WPS6373

Policy Research Working Paper 6373

Weight Calculations for Panel Surveys with Sub-Sampling and Split-off Tracking

Kristen Himelein

The World Bank
Development Research Group
Poverty and Inequality Team
February 2013

Policy Research Working Paper 6373

Abstract

The Living Standards Measurement Study—Integrated Surveys on Agriculture project collects agricultural and livelihood data in seven countries in Sub-Saharan Africa. In order to maintain representativeness as much as possible over multiple rounds of data collection, a sub-sample of households are selected to have members that have left the household tracked and interviewed in their new location with their new household members. Since the sub-sampling occurs at the level of the household but tracking occurs at the level of the individual, a number of issues arise with the correct calculation for the sub-sampling and attrition corrections. This paper is based on the panel weight calculations for the initial rounds of the Integrated Surveys on Agriculture surveys in Uganda and Tanzania, and describes the methodology used for calculating the weight components related to sub-sampling, tracking, and attrition, as well as the criteria used for trimming and post-stratification. It also addresses complications resulting from members previously classified as having attrited from the sample returning in later rounds.

This paper is a product of the Poverty and Inequality Team, Development Research Group. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at khimelein@worldbank.org.

Weight Calculations for Panel Surveys with Sub-Sampling and Split-off Tracking

Kristen Himelein[1]

Development Research Group, World Bank

1818 H Street NW, Washington DC, 20433, USA

Keywords: survey weights, panel surveys, attrition, tracking surveys

JEL classification: C83

Sector: POV

[1] This paper was written as part of the Living Standards Measurement Study – Integrated Surveys on Agriculture project and benefited from comments from Peter Lynn, Steve Herringa, Stephanie Eckman, Kathleen Beegle, Jonathan Kastelic, and Raka Banerjee, as well as background research by Anthony Tamasuza.

Introduction

This note serves as a guide to the calculation of weights for multiple waves of a household panel survey based on the experience of the Living Standards Measurement Study – Integrated Surveys on Agriculture (LSMS-ISA) project, located within the Development Research Group of the World Bank. The LSMS-ISA project is a multi-year project funded primarily by a grant from the Bill and Melinda Gates Foundation to collect panel data on households in Sub-Saharan Africa. The project is currently active in seven countries: Ethiopia, Malawi, Mali, Niger, Nigeria, Tanzania, and Uganda.

Although there is some variation between individual country projects, the main objective of the LSMS-ISA project is to provide public-use datasets to track income dynamics and changes in socioeconomic well-being with multiple rounds of data collection, while also providing as representative as possible a cross-section of households at each round. In the LSMS-ISA surveys, the initial rounds of all surveys were two-stage stratified cluster samples. Census enumeration areas (EAs) were selected in the first stage with probability proportional to size, then a household listing operation was conducted from which the second stage households were selected. As with all panel surveys, the LSMS-ISA surveys are only considered completely representative of the target population at the time of the selection of round 1. In subsequent rounds, the level of representativeness decays as the panel ages and there is sample attrition.[2] To minimize these effects, the LSMS-ISA surveys track both households that have changed location as well as a sub-sample of household members that have left households, surveying them in subsequent rounds in their new households. This methodology minimizes attrition by reducing the loss of shifted households, and allows new units to enter into the sample. With perfect implementation, the introduction of new households would match the natural changes in the overall cross-sectional population. In this case, the only loss of representativeness for the follow-up rounds would be from new households being formed by members not living in a particular country at the time of the round 1 survey.[3]

This paper is primarily based on the experience of calculating the round 2 weights in the LSMS-ISA surveys from Tanzania and Uganda, and the round 3 weights in Uganda. The methodology described in this note builds upon published documentation from established panel surveys, such as the Panel Study of Income Dynamics [PSID], conducted since 1968 by the Institute for Social Research at the University of Michigan, and the British Household Panel Survey [BHPS], conducted since 1991 by the Institute for Social and Economic Research at the University of Essex (Gouskova et al, 2008, Taylor, 2010). Both the PSID and the BHPS are nationally-representative panel surveys in the USA and the UK respectively.

Calculating Weights for Round 2 of a Panel Survey

The methodology for calculating weights for a panel survey is developed in the following eight steps:

1) Begin with the base weights (i.e. those calculated for round 1 of the survey) for round 2, or the 'shadow weights' for subsequent rounds (see section on the calculation of panel weights subsequent to round 2 for details on the calculation of shadow weights);
2) incorporate the probability of sub-selection of a round 1 unit into the round 2 sample;
3) incorporate the probability of sub-selection into tracking (if applicable);
4) derive fair-share factors for all household composition changes;
5) pool the weights in (1), (2) and (3) together;

[2] It is also possible that the survey loses representativeness through changes in the population resulting from in-migration, although this is not believed to be a major threat in the LSMS-ISA survey context.

[3] Exceptions to the general methodology described above include a rural-only sample design for the Ethiopia Rural Socio-Economic Survey, which also includes additional second-stage stratification, and not tracking split-off household members in the panel component of the Nigerian General Household Survey.

6) derive attrition-adjusted weights for all individuals, including split-off [4] households, then aggregate these weights to the level of the round 2 household;
7) trim these weights;
8) post-stratify the pooled weights to known population totals.

1) Base weights from round 1

The panel weight calculations are based on the household weights from round 1 of the survey (W_1). The round 1 weights incorporate a selection weight factor equal to the reciprocal of the probability of selection of each unit into the round 1 sample, as well as any non-response correction or post-stratification adjustments. These weights are nonzero for all round 1 respondents and form the base of the round 2 calculations.

$$W_1 = W_{round\ 1}$$

2) Probability of EA selection into round 2

In many panel surveys, it is often not possible, either for reasons of cost or logistics, to re-survey all responding round 1 households in round 2. It is therefore necessary to choose a sub-sample. The sub-sample can be chosen as a probability sample of all eligible round 1 households or enumeration areas, or households in certain regions can be disproportionately subsampled compared to their representation in round 1. If, for example, the capital is determined to be the area of highest variation or the most interest analytically, all responding households in round 1 can be selected into round 2 in the capital, and then only a sub-sample of enumeration areas in other regions.

The probability of selection for each household in stratum *h* is:

$$p_1 = \frac{m_h}{M_h}$$

where *m* is the number of households sub-sampled into round 2 in stratum *h* and *M* is the total number of households in the round 1 sample in stratum *h*. [5] For the round 1 households in the hypothetical capital, which were retained for round 2 with certainty, p_1 would equal one. (Note that if whole EAs are selected for round 2, then $p_1 = \frac{l_h}{L_h}$, where l_h is the number of EAs sub-sampled in stratum *h* and L_h is the total number of EAs in the round 1 sample in stratum *h*.)

If all round 1 households are included in the sample for round 2, then p_1 is one for all households.

3) Probability of household selection into tracking

In panel surveys, some households from round 1 will evolve into more than one household in round 2. That is, the set of household members from round 1 will no longer all be residing together by round 2. For the purposes of administration in the LSMS-ISA surveys, one of the round 2 households will be

[4] For the purposes of this note, 'parent' generally refers to the household found at the same location as the previous round of data collection, and 'split-off' refers to new households entering the sample through an individual originally resident in a parent household during a previous round. Note that this difference between 'parent' and 'split' is purely administrative.

[5] For the simplification of notation the subscript *h*, representing each stratum *h*, is generally dropped in the subsequent formulas despite the fact that the LSMS-ISA surveys all have stratified designs.

designated as the 'parent' household and the others as 'split-off' households.[6] For example, consider a husband and wife cohabiting in round 1, but who are now living in separate households in round 2. The parent household will typically be the one occupying the same dwelling as in round 1 or the household in a dwelling nearest the location of the round 1 dwelling. The other household(s) would be designated as a split-off household(s). The split-off households are then, ideally, tracked and interviewed in their new location. Often the designation between the two is somewhat arbitrary and one of convenience for field work.

Again, due to cost and logistics considerations, it may not be possible to track all split-off households. In this case, eligible movers are stratified and a sub-sample is selected for tracking. It is necessary to integrate this selection into the weight calculations. The rules on which movers to track need to be established (and well-enforced) based on the expected analytical primary uses of the data.

For example, in the Tanzania National Panel Survey (TZNPS), all regular adult household members were deemed as eligible tracking targets. This excluded children under the age of 15, servants, and guests. In the Ugandan National Panel Survey (UNPS), two households within each EA were randomly assigned to have all split-offs tracked, regardless of the age of the member who resided in a new household. Since this selection was done *ex ante* at the statistics bureau headquarters, it was possible that some households selected for tracking did not have any members who formed a split-off, or that no one from the household was located in round 2.

In the TZNPS, there are no additional calculations necessary at this step because the tracking rules were defined at the individual level and therefore all households were eligible for tracking. In the UNPS, the sub-sampling of households complicates the weighting calculations. Ideally, the parent household would receive a tracking weight of 1, because it is selected into the sample with certainty. The weight for split-off households that are subsampled to be followed would include a weight factor to account for the probability of selection into tracking. For the example of an EA with 10 households, this tracking probability would be 2/10. The appropriate expansion factor for each split–off that was tracked would then be $\left(\frac{2}{10}\right)^{-1}$ or 5.

The uncertainty arises, however, from the fact that the difference between a parent and a split-off is a somewhat arbitrary distinction, driven mostly by location. As noted above, among the set of households in round 2 with members from the same round 1 household, the parent household is typically the one whose dwelling is in the same location as in round 1. This would mean if one child remained at the original location and the other five family members moved to another location, the child would be the parent household and the other five members the split-off.[7] If only the child re-located, the five original members would be the parent and the child the split-off. The designation becomes even more arbitrary if there are no members present in the original dwelling. In this case, field supervisors could choose to designate the parent household based on the number of original members, the location of the household head, or some other criteria.

Since the designation is to a great extent arbitrary, there should be no mathematical difference in the probability of selection between the parent and the split-off household. The probabilities of selection are therefore pooled and averaged over all households originating from a single original household. Consider

[6] Note that 'parent' household does not literally refer to familial relationships, as the split-offs may themselves contain parents of members in the 'parent' household.

[7] This also assumes that children are eligible tracking targets, as they are in the case of the UNPS. If the same case had arisen in Tanzania, where children under 15 are not eligible tracking targets, the remaining child would not be counted. Since all eligible household members had relocated, the entire household would be considered to have relocated. The new location would be considered the parent household and there would be no split-offs in this case.

the following examples. If a split-off household is not selected into tracking, the parent household is followed with certainty. The probability of selection (p_2 in the equation below) for this household is one. Similarly, if the household is selected into tracking but did not split, the parent household is followed with certainty. The probability of selection for this household is also one.

The situation becomes more complex when the household selected into tracking has split. In that case, the probabilities of selection must be pooled and averaged. It may be more intuitive to think of this in terms of expansion factors, which are the reciprocal of the probability of selection. If a household from an EA with 10 households is selected into tracking and one member splits from the original household, the parent household would just represent itself, and therefore have an expansion factor of one. Since there are only two original households selected out of the ten possible, any split-offs from the selected household would represent the potential split-offs of five households in the EA. Since the definition of parent and split-off is arbitrary, however, the expansion factors (one and five respectively) are averaged over two households, giving each a value of three. The probability of selection would then be the reciprocal or $\frac{1}{3}$.

Similarly, if a household from an EA with 10 households is selected into tracking and splits into two additional households, the parent household would still have an expansion factor of one and each split-off five. In this case, it would be averaged over three households, giving each a value of $\frac{11}{3}$. The probability would then be $\frac{3}{11}$ for each household. Additionally, the calculations remain the same regardless of whether the split-off household is found and interviewed or if it becomes part of the attrition calculations, although in the case of attrition, it is necessary to make assumptions about the number of households formed by the lost members.

The generalized formula for the probability of selection at this step is:

$$p_2 = \begin{cases} 1 & \text{if not selected for tracking, or selected but no split} - \text{offs} \\ \dfrac{(z+1)}{1+z\left(\frac{q}{2}\right)} & \text{otherwise} \end{cases}$$

where z is the number of new households entering the survey from a given parent household (as multiple split-off members can form either a single new or multiple new households) and q is the number of households in the EA in round 1.

It may seem somewhat counterintuitive that the sum of the conditional probabilities of selection into tracking for the parent and split-off households is not equal to one, since they stemmed from the same initial household. Since these probabilities total less than one, when the inverse is taken, the weight will increase by some number greater than one. When this step is incorporated into the final weight calculations, it will increase the total sum of the weights in such a way as to mirror the natural population growth of new households created between survey rounds.

4) Fair share correction

Since the follow-up rules for the panel survey include interviewing split-off members in their new households along with all the current members of that household, the weights must allow for the incorporation of those who now live with original sample members. For example, if a young man living with his parents in round 1 has formed a new household in round 2 and resides with his wife and newborn child, the wife and infant will be incorporated into the survey and thus require a probability of selection. Such corrections, known as fair share corrections, are routinely used to distribute weight to new sample

members in panel surveys. See Rendtel and Harms (2009) for a discussion of several different methods of weight correction.

Therefore, for any given household with new members, there are two ways that a household can become part of the survey: either by being selected initially for round 1 (and during the subsequent rounds of sub-selection), or by receiving a member that came from a household that was selected for round 1 (and during the subsequent rounds of sub-selection). In order to properly account for the household's current probability of selection, it would be necessary to calculate the exact probability of selection for each new member brought into the household, and adjust the weights accordingly. This adjustment is necessary since households receiving members have higher probabilities of selection (and therefore necessarily lower weights) because the household could have been selected into the survey in multiple ways. The precise calculation, however, would be impossible in all but the rarest of circumstances.[8]

Since it is not possible to compute round 1 probabilities of selection for every member of each sample household in a subsequent round, simplifying assumptions are necessary to continue with the weight calculations. The first simplifying assumption is that the new members to the survey arrived together from one other household. This would be the situation in most common cases, for example, a man and woman marrying and setting up a new household, or an older relative moving in with adult children. In certain cases, however, arriving members come from more than one household. Therefore, assuming only two source households slightly underestimates the probability of selection (and therefore over-estimates the weights) for those cases where members actually came from more than two households. However, as long as the incidence of these cases is believed to be relatively rare, any resulting bias should be negligible.

The second necessary simplifying assumption is that the arriving members have the same probability of selection, on average, as those original members living in the household to which they come. This would not be true on a case-by-case basis but would be true in the aggregate. With these simplifying assumptions, the household probability of selection is increased by a factor of 2 for all households that have new members arriving from other households, whether the receiving household is a parent or split-off, and regardless of if they have been selected for tracking. (Note that children born since the previous round of data collection was completed are not taken as new members from this perspective because they could not have been selected in round 1.) This takes into account the fact that they could have been selected in two ways, and assumes that the probability of selection is equal.

$$k = \begin{cases} 2 & \text{if any new members} \\ 1 & \text{otherwise} \end{cases}$$

A limitation of the panel methodology used in the LSMS-ISA surveys is that the represented round 2 population is not completely representative, as it does not include those solely from outside the original universe of selection who formed households. In this case, it would not include households completely formed by those leaving an institutionalized population, or immigrants from outside the study country. Inclusion of such households would necessitate refreshing the sample with new households. However, in cases where this type of in-migration into the household population is relatively rare, the represented population is close enough to the actual population to permit the desired estimates.

5) Pooling

[8] One example of a rare circumstance where the exact calculation would be possible would be if two (or more) eligible members of existing separate panel households merged into a single household, for example if the children of two panel households in the same village married and established a new household. In this case the probability of selection for both the husband and wife would be known, and the exact probability could be calculated.

At this point, the first four steps encompass the calculations to calculate the panel weights in the absence of attrition. The round 2 household weight would be:

$$W_2 = W_1 * (p_1 * p_2 * k)^{-1}$$

6) Attrition correction factor

Household panel survey weights need to address the problem of attrition, i.e. where survey observations (household or individual members) are selected for re-interview but cannot be located or refuse to be interviewed. The methodology used to adjust weights for attrition in the LSMS-ISA surveys follows Rosenbaum & Rubin (1984), which has also been adopted in the PSID; see for example, Gouskova et al (2008). Predicted response probabilities from a logistic regression model based on the covariates are used to form the weighting classes or cells.

Due to the somewhat unusual design of the panel, the attrition correction in the case of the LSMS-ISA surveys needs to take into account two distinct sources of attrition: entire households that are not found, and split-off individuals that are selected for tracking but not found. The two potential options for the calculations given these two conditions are (1) to treat the split-off households as household heads and perform the calculations at the level of the household, or (2) to treat the households that are not found as individuals and perform the calculations at the individual level. The first option is problematic if the characteristics of household heads are dissimilar to the characteristics of split-offs. The LSMS-ISA surveys have generally found that the groups most likely to attrite from the sample are young people, generally but not always male, often residing in urban areas. Therefore for the LSMS-ISA surveys, the second methodology is employed.

To obtain the attrition adjustment factor, the probability that a sample household was successfully re-interviewed in round 2 is modeled with the linear logistic model at the level of the individual. A binary response variable is created by coding the response disposition for eligible household members that are not interviewed in round 2 as zero, and household members that are interviewed as one. The eligibility requirements vary between surveys (for example, excluding children under the age of 15 in the Tanzania LSMS-ISA calculations). In all surveys, the dead are excluded from the calculations because they cannot be tracked. Members who moved abroad are generally excluded from tracking but remain eligible, as they may return in later rounds.

These calculations use a logistic response propensity model, including the household and individual characteristics measured in the previous round as covariates. The details of this model will vary, perhaps dramatically in other types of surveys or those implemented in different contexts. General model building techniques should be applied to determine the most appropriate set of covariates for a given dataset. For the purposes of illustration, the following round 1 covariates were among those included in the LSMS-ISA survey attrition calculations:

- being a split-off member targeted for tracking
- demographic characteristics (age, gender, marital status, etc.)
- residence of one or both parents in household
- years of education
- current school attendance
- labor force participation
- household consumption
- household assets
- household size

- household livelihood characteristics (agriculture, livestock, non-farm enterprises)
- financial characteristics (receiving transfer income, savings, etc.)
- dwelling characteristics (ownership, improved roof/walls/floor)
- household mobile phone ownership
- geographic characteristics (rural / urban status, district of residence)

The first characteristic, i.e. whether the individual is a member of the original household or a tracking target, is particularly important in cases where the implementation of the tracking methodology is of questionable quality. The described methodology thus far assumes perfect implementation of the survey plan. However, in some cases, the field teams may not search for all eligible tracking cases. In an ideal case, it would be possible to know which cases were attempted and construct a separate model at this step to incorporate those probabilities of selection. For example, perhaps the tracking team decided only to look for those split-off members that they thought would be easy to locate. If the analyst knows which households were considered easy and therefore actually targeted, whether or not they were located, the attrition calculations can be done separately for those targeted and those not targeted, as the most important determining factor of attrition in those cases would be the classification as easy. If the explicit division was not recorded, but the tracking team can inform the analyst roughly which criteria they used in making the determination about whether the split-off members would be easy to find, the designation can be modeled and the division predicted. In the absence of any information, the split-off indicator variable is the best way to capture this element of attrition.

The estimated logistic model is used to obtain a predicted probability of response for each household member in the round 2 survey.[9] These response probabilities are then aggregated to the household level (by calculating the mean). Then, using the household-level predicted response probabilities as the ranking variable, all households are ranked into 10 equal groups (deciles). An attrition adjustment factor (*ac*) is then defined as the reciprocal of the mean empirical response rate for the household-level propensity score decile.

Incorporating the attrition correction, the adjusted weights would be:

$$W_3 = W_2 * (ac)$$

7) Trimming

Complex weight calculations have the potential to produce outlier weights, which increase the standard errors of estimates. A common practice is therefore to 'trim' the weights at this stage to eliminate the outlier weights (see Little et al, 1997). Trimming introduces a small amount of bias into the estimates, but allows estimates to be more efficient. Common values for trimming range between one and five percentage points at the top and bottom of the distribution, and the LSMS-ISA weights are trimmed below the 2nd percentile and above the 98th percentile. Those weights identified as outliers on the top of the distribution are replaced with the highest value of a weight below the cut-off. For example, with a two-percent trim, weights in the 99th and 100th percentiles would be replaced with the highest weight

[9] In some cases, values of round 1 variables may be missing. These values must be imputed using multivariate regression or logistic regression techniques. There are many suitable imputation command packages available in statistical analysis software, such as the multiple imputation (*mi*) commands in Stata or (*proc mi*) in SAS. Note that the resulting values from these imputations do not necessarily need to become part of the published dataset, and in fact the LSMS-ISA surveys leave missing values as missing to be treated at the discretion of the end user. The imputations are necessary only for the weight calculations so that there are no missing values in the propensity score model and values can be predicted for all observations.

value in the 98th percentile. Similarly, weights in the 1st and 2nd percentile would be replaced with the lowest value in the 3rd percentile.

After trimming W_3, we have the adjusted weights W_4.

8) Post-stratification

To reduce the overall standard errors and weight the population totals up to the known population figures, a post-stratification correction (W_{ps}) is applied. This correction also reduces overall standard errors (see Little et al, 1997). The level of disaggregation for the post-stratification correction should be at the lowest reliable level available from an auxiliary data source, regardless of the level of representativeness for which the survey was designed. For example, if reliable census projections are available for the sub-regional level, these should be used even if the sample was designed only to be representative at the regional level. Note that close attention should be paid to the reliability of the outside data source underlying the post-stratification calculations. With proper design and implementation, the survey population estimates should be close to the actual population estimates. Post-stratification should therefore be seen more as a fine-tuning adjustment rather than a major realignment. If the weights are adjusted using poor quality auxiliary information, there is the possibility of reducing precision or introducing bias into the estimates.

In order to calculate the post-stratification adjustment for the LSMS-ISA surveys, census projections of household populations at the regional–urban/rural level are generally used. These projections are provided by the national statistics bureaus. The adjustment is done at the household level since households are the units of analysis which are sampled and for which the weights are calculated. If the census projections give the population totals in terms of individuals, the average household size is calculated from the dataset for each category in the disaggregated data and the totals calculated in terms of households. Within each category the weighted total number of observations is calculated and the census projection is divided by this number for the corresponding category. This correction is then applied to all the observations in the category. For example, if the weighted total of households in the capital city is calculated from the survey to be 157,304, but the census projections give a total of 193,484, the post-stratification adjustment factor would be 1.23. The weight of each household in the capital city would then be multiplied by 1.23.

The final panel weight would then be:

$$W_{final} = W_4 * W_{ps}$$

Weight Calculations for Rounds Subsequent to Round 2

The weights for subsequent rounds of panel data follow the same steps as round 2 with one important difference. In round 2, each household member was classifiable into one of four categories: a re-interviewed household member (observed in round 1), a new household member (new in round 2), a round 1 member who has died, or an attriter (observed in round 1 but not re-interviewed in round 2). In round 3, there would be an additional category of returning household member, which would apply to those members that were present in round 1, not found and thus classified as attriters in round 2, and then present in round 3. In the course of the calculation of round 2 weights, an attrition adjustment was applied to compensate for the departure of these members. As these members have now returned, simply using the weights from the previous round as the base weights for the panel calculations would overestimate the

weights for these households as the attrition calculations are now incorrect. Therefore, in order to have accurate base weights for the calculation of round 3, round 2 weights must be re-calculated using the third round re-interview information. Returning members should be marked as present in round 2, and the attrition correction re-calculated based on this information. These re-calculated 'shadow' weights from round 2 become the base weights for the round three calculations.

Note that when the round 3 attrition calculations are done, the necessary data from round 2 will be missing for returning household members. These missing values will either have to be recaptured using the data from the 1st and 3rd rounds, or imputed. For example, if an adult household member reports having the same number of years of completed education in rounds 1 and 3, it is probably safe to assume they had the same number of years of education in round 2. If the member lives in a household with a traditional roof in rounds 1 and 3, he can be assumed to also have had a traditional roof in round 2. This is slightly different from the case of years of education because while the years of education could not have increased and then decreased for the missing year, it is possible that the returning member lived in a house with an improved roof while he was absent from the original household. This distinction is unimportant, however, because we are not trying to impute the values for the returning member's life while they were missing, only what it would have been had they remained in the household. This means that the dataset compiled to construct the shadow round 2 weights has no real analytical meaning beyond these calculations.

Other Weights

For each round of the panel, there are *2n-1* sets of weights that can be calculated from the data. In round 1, there is only one set of weights - the cross sectional weights for that round. In round 2, there are three sets of weights: one set of cross sectional weights for each round and then the panel weights to be used for combined analysis. By the third round, there are seven sets of weights. The first three are cross-sectional weights, one for each round. There would also be the combined panel weights (to be used when all three rounds of data are analyzed), a first-and-second only set, a second-and-third only set, and a first-and-third only set. In most cases, it is not necessary to calculate all possible combinations of weights, only those which are necessary for the given analysis. Note also that the calculation of different combinations of weights becomes more complicated in later rounds of the panel, as the shadow weights must take into account the proper combination of absent and returning members in order to form an unbiased base for the calculations.

Using Panel Weights

Regardless of which set of weights are appropriate, all analysis using data from a stratified cluster sample should take into account the complex design features of the data to correctly estimate standard errors. The can be done using the *svy* package in Stata software or the *PROC SURVEY* commands in SAS. The cluster and stratification identifier variables should be based on the initial selection in round 1, even if households are no longer physically located in the same cluster. For example, all households found in cluster 1 in stratum 1 in round 1 of the survey should use the cluster identifier 1 and stratum identifier 1 in analysis in subsequent rounds, even if they have now physically relocated to cluster 8 in stratum 3. This is also true of split-off households, which should be classified as part of the cluster where the source member was surveyed in round 1, even if they have never lived in that cluster as a complete household.[10]

[10] In terms of using the later rounds of a panel survey in sample design for other data collection efforts, some caution should be used with the intracluster correlation coefficient (icc). The icc for later rounds of the panel survey represents the correlation amongst persons or households based on their cluster location at the time the original sample was selected. As time passes, members may have shifted out of the original cluster location physically but retain the original identifiers for the purpose of correctly estimating standard errors in panel analysis. There would

most likely be an increase in the variation, particularly for any variable that could be correlated with location and therefore decrease the icc. This could lead to an underestimation of the design effects and potentially an underestimation of sample size requirements for any new surveys.

References

Gouskova, E, Heeringa, S, McGonagle, K, Schoeni, R, and Stafford, F. (2008) Panel study of income dynamics revised longitudinal weights 1993–2005. Panel Study of Income Dynamics, Technical Series Paper #08-05, Ann Arbor (Available from https://psidonline.isr.umich.edu/Publications/Papers/tsp/2008-05_PSID_Revised_Longitudinal_Weights_1993-2005%20.pdf)

Little, R. J. A., Lewitzky, S, Heeringa, S, Lepkowski, J, and Kessler, R. C. (1997) Assessment of weighting methodology for the National Comorbidity Survey. American Journal of Epidemiology. vol. 146, no. 5: 439-449.

Rendtel, U. and Harms, T. (2009) Weighting and Calibration for Household Panels. In Methodology of Longitudinal Surveys (ed P. Lynn), John Wiley & Sons, Ltd, Chichester, UK. (Available from doi: 10.1002/9780470743874.ch15).

Rosenbaum, P. R., and Rubin, D. B. (1984) Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association. vol 79, no. 387: 516-524.

Taylor, M. F. (ed). with Brice, J, Buck, N, and Prentice-Lane, E. (2010) British Household Panel Survey User Manual Volume A: Introduction, Technical Report and Appendices. Colchester: University of Essex. (Available from https://www.iser.essex.ac.uk/bhps/documentation/vola/vola.html).