A SURVEY ERROR MODEL WHICH INCLUDES EDIT AND IMPUTATION ERROR

Paul P. Biemer, Bureau of the Census

1. INTRODUCTION

In an effort to improve the quality of the 1980 Census data, every mail-returned questionnaire and, in many census districts, every enumerator filled questionnaire undergoes and extensive clerical edit in the district office. The main purpose of this edit is to identify the questionnaire items still requiring responses; such questions have either not been answered or have been answered incompletely or inconsistently. After these unacceptable items have been identified, those questionnaires exceeding a specified tolerance with regard to the number of unresolved questions are sent on to a second phase followup. This second phase followup operation consists of either an office telephone interview, or, if telephone followup is not possible, a personal interview to recover the missing information. The information not obtained during the second phase followup is usually imputed from a similar record during the final computer editing and imputation procedure.

Due to the complexity of the editing instructions and the unavailability of qualified edit clerks (or editors), high error rates are expected for the edit operation. The most common error of edit clerks is failure to mark an item requiring further followup for followup. As a result, the missing information, which might have otherwise been obtained, is ultimately imputed during the final stages of processing and, thus, imputation error is introduced in the final census results.

In order to assess the impact of editing errors on the total mean square error of census statistics, a survey error model is needed which practically describes the complex nature of the errors. The usual models for nonsampling errors, such as those proposed by Hansen, Hurwitz, and Bershad (1961), Fellegi (1964), or Hartley and Biemer (1978), do not sufficiently describe the interrelationships between editor errors, response errors and imputation errors. For example, the interpretation of a small or large correlated component for editors (or editor variance) is not clear in the context of these models; nor is the effect of high editor error rates on total mean square error.

This paper presents a survey error model which describes the interrelationships of editor error, imputation error, response error and sampling error in surveys similar to the census long-form survey. Under this proposed model, the total mean square error of the sample mean is derived and examined. In addition, the expected value of the usual correlated component estimator (or, equivalently, the ANOVA between editor variance estimator) is derived and the usefulness of the model for interpreting the magnitude of this estimate is demonstrated.

The results presented are applicable to most nonresponse imputation procedures including the census "hot-deck" procedure. Although the assumptions for the type of survey and the survey error was specified for census application, the model can be adapted to other types of surveys employing a clerical edit as a quality control. Moreover, this technique for modeling the errors of editors is applicable to other sources of error of similar nature such as interviewer error (see section 5.3).

2. THE MODEL

2.1 Notation and Assumptions

Consider a population of N units and one particular item in the survey. Suppose that the population is divided into two distinct strata. Let Stratum 1, referred to as the "fail-edit stratum," consist of all units in the population for which either

a) no measurements would be obtained if the unit happened to fall into the sample, or
b) measurements would be obtained that would require some type of action by an edit clerk - for example, inconsistent responses.

Let Stratum 2 consist of those units for which measurements would be obtained requiring no action by an edit clerk.

This division of the population into two strata is a generalization of the concept of a nonresponse stratum found in the sample survey literature on nonresponse adjustment (see eg. Cochran (1977), page 360). More sophisticated models for nonresponse have been proposed (see, eg., Platek, R., Singh, M.P., and Tremblay, V., (1977); however, the simplicity of the "stratum" concept is well-suited for the purposes of this investigation.

For 1980 Census application, Stratum 1 and Stratum 2 refer only to nonresponses and responses obtained up to the time of the edit operation. For example, Stratum 1 includes respondents who initially would not respond to the census item, but would respond at the second phase of enumerator followup which follows the clerical editing operation.

Suppose a simple random sample of n units is drawn from the population and let K denote the number of edit clerks that are available for the survey. The sample is split randomly into K mutually exclusive and exhaustive subsamples, each containing m = n/K units (it is assumed for convenience that K divides n evenly). Each subsample is assigned to one of the K edit clerks for editing.

The following notation may now be defined:

N_1 = number of units in Stratum 1 for the population

n_1 = number of units in Stratum 1 for the sample,

m_{1k} = number of units in Stratum 1 in the k-th edit clerk's assignment,

m_{2k} = $m - m_{1k}$, number of units in Stratum 2 in the k-th edit clerks assignment

$\hat{\pi} = \frac{n_1}{n}$, the proportion of sample units belonging to Stratum 1; referred to as the "sample edit failure rate."

The quantities n_1, $\hat{\pi}$ and m_{1k} (k=1,..,K) are random variables which depend upon the particular sample selected. The quantities m_{1k} (k=1,..,K) also depend upon the particular split of the sample into K subsamples.

2.2 Definition of the Model

To fix the ideas, suppose that the prescribed edit action for each unit in Stratum 1 is to mark the item for further enumerator followup.

If the edit clerk correctly marks a unit in Stratum 1 for followup, the recorded value for the unit is still subject to some nonsampling error. For example, a response error may be committed by the enumerator, the respondent, the coder, etc., or the followup enumerator may not be successful in obtaining the required information in which case the value for the unit is imputed and an imputation error is committed.

If the edit clerk fails to mark a unit in Stratum 1 for followup, the recorded value for the unit is again subject to nonsampling error. For example, the values for these units may be imputed during the final stages of processing and an imputation error committed.

Hence, the final recorded value for a unit in Stratum 1 may differ from the true value of the unit by some nonsampling error which depends upon the action of the edit clerk.

Now, consider those units in the population belonging to Stratum 2, i.e., the units requiring no edit action. The edit clerk may incorrectly mark a good response for followup; however, this is not a serious error since, even if the unit were followed-up, the response would probably not change appreciably. It is, therefore, assumed that these units are only subject to response errors committed by the respondent, the enumerator, etc. which do not depend upon the action of the edit clerk.

These considerations can be expressed in terms of the following general model. Let y_{kj} denote the final recorded value for the j-th unit assigned to editor k and let x_{kj} denote the corresponding true value for the unit, usually unknown. Then

$$y_{kj} = x_{kj} + \delta_{kj}\,\varepsilon_{kj}^{(I)} + (1 - \delta_{kj})\,\varepsilon_{kj}^{(F)}$$

if unit (k, j) belongs to Stratum 1

$$y_{kj} = x_{kj} + \varepsilon_{kj}^{(R)} \qquad (2.2.1)$$

if unit (k, j) belongs to Stratum 2

where

δ_{kj} = 0 if edit clerk k takes the prescribed action, 1 if edit clerk k does not take the prescribed action

$\varepsilon_{kj}^{(I)}$ = the deviation of y_{kj} from x_{kj} if editor k fails to take the prescribed action (referred to as "imputation error"),

$\varepsilon_{kj}^{(F)}$ = the deviation from x_{kj} if editor takes the prescribed action (referred to as "followup error"), and

$\varepsilon_{kj}^{(R)}$ = the response error for unit (k, j) in Stratum 2

The error term $\varepsilon_{kj}^{(F)}$ may be a combination of two errors - imputation error, $\varepsilon_{kj}^{(I)}$ and response error, $\varepsilon_{kj}^{(R)}$. For example, assuming that the prescribed action for each unit in Stratum 1 is to mark the item for enumerator followup, two outcomes can occur when an editor clerk takes the prescribed action:

a) the unit is followed-up successfully in which case the old value of the content item is replace by a new value which is subject to response errors, or
b) the unit is not followed-up in which case the value of the content item is imputed.

Hence, $\varepsilon_{kj}^{(F)}$ may be expressed as

$$\varepsilon_{kj}^{(F)} = \lambda_{kj}\,\varepsilon_{kj}^{(I)} + (1 - \lambda_{kj})\,\varepsilon_{kj}^{(R)} \qquad (2.2.2)$$

where

$$\lambda_{kj} = \begin{cases} 0 \text{ if unit (k, j) is followup successfully} \\ 1 \text{ if unit (k, j) is not followed-up successfully.} \end{cases}$$

Assumptions Made

The assumptions made for the model are

1. For a given editor k, δ_{kj} and δ_{ki} are independent for all $i \neq j$.
2. $\Pr(\delta_{kj} = 1 | k) = \phi_k$ (k = 1,.., K) for all j, referred to as the "error rate" of the k-th editor.
3. $\phi_1,..,\phi_K$ is a random sample from an infinite population of editor error rates with mean ϕ and variance σ^2_ϕ.
4. The nonsampling errors $\varepsilon_{kj}^{(I)}$, $\varepsilon_{kj}^{(F)}$ and $\varepsilon_{kj}^{(R)}$ are random variables with conditional expectations, variances, and covariances given the sample edit failure rate, $\hat{\pi}$, as follows:

(i) $E(\varepsilon_{kj}^{(I)} | \hat{\pi}) = B_I$;

$\mathrm{Var}(\varepsilon_{kj}^{(I)} | \hat{\pi}) = \sigma^2_I$,

(ii) $E(\varepsilon_{kj}^{(F)} | \hat{\pi}) = B_F$;

$\mathrm{Var}(\varepsilon_{kj}^{(F)} | \hat{\pi}) = \sigma^2_F$,

(iii) $E(\varepsilon_{kj}^{(R)} | \hat{\pi}) = B_R$;

$Var(\varepsilon_{kj}^{(R)} | \hat{\pi}) = \sigma^2_R$,

(iv) $Cov(\varepsilon_{kj}^{(I)}, \varepsilon_{hi}^{(F)} | \hat{\pi}) =$

$\rho_{IF}\sigma_I\sigma_F$, $(k, j) \neq (h, i)$,

(v) $Cov(\varepsilon_{kj}^{(I)}, \varepsilon_{hi}^{(R)} | \hat{\pi}) =$

$\rho_{IR}\sigma_I\sigma_R$, $(k,j) \neq (h,i)$,

(vi) $Cov(\varepsilon_{kj}^{(F)}, \varepsilon_{hi}^{(R)} | \hat{\pi}) =$

$\rho_{FR}\sigma_F\sigma_R$, $(k,j) \neq (h,i)$,

(vii) $Cov(x_{kj}, \varepsilon_{hi}^{(I)} | \hat{\pi}) = \rho_{xI}\sigma_x\sigma_I$ if $(k,j) \neq (h,i)$, $= \rho^*_{xI}\sigma_x\sigma_I$ if $(k,j) = (h,i)$

(viii) $Cov(x_{kj}, \varepsilon_{hi}^{(F)} | \hat{\pi}) = \rho_{xF}\sigma_x\sigma_F$ if $(k,j) \neq (h,i)$, $= \rho^*_{xF}\sigma_x\sigma_F$ if $(k,j) = (h,i)$,

(ix) $Cov(x_{kj}, \varepsilon_{hi}^{(R)} | \hat{\pi}) = \rho_{xR}\sigma_x\sigma_R$ if $(k,j) \neq (h,i)$ $= \rho^*_{xR}\sigma_x\sigma_R$ if $(k,j) = (h,i)$

The expectations in (4) are taken over: (i) all possible samples of size n having n_1 units in Stratum 1, (ii) all possible splits of the sample into K subsamples of m units and resulting values of m_{1k} (k=1,..,K), (iii) all possible samples $\{\phi_1,..,\phi_K\}$ from the infinite population of editor error rates, (iv) the infinite population of hypothetical edit trials for a given unit, and (v) the infinite population of response errors resulting from enumerators, coders and other field and office personnel whose effects are present in all three error terms $\varepsilon_{kj}^{(I)}$, $\varepsilon_{kj}^{(F)}$ and $\varepsilon_{kj}^{(R)}$.

Finally, in order to simplify the subsequent derivations and the interpretations of the results, it is further assumed that

5. δ_{kj} is independent of the errors $\varepsilon_{hi}^{(R)}$, $\varepsilon_{hi}^{(I)}$ and $\varepsilon_{hi}^{(F)}$ for all (k,j) and (h,i)

However, the last assumption is relaxed in section 3.2 and the consequential changes to the derived formulae are discussed in the main paper.

3. THE MEAN SQUARE ERROR OF THE SAMPLE MEAN

The true mean of the population is

$$\bar{X} = \frac{1}{N}\sum_{t=1}^{N} x_t \qquad (3.1)$$

for which the usual sample estimator is

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{m} y_{kj}, \qquad (3.2)$$

the sample mean.

Denote by $y_{kj}^{(1)}$, the j-th unit in the k-th edit clerk's assignment belonging to Stratum 1 and denote by $y_{kj}^{(2)}$, the j-th unit of the k-th edit clerk's assignment belonging to Stratum 2.

Then (3.2) may be rewritten as

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{m_{1k}} y_{kj}^{(1)} + \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{m_{2k}} y_{kj}^{(2)} \qquad (3.3)$$

where m_{1k} and m_{2k} are as defined in section 2.1.

In the remainder of this section the mean square error (MSE) of (3.3) is derived using the following formula:

$$MSE(\bar{y}) = E(\bar{y} - \bar{X})^2$$
$$= E_{\hat{\pi}}(B_y^2) + Var(\bar{x}) + E_{\hat{\pi}}(Var_\varepsilon) + 2E_{\hat{\pi}}(B_y E(\bar{x} - \bar{X} | \hat{\pi})) \qquad (3.4)$$

where

$B_y = E(\bar{y} - \bar{x} | \hat{\pi})$ and

$Var_\varepsilon = Var(\bar{y} | \hat{\pi}) - Var(\bar{x} | \hat{\pi})$.

The main emphasis of this study is on investigating the structure of B_y, the nonsampling bias term, and Var_ε, the nonsampling variance term. The last term in (3.4) is not examined. However, if it can be assumed that the last term is (3.4) is small relative to the remaining terms (an assumption that seems justified for large n) then the mean square error of y may be written approximately as

$$MSE(\bar{y}) \doteq E_{\hat{\pi}}(B_y^2) + Var(\bar{x}) + E_{\hat{\pi}}(Var_\varepsilon)$$
$$= (Bias)^2 + \text{Sampling Variance} + \text{Nonsampling Variance} \qquad (3.5)$$

3.1 The Bias of the Sample Mean

From (2.2.1) and (3.3) it follows that

$$E(\bar{y} - \bar{x} | \hat{\pi}) = \frac{1}{n} E\left\{\sum_{j}^{K}\sum_{j}^{m_{1k}} (\delta_{kj}\varepsilon_{kj}^{(I)} + (1 - \delta_{kj})\varepsilon_{kj}^{(F)}) | \hat{\pi}\right\}$$
$$+ \frac{1}{n} E\left(\sum_{k=1}^{K}\sum_{j=1}^{m_{2k}} \varepsilon_{kj}^{(R)} | \hat{\pi}\right) \qquad (3.1.1)$$

Letting $\underset{\sim}{m}_1 = (m_{11},\ldots,m_{1K})$ it follows from the relationships in the appendix that the first term on the right above is

$$\frac{1}{n} E_{\underset{\sim}{m}_1}\sum_{k}^{K}\sum_{j}^{m_{1k}} E(\delta_{kj}\varepsilon_{kj}^{(I)} + (1-\delta_{kj})\varepsilon_{kj}^{(F)} | \underset{\sim}{m}_1, \hat{\pi})$$
$$= \frac{1}{n} E_{\underset{\sim}{m}_1}(n_1\phi B_I + n_1(1-\phi)B_F | \hat{\pi}) \qquad (3.1.2)$$

where $E(.|\underset{\sim}{m}_1, \hat{\pi})$ denotes the expectation over all sources of variation named in section (2.2) with $\underset{\sim}{m}_1$ and $\hat{\pi}$ held fixed and $E_{\underset{\sim}{m}_1}$ denotes expectation over all possible values of $\underset{\sim}{m}_1$ with $\hat{\pi}$ held constant. But B_I and B_F do not depend upon the split of the sample into edit assignments and are, therefore, independent of $\underset{\sim}{m}_1$ so that (3.1.2) is

$$\hat{\pi}(\phi B_I + (1-\phi) B_F). \qquad (3.1.3)$$

Similarly, the second term on the right in (3.1.1) is

$$\frac{1}{n} E_{\underset{\sim}{m}_1} E\left(\sum_k^K \sum_j^{m_{2k}} \varepsilon_{kj}^{(R)} \mid \underset{\sim}{m}_1, \hat{\pi}\right) = (1-\hat{\pi})B_F. \qquad (3.1.4)$$

Hence, combining (3.1.3) and (3.1.4)

$$\text{Bias}(\bar{y}) = E_{\hat{\pi}} \{B_y\}$$

where $B_y = \hat{\pi}(\phi B_I + (1-\phi) B_F) + (1-\hat{\pi}) B_R \qquad (3.1.5)$

This result shows that the usual sample estimator of $\bar{X}$ is biased and that this bias is a weighted sum of the imputation bias, the followup bias and the response bias.

An important special case of this result asummes that the followup error, $\varepsilon_{kj}^{(F)}$, has the structure (2.2.2). If it is also assumed that the random variables λ_{kj} are independent of each other and of the errors $\varepsilon_{hi}^{(I)}$ and $\varepsilon_{hi}^{(F)}$ for all (k,j) and (h,i) then the conditional bias, B_F, may be written as

$$B_F = \lambda B_I + (1-\lambda) B_R \qquad (3.1.6)$$

where $\lambda = \Pr(\lambda_{kj} = 1)$ is the probability that an item marked for followup is not successfully followed-up, i.e., the "followup error rate." Now using (3.1.6),

$$B_y = \hat{\pi}(\phi + (1-\phi)\lambda) B_I + [1-\hat{\pi}(\phi + (1-\phi)\lambda)]B_R \qquad (3.1.7)$$

This result shows that for this special case the bias of $\bar{y}$ is due only to response bias, B_R, and imputation bias, B_I. Furthermore, the weight given to each bias depends upon the edit failure rate, $\hat{\pi}$, the average editor error rate, ϕ, and the followup error rate, λ.

3.2 The Nonsampling Variance of the Sample Mean

Now consider the nonsampling variance term in (3.5).

$$\text{Var}(\bar{y} \mid \hat{\pi}) \doteq \text{Var}_{\underset{\sim}{m}_1}(E(\bar{y} \mid \underset{\sim}{m}_1, \hat{\pi}) \mid \hat{\pi}) + E_{\underset{\sim}{m}_1}(\text{Var}(\bar{y} \mid \underset{\sim}{m}_1, \hat{\pi}) \mid \hat{\pi}). \qquad (3.2.1)$$

But, from section 3.1, $E(\bar{y} \mid \underset{\sim}{m}_1, \pi)$ does not depend upon $\underset{\sim}{m}_1$ so the first term on the right in (3.2.1) is zero.

Since simple random sampling is assumed

$$\text{Var}(\bar{x}) = (1 - \frac{n}{N}) S_x^2/n$$

$$S_x^2 = \sum_{t=1}^{N} (x_t - \bar{X})^2 / (N-1). \qquad (3.2.2)$$

It is shown in the main paper that

$$\text{MSE}(\bar{y}) = E_{\hat{\pi}}\left\{\hat{\pi}(\phi B_I + (1-\phi) B_F) + (1-\hat{\pi}) B_R\right\}^2 + (1-\frac{n}{N}) S_x^2/n + E_{\hat{\pi}}\ \hat{\pi}^2 (\frac{1}{n_1} V_1 + \frac{m-1}{m} \frac{1}{K} C_E + \frac{n_1 - 1}{n_1} C_{11}) + (1-\hat{\pi})^2 (\frac{1}{n_2} V_2 + \frac{n_2 - 1}{n_2} C_{22}) + \qquad (3.2.3)$$

$$2\hat{\pi}(1-\hat{\pi}) C_{12}$$

where

$$V_1 = \phi(1-\phi)(B_I - B_F)^2 + (1-\phi)\sigma_F^2 + 2\phi \rho_{xI}^* \sigma_x \sigma_I + 2(1-\phi) \rho_{xF}^* \sigma_x \sigma_F,$$

$$C_E = \sigma_\phi^2 [(B_I - B_F)^2 + \rho_I \sigma_I^2 + \rho_F \sigma_F^2 - 2\rho_{IF} \sigma_I \sigma_F],$$

$$C_{11} = \phi^2 \rho_I \sigma_I^2 + (1-\phi)^2 \rho_F \sigma_F^2 + 2\phi(1-\phi) \rho_{IF} \sigma_I \sigma_F + 2\phi \rho_{xI} \sigma_x \sigma_I + 2(1-\phi) \rho_{xF} \sigma_x \sigma_F,\ V_2 = \sigma_R^2 + 2\rho_{xR}^* \sigma_x \sigma_R,$$

$$C_{22} = \rho_R \sigma_R^2 + 2\rho_{xR} \sigma_x \sigma_R, \text{ and}$$

$$C_{12} = \rho_{xR} \sigma_x \sigma_R + \phi(\rho_{IR} \sigma_I \sigma_R + \rho_{xI} \sigma_x \sigma_I) + (1-\phi)(\rho_{FR} \sigma_F \sigma_R + \rho_{xF} \sigma_x \sigma_F).$$

The term V_1 is analogous to what is usually termed the simple response variance for the fail-edit stratum. Likewise, V_2 is the simple response variance for Stratum 2. The term C_E represents the between editor variability, sometimes referred to as the editor correlated component of response variance. The terms C_{11} and C_{22} represent the correlated measurement variances for Stratum 1 and Stratum 2, respectively. C_{11} reflects the contribution to the total variance due to imputation and followup errors. C_{22} reflects the contribution due to response errors. Finally, C_{12} is the between strata nonsampling covariance.

In section 2.2, it was assumed that the editor indicator variables, δ_{kj}, were independent of the errors $\varepsilon_{hi}^{(R)}$, $\varepsilon_{hi}^{(I)}$ and $\varepsilon_{hi}^{(F)}$ for all (k,j) and (h,i). In some cases, however, this assumption does not hold. For example, with similar record substitution procedures, such as the "hot-deck" method of imputation, the imputation error depends upon the pattern of response for the sample. This pattern of response is partially determined by the actions of the edit clerks, or equivalently, by the δ_{kj}'s. Therefore, in order for the model to be generally applicable, assumption (5) is now relaxed.

Instead of (5) in section 2.2, it is now assumed that

5′. the editor error rates ϕ_k, (k=1,...,K) are independent of the errors $\varepsilon_{kj}^{(R)}$, $\varepsilon_{kj}^{(I)}$ and $\varepsilon_{kj}^{(F)}$ for all (k,j).

It is shown in the appendix that after replacing assumption (5) with (5′), the formula for MSE $(\bar{y})$ in (3.2.3) remains essentially unchanged - only the interpretation of the affected components of bias and variance change.

Define, for all (k,j),

$$\tilde{B}_I = E(\varepsilon_{kj}^{(I)} \mid \delta_{kj} = 1, \hat{\pi}); \quad \tilde{\sigma}_I^2 = \text{Var}(\varepsilon_{kj}^{(I)} \mid \delta_{kj}=1, \hat{\pi})$$

$$\tilde{B}_F = E(\varepsilon_{kj}^{(F)} \mid \delta_{kj} = 0, \hat{\pi}); \quad \tilde{\sigma}_F^2 = \text{Var}(\varepsilon_{kj}^{(F)} \mid \delta_{kj}=0, \hat{\pi})$$

$$\tilde{\rho}_{xI}^* \sigma_x \tilde{\sigma}_I = E(\varepsilon_{kj}^{(I)} x_{kj} \mid \delta_{kj} = 1, \hat{\pi})$$

$$\tilde{\rho}^{*}_{xF}\,\sigma_x\,\tilde{\sigma}_F = E\,(\epsilon^{(F)}_{kj}\,x_{kj}\,|\,\delta_{kj}=0,\ \hat{\pi}) \qquad (3.2.4)$$

and, for all $(k,j) \neq (h,i)$

$$\tilde{\rho}_I\tilde{\sigma}_I^2 = \mathrm{Cov}\,(\epsilon^{(I)}_{kj},\ \epsilon^{(I)}_{hi}\,|\,\delta_{kj}=\delta_{hi}=1,\ \hat{\pi})$$

$$\tilde{\rho}_F\tilde{\sigma}_F^2 = \mathrm{Cov}\,(\epsilon^{(F)}_{kj},\ \epsilon^{(F)}_{hi}\,|\,\delta_{kj}=\delta_{hi}=0,\ \hat{\pi})$$

$$\tilde{\rho}_{IF}\,\tilde{\sigma}_I\tilde{\sigma}_F = \mathrm{Cov}\,(\epsilon^{(I)}_{kj},\ \epsilon^{(F)}_{hi}\,|\,\delta_{kj}=1,\ \delta_{hi}=0,\ \hat{\pi})$$

with analogous definitions for

$\tilde{\rho}_{xI}\sigma_x\tilde{\sigma}_I$, $\tilde{\rho}_{xF}\sigma_x\tilde{\sigma}_F$, $\tilde{\rho}_{IR}\tilde{\sigma}_I\sigma_R$ and $\tilde{\rho}_{FR}\tilde{\sigma}_F\sigma_R$.

The term $\tilde{B}_I$ is the average value of $\epsilon^{(I)}_{kj}$ over those edit trials for unit (k,j) in which the edit clerk does not take the prescribed action. The term $\tilde{\rho}_F\sigma_F^2$ is the covariance between the followup errors for two units over all edit trials in which the edit clerk(s) took the prescribed action. Similar interpretations apply to the remaining terms.

In the appendix, it is shown that if assumption (5) is replaced by (5′), the resulting formula for MSE $(\bar{y})$ is still (3.2.3) except that the corresponding components are replaced by those in (3.2.4). However, the notation of (3.2.3) will continue to be used in the sequel as a matter of convenience.

4. THE EDITOR COVARIANCE COMPONENT

4.1 Interpretation

The component of variance which can be attributed to the correlation between units within an edit clerk's assignment is

$$C_E = \sigma_\phi^2\,[(B_I - B_F)^2 + \rho_I\sigma_I^2 + \rho_F\sigma_F^2 - 2\rho_{IF}\,\sigma_I\,\sigma_F\,]. \qquad (4.1.1)$$

C_E may also be interpreted as the between editor variance component. Note that this component will be zero if $\sigma_\phi^2 = 0$, i.e. if all editors have the same error rate, ϕ_k.

Consider the special case in which $\epsilon^{(F)}_{kj}$ has the structure given by (2.2.2). Assuming a uniform followup error rate λ, defined in (3.1.6), applies to all enumerators, it can be shown that under assumptions similar to those in section 2.2,

$$C_E = \sigma_\phi^2\,(1-\lambda)^2\,[(B_I - B_R)^2 + \rho_I\sigma_I^2 + \rho_R\sigma_R^2 - 2\rho_{IR}\,\sigma_I\sigma_R] \qquad (4.1.2)$$

The term $(B_I - B_R)^2$ is large in many cases since this is the squared difference between imputation bias and response bias. Hence, one can usually assume the term in brackets is large and, therefore, if C_E is small, the product $(1-\lambda)^2\,\sigma_\phi^2$ is small.

Thus, if there is considerable variability between edit clerks in their error rates and if a large proportion of those units marked for followup are successfully followed-up, C_E will be large. Conversely, a small C_E implies that either (a) there is little variability between edit clerks, (b) the second phase followup is not effective or (c) both (a) and (b).

The quantity λ can be easily estimated from the survey data by

$$\hat{\lambda} = 1 - \frac{\#\text{ of units successfully followed-up}}{\#\text{ of units marked for followup}} \qquad (4.1.3)$$

However, in order to estimate σ_ϕ^2, the edit error rates, ϕ_k, must be estimated which essentially requires re-editing a sample of units from each editor's assignment.

4.2 An Estimator of Editor Covariance

The usual analysis of variance estimator of the between editor component is

$$\hat{\sigma}_E^2 = \sum_{k=1}^{K}\frac{(\bar{y}_{k.} - \bar{y})^2}{K-1} - \frac{1}{n}\sum_{k=1}^{K}\sum_{j=1}^{m}\frac{(y_{kj} - \bar{y}_{k.})^2}{m-1} \qquad (4.2.1)$$

where $\bar{y}_{k.} = \frac{1}{m}\sum_{j=1}^{m} y_{kj}$. This estimator is now shown to provide biased estimate of $\hat{\pi}^2\,C_E$. However, since the relative bias of $\hat{\sigma}_E^2$ is of order $\frac{1}{n}$, it can be ignored for sufficiently large samples.

In the main paper, it is shown that the expected value of the between editor variance component, $\hat{\sigma}_E^2$, is

$$E(\hat{\sigma}_E^2\,|\,\hat{\pi}) = \hat{\pi}^2 C_E + \frac{\hat{\pi}(1-\hat{\pi})}{n}\left(\frac{K}{K-1}\right)[C_{11} + C_{22} - 2C_{12} + (\phi B_I + (1-\phi)B_F - B_R)^2]$$

or, for large n,

$$\doteq \hat{\pi}^2\,C_E. \qquad (4.2.2)$$

Combining the results of the previous section with (4.2.2), one notes that a small value for σ_E^2 does not necessarily indicate that there is little variability between editors. In fact, a small σ_E^2 could be the result of (a) a small fail-edit stratum in the population and consequently a small $\hat{\pi}$, (b) a small success rate $(1-\lambda)$ for the post-edit followup enumeration, (c) homogeneity among editors in their failure rates, $(\sigma_\phi^2 = 0)$, or (d) any combination of the above. However, if $\hat{\pi}$ can be determined from the sample and if λ is estimated as in (4.1.3) then, for $\hat{\pi}(1-\hat{\lambda}) \neq 0$, $\hat{\sigma}_E^2/[\hat{\pi}(1-\hat{\lambda})]^2$ is an estimator of

$$\sigma_\phi^2\,[(B_I - B_R)^2 + \rho_I\sigma_I^2 + \rho_R\rho_R^2 - 2\rho_{IR}\sigma_I\sigma_R] \qquad (4.2.3)$$

assuming that (2.2.2) holds. Thus, under the assumptions for this model, a small value for $\hat{\sigma}_E^2/[\hat{\pi}(1-\hat{\lambda})]^2$ could imply homogeneity among editors in their failure rates.

5. SOME APPLICATIONS OF THE MODEL

5.1 The Nature of Edit Error

As part of the 1980 Census Evaluation Program, an experiment was performed during the

census in a sample of centralized district offices to estimate the components of variance due to editors, telephone followup clerks and their interaction, (see Katzoff and Biemer (1980)). Using the method of interpenetrated work assignments, data was collected which will allow the estimation of these components when the census long-form data becomes available.

A model such as the one proposed in this paper will be used to interpret the ANOVA estimates of the target components from that experiment. In this way, survey operations specialists can acquire some insight into the nature of the errors.

For example, suppose that for an item known to have a high error rate from previous experience, an insignificant editor component is observed. Since this indicates homogeneous error rates among editors, some factors which affect all the editors uniformly might be suspected as the dominating cause of editor error, (e.g. editor training). Conversely, a statistically significant editor component would indicate that individual error rates vary among the editors. This could suggest that the high overall editor error rate is due to a nonuniformity in the quality of personnel hired for the office edit operation. This broad classification of editing error could be the first step toward improving the clerical edit in future censuses and surveys.

5.2 Human vs Computer Editing

The proposed model can be a useful device for comparing various alternative methods for editing survey data. As an example, consider comparing a computer automated edit operation against the present human edit operation for future census use.

Note that $\sigma^2_\phi = 0$ for the computer editing method so that $C_E = 0$ in (3.2.3). However, ϕ_A, the error rate for automated editing may be considerably higher than ϕ_H, the average error rate for human editing, for the same fixed cost. But even if $\phi_A > \phi_H$, the mean square error for automated editing may still be less than that for human editing. Some study which, for fixed cost compares (3.2.3) for the two methods or, for fixed values of (3.2.3) compares the relative costs of the two procedure could be helpful in determining the more efficient method of editing.

5.3 A Model for Interviewer Error

Besides biasing the responses of respondents, interviewers may also discourage responses of any type from a respondent, i.e. they may encourage nonresponses. In some cases, the interviewer may accept a refusal or a "don't know" too readily, or fail to followup a "not-at-home". The impact of these interviewer errors can be examined by postulating a model for the population similar to the model of section 2.

The method for modifying (2.2.1) for use in describing the errors of survey interviewers is now briefly mentioned. Consider a population of N units and a simple random sample of n units. It is not necessary to assume two strata for the population. Let the sample be split as before into K subsamples and suppose each subsample is assigned to one of K interviewers for the survey.

Consider the model, similar to (2.2.1),

$$y_{kj} = x_{kj} + \delta_{kj}\,\epsilon_{kj}^{(I)} + (1 - \delta_{kj})\,\epsilon_{kj}^{(R)} \qquad (5.3.1)$$

where y_{kj}, x_{kj}, $\epsilon_{kj}^{(I)}$ and $\epsilon_{kj}^{(R)}$ are defined in analogy to (2.2.1) for the j-th unit in interviewer k's assignment and

$$\delta_{kj} = \begin{cases} 1 \text{ if interviewer k fails to obtain a response for unit (k,j)} \\ 0 \text{ if interview k obtains a response for unit (k,j).} \end{cases} \qquad (5.3.2)$$

The derivation and interpretation of the bias and variance components of $MSE(\bar{y})$ will closely follow that of section 3 and will be presented in a subsequent paper. Now the interpretation of the so-called correlated component for interviewers is somewhat changed from the usual interpretation (see Biemer (1978)). This model allows the study of the impact of the interviewer "failure rate" on total variance.

APPENDIX

The appendix has been deleted due to lack of space. However, the main paper, which includes the appendix, is available upon request from the author.

REFERENCES

Biemer, P. P. (1978). "The estimation of non-sampling variance components in sample surveys", unpublished Ph.D dissertation, Institute of Statistics, Texas A&M University.

Cochran, W. G. (1977). Sampling Techniques, Third Edition, John Wiley & Sons, New York.

Fellegi, I. P. (1964). "Response variance and its estimation," Journal of American Statistical Assn., 59, 1016-1041.

Hansen, M. N., Hurwitz, W. N. and Bershad, M. A. (1961). "Measurement errors in censuses and surveys." Bull. International Stat. Inst., 38, 359-374.

Hartley, H. O. and Biemer, P. P. (1978). "The estimation of non-sampling variance in current surveys." Proceedings of the ASA (San Diego, 1978).

Katzoff, E. B. and Biemer, P. P. (1980). "Estimation of nonsampling error due to selected office operations of the 1980 census." To be presented at the Joint Statistical Meetings of the ASA, Houston, TX.

Lessler, J. T. (1979). "An expanded survey error model." Prepared for the Symposium on Incomplete Data, Washington, D.C.

Platek, R., Singh, M. P. and Tremblay, V. (1977). "Adjustment for nonresponse in surveys." Survey Methodology, Vol. 3, No. 1, pp 1-24.