HelpAge International: 2011 India NSS Analysis

Stewart Kerr & Rick Griffin

2021-02-15

1. Summary

2. Introduction

HelpAge International wants to challenge established norms for statistical reporting on older persons by proving that data disaggregation to a lower, more granular level is possible and statistically robust. Nongranular statistics reinforces an oversimplified picture of inequalities and the inadequate data itself becomes a barrier to the inclusion of at-risk and marginalized groups in policy and program responses.

To serve this goal, we analyzed data from the employment and unemployment surveys included in the 2011 India National Sample Survey (NSS) to determine the lowest level of disaggregation that was possible while maintaining statistical robust estimates of average wage. The employment and unemployment surveys of the NSS aim to get estimates of various employment characteristics at the national and state level. In addition to employment related variables, individual characteristics such as region, age, sex, industry, education, and others are collected by the survey. In accordance with HelpAge's statement of work, we focused how earnings of employees varied by an individual's age, sex, employment industry, and region (urban/rural). Disability status was only collected in relationship to employment (i.e. unable to work due to disability) and was not available to analyze in relation to average earnings.

In our analysis, we sought to answer three specific research questions related to data disaggregation:

1. What is the most granular level of disaggregation of age, sex, and employment industry? What are the most appropriate age bands (i.e. 5 year groupings or 10 year groupings) and upper age cohort (i.e. 80+, 85+, etc)? How do results differ going from braoder to more granular disaggregation?
2. What is the most granular level of disaggregation of age, sex, and employment industry when we also include geographic location (urban/rural)?

3. Based on these results, what general recommendations or considerations can be made on data disaggregation for similar surveys?

3. Materials and Methods

Data Collection

The 2011 India NSS used a stratified multi-stage design. First villages in rural communities were selected by probability proportional to size with replacement while blocks in urban areas were selected by simple random sampling without replacement. An equal number of villages and blocks were selected. Next, if the village or block contained more than 1200 people, it was divided into subgroups containing roughly the same amount of people. Then, households within each subgroup were stratified into 3 groups according to to measures of wealth. Lastly, households from each strata were selected by simple random sampling without replacement and all individuals within the household were surveyed. Samplings weights were calculated and provided for each individual by the India Ministry of Statistics & Programme Implementation.

We extracted the raw survey data in .sav (SPSS) format from the file provided by HelpAge using the required Nesstar Explorer software. For our analysis, we needed to extract the data files `Block_4_Demographic particulars of household members` and `Block 5_3_Time disposition during the week ended on`. This data was then loaded into R using the `haven` package and processed using the `tidyverse` set of R packages.

Data Processing

First, in order to join demographic data in the "Block 4" (B4) dataset to employment data for the past week in the "Block 5_3" (B53) dataset, we had to create a unique ID. This was accomplished by concatenating the following variables for each dataset: * `FSU_Serial_No`, `Stratum`, `Sub_Stratum_No`, `Hamlet_Group_Sub_Block_No`, `Second_Stage_Stratum_No`, `Sample_Hhld_No`, and `Person_Serial_No`

Then, before joining B4 and B53, the following variables were processed or created from the B53 dataset: * `employment_status` - Takes either "employed", "unemployed" or "not in labor force" depending on the value of `current_weekly_activity_status` * `weekly_earnings` - An individual can have multiple entries in the B53 dataset if they performed multiple jobs during the week. Thus, for each person, we get their total weekly earnings by summing their earnings across the last 7 days. * `industry` - There are many industries reported in the `current_weekly_activity_NIC_2008` variable. We collapsed the industries into 4 groups based on sample size considerations: "farming, forestry, or fishing", "manufacturing", "construction", or "other".

After creating these variables, we joined the B4 and B53 datasets as our "analysis dataset." As we are primarily interested in the average earnings of different groups, we focused only on employed individuals. However, there are many employed people in the dataset that did not report any earnings in the previous 7 days. Nevertheless, **we chose to keep those individuals in our analysis dataset.** Table 1 presents the counts of the people included in our analysis dataset.

Table 1. Counts of people in various groupings of the 2011 India NSS survey..

	Employed		Unemployed		Not in labor force		
	Male	Female	Male	Female	Male	Female	Overall
Under 60	113848	37748	4828	2452	96191	163904	**418971**
60-64	4898	1468	46	21	2368	5782	**14583**
65-69	2825	715	19	11	2291	4433	**10294**
70-74	1298	258	10	0	2050	2905	**6521**
75-79	448	65	2	0	1146	1614	**3275**
80+	248	33	4	0	1284	1786	**3355**
Farming, forestry, or fishing	36486	20314	0	0	0	0	**56800**
Manufacturing	14880	5562	0	0	0	0	**20442**
Construction	16081	2301	0	0	0	0	**18382**
Other	56118	12110	0	0	0	0	**68228**
Urban	47723	11969	2079	1146	40926	72393	**176236**
Rural	75842	28318	2830	1338	64404	108031	**280763**
Total	123565	40287	4909	2484	105330	180424	**456999**

Statistical Analysis and Methodology

4. Discussion

Research question 1

Research question 2

Research question 3

5. Conclusion

References

1. Hillman, N. W., Fryar, A. H., Crespin-Trujillo, V. (2017) *Evaluating the Impact of Performance Funding in Ohio and Tennessee* **American Educational Research Journal**