1352.0.55.063

Australian Bureau of Statistics

Research Paper

Estimation for the Household Income and Expenditure Survey

New Issue

Research Paper

Estimation for the Household Income and Expenditure Survey

Sybille McKeown

Statistical Services Branch

Methodology Advisory Committee

18 June 2004, Canberra

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) TUE 6 JUL 2004

ABS Catalogue no. 1352.0.55.063
ISBN 0 642 48167 9

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Ms Sybille McKeown, Statistical Services Branch on Canberra (02) 6252 7311 or email <sybille.mckeown@abs.gov.au>.

ESTIMATION FOR THE HOUSEHOLD INCOME AND EXPENDITURE SURVEY

Sybille McKeown
Statistical Services

EXECUTIVE SUMMARY

Historically, the ABS has collected income and expenditure statistics via the Survey of Income and Housing Costs (SIHC) and the Household Expenditure Survey (HES). For the 2003/2004 financial year, these two surveys were combined to form the Household Income and Expenditure Survey (HIES). The HIES can be viewed as a two-phase survey with the first phase collecting income data and a smaller second phase collecting expenditure data. This paper presents the proposed estimation methodology for HIES, namely two-phase calibration, and discusses the associated issues including the selection of auxiliary variables, dealing with non-response, and variance estimation.

DISCUSSION POINTS FOR MAC

There are specific questions presented in the text for MAC's consideration. However, the main areas for discussion can be summarised as follows.

- Are there limitations to the two-phase calibration methodology that have not been addressed?
- What are MAC's views on the proposed methodology for choosing auxiliary variables?
- What are MAC's views on the proposed methodology for addressing non-response bias?
- Are there any limitations to using delete-a-group jackknife variance estimation for this scenario?

CONTENTS

ESTIMATION FOR THE HOUSEHOLD INCOME AND EXPENDITURE SURVEY

Sybille McKeown
Statistical Services

1. INTRODUCTION

1 The Australian Bureau of Statistics regularly collects data on household income and expenditure. In the past, income data has been collected by the Survey of Income and Housing Costs (SIHC), and expenditure data has been collected by the Household Expenditure Survey (HES). SIHC was run as a supplementary survey to the Labour Force Survey, and HES was run as a special supplementary survey.

2 For the 2003/2004 financial year, these two surveys were combined to form the Household Income and Expenditure Survey (HIES), run as a special supplementary survey. The HIES collects income data (consistent with what was previously collected in SIHC) for all households selected in the sample, and expenditure data (consistent with what was previously collected in HES) for a sub-sample of households. The data is collected over a twelve month period.

3 While the income and expenditure data is being collected together for the 2003/2004 financial year, separate publications for the SIHC and HES components are still required. In previous cycles of HES, basic income statistics were collected and published. In 2003/2004, detailed income will be available for all households in the HIES sample, not just for the households in the HES sample.

4 This presents a number of options for producing income estimates for the HES publication. Income estimates for the HES publication could be based solely on the subsample of households that received the HES questionnaire. Alternatively, the income estimates could be based on all households that received the SIHC questionnaire.

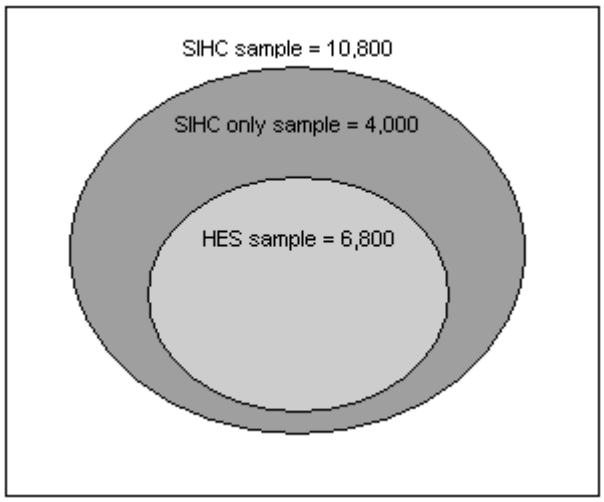
5 Users of the HES data often compare income and expenditure estimates, so maintaining consistency in the relationship between a household's income and expenditure is a requirement of the estimation methodology. Coherency in estimates between ABS publications is also a requirement of the methodology. That is, income estimates in the HES publication should be the same as income estimates in the SIHC publication.

6 Treating HIES as a two-phase survey and using two-phase calibration estimation will enable both requirements to be met. The method allows us to calibrate the second phase sample estimates to match the chosen first phase sample estimates, in

addition to population auxiliary variables. If we use household income as one of these variables, we will maintain consistency in the relationship between a household's income and expenditure, and ensure coherency between estimates in ABS publications.

7 The following diagram shows how the HIES can be viewed as a two-phase sample. The first phase sample consists of approximately 10,800 responding households and collects income and housing data. The second phase sample consists of approximately 6,800 responding households and collects expenditure data. Throughout this paper, the first phase will be referred to as the Survey of Income and Housing Costs (SIHC), and the second phase will be referred to as the Household Expenditure Survey (HES). The households selected in the first phase, but not in the second, will be referred to as the SIHC only sample.

1.1 Pictorial representation of the Household Income and Expenditure Survey sample



8 The Statistical Services Branch is proposing to use two-phase calibration estimation to produce estimates for the HIES. The purpose of this paper is to:

- present the proposed two-phase calibration approach for HIES estimation, including choosing auxiliary variables, adjusting for non-response and variance estimation;
- discuss the inherent assumptions and potential problems with using this approach;
- obtain MAC's input into the validity of the proposed approach.

2. SAMPLE DESIGN AND DATA COLLECTION SUMMARY

9 The HIES was designed to achieve a fully responding sample size of 10,800 households for the SIHC component, and a fully responding subsample of 6,800 households for the HES component. The sample was allocated at the state by part of state (metropolitan areas and non-metropolitan areas) level, to meet relative standard error constraints in line with the design.

10 The SIHC component of the survey collected information on income, housing costs and wealth. The HES component of the survey collected information on the expenditure on goods and services. The majority of data items were collected by an ABS interviewer in a face-to-face interview, both with individuals and with the household. The detailed expenditure items were collected via a self-completion diary which covered a two week period.

11 Households were selected through a multi-stage cluster design. Selected clusters were split such that approximately one third of households in the cluster received only the SIHC component of the survey, and two thirds of households in the cluster received the SIHC and HES components of the survey. Selections were distributed randomly across a twelve month enumeration period in order to capture changes in income and expenditure patterns across a year.

3. ESTIMATION METHODOLOGY

3.1 Literature review

12 In the literature (e.g. Cochran, 1977) two-phase sample designs are generally discussed in the context of taking a large, cheaper first phase sample to enable improved estimates of the variables of interest collected in the second phase. The variables collected in the first phase can be used in a number of ways. For example, they may be used to stratify the second phase sample, or as auxiliary variables in estimation.

13 Dupont (1995) considered different estimation strategies for making use of auxiliary data in a two-phase design. Three possibilities in the relationship between the auxiliary variables available at the population and first phase levels were considered, as were two estimation methodologies (calibration estimation and a regression model approach). Dupont discussed seven possible estimators, the links between them, and methods for variance estimation.

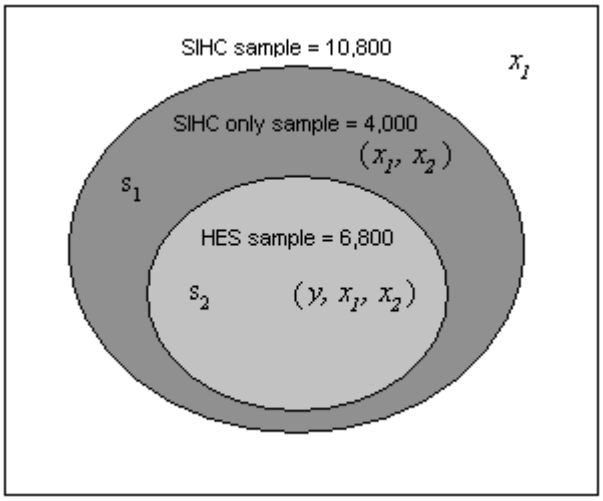
14 Hidiroglou and Särndal (1998) presented a more generalised theory for two-phase sampling with auxiliary information. The authors stated that some of Dupont's strategies could be considered as special cases of their general approach. Two estimation methods, namely two-phase calibration and two-phase generalised regression, were discussed. The authors showed that the generalised regression method can actually be considered as a special case of the two-phase calibration method. The authors showed that gains in variance can be made if there is a strong correlation between the auxiliary variables and the variable of interest.

15 The two-phase calibration method, as discussed in Hidiroglou and Särndal will be described in more detail below, using the following notation.

Population	$U = \{1, \ldots, k, \ldots, N\}$
First phase probability sample	s_1
First phase probability of selection	$\pi_{1k} = P(k \in s_1)$
First phase sampling weight	$w_{1k} = 1/\pi_{1k}$
First phase calibrated weight	$\tilde{w}_{1k}$
Second phase probability sample	s_2
Second phase probability of selection	$\pi_{2k} = P(k \in s_2 \mid s_1)$
Second phase sampling weight	$w_{2k} = 1/\pi_{2k}$
Overall sampling weight	$w_k^* = w_{1k}\, w_{2k}$
Second phase calibrated weighted	$\tilde{w}_k^*$
Auxiliary information	$\mathbf{x}_k = \left(\mathbf{x}_{1k}^{'}, \mathbf{x}_{2k}^{'}\right)$

16 Note that $\mathbf{x}_{1k}$ is available for the entire population and $\mathbf{x}_{2k}$ is available for units selected in the sample only. Typically, we do not know $\mathbf{x}_{1k}$ for each unit, but it is sufficient to know the population total of $\mathbf{x}_1$. The diagram below attempts to consolidate this notation.

3.1 Pictorial representation of the two-phase notation



17 Two-phase calibration involves two successive calibrations to produce the final calibrated weights for the units in the second phase sample. As in single phase calibration, the first calibration goes from the level of the first phase sample to the population, with population auxiliary variables, $\mathbf{x}_{1k}$, as constraints. That is, we want to minimise the distance between the initial Horvitz–Thompson weights and the calibrated weights, subject to the constraint that the estimates of auxiliary variables using the calibrated weights will match known population totals. The initial weight is the first phase sampling weight w_{1k}. This can be expressed as follows:

Minimise the distance function D_1:

$$D_1 = \frac{1}{2} \sum_{k \in s_1} C_{1k} \frac{(\tilde{w}_{1k} - w_{1k})^2}{w_{1k}}$$

(where C_{1k} are pre-specified positive factors) subject to:

$$\sum_{s_1} \tilde{w}_{1k} \mathbf{x}_{1k} = \sum_{U} \mathbf{x}_{1k}$$

The resulting weights are

$$\tilde{w}_{1k} = w_{1k}\, g_{1k}$$

where g_{1k} is the usual g-weight used in the calibration framework.

18 The second calibration goes from the level of the second phase sample to the first phase sample, with weighted auxiliary variables, $\mathbf{x}_k$, in the constraint. The aim of the calibration is to minimise the distance between the initial weight and the final calibrated weight. The initial weight this time is the first phase calibrated weight multiplied by the second phase sample weight. This can be expressed as follows:

Minimise the distance function D_2:

$$D_2 = \frac{1}{2} \sum_{k \in s_2} C_{2k} \frac{(\tilde{w}_k^* - \tilde{w}_{1k}\, w_{2k})^2}{\tilde{w}_{1k}\, w_{2k}}$$

(where C_{2k} are pre-specified positive factors) subject to:

$$\sum_{s_2} \tilde{w}_k^* \mathbf{x}_k = \sum_{s_1} \tilde{w}_{1k} \mathbf{x}_k$$

Denoting the g-weights for the second phase calibration as g_{2k}, then the resulting weights are

$$\begin{aligned} \tilde{w}_k^* &= g_{1k}\, g_{2k}\, w_{1k}\, w_{2k} \\ &= g_k^*\, w_k^* \end{aligned}$$

where

$$g_k^* = g_{1k}\, g_{2k}$$

19 Estevao and Särndal (2002) considered ten cases of auxiliary information for calibration estimation in two-phase sampling. Again, they assumed that $\mathbf{x}_1$ is available for all units in the population and that $\mathbf{x}_2$ is available for all units in the first phase. The ten cases varied in the extent to which the auxiliary data was used. They ranged from the complete use of the data as described in Hidiroglou and Särndal, to not using any auxiliary data at all, which was shown to be equivalent to the double expansion estimator.

20 The authors carried out a simulation study to assess the ten estimators and found that using the complete auxiliary information did not always give the most efficient estimator. However, the authors did suggest that using the complete auxiliary information would be a reasonable approach when there are many variables of interest collected in the second phase.

3.2 Proposed methodology for HIES

21 In addition to obtaining accurate estimates, there are three requirements for the HIES estimation strategy – continuity, consistency and coherence. Users desire continuity in the methodology used for SIHC estimates, specifically, the use of the same auxiliary variables in the calibration. This is to allow comparability with previous SIHC estimates and SIHC estimates in the future when it is run without HES. Further, users are interested in relationships between the income and expenditure of a household, so maintaining the consistency in this relationship is important. Finally, the coherence between estimates presented in the SIHC and HES publications is highly desirable.

22 Using the two-phase calibration methodology presented by Hidiroglou and Särndal (1998), will allow us to achieve all three requirements. We are free to choose the auxiliary variables, so we can ensure that the x_1 are the same as in previous cycles of SIHC. Further, by using variables that are presented as estimates in both the SIHC and HES publications in x_2 (such as income), the requirements of consistency and coherency will be met, due to the constraints in the second calibration. For these reasons, we propose to use the two phase calibration methodology to produce estimates for the HIES 2003/2004.

23 The two-phase methodology also presents us with the opportunity to obtain more precise HES estimates, in comparison with the approach of treating HES as a separate sample (that is, calibrating only to population auxiliary variables and not drawing strength from correlations between income and expenditure). We recognise

that great gains in precision may not eventuate since the population auxiliary variables are being chosen with SIHC requirements in mind, rather than being chosen to give the best HES estimates. This is where the HIES case differs to the two-phase scenario generally discussed in the literature, as first phase estimates are equally as important as second phase estimates.

24 As discussed in Estevao and Särndal (2002), it is not necessary to use the complete auxiliary information in estimation, and in fact, using the complete auxiliary information does not always give the most efficient estimator. However, we do propose to use the complete auxiliary information, for coherency reasons, and also because there are many variables of interest in the second phase. [An interesting direction for future work would be to investigate any changes to efficiency by not using the complete auxiliary information for producing HES estimates.]

25 Existing estimation tools within the ABS can be used to implement the two-phase calibration methodology with relative ease. The GREGWT macro, which performs calibration, is currently used to produce estimates for ABS household surveys. The macro will simply need to be run successively to produce the second phase calibrated weights. The macro also caters for the use of both household and person level auxiliary variables through integrated weighting.

26 Integrated weighting enables the use of group (e.g. household) and person level auxiliary variables by firstly collapsing the person level data to the group level. The GREGWT macro then computes the number of times a group contributes to each constraint (e.g. once to the group level constraints and more than once to the person level constraints). Next, the group weights are calculated to meet the calibration constraints, which are now at the group level. Finally, the group weights are merged back to the person level file.

27 For more details on the GREGWT macro, see Bell (2000).

- Can MAC see any limitations with the two-phase calibration methodology that have not been addressed?

4. SELECTING AUXILIARY VARIABLES

4.1 Literature review

28 Skinner and Nascimento Silva (1997) provided a method for choosing auxiliary variables for generalised regression (GREG) estimation in the presence of non-response. The authors recognised that while the GREG estimator is asymptotically design unbiased under simple random sampling, bias can be introduced through non-response. As such, they proposed a method which chooses a set of auxiliary variables to minimise the mean squared error (MSE), giving the best trade-off between variance and bias caused by non-response.

29 The method assumes that the regression estimator that uses all available auxiliary variables is free from non-response bias, but may have higher variance than other estimators due to the inclusion of all auxiliary variables. A further assumption is that an estimator using a subset of auxiliary variables may include some non-response bias, but possibly a lower variance. The MSE of GREG estimators based on subsets of auxiliary variables, is used to determine which subset gives the best trade-off between variance and bias.

30 The following notation is used:

GREG estimator of the mean of *y* using all auxiliary variables	$\bar{y}_{reg}$
GREG estimator using a subset, *A*, of auxiliary variables	$\bar{y}_{reg,A}$
Estimated variance of the GREG estimate using subset *A*	$\hat{v}\left(\bar{y}_{reg,A}\right)$
Estimated bias of the GREG estimate using subset *A*	$\hat{B}\left(\bar{y}_{reg,A}\right) = \bar{y}_{reg,A} - \bar{y}_{reg}$
Variance of the estimated bias using subset *A*	$\hat{v}\left[\hat{B}\left(\bar{y}_{reg,A}\right)\right]$
MSE of the GREG estimate using subset *A*	$M\hat{S}E\left(\bar{y}_{reg,A}\right)$

31 The following formula is given for estimating the MSE:

$$M\hat{S}E\left(\bar{y}_{reg,A}\right) = \hat{v}\left(\bar{y}_{reg,A}\right) + \hat{B}\left(\bar{y}_{reg,A}\right)^2 - v\left[\hat{B}\left(\bar{y}_{reg,A}\right)\right]$$

32 Using this diagnostic, the authors presented three methods of choosing the subsets A, namely, the Best Subset method, the Forward Selection method, and the Standard Forward Selection Method.

33 The Best Subset Method involved calculating the MSE for every possible subset of auxiliary variables. Theoretically, this was the preferred approach. However, the authors acknowledged that this method would be unsuitable when a large number of auxiliary variables are available.

34 The method of Forward Selection was presented as an alternative to the Best Subset approach. It involved starting with the subset A which contains only a constant term, and adding variables one by one on the basis of the MSE. This continues until no further gains in the MSE can be made by increasing A, in comparison to the MSE of the GREG estimator using only the constant term.

35 The Standard Forward Selection method as implemented in SAS PROC REG (see SAS, 1990) was also considered. It was presented as a simplified method that is easy to apply.

36 Skinner and Nascimento Silva carried out a simulation study to investigate which of the methods provided the best results. The authors found little loss in precision by using the Forward Selection method instead of the Best Subset method. The Standard Forward Selection method displayed no noticeable bias, but had slightly higher MSE when compared with the other methods. This method also yielded smaller subsets a majority of times, which was expected as the Standard Forward Selection method looked only at the impact of adding variables on variance, not the MSE. So any variables required to counter bias would not be included.

37 The simulation study was also presented as support to the assumption that the saturated model is free from non-response bias. In the study, the non-response was artificially created by a model. The variables used in the response model were the same as the variables used in the saturated model for the regression estimator. As such, it was not surprising that the assumption of no bias in the saturated model held.

4.2 Proposed methodology for HIES

38 As discussed, users require that the population auxiliary variables, $\mathbf{x}_1$, are the same as the variables used in calibration for previous cycles of SIHC. However, this may not be methodologically the best option. For example, if the undercoverage and non-response characteristics in HIES are very different to the case under the previous design of SIHC, then there would be a strong argument for changing the $\mathbf{x}_1$. This will be investigated before the $\mathbf{x}_1$ are finalised.

39 There is more room for choice in selecting the auxiliary variables $\mathbf{x}_2$. Clearly, we want to include the variables that are in both the HES and SIHC publications, for the consistency and coherency requirements. But how do we choose other variables that may yield gains in precision?

40 Skinner and Nascimento Silva (1997) provided a useful method for choosing auxiliary variables in the case of single phase regression estimation in the presence of non-response. Hidiroglou and Särndal (1998) showed that the two-phase regression estimator is equivalent to the two-phase calibration estimator in some situations. This suggests that if we modify the Skinner and Nascimento Silva method for the two-phase scenario, then we can use it to choose our auxiliary variables for two-phase calibration estimation.

41 The proposed modification is as follows. First we assume that the $\mathbf{x}_1$ have been chosen and we are only interested in the relationship between $\mathbf{x}_2$ and the variable of interest y (likely to be an aggregate expenditure variable). In terms of the regression estimator we are only interested in the second term in the assumed model, as the first term is already set:

$$E\left(Y_k\right) = \mathbf{x}_{1k}^{\mathrm{T}}\beta_1 + \mathbf{x}_{2k}^{\mathrm{T}}\beta_2$$

42 [An area for future work will be to investigate auxiliary variable selection for the two-phase scenario when both sets of auxiliary variables need to be chosen. As yet, we have been unable to find any papers in the literature that provide a possible theoretical framework for this scenario.]

43 From here we propose to proceed as per the Skinner and Nascimento Silva Forward Selection method described above, because it's easier to implement than the Best Subset method without much loss in precision. This means that we need to determine the maximal set of auxiliary variables to enable the calculation of the saturated model, and then calculate the estimates and MSEs for subsets A. With regard to calculating the variance components of the MSE of the subset estimates, this can be done via the ABS GREGWT macro which implements the delete-a-group jackknife method. That is, we can calculate the jackknife variance of the subset estimate and the bias estimate.

44 Bell (2000) undertook a simple simulation study to compare the delete-a-group jackknife variance estimator with other methods such as weighted residuals. In comparison to the weighted residuals method, the jackknife estimates with 30 replicates were more variable. Using 60 replicates improved the stability of the estimates, but was still not as good as the weighted residuals method. Bell concluded that the jackknife was a suitable approach for estimating variances for general purposes (e.g. for publication), but that other methods such as the weighted residuals would be better when the precision of the variance estimates is more important (e.g. when trying to detect small differences while evaluating alternative methodologies).

45 This suggests that the weighted residuals method may be more suitable for the MSE calculations described above. However, it is unclear how the residuals should be

calculated in the two-phase scenario, and further, it would be more time consuming than the jackknife if we want to assess the MSE of more than one variable of interest (as residuals would need to be calculated for each variable). As a compromise, we propose to stick with the jackknife method, but increase the number of replicates from the usual 30 to at least 60.

46 Choosing the variables to be included in the saturated model is not straightforward. SIHC questionnaires can take up to one hour to administer, and as such there are many variables that we could potentially use. Practically, this is not possible because we are limited by computing power (GREGWT does not converge when a large number of auxiliary variables are used). We propose to use at least all SIHC variables to be included in the HES publication, namely, at the state level:

- average weekly household income,
- source of income (employee, own business, government pensions and allowances, other),
- tenure type (owners without a mortgage, owners with a mortgage, renters from state or territory housing authority, renters–other, other).

47 We also propose to undertake some preliminary analysis of relationships between SIHC and HES to see if there are other variables that would be sensible to include in the saturated model. This will require some degree of subjective decision making.

48 Further, in the HIES case, it is not realistic to assume that the saturated model is free from bias, as we do not know the underlying response mechanism. Perhaps a reasonable alternative assumption is that the saturated model contains an acceptable level of bias, and we are simply trying to make gains in variance without introducing too much additional bias.

49 In undertaking the above analysis to determine the appropriate auxiliary variables, $\mathbf{x}_2$, we may be faced by some practical limitations. At the time of analysis, only the first six months of HIES data will be available. Problems such as there being insufficient data for the GREGWT macro to converge is an example of what may occur.

- Is the proposed method for choosing x_2 reasonable?
- Are there more appropriate alternative methods of choosing auxiliary variables that we should be considering?

5. NON-RESPONSE ISSUES

50 Non-response bias in ABS household survey estimates is generally dealt with in one of two ways. The first is an explicit adjustment to the initial weight by the probability of response, and the second is an implicit adjustment undertaken in the calibration step. Both methods first require understanding the underlying response mechanism.

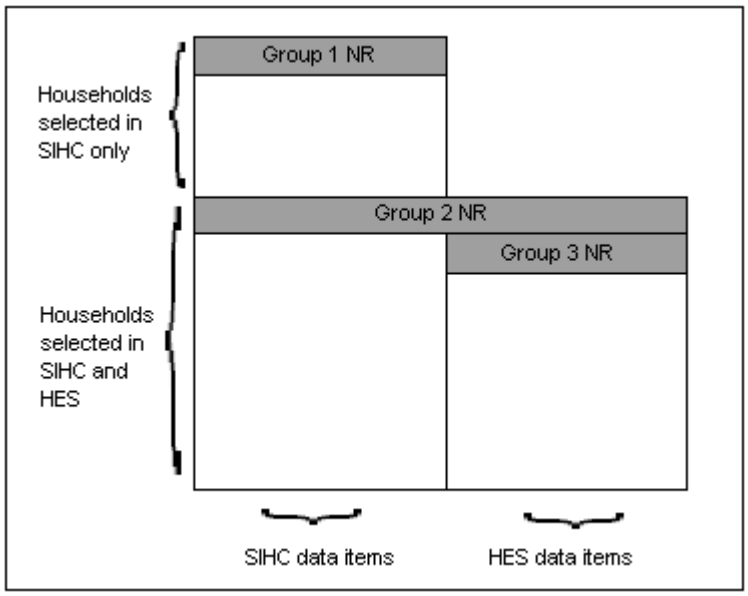
51 In order to understand the underlying response mechanism, an analysis of auxiliary variables is usually undertaken to identify characteristics of respondents and non-respondents. It is then assumed that response probabilities are uniform among groups of units with the same characteristics, with respect to the auxiliary variables. Non-response bias can then be adjusted for explicitly within these groups, or in the calibration step by including the auxiliary variables found to be associated with non-response.

52 Non-response for the HIES is more complicated than most ABS household surveys. Basically, non-respondents can be split into three groups:

- households selected for the SIHC only component that did not respond
- households selected for SIHC and HES that did not respond to either
- households selected for SIHC and HES that responded to SIHC, but not to HES.

This has been presented pictorially below.

5.1 Non-response groups in HIES



53 Non-response bias from group three will be adjusted for implicitly in the calibration step, as the auxiliary variable selection method takes non-response bias into account. The first phase calibration may also implicitly adjust for non-response bias from groups one and two, if the auxiliary variables reflect characteristics of non-respondents. The auxiliary variables used in the calibration include almost all population auxiliary variables available anyway, and as such, it is not possible to further investigate non-response bias related to population auxiliary variables.

54 One unusual question is whether a household's probability of responding in the first phase is affected by whether or not it is selected in the second phase. In other words, is a household more or less likely to respond to SIHC if it has also been selected in HES? In the diagram above, this relates to understanding any differences between group one and group two.

55 If there is a difference in response probabilities between the SIHC only and SIHC and HES samples, the implicit adjustment through the calibration will not take it into account. An explicit adjustment to the initial weight, before the calibration, may be required. If an explicit adjustment is required, we propose to use the sample based weighting cell method (Allen, 1998).

56 The sample based weighting cell method divides the sample into mutually exclusive groups, or non-response adjustment cells.The respondents in each cell are weighted by the probability of responding to compensate for non-respondents. The probability of responding, φ_r, for units in group r is estimated by:

$$\hat{\varphi}_r = \frac{\sum_{i \in s_r} w_{ri}}{\sum_{i \in s_r} w_{ri} + \sum_{i \in s_{nr}} w_{ri}}$$

where

s_r is the sample of respondents,

s_{nr} is the sample of non-respondents, and

w_{ri} is the initial weight, calculated on the basis of selection probabilities.

57 The non-response adjusted weight would be calculated by multiplying the initial weight by the inverse of the response probability for group r. For the HIES scenario, r would be determined by membership of the SIHC only selections and the SIHC and HES selections groups. Before any such adjustment is made we need to investigate whether or not there is in fact a difference in the probability of responding between the two groups. We can do this by modelling reponse probabilities, and including an indicator of group membership as one of the explanatory variables.

58 The limitation of this approach is that we can only include variables in the model that we know about respondents and non-respondents. This limits the choice of variables to state, part of state (metropolitan or non-metropolitan area) and group membership. We propose to include these three explanatory variables as well as interaction terms in the response model. If the coefficient of the group membership variable is found to be significant, then we will conclude that an explicit non-response adjustment as described above is necessary.

- What are MAC's thoughts on the proposed method for adjusting for non-response if it is required?
- What are MAC's thoughts on the proposed method for determining if a non-response adjustment is required?

6. VARIANCE ESTIMATION

6.1 Literature review

59 Sitter (1997) investigated a number of variance estimators for the regression estimator under a two-phase sampling scenario. The scenario assumed simple random sampling at both phases. Five variance estimators were investigated, including a jackknife method where single units are dropped when calculating replicates. The jackknife variance estimator was found to be design consistent for large second phase samples.

60 A limited simulation study was undertaken to assess the various methods. The simulation study showed that the jackknife method worked well conditionally, but not so well unconditionally. However, it's poorer performance unconditionally did not seem to affect the coverage probability of the confidence intervals based on it. Sitter suggested that the jackknife might be one of the preferred methods to use due to its good conditional performance and the fact that it is more operationally desirable.

61 Kott and Stukel (1997) investigated the appropriateness of the jackknife variance estimator for two-phase sampling. The sample design considered involved a first phase that was a stratified with-replacement cluster sample, and a second phase that was a stratified simple random sample without replacement. The main focus of the paper is the re-weighted expansion estimator and the double expansion estimator, but the extension to regression estimation was discussed. The authors stated that the jackknife has an approximate upward bias in their scenario and is therefore likely to be a conservative estimator of variance.

62 Kott (1998, revised 2001) presented justification for using the delete-a-group jackknife variance estimator for a range of complex estimation strategies that included multi-phase designs. He found that the delete-a-group jackknife is a nearly unbiased estimator of variance when the first phase sampling fractions are small. When the first phase sampling fractions are not small, the jackknife gives an upward biased estimate of variance.

6.2 Proposed methodology for HIES

63 The GREGWT macro caters for the delete-a-group jackknife method of variance estimation for single phase and two-phase samples. The findings in the literature on jackknife on similar scenarios suggest that this method is appropriate. As such, we propose to use this method for calculating the variances of HIES estimates.

64 Replicate weights will be calculated for the first phase, and used to calculate variances for SIHC estimates. The first phase replicate weights are used as input into the second phase calibration, and an additional set of replicate weights will be

calculated for the second phase sample (note that the second phase replicates are produced by droppoing the same set of clusters). These replicate weights will be used to calculate variance estimates for the HES estimates. The variance estimation for the first phase is consistent with how SIHC would be treated as a single phase survey.

65 The default number of replicates used in ABS household surveys is 30. Due to practical limitations, we propose to use 30 replicates in variance estimation for HIES. However, we would like to undertake investigations into whether or not 30 replicates are sufficient. We may do this by simply observing the change in variance estimates when we alter the number of replicates, or by more complex methods like assessing the variance of the variance estimates as the number of replicates change.

- In MAC's view, are there any limitations on using the delete-a-group jackknife method for variance estimation in a multi-stage, multi-phase cluster design?
- What are MAC's suggestions on assessing the required number of replicates?

7. REFERENCES

Allen, S. (1998) *Weighting Background Paper*, Internal Paper, Australian Bureau of Statistics, Canberra.

Bell, P. (2000) *Weighting and Standard Error Estimation for ABS Household Surveys*, http://www.abs.gov.au/websitedbs/D3110122.NSF/4a255eef008309e44a255eef00061e57/43677955ac4fa287ca2569920080c1ec!OpenDocument

Cochran, W.G. (1977) *Sampling Techniques*, 3rd ed., Wiley, New York.

Dupont, F. (1995) "Alternative Adjustments Where There Are Several Levels of Auxiliary Information", *Survey Methodology*, 21(2), pp. 125–135.

Estevao, V.M. and Särndal,C.-E. (2002) "The Ten Cases of Auxiliary Information for Calibration in Two-Phase Sampling",*Journal of Official Statistics*, 18(2), pp. 233–255.

Hidiroglou, M.A. and Särndal,C.-E. (1998) "Use of Auxiliary Information for Two-Phase Sampling", *Survey Methodology*, 24(1), pp. 11–20.

Kott, P.S. and Stukel, D.M. (1997) "Can the Jackknife Be Used With a Two-Phase Sample?", *Survey Methodology*, 23(2), pp. 81–89.

Kott, P.S. (1998, revised 2001) *Using the Delete-a-group Jackknife Variance Estimator in NASS Surveys*, www.nass.usda.gov/ research/allreports

SAS Institute Inc. (1990) *SAS/STAT User's Guide, Version 6*, vol. 2, 4th ed., SAS Institute Inc., Cary, NC.

Sitter, R.R. (1997) "Variance estimation for the regression estimator in two-phase sampling",*Journal of the American Statistical Association*, 92, pp. 780–787.

Skinner, C.J. and Nascimento Silva, P.L. (1997) "Variable Selection for Regression Estimation in the Presence of Nonresponse", *ASA Proceedings of the Section on Survey Research Methods*, pp. 76–82.

FOR MORE INFORMATION . . .

INTERNET **www.abs.gov.au** the ABS web site is the best place for data from our publications and information about the ABS.

LIBRARY A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our web site for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our web site, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE 1300 135 070

EMAIL client.services@abs.gov.au

FAX 1300 135 211

POST Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All ABS statistics can be downloaded free of charge from the ABS web site.

WEB ADDRESS www.abs.gov.au

2000001524435
ISBN 0 642 48167 9

RRP $11.00