2002

The Effect of using Household as a Sampling Unit

Robert Graham Clark

University of Wollongong, rclark@uow.edu.au

David G. Steel

University of Wollongong, dsteel@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Clark, Robert Graham and Steel, David G.: The Effect of using Household as a Sampling Unit 2002.
<https://ro.uow.edu.au/infopapers/552>

The Effect of using Household as a Sampling Unit

Abstract

The effect of sampling people through households is considered. Results on design effects for two stage surveys are reviewed and applied to give design effects of household samples. The main factors that determine the design effect are identified for the designs in which one person, or all people, are selected from each selected household. Within household correlation is one factor. We show that the relationships between household size and the mean and variance within households are also important factors. Census and survey data are used to empirically compare the design effects for a range of estimators, variables and designs.

Keywords

Cluster sampling, Design effect, Generalized regression estimator, Household surveys, Intra-class correlation, Sample design, Survey design

Disciplines

Physical Sciences and Mathematics

Publication Details

This article was originally published as: Clark, RG & Steel, DG, The Effect of using Household as a Sampling Unit, International Statistical Review, 2002, 70(2), 289-314. Copyright 2002 International Statistical Institute/Blackwell Publishing. The definitive published version is available at www.blackwell-synergy.com.

The Effect of using Household as a Sampling Unit

Robert G. Clark and David G. Steel ¹

Abstract

The effect of sampling people through households is considered. Results on design effects for two stage surveys are reviewed and applied to give design effects of household samples relative to direct sampling of people. The main factors that determine the design effect are identified for the designs in which one person, or all people, are selected from each selected household. Within household correlation is only one factor. We show that the relationships between household size and the mean and variance within households are also important factors. Census and survey data are used to empirically compare the design effects for a range of estimators, variables and designs.

Keywords: cluster sampling; design effect; household surveys; intra-class correlation; sample design; survey design

1 Introduction

Household as a Sampling Unit

Samples of the general population are often selected by selecting a sample of households and selecting some or all of the people in the selected households. This is because, in many countries, there is no readily available list of people in the general population which can be used as a sampling frame. Even if a list of people is maintained it may not be available outside the organization that compiles it or to private survey organizations. It will suffer to some degree from the problems of omissions, duplicates and out-of-dateness (see for example Kish, 1967, pp.53-59). Contact information may be unavailable, out-of-date or suited only to a mail survey, which may have a low response rate or be unsuitable for

¹School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522 Australia. E-mail: David.Steel@uow.edu.au.

the survey topic. In contrast, reasonably complete lists of households are often available either for the whole population, for example a list of telephone number ranges or a national list of addresses such as the British Postal Address File (Foster, 1994), or within selected areas by a field listing exercise.

This article is concerned with the effect of using the household as a sampling unit. It is assumed that households are selected directly from the population of households without any intervening stages of selection. This is the case for telephone based surveys, which select either from lists of subscribers or use some form of random digit dialing (Lepowski, 1988). The latter is used to avoid undercoverage due to unlisted numbers. Surveys which involve some face-to-face interviewing will normally geographically cluster the selected households to reduce travel costs, by using an initial area-based stage of selection. This article is relevant even for area-based surveys, because clustering within areas often does not have a large effect unless many households are selected within each area-based unit. Even if geographic clustering has a major impact, it is still useful to consider the component of the overall variance that is due to the use of households as sampling units.

In practice, survey information is collected for either all people or one randomly selected person, in each selected household. In the former case, the information for all people in the household may be provided by a single contact person or proxy; this does not affect sampling error which is the concern of this article. Proxy interviewing is known to increase measurement error in many but

not all cases (Groves, 1989, pp.414-417), however it reduces respondent burden and improves response rates.

Cost, operational issues, timeliness, response rate, data quality and sampling variance must all be considered in choosing the sampling method within households and the sample size overall and for important domains; see e.g. Kish (1967, pp.396-404). While the factors affecting cost and nonsampling error are influential, it is also important to consider the sampling variance that will result from different sample sizes and designs, based on pilot data or past experience. An indication of the standard errors expected is needed to decide on the total sample size, to inform the survey users and sponsors and to consider the trade-offs between the operational issues and precision of various sample design options.

One important design choice is whether to sample one or all people in each selected household. Selecting one person from a household may remove any effect of within household clustering, having a beneficial effect on sampling variance, but result in unequal probabilities of selection which may have a detrimental effect. It is unclear what the net result of these two effects will be for a particular variable, particularly if a complex estimator is to be used.

Design Effects Relative to Direct Sampling of People

When developing the sample design for a survey, there may be little or no information with which to calculate the likely standard errors of key estimates for different design options. It is often possible to calculate the likely standard errors that a simple random sample without replacement (SRSWOR) of people would

produce. The design effect (deff) for a particular sample design and estimator is defined here as the ratio of the variance of the estimator for that design to the variance of the simple expansion estimator for a SRSWOR of people with the same expected sample size (Kish, 1967). Hence if a reasonable idea of the deff for a design can be developed we can get a reasonable idea of the variances that it will produce. Knowledge of the deffs for different designs and variables can indicate the likely standard errors of estimates for the designs under consideration.

For example, many surveys are designed to estimate the number of people in certain subpopulations with a given precision. An estimate from a SRSWOR of people would have variance

$$\frac{N^2}{n} \left(1 - \frac{n}{N}\right) \frac{N}{N-1} P(1-P) \approx \frac{N^2}{n} \left(1 - \frac{n}{N}\right) P(1-P) \quad (1)$$

where N is the population size which is assumed to be known and P is the proportion of people in the subpopulation. Provided that the survey designer has an idea of P , the expression (1) can be multiplied by the deff to give the variance that would be achieved from a particular sample design.

Even after the survey has been conducted, the deff may be useful if there are many variables of interest (Kish, 1995). The design effect may be similar across variables, at least within some broad groupings, making it easier to compare deffs than variances. A single deff could then be estimated for a group of similar variables, simplifying the presentation of variances and reducing the sampling variability of variance estimates. A similar approach is to relate the variance of

estimators $\hat{Y}$ to the total T_y of the survey variable using a "generalized variance function" or GVF (see eg Valliant, 1987). For example, the Australian Bureau of Statistics and Statistics Canada have used models of the form

$$\log \left(\text{var} \left(\hat{Y} \right) \right) = A + B \log (T_y) + C (\log (T_y))^2$$

while the US Current Population Survey and the National Health Interview Survey have used $\log \left(\text{var} \left(\hat{Y} \right) \right) = A + B/T_y$. The parameters A, B, C are estimated for large classes of similar variables, for example using least squares. The functions are primarily empirically determined although Valliant (1987) has provided some supporting theory in a model-based framework. GVFs constructed from a previous survey could be used for designing a new survey, but the GVF would need to be updated to allow for changes in the sample design, if any. If we modify the GVF by multiplying by a factor not depending on T_y , then the factor should equal the ratio of the deffs of the two designs. Hence we concentrate on deffs.

Skinner (1986) discussed "mis-specification effects" (meffs) which have been called design effects by some authors. Survey data can be used to estimate finite population quantities or model parameters, for example a regression parameter β . Suppose a working model assumes independent observations, but the true model has positive correlations within households. The estimator of β based on the working model may be unbiased but the estimator of $\text{var} \left(\hat{\beta} \right)$ will usually have a negative bias, depending on the correlation within households. The mis-

specification effect is defined as

$$\text{var}_M(\hat{\beta}) / E_M \widehat{\text{var}}(\hat{\beta})$$

where $\widehat{\text{var}}(\hat{\beta})$ is an estimate of $\text{var}(\beta)$ based on the working model and E_M and var_M are the expected value and variance under the true model. The meff measures the effect of the sample design and population structure on estimators of variance which are based on an incorrect model. The deff measures the effect of the sample design and population structure on the true sampling variance.

Outline of this Article

This article is concerned with design effects for two stage designs with SR-SWOR at both stages, where primary sampling units(PSUs) are households and final sampling units(FSUs) are people. Probability proportional to size selection of PSUs, which is often used with geographically defined PSUs (eg Hansen et al., 1953), is generally not feasible for households because household sizes are normally not available prior to sampling. Section 2 reviews the existing literature on deffs for two-stage surveys for two simple estimators of a population total. Generally the deff is related to some population parameters, for example measures of within-PSU correlation. These parameters do not depend on the sample design, so that deffs can be calculated for several alternative sample designs, either exactly or approximately.

It is often assumed, explicitly or implicitly, that variation in PSU size is negligible, or that within-PSU sample sizes are large. These assumptions are

not realistic where PSUs are households. Section 3 contains expressions for the deff for the one/household and all/household designs, in terms of population parameters which have a clear interpretation even where PSU sizes are small and vary significantly. Some of these population parameters would be difficult to estimate in practice, however, and we suggest some simpler approximations motivated by empirical results and a superpopulation model.

In Section 4 we discuss the deffs of generalized regression estimators (GREGs) and relate the deff of the GREG to the deffs discussed in Sections 2 and 3. In Section 5, the deff is estimated for the one/household and all/household designs for a number of estimators and variables from the 1991 Australian Census of Population and Housing and the 1995 Australian National Health Survey. Other within-household sampling methods are also evaluated. Section 6 contains a summary and conclusions.

2 Review of Deffs for Two Stage Sampling

This review is restricted to the case where PSUs, and FSUs within selected PSUs, are selected by SRSWOR. The following notation will be used:

U_1, s_1 = population and sample of PSUs, of sizes M and m respectively,

U_2, s_2 = population and sample of FSUs, of sizes N and n respectively,

U_{2h}, s_{2h} = population and sample of FSUs in PSU h , of sizes N_h and n_h respectively,

$$\pi_{hi} = P(hi \in s_2) = \frac{m}{M} \frac{n_h}{N_h},$$

$\bar{N} = N/M =$ average number of FSUs in a PSU in the population,

Y_{hi} = variable of primary interest for FSU i in PSU h , observed for $hi \in s_2$,

$T_y = \sum_{hi \in U_2} Y_{hi} =$ the population total of interest,

$\bar{Y} = T_y/N =$ population mean per FSU,

$\bar{Y}_h = N_h^{-1} \sum_{i \in U_{2h}} Y_{hi} =$ population mean for PSU h ,

E_p and var_p will denote expectation and variance over repeated probability sampling from a fixed population. Inference on this basis is often called design-based inference and does not rely on a model for how $\{Y_{hi}\}$ are generated. This article is mainly concerned with design-based inference, although a model will be used to motivate some estimators in Sections 3.2 and 4. This general approach is sometimes called model-assisted (Sarndal et al., 1992).

We will focus on the estimation of the population total T_y . The sample sizes n_h are assumed to be fixed numbers for each $h \in U_1$, which do not depend on s_1 . The expected sample size is

$$n_0 = E_p(n) = \frac{m}{M} \sum_{h \in U_1} n_h$$

and $\bar{n} = n_0/m$ is the average number of selected FSUs in a PSU in the sample. It is assumed for the remainder of the article that terms in $\left(\frac{n_0}{N}\right)$, $\left(\frac{m}{M}\right)$, n_0^{-1} and m^{-1} are negligible whereas the sampling fractions within PSUs (n_h/N_h) may not be. All approximations made in this article will be based on this assumption which would hold for almost all household surveys.

A simple design-unbiased estimator of the population total T_y is the π -weighted or Horvitz-Thompson estimator.

$$\hat{Y}_\pi = \sum_{hi \in s_2} \pi_{hi}^{-1} Y_{hi} = \frac{M}{m} \sum_{h \in s_1} \frac{N_h}{n_h} \sum_{i \in s_{2h}} Y_{hi} \quad (2)$$

This estimator does not make use of any population information except for M and $\{N_h : h \in s_1\}$ and would be used if other auxiliary information such as the total number of people, N , is unavailable. For example, sometimes a frame of households is available but not a list or count of people. One of the outputs of the survey would then be an estimator of N :

$$\hat{N}_\pi = \frac{M}{m} \sum_{h \in s_1} \frac{N_h}{n_h} \sum_{i \in s_{2h}} 1 = \frac{M}{m} \sum_{h \in s_1} N_h.$$

If the population size N is known, then a more efficient estimator in most cases is the Hajek estimator

$$\hat{Y}_1 = N \frac{\hat{Y}_\pi}{\hat{N}_\pi} \quad (3)$$

(eg Sarndal et al., 1992, p.258). The variances of $\hat{Y}_\pi$ and $\hat{Y}_1$ over repeated sampling are

$$var_p(\hat{Y}_1) \approx \frac{M^2}{m} \bar{N}^2 S_b^2 + \frac{M}{m} \sum_{h=1}^M \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_h^2 \quad (4)$$

$$var_p(\hat{Y}_\pi) \approx var_p(\hat{Y}_1) + \frac{M^2}{m} \bar{N}^2 \bar{Y}^2 (C_N^2 + 2C_{NY}) \quad (5)$$

where

$$S_h^2 = \frac{1}{N_h - 1} \sum_{i \in U_{2h}} (Y_{hi} - \bar{Y}_h)^2$$

$$S_b^2 = \bar{N}^{-2} (M - 1)^{-1} \sum_{h \in U_1} (Y_h - N_h \bar{Y})^2$$

$$C_N^2 = M^{-1} \sum_{h \in U_1} (N_h - \bar{N})^2 / \bar{N}^2$$

$$C_{NY} = \sum_{h \in U_1} \frac{N_h}{N} (N_h - \bar{N}) (\bar{Y}_h - \bar{Y}) / (\bar{Y} \bar{N})$$

C_N^2 is the population relative variance of N_h , so that C_N is the coefficient of variation (CV) of the PSU sizes. C_{NY} is the population relative covariance of N_h and $\bar{Y}_h$, weighted by N_h/N . Results (4) and (5) are due to Hansen et al. (1953), who commented that C_{NY} would often be negligible in practice. We would expect $\hat{Y}_1$ to have lower variance than $\hat{Y}_\pi$ in most cases, because it uses more information, however $\hat{Y}_1$ has higher variance than $\hat{Y}_\pi$ when $C_{NY} < -C_N^2/2$. For this to occur, C_{NY} must be negative, which may be the case for some variables. This condition is a special case of the well known condition for a ratio estimator to be worse than an expansion estimator (see e.g. Cochran, 1977, pp.157-158).

The variance of a π -weighted estimator from a simple random sample without replacement of size n_0 , $\hat{Y}_{\pi(srs)}$, is $\frac{N^2}{n_0} \left(1 - \frac{n_0}{N}\right) S^2$ where

$$S^2 = \frac{1}{N-1} \sum_{hi \in U_2} (Y_{hi} - \bar{Y})^2.$$

The design effects of $\hat{Y}_1$ and $\hat{Y}_\pi$ are

$$D_1 = \frac{var_p(\hat{Y}_1 - Y)}{var_p(\hat{Y}_{\pi(srs)} - Y)} \approx \bar{n} \frac{S_b^2}{S^2} + \bar{n} \frac{1}{M \bar{N}^2} \sum_{h \in U_1} \frac{N_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) S_h^2 / S^2 \quad (6)$$

$$D_\pi = D_1 + \bar{n} \frac{C_N^2 + 2C_{NY}}{C_Y^2} \quad (7)$$

where $C_Y^2 = S^2 / \bar{Y}^2$ is the CV of Y .

Deff Expressions with Portable Parameters

In developing a sample design, estimates of the deff are often required for several alternative designs being considered. It is useful to distinguish between quantities which summarize the sample design and population parameters. D_1 depends on the sample design through $\bar{n}$ and $n_h (h = 1, \dots, M)$ and the population structure through the parameters S_b^2/S^2 , S_h^2/S^2 and $N_h (h = 1, \dots, M)$. The population parameters are specific to the variable and population and the PSUs into which the population is partitioned. If these parameters are known or can be accurately estimated, then we can calculate the deffs associated with different sample design options. However, when developing a sample design such detailed information about the population is often not available and must be estimated or guessed from the limited information available from a pilot test, a previous survey or the most recent population census. Sometimes, the only available information refers to similar but not identical variables, or to a different set of PSUs from those to be used in the current survey. Using such information is made easier if the deff can be related to a small number of population parameters, perhaps just one. It is also useful if the parameters are approximately constant, or vary within a known small interval, for a set of variables or the same variable measured at different times. Such parameters have been called transferable or generalizable parameters (Kish, 1995) or portable parameters (Brewer et al., 1977).

Developing the survey design is simpler if the effect of the sample design is mainly determined by the total sample size, n , and the average number of sample

FSUs per PSU, $\bar{n}$. The simplest case is when all the PSUs are the same size $\bar{N}$ and the same sample size $\bar{n}$ is used in each selected PSU. Then

$$D_1 = 1 + (\bar{n} - 1) R \quad (8)$$

where

$$R = S^{-2} (S_b^2 - \bar{N}^{-1} S_w^2) \quad (9)$$

is a measure of the intra-PSU correlation and $S_w^2 = \sum_{h \in U_1} \frac{N_h}{\bar{N}} S_h^2$ (Hansen et al., 1953). R is the correlation between the value of the variable of interest for two units picked at random from the whole population conditional on their belonging to the same PSU, and must lie in the range $\left[-(\bar{N} - 1)^{-1}, 1\right]$. If PSUs are as heterogeneous as the population as a whole, then $R = 0$ and the deff is 1. If PSUs are perfectly homogeneous, then $R = 1$ and $D_1 = \bar{n}$. $R > 0$ for most variables and for area based PSUs, $R < 0.1$ for many variables (Verma et al., 1980).

Consider equal probability sample designs, where PSUs need not be equal in size and a uniform sampling fraction is used within PSUs so that $n_h/N_h = \bar{n}/\bar{N}$. Then

$$D_1 = K (1 + (\bar{n} - 1) \delta_1) \quad (10)$$

(Hansen et al., 1953) where

$$\delta_1 = (S_b^2 - \bar{N}^{-1} S_w^2) / \left(S_b^2 + \frac{\bar{N} - 1}{\bar{N}} S_w^2 \right)$$

$$K = \left(S_b^2 + \frac{\bar{N} - 1}{\bar{N}} S_w^2 \right) / S^2$$

Both K and δ_1 are population parameters. The parameter δ_1 is called the measure

of within PSU homogeneity. If PSUs do not vary significantly in size, then $K \approx 1$; in general K can be much larger than 1 because S_b^2 can be large.

For unequal sized PSUs and unequal probability designs, Kish and Frankel (1974) suggested the following expression motivated by (8):

$$D_1 = 1 + (\bar{n} - 1) \delta_2 \quad (11)$$

where δ_2 is defined to make this equation true. However, δ_2 is not a population parameter because it depends on the type of unequal probability sampling design. The use of δ_2 implies the assumption that δ_2 does not depend on $\bar{n}$ and is approximately portable over designs which have different $\bar{n}$ but are similar in other respects.

The deff of the mean of a variable for a subpopulation may be related to the deff of the grand mean. Kish (1995) defined the deff for a subpopulation $U_{2c} \subseteq U_2$ by

$$D_c = \frac{\text{var}(\bar{y}_c : \text{clustered sample from } U_{2c})}{\text{var}(\bar{y}_c : \text{SRSWOR from } U_{2c})}$$

A measure of of homogeneity for U_{2c} , δ_{2c} , is defined by

$$D_c = 1 + (\bar{n}_c - 1) \delta_{2c}$$

where $\bar{n}_c = E_p(n_c) / E_p(m_c)$ and n_c and m_c are the number of selected FSUs and PSUs respectively in subpopulation c . Based on empirical evidence, Kish (1995) suggested that the subpopulation δ_{2c} can be approximated by $F\delta_2$, where $F = 1$ for demographic classes such as agegroup by sex and $F = 1.2$ or $F = 1.3$ for socio-economic classes which would usually be more clustered.

Brewer et al. (1977) suggested a portable parameter which allows for designs using different sets of PSUs. Using the empirical observation that δ_2 tends to decrease with the average PSU size $\bar{N}$, they discussed a number of models for δ_2 , in particular

$$\delta_2 \approx \bar{N}^{2(\gamma-1)} \quad (12)$$

where γ was between 0.55 and 0.88 for 7 variables relating to labour force status and housing information. Further empirical study was suggested, to confirm the portability of γ .

Verma et al. (1980) considered an improvement on (11) to give population parameters which may be portable over designs with different estimation weights. They noted that for stratified single stage sample designs where the sampling fractions are unrelated to S_h^2 , the design effect is given by $1 + C_w^2$ where $w_h = n_h/N_h$, $\bar{w} = \sum_{h \in U_1} n_h w_h / \sum_{h \in U_1} n_h$ and

$$C_w^2 = \left[\sum_{h \in U_1} n_h (w_h - \bar{w})^2 / \sum_{h \in U_1} n_h \right] / \bar{w}^2.$$

C_w^2 is the relative variance of w_h over U_1 , weighted by n_h . For equal probability two stage designs, the deff expression (11) is often used. If the effect of clustering and weighting are multiplicative, then δ_3 may be reasonably portable where

$$D_1 = (1 + C_w^2) (1 + (\bar{n} - 1) \delta_3) \quad (13)$$

Verma et al. (1980) estimated D_1 and δ_3 for many variables and countries in the World Fertility Survey using replication methods. PSUs were areas of at least 20

final units. They found that the values of δ_3 were in the range 0.01 to 0.17 and were fairly similar for variables within some broad groupings.

Sarndal et al. (1992, p.132) derived an expression for D_π for cluster samples, where $n_h \equiv N_h$. In this case,

$$D_\pi \approx 1 + (\bar{N} - 1) \delta_4 + \bar{N}^{-1} S^{-2} C \quad (14)$$

where

$$\begin{aligned} \delta_4 &= 1 - S^{-2} \sum_{h \in U_1} \frac{N_h - 1}{N - M} S_h^2 \\ C &= (M - 1)^{-1} \sum_{h \in U_1} N_h (N_h - \bar{N}) \bar{Y}_h^2. \end{aligned} \quad (15)$$

Sarndal et al. (1992) commented that in practice, N_h may be positively correlated with $N_h \bar{Y}_h^2$ so that D_1 may be quite large.

For equal PSU sizes, $\delta_1 = \delta_2 = \delta_3 = \delta_4 = R$ and $K = 1$ so that expressions (10), (11) and (13) for D_1 collapse to (8). The expression most useful in practice depends mainly on the range of sample designs of interest. Expression (8) is relevant if PSU sample and population sizes are equal. Expression (11) can be used for complex sample designs and is widely used in practice. Notice that δ_2 can be treated as a population parameter provided that only $\bar{n}$ is to be varied and the effect of other aspects such as unequal probability sampling does not depend on $\bar{n}$. δ_3 in expression (13) may vary less than δ_2 across designs with different estimation weights, provided that the effects of weighting and clustering are multiplicative. Expression (14) can be used for cluster samples only. The parameter γ in (12) is useful in considering alternative groupings of FSUs into PSUs with different $\bar{N}$.

An important issue is whether a portable parameter is roughly constant for similar variables or the same variable measured at different times. Population parameters which tend to lie in a fairly narrow range are particularly valuable. For example, for equal sized PSUs, the intra-class correlation is often between 0 and 0.1 and must always lie in $\left[-\left(\bar{N}-1\right)^{-1}, 1\right]$. If there is little variation in PSU size and estimation weight, then δ_2 behaves similarly, however if PSU size varies greatly then it may be very difficult to estimate δ_2 (and δ_3). An implicit assumption often made in estimating δ_2 and δ_3 is that variation in PSU size is small, so that these parameters lie in a fairly small range similar to R . This may be reasonable for area-based PSUs but not if PSUs are households.

Even if PSU sizes vary, δ_4 should have properties similar to R . Like R , $\delta_4 \in \left[-\left(\bar{N}-1\right)^{-1}, 1\right]$ for large M , and $(1-\delta_4)$ is the ratio of the mean square residual of a household mean model to the mean square residual of a grand mean model:

$$1-\delta_4=\left[(N-M)^{-1} \sum_{hi \in U_2}\left(Y_{hi}-\bar{Y}_h\right)^2\right] /\left[(N-1)^{-1} \sum_{hi \in U_2}\left(Y_{hi}-\bar{\bar{Y}}\right)^2\right] .$$

Expression (14) gives the deff of $\hat{Y}_\pi$ in terms of δ_4 for cluster sampling, however the parameter C is also required. In Section 3 we derive similar expressions for D_1 . We focus on D_1 because the population total is often available for use in estimation. If $\hat{Y}_\pi$ has to be used then we can employ (7) or (14).

3 Deffs for Household Surveys

3.1 Design-Based Deffs

In this section, PSUs are assumed to be households. Expression (6) becomes

$$D_1 = D_{1(all)} = \bar{N} \frac{S_b^2}{S^2}$$

for the all/household design ($n_h = N_h$) and

$$D_1 = D_{1(one)} = \frac{S_b^2}{S^2} + \frac{M}{N^2} \sum_{h \in U_1} N_h (N_h - 1) S_h^2 / S^2$$

for the one/household design ($n_h = 1$). Our aim in this section is to express D_1 in terms of as few population parameters as possible, which can hopefully then be estimated using judgement or limited data.

The expressions (10), (11), (13) and (14), which were discussed in Section 2, are unsuitable for this purpose. Expression (14) applies to cluster sampling only. Expressions (11) and (13) do not apply to the one/household case because they cannot be inverted to give δ_2 and δ_3 ; as result they assume that $D_{1(one)} = 1$ or $D_{1(one)} = 1 + C_N^2$. Expression (10) can be used, however δ_1 determines the comparison between $D_{1(all)}$ and $D_{1(one)}$, and this parameter does not have a simple interpretation and may be difficult to estimate. We will instead express $D_{1(all)}$ and $D_{1(one)}$ in terms of δ_4 and other interpretable parameters, and look for sensible approximations to reduce the number of population parameters involved.

With a little manipulation, the deffs become

$$D_{1(all)} \approx 1 + (\bar{N} - 1) \delta_4 + \bar{N}^{-1} S^{-2} C_{NY2} \quad (16)$$

$$D_{1(one)} \approx 1 + \bar{N}^{-2} S^{-2} C_{NY2} + \bar{N}^{-2} S^{-2} C_{NS2} \quad (17)$$

where

$$C_{NY2} = M^{-1} \sum_{h \in U_1} (N_h - \bar{N}) N_h (\bar{\bar{Y}}_h - \bar{\bar{Y}})^2$$

is the covariance between N_h and $N_h (\bar{\bar{Y}}_h - \bar{\bar{Y}})^2$ over U_1 , and

$$C_{NS2} = M^{-1} \sum_{h \in U_1} (N_h - \bar{N}) (N_h - 1) S_h^2$$

is the covariance between N_h and $(N_h - 1) S_h^2$ over U_1 . As expected $D_{1(all)}$ involves the within household homogeneity as measured by δ_4 multiplied by $(\bar{N} - 1)$, but it also has a term involving C_{NY2} . Comparing (16) with (14) we see that using $\hat{Y}_1$ instead of $\hat{Y}_\pi$ for the all/household design leads to C being replaced by C_{NY2} , i.e. $\bar{\bar{Y}}_h$ is replaced by $\bar{\bar{Y}}_h - \bar{\bar{Y}}$. Comparing (16) with (17) shows that by taking only one person per household the term directly determined by the within household homogeneity, $(\bar{N} - 1) \delta_4$, disappears and the impact of C_{NY2} is multiplied by $\bar{N}^{-1}$, but the term involving C_{NS2} is added. If the relative variation in PSU size is small then C_{NY2} and C_{NS2} may be negligible. Household sizes vary significantly, so the potential effect of these terms cannot be ignored for household surveys.

Expressions (16) and (17) would often be difficult to apply in developing a sample design, because the terms C_{NY2} and C_{NS2} would usually be difficult to estimate or predict. If it is assumed that $(\bar{\bar{Y}}_h - \bar{\bar{Y}})^2$ has zero population covariance with N_h and N_h^2 over U_1 then

$$C_{NY2} = \bar{N} S^2 (1 + (\bar{N} - 1) \delta_4) C_N^2 \quad ; \quad \text{and}$$

$$D_{1(all)} = (1 + (\bar{N} - 1) \delta_4) (1 + C_N^2).$$

If the population covariance between S_h^2 and N_h over U_1 is zero then

$$C_{NS2} = \bar{N}^2 C_N^2 S^2 (1 - \delta_4).$$

If all of these conditions apply, then

$$D_{1(one)} = 1 + C_N^2 + \bar{N}^{-1} (1 - \delta_4) C_N^2$$

which is greater than $1 + C_N^2$ implied by (13) for this case, and still includes δ_4 .

$D_{1(one)}$ is decreasing in δ_4 , unlike $D_{1(all)}$, because high δ_4 means that people within households are similar so there is less sampling error associated with estimating the values of the unselected people in selected households. The empirical study in Section 4 shows that the approximation for C_{NS2} is reasonable in practice but not the approximation for C_{NY2} .

3.2 Model-Assisted Deffs

An alternative way of deriving approximations to C_{NY2} and C_{NS2} and hence $D_{1(all)}$ and $D_{1(one)}$ is to use a superpopulation model. Because $D_{1(one)}$ and $D_{1(all)}$ are population quantities based on many observations they should closely approximate their model-expectations, provided the model is correct. The model-expectation of the design-variance, or *anticipated variance*, is an important consideration in model-assisted sample theory particularly in sample design, see for example Sarndal et al. (1992, pp.450-452).

Assume a model with constant mean and variance. This model will be seen to be a poor approximation to reality in some respects, however it will be useful

to motivate some decompositions of C_{NY2} and C_{NS2} .

$$\left. \begin{aligned} E_M Y_{hi} &= \mu \\ var_M Y_{hi} &= \sigma^2 \\ cov_M(Y_{hi}, Y_{hj}) &= \rho \sigma^2 (\forall i \neq j) \\ cov_M(Y_{gi}, Y_{hj}) &= 0 (\forall g \neq h) \end{aligned} \right\} \quad (18)$$

The model expectations of the relevant terms in (16) and (17) are

$$\begin{aligned} E_M S^2 &\approx \sigma^2 \\ E_M \delta_4 &\approx \rho \\ E_M C_{NY2} \bar{N}^{-1} S^{-2} &\approx \bar{N} \rho C_N^2 \\ E_M C_{NS2} \bar{N}^{-2} S^{-2} &\approx (1 - \rho) C_N^2 \end{aligned}$$

for large M , and hence

$$D_{1(all)} \approx E_M D_{1(all)} \approx 1 + \left(\bar{N} (1 + C_N^2) - 1 \right) \rho \quad (19)$$

$$D_{1(one)} \approx E_M D_{1(one)} \approx 1 + C_N^2. \quad (20)$$

Notice that the difference between the design effects obtained under the assumptions of zero covariances and these obtained under this model are $(1 - \delta_4) C_N^2$ and $\bar{N}^{-1} (1 - \delta_4) C_N^2$ respectively. This is because $(\bar{Y}_h - \bar{Y})^2$ is proportional to N_h^{-1} in model expectation and so is negatively correlated with N_h and N_h^2 . Hence the model approximation to C_{NY2} is lower than the approximation based on zero covariances. The model approximation to C_{NS2} is equal to the approximation based on zero covariances, because $E_M S_h^2 = \sigma^2(1 - \rho)$ for $N_h > 1$ which does not depend on N_h .

If the model holds then (19) and (20) can be used to approximate the deffs, with ρ replaced by δ_4 . To examine how the deffs behave when the model does not hold, it is useful to decompose C_{NY2} and C_{NS2} . The first term of each decomposition will be the model expectation and the subsequent terms will measure the

effect of specific model failures. We define U_{1a} and U_{2a} to be the subsets of U_1 and U_2 corresponding to households of size a , for $a = 1, \dots, A$. We define $\bar{\bar{Y}}_a$, S_a^2 and δ_{4a} analogously to $\bar{\bar{Y}}$, S^2 and δ_4 :

$$\begin{aligned}\bar{\bar{Y}}_a &= N_a^{-1} \sum_{hi \in U_{2a}} Y_{hi} \\ S_a^2 &= N_a^{-1} \sum_{hi \in U_{2a}} (Y_{hi} - \bar{\bar{Y}}_a)^2 \\ \delta_{4a} &= 1 - \sum_{h \in U_{1a}} \frac{N_h - 1}{N_a - M_a} S_h^2 / S_a^2\end{aligned}$$

Notice that δ_{4a} is an intra-class correlation as defined in (9), because household size is constant within U_{1a} . We derive our decomposition of C_{NY2} by breaking up the sums $\sum_{h \in U_1}$ into $\sum_{a=1}^A \sum_{h \in U_{1a}}$ and expanding $\bar{\bar{Y}}_h$ into $\bar{\bar{Y}} + (\bar{\bar{Y}}_h - \bar{\bar{Y}})$. With a little manipulation, this gives

$$C_{NY2} = M^{-1} \sum_{a=1}^A (a - \bar{N}) a \left\{ \sum_{h \in U_{1a}} (\bar{\bar{Y}}_h - \bar{\bar{Y}}_a)^2 + M_a (\bar{\bar{Y}}_a - \bar{\bar{Y}})^2 \right\} \quad (21)$$

Household sizes are equal within each U_{1a} , hence

$$M_a^{-1} \sum_{h \in U_{1a}} (\bar{\bar{Y}}_h - \bar{\bar{Y}}_a)^2 \approx S_a^2 - M_a^{-1} \sum_{h \in U_{1a}} S_h^2 \quad (22)$$

using the well known analysis of variance decomposition (e.g. Johnson & Wichern, 1992, formula 6-31, p.317). Substituting into (21) and further expanding S_a^2 into $S^2 + (S_a^2 - S^2)$ and δ_{4a} into $\delta_4 + (\delta_{4a} - \delta_4)$ gives

$$C_{NY2} = C_{NY2}^{(1)} + C_{NY2}^{(2)} + C_{NY2}^{(3)} + C_{NY2}^{(4)} + C_{NY2}^{(5)}$$

where

$$C_{NY2}^{(1)} = S^2 \bar{N}^2 C_N^2 \delta_4$$

$$\begin{aligned}
C_{NY2}^{(2)} &= \left(1 + (\bar{N} - 1) \delta_4\right) \sum_{a=1}^A \frac{M_a}{M} (a - \bar{N}) (S_a^2 - S^2) \\
C_{NY2}^{(3)} &= \delta_4 \sum_{a=1}^A \frac{M_a}{M} (a - \bar{N})^2 (S_a^2 - S^2) \\
C_{NY2}^{(4)} &= \sum_{a=1}^A \frac{M_a}{M} (a - \bar{N}) S_a^2 (a - 1) (\delta_{4a} - \delta_4) \\
C_{NY2}^{(5)} &= \sum_{a=1}^A \frac{M_a}{M} (a - \bar{N}) (\bar{Y}_a - \bar{Y})^2
\end{aligned}$$

We can decompose C_{NS2} by breaking up $\sum_{h \in U_1}$ into $\sum_{a=1}^A \sum_{h \in U_{1a}}$ and using (22).

Then expanding S_a^2 into $S^2 + (S_a^2 - S^2)$ gives

$$C_{NS2} = C_{NS2}^{(1)} + C_{NS2}^{(2)}$$

where

$$\begin{aligned}
C_{NS2}^{(1)} &= S^2 \bar{N}^2 C_N^2 (1 - \delta_4) \\
C_{NS2}^{(2)} &= \sum_{a=1}^A \frac{M_a}{M} (a - \bar{N}) \left[S_a^2 (1 - \delta_{4a}) - S^2 (1 - \delta_4) \right].
\end{aligned}$$

If the model (18) is correct, then $C_{NY2}^{(2)}, C_{NY2}^{(3)}, C_{NY2}^{(4)}, C_{NY2}^{(5)}$ and $C_{NS2}^{(2)}$ are all negligible. These terms depend on the covariation of N_h with S_a^2 , δ_{4a} and $\bar{Y}_a$. For example, $C_{NY2}^{(2)}$ measures the covariation of N_h and S_a^2 . In the empirical study, which is detailed in Section 5, it was found that this term generally dominated $C_{NY2}^{(3)}, C_{NY2}^{(4)}$ and $C_{NY2}^{(5)}$ and hence $C_{NY2} \approx C_{NY2}^{(1)} + C_{NY2}^{(2)}$. It was further found that $C_{NS2}^{(1)} \gg C_{NS2}^{(2)}$ so that $C_{NS2} \approx C_{NS2}^{(1)}$.

Useful approximations can be derived if we consider the case where Y_{hi} is an indicator variable for some subpopulation,. Then $\bar{Y} = P$ and $\bar{Y}_a = P_a$ are the proportions in the subpopulation overall and for households of size a , respectively.

If P and P_a are small, then $S^2 \approx P$ and $S_a^2 \approx P_a$. Empirical results suggested that the main variation in P_a was between households of size 1 and 2, and households of size 3 or more, so that $P_1 \approx P_2 = P_{G1}$ and $P_a \approx P_{G2}$ for $a \geq 3$. Hence

$$C_{NY}^{(2)} \bar{N}^{-1} S^{-2} \approx \left(1 + (\bar{N} - 1) \delta_4\right) \frac{M_{G1}}{M} \frac{M_{G2}}{M} \frac{P_{G1} - P_{G2}}{P} (\bar{N}_{G1} - \bar{N}_{G2}). \quad (23)$$

The deffs are then approximated by

$$D_{1(all)} \approx 1 + \left(\bar{N} (1 + C_N^2) - 1\right) \delta_4 + \left[1 + (\bar{N} - 1) \delta_4\right] \frac{M_{G1}}{M} \frac{M_{G2}}{M} \frac{P_{G2} - P_{G1}}{P} (\bar{N}_{G2} - \bar{N}_{G1}) \quad (24)$$

$$D_{1(one)} \approx 1 + C_N^2 + \left[1 + (\bar{N} - 1) \delta_4\right] \frac{M_{G1}}{M} \frac{M_{G2}}{M} \left(\frac{P_{G2} - P_{G1}}{P}\right) \left(\frac{\bar{N}_{G2} - \bar{N}_{G1}}{\bar{N}}\right) \quad (25)$$

If the proportion of people in the subpopulation is higher in large households then $P_{G2} > P_{G1}$ and both deffs are increased; the increase in $D_{1(all)}$ is $\bar{N}$ times the increase in $D_{1(one)}$. Notice that both deff expressions include δ_4 . If approximations (24) and (25) can be used, then the deffs can be estimated provided δ_4 , P , P_{G1} and P_{G2} can be estimated.

Brewer et al. (1977) commented that it is "difficult to estimate variance components (*e.g.* δ_4) accurately unless the pilot-test sample is large, usually larger than warranted or necessary for other purposes". This is true for area-based PSUs because a pilot sample would contain only a small number of PSUs. If PSUs are households, then even a small pilot may contain sufficient PSUs to estimate δ_4 adequately, because $\bar{N}$ would be less than 3. A more difficult problem may be the estimation of P_{G2} as there may be insufficient large households; for the Census data used in our empirical study, only 19% of households had 3 or more adults.

4 GREG Estimators

4.1 Deffs for GREG Estimators

Improvements in the variances of $\hat{Y}_\pi$ and $\hat{Y}_1$ may be possible if population data for some additional auxiliary variables are available. Let $\mathbf{X}_{hi}$ be the vector of all of auxiliary variables associated with unit hi . The values $\{\mathbf{X}_{hi} : hi \in s_2\}$ and the vector of population totals $\mathbf{T}_x = \sum_{h \in U_2} \mathbf{X}_{hi}$ must be observed. The generalized regression estimator (GREG) is defined by

$$\hat{Y}_{reg} = \sum_{hi \in U_2} \hat{\beta}^T \mathbf{X}_{hi} + \frac{M}{m} \sum_{h \in s_1} \frac{N_h}{n_h} \sum_{i \in s_{2h}} (Y_{hi} - \hat{\beta}^T \mathbf{X}_{hi}) \quad (26)$$

where $\hat{\beta}$ is an ordinary least squares (OLS) estimator of the linear regression of Y_{hi} on $\mathbf{X}_{hi}$ calculated from the sample:

$$\hat{\beta} = \left(\frac{M}{m} \sum_{h \in s_1} \frac{N_h}{n_h} \sum_{i \in s_{2h}} \mathbf{x}_{hi} \mathbf{x}_{hi}^T \right)^{-1} \left(\frac{M}{m} \sum_{h \in s_1} \frac{N_h}{n_h} \sum_{i \in s_{2h}} \mathbf{x}_{hi} Y_{hi} \right). \quad (27)$$

See for example Sarndal et al. (1992, ch.6). The asymptotic design variance of the GREG is the same as the variance of $\hat{Y}_\pi$, with Y_{hi} replaced by $R_{hi} = Y_{hi} - \mathbf{B}^T \mathbf{X}_{hi}$, where $\mathbf{B}$ is the finite population OLS regression parameter of Y on $\mathbf{X}$. Hence the results in Sections 2 and 3 apply to the deff of a GREG, using the variable R_{hi} instead of Y_{hi} . This deff,

$$D_{reg} = var(\hat{Y}_{reg}) / var(\hat{Y}_{reg(srs)}),$$

summarizes the effect of using household sampling relative to regression estimation for a simple random sample of people.

The gain in using a GREG for a simple random sampling of people is

$$\frac{\text{var}(\hat{Y}_{reg(srs)})}{\text{var}(\hat{Y}_{1(srs)})} = \frac{S_r^2}{S^2}$$

where $S_r^2 = (N-1)^{-1} \sum_{hi \in U_2} (Y_{hi} - \mathbf{B}^T \mathbf{X}_{hi})^2$. The combined effect of using household sampling and regression estimation is

$$\frac{\text{var}(\hat{Y}_{reg})}{\text{var}(\hat{Y}_{1(srs)})} = D_{reg} \frac{S_r^2}{S^2}.$$

In some situations there may be available information that allows assessment of D_1 and S_r^2/S^2 but not D_{reg} .

It would be useful in practice to relate D_{reg} to D_1 , that is the deffs for the variables Y_{hi} and R_{hi} respectively. $D_{reg} \approx D_1$ may be assumed, however it is not clear when this assumption is reasonable, especially as in general D_1 depends on C_{NY2} and C_{NS2} as well as δ_4 . Let δ_{4r} , $C_{NY2(r)}$, $C_{NS2(r)}$, $C_{NY2(r)}^{(u)}$, $C_{NS2(r)}^{(u)}$ and $S_{a(r)}^2$ be the values of δ_4 , C_{NY2} , C_{NS2} , $C_{NY2}^{(u)}$, $C_{NS2}^{(u)}$ and S_a^2 , respectively, for the variable R_{hi} . As in Section 3.2, assume that $C_{NY2(r)} \approx C_{NY2(r)}^{(1)} + C_{NY2(r)}^{(2)}$ and $C_{NS2(r)} \approx C_{NS2(r)}^{(1)}$, and that Y_{hi} is an indicator variable with small mean P . If, in addition, $S_a^2/S^2 = S_{a(r)}^2/S_r^2$, i.e. the variance of R_{hi} within households of a given size is proportional to the corresponding variance of Y_{hi} , then the deffs of $\hat{Y}_{reg}$ for the one/household and all/household design are given by (24) and (25), with δ_4 replaced by δ_{4r} .

The remainder of this subsection focuses on how δ_{4r} relates to δ_4 ; if they are equal then the deffs of $\hat{Y}_{reg}$ may be equal to the deffs of $\hat{Y}_1$. The values of S^2 , S_h^2

and δ_4 for R_{hi} will be denoted by S_r^2 , S_{hr}^2 and δ_{4r} respectively. The corresponding values for the variable $\mathbf{B}^T \mathbf{X}_{hi}$ will be denoted by S_x^2 , S_{hx}^2 and δ_{4x} .

$$\begin{aligned}\delta_{4r} &= 1 - S_r^{-2} \sum_{h \in U_1} \frac{N_h - 1}{N - M} S_{hr}^2 \\ &= 1 - S_r^{-2} \sum_{h \in U_1} \frac{N_h - 1}{N - M} (S_h^2 + S_{hx}^2 - 2S_{hyx})\end{aligned}$$

where $S_{hyx} = (N_h - 1)^{-1} \sum_{hi \in U_{2h}} (Y_{hi} - \bar{Y}_h) (\mathbf{B}^T \mathbf{X}_{hi} - \mathbf{B}^T \bar{\mathbf{X}}_h)$. $\mathbf{B}$ is the population OLS regression parameter, so $S^2 = S_r^2 + S_x^2$. We can expand S_{hyx} as

$$\begin{aligned}S_{hyx} &= (N_h - 1)^{-1} \sum_{hi \in U_{2h}} Y_{hi} (\mathbf{B}^T \mathbf{X}_{hi} - \mathbf{B}^T \bar{\mathbf{X}}_h) \\ &= (N_h - 1)^{-1} \sum_{hi \in U_{2h}} \left\{ (Y_{hi} - \mathbf{B}^T \mathbf{X}_{hi}) (\mathbf{B}^T \mathbf{X}_{hi} - \mathbf{B}^T \bar{\mathbf{X}}_h) + (\mathbf{B}^T \mathbf{X}_{hi} - \mathbf{B}^T \bar{\mathbf{X}}_h)^2 \right\} \\ &= (N_h - 1)^{-1} \sum_{hi \in U_{2h}} \left\{ (\mathbf{B}^T \mathbf{X}_{hi} - \mathbf{B}^T \bar{\mathbf{X}}_h)^2 - (Y_{hi} - \mathbf{B}^T \mathbf{X}_{hi}) \mathbf{B}^T \bar{\mathbf{X}}_h \right\}\end{aligned}$$

Hence

$$\begin{aligned}\sum_{h \in U_1} \frac{N_h - 1}{N - M} S_{hyx} &= \sum_{h \in U_1} \frac{N_h - 1}{N - M} S_{hx}^2 - (N - M)^{-1} \sum_{hi \in U_2} (Y_{hi} - \mathbf{B}^T \mathbf{X}_{hi}) \mathbf{B}^T \bar{\mathbf{X}}_h \\ &= \sum_{h \in U_1} \frac{N_h - 1}{N - M} S_{hx}^2 - \frac{N - M}{N} S_{xr}\end{aligned}$$

where $S_{xr} = (N - 1)^{-1} \sum_{hi \in U_{2h}} R_{hi} \mathbf{B}^T \bar{\mathbf{X}}_h$ is the finite population covariance of

R_{hi} and $\mathbf{B}^T \bar{\mathbf{X}}_h$. Hence

$$\begin{aligned}\delta_{4r} &= 1 - S_r^{-2} \sum_{h \in U_1} \frac{N_h - 1}{N - M} (S_h^2 - S_{hx}^2) + 2S_r^{-2} (1 - \bar{N}^{-1}) S_{xr} \\ &= 1 - S_r^{-2} S^2 (1 - \delta_4) + S_r^{-2} S_x^2 (1 - \delta_{4x}) + 2 (1 - \bar{N}^{-1}) S_r^{-2} S_{xr} \\ &= \frac{S^2}{S_r^2} \delta_4 + \left(1 - \frac{S^2}{S_r^2} \right) \delta_{4x} + 2 \frac{\bar{N} - 1}{\bar{N}} S_r^{-2} S_{xr} \\ &= \delta_4 + \left(\frac{S^2}{S_r^2} - 1 \right) (\delta_4 - \delta_{4x}) + 2 \frac{\bar{N} - 1}{\bar{N}} S_r^{-2} S_{xr}\end{aligned}\tag{28}$$

Notice that S_{xr} measures the covariance between Y_{hi} and $\mathbf{B}^T \bar{\mathbf{X}}_h$, after removing the linear effect of $\mathbf{X}_{hi}$ on Y_{hi} . Suppose S_{xr} is negligible, which will be the case unless Y_{hi} depends on the contextual variable $\bar{\mathbf{X}}_h$ even after correcting for $\mathbf{X}_{hi}$. If Y_{hi} and $\mathbf{B}^T \mathbf{X}_{hi}$ have the same within-household correlation (i.e. $\delta_4 = \delta_{4x}$) then $\delta_{4r} = \delta_4$. If the fitted value based on the auxiliary data, $\mathbf{B}^T \mathbf{X}_{hi}$, has higher within-household correlation than Y (i.e. $\delta_4 < \delta_{4x}$), then $\delta_{4r} < \delta_4$. The empirical study will show that the latter is generally true so that the deff for all/household surveys is slightly less for $\hat{Y}_{reg}$ than for $\hat{Y}_1$.

4.2 $\hat{Y}_1$ and $\hat{Y}_\pi$ as GREGs

Estimators $\hat{Y}_1$ and $\hat{Y}_\pi$ are GREGs under two special cases of the model

$$E_M Y_{hi} = \alpha + \beta N_h^{-1}. \quad (29)$$

$\hat{Y}_\pi$ is the case where α is known to be 0 and $\hat{Y}_1$ is the case where β is known to be 0. The auxiliary variables for model (29) are 1 and N_h^{-1} , which sum to M and N over the population of people U_2 respectively. So $\hat{Y}_\pi$ can be calculated if M is known and $\hat{Y}_1$ can be calculated if N is known.

In Section 2, $\hat{Y}_1$ was shown to have higher variance than $\hat{Y}_\pi$ in some conditions, even though $\hat{Y}_1$ uses N and $\hat{Y}_\pi$ does not. This is because the model underlying $\hat{Y}_1$ is not a generalization of the model underlying $\hat{Y}_\pi$. Hence $\hat{Y}_1$ may perform worse if the inverse of household size is a good predictor of person responses, which will lead to negative covariance between $\bar{Y}_h$ and N_h so that C_{NY} will be negative. The GREG estimator based on the full model (29) always has asymptotic variance

less than or equal to the asymptotic variance of both $\hat{Y}_1$ and $\hat{Y}_\pi$. This GREG is equal to

$$\hat{Y}_{reg} = \hat{Y}_1 + \hat{\beta}M(1 - N/\hat{N}_\pi) \quad (30)$$

where $\hat{\beta} = [m^{-1} \sum_{h \in s_1} \bar{y}_{s2h} - \hat{Y}_\pi/\hat{N}_\pi] / [m^{-1} \sum_{h \in s_1} N_h^{-1} - M/\hat{N}_\pi]$

5 Empirical Study

An empirical study was conducted to compare the deffs for all/household and one/household sampling and examine the model-assisted approximations of Section 3 and the effect of regression estimation. Data were taken from the 1% sample file of the 1991 Australian Census of Population and Housing and the 1995 Australian National Health Survey. Both surveys were conducted by the Australian Bureau of Statistics (ABS). The Census 1% sample was an equal probability sample of households and the Health sample was a sample of households with probability of selection varying by state. Both surveys included all people in the selected households for the variables included here. Households were defined slightly differently for the two surveys; the Health sample had $\bar{N} = 1.90$ and $C_N^2 = 0.171$ and the Census sample had $\bar{N} = 2.02$ and $C_N^2 = 0.187$. The deffs for the one/household and all/household designs and population quantities such as δ_4 were estimated using sample weights supplied by the ABS.

Table 1 shows the deffs of $\hat{Y}_1$ and related quantities for all/household sampling. These deffs vary from 1.099 to 2.611, with most less than 1.41. Equation (16) decomposes the deff into a component involving δ_4 and a component involving

C_{NY2} . Both of these components contributed appreciably to the deffs in Table 1, indicating that the deff is driven by both the correlation within households and the covariance between household size and $\left(\bar{\bar{Y}}_h - \bar{\bar{Y}}\right)^2$. As a result δ_2 , which is a combination of both components, varies greatly across variables. It had a range of 0.108 to 1.577 with most 0.4 or less. For two variables δ_2 exceeds 1 and clearly should not be interpreted as a measure of within household correlation. δ_4 is positive for all variables, varying from 0.027 to 0.863 with most values less than 0.33.

If the population is generated by a model of the form (18), then $E_M C_{NY2}$ is approximately equal to the simpler expression $C_{NY2}^{(1)}$, however the last two columns of Table 1 show that this is not the case. Hence richer models are needed for the relationship between Y_{hi} and N_h , and δ_4 cannot be used to predict the deff without additional information.

Table 2 shows the deffs of $\hat{Y}_1$ for one/household sampling. The approximation discussed in Section 3, $D_1 \approx 1 + C_N^2$ which equals 1.171 and 1.187 for the Health Survey and Census respectively, holds reasonably well on average but the deff ranges between 1.049 and 1.520 over all variables. The deff is driven by the covariances between $\left(\bar{\bar{Y}}_h - \bar{\bar{Y}}\right)^2$ and N_h and N_h^2 , and between N_h and S_h^2 . In most cases, these covariances are positive and hence increase the deff. C_{NS2} is generally well approximated by its expectation under the model, $C_{NS2}^{(1)}$.

Comparing Table 1 and Table 2, we see that generally $D_{1(all)}$ is greater than $D_{1(one)}$. The ratio of the deffs $D_{1(all)}/D_{1(one)}$ varies from 0.93 to 1.87 with most

Table 1: Decomposing Design Effect of $\hat{Y}_1$ for All/HH Design

Survey Variable	$\bar{Y}$	$DEFF$	δ_2	δ_4	$(C_{NY2}/\bar{N}S^2)$	$(C_{NY2}^{(1)}/\bar{N}S^2)$
Alcohol Last 7 Days	0.281	1.717	0.792	0.588	0.185	0.192
Arthritis	0.194	1.184	0.203	0.259	-.051	0.085
Eng 2nd lang	0.156	2.537	1.504	0.863	0.655	0.326
FT student	0.039	1.406	0.397	0.027	0.379	0.010
Health Fair or Poor	0.070	1.233	0.257	0.289	-.029	0.094
Health Poor	0.015	1.099	0.109	0.254	-.131	0.083
High BPR	0.143	1.208	0.230	0.296	-.060	0.097
Indigenous	0.011	2.611	1.577	0.674	0.922	0.254
Post School	0.282	1.259	0.253	0.203	0.052	0.076
Pub Trans	0.038	1.163	0.159	0.087	0.074	0.033
Smoker	0.237	1.326	0.360	0.311	0.045	0.102
employment	0.571	1.402	0.393	0.330	0.065	0.124
income	17077	1.161	0.158	0.117	0.041	0.044
low income	0.278	1.256	0.251	0.170	0.082	0.064
unemployment	0.074	1.324	0.317	0.158	0.162	0.060

Table 2: Decomposing Design Effect of $\hat{Y}_1$ for One/HH Design

Survey Variable	$\bar{Y}$	$DEFF$	δ_4	$(C_{NY2}/\bar{N}^2S^2)$	$(C_{NS2}/\bar{N}^2S^2)$	$(C_{NS2}^{(1)}/\bar{N}^2S^2)$
Alcohol Last 7 Days	0.281	1.175	0.588	0.097	0.078	0.071
Arthritis	0.194	1.090	0.259	-.027	0.117	0.127
Eng 2nd lang	0.156	1.357	0.863	0.324	0.033	0.026
FT student	0.039	1.519	0.027	0.187	0.332	0.182
Health Fair or Poor	0.070	1.113	0.289	-.015	0.128	0.122
Health Poor	0.015	1.051	0.254	-.069	0.119	0.128
High BPR	0.143	1.086	0.296	-.031	0.117	0.121
Indigenous	0.011	1.520	0.674	0.456	0.064	0.061
Post School	0.282	1.167	0.203	0.026	0.141	0.149
Pub Trans	0.038	1.244	0.087	0.037	0.207	0.170
Smoker	0.237	1.156	0.311	0.023	0.132	0.118
employment	0.571	1.177	0.330	0.032	0.145	0.125
income	17077	1.157	0.117	0.020	0.137	0.165
low income	0.278	1.209	0.170	0.041	0.168	0.155
unemployment	0.074	1.294	0.158	0.080	0.214	0.157

less than 1.20. Using the all/household design will be more cost effective than selecting one person per household if the ratio of the average cost per interview for the one/household to the all/household design exceeds $D_{1(all)}/D_{1(one)}$.

As discussed in Section 3.2, C_{NY2} may be decomposed into the model-assisted approximation plus some remainder terms. Table 3 contains this decomposition; it may be seen that $C_{NY2}^{(1)}$ and $C_{NY2}^{(2)}$ are the dominant terms. $C_{NY2}^{(1)}$ is the model-assisted approximation and $C_{NY2}^{(2)}$ measures the covariation of N_h and S_a^2 . Examination of S_a^2 for each $a = 1, \dots, A$ suggested that the major differences were between small households (1-2 adults) and large households (3 or more adults). As discussed in Section 3.2, this implies the deffs can be approximated by (24) and (25) for the important special case of estimating a subpopulation total. These expressions make use of the proportion of people in small households who are also in the subpopulation (P_{G1}), and the proportion of people in large households who are also in the subpopulation (P_{G2}). These approximations for the deffs are labeled $\widehat{DEFF}^b$ in Table 4 and the simple model-assisted approximations are labeled $\widehat{DEFF}^a$. In general, $\widehat{DEFF}^b$ is a better approximation to the deffs, with roughly half the mean relative absolute error (mean over all variables of $|\widehat{DEFF} - DEFF|/DEFF$.)

For example, consider the variable "FT student". The within-household correlation is small for this variable with $\delta_4 = 0.027$. Hence the simple model-assisted approximation $\widehat{DEFF}^a$ is also small for the all/household design (1.038). However, the actual deff is quite large (1.406). The main reason is that 1.9% of

residents of small households are full-time students, compared with 8.1% of residents of large households. This means that the variance S_a^2 is higher in large households. The improved approximation $\widehat{DEFF}^b$ is much closer (1.256) to the true $DEFF$, although still not as close as we would like for a precise determination of the sample size required for a given precision.

Table 5 shows the design effects of a number of GREG estimators, i.e. their variances compared to a corresponding GREG calculated on a SRSWOR of people. Five types of GREG are included:

- $\hat{Y}_1$ which corresponds to $X_{hi} = 1$;
- $\hat{Y}_\pi$ which corresponds to $X_{hi} = N_h^{-1}$;
- a GREG labeled "1/size" corresponding to $\mathbf{X}_{hi} = (1, N_h^{-1})^T$;
- a person level GREG, $\hat{Y}_{person}$, using the model age(10 categories) * sex * region (20 regions) + household composition(1,2,3+ adults by 0,1+ children);
- an integrated weights estimator, $\hat{Y}_{integ}$, (eg Lemaitre & Dufour, 1987) using the same model. This is the GREG using both $\mathbf{X}_{hi}$ and $\bar{\mathbf{X}}_h$ as auxiliary variables - see Steel & Clark, 2007. For the all/household design this equals a GREG applied to the household totals of the variables, Y_h and $\mathbf{X}_h$.

The parameters δ_{4r} and δ_{4x} were calculated in the same way as δ_4 , using R_{hi} and the model fitted values, respectively, in place of Y_{hi} . Notice that $\delta_{4r} = \delta_4$ for $\hat{Y}_1$; $\delta_{4x} = 1$ for $\hat{Y}_\pi$ and (1/size) because the auxiliary variables are constant within households for these estimators; and $\delta_{4x} = 0$ for $\hat{Y}_1$ because the auxiliary

Table 3: Decomposition of C_{NY2}

Survey Variable	$\frac{C_{NY2}}{NS^2}$	$\frac{C_{NY2}^{(1)}}{NS^2}$	$\frac{C_{NY2}^{(2)}}{NS^2}$	$\frac{C_{NY2}^{(3)}}{NS^2}$	$\frac{C_{NY2}^{(4)}}{NS^2}$	$\frac{C_{NY2}^{(5)}}{NS^2}$	$\frac{C_{NY2}^{(1)} + \hat{C}_{NY2}^{(2)}}{NS^2}$
Alcohol Last 7 Days	0.185	0.192	0.016	-.008	-.016	0.001	-.001
Arthritis	-.051	0.085	-.092	-.004	-.043	0.003	-.050
Eng 2nd lang	0.655	0.326	0.171	0.062	0.008	0.089	0.177
FT student	0.379	0.010	0.213	0.006	0.116	0.033	0.219
Health Fair or Poor	-.029	0.094	-.097	0.010	-.036	-.001	-.013
Health Poor	-.131	0.083	-.170	0.009	-.052	-.000	-.046
High BPR	-.060	0.097	-.105	-.002	-.050	0.001	-.051
Indigenous	0.922	0.254	0.198	0.238	0.149	0.083	0.190
Post School	0.052	0.076	-.004	-.008	-.018	0.005	-.031
Pub Trans	0.074	0.033	0.033	0.005	0.003	0.002	0.042
Smoker	0.045	0.102	-.028	0.004	-.031	-.002	-.004
employment	0.065	0.124	-.003	-.002	-.046	-.009	0.015
income	0.041	0.044	-.004	-.007	-.001	0.009	-.019
low income	0.082	0.064	0.021	0.001	-.007	0.003	0.021
unemployment	0.162	0.060	0.067	0.014	0.011	0.010	0.076

Table 4: Design Effect Approximations

Survey Variable	All/Household			One/Household			$P_{1,2}$	P_{3+}
	DEFF	$D\widehat{EFF}^a$	$D\widehat{EFF}^b$	DEFF	$D\widehat{EFF}^a$	$D\widehat{EFF}^b$		
Alcohol Last 7 Days	1.717	1.724	1.723	1.175	1.171	1.171	0.281	0.280
Arthritis	1.184	1.319	1.269	1.090	1.171	1.145	0.212	0.143
Eng 2nd lang	2.537	2.207	2.384	1.357	1.187	1.274	0.121	0.229
FT student	1.406	1.038	1.256	1.519	1.187	1.295	0.019	0.081
Health Fair or Poor	1.233	1.356	1.343	1.113	1.171	1.165	0.072	0.065
Health Poor	1.099	1.312	1.266	1.051	1.171	1.147	0.016	0.011
High BPR	1.208	1.364	1.313	1.086	1.171	1.145	0.156	0.106
Indigenous	2.611	1.943	2.133	1.520	1.187	1.281	0.008	0.017
Post School	1.259	1.283	1.252	1.167	1.187	1.171	0.299	0.245
Pub Trans	1.163	1.121	1.163	1.244	1.187	1.207	0.035	0.045
Smoker	1.326	1.383	1.379	1.156	1.171	1.170	0.239	0.233
employment	1.402	1.462	1.476	1.177	1.187	1.194	0.556	0.602
low income	1.256	1.238	1.259	1.209	1.187	1.197	0.266	0.303
unemployment	1.324	1.221	1.298	1.294	1.187	1.224	0.062	0.099
Mean Rel. Error (%)		0.102	0.063		0.077	0.053		

variable is equal to 1. For the all/household design, the deffs of the person level GREG are generally a little lower or the same than the deffs for $\hat{Y}_1$. As discussed in Section 4.1, this can be explained by $\delta_{4x} > \delta_4$ i.e. the auxiliary variables are more correlated within households than the survey variables. The contextual effect $2\frac{\bar{N}-1}{N}S_r^{-2}S_{xr}$, which appears in equation (28), was very small in all cases. For the one/household design, the deffs of $\hat{Y}_1$ and $\hat{Y}_{person}$ were generally similar.

Table 5 also shows that $\hat{Y}_1$ had lower variance than $\hat{Y}_\pi$ for all variables except "High BPR" and "arthritis". The estimator "(1/size)" which is defined in (30) had lower variance than both $\hat{Y}_\pi$ and $\hat{Y}_1$, however these three estimators had very similar variance for all estimators considered. Integrated weighting had around the same variance as the person level GREG for both designs and all variables.

Table 6 is similar to Table 5 but includes the deff relative to regression estimation for a SRSWOR, (D_{reg}) , and the deff relative to $\hat{Y}_1$ for a SRSWOR.

This article has focused on the widely used one/household and all/household designs, however other designs could be used. Table 7 shows the deff of $\hat{Y}_1$ for the one/household and all/household designs, and the designs where (up to) 2,3 and 4 adults/household are selected, and the design where half the adults in each household are selected, rounding up for odd numbers of people. The "a/household" designs with $a > 1$ were always out-performed by either the one/household or the all/household design. The half/household design was efficient for many variables; the variance was usually at least 5% lower than both the one/household and all/household designs. This can be explained by comparing the variance ex-

Table 5: DEFF of GREGs for Several Designs

Survey Variable	Type of GREG	Deff		S_r^2/S^2	δ_{4r}	δ_{4x}	$2\frac{N-1}{N}\frac{S_{xr}}{S_r^2}$
		All/HH	One/HH				
Arthritis	(1/size)	1.171	1.091	0.985	0.248	1.000	-0.000
Arthritis	Integrated	1.080	1.099	0.785	0.150	0.657	-0.000
Arthritis	$\hat{Y}_1$	1.184	1.090	1.000	0.259	0.000	0.000
Arthritis	Person	1.077	1.100	0.795	0.145	0.621	-0.021
Arthritis	$\hat{Y}_\pi$	1.176	1.090	0.991	0.252	1.000	-0.000
Eng 2nd lang	(1/size)	2.503	1.342	0.984	0.861	1.000	-0.000
Eng 2nd lang	Integrated	2.440	1.320	0.885	0.847	0.983	-0.000
Eng 2nd lang	$\hat{Y}_1$	2.537	1.357	1.000	0.863	0.000	0.000
Eng 2nd lang	Person	2.440	1.324	0.890	0.842	0.941	-0.011
Eng 2nd lang	$\hat{Y}_\pi$	2.623	1.394	1.054	0.870	1.000	0.000
FT student	(1/size)	1.375	1.512	0.989	0.016	1.000	0.000
FT student	Integrated	1.367	1.481	0.821	0.037	-0.018	-0.000
FT student	$\hat{Y}_1$	1.406	1.519	1.000	0.027	0.000	0.000
FT student	Person	1.366	1.482	0.827	0.037	0.018	0.008
FT student	$\hat{Y}_\pi$	1.457	1.532	1.015	0.041	1.000	0.000
Health Fair or Poor	(1/size)	1.233	1.115	0.997	0.287	1.000	0.000
Health Fair or Poor	Integrated	1.221	1.117	0.941	0.273	0.545	-0.000
Health Fair or Poor	$\hat{Y}_1$	1.233	1.113	1.000	0.289	0.000	0.000
Health Fair or Poor	Person	1.215	1.116	0.953	0.269	0.558	-0.007
Health Fair or Poor	$\hat{Y}_\pi$	1.239	1.117	1.000	0.289	1.000	-0.000
High BPR	(1/size)	1.201	1.086	0.990	0.288	1.000	-0.000
High BPR	Integrated	1.068	1.099	0.715	0.144	0.675	-0.000
High BPR	$\hat{Y}_1$	1.208	1.086	1.000	0.296	0.000	0.000
High BPR	Person	1.065	1.099	0.727	0.140	0.695	-0.006
High BPR	$\hat{Y}_\pi$	1.205	1.087	0.994	0.291	1.000	-0.000
Post School	(1/size)	1.257	1.166	1.000	0.202	1.000	-0.000
Post School	Integrated	1.270	1.169	0.902	0.209	0.140	0.000
Post School	$\hat{Y}_1$	1.259	1.167	1.000	0.203	0.000	0.000
Post School	Person	1.269	1.169	0.908	0.209	0.133	-0.000
Post School	$\hat{Y}_\pi$	1.290	1.156	1.050	0.241	1.000	-0.000
Smoker	(1/size)	1.327	1.157	0.998	0.310	1.000	0.000
Smoker	Integrated	1.317	1.165	0.935	0.290	0.609	0.000
Smoker	$\hat{Y}_1$	1.326	1.156	1.000	0.311	0.000	0.000
Smoker	Person	1.313	1.166	0.947	0.285	0.465	-0.017
Smoker	$\hat{Y}_\pi$	1.368	1.167	1.024	0.327	1.000	0.000
employment	(1/size)	1.410	1.188	0.988	0.322	1.000	-0.000
employment	Integrated	1.357	1.232	0.713	0.225	0.590	-0.000
employment	$\hat{Y}_1$	1.402	1.177	1.000	0.330	0.000	0.000
employment	Person	1.354	1.232	0.722	0.223	0.498	-0.043
employment	$\hat{Y}_\pi$	1.636	1.195	1.255	0.466	1.000	0.000
income	(1/size)	1.157	1.155	0.999	0.117	1.000	-0.000
income	Integrated	1.279	1.159	0.772	0.218	-0.225	-0.000
income	$\hat{Y}_1$	1.161	1.157	1.000	0.117	0.000	0.000
income	Person	1.280	1.155	0.788	0.222	-0.102	0.046
income	$\hat{Y}_\pi$	1.301	1.153	1.145	0.229	1.000	-0.000
low income	(1/size)	1.253	1.209	0.998	0.169	1.000	0.000
low income	Integrated	1.287	1.214	0.878	0.193	0.001	-0.000
low income	$\hat{Y}_1$	1.256	1.209	1.000	0.170	0.000	0.000
low income	Person	1.286	1.211	0.889	0.195	0.094	0.015
low income	$\hat{Y}_\pi$	1.366	1.227	1.067	0.222	1.000	-0.000

Table 6: Two Definitions of the DEFF of GREGs

Survey Variable	Type of GREG	$D_{reg} = \frac{var(\hat{Y}_{reg})}{var(\hat{Y}_{reg(srs)})}$		$DEFF = \frac{var(\hat{Y}_{reg})}{var(\hat{Y}_{\pi(srs)})}$	
		All/HH	One/HH	All/HH	One/HH
Arthritis	(1/size)	1.171	1.091	1.153	1.074
Arthritis	Integrated	1.080	1.099	0.848	0.862
Arthritis	$\hat{Y}_1$	1.184	1.090	1.184	1.090
Arthritis	Person	1.077	1.100	0.857	0.875
Arthritis	$\hat{Y}_\pi$	1.176	1.090	1.165	1.080
Eng 2nd lang	(1/size)	2.503	1.342	2.464	1.321
Eng 2nd lang	Integrated	2.440	1.320	2.160	1.168
Eng 2nd lang	$\hat{Y}_1$	2.537	1.357	2.537	1.357
Eng 2nd lang	Person	2.440	1.324	2.172	1.179
Eng 2nd lang	$\hat{Y}_\pi$	2.623	1.394	2.763	1.469
FT student	(1/size)	1.375	1.512	1.360	1.496
FT student	Integrated	1.367	1.481	1.122	1.215
FT student	$\hat{Y}_1$	1.406	1.519	1.406	1.519
FT student	Person	1.366	1.482	1.129	1.225
FT student	$\hat{Y}_\pi$	1.457	1.532	1.479	1.555
Health Fair or Poor	(1/size)	1.233	1.115	1.230	1.111
Health Fair or Poor	Integrated	1.221	1.117	1.149	1.050
Health Fair or Poor	$\hat{Y}_1$	1.233	1.113	1.233	1.113
Health Fair or Poor	Person	1.215	1.116	1.158	1.064
Health Fair or Poor	$\hat{Y}_\pi$	1.239	1.117	1.239	1.116
High BPR	(1/size)	1.201	1.086	1.188	1.075
High BPR	Integrated	1.068	1.099	0.764	0.786
High BPR	$\hat{Y}_1$	1.208	1.086	1.208	1.086
High BPR	Person	1.065	1.099	0.774	0.798
High BPR	$\hat{Y}_\pi$	1.205	1.087	1.197	1.080
Post School	(1/size)	1.257	1.166	1.257	1.166
Post School	Integrated	1.270	1.169	1.146	1.055
Post School	$\hat{Y}_1$	1.259	1.167	1.259	1.167
Post School	Person	1.269	1.169	1.153	1.061
Post School	$\hat{Y}_\pi$	1.290	1.156	1.355	1.214
Smoker	(1/size)	1.327	1.157	1.325	1.155
Smoker	Integrated	1.317	1.165	1.231	1.089
Smoker	$\hat{Y}_1$	1.326	1.156	1.326	1.156
Smoker	Person	1.313	1.166	1.243	1.104
Smoker	$\hat{Y}_\pi$	1.368	1.167	1.400	1.195
employment	(1/size)	1.410	1.188	1.393	1.173
employment	Integrated	1.357	1.232	0.968	0.879
employment	$\hat{Y}_1$	1.402	1.177	1.402	1.177
employment	Person	1.354	1.232	0.978	0.890
employment	$\hat{Y}_\pi$	1.636	1.195	2.054	1.500
income	(1/size)	1.157	1.155	1.156	1.155
income	Integrated	1.279	1.159	0.988	0.895
income	$\hat{Y}_1$	1.161	1.157	1.161	1.157
income	Person	1.280	1.155	1.008	0.910
income	$\hat{Y}_\pi$	1.301	1.153	1.491	1.320
low income	(1/size)	1.253	1.209	1.251	1.207
low income	Integrated	1.287	1.214	1.129	1.066
low income	$\hat{Y}_1$	1.256	1.209	1.256	1.209
low income	Person	1.286	1.211	1.143	1.077
low income	$\hat{Y}_\pi$	1.366	1.227	1.458	1.309

pression (4) for designs which have the same $\bar{n}$ but different n_h . The first term of this expression depends on $\bar{n}$ but not on n_h . The second term is minimized for fixed n_0 by choosing $n_h \propto N_h S_h$, so if S_h^2 vary little then $n_h \propto N_h$ is optimal. In this case the half/household design should be more efficient than the one/household design. More investigation of allocation rules $n_h = f(N_h)$ would be worthwhile, to establish even more efficient alternatives, and to examine cost, operational and data quality implications of this type of design.

6 Conclusions

It is usually assumed that the deff for all/household sampling is mainly due to within-household correlations; and the deff for one/household sampling is mainly due to variation in household size (C_N^2). This is generally not the case. The covariation between household size and S_h^2 and $(\bar{Y}_h - \bar{Y})^2$ are also important and this effect is inconsistent across survey variables. Thus assessing the deff associated with using the household as a sampling unit involves assessing several population parameters, not just the within household correlation.

Expressions for the deff were derived in Section 3, which depended on an intra-household correlation measure δ_4 , and on C_{NS2} and C_{NY2} which measure covariations with household size. The effect of these covariation parameters was found in the empirical study to be as great as the effect of δ_4 , but they would be difficult to predict in practice. Simple approximations based on a model worked well for C_{NS2} but poorly for C_{NY2} : the latter term cannot be predicted using

only δ_4 and C_N^2 . The all/household design generally had a higher deff than the one/household design, although not for all variables and often not by much. The all/household design is generally cheaper per interview so that, if feasible, this design would be preferred for most variables. For sensitive or lengthy interviews, the one/household design may be necessary to ensure respondent cooperation.

Some approximations to the deffs, (24) and (25), were found to work well for estimators of the size of a subpopulation. Instead of C_{NS2} and C_{NY2} , the proportion of the population falling in the subpopulation in small households (1-2 adults) and large households (3 or more adults) is required. These proportions could be estimated from a pilot study, or predicted based on experience or judgement, to assist the choice of a sample design.

The deffs for GREG estimators were generally lower than the deffs for the Hajek estimator $\hat{Y}_1$, particularly for the all/household design. Equivalently, the benefit from GREG estimation is less for the all/household design than for simple random sampling of people. The empirical results and the argument in Section 4.1 suggested that this was due to the fitted values being more highly correlated within households than the survey variables. The Hajek estimator and the π -estimator can be expressed as special cases of the GREG, which explains why the Hajek estimator has higher variance for some variables.

This paper gives results relating deffs to a small number of population parameters, which can be used to calculate approximations to the deffs based on the limited data often available for sample design. After the survey has been

conducted, replication-based methods can be used to fully account for complex estimators and sample designs. The decompositions derived here and elsewhere can complement these methods, by providing simple checks on direct variance estimates, and allowing deffs to be extrapolated to alternative sample designs for future surveys.

Acknowledgements: This work was jointly supported by the Australian Research Council and the Australian Bureau of Statistics. The views expressed here do not necessarily reflect the views of either organization. The authors thank Frank Yu and John Kiely for their helpful comments.

References

- Brewer, K., Foreman, E., & Trewin, D. (1977). Use of experimental design and population modelling in survey sampling. *Bull.Int.Statist.Inst.*, 47(Book 3), 173–190.
- Cochran, W. G. (1977). *Sampling Techniques* (3 ed.). New York: Wiley.
- Foster, K. (1994). The coverage of the postcode address file as a sampling frame. *OPCS Survey Methodology Bulletin*, 34, 9–18.
- Groves, R. M. (1989). *Survey errors and survey costs*. New York: Wiley.
- Hansen, M., Hurwitz, W., & Madow, W. (1953). *Sample Survey Methods and Theory Vol.1 and 2*. New York: Wiley.
- Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). London: Prentice-Hall International.
- Kish, L. (1967). *Survey Sampling*. New York: Wiley.
- Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11, 55–77.
- Kish, L., & Frankel, M. (1974). Inference from complex samples (with discussion). *Journal of the Royal Statistical Society Series B*, 36, 1–37.
- Lemaitre, G., & Dufour, J. (1987). An integrated method for weighting persons and families. *Survey Methodology*, 13(2), 199–207.
- Lepowski, J. M. (1988). Telephone sampling methods in the united states. In R. M. Groves, P. P. Biemer, L. E. Lyberg, J. T. Massey, W. L. N. 2, & J. Waksberg (Eds.), *Telephone survey methodology* (pp. 73–98). Wiley.

- Sarndal, C., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling*. New York: Springer-Verlag.
- Skinner, C. (1986). Design effects of two-stage sampling. *Journal of the Royal Statistical Society Series B*, 48, 88–99.
- Steel, D. G., & Clark, R. G. (2007). Person-level and household-level regression estimation in household surveys. *Survey Methodology*, 33(1), 51–60.
- Valliant, R. (1987). Generalized variance functions in stratified two-stage sampling. *Journal of the American Statistical Association*, 82(398), 499–508.
- Verma, V., Scott, C., & O’Muircheartaigh, C. (1980). Sample designs and sampling errors for the world fertility survey. *Journal of the Royal Statistical Society Series A*, 143, 431–473.

Table 7: Design Effect of $\hat{Y}_1$ for Alternative Designs

Survey Variable	Sample Design					
	One/HH	Two/HH	Three/HH	Four/HH	All/HH	Half/HH
employment	1.177	1.384	1.392	1.399	1.402	1.133
unemployment	1.294	1.396	1.344	1.330	1.324	1.145
low income	1.209	1.287	1.260	1.257	1.256	1.123
income	1.157	1.170	1.155	1.158	1.161	1.115
Eng 2nd lang	1.357	2.226	2.442	2.514	2.537	1.578
Indigenous	1.520	2.326	2.525	2.592	2.611	1.716
FT student	1.519	1.607	1.464	1.420	1.406	1.187
Pub Trans	1.244	1.251	1.187	1.169	1.163	1.095
Post School	1.167	1.258	1.251	1.255	1.259	1.122
Smoker	1.156	1.320	1.321	1.325	1.326	1.111
Health Fair or Poor	1.113	1.233	1.225	1.231	1.233	1.061
Health Poor	1.051	1.102	1.092	1.095	1.099	1.011
Alcohol Last 7 Days	1.175	1.613	1.688	1.712	1.717	1.228
Arthritis	1.090	1.175	1.175	1.182	1.184	1.062
High BPR	1.086	1.197	1.199	1.206	1.208	1.055