Centre for Statistical and Survey Methodology

The University of Wollongong

Working Paper

05-08

Adaptive Calibration for Prediction of Finite Population Totals

Robert G. Clark and Raymond L. Chambers

Centre for Statistical & Survey Methodology, University of Wollongong, Wollongong NSW 2522. Phone +61 2 4221 5435, Fax +61 2 4221 4845. Email: anica@uow.edu.au

Adaptive Calibration for Prediction of Finite Population Totals

Robert G. Clark and Raymond L. Chambers [1]

Abstract

Sample weights can be calibrated to reflect the known population totals of a set of auxiliary variables. Predictors of finite population totals calculated using these weights have low bias if the these variables are related to the variable of interest, but can have high variance if too many auxiliary variables are used. This article develops an "adaptive calibration" approach, where the auxiliary variables to be used in weighting are selected using sample data. Adaptively calibrated estimators are shown to have lower mean squared error and better coverage properties than non-adaptive estimators in many cases.

Key Words: sample surveys, sample weighting, prediction approach, ridge estimation, model selection, stepwise procedures

1. Introduction

Predictors of finite population totals are commonly calculated by weighted sums of sample values. Auxiliary variables are often available, whose sample values and population totals are known. Weights can be constructed so that weighted sums of auxiliary variables agree with the known population totals, a process called calibration (Deville & Sarndal, 1992). Predictors of finite population totals based on calibrated weights generally have much lower prediction bias than predictors calculated without auxiliary information.

Existing literature on finite population prediction essentially assumes that a set of useful auxiliary variables is chosen without reference to sample data. In practice, however, there may be a large set of potential auxiliary variables, not all

[1] Centre for Statistical and Survey Methodology, University of Wollongong, NSW 2522 Australia. E-mail: Robert_Clark@uow.edu.au

of which should be used. Using additional auxiliary variables generally reduces the bias of calibrated predictors but increases the variance, so that using too many auxiliary variables can actually increase the mean squared error of calibrated predictors. The choice of which auxiliary variables to use is often not obvious, and sample data may be required to determine which set of auxiliary variables is appropriate for predictors of the totals of particular variables of interest. This paper develops methods for making this determination. Our approach may be called adaptive calibration, because the set of variables is chosen adaptively from sample data, rather than statically without reference to the sample at hand.

The prediction (or model-based) framework to finite population estimation will be used (see for example Brewer, 1963; Royall, 1970; Valliant et al., 2000). In this approach, the population values of the variables of interest are treated as random variables. The aim is to predict the population total (which is also a random variable) or other finite population quantities using sample data on the variable of interest, and population data on some auxiliary variables. The sample may have been selected using probability sampling or some other method, and is conditioned upon in inference. A stochastic model for the variable of interest is a central feature. One feature of the prediction framework is that mis-specification of the model, for example due to omitting important auxiliary variables, can lead to substantial bias.

An alternative framework is the model-assisted approach (Sarndal et al., 1992). In this approach, a stochastic model is used but the model plays a less crucial role. The randomized nature of sampling is exploited to ensure that estimators are approximately unbiased even if the model is incorrect. When the model is correct, both approaches give approximately unbiased estimators, but the model-based approach would generally give lower variances of estimators of interest. If the model is mis-specified, then model-based predictors and variance estimators may be more biased, however robust model-based methods have been developed to combat this problem. For example Royall and Herson (1973a, 1973b) discuss robust prediction, and Royall and Cumberland (1981a, 1981b) develop variance estimators which are robust to heteroskedasticity. For comparisons of the prediction and model-assisted frameworks, see Smith (1976) and Hansen et al. (1983).

The problem of selecting a set of auxiliary variables in the model-assisted framework was considered by Silva and Skinner (1997) and Skinner and Silva (1997). They found that adding calibration variables reduces the mean squared error (MSE) up to a point, after which adding further variables increases the MSE. Choosing calibration variables adaptively, based on sample data, gave better estimates than either calibrating on all variables or no variables. The applicability of this work to model-based prediction is not clear, because the role of the model is very different in the two frameworks. Mis-specified models can lead to substantially biased model-based predictors, whereas model-assisted estimators are approximately unbiased even if important variables are omitted. As a result, different strategies for model selection could be appropriate in the two frameworks. Moreover, the differences between alternative approaches would be expected to be more pronounced in the prediction framework than in the model-assisted framework.

Chambers et al. (1999) proposed an approach for selecting calibration variables in the prediction framework, using forward, backward or stepwise selection. (This paper will henceforth be referred to as CSW.) The decision whether to omit (or add) a variable at each step was based on minimizing the estimated squared error of prediction (MSEP) for the predictor of interest. The approach was not evaluated by simulation study, and the estimators of MSEP used were not robust to heteroskedasticity.

The purpose of this paper is to develop the basic approach of CSW to apply to a wider range of situations, including heteroskedastic populations and multi-stage samples, and to evaluate the approach using realistic simulation studies. Estimators of the MSEP which are robust to heteroskedasticity, and to correlation in the case of multi-stage surveys, will be used. The performance of the estimators will be evaluated by simulation from two populations generated from financial data from a farm survey and labour force data from a population census.

Following CSW, the basic approach will be to build a set of auxiliary variables using stepwise selection of variables, starting with some initial set. This algorithm builds up a set of auxiliary variables by a sequence of many decisions between two nested sets of variables. We compare several alternative criteria for deciding between two nested sets, including statistical significance and a number of alternative estimators of the mean squared error of prediction (MSEP). Three alternative estimators of MSEP are considered: a non-robust estimator; an estimator of MSEP which is robust to heteroskedasticity; and an estimator which is robust both to heteroskedasticity and correlations within primary sampling units in multi-stage sampling.

Section 2 contains notation and definitions. Section 3 derives the difference in the MSEP of two predictors based on nested models, and develops several alternative estimators of this difference. Section 4 contains simulation results for a farm survey and a multi-stage household survey. Section 5 is a discussion. We conclude that adaptive calibration generally performs better than static calibration, provided that a non-robust estimator of the MSEP is used as the objective in model selection.

2. Notation and Definitions

A variable of interest Y_i is observed for a sample s of n units, which is a subset of a finite population U containing N units. The aim is to estimate the population total $T_Y = \sum_{i \in U} Y_i$ and other finite population quantities of Y. A vector of auxiliary variables $\boldsymbol{x_i}$ is available for $i \in s$, with known population total $\boldsymbol{T}_x = \sum_{i \in U} \boldsymbol{x_i}$.

Weighted estimators of T_Y are given by $\hat{T}_Y = \sum_{i \in s} w_i Y_i$, where w_i can depend on the auxiliary variables but not on the variable of interest. A set of weights is said to be calibrated on $\boldsymbol{x_i}$ if

$$\sum_{i \in s} w_i \boldsymbol{x_i} = \boldsymbol{T_x}.$$

The best linear unbiased predictor (BLUP) is one example of a calibrated estimator. The most commonly used BLUP is based on the model

$$\begin{array}{l} E\left[Y_i\right] = \boldsymbol{\beta^T} \boldsymbol{x_i} \\ var\left[Y_i\right] = \sigma_i^2 = v_i \sigma^2 \\ cov\left[Y_i, Y_j\right] = 0 (i \neq j) \end{array} \tag{1}$$

(with v_i assumed to be known) and is given by

$$\hat{T}_Y = \sum_{i \in s} Y_i + \sum_{i \in r} \boldsymbol{\hat{\beta}^T} \boldsymbol{x_i} \tag{2}$$

where $r = U - s$ is the set of non-sample units and

$$\boldsymbol{\hat{\beta}} = \left\{ \sum_{i \in s} v_i^{-1} \boldsymbol{x_i} \boldsymbol{x_i^T} \right\}^{-1} \sum_{i \in s} v_i^{-1} \boldsymbol{x_i} Y_i$$

is a weighted least squares estimator of $\boldsymbol{\beta}$. The BLUP can also be written in weighted form as

$$\hat{T}_Y = \sum_{i \in s} w_i Y_i$$

where the weights w_i are given by

$$w_i = 1 + \boldsymbol{T_{xr}}^T \left\{ \sum_{j=1}^{n} v_j^{-1} \boldsymbol{x_j} \boldsymbol{x_j^T} \right\}^{-1} v_i^{-1} \boldsymbol{x_i} \tag{3}$$

where $\boldsymbol{T_{xr}} = \sum_{i \in r} \boldsymbol{x_i}$. It is straightforward to verify that $\sum_{i \in s} w_i \boldsymbol{x_i} = \boldsymbol{T_x}$.

For heteroskedastic data, it is usually difficult to model v_i reliably. In this case, robust estimators of the prediction variance of the BLUP are available, which do not rely on knowledge of v_i (Royall & Cumberland, 1981b). For multi-stage samples, the assumption of independence may be violated. In this case, the BLUP based on (1) may still be used, and a robust ultimate cluster variance estimator of its prediction variance can be used (e.g. Valliant et al., 2000, Chapter 9). An alternative approach, which will not be considered here, would be to construct a BLUP based on a model for the within-cluster correlations (Royall, 1976). Section 3 will discuss robust and non-robust estimation of the mean squared error of prediction of the BLUP in more detail.

A decision needs to made on what to include in $\boldsymbol{x_i}$ in the BLUP. Stepwise selection, forward selection and backward selection are algorithms that can be used to decide which subset of the available auxiliary variables should be used. All three algorithms include many choices between two nested sets of auxiliary variables. Suppose the choice is between (A) using a predictor $\hat{T}_A$ based on $\boldsymbol{x_i}$ and (B) using a predictor $\hat{T}_B$ based on a subvector $\boldsymbol{x_{1i}}$. We can partition $\boldsymbol{x_i}$ as $\boldsymbol{x_i} = \left(\boldsymbol{x_{1i}^T}, \boldsymbol{x_{2i}^T}\right)^T$. The number of elements of $\boldsymbol{x_i}$, $\boldsymbol{x_{1i}}$ and $\boldsymbol{x_{2i}}$ are denoted by p, p_1 and p_2, respectively.

We similarly partition $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = \left(\boldsymbol{\beta_1^T}, \boldsymbol{\beta_2^T}\right)^T$. Predictor $\hat{T}_A$ is unbiased under model A:

$$E\left[Y_i\right] = \boldsymbol{\beta^T x_i} = \boldsymbol{\beta_1^T x_{1i}} + \boldsymbol{\beta_2^T x_{2i}}. \tag{4}$$

The predictor $\hat{T}_B$ is unbiased for model B,

$$E\left[Y_i\right] = \boldsymbol{\beta_1^T x_{1i}}, \tag{5}$$

which is the special case of model A where $\boldsymbol{\beta_2} = \mathbf{0}$.

3. Estimation of the Difference in the MSEP

3.1 Comparing Predictors from Nested Models

Following CSW, our approach is to estimate the difference in the MSEPs of the two estimators:

$$\Delta = E\left[\left(\hat{T}_A - T_Y\right)^2\right] - E\left[\left(\hat{T}_B - T_Y\right)^2\right]$$

where the expectations are evaluated with respect to model (4), because model (5) is a special case of this model. Typically, $\hat{T}_A$ will be less biased than $\hat{T}_B$ but have higher variance. Either predictor can have higher or lower MSEP depending on the particular population and sample.

For single stage sampling, it is usually reasonably to assume Y_i and Y_j independent for all $i \neq j$, with known or approximately known variance. Section 3.2 will derive Δ and an estimator of it in this case. Section 3.3 will describe the instructive special case where variances are equal and BLUPs are used; this was the case considered by CSW. Section 3.4 extends this by describing a heteroskedasticity-robust estimator of Δ. Section 3.5 further extends the approach by deriving Δ and an estimator of it for multi-stage sampling where there may be correlations between values from the same cluster.

3.2 Estimating Δ in Single-Stage Sampling with Known Variance

In addition to model (4), we assume in this subsection that Y_i and Y_j are independent for $i \neq j$ and that $var\left[Y_i\right] = \sigma_i^2 = \sigma^2 v_i$ where v_i are known. In this case, the MSEP of any predictor $\hat{T} = \sum_{i \in s} w_i Y_i$ is given by

$$\begin{aligned} MSEP\left[\hat{T}\right] &= E\left[\left(\hat{T} - T_Y\right)^2\right] \\ &= \left\{E\left[\sum_{i \in s} w_i Y_i - \sum_{i \in U} Y_i\right]\right\}^2 + var\left[\sum_{i \in s} (w_i - 1) Y_i - \sum_{i \in r} Y_i\right] \\ &= \left\{\boldsymbol{\beta^T}\left(\sum_{i \in s} w_i \boldsymbol{x_i} - \sum_{i \in U} \boldsymbol{x_i}\right)\right\} + \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2 \end{aligned}$$

Writing $\boldsymbol{d} = \sum_{i \in s} w_i \boldsymbol{x_i} - \boldsymbol{T_x}$, we can rewrite the MSEP as

$$MSEP\left[\hat{T}\right] = \boldsymbol{d^T}\left(\boldsymbol{\beta\beta^T}\right)\boldsymbol{d} + \sum_{i \in s} (w_i - 1)^2 \sigma_i^2 + \sum_{i \in r} \sigma_i^2$$

Let $\boldsymbol{d_A} = \sum_{i \in s} w_{Ai} \boldsymbol{x_i} - \boldsymbol{T_x}$ and $\boldsymbol{d_B} = \sum_{i \in s} w_{Bi} \boldsymbol{x_i} - \boldsymbol{T_x}$. Then Δ is given by:

$$\begin{aligned} \Delta &= MSEP\left[\hat{T}_A\right] - MSEP\left[\hat{T}_B\right] \\ &= \boldsymbol{d_A^T}\left(\boldsymbol{\beta\beta^T}\right)\boldsymbol{d_A} - \boldsymbol{d_B^T}\left(\boldsymbol{\beta\beta^T}\right)\boldsymbol{d_B} + \sum_{i \in s} (w_{Ai} - 1)^2 \sigma_i^2 - \sum_{i \in s} (w_{Bi} - 1)^2 \sigma_i^2 \end{aligned} \tag{6}$$

To estimate Δ, we first consider how to estimate $\boldsymbol{\beta}$ and the variance of $\hat{\boldsymbol{\beta}}$. The usual weighted least squares estimator is $\hat{\boldsymbol{\beta}} = S_x^{-1} S_{xy}$ where $S_x = \sum_{i \in s} v_i^{-1} \boldsymbol{x_i} \boldsymbol{x_i^T}$ and $S_{xy} = \sum_{i \in s} v_i^{-1} \boldsymbol{x_i} Y_i$. The usual estimator of the variance of $\hat{\boldsymbol{\beta}}$ is $v\hat{a}r\left[\hat{\boldsymbol{\beta}}\right] = \hat{\sigma}^2 S_x^{-1}$ where $\hat{\sigma^2} = \sum_{i \in s}\left(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{x_i}\right)^2 / (n - p)$.

We can estimate $\left(\boldsymbol{\beta\beta^T}\right)$ unbiasedly by $\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - v\hat{a}r\left[\hat{\boldsymbol{\beta}}\right]\right)$. Hence the following is an unbiased estimator of Δ:

$$\begin{aligned} \hat{\Delta} &= \boldsymbol{d_A^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - v\hat{a}r\left[\hat{\boldsymbol{\beta}}\right]\right)\boldsymbol{d_A} - \boldsymbol{d_B^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - v\hat{a}r\left[\hat{\boldsymbol{\beta}}\right]\right)\boldsymbol{d_B} \\ &\quad + \sum_{i \in s} (w_{Ai} - 1)^2 \hat{\sigma}^2 v_i - \sum_{i \in s} (w_{Bi} - 1)^2 \hat{\sigma}^2 v_i \end{aligned} \tag{7}$$

Expression (6) applies, and estimator (7) is an unbiased estimator of it, for any weighted predictor. We are concerned with the special case where $\hat{T}_A$ and $\hat{T}_B$ are BLUPs. In this case, $\hat{T}_A$ is calibrated to $\boldsymbol{T_x}$ so that $\boldsymbol{d_A} = \mathbf{0}$, and so (7) simplifies to

$$\hat{\Delta} = -\boldsymbol{d_B^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - v\hat{a}r\left[\hat{\boldsymbol{\beta}}\right]\right)\boldsymbol{d_B} + \sum_{i\in s}\left(w_{Ai}-1\right)^2\hat{\sigma}^2 v_i - \sum_{i\in s}\left(w_{Bi}-1\right)^2\hat{\sigma}^2 v_i. \quad (8)$$

3.3 An Important Special Case

In this Subsection, we make the assumptions stated in Section 3.2, and further assume that $v_i = 1$ for all i. We also assume that the dimension of $\boldsymbol{x_{2i}}$ is 1, i.e. that we are considering whether or not one particular auxiliary variable from $\boldsymbol{x_i}$ is to be used in prediction. Expressions (6) and (8) simplify in this case.

Let u_i be the residual of a regression of X_{2i} on $\boldsymbol{X_{1i}}$:

$$u_i = x_{2i} - C^T\boldsymbol{x_{1i}}$$
$$C = \left(\sum_{i\in s}\boldsymbol{x_{1i}}\boldsymbol{x_{1i}^T}\right)^{-1}\sum_{i\in s}\boldsymbol{x_{1i}}x_{2i}.$$

Using straightforward linear algebra operations, it can be shown that

$$\boldsymbol{\beta^T d_B} = \boldsymbol{\beta_2}\sum_{i\in r}u_i$$

and that

$$\sum_{i\in s}\left(w_{Ai}-1\right)\sigma_i^2 - \sum_{i\in s}\left(w_{Bi}-1\right) = \left(\sum_{i\in r}\boldsymbol{u_i}\right)^T S_u^{-1}\left(\sum_{i\in r}\boldsymbol{u_i}\right)$$

where $S_u = \sum_{i\in s} u_i^2$.

Hence (6) becomes

$$\Delta = \left(\sum_{i\in r} u_i\right)^2 S_u^{-1} - \left(\sum_{i\in r} u_i\right)^2 \beta_2^2.$$

CSW show that $d_B^T v\hat{a}r\left[\hat{\beta}_2\right] d_B = \left(\sum_{i\in r} u_i\right)^2 S_u^{-1}$. Hence (8) becomes

$$\hat{\Delta} = \left(\sum_{i\in r} u_i\right)^2 \left(2S_u^{-1} - \hat{\beta}_2^2\right).$$

It is proposed that $\hat{T}_A$ be adopted when $\hat{\Delta} < 0$, and $\hat{T}_B$ be used otherwise. It follows that we adopt $\hat{T}_A$ whenever $\hat{\beta}^2 > 2S_u^{-1}$. As noted by CSW, this is equivalent to adopting $\hat{T}_A$ whenever $F = \hat{\beta}^2/S_u^{-1}$ is greater than 2. Notice that F is the usual F-statistic for testing the null hypothesis that $\beta_2 = 0$. For large n, the cutoff for the F-test at the 5% significance level is 3.96, whereas we have arrived at a cutoff of 2 for adopting the larger set of variables. Thus, the decision to use A instead of B on the basis of a test of significance requires more evidence against B than a simple comparison of the value of the estimated MSEPs of $\hat{T}_A$ and $\hat{T}_B$ would suggest. Therefore, using $\hat{\Delta}$ leads to larger models compared to using significance testing.

3.4 Heteroskedasticity-Robust Estimation of Δ

The estimators of Δ in Sections 3.2 and 3.3 relied on knowing $var\left[Y_i\right]$ at least up to a constant of proportionality. In practice, variances are at best known approximately, and methods which do not rely on an assumption of known variance may perform better. We will use an estimator of σ_i^2 which, assuming model (4), is approximately unbiased for σ_i^2 in general, and exactly unbiased if $\sigma_i^2 = \sigma^2$:

$$\hat{\sigma}_i^2 = \frac{n}{n-p}\left(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x_i}\right)^2.$$

(An alternative estimator would be $\hat{\sigma}_i^2 = \left(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x_i}\right)^2$, as in Royall & Cumberland, 1981b.)

The estimator of $\boldsymbol{\beta}$ would still be as in Section 4.2. The variance of $\hat{\boldsymbol{\beta}}$ is

$$\begin{aligned} var\left[\hat{\boldsymbol{\beta}}\right] &= var\left[S_x^{-1} S_{xy}\right] \\ &= var\left[S_x^{-1} \sum_{i\in s} \boldsymbol{x_i} Y_i\right] \\ &= S_x^{-1}\left(\sum_{i\in s} \boldsymbol{x_i}\boldsymbol{x_i^T}\sigma_i^2\right) S_x^{-1} \end{aligned}$$

This can be estimated by

$$v\hat{a}r_{robust}\left[\hat{\boldsymbol{\beta}}_{\mathbf{1}}\right] = S_x^{-1}\left(\sum_{i\in s} \boldsymbol{x_i}\boldsymbol{x_i^T}\hat{\sigma}_i^2\right) S_x^{-1}$$

Hence we can estimate Δ by

$$\begin{aligned} \hat{\Delta} &= \boldsymbol{d_A^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - v\hat{a}r_{robust}\left[\hat{\boldsymbol{\beta}}\right]\right)\boldsymbol{d_A} - \boldsymbol{d_B^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - v\hat{a}r_{robust}\left[\hat{\boldsymbol{\beta}}\right]\right)\boldsymbol{d_B} \\ &\quad + \sum_{i\in s}(w_{Ai}-1)^2\hat{\sigma}_i^2 - \sum_{i\in s}(w_{Bi}-1)^2\hat{\sigma}_i^2. \end{aligned} \tag{9}$$

3.5 Estimation of Δ in Multi-Stage Sampling

The estimators of Δ in Sections 3.2, 3.3 and 3.4 all assumed that the values of Y are independent for different units. In multi-stage sampling, a sample of primary sampling units (PSUs) is initially selected. A sample of units within the selected PSUs is then selected. For example, PSUs may be areas and units may be households or people; or PSUs could be schools and units could be students. Typically units from the same PSUs tend to be similar, so that values of Y_i and Y_j may be correlated if i and j belong to the same PSU. This subsection develops an estimator of Δ which is approximately unbiased even when there are correlations between values of Y within the same PSU.

Let s_1 be the sample of PSUs, selected from the population U_1. Let s_g be the sample of units from PSU g, where $g \in s_1$. Let $r_1 = U_1 - s_1$ and $r_g = U_g - s_g$. We assume model (4), and further assume that Y_i and Y_j are uncorrelated for $i \in g_1$ and $j \in g_2$ if $g_1 \neq g_2$. The values Y_i and Y_j may be correlated if $i \neq j$ with $i, j \in U_g$.

Let $\hat{T} = \sum_{i \in s} w_i Y_i$ be any predictor and let $\boldsymbol{d} = \sum_{i \in s} w_i \boldsymbol{x_i} - \boldsymbol{T_x}$. The bias of $\hat{T}$ is $\boldsymbol{\beta^T d}$, as in Section 3.2. The variance of $\left(\hat{T} - T_Y\right)$ is

$$
\begin{aligned}
var\left[\hat{T} - T_Y\right] &= var\left[\sum_{i \in s} (w_i - 1) Y_i - \sum_{i \in r} Y_i\right] \\
&= var\left[\sum_{g \in s_1} \left(\sum_{i \in s_g} (w_i - 1) Y_i - \sum_{i \in r_g} Y_i\right) - \sum_{g \in r_1} \sum_{i \in U_g} Y_i\right] \\
&= \sum_{g \in s_1} var\left[\sum_{i \in s_g} (w_i - 1) Y_i - \sum_{i \in r_g} Y_i\right] + \sum_{g \in r_1} var\left[\sum_{i \in U_g} Y_i\right]
\end{aligned}
$$

It is further assumed that the variance of $\sum_{i \in r_g} Y_i$ is negligible relative to other terms. This is the case if cluster sampling is used (because in this case $s_g = U_g$ and r_g is empty) or if the sampling fraction within PSUs is small. The variance becomes

$$
var\left[\hat{T} - T_Y\right] \approx \sum_{g \in s_1} var\left[\sum_{i \in s_g} (w_i - 1) Y_i\right] + \sum_{g \in r_1} var\left[\sum_{i \in U_g} Y_i\right].
$$

Applying this to Δ, we get:

$$\begin{aligned}\Delta &= MSEP\left[\hat{T}_A\right] - MSEP\left[\hat{T}_B\right] \\ &= \boldsymbol{d_A^T}\left(\boldsymbol{\beta\beta^T}\right)\boldsymbol{d_A} - \boldsymbol{d_B^T}\left(\boldsymbol{\beta\beta^T}\right)\boldsymbol{d_B} + \sum_{g\in s_1} var\left[\sum_{i\in s_g}(w_{Ai}-1)\,Y_i\right] \\ &\quad - \sum_{g\in s_1} var\left[\sum_{i\in s_g}(w_{Bi}-1)\,Y_i\right] \end{aligned} \tag{10}$$

To estimate Δ, we need estimators of the variances of $\hat{\boldsymbol{\beta}}$ and $\left(\sum_{i\in s_g}(w_{Bi}-1)\,Y_i\right)$.

Firstly, notice that

$$var\left[\hat{\boldsymbol{\beta}}\right] = var\left[S_x^{-1}\sum_{g\in s_1}\sum_{i\in s_g}\boldsymbol{x_i}Y_i\right] = S_x^{-1}\sum_{g\in s_1} var\left[\sum_{i\in s_g}\boldsymbol{x_i}Y_i\right]S_x^{-1}$$

This can be estimated using the "ultimate cluster variance" method by

$$v\hat{a}r_{ucv}\left[\hat{\boldsymbol{\beta}}\right] = S_x^{-1}\sum_{g\in s_1}\left(\sum_{i\in s_g}\boldsymbol{x_i}\left(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x_i}\right)\right)^2 S_x^{-1}$$

This is a well known estimator of the variance of a weighted sum from clustered data, and is equivalent to Valliant et al. (2000, 9.5.5,p.312). The variance estimator has been called a "sandwich level variance estimator using the cluster-level residuals" (Valliant et al., 2000) and an "ultimate cluster variance" estimator (e.g. Wolter, 1985 describes essentially the same idea in a randomization framework).

The variance of $\left(\sum_{i\in s_g}(w_i-1)\,Y_i\right)$ can also be estimated by the ultimate cluster variance method:

$$v\hat{a}r\left[\sum_{i\in s_g}(w_i-1)\,Y_i\right] = \left\{\sum_{i\in s_g}(w_i-1)\left(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{x_i}\right)\right\}^2$$

Hence we can estimate Δ by

$$\hat{\Delta}_{ucv} = \boldsymbol{d_A^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - v\hat{a}r_{ucv}\left[\hat{\boldsymbol{\beta}}\right]\right)\boldsymbol{d_A} - \boldsymbol{d_B^T}\left(\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}^{\boldsymbol{T}} - v\hat{a}r_{ucv}\left[\hat{\boldsymbol{\beta}}\right]\right)\boldsymbol{d_B}$$

$$+ \sum_{g \in s_1} \left\{ \sum_{i \in s_g} (w_{Ai} - 1) \left(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{x_i} \right) \right\}^2 + \sum_{g \in s_1} \left\{ \sum_{i \in s_g} (w_{Bi} - 1) \left(Y_i - \hat{\boldsymbol{\beta}}^{\boldsymbol{T}} \boldsymbol{x_i} \right) \right\}^2 \tag{11}$$

4. Simulation Study

4.1 Simulation of Farm Survey

Population and Sampling Scheme

The population distribution of the auxiliary variables, the sample and population size, and heteroskedasticity and other properties of the variable of interest would all be expected to play a part in the performance of the adaptive BLUPs. To make a realistic assessment of the performance of these estimators, a simulation study based on a large, realistic population is needed.

We generated a simulation population of 80,000 units, using sample data on 1652 farms from the 1988 Australian Agricultural and Grazing Industry Survey (AAGIS) as a starting point. Total cash crop was used as the survey variable of interest, and potential auxiliary variables included DSE (derived size estimate), number of sheep, crops area, number of beef cattle, region (29 regions) and industry (5 industries). The dataset also contained a sampling weight which was approximately equal to the inverse of the selection probability. 27 outliers with very large values of DSE were removed, as these would normally be placed in a completely enumerated stratum in a survey. A population of 80,000 was then constructed by probability proportional to size sampling with replacement, with probabilities proportional to the estimation weight on the original sample file.

250 samples were then selected without replacement from the simulation population. The samples were stratified by Region and DSE, with DSE divided into four categories, to give 20 strata. The stratum boundaries were set such that the stratum sums of DSE were all equal. Total sample sizes of 500, 1000 and 1500 were simulated. The stratum sample sizes were proportional to the original AAGIS sample sizes by Region and DSE.

Auxiliary Variables and Stepwise Selection Method

Auxiliary variables were included corresponding to the model containing: an intercept; sheep (x1); crops area (x2); beef cattle (x3); Industry; interaction of Industry and x1, x2 and x3; and Region. This gives a total of 52 potential auxiliary variables (some of which are redundant). We also considered the set of 139 auxiliary variables which included this set as well as the interaction of Region and x1, x2 and x3. Models were constructed by stepwise selection starting with the model Intercept + x1 + x2 + x3. Variables were added or removed based on which step most reduced the estimated MSEP, for several alternative estimators of Δ. An adaptive BLUP was also calculated based on statistical significance, with $p < 0.05$ being the cutoff for inclusion.

A number of modifications were needed for the stepwise selection algorithm to work reliably. Firstly, auxiliary variables were not added to the model if they were highly correlated with any variables already in the model (> 0.95). Secondly, variables were not added if this would result in the calibration equations not being solvable. Finally, if the stepwise selection began cycling (for example, adding x1, then adding x2, then removing x1, then removing x2, then adding x1, etc), then the model building process stopped, and the the current model was used as the final model.

Estimators Used

Several BLUPs were calculated: with all auxiliary variables included; with just Intercept+x1+x2+x3; and with auxiliary variables chosen by stepwise selection using the non-robust or robust estimator of Δ, from either the set of 52 or the set of 139 potential auxiliary variables.

Ridge estimators (e.g. Bardsley & Chambers, 1984) are an alternative approach to the problem of variable selection, so we included them in the simulation to compare their performance to that of the adaptive BLUPs. The estimators we have so far considered either include or exclude each variable. If a variable is included, then the weights must calibrate on that variable exactly, in the sense that $\sum_{i \in s} w_i \boldsymbol{x_i} = \boldsymbol{T_x}$. Ridge regression introduces a penalty for non-calibration, but does not necessarily require that the weights provide perfect calibration for all variables. In ridge regression, the vector of sample weights $\boldsymbol{w}$ is chosen to minimise

$$\sum_{i \in s} (w_i - 1)^2 v_i^{-1} + \sum_{j=1}^{p} c_j^{-1} \left(\sum_{i \in s} w_i x_{ij} - T_{xj} \right)^2 .$$

The c_j are non-negative cost coefficients indicating the priority to be placed on meeting calibration constraint j. A value of 0 indicates that the constraint must be met precisely and larger cost coefficients result in placing less weight on the constraint. Thus the ridge estimator allows for a smooth reduction in the effective dimension of the model, by effectively interpolating between including a calibration variable ($c_j = 0$) and excluding it ($c_j = \infty$).

Typically the c_j in ridge regression are set to λc_j^*, where c_j^* reflect a somewhat subjective assessment of the relative importance of each constraint, and λ is chosen to ensure that the final weights w_i have reasonable properties, for example are all greater than or equal to 0, or to 1. We set c_j^* to 0 for the constant (reflecting an intercept in the model), to 1 for $x1$, $x2$ and $x3$, to 10 for the region indicators, to 5 for the industry indicators, and to 100 for interactions. The choice of c_j^* was based on which variables were thought to be likely to be most useful. The value of λ was numerically determined for each sample to be the smallest value such that all weights were greater than or equal to 1.

All of the methods considered used a working variance model with v_i equal to DSE. This is only a rough working model, and the true variance is likely to depend on DSE and other variables in a more complex way.

Results

Table 1 shows the Relative Root Mean Squared Error (RRMSE) of the various calibrated predictors. The first four rows of the table are for the set of auxiliary variables corresponding to the first set of auxiliary variables (52 potential variables). The last four rows of the table are for the second set (139 potential auxiliary variables).

For the smaller set of 52 potential variables, the three adaptive BLUPs had

RRMSE in between that of the two non-adaptive BLUPs ("all" and "first 4"). For the larger set of 139 auxiliary variables, all of the adaptive estimators performed better than the non-adaptive BLUPs. The usual strategy in current use would be to use all auxiliary variables, and the adaptive BLUPs gave very useful gains over this approach. For example, for $n = 1500$, the BLUP using all 139 variables had RRMSE of 1.73%, whereas the adaptive BLUP based on nonrobust $\hat{\Delta}$ had RRMSE of 1.52%. This is a 12% improvement in RRMSE, or a 23% improvement in mean squared error, corresponding to an improvement equivalent to increasing sample size by 23%. Given the expense involved in surveying 1500 farms, the practical benefits of this improvement would be substantial.

The adaptive BLUP based on the non-robust $\hat{\Delta}$ and the adaptive BLUP based on significance testing performed similarly. Both had lower RRMSE than the adaptive BLUP based on the robust $\hat{\Delta}$. The Ridge estimator had higher RRMSE than the adaptive BLUPs when there were 52 auxiliary variables. When there were 139 potential variables, the ridge estimator performed worse than the adaptive BLUP based on non-robust $\hat{\Delta}$ but slightly better than the other two adaptive BLUPs.

Insert Tables 1 and 2 about here

Table 2 shows how many auxiliary variables were selected for the two adaptive BLUPs. The robust $\hat{\Delta}$ led to larger sets of auxiliary variables than the non-robust, with about 10 more auxiliary variables selected.

Table 3 shows the confidence interval (CI) non-coverage of the various predictors. 90% CIs were defined as the estimator +/- 1.64 standard errors, where the variance was estimated using a robust variance estimator. The simulation estimates of the non-coverage rates are based on only 250 simulations so the non-coverage percentages would be expected to have simulation standard errors of approximately 1.9%. A larger simulation study could be used to give more precise estimates of coverage, but this was not pursued due to the computationally intensive nature of the stepwise selection process. This simulation study was sufficient to show that: the BLUP using the first 4 variables had non- coverage close to the nominal 10%; the adaptive BLUP based on robust $\hat{\Delta}$ had very high non-coverage; and the other estimators had non-coverage slightly higher than 10%.

Total cash crops is a major variable of interest in the AAGIS survey, but the totals of other variables are also important, including Farm Equity. For practical reasons, a single set of weights is normally used for all variables. Table 4 shows how well the adaptive calibration weights designed for the Total Cash Crops (TCC) variable performed when used to estimate the total of Farm Equity. For the case of 52 potential auxiliary variables, the adaptive BLUP weights chosen based on TCC (using non-robust $\hat{\Delta}$) performed reasonably well, as did the ridge estimator. Substantial improvements could be made, however, by choosing auxiliary variables based on Equity (again using non-robust $\hat{\Delta}$). This shows that the best set of auxiliary variables can differ greatly depending on the variable of interest.

Insert Tables 3 and 4 about here

4.2 Simulation of Labour Force Survey

Population and Sampling Scheme

A simulation population was constructed by selecting a simple random sample without replacement of 30000 people aged 15-64 from the 1% sample file of the 1991 Australian Census of Population and Housing. The variable of interest was Employment (1 for employed people, 0 for others). The simulation population was divided into simulated primary sampling units (PSUs) containing 75 people each, in such a way that the intra-cluster correlation was 0.05. (This is a fairly typical intra-class correlation for the employment variable within primary sampling units in a household survey. See for example Clark & Steel, 2002). The algorithm for defining clusters was to sort the data by a randomly generated $N(0, \gamma^2)$ variable plus the employment variable, then to define clusters as sequential sets of 75 people, where γ was chosen so as to give the desired intra-cluster correlation.

The simulation consisted of 250 repeated two-stage samples. The first stage was a simple random sample without replacement of m PSUs and the second stage was a simple random sample of replacement of 20 people from each selected PSU. The total sample size was set to be $n = 200, 400$ and 1000 people. Most national household surveys have sample sizes much larger than this, but it is common to construct estimation post-strata within states or provinces, and the sample sizes for these areas would often be in the range 200-1000.

The potential auxiliary variables were age by sex, where age was recorded in single years for 16-24 year olds, then in five year age groups 25-29, 30-34, ..., 55-59 year olds, and 60+ year olds.

Non-Response

One of the main reasons why age and sex are used as auxiliary variables in household surveys is that non-response is known to depend age and sex. For example, young men are typically the group with the lowest response rates. Non-response was simulated by assuming that the logit of the probability of response was equal to $1.8 - \left(\frac{age-50}{25}\right)^2$ for men, and $2 - 0.7\left(\frac{age-50}{25}\right)^2$ for women. This model gave a response rate of 75%. The initial sample size was increased so that the final responding sample size was equal to $n = 80, 160, 400$ or 1000.

Auxiliary Variables and Stepwise Selection Method

The potential auxiliary variables were based on age by sex cells. The definition of the x-variables is shown in Table 5. This parameterization was chosen so that the auxiliary variables corresponding to specific ages or agegroups can be dropped while still giving a sensible model. For example, if all auxiliary variables were included except for x_{4i}, then the model expected value for people aged 17 would be the same as those aged 16, rather than being equal to the intercept parameter. Even better results might be obtained from using more sophisticated parameterizations such as spline models and this will be investigated in a future study.

Insert Table 5 about here

The stepwise selection algorithm was the same as described in Section 4.1.

The estimators of Δ used were the non-robust estimator, the robust (to heteroskedasticity) estimator and the ultimate cluster variance (UCV) estimator which is robust to heteroskedasticity and correlations within PSUs. Significance tests were not used as they would need to incorporate correlations within PSUs to be realistic. Results for the ridge estimator are not shown because negative weights rarely occurred in this simulation, so that this estimator performed very similarly to the BLUP using all auxiliary variables.

Results

Table 6 shows the RRMSE of the various adaptive and non-adaptive BLUPs. There was relatively little difference in RRMSE between the BLUP with intercept only and the BLUP with all auxiliary variables. It is therefore not surprising that at best minor gains were made by using the adaptive BLUPs rather than using the BLUP with all variables. The adaptive BLUP using the non-robust $\hat{\Delta}$ gave the lowest RRMSE in all cases.

Table 7 shows the mean number of variables selected for each of the adaptive BLUPs. Of the 36 potential auxiliary variables, between about 5 and 9 variables were selected based on the non-robust $\hat{\Delta}$. The number of variables selected increased as the sample size increased. The non-robust criterion resulted in larger sets of auxiliary variables, and the UCV criterion gave even larger sets.

Insert Tables 6 and 7 about here

Table 8 shows the confidence interval (CI) non-coverage of the various predictors. 90% CIs were defined as the estimator +/- 1.64 standard errors, where the variance was estimated using a UCV variance estimator. Table 8 shows that the BLUP using all auxiliary variables had high non-coverage for $n = 200$ and 400. The adaptive BLUP using nonrobust $\hat{\Delta}$ had reasonably close to nominal coverage, while the other adaptive BLUPs had high non-coverage.

Table 9 shows how well the various weights performed when used to estimate a different variable, unemployment (equal to 1 for unemployed people and 0 otherwise). Adaptive BLUPs were calculated using the non-robust $\hat{\Delta}$, with the variable of interest given by Employment, and by Unemployment. The adaptive BLUP with variables chosen for unemployment had RRMSE between the non-adaptive BLUP with all variables and the non-adaptive BLUP with intercept only. This suggests that this adaptive BLUP gives reasonable results even when applied to variables other than employment. The adaptive BLUP based on Unemploment actually had higher RRMSE. This may be because the auxiliary variables had little or no predictive power for unemployment, so that attempting to tailor the choice of auxiliary variables for this variable of interest did not work well.

Insert Tables 8 and 9 about here

5. Discussion

Adaptive BLUPs generally showed themselves to be superior to the non-adaptive BLUPs in the two simulation studies described here. The adaptive BLUPs based on a non-robust estimator of Δ and based on significance testing were both clearly better than the adaptive BLUP based on a robust criterion, in terms of mean squared error and confidence interval coverage.

In both the farm survey and the labour force survey simulations, the adaptive BLUPs based on a nonrobust estimator of Δ and based on significance testing both had lower MSEP than non-adaptive estimators in almost all cases. In the case of the farm survey, the gains were substantial when there were a large number of potential auxiliary variables. In the case of the labour force survey, the gains were minor. The adaptive BLUPs also gave reasonable confidence interval coverage.

The adaptive BLUPs based on the robust and UCV criteria performed much worse than the other adaptive BLUPs. This is surprising, as the AAGIS data is known to be heteroskedastic and only a crude untested variance model was used in the nonrobust criteria. The Labour Force data was known to be clustered suggesting that the UCV criteria should have given good results. Apparently the use of these more robust criteria resulted in extra variability in the selection of variables, with the result that larger than optimal sets of auxiliary variables were selected.

Ridge estimators also performed reasonably well in terms of RRMSE and confidence interval coverage. They generally gave worse results than the adaptive BLUPs for estimating the total of the variable of interest when the choice of auxiliary variables was based on this variable. However, when the adaptive BLUP weights were applied to different variables, the ridge estimators performed slightly better. An even better approach may be to adaptively choose both which auxiliary variables to include and how to apply ridging, based on some criterion calculated from the sample. This will be the topic of future research.

A related subject for future research will be to apply sparse data approaches, for example the false discovery rate (Benjamini & Hochberg, 1995; Abramovich et al., 2006), to try to choose the best set of auxiliary variables, and possibly the best choice of ridging parameters. This may give superior performance, particularly when there are large numbers of potential auxiliary variables.

One barrier to wide adoption of the prediction approach to finite population sampling is concern that it may be less robust than the randomization framework. In particular, Hansen et al. (1983) argued that the prediction framework implies inappropriately omitting design variables from calibration in some cases, thereby leading to bias. The adaptive calibrated predictors developed here address this concern. In our simulations from farm economic data and social data, the adaptive predictors had low bias and lower mean-squared error than the non-adaptive estimators in most of the wide range of cases in our simulation study, and were never substantially worse. Provided that all design variables are considered as potential auxiliary variables, adaptive calibration provides a robust and efficient strategy for finite population prediction.

Acknowledgements: This work was jointly supported by the Australian Research Council and the Australian Bureau of Statistics.

References

Abramovich, F., Benjamini, Y., Donoho, D. L., & Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Annals of Statistics*, *34*(2), 584–653.

Bardsley, P., & Chambers, R. (1984). Multipurpose estimation from unbalanced samples. *Applied Statistics*, *33*(3), 290–299.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, *57*(1), 289–300.

Brewer, K. R. W. (1963). Ratio estimation and finite populations: some results deductible from the assumption of an underlying stochastic process. *Australian Journal of Statistics*, *5*, 93–105.

Chambers, R., Skinner, C., & Wang, S. (1999). Intelligent calibration. *Bulletin of the International Statistical Institute*, *58*(2), 321–324.

Clark, R. G., & Steel, D. G. (2002). The effect of using household as a sampling unit. *International Statistical Review*, *70*(2), 289–314.

Deville, J.-C., & Sarndal, C. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, *87*(418), 376–382.

Hansen, M., Madow, W., & Tepping, B. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys. *Journal of the American Statistical Association*, *78*, 776–793.

Royall, R. M. (1970). On finite population sampling theory under certain linear regression models. *Biometrika*, *57*(2), 377–387.

Royall, R. M. (1976). The linear least squares prediction approach to two-stage sampling. *Journal of the American Statistical Association*, *71*(355), 657–664.

Royall, R. M., & Cumberland, W. G. (1981a). An empirical study of the ratio estimator and estimators of its variance. *Journal of the American Statistical Association*, *76*(373), 66–80.

Royall, R. M., & Cumberland, W. G. (1981b). The finite-population linear regression estimator and estimators of its variance - an empirical study. *Journal of the American Statistical Association*, *76*(376), 924–930.

Royall, R. M., & Herson, J. (1973a). Robust estimation in finite populations 1. *Journal of the American Statistical Association*, *68*(344), 880–889.

Royall, R. M., & Herson, J. (1973b). Robust estimation in finite populations 2: Stratification on a size variable. *Journal of the American Statistical Association*, *68*(344), 890–893.

Sarndal, C., Swensson, B., & Wretman, J. (1992). *Model assisted survey sampling.* New York: Springer-Verlag.

Silva, P. L. N., & Skinner, C. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, *23*, 23–32.

Skinner, C., & Silva, P. L. N. (1997). Variable selection for regression estimation in the presence of nonresponse. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 76–81.

Smith, T. M. F. (1976). The foundations of survey sampling: a review. *Journal of the Royal Statistical Society Series A*, *139*(2), 183–202.

Valliant, R., Dorfman, A. H., & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach.* New York: Wiley.

Wolter, K. M. (1985). *Introduction to variance estimation.* New York: Springer-Verlag.

Table 1: RRMSE (%) of AAGIS Predictors of Total Cash Crops

# Vars	n	BLUP		Adaptive BLUP			Ridge
		all	first4	nonrobust $\hat{\Delta}$	robust $\hat{\Delta}$	Sig.Test	
52	500	2.34	2.85	2.42	2.62	2.39	2.67
	1000	1.65	2.05	1.71	1.86	1.71	2.12
	1500	1.46	1.83	1.51	1.61	1.48	1.88
139	500	3.49	2.85	2.39	2.85	2.39	2.42
	1000	1.89	2.05	1.74	1.85	1.78	1.82
	1500	1.73	1.83	1.52	1.49	1.51	1.59

Table 2: Mean (Interquartile Range) of Number of Auxiliary Variables Selected in AAGIS

# Vars	n	nonrobust $\hat{\Delta}$	robust $\hat{\Delta}$	Sig.Test
52	500	16.4 (14.0-18.0)	25.7 (23.0-28.0)	11.2 (10.0-12.0)
	1000	20.7 (19.0-23.0)	27.9 (25.0-31.0)	14.3 (13.0-16.0)
	1500	23.8 (22.0-25.0)	30.5 (28.0-33.0)	17.3 (16.0-19.0)
139	500	36.3 (32.0-40.0)	61.7 (57.0-66.0)	22.9 (20.0-26.0)
	1000	44.1 (41.0-47.8)	65.6 (62.0-70.0)	29.8 (27.0-33.0)
	1500	50.6 (47.0-54.0)	68.7 (65.3-72.0)	34.9 (32.0-38.0)

Table 3: Confidence Interval Non-Coverage in AAGIS

# Vars	n	BLUP		Adaptive BLUP			Ridge
		all	first4	nonrobust $\hat{\Delta}$	robust $\hat{\Delta}$	Sig.Test	
52	500	9.2	7.6	13.2	22.0	12.8	19.2
	1000	9.2	6.4	9.2	19.6	12.4	21.6
	1500	13.2	8.4	12.4	16.0	12.4	23.6
139	500	16.8	7.6	13.6	28.0	13.6	14.8
	1000	12.8	6.4	10.8	19.2	14.0	13.6
	1500	17.6	8.4	13.2	18.4	15.2	14.8

Table 4: RRMSE (%) of AAGIS Predictors of Total Equity

# Vars	n	BLUP		Adaptive BLUP(nonrobust $\hat{\Delta}$)		Ridge
		all	first4	based on TCC	based on Equity	
52	500	4.14	6.02	4.78	4.07	4.45
	1000	3.26	5.70	4.08	3.29	3.73
	1500	3.04	5.44	3.79	3.00	3.46
139	500	5.37	6.02	4.18	3.97	4.13
	1000	3.43	5.70	2.99	2.90	3.09
	1500	2.75	5.44	2.51	2.42	2.58

Table 5: Potential Auxiliary Variables in Labour Force Survey Simulation

Variable	Definition
x_{1i}	1 (corresponding to intercept in model for Y)
x_{2i}	1 if person i male -1 if female
x_{3i}	1 if person i aged 16 or over
x_{4i}	1 if person i aged 17 or over
$\vdots$	$\vdots$
$x_{12,i}$	1 if person i aged 25 or over
$x_{13,i}$	1 if person i aged 30 or over
$\vdots$	$\vdots$
$x_{19,i}$	1 if person i aged 60 or over
$x_{20,i}$	x_{3i} if person i male $-x_{3i}$if female
$\vdots$	$\vdots$
$x_{36,i}$	x_{2i} if person i male $-x_{19,i}$if female

Table 6: RRMSE of Labour Force Survey Predictors of Employment

n	BLUP		Adaptive BLUP		
	all	intercept	nonrobust $\hat{\Delta}$	robust $\hat{\Delta}$	UCV $\hat{\Delta}$
200	6.54	6.77	6.44	7.06	6.96
400	4.72	4.76	4.61	4.72	4.65
1000	2.45	2.70	2.43	2.45	2.49

Table 7: Mean (Interquartile Range) of Number of Auxiliary Variables Selected in Labour Force Simulation

n	Variable Selection Method		
	nonrobust	robust	UCV
200	6.5 (5.0- 8.0)	13.4 (10.0-16.0)	16.1 (13.0-19.0)
400	7.4 (6.0- 8.0)	12.1 (9.0-15.0)	14.5 (12.0-17.0)
1000	8.6 (7.0-10.0)	11.6 (10.0-13.0)	14.2 (12.0-17.0)

Table 8: Confidence Interval Non-Coverage (%) for Predictors of Employment

n	BLUP		Adaptive BLUP		
	all	intercept	nonrobust $\hat{\Delta}$	robust $\hat{\Delta}$	UCV $\hat{\Delta}$
200	17.6	12.0	12.0	20.0	24.0
400	17.2	12.0	14.8	16.8	17.6
1000	6.4	11.6	7.6	6.8	9.6

Table 9: RRMSE of Labour Force Survey Predictors of Unemployment

n	BLUP		Adaptive BLUP	
	all	intercept	based on emp	based on unemp
200	36.3	32.6	34.5	36.0
400	24.1	21.7	22.8	23.7
1000	14.5	14.2	14.1	14.2