gamma. The values of the parameters will determine the shape of the posterior distribution. The posterior distribution (normal-gamma) can be positively skewed, negatively skewed or symmetric depending on the parameter values.

Empirical Study

Simulation Study

In order to investigate the performance of the proposed modified BGREG and to compare with the existing GREG and calibration models for large domains, we carried out a simulation study. In so doing, artificial observations are generated using the MCMC methods for fitting realistically complex modes.

The computational algorithm for estimating the parameters of the models is outlined as follows:

Step 1. Set seed (1235)

Step 2. If n = 10, then

Step 3. Fit gamma distribution for y_1

Step 4. Fit gamma distribution for y_2

Step 5. Generate normal variates y_1 and y_2

Step 6. Generate gamma variates y_1 and y_2

Step 7. Generate normal variate for x

Step 8. Fit generalized model for y_1 and y_2 in step 5

Step 9. Fit generalized model for y_1 and y_2 in step 6

Step 10. Fit a bayes generalized linear model for y_1 and y_2

Step 11. Fit linear model for calibration for y_1 and y_2

Step 12. Summarize models 1, 2, 3, 4, 5 and 6

Step 13. Find the Akaike Information Criterion (AIC) for model 1 to 6

Step 14. If $AIC_i < AIC_j$, then

Step 15. Select $AIC_i \; \forall \; i \; ^1 \; j$

Step 16. Find the Bayesian Information Criterion (BIC) for model 1 to 6

Step 17. If $BIC_i < BIC_j$, then

Step 18. Select BICi $\forall$ i ^{1} j

Step 19. Find the Mean Square Error (MSE) for model 1 to 6

Step 20: If $MSE_i < MSE_j$, then

Step 21. Select MSEi $\forall$ i ^{1} j

Step 22. Print AIC, BIS and MSE

Step 23. Else, if n = 20 then

Step 24. Repeat steps 1 to 22

Step 25. Do for n = 25, 50, 100, 250, 500, 750, 1000.

We assume that the two sensitive questions and innocuous questions are from a bivariate normal gamma distribution with μ = (10.368, 7.02, 4.439) and s = (19.121, 5.470,1.889) respectively. The two sensitive variables are not normally distributed, they are assumed to follow the gamma distribution. However, the innocuous variables are normally distributed. The BGREG is a mixture of both gamma and normal distributions, thereby, resulting to bivariate normal-gamma distribution.

The result of the simulated data is given in the tables below.

Table 1: Parameter Estimation from Simulation Study for n = 10, 20, 25.

	Model	Parameter		Standard Error		P-value				
n=10	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	1.605	-0.041	0.0179	0.0029	0.0000	0.0000	59.964	62.056	18.030
	GREG	-44.834	8.542	4.512	0.756	0.0000	0.0000	63.948	64.855	19.238
	CALIB	-1.292	5.929	2.645	1.289	0.6384	0.0018	223.089	75.535	66.755
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	0.754	-0.034	0.023	0.004	0.0000	0.0000	19.205	18.297	5.164
	GREG	-7.162	2.415	2.549	0.427	0.0000	0.0000	52.528	53.436	6.141
	CALIB	5.495	2.841	0.700	0.341	0.0000	0.0000	157.927	168.951	7.677
		Parameter		**Standard Error**		**P-value**				
n=20	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	0.635	-0.083	0.078	0.011	0.0000	0.0000	115.030	118.021	24.314
	GREG	-21.58	7.025	3.231	0.619	0.0000	0.0000	130.390	133.373	29.412
	CALIB	20.825	6.096	0.948	0.459	0.0000	0.0000	223.089	225.412	109.834
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	0.606	-0.075	0.034	0.005	0.0000	0.0000	55.418	58.406	19.653
	GREG	-8.266	3.149	1.227	0.235	0.0000	0.0000	91.65	94.637	42.240
	CALIB	8.439	1.548	0.275	0.133	0.0000	0.0000	157.927	161.944	17.790
		Parameter		**Standard Error**		**P-value**				
n=25	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	3.659	-0.347	0.164	0.027	0.0000	0.0000	61.873	64.216	23.273
	GREG	-21.580	7.025	3.231	0.619	0.0000	0.0000	130.390	133.373	23.529
	CALIB	10.973	9.865	1.025	0.702	0.0000	0.0000	223.089	226.232	416.744
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	1.657	-0.171	0.089	0.014	0.0000	0.0000	2.733	5.923	0.369
	GREG	-9.176	3.139	1.317	0.266	0.0000	0.0000	107.990	111.647	3.462
	CALIB	7.088	3.693	0.272	0.186	0.0000	0.0000	157.927	160.847	6.049

Table 1 shows that BGREG is the best model for variables Y_1 and Y_2 using AIC, BIC and the MSE criteria for n = 10, 20 and 25. Hence, the BGREG outperformed the GREG and Calibration models for small sample size estimation. All the parameter estimates are significant ($p < 0.05$).

Table 2: Parameter Estimation from Simulation Study for n = 50, 100, 250.

	Model	Parameter		Standard Error		P-value				
n=50	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	3.6586	-0.3469	0.1636	0.0273	0.0000	0.0000	58.2160	61.873	0.194
	GREG	0.2733	2.8826	0.6024	0.1293	0.0000	0.0000	161.234	188.116	10.842
	CALIB	12.0443	13.2239	0.7866	0.5294	0.0000	0.0000	223.0S9	323.187	453.7216
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	0.3972	-0.05036	0.161	0.04236	0.0000	0.0000	125.46	131.199	1.501
	GREG	2.5176	19.8570	0.1732	0.06236	0.0000	0.0000	135.16	142.998	1.512
	CALIB	7.0739	3.7454	0.182 4	0.0958	0.0000	0.0000	157.927	152.247	47.804
		Parameter		**Standard Error**		**P-value**				
n=100	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	0.31291	0.0334	0.0214	0.0025	0.0000	0.0000	190.960	198.774	0.161
	GREG	3.1958	29.9401	0.0636	0.0633	0.0000	0.0000	195.906	204.742	0.196
	CALIB	8.1710	9.5515	0.2466	0.1322	0.0000	0.0000	223.089	476.0872	369.520
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	0.34685	-0.03548	0.1121	0.0181	0.0000	0.0000	130.280	138.100	1.501
	GREG	3.1958	29.9401	0.7290	0.0500	0.0000	0.0000	210.056	218.651	0.177
	CALIB	6.60994	2.71774	0.7575	0.0406	0.0000	0.0000	157.927	240.047	49.723
		Parameter		**Standard Error**		**P-value**				
n=250	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	3.1518	-0.2572	0.0310	0.0052	0.0000	0.0000	155.770	245.208	2.644
	GREG	0.3173	3.8883	0.2269	0.0349	0.0000	0.0000	171.347	299.729	2.909
	CALIB	10.6402	10.6809	0.3252	0.06691	0.0000	0.0000	223.089	1065.272	494.5805
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	1.27111	-0.10640	0.05165	0.0140	0.0000	0.0000	116.390	125.829	0.617
	GREG	0.7867	9.3985	0.0628	0.0221	0.0000	0.0000	128.029	138.412	0.678
	CALIB	7.66703	2.95262	0.08663	0.0695	0.0000	0.0000	157.927	450.911	66.697

Table 2 shows that BGREG is the best model for variables Y_1 and Y_2 using AIC, BIC and the MSE criteria for n = 50 and 250. While GREG is the best using the MSE criteria for sample size n = 100. However, the BGREG outperformed the GREG and Calibration models for moderately large samples. All the parameter estimates are significant ($p < 0.05$). Note that the smaller the value of the criterion, the better the performance of the model.

Table 3: Parameter Estimation from Simulation Study for n = 500, 750, 1000.

	Model	Parameter		Standard Error		P-value				
n=500	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	2.98276	-0.2370	0.01889	0.003 I89	0.0000	0.0000	1572.100	1559.47	2.319
	GREG	0.3353	4.2194	0.0208	0.0035	0.0000	0.0000	1729.310	1715.417	2.550
	CALIB	11.1772	10.1962	0.0573	0.0310	0.0000	0.0000	1623.089	1684.151	460.406
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	0.3972	-0.05036	0.04061	0.0023	0.0000	0.0000	125.46	131.199	1.5012
	CALIB	7.0739	3.7454	0.142 4	0.0958	0.0000	0.0000	1 57.927	152.2469	47.804
						0.0000	0.0000			
		Parameter		**Standard Error**		**P-value**				
n=750	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	3.1533	-0.2574	0.0165	0.0025	0.0000	0.0000	2220.400	2206.560	2.507
	GREG	0.3171	3.8850	0.0182	0.0028	0.0000	0.0000	2442.440	2427.216	2.758
	CALIB	10.4583	10.2184	0.0476	0.0262	0.0000	0.0000	2223.0S9	2544.587	434.931
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	1.4080	-0.1243	0.0197	0.0017	0.0000	0.0000	399.590	386.942	0.4561
	GREG	0.7102	8.0451	0.0118	0.0018	0.0000	0.0000	439.549	425.636	0.501
	CALIB	6.8139	2.9841	0.0153	0.0084	0.0000	0.0000	457.927	846.513	54.724
		Parameter		**Standard Error**		**P-value**				
n=1000	**Variable Y_1**	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	3.10321	-0.25076	0.01365	0.00214	0.0000	0.0000	303.600	302.850	2.427
	GREG	0.3222	3.9879	0.0150	0.0024	0.0000	0.0000	333.960	333.135	2.669
	CALIB	10.9573	9.8678	0.03625	0.01939	0.0000	0.0000	323.081	386.942	442.621
	Variable Y_2	**b_0**	**b_1**	**b_0**	**b_1**	**b_0**	**b_1**	**AIC**	**BIC**	**MSE**
	BGREG	1.4081	-0.1243	0.0187	0.0019	0.0000	0.0000	399.590	386.942	0.587
	GREG	0.7102	8.0438	0.0198	0.0028	0.0000	0.0000	439.549	425.636	0.646
	CALIB	7.0448	2.9343	0.0895	0.0051	0.0000	0.0000	457.927	465.962	58.312

Table 3 shows that BGREG is the best model for variables Y_1 and Y_2 using AIC, BIC and the MSE criteria for n = 500, 750 and 1000. Hence, the BGREG outperformed the GREG and Calibration models for large sample domain estimation. All the parameter estimates are significant ($p < 0.05$).

Fig. 1 and Fig. 2 below shows the asymptotic and consistency properties of standard error of the estimates for both the slope and intercept of the regression model. The standard errors approaches zero as the sample size increases.

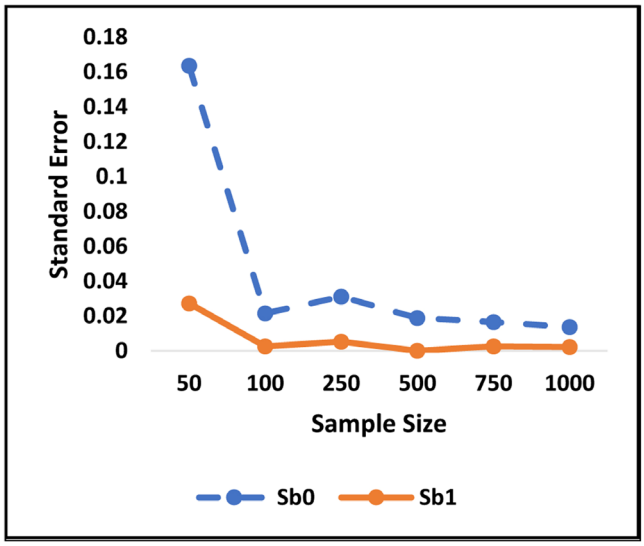

Figure 1: y_1 BGREG Estimate Consistency Property

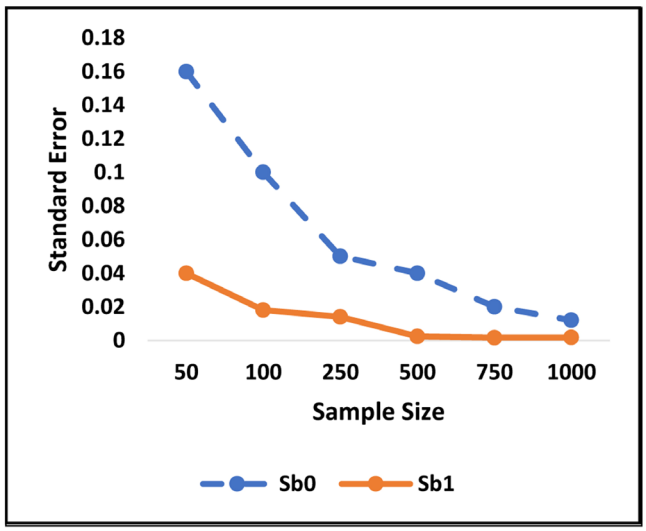

Figure 2: y_2 BGREG Estimate Consistency Property

RESULTS

A study was carried out on 75 sampled respondents who are students from tertiary institutions in Nigeria using indirect questioning IST, consisting of two sensitive and two innocuous questions. The two sensitive questions include the quantity of substance abused by students and the amount spent on internet gambling. These are the variables of interest. The sampling design used is the two-stage cluster sampling. Two independent samples of such were carried out. The first sample (S_1) consists of only innocuous variables without any sensitive variable, the second sample (S_2) consist of the innocuous variables and two sensitive variables. The first innocuous question (variable) is how many hours on average in the last 30 days did you spend reading books for leisure? and the second innocuous question is, during the last 24 hours, how many hours did you spend watching television? Exploratory Data Analysis (EDA), Estimation of the parameters using the BGREG, GREG and Calibration.

Exploratory Data Analysis

The exploratory data analysis was carried out to show some hidden features of the data set. An Empirical test for normality called the Shapiro-Wilk normality test was carried out to see if the variables measured are Gaussian or not. Let Y_1 be the quantity of substance abuse, Y_2 be the amount spent on internet gambling among students and X be the auxiliary variable. Then the Summary Statistics and Shapiro- Wilk Normality Test are given in table 4 and 5 respectively.

Table 4 shows that the variables y_1 and y_2 are not normally distributed, while the auxiliary variable x is approximately Gaussian.

Table 4: Summary Statistic from Sample Survey.

Variable	K	Q_1	Median	Mean	Q_3	Max	Skewness	Kurtosis
Y_1	1.200	2.200	3.400	10.370	7.000	88.500	3.140	12.563
Y_2	1.500	3.333	5.833	7.020	9.000	22.333	1.421	4.282
X_1	1.000	3.000	4.667	4.439	6.000	7.667	-0.086	2.167

The table is reprinted from Alakija *et al.* (2019)

Table 5: Shapiro-Wilk Normality Test from sample survey data.

Variable	W	P-value
Y_1	0.49904	0.0000000366
Y_2	0.93237	0.0008299
X_1	0.96615	0.5497

The table is reprinted from Alakija *et al.* (2019).

Table 5 shows that the variables y_1 and y_2 are not normally distributed while the auxiliary variable x is Gaussian, since its p-value is less than and greater than 0.05 respectively. Note that all tests are carried out at 5% level of significance.

The varaibles y1 and y2 are normalized using the transformation in equation (15)

$$y_i^* = \left| \frac{y_i - \bar{y}}{y_{\max} - y_{\min}} \right| \quad (15)$$

where yi* is the transformed variable, which is normally distributed, yi is the original variable, which is not normally distributed, ymax and ymin are the maximum and minimum values of y respectively. The numerator is the deviation of each value of y from the mean, while the denominator is the range.

Thus, the values of y1 and y2 are transformed. It was the transformed data that was used for all the subsequent analysis.

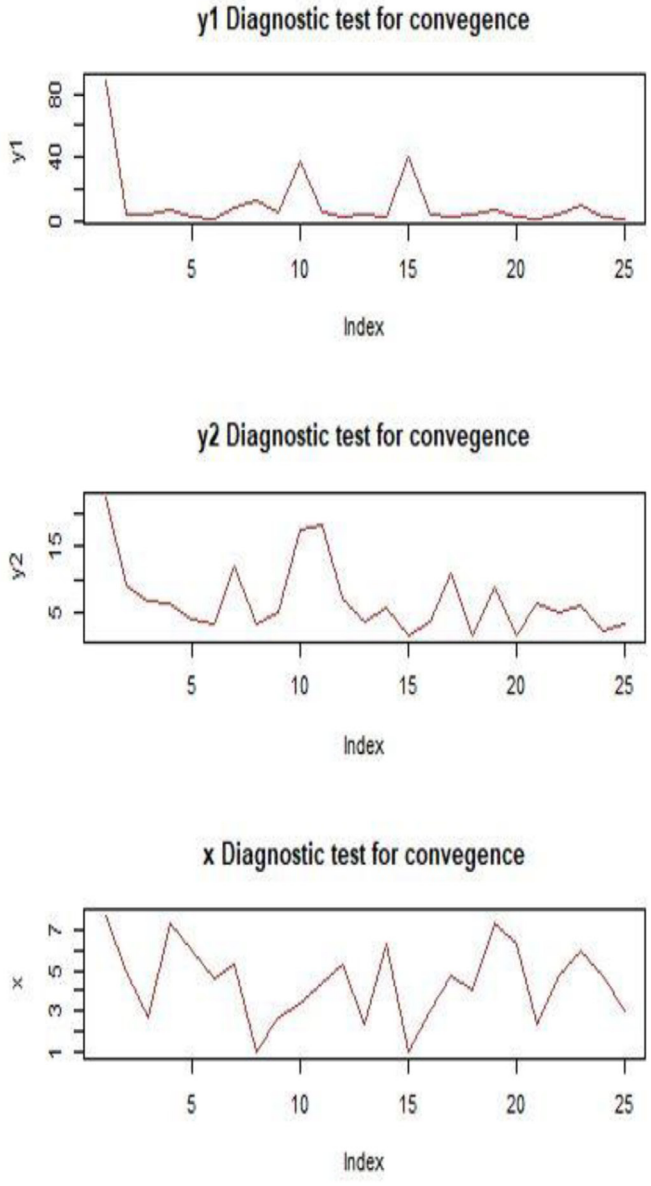

Figure 3: Convergence test

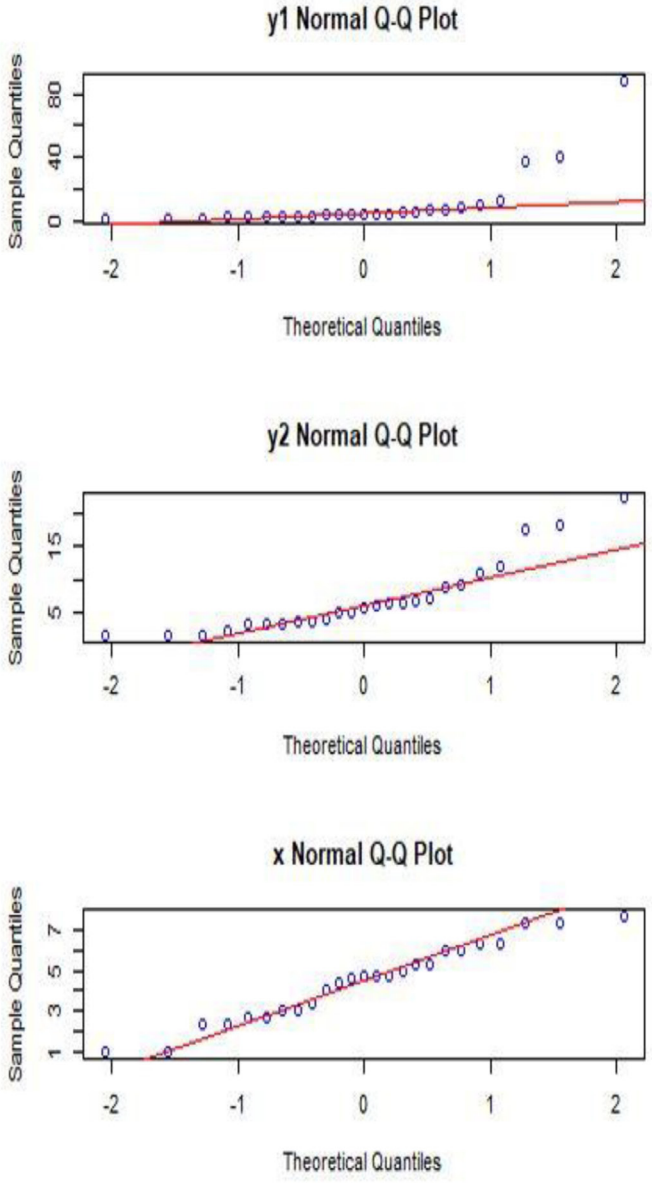

Figure 4: Q-Q Plot for Normality Test

Fig. 3 and Fig. 4 are the diagnostic test for convergence and normal Q-Q plots of y_1, y_2 and x respectively and they depict that the variables are not normal but approximately normal, since all the points are not on the Q-Q line looking at Fig 4.

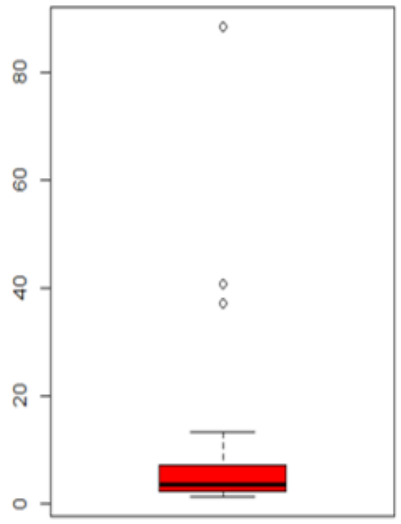

Figure 5a: Box plot y_1

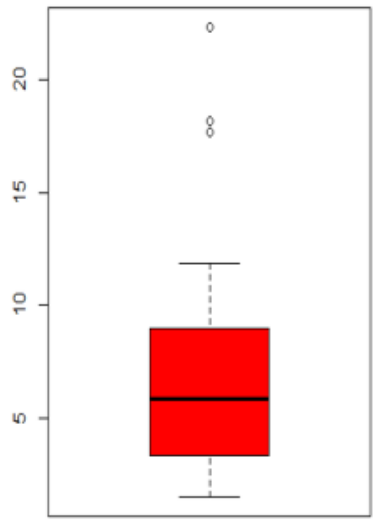

Figure 5b: Box plot y_2

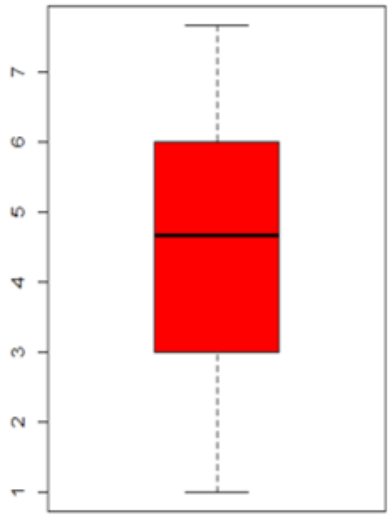

Figure 5c: Box plot x e Q-Q line

Fig. 5a, Fig. 5b and Fig. 5c are the box plots of y_1, y_2 and x respectively. Fig.5a and Fig.5b depicts that the variables y_1 and y_2 have outliers while x does not have outliers.

Estimation of Model Parameters

In this study, three models were fitted using the BGREG, GREG and calibration estimators for comparison and to determine the best of the three models. The difference between calibration and BGREG is that the calibration estimator assumes that y_1 and y_2 are normally distributed, while the BGREG assumes that y_1 and y_2 need not be necessarily normally distributed. So, with the posterior distribution of normal-gamma, the parameters are varied to be distributed as normal or as gamma distribution.

Table 6: Parameter estimation from sample survey.

Model	**Variable**	**Parameter**		**Standard Error**		**P-value**		**AIC**	**BIC**	**MSE**
	Y_1	b_0	b_1	b_0	b_1	b_0	b_1			
Model 1	BGREG	0.153	-0.011	0.098	0.018	0.032	0.031	171.07	174.73	326.48
Model 2	GREG	4.881	1.236	0.172	2.094	0.033	0.041	222.98	225.646	344.64
Model 3	CALIB	10.368	1.236	0.877	2.094	0.044	0.046	223.09	226.75	345.74
Model	**Variable**	**Parameter**		**Standard Error**		**P-value**		**AIC**	**BIC**	**MSE**
	Y_2	b_0	b_1	b_0	b_1	b_0	b_1			
Model 1	BGREG	0.240	-0.020	0.006	0.062	0.001	0.001	143.92	147.57	22.03
Model 2	GREG	2.721	0.968	0.053	0.019	0.021	0.011	156.83	159.58	24.41
Model 3	CALIB	7.020	0.968	0.153	0.044	0.035	0.023	157.93	160.48	25.51

Table 6 shows that the BGREG performed better than the GREG and Calibration using the AIC, BIC and MSE criteria. The model that minimizes the three criterion should be preferred.

CONCLUSION

In assessing the prevalence of a sensitive attribute in survey and to combat respondent underreporting of stigmatizing attributes, investigators have tried various techniques. The IST is a recently used technique and it is a quantitative response alternative of the ICT. The IST method can be used in large scale surveys where the means of getting information is via a structured questionnaire. The use of sensitive questions in surveys could result into non-response or falsified information such that; it will require a very crucial method since such questions could have a significant effect on the analysis of real-life situation. However, due to the increase in demand for reliable and efficient small area statistics, there had been a lot of focus and attention on the methods of estimation for small area surveys. This study broadens the IST by proposing the BGREG method for its finite population estimation. The study further compared the BGREG to the current existing GREG and calibration methods to determine its efficiency. The robustness of the methods was further evaluated using simulation study. The simulation study indicates that the parameter estimates are consistent and significant while the standard errors approach zero as the sample size increases. The performance of the estimators used depends on the distribution of the data. It was observed that the GREG and calibration methods are good methods of estimation, but can only be used for large domains, but the BGREG method of estimation fits well for both large and small area statistics depending on the posterior distribution.

REFERENCES

Alakija, T. O., Adeleke, I. A., Adekeye, K. S., and Adamu, M. O. (2019). A Generalized Regression Estimation of the Item Sum Technique in Sensitive Surveys. *International Journal of Mathematical Analysis and Optimization: Theory and Applications*, *2019*(1), 512-520.

Chaudhuri, A., and Christofides, T. C. (2013). *Indirect questioning in sample surveys.* Springer Verlag, Berlin, Heidelberg, DE.

Dalton, D. R., Wimbush, J. C., and Daily, C. M. (1994). 'Using the unmatched count technique (uct) to estimate the base rates for sensitive behavior'. *Personnel Psychology*, *47,* 817–828.

Deville, J. C., and Sarndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of American Statistical Association. 87*, 376–382.

Deville, J. C., Sarndal, C. E., and Sautory, O. (1992). Generalized raking procedures in survey sampling. *Journal of American Statistical Association*, *88*, 1013–1020.

Guo-Liang, T., and Man-Lai, T. (2014). *Incomplete categorical data design. Non-randomized response techniques for sensitive questions in surveys.* Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences Series. New York, USA.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods.* Springer Science & Business Media.

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of American Statistical Association*, *47*, 663–685.

Kuha, J., and Jackson, J. (2014). The item count method for sensitive survey questions: modelling criminal behavior. *Journal of the Royal Statistical Society, Series C, Applied Statistics, 63*(2). DOI:10.2139/ssrn.2119238.

Miller, J. D. (1984). *A new survey technique for studying deviant behavior*. Ph.D. thesis. The George Washington University, Washington, DC.

Perri, P. F., Rueda, M. M. G., and Cobo, B. R. (2017). Multiple sensitive estimation and optimal sample size allocation in the item sum technique. *Biometrical Journal*, 1–19.

Priyanka, K., and Trisandhya, P. (2019). Item sum techniques for quantitative sensitive estimation on successive occasions. *Communication for statistical applications and methods*, *2019*(26), 175–189. https://doi.org/10.29220/CSAM.2019.26.2.175.

Raghavarao, D., and Federer, W. T. (1979). Block total response as an alternative to the randomized response method in surveys. *Journal of the Royal Statistical Society,* B *41*(1), 40–45.

Rueda, M. M. G., Perri, P. F., and Cobo, B. R. (2017). Advances in estimation by the item sum technique using auxiliary information in complex surveys. *Springer-Verlag Germany, part of Springer Nature. ASTA Advances in Statistical Analysis, Springer; German Statistical Society*, *102*(3), 455–478.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology, 33*(2), 99–119.

Trappmann, M., Krumpal, I., Kirchner, A., and Jann, B. (2014). Item sum: a new technique for asking quantitative sensitive questions. *Journal of Survey Statistics and Methodology, 2*(1), 58–77.

Trisandhya, P. (2020). Application of item sum technique for estimating quantitative sensitive mean on successive moves using auxiliary information. *Communications in Statistics - Theory and Methods*. DOI:10.1080/03610918.2020.1722839. [Print ISSN 0361-0918, Online ISSN 1532-4141]

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of American Statistical Association, 60*, 63–69.

Wolter, F., and Herold, L. (2018). Sensitive questions

in surveys: Testing the Item Sum Technique (IST) to tackle Social Desirability Bias. *Sage Research Methods.* DOI**:** https://dx.doi.org/,10.4135/9781526441928.

Zawar, H., and Naila, S. (2017). On some new randomized item sum techniques. *Communications in Statistics-Theory and Methods*, *1*(1), 1–14.