1351.0.55.019

Australian Bureau of Statistics

Research Paper

An Estimating Equation Approach to Census Coverage Adjustment

www.abs.gov.au

New Issue

Research Paper

An Estimating Equation Approach to Census Coverage Adjustment

Philip A. Bell, Claire F. Clarke and Julian P. Whiting

Methodology Division

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) MON 7 MAY 2007

ABS Catalogue no. 1351.0.55.019
ISBN 978 0 64248 300 3

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper.
For further information, please contact Mr Philip Bell, Methodology Division, on Adelaide (08) 8237 7304 or email <methodology@abs.gov.au>.

CONTENTS

AN ESTIMATING EQUATION APPROACH TO CENSUS COVERAGE ADJUSTMENT

Philip A. Bell, Claire F. Clarke and Julian P. Whiting
Methodology Division

ABSTRACT

A Census Post-Enumeration Survey (PES) is conducted after each Australian Population Census. The PES provides a measure of the net under-count of the Census, and is a key input into the production of the official Estimated Resident Population counts.

This paper describes estimation of net under-count in the context of the 2006 Census and PES. It introduces a new estimator, the prediction regression (PREG) estimator, as an extension of the dual system estimator (DSE) standardly used in estimating Census under-count. In contrast to the DSE, the PREG estimator can use a variety of benchmark variables without the need to form non-overlapping post-strata. It can also adjust appropriately for persons that report different categories in the PES than were recorded in the Census.

1. BACKGROUND

The Australian Census Post-Enumeration Survey (PES) is run about four weeks after each Australian Population Census. Its purpose is to provide a check on the accuracy of the population figures from the Census, and in particular to provide estimates of the number of persons incorrectly missed in the Census (under-count), and the net under-count (under-count minus any over-count from persons incorrectly counted or counted multiple times). Responding persons in the PES are matched to the Census to determine how many times they were counted in the Census (and whether they should have been counted).

The classical approach to estimating the true population size in this situation is the dual system estimator (DSE). This proceeds by using the PES sample to estimate a Census coverage rate for each of a set of predefined categories (the post-strata). The observed Census counts in each post-stratum are then assumed to have arisen by applying that coverage rate to the true population – hence the true population counts for a post-stratum are estimated by the Census counts divided by the estimated coverage rate.

The DSE is used in overseas agencies e.g. in the US Census Accuracy and Coverage Survey (ACE) (Hogan, 2001). This approach was also the basis for estimation in the Australian PES up to 1996. The need to post-stratify the population into non-overlapping categories is a drawback of the method: to avoid post-strata having very small sample counts it may be necessary to ignore some potential benchmark variables that are in fact likely to be related to Census or PES response and which could therefore improve the accuracy of the estimates.

For the 2001 Census, the Australian Bureau of Statistics (ABS) applied a generalised regression (GREG) estimation methodology (Deville and Särndal, 1992) to give PES weights that reproduce a set of Census counts for benchmark categories. This approach avoided the restrictions of post-stratification, allowing indigenous status to be used alongside the 448 demographic categories (14 regions by 16 age groups by 2 sexes) that were previously used as post-strata. The work in this paper develops an estimator more theoretically comparable with the DSE estimator, yet that retains the desirable properties of the GREG approach. In particular, the estimator can use a variety of benchmark variables without the need to form non-overlapping post-strata. It can also cater for persons that report different categories in the PES than were recorded in the Census.

Section 2 describes the use of the DSE in the context of weighting the dwellings in the PES sample to represent all private dwellings in Australia. The resulting 'dwelling weights' are used as an input to the weighting of persons in the PES to represent the Australian population at Census time. Section 3 sets up a theoretical framework for this problem. This framework is used for the application of an estimating equation approach in Section 4 to develop an estimator dubbed the PREG estimator (for prediction regression). This estimator is a new application of instrumental variables regression (presented, for example, by Sargan (1958) and White (1982)). Section 5 shows how this estimator can also be seen as an application of the GREG methodology to predicted Census counts for each unit. A variance estimator for the PREG estimator is presented in Section 6.

The next sections describe an evaluation of the PREG estimator using simulations. Section 7 describes the creation of a simulated population from which repeated samples can be drawn. Section 8 evaluates the bias and variance of a variety of estimators including the PREG estimator based on samples drawn from the simulated population under various coverage and response models, and demonstrates the performance of the proposed variance estimators. Finally, Section 9 presents conclusions from the study, and discusses various details of the application of the PREG estimator proposed for the 2006 PES.

2. WEIGHTING DWELLINGS USING THE DUAL SYSTEM ESTIMATOR

2.1 A weight for dwellings based on the dual system estimator

The PES sampling scheme provides a sample of dwellings from the population of private dwellings in Australia. These dwellings can be matched to the list of dwellings counted in the Census, so that the number of times each PES dwelling was counted in the Census can be determined. The first stage of PES weighting provides a dwelling weight for each dwelling selected in PES, so that the PES selected dwellings represent the full population of private dwellings in Australia at PES time. This provides a good illustration of the dual system estimator (DSE).

The DSE assumes that within a post-stratum p dwellings fall into four categories at random, as follows: they have probability p_p^{CP} of being counted in both the Census and (if selected) in the PES, p_p^{UP} of being uncounted in the Census but found in the PES, probability p_p^{CM} of being counted in the Census but missed in the PES, and probability p_p^{UM} of being missed in both collection. These probabilities are shown in table 2.1 (with the subscript p dropped for clarity.

2.1 Assumed probabilities of being counted in Census and being found if selected in PES

	Counted in Census	*Missed in Census*	*All Dwellings*
Found in PES	p^{CP}	p^{UP}	$\frac{p^{\mathrm{CP}}}{p^{\mathrm{CP}}+p^{\mathrm{CM}}}$ (assumed)
Missed in PES	p^{CM}	$p^{\mathrm{UM}} = \frac{p^{\mathrm{CM}}p^{\mathrm{UP}}}{p^{\mathrm{CP}}}$ (assumed)	$\frac{p^{\mathrm{CM}}}{p^{\mathrm{CP}}+p^{\mathrm{CM}}}$ (assumed)
All dwellings	$\frac{p^{\mathrm{CP}}}{p^{\mathrm{CP}}+p^{\mathrm{UP}}}$ (assumed)	$\frac{p^{\mathrm{UP}}}{p^{\mathrm{CP}}+p^{\mathrm{UP}}}$ (assumed)	1

The PES provides an estimate $\hat{p}_p^{\mathrm{C|P}}$ of the probability of a dwelling being counted in the Census given that it is found in the PES. The DSE assumes that this probability applies to the whole population of the post-stratum. This corresponds to assuming that $p_p^{\mathrm{UM}} = p_p^{\mathrm{CM}}p_p^{\mathrm{UP}}/p_p^{\mathrm{CP}}$ – i.e. that being counted in the Census and being found in the PES are independent events. This is a key assumption underpinning the DSE (and other) under-count estimators (see Section 3.5 for further discussion). Knowing the post-stratum Census count of dwellings N_p, the DSE estimate of the post-stratum number of dwellings is then $\hat{N}_p^{\mathrm{DSE}} = N_p/\hat{p}_p^{\mathrm{C|P}}$.

In the ABS, the post-strata used for this initial dwelling weight are based on region and dwelling structure (e.g. separate house, other). Each selected dwelling j (including vacant or non-responding dwellings) is given a selection weight π_j^{-1} equal to the inverse of their selection probability. If dwelling j was counted m_j times in the Census, the DSE estimate is given by

$$\hat{N}_p^{\mathrm{DSE}} = N_p \frac{\Sigma_{j\in p}\,\pi_j^{-1}}{\Sigma_{j\in p}\,\pi_j^{-1} m_j} = \frac{N_p}{\hat{p}_p^{\mathrm{C|P}}} \qquad \text{for } \hat{p}_p^{\mathrm{C|P}} = \frac{\Sigma_{j\in p}\,\pi_j^{-1} m_j}{\Sigma_{j\in p}\,\pi_j^{-1}}$$

Table 2.2 shows the counts that are estimated by the DSE for dwellings in the various categories. The key thing to note is that the DSE provides an estimate for dwellings missed in both the Census and the PES based on the assumption that being missed in the Census and being missed in the PES are independent events.

2.2 Estimated counts in various categories under the DSE

	Counted in Census	*Missed in Census*	*All Dwellings*
Found in PES	$\Sigma_{j\in p}\,\pi_j^{-1} m_j$	$\Sigma_{j\in p}\,\pi_j^{-1}(1-m_j)$	$\Sigma_{j\in p}\,\pi_j^{-1}$
Missed in PES	$N_p - \Sigma_{j\in p}\,\pi_j^{-1} m_j$	$\Sigma_{j\in p}\,\pi_j^{-1}(1-m_j)$ $\times\left(\frac{N_p - \Sigma_{j\in p}\,\pi_j^{-1} m_j}{\Sigma_{j\in p}\,\pi_j^{-1} m_j}\right)$	$\Sigma_{j\in p}\,\pi_j^{-1}$ $\times\left(\frac{N_p - \Sigma_{j\in p}\,\pi_j^{-1} m_j}{\Sigma_{j\in p}\,\pi_j^{-1} m_j}\right)$
All dwellings	N_p	$N_p \frac{\Sigma_{j\in p}\,\pi_j^{-1}(1-m_j)}{\Sigma_{j\in p}\,\pi_j^{-1} m_j}$	$\hat{N}_p^{\mathrm{DSE}} = N_p \frac{\Sigma_{j\in p}\,\pi_j^{-1}}{\Sigma_{j\in p}\,\pi_j^{-1} m_j}$

2.2 Details of dwelling weighting in the ABS

The DSE described above applies an initial dwelling weight for dwelling j in PES post-stratum p of $w_j^{\mathrm{I}} = \pi_j^{-1} N_p / \Sigma_{j^*\in p}\,\pi_{j^*}^{-1} m_{j^*}$. The ratio adjustment applied to a PES dwelling's selection weight depends only on its post-stratum, not on whether it was counted in the Census. This illustrates the fundamental weighting principle of PES weighting:

> *Fundamental principle of PES weighting*:
>
> Identical weight changes should apply to all units selected in the PES with the same characteristics, *regardless* of whether they were counted in the Census.

Applying this principle would force the use of post-strata based only on items known for all dwellings, even those missed in the Census. In practice, DSE weights are produced under a broad and a fine post-stratification. The broad post-stratification is applied first, and is based solely on PES information. Then, for those dwellings found in both PES and Census (the vast majority of dwellings), a finer post-stratification is applied to ensure that the PES sampled dwellings represent the range of Census response types (single occupant, multiple occupant, vacant on Census night, not contacted in the Census by the start of PES) and also dwelling types (separate house, flat etc.) as accurately as possible.

The effect is that the weight adjustment for dwellings missed in the Census is a weighted average of weight adjustments for the corresponding fine post-strata containing dwellings found in both PES and Census.

This initial weighting provides a weight to all selected dwellings, even those for which PES does not obtain a response from persons in that dwelling.

The ABS produces a final dwelling weight by applying a non-response adjustment, so that the responding PES dwellings are appropriately weighted up to represent other dwellings from which no PES response was obtained but that are non-vacant. Like the initial dwelling weighting, the non-response adjustment is performed within post-strata based on region, dwelling structure and the Census response provided by that dwelling. This sort of non-response adjustment is routinely required in household surveys, and details are not provided here. The ABS achieves very high contact and response rates in the PES.

3. A FRAMEWORK FOR WEIGHTING OF PERSONS IN THE PES SURVEY

3.1 The need for a person weight

The PES is enumerated at a sample of dwellings from the population of private dwellings (PDs) at the time of the PES. These are weighted as described above so as to represent the full population of private dwellings in Australia at PES time.

This sample of dwellings can also be seen as a sample of persons. Information is sought from selected dwellings for all persons associated with the dwelling at PES time by standard 'scope and coverage' rules; basically, this includes all persons whose usual residence is that selected dwelling, as well as any visitors from any other usual residence in Australia that is unoccupied during the survey period. The key items of interest for each person are whether they *should* have been counted in the Census and how many times they *were* actually counted.

The PES only covers persons associated with dwellings that are available for selection at PES time. The population of interest, however, is all persons in Australia on Census night. Clearly there are a number of persons in this population of interest that are not covered in the PES. Some such persons are in other types of dwellings at PES time, such as hotels, hospitals and jails. The PES does not cover these 'non-private dwellings' for practical reasons, including the difficulty of contacting them in a manner that is independent of the Census. Other classes of persons who should have been counted in the Census but are not covered in PES would be persons in Australia on Census night but overseas at the time of the PES, and persons who died between the Census and the PES.

The dwelling weights, if applied to responding PES persons, provide an estimate that represents only that subset of the Census-night population that is found in private dwellings at PES time. A second stage of weighting is required to provide person weights that represent the whole Census-night population, using the assumption that the behaviour of the PES sample is representative of this population. For example, the under-coverage rate of young males observed in the PES sample is assumed also to apply to young males not accessible in the PES sample because they were in non-private dwellings or overseas at PES time (but not at Census time).

3.2 The treatment of PES-prompted Census returns and of the Census non-contact sector

The PES is conducted only four weeks after Census night, and a small number of Census forms are still being solicited and received at the time of the PES. A problem arises if being selected in the PES sample prompts the dwelling to return a Census form when otherwise they would not have done so. This would lead to the PES

sample having an unrepresentatively high rate of completed Census forms, and a correspondingly low rate of dwellings not completing a form. The particular forms prompted by PES cannot be distinguished from other Census returns arriving late for reasons unrelated to the PES.

To deal with this situation, PES estimation conceptually uses the Census on an 'at start of PES' basis. Census forms that were returned by mail or the Internet after a specified date are classified as late returns (LR). For the purpose of PES estimation these LR dwellings are treated as though they had not been contacted in the Census, and are classified to the 'non-contact sector' of the Census.

The non-contact sector also contains dwellings which Census classed as non-contacts - that is, dwellings where the Census never obtained a return, and which could not be established as having been unoccupied on Census night. These non-contact dwellings are given imputed values during Census processing, based in many cases on information provided by the Census collector about the dwelling and its residents. Inevitably, the imputed values, at the dwelling and aggregate level, differ from the true, but unknown, values. The imputed records constitute the majority of the Census non-contact sector records; late returns (as defined here) are only a small component of the overall Census non-contact sector. Given that late returns prompted by PES would otherwise have been classed as non contacts, the PES sample is representative of the whole non-contact sector, even though it cannot split late returns from non-contacts in a manner comparable to the Census.

In previous Censuses, only the Census contact sector was corrected for under- and over-count by using the PES estimates. Effectively late returns and imputed records (the Census non-contact sector for PES purposes) were treated as being reported accurately. While this assumption is imperfect, this was considered the most feasible way to calculate accurate net under-count estimates at the level of detail needed for producing estimated resident population counts in Australia.

Under the methodology detailed in this paper, the person weighting step in PES processing calculates weights for all PES records, including those that correspond to the Census non-contact sector. So the PES can provide an estimate of the total population in Census late return and non-contact dwellings on Census night. This is a change from previous PES surveys, in which persons selected in non-contact sector dwellings were excluded from matching and from estimation. The inclusion of these persons in the 2006 PES is an innovation made possible by the development of appropriate methods for representing them in estimation.

Estimates for the non-contact sector have relatively high sampling errors because of the small sample size (there are relatively few non-contact dwellings selected by chance in the PES sample); and also because person counts for this sector are not available to use as a weighting 'benchmark'. This lack of Census person counts also

means that, while the dwelling weights used for the non-contact sector are estimated from the sector itself, the adjustments applied to provide final person weights depend strongly on information observed in the contact sector. This is a potential source of non-sampling error, as is any bias arising from peculiarities of the non-respondents in this sector. Both these sources of non-sampling error are expected to be small compared to the sampling error of the non-contact sector estimates.

Using PES estimates for the population of the non-contact sector will lead to a noticeable rise in the standard error of the overall population estimates, compared to the alternative in which this sector is not measured by PES but is treated as accurately represented by the Census figures. This alternative could, however, have a bias associated with Census imputation of non-contacts. Since the standard error for the non-contact sector estimate can be calculated, the ABS will be able to make a scientifically considered judgement about the comparative accuracy of the estimate based on PES, and the Census count for the late return and imputed dwellings. This gives the option of using the PES estimate for this sector, and thus to adjust for inaccuracies in the Census imputation process. Note that, should this prove to be desirable, it would be conceptually separate from the part of the net under-count adjustment that applies to the Census contact sector. The underlying processes giving rise to net under-count in the Census contact sector (basically reporting errors by individuals, and collection errors in the Census) are quite distinct from those causing errors in the imputation process for Census non-contacts, and are likely to have different net under- or over-count rates.

3.3 A model for PES coverage and response

As noted previously, the key items of interest for each PES unit are the number of times persons should have been counted in the Census and the number of times that they actually were counted. Aggregating these items, using the dwelling weights, gives estimates of total persons and number of persons counted for that subset of the Australian population on Census night who were in private dwellings at PES time. The second, person weighting stage modifies these weights so that they represent the whole Australian population on Census night. It is important to note that the weights perform a dual function: to enable inference from the PES sample to the full target PES population, and to estimate the population not covered by either PES or Census in the spirit of dual system estimation.

Define the PES target population U as consisting of the set of persons in private dwellings at the time of PES plus any persons who should have been or were counted in the Census but who were not in private dwellings at the time of PES. Ideally the PES should provide a probability sample from this population.

However when weighted with the dwelling weights the PES sample actually only provides an estimate for the covered population U^C consisting of persons in private dwellings at PES time. Furthermore, it is known that particular classes of people (for example, indigenous persons or males aged 25–34) are likely to be under-represented by the responding PES sample. The objective of person weighting is to adjust the dwelling weights d_i attached to each sampled person i to give new weights w_i that adjust for these inadequacies and appropriately represent the complete PES target population U.

For unit i in the complete target population U let $\boldsymbol{y}_i = (y_{i1}, ..., y_{iL})$ be a row vector with an element y_{il} for each of L PES items describing characteristics of the unit. For example, the lth element y_{il} could give the number of persons from unit i in a PES category l (for example, females aged 15 to 19 years with usual residence in Hobart). If the units are persons, y_{il} takes values 0 or 1, while if the units are dwellings, y_{il} is an integer greater than or equal to zero.

Let $\hat{\boldsymbol{y}}^D = \Sigma_{i \in S} d_i \boldsymbol{y}_i$ for S a PES sample. This dwelling-weighted estimate $\hat{\boldsymbol{y}}^D$ is expected to differ from the Australian population total vector $\boldsymbol{Y}^U = \Sigma_{i \in U} \boldsymbol{y}_i$, and it is this expected difference that the person weighting seeks to allow for. To formalise this requires proposing a probability mechanism, M, for PES coverage and response.

Consider the complete target population U as a realisation of a super-population model in which potential dwellings and persons of many different types are generated and given values. Consider further that for each unit i in U, membership of the covered population U^C, selection in the PES sample and finally response in PES all occur at random according to some overall probability mechanism, M. Then the dwelling weights d_i account for the probability of selection in the PES, dwelling non-response, and any under-coverage of the population of dwellings at PES time. They don't account for the remaining under-coverage (people not in private dwellings, gone overseas at PES time etc.) and PES person level non-contact and non response mechanisms.

Under this set-up, a 'coverage-response adjustment' for units $i \in C$ for some class C is defined as the ratio $R_C = E_{\mathbf{M}}(\Sigma_{i \in U,C}\ 1) / E_{\mathbf{M}}(\Sigma_{i \in S,C}\ d_i)$, where expectations are across the probability mechanism M described in the previous paragraph. The adjustment R_C is the ratio of the expected overall number of units in class C in U to the expected estimated number of units in C based on the dwelling weighting.

Suppose that the coverage-response adjustment for a class C depends only on the values of the $\boldsymbol{y}_i$ vector, for the units i in C. Then as an alternative to writing this as an adjustment R_C for the whole class C we can write it as an adjustment R_i, for each unit i in C. Formalising this, we define the PES coverage-response adjustment R_i for a unit i as the adjustment for the class of units having the same $\boldsymbol{y}_i$ characteristics as unit i.

The dependence of the adjustment on $\boldsymbol{y}_i$ will be described by a function $R_i = \mathrm{R}(\boldsymbol{y}_i, \boldsymbol{\theta})$ for some row vector of parameters $\boldsymbol{\theta}$.

3.4 Estimating PES coverage and response adjustments using benchmark variables from the Census

Let $\boldsymbol{x}_i = (\mathrm{x}_{i1}, ..., x_{iK})$ be a row vector with an element for each of K Census items. For example, the kth element x_{ik} could give the number of times persons from unit i were actually counted in the Census in a Census category k. So if i represents persons, x_{ik} could be 0, 1 or greater, as these items give what was observed in the Census, which may under-count or over-count the actual population at Census time.

The totals of these variables across the whole population U are known from the Census, given in the vector $\boldsymbol{X} = (X_1, ..., X_K) = \Sigma_{i \in \mathrm{U}} \boldsymbol{x}_i$, where element k could give the Census count of persons in Census category k. The dwelling weights can be used to produce a corresponding vector of estimates $\hat{\boldsymbol{x}}^{\mathrm{D}} = \Sigma_{i \in \mathrm{S}} d_i \boldsymbol{x}_i$ for these variables from the PES sample. Person weighting for PES is based on adjusting for the difference between $\boldsymbol{X}$ and $\hat{\boldsymbol{x}}^{\mathrm{D}}$. In simple terms, we construct person weights for the PES in such a way that, if we were to use them to estimate what was reported in the Census (the $\hat{\boldsymbol{x}}$ variables), we would get the actual total Census count $\boldsymbol{X}$. We can then apply these weights in the estimation of other variables which were reported in the PES.

The variables in $\boldsymbol{x}_i$ are known as the benchmark variables for PES person weighting, and $\boldsymbol{X}$ contains the corresponding benchmark totals (or 'benchmarks'). For a good choice of benchmark variables the values $\boldsymbol{x}_i$ for each unit $i \in$U will be strongly related to the $\boldsymbol{y}_i$ values that determine the PES coverage-response adjustment. This can be expressed as a dependence e.g. an element of $\boldsymbol{x}_i$ containing the count of times a person actually was counted in a category in the Census depends on the element of $\boldsymbol{y}_i$ giving whether they really belong in that category. It may be helpful to think of this as a 'measurement error model' where the $\boldsymbol{x}_i$ take values generated at random with some probability density conditional on the PES item values $\boldsymbol{y}_i$ for the unit – but such is not required for the estimation developed here.

The objective of Census adjustment is to obtain an estimate of various population totals $T = \Sigma_{i \in \mathrm{U}} t_i$ of items t_i known for the PES sample. For example, T could be the total Australian population at Census time, for t_i the total number of persons from unit i that should have been counted in the Census. The estimators implicitly assume that any estimated item t_i is actually measured correctly by the PES. The benchmark variables $\boldsymbol{x}_i$ are used as auxiliary information to assist with the estimation of totals T using the PES sample.

To summarise the notation:

- $\boldsymbol{x}_i$ are variables which contain information about how many times the Census actually counted a person in a particular Census category;
- $\boldsymbol{y}_i$ are variables in PES that contain information determining a person's response and coverage weighting adjustments, so that the PES sample correctly represents the full Census night population; and
- t_i are the 'true' variables to be estimated in PES, that contain information about how many times the Census should have counted a person in a particular category.

3.5 Independence between Census and PES response

Importantly, the Census responses are assumed to occur independently of PES coverage and response, conditional on the PES values $\boldsymbol{y}_i$. This is critical to obtaining unbiased estimates from the PES.

ABS has control over both Census and PES operations, and so can ensure that there is 'operational independence' between them. Field operations for the PES are carefully designed to ensure that the PES collection is independent of the Census collection as far as possible. However, characteristics of a unit itself that affect both PES and Census response, unrelated to ABS activities, can lead to lack of independence between responses in the two collections, leading to a bias in estimates. This form of 'correlation bias' can be minimised by including appropriate variables in $\boldsymbol{y}_i$. For example, because being indigenous is expected to decrease the likelihood of being counted in both Census and in PES it is helpful to include one or more variables in $\boldsymbol{y}_i$ indicating indigenous persons. The desire to extend the $\boldsymbol{y}_i$ information beyond a simple post-stratum indicator is one motivation for the extensions of the DSE given in this paper.

Unfortunately, there always remains the possibility of some correlation between Census response and PES response that is not controlled for by the variables $\boldsymbol{y}_i$. For example, people may avoid the PES interviewer specifically because they don't want to own up to having missed filling in their Census form. This results in another form of correlation bias that cannot be dealt with by extending $\boldsymbol{y}_i$.

Demographic analysis of the PES estimates, such as comparisons of numbers of males and females in various age groups, can shed light on the size of this final potential correlation bias. It is ABS practice to perform estimation first, and then apply demographic analysis to identify any potential problems. Potentially this could lead to some adjustment in how the PES estimates impact the published Estimated Resident Population (ERP) figures.

An alternative would be to propose PES estimators that incorporate an adjustment for correlation bias, as in Isaki and Schultz (1986) and Wolter (1990). Choi, Steel and Skinner (1988) evaluated some of Wolter's estimators for the 1986 Australian PES, and observed some erratic results, concluding that the estimators may be useful for evaluation but not recommending their use in PES estimation. There would seem to be some danger in automatically treating an apparently anomalous figure as though it were due to correlation bias; other issues may be responsible, such as systematic errors arising in matching Census and PES persons.

4. AN ESTIMATING EQUATION APPROACH TO PES ADJUSTMENT

4.1 An initial estimate based on dwelling weights

The dwelling weights (as developed in Section 2) can be used as weights for the individual persons in those dwellings; thus dwelling weights d_i are available for all units i in S, whether person or dwelling is used as the unit. These weights provide an estimator $\hat{t}^{\mathrm{D}} = \Sigma_{i \in \mathrm{S}}\, d_i^{-1} t_i$ that is biased as an estimate of the full population total $T = \Sigma_{i \in \mathrm{U}}\, t_i$ of an item t_i. This unavoidable bias was discussed in Section 3.1, and arises essentially because the PES only covers persons in private dwellings at the time of the PES. The estimation process seeks to obtain a new set of weights that give less biased estimates for the population U. This section explains the approach to person weighting used in the Australian PES as an application of a technique known as 'estimating equations'.

4.2 Estimation by modelling the coverage-response adjustment

If the coverage-response adjustment R_i was known for every unit in the sample S then a Horvitz–Thomson estimator of the Census totals $\boldsymbol{X}$ would be available. The combined probability of being in sample for unit i under the probability sampling scheme and the PES coverage and response model M is given by $(d_i R_i)^{-1}$. The resulting Horvitz–Thomson estimator is given by $\hat{\boldsymbol{x}}^{\mathrm{H}} = \Sigma_{i \in \mathrm{S}}\, d_i R_i \boldsymbol{x}_i$. The expectation of this estimator under M and the probability sampling scheme is written $\mathrm{E}_{\mathbf{M}}(\hat{\boldsymbol{x}}^{\mathrm{H}}) = \boldsymbol{X}$.

In practice, the coverage-response probabilities are not known, and need to be estimated. The estimating equation approach proposes a weight adjustment function $R_i = \mathrm{R}(\boldsymbol{y}_i, \boldsymbol{\theta})$ — this function links the coverage-response adjustments R_i to characteristics $\boldsymbol{y}_i$ of the unit (available for all survey responses) and the weights d_i available from an initial stage of weighting. The weight adjustment function introduces a row vector $\boldsymbol{\theta}$ of unknown parameters. These parameters are estimated by choosing values $\hat{\boldsymbol{\theta}}$ for which the resulting estimator $\hat{\boldsymbol{x}}^{\mathrm{H}}(\hat{\boldsymbol{\theta}}) = \Sigma_{i \in \mathrm{S}}\, d_i \mathrm{R}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}}) \boldsymbol{x}_i$ reproduces the Census totals $\boldsymbol{X}$.

This approach leads to weights $w_i(\hat{\boldsymbol{\theta}}) = d_i r_i$, for weight adjustment $r_i = \mathrm{R}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}})$. These weights are used to produce estimates of totals based on the PES sample.

4.3 Calibration of weights without biasing estimates for the total population

Adjusting an initial set of weights so that they reproduce a set of known benchmark totals is a process known as calibration. Calibration of weights can have benefits in terms of reducing the sampling error of estimates, even if the initial weights were unbiased for the true population. So weight adjustment in the PES should be seen not

only as adjusting for differential coverage-response probabilities between units, but also as reducing sampling error by adjusting for imbalances in the sample occurring by chance.

In the subsections below, we shall introduce an estimator in which the weight adjustment applied to a particular unit does not depend on its Census response $\boldsymbol{x}_i$ but only on characteristics $\boldsymbol{y}_i$ of the unit known regardless of whether the unit was counted in the Census. This is a distinctive property of the estimator developed here, and is required for unbiased estimation of totals in the PES context.

4.4 The dual system estimator derived using estimating equations

The classical approach to PES weighting uses a post-stratification of the PES units. Each unit i contributes to a single post-stratum, indicated by a one in the appropriate element of $\boldsymbol{y}_i$. Persons either respond in Census correctly, or fail to respond in Census at all – thus either $\boldsymbol{x}_i = \boldsymbol{y}_i$ or $\boldsymbol{x}_i = \boldsymbol{0}$.

In this setting, the PES coverage and response model will assign all units in a post-stratum k the same weight adjustment $r_i = \boldsymbol{y}_i \boldsymbol{\theta}^{\mathrm{DSE}\prime}$, where $\boldsymbol{\theta}^{\mathrm{DSE}} = (\theta_1^{\mathrm{DSE}}, ..., \theta_K^{\mathrm{DSE}})$ contains weight adjustments for each post-stratum. The estimating equations are $\Sigma_{i \in \mathrm{S}} \, d_i r_i \boldsymbol{x}_i = \boldsymbol{X}$; this reduces to the K equations $\hat{\theta}_k^{\mathrm{DSE}} = X_k / \hat{x}_k^{\mathrm{D}}$, where $\hat{x}_k^{\mathrm{D}}$ denotes the kth element of $\hat{x}^{\mathrm{D}} = \Sigma_{i \in \mathrm{S}} \, d_i \boldsymbol{x}_i$. This leads to applying a unit weight given by $w_i^{\mathrm{DSE}} = d_i r_i = d_i X_k / \hat{x}_k^{\mathrm{D}}$ for i in post-stratum k.

The resulting estimator $\hat{t}^{\mathrm{DSE}} = \Sigma_{i \in \mathrm{S}} \, w_i^{\mathrm{DSE}} t_i = \Sigma_k (X_k / \hat{x}_k^{\mathrm{D}}) \, \Sigma_{i \in k} \, d_i t_i$ corresponds to the dual system estimator (DSE). This approach, with person as the unit, was the basis for PES estimation up to 1996, though the ABS application expands slightly on the original DSE in that the Census response of a person may be any multiple of its PES response, to account for persons counted multiple times in the Census.

Note that the weight adjustment in a post-stratum can be estimated because it is constant within a post-stratum i.e. given the $\boldsymbol{y}_i$ which defines which post stratum a unit i is in. If the coverage-response adjustment also depended on $\boldsymbol{x}_i$ (as it would if Census and PES were not independent conditional on $\boldsymbol{y}_i$) it would not have been possible to estimate the weight adjustments (except by proposing a correlation model in which the correlation parameters were known).

4.5 The prediction regression estimator

The same estimating equation approach can be used to extend the DSE by assuming a more sophisticated model for the coverage-response probability R_i. In this model the coverage-response adjustment R_i still depends only on $\boldsymbol{y}_i$, through a model $R_i = \mathrm{R}(\boldsymbol{y}_i, \boldsymbol{\theta})$ with a row vector of unknown parameters $\boldsymbol{\theta}$. But the definition of $\boldsymbol{y}_i$ has been greatly expanded; it is no longer restricted to indicating post-stratum

membership, but can indicate a variety of information about a unit i that could be related to coverage-response. (In practice we will choose variables $\boldsymbol{y}_i$, and a function $\mathrm{R}(\boldsymbol{y}_i, \boldsymbol{\theta})$, that give the estimator the properties we desire – low variance, unbiasedness and calibration to known Census totals). Estimates $\hat{\boldsymbol{\theta}}$ of the model parameters will be chosen so that the resulting weight adjustments $\hat{r}_i = \mathrm{R}(\boldsymbol{y}_i, \hat{\boldsymbol{\theta}})$ give an estimator $\hat{\boldsymbol{x}}^{\mathrm{P}} = \Sigma_{i \in \mathrm{S}}\, d_i \hat{r}_i \boldsymbol{x}_i$ that reproduces the Census counts i.e. $\hat{\boldsymbol{x}}^{\mathrm{P}} = \boldsymbol{X}$.

This paper will look at a model of the form $r_i = 1 + \boldsymbol{\theta} \boldsymbol{y}_i' / c_i$ for some penalty values c_i. This form of model results in an estimator that resembles a generalised regression (GREG) estimator (Deville and Särndal, 1992). The estimator resulting from this model will be termed the 'prediction regression' or PREG estimator.

The constant 1 in the model $r_i = 1 + \boldsymbol{\theta} \boldsymbol{y}_i' / c_i$ is included to ensure that a unit with $\boldsymbol{y}_i = \boldsymbol{0}$ gets a default weight adjustment of 1 i.e. no change to its weight. This constant makes no difference to the resulting weights provided that the model contains an intercept term i.e. if there is a linear combination γ for which $\gamma \boldsymbol{y}_i' = c_i$ for all units i. This is guaranteed if $\boldsymbol{y}_i$ includes variables for person counts in a mutually exclusive and exhaustive set of categories, and the total person count for the unit is used for the penalty c_i. In practice, a situation in which there are units with $\boldsymbol{y}_i = \boldsymbol{0}$ can be avoided by changing or extending the $\boldsymbol{y}_i$ vector, perhaps by adding a separate element of the vector to indicate these units.

4.6 Estimating equations for the PREG model

If $\boldsymbol{\theta}$ were known, the relationship $r_i = 1 + \boldsymbol{\theta} \boldsymbol{y}_i' / c_i$ gives a Horvitz–Thomson estimator $\hat{\boldsymbol{x}}^{\mathrm{H}} = \hat{\boldsymbol{x}}^{\mathrm{D}} + \Sigma_{i \in \mathrm{S}}\, d_i \boldsymbol{\theta} \boldsymbol{y}_i' / c_i \boldsymbol{x}_i = \boldsymbol{X}$. Thus the estimating equations are given by

$$\boldsymbol{\theta} \boldsymbol{B} = \boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}} \qquad \text{for } \boldsymbol{B} = \Sigma_{i \in \mathrm{S}}\, d_i \boldsymbol{y}_i' \boldsymbol{x}_i / c_i. \tag{1}$$

Any value $\hat{\boldsymbol{\theta}}$ for which this relationship holds will result in an estimator that reproduces the Census totals $\boldsymbol{X}$ when applied to the PES sample values $\boldsymbol{x}_i$. Setting $\hat{\boldsymbol{\theta}}$ to a solution of these equations gives an estimator of a total $T = \Sigma_{i \in \mathrm{U}}\, t_i$ given by

$$\hat{t}^{\mathrm{R}}(\hat{\boldsymbol{\theta}}) \qquad = \hat{t}^{\mathrm{R}} d_i \hat{r}_i t_i = \hat{t}^{\mathrm{D}} + \hat{\boldsymbol{\theta}} \Sigma_{i \in \mathrm{S}}\, d_i \boldsymbol{y}_i' t_i / c_i \tag{2}$$

The weight applied by the estimator to unit i is thus

$$w_i^{\mathrm{R}}(\hat{\boldsymbol{\theta}}) \qquad = d_i (1 + \hat{\boldsymbol{\theta}} \boldsymbol{y}_i' / c_i) \tag{3}$$

In practice, there may be more parameters in $\boldsymbol{\theta}$ than there are counts in $\boldsymbol{X}$. In this case, there are a range of values that $\hat{\boldsymbol{\theta}}$ could take to solve the estimating equations.

Another way to look at this is that there is too much information in $\boldsymbol{y}_i$ to uniquely specify $\boldsymbol{\theta}$ – in fact there are different estimators corresponding to replacing $\boldsymbol{y}_i$ by various K-dimensional combinations of the form $\boldsymbol{z}_i = \boldsymbol{y}_i \boldsymbol{Z}$.

4.7 Optimality criterion defining the PREG estimator

The PREG estimator arises by choosing a value of $\hat{\boldsymbol{\theta}}$ that minimises a distance function for the change in weights, subject to constraints given by the estimating equations (1). The distance function to be minimised is

$$\text{distance} \quad = \Sigma_{i \in S} \, d_i c_i \left(\frac{w_i - d_i}{d_i}\right)^2 \tag{4}$$

This is the same distance function that is used by the standard generalised regression (GREG) estimator (Deville and Särndal, 1992). Substituting $w_i^R(\hat{\boldsymbol{\theta}})$ from equation (3) for the weights in (4) results in a function $D(\hat{\boldsymbol{\theta}})$ giving the distance for a particular choice of $\hat{\boldsymbol{\theta}}$, as follows.

$$\begin{aligned} D(\hat{\boldsymbol{\theta}}) &= \Sigma_{i \in S} \, d_i c_i [\hat{\boldsymbol{\theta}} \boldsymbol{y}_i' / c_i][\hat{\boldsymbol{\theta}} \boldsymbol{y}_i' / c_i]' \\ &= \hat{\boldsymbol{\theta}} \boldsymbol{A} \hat{\boldsymbol{\theta}}' \qquad \text{for } \boldsymbol{A} = \Sigma_{i \in S} \, d_i \boldsymbol{y}_i' \boldsymbol{y}_i / c_i' \end{aligned} \tag{5}$$

To obtain the minimum distance subject to the condition $\boldsymbol{\theta B} = \boldsymbol{X} - \hat{\boldsymbol{x}}^D$, apply the Lagrange multipliers method. The function to be minimised by choice of row vectors $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ is then

$$f(\boldsymbol{\theta}, \boldsymbol{\lambda}) \qquad = D(\boldsymbol{\theta}) + (\boldsymbol{X} - \hat{\boldsymbol{x}}^D - \boldsymbol{\theta B}) \boldsymbol{\lambda}' \tag{6}$$

Setting the derivatives of this to 0 gives

$$f_\theta(\boldsymbol{\theta}, \boldsymbol{\lambda}) \qquad = 2\boldsymbol{A\theta}' - \boldsymbol{B\lambda}' = 0 \text{ so that } \boldsymbol{\theta} = \tfrac{1}{2} \boldsymbol{\lambda B}' \boldsymbol{A}^{-1} \tag{7}$$

$$f_\lambda(\boldsymbol{\theta}, \boldsymbol{\lambda}) \qquad = (\boldsymbol{X} - \hat{\boldsymbol{x}}^D) - \boldsymbol{\theta B} = 0 \tag{8}$$

Substituting (7) into (8) gives

$$(\boldsymbol{X} - \hat{\boldsymbol{x}}^D) = \tfrac{1}{2} \boldsymbol{\lambda B}' \boldsymbol{A}^{-1} \boldsymbol{B} \tag{9}$$

Finally, substituting (9) into (7) gives the parameter $\hat{\boldsymbol{\theta}}^{\mathrm{PR}}$ that minimises $\mathrm{D}(\boldsymbol{\theta})$

$$\hat{\boldsymbol{\theta}}^{\mathrm{PR}} = (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}})(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}^{-1} \tag{10}$$

This corresponds to a standard situation described by Rao (1973, p. 60) in which a vector $\boldsymbol{\theta}$ is chosen to minimise a quadratic form $\boldsymbol{\theta}\boldsymbol{A}\boldsymbol{\theta}'$ subject to linear constraints, in this case $\boldsymbol{\theta}\boldsymbol{B} = \boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}}$. If there are dependent rows in $\boldsymbol{A}$ or in $(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})$ the inverse can be replaced by any generalised inverse. The resulting estimator will be called the prediction regression estimator, denoted

$$\hat{t}^{\mathrm{PR}} = \hat{t}^{\mathrm{D}} + \hat{\boldsymbol{\theta}}^{\mathrm{PR}} \Sigma_{i \in \mathrm{S}}\, d_i \boldsymbol{y}_i' t_i / c_i \tag{11}$$

The weight applied by the prediction regression estimator to unit i is then

$$w_i^{\mathrm{PR}} = d_i[1 + (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}})(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{y}_i'/c_i] \tag{12}$$

4.8 Application of the PREG estimator in the PES

In the 2006 PES, the PREG estimator is applied with person as the weighting unit. The predictor variables used are the row vector $\boldsymbol{y}_i$ in which each element is a 0-1 variable representing whether person i is in a specified category as reported in the PES; these categories are based on items such as region of usual residence, sex, age and indigenous status. The variables in $\boldsymbol{y}_i$ do not depend on whether a person should have been or was counted in the Census, or whether their Census dwelling was a late return or imputed dwelling. This means that the same response-coverage adjustment applies to persons not counted in the Census as is applied for a similar person who was counted.

The benchmark variables $\boldsymbol{x}_i$ are a corresponding vector containing how many times the person was counted in the Census in each category (as reported in the Census), at private dwellings other than late returns or imputed dwellings. The corresponding total Census counts form the benchmarks $\boldsymbol{X}$.

This process gives weights that represent all persons in private dwellings at Census time—even those in the non-contact sector, since persons who should have been counted in late returns or imputed dwellings receive a sensible response-coverage adjustment (equal to that used for a contact-sector person with the same $\boldsymbol{y}_i$ values).

In practice, a second step of PREG estimation is used to adjust the above weights to reproduce overall Census counts including non-private dwellings. The predictor variables $\boldsymbol{y}_i$ for this second step use categories based on region, sex and age only,

since other variables are not reported consistently in the Census for non-private dwellings. The predictor variables at this second weighting step are set to zero for people who should have been counted in late returns and imputed dwellings—this ensures that the PES estimate of persons who should have been counted in this non-contact sector are unaffected by this second step of weighting (which is appropriate given that there are no non-private dwellings in this sector).

The benchmark variables $\boldsymbol{x}_i$ at this second step contain how many times the person was counted in the Census in each region by sex by age group category, at both private and non-private dwellings (other than late returns and imputed dwellings). The benchmark vector $\boldsymbol{X}$ contains the corresponding overall Census counts.

5. COMPARING THE PREG ESTIMATOR TO A GREG ESTIMATOR

5.1 Viewing PES estimation as a weight adjustment problem

This section presents an alternative development of the PREG estimator. Section 4 derived it as an application of the 'estimating equation' approach. This section shows how it can equally well be viewed as an extension of the well-known generalised regression (GREG) estimator.

As noted previously, using only the dwelling weights d_i gives a biased estimator $\hat{t}^{\mathrm{D}} = \Sigma_{i\in\mathrm{S}}\, d_i t_i$ for a total $T = \Sigma_{i\in\mathrm{U}}\, t_i$ over the population U. The PES estimation problem is to provide weights w_i for which $\mathrm{E}_{\mathbf{M}}(\Sigma_{i\in\mathrm{S}}\, w_i t_i) = T$ for any item t_i. A basic condition for this is that $\mathrm{E}_{\mathbf{M}}(\Sigma_{i\in\mathrm{S}}\, w_i \boldsymbol{x}_i) = \boldsymbol{X}$. The only information available about the uncovered portion of the population is the difference between the estimates $\hat{\boldsymbol{x}}^{\mathrm{D}} = \Sigma_{i\in\mathrm{S}}\, d_i \boldsymbol{x}_i$ and the Census totals $\boldsymbol{X} = \Sigma_{i\in\mathrm{U}}\, \boldsymbol{x}_i$. It is natural to seek a set of weights for which $\Sigma_{i\in\mathrm{S}}\, w_i \boldsymbol{x}_i = \boldsymbol{X}$ and use them in estimating any item of interest t_i.

Unfortunately, not every set of weights fulfilling this condition on the $\boldsymbol{X}$ vector will give an unbiased estimator for other items. The following discusses some biased approaches to this problem before arriving at another derivation of the PREG estimator.

5.2 The GREG estimator

The generalised regression (GREG) estimator chooses weights w_i^{GR} that add to the benchmarks $\boldsymbol{X}$ while remaining as close as possible to the design weights d_i. For the linear distance function (5) introduced above, the GREG estimator gives the weights

$$w_i^{\mathrm{GR}} = d_i[1 + (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}})\boldsymbol{C}^{-1}\boldsymbol{x}_i'/c_i] \text{ for } \boldsymbol{C} = \Sigma_{i\in\mathrm{S}}\, d_i \boldsymbol{x}_i'\boldsymbol{x}_i/c_i \qquad (13)$$

This solution unfortunately gives biased estimates, since it allows the weight adjustment ratio w_i^{GR}/d_i to depend on the Census response $\boldsymbol{x}_i$. One problematic result of this is that units with zero contribution to any Census category do not have their weights adjusted at all, since, whatever value of c_i these units are assigned, setting $w_i = d_i$ for these units minimises the distance function without affecting the property $\hat{\boldsymbol{x}}^{\mathrm{GR}} = \boldsymbol{X}$.

In simple terms, the basic GREG estimator has no 'incentive' to adjust the weights of PES records i which the Census failed to count, so they retain their simple dwelling weight d_i. This leads to a bias, since the weight adjustment was intended to account for the fact that the PES only covers people in private dwellings when the PES is run, not the whole population, and it seems likely that the need for an adjustment applies equally to any units not counted in the Census.

The estimator $\hat{t}^{GR} = \Sigma_{i \in S} w_i^{GR} t_i$ thus has a downward bias, in that the units in the uncovered portion of U (those not in U^C) are represented only by adding weight to PES units with $\boldsymbol{x}_i \neq \mathbf{0}$. It is likely that some units in the uncovered portion will in fact have $\boldsymbol{x}_i = \mathbf{0}$ as well as non-zero t_i. By not accounting for this the estimates $\hat{t}^{GR}$ will understate the true population total T.

5.3 A penalised GREG estimator

One sensible alternative to the GREG estimator is to apply the GREG weighting above for units with $\boldsymbol{x}_i \neq \mathbf{0}$, but to adjust the weights of units with $\boldsymbol{x}_i = \mathbf{0}$ and $\boldsymbol{y}_i \neq \mathbf{0}$ by the same proportionate amount as other units i^* that had $\boldsymbol{x}_{i^*} = \boldsymbol{y}_i$. This can be achieved by replacing $\boldsymbol{x}_i$ with a value $\boldsymbol{x}_i^* = \delta \boldsymbol{y}_i$ whenever $\boldsymbol{x}_i = \mathbf{0}$ (with $\boldsymbol{x}_i^* = \boldsymbol{x}_i$ otherwise) and applying a penalty $c_i = \delta$ to these units. Here δ is some small enough value that the units have negligible impact on the weighting applied to other units (e.g. $\delta = 10^{-12}$). This approach 'tricks' the GREG into adjusting the weights of these units with $\boldsymbol{x}_i = \mathbf{0}$, since the distance penalty for adjusting the weights is commensurate with the tiny size of their $\boldsymbol{x}_i^*$ values.

In the post-stratified case, this weighting will adjust all units in a post-stratum by the same ratio. It thus reproduces the DSE in this case. In the more general setting we will refer to the estimator as the penalised GREG estimator $\hat{\boldsymbol{y}}^{PG}$. The penalty for a unit is its total contribution in persons to the Census, or the value δ if this contribution is zero. Applying this penalty should reduce the bias of the GREG approach, since some portion of the uncovered population (units in U but not in U^C) will now be represented by units in the sample with $\boldsymbol{x}_i = \mathbf{0}$.

5.4 The PREG estimator obtained by applying GREG to predicted values

The basic problem with both GREG and penalised GREG approaches is that the weight applied to a unit depends to some extent on its Census responses $\boldsymbol{x}_i$. This leads to a biased estimate, although the bias is smaller for the penalised GREG. To avoid the dependence on $\boldsymbol{x}_i$, consider applying GREG to predictions $\hat{\boldsymbol{z}}_i$ of the $\boldsymbol{x}_i$ values, based only on the PES variables $\boldsymbol{y}_i$.

Assume a linear model $E(\boldsymbol{x}_i) = \boldsymbol{y}_i \boldsymbol{Z}$, so that the expected Census response for a PES respondent under the model is a linear combination of the true values $\boldsymbol{y}_i$. We can estimate $\boldsymbol{Z}$ as the parameters of a linear regression of $\boldsymbol{x}_i$ on $\boldsymbol{y}_i$ using weight d_i/c_i, where the regression includes all units i in PES. This gives the matrix of parameter estimates

$$\hat{\boldsymbol{Z}} = (\Sigma_{i \in S} d_i \boldsymbol{y}_i' \boldsymbol{y}_i / c_i)^{-1} \Sigma_{i \in S} d_i \boldsymbol{y}_i' \boldsymbol{x}_i / c_i = \boldsymbol{A}^{-1} \boldsymbol{B} \qquad (14)$$

This model result can be applied to all units to give predicted Census responses $\hat{\boldsymbol{z}}_i = \boldsymbol{y}_i \hat{\boldsymbol{Z}}$ for all units in the PES sample. In a simple post-stratified situation, the predictions for all units in the post-stratum would be equal, set to the proportion of the units in the post-stratum that were counted in the Census. (In this case all units in a post-stratum will get the same adjustment i.e. this gives the DSE.)

Generalised regression can then be used to calibrate the weights so that these predicted Census responses $\hat{\boldsymbol{z}}_i$ aggregate to the Census totals i.e. $\hat{\boldsymbol{z}}^{\mathrm{Z}} = \Sigma_{i \in \mathrm{S}} w_i^{\mathrm{Z}} \hat{\boldsymbol{z}}_i = \boldsymbol{X}$. This prevents any dependence of the weight of a unit on its individual $\boldsymbol{x}_i$ value (except to the extent that the unit's values influence the fitting of the model). This turns out to produce the PREG estimator, provided that there is a linear combination $\boldsymbol{\gamma}$ for which $\boldsymbol{\gamma} \boldsymbol{y}_i' = c_i$ for all units i (a condition discussed previously for the PREG). Under this condition, $\hat{\boldsymbol{x}}^{\mathrm{D}} = \Sigma_{i \in \mathrm{S}} d_i \boldsymbol{x}_i = \boldsymbol{\gamma} \Sigma_{i \in \mathrm{S}} d_i \boldsymbol{y}_i' \boldsymbol{x}_i / c_i = \boldsymbol{\gamma} (\Sigma_{i \in \mathrm{S}} d_i \boldsymbol{y}_i' \boldsymbol{y}_i / c_i) \hat{\boldsymbol{Z}} = \boldsymbol{\gamma} \Sigma_{i \in \mathrm{S}} d_i \boldsymbol{y}_i' \hat{\boldsymbol{z}}_i / c_i = \hat{\boldsymbol{z}}^{\mathrm{D}}$, and the GREG estimator applied to $\hat{\boldsymbol{z}}_i$ gives the PREG estimator.

$$
\begin{aligned}
w_i^{\mathrm{PR}} &= d_i [1 + (\boldsymbol{X} - \hat{\boldsymbol{z}}^{\mathrm{D}}) (\Sigma_{i \in \mathrm{S}} d_i \hat{\boldsymbol{z}}_i' \hat{\boldsymbol{z}}_i / c_i)^{-1} \hat{\boldsymbol{z}}_i' / c_i] \\
&= d_i [1 + (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}}) (\boldsymbol{B}' \boldsymbol{A}^{-1} \boldsymbol{B})^{-1} \boldsymbol{B}' \boldsymbol{A}^{-1} \boldsymbol{y}_i' / c_i] \qquad \text{as at (12)}
\end{aligned}
$$

5.5 Relationship to instrumental variables regression

The standard GREG estimator of the total T can be rewritten to show its relationship to a regression parameter relating the variable of interest t_i to the auxiliary variable $\boldsymbol{x}_i$, as follows.

$$
\hat{t}^{\mathrm{GR}} = \hat{t}^{\mathrm{D}} + (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}}) \hat{\boldsymbol{\beta}}^{\mathrm{GR}} \qquad \text{for } \hat{\boldsymbol{\beta}}^{\mathrm{GR}} = \boldsymbol{C}^{-1} (\Sigma_{i \in \mathrm{S}} d_i \boldsymbol{x}_i' t_i / c_i)
$$

Correspondingly, the PREG estimator can be rewritten as based on an instrumental variables regression relating t_i to $\boldsymbol{x}_i$ using the vector $\boldsymbol{y}_i$ as the instrumental variables. This gives the following expression, with $\hat{\boldsymbol{\beta}}^{\mathrm{PR}}$ the instrumental variables regression parameter. A useful reference discussing instrumental variables estimation is Hall (1993).

$$
\hat{t}^{\mathrm{PR}} = \hat{t}^{\mathrm{D}} + (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}}) \hat{\boldsymbol{\beta}}^{\mathrm{PR}} \qquad \text{for } \hat{\boldsymbol{\beta}}^{\mathrm{PR}} = (\boldsymbol{B}' \boldsymbol{A}^{-1} \boldsymbol{B})^{-1} \boldsymbol{B}' \boldsymbol{A}^{-1} (\Sigma_{i \in \mathrm{S}} d_i \boldsymbol{y}_i' t_i / c_i)
$$

6. VARIANCE ESTIMATOR FOR THE PREG ESTIMATOR

6.1 Linearising the estimator

The PREG estimate is a complex combination of sample estimates. The linearisation approach will be used to obtain a variance estimator for the PREG estimate. Writing $\boldsymbol{U} = \Sigma_i \, d_i \boldsymbol{y}_i' t_i / c_i$, the PREG estimator for an item t_i is

$$\hat{t}^{\mathrm{PR}} \quad = \hat{t}^{\mathrm{D}} + (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}})(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{U} \tag{16}$$

The values $\hat{t}^{\mathrm{D}}, \hat{\boldsymbol{x}}^{\mathrm{D}}, \boldsymbol{A}, \boldsymbol{B}$ and $\boldsymbol{U}$ are sample estimates. Write $\tilde{\boldsymbol{A}} = \mathrm{E}(\boldsymbol{A})$, $\tilde{\boldsymbol{B}} = \mathrm{E}(\boldsymbol{B})$, $\tilde{\boldsymbol{U}} = \mathrm{E}(\boldsymbol{U})$ and $\boldsymbol{X}^{\mathrm{D}} = \mathrm{E}(\hat{\boldsymbol{x}}^{\mathrm{D}})$, and replace $\hat{t}^{\mathrm{PR}}$ by a linear approximation based on the derivative with respect to $\hat{\boldsymbol{x}}^{\mathrm{D}}$, $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{U}$ evaluated at their expectations. A key component is an approximation for the inverse of a matrix, given by $\boldsymbol{A}^{-1} \doteq \tilde{\boldsymbol{A}}^{-1} - \tilde{\boldsymbol{A}}^{-1}(\boldsymbol{A} - \tilde{\boldsymbol{A}})\tilde{\boldsymbol{A}}^{-1}$. The whole linear approximation for $\hat{t}^{\mathrm{PR}}$ follows.

$$\begin{aligned}
\hat{t}^{\mathrm{PR}} \quad \doteq \; & \underset{(1)}{\hat{t}^{\mathrm{D}}} + (\boldsymbol{X} - \underset{(3)}{\hat{\boldsymbol{x}}^{\mathrm{D}}})(\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}} + (\boldsymbol{X} - \boldsymbol{X}^{\mathrm{D}})(\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}})^{-1}\{ \\
& -[\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}(\underset{(4)}{\boldsymbol{B}} - \tilde{\boldsymbol{B}}) + (\underset{(5)}{\boldsymbol{B}} - \tilde{\boldsymbol{B}})'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}} - \tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}(\underset{(6)}{\boldsymbol{A}} - \tilde{\boldsymbol{A}})\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}}'](\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}} \\
& +(\underset{(5)}{\boldsymbol{B}} - \tilde{\boldsymbol{B}})'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}} - \tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}(\underset{(6)}{\boldsymbol{A}} - \tilde{\boldsymbol{A}})\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}} + \tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}(\underset{(2)}{\boldsymbol{U}} - \tilde{\boldsymbol{U}}) \; \}
\end{aligned}$$

Terms in this approximation corresponding to constants can be ignored in taking the variance. The non-constant terms have been numbered underneath to show where they appear in the variance expression that follows.

$$\begin{aligned}
\mathrm{var}(\hat{t}^{\mathrm{PR}}) \quad & \doteq \mathrm{var}([\underset{(1)}{\hat{t}^{\mathrm{D}}} + (\boldsymbol{X} - \boldsymbol{X}^{\mathrm{D}})(\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\underset{(2)}{\boldsymbol{U}}] \\
& \quad - [\underset{(3)}{\hat{\boldsymbol{x}}^{\mathrm{D}}} + (\boldsymbol{X} - \boldsymbol{X}^{\mathrm{D}})(\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\underset{(4)}{\boldsymbol{B}}](\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}} \\
& + (\boldsymbol{X} - \boldsymbol{X}^{\mathrm{D}})(\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}}^{-1})^{-1}(\underset{(5)}{\boldsymbol{B}}' - \tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\underset{(6)}{\boldsymbol{A}})(\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}} - \tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}}(\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}}) \;)
\end{aligned} \tag{17}$$

Consider the third line of this expression. The vector $\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}}$ is a regression parameter for predicting t_i from y_i directly, while the vector $\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}}(\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{B}})^{-1}\tilde{\boldsymbol{B}}'\tilde{\boldsymbol{A}}^{-1}\tilde{\boldsymbol{U}}$ is the regression parameter of an 'instrumental variable' regression, in which $\boldsymbol{y}_i$ are used to make predictions $\boldsymbol{z}_i$ of the instrumental variables $\boldsymbol{x}_i$, and the t_i are then predicted based on these $\boldsymbol{z}_i$. These two regression parameters are identical if $\tilde{\boldsymbol{B}}$ is invertible (i.e. if there is a $\tilde{\boldsymbol{B}}^{-}$ for which $\tilde{\boldsymbol{B}}^{-}\tilde{\boldsymbol{B}}$ is the identity matrix), in which case the third line of (17) is zero. In practical applications the third line of (17) will be negligible and can be dropped for variance calculations.

All the population values $\boldsymbol{X}^{\mathrm{D}}$, $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{B}}$ and $\tilde{\boldsymbol{U}}$ are unknown. For the purpose of variance estimation they will be replaced by their sample estimates, which are treated as constants for this purpose. Performing this substitution of sample estimates for the unknown values $\boldsymbol{X}^{\mathrm{D}}$, $\tilde{\boldsymbol{A}}$, $\tilde{\boldsymbol{B}}$ and $\tilde{\boldsymbol{U}}$ and treating these as fixed leads to the *weighted residuals* variance estimator. Writing

$$
\begin{aligned}
r_i^{\mathrm{PR}} &= 1 + (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}})(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{y}_i'/c_i, \\
\hat{\gamma} &= (\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{U}, \\
\theta^{\mathrm{X}} &= (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}})(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}, \\
\Delta^{\mathrm{T}} &= \boldsymbol{A}^{-1}(\boldsymbol{U} - \boldsymbol{B}\hat{\gamma})
\end{aligned}
$$

and recalling

$$
\hat{\theta}^{\mathrm{PR}} = (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}})(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{B}'\boldsymbol{A}^{-1}
$$

the variance becomes

$$
\begin{aligned}
&\mathrm{var}(\hat{t}^{\mathrm{PR}}) \\
&\quad \doteq \mathrm{var}(\Sigma_{i\in hg}\, d_i r_i^{\mathrm{PR}}(t_i - \boldsymbol{x}_i\hat{\gamma}) + (\boldsymbol{X} - \hat{\boldsymbol{x}}^{\mathrm{D}})(\boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}[d_i(\boldsymbol{x}_i' - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{y}_i')\boldsymbol{y}_i/c_i]\boldsymbol{A}^{-1}(\boldsymbol{U} - \boldsymbol{B}\hat{\gamma})) \\
&\quad = \mathrm{var}(\Sigma_{i\in hg}\, d_i[r_i^{\mathrm{PR}}(t_i - \boldsymbol{x}_i\hat{\gamma}) + (\theta^{\mathrm{X}}\boldsymbol{x}_i' - \theta^{\mathrm{PR}}\boldsymbol{y}_i')\boldsymbol{y}_i\Delta^{\mathrm{T}}/c_i]) \qquad (18)
\end{aligned}
$$

This form, which incorporates the term in the third line of (17), was used in the evaluations presented in this paper. In all these evaluations this term proved to be entirely negligible, as was expected given the discussion above. Thus the following simpler form is recommended for variance calculations for the PREG estimator in the PES situation.

$$
\mathrm{var}(\hat{t}^{\mathrm{PR}}) \quad \doteq \mathrm{var}[\Sigma_i\, d_i r_i^{\mathrm{PR}}(t_i - x_i\hat{\gamma})] \qquad (19)
$$

6.2 The weighted residuals variance estimator

For the purpose of variance estimation the weight adjustments r_i^{PR} and the vector $\hat{\gamma}$ are treated as known and fixed. If the initial weights d_i are also fixed then (19) can be used to approximate the variance of the PREG estimator by the variance of a simple linear estimator, using a variance estimator appropriate for the sample design of the survey. The sample design of the PES is multi-stage within strata. Variance estimation proceeds by forming variance groups from the first-stage sampling units in each stratum—typically a variance group for each primary sampling unit (PSU).

Let G_h be the number of variance groups in stratum h. The weighted residual for a variance group g is given by

$$e_{hg} = \Sigma_{i \in hg} w_i^{\mathrm{PR}} (t_i - \boldsymbol{x}_i \hat{\gamma}) \tag{20}$$

The weighted residuals variance estimator is then computed as

$$\mathrm{v}^{\mathrm{WR}}(\hat{t}^{\mathrm{PR}}) = \sum_h \frac{G_h}{G_h - 1} \sum_{g=1}^{G_h} \left(e_{hg} - \bar{e}_h\right)' \left(e_{hg} - \bar{e}_h\right) \text{ for } \bar{e}_h = \frac{1}{G_h} \Sigma_g \, e_{hg} \tag{21}$$

A comparable expression to (20) gives the weighted residuals variance estimator for the GREG (and penalised GREG) estimators. Writing

$$\hat{\beta}^{\mathrm{GR}} = (\Sigma_{i \in \mathrm{S}} \, d_i \boldsymbol{x}_i' \boldsymbol{x}_i / c_i)^{-1} \, \Sigma_{i \in \mathrm{S}} \, d_i \boldsymbol{x}_i' t_i / c_i$$

the weighted residual for the GREG estimator takes the form

$$e_{hg}^{\mathrm{GR}} = \Sigma_{i \in hg} w_i^{\mathrm{GR}} (t_i - \boldsymbol{x}_i \hat{\beta}^{\mathrm{GR}}) \tag{22}$$

6.3 Variance for multiple steps of weighting

If the d_i are the result of a previous estimation step then they are not fixed values. Theoretically this may induce some additional bias in the above variance estimator. In this paper this aspect of the approximation has been assumed to have negligible impact.

The alternative would be to calculate the values $v_i = r_i^{\mathrm{PR}} (t_i - x_i \hat{\gamma})$ for all units in the sample, and then estimate the variance of $\hat{v}^{\mathrm{D}} = \Sigma_i \, d_i v_i$ using the appropriate weighted residuals variance estimator for the estimator used at the previous step. From (19) it follows that this is approximately the variance of $\hat{t}^{\mathrm{PR}}$.

In the PES situation, the dwelling weighting is itself a DSE. The overall variance estimator would take the same form (21) but with a slightly modified weighted residual taking the form $e_{hg}^{V} = \Sigma_{i \in hg} w_i^{\mathrm{PR}} (t_i - \boldsymbol{x}_i \hat{\gamma} - \hat{v}_i)$, in which $\hat{v}_i$ would be a prediction of the unit's v_i value based only on the post-stratum of the unit as used in the dwelling weighting step. Given that the key PES items of interest t_i are unlikely to be well predicted by these post-strata, this more complicated variance estimator will have expectation nearly identical to that of the proposed simpler estimator. This reasoning justifies the decision to use the variance estimator based on the PREG step only, as given by (20) and (21).

The use of two steps of PREG weighting as described in section 4.8 does, however, need to be accounted for in variance estimation. Recall that the second step of weighting is only a minor adjustment to the weights from the first step, so as to represent private and non-private dwellings rather than private dwellings only. Furthermore the predictor variables used at the second step are a subset of those used at the first step (being those classified by region, sex and age group only).

In order to approximate the full effect of both steps of weighting, the $\boldsymbol{y}_i$ vector from the first step is used, with the $\boldsymbol{x}_i$ elements from the first stage used for any elements of $\boldsymbol{y}_i$ not used at the second stage. These vectors are then used along with the dwelling weight d_i in calculating the component $\boldsymbol{x}_i\hat{\gamma}$ of the weighted residual formula (20). The weight w_i^{PR} used in this formula is the final PES weight from the second step of weighting.

Clearly there could be alternative choices in defining a variance estimator for the PES estimates. Experimentation shows that using the predictor variables from the second step only would give almost identical variance estimates on under-count rates for a variety of categories, and for estimates of population by region, sex or age group. Using the approach described above, however, gives lower (and more realistic) variance estimates for estimates of the population of other categories used in weighting at the first stage, such as estimates of indigenous persons.

6.4 Variance for a ratio of two estimates

A weighted residuals estimator can also be obtained for a function of two or more estimated totals. In the PES, a key quantity is the Census coverage ratio, which is the number of people who were counted for a category divided by the number of people who should have been counted in that category. This takes the form of a ratio of two estimates.

A ratio T_1/T_2 of two totals is estimated by $\hat{t}_1^{\mathrm{PR}}/\hat{t}_2^{\mathrm{PR}}$. Its variance can be estimated using the following approximation.

$$\begin{aligned}\mathrm{var}(\hat{t}_1^{\mathrm{PR}}/\hat{t}_2^{\mathrm{PR}}) &\doteq \mathrm{var}((\hat{t}_1^{\mathrm{PR}} - \frac{T_1}{T_2}\hat{t}_2^{\mathrm{PR}})/T_2) \\ &\doteq \mathrm{var}(\Sigma_i\, w_i^{\mathrm{PR}}\left(t_{1i} - x_i\hat{\gamma}_1 - \frac{T_1}{T_2}(t_{2i} - x_i\hat{\gamma}_2)\right)/T_2) \text{ from (19)} \qquad (23)\end{aligned}$$

Here t_{1i} and t_{2i} are two item values for unit i and $\hat{\gamma}_1$ and $\hat{\gamma}_2$ are the $\hat{\gamma}$ values for the two items. The variance is then computed by replacing T_1 and T_2 by their estimates (treated as fixed quantities) and using the weighted residuals variance estimator.

7. A FRAMEWORK FOR SIMULATING THE DRAWING OF PES SAMPLES

Evaluation of the PREG estimator against alternatives was conducted based on a simulated population framework. The 2001 Census unit data was used as the basis for simulating the repeated drawing of PES-like samples. By drawing a large set of PES-like samples, the bias of alternative PES estimators and accompanying variance estimators could be assessed.

7.1 Setting up the PES population framework

The 2001 Census data was treated as the augmented population U as described in Section 2. The dwellings appearing on U were treated as containing persons at PES time, with the values reported on the 2001 Census providing their demographic characteristics *at PES time*. The vector $\boldsymbol{u}_i$ gives numbers of persons in various categories for each dwelling i in U. Corresponding counts $\boldsymbol{u}_{ij}$ give the contribution to $\boldsymbol{u}_i$ of each person j in U. For these simulations, $\boldsymbol{u}_i$ included counts of persons in various cross classifications of 14 regions of usual residence, 2 sexes, 16 age groups (5-year age groups between 0 and 74 and a group for 75+), 2 indigenous statuses, 2 marital status groups (married and other) and 2 country of birth groups (born in Australia and born overseas). Counts for 448 region–sex–age groups and a count for Indigenous were essential for the weighting process.

From the available information on the $\boldsymbol{u}_i$ vector, values for two vectors $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ for each dwelling i in U were generated. Corresponding counts $\boldsymbol{y}_{ij}$ give the contribution to $\boldsymbol{y}_i$ of each person j in dwelling i, and similarly for $\boldsymbol{x}_{ij}$. For these simulations, $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ contained counts of persons by the 448 region–sex–age groups and a count for Indigenous. The values in $\boldsymbol{x}_i$ are of persons counted in the Census from that category, while the values in $\boldsymbol{y}_i$ are of how many persons should have been counted from that category. In addition, $\boldsymbol{y}_i$ had an extra element that took the value one if the remaining elements of $\boldsymbol{y}_i$ were all zero.

7.2 Steps in fitting models based on PES survey data

The 2001 PES survey file contains a sample of dwellings that contained at least one person at PES time, and demographics for each such person. This file provided data to derive models that were then used to randomly generate the data $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ for each PES dwelling in U.

Step 1: Determine explanatory variables

An exploratory data analysis was used to determine the set of explanatory variables $\boldsymbol{u}^*_{ij}$ at person level and $\boldsymbol{u}^*_i$ at dwelling level to use in the modelling. In choosing $\boldsymbol{u}^*_{ij}$ a variety of interaction effects were explored using a logistic regression for the

probability of a PES person contributing to Census. The variables $\boldsymbol{u}_i^*$ were simply the totals of the chosen person indicators at dwelling level.

Using the chosen predictors, the values $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ for dwellings in U were generated based on models developed from the 2001 PES data. Most of the modelling work involved logistic regression models, all of which were done unweighted. This is defensible here as the focus is on estimating relationships rather than on reproducing a population total. Programs to fit logistic models with random effects typically perform an unweighted analysis.

Step 2: Generate whether (any persons in) the dwelling should contribute to Census

A logistic model was used for this step.

Step 3: Generate whether each person should contribute, if their dwelling should

For dwellings with multiple persons a logistic model was used to obtain probabilities of this item for each person. Random generation of each individual person's value was performed in a manner that achieved the modelled probabilities while ensuring that at least one person was generated for the dwelling.

Step 4: Generate whether (any persons in) a dwelling did contribute to the Census

For dwellings that should have contributed to the census, a logistic model was fitted to the 2001 data for whether they did in fact contribute. The model allowed for a random intercept common to all persons in a Census collector's district (CD); this proved to be significant. This model was used to generate this item. A separate, simple logistic model was used to generate a small number of dwellings contributing even though they should not have done so.

Step 5: Generate whether a person contributed given that their dwelling did

A logistic model was used to obtain probabilities that an individual in a multi-person dwelling did contribute to the Census, given that the dwelling did, with separate models for persons who should have contributed and those who should not have. The model for persons who should have contributed included a random effect at the dwelling level, to account for the fact that if one person contributes there is an increased likelihood of others in their dwelling contributing.

Step 6: Generate whether a person (who contributed) contributed twice

A logistic model was used for this item.

Step 7: Finalise the simulated population by allowing for misclassification

The final values of $\boldsymbol{y}_{ij}$ were copies of the appropriate elements of $\boldsymbol{u}_{ij}$ for persons that should have been counted, and zero otherwise. The values of $\boldsymbol{y}_i$ are sums of $\boldsymbol{y}_{ij}$ to dwelling level.

Separate logistic models were used for region, sex, age and indigenous status to generate whether each person who was counted was given a different category than that given in the elements of $\boldsymbol{u}_{ij}$. Where this occurred, misclassification was assumed to be to an adjacent category, except for region, where the distribution of misclassified region values observed in the 2001 PES was used to randomly assign a region value. The final values of $\boldsymbol{x}_{ij}$ were obtained by assigning the number of times the person was counted to the elements of the vector corresponding to the person's categories (from $\boldsymbol{u}_{ij}$ or generated). The values of $\boldsymbol{x}_i$ are sums of $\boldsymbol{x}_{ij}$ to dwelling level.

The resulting population U

The result is a population for which the true totals $\boldsymbol{Y} = \Sigma_{i \in \mathrm{U}} \boldsymbol{y}_i$ are known, as well as the totals $\boldsymbol{X} = \Sigma_{i \in \mathrm{U}} \boldsymbol{x}_i$ observed for the population in the Census. Both variables show low intra-class correlation coefficients at CD level (0.01 for both $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$). They show much higher coefficients at dwelling level (0.27 for $\boldsymbol{y}_i$, 0.31 for $\boldsymbol{x}_i$). This reflects the fact that persons who do not contribute to Census frequently belong to dwellings in which no person contributed, and similarly persons who should not have contributed to Census belong to dwellings in which no person should have contributed.

7.3 Determining coverage and response for PES simulations

In PES estimation, inference about the whole population U is made based on a probability sample from U^{C}, the covered, responding portion of the population. Simulations were performed for U^{C} generated under three different coverage-response models.

For the purpose of modelling on 2001 PES data, coverage-response probabilities were estimated as the inverse of the response coverage adjustment required to adjust the selection weights of units to final weights that reproduce a set of Census benchmarks. The PREG estimator was used for this modelling. The resulting models were adjusted in an ad hoc manner to ensure that probabilities lie in the range (0,1).

Only the overall features of the models will be presented here. The aim is to produce a framework for simulation in which the PREG estimator can be compared to various alternatives. This does not require that the simulations correspond precisely to the situation in PES – for instance, the estimators tested will not have a separate stage of dwelling weighting.

Model 1: Inclusion based on a dwelling-level model

Coverage-response probabilities for dwellings were modelled as a function of the number of persons in the dwelling who should have been counted, with probabilities depending on sex, 16 age groups and indigenous status. Inclusion in U^C was then generated independently for each dwelling with the modelled probability.

Model 2: Inclusion based on a person-level model

Coverage-response probabilities for persons were modelled in a similar manner to model 1 but with person as the unit. This person-level model is motivated the possibility of dwellings being only partially covered or partially responding in the PES. This simulation thus tests, for example, whether an estimator which assumes a dwelling-level coverage-response model will perform adequately even in this extreme case in which individuals are included in U^C independently.

This model is unrealistic, however, since in reality much non-response and non-coverage is driven by factors applying for a whole dwelling.

Model 3: A combined person and dwelling-level response model

A more sophisticated model was set up, using a combination of person-level and dwelling-level models. The dwelling level component of the coverage-response model was based around model 1, but effects were then introduced for the following items: born outside of Australia, enumerated in a non-private dwelling, enumerated in an indigenous community, enumerated outside state of usual residence (all decreasing coverage-response probability), and married (increasing coverage-response probability. The value of these parameters were chosen subjectively. An additional random effect was introduced for each census collector's district (CD).

This dwelling-level model was used to generate membership of U^C for 80% of the population U, with the remaining 20% of the population generated by the person-level model 2. Which of the two coverage-response models to use was determined at random for each dwelling.

The point of this more sophisticated model is not that it corresponds closely to effects occurring in the actual Australian population, but that it is realistically complex. In particular it ensures that the actual determination of response and coverage depends on variables that have not been included in the PREG estimator that is being tested.

7.4 Drawing samples from the population framework

For each of the coverage-response situations to be examined, a single realisation of the covered, responding population U^C was generated. Given the large size of the

Census, the outcomes from different realisations of U^C can be expected to be very similar; so there was little to gain from obtaining multiple realisations for each coverage-response situation. In practice, a realisation of U^C was generated by assigning a permanent random number from U(0,1) to each unit on the population file U. Inclusion of a unit in U^C was determined by comparing its random number with the model-predicted value for the given coverage-response model.

Repeated samples were drawn from the population file U, using a multi-stage sampling scheme that mimics the one used in the actual PES. The first stage of this sampling scheme divides the Census Collector's Districts (CDs) on the framework into strata, and chooses a sample of these CDs with probability proportional to the number of dwellings in the CDs. The second stage divides the CDs into blocks, and selects a block at random. For the framework here the blocks are contiguous dwellings on the Census file; this is a crude substitute for actually identifying blocks based on physical layout of the CD as is done in the actual PES. Finally, a cluster of dwellings is selected within each selected block by skipping through the list of dwellings from a random start. The skips used are chosen to give an equal probability of selection to all dwellings in a state.

Note that this sampling scheme cannot be perfectly realistic. For example, the sampling scheme in the simulations was applied to dwellings on the Census file, which excludes vacant dwellings, whereas in the actual PES vacant dwellings are selected (but provide no persons to the sample). Cluster sizes (numbers of non-vacant dwellings selected from the same CD) were thus likely to be less variable in the simulations than in the real PES.

The appropriate subset of a sample from U was used as the corresponding sample from each realisation of U^C. This should reduce the sampling error on comparisons between estimates under the various coverage-response models. Note that different coverage-response models may result in different overall sample sizes in the simulations – this may have to be taken into account in comparing results from different coverage-response models.

7.5 Measuring bias and variance using the simulations

For each sample and each realisation of U^C, estimates and variance estimates were obtained for a variety of estimators. The average across samples of the estimates is used to measure the bias of the estimator. The variance across samples of the estimates provides an estimate of the true variance of the estimator – this is compared to the average across samples of the variance estimates to detect any bias of the proposed variance estimator. In cases where the variance estimator does not display a significant bias, the average of the variance estimates across samples is used as the best estimate of the variance of an estimator in a given situation.

8. EVALUATION OF ALTERNATIVE ESTIMATORS USING SIMULATION

8.1 Estimators evaluated using simulation

A number of estimators were investigated by simulation. For simplicity, these estimators will apply a single stage of weighting based on the selection weight (the inverse of the probability of selection of the unit). The estimators will be identified by a short code: all the codes introduced for simulations include a suffix S to signify that the are based on the selection weights.

GS: The GREG estimator applied at dwelling level

The 2001 PES weighting was applied at dwelling level, rather than person level, so that individual persons inherit their weight from their PES dwelling. This 'integrated weighting' can be used to achieve consistency between dwelling estimates and lower level estimates such as household or person estimates that are obtained using the same weights. This is not a key requirement for the 2006 PES estimator; the decision to apply weighting at dwelling level in PES 2006 will depend on the quality of the resulting estimates as investigated later in this paper.

The code GS will denote the GREG estimator at dwelling level based upon the selection weights π_i^{-1}. The adjustment categories comprised 14 geographic regions (state by part of state, with the Northern Territory and the ACT considered as single geographic regions with no part of state division for this purpose) classified by sex and 16 age groups. An additional adjustment category of Indigenous status was also used, giving 449 elements in the $\boldsymbol{x}_i$ vector and the corresponding vector $\boldsymbol{X}$ of Census counts.

For the GS estimator no penalty was applied, so $c_i = 1$ for all dwellings.

GPS: The penalised GREG estimator applied at dwelling level

The code GPS will denote the penalised GREG estimator introduced in Section 4. The penalty used was the number of times that persons in the dwelling were counted in the Census, provided that this is non-zero. Dwellings that were counted 0 times in the Census (i.e. with $\boldsymbol{x}_i = 0$) were given a penalty $c_i = \delta = 10^{-12}$, and their $\boldsymbol{x}_i$ values were replaced by $\delta \boldsymbol{y}_i$.

The GREG estimator can be expected to under-estimate totals, as discussed previously. The penalised GREG estimator will still give somewhat more weight to dwellings that fully respond in the Census than to similar dwellings that partially respond. This could result in a slight under-estimate of totals.

PS: The PREG estimator applied at dwelling level

The code PS will denote the PREG estimator at dwelling level based upon the selection weights π_i^{-1}. The $\boldsymbol{x}_i$ vector used was the same as used in the GREG estimator, containing counts of the number of times persons from the dwelling were counted in each of 449 categories. The $\boldsymbol{y}_i$ vector contained counts for how many times persons from the dwelling should have been counted in each of the categories. A further element was introduced to the $\boldsymbol{y}_i$ vector taking the value 1 for units for which all the other elements of $\boldsymbol{y}_i$ are zero i.e. for dwellings in which no persons should have been counted in the Census.

The PS estimator uses a penalty of the number of times that persons in the dwelling should have been counted in the Census, where this is positive, and a penalty $c_i = 1$ otherwise.

PS1: The PREG estimator with a penalty of one

The penalty chosen for the PS estimator was suggested in Section 3. The PS1 estimator is identical to the PS estimator except that it neglects this penalty and sets $c_i = 1$ for all dwellings.

PSP: The PREG estimator applied at person level

It is not clear that a dwelling-level coverage-response model is ideal. The code PSP will be used to denote the PREG estimator applied as above but with the unit being person rather than dwelling. For this estimator the penalty $c_i = 1$ for all persons. The model underlying the PSP assumes that individual persons in a dwelling respond and are covered independently from other persons in the same dwelling, rather than together as is assumed in the PS estimator.

DSP: The dual system estimator applied at person level

The DSE was applied at person level to the selection weights, using the 448 region by age by sex categories as post-strata. The DSE assumes that these categories do not change between the Census and the PES. This is not in fact the case, and different estimators result depending on which category is used.

The code DSP represents the DSE applied to post-strata based on the category the person reported in the Census, where this is known. For persons that were not counted in the Census their PES category is substituted. The comparison between the PSP and DSP estimators should highlight the impact of including indigenous status as an indicator of under-count.

DSPY: The dual system estimator with categories defined by PES response

The code DSPY represents the DSE applied to post-strata based on the category reported in PES. The Census counts X are treated as applying for these post-strata, even though they are actually based on Census categories. This gives a potential for significant bias if Census and PES reporting of categories is different.

In previous PES surveys, it has been the practice to define categories based on the PES responses, even though the Census totals are based on Census responses. This leads to variations on the generalised regression estimators denoted GSY and GPSY. While an evaluation of these estimators was conducted, it is not presented here – they performed less well overall than the corresponding estimators GS and GPS.

8.2 Evaluation of estimators under a dwelling-level coverage-response model (model 1)

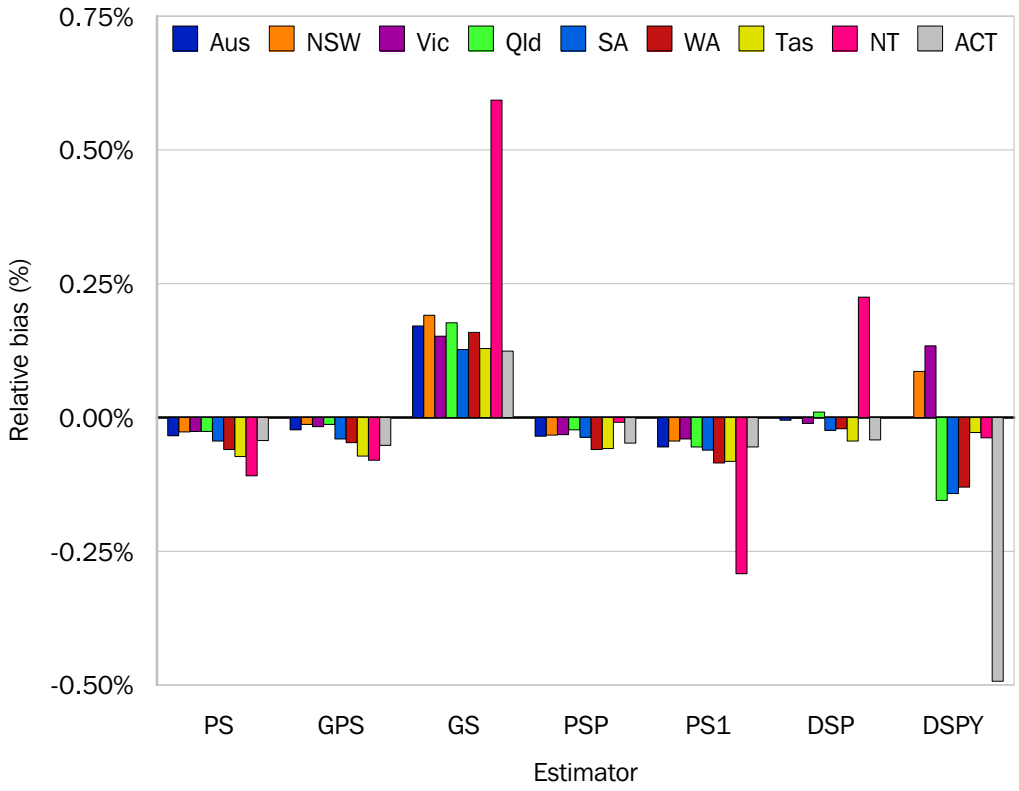
It is important to bear in mind in reading this section that all the results are from a simulated population, generated as described in Section 7. The classifications presented (Australia, states and territories etc.) are for the categories of that simulated population framework, and are affected by various arbitrary choices made to add realistic complexity to the simulated situation. The results do not correspond to what would be obtained using the methods if they were applied in the real world and with the inclusion of a dwelling weighting step (not used in these simulations).

In particular, it would be incorrect to regard any of these results as portraying the actual bias, RSE etc. for the published estimates from the 2001 PES. Even though the simulated population was built using 2001 census data, the coverage and non-response mechanisms described as Model 1, 2 and 3 are not real world facts. They were constructed to represent the type of factors that are likely to affect under-coverage and non-response to the PES, but specific details are entirely artificial.

Graph 8.1 presents a measure of the relative bias of the various estimators introduced above when measuring Census coverage ratio for Australia and the states and territories. The Census coverage ratio is the number of persons who were counted in the Census in a given category divided by the number of persons who should have been counted.

The relative bias was measured by averaging the difference between the estimate and the population value across a set of 970 simulations, and expressing this as a percentage of the population value. The simulations were conducted using coverage response model 1, so that dwellings respond or are missed in the PES sample with a probability depending on the age, sex and indigenous status of their inhabitants.

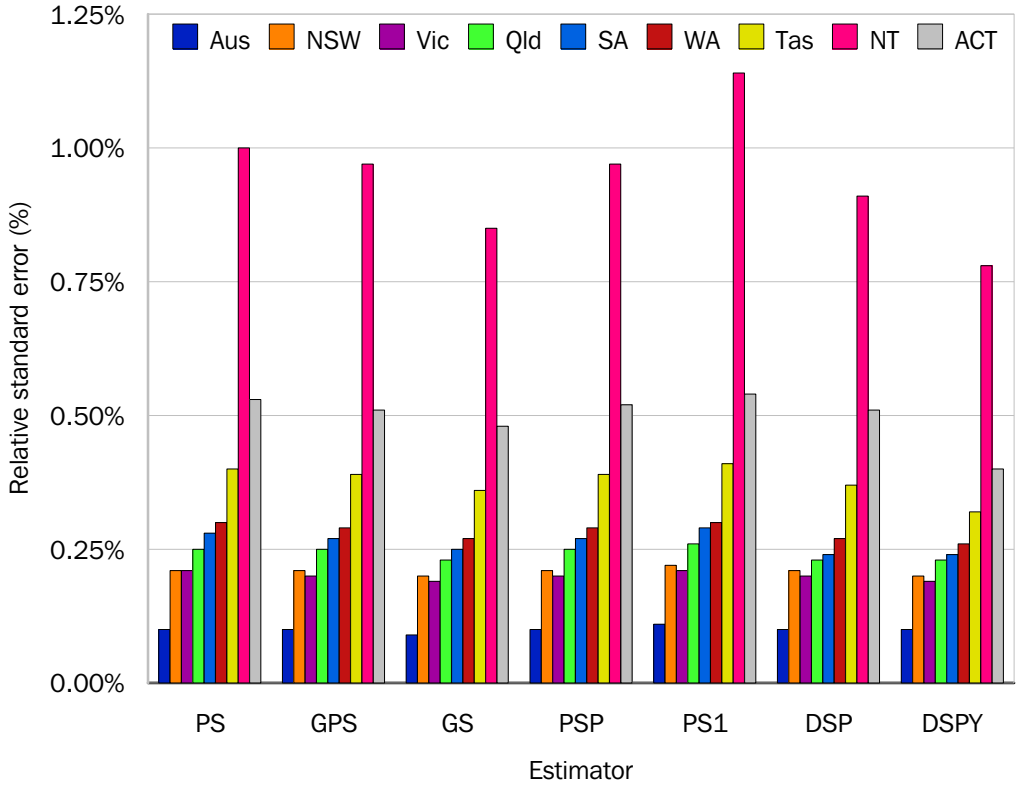
8.1 Relative bias of various Census coverage ratio estimates by state



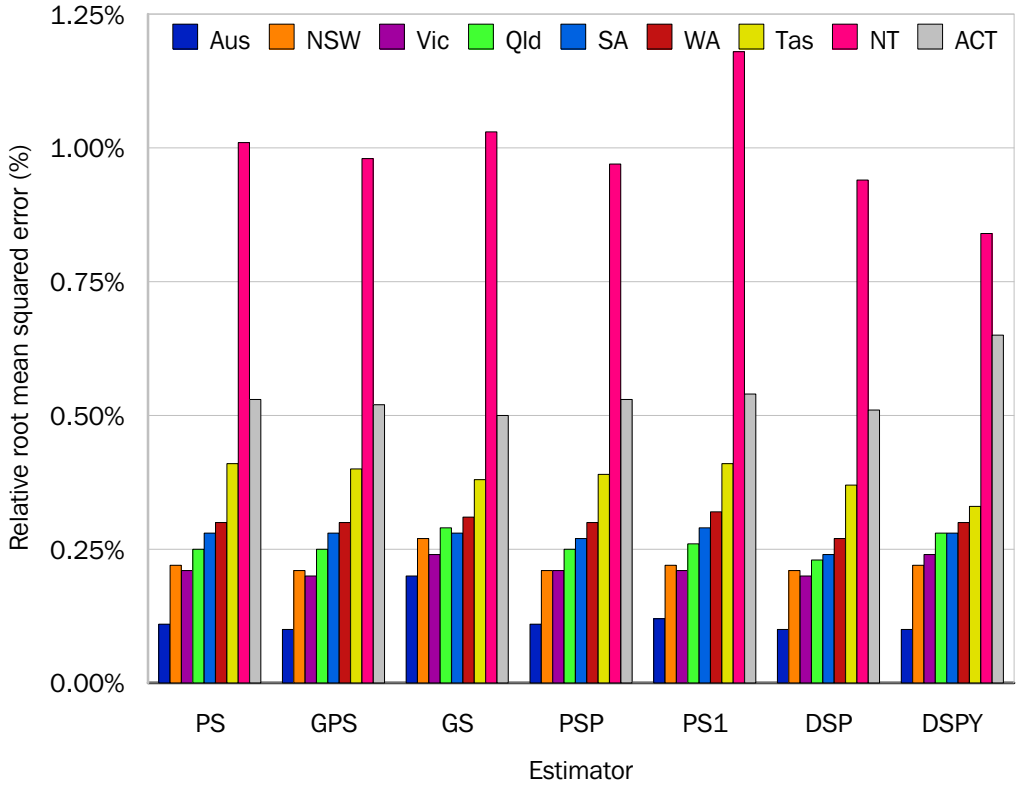
The most obvious feature in the graph is the upward bias of the GS estimator – it over-estimates how well the Census counted people, or put the other way, it leads to an under-estimate of the number of people who should have been counted in the Census. This is in line with what was expected. Other larger biases show up for individual states, where there appear to be larger biases for the PS1, DSP and DSPY estimators. The DSP bias in NT can be attributed to the failure to account for the different under-count rate of indigenous. The other three estimators PS, GPS and PSP appear to be working well across the board, with a slight downward bias.

Graph 8.2 shows the sampling error (expressed as relative standard errors) for this range of estimators, and graph 8.3 shows the relative root mean squared error (which combines the effect of the bias and sample errors). The similarity of these graphs shows that sampling error tends to overwhelm bias for most states and estimators. The clearest exception is the GS estimator, which has a large enough bias to noticeably increase the root mean squared error of estimates.

8.2 Relative standard error of various Census coverage ratio estimates by state

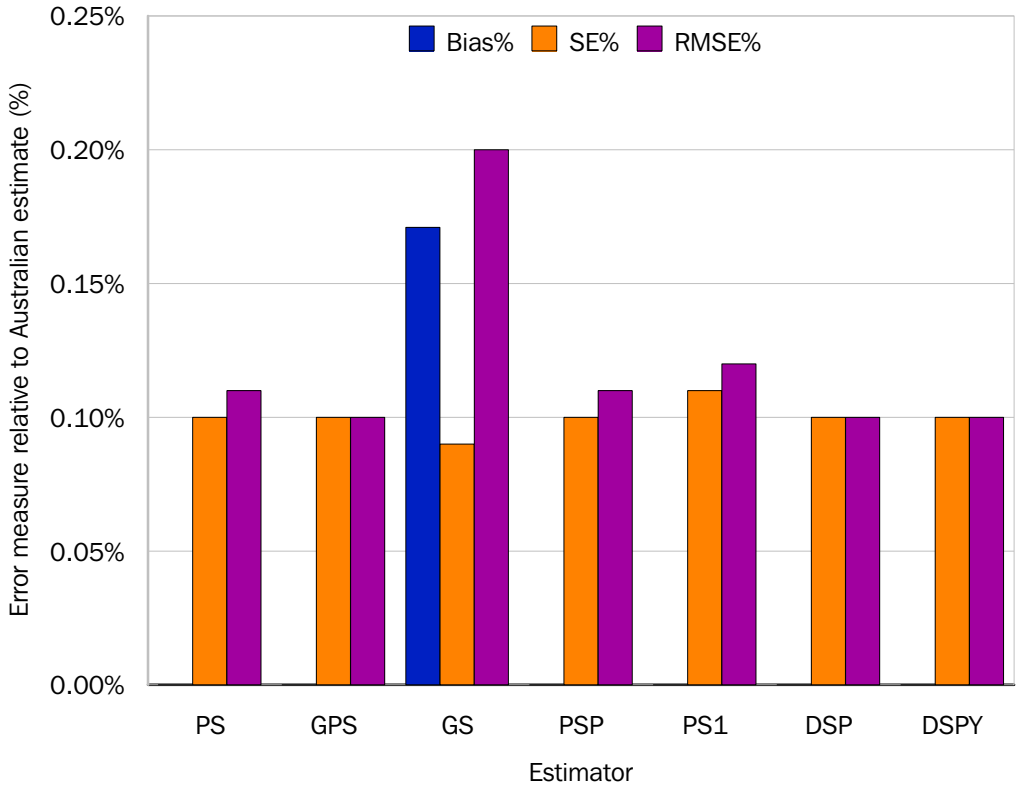


8.3 Relative root mean squared error of various Census coverage ratio estimates by state



Graph 8.4 shows (for Australian level estimates) the relative bias, SE, and root mean squared error side by side. For the GS estimator, it is clear that the bias term is the major contributor to the overall error, while for all the other estimators, where the bias is much smaller, it is the SE which drives the overall precision of the estimator.

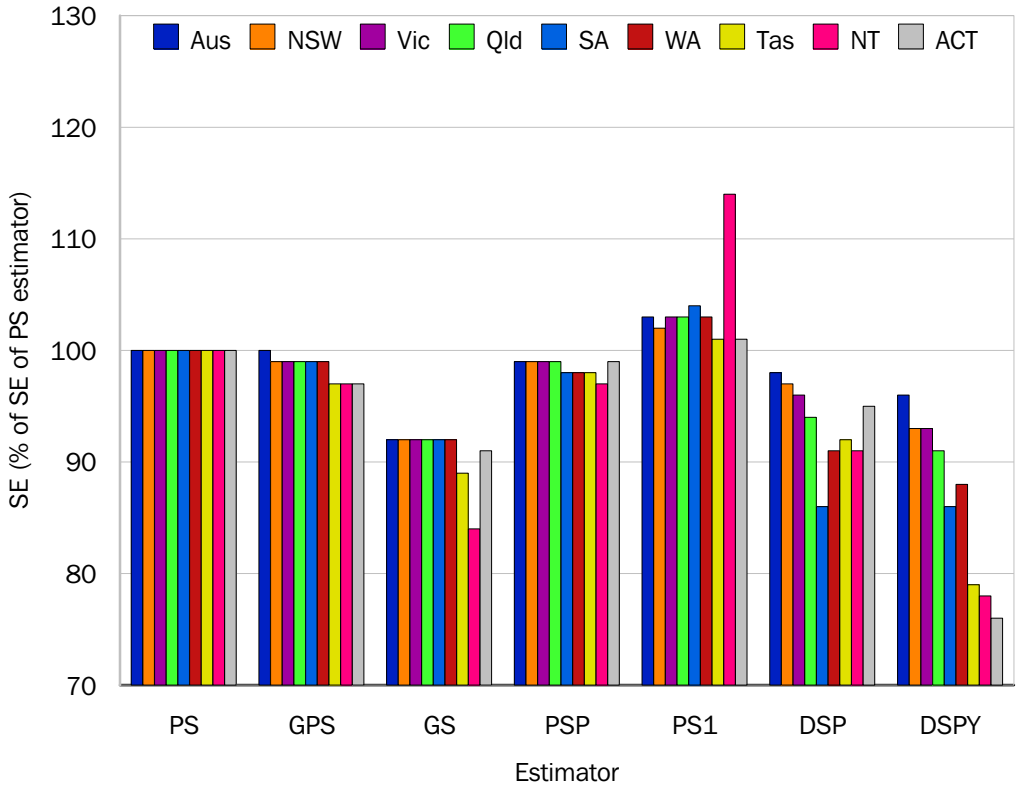
8.4 Relative Bias, Standard Error and Root mean squared error for various Census coverage ratio estimates, Australia



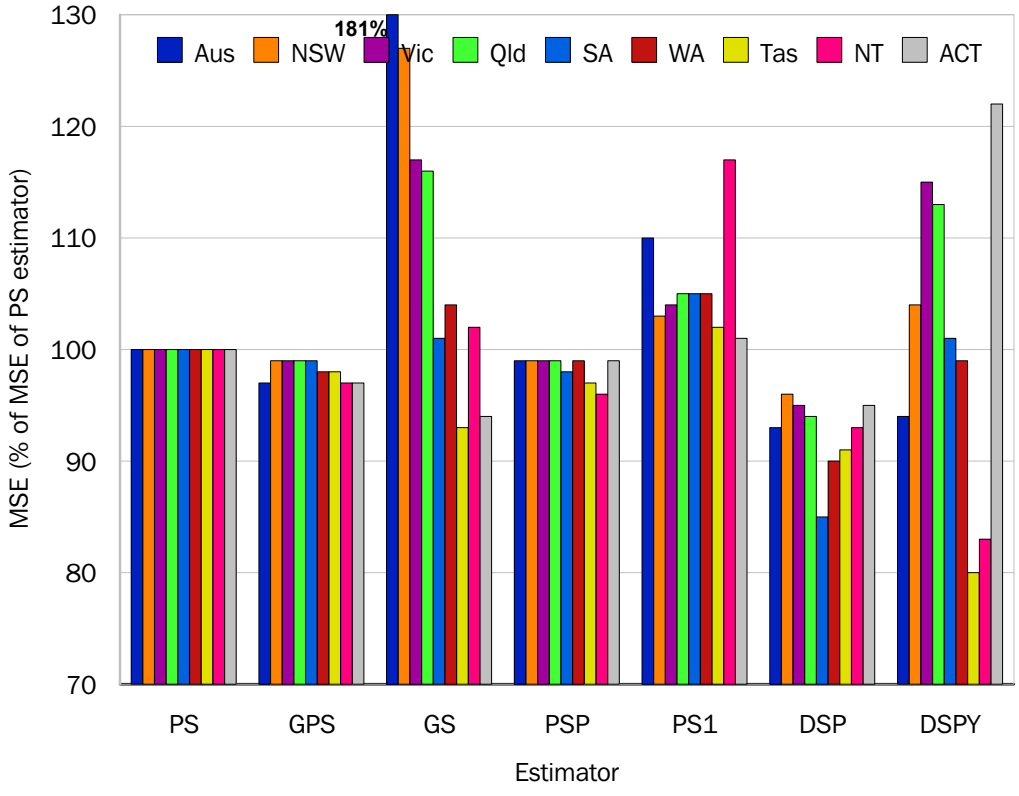
Since the main objective of this simulation study was to discover which estimator has the best overall properties, the remaining graphs in this section show the behaviour of each estimator relative to our "candidate" PREG estimator (labelled as the PS estimator in these graphs).

Graph 8.5 compares the standard errors of the estimates shown in graph 8.1 to those SE of the PS estimator. It appears that the lower bias of the PS, GPS and PSP estimators comes at the expense of a larger standard error. This rise in standard error increases the overall mean squared error (MSE) of these estimates relative to the DSP estimators. This is seen in graph 8.6.

8.5 Standard error of various Census coverage ratio estimates by state (as a percentage of the SE of the PS estimator for that state)

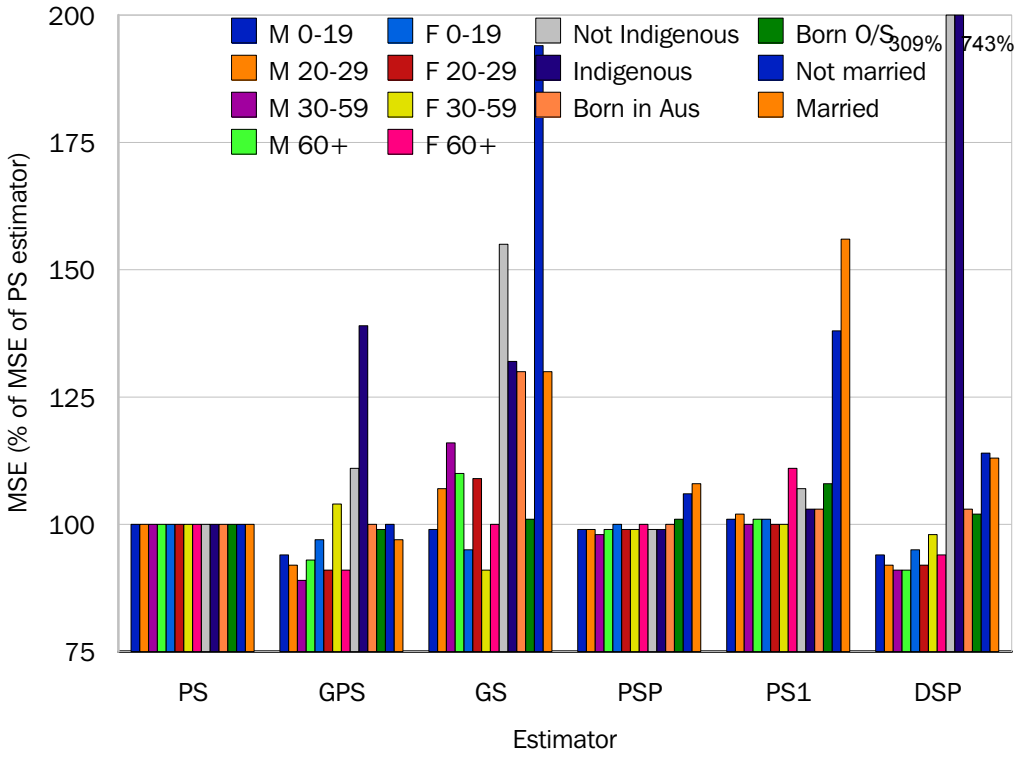


8.6 Mean squared error of various Census coverage ratio estimates by state (as a percentage of the MSE of the PS estimator for that state)



Graph 8.7 is a similar graph that compares mean squared errors of Census coverage ratio estimates by a variety of classifications other than state. (The DSPY has been omitted to allow the display of a larger number of classifications.) Fourteen classes are used: sex (denoted M or F) by four age groups (0–19, 20–29, 30–59, 60+); not indigenous, indigenous; born in Australia, born overseas; not married, married). The DSP has a large bias for indigenous status, resulting in huge increases in MSE (beyond the limits of the graph) for estimates of not indigenous (309%) and indigenous (743%). Looking across this range of estimates, the PS and PSP estimators perform well across the board, although the GPS estimator has lower MSE for most classes other than indigenous.

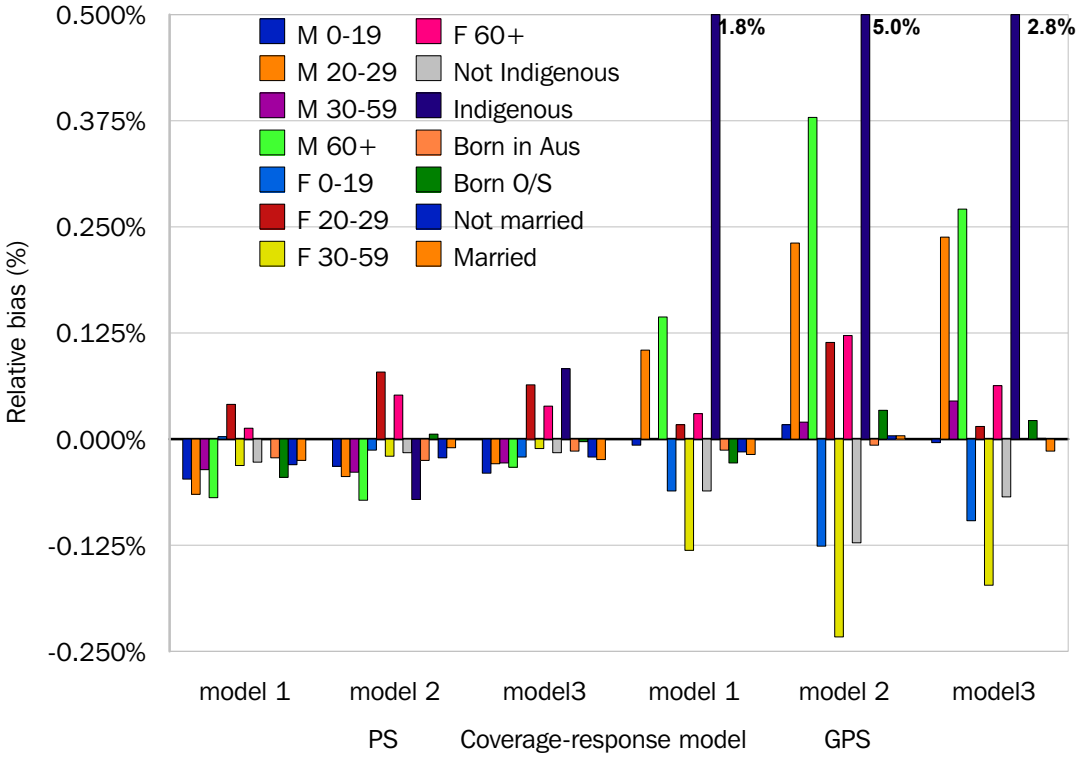
8.7 Mean squared error of Census coverage ratio estimates by various classifications (as a percentage of the MSE of the PS estimator for that classification)



8.3 Evaluation under alternative coverage-response models

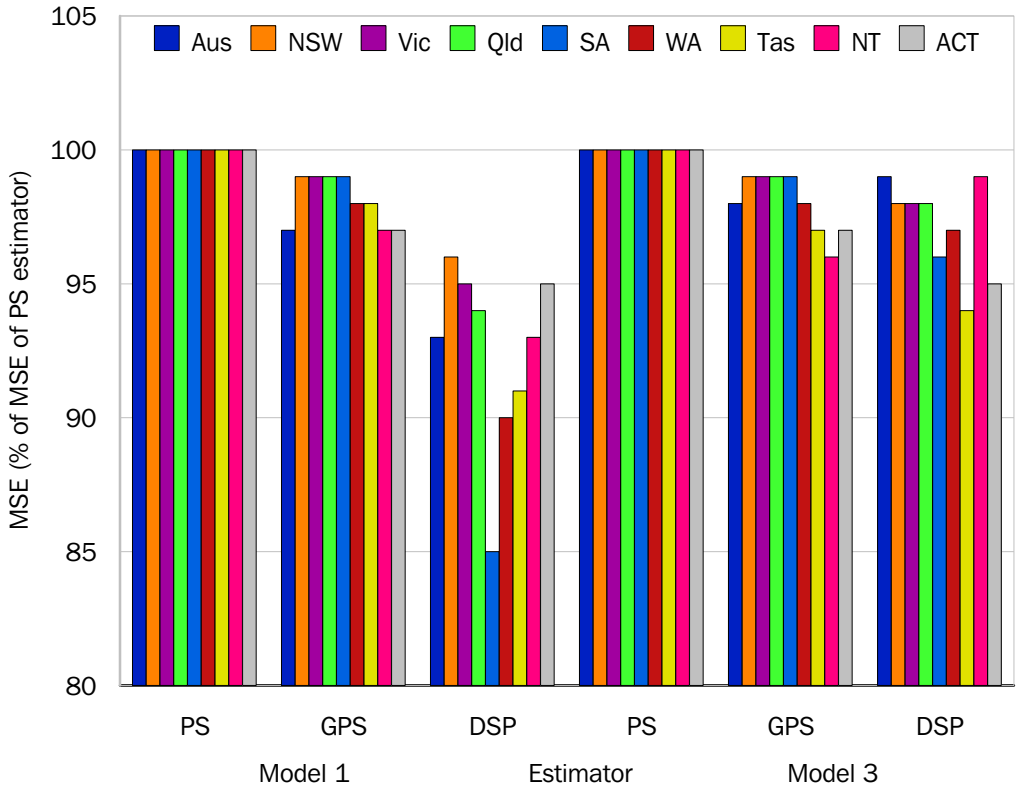
Graph 8.8 shows the relative biases of the PS and GPS estimators of Census coverage ratio under different coverage-response models: the dwelling-level model used above (model 1), the person-level model (model 2) and the sophisticated model (model 3). The bias of the PS estimator is not much worse under either alternative model, whereas the GPS appears to be considerably more biased. Indigenous estimates have a large positive biased under the GPS – the bars going off the graph are 1.8%, 5.0% and 2.8%. The estimate with the largest negative bias is females aged 30 to 59 years.

8.8 Relative bias of Census coverage ratio for PS and GPS estimates by various classifications for three coverage-response models



For estimates at state level, the simulations based on model 1 showed that the DSP estimator had lower mean squared error than the PS estimator (see graph 8.6). Graph 8.9 compares the mean squared error of the PS, GPS and DSP for state estimates using the simple model 1 alongside a similar comparison for the more sophisticated model 3. The gain in mean squared error for the DSP estimator is lower under the more sophisticated model, reflecting higher bias of the DSP even for these state estimates. Given the potential for bias in the DSP estimates for categories related to under-count, evidenced by the high mean squared errors for indigenous shown in graph 8.7, it appears that the PS estimator will be a better choice of estimator for many purposes.

8.9 Mean squared error of Census coverage ratio estimates by state (as a percentage of the MSE of the PS estimator for that classification) for coverage-response models 1 and 3



8.4 Evaluation of weighted residuals variance estimators

Table 8.10 presents an analysis of the weighted residuals (WR) variance estimator for PS estimates of the number of people who should have been counted in the Census, based on 4800 simulated samples. The number of simulations used was sufficient to identify a statistically significant positive bias in the weighted residual variance estimator for the smaller States.

The last two columns demonstrate the coverage of a symmetric confidence interval based on the weighted residuals variance estimates. The positive bias in the estimates of population count means that the true population value is less than the lower bound of the confidence interval more often than it is greater than the upper bound. The NT estimate breaks this pattern because for NT the SE is much larger than the bias. The larger SE estimates tend to occur with high values of the estimate, and in the NT this effect is sufficient to lead to the observed situation, in which the confidence interval falls below the population value more often than it falls above.

The coverage of the confidence intervals is over 95% for the smaller states because of the upward bias of the SE estimates. In other words, the weighted residual variance estimator is conservative; the estimated variance is slightly larger than the 'true' variance (as established by the repeated simulations study).

8.10 Properties of weighted residuals variance estimator, for PS estimates of population count

	Direct SE	Mean WR SE – direct SE (% of direct SE)	Proportion of samples with population value: less than estimate – 2 WR SE	Proportion of samples with population value: greater than estimate + 2 WR SE
Australia	18,987	0.0%	3.85%	1.35%
NSW	13,512	–1.0%	2.60%	2.48%
Vic	9,397	0.6%	2.52%	2.06%
Qld	8,617	0.6%	2.21%	2.35%
SA	3,923	*2.7%	2.06%	2.23%
WA	5,321	0.1%	2.56%	2.15%
Tas	1,760	*3.9%	2.13%	1.58%
NT	1,798	*3.3%	1.44%	2.56%
ACT	1,600	*2.0%	2.46%	1.85%

* signifies significantly different from zero at the 95% confidence level.

8.5 Summary of results

It should be reiterated that these simulations cannot provide information about the situation in the real world for the various category variables. The simulations do show that the PREG estimators have good bias properties even when the coverage-response model generating the data is more complex than the one used to justify the estimators. The good bias properties come at the expense of a sampling error which is a few percent higher relative to the sampling error of the penalised GREG estimator, which is the most appropriate estimator other than the PREG.

The simulations have also shown that the weighted residuals variance estimator proposed in Section 4 performs well as an estimator of the variance of the PREG estimator.

9. FURTHER ISSUES AND CONCLUSION

9.1 Conclusion

This paper has introduced a theoretical framework for PES estimation. This framework provides a new estimator, the prediction regression or PREG estimator, as a natural extension of the classical dual system estimator.

The PREG estimator allows weighting to take account of a variety of item values for each unit that could potentially affect Census and PES response. This is in contrast to the DSE, in which each unit's weighting depends on a single post-stratum value. The weighting can also take account of any differences in the category values reported for a unit in the Census and the PES.

Simulation studies have shown that the new estimator will have reduced 'correlation bias' compared to the DSE, through the use of a coverage-response model that can account for a greater variety of factors that will affect PES and Census response. The simulations have also shown that the PREG estimator is quite robust to a range of plausible response-coverage mechanisms, giving it an advantage over several other possible estimators such as the penalised GREG estimator, which were also examined in detail, but which are not described so fully in this paper. This may be achieved at the cost of a slightly higher standard error; but given the reduced bias, the measured standard error is expected to give a better indication of the overall quality of the estimates than would be the case if the DSE were used.

9.2 Further practical implementation issues

Choice of benchmark categories for person weighting

Examining the 2001 PES data, for example using logistic modelling as outlined in Section 7, demonstrates that the propensity to be counted in the Census is related to a number of categorical variables available in both PES and Census. The PREG estimator allows a rich model for the coverage-response adjustment that can incorporate effects from the variety of variables that are expected to affect PES and Census response.

For the 2006 PES estimation, benchmark variables will be included at the state or region level for a sex by age interaction, indigenous status, marital status and country of birth category. Other benchmark variables may also be included for persons counted in CDs to which special enumeration practices were applied in the Census and PES e.g. indigenous communities.

Limiting large weight changes

Given the number of benchmark variables to be used, there is potential for a few units in the PES sample to have modelled coverage-response adjustments very different from one. This can lead to very low or very high weights, and this could result in an increased variance for the PREG estimator. These large adjustments are also unrealistic, resulting from using a linear model for the coverage-response adjustment in which the effects of particular categories a unit belongs to are added together.

The proposed estimator for the 2006 PES will limit weight adjustments to a range around one, so as to minimise these problems. An iterative approach will be used. Units with a weight adjustments outside the allowable range at a first iteration of PREG weighting will have their weight adjustments set to the boundary of the range. A second iteration of PREG estimation will then be performed excluding these units, using adjusted benchmark totals that ensure that overall the original benchmarks are achieved by the weighted dataset. Multiple iterations of this procedure may be required to give a final weighting in which all weight adjustments fall within the allowable range. A similar approach was presented in the context of calibration estimators by Singh and Mohl (1996), who called it the Linear Truncated method.

This process can break down if there are benchmark categories with few contributors. These cases can be dealt with by collapsing the fine categories into broader categories for which extreme adjustments will not be required in order to meet the benchmark constraints.

ACKNOWLEDGEMENTS

The authors would like to acknowledge David Steel of the University of Wollongong for his role in reviewing an early draft of this paper, and for the number of helpful suggestions generated during discussions with him. We also would like to thank Bill Gross for his guidance in the development of the methodology, particularly regarding details of its application in the PES context. Finally, we extend thanks to Stephen Carlton and Frank Yu for pointing us in appropriate directions and linking us to relevant literature.

REFERENCES

Bell, P.A. (2000) *Weighting and Standard Error Estimation for ABS Household Surveys*, Paper prepared for ABS Methodology Advisory Committee, July 2001.

Bell, W. (2001) *ESCAP II: Estimation of Correlation Bias in 2000 A.C.E. Estimates using Revised Demographic Analysis Results*, Report No. 10 to the Executive Steering Committee for A.C.E. Policy II, U.S. Bureau of the Census.

Choi, C.Y., Steel, D.G. and Skinner, T.J. (1988) "Adjusting the 1986 Australian Census Count for Under-enumeration", *Survey Methodology*, 14(2), pp. 173–189.

Deville, J.C. and Särndal, C.E. (1992) "Calibration Estimators in Survey Sampling", *Journal of the American Statistical Association*, 87, pp. 376–382.

Estevao, V.M. and Särndal, C.E. (2000) "A Functional Form Approach to Calibration", *Journal of Official Statistics*, 16, pp. 379–399.

Hall, A. (1993) "Some Aspects of Generalized Method of Moments Estimation", in Maddala, G.S., Rao, C.R. and Vinod, H.D. (editors), *Handbook of Statistics, Volume 11*, Elsevier Science Publishers, B.V., pp. 393–417.

Hogan, H. and Wolter, K. (1988) "Measuring Accuracy in a Post-Enumeration Survey", *Survey Methodology*, 14(1), pp. 99–116

Isaki, C.T. and Schultz L.K. (1986) "Dual System Estimation using Demographic Analysis Data", *Journal of Official Statistics*, 2(2), pp. 169–179.

Kott, P.S. (2003) "A Practical Use for Instrumental-Variable Calibration", *Journal of Official Statistics*, 19(3), pp. 265–272.

Rao, C.R. (1973) *Linear Statistical Inference and its Applications*, Second edition, John Wiley and Sons, New York.

Sargan, J.D. (1958) "The Estimation of Economic Relationships using Instrumental Variables", *Econometrica*, 26, pp. 393–415.

Singh, A.C. and Mohl, C.A. (1996) "Understanding Calibration Estimators in Survey Sampling", *Survey Methodology*, 22(2), pp. 107–115.

Wachter, K.W. and Freedman, D.A. (2000) "The Fifth Cell: Correlation Bias in U.S. Census Adjustment", *Evaluation Review*, 24, pp. 191–211.

White, H. (1982) "Instrumental Variables Regression with Independent Observations", *Econometrica*, 50(2), pp. 483–499.

Wolter, K.M. (1990) "Capture–Recapture Estimation in the Presence of a Known Sex Ratio", *Biometrics*, 46, pp. 157–162.

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

WEB ADDRESS	www.abs.gov.au

2000001561102
ISBN 9780642483003

RRP $11.00