Notes on Reweighting for the High Frequency Phone Survey in the Philippines (HFPS 2020)

The outbreak of the COVID-19 pandemic has posed great challenges to the traditional practice of data collection, i.e., face-to-face interviews. Recently, a more commonly adopted strategy for collecting household survey data is through phone surveys, which do not require face-to-face interactions. One concern with the phone surveys though, is the lack of national representativeness. Presumably, people who could be more easily reached by phone should have very different characteristics from people with no phone. For example, it is likely that households who own a phone are wealthier than those without. Additionally, households with a phone installed are more likely to reside in urban areas with better infrastructure, whereas households with no phone are more likely to be located in remote/rural areas. Therefore, phone surveys only represent a certain group of households with particular characteristics, thereby failing to be nationally representative.

To address the above-mentioned caveats with phone surveys, we develop a reweighting procedure in which - by calibrating the phone survey against a reference survey that is known to be nationally representative ex ante - allows one to re-adjust the phone survey and make it nationally representative.

In this report, we will describe the reweighting procedure used for the high frequency phone survey conducted in the Philippines this year. We first discuss the "ingredients" needed for implementing the reweighting exercise, i.e., the required data and variables. Then we discuss the three-step procedure for creating weights for the Philippines. Along the way, we will discuss some details that need to be noted for future reweighting practice.

Inputs for reweighting

- Surveys Needed
  - 1) **Reference survey that is nationally representative**. For the Philippines study, we used the survey conducted in 2018, i.e., Family Income and Expenditure Survey 2018 (FIES 2018)
  - 2) **High Frequency Phone Survey** conducted in 2020
- Key Variables Needed: **time-invariant variables (listed below)**
  - Our goal is to make the phone survey resemble the distribution of the nationally representative survey as much as possible. To achieve this goal, we need to compare variables that are time-invariant between the two surveys. If these variables are close enough across the two surveys, we can safely conclude that the phone survey has resembled the reference survey quite well, or, the reweighting has been implemented successfully.
  - In the Philippines reweighting procedure, we used the following time-invariant variables as targets to be matched across surveys:
    - **household size**
    - **household size squared**
    - **dependent share**
    - **urban/rural shares**
    - **district-level population sums**
    - **highest educational attainment of household heads**
    - **the age of household head**
  - From both surveys, we also need the initial weights created before data collection. These weights serve as a starting point for weight adjustment (household weight – **WT,** population weight – **popweight**).

Three-Step Reweighting Procedure

Propensity Score Weighting

1. Once we have chosen a reference survey against which we could reweigh the phone survey, we can append the two datasets. Generate a variable named **"append"**, which takes the value of 1 if an observation is from the phone survey, and takes the value of 0 if it comes from the reference survey.
2. Utilizing logit regressions, put the variable "append" on the left-hand side and regress this dependent variable on a series of variables that are correlated with the respondent's likelihood of being reached by phone. In the Philippines study, the regressors we used in the logit regression were:
    i. **household size**
    ii. **household size squared**
    iii. **dependent share**
    iv. **urban/rural information**
3. Divide the appended data set into five quintiles based on the predicted probability.
4. Compute the **quintile-level sum** of predicted probability **for the reference and phone surveys**, respectively.
5. Compute **the sum** of predicted probability **for both the reference and phone surveys**, respectively.
6. Divide the quintile-level sum by the survey-level sum of predicted probability for both surveys.
7. Divide the quintile-to-total ratio from the reference survey by the quintile-to-total ratio from the phone survey, and obtain a new ratio which we name as "**coefficient**".
8. Generate a new household weight by multiplying the initial household weight from the phone survey, WT, by the coefficient: $\boldsymbol{WT_{psm} = WT \times coefficient}$.
9. Generate a new population weight by multiplying the initial population weight from the phone survey, "popweight", by the coefficient: $\boldsymbol{popweight_{psm} = popweight \times coefficient.}$

Post-Stratification on district-level population sums

- While the propensity-score-matching-based procedure, by overweighing the group of people that were hard to be reached by phone, makes the phone survey closer to "being nationally representative"; however, the district-level population in the phone survey may still differ from the reference survey to a great extent. At this stage, we implement a procedure named post-stratification in order to exactly match **the district-level population sums** between the reference and phone surveys.
- The steps for executing post-stratification are listed as follows:
    1. Create **district-level population sums** applying the newest weights (original weights for the reference and PSM-based weights for the phone survey) for **both the reference and the phone surveys,** respectively.
    2. Create a coefficient by dividing the district-level population sum from the reference survey by the counterpart in the phone survey, for all districts, and name this ratio as **"coefficient2"**
    3. Generate a new household weight by multiplying the PSM-based weight from the last step by this coefficient, namely,

$$\boldsymbol{WT_{post} = WT_{psm} \times coefficient2}$$
$$\boldsymbol{popweight_{post} = popweight_{psm} \times coefficient2}$$

Maxentropy

- **Maxentropy** was the last command we used in the reweighting process for the Philippines high-frequency phone survey, which serves as a powerful tool that helps align the means of time-invariant variables in an **exact** manner.
- The variables below were included into the reweighting procedure for the Philippines phone survey, the means of which were matched exactly between surveys after the execution of **maxentropy**.
    - **Household size**
    - **Household size squared**
    - **Dependent share**
    - **Urban/rural dummy**
    - **District-level population sums (eight regions out of seventeen regions. See footnote.)**
    - **Household head highest educational attainment (five levels out of six. See footnote.)**
    - **Household head age**
- Sometimes, incorporating all variables to the maxentropy procedure may result in non-convergence errors. One way to circumvent the issue is to firstly sort the district-level population totals from the reference survey, and only include the top few most populous districts into the maxentropy process. Starting from a full set of district-level population shares, one can incrementally drop the least populous district in each attempt until the maxentropy algorithm converges.
- The maxentropy command usually generates much smaller weights in magnitude compared to the initial weights. While directly applying the maxentropy weights to the phone survey could generate the same means of target variables as in the reference survey, it cannot replicate the sums of population. To deal with this issue, we scaled up the maxentropy weights by a coefficient, which was computed by using the population sum from **district 3** in the reference survey to be divided by the counterpart from the phone survey. In this manner, we were able to match district-level sums between the reference and phone surveys quite closely in magnitude. As per expectation, as can be seen in Table 2, the population sums of district 3 for the two surveys are identical after the scale-up process.

Weights Performance

In this section, we show that the weights we constructed for the Philippines perform well in matching target variables between the reference survey (FIES 2018) and the weighted phone survey.

Table 1 displays the means of target variables and some asset variables from the two surveys. Target variables that are highlighted in green were identical when applying the constructed weights to the phone survey. Some variables in yellow were not identical though, which is very likely due to their time-varying property. Examples include the ownership of stereo, refrigerator, air conditioner, oven, motorcycle, and certain types of walls.

Table 2 displays the district-level totals for both the reference survey and the weighted phone survey. The eight districts included into the maxentropy process are matched relatively well. Nevertheless, as per expectation, districts that were not included into the procedure (highlighted in green) were not matching perfectly, but this is the maximal number of districts (that are most populous) we could include. Overall, the weights are doing a good job in aligning the target variables and the population sums of the most populous regions across surveys.

Table 1 Target Variable Means

	FIES 2018		High Frequency Phone Survey 2020			
Variable	**Obs**	**Mean**	**Obs**	**Mean**	**Min**	**Max**
tv	147,717	0.82	9,448	0.84	0	1
cd	147,717	0.32	9,448	0.32	0	1
stereo	147,717	0.17	9,448	0.30	0	1
ref	147,717	0.46	9,448	0.54	0	1
wash	147,717	0.44	9,448	0.48	0	1
aircon	147,717	0.14	9,448	0.23	0	1
oven	147,717	0.16	9,448	0.50	0	1
motorcycle	147,717	0.34	9,448	0.39	0	1
wall1	147,717	0.74	9,448	0.40	0	1
wall2	147,717	0.11	9,448	0.14	0	1
wall3	147,717	0.15	9,448	0.46	0	1
hhsize	147,717	4.46	9,448	4.46	1	25
depend	147,717	0.26	9,448	0.26	0	0.96
urban	147,717	0.52	9,448	0.52	0	1
hhsize2	147,717	24.28	9,448	24.28	1	625
head_age	147,717	50.26	9,448	50.26	18	100
highest_grade1	147,717	0.02	9,448	0.02	0	1
highest_grade2	147,717	0.17	9,448	0.17	0	1
highest_grade3	147,717	0.17	9,448	0.17	0	1
highest_grade4	147,717	0.11	9,448	0.11	0	1
highest_grade5	147,717	0.28	9,448	0.28	0	1
highest_grade6	147,717	0.24	9,448	0.24	0	1

Table 2 District-level Population Totals

Reference		**Maxentropy**	
I - Ilocos Region	27049332.69	I - Ilocos Region	27090460.13
II - Cagayan Valley	18152008.58	II - Cagayan Valley	264723.0819
III - Central Luzon	**64353743.26**	**III - Central Luzon**	**64353743.38**
IVA- CALABARZON	82142579.77	IVA- CALABARZON	81908859.81
V - Bicol Region	34440872.63	V - Bicol Region	34493238.25
VI - Western Visayas	41881876.88	VI - Western Visayas	41945555.93
VII - Central Visayas	43187203.03	VII - Central Visayas	43252867.03
VIII - Eastern Visayas	25603240.82	VIII - Eastern Visayas	343411.3685
IX - Zamboanga Peninsula	21128216.99	IX - Zamboanga Peninsula	313583.5912
X - Northern Mindanao	26697722.37	X - Northern Mindanao	26738315.04
XI - Davao Region	26481322.3	XI - Davao Region	26521585.71
XII - SOCCSKSARGEN	25707173.16	XII - SOCCSKSARGEN	25709797.04
NCR	71772612.45	NCR	71812998.88
CAR	9625651.64	CAR	135230.2016

ARMM	26281278.59	ARMM	26321238.01
Carga	15203828.02	Carga	239729.6865
IVB - MIMAROPA	16204725.57	IVB - MIMAROPA	349902.899