Thomas W. Pullum

Statistical Models to Assess Data Quality in DHS Surveys

Abstract:

Indicators to assess the quality of demographic data have typically been calculated from distributions or tables, typically originating with census data. They have been limited in number and scope and outside of a framework for statistical modeling. With survey data, such as that collected by The Demographic and Health Surveys Program (DHS), a wide range of individual-level indicators of data quality can be constructed with methods that allow for multivariate analysis. This paper uses approximately 50 binary indicators on topics such as nonresponse; potential omission of children who died; age incompleteness, inconsistency, heaping, and displacement; and process indicators such as implausibly short interviews. The responses are analyzed according to interviewer ID, characteristics of the respondent, and when possible, characteristics of the interviewer. Associations among indicators can be described and differences between surveys can be tested. The focus is on the statistical models but illustrative results are also included.

Extended Abstract

The proposed paper will be a major extension of a DHS report published in September 2018 (i.e. earlier this month), referred to as MR24:

Pullum, Thomas W., Christina Juan, Nizam Khan, and Sarah Staveteig. 2018. The Effect of Interviewer Characteristics on Data Quality in DHS Surveys. DHS Methodological Reports No. 24. Rockville, Maryland, USA: ICF. https://www.dhsprogram.com/pubs/pdf/MR24/MR24.pdf

It is likely that co-authors, probably some of the co-authors of MR24, will be added, but at this time I cannot be sure who they will be.

Here is the abstract for MR24:

As part of a continuous effort to maintain and improve the quality of data in DHS surveys, this report examines whether variation in 25 indicators of data quality, in 15 recent DHS surveys, can be attributed to interviewers and their characteristics. The analysis is based on interviewer ID codes that appear at several points in DHS data files, and information about the interviewers obtained in a Fieldworker Survey that is now a standard component of all DHS surveys. All of the data files are publicly available.

The 25 indicators are in three broad categories: nonresponse and refusals; reported age at death of young children; and ages and dates. The third includes five subgroups or domains: incompleteness of age, which usually takes the form of a missing month of birth; inconsistency between age in the household survey and age in the individual surveys of women or men; heaping on ages that end in 0 or 5; displacement of age across boundaries for eligibility; and a new indirect indicator of over-dispersion of children's age derived from flagging of the height-for-age and weight-for-age scores. All indicators are defined at the level of the individual, with outcome "1" for a problematic or potentially problematic response, and otherwise either "0" or "Not Applicable". Because the outcomes are binary, they can be easily analyzed with logit regression and related versions of generalized linear models. Combinations of indicators and surveys are judged to be problematic if the level or prevalence of the outcome "1" is relatively far from an

acceptable level and there is highly significant variation in the outcome across interviewers. Many such combinations are identified, with systematic in-depth investigation of several examples. It is found that when there is a high degree of variation across interviewers, in terms of a data quality indicator, the bulk of that variation can often be traced to a handful of interviewers on the same team or on different teams.

To investigate the potential effect of the covariates in the Fieldworker Survey, similar indicators are pooled and all the surveys are pooled. There are exceptions, but it is generally found that interviewers who are older and better educated have lower levels of problematic outcomes. Prior experience with a DHS survey or with other surveys is often statistically significant, and often—but not always—in the direction of better quality data. There is concern when previous experience may lead to worse, rather than better, data.

The most important limitation is that interviewer assignments are almost always to just one or two geographic regions within a country, and the quality of the data they collect is confounded with potentially relevant characteristics of the regions and the composition of potentially relevant characteristics of the respondents. For example, the respondents' level of education is associated with the accuracy of their stated age, and interviewers assigned to a region with a low level of education cannot be expected to obtain the same quality of responses as interviewers who are assigned to other regions.

Further analysis is planned that will include characteristics of the respondents along with those of the interviewer, and possible statistical interactions that reflect the social distance between interviewers and respondents. The methods and findings of this study are relevant to ongoing efforts to improve the training of interviewers and the monitoring of fieldwork.

An earlier DHS report on fieldwork effects is AS19:

Johnson, K., M. Grant, S. Khan, Z. Moore, A. Armstrong, and Z. Sa. 2009. *Fieldwork-Related Factors and Data Quality in the Demographic and Health Surveys Program*. DHS Analytical Studies No. 19. Calverton, Maryland, USA: ICF Macro. http://dhsprogram.com/pubs/pdf/AS19/AS19.pdf.

A recent (September 6, 2018) non-DHS article of relevance is the following:

Amos, Mark. 2018. Interviewer effects on patterns of nonresponse: Evaluating the impact on the reasons for contraceptive nonuse in the Indonesia and the Philippines DHS. Demographic Research, Vol 39 Article 14. DOI: 10.4054/DemRes.2018.39.14.

DHS is developing a complete package to assess potential data quality problems in any DHS survey after the files are complete, to complement the tracking of indicators during fieldwork and potentially to feed back into improvements in fieldwork.

The model under development is alluded to in the last paragraph of the abstract for MR24. It will have the following components, although not all of them are available in all surveys:

Y: A set of binary respondent-level outcomes that identify problems, such as incompleteness of age, or potential problems, such as age ending in 0 or 5. MR24 included 25 indicators of type Y. That list will be expanded to include anthropometry, nutrition, maternal mortality, and other topics for which refusals or incompleteness or other problematic responses are possible.

X: A small set of individual-level covariates that may be related to non-response, such as place and region of residence, age, level of education, recall interval. Such variables are included in the Amos article but not in MR24.

Z: Interview process indicators, such as length of interview, number of visits, position of the interview—early or late—in the sequence of interviews for the interviewer or within the cluster. The process indicators are intermediate, in the structure of the model, and mediate the effect of the interviewer variables on the outcome variables. Some such variables were included in AS19 but not MR24.

ID: Interviewer id code, treated as a categorical variable. There is strong evidence that it is risky to treat this as a random effect, because MR24 finds that there are typically a small number of interviewers who are outliers, violating an assumption of normality. Random effect (multi-level) models will be included, however, if/when possible. Such variables were included in MR24.

FW: Characteristics of the interviewers, when there is a distinct Fieldworker Survey (especially age, sex, education, and previous survey experience), or characteristics that can usually be inferred from the id code, such as team ID and supervisor ID. The Fieldworker Surveys are recent and standard but only about 20 are available at this time. FW variables were included in MR24.

SD: When variables of type FW are present, it may be possible to construct indicators of the social distance between the respondent and the interviewer, such as differences in age or education. Interactions such as these have virtually never been checked with DHS. (Differences in language or ethnicity would also be relevant, but due to the design of DHS surveys they should be very rare.)

The statistical model will be a variant of logit regression, similar to that used in MR24 (and described in Appendix 2 of MR24).

The full model can only be applied to the relatively small number of surveys for which full sets of all the types of variables are available. Several of the DHS surveys that have included the Fieldworker survey were Malaria or AIDS Indicator Surveys that had truncated birth histories and omitted the anthropometry variables (and in general have a reduced set of potential Y variables). Some of these surveys have other problems with the ID codes—for example, Afghanistan and Mozambique—because the transition to actually using interviewer ID codes as part of data quality analysis is relatively recent. At most about ten surveys will be available for a complete analysis. It is expected that the empirical findings in the PAA presentation would be limited to an in-depth analysis of a single illustrative survey, and that survey would be selected because of substantial evidence of reporting problems. In that sense, the example will not be representative of all DHS surveys but will have an atypically high level of data quality problems.