The construction of the PIP dataset

Michail Moatsos

31 Mai 2022

This document explains how I constructed the PIP dataset using the script GetPIPDistro.R. PIP dataset consists of the underlying distributions behind the global poverty statistics reported by the World Bank. This dataset can be used to dive deeper in the statistics and replicate and investigate the World Bank results.

Note: a workstation with more than 64 GB of RAM would be useful to hold all the fetched data in its memory without issues. Some memory optimization would improve this naturally. Currently the peak memory used is above 70GB.

1 Fetching the PIP data in a few steps

The method with which we get the household survey (HHS) data from PIP is via its API (for the commands to fetch the data see document PIP_Structure_and_how_the_API_works.pdf). I fetch the data at various points across the distribution and test for inconsistencies (mainly monotonicity issues, see document PIP_Issues.pdf for a discussion on this; and proper correspondance of the requested poverty line with the returned poverty line, see below). I am fetching the interpolated version of the data, and I am then flagging the interpolated and the original HHS as such by comparing them with one slice of original data obtain from the API with the flag fill_gaps=FALSE (or by simply not including the flag in the API request).

1.1 Decide the points to sample the distributions

We wish to get a copy of the underlying distributions from PIP, with minimum loss of information. This is why we use a high sampling frequency. This means that we use many points across the distribution especially at the low end (lower than \$5/day or \$60/day). The sampling interval is at 0.001 of a dollar below 5 dollars, 0.01 of a dollar between 5 and 60, 0.1 of a dollar between 60 and 150, 0.5 of a dollar between 150 and 1400, 5 dollars between 1400 and 3000, 10 dollars between 3000 and 10000. 10000 is the maximum poverty line allowed by the PIP API to be requested, which means that several distributions from high income countries will not reach 100%. The sampling vector is called IdealPLs in the scripts. In total the data are being fetch at 14920 distinct poverty lines.

1.2 Request the data and perform an initial check

Next step is to make a basic test comparing the poverty line requested in the command with the poverty line that is available in the returned dataset. If the two are not identical the particular slice of data is dropped and the data are requested again (up to 8 times). Every 100 requested poverty lines the data are saved on a file. These files will then be read in, checked and exported accordingly.

1.3 Integrity check

Each saved file with the HHS data is read in and an integrity check is performed. First, the test described above is performed once again. Second, entries that do not have the proper PPP value are also dropped. Other tests that were envisaged are not performed as were deemed unnecessary (such as that some columns must be numeric or convertible to numeric without warnings, interpolated values must be between the

benchmark values used in the interpolation, and coverage type must be from the range included in HHS). Third, the all saved datasets are combined into one big dataset.

1.4 Missing Poverty Lines

Any poverty lines found missing after the above procedure are requested again and tested with the same steps as above. The end product is a complete dataset covering all country-years at all 14920 requested poverty lines.

1.5 Country level export

In the same script I export all unique household surveys (original and interpolated) in two versions. One containing all the data that have been downloaed, meaning at all 14920 poverty lines. And another, more compact version, only at the poverty lines where a change in the headcount ratio takes place (see PIP_Structure_and_how_the_API_works.pdf for the definition of this and all other variables). This is because the other points carry no information, and the complete distribution can be re-structured without them.

1.6 Regional Level

The above procedure is repeated for the 7 global regions that the World Bank is using.