SQUAT Survey 2013-14

Documentation for Cleaned dataset

Household vs. individual: The dataset contains information both at the household level and at the individual level. The household roster on page 2 of the questionnaire collects data for all members of the household. After this section, we randomly selected one respondent by age and sex with whom we could conduct the rest of the interview. The remaining sections of the questionnaire were conducted with the selected respondent, for whom the variable selected = 1.

All individuals in a household have the same values for variables after the household roster. Therefore, if you would like to do analysis at the household level or just for the selected respondent, then you will need to restrict the dataset to those for which selected=1. If you would like to do analysis at the individual level, you can use the dataset as is. The variable hh_sno gives the serial number code for the member of the household.

Men vs. women: The men's and women's questionnaires are exactly the same except that only the women's questionnaires contained sections M and N (while the men's did not), and only the men's questionnaires contained sections P and Q (while the women's did not). The dataset merges both types of questionnaires into one dataset. So only female respondents would have answered M and N and only male respondents would have answered P and Q.

Labels: All variable and values labels have been labeled following the survey schedule. Variable names follow question numbers in the survey schedule (available online).

Missing Values: Right now, they are marked as "." in the dataset. Later on, they would be classified as valid skip (skipped because the section or the question was to be skipped, or didn't apply) and as invalid missing (missing because information that should have been asked was not found in the schedule).

"Don't know" and "Others": "Don't know" or "Can't say" responses have been labeled "9999", following the questionnaire. If the respondent gave any "Other" responses, they have been provided in a separate variable, just below the main question variable.

Calculated variables: Some variables in the dataset have been calculated based on the responses to other variables. A partial list of such variables includes the following. But in general, if the variable is not in the questionnaire, but is in the dataset, then you can assume that it has been calculated by us.

id	household id, this is unique
hh_sno	individual serial number within household
e12_A to e12_EE	preferences for assets if toilet owned
e13_A to e13_EE	preferences for assets if toilet not owned
asset_count	asset index
selected	equals 1 if respondent chosen for interview
pit_size	volume of pit in cubic feet

Known Issues: While cleaning data, we encountered the following issues.

1. Surveyor codes are incorrect for 10 households.
2. Month codes are incorrect for 5 households.

3. For section F’s questions, some questions have zero as a response. It should be treated as an invalid skip. This is less than 30 in all cases in section F.

We welcome feedback on the dataset, and if you have any questions, feel free to get in touch with us at aashish@riceinstitute.org.