Improved Logit Estimation through Mangat Randomized Response Model

Zawar Hussain*

Sidra Anjum**

Javid Shabbir***

*Department of Statistics, Quaid-i-Azam University
45320, Islamabad 44000, Pakistan.
E-mail: zhlangah@yahoo.com,* Sidra_anjum86@yahoo.com,** jsqau@yahoo.com***

Abstract

Using the Mangat (1994) randomized response model improved logit estimation is proposed and it is compared with the Hussain and Shabbir (2008) logit estimation at equal level of privacy protection. The proposed logit estimation has also been compared with ordinary logit estimation. The proposed estimation is observed to be better than Hussain and Shabbir (2008) logit estimation. The case of missing observations has also been dealt with through EM algorithm.

Key Words: logit estimation, randomized response model, privacy protection, sensitive binary variable, missing observations and EM algorithm.

1 Introduction

In an analysis using regression models, we generally assume that the data at hand have been generated by a mathematical model (function) whose parameters are unknown, and we must estimate those underlying parameters. Ordinary regression is used to forecast a continuous outcome. Logistic Regression is used when the outcome (regressand) is dichotomous: yes/no. Logistic Regression has been used by many authors in different fields of Psychology like Psychopathology (Clark and Beck (1991)), Pediatric Psychology (Freidrich et al. (1986)), Clinical Psychology (Waldman and Rhee (2002)), Cognitive Psychology (Bradley (1988)), Community Psychology (Hedeker et al. (1994)), Exercise and Sports studies (Capel (1986)), Human Genetics (Waldman et al. (1999)) and many others. In such studies, certainly, the outcome variables may be more likely to be sensitive, e.g. sexual abuse, addiction to Drugs, committing a fraud, Shoplifting, tax evasion, induced abortion, illegal Sex, etc. In summary, Logistic Regression provides a powerful tool for improved understanding of how the independent factors can be used accurately to assess likelihood of occurrence of a category of a dichotomous variable.

In Psychological or Behavioral studies dichotomies are attention-grabbing. Some are so appealing that we go to a great length to cram them. Did the person use a drug? Did the voter vote for Party A? Did the player use his best strategy in a contest? Did a War break out? Had a woman induced abortion? Did a person pay full tax? Had a person involved in fraud? Had a person ever visited adult website? etc. In studies like asking about the preference of ice cream flavors, books, late-night hosts, cooking oil, fashion, etc. we have a little or no reason to suspect the truthfulness of the responses and it is assumed that what we have observed is what had happened in reality. But when a question of sensitive nature is posed to the respondents they may disapprove their affirmative actions about these sensitive traits. For instance, asking about visiting adult website, committing a fraud, using marijuana, may result in an evasive answer because affirmative response on visiting adult website is simple embarrassing or socially unacceptable, affirmative action on tax evasion is illegal, and so on. In these situations expecting an honest response may be little bit optimistic.

For self-reported data, assumption of true responses on sensitive issues may be questionable because the respondents have incentives in reporting less than truthful answers. Moreover, it may happen in a sensitive study that a respondent refuses to respond at all. Thus procurement of trustworthy data in sensitive psychological surveys is an important issue.To increase the response rate and lessen the evasive answer bias in the estimators, Warner (1965), for the first time, proposed an ingenious survey method, called Randomized Response Technique (*RRT*). Warner's model consists of two complimentary questions A and A^c to be answered on probability basis, where A is "do you possess the sensitive trait", and A^c is "do you not possess the sensitive trait". The two questions A and A^c are offered to respondents with preset probabilities P and $1-P$ respectively. A big amount of developments and variants of Warner's RRT have been suggested by a number of researchers. Greenberg et al. (1969), Mangat and Singh (1990), Mangat (1994), Singh et al. (1998), Christofides (2003), Kim and Warde (2004), Hussain and Shabbir (2007, 2010) are some of the many to be cited.

For a comprehensive note the attracted readers may be referred to Chaudhuri and Mukerjee (1998) and Tracy and Mangat (1996). Using the Warner (1965) *RRT*, Hussain and Shabbir (2008) proposed a hidden logit estimation method and established that hidden logit estimation through Warner (1965) *RRT* give the estimates closer to the results what could had been observed through the true data, but using *RRT* resulted in decreased precision of the estimates. In this article, we intend to use Mangat (1994) *RRT* in logit estimation due to its simplicity and increased precision compared to Warner (1965) *RRT* with and without taking care of privacy protection to the respondents. The case when some observations are lost due to any reason is also the subject of our study. The paper is organized as: Section 2 describes Hussain and Shabbir (2008) hidden logit estimation, briefly. In Section 3, we present the derivation of hidden logit estimation through Mangat (1994) *RRT*. Section 4 is comprised of a comparative study followed by a hidden logit estimation using EM algorithm in Section 5. Finally Section 6 concludes the findings of the study.

2 Hussain and Shabbir hidden logit estimation

Hussain and Shabbir (2008) used Warner (1965) *RRT* to estimate the hidden logits. As discussed earlier, Warner's model consists of two complimentary questions A and A^c to be answered on probability basis, where A is "do you possess the sensitive trait", and A^c is "do you not possess the sensitive trait". The two questions A and A^c are offered to respondents with preset probabilities P and $1-P$ respectively. Using simple random sampling without replacement (SRSWOR) sampling, the i^{th} selected respondent is asked to select a question (A or A^c) and report *yes* if his/her actual status matches with selected question and *no* otherwise. Suppose the population proportion of individuals with sensitive trait is π. Then the probability of a *yes* response for a particular respondent is then given by

$$P_r(yes) = \theta = P\pi + (1-P)(1-\pi), \tag{1}$$

where P is the probability of selecting question A.
From (1) it can be written that

$$\pi = \frac{\theta-(1-P)}{2P-1}. \tag{2}$$

As the interest of their study was to model the true probability of an event of interest (using marijuana) as a function of a set of independent psychological, social, or economic factors such that $\pi = f\left(\underline{X}, \underline{\beta}\right)$ and the true probability π is unobservable, they assumed logistics regression function for θ and modeled the log-odds ratio of the θ as

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \underline{X}_i'\underline{\beta}, \tag{3}$$

where $\theta_i = \dfrac{Pe^{\underline{X}_i\underline{\beta}}+(1-P)}{1+e^{\underline{X}_i\underline{\beta}}}$, $\underline{X}_i$ is the vector of i^{th} observation on independent factors and $\underline{\beta}$ is the unknown vector of parameters.

For $P=1$, (3) becomes an ordinary logistic regression function derived from direct responses. They established that maximum likelihood estimator of β cannot be derived because setting derivative of log-likelihood function i.e. $\frac{\partial \ell}{\partial \underline{\beta}} = 0$ (ℓ is log-likelihood function of the observed data), cannot be solved analytically.

Therefore, they obtained the numerical estimates of $\underline{\beta}$.

3 Proposed Hidden Logit estimation

Consider the *RRT* proposed by Mangat (1994) which requires a respondent to say *yes* if he/she possesses the sensitive trait and to use Warner (1965) *RRT* with probability of a sensitive question as P_1, otherwise. The probability of a *yes* response from a respondent is now given by

$$P(yes) = \lambda = \pi + (1-\pi)(1-P_1). \tag{4}$$

Solving (4) for π we get

$$\pi = \frac{\lambda-(1-P_1)}{P_1}. \tag{5}$$

In Case of ordinary logit estimation we could have model π_i as

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = e^{\underline{X}_i'\underline{\beta}}. \tag{6}$$

But again π_i is unobservable and data are available only on λ and P_1, we cannot estimate $\underline{\beta}$ from (6). For estimating $\underline{\beta}$, using (5) and (6) we can write

$$\lambda_i = \frac{e^{\underline{X}_i'\underline{\beta}} + (1-P_1)}{e^{\underline{X}_i'\underline{\beta}}}. \tag{7}$$

Similar to the Hussain and Shabbir (2008), for $P_1 = 1$, it reduces to ordinary logit estimation.

Consider a random variable Y_i which takes on the value 1 with probability λ_i and the value 0 with probability $(1-\lambda_i)$. Then the likelihood function for $\underline{\beta}$ is given by

$$L(\underline{\beta} \mid \underline{Y}) = \prod_{i=1}^{n} \lambda_i^{Y_i} (1-\lambda_i)^{1-Y_i}, \tag{8}$$

or

$$\log L(\underline{\beta} \mid \underline{Y}) = \ell = \sum_{i=1}^{n} \left\{ Y_i \lambda_i + (1-Y_i) \log(1-\lambda_i) \right\}$$

The first derivative of this log-likelihood function with respect to $\underline{\beta}$ is given by

$$\frac{\partial \ell}{\partial \underline{\beta}} = \sum_{i=1}^{n} \left[\frac{Y_i e^{\underline{X}_i'\underline{\beta}}}{e^{\underline{X}_i'\underline{\beta}} + (1-P_1)} + \left(1 + e^{\underline{X}_i'\underline{\beta}}\right)^{-1} \right] \underline{X}_i'. \tag{9}$$

As discussed by Hussain Shabbir (2008), setting this derivative equal to zero does not give us the closed form solution of $\underline{\beta}$. So, we solve this expression numerically and find out the estimates of $\underline{\beta}$ for different values of P_1 and compare them with the ordinary logit estimates of $\underline{\beta}$ when $P_1 = 0$.

4 Comparisons

Now we compare the proposed logit estimation with the ordinary logit estimation and Hussain and Shabbir (2008) hidden logit estimation at equal level of privacy.

(i) proposed logit estimation versus ordinary logit estimation

To compare the two logit estimates of $\underline{\beta}$ we do a small simulation study because maximum likelihood estimators have certain properties such as consistency, normality and efficiency for large samples (King 1998), and Greene (2000)). As mentioned above for $P_1 = 1$, hidden logit estimates of $\underline{\beta}$ reduce to ordinary logit estimates, we compare them for different values of P_1. For this purpose data were generated as follows.

A three regressors equation with no constant term is assumed. For simplicity we assumed $n = 5000$, $\beta_j = 1$ and X_j follows $U(-3,3)$ for $j = 1,2,3$. The results for various values of P_1 are shown in Table 1. To examine the behavior of the estimates $\hat{\beta}_j$'s of β_j's we have plotted $\hat{\beta}_j$'s and their standard errors against different values of P_1 in Figure 1 and Figure 2 respectively. From Figure 1 it can be seen that as the value of P_1 increases the estimates of proposed hidden logit approximately approach to the true parametric values. Form Figure 2, it is observed that the standard errors of the estimates decrease as the value of P_1 increases and the standard errors are at their minimum when $P_1 = 1$. Thus it is evident that we gain in terms of trustworthy data and almost unbiased estimates at the cost of decreased precision. Thus it is trade off between honest data and the precision.

(ii) Proposed versus Hussain and Shabbir (2008) at equal privacy level

Leysieffer and Warner (1976) suggested a measure of privacy called jeopardy measure. According to them population is divided into two complementary groups A and A^c having proportions π and $(1-\pi)$ respectively. In case of *yes/no* responses a response R is either *yes* or *no*. Using the conditional probabilities of response R, $P(R \mid A)$ and $P(R \mid A^c)$, jeopardy functions are defined as

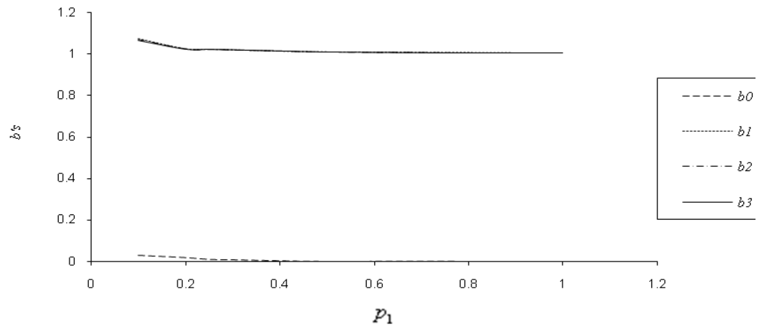

Figure 1: Estimates of $\beta's\left(b's\right)$ against P_1 for n = 5000

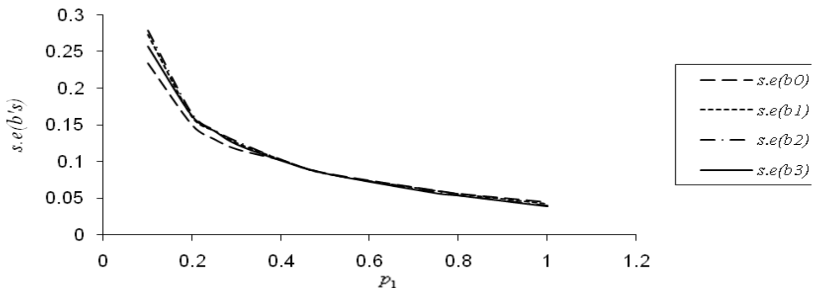

Figure 2: Standard errors of estimates for several values of P_1 when n = 5000

$$g\left(R|A\right)=\frac{P\left(R|A\right)}{P\left(R|A^c\right)}. \tag{10}$$

$$g\left(y|A\right)=\frac{P\left(y|A\right)}{P\left(y|A^c\right)}\ . \tag{11}$$

Also,

$$g\left(n|A^c\right)=\frac{P\left(n|A^c\right)}{P\left(n|A\right)}. \tag{12}$$

In order to obtain logit estimates for proposed and Hussain and Shabbir (2008) at equal privacy level, we proceed as follows:

Jeopardy function for Hussain and Shabbir (2008) is given by

$$g_W(y|A)=\frac{P}{1-P}=k_1(say). \tag{13}$$

$$g_W(n|A^c)=\frac{P}{1-P}=k_2(say). \tag{14}$$

The jeopardy functions for M_1 model are

$$g_{M_1}\left(y|A\right)=\frac{1}{1-P_1}=k_1, \tag{15}$$

and

$$g_{M_1}\left(n|A^c\right)=k_2. \tag{16}$$

At equal privacy level (13) must be equal to (15). So after equating these expressions we get:

$$g_{M_1}\left(y|A\right)=g_W\left(y|A\right),$$

$$\frac{1}{1-P_1} = \frac{P}{1-P} ,$$

$$P = \frac{1}{2-P_1} . \tag{17}$$

From (17) we get P in terms of P_1. Now we do logit estimation with Warner (1965) model for those values of P which we get using relation (17). Now we are in a position to compare proposed and Hussain and Shabbir (2008). The values of P are given below.

Table 2: Values of P evaluated using relation (17)

P_1	0.1	0.2	0.25	0.3	0.4	0.5	0.75	0.8	1
$P = \frac{1}{2-P_1}$	0.5263	0.5556	0.5714	0.5882	0.625	0.6667	0.8	0.8333	1

We know from Hussain and Shabbir (2008) that the estimates become close to parametric values when P moves from 0.50 to 1.0, and the behavior of standard errors of estimates is symmetric around $P = 0.5$. The results for various values of P are shown in Table 3. To examine the behavior of the estimates $\hat{\beta}_j$'s of β_j's we have plotted $\hat{\beta}_j$'s and their standard errors against different values of P in Figure 3 and Figure 4 respectively. From Figure 3 it is observed that for P greater than 0.5263 estimates becomes closer to the parametric values. The standard errors of estimates decrease as P moves away from P =0.5714 (see Figure 4). When P_1 = 0.10, the estimate b_0 from proposed model is 0.0282 and its standard error is 0.2348. At equal privacy level using the relation (17) we have *P*=0.5263 .When *P*=0.5263 the estimate b_0 for Warner (1965) model is 0.1790 and standard error of estimate is 0.3358.

It is clear from the above values that at equal privacy level estimates from proposed model are more close to the parameter value as compared to the Warner (1965) model. Also, we can examine from above mentioned values that the standard errors for proposed model are less than Warner (1965) model. When P_1 = 0.80 the estimate (b_1) for proposed model is 1.0056 and its standard error is 0.0556. For P_1 = 0.80, at equal privacy level we get *P* = 0.8333. When *P* = 0.8333 the estimated value (b_1) of β_1 for Warner (1965) model is 1.0099 and its standard is 0.0709. It can be observed from study that estimates obtained from proposed model are more close to the parametric values and their standard errors are small as compared to that obtained from Warner (1965) model.

5 Hidden logit estimation using EM algorithm

Now we move towards the situation where some responses are lost due to some reason. One solution is to discard the cases with missing values and analyze the rest of the cases by standard procedures treating it as complete data. Advantage of this approach is simplicity since standard complete data statistical analysis can be applied without modifications. Drawback of this approach is the loss of information, that is, we sacrifice the information stored in independent variables. To handle this problem here we use Expectation Maximization (*EM*) algorithm. The *EM* algorithm is a very general iterative

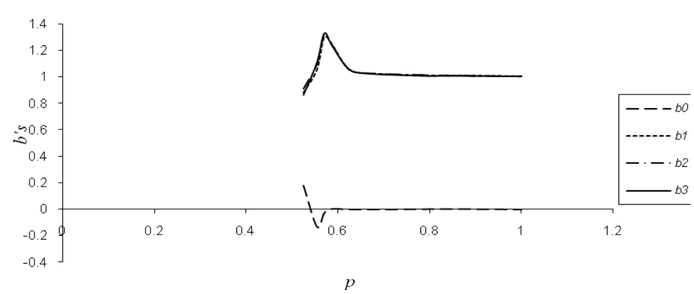

Figure 3: Estimates for several values of P when *n* = 5000

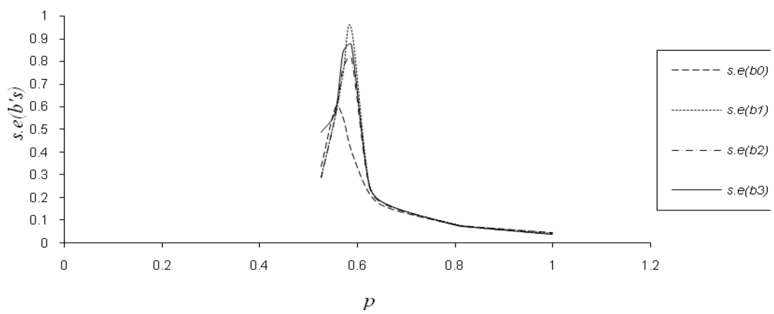

Figure 4: Standard errors of estimates for several values of P when *n* = 5000

algorithm for maximum likelihood (*ML*) estimation in incomplete-data problems explained by Dempster et al. (1977), Little and Rubin (2002), Longford (2005) etc. Little and Rubin (2002) stated that: "*The EM algorithm formalizes a relatively old ad hoc idea for handling missing data, i.e. (1) replace missing values by estimated values, (2) estimate parameters, (3) re-estimate the missing values assuming the new parameter estimates are correct, (4) re-estimate parameters, and so forth, iterating until convergence. Such methods are EM algorithms for models where the complete-data log-likelihood* $l\left(\theta \middle| Y_{obs}, Y_{mis}\right) = \ln l\left(\theta \middle| Y_{obs}, Y_{mis}\right)$". Each iteration of *EM* consists of an *E* step (expectation step) and an *M* step (maximization step). The steps involved in estimating *β's* are described as below:

- **Step 1:** We set some initial estimates of $\underline{\beta}$ say $\underline{\beta}^0$.
- **Step 2:** Calculate the expected complete data log-likelihood, which we find by conditional expectation based on observed data and initial estimates, $\underline{\beta}^0$.
- **Step 3:** As we have complete data log-likelihood so now we find the maximum likelihood estimates for $\underline{\beta}$ using logit estimation procedure as we have done before.
- **Step 4:** Based on these current estimates repeat step 2 and 3 until convergence occurs.

Step 2 is executed here by simply replacing missing entries by their conditional expectation on *X's* and initial parameters $\underline{\beta}^0$ that is given by:

$$E\left(Y_j \middle| \underline{X}_i', \underline{\beta}^0\right) = P\left(Y_j = 1 \middle| \underline{X}_i', \underline{\beta}^0\right) = \frac{\varphi + e^{\underline{X}_i'\underline{\beta}}}{1 + e^{\underline{X}_i'\underline{\beta}}}. \qquad (18)$$

As the factor φ is included to incorporate the effect of *RR* procedure, due to which it is known as hidden logit estimation procedure. Here $\varphi = 1 - P_1$. So, expression involves in step 2 for this case becomes:

$$E\left(Y_j \middle| \underline{X}_i', \underline{\beta}^0\right) = P\left(Y_j = 1 \middle| \underline{X}_i', \underline{\beta}^0\right) = \frac{\left(1 - P_1\right) + e^{\underline{X}_i'\underline{\beta}}}{1 + e^{\underline{X}_i'\underline{\beta}}}. \qquad (19)$$

It means that we replace each missing observation by (19). Here we put some function of missing data instead of missing value. When $P_1 = 1$, the above expression becomes the same as that for ordinary logit.

To compare the two logit estimates of $\underline{\beta}$ when some observations are missing and EM algorithm is used to over come this problem we do a small simulation study. The results of this simulated study for sample size 5000 for different values of P_1 are shown in Table 3. To examine the behavior of the estimates we have plotted estimates and their standard errors against different values of P_1 in Figure 5 and Figure 6 respectively. It is examined form Figure 5 that the values of the estimators approach, approximately, to the true value of the parameters as P_1 increases, and are more close to parametric values for P_1 =1, that is, for ordinary logit. But ordinary logit do not provide the required results. As respondent might refuse to respond or gave wrong information about his/her true status in sensitive psychological surveys.

So we prefer hidden logit to gain true information. In this way we are in better position to draw conclusion about any sensitive or stigmatizing characteristic of the respondent.

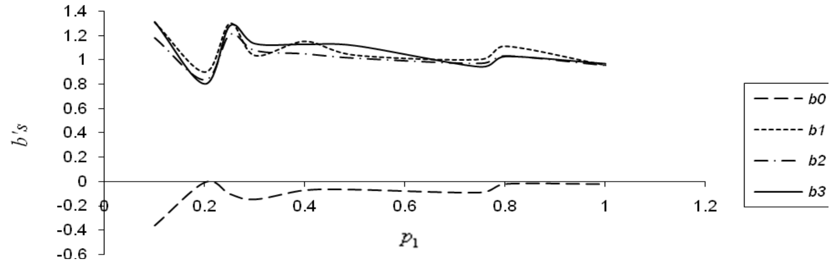

Figure 5: Estimates of $\beta's$ for different values of P_1 when *n = 5000*

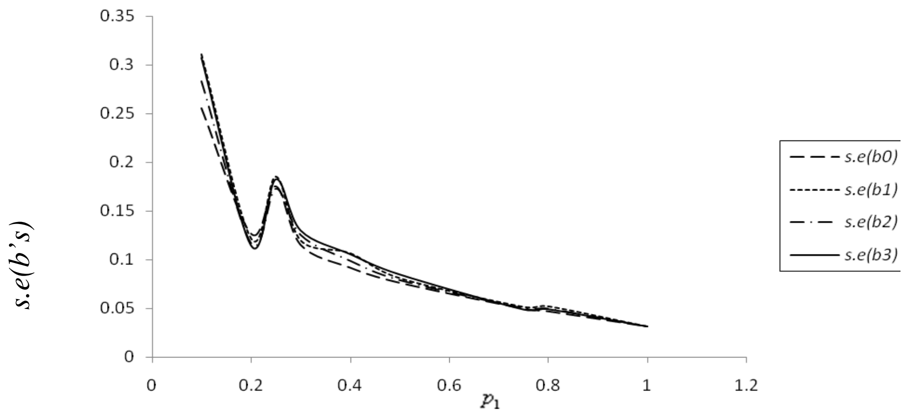

Figure 6: Standard errors of estimates for different values of P_1 when *n = 5000*

6 Conclusion

We find that the estimates of hidden logit approaches to true parametric values as the design probability P_1 increases. Also we examine that the standard errors of estimates decreases as P_1 increases, and is least for ordinary logit that is $P_1 = 1$. Also we observe that hidden logit estimates for Mangat (1994) are closer to the true parametric values as compare to the Warner (1965). Also they show increase in precision. So hidden logit estimation using Mangat (1994) is more appropriate to obtain true estimates of population proportion. If there are some missing responses than it is beneficial to use *EM* algorithm. In this way at least the information stored in independent variables can be utilized. Although standard errors of estimator are least for ordinary logit as compared to hidden logit even then it is preferable to use hidden logit in studying embarrassing/stigmatizing characteristics. Because if we do not provide the respondent sufficient anonymity in sensitive psychological surveys he/she will either refuse to give response or not respond truthfully. So we prefer to attain more accurate information through *RR design.*

References

1 Bradely, L. (1988). Making connections in learning to read and to spell. Applied Cognitive Psychology. 2(1), 3-18.

2 Capel, S. A. (1986). Psychological and organizational factors related to burnout in athletic trainers. Research Quarterly for Exercise and Sport, 57, 321-328.

3 Chaudhuri, A. and Mukerjee, R. (1988). Randomized response: Theory and Methods. Marcel- Decker, New York.

4 Christofides, T. C. (2003). A generalized randomized response technique. *Metrika*, 57, 195-200.

5 Clark, D.A. and Beck, A.T., 1991. Personality factors in dysphoria: A psychometric refinement of Beck's Sociotropy-Autonomy Scale. *Journal of Psychopathology and Behavioral Assessment* 13, pp. 369–388.

6 Clark, D.A. and Beck, A.T., 1991. Personality factors in dysphoria: A psychometric refinement of Beck's Sociotropy-Autonomy Scale. *Journal of Psychopathology and Behavioral Assessment* 13, pp. 369–388.

7 Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from Incomplete data via EM algorithm. *Journal of Royal Statistical Society, B*, 39 (1), 1-38.

8 Friedrich, W. N., Urquiza, A. J. and Beilke, R. L. (1986). Behavior Problems in Sexually Abused Young Children. Journal of Pediatric Psychology, 11(1), 47-57.

9 Greenberg, B. G., Kuebler, R. R., Jr., Abernathy, J. R. and Horvitz, D. G. (1969). The unrelated question randomized response model: theoretical framework. *Journal of the American Statistical Association*, 64, 520-539.

10 Hedeker, D., McMahon, S. D., Jason, L. A., & Salina, D. (1994). Analysis of clustered data in community psychology: with an example from a worksite smoking cessation project. American Journal of Community Psychology, 22, 595-615.

11 Hussain, Z. and Shabbir, J, 2010. Three stage quantitative randomized response model. Journal of Probability and Statistical Sciences, 8(2): 223-235.

12 Hussain, Z. and Shabbir, J. (2008): Logit estimation using Warner's randomized response model, *Journal of Modern Applied Statistical Methods*, 7 (1), 140-151.

13 Hussain, Z., Shabbir, J. and Gupta S, 2007. An alternative to Ryu et al. randomized response model. Journal of Statistics & Management Systems, 10(4): 511-517.

14 Kim, J. M. and Warde, D. W. (2004). A stratified Warner's randomized response model. *Journal of Statistical Planning and Inference*, 120 (1-2), 155-165.

15 Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd edition, New York, John Wiley.

16 Longford, N. T. (2005). Missing data and small-area estimation : Modern Analytical equipment for the survey statistician. New York, Springer.

17 Mangat, N. S. (1994). An improved randomized response strategy. *Journal of Royal Statistical Society*, B, 56(1) 93-95.

18 Mangat, N.S. and Singh R. (1990). An alternative randomized response procedure. *Biometrika*, 77, 439-442.

19 Singh, S., Horn, S., Chowdhuri, S. (1998). Estimation of stigmatized characteristics of a hidden gang in finite population. *Australian and New Zealand Journal of Statistics*, 40(3), 291-297.

20 Tracy, D. and Mangat, N. (1996). Some development in randomized response sampling during the last decade-a follow up of review by Chaudhuri and Mukerjee. *Journal of Applied Statistical Science*, 4, 533-544.

21 Waldman I. D., Robinson, B. F. & Rowe, D. C. (1999). A logistic regression based extension of the TDT for continuous and categorical traits. Annals of Human Genetics, 63,329-340.

22 Waldman, I.D. & Rhee, S.H. (2002). Behavioral and molecular genetic studies of ADHD. In Sandberg, S. (Ed.) *Hyperactivity and Attention Disorders in Childhood* (2nd Ed.). New York: Cambridge University Press.

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

APPENDIX

Derivation of hidden logit for M_1 model

As

$$\lambda = p_1\pi + (1 - p_1) \quad (1)$$

$$\pi = \frac{\lambda - (1 - p_1)}{p_1} \quad (2)$$

$$\ln\left(\frac{\pi}{1-\pi}\right) = X_i\beta \quad (3)$$

Putting (2) in (3) we get:

$$\ln\left[\frac{\lambda-(1-p_1)}{p_1}\Big/1-\frac{\lambda-(1-p_1)}{p_1}\right]=X_i\beta \tag{4}$$

$$\ln\left[\frac{\lambda-(1-p_1)}{p_1-\lambda+(1-p_1)}\right]=X_i\beta \tag{5}$$

$$\ln\left[\frac{\lambda-(1-p_1)}{1-\lambda}\right]=X_i\beta \tag{6}$$

$$\frac{\lambda-(1-p_1)}{1-\lambda}=e^{x_i\beta} \tag{7}$$

$$\lambda-(1-p_1)=e^{X_i\beta}-\lambda e^{X_i\beta} \tag{8}$$

$$\lambda\left(1+e^{X_i\beta}\right)=e^{X_i\beta}+(1-p_1) \tag{9}$$

$$\lambda=\frac{e^{X_i\beta}+(1-p_1)}{1+e^{X_i\beta}}\ . \tag{10}$$

Derivation of the first derivative

$$\ln L=\sum_{i=1}^{n}\left\{y_i\ln\lambda_i+(1-y_i)\ln(1-\lambda_i)\right\} \tag{11}$$

Differentiate w.r.to β

$$\frac{\partial\ln L}{\partial\beta}=\sum_{i=1}^{n}\left\{y_i\frac{\partial\ln\lambda_i}{\partial\beta}+(1-y_i)\frac{\partial\ln(1-\lambda_i)}{\partial\beta}\right\} \tag{12}$$

$$=\sum_{i=1}^{n}\left\{y_i\frac{1}{\lambda_i}\frac{\partial\lambda_i}{\partial\beta}+(1-y_i)\frac{1}{(1-\lambda_i)}(-1)\frac{\partial\lambda_i}{\partial\beta}\right\} \tag{13}$$

Consider $\frac{\partial\lambda_i}{\partial\beta}$

$$\frac{\partial\lambda_i}{\partial\beta}=\frac{\partial}{\partial\beta}\left(e^{X_i\beta}+(1-p_1)\right)\left(1+e^{X_i\beta}\right)-\left(e^{X_i\beta}+(1-p_1)\right)\frac{\partial}{\partial\beta}\left(1+e^{X_i\beta}\right)\Big/\left(1+e^{X_i\beta}\right)^2 \tag{14}$$

$$=\frac{X_ie^{X_i\beta}\left(1+e^{X_i\beta}\right)-\left(e^{X_i\beta}+(1-p_1)\right)e^{X_i\beta}X_i}{\left(1+e^{X_i\beta}\right)^2} \tag{15}$$

$$=X_ie^{X_i\beta}\left[\frac{p_1}{\left(1+e^{X_i\beta}\right)^2}\right]. \tag{16}$$

Put this in (13) and simplifying we get:

$$\frac{\partial\ln L}{\partial\beta}=\sum_{i=1}^{n}\left[\frac{y_ip_1X_ie^{X_i\beta}}{\left(e^{X_i\beta}+(1-p_1)\right)\left(1+e^{X_i\beta}\right)}-\frac{(1-y_i)e^{X_i\beta}X_i}{1+e^{X_i\beta}}\right] \tag{17}$$

$$=\sum_{i=1}^{n}\left[\frac{y_ip_1}{\left(e^{X_i\beta}+(1-p_1)\right)\left(1+e^{-X_i\beta}\right)}-\frac{(1-y_i)}{1+e^{-X_i\beta}}\right]X_i \tag{18}$$

$$=\sum_{i=1}^{n}\left[\frac{y_ip_1}{\left(e^{X_i\beta}+(1-p_1)\right)\left(1+e^{-X_i\beta}\right)}+\frac{y_i}{1+e^{-X_i\beta}}-\frac{1}{1+e^{-X_i\beta}}\right]X_i \tag{19}$$

$$=\sum_{i=1}^{n}\left[y_i\left\{\frac{p_1}{\left(e^{X_i\beta}+(1-p_1)\right)\left(1+e^{-X_i\beta}\right)}+\frac{1}{1+e^{-X_i\beta}}\right\}-\left(1+e^{-X_i\beta}\right)^{-1}\right]X_i \tag{20}$$

$$=\sum_{i=1}^{n}\left[y_i\left\{\frac{p_1+e^{X_i\beta}+(1-p_1)}{\left(e^{X_i\beta}+(1-p_1)\right)\left(1+e^{-X_i\beta}\right)}\right\}-\left(1+e^{-X_i\beta}\right)^{-1}\right] \tag{21}$$

$$=\sum_{i=1}^{n}\left[y_i\left\{\frac{e^{X_i\beta}\left(1+e^{-X_i\beta}\right)}{\left(e^{X_i\beta}+(1-p_1)\right)\left(1+e^{-X_i\beta}\right)}\right\}-\left(1+e^{-X_i\beta}\right)^{-1}\right] \tag{22}$$

$$=\sum_{i=1}^{n}\left[y_i\left\{\frac{e^{X_i\beta}}{\left(e^{X_i\beta}+(1-p_1)\right)}\right\}-\left(1+e^{-X_i\beta}\right)^{-1}\right]X_i\ . \tag{23}$$

Table 1*: Estimates and their standard errors of ordinary and hidden logit models for various values of p_1 for M_1 model when n = 5000*

Entries in the table represent estimates of coefficients and respective standard errors in parenthesis.

	β_i	(ordinary logit)	$p_1=0.10$	$p_1=0.20$	$p_1=0.25$	$p_1=0.30$	$p_1=0.40$	$p_1=0.50$	$p_1=0.75$	$p_1=0.80$
X_0	0	0.0002 (0.0445)	0.0282 (0.2348)	0.0182 (0.1499)	0.0103 (0.1307)	0.0085 (0.1171)	0.0040 (0.1021)	0.0011 (0.0844)	0.0018 (0.0605)	0.0009 (0.0563)
X_1	1	1.0034 (0.0419)	1.0746 (0.2738)	1.0247 (0.1605)	1.0222 (0.1420)	1.0191 (0.1254)	1.0129 (0.1025)	1.0092 (0.0833)	1.0061 (0.0598)	1.0056 (0.0556)
X_2	1	1.0032 (0.0405)	1.0695 (0.2786)	1.0240 (0.1645)	1.0206 (0.1430)	1.0182 (0.1278)	1.0130 (0.1031)	1.0086 (0.0843)	1.0045 (0.0586)	1.0039 (0.0544)
X_3	1	1.0022 (0.0394)	1.0655 (0.2572)	1.0215 (0.1621)	1.0209 (0.1421)	1.0184 (0.1238	1.0128 (0.1018)	1.0073 (0.0843)	1.0028 (0.0573)	1.0027 (0.0542)

Table 2*: Estimates and their standard errors of ordinary and hidden logit models for Warner's (1965) model at equal privacy level to M_1 model when n = 5000*

Entries in the table represent estimates of coefficients and respective standard errors in parenthesis.

	βi	ordinary logit	p =0.5263	p =0.5556	p =0.5714	p =0.5882	p =0.625	p =0.6667	p =0.8	p =0.8333
X_0	0	-0.0029 (0.0460)	0.1789 (0.3357)	-0.1338 (0.5959)	-0.0304 (0.5504)	0.0011 (0.4055)	-0.0035 (0.2182)	-0.0033 (0.1526)	-0.0004 (0.0813)	0.0003 (0.0730)
X_1	1	1.0043 (0.0404)	0.8782 (0.2897)	1.0535 (0.5605)	1.3025 (0.7486)	1.2465 (0.9440)	1.0539 (0.2563)	1.0255 (0.1628)	1.0097 (0.0818)	1.0099 (0.0709)
X_2	1	1.0032 (0.0396)	0.9078 (0.2881)	1.0952 (0.5612)	1.3211 (0.7716)	1.2412 (0.8168)	1.0570 (0.2487)	1.0259 (0.1622)	1.0115 (0.0839)	1.0082 (0.0715)
X_3	1	1.0036 (0.0391)	0.8642 (0.4870)	1.1070 (0.5799)	1.3258 (0.8414)	1.2344 (0.8705)	1.0572 (0.2531)	1.0240 (0.1657)	1.0064 (0.0799)	1.0080 (0.0707)

Table 3*: Estimates and their standard errors of ordinary and hidden logit models for M_1 model using EM algorithm in case of missing values when n = 5000*

Entries in the table represent estimates of coefficients and respective standard errors in parenthesis.

	β_i	Ordinary logit	$p_1=0.10$	$p_1=0.20$	$p_1=0.25$	$p_1=0.30$	$p_1=0.40$	$p_1=0.50$	$p_1=0.75$	$p_1=0.80$
X_0	0	-0.0181 (.0316)	-0.3586 (0.2553)	-0.0044 (0.1268)	-0.0993 (0.1753)	-0.1442 (0.1153)	-0.0683 (0.0919)	-0.0638 (0.0764)	-0.0871 (0.0499)	-0.0171 (0.0471)
X_1	1	0.9543 (.0314)	1.3079 (0.3104)	0.8967 (0.1217)	1.2999 (0.1851)	1.0343 (0.1197)	1.1517 (0.1066)	1.0371 (0.0810)	1.0021 (0.0515)	1.1116 (.0521)
X_2	1	0.9545 (.0316)	1.1827 (0.2830)	0.8343 (0.1171)	1.2099 (0.1729)	1.0784 (0.1255)	1.0504 (0.0990)	1.0151 (0.0792)	0.9710 (0.0504)	1.0334 (0.0499)
X_3	1	0.9698 (.0317)	1.3142 (0.3075)	0.8026 (0.1149)	1.2822 (0.1828)	1.1345 (0.1303)	1.1290 (0.1060)	1.1184 (0.0851)	0.9436 (0.0492)	1.0288 (0.0494)