Institute for Fiscal Studies

Refining the bootstrap methodology for HBAI statistics

IFS Mimeo

Mike Brewer
Olga Gdula
Robert Joyce

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Refining the bootstrap methodology for HBAI statistics

Mike Brewer
University of Essex
Institute for Fiscal Studies

Olga Gdula
University of Oxford

Robert Joyce
Institute for Fiscal Studies

January 2017

Correspondence to: robert_j@ifs.org.uk

This work was funded by the Department for Work and Pensions. The authors are grateful to various people within DWP, ONS and the Department of Finance and Personnel in Northern Ireland, for providing helpful information about details of the FRS sampling design – in particular Fola Ariyibi, Greg Ceely, Dermot Donnelly, Paul March and Salah Merad. They would also like to thank Stephen Jenkins for helpful comments on an earlier draft and Stas Kolenikov for a helpful discussion. FRS and HBAI data were made available by DWP, which bears no responsibility for its analysis or interpretation.

1. Introduction

The Family Resources Survey (FRS) is an annual survey of about 20,000 households in the UK. It is widely used by statisticians and researchers for its detailed information on household incomes, and it underlies the official Households Below Average Income (HBAI) series produced by the Department for Work and Pensions (DWP). HBAI contains high profile statistics about the UK income distribution, including average incomes, measures of income inequality, and poverty rates.

Because the FRS is a sample from the population, any statistics derived from it are only estimates of the statistics for the whole population. It is important to quantify the extent of uncertainty around those estimates arising from sampling variation. The FRS, like many household surveys, employs a complex sampling design. Standard techniques which do not account for these complexities will therefore yield imperfect estimates of the extent of uncertainty.

There is more than one type of approach to quantifying uncertainty around sample estimates of a population statistic. The most common alternatives to bootstraps are 'analytical', formula-based methods. A limitation of those is that they tend to be less flexible: they are typically based on assumptions or approximations that are not appropriate if the sampling design or the statistic in question is too complex, or if the sample is too small. For example, many such approaches are valid only for statistics that are linear functions of the data. This would exclude many of the major statistics used in HBAI, such as median income or relative poverty rates. This was a key reason for the recent shift in the HBAI publication towards using a bootstrap methodology throughout, rather than the previous 'estimating function' approach.[1]

The next step, explored in this paper, is to refine the bootstrap methodology so that it accounts as fully as possible for the relevant complexities of the FRS sampling design - a task for which the flexibility of bootstrap techniques is again an advantage. Here we set out work undertaken for DWP with that aim. Essentially, the key features of the sampling design accounted for are stratified sampling, clustered sampling, the fact that the sample is weighted *ex post* using external information about the characteristics of the population[2], and non-independence of the samples in consecutive years (relevant for statistics such as year-on-year changes in incomes). These are explained more fully in Section 2.

To some extent these features of the survey design should have offsetting effects on the extent of uncertainty (in a way that will vary depending on the statistic of interest). It is therefore theoretically unclear whether, in combination, they increase or decrease the true degree of uncertainty around a given population statistic. This also implies that accounting for only some of these features could lead to less accurate quantification of

[1] See Section 8.4 of DWP (2015a), and DWP (2014).

[2] Technically this is not a feature of the sampling design, as it concerns what happens once the sample has been drawn; but it is an important feature of the HBAI methodology which is relevant for the uncertainty around sample estimates. For brevity, we abstract from this distinction in the rest of this note.

uncertainty than accounting for none. For this reason we show systematically both the individual and cumulative effects of incorporating each of these complexities.

The paper is structured as follows. In Section 2 we outline the FRS sampling design and weighting methodology, and in Section 3 we explain how these are relevant for the quantification of uncertainty. Section 4 describes the bootstrap methodology we use to account for this. Section 5 sets out the results, which are essentially sets of 95% confidence intervals for key HBAI statistics using bootstraps of differing degrees of refinement. Section 6 concludes by highlighting recommendations for the methodology used in future vintages of HBAI.

2. FRS sampling design and grossing methodology

Although the main statistics used in the HBAI series take individuals as the unit of analysis[3], the ultimate sampling units in the FRS are households: a sampling design is used to select a set of households in each year, and an individual features in the survey if he or she is in one of those selected households. Households are selected using different sampling designs for Great Britain and Northern Ireland, which we summarise in turn.

Sampling in Great Britain

For households in Great Britain, a stratified clustered probability sampling design is used. A comprehensive list of postcodes of private households is grouped into over 12,000 postcode sectors. These sectors become what are known as Primary Sampling Units (PSUs). The sampling scheme can broadly be thought of as containing two steps: the sampling of PSUs, and the sampling of households within each selected PSU. In 2013-14 (the most recent year of published data), 1417 PSUs were sampled, and 24 households per selected PSU were in the target sample (though note that this is larger than the actual sample of households, due mainly to non-response).

To select PSUs, Great Britain is split into 27 regions ("major strata"). Within each major stratum, an ordered list of PSUs is constructed using information about PSUs from external (Census-based) sources. PSUs are split into 8 equally-sized bands ranked by the proportion of households in which the Household Reference Person has a National Statistics Socio-Economic Classification 1-3; further subdivided in half based on the proportion of economically active adults aged 16-74, to create 16 groups within each major stratum; and then, within each of those 16 groups, ranked according to the proportion of economically active men aged 16-74 who are unemployed.

The resulting ordered list of PSUs in each major stratum is then split into a discrete number of subgroups (i.e. subgroups contain PSUs adjacent to each other in the list, and hence sharing relatively similar socio-economic characteristics), such that the number of households in each subgroup is approximately equal and the number of subgroups is

[3] For example, median household income is calculated as the household income of the median individual (not the household income of the median household).

the number of PSUs that will be sampled from that major stratum. These subgroups are the "minor strata". One PSU per minor stratum is then randomly sampled. Finally, a random sample of households is selected from each sampled PSU.

The samples from consecutive years are not independent of each other. This is because, within each major stratum, half of the PSUs are retained each year (i.e. PSUs are only sampled afresh from half of the minor strata). All PSUs are retained for exactly two years before being rotated out of the survey. Note, however, that a fresh sample of households is always drawn from every PSU each year, even if the PSU in question was retained from the previous year.

Sampling in Northern Ireland

The FRS sampling design in Northern Ireland is known as a systematic stratified sample design. It is simpler because there are no PSUs, so it is a 1-step design rather than a 2-step design; and because samples are essentially entirely independent[4] from one year to the next.

All eligible private households are put into an ordered list based on the District Council area and electoral ward to which they belong. There are 26 District Council areas and 582 electoral wards. Starting from a random point in the list, every n^{th} address is selected (where 1/n is the proportion of eligible addresses that will be sampled). In 2013–14, 3,600 addresses were sampled out of approximately 680,000 eligible ones.

Ex post weighting of households

Once the samples of households have been selected, an important final step is to weight them. The most basic function served by this step is to ensure that the sample scales up to the size of the overall population (hence why this step is sometimes referred to as "grossing up"). However, the aim is also to ensure that the characteristics of the weighted sample conform to known (or government- or ONS- produced projections of) population characteristics based on external data sources, such as the age, gender and region distribution, the number of dependent children and single parent families, and the number of very high income individuals recorded by administrative tax records. In other words, the aim is to gross up not just to the total population size, but to the total number of people belonging to various (not necessarily mutually exclusive) sub-populations. Weights are derived such that the weighted sample is consistent with a set of "control totals" of the number of people, benefit units or households with certain characteristics in the population as a whole. Full details about the control totals used are in DWP (2015b; see Tables 1 and 2).

[4] The only sense in which the samples drawn in Northern Ireland are not independent from one year to the next is that a household is not included in the sample if it was in the sample in the previous year (to ease respondent burden). This is a very small degree of non-independence and is ignored in the rest of this note.

The distinction that we focus on in this note is *not* the distinction between ex post weighting and no weighting at all. The case of no weighting at all is not the relevant comparison - this would not only have implications for the variability of sample estimates (discussed in the next section), but would also be likely to result in estimates of the income distribution that are systematically biased (for example, Northern Ireland is deliberately under-sampled, and high-income individuals are known to be under-represented in the survey; without any weighting there would be no way to account for this). Rather, we take as given the fact that the data are weighted, but highlight the relevance of the fact that the weighting procedure is an *ex post* one. In other words, the weights are not simply determined in advance for each household based on its characteristics (to correct, for example, for the fact that the proportion of households sampled who are from Northern Ireland is deliberately larger than the proportion of all UK households that are in Northern Ireland). Instead, the weights are calculated *after* the sample of households for that year has been obtained, and a given household's weight will depend on which other households happen to have been sampled that year.

3. Implications for quantifying uncertainty

Here we give a brief, non-technical description of the basic implications of the survey features described above for statistical inference. A fuller treatment is in Deaton (1997).

Stratified sampling

The fact that the samples are stratified means that, by construction, they can be constrained to look similar to the population as whole in terms of the socio-economic or geographic variables used for stratification. Hence stratification increases the precision of sample estimates: it should reduce uncertainty arising from sampling variability and make the true confidence intervals narrower. All else equal, we would expect that ignoring this feature of the sampling design would lead to inference that is too conservative: overstating the true degree of uncertainty around statistics about the income distribution (e.g. producing confidence intervals that are too wide).

Clustered sampling

The 2-stage design of the Great Britain sampling procedure, with the samples of households drawn from "clusters" of nearby households (the PSUs) which were sampled in the first stage, should have the opposite effect. Given that nearby households are likely to have relatively similar income levels (i.e. given that there is likely to be a positive correlation between incomes within clusters), sampling multiple households per cluster provides less independent information about the true population than if all households were sampled from different clusters. For this reason, clustering is often described as akin to reducing the effective sample size. To give a more concrete example: if a relatively rich PSU is randomly selected rather than a PSU of average affluence within its stratum, then a whole cluster of households who tend to have relatively high incomes have been sampled. Hence clustering increases sampling

variation. All else equal, ignoring it will tend to lead one to understate how much uncertainty there is around a population statistic estimated from the sample (e.g. producing confidence intervals that are too narrow).

Ex post weighting

Relative to not weighting the data at all - a scenario that we do *not* explore in this paper - the weighting procedure could increase the variability of sampling statistics due to the variability of the weights (though the effect will depend on the correlation between incomes and the weights). However, not weighting at all would create systematic biases in sample estimates (see Section 2) as well as having implications for their variability. In this note we consider the impact of accounting for the fact that the weights are computed *ex post* rather than *ex ante*: weights are computed after that year's sample has been selected, so that the weighted sample conforms to a set of known or projected population characteristics. In other words, each household's weight is not determined in advance of the sampling, but instead depends on which other households happen to have been sampled that year.

The effect of the ex post nature of the weighting methodology is to ensure that, by construction, the weighted sample always conforms to a set of known population characteristics, regardless of what the unweighted sample looks like. Hence it effectively corrects for some of the random sampling variation in the unweighted data. We would therefore expect the ex post nature of the weighting to reduce the true degree of uncertainty around population statistics estimated from the (weighted) sample. All else equal, ignoring it should lead to inference that is too conservative.

Non-independence of samples in consecutive years

The non-independence of samples in Great Britain in consecutive years, due to the retention of half of the PSUs from one year to the next, is only relevant for inference about statistics that combine more than one year of data. The most prominent of these would be year-on-year changes (e.g. growth in median income or the change in the poverty rate).

For year-on-year changes, non-independence should reduce the degree of sampling variation because some of it is effectively "differenced out". For example, in any one year, random sampling variation means that the set of sampled PSUs might be a set of permanently relatively rich PSUs, biasing estimates of average income upwards. But to the extent that the same PSUs are sampled next year, a roughly similar bias will appear in the next year's estimate of the level of average income. When comparing income levels across the two years, the same bias is present in both years so it will cancel out. To put it another way, with completely independent sampling from one year to the next, one of the sources of uncertainty is that part of any estimated year-on-year change might just be caused by random replacement of one set of PSUs in one year with a

different set of PSUs in the next year (i.e. not comparing 'like with like' across years). Non-independence reduces this source of uncertainty.[5]

All else equal, we would therefore expect that ignoring this feature of the sampling design would lead to inference that is too conservative when estimating year-on-year changes, overstating the true degree of uncertainty (e.g. producing confidence intervals that are too wide).

Non-independence could be relevant for other kinds of statistics too, however, with different implications. For example, estimates of income statistics at the regional level in HBAI are produced by pooling 3 consecutive years of data, and taking the average of the statistic in question across the 3 years. Here the effect of non-independence will be the opposite to above. It means that any random variation in the types of PSUs selected in one year will be partly replicated in the next year too. Rather than being differenced out as in the case of year-on-year changes, we would expect non-independence to magnify the impact of random sampling variation on 3-year averages.

Summary

Overall then, the various complexities of the FRS/HBAI methodology should have different and contrasting impacts on the true width of confidence intervals; and the details may also differ depending on the statistic being calculated. The main objective of this paper is to quantify the importance in practice of these different features, and to assess their net effects when considered together.

4. The bootstrap

The bootstrap is one of many approaches to statistical inference. It is often employed because of its flexibility relative to computationally simpler analytical procedures, which are typically based on assumptions or approximations that are not appropriate if the sampling design or the statistic in question is too complex, or if the sample is too small.

The bootstrap is therefore a natural option in the context of HBAI, which draws on a complex survey design and features a range of complex statistics (including poverty rates using poverty lines calculated from within the sample, average incomes within subgroups defined by income cutoffs calculated within the sample, Gini coefficients, and assorted cross-time changes). There is a growing literature developing analytical alternatives to account for particular complex sampling features for particular types of statistics (recent examples include Preston (2009) and Berger and Priam (2015)). To our knowledge, though, there is no analytical alternative to the bootstrap which can reliably account for all the features outlined above for the full range of HBAI statistics.

[5]Note also that the effect of non-independence in this respet will depend on how correlated incomes are within a PSU. The more highly correlated, the more different PSUs will tend to be from each other in terms of their income levels. Therefore, the more year-to-year variability will be 'differenced out' in the way described above if some PSUs are the same in consecutive years.

There are many thorough textbook treatments of bootstrapping techniques, which we do not repeat here. The basic principle, though, is that we can estimate how variable the sample estimates of a population statistic are by repeatedly approximating (or "simulating") the process which produced the sample data set and sample statistics from the underlying population. The approximation is to treat the sample data set as though it were itself the underlying population, and to resample from that.

Correct bootstraps will therefore incorporate in the simulations the features of the process used to produce the original sample statistics from the population (e.g. stratified or clustered sampling, or ex post grossing using external information). Below we outline how the bootstrap employed in this paper accounts for the features of the HBAI methodology described in Section 2.

Overview of bootstrap methods used in this paper

From a practical point of view it is worth being explicit that we start by simply collapsing the dataset to the household level, because households are the ultimate sampling units in the FRS (see Section 2). All that is required for the bootstrap resampling is a list of household identifiers, plus information on strata and PSU for each household if using that information in the bootstrap (see below). Once each bootstrap resample of households has been drawn, then of course benefit unit- or individual- level data from the FRS can be merged back in before the statistic being bootstrapped is calculated (and, if applicable, before weights are re-computed based partly on benefit unit- and individual-level information – see below).

In order to reflect the differences in sampling design described in Section 2, we apply separate procedures to resample households in Great Britain and Northern Ireland, and then combine the two resamples at each bootstrap replication.

For statistics based on a single year of data, we apply a series of bootstraps which account for progressively more features of the HBAI methodology, outlined below. In all but the last bootstrap, the weights used to calculate the statistic in question at each bootstrap replication are simply held fixed at their original values for each household.

- The 'baseline' bootstrap: simple random sampling of households, not accounting for stratification, clustered sampling or ex post weighting;
- stratification based only on 27 Great Britain regions and 26 Northern Irish District Councils ("major strata");
- stratification based also, as far as possible, on the full set of stratification variables used in the FRS in Great Britain ("minor strata");
- clustered sampling in Great Britain (based on PSUs), as well as stratification;

- ex post weighting of the sample in light of external information, in addition to stratification and clustering.

The second of these specifications, accounting for stratification based on major strata only, involves random resampling of households (with replacement) within each major stratum (i.e. so that the number of resampled households in each Great Britain region and Northern Irish District Council is the same as in the original sample). The weights assigned to each resampled household are simply the original survey weights for those households (as is the case for all the bootstraps described below, except for the last, where we account for ex post weighting of the data).

The next specification essentially mirrors this, but with more finely defined strata in Great Britain. In reality, each minor stratum in Great Britain is represented by (at most) one postcode sector in the FRS sample (see Section 2). For reasons that will become clear when we discuss the incorporation of clustered sampling below, we approximate this by grouping adjacent minor strata together into pairs, and treating the resulting pairs of postcode sectors in the FRS data as if they were the minor strata. Hence we are really using "pseudo-strata" here.[6] (In major strata containing an odd number of minor strata, one of our pseudo-strata is a group of 3 postcode sectors.) For Northern Ireland, no further stratification (beyond District Councils) is accounted for, because data on electoral wards were not available to the research team.

Next we also account for the fact that, within minor strata in Great Britain, clusters of households are sampled: a postcode sector, or PSU, is selected in a first stage, and all sampled households in that minor strata come from that single PSU. Here we have to deal with the fact that the bootstrap cannot perfectly replicate this first stage. Because the FRS sample only contains one cluster (PSU) per minor stratum, resampling clusters from the FRS data within each minor stratum would simply involve resampling the originally sampled cluster. In other words the sample data do not contain enough information for the bootstrap to approximate the true sampling variability from the underlying population (in which each minor strata actually contains many PSUs), and each resample would be the same by definition. Instead, as described above, we define pseudo-strata based on groupings of 2 or 3 adjacent clusters of PSUs in the FRS data, and then resample clusters from the pseudo-strata. In practice, as we show in Section 5, we find that the degree of stratification has little effect on confidence intervals, so the need for this approximation is highly unlikely to matter much.

2 or 3 clusters per (pseudo-) stratum is still a small number and the theoretical justification behind bootstraps is "asymptotic" - that is, it relies on having a large number of units from which to resample. With few units, small sample corrections are required for a robust bootstrap, and without them bootstraps will tend to underestimate the true degree of sampling variability. We follow Kolenikov (2010), who discusses the

[6] On practical grounds we also define pseudo-strata after first partitioning PSUs into two groups based on whether or not they were sampled in the previous year, so that single-year and consecutive-year statistics do not require entirely separate bootstraps. We explain this in more detail below.

literature on this issue and recommends a correction which requires that the number of units resampled is one fewer than the number in the original sample.[7] For example, given that there are either 2 or 3 PSUs in the FRS data in each pseudo-stratum, we resample either 1 or 2 PSUs (with replacement) from each pseudo-stratum. Weights for each household in each resampled PSU are then simply scaled up, by factors of 2 and 3/2 respectively, so that they gross up to the totals that would have applied if the original number of PSUs had been resampled within each pseudo-stratum. (If doing the re-weighting step afterwards - see below - this means that the weights in memory at this stage should simply be 2 or 3/2 for each resample of each household. If not re-weighting, the weights are 2 or 3/2 multiplied by the original survey weight.)

The resulting list of resampled PSUs is then used in a second stage, to draw, at random and with replacement, a sample of households within each resampled PSU, again applying the small sample correction including rescaling of weights (the number of sampled households per PSU is larger than the number of sampled PSUs per pseudo-stratum, but still relatively small). Note that, if in the first stage the same PSU happens to be resampled twice, in the second stage two separate independent samples of households are redrawn from that PSU.

Finally, our richest bootstrap specification for single-year statistics also includes ex post re-weighting of the data. At each replication, once the sample of UK households has been drawn as described above, a fresh set of weights is computed using the same set of control totals used to compute the weights used for the original HBAI sample (rather than holding the weights fixed based on their original values – multiplied by the scaling factors as above - for each resampled household). We use the algorithm set out in Gomulka (1992), which has been implemented in Stata,[8] to compute the weights. Like the CALMAR program used by DWP, it minimises a measure of difference between a set of starting weights and the new weights, subject to the constraint that the new weights are consistent with the specified control totals. The starting weights should simply be the bootstrap weights in memory (i.e. for each household in the original survey, the starting weight should be the number of times it was resampled in the current bootstrap replication, multiplied by its scaling factors as defined in the previous paragraphs). In particular (as pointed out by Kolenikov, 2010), one should not use the original survey weights as the starting weights - even though it may save computational time - as this would introduce dependence between the weights used in different bootstrap replications.

When the statistics of interest involve more than one year of data and the years are consecutive (e.g. annual growth in median income), we also need to account for the non-independence of consecutive-year samples in Great Britain (see Section 3). For simplicity, in the case of year-on-year changes we describe only one bootstrap

[7] This correction belongs to the class of 'rescaling bootstrap' procedures proposed by Rao and Wu (1988). See pages 178-179 of Kolenikov (2010) for details.

[8] We use the Stata command 'reweight2', written by Browne (2012).

procedure, which incorporates this dependence in addition to all the features already described.

Due to the retention of half of the PSUs from one year to the next, one can partition data from any two consecutive years into three groups of PSUs: those which only appear in the first year (group 1), those which appear in both years (group 2), and those which only appear in the second year (group 3).

The resampling of PSUs in the first stage can simply be done separately for each of the three groups: define pseudo-strata based on adjacent PSUs within each of the major strata *within each of the three groups*, and then resample PSUs from each pseudo-stratum with a small sample correction as above. The same resample of PSUs from group 2 is then joined to the group 1 resample (creating a bootstrap resample of PSUs for the first year) and to the group 3 resample (creating a bootstrap resample of PSUs for the second year), mirroring the dependence of consecutive-year samples in the original sampling scheme. With each year's bootstrap sample of PSUs now created, households are then resampled from within each resampled PSU, independently in each year. Re-weighting is then applied separately for each year.

All of our bootstraps use 2000 replications (see below for discussion). 95% confidence intervals are computed using the 2.5th and 97.5th percentiles of the distribution of estimated statistics across all the bootstrap replications. The simple 'percentile method', which just uses those percentiles directly as the lower and upper confidence interval limits respectively, can be incorrect if the sampling distribution is asymmetric (or if it is biased). We therefore use a refinement which is robust to this. It is set out formally in Hansen (2015, pp. 233-234) but, in words, the solution is to subtract (in turn) the 97.5th and 2.5th percentiles of the distribution of estimated statistics from twice the original sample estimate, and to use those as the lower and upper 95% confidence interval limits respectively.

Practical considerations

In choosing the number of bootstrap replications there is of course a trade-off between accuracy (more replications means that 'simulation error' will average out to a greater degree) and computational time. As the aim of this paper was to come to robust and confident conclusions about the relative properties of different bootstraps in order to make recommendations about future methodology, a quite large number of replications was chosen (2000). The precise number of replications is however something that could reasonably be varied depending on time or computational constraints.

Accounting for all of the features of the FRS/HBAI methodology described, and in particular re-weighting, is computationally intensive - but we show in Section 5 that accounting for re-weighting is important. Hence it is worth elaborating a little more on the computational implications.

The procedure needs to be run only once per year. All that is necessary is to save a dataset of weights - each household in the original HBAI sample is listed alongside a

large number of weights, where each weight is the weight that the household in question ended up with after a particular bootstrap replication (hence weights will often be zero, where households were not resampled in a particular replication, and they will often be particularly high where a household was sampled multiple times). Once created and saved, these weights can then be used to compute confidence intervals for any statistic, by simply calculating the statistic repeatedly using each set of weights.

Because of the additional partitioning of PSUs according to the combination of years in which they were sampled, pseudo-strata defined when computing consecutive-year statistics will be slightly different for any given year than if they were defined as described above when simply bootstrapping statistics based on one year of data (or non-consecutive years of data). In Section 5 we show that, in practice, the two different ways of defining pseudo-strata make no appreciable difference to confidence intervals for single-year statistics. Given that both single-year and consecutive-year statistics are always of interest, we therefore recommend on practical grounds simply running one bootstrap each year as though one were interested in consecutive-year statistics. This one bootstrap can then also be used to obtain robust estimates for single-year statistics.

Bootstrapping 3-year averages for statistics at the region or nation level

The official HBAI publication includes statistics at the region or nation level, based on averages over 3 consecutive years of data to ensure sufficient sample sizes. The dependence of consecutive years of data in Great Britain now needs to be accounted for in a 3-year, rather than 2-year, setting. In the interests of time we have not included bootstraps of such statistics in this paper, but the generalisation from 2 years to 3 (or more) years is straightforward conceptually.

Rather than partitioning the sets of PSUs into three groups as in the 2-year case, one would partition the PSUs into (y+1) groups, where y is the number of consecutive years of data being used. For example, in the 3-year case, the groups would be those PSUs appearing in year 1 only, those appearing in years 1 and 2, those appearing in years 2 and 3, and those appearing in year 4 only. One then defines pseudo-strata within each group. For the first and last groups, PSUs will be resampled and then households resampled from the relevant years of data; and the resampled households will be added to the bootstrap samples for the first and last years respectively. For all other groups, PSUs will be resampled, and then households will be resampled from the resampled PSUs twice - from both of the years of data in which those PSUs appear.

Again, in practice the implementation would ultimately be more straightforward. With each new year of data, one simply has two groups of Great Britain PSUs: those that appeared in last year's sample and those that did not (and will therefore appear next year). For the first group, bootstrap resamples of PSUs have already been produced (in last year's bootstrap). One simply needs to resample households in the new year of data for each of those resampled PSUs in each replication. For the second group, one resamples PSUs and then households afresh in each replication. Joining the two sets of resampled households yields the new bootstrap samples for the latest year, which

account properly for the non-independence with last year's sample. These can then simply be appended to the bootstrap samples drawn in all previous years.

5. Results

Table 1 sets out the estimated 95% confidence intervals around key HBAI statistics from the latest year of data, 2013–14, using different bootstraps. All statistics are based on incomes measured before deducting housing costs (BHC). These bootstraps vary in how comprehensively they account for the complexities of the FRS survey design. The least comprehensive 'baseline' bootstrap is on the left : this simply resamples households randomly and independently and recomputes the survey statistic using the original survey weights at each replication: not accounting for stratification, clustered sampling or ex post weighting. For a detailed outline of what the different bootstraps account for, see Section 4. The final column of the Table simply changes the definition of the pseudo-strata to the one that would be used if first separating PSUs into those retained from the previous year (2012–13) and those new in the year in question (2013–14), to see whether this has a material impact (see Section 4).

The important points highlighted by Table 1 are:

- Accounting for stratification (both major and minor strata) does very little to estimated confidence intervals, suggesting that quantitatively the stratified sampling design plays only a minor role in affecting the precision of survey estimates.

- Accounting for the clustered sampling design in Great Britain is quantitatively more important. As expected (see Section 3), it acts to widen the estimated confidence intervals significantly, reflecting the fact that clustered sampling makes survey estimates less precise.

- Accounting for the ex post nature of the weighting regime, by re-computing the weights based on a constant set of control totals at each replication, also makes a notable difference to some of the estimated confidence intervals, with the effect in those cases being – again as expected (see Section 3) - to narrow the confidence intervals.

- The importance of accounting for ex post weighting varies according to the statistic being calculated, to a greater degree than the importance of accounting for stratification or clustered sampling. In particular, the effect of narrowing the confidence interval is particularly large for mean income and the Gini coefficient. This is almost certainly because those statistics are sensitive to incomes towards the very top of the income distribution, and the control totals used to compute weights include explicit totals for the number of very high-income individuals.

- As a result, when compared to the baseline 'naive' bootstrap, the most sophisticated of the bootstraps results in narrower confidence intervals for mean income and the Gini coefficient (because the effect of accounting for the ex post nature of the weighting regime dominates the effect of clustered sampling), but wider confidence intervals for the other statistics (because the effect of clustered sampling dominates).

- The final column shows that changing the definition of pseudo-strata, by first separating PSUs that were and were not in the previous year's sample (as one would do if bootstrapping a statistic based on two consecutive years of data), has essentially no impact on the confidence intervals. This has the practically convenient implication that one can use the same set of bootstrap resamples to calculate confidence intervals for single-year and consecutive-year statistics (rather than requiring entirely separate bootstrap procedures for the two).

Table 2 sets out the estimated confidence intervals around year-on-year changes in key HBAI statistics between 2012–13 and 2013–14. Three different bootstraps are presented for comparison. Again, in the left-hand column the most basic bootstrap is included as a baseline: in each bootstrap replication, households are simply resampled randomly and independently, separately for each year of data, and then the year-on-year change in the statistic is recalculated. In the next column, the bootstrap essentially replicates the procedure used in the last column of Table 1 for each year of data independently – accounting for stratification, clustered sampling and ex post weighting. In the final column, the bootstrap additionally accounts for the non-independence of the samples in consecutive years of data.

The important points highlighted by Table 2 are:

- As for single year statistics, the net effect of accounting for stratification, clustered sampling and ex post weighting is to narrow confidence intervals for the statistics that are sensitive to incomes at the very top of the distribution, and to widen the confidence intervals for the other statistics.

- Accounting for non-independence of consecutive-year samples has only a very modest additional impact on confidence intervals.

Table 1: 95% confidence intervals for single-year statistics (2013–14) using different bootstraps

Statistic	Memo: sample estimates	Bootstrap											
		'Naive': simple random sampling		Major strata		Minor (pseudo) strata		+ clustered sampling		+ ex post reweighting		+ defining strata as for consecutive year bootstrap	
Mean Income (£ p/w)	561.3	546.2	575.7	546.0	575.5	546.6	574.6	533.5	584.3	552.5	570.7	553.1	570.7
Median Income (£ p/w)	453.1	448.0	458.2	448.0	458.2	447.6	457.9	443.3	462.0	445.6	461.8	445.5	461.3
Gini Coefficient	0.342	0.327	0.357	0.327	0.357	0.327	0.356	0.316	0.366	0.334	0.349	0.335	0.349
90/10 Ratio	3.813	3.706	3.905	3.710	3.900	3.714	3.904	3.620	3.956	3.629	3.943	3.641	3.945
Relative poverty rate	0.152	14.6%	15.8%	14.6%	15.8%	14.5%	15.8%	14.1%	16.3%	14.1%	16.4%	14.2%	16.3%
Relative child poverty rate	0.170	15.7%	18.1%	15.8%	18.1%	15.8%	18.2%	14.8%	19.1%	14.9%	19.2%	15.0%	19.1%

Note: Bootstraps use 2000 replications. Incomes measured before deducting housing costs. See text for details of the bootstraps.

Table 2: 95% confidence intervals for year-on-year changes (2012–13 to 2013–14) using different bootstraps

Statistic	Memo: sample estimates	Bootstrap					
		'Naive': simple random sampling		Stratification, clustering and ex post reweighting		+ non-independence of consecutive years	
Mean Income (£ p/w)	1.83%	-1.53%	5.16%	-0.47%	4.15%	-0.36%	4.22%
Median Income (£ p/w)	0.07%	-1.69%	1.75%	-2.38%	2.92%	-2.22%	2.78%
Gini Coefficient	0.006	-0.011	0.025	-0.006	0.014	-0.006	0.014
90/10 Ratio	-0.079	-0.234	0.064	-0.337	0.117	-0.326	0.123
Relative poverty rate (ppts)	-0.2	-1.2	0.7	-1.8	1.4	-1.8	1.2
Relative child poverty rate (ppts)	-0.4	-2.2	1.2	-3.4	2.5	-3.3	2.4

Note: Bootstraps use 2000 replications. Incomes measured before deducting housing costs. See text for details of the bootstraps.

6. Conclusions

The analysis undertaken in this paper suggests that clustered sampling (in Great Britain) and the ex post nature of the weighting regime are the key features of the HBAI methodology to account for when estimating confidence intervals.

These two features have opposite effects on the width of confidence intervals, so the net impact of accounting for both of them depends on which effect dominates. This turns out to vary depending on the statistic in question: confidence intervals sometimes become narrower, and sometimes become wider.

Accounting for clustered sampling but not ex post weighting will tend to result in inference that is too conservative: the degree of uncertainty around, for example, median income or the change in poverty rates will be overstated. Accounting for ex post weighting but not clustered sampling will tend to result in the opposite: uncertainty will be understated. Note also the implication that, because these features have offsetting effects, accounting for just one of them could easily be "worse" than accounting for neither - in the sense that it could return an estimated confidence interval less close to the true one.

Stratification and non-independence of consecutive-year samples turn out to have much smaller impacts on confidence intervals.

References

Berger, Y. G. and Priam, R. (2015), 'A simple variance estimator of change for rotating repeated surveys: an application to the European Union Statistics on Income and Living Conditions household surveys', *Journal of the Royal Statistical Society: Series A (Statistics in Society*).

Browne, J. (2012), 'Reweight2', user-written Stata software (http://www.ifs.org.uk/publications/6270).

Deaton, A. (1997), 'The analysis of household surveys: a microeconometric approach to development policy', Baltimore: John Hopkins University Press (available online at http://documents.worldbank.org/curated/en/1997/07/694690/analysis-household-surveys-microeconometric-approach-development-policy#)

DWP (2014), 'Uncertainty in Family Resources Survey-based analysis' (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/321821/uncertainty-family-resources-survey-based-analysis.pdf).

DWP (2015a), 'Households Below Average Income: An analysis of the income distribution 1994/95 – 2013/14' (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/437246/households-below-average-income-1994-95-to-2013-14.pdf).

DWP (2015b), 'Households Below Average Income (HBAI) Quality and Methodology Information Report – 2013/14' (https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/437251/households-below-average-income-quality-methodology-2013-14.pdf).

Gomulka, J. (1992), 'Grossing-up revisited', in R. Hancock and H. Sutherland (eds), *Microsimulation Models for Public Policy Analysis: New Frontiers*, STICERD Occasional Paper, London: London School of Economics.

Kolenikov, S. (2010), 'Resampling variance estimation for complex survey data', *Stata Journal*, 10(2), pp. 165-199 (http://www.stata-journal.com/article.html?article=st0187).

Hansen, B. (2015), Econometrics, (http://www.ssc.wisc.edu/~bhansen/econometrics/Econometrics.pdf).

Preston, J. (2009), 'Rescaled bootstrap for stratified multistage sampling', *Survey Methodology*, 35(2), pp. 227-234.

Rao, J. and Wu, C. (1988), 'Resampling inference with complex survey data', *Journal of the American Statistical Association* 80: pp. 620-630.