

COVID-19 Shocks in Rural India (Round 1)

Dataset Description

This document provides an overview of the organization of the dataset from the World Bank Round 1 COVID survey. The dataset includes both the raw data, and the defined indicators. There are 4,550 observations and 234 variables.

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**. In the dataset, finished indicators follow immediately after the questionnaire item(s) used to create them, and follow the same naming convention with the addition of a suffix for type of indicator (proportion, mean, or ratio): **module_description_type**.

- The **_module** prefix refers to an abbreviation of the module in the questionnaire
 - Geo = Geographic information (Module 0)
 - Demo = Demographic information (Module 0)
 - Mig = Migration (Module 1)
 - Con = Consumption (Module 3)
 - Lab = Labor and Income (Module 2)
 - Agr = Agriculture (Module 4)
 - Rel = Relief (Module 5)
 - Hea = Health (Module 6)
- 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

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.



Indicator Analysis

SHRIDS

IDinsight conducted a spatial merge based on the GPS coordinates collected from each household when our surveyors conducted in-person interviews. We were unable to obtain SHRIDS for 70 of the 4550 observations in this dataset.

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 all 7 attempts at reaching the household).

Topcoding


Indicators where type = `_mean` are topcoded at the 95th percentile. Because most 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 preliminary 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.

This dataset contains two sets of weights. Either can be assigned to `weight_hh`. The first is the original sample weight (if it existed) from the source dataset, saved under the variable name `weight_base`. The second is a post-stratified weight, which is the base weight 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, and they are saved as `raked_wgt`. Within states, the raked 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 statewide population totals, the raked weights correct for the fact that we have much larger samples in some states than others.

There were 26 observations for which we did not have appropriate weights data on, which have been dropped from this dataset. Therefore, this dataset has a total of 4,550 observations.