

World Bank Round 3 - Dataset Description

This document provides an overview of the organization of the dataset from the World Bank Round 3 COVID survey. There are 5,200 unique households reached in World Bank Round 3. The dataset includes both the raw data and the defined indicators.

Responded in R3 only	1,460 unique households
Responded in R1 and R3	654 unique households
Responded in R2 and R3	2,018 unique households
Responded in R1, R2, and R3	1,068 unique households
Total observations	5,200 unique households

Variables

The variables in this dataset easily cross-reference to the questionnaire. To ensure that we have organized naming conventions, we named and labeled variables in the dataset according to the questionnaire.

Variable Names

There are three main types of variables: questionnaire items, finished indicators, and identifiers. Each questionnaire item follows the following convention: **module_description_r#**. In the dataset, finished indicators follow the same naming convention with the addition of a suffix for type of indicator (proportion, mean, or ratio): **module_description_type_r#**.

- The **_module** prefix refers to an abbreviation of the module in the questionnaire
 - Demo = Demographic information
 - Mig = Migration
 - Con = Consumption
 - Lab = Labor and Income
 - Agr = Agriculture
 - Rel = Relief
 - Hea = Health
- The **_description** characterizes the content of the variable
- The **_type** suffix denotes the estimation command to be used in indicator analysis: **_prop** (for binary variables to indicate proportions), **_mean** (for averages), and **_ratio** (for ratios). Raw questionnaire items do not have a **_type** suffix

- The **_r#** suffix denotes round. We merged in the variables and indicators which overlapped between rounds 1, 2, and 3. To delineate these, we added the suffixes **_r1**, **_r2**, and **_r3** to different variables respectively.

Variable Labels

The variable labels correspond to the numbers in the questionnaire linked above. For indicators, they are descriptive and correspond to previously shared indicator definitions and output.

Other Important Variables

- **sampling_frame** indicates which sampling frame the household was from
- **module_complete_r3** indicates whether or not all questions were fully completed in the module or not. This can help identify where the partially complete surveys ended.
- **fully_complete_r3** indicates whether or not the responses we have for the particular household represent a situation in which the respondent answered the entire survey or whether or not the survey was partially completed due to an emergency/call drop
- **r3** is a binary variable that indicates whether or not the household was reached in R3 ONLY
- **r2_r3** is a binary variable that indicates whether or not the household was reached in BOTH round 2 and round 3
- **r1_r3** is a binary variable that indicates whether or not the household was reached in BOTH round 1 and round 3
- **r1_r2_r3** is a binary variable that indicates whether or not the household was reached in all three rounds

The demographic variables most relevant to round 3 are:

- demo_hh_size
- demo_ag_hh
- demo_hoh_gender_r3
- demo_ag_resp_gender_r3
- demo_nonag_resp_gender_r3
- demo_resp_age
- demo_fem_hoh_edu
- demo_caste
- demo_religion
- demo_hoh_gender_r3
- demo_shg_status
- rel_ration_card_prop_r3
- rel_ration_card_type_r3
- demo_asset_cycle
- demo_asset_refrigerator
- demo_asset_stove
- demo_asset_pressure_cook

- demo_asset_tv
- demo_asset_fan
- demo_asset_almirah
- demo_asset_furniture

The variables to indicate completeness of each section in round 3 are:

- demo2_complete_r3
- agr_complete_r3
- mig_complete_r3
- con_complete_r3
- lab_complete_r3
- rel_complete_r3
- hea_complete_r3
- fully_complete_r3

Indicator Analysis

Missing Values

Values are coded as missing if the question was not asked to the respondent (either because it was not relevant, or because the call-dropped and the survey was only partially completed after exhausting attempts at reaching the household).

Please note that the variable `agr_pmkisan_amt_mean` has 1,333 missing values for households who are eligible for PM Kisan due to an error in the survey form which skipped the question for certain respondents. We conducted a corrections survey on October 3-5, 2020 in order to collect as much missing data as possible. We were able to collect enough data to make reasonable cross-state comparisons of this indicator, but are still missing some values.

Additionally, note that the variable `agr_sell_location_prevyr_r3` only includes 632 observations, even though we expected roughly double this sample size. This question was mistakenly skipped for households that were added in round 3. In addition, the question was skipped for households that did not intend to market their Kharif harvest, or who were undecided about where to sell.

Topcoding

Indicators where type = _mean are topcoded at the 95th percentile. Because many of these indicators are estimated as a mean of household percent change, they have a natural lower bound at -1 (100%), and therefore are not trimmed on the left tail. The questionnaire items that underlie the indicators retain all outliers.

Strata, Primary Sampling Units, & Weights

To correctly account for the sampling strategy employed in this survey, we recommend analyzing this data in Stata using `svyset` data and `svy` prefixed estimation commands. The `svyset` command requires a stratum identifier, a primary sampling unit identifier to account for clustering, and a probability weight. We used the following `svyset` command:

```
svyset psu [pw=weight_hh], strata(strata_id) singleunit(scaled)
```

The primary sampling unit variable (`psu`) is included in this dataset, and is unique within states and strata. As we mentioned in the technical note, the four datasets that provided the phone numbers for this survey had different sampling strategies. The included `strata_id` contains the correct stratum identifier for each state/sample subset of the data.

The variable `weight_hh` is a post-stratified weight that is scaled to state-level marginal totals of caste and religion categories from the 2011 population census. We used a “raking to margins” process to generate these weights from a base weight. The base weight is taken from the source dataset where such a weight exists (AP, Rajasthan, Jharkhand), and is constant for states where the source dataset was not a probability sample (Bihar, MP, UP). In certain states availability of baseline covariates also allowed us to perform adjustments for noncoverage and nonresponse (AP, Bihar, MP, UP).

Within states, the post-stratified weights attempt to correct any bias that might result from imbalance along caste or religious lines. In addition, because post-stratification forces the sum of weights to equal statewise population totals, the raked weights correct for the fact that we have much larger samples in some states than others and allow each state sample to be pooled together for overall estimates.