8 Construction of replicate weights for variance estimation

8.1 Introduction

The use of the final survey weights described in chapter 7 is sufficient when estimating population parameters. However, calculating the corresponding correct variances or standard errors of these estimators requires replicate weights, which are described here. HFCS sampling involves a variety of complex features, such as stratification, multistage sampling, proportional-to-size sampling in the first stage or sampling without replacement in the second stage. In addition, the design weights are adjusted for nonresponse and poststratification. Ignoring these features in statistical analysis will bias the estimated variances of point estimators. For example, if stratification is ignored, the standard errors will be too large, and if clusters are ignored, the standard errors will be too small. Furthermore, if design weights are ignored, the sampling distributions of the statistics underrepresent the observations with a low selection probability and overrepresent those with a high selection probability (see Kolenikov, 2010).

A problem that occurs frequently when statistical analysis takes into account a complex survey design with all its features is that the mathematical functions of the variance estimators are unknown. Therefore, performing a statistical analysis requires methods developed especially for the purpose of variance estimation. There are two general categories of variance estimation methods: *replicate weight methods* (also called *replication* or *resampling methods*) and *linearization*.[1]

Until recently, literature preferred linearization to replication, as it requires less computational power. However, an important disadvantage of linearization is that data protection regulations prevent the required information necessary for linearization from being provided. One way to avoid the problem that certain information is not available for privacy reasons is to use replicate weights. Since replicate weights consist of many variables and their values are based on information not provided to the user of the dataset – e.g. stratum and primary sampling unit (PSU) variables – it is not possible for the data user to identify a given respondent (see Stata Library, 2016).

Moreover, the linearization method is unsuitable for estimating the variance of nonlinear statistics (medians, quartiles, etc.), as it requires computing derivatives of continuous functions; however, quantile functions, for instance, are discontinuous. Replicate weights, by contrast, are well suited for estimating the variance of such statistics (see Heeringa et al., 2010).

Given the data protection requirements mentioned above and because the HFCS data facilitate in particular the analysis of distributional parameters such as medians and quantiles, we decided that the variance estimation method to be employed for the HFCS should be based on the use of replicate weights.[2] In the following section, we describe how replicate weights were constructed for the HFCS in Austria.

[1] *For a comprehensive overview of variance estimation methods, see Levy and Lemeshow (2008) or Heeringa et al. (2010).*

[2] *In combination with multiple imputations, variance estimation of nonlinear statistics by means of resampling weights is still largely unexplored.*

8.2 Construction of replicate weights

8.2.1 The replication method

The replication method aims to estimate the variance of an estimated population parameter. The idea behind this is to estimate population parameters for individual subsets (so-called replicates) of the sample observations. The variability of these estimated population parameters across all replicates is subsequently calculated, resulting in the desired variance of the estimated population parameter (see Levy and Lemeshow, 2008).

Instead of saving a whole sample for each replicate, it is more practical to vary the final survey weights. For example, instead of removing a sample observation to construct a certain replicate, it can be given a weight of zero in the replicate. Then the weights of the other observations in the same stratum need to be increased to ensure that the totals are unbiased for each replicate *r* (see Kolenikov, 2010). The replicate weights $w_i^{(r)}$ for $r=1,...,R$ are published together with the HFCS dataset.

There are different methods to form such replicates. The three major replication methods used in survey literature are *balanced repeated replication, jackknife repeated replication* and *bootstrap replication.* Although in most cases, the estimators of the variance of all replication methods converge toward one another as the sample size increases, simulation studies have shown that bootstrap and balanced repeated replication are better suited to quantile estimation than jackknife (see Kovar et al., 1988). Finally, as balanced repeated replication works only in designs with exactly two PSUs per stratum, which is not the case in the HFCS in Austria, we decided to use the *(rescaling) bootstrap procedure* proposed by Rao and Wu (1988) and enhanced by Rao et al. (1992). This procedure is also in line with the provisions of the ECB's Household Finance and Consumption Network.

The bootstrap procedure forms replicates based on repeated with-replacement sampling of the PSUs within a stratum. The idea is to mimic the original sampling procedure in order to obtain approximations for the sampling distributions of the relevant statistics.

8.2.2 Sampling error calculation model

To mimic the original sampling procedure, we create a sampling error calculation model that is a simplification (see Heeringa et al., 2010) of the actual complex sample design (see chapter 6).

In the HFCS in Austria, one necessary simplification of the sampling error calculation model compared with the original sampling procedure is to collapse, i.e. merge, strata with one single PSU because the bootstrap procedure requires at least two PSUs per stratum. Due to the specific stratification of the HFCS sample design, single-PSU strata are quite common in the sample: Only one PSU was drawn in 50 out of 185 strata. For the sampling error calculation model, every single-PSU stratum is paired with the geographically nearest stratum to form a single pseudo stratum, taking into account how many PSUs are in this stratum. Aggregation is carried out with the nearest stratum containing a smaller number of PSUs, reducing the frequency of necessary aggregations. Although collapsing the strata produces an upward bias in the estimated variance, this bias is kept as small as possible by collapsing geographically close strata, which keeps the PSUs within one pseudo stratum very homogeneous. In this context it must be pointed out that upward biases of standard errors lead to a loss in statistical power. In general, however, this is more acceptable than downward biases of standard errors, which lead to results that are too often considered statistically significant.

Table 17 shows how stratum size (in terms of the number of PSUs drawn per stratum) changes when the HFCS sampling error calculation model is used instead of the original HFCS sample design: When collapsing strata in the sampling error calculation model, their number decreases from 185 to 135, which means stratification is still very high. Moreover, the mean stratum size increases from 3.3 PSUs to 4.6 PSUs per stratum.

Table 17

Comparison of HFCS design strata and HFCS pseudo strata

	Design strata	Pseudo strata
Number of strata	185.0	135.0
Mean size	3.3	4.6
Median size	2.0	2.0
Minimum size	1.0	2.0
Maximum size	37.0	37.0

Source: HFCS Austria 2014, OeNB.

Note: Stratum size as measured by PSUs drawn per stratum.

Another simplification performed in the HFCS sampling error calculation model in contrast to the original sample design is to assume that sampling variance stems mostly from the first stage of sampling (i.e. the selection of PSUs, and not that of households within each PSU). Therefore, two-stage sampling is reduced to single-stage sampling where all gross sample households within drawn PSUs are selected in the replicate sample.

In addition, all PSUs have the same probability of being selected in the replicate sample. Thus, the sampling error calculation model simplifies sampling by making a PSU's probability of being drawn independent of its size as measured by the number of households.

No further simplifications are required by the sampling error calculation model. The nonresponse and poststratification weight adjustments are implemented in the same way as in the original weighting procedures (see chapter 7), and a finite population correction[3] is performed.

8.2.3 Calibration of replicate weights

The algorithm used to construct the HFCS replicate weights comprises the following steps:

Step 1: Draw m_h PSUs with replacement within each pseudo stratum h.

Step 2: Adjust the final survey weights of the drawn observations to create a new set of replicate weights. In particular, apply the same nonresponse and poststratification weight adjustments (sections 7.2.3 and 7.2.4) as for the final design weights and perform a finite population correction.

Step 3: Repeat steps 1 and 2 R times to obtain $r = 1,...,R$ sets of replicate weights.

In step 1, the number of PSUs m_h drawn in each stratum of size n_h is set to $m_h = n_h - 1$. This decision is taken often in order to ensure the efficiency of the bootstrap estimators without violating the natural parameter ranges (see Kolenikov, 2010).

In step 2, the final survey weights must be adjusted because some PSUs may be duplicates and some may not have been drawn at all. As a consequence, each

[3] *The finite population correction accounts for the reduction in variance that occurs when sampling without replacement from a finite population. This type of sampling is used in the sample design of the second stage of the HFCS in Austria.*

replicate will be biased with respect to the target population and therefore, to obtain the replicate weights, the design weights must be adjusted in the same way they were adjusted when constructing the final survey weights (see chapter 7). In addition, a finite population correction is required, as SSUs are sampled without replacement in the original HFCS sample design (see footnote 3).[4]

Finally, in step 3, the higher the number of replicates *R* is, the more precise the standard error estimates are. We choose $R = 1{,}000$, which lies in the upper bound of the usual recommendations found in literature (see Kolenikov, 2010).

Table 18 shows some descriptive statistics of a selection of HFCS replicate weights. We can see that owing to the homogeneous weighting adjustments, the mean and the total sum of replicate weights remain unchanged. Moreover, compared with the final survey weights in the HFCS, the replicate weights have smaller minimum values, however none are equal to zero. These values correspond to the nonselected PSUs, which, instead of being assigned a weight equal to zero, are assigned a small positive weight in the finite population correction. The fact that the replicate weights also have larger maximum values than the final survey weights can be explained by the weight adjustments that were carried out: As some PSUs are not drawn in the replicates, and in order to obtain the same estimated population sizes as in the original sample, the weights of the observations in the drawn PSUs must be increased.

Table 18

Selected HFCS replicate weights

	Mean	Median	Minimum	Maximum	Total
Final survey weights	1,289	1,207	287	4,360	3,862,526
1st set of replicate weights	1,289	1,040	7	14,374	3,862,526
2nd set of replicate weights	1,289	989	10	11,418	3,862,526
3rd set of replicate weights	1,289	1,023	8	10,852	3,862,526
998th set of replicate weights	1,289	1,104	10	8,369	3,862,526
999th set of replicate weights	1,289	985	6	11,201	3,862,526
1,000th set of replicate weights	1,289	974	7	10,349	3,862,526

Source: HFCS Austria 2014, OeNB.

Note: Statistics refer to successfully interviewed households only.

8.3 Concluding remarks

We constructed 1,000 sets of replicate weights to enable HFCS data users to correctly estimate the standard errors of point estimators in the HFCS. This is necessary because the complex features of the HFCS survey design, which comprises amongst other things stratification, several stages of cluster sampling and weighting adjustments, bias the variance estimators if data users ignore them.

While it is true that correctly calculating the standard errors by using replicate weights requires more computational power than analyzing the data without using

[4] *In the HFCS sample design, PSUs are drawn with replacement, SSUs without. Although the sampling error calculation model ignores the second stage, a finite population correction was performed to allow for the fact that households are not allowed to appear twice in the sample. Finite population correction reduces the bias of a higher variability of replicate weights.*

replicate weights, in practice it is not necessary to use all 1,000 sets of replicate weights for variance estimation. Thus, for example, it is possible to perform variance estimations using fewer replicates more quickly but less precisely. The number of replicates used depends on the type of estimator and the size of the population surveyed. For instance, estimating the means for the total population will, as a rule, require fewer replicates than estimating the medians for specific population subgroups.

See the HFCS User guide (chapter 9) for an explanation of how to use the replicate weights correctly in Stata®.