© Kamla-Raj 2009 J Soc Sci, 19(1): 57-61 (2009)
PRINT: ISSN 0971-8923 ONLINE: ISSN 2456-6756 DOI: 10.31901/24566756.2009/19.01.08

A Multivariate Statistical Analysis of Reporting Error in Age Data of India

Barun Kumar Mukhopadhyay* and Prasanta Kumar Majumdar

Population Studies Unit, Indian Statistical Institute, 203 B.T.Road, Kolkata 700 108,
West Bengal, India
**E-mail: barun_mukhopadhyay@yahoo.com*

KEYWORDS Myers' Index. Backward Elimination. Commonality Factors

ABSTRACT It is now a general concept that education improves the quality of age reporting. However, in the paper an attempt has been made to analysis the age reporting error vis-à-vis different socio economic cultural and developmental factors in addition to literacy factor in order to find out the different unique factors and their commonalities in influencing the variations in the age reporting error in India. The data for this type of analysis have been obtained from the reports of National Family Health Survey (NFHS) for two periods 1992-93 and 1998-99. In those reports a measure of age reporting error separately for male and female populations has been given by Myers' indices. Since handling with large number of variables which may not be significant or may not be worth at all, at the initial stage of analysis all the variables were undergone through backward regression process where a number of variables were deleted as they were found insignificant at a certain level of significance. The remaining variables then were analyzed through commonality analysis which gave some interesting results. Still literacy of male played a significant role in improving the quality of age data. In female part, scheduled tribe population has some influence on the reporting error. Mean household size also played some role in influencing the age reporting error. Apart from these some commonalities have been found between urban households and mean household size or between household response rate and many other. Apart from these there was some difference between the two periods of 1992-93 and 1998-99.

INTRODUCTION

Age reporting error has been a perennial cause of concern for many social scientists including demographers in particular and actuaries, sociologist, economist etc., especially in developing countries. India's position even much worse than some of its neighbouring countries. The importance of age is well known if it is correctly available or properly adjusted .The studies on age reporting error need special attention since the errors in the age distribution particularly in censuses are examined more intensively than any other information (Shryock et al.1973). Since the social and economic characteristics vary so much with age and also vary in time and place, population can not be meaningfully compared with respect to these unless age has been controlled. Accordingly, age is considered to be the variable of highest priority in demographic analysis (Srinivasan 1998). The importance of age is further quoted (Kerr 2003) as "it is advisable to adjust the age data before its use in many fields of research". Keeping in view of the above facts, there is always a need to evaluate the Indian age data. In evaluating the data, it is important to investigate the direction and magnitude of error, so that necessary adjustment could be made. Nevertheless, it is also important to know the reasons for the deficiencies in the quality of the data. It is more desirable to get accurate data rectifying the deficiencies than to make adjustment on faulty data. It is pointed out (Ambannavar and Visaria 1975) that the improvement in the quality of age data, especially the age heaping can be explained in favour of rise of education.

Heaping is defined here as the over reporting of ages at some preferred digits of ages. A statistically tested heaping index has been proposed (Mukhopadhyay and Muherjee 1988) and applied on a number of countries with some good results. But apart from education there may be other socio cultural, developmental factors which may influence the quality of age data. In this context, the present paper attempts to study the quality of age data in terms of digit preference error vis à vis the different socio, economic, cultural and developmental factors affecting the age reporting error in India through some multivariate statistical analysis. In an earlier attempt (Mukhopadhyay 1988) a similar kind of study though not statistically tested has been made to know the socio economic classification in age reporting error in rural West Bengal on the basis of data from a household survey conducted in rural areas of districts around calcutta (Sarkar 1981).

DATA AND METHODOLOGY OF STUDY

In order to study this kind of detailed statistical analysis, National Family Health Survey (NFHS) data would be appropriate in the sense that this is a huge survey covering all the states of India on a random sample basis. NFHS-1 (IIPS 1995) was conducted during 1992-93 and NFHS-2 (IIPS 2000) was conducted during 1998-99 period. These surveys were conducted under a joint collaboration among the Government of India, New Delhi, International Institute for Population Sciences, Mumbai and the Macro International company of the United States of America.

NFHS (I) is comprised of 25 states namely Delhi, Haryana, Himachal Pradesh, Jammu & Kashmir, Punjab, Rajasthan, Madhya Pradesh, Uttar Pradesh, Bihar, Orissa, West Bengal, Arunachal Pradesh, Assam, Manipur, Meghalaya, Mizoram, Nagaland, Goa, Gujarat, Maharashtra, Andhra Pradesh, Karnataka, Kerala, Tamil Nadu, and Tripura, and NFHS (2) also comprised of all these states including another state, Sikkim. As such 26 states in NFHS (2) during 1998-99 are available. Since in the present study there was no interest to individually study the states rather the aim of the study is to find out the different significant factors and their combinations to influence the age reporting error, therefore only 25 study units in 1992-93 and 26 in 1998-99 have been considered separately. Combining these two units there are altogether 51 units of the study design in order to have more valued results from higher sample units .

In order to quantify the quality of age reporting error different indices are in vogue in demographic research. In the present case, the age reporting error is measured through an important and useful index like Myers' index (Myers 1940) which in the present case, is the dependent variable and the calculated figures are already available by two sexes separately in the different reports pertaining to different states of India of those two surveys(i) NFHS(1),1992-93 and (ii) NFHS (2), 1998-99. It has to be kept in mind that female data in NFHS survey are superior to male counterpart since special attention was paid to female data collection taking only eligible females of 13-49 age groups. Hence whatever values for Myers' indices have been found for female data might not be comparable with male figure. However, their separate study is quite possible and contributions of different factors with their combinations are worth to be studied.

The different independent variables considered for the study were available in the reports included, (i) per cent urban households, (ii) per cent illiterate male, (iii) per cent illiterate female, (iv) per cent scheduled caste population, (v) per cent scheduled tribe population, (vi) mean household size, (vii) per cent of persons living in less than 3 persons per room, (viii) per cent of persons with low living index, (ix) per cent of population having no agricultural land, (x) household response rate, and (xi) eligible women's response rate. These independent variables have been considered according to the availability from each of the survey periods of 1992-93 and 1998-99. As for example, there was no information on the low living index in the earlier period of 1992-93.

For statistical analysis all the information are entered in code sheets for both the periods separately. Different codes, like 1 for 1992-93 and 2 for 1998-99 surveys have been assigned. In all other cases the actual values with one place of decimal have been put in consecutive cells for the two different data sheets. As the study has been done separately for male and female populations, literacy variable also has been done correspondingly male to male and female to female. Similarly household response rate has been taken into account for male population and eligible women's response rate for female population. As the entire calculation was done in computer through SPSS package with 7.5 version, all the data were transcribed into the SPSS data layout with suitable variable names putting on each column head into the categories, such as one for 1992-93, another for 1998-99 and finally one combined form with 51 rows.

The multivariate statistical analysis has been performed at two stages, firstly eliminating some unimportant and insignificant independent variables out of the total list considered through backward regression process and secondly the remaining variables have been analyzed through commonality analysis (Kerlinger and Pedhazur 1973).

ACTUAL ANALYSIS OF THE DATA

To start with actual analysis of the data under three categories such as data for the period 1992-93 and 1998-99 combined (Table 1), then each 1992-93 and 1998-99, different list of variables for these three categories have been considered separately for male and female populations and are given with their abbreviated names (Tables 2, 3).

Table 1: Name of variables separately for male and female populations for NFHS, 1992-93 & 1998-99 combined

Dependent Variable	*Independent Variables*
mi	Puh,pillm,psc,pst,mhhs,pllt3r,pnagl,hrr
mf	puh,pillf,psc,pst,mhhs,pllt3r,pnagl,ewrr

Note: mi= Myers' index (male); mf= Myers' index (female)
puh= per cent of urban households; pillm= per cent of illiterate male
pillf= per cent of illiterate female; psc= per cent of scheduled caste
pst = per cent of scheduled tribe; mhhs = mean household size
pllt3r = per cent of population living in less then 3 persons per room
pnagl = per cent of population having no agricultural land
hrr = household response rate; ewrr = eligble women's response rate

Table 2: Name of variables separately for male and female populations for NFHS, 1992-93

Dependent variable	*Independent variables*
mi	puh,pillm,psc,pst,mhhs,pllt3r,pnagl,hrr
mf	puh,pillf,psc,pst,mhhs,pllt3r,pnagl,ewrr

Note: foot note is same as that of table 1

Table 3: Name of variables separately for male and female populations for NFHS, 1998-99

Dependent variable	*Independent variables*
mi	puh,pillm,psc,pst,mhhs,pllt3r,pnagl,hrr
mf	puh,pillf,psc,pst,mhhs,pllt3r, pplli, pnagl,ewrr

Note: foot note is same as that of table 1, the only additional variable, pplli = per cent of population with low living index

Considering all the independent variables pertaining to male counterpart of the combined periods of 1992-93 and 1998-99, backward elimination was run in the computer with SPSS package of 7.5 version with acceptance level of 0.05 and rejection level of 0.10, most of the independent variables were found to be insignificant except pillm and hrr for which the following commonality analysis was performed (Table 4).

Table 4: Summary of commonality analysis of pillm and hrr for male population of India, 1992-93 & 1998-99

Different combinations	*Variables*	
	pillm	*hrr*
Unique to 1, pillm	0.235	
Unique to 2, hrr		0.152
Common to 1&2	0.104	0.104
Σ	0.339	0.256

From the above table pillm (per cent illiterate male) and hrr (household response rate) explained around 50 per cent of variation in the dependent variable, mi (Myers'index male). Out of these two, pillm singly contributed around 24 per cent, whereas hrr explained relatively less variation (15 per cent). Commonality of both the independent variables turned out to be around 10 per cent. Illiteracy of male or literacy factor is the most significant variable influencing the quality of age reporting in India followed by household response rate.

Similar analysis when conducted for female population of the combined period of 1992-93 and 1998-99, out of total list of independent variables, only pst (per cent of scheduled tribe) and mhhs (mean household size) were found contributing variation in the Myers' index for female. The different combinations are given in Table 5.

From the table 5, pst (per cent scheduled tribe) and mhhs (mean household size) explained around 24 per cent of variation in the dependent variable, i.e., Myers' index. Mean household size, i.e., mhhs singly contributed around 17 per cent of variation followed by pst, i.e., per cent scheduled tribe explaining less variation (6 per cent). There was almost no commonality of these two factors (less then 1 per cent). In female reporting, error is expected to be less if household size is smaller. Scheduled tribe population in some states of India might be influencing age reporting error. However on the average based on all India figures their effect is less.

After analyzing the data corresponding to the combined period of 1992-93 and 1998-99, it would

Table 5: Summary of commonality analysis of pst and mhhs for female population of India, 1992-93 & 1998-99

Different combinations	*Variables*	
	pst	*mhhs*
Unique to 1, pst	0.063	
Unique to 2, mhhs		0.167
Common to 1&2	0.011	0.011
Σ	0.074	0.178

be interesting to study the individual period separately to find out which factors and their combinations are responsible for errors in age reporting. While studying the data for male counterpart during 1992-93, the backward process gives only illiteracy factor (pillm) contributing variation of the order of around 57 per cent in the Myers' index. All other variables were found either insignificant or meaningless.

The female data for the same period give per cent scheduled tribe (pst) as the most significant factor followed by per cent illiterate female (pillf). The following table gives the details of commonality analysis on these two variables (Table 6).

Table 6: Summary of commonality analysis of pillf and pst, for female population of India,1992-93

Different	*Variables*	
combinations	*Pillf*	*pst*
Unique to 1, pillf	0.11	
Unique to 2, pst		0.433
Common to 1&2	-0.092	-0.092
Σ	0.019	0.341

From the table 6, pilllf, i.e., per cent illiterate females and pst, per cent scheduled tribe together explained variation of the order of 45 per cent in the Myers' index. Out of these two, pst alone contributed about 80 per cent variation followed by 20 per cent variation by pillf. Commonality factor jointly explained variation in the negative direction (9 per cent). In 1992-93, both male and female populations showed illiteracy factor effective in explaining variation in Myers' index, though contribution is higher for male. The contribution of scheduled tribe population for female data was at a higher order (43 per cent).

Finally the analysis has been done on NFHS (2) data for the period, 1998-99. After analysis of the data for male population through backward regression process, four variables namely per cent urban household (puh), per cent scheduled caste population (psc), mean household size (mhhs) and household response rate (hrr) were found significant and meaningful. Commonality analysis has been done only on these four variables and the corresponding table 7 has been given.

From the analysis of the Table 7 it was quite evident that those four variables, puh, psc, mhhs and hrr altogether explained variation of the order of 56 per cent in the dependent variable, Myers' index. Mean household size uniquely contributed maximum out of these four variables with 19 per cent variation. Scheduled caste population came next in this regard with 15 per cent variation followed by household response rate with 9 per cent variation. Finally per cent of urban household contributed minimum with around 7 per cent variation in Myers' index.

Table 7: Summary of commonality analysis of puh, psc, mhhs and hrr for male population of India, 1998-99

Different	*Variables*			
combinations	*puh*	*psc*	*mhhs*	*hrr*
Unique to 1, puh	0.074			
Unique to 2, psc		0.148		
Unique to 3, mhhs			0.19	
Unique to 4, hrr				0.089
Common to 1&2	-0.007	-0.007		
Common to 1&3	0.075		0.075	
Common to 1&4	-0.03			-0.030
Common to 2&3		0.016	0.016	
Common to 2&4		-0.050		-0.050
Common to 3&4			0.089	0.089
Common to1,2&3	0.000	0.000	0.000	
Common to 1,2,&4	0.010	0.010		0.010
Common to2,3,&4		-0.026	-0.026	-0.026
Common to 1,3&4	-0.021		-0.021	-0.021
Common to1,2,3&4	0.003	0.003	0.003	0.003

These four variables were also found contributing jointly in explaining the variation in the Myers' index. A few cases were worth to be noted, otherwise, in maximum cases there were no joint effects. As for example, the maximum contribution of the order of only 9 per cent variation was found jointly with mean household size and household response rate. The next combination in per cent urban households and mean household size with around 7 and half per cent variation was observed from table 8. The values for the other combinations were found to be very insignificant.

In so far as female populations in 1998-99 were concerned, mean household size was the only variable which was retained from backward regression process in the analysis after deleting all other variables which were insignificant and meaningless. Mean household size (mhhs) was found explaining variation of the order of 25 per cent in the dependent variable, Myers' index.

CONCLUSION

From the statistical analysis of the Indian age data during three periods namely, 1992-93 and 1998-99 combined, and each 1992-93 and 1998-

99, some pattern as well as the factors responsible for misreporting of age data were tried to be investigated in the present paper. From the overall analysis of the combined periods of 1992-93 and 1998-99 of the Indian age data (NFHS), first of all, in case of male reporting error, illiteracy of the males played some significant role in explaining the variation in the Myers' index. Next to this factor, household response rate contributed towards variation in the Myers' index. They jointly contributed about one tenth of the population. Apart from these there were no other significant factors.

In case of female data for the same period, only mean household size and per cent scheduled tribe played some significant role barring all other variables. Moreover, out of these two, mean household size contributed more (about 17 per cent) than per cent scheduled tribe (around 6 per cent). Commonality of these two was found negligible. A point must be noted here that as the household size became larger the female age reporting error was also higher.

While studying the pattern between the two periods of time gap of about six years, male data in earlier period of 1992-93 showed illiteracy as the only significant factor influencing age reporting. Other variables were deleted from backward regression process. On the other hand after six years, four factors namely per cent urban households (puh), per cent scheduled caste population (psc), mean household size (mhhs) and household response rate (hrr) were found significant and meaningful in 1998-99. Moreover, out of these four factors mean household size (mhhs) played the most significant role in explaining variation in the Myers' index. Per cent scheduled caste population took the second position followed by household response rate. Lastly per cent urban households contributed to some extent in explaining the variation in the index. A few combined effects were noticed barring most of the other combinations. In case of female data of the two periods of 1992-93 and 1998-99, when earlier period showed per cent scheduled tribe population played most significant role in explaining variation in Myers' index, latter period showed mean household size played the most significant role. Illiteracy factor played, to some extent, in the earlier period.

In sum, it may be concluded that first of all, illiteracy in a true sense literacy factor played some role in controlling the quality of age data both for male and female data. Similarly another important factor, mean household size played some significant role in explaining the quality of age data. It is quite possible that if the number of household members are large, respondents might very well be annoyed to report correctly the ages of all the members of the households. It is also noticed that scheduled caste and scheduled tribe populations were found significant in numbers in some states to report their ages wrongly. However, no development or economic factors considered in the study were found effective when at the initial stage of the analysis they were dropped from the study by backward regression process. There were a few commonality factors playing some role. Seeing all these findings, more and more studies are still necessary to find out more clear picture of problem in order to prescribe important variable for future improvement of this type of more specific factors.

REFERENCES

Ambannavar JP, Visaria P 1975. Influence of literacy and education on the quality of age returns. *Demography India*, 4: 11-15. International Institute for Population Sciences (IIPS) 2000. *National Family Health Survey (MCH and Family Planning), India 1998-99*, Mumbai: IIPS

International Institute for Population Sciences (IIPS) 1995. *National Family Health Survey (MCH and family planning), India 1992-93*. Mumbai: IIPS.

Kerlinger FN, Pedhazur J 1973. *Multivariate Regression in Behavioral Research*. New York: Holt, Holt Rinehart and Winstin, Inc.

Kerr D 2003. An alternative strategy for evaluating and generating censal estimates. *Genus*, Vol-LIX, No.3-4: via Nomentana, 41,00161-Roma, Italy, pp.71-89.

Mukherjee BN, Mukhopadhyay, BK 1988. A study of digit preference and quality of age data Turkish data da censuses. *Genus*, Vol. XLIV- n(I-2): via Nomentana, 41,00161 -Roma Italy, pp. 201-228.

Mukhopadhyay BK1988. A study of social classification in age reporting in rural West Bengal *Rural Demography*, Vol. XV(Nos. 1 & 2): Institute of Statistical Research and Training, University of Dhaka, Bangladesh, pp. 47-58

Myers RJ1940. Errors and bias in the reporting of ages in census data. *Transactions of Acturial Society of America*, XLI (Part 2, No.104): 395-415.

Sarkar BN 1981. *Report on Education and Family Planning, Rural Areas of Districts Around Calcutta*. Calcutta: Demography Research Unit, Indian Statistical Institute.

Shryock HS, Siegel JS et al 1973. *The Methods and Materials of Demography*, Vol. No. 1, USA: U.S Department of Commerce.

Srinivasan K1998. *Basic Demographic Techniques and Applications*. New Delhi: Sage Publications.