ESTIMATION OF MEASUREMENT BIAS USING A MODEL PREDICTION APPROACH

Paul Biemer, Research Triangle Institute, and Dale Atkinson, National Agricultural Statistics Service
P.O. Box 12194, Research Triangle Park, NC 27709

KEY WORDS: Reinterview; Repeated measures; response error; bootstrap

1. INTRODUCTION

It is well-known in the survey literature that when responses are obtained from respondents in sample surveys, the actual observed values of measured characteristics may differ markedly from the true values of the characteristics. Evidence of these so-called measurement errors in surveys has been collected in a number of ways. For example, the recorded response may be checked for accuracy against administrative records or legal documents within which the true (or at least a more accurate) value of the characteristic is contained. An alternative means relies on revised reports from respondents via reinterviews. In a reinterview, a respondent is recontacted for the purpose of conducting a second interview regarding the same characteristics measured in the first interview. Rather than simply repeating the original questions in the interview, there may be extensive probes designed to elicit a more accurate response, or the respondent may be instructed to consult written records for the "book values" of the characteristics. For some reinterview surveys, descrepancies between the first and second interviews are reconciled with the respondent until the interviewer is satisfied that a correct answer has been obtained. Forsman and Schreiner (1991) provide an overview of the literature for these types of reinterviews. Other means of checking the accuracy of survey responses include: a) comparing the survey statistics (i.e., means, totals, proportions, etc.) to statistics from external sources that are more accurate; b) using experimental designs to estimate the effects on survey estimates of interviewers and other survey personnel; and c) checking the results within the same survey for internal consistency.

The focus of the current work is on estimators of measurement bias from data collected in true value remeasurement studies, i.e., record check and reinterview studies, where the objective is to obtain the true value of the characteristic at, perhaps, a much greater cost per measurement than the original survey.

Because of the high costs typically involved in conducting reinterview studies, repeated measurements are usually obtained for only a small fraction of the original survey sample. While the sample size may be quite adequate for estimating biases at the national and regional levels, they may not be adequate for estimating the error associated with small subpopulations or rare survey characteristics. In this paper, our objective is to consider estimators of response bias having better mean squared error properties than the traditional estimators. The basic idea behind our approach can be described as follows.

In a typical remeasurement study, a random subsample of the survey respondents are selected and, through some means, the true values of the characteristics of interest are ascertained. Let n_1 denote the number of respondents to the first survey and let n_2 denote the number selected for the subsample or *evaluation sample.* The usual estimator of response bias is the *net difference rate*, computed for the n_2 respondents in the evaluation sample as

$$NDR = \bar{y}_2 - \bar{\mu}_2 \qquad (1.1)$$

where $\bar{y}_2$ is the sample mean of original responses and $\bar{\mu}_2$ is the sample mean of the true measurements. A disadvantage of the *NDR* is that it excludes information on the $n_1 - n_2$ units in the original survey who were not included in the remeasurement study. Further, the estimator does not incorporate information on auxilliary variables, x, which may be combined with the information on y and μ available from the survey to provide a more precise estimator of response bias.

Given that we have a stratified, two phase sample design and resulting data (y,μ,x), our objective is to determine the "best" estimator of measurement bias given these data. Our essential approach is to identify a model for the true value, μ_i, which is a fi.action of the observed values, y_i, $i=1,...,n_1$, and any auxilliary information, x, that may be available for the population. The model is then used to predict μ_i for all units in the population for which μ_i is unknown. These predictions can then be used to obtain estimates of the true population mean, total, or proportion. Thus, estimators of the response bias for these parameters can be derived from the main survey.

Since the approach provides a prediction equation for μ_i which is a function of the observations, estimators of response bias can be computed for areas having small sample sizes. In this case, the prediction equation for μ_i may be augmented by other geographic and respondent variables such as: demographic characteristics, type of unit, unit size, geographic characteristics, and so on.

The basic estimation and evaluation theory for a prediction approach to the estimation of response bias is presented in the following sections. Under stratified random sampling, estimators of means and totals, their variances and their mean squared errors are provided. Results from application to National Agricultural Statistics Service (NASS) data are also presented.

2. METHODOLOGY FOR ESTIMATION AND EVALUATION

2.1 The Measurement Error Model

To fix the ideas, we shall consider the case of simple random sampling without replacement (SRSWOR) from a single population. Generalizations to stratified random sampling are straightforward and will be considered subsequently.

Let $U = \{1,2,...,N\}$ denote the label set for the population and let $S_1 = \{1,2,...,n_1\}$, without loss of generality, denote the label set for the first phase SRSWOR sample of n_1 units from U. For y_i, $i \in S_1$, assume the model

$$y_i = \gamma_0 + \gamma_1 \mu_i + \varepsilon_i \tag{2.1}$$

where μ_i is the true value of the measured characteristic, γ_0 and γ_1 are constants, and ε_i is an independent error term having zero expectation and conditional variance, $\sigma^2_{\varepsilon i}$.

Since the focus of our investigation is on the bias associated with the measurements y_i, consider the expectation of y_i. For a given unit, i,

$$E(y_i|i) = \gamma_0 + \gamma \mu_i \tag{2.2}$$

and, hence, the unconditional expectation is

$$E(y_i) = \gamma_0 + \gamma \bar{M} \tag{2.3}$$

where $\bar{M} = \sum_{i=1}^{N} \mu_i/N$. Thus, the measurement bias is

$$B = E(y_i - \mu_i) = \gamma_0 + (\gamma - 1)\bar{M}. \tag{2.4}$$

The parameter, γ_0, is a constant bias term that does not depend upon the magnitude of $\bar{M}$. Note that γ_0 is consistent with the usual definition of measurement bias obtained from the simple model

$$y_i = \mu_i + \varepsilon_i \tag{2.5}$$

with $\varepsilon_i \sim (\gamma_0, \sigma^2_{\varepsilon i})$. (See, for example, Biemer and Stokes, 1991.)

Consider the estimation of B. Assume that a subsample of size n_2 of the original n_1 sample units is selected and the true value, μ_i, is measured for these n_2 units. The true value may be ascertained either by a reinterview, a record check, interviewer observation, or some other means. Let $S_2 \subseteq S_1$ denote this so-called second phase sample. The usual estimator of the measurement bias is the *NDR* defined in (1.1). If the assumption that "the true value, μ_i, is observed in phase 2, for all $i \varepsilon S_2$" is satisfied, then *NDR* is an unbiased estimator of B. It may further be shown that the variance of *NDR* is

$$E\{\left(1-\frac{n_2}{n_1}\right)\frac{s_\mu^2}{n_2}\left(1-\frac{s_{\mu y}^2}{s_y^2 s_\mu^2}\right)+\left(1-\frac{n_2}{n_1}\right)\frac{s_y^2}{n_2}(1-b)^2\} \tag{2.6}$$

where $s_\mu^2 = \sum_{j \varepsilon S_1} (\mu_j - \bar{\mu}_1)^2/(n_1 - 1)$ with analogous definitions for s_y^2 and $s_{\mu,y}$, and $b = s_{\mu y}/s_y^2$.

The *NDR* may be suboptimal in a number of situations which occur with some frequency. To see this, consider estimators of the form

$$\hat{B}_{ga} = \bar{y}_g - \bar{\mu}_{Ra} \tag{2.7}$$

where $\bar{y}_g = \sum_{j \varepsilon S_g} y_j/n_g$, $g = 1,2$,

$$\bar{\mu}_{Ra} = \bar{\mu}_2 + a(\bar{y}_1 - \bar{y}_2) \tag{2.8}$$

and $\bar{\mu}_2 = \sum_{j \varepsilon S_2} \mu_j/n_2$, for a a constant given the subsample, S_1. It can be shown that the value of a that minimizes $\mathrm{Var}(\hat{B}_{ga})$ is

$$\begin{aligned} a &= b \quad \text{for } g=1, \text{ or} \\ &= b\text{-}1 \quad \text{for } g=2. \end{aligned} \tag{2.9}$$

Thus, for $g = 1$ or 2, the "optimal" choice of $\hat{B}_{ga}$ is

$$\hat{B}_{opt} = \bar{y}_1 - [\bar{\mu}_2 + b\ (\bar{y}_1 - \bar{y}_2)] \tag{2.10}$$

which differs from *NDR* by the term $(b - 1)\ (\bar{y}_1 - \bar{y}_2)$. Since, in general, $\bar{y}_1 \neq \bar{y}_2$, *NDR* is optimal only if $b = 1$. It can be shown that this corresponds to the case where γ_1 in (2.1) is 1.

In this paper we shall explore alternatives to *NDR* which incorporate information on y for units in the set $S_1 \sim S_2$ as well as information on some auxilliary variable,x. Our objective is to consider "no-intercept" linear models initially, i.e., $\gamma_0 = 0$ in (2.1). However, a subsequent paper will examine both "intercept" and "no-intercept" models.

2.2 Model Prediction Approaches To Estimation

Model prediction approaches to the estimation of population parameters in finite population sampling are well-documented in the literature. Cochran (1977) and other authors have demonstrated the model-based foundations of the ubiquitous ratio estimator. There is also a considerable literature on the choice between using weights that are derived from explicit model assumptions in estimation for complex surveys or eliminating the sample weights. Proponents of so-called model-based estimation recommend against the use of weights in parameter estimation (see, for example, Royall and Herson, 1973; and Royall and Cumberland, 1981). They contend that the probabilities of selection in finite population sampling, whether equal or unequal, are irrelevant once the sample is produced. The reliability criteria used by model-based samples are derived from the model distributional assumptions rather than sampling distributions. If an appropriate model is chosen to describe the relationship between the response variable and other measured survey variables, "model-unbiased" estimators of the population parameters may be obtained which have greater reliability than estimators which incorporate weights.

On the other side of the controversy are the design-based samplers. Instead of the model-based assumptions, design-based samplers assume that an estimator from a survey is a single realization from a large population of potential realizations of the estimator, where each potential realization depends upon the selected sample. The distribution of the values of the estimator when all possible samples that may be selected by the sampling scheme are considered is referred to as the *sampling distribution of the estimator*. Criteria for evaluating estimators under the design-based approach then consider the properties of the sampling distributions of the estimators. Under this approach, weighting of the estimators is required to achieve unbiasedness if unequal probability sampling is used.

Although the estimators of B considered here are representative of all three classes of estimators, it is not a major objective of this paper to compare design-based, model-assisted, and model-based estimators. More importantly we first seek to develop a systematic approach for evaluating alternative estimators for a given two-phase sample design. The major problem considered is the following: Given a two-phase sample design and estimators of B denoted by $\hat{B}_1$, $\hat{B}_2$, ..., $\hat{B}_p$, how does an analyst identify which estimator minimizes the mean squared error? A second objective of the article is to specify a number of alternative estimators, and apply a systematic approach for evaluating the estimators. As an illustration, the methodology will be applied to data from the December 1990 Agricultural Survey.

2.3 The Estimators Considered in Our Study

Extending the previously developed notation to stratified, two-phase designs, let N_h denote the size of the *hth* stratum, for $h = 1,...,L$. A two-phase sample is selected in each stratum using simple random sampling at each phase. Let n_{1h} and $n_{2h} \le n_{1h}$ denote the phase 1 and phase 2 sample sizes, respectively, in stratum h. Let S_{1h} and$S_{2h} \subseteq S_{1h}$ denote the label sets for the phase 1 and phase 2 samples, respectively, in stratum h. Assume the following data is either observed or otherwise known:

outcome variables:	y_i	$\forall\ i \in S_{1h}$
true values:	μ_i	$\forall\ i \in S_{2h}$
auxilliary variables:	x_i	$\forall\ i \in S_{1h}$

Further assume that $X_h = \sum_{i \in U_h} x_i$ is known for $h = 1,...,L$ where U_h is the label set for the *hth* stratum.

Weighted Estimators of M and B

The usual estimator of M = $N\bar{\text{M}}$ is the unbiased stratified estimator given by

$$\hat{\text{M}}_{2st} = \sum_h N_h \bar{\mu}_{2h} \qquad (2.11)$$

where $\bar{\mu}_{2h} = \sum_{i \in S_{2h}} \mu_i / n_{2h}$. The corresponding estimator of B is *NDR* defined in (1.1). For stratified samples, it is

$$\hat{B}_{2st} = \hat{Y}_{2st} - \hat{\text{M}}_{2st} \qquad (2.12)$$

where $\hat{Y}_{2st} = \sum_h N_h \bar{y}_{2h}$ and $\bar{y}_{2h} = \sum_{i \in S_{2h}} y_i / n_{2h}$. Note that (2.12) does not incorporate the information on y for units with labels $i \in S_{1h} \sim S_{2h}$. An alternative estimator that uses all the data on y is

$$\hat{B}_{12st} = \hat{Y}_{1st} - \hat{M}_{2st} \qquad (2.13)$$

where $\hat{Y}_{1st} = \sum_h N_h \bar{y}_{1h}$ and $\bar{y}_{1h} = \sum_{i \in S_{1h}} y_i / n_{1h}$.

A number of model-assisted estimators can be specified for two-phase stratified designs. These may take the form of either separate or combined estimators (see, for example, Cochran, 1977, pp.327-330). Further, the ratio adjustments may be applied to either phase 1 or phase 2 stratum-level estimators. Because stratum sample sizes are typically small in two phase samples, only combined estimators shall be considered here.

Consider a special case of the model (2.1) as follows. Letting $\gamma_0 = 0$, we have

$$y_i = \gamma \mu_i + \varepsilon_i \qquad (2.14)$$

where γ is an unknown constant and we assume $\varepsilon_i \sim (0, \sigma_\varepsilon^2 \mu_i)$. The least squares estimator of γ is $\hat{\gamma} = \bar{y}_{2st} / \bar{\mu}_{2st}$. Thus, a model estimator of μ_i is $y_i / \hat{\gamma} = \bar{\mu}_{2st} y_i / \bar{y}_{2st}$ and an estimator of M is

$$\hat{M}_{2stR} = \frac{\hat{M}_{2st}}{\hat{Y}_{2st}} \hat{Y}_{1st}. \qquad (2.15)$$

Using this estimator of M, two estimators of B corresponding to (2.12) and (2.13) are

$$\hat{B}_{2stR} = \hat{Y}_{2st} - \hat{M}_{2stR} \qquad (2.16)$$

and

$$\hat{B}_{12stR} = \hat{Y}_{1st} - \hat{M}_{2stR} \qquad (2.17)$$

A third estimator of B can be obtained via the model

$$y_i = \beta x_i + e_i \qquad (2.18)$$

where β is a constant and $e_i \sim (0, \sigma_e^2 x_i)$. This leads to a ratio estimator of Y,

$$\hat{Y}_{xstR} = \frac{\bar{y}_{1st}}{\bar{x}_{1st}} X \qquad (2.19)$$

Thus, the corresponding estimator of B is

$$\hat{B}_{x2stR} = \hat{Y}_{xstR} - \hat{M}_{2stR} \qquad (2.20)$$

Finally, Särndal, Swensson, and Wretman (1992, p. 360) suggest a general estimator of M in two phase sampling. Adapting their estimator to the stratified random sampling design yields

$$\hat{M}_{SSW} = \hat{M}_{2stR} + \frac{\bar{\mu}_{2st}}{\bar{x}_{2st}} (X - \hat{X}_{1st}) \qquad (2.21)$$

Note that the addition of the unbiased estimator of zero to the ratio estimator $\hat{M}_{2stR}$ in (2.18) results in an estimator which may have smaller variance than $\hat{M}_{2stR}$ if this term is negatively correlated with $\hat{M}_{2stR}$. Likewise, their estimator of Y reduces to $\hat{Y}_{xstR}$ defined in (2.19). Thus the corresponding estimator of B is

$$\hat{B}_{SSW} = \hat{Y}_{xstR} - \hat{M}_{SSW} \qquad (2.22).$$

Note that $\hat{B}_{SSW} = B_{x2stR}$ - (the additive term in (2.21)).

Unweighted Estimators of M and B

Rewrite M as

$$M = \sum_{i \in S_2} \mu_i + \sum_{i \in S_1 - S_2} \mu_i + \sum_{i \in U - S_1} \mu_i \qquad (2.23)$$

$$= M_{(2)} + M_{(1-2)} + M_{(-1)},$$

say, where $S_g = \bigcup_{h=1}^{L} S_{gh}$, g=1,2. The strategy for unweighted, model-based estimation is to replace μ_i in $M_{(1-2)}$ and $M_{(-1)}$ by a prediction, $\hat{\mu}_i$, obtained from a model.

Using the model in (2.14), an estimator of μ_i is

$$\hat{\mu}_i = y_i / \hat{\gamma}$$

where now $\hat{\gamma} = \bar{y}_2 / \bar{\mu}_2$. Thus an estimator of $M_{(1-2)}$ is

$$\hat{M}_{(1)} = n_1 \frac{\bar{\mu}_2}{\bar{y}_2} \bar{y}_1 \qquad (2.24)$$

where $\bar{y}_g = \sum_{i \in S_g} y_i / n_g$, $\bar{\mu}_2 = \sum_{i \in S_2} \mu_i / n_2$, and $n_g = \sum_h n_{gh}$, *for* $g = 1,2$. Further, using the model

$$\mu_i = \delta x_i + \xi_i \qquad (2.25)$$

where δ is a constant and $\xi_i \sim (0, \sigma_\xi^2 x_i)$, we obtain

$$\hat{M}_{(-1)} = \frac{\bar{\mu}_2}{\bar{x}_2} X_{U-S_1} \qquad (2.26)$$

where $X_{U-S_1} = \sum_{i \in U - S_1} X_i$. Thus, a model-based estimator of M is

$$\hat{M}_M = M_{(2)} + \hat{M}_{(1-2)} + \hat{M}_{(-1)} \qquad (2.27)$$

Likewise, Y can be rewritten as

$$Y = \sum_{i \in S_1} y_i + \sum_{i \in U-S_1} y_i$$

$$= Y_{(1-2)} + Y_{(-1)} \qquad (2.28)$$

and we wish to predict y_i in $Y_{(-1)}$. Using the model in (2.18) a model-based estimator of $Y_{(-1)}$ is

$$\hat{Y}_{(-1)} = \frac{\bar{y}_1}{\bar{x}_1} X_{U-S}$$

and, thus, an estimator of Y is

$$\hat{Y}_M = Y_{(1-2)} + \hat{Y}_{(-1)}. \qquad (2.29)$$

Thus, B is estimated as

$$\hat{B}_M = \hat{Y}_M - \hat{M}_M \qquad (2.30)$$

In addition to these estimators, robust versions of $\hat{B}_{2stR}, \hat{B}_{12stR}, \hat{B}_{x2stR}$, and $\hat{B}_M$ were evaluated. These estimators, denoted by $\tilde{B}_{2stR}, \tilde{B}_{12stR}, \tilde{B}_{x2stR}$, and $\tilde{B}_M$, respectively, were obtained by eliminating regression outliers from the model-based or model-assisted estimators. To illustrate, consider the estimator $\hat{M}_{2stR}$ in (2.15). For this estimator, we computed

$$(n_{2h} - 1)s^2_{res,h} = \sum_{\mu_{hi} \neq 0} \frac{(y_{hi} - \hat{\gamma}\mu_{hi})^2}{\mu_{hi}}, \qquad (2.31)$$

the sum of squares of residuals for the model (2.14). Then, only those units $i \in \tilde{S}_{2h} =$ $= \{ i \in S_{2h}: |y_{ih} - \hat{\gamma}\mu_{ih}| \leq 3s_{res,h}\sqrt{\mu_{hi}} \}$ were included in the calculation of the estimator of γ. Denoting this estimator of γ as $\tilde{\gamma}$, the estimator of M is $\tilde{M}_{2stR} = \tilde{\gamma}\hat{Y}_{1st}$ where $\tilde{\gamma} = \tilde{y}_{2st}/\tilde{\mu}_{2st}$ and $\tilde{\mu}_{2st}$ and $\tilde{y}_{2st}$ are the stratified means of μ_i and y_i for $i \varepsilon \tilde{S}_{2h}$. The other robust model prediction estimators are computed analogously.

Many other unweighted, model-based estimators may be explored in the context of our two phase design. For example, an intercept term may be added to models (2.14), (2.18), and (2.26). Further, slope and intercept parameters may be specified separately for each stratum or combination of strata.

2.4 Estimation of Mean Squared Errors Using Bootstrap Estimators

Although it is possible, under the appropriate design-based or model-based assumptions, to derive closed form analytical estimates of the variance of the estimators we are considering in this study, we have elected instead to use a computer-intensive resampling method. First, we seek a method which is easy to apply since there are potentially many estimators which will be considered in our study. Secondly, it is important to evaluate each estimator using the same criteria and a consistent method of variance estimation is essential to achieving this objective. Thus, it is essential that we employ a variance estimation method which can be applied to estimators of any complexity, under assumptions which are consistent and which do not rely upon any model assumptions. It is well-known that model-based variance estimation approaches are quite sensitive to model failure (see, for example, Royall and Herson, 1973; Royall and Cumberland, 1978; and Hansen, Madow, and Tepping, 1983.) Royall and Cumberland (1981) discuss several bias relevant alternatives including the jackknife variance estimator.

Our approach is similar to that of Royall and Cumberland except rather than using a jackknife estimator, we employ a bootstrap estimator of the variance. For independent and identically distributed observations, Efron and Gong (1983) show that the bootstrap and the jackknife variance estimators differ by a factor of $n/(n-1)$ for samples of size n. Thus, the robustness properties Royall and Cumberland demonstrate for the jackknife estimator also hold for the bootstrap estimator.

Other properties of the bootstrap estimator have led us to choose it above other resampling methods. The jackknife and balance repeated replication (BRR) methods are not easily modified for the two-phase sampling design of our study. However, the bootstrap is readily adaptable to two-phase sampling. Further, Rao and Wu (1988) provide evidence from a simulation study that the coverage properties of bootstrap confidence intervals in complex sampling compare favorably to the jackknife and BRR.

Our general approach extends the method developed by Bickel and Freedman (1984) for single phase, stratified sampling, to two-phase stratified sampling. Since the bootstrap procedure is implemented independently for each stratum, we shall, for simplicity, describe the method for the single stratum case.

2.4.1 Estimation of Variance

Let S_1 and S_2 denote the phase 1 and phase 2 samples, respectively, selected from U using SRSWOR. Let S_{1-2} denote the label set, $S_1 - S_2$. Let $\hat{\theta} = \hat{\theta}(S_{1-2}, S_2)$ denote an estimator of θ which may be a function of the observations corresponding to units in both S_2 and S_{1-2}. Define N, n_1, n_2 and n_{1-2} as the sizes of sets U, S_1, S_2 ,and S_{1-2}, respectively. Consider how the bootstrap is applied to obtain estimates of Var $(\hat{\theta})$.

The simplest case is when N/n_1 is an integer, say k. First, we form the psuedo-population label set

$$U_A^* = U_{A(2)}^* \cup U_{A(1-2)}^*$$

where $U_{A(2)}^*$ consists of k copies of the units in S_2 and $U_{A(1-2)}^*$ consists of k copies of the units in S_{1-2}. We then perform the following three steps:

1. Draw a SRSWOR of size n_2 from $U_{A(2)}^*$ and denote this set by S_2^*.

2. Draw a SRSWOR of size n_{1-2} from $U_{A(1-2)}^*$ and denote this set by S_{1-2}^*.

3. Compute $\hat{\theta}_1^* = \hat{\theta}_1(S_{1-2}^*, S_2^*)$ which has the same functional form as $\hat{\theta}(S_{1-2}, S_2)$, but is computed for the $n_1 = n_{1-2} + n_2$ units in $S_1^* = S_{1-2}^* \cup S_2^*$.

Repeat steps 1 to 3 some large number, Q, times to obtain $\theta_1^*, \ldots, \theta_Q^*$. Then, an estimator of Var$(\hat{\theta})$ is

$$\mathrm{var}_{\mathrm{BSS}}(\hat{\theta}) = \sum_{q=1}^{Q} \frac{(\hat{\theta}_q^* - \hat{\theta}_\cdot^*)^2}{Q-1}$$

where $\hat{\theta}_\cdot^* = \sum_{q=1}^{Q} \hat{\theta}_q^* / Q$.

Using the methods of Rao and Wu (1988), it can now be shown that $\mathrm{var}_{\mathrm{BSS}}(\hat{\theta})$ is a consistent estimator of Var$(\hat{\theta})$.

If $N = kn_1 + r$, where $0 < r < n_1$, the procedure is modified as follows using the Beckel and Freedman procedure. First, form the pseudo-population U_A^* as above consisting of kn_1 units. In addition, form the pseudo population $U_B^* = U_{B(1-2)} \cup U_{B(2)}^*$ of size $(k+1)n_1$ where $U_{B(1-2)}^*$ and $U_{B(2)}$ consist of $k+1$ copies of the labels in S_{1-2} and S_2, respectively. Then, for αQ of the bootstrap samples, select $S_1^* = S_{1-2}^* \cup S_2^*$ from U_A^* and for $(1-\alpha)Q$ samples, select S_1^* from the psuedo-population, U_B^* using the three-step procedure described above, where

$$\alpha = (1 - \frac{r}{n_1})\ (1 - \frac{r}{N-1})$$

2.4.2 Estimation of Bias and MSE

The bootstrap procedure can also provide an estimate of estimator bias. The usual bootstrap bias estimator (see Efron and Gong, 1983; Rao and Wu, 1988) is $b(\hat{\theta}) = \hat{\theta}_\cdot^* - \hat{\theta}$ where $\hat{\theta}_\cdot^* = \sum_q \hat{\theta}_q^* / Q$ and $\hat{\theta}$ is the estimate computed from the sample. Note that $\hat{\theta}_q^*$ (q=1, ...Q) and $\hat{\theta}$ have the same functional form and are based upon the same model assumptions. Thus $b(\hat{\theta})$ does not reflect the contribution to bias due to model failure. We propose an alternative estimator of bias which we conjecture is an improvement over $b(\hat{\theta})$.

Recall from (2.4) that $B = E(y_i - \mu_i)$ where E() denotes expectation over both the measurement error and sampling error distributions. Thus, B may be rewritten as $B = \sum_{i=1}^{N} (Y_i - \mu_i)/N$ where $Y_i = E(y_i | i)$. Unfortunately, Y_i and μ_i are unknown for all $i \in U$. Therefore, we shall construct a pseudo population resembling U, denoted by U^*, such that $B^* = E^*(y_i - \mu_i)$ is known, where $E^*()$ is expected value with respect to the measurement error distribution and the sampling distribution associated with U^*.

Let $U^* = \bigcup_{h=1}^{L} U_h^*$ where U_h^* consists of $k_h = N_h / n_{1h}$ copies of the units in S_{1h}. Here we have assumed k_h is an integer, but will subsequently relax the assumption. Further, define y_i^* for $i \in U$ as y_i for the corresponding unit in S_{1h}. Thus, the population total of the y_i^* is $Y^* = \sum_{i \in U^*} y_i^* = \hat{Y}_{1st}$ for $\hat{Y}_{1st}$ defined in (2.13). Analogously, define the true value for unit $i \epsilon U^*$ as $\mu_i^* = \mu_j$ for $i \epsilon U^*$ corresponding to $j \epsilon S_2$. For $j \epsilon S_{1-2}$, μ_j is unknown; however, for our pseudo-population we could generate pseudo-values for the μ_i^* such that $M^* = \sum_{i \epsilon U^*} \mu_i^* = \hat{M}_{2st}$ where $\hat{M}_{2st}$ is defined in (2.13). Thus, for U^*, $B^* = \bar{y}_{1st} - \bar{\mu}_{2st} = \hat{B}_{12st}$ defined in (2.13). As we shall see, it is not necessary to generate the pseudo-values for μ_i^* in order to evaluate the bias in the estimators of B^*.

Note that under stratified sampling, $U^* = U_A^*$ defined in Section 2.4. Further, the bootstrap procedure described in this section is equivalent to repeated sampling from U^* and the alternative estimators $\hat{\theta}_1,\ldots,\hat{\theta}_p$ of B may also be considered estimators of B^*. Since B^* is known, the bias of $\hat{\theta}$ as an estimator of B^* is $\hat{B}^* = \hat{\theta} - B^*$ and the corresponding MSE may be estimated as

$$M\hat{S}E^* = \sum_q (\hat{\theta}_q - B^*)^2/Q$$

$$= var_{BSS}(\hat{\theta}) + (\hat{\theta}_.^* - B^*)^2$$

where $var_{BSS}(\hat{\theta})$, $\hat{\theta}_q$, and $\hat{\theta}_.^*$ are defined in Section 2.4. It can be easily verified that these results still hold when k_h is non-integer.

Thus, the bootstrap procedure provides a method for evaluating the *MSE* of alternative estimators for estimating B^*. Further, the pseudo-population U^* is a reconstruction of U based upon copies of the values for the units in S_1 and S_2. Thus, it is reasonable to use $\hat{B}^*$ and $M\hat{S}E^*$ to evaluate alternative estimators of B.

3.0 APPLICATION TO THE AGRICULTURAL SURVEY

3.1 Description of the Survey

Each year the National Agricultural Statistics Service (NASS) conducts a series of surveys, collectively referred to as the Agricultural Survey (AS) program, to estimate specific agricultural commodities at the state and national levels. Reinterview studies designed to measure response bias in Computer Assisted Telephone Interviewing (CATI) collected data were conducted in Indiana, Iowa, Minnesota, Nebraska, Ohio, and Pennsylvania in December 1988-1990.

The reinterview techniques used by NASS are similar to those of other organizations (i.e., the U.S. Census Bureau). The NASS focus, however, is on response bias rather than response variance or consistency of response. For the reinterview surveys NASS used supervisory or experienced field interviewers for face-to-face reinterviewing of selected items from a subsample of AS respondents. All reinterviews were conducted within 10 days of the AS CATI interview. Any differences between the original AS and reinterview responses were reconciled to determine the "true" value. This use of the reconciled value in bias calculations assumes that it represents a reasonable proxy for the truth. Considerable effort is expended in procedural development, training, and supervision to ensure that this is the case.

The reinterview samples were chosen from CATI respondents to the AS because CATI accounts for a large percentage of the AS data collected, provides considerable control of the reinterview process, and affords flexibility in the computer generation of reconciliation forms. Parent survey (AS) CATI interviews were completed in the state offices of the states in the reinterview study. A separate corps of supervisory and/or experienced field interviewers was used to conduct the followup face-to-face reinterviews.

Interviewers were instructed to complete the reinterview and reconciliation within 10 days of the original CATI interview to minimize recall problems. In general, the questions reinterviewed relate to values of a particular item as of the first of the month. The average time between the original CATI interview and the reinterview ranged from 6.4 days in March 1988 to 5.9 days in December 1989.

Questionnaires used in the reinterview were similar to the AS questionnaires with respect to question wording. However, not all questions asked on the original interview were reasked on the reinterview. The goal of the reinterview was to obtain the best possible information for the subsampled operation; therefore, interviewers were to contact the person most knowledgeable about the operation. Since it was not the purpose of the study to investigate response variance, it was not necessary to recontact the same individual originally interviewed on the AS.

The Sample:

The December AS is a stratified random sample survey based on a multiple frame survey design that uses independent list and area frames. The reinterview subsamples were drawn from the portion of each state's AS list sample that was completed on CATI. Samples eligible for reinterview included completed interviews, out-of-business operations, and interviews with operations that could not (or would not) report for some items but did report for other items. The reinterview response rate was 87%.

Table 1 presents the reinterview sample sizes for the December 1990 Reinterview Survey whose data are analyzed in this report.

Table 1 Sample Sizes by Survey Item

Item	x	y	μ
	U	S_1	S_2
All Wheat Stocks	108,267	8,176	1,157
Corn Planted Acres	225,269	8,211	1,157
Corn Stocks	225,269	7,990	1,115
Cropland Acreage	278,045	8,274	1,141
Grain Storage Capacity	207,460	8,126	1,104
Soybean Planted Acreage	171,761	8,211	1,156
Soybean Stocks	171,761	8,113	1,130
Total Land in Farm	276,450	8,309	1,159
Total Hog/Pig Inventory	248,571	8,247	1,142
Winter Wheat Seedlings	108,267	8,211	1,150

3.2 Comparison of the Estimators of M and B

Using the December 1990 Agricultural Survey, the estimators developed in the previous section were compared. Estimates of standard errors and mean squared errors were computed using the Bickel-Freedman bootstrap procedure described in Section 2.4, with Q = 300 bootstrap samples. Table 3.2 displays the results for six of the estimators: $\hat{B}_{2st}$, the traditional difference estimator; $\hat{B}_{x2stR}$, the weighted ratio estimator; $\tilde{B}_{x2stR}$, the robust (outlier deletion) version of $\hat{B}_{x2stR}$; $\hat{B}_{SSW}$, the Särndal, Swensson, and Wretman estimator; $\hat{B}_M$, the unweighted model-based estimator; and $\tilde{B}_M$, the robust (outlier deletion) version of $\hat{B}_M$.

3.3 Summary of Results

Table 2 presents a summary of the results from our study. The first data column is the known value of $B^* = E\,(y_i^* - \mu_i^*)$, the bias parameter for the pseudo-population, U^*. The other data columns contain the values of the estimators with their standard errors in parentheses, where s.e. $(\hat{\theta}) = \sqrt{var_{BSS}\,(\hat{\theta})}$. The last four rows of the table correspond, respectively, to : a) the number of items (out of 10) for which a 95% confidence interval would contain B^*; b) the average coefficient of variation (C.V.); c) the average square root of$M\hat{S}E^*$ (RMSE); and d) the average absolute relative bias.

A striking feature of these results is the large disparity among the six estimators across all commodities; particularly for All Wheat Stocks. For this commodity, the range of estimates is -94.2 to 103.2.

Also indicated (by the ‡ symbol) in Table 2 is whether a 95% confidence interval, i.e., $[\hat{\theta} - 2\,s.e.(\hat{\theta}),\ \hat{\theta} + 2\,s.e.(\hat{\theta})]$, covers the parameter B^*. The best performer for parameter coverage is $\hat{B}_{SSW}$ which produced confidence intervals that covered B^* for eight out of ten commodities. $\hat{B}_{2st}$ was the next best with six and $\hat{B}_M$ was third with five. The traditional ratio estimator and its robust version were the worst performers with only one commodity having a confidence interval covering B^*.

The mean square error criteron tells a different story. Here, $\tilde{B}_M$ emerged as the estimator having the smallest average root *MSE*. However, $\hat{B}_{SSW}$ and$\hat{B}_{2st}$ are not much greater. Further, $\hat{B}_{SSW}$ was the estimator having the smallest average absolute relative bias. Only two commodities were estimated with significant biases using this estimator. Thus, it appears from these results that $\hat{B}_{SSW}$ is the preferred estimator using overall performance as the evaluation criterion.

4. CONCLUSIONS AND RECOMMENDATIONS

In this article, we proposed a number of weighted and unweighted model-based estimators of measurement bias for stratified random, two-phase

Table 2 Comparison of Estimators With, B^*, the Pseudo-Population Value of the Bias†

Characteristic	B^*	$\hat{B}_{2st}$	$\hat{B}_{x2stR}$	$\tilde{B}_{x2stR}$	$\hat{B}_{SSW}$	$\hat{B}_M$	$\tilde{B}_M$
All Wheat Stocks	42.3	-6.1 (12.3)	103.2 (17.6)	-94.2 (16.5)	-0.9‡ (24.8)	19.2‡ (16.5)	10.6‡ (16.7)
Corn Planted Acreage	-1.8	1.1‡ (1.1)	11.7 (1.3)	10.1 (1.1)	0.3‡ (1.2)	-4.7‡ (1.9)	-5.0 (1.5)
Corn Stocks	-6.4	-5.4‡ (1.5)	2.4 (1.6)	0.2 (1.3)	-6.5‡ (1.6)	-7.9‡ (2.4)	-9.3‡ (2.2)
Cropland Acreage	27.0	-19.6 (8.3)	-15.0 (8.3)	7.0 (3.1)	-19.6 (8.2)	-36.8 (11.0)	-12.8 (4.0)
Grain Storage Capacity	-3.37	1.4‡ (3.7)	32.3 (3.7)	29.5 (2.6)	-0.1‡ (3.9)	-6.9 (3.0)	-6.8 (2.5)
Soybean Planted Acreage	-4.4	.8 (.8)	13.0 (1.0)	9.9 (0.9)	-0.3 (1.0)	-2.9 (1.1)	-2.7 (1.0)
Soybean Stocks	-0.01	2.8‡ (3.1)	21.3 (2.9)	5.0 (2.3)	0.2‡ (3.5)	-11.0 (3.6)	-8.9 (3.4)
Total Land in Farm	-20.0	-24.7‡ (10.4)	-18.8‡ (12.5)	-2.6 (7.6)	-25.7‡ (10.7)	-44.5‡ (13.4)	-21.2 (5.8)
Total Hogs/Pigs Inventory	-0.1	-2.1 (0.9)	3.4 (1.1)	-0.0‡ (1.0)	-2.2‡ (1.1)	-2.5‡ (1.3)	-1.6‡ (1.0)
Winter Wheat Seedlings	-0.6	-0.5‡ (0.4)	3.8 (0.6)	1.8 (0.5)	-1.2‡ (0.6)	1.1 (0.4)	1.1 (0.4)
Number of Items where C.I. covers B^*		6	1	1	8	5	3
Average C.V.		1.01	.30	11.1	9.5	.41	.48
Average RMSE		13.2	22.4	25.2	12.9	14.9	10.8
Average |Relbias|		30.8	220.0	53.4	4.9	113.1	91.3

† standard errors in parentheses

‡ 95 percent confidence interval covers the pseudo population parameter

sample designs. The proposed estimators incorporate information on the observations, y_i, from the first phase sample, and an auxilliary variable, x. Our aim is to identify estimators which make optimal use of the data (y, μ, x). For the current study, the models proposed for y_i and μ_i were confined to single variable, no-intercept models.

We further proposed evaluation criteria based upon estimates of bias, variance, and mean squared error which utilized a bootstrap methodology. The method of Bickel and Freedman was extended to two-phase sampling for this purpose. It was shown both analytically and empirically that the usual *NDR* estimator is not optimal under the model prediction approach to estimating measurement bias. Our analyses found that an estimator derived from the work of Särdnal, Swensson, and Wretman was the best overall estimator among the six estimators considered under the proposed bootstrap evaluation criteria.

For future research, we intend to incorporate multivariate intercept models in the estimation of measurement bias. Since the bootstrap evaluation criteria developed in this article is completely general, no changes in the evaluation methodology are required to handle the addition of variables in the estimation models. Further, the model assumptions and the methods for handling outliers will be refined and evaluated in a subsequent paper.

References

Bickel, P. and Freedman, (1984). "Asymptotic Normality and the Bootstrap in Stratified Sampling," *The Annals of Statistics*, Vol. 12, No. 2, 470-482.

Biemer, P. and L. Stokes, (1991). "Approaches to the Modeling of Measurement Errors," in P. Biemer, et. al., (eds.) *Measurement Errors in Surveys*, John Wiley & Sons, Inc., N.Y.

Cochran, W., (1977). *Sampling Techniques*, John Wiley & Sons, Inc., N.Y.

Efron, B. and G. Gong, (1983). "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician*, Vol. 31, No. 1, 36-48.

Forsman, G., and I. Schreiner, (1991). "The Design and Analysis of Reinterview: An Overview," in P. Biemer, et. al. (eds.) *Measurement Errors in Surveys*, John Wiley & Sons, Inc., N.Y.

Hansen, M., W. Madow, and B. Tepping, (1983). "An Evaluation of Model-Dependent and Probability Sampling Inferences in Sample Surveys," *Journal of the American Statistical Association*, Vol. 78, No. 384, 776-793.

Rao, J.N.K., and C. Wu, (1988). "Resampling Inference with Complex Survey Data," *Journal of the American Statistical Association*, Vol. 83, No. 401, 231-241.

Royall, R. and J. Herson, (1973). "Robust Estimation in Finite Populations I," *Journal of the American Statistical Association*, Vol. 68, No. 344, 880-893.

Royall, R. and W. Cumberland, (1978). "Variance Estimation in Finite Population Sampling," *Journal of the American Statistical Association*, Vol. 73, No. 362, 351-361.

Särndal, C., B. Swensson, and J. Wretman, (1991). *Model Assisted Survey Sampling*, Springer-Verlag, N.Y.