Rapid #: -21597740

CROSS REF ID: **6300936**

LENDER: **IPL (Purdue University) :: Main Library**

BORROWER: **HUL (Harvard University) :: Widener Library**

TYPE:	Book Chapter
BOOK TITLE:	Sampling statistics
USER BOOK TITLE:	Sampling statistics
CHAPTER TITLE:	Models Used in Conjunction with Sampling
BOOK AUTHOR:	Fuller, Wayne A
EDITION:	
VOLUME:	
PUBLISHER:	Wiley
YEAR:	2009
PAGES:	NA
ISBN:	9780470454602
LCCN:	
OCLC #:	

Processed by RapidX: 11/13/2023 7:00:55 AM

CHAPTER 5

MODELS USED IN CONJUNCTION WITH SAMPLING

5.1 NONRESPONSE

5.1.1 Introduction

Most surveys of human respondents suffer from some degree of nonresponse. One reason for nonresponse is a failure to contact some elements of the sample. In addition, some people may refuse to participate or fail to respond to certain items in the data collection instrument. Nonresponse is also common in other surveys. An instrument used to record physical data may fail or it may be impossible to record certain data. For example, in an aerial survey of land use it may not be possible to photograph certain selected sampling units where the air space is restricted.

Nonresponse is generally placed in two categories: unit nonresponse and item nonresponse. *Unit nonrespondents*, as the name implies, are those sample elements for which none of the questionnaire information is collected. However, often some information is available. For example, the address of the household is generally available in household surveys, and other information,

Sampling Statistics. By Wayne A. Fuller

such as the physical condition of the residence or the number of residents, may be collected.

Item nonresponse occurs when responses for some items are missing from a questionnaire that is generally acceptable. Such nonresponse is common in self-administered surveys where the respondent can skip questions or sections. Some collection procedures are designed recognizing that people may be reluctant to answer certain questions. For example, questions about income may be placed near the end of the interview and interval categories given as possible answers.

We have introduced the topic of this section using examples of data collected from human respondents, where some people neglect, or refuse, to report for some items. Often, as in two-phase sampling, some data are missing on the basis of the design. Such missing data is called *planned*, or *designed*, *missingness*. Also, some data collection may not require active participation from the sample unit, as in photo interpretation of an area segment. Nonetheless, with analogy to human respondents, we call an element with a reported value a *respondent* and call an element with a missing value a *nonrespondent*.

The analysis of data with unplanned nonresponse requires the specification of a model for the nonresponse. Models for nonresponse address two characteristics: the probability of obtaining a response and the distribution of the characteristic. In one model it is assumed that the probability of response can be expressed as a function of auxiliary data. The assumption of a second important model is that the expected value of the unobserved variable is related to observable auxiliary data. In some situations models constructed under the two models lead to the same estimator. Similarly, specifications containing models for both components can be developed.

5.1.2 Response models and weighting

A model specifying the probability of responding is most common for unit nonresponse, with the complexity of the model depending on the data available. In two-phase estimation in which the vector $(\mathbf{x}, \mathbf{y})$ is collected on phase 2 units but only $\mathbf{x}$ is observed on the remainder of the phase 1 sample, the probabilities of observing y given $\mathbf{x}$ are known. If the nonresponse is unplanned, it is common to assume that the probability of response is constant in a subpopulation, often called a *cell*. The response cell might be a geographic area or a subpopulation defined by demographic characteristics.

Under the cell response model, the sample is formally equivalent to a two-phase sample and we use the notation of Section 3.3 in our discussion. Assume that the original sample was selected with selection probabilities π_{1i}, that the population is divided into G mutually exclusive and exhaustive response cells,

and that every element in a cell has the same probability of responding. Then the two-phase estimated mean of the form (3.3.13) is

$$\bar{y}_{2p,reg} = \sum_{g=1}^{G} \bar{x}_{1\pi,g} \bar{y}_{2\pi,g}, \tag{5.1.1}$$

where

$$\bar{x}_{1\pi,g} = \left(\sum_{g=1}^{G} \sum_{j \in A_g} \pi_{1j}^{-1} \right)^{-1} \sum_{j \in A_g} \pi_{1j}^{-1},$$

$$\bar{y}_{2\pi,g} = \left(\sum_{j \in A_{Rg}} \pi_{1j}^{-1} \right)^{-1} \sum_{j \in A_{Rg}} \pi_{1j}^{-1} y_j,$$

A_g is the set of sample indices in cell g, A_{Rg} is the set of indices for the respondents in cell g, and $\bar{x}_{1\pi,g}$ is the estimated fraction of the population in cell g. Under the cell response model, the estimated variance of (5.1.1) can be computed with the two-phase formulas of Section 3.3. Of course, the validity of the variance estimator rests on the validity of the cell response model.

If the fractions of the population in the cells are known, the estimated mean

$$\bar{y}_r = \sum_{g=1}^{G} N^{-1} N_g \bar{y}_{2\pi,g} \tag{5.1.2}$$

can be treated as a poststratified estimator under the cell response model. See Section 2.2.3 for variance formulas.

The cell mean model is a special case of the regression model and (5.1.2) is the corresponding special case of the regression estimator. To consider general regression estimation, let a vector of auxiliary variables, $\mathbf{x}$, be available for both respondents and nonrespondents, and let the population mean of $\mathbf{x}$, denoted by $\bar{\mathbf{x}}_N$, be known. Then a regression estimator using the inverses of the original probabilities as weights is

$$\bar{y}_{reg} = \bar{\mathbf{x}}_N \hat{\boldsymbol{\beta}}, \tag{5.1.3}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_R' \mathbf{D}_{\pi R}^{-1} \mathbf{X}_R \right)^{-1} \mathbf{X}_R \mathbf{D}_{\pi R}^{-1} \mathbf{y}_R,$$

$\mathbf{X}_R$ is the $n_R \times k$ matrix of observations on the respondents, n_R is the total number of respondents, $\mathbf{D}_{\pi R}$ is the diagonal matrix of original selection probabilities for respondents, and $\mathbf{y}_R$ is the vector of observations for respondents.

Assume that there is a vector $\boldsymbol{\alpha}$ such that

$$\mathbf{x}_i\boldsymbol{\alpha} = \pi_{2i|1i}^{-1}, \tag{5.1.4}$$

where $\pi_{2i|1i}$ is the conditional probability that element i responds given that it is selected for the original sample. Note that condition (5.1.4) holds if a vector of indicator variables is used to construct estimator (5.1.2). Given (5.1.4), the regression estimator (5.1.3) is consistent. Furthermore, there is an appropriate regression estimator of variance if the finite population correction can be ignored.

Theorem 5.1.1. Let a sequence of finite populations and samples be such that the variance of the Horvitz–Thompson estimator of a mean for a complete sample has a variance that is $O_p(n^{-1})$, the Horvitz–Thompson estimator of the variance of a mean for a complete sample has a variance that is $O_p(n^{-3})$, and the limiting distribution of the properly standardized Horvitz–Thompson mean of a complete sample is normal. Assume that for a sample with nonresponse, (5.1.4) holds and that

$$K_L < \pi_{2i|1i} < K_U \tag{5.1.5}$$

for positive constants K_L and K_U. Assume that responses are independent, that $\bar{\mathbf{x}}_N$ is known, that there is a $\boldsymbol{\lambda}$ such that $\mathbf{x}_i\boldsymbol{\lambda} = 1$ for all i, and let the regression estimator be defined by (5.1.3). Then

$$\bar{y}_{reg} - \bar{y}_N = N^{-1}\sum_{i\in A_R}\pi_{2i}^{-1}e_i + O_p(n^{-1}), \tag{5.1.6}$$

where A_R is the set of indices of the respondents, $e_i = y_i - \mathbf{x}_i\beta_N, \pi_{2i} = \pi_{1i}\pi_{2i|1i}$, π_{1i} is the probability that element i is included in the original sample, and

$$\boldsymbol{\beta}_N = \left(\sum_{i\in U}\mathbf{x}_i'\pi_{2i|1i}\mathbf{x}_i\right)^{-1}\sum_{i\in U}\mathbf{x}_i'\pi_{2i|1i}y_i. \tag{5.1.7}$$

Furthermore,

$$\hat{V}_{HT}\left\{n^{1/2}\sum_{j\in A_R}\hat{b}_j\right\} = V_\infty\{n^{1/2}(\bar{y}_{reg} - \bar{y}_N) \mid \mathcal{F}_N\} - N^{-2}\sum_{i\in U}(\pi_{1i} - \pi_{2i})\pi_{2i}^{-1}e_i^2 + O_p(n^{-3/2}), \tag{5.1.8}$$

where

$$\hat{b}_j = \bar{\mathbf{x}}_N \left(\sum_{i \in A_R} \mathbf{x}_i' \pi_{1i}^{-1} \mathbf{x}_i \right)^{-1} \mathbf{x}_j' \pi_{1j}^{-1} \hat{e}_j,$$

$$\hat{V}_{HT} \left(\sum_{j \in A_R} \hat{b}_j \right) = \sum_{i \in A_R} \sum_{j \in A_R} \pi_{1ij}^{-1} \left(\pi_{1ij} - \pi_{1i} \pi_{1j} \right) \hat{b}_i \hat{b}_j, \tag{5.1.9}$$

$\hat{e}_j = y_j - \mathbf{x}_j \hat{\boldsymbol{\beta}}$, and $V_\infty \left\{ n^{1/2} \left(\bar{y}_{reg} - \bar{y}_N \right) \mid \mathcal{F}_N \right\}$ is the variance of the limiting distribution of $n^{1/2} \left(\bar{y}_{reg} - \bar{y}_N \right)$ conditional on $\mathcal{F}_N$.

Proof. The conditional expectations of the components of $\hat{\boldsymbol{\beta}}$ of (5.1.3) are

$$E \left\{ \sum_{i \in A_R} \mathbf{x}_i' \pi_{1i}^{-1} \mathbf{x}_i \mid \mathcal{F} \right\} = \sum_{i \in U} \mathbf{x}_i' \pi_{2i|1i} \mathbf{x}_i$$

and

$$E \left\{ \sum_{i \in A_R} \mathbf{x}_i' \pi_{1i}^{-1} y_i \mid \mathcal{F} \right\} = \sum_{i \in U} \mathbf{x}_i' \pi_{2i|1i} y_i.$$

By the assumption that the Horvitz–Thompson estimators of means have errors that are $O_p \left(n^{-1/2} \right)$,

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N = O_p(n^{-1/2}). \tag{5.1.10}$$

By (5.1.4), $\Sigma_{i \in U} e_i = 0$, and by Theorem 2.2.1, $n^{1/2} \left(\bar{y}_{reg} - \bar{y}_N \right)$ has a normal distribution in the limit.

To obtain representation (5.1.6), assume, without loss of generality, that the first element of $\mathbf{x}$ is $\pi_{2i|1i}^{-1}$. Define a transformation of the original $\mathbf{x}$-vector by $\mathbf{z}_i = \mathbf{x}_i \hat{\boldsymbol{\Lambda}}$, where

$$\begin{aligned} z_{1i} &= \pi_{2i|1i}^{-1}, \\ z_{ji} &= x_{ji} + z_{1i} \hat{\lambda}_{1j} \end{aligned}$$

and

$$\hat{\lambda}_{1j} = - \left(\sum_{i \in A_R} z_{1i}^2 \pi_{1i}^{-1} \right)^{-1} \sum_{i \in A_R} z_{1i} x_{ji} \pi_{1i}^{-1}$$

for $j = 2, 3, \ldots, k$. Then

$$\hat{\boldsymbol{\Lambda}}^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N\right) = \begin{pmatrix} \sum_{i \in A_R} z_{1i}^2 \pi_{1i}^{-1} & \mathbf{0} \\ \mathbf{0}' & \sum_{i \in A_R} \mathbf{z}_{2i}' \pi_{1i}^{-1} \mathbf{z}_{2i} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i \in A_R} z_{1i} \pi_{1i}^{-1} e_i \\ \sum_{i \in A_R} \mathbf{z}_{2i}' \pi_{1i}^{-1} e_i \end{pmatrix},$$

where $\bar{y}_{reg} - \bar{y}_N = \bar{\mathbf{x}}_N \hat{\boldsymbol{\Lambda}} \hat{\boldsymbol{\Lambda}}^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_N)$, and $\mathbf{z}_i = (z_{1i}, \mathbf{z}_{2i}) = (z_{1i}, z_{2i}, \ldots, z_{ki})$. Now

$$\begin{aligned} \hat{\lambda}_{1j} &= -\left(\textstyle\sum_{i \in U} \pi_{2i|1i}^{-1}\right)^{-1} \sum_{i \in U} x_{ji} + O_p\left(n^{-1/2}\right) \\ &= -\bar{z}_{1,N}^{-1} \bar{x}_{j,N} + O_p\left(n^{-1/2}\right) \end{aligned}$$

for $j = 2, 3, \ldots, k$, and

$$\bar{\mathbf{x}}_N \hat{\boldsymbol{\Lambda}} = (\bar{z}_{1,N}, \mathbf{0}) + O_p(n^{-1/2}).$$

It follows that

$$\bar{y}_{reg} - \bar{y}_N = N^{-1} \sum_{i \in A_R} \pi_{1i}^{-1} z_{1i} e_i + O_p(n^{-1})$$

because

$$\begin{aligned} \bar{z}_{1,N} \left(\sum_{i \in A_R} \pi_{1i}^{-1} z_{1i}^2 \right)^{-1} &= N^{-1} \bar{z}_{1,N} \bar{z}_{1,HT}^{-1} \\ &= N^{-1}[1 + O_p(n^{-1/2})], \end{aligned}$$

where

$$\bar{z}_{1,HT} = N^{-1} \sum_{i \in A_R} \pi_{2i}^{-1} z_{1i},$$

and result (5.1.6) is proven.

To prove (5.1.8), note that

$$\sum_{i \in A_R} N^{-1} \pi_{1i}^{-1} z_{1i} e_i = \sum_{i \in A_R} w_{2i} e_i,$$

where $w_{2i} = N^{-1} \pi_{2i}^{-1}$, is a design linear estimator and

$$V\left\{ \sum_{i \in A_R} w_{2i} e_i \mid \mathcal{F} \right\} = \sum_{i \in U} \sum_{j \in U} (\pi_{2ij} - \pi_{2i} \pi_{2j}) w_{2i} w_{2j} e_i e_j$$

$$= \sum_{i\in U}\sum_{\substack{j\in U\\ j\neq i}} \pi_{2i|1i}\pi_{2j|1j}\left(\pi_{1ij}-\pi_{1i}\pi_{1j}\right) w_{2i}w_{2j}e_ie_j + \sum_{j\in U}\left(\pi_{2j}-\pi_{2j}^2\right) w_{2j}^2e_j^2,$$

where π_{1ij} is the probability that elements i and j are in the original sample, $\pi_{2i} = \pi_{1i}\pi_{2i|1i}$, and π_{2ij} is the probability that both i and j are in the sample and both respond. We have $\pi_{2ij} = \pi_{1ij}\pi_{2i|1i}\pi_{2j|1j}$ because responses are independent.

The expectation of the Horvitz–Thompson variance estimator for $\Sigma w_{2i}e_i$ constructed with π_{1i} and π_{1ij} is

$$E\left\{\sum_{i\in A_R}\sum_{j\in A_R}\pi_{1ij}^{-1}\left(\pi_{1ij}-\pi_{1i}\pi_{1j}\right)w_{2i}e_iw_{2j}e_j \mid \mathcal{F}\right\}$$

$$= \sum_{i\in U}\left(\pi_{1i}-\pi_{1i}^2\right)\pi_{2i|1i}w_{2i}^2e_i^2 + \sum_{i\in U}\sum_{\substack{j\in U\\ i\neq j}}\pi_{2i|1i}\pi_{2j|1j}\left(\pi_{1ij}-\pi_{1i}\pi_{1j}\right)w_{2i}e_iw_{2j}e_j$$

$$= \sum_{i\in U}\sum_{j\in U}\left(\pi_{2ij}-\pi_{2i}\pi_{2j}\right)w_{2i}e_iw_{2j}e_j + \sum_{i\in U}\pi_{2i}\left(\pi_{2i}-\pi_{1i}\right)w_{2i}^2e_i^2.$$

The variance estimator constructed with $\hat{e}_t$ is asymptotically equivalent to that constructed with e_t. See the proofs of Theorems 2.2.1 and 2.2.2. Therefore, result (5.1.8) is proven. ■

In the variance estimator (5.1.9), $\hat{b}_i$ is of the form $\tilde{w}_i\hat{e}_i$, where $\tilde{w}_i$ is the regression weight. The $\tilde{w}_i$ must be retained in the variance calculations because the error in $\hat{\beta}$ contributes an $O(n^{-1})$ term to the variance.

The second term in (5.1.8) can be written as

$$-N^{-2}\sum_{i\in U}\pi_{2i|1i}^{-1}\left(1-\pi_{2i|1i}\right)e_i^2 \tag{5.1.11}$$

and will be relatively small for small sampling rates. Given (5.1.4), the $\pi_{2i|1i}$ can be estimated by expressing the response indicator as a function of $(\mathbf{x}_i\boldsymbol{\alpha})^{-1}$ and estimating $\boldsymbol{\alpha}$. Then, using the $\hat{e}_i$ of (5.1.9), expression (5.1.11) can be estimated.

5.2 IMPUTATION

5.2.1 Introduction

If a modest number of variables are missing from otherwise complete questionnaires, one method of implementing estimation is to replace individual missing values with "estimates." The objective is to use the replacement values as if they were observed values in a full-sample estimation procedure. The replacement values are called *imputed values*.

A goal of the imputation procedure is to construct imputed values that will yield efficient estimators of parameters for which estimators would be available from the full sample. Second, it should be possible to estimate the variance of the imputed estimators. In a typical survey situation, the survey statistician makes available to analysts a data set with weights and the values of a set of characteristics for the sample elements. The statistician may know some of the estimates that will be constructed from the data set, but seldom will the full set of possible estimates be known. Thus, the objective is to design an imputation procedure such that the imputed data set will be appropriate for both planned and unplanned estimates.

One may ask; "If one must build a model for the imputation, why not simply use the estimator obtained from the model?" The answer is in the many ways in which survey sample data are used. If a single variable is of importance, a model will be developed for that variable and estimates generated directly from the model. If the objective is to create a data set for general use, replacing the missing values with model-imputed values gives such a data set. Of course, the imputed values must be identified, and the model used for imputation must be made available to the end users.

Consider a simple random sample of n elements in which the y value for m elements is not observed and $r = n - m$ are observed. Assume that the fact that an element is not observed is independent of y. Then the r observations are a simple random sample of size r and the natural estimator of the mean of y is

$$\hat{\mu}_y = r^{-1} \sum_{i \in A_R} y_i, \tag{5.2.1}$$

where A_R is the set of indices of units observed and responding. Now assume that we wish to impute values for the missing values so that estimates based on the entire set of n elements will be equal to estimates based only on the responding units. If the only parameter to be estimated is the mean, replacing the missing values with the mean of the responding values will give a mean of the completed sample that is equal to the mean of the responding units.

However, if other characteristics of the distribution are of interest, estimates based on the mean-imputed data set will be seriously biased. For example, the large fraction of imputed values equal to the mean will bias all estimated quantiles. To meet the goal of multiple use, the imputed data set should provide a good estimate of any function of the variables. That is, the imputed data set should give a good estimate of the distribution function.

There are a number of imputation procedures that furnish good estimates of the distribution function. For a simple random sample and random nonresponse, one procedure is to choose randomly one of the respondents for each nonrespondent and use the respondent value for the missing value. Let y_{iI}, $i = r+1, r+2, \ldots, n$, be the m imputed values and let the mean computed with imputed data be

$$\begin{aligned} \bar{y}_I &= n^{-1}\left(\sum_{i=1}^{r} y_i + \sum_{i=r+1} y_{iI}\right) \\ &=: n^{-1}\left(r\bar{y}_r + m\bar{y}_{m,I}\right), \end{aligned} \tag{5.2.2}$$

where

$$\bar{y}_{m,I} = m^{-1}\sum_{i=r+1}^{n} y_{iI}.$$

Because y_{iI} is a random selection from the respondents, the expected value for any percentile is that for the respondents.

Procedures that use values from the sample as imputed values are called *hot deck imputation procedures*. In a situation such as that just described, the element with a missing value is called the *recipient* and the element providing the value for imputation is called the *donor*. The hot deck name was originally used by the U.S. Census Bureau to describe an imputation procedure when computer cards were used in processing data. The donor was an element that was close to the recipient in the deck of cards. An advantage of hot deck procedures is that the imputed values are values that appear in the data set. It is possible for some imputation procedures to generate impossible responses.

The random selection of donors gives an imputed data set with the correct expectation under the model, but the random selection increases the variance of an estimator relative to an estimator constructed directly from the respondents. If response is independent of y, and we use a random replacement selection of donors for a simple random sample, the conditional variance of the mean of the imputed values is

$$V\{\bar{y}_{m,I} - \bar{y}_r \mid \mathbf{y}_r\} = r^{-1}m^{-1}\sum_{i=1}^{r}(y_i - \bar{y}_r)^2, \tag{5.2.3}$$

where $\mathbf{y}_r = (y_1, y_2, \ldots, y_r)$ is the set of respondents and $\bar{y}_r$ is the mean of the respondents. See Section 1.2.5. Then

$$\begin{aligned} V\{\bar{y}_I \mid (\mathcal{F}, m)\} &= V\{E(\bar{y}_I \mid \mathbf{y}_r)\} + E\{V(\bar{y}_I \mid \mathbf{y}_r)\} \\ &= \left(r^{-1} - N^{-1}\right) S_y^2 + n^{-2} m r^{-1} (r-1) S_y^2, \quad (5.2.4) \end{aligned}$$

where the conditioning notation denotes the variance for samples of size n with exactly m missing.

There are a number of ways to select donors to reduce the imputation variance. One possibility is to use a more efficient sampling method, such as nonreplacement sampling, to select the donors. If m is an integer multiple of r, the imputed estimator of the mean based on without-replacement sampling is equal to the mean of the y values for the respondents. If m is not an integer multiple of r, there is an increase in variance relative to the mean of the respondents. See Exercise 1. Also, one can reduce the variance by using stratified or systematic selection of donors.

Another way to reduce the variance due to imputation is to impute more than one value for each respondent. In a procedure proposed by Rubin (1987) and called *multiple imputation*, the imputation operation is repeated a number of times to create multiple sets of imputed data. Also see Little and Rubin (2002) and Schafer (1997). In the next section we consider a procedure called *fractional imputation*, suggested by Kalton and Kish (1984).

5.2.2 Fractional imputation

In fractional imputation, a number, say M, of donors is used for each recipient and each donor is given a fractional weight, where the fractions sum to 1. Consider a simple random sample with random nonresponse and r respondents. Instead of selecting a single donor for each recipient we assign all respondents to each recipient and give a relative weight of r^{-1} to each donor value. The resulting data set has $r + mr$ vectors where there are now r vectors for each of the elements with missing y. For such a data set, the estimate for any function of y is exactly the same as that obtained by tabulating the sample composed of the respondents. Kim and Fuller (2004) call the procedure *fully efficient fractional imputation* because there is no variance due to the selection of imputed values. Fully efficient fractional imputation is not common because of the size of the resulting data set. However, very efficient procedures can be constructed with two to five imputed values per respondent.

Table 5.1 Sample with Missing Data

Observation	Weight	Cell for x	Cell for y	x	y
1	0.10	1	1	1	7
2	0.10	1	1	2	M
3	0.10	1	2	3	M
4	0.10	1	1	M	14
5	0.10	1	2	1	3
6	0.10	2	1	2	15
7	0.10	2	2	3	8
8	0.10	2	1	3	9
9	0.10	2	2	2	2
10	0.10	2	1	M	M

Example 5.2.1. We use a small artificial data set to illustrate the use of fractional imputation for the calculation of fully efficient estimators and for the calculation of estimated variances. Assume that the data in Table 5.1 constitute a simple random sample and ignore any finite population correction. Variable x is a categorical variable with three categories, identified as 1, 2, and 3. The sample is divided into two imputation cells for this variable. In imputation cell 1 the fraction in the three categories is 0.50, 0.25, and 0.25 for categories 1, 2, and 3, respectively. In imputation cell 2 the fractions are 0.00, 0.50, and 0.50 for categories 1, 2, and 3, respectively. For the missing value of x for observation 4, we impute three values, one for each category, and assign weights for the fractions equal to the observed fractions. All other data are the same for each "observation" created. See the three lines for original observation 4 in Table 5.2. The estimated fraction in a category for imputation cell 1 calculated using the imputed data and the fractional weights is the same as the fraction for the respondents.

The fully efficient fractional imputation of y for y-imputation cell 1 would require four imputed values. That would not be a problem for this small data set, but to illustrate the computation of efficient estimators with a sample of donors, we select a sample of three of the four available donors. See the imputed values for observation 3 in Table 5.2.

Several approaches are possible for the situation in which two items are missing, including the definition of a third set of imputation cells for such cases. Because of the small size of our illustration, we impute under the assumption that x and y are independent within cells. Thus, we impute four values for observation 10. For each of the two possible values of x we impute two possible values for y. One of the pair of imputed y values is chosen to be

less than the mean of the responses, and one is chosen to be greater than the mean. See the imputed values for observation 10 in Table 5.2.

Table 5.2 Fractionally Imputed Data Set

Observation	Donor x	Donor y	w^*_{ij0}	Final Weight	Cell for x	Cell for y	x	y
1	0	0	–	0.1000	1	1	1	7
2	0	1	0.3333	0.0289	1	1	2	7
	0	6	0.3333	0.0396	1	1	2	15
	0	8	0.3333	0.0315	1	1	2	9
3	0	5	0.3333	0.0333	1	2	3	3
	0	7	0.3333	0.0333	1	2	3	8
	0	9	0.3333	0.0333	1	2	3	2
4	†	0	0.5000	0.0500	1	1	1	14
	†	0	0.2500	0.0250	1	1	2	14
	†	0	0.2500	0.0250	1	1	3	14
5	0	0	–	0.1000	1	2	1	3
6	0	0	–	0.1000	2	1	2	15
7	0	0	–	0.1000	2	2	3	8
8	0	0	–	0.1000	2	1	3	9
9	0	0	–	0.1000	2	2	2	2
10	†	8	0.2500	0.0225	2	1	2	9
	†	4	0.2500	0.0275	2	1	2	14
	†	1	0.2500	0.0209	2	1	3	7
	†	6	0.2500	0.0291	2	1	3	15

†All relevant values of x are imputed for every missing observation.

To create fully efficient estimates of the mean of y, the cell mean of the imputed data should be the same as the mean of the respondents in the cell. To define such a data set, we use the regression estimator and require the fractional weights to sum to 1 for each observation. For observation 10, we require the two weights for each category to sum to the fraction (0.5) for the category. In using regression to adjust the fractional weights, one can adjust all weights subject to the restriction that the sum of the fractional weights is 1 for each person or one can adjust the weights for each person. Because of the small number of imputed values per person we use the second approach.

Let B_g be the set of indices of elements in cell g that have at least one characteristic imputed, let $\mathbf{z}_{g[i]j}$ be the ith imputed vector of characteristics for which at least one value has been imputed, let w_j be the weight for observation j, and let w^*_{ij0} be an initial fractional weight for the ith imputed

vector for element j, where

$$\sum_{i \in A_{Ij}} w^*_{ij0} = 1$$

and A_{Ij} is the set of indices of imputed values for observation j. The fractional weight for imputed value i of observation j in cell g is

$$w^*_{ij} = w^*_{ij0} + (\bar{\mathbf{z}}_{FE,g} - \bar{\mathbf{z}}_g)\mathbf{S}^{-1}_{\mathbf{zz}g} w^*_{ij0}(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})', \qquad (5.2.5)$$

where

$$\mathbf{S}_{\mathbf{zz}g} = \sum_{j \in B_g} \sum_{i \in A_{Ij}} w^*_{ij0}(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j})'(\mathbf{z}_{g[i]j} - \bar{\mathbf{z}}_{g\cdot j}),$$

$$\bar{\mathbf{z}}_{g\cdot j} = \sum_{i \in A_{Ij}} w^*_{ij0}\mathbf{z}_{g[i]j},$$

$$\bar{\mathbf{z}}_g = \left(\sum_{j \in B_g} w_j\right)^{-1} \sum_{j \in B_g} w_j \bar{\mathbf{z}}_{g\cdot j},$$

$\bar{\mathbf{z}}_{FE,g}$ is the fully efficient weighted mean of the respondents in cell g, and $\bar{\mathbf{z}}_{g\cdot j}$ is the mean of imputed values for observation j. If the rth characteristic is observed, $\bar{\mathbf{z}}_{g\cdot jr}$ is the value observed and $\bar{\mathbf{z}}_{g[i]jr} - \bar{\mathbf{z}}_{g\cdot jr} = 0$.

Because all values of y in cell 2 were used to impute for observation 3, we need only compute weights for cell 1 for y. The mean of imputed values for observation 2 is 10.333 and the two means for observation 10 are 11.500 and 11.000. The mean of the observed values of y for cell 1 is 11.25, the weighted mean of the imputed values is $\bar{z}_1 = 10.7917$, and the weighted sum of squares is 23.1618. The adjusted fractions are 0.2886, 0.3960, and 0.3154 for observation 2 and 0.2247, 0.2753, 0.2095, and 0.2905 for observation 10, in the order that they appear in the table. The weights, called *final weights* in the table, are the products $w^*_{ij}w_j$.

We use jackknife replicates to illustrate variance estimation with fractional imputation. The procedure is analogous to that for two-phase samples. As the first step we create a standard jackknife replicate by deleting an observation. Tables 5.3 and 5.4 give the estimates for the cell means for the observed values for the 10 replicates. For example, when observation one is deleted, the replicate mean of y for y-cell 1 is 12.67, and the fractions for x-cell 1 are $0.3\bar{3}$, $0.3\bar{3}$, and $0.3\bar{3}$ for categories 1, 2, and 3, respectively.

Using the regression procedure, the fractional weights of each replicate are adjusted to give the mean for that replicate. For example, the fractional weights of the imputed y-values for observations 2 and 10 of replicate 1 are

modified so that

$$\frac{\hat{y}_{2I} + y_3 + y_6 + y_8 + \hat{y}_{10}}{5} = 12.67,$$

where $\hat{y}_{jI} = \Sigma_{i \in A} w_{ij}^* y_i$. Table 5.5 contains the replicate weights.

Table 5.3 Jackknife Replicate Cell Means for y-Variable

	Replicate									
Cell	1	2	3	4	5	6	7	8	9	10
1	12.67	11.25	11.25	10.33	11.25	10.00	11.25	12.00	11.25	11.25
2	4.33	4.33	4.33	4.33	5.00	4.33	2.50	4.33	5.50	4.33

Table 5.4 Jackknife Replicate Fractions for x-Categories

	Cat.	Replicate									
Cell	of x	1	2	3	4	5	6	7	8	9	10
1	1	0.33	0.67	0.67	0.50	0.33	0.50	0.50	0.50	0.50	0.50
	2	0.33	0.00	0.33	0.25	0.33	0.25	0.25	0.25	0.25	0.25
	3	0.33	0.33	0.00	0.25	0.33	0.25	0.33	0.25	0.25	0.25
2	1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	2	0.50	0.50	0.50	0.50	0.50	0.33	0.67	0.67	0.33	0.50
	3	0.50	0.50	0.50	0.50	0.50	0.67	0.33	0.33	0.67	0.50

The final data set with the weights of Table 5.2 and the replicate weights of Table 5.5 can be used to compute all estimators and all estimated variances for which the jackknife is appropriate. For example, the estimated cumulative distribution function for y and its variance could be computed.

The jackknife estimated variance for the mean of y is

$$\hat{V}_{JK}(\bar{y}_{FI}) = \sum_{k=1}^{10} 0.9(\bar{y}_{FI}^{(k)} - \bar{y}_{FI})^2 = 3.1095$$

and the two-phase variance estimator is

$$\hat{V} = \frac{1}{n}\sum_{g=1}^{2} \frac{n_g}{n}(\bar{y}_{Rg} - \bar{y}_{FE})^2 + \sum_{g=1}^{2} \left(\frac{n_g}{n}\right)^2 \frac{1}{r_g} s_{Rg}^2 = 3.043,$$

where s^2_{Rg} is the within-cell sample variance for cell g. The two estimates differ by the amount

$$\sum_{g=1}^{2} \left[(r_g - 1)^{-1} r_g (n-1)^{-1} n - 1\right] s^2_{Rg}.$$

See Section 4.4. ■■

Table 5.5 Jackknife Weights† for Fractionally Imputed Data

	Replicate									
Obs	1	2	3	4	5	6	7	8	9	10
1	0	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
2	0.17	0	0.32	0.42	0.32	0.46	0.32	0.24	0.32	0.27
	0.66	0	0.44	0.30	0.44	0.25	0.44	0.55	0.44	0.51
	0.29	0	0.35	0.39	0.35	0.40	0.35	0.32	0.35	0.33
3	0.37	0.37	0	0.37	0.32	0.37	0.50	0.37	0.29	0.37
	0.37	0.37	0	0.37	0.50	0.37	0.01	0.37	0.60	0.37
	0.37	0.37	0	0.37	0.29	0.37	0.60	0.37	0.22	0.37
4	0.37	0.74	0.74	0	0.37	0.56	0.56	0.56	0.56	0.56
	0.37	0	0.37	0	0.37	0.28	0.28	0.28	0.28	0.28
	0.37	0.37	0	0	0.37	0.28	0.28	0.28	0.28	0.28
5	1.11	1.11	1.11	1.11	0	1.11	1.11	1.11	1.11	1.11
6	1.11	1.11	1.11	1.11	1.11	0	1.11	1.11	1.11	1.11
7	1.11	1.11	1.11	1.11	1.11	1.11	0	1.11	1.11	1.11
8	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0	1.11	1.11
9	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0	1.11
10	0.16	0.28	0.28	0.31	0.28	0.23	0.35	0.30	0.15	0
	0.39	0.28	0.28	0.25	0.28	0.14	0.39	0.44	0.22	0
	0.09	0.28	0.28	0.32	0.28	0.44	0.15	0.07	0.32	0
	0.46	0.28	0.28	0.23	0.28	0.30	0.22	0.30	0.42	0

†Multiply entries by 0.10 for mean estimation. Weights are rounded.

5.2.3 Nearest-neighbor imputation

Nearest-neighbor imputation is a hot deck procedure in which a distance measure, defined on the basis of observed characteristics, is used to define the donor. The respondents closest to the element with a missing value act as donors. It is common practice to use a single donor, but we suggest that two

or more donors be used for each recipient. The use of more than one donor facilitates variance estimation and generally improves efficiency.

Assume that the finite universe is generated by a stochastic mechanism and that a distance measure is defined for the elements. Let a neighborhood of element g be composed of elements that are close to element g, and let

$$\begin{aligned} \mu_g &= E\{y_j \mid j \in B_g\}, \\ \sigma_g^2 &= E\{(y_j - \mu_j)^2 \mid j \in B_g\}, \end{aligned}$$

where B_g is the set of indices for the elements in the neighborhood of element g. One might suppose that there would be some correlation among elements in the neighborhood, with elements that are close having a positive correlation, but we will assume that neighborhoods are small enough so that the correlation can be ignored. We assume that an adequate approximation for the distribution of elements in the neighborhood is

$$y_j \sim ii(\mu_g,\ \sigma_g^2), \qquad j \in B_g, \tag{5.2.6}$$

where $\sim ii$ denotes independent identically distributed. We assume that response is independent of the y values so that the distribution (5.2.6) holds for both recipients and donors. Our results are obtained under the working assumption (5.2.6). For the assumption to hold exactly for every neighborhood, the assumption must hold globally or the neighborhoods must be mutually exclusive. See Chen and Shao (2000, 2001) for conditions under which it is reasonable to use (5.2.6) as an approximation.

Let a probability sample be selected from the finite universe with selection probabilities π_j. Let $\hat{\theta}_n$ be a design linear estimator based on the full sample,

$$\hat{\theta}_n = \sum_{i \in A} w_i y_i, \tag{5.2.7}$$

and let $V\{\hat{\theta}_n\}$ be the variance of the full-sample estimator. Under model (5.2.6) we can write

$$y_i = \mu_i + e_i, \tag{5.2.8}$$

where the e_i are independent $(0,\ \sigma_i^2)$ random variables and μ_i is the neighborhood mean. Then, under model (5.2.6), the variance of $\hat{T}_y = \Sigma_{i \in A} w_i y_i$ is

$$V\left\{\sum_{i \in A} w_i y_i - T_y\right\} = V\left\{\sum_{i \in A} w_i \mu_i - T_\mu\right\} + E\left\{\sum_{i \in A} (w_i^2 - w_i)\sigma_i^2\right\}, \tag{5.2.9}$$

where T_y is the population total of the y_i and T_μ is the population total of the μ_i. Note that the variance is an unconditional variance.

Assume that y is missing for some elements and assume that there are always at least M observations on y in the neighborhood of each missing value. Let an imputation procedure be used to assign M donors to each recipient. Let w_{ij}^* be the fraction of the weight allocated to donor i for recipient j. Then

$$\alpha_i = \sum_{j \in A} w_j w_{ij}^* \tag{5.2.10}$$

is the total weight for donor i, where it is understood that $w_{ii}^* = 1$ for a donor donating to itself. Thus, the imputed linear estimator is

$$\hat{\theta}_I = \sum_{j \in A} w_j y_{Ij} = \sum_{i \in A_R} \alpha_i y_i, \tag{5.2.11}$$

where A_R is the set of indices of the respondents, the mean imputed value for recipient j is

$$y_{Ij} = \sum_{i \in A} w_{ij}^* y_i, \tag{5.2.12}$$

and $y_{Ij} = y_j$ if j is a respondent. Then, under model (5.2.6),

$$V\{\hat{T}_{yI} - T_y\} = V\left\{\sum_{i \in A} w_i \mu_i - T_\mu\right\} + E\left\{\sum_{i \in A_R} (\alpha_i^2 - \alpha_i)\sigma_i^2\right\}, \tag{5.2.13}$$

where A_R is the set of indices of the respondents and $\hat{T}_{yI}$ is the estimated total based on imputed data. The increase in variance due to imputing for missing values is, from (5.2.9),

$$\sum_{i \in A_R} (\alpha_i^2 - \alpha_i)\sigma_i^2 - \sum_{i \in A} (w_i^2 - w_i)\sigma_i^2.$$

To use replication to estimate the variance of the imputed estimator, let a replication variance estimator for the complete sample be

$$\hat{V}_1\{\hat{\theta}\} = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2, \tag{5.2.14}$$

where $\hat{\theta}$ is the full-sample estimator, $\hat{\theta}^{(k)}$ is the kth estimate of θ_N based on the kth replicate, L is the number of replicates, and c_k is a factor associated

with replicate k determined by the replication method. If imputed data are used in (5.2.14) for $\hat{\theta} = \hat{T}_{yI}$,

$$E\{\hat{V}_1(\hat{T}_{yI} - T_y)\} = V\left\{\sum_{i\in A} w_i\mu_i - T_\mu\right\} + E\left\{\sum_{k=1}^{L}\sum_{i\in A_R} c_k(\alpha_{i1}^{(k)} - \alpha_i)^2\sigma_i^2\right\}, \quad (5.2.15)$$

where $\alpha_{i1}^{(k)} = \sum_j w_j^{(k)} w_{ij}^*$ and $w_j^{(k)}$ is the weight for element j in replicate k. The estimator $\hat{V}_1(\hat{\theta})$ computed as if imputed data were observed is sometimes called the *naive variance estimator*. We outline a replication procedure that produces unbiased variance estimates.

For nearest-neighbor imputed data, there are three types of observations in the data set:

1. Respondents that act as donors for at least one recipient
2. Respondents that are never used as donors
3. Recipients

The original full-sample replicate weights will be used for the last two types. For donors, the initial fractional weights w_{ij}^* in replicate k will be modified so that we obtain the correct expectation. Let superscript k denote the replicate where element k is in the deleted set. Following Kim and Fuller (2004), the fractions assigned to donor k are changed so that the expected value of the sum of squares is changed by the proper amount. First, the full-sample replicates for the variance estimator (5.2.14) are computed, and the sum of squares for element i computed as

$$\sum_{k=1}^{L} c_k(\alpha_{i1}^{(k)} - \alpha_i)^2 = \Phi_i, \qquad i \in A_R, \quad (5.2.16)$$

where $\alpha_{i1}^{(k)}$ is defined following (5.2.15).

In the second step, the fractions for replicates for donors are modified. Let R_k be the set of indices of recipients for which k is a donor. We use k as the index for the replicate and for the donor. Let the new fractional weight in replicate k for the value donated by element k to recipient j be

$$w_{kj}^{*(k)} = w_{kj}^* b_k, \quad (5.2.17)$$

where b_k is to be determined. For two donors to each recipient the new fractional weight for the other donor, denoted by t, is

$$w_{tj}^{*(k)} = 1 - b_k w_{kj}^*. \tag{5.2.18}$$

For $w_{kj}^* = 0.5$, $w_{kj}^{*(k)} = 0.5b_k$ and $w_{tj}^{*(k)} = (1 - 0.5b_k)$. Then, by (5.2.15), the b_k that gives the correct sum of squares is the solution to the quadratic equation

$$\begin{aligned} & c_k \left(w_k^{(k)} + b_k \sum_{\substack{j \in R_k \\ j \neq k}} w_j^{(k)} w_{ij}^* - \alpha_k \right)^2 \\ & + \sum_{t \in D_{Rk}} c_t \left(w_t^{(k)} + \sum_{j \in R_t \cap R_k} w_j^{(k)} (1 - w_{kj}^* b_k) - \alpha_t \right)^2 \\ & - c_k (\alpha_{k1}^{(k)} - \alpha_k)^2 - \sum_{j \in D_{Rk}} c_j (\alpha_{j1}^{(k)} - \alpha_j)^2 \\ & = \alpha_k^2 - \alpha_k - \Phi_k, \end{aligned} \tag{5.2.19}$$

where t is used as the index for the donors other than k that donate to j, and D_{Rk} is the set of donors other than k that donate to recipients that receive a value from donor k. The difference $\Phi_k - (\alpha_k^2 - \alpha_k)$ is the difference between the sum of squares for the naive estimator and the sum of squares desired for observation k. Under the assumption of a common variance in a neighborhood and the assumption that the full-sample variance estimator $\hat{V}_1(\hat{\theta})$ of (5.2.14) is unbiased, the variance estimator with b_k defined by (5.2.19) is unbiased for the variance of the mean of the imputed sample. The procedure corrects weights within each replicate and does not force the sum of squares over replicates for observation i to be equal to α_i^2.

Example 5.2.2. Table 5.6 contains an illustration data set of six observations. The variable x_i is observed on all six, but the variable y is missing for observations 3 and 6. The variable x is used to determine distance and, in Euclidean distance, observation 2 is closest to observation 3. Therefore, using the nearest-neighbor rule, we replace the missing value for observation 3 by the value of observation 2. In the same way, observation 5 is closest to observation 6, so the missing value for observation 6 is replaced by 2.3, the value of y for observation 5. If only the nearest neighbor is used for imputation, we obtain the imputed data set of the last column. The weight of 1/6 would be the weight for a simple mean.

Table 5.6 Data Set

Obs.	Weight	x_i	y_i	y_{Ii}
1	$0.16\bar{6}$	0.9	0.7	0.7
2	$0.16\bar{6}$	1.1	1.0	1.0
3	$0.16\bar{6}$	1.3	M	1.0
4	$0.16\bar{6}$	2.2	1.9	1.9
5	$0.16\bar{6}$	2.6	2.3	2.3
6	$0.16\bar{6}$	3.1	M	2.3

Table 5.7 contains the imputed observations when two imputations are made for each missing value. The second nearest neighbor for observation 2 is observation 4, where $\mid x_2 - x_4 \mid = 0.6$. Observation 6 has the largest x-value, so the two nearest neighbors are observations 4 and 5. In some situations one might impose the restriction that a donor is used only once when that is possible. We use the strict nearest-neighbor rule and use observation 4 as a donor for both observations 3 and 6. Each imputed value is weighted by one-half of the weight of the original observation. For observation 3 there are two new lines in the imputed data set, each with a weight of 1/12. The x-value is the same for both lines. If we had additional variables in the data set, those data are also repeated for the two lines.

Table 5.7 Jackknife Data

				Naive Replicate Weights					
Obs.	Donor	Weight	y_{Ii}	1	2	3	4	5	6
1	–	$0.16\bar{6}$	0.7	0	0.2	0.2	0.2	0.2	0.2
2	–	$0.16\bar{6}$	1.0	0.2	0	0.2	0.2	0.2	0.2
3	2	$0.08\bar{3}$	1.0	0.1	0.1	0	0.1	0.1	0.1
	4	$0.08\bar{3}$	2.2	0.1	0.1	0	0.1	0.1	0.1
4	–	$0.16\bar{6}$	2.2	0.2	0.2	0.2	0	0.2	0.2
5	–	$0.16\bar{6}$	2.3	0.2	0.2	0.2	0.2	0	0.2
6	4	$0.08\bar{3}$	2.2	0.1	0.1	0.1	0.1	0.1	0
	5	$0.08\bar{3}$	2.3	0.1	0.1	0.1	0.1	0.1	0

We construct jackknife replicate weights for variance estimation. The weights for six naive jackknife replicates are given in Table 5.8, where the weights are constructed as if the sample were complete. The two imputed values for an observation are treated as two observations from a primary

Table 5.8 Naive Weights for Respondents

Obs.	α_i	Naive Respondent Replication Weights $\alpha_{1i}^{(1)}$	$\alpha_{1i}^{(2)}$	$\alpha_{1i}^{(3)}$	$\alpha_{1i}^{(4)}$	$\alpha_{1i}^{(5)}$	$\alpha_{1i}^{(6)}$
1	$0.16\bar{6}$	0.0	0.2	0.2	0.2	0.2	0.2
2	0.250	0.3	0.1	0.2	0.3	0.3	0.3
4	$0.33\bar{3}$	0.4	0.4	0.3	0.2	0.4	0.3
5	0.250	0.3	0.3	0.3	0.3	0.1	0.2

sampling unit. Ignoring the finite population correction, $c_k = 5/6$. The full-sample replicate weights for the respondents are given in Table 5.8. We have

$$(5/6)\sum_{k=1}^{L}(\alpha_{1i}^{(k)} - \alpha_i)^2 \quad = \quad (0.0278,\ 0.0292,\ 0.0278,\ 0.0292)$$

for $i = 1,\ 2,\ 4,\ 5$, respectively. From Table 5.8,

$$(\alpha_1^2, \alpha_2^2, \alpha_3^2, \alpha_4^2) = (0.0278, 0.0625, 0.1111, 0.0625).$$

The use of the naive replicates severely underestimates most of the coefficients for σ_i^2. Only for observation 1, the observation not used as a donor, is the sum of squares from the naive replicates equal to α_i^2.

Using the Φ_2 defined in (5.2.16) and the $w_{tj}^{*(k)}$ of (5.2.17), the quadratic equation for b_2 is

$$[0 + b_2(0.2)(0.5) - 0.25]^2 + [0.2 + 0.2(0.5) + (1 - 0.5b_2)(0.2) - 0.3333]^2 - 0.0350 - 0.0269 \quad = \quad 0.0750,$$

where $0.0750 - c_2^{-1}\alpha_2^2$, $c_k = 5/6$, and $\Phi_2 = 0.0350$. The simplified quadratic equation is

$$0.02b_2^2 - 0.0833b_2 + 0.0234 = 0$$

and $b_2 = 0.3033$. The equation for b_5 is the same as that for b_2. The quadratic equation for b_4 is

$$[0 + 2(0.1)b_4 - 0.3333]^2 + 2[0.2 + (1 - 0.5b_4)(0.2) - 0.25]^2 + 0.0333 - 0.0228 \quad = \quad 0.1333$$

and $b_4 = 0.1827$. The final jackknife replicates are given in Table 5.9 and the respondent weights in Table 5.10.

Table 5.9 Jackknife Weights for Fractional Imputation

Obs.	Donor	Weight	Weights for Unbiased Variance Estimator 1	2	3	4	5	6
1	–	$0.16\bar{6}$	0	0.2	0.2	0.2	0.2	0.2
2	–	$0.16\bar{6}$	0.2	0	0.2	0.2	0.2	0.2
3	2	$0.08\bar{3}$	0.1	0.030	0	0.183	0.1	0.1
	4	$0.08\bar{3}$	0.1	0.170	0	0.017	0.1	0.1
4	–	$0.16\bar{6}$	0.2	0.2	0.2	0	0.2	0.2
5	–	$0.16\bar{6}$	0.2	0.2	0.2	0.2	0	0.2
6	4	$0.08\bar{3}$	0.1	0.1	0.1	0.017	0.170	0
	5	$0.08\bar{3}$	0.1	0.1	0.1	0.183	0.030	0

Table 5.10 Final Weights for Respondents

Obs.	α_i	Final Respondent Replication Weight $\alpha_i^{(1)}$	$\alpha_i^{(2)}$	$\alpha_i^{(3)}$	$\alpha_i^{(4)}$	$\alpha_i^{(5)}$	$\alpha_i^{(6)}$
1	$0.16\bar{6}$	0.0	0.2000	0.2	0.2000	0.2000	0.2
2	0.250	0.3	0.0303	0.2	0.3817	0.3000	0.3
4	$0.33\bar{3}$	0.4	0.4697	0.4	0.0366	0.4697	0.3
5	0.250	0.3	0.3000	0.3	0.3817	0.0303	0.2

The reader may check that

$$(5/6)\sum_{i\in A_R}\sum_{k=1}^{6}(\alpha_i^{(k)}-\alpha_i)^2 \;=\; \sum_{i\in A_R}\alpha_i^2,$$

where $A_R = (1,\ 2,\ 4,\ 5)$. Only for $i = 1$ is $\sum_k(\alpha_1^{(k)}-\alpha_1)^2 = \alpha_1^2$. For other observations the individual sums deviate slightly. Under our assumptions the neighborhoods that share a common donor have the same variance and hence the variance estimator is unbiased. Of course, the unbiased result requires the model assumptions of (5.2.6). ■■

5.2.4 Imputed estimators for domains

One of the reasons imputation is used in place of weighting is to improve estimates for domains. If the imputation model includes items such as age

and gender but not local geography, it is reasonable to believe that imputation will give an estimator for a small geographic area that is superior to the mean of the respondents in that area. If the model used for imputation is true, the imputed estimator for the small area may be superior to the simple estimator constructed from the full sample.

To illustrate the last point, consider a simple random sample and assume that the imputation model is

$$y_i = \mu + e_i, \tag{5.2.20}$$

where the e_i are $iid(0, \sigma^2)$ random variables. Let there be m nonrespondents and let the imputed value for each nonrespondent be the mean of the respondents. Let z_i be an indicator variable for membership in a domain, where a domain might be a cell in a two-way table. Assume that z_i is observed for all elements of the sample, and let the imputed estimator for domain a be

$$\hat{\mu}_a = \left(\sum_{i \in A} z_i \right)^{-1} \sum_{i \in A} z_i y_{iI}, \tag{5.2.21}$$

where y_{iI} is the imputed value for the rth element and $y_{iI} = y_i$ for respondents. Let domain a contain r_a respondents and m_a nonrespondents. Then the estimated domain mean based on imputed data is

$$\hat{\mu}_a = (m_a + r_a)^{-1} \left(\sum_{i \in A_{R,a}} y_i + m_a \bar{y}_r \right), \tag{5.2.22}$$

where $A_{R,a}$ is the set of indices of respondents in the domain and $\bar{y}_r$ is the mean of all respondents.

The model (5.2.20) is assumed to hold for all observations and hence holds for observations in the domain. Under model (5.2.20), mean imputation, and a negligible finite population correction,

$$\begin{aligned} V\{\hat{\mu}_a\} &= (m_a + r_a)^{-2}(r_a + 2m_a r_a r^{-1} + m_a^2 r^{-1})\sigma^2 \\ &= [(m_a + r_a)^{-1} + (m_a + r_a)^{-2} r^{-1} m_a (m_a + 2r_a - r)]\sigma^2 \\ &= [r^{-1} + (m_a + r_a)^{-2}(r_a - r_a^2 r^{-1})]\sigma^2. \end{aligned} \tag{5.2.23}$$

The second set of the expressions in (5.2.23) demonstrates that the imputed domain estimator is superior to the full-sample estimator if $r > m_a + 2r_a$, a condition easy to satisfy if the domain is small.

Under the model, the best estimator for the domain is $\bar{y}_r$. The last expression in (5.2.23) contains the increase in variance for the imputed domain estimator relative to the grand mean of the respondents. Often, practitioners are willing to use the model for imputation but unwilling to rely on the model to the degree required to use the model estimated mean for the cell. When the practitioner is willing to use the model estimator for the domain, the procedure is more often called *small area estimation.* See Section 5.5.

The use of donors from outside the domain produces a bias in the fractional replicated variance estimator for the domain. For nearest-neighbor imputation and an estimator linear in y, we constructed replicate weights that met the unbiasedness requirement (5.2.16) or an equivalent requirement. Because the weights for a domain estimator are not the same as the weights for the overall total, (5.2.16) will, in general, not hold for the domain mean.

Example 5.2.3. We use the imputed data of Table 5.7 to illustrate the nature of domain estimation. Assume that observations 1, 2, and 3 are in a domain. Then the imputed estimator for the domain total of y is

$$\hat{T}_d = N \sum_{j \in A} \sum_{i \in A_{Ij}} w_{ij}^* y_{[i]j} \delta_{dj}$$

and the imputed estimator for the domain mean is

$$\bar{y}_d = \left(\sum_{j \in A} \sum_{i \in A_{Ij}} w_{ij}^* \delta_{dj} \right)^{-1} \sum_{j \in A} \sum_{i \in A_{Ij}} w_{ij}^* y_{[i]j} \delta_{dj},$$

where $y_{[i]j}$ is the imputed value from donor i to recipient j, A_{Ij} is the set of indices of donors to j, w_{ij}^* are the weights of Table 5.7, and

$$\begin{aligned} \delta_{dj} &= 1 \quad \text{if observation } j \text{ is in domain } d \\ &= 0 \quad \text{otherwise.} \end{aligned}$$

Table 5.11 Respondent Weights for Alternative Estimators

	Observation			
Estimator	1	2	4	5
Total ($N = 1000$)	167	250	333	250
Imputed domain	167	250	83	
Poststrat. domain	250	250		

The line "imputed domain" in Table 5.11 gives the weights for the three respondents that contribute to the estimate for the domain. The weights are for a total under the assumption that $N = 1000$. Although observation 4 is not in the domain, it contributes to the domain estimate because, under the model, observation 4 has the same expectation as observation 2, which is in the domain. If only the two observations that fall in the domain are used to estimate the domain total, they receive the weights given in the last line of Table 5.11. Clearly, the weights of the second line of Table 5.11 will give a smaller sum of squares than those of the third line. ■■

5.3 VARIANCE ESTIMATION

Systematic sampling and one-per-stratum sampling produce unbiased estimators of totals, but design-unbiased variance estimation is not possible because some joint probabilities of selection are zero. One approach to variance estimation in these cases is to postulate a mean model and use deviations from the fitted model to construct a variance estimator.

Models can be divided into two types: local models and global models. For a population arranged in natural order, on a single variable x, a local model for y given x in the interval $q_j = (x_{Lj}, x_{Uj})$, is

$$y_i = g\left(x_i, \boldsymbol{\beta}_j\right) + e_i, \; x_i \in q_j, \qquad (5.3.1)$$
$$e_i \sim ind(0, \sigma_i^2).$$

The model is often simplified by letting x be the order number of the sample observations. A global model assumes that $g\left(x_i, \boldsymbol{\beta}\right)$ holds for the entire data set.

The most common local model assumption for a one-per-stratum design is that the means of two strata are the same. Thus, for an ordered set of an even number of strata,

$$g\left(x_i, \mu_h\right) = \mu_h, \qquad i = 1, \; 2, \; \ldots, \; 0.5n, \qquad (5.3.2)$$

for $x_i = 2h$ and $x_i = 2h - 1$, where x_i is the order identification of the strata. Use of the mean model (5.3.2) leads to the procedure of collapsed strata discussed in Section 3.1.3.

Of the many variance estimation procedures that have been suggested for systematic sampling, the most popular is to form pairs of sample elements and assume a common mean for the pair. The pair is then treated as a set of two observations from a stratum. The created strata are sometimes called *pseudostrata*. Unlike the result for collapsed strata with one-per-stratum

sampling, the estimated variance for a systematic sample based on model (5.3.2) is not guaranteed to be an overestimate.

For a systematic sample or a one-per-stratum sample from a population arranged in natural order, a local model can be used to increase the number of degrees of freedom for the variance estimator relative to that for the collapsed strata procedure. Let $y_{[i]}$ denote the ith observation, where the order is that used in the sample selection, and let $x_i = i$. Then a local model that specifies adjacent observations to have the same mean is

$$\begin{aligned} y_{[i]} &= \mu_j + e_i, \\ e_i &\sim ind(0, \sigma_i^2), \end{aligned} \tag{5.3.3}$$

for $i \in [j, j+1]$. The associated variance estimator is

$$\begin{aligned} \hat{V}\{\bar{y}_\pi\} &= 0.5w_1^2(1-\pi_1)(y_{[2]} - y_{[1]})^2 \\ &\quad + 0.25\sum_{i=2}^{n-1} w_i^2(1-\pi_i)\left[(y_{[i-1]} - y_{[i]})^2 + (y_{[i]} - y_{[i+1]})^2\right] \\ &\quad + 0.5w_n^2(1-\pi_n)(y_{[n]} - y_{[n-1]})^2, \end{aligned} \tag{5.3.4}$$

where $w_i = N^{-1}\pi_i^{-1}$. If $w_i = n^{-1}$, the estimator reduces to

$$\begin{aligned} V\{\bar{y}\} &= N^{-1}(N-n)n^{-2}\{0.25\left[(y_{[2]} - y_{[1]})^2 + (y_{[n]} - y_{[n-1]})^2\right] \\ &\qquad + 0.5\sum_{i=2}^{n}(y_{[i]} - y_{[i-1]})^2\} \\ &\doteq N^{-1}(N-n)0.5n^{-1}(n-1)^{-1}\sum_{i=2}^{n}(y_{[i]} - y_{[i-1]})^2. \end{aligned} \tag{5.3.5}$$

The estimator (5.3.4) has nearly twice as many degrees of freedom as the collapsed strata procedure.

A second local model assumes a linear model for the center observation in a set of three adjacent observations. The model is

$$\begin{aligned} y_{[i]} &= \beta_{0j} + x_i\beta_{1j} + e_i \\ e_i &\sim ind(0, \sigma_i^2) \end{aligned} \tag{5.3.6}$$

for $x_i \in \{j-1,\ j,\ j+1\}$, where, as before, $x_i = i$ is the order identification. The use of local models for intervals greater than 2 usually requires an adjustment for the end observations. With model (5.3.6), for variance estimation purposes, we assume that the superpopulation mean associated with the first observation is equal to the superpopulation mean associated with the second

Table 5.12 Weights for Replicate Variance Estimation

Observation	Full Sample	Replicate 1	2	···	9	10
1	0.1000	0.1671	0.1387		0.1000	0.1000
2	0.1000	0.0329	0.0226		0.1000	0.1000
3	0.1000	0.1000	0.1387		0.1000	0.1000
4	0.1000	0.1000	0.1000		0.1000	0.1000
5	0.1000	0.1000	0.1000		0.1000	0.1000
6	0.1000	0.1000	0.1000		0.1000	0.1000
7	0.1000	0.1000	0.1000		0.1000	0.1000
8	0.1000	0.1000	0.1000		0.1387	0.1000
9	0.1000	0.1000	0.1000		0.0226	0.1671
10	0.1000	0.1000	0.1000		0.1387	0.0329

observation. Making the same assumption for the last two observations, the estimated variance of an estimator of the form $\Sigma_{i\in A} w_i y_i$ is

$$\begin{aligned}\hat{V}\{\bar{y}\} &= N^{-1}(N-n)[0.5w_1^2(y_{[1]}-y_{[2]})^2 \\ &\quad + \sum_{i=2}^{n-1} w_i^2 6^{-1}(y_{[i-1]}-2y_{[i]}+y_{[i+1]})^2 \\ &\quad + 0.5w_n^2(y_{[n-1]}-y_{[n]})^2]. \end{aligned} \tag{5.3.7}$$

The linear combination $(y_{[i-1]} - 2y_{[i]} + y_{[i+1]})$ is a multiple of the deviation from fit for the linear model estimated with the three observations $(y_{[i-1]}, y_{[i]}, y_{[i+1]})$.

The estimator (5.3.7) has a positive bias for the design variance of the one-per-stratum design. The bias expression can be obtained by replacing y_j with μ_j in (5.3.7), where μ_j is the mean for the jth stratum.

Example 5.3.1. Replication can be used to construct estimator (5.3.7). To illustrate, assume that a sample of 10 observations is selected, either by systematic sampling or by equal-probability one-per-stratum sampling, from an ordered population of 100 elements.

The replicate weights in Table 5.12 are such that the kth deviation $\bar{y}^{(k)} - \bar{y}$ is the kth linear combination in (5.3.7) normalized so that the square has the correct expectation. Thus, the entries in the first column for replicate 1 give

$$\begin{aligned}\bar{y}^{(1)} - \bar{y} &= \left[0.5N^{-1}(N-n)\right]^{0.5} n^{-1}(y_{[1]}-y_{[2]}) \\ &= 0.0671(y_{[1]}-y_{[2]}).\end{aligned}$$

and the entries in the second column give

$$\begin{aligned}\bar{y}^{(2)} - \bar{y} &= \left[6^{-1}N^{-1}(N-n)\right]^{0.5} n^{-1}(y_{[1]} - 2y_{[2]} + y_{[3]}) \\ &= 0.0387y_{[1]} - 0.0774y_{[2]} + 0.0387y_{[3]}.\end{aligned}$$

■■

Models for the covariance structure of the population can be used to estimate the variance of one-per-stratum designs. Let the population size for stratum h be N_h, and let U_h be the set of indices for stratum h. Let a simple random sample of size 1 be selected in each stratum. Then the variance of the estimated total for stratum h is

$$V\{\bar{y}_{st}\} = \sum_{h=1}^{H}(1 - N_h^{-1})W_h^2 S_h^2, \tag{5.3.8}$$

where $W_h = N^{-1}N_h$. Assume that the finite population is a realization of a stationary stochastic process, where the covariance is a function of the distance between observations. That is,

$$C\{y_j, y_k\} = \gamma(d), \tag{5.3.9}$$

where d is the distance between j and k and $\gamma(d) = \gamma(-d)$ is the covariance between two units that are a distance d apart. The distance can be defined in terms of auxiliary information and is often the distance between the indexes of a population arranged in natural order.

Under the model,

$$E\{S_h^2\} = (N_h - 1)^{-1}\left(N_h\gamma(0) - N_h^{-1}\sum_{d=-(N_h-1)}^{N_h-1}(N_h - d)\gamma(d)\right)$$

and

$$E\{V(y_j - \bar{y}_{h,N} \mid j \in U_h, \mathcal{F})\} = \gamma(0) - N_h^{-2}\sum_{d=-(N_h-1)}^{N_h-1}(N_h - d)\gamma(d). \tag{5.3.10}$$

If a parametric model for $\gamma(d)$ can be estimated, the estimated values for $\gamma(d)$ can be substituted in (5.3.10) and (5.3.8) to obtain an estimated variance for $\bar{y}_{st}$. It may be simpler to estimate the parameters of

$$E\{(y_j - y_k)^2\} = \psi(d, \boldsymbol{\theta}), \tag{5.3.11}$$

where $\psi(d,\ \boldsymbol{\theta})$ is called the *variogram* and $\boldsymbol{\theta}$ is the parameter vector.

For the one-per-stratum procedure, only differences with $d \leq N_M - 1$ enter the variance expression, where N_M is the maximum of the N_h. Thus, a simple procedure is to assume a constant variogram for $d \leq N_M - 1$ and estimate the variance with

$$\hat{V}\{\bar{y}_{st}\} = 0.5 \sum_{h=1}^{H} (1 - N_h^{-1}) W_h^2 n_\delta^{-1} \sum_{j \in A} \sum_{k \in A} (y_j - y_k)^2 \delta_{j,k}, \quad (5.3.12)$$

where

$$\begin{aligned} \delta_{j,k} &= 1 \quad \text{if } 0 < \mid j - k \mid \leq N_M - 1 \\ &= 0 \quad \text{otherwise} \end{aligned}$$

and

$$n_\delta = \sum_{j \in A} \sum_{k \in A} \delta_{j,k}.$$

For a variogram that increases with distance, the estimator (5.3.12) will have a positive bias because differences with large d appear more frequently in the estimator than in the population.

A three-parameter variogram model based on the first-order autoregressive process is

$$\psi(d, \boldsymbol{\theta}) = \theta_0 + \theta_1 \theta_2^{|j-k|}, \quad (5.3.13)$$

where $\theta_0 \geq \theta_1$ and $\mid \theta_2 \mid < 1$. For the stationary first-order autoregressive process, $\theta_0 = \theta_1$. For a process with measurement error, $\theta_0 > \theta_1$. The parameters can be estimated using a nonlinear least squares procedure or by maximum likelihood. A large number of observations is required to obtain good estimates for the parameters of variance models. See Cressie (1991, Chapter 2) for a discussion of the variogram and for variogram models.

5.4 OUTLIERS AND SKEWED POPULATIONS

The problem of estimating the mean, or total, using a sample containing a few "very large" observations will be faced by almost every sampling practitioner. The definition of "very large" must itself be part of a study of estimation for such samples and the definition of "very large" that appears most useful is a definition that separates cases wherein the sample mean performs well as an estimator from those cases wherein alternative estimators are markedly

superior to the mean. Most editing procedures will have rules that identify unusual observations as part of the checking for errors. In this section we are interested in situations where the extreme observation, or observations, have been checked and the data are believed to be correct. These are typically situations where the population sampled is very skewed. Personal income and size measures of businesses are classical examples.

One approach to estimation for skewed populations is to specify a parametric model for the superpopulation and estimate the parameters of that model. Parametric estimation is discussed in Chapter 6. Our experience suggests that it is very difficult to specify a relatively simple model for the entire distribution. Robust procedures, as described by Huber (1981) and Hempel et al. (1986), are related estimation procedures but have heavy emphasis on symmetric distributions.

In applications, an observation can be extreme because the value of the characteristic is large, because the weight is large, or both. Estimators that make adjustments in the largest observations can be made by modifying the value or by modifying the weight. Because the value is believed to be correct, the modification is most often made by modifying the weight. We begin by considering the general estimation problem for simple random samples.

We present the procedure of Fuller (1991), in which it is assumed that the right tail of the distribution can be approximated by the right tail of a Weibull distribution. The Weibull density is

$$\begin{aligned} f(y;\alpha,\lambda) &= \alpha\lambda^{-1}y^{\alpha-1}exp\{-\lambda^{-1}y^{\alpha}\} && \text{if } y > 0 \\ &= 0 && \text{otherwise,} \end{aligned} \quad (5.4.1)$$

where $\lambda > 0$ and $\alpha > 0$. If x is defined by the one-to-one transformation $x = y^{\alpha}$, x is distributed as an exponential random variable with parameter λ. Conversely, the Weibull variable is the power of an exponential variable, x^{γ}, where $\gamma = \alpha^{-1}$. If $\alpha \leq 1$, the sample mean will perform well as an estimator of the population mean. If α is much larger than 1, there are alternative estimators that will perform better than the sample mean. We use the order statistics to test the hypothesis that $\alpha = 1$ against the alternative that $\alpha > 1$.

The distribution of the differences of order statistics from the exponential distribution are distributed as exponential random variables. Let $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ be the order statistics of a sample of size n selected from an exponential distribution with parameter λ, and let $x_{(0)} = 0$. Then the random variables

$$z_k = (n - k + 1)(x_{(k)} - x_{(k-1)}), \qquad k = 1, 2, \ldots, n, \quad (5.4.2)$$

are *iid* exponential random variables with parameter λ. See David (1981, p. 20). Postulating the exponential model for the largest m observations, we

construct the test

$$F_{mj} = \left((m-j)^{-1} \sum_{i=n-m+1}^{n-j} Z_i \right)^{-1} j^{-1} \sum_{i=n-j+1}^{n} z_i. \qquad (5.4.3)$$

If $\alpha = 1$, F_{mj} is distributed as Snedecor's F with $2j$ and $2(m-j)$ degrees of freedom. If the test rejects $\alpha = 1$, one has reason to believe that there are estimators superior to the mean. A relatively simple estimator constructed with the order statistics is

$$\begin{aligned} \hat{\mu}_{mj} &= \bar{y} && \text{if } F_m < K_j \\ &= n^{-1} \left(\sum_{i=1}^{n-j} y_{(i)} + j(y_{n-j} + K_j \bar{d}_{mj}) \right) && \text{otherwise,} \end{aligned} \qquad (5.4.4)$$

where F_m is defined in (5.4.3), K_j is a cutoff value, and

$$\bar{d}_{mj} = (m-j)^{-1} \left(\sum_{i=j}^{m-1} (y_{(n-i)} - y_{(n-m)}) + j(y_{(n-j)} - y_{(n-m)}) \right).$$

The estimator of (5.4.4) is a test-and-estimate procedure in which the estimator is a continuous function of the sums formed from different sets of order statistics. The sample mean is a special case of estimator (5.4.4) obtained by setting K_j equal to infinity.

It is difficult to specify the number of tail observations, m, the number of large order statistics, j, and the cutoff values, K_j, to use in constructing the estimator for the tail portion. It would seem that m approximately equal to one-fifth to one-third of the observations is reasonable for many populations and sample sizes. It also seems that one can reduce this fraction in large $(n > 200)$ samples. When the sample is large, setting $m \doteq 30$ seems to perform well.

In many applications $j = 1$, and K_j equal to the 99.5 percentile of the F distribution works well. The large value required for rejection means that the procedure has good efficiency for populations with modest skewness. See Fuller (1991a). The results of Rivest (1994) also support the use of $j = 1$.

5.5 SMALL AREA ESTIMATION

Sometimes estimates for a set of small domains are of considerable interest, but the sample sizes in the individual domains are not large enough to provide direct estimates with acceptable standard errors. In such situations models

and auxiliary variables can be used to construct improved estimates for the domains. The domains are often geographic areas, and the term *small area estimation* is used as a generic expression for such procedures in the survey literature. Rao (2003) describes a large number of procedures and models. We present one frequently used model.

Let the population be divided into M mutually exclusive and exhaustive areas. Let survey estimates be available for m, $m \leq M$, of the areas. Let $\bar{y}_g$ be the estimate of the mean for the gth area, and let $\bar{\mathbf{x}}_{gN}$ be a vector of known means for a vector of auxiliary variables for the gth area. For example, the areas might be metropolitan areas, and the means might be means per household. Assume that the $\bar{y}_g$ satisfy the model

$$\bar{y}_g = \theta_g + \bar{e}_g \tag{5.5.1}$$

and

$$\theta_g = \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} + u_g,$$

where u_g is the area effect, $\bar{e}_g$ is the sampling error,

$$u_g \sim ii(0, \sigma_u^2),$$

$$\bar{e}_g \sim ind(0, \sigma_{eg}^2),$$

and u_g is independent of $\bar{e}_h$ for all h and g. This model is also called a *mixed model* because the mean of y for area g is assumed to be the sum of a fixed part $\bar{\mathbf{x}}_{gN}\boldsymbol{\beta}$ and a random part u_g. The unknown mean for area g is $\theta_g = \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} + u_g$.

To be comfortable with model (5.5.1) the analyst should feel that the observable important differences among areas are included in the vector $\bar{\mathbf{x}}_{gN}$. That is, after adjusting for $\bar{\mathbf{x}}_{gN}$, there is no reason to believe that any area is particularly unusual relative to the others. We state our model for means, but the nature of the data will vary for different problems. The model could be stated in terms of mean per primary sampling unit or mean per element. The model could also be defined in terms of small area totals, but we find the model (5.5.1) more appealing when expressed in terms of means.

To introduce the estimation procedure, assume $\boldsymbol{\beta}$, σ_u^2, σ_{eg}^2 are known. Then $\bar{\mathbf{x}}_{gN}\boldsymbol{\beta}$ and $u_g + \bar{e}_g = \bar{y}_g - \bar{\mathbf{x}}_{gN}\boldsymbol{\beta}$ are known for the m sampled areas. If $(u_g, \bar{e}_g)$ is normally distributed, then $(u_g + \bar{e}_g, u_g)$ is normally distributed and the best predictor of u_g, given $u_g + \bar{e}_g$, is

$$\hat{u}_g = \gamma_g(u_g + \bar{e}_g), \tag{5.5.2}$$

where

$$\gamma_g = (\sigma_u^2 + \sigma_{eg}^2)^{-1}\sigma_u^2$$

is the population regression coefficient for the regression of u_g on $(u_g + \bar{e}_g)$. If $(u_g, \bar{e}_g)$ is not normal, (5.5.2) is the best linear unbiased predictor of u_g. Therefore, a predictor of the mean of y for the gth area is

$$\begin{aligned} \tilde{\theta}_g &= \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} + \gamma_g(\bar{y}_g - \bar{\mathbf{x}}_{gN}\boldsymbol{\beta}) \qquad \text{if } g \in A \\ &= \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} \qquad \text{if } g \notin A, \end{aligned} \tag{5.5.3}$$

where A is the index set for small areas in which $\bar{y}_g$ is observed.

The terms *small area estimator* and *small area predictor* are both used in the literature. We prefer to describe (5.5.3) as a predictor because u_g is a random variable. The variance of the prediction error is

$$\begin{aligned} V\{\tilde{\theta}_g - \theta_g\} &= (\sigma_u^2 + \sigma_{eg}^2)^{-1}\sigma_u^2\sigma_{eg}^2 \qquad \text{if } g \in A \\ &= \sigma_u^2 \qquad \text{if } g \notin A. \end{aligned} \tag{5.5.4}$$

See Exercise 8.

If $\boldsymbol{\beta}$ is unknown but σ_u^2 and σ_{eg}^2 are known, the generalized least squares estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{g=1}^{m} \bar{\mathbf{x}}_{gN}' \left(\sigma_u^2 + \sigma_{eg}^2\right)^{-1} \bar{\mathbf{x}}_{gN} \right)^{-1} \sum_{g=1}^{m} \bar{\mathbf{x}}_{gN}' \left(\sigma_u^2 + \sigma_{eg}^2\right)^{-1} \bar{y}_g. \tag{5.5.5}$$

The estimator (5.5.5) of $\boldsymbol{\beta}$ can be substituted for $\boldsymbol{\beta}$ in (5.5.3) to obtain the unbiased predictor,

$$\begin{aligned} \hat{\theta}_g &= \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}} + \gamma_g(\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}) \qquad \text{if } g \in A \\ &= \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}} \qquad \text{if } g \notin A. \end{aligned} \tag{5.5.6}$$

The prediction variance has an added term due to the estimation of $\boldsymbol{\beta}$,

$$\begin{aligned} V\{\hat{\theta}_g - \theta_g\} &= \gamma_g\sigma_{eg}^2 + (1-\gamma_g)^2\bar{\mathbf{x}}_{gN}V\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}_{gN}' \quad \text{if } g \in A \\ &= \sigma_u^2 + \bar{\mathbf{x}}_{gN}V\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}_{gN}' \qquad \text{if } g \notin A, \end{aligned} \tag{5.5.7}$$

where

$$V\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{g=1}^{m} \bar{\mathbf{x}}_{gN}' \left(\sigma_u^2 + \sigma_{eg}^2\right)^{-1} \bar{\mathbf{x}}_{gN} \right)^{-1}. \tag{5.5.8}$$

See Exercise 9.

Estimation becomes even more difficult for the realistic situation in which σ_u^2 is unknown. Although an estimator of σ_{eg}^2 is often available, estimation of

σ_u^2 and $\boldsymbol{\beta}$ requires nonlinear estimation procedures. A number of statistical packages contain estimation algorithms for both Bayesian and classical procedures. For classical estimation, one procedure uses estimators of $\boldsymbol{\beta}$ and σ_{eg}^2 to construct an estimator of σ_u^2 and then uses the estimator of σ_u^2 and estimators of σ_{eg}^2 to construct an improved estimator of $\boldsymbol{\beta}$. The predictor is (5.5.3) with the estimators of σ_{eg}^2, σ_u^2, and $\boldsymbol{\beta}$ replacing the unknown parameters.

An estimator of the prediction mean square error (MSE) is

$$\begin{aligned}\hat{V}\{\hat{\theta}_g - \theta_g\} &= \hat{\gamma}_g\hat{\sigma}_{eg}^2 + (1-\hat{\gamma}_g)^2\bar{\mathbf{x}}_{gN}\hat{V}\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}_{gN}' \\ &\quad + 2(\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)\hat{V}\{\hat{\gamma}_g\} \qquad \text{if } g \in A \\ &= \hat{\sigma}_u^2 + \bar{\mathbf{x}}_{gN}\hat{V}\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}_{gN}' \qquad \text{if } g \notin A, \quad (5.5.9)\end{aligned}$$

where

$$\hat{V}\{\hat{\gamma}_g\} = (\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)^{-4}[\,\hat{\sigma}_u^4\hat{V}\{\hat{\sigma}_{eg}^2\} + \hat{\sigma}_{eg}^4\hat{V}\{\hat{\sigma}_u^2\}],$$

$$\hat{V}\{\hat{\boldsymbol{\beta}}\} = \left(\sum_{g=1}^{m}\bar{\mathbf{x}}_{gN}'(\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)^{-1}\bar{\mathbf{x}}_{gN}\right)^{-1},$$

$\hat{\sigma}_{eg}^2$ is an estimator of σ_{eg}^2 based on d_g degrees of freedom, and $d_g + 1$ is typically the number of primary sampling units in the small area. Many computer programs will provide an estimate of the variance of σ_u^2, where the estimated variance of $\hat{\sigma}_u^2$ will depend on the particular algorithm used.

One estimator of σ_u^2 is

$$\hat{\sigma}_u^2 = \sum_{g=1}^{m}\kappa_g[(m-k)^{-1}m(\bar{y}_g - \bar{\mathbf{x}}_g\hat{\boldsymbol{\beta}})^2 - \hat{\sigma}_{eg}^2],$$

with the estimated variance

$$\hat{V}\{\hat{\sigma}_u^2\} = \sum_{g=1}^{m}\kappa_g^2[\,2(\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)^2 + \hat{V}\{\hat{\sigma}_{eg}^2\}],$$

where k is dimension of $\mathbf{x}_g$ and

$$\kappa_g = \left(\sum_{j=1}^{m}\left[\hat{\sigma}_u^2 + d_j^{-1}(d_j+2)\hat{\sigma}_{ej}^2\right]^{-1}\right)^{-1}[\,\hat{\sigma}_u^2 + d_g^{-1}(d_g+2)\hat{\sigma}_{eg}^2]^{-1}.$$

Construction of the estimator requires iteration because κ_g depends on $\hat{\sigma}_u^2$. See Prasad and Rao (1990), Rao (2003), and Wang and Fuller (2003) for derivations and alternative estimators.

Example 5.5.1. We illustrate small area estimation with some data from the U.S. National Resources Inventory (NRI). The NRI was described in Example 1.2.2.

In this example we use data on wind erosion in Iowa for the year 2002. The analysis is based on that of Mukhopadhyay (2006). The data had not been released at the time of this study, so the data in Table 5.13 are a modified version of the original data. The general nature of the original estimates is preserved, but published estimates will not agree with those appearing in the table. The N_g is the population number of segments in the county and n_g is the sample number of segments. The variable y is the cube root of wind erosion. This variable is not of subject matter interest in itself but is used for illustrative purposes. There are 44 counties in Iowa for which wind erosion is reported. There were observations in all 44 counties in the study, but for purposes of this illustration we assume that there are four additional counties with no sample observations.

Wind erosion is a function of soil characteristics. The soils of Iowa have been mapped, so population values for a number of soil characteristics are available. The mean of the soil erodibility index for the county is used as the explanatory variable in our model. For our purposes, the sample of segments in a county is treated as a simple random sample. A preliminary analysis suggested that the assumption of a common population variance for the counties was reasonable. Therefore, we assume that the variance of the mean wind erosion for county g is $n_g^{-1}\sigma_e^2$, where n_g is the number of segments in county g and σ_e^2 is the common variance. We treat $\sigma_e^2 = 0.0971$ as known in our analysis. Thus, our model is

$$\bar{y}_g = \theta_g + e_g, \tag{5.5.10}$$

$$\theta_g = \bar{\mathbf{x}}_{gN}\boldsymbol{\beta} + u_g,$$

$$u_g \sim ii(0, \sigma_u^2),$$

$$e_g \sim ind(0, n_g^{-1}\sigma_e^2),$$

where u_g is independent of e_j for all g and j, $\bar{\mathbf{x}}_{gN} = [1, 0.1(\bar{r}_{1,g,N} - 59)]$, $\bar{r}_{1,g,N}$ is the population mean erodibility index for county g and $\bar{y}_g$ is the estimated mean wind erosion for county g. The erodibility index was reduced by 59 in the regression to facilitate the numerical discussion.

Using a program such as PROC Mixed of SAS, the estimated model parameters are

$$(\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_u^2) = \underset{(0.026)}{(0.770,} \quad \underset{(0.024)}{0.155,} \quad \underset{(0.0062)}{0.0226),}$$

where the estimates are based on the 44 counties with erosion values and the estimator of σ_u^2 is the maximum likelihood estimator.

The estimated parameters were used to construct the predictions of the erosion measure given in Table 5.13. For the first county in the table,

$$\begin{aligned} \hat{\theta}_3 &= \hat{\gamma}_3 \bar{y}_3 + (1 - \hat{\gamma}_3)\bar{\mathbf{x}}_{3N}\hat{\boldsymbol{\beta}}, \\ &= 0.466, \end{aligned}$$

where $\hat{\gamma}_3 = [0.0226 + (13)^{-1}0.0971]^{-1}(0.0226) = 0.7516$ and $\bar{\mathbf{x}}_{3N} = (1, -1.232)$. The standard errors of Table 5.13 were computed using (5.5.9) treating σ_e^2 as known. For county 3 the model standard error of 0.077 is about 89% of the design standard error of 0.086. County 167 has 44 observations, $\hat{\gamma}_{167} = 0.942$, and the model standard error is about 99% of the design standard error. The difference between the design standard errors and the prediction standard errors are modest because the σ_{eg}^2 are small relative to $\hat{\sigma}_u^2$.

Table 5.13: **Data on Iowa Wind Erosion**

County	N_g	n_g	Erodibility Index	$\bar{y}_g$	$\hat{\theta}_g$	S.E. of Prediction
3	1387	13	46.683	0.429	0.466	0.077
15	2462	18	58.569	0.665	0.684	0.067
21	2265	14	65.593	1.083	1.034	0.074
27	2479	19	47.727	0.788	0.753	0.066
33	2318	18	52.802	0.869	0.831	0.067
35	1748	12	72.130	1.125	1.085	0.079
41	2186	16	59.079	0.683	0.701	0.070
47	3048	19	49.757	0.408	0.448	0.066
59	1261	12	53.694	0.839	0.799	0.079
63	1822	15	63.563	0.754	0.773	0.072
67	1597	11	44.947	0.690	0.651	0.082
71	1345	15	56.807	0.927	0.885	0.072
73	1795	12	54.182	0.945	0.879	0.079
75	2369	13	40.951	0.619	0.587	0.077
77	2562	15	48.605	0.475	0.504	0.072
79	1899	11	74.981	0.790	0.854	0.082
83	2486	16	57.455	0.647	0.668	0.070
85	2241	19	66.700	0.727	0.757	0.066
91	2066	15	56.118	1.120	1.032	0.072

Continued

County	N_g	n_g	Erodibility Index	$\bar{y}_g$	$\hat{\theta}_g$	S.E. of Prediction
93	1385	10	61.830	0.677	0.718	0.084
109	2752	18	64.255	0.968	0.945	0.067
119	1753	29	61.605	0.703	0.717	0.055
129	1270	12	58.739	0.616	0.656	0.079
131	1232	10	48.739	0.422	0.478	0.084
133	2943	24	73.121	1.045	1.037	0.060
135	1190	15	45.417	0.363	0.407	0.072
141	1567	11	81.911	1.424	1.340	0.083
143	1511	10	56.229	0.975	0.900	0.084
145	1772	16	40.862	0.451	0.459	0.071
147	2716	17	60.811	0.945	0.915	0.069
149	3877	16	80.541	1.065	1.073	0.071
151	1823	10	63.190	0.918	0.893	0.084
153	1580	18	48.503	0.670	0.658	0.067
155	4405	21	62.348	0.619	0.653	0.063
157	2121	13	44.462	0.578	0.570	0.077
161	2423	16	66.551	0.719	0.754	0.070
165	2327	12	47.496	0.376	0.432	0.079
167	3180	44	72.262	0.954	0.956	0.045
169	1862	16	54.794	0.583	0.609	0.070
187	3011	15	58.420	0.874	0.849	0.072
189	1644	10	76.335	1.256	1.191	0.085
193	2319	17	75.142	0.905	0.928	0.069
195	1290	16	46.488	0.599	0.594	0.071
197	1754	11	71.380	0.577	0.685	0.082
201	1822		63.563		0.841	0.153
202	1511		56.229		0.727	0.153
203	3877		80.541		1.104	0.161
204	3011		58.420		0.761	0.153

The prediction for a county with no observations is $\bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$ and has a variance

$$V\{\hat{\theta}_g - \theta_g\} = V\{\bar{\mathbf{x}}_g\hat{\boldsymbol{\beta}}\} + \sigma_u^2,$$

estimated by

$$\hat{V}\{\hat{\theta}_g - \theta_g\} = \bar{\mathbf{x}}_{gN}\hat{V}\{\hat{\boldsymbol{\beta}}\}\bar{\mathbf{x}}'_{gN} + \hat{\sigma}_u^2,$$

where the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\hat{V}\{\hat{\boldsymbol{\beta}}\} = \begin{pmatrix} 6.65 & 0.04 \\ 0.04 & 5.77 \end{pmatrix} \times 10^{-4}.$$

Thus, the estimated prediction variance for county 201 is

$$\begin{aligned}\hat{V}\{\hat{\theta}_{201} - \theta_{201}\} &= (1, 0.456)\hat{V}\{\hat{\boldsymbol{\beta}}\}(1, 0.456)' + \hat{\sigma}_u^2 \\ &= 7.8894(10^{-4}) + 0.0226 = 0.0234.\end{aligned}$$

The estimated variance of u_g is the dominant term in the estimated prediction variance when there are no observations in the county. ■■

In many situations the standard error of the direct survey estimate for the overall total is judged to be acceptable, whereas those for the small areas are judged to be too large. In such situations the practitioner may prefer small area estimates that sum to a design consistent survey estimate of the total. That is, it is requested that

$$\sum_{g=1}^{M} N_g \hat{\theta}_g = \hat{T}_y, \tag{5.5.11}$$

where N_g is the number of elements in small area g, $\hat{\theta}_g$ is the small area predictor, and $\hat{T}_y$ is a design-consistent estimator of the total of y. If (5.5.11) is satisfied, the predictions are said to be *benchmarked* and the small area procedure becomes a method for allocating the design-consistent estimated total to the small areas. Two situations for benchmarking can be considered. In one the design-consistent estimator has been constructed using information not used for the small area estimation. Procedures appropriate for this situation have been reviewed by Wang, Fuller, and Qu (2009).

We consider benchmarking for the situation in which information to be used to construct the design consistent estimator is that used in the small area estimation. Under model (5.5.1), $\bar{x}_{gN}$ is known for all small areas and it follows that the population total

$$\mathbf{T}_x = \sum_{g=1}^{M} N_g \bar{\mathbf{x}}_{gN} \tag{5.5.12}$$

is known. Given the information available on $\mathbf{x}$, it is natural to use the regression estimator as an estimator for the total of y. If σ_{eg}^2 and σ_u^2 are known, the generalized least squares estimator (5.5.5) is the preferred estimator for $\boldsymbol{\beta}$, and a regression estimator of the total of y is

$$\hat{T}_{y,reg} = \mathbf{T}_x \hat{\boldsymbol{\beta}}, \tag{5.5.13}$$

where

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y},$$

$\mathbf{V} = \text{diag}(\sigma_u^2 + \sigma_{eg}^2)$, $\mathbf{X}$ is the $m \times k$ matrix with gth row equal to $\bar{\mathbf{x}}_{gN}$, $\mathbf{y}' = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_g, \ldots, \bar{y}_m)$, and the estimator of $\boldsymbol{\beta}$ of (5.5.13) is identically equal to the estimator $\hat{\boldsymbol{\beta}}$ of (5.5.5). We investigate the design consistency of estimator (5.5.13) permitting the number of small areas with a direct observation $\bar{y}_g$ to be less than M. Let π_g denote the probability that area g is observed and assume that $(\bar{y}_g, \bar{\mathbf{x}}_g)$ is design unbiased for $(\bar{y}_{gN}, \bar{\mathbf{x}}_{gN})$. Then a design-unbiased estimator of the vector of totals is

$$(\hat{T}_y, \hat{\mathbf{T}}_x) = \sum_{g \in A} N_g \pi_g^{-1} (\bar{y}_g, \bar{\mathbf{x}}_g),$$

It follows that the regression estimator (5.5.13) will be design consistent for the total of y if there is a vector $\mathbf{c}_1$ such that

$$\bar{\mathbf{x}}_{gN} \mathbf{c}_1 = \pi_g^{-1} N_g (\sigma_u^2 + \sigma_{eg}^2) \tag{5.5.14}$$

for all g. See Corollary 2.2.3.1.

For the weighted sum of the small area predictors to be equal to the regression estimator of the total, we require that

$$\begin{aligned} \hat{T}_{y,reg} &= \sum_{g=1}^{M} N_g \hat{\theta}_g \\ &= \sum_{g=1}^{M} N_g [\bar{\mathbf{x}}_{gN} \hat{\boldsymbol{\beta}} + \gamma_g (\bar{y}_g - \bar{\mathbf{x}}_{gN} \hat{\boldsymbol{\beta}})], \end{aligned} \tag{5.5.15}$$

where it is understood that the predictor is $\bar{\mathbf{x}}_{jN} \hat{\boldsymbol{\beta}}$ if area j is not observed. Because $\mathbf{T}_x = \sum_{g=1}^{M} N_g \bar{\mathbf{x}}_{gN}$, the requirement (5.5.15) becomes

$$\sum_{g=1}^{m} N_g \gamma_g (\bar{y}_g - \bar{\mathbf{x}}_{gN} \hat{\boldsymbol{\beta}}) = 0,$$

where m is the number of small areas observed, and the requirement (5.5.15) is satisfied for estimator (5.5.13) if there is a vector $\mathbf{c}_2$ such that

$$\bar{\mathbf{x}}_{gN} \mathbf{c}_2 = N_g \tag{5.5.16}$$

for all g. See Exercise 12. Thus, if N_g and $\pi_g^{-1} N_g (\sigma_u^2 + \sigma_{eg}^2)$ are in the column space of $\mathbf{X}$, the weighted sum of the small area predictors is equal to the design-consistent regression estimator of the total.

Typically, σ_{eg}^2 will not be known for the unobserved areas, but $\pi_g^{-1} N_g$ will be known. In some situations the rule defining n_g as a function of N_g

is known and one may be willing to assume that σ_{eg}^2 is of the form $n_g^{-1}\sigma_e^2$. Then $\pi_g^{-1}N_g\sigma_{eg}^2 = \pi_g^{-1}N_g n_g^{-1}\sigma_e^2$ can be treated as known for the unobserved areas.

To consider the case where σ_{eg}^2 is unknown, we adopt some of the notation of Section 2.3 and let $\mathbf{z}_g = (\bar{\mathbf{x}}_{0,gN}, \mathbf{z}_{dg})$, where

$$\mathbf{z}_d = \mathbf{x}_d - \mathbf{X}_0(\mathbf{X}_0'\mathbf{V}^{-1}\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{V}^{-1}\mathbf{x}_d, \tag{5.5.17}$$

$\mathbf{x}_d$ is the vector of observations on $x_{d,g} = \pi_g^{-1}N_g(\sigma_u^2 + \sigma_{eg}^2)$, $\mathbf{X}_0$ is the matrix of observations on the other explanatory variables, and $\mathbf{V}$ is defined in (5.5.13). The population mean of $\bar{\mathbf{x}}_0$ is known, but the population mean of z_d is unknown. Following the development of Section 2.3, we let

$$\hat{\boldsymbol{\alpha}} = (\hat{\boldsymbol{\alpha}}_0', \hat{\alpha}_d)' = (\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{y}, \tag{5.5.18}$$

where $\mathbf{Z} = (\mathbf{X}_0, \mathbf{z}_d)$. Then the regression estimator (2.3.22) is

$$\bar{y}_{reg} = (\bar{\mathbf{x}}_{0,\cdot,N}, \bar{z}_{d\pi})\hat{\boldsymbol{\alpha}} = \bar{y}_\pi + (\bar{\mathbf{x}}_{0,\cdot,N} - \bar{\mathbf{x}}_{0\pi})\hat{\boldsymbol{\alpha}}_0, \tag{5.5.19}$$

where

$$(\bar{y}_\pi, \bar{\mathbf{x}}_{0\pi}, \bar{z}_{d\pi}) = \left(\sum_{g\in A} N_g\pi_g^{-1}\right)^{-1} \sum_{g\in A} N_g\pi_g^{-1}(\bar{y}_g, \bar{\mathbf{x}}_{0g}, z_{dg})$$

and $\bar{\mathbf{x}}_{0,\cdot,N}$ is the population mean of $\mathbf{x}_0$. The small area predictors are

$$\begin{aligned}\hat{\theta}_g &= \bar{\mathbf{z}}_{gN}\hat{\boldsymbol{\alpha}} + \hat{\gamma}_g(\bar{y}_g - \bar{\mathbf{z}}_{gN}\hat{\boldsymbol{\alpha}}) \\ &= \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}} + \hat{\gamma}_g(\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}) \qquad \text{if } g \in A\end{aligned}$$

and

$$\hat{\theta}_g = (\bar{\mathbf{x}}_{0,g,N}, \bar{z}_{d\pi})\hat{\boldsymbol{\alpha}} \qquad \text{if } g \notin A. \tag{5.5.20}$$

When $\bar{y}_g$ is observed, the estimated MSE of $\hat{\theta}_g$ is of the form (5.5.9) and, in the current notation, is

$$\begin{aligned}\hat{V}\{\hat{\theta}_g - \theta_g\} &= \hat{\gamma}_g\hat{\sigma}_{eg}^2 + (1-\hat{\gamma}_g)^2\bar{\mathbf{z}}_{gN}\hat{V}\{\hat{\boldsymbol{\alpha}}\}\bar{\mathbf{z}}_{gN}' \\ &\quad + 2(\hat{\sigma}_u^2 + \hat{\sigma}_{eg}^2)\hat{V}\{\hat{\gamma}_g\}.\end{aligned}$$

If $\bar{y}_g$ is not observed, the estimation error is

$$\hat{\theta}_g - \theta_g = (\bar{\mathbf{x}}_{0,g,N}, \bar{z}_{d\pi})(\hat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}) - (\bar{z}_{d,g} - \bar{z}_{d\pi})\alpha_d - u_g$$

and an estimator of the MSE is

$$\hat{V}(\hat{\theta}_g - \theta_g) = (\bar{\mathbf{x}}_{0,g,N}, \bar{z}_{d\pi})\hat{V}\{\hat{\boldsymbol{\alpha}}\}(\bar{\mathbf{x}}_{0,g,N}, \bar{z}_{d\pi})' + \hat{\alpha}_d^2 s_{zd}^2 + \hat{\sigma}_u^2, \quad (5.5.21)$$

where

$$s_{zd}^2 = \left(\sum_{g\in A} \pi_g^{-1}\right)^{-1} \sum_{g\in A} \pi_g^{-1}(\bar{z}_{dg} - \bar{z}_{d\pi})^2.$$

Example 5.5.2. To illustrate the construction of small area estimates constrained to match the regression estimator, we use the data of Table 5.13. We assume, for the purposes of this example, that the sample of 44 counties is a simple random sample selected from a population of 48 counties. Then $\pi_g = 44/48$ for all counties. To satisfy (5.5.14) and (5.5.16), we add multiples of $\pi_g^{-1}N_g(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)$ and N_g to model (5.5.10), letting

$$\begin{aligned}\bar{\mathbf{x}}_{gN} &= (1, \bar{x}_{2,g,N}, \bar{x}_{3,g,N}, \bar{x}_{4,g,N}) \\ &=: [1, 0.1(\bar{r}_{1,g,N} - 59), 0.01N_g, 0.01N_g(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)].\end{aligned}$$

We iterate the estimation procedure redefining $\bar{\mathbf{x}}_{gN}$ at each step and using $\hat{\sigma}_u^2$ from the previous step until

$$(m-k)^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = 1.00, \quad (5.5.22)$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y},$$

$\hat{\mathbf{V}} = \text{diag}(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)$, the gth row of $\mathbf{X}$ is $\bar{\mathbf{x}}_{gN}$, and $m - k = 40$. The vector of estimates is

$$\begin{aligned}(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\sigma}_u^2) = (&0.800, \quad 0.160, -0.016, \quad 0.452, \quad 0.0256), \\ &(0.111)\ (0.026)\ \ (0.026)\ \ (0.897)\ (0.0066)\end{aligned} \quad (5.5.23)$$

where the standard errors for the elements of $\hat{\boldsymbol{\beta}}$ are the square roots of the diagonal elements of $(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}$. The variance of $\hat{\sigma}_u^2$ was estimated by

$$\begin{aligned}\hat{V}\{\hat{\sigma}_u^2\} &= 2(m-k)^{-1}\left(m^{-1}\sum_{g=1}^{m}(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)^{-1}\right)^{-2} \\ &= (20)^{-1}(33.8031)^{-2} = (0.0066)^2, \quad (5.5.24)\end{aligned}$$

obtained from a Taylor expansion of equation (5.5.2). See Exercise 16. The estimate of σ_u^2 differs from that of Example 5.5.1 because of the additional

explanatory variables and because the current estimator differs from the maximum likelihood estimator used in Example 5.5.1.

If the small areas are not selected with equal probability, the estimator of σ_u^2 should recognize this fact. See Pfeffermann et al. (1998).

Because $x_{d,g} = x_{4,g}$ is unknown for the unobserved counties, we construct the design-consistent regression estimator (5.5.19) for the mean of y. We define $z_4 = z_d$ to be the deviations from the weighted regression of x_4 on $(1, \bar{x}_{2,gN}, \bar{x}_{3,g,N})$. Then the vector of regression coefficients for the weighted regression of y on $(1, \bar{x}_{2,gN}, \bar{x}_{3,g,N}, z_{4g})$, the $\hat{\boldsymbol{\alpha}}$ of (5.5.18), is

$$\begin{array}{rcl} \hat{\boldsymbol{\alpha}}' & = & (0.831, \ \ 0.161, \ -0.003, \ \ 0.453). \\ & & \ (0.090) \ (0.026) \ \ (0.004) \ (0.897) \end{array} \tag{5.5.25}$$

A design-consistent estimator of the mean of z_d is

$$\begin{aligned} \bar{z}_{d\pi} &= \left(\sum_{g=1}^{44} N_g \right)^{-1} \sum_{g=1}^{44} N_g z_{dg} \\ &= -0.0013 \end{aligned}$$

and the regression estimator (5.5.19) of the population mean of y is

$$\begin{aligned} \bar{y}_{reg} &= (1, \ \bar{x}_{2,\cdot,N}, \ \bar{x}_{3,\cdot,N}, \ \bar{z}_{d\pi})\hat{\boldsymbol{\alpha}} \\ &= (1, \ 0.1601, \ 23.9906, \ -0.0013)\hat{\boldsymbol{\alpha}} \\ &= 0.7887. \end{aligned}$$

The regression estimated mean of 0.7887 corresponds to an estimated total of 81,441.28.

The county predictions of (5.5.20) are given in Table 5.14. The values predicted differ slightly from those of Table 5.13, primarily because of the difference in $\hat{\sigma}_u^2$. The standard errors for the predictions for counties with a y observation are computed using (5.5.9), and the standard errors for counties with no y observation are computed using (5.5.21).

The term $\hat{\alpha}_d^2 s_{zd}^2 = (0.453)^2(0.00099)$ adds little to the variance of predictions for areas with no y observations because s_{zd}^2 is small. The standard errors of Table 5.14 are sometimes slightly larger than those of Table 5.13 because more parameters are being estimated and the estimate of σ_u^2 is larger than that used to construct Table 5.13.

Table 5.14: **Predictions with Sum Constrained to Regression Estimator**

County	x_{4g}	$\bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$	$\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$	$\hat{\theta}_g$	S.E. of Prediction
3	0.4084	0.5906	–0.1615	0.466	0.077
15	0.6739	0.7481	–0.0831	0.679	0.067
21	0.6549	0.8799	0.2034	1.040	0.074
27	0.6716	0.5711	0.2173	0.752	0.066
33	0.6345	0.6584	0.2105	0.832	0.067
35	0.5256	0.9995	0.1250	1.094	0.079
41	0.6131	0.7679	–0.0854	0.699	0.070
47	0.8257	0.5924	–0.1842	0.439	0.066
59	0.3792	0.7075	0.1312	0.807	0.079
63	0.5184	0.8486	–0.0950	0.773	0.072
67	0.4920	0.5708	0.1189	0.659	0.082
71	0.3827	0.7469	0.1802	0.891	0.072
73	0.5398	0.7120	0.2327	0.889	0.079
75	0.6976	0.4901	0.1287	0.590	0.077
77	0.7290	0.5993	–0.1246	0.500	0.073
79	0.5850	1.0505	–0.2605	0.857	0.083
83	0.6973	0.7374	–0.0902	0.664	0.070
85	0.6071	0.8793	–0.1525	0.752	0.066
91	0.5878	0.7262	0.3936	1.040	0.072
93	0.4389	0.8471	–0.1703	0.724	0.085
109	0.7533	0.8337	0.1339	0.944	0.067
119	0.4440	0.7934	–0.0907	0.713	0.055
129	0.3819	0.7881	–0.1719	0.658	0.079
131	0.3904	0.6374	–0.2157	0.481	0.085
133	0.7659	0.9541	0.0912	1.033	0.060
135	0.3386	0.5667	–0.2033	0.404	0.073
141	0.4827	1.1623	0.2614	1.357	0.083
143	0.4788	0.7576	0.2175	0.913	0.085
145	0.4970	0.4828	–0.0314	0.457	0.071
147	0.7521	0.7834	0.1619	0.915	0.069
149	1.0874	1.0855	–0.0208	1.069	0.072
151	0.5777	0.8694	0.0488	0.905	0.085
153	0.4325	0.6032	0.0670	0.658	0.067
155	1.1719	0.7575	–0.1386	0.640	0.064
157	0.6246	0.5485	0.0294	0.571	0.077

Continued

County	x_{4g}	$\bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$	$\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}$	$\hat{\theta}_g$	S.E. of Prediction
161	0.6796	0.8839	–0.1653	0.750	0.070
165	0.6997	0.6018	–0.2261	0.430	0.080
167	0.7691	0.9081	0.0462	0.951	0.046
169	0.5223	0.7043	–0.1215	0.606	0.070
187	0.8567	0.7503	0.1239	0.849	0.073
189	0.5210	1.0795	0.1767	1.208	0.086
193	0.6422	1.0192	–0.1143	0.926	0.069
195	0.3618	0.5802	0.0189	0.595	0.071
197	0.5403	0.9933	–0.4164	0.684	0.082
201	0.5282	0.8528	NA	0.853	0.164
202	0.4504	0.7434	NA	0.743	0.165
203	1.0408	1.0683	NA	1.068	0.180
204	0.8240	0.7362	NA	0.736	0.167

Because

$$\sum_{g\in A} N_g\hat{\gamma}_g(\bar{y}_g - \bar{\mathbf{x}}_{gN}\hat{\boldsymbol{\beta}}) = 0,$$

the weighted sum of the predicted values for the 48 counties in Table 5.14 is

$$\sum_{g=1}^{48} N_g\hat{\theta}_g = 81,441.28,$$

equal to the regression estimator of the total. The sum of the predictions in Table 5.13 is $81,605$.

In this example wind erosion has a small correlation with the sampling weight, so the weighted sum of predictions constructed with N_g and $N_g(\hat{\sigma}_u^2 + n_g^{-1}\sigma_e^2)$ included in the set of explanatory variables differs little from the weighted sum constructed with only $(1, \bar{x}_{1,g,N})$ as the explanatory vector. ■■

5.6 MEASUREMENT ERROR

5.6.1 Introduction

Essentially all data are collected subject to measurement error, and the design of collection instruments to minimize measurement error is an important part of the discipline of survey sampling. The mean and variance of the measurement process are both important. The mean is typically the most

difficult to evaluate, because some determination of the "truth" is required. In survey sampling, external sources, perhaps available at a later time, can sometimes be used to estimate bias. If a reliable external source is available, the collection instrument can be recalibrated.

The variance of the measuring operation can sometimes be estimated by repeated independent determinations on the same element. One measure of the relative magnitude of the variance of measurement error is the correlation between two independent determinations on a random sample. This measure is called the *attenuation coefficient* and is denoted by κ_{xx} for the variable x. We adopt the convention of using a lowercase letter to denote the true value and a capital letter to denote the observed value, where the observed value is the sum of the true value and the measurement error. The κ_{xx} gives the relative bias in the simple regression coefficient of y on X as an estimator of the population regression of y on x. For continuous variables, κ_{xx} is the ratio of the variance of the true x to the variance of observed X.

Fuller (1987b, p. 8) reports on a large study conducted by the U.S. Census Bureau in which determinations were made on a number of demographic variables in the Decennial Census and in the Current Population Survey. The two surveys gave two nearly independent determinations. The attenuation coefficient for education was 0.88, that for income was 0.85, and that for fraction unemployed was 0.77. Thus, for example, of the variation observed in a simple random sample of incomes, 15% is due to measurement error.

5.6.2 Simple estimators

Consider a simple measurement error model in which the observation is the true value plus a zero-mean measurement error. Let X_i be the observed value, and x_i be the true value, so that

$$\begin{aligned} X_i &= x_i + u_i, \\ u_i &\sim (0, \sigma_u^2), \end{aligned} \tag{5.6.1}$$

where u_i is the measurement error. Assume that u_i is independent of x_j for all i and j. Assume that a sample of size n is selected from a finite population of x values and that the measurement process satisfies (5.6.1). Let the Horvitz–Thompson estimator of the total be constructed as

$$\hat{T}_X = \sum_{i \in A} w_i X_i, \tag{5.6.2}$$

where $w_i = \pi_i^{-1}$. Because the estimator based on the true values,

$$\hat{T}_x = \sum_{i \in A} w_i x_i,$$

is unbiased for T_x, $\hat{T}_X$ is also unbiased for T_x. That is,

$$E\{E(\hat{T}_X \mid \mathcal{F}_X) \mid \mathcal{F}_x\} = E\{T_X \mid \mathcal{F}_x\} = T_x,$$

where $\mathcal{F}_X = (X_1, X_2, \ldots, X_N)$ and $\mathcal{F}_x = (x_1, x_2, \ldots, x_N)$ is the set of true x values.

Given that x_i is independent of u_j for all i and j.

$$V\{\hat{T}_X - T_x \mid \mathcal{F}_x\} = V\{\hat{T}_x - T_x \mid \mathcal{F}_x\} + V\left\{\sum_{i\in A} w_i u_i\right\}, \tag{5.6.3}$$

where $\hat{T}_x = \Sigma_{i\in A} x_i$. The variance of $\hat{T}_u = \Sigma_{i\in A} w_i u_i$ is a function of the covariance structure of $\mathbf{u} = (u_i, u_2, \ldots, u_n)$, and if $u_i \sim ind(0, \sigma_u^2)$, independent of (x_j, w_j) for all i and j, then

$$\begin{aligned} V\{\hat{T}_X - T_x \mid \mathcal{F}_x\} &= V\{\hat{T}_x - T_x \mid \mathcal{F}_x\} + E\left\{\sum_{i\in A} w_i^2 \sigma_u^2\right\} \\ &= V\{\hat{T}_x - T_x \mid \mathcal{F}_x\} + \sum_{i\in U} w_i \sigma_u^2. \end{aligned} \tag{5.6.4}$$

Furthermore,

$$\begin{aligned} E\{\hat{V}_{HT}(\hat{T}_X - T_X \mid \mathcal{F}_X) \mid \mathcal{F}_x\} &= V\{\hat{T}_x - T_x \mid \mathcal{F}_x\} \\ &\quad + E\left\{\sum_{k\in A}\sum_{j\in A} \pi_{jk}^{-1}(\pi_{jk} - \pi_j\pi_k) w_j u_j w_k u_k\right\} \\ &= V\{\hat{T}_x - T_x\} + \sum_{i\in U}(1 - \pi_i) w_i \sigma_u^2, \end{aligned} \tag{5.6.5}$$

where

$$\hat{V}_{HT}\left(\hat{T}_X - T_X \mid \mathcal{F}_X\right) = \sum_{k\in A}\sum_{j\in A} \pi_{jk}^{-1}(\pi_{jk} - \pi_j\pi_k) w_j X_j w_k X_k.$$

It follows that

$$E\{\hat{V}_{HT}(\hat{T}_X - T_X \mid \mathcal{F}_X) \mid \mathcal{F}_x\} - V\{\hat{T}_X - T_x \mid \mathcal{F}_x\} = -N\sigma_u^2. \tag{5.6.6}$$

Thus, if the measurement errors have zero means, are independent of x and w, and are independent, the expectation of a design linear estimator in the observed values is equal to the expectation of the estimator in the true variables.

Furthermore, the bias in the Horvitz–Thompson estimator of variance is small if the sampling rates are small. See Exercise 6.

The assumption of independent measurement errors is critical for result (5.6.6) and the result does not hold with personal interviews, where each interviewer collects data from several respondents. See Hansen et al. (1951). The traits of the interviewer lead to correlation among responses obtained by that interviewer. To illustrate the effect of correlated measurement error, assume that a simple random sample of size $n = mk$ is selected and each of k interviewers is given an assignment of m interviews, where assignment is at random. Assume that interviewers have an effect on the responses and that it is reasonable to treat the k interviewers as a random sample from the population of interviewers. With these assumptions a representation for the observations is

$$Y_{gj} = \mu + e_j + \alpha_g + \epsilon_{gj}, \tag{5.6.7}$$

where μ is the superpopulation mean, $e_j = y_j - \mu$, y_j is the true value, α_g is the interviewer effect for interviewer g, and ϵ_{gj} is the measurement error for person j interviewed by interviewer g. We consider the relatively simple specification

$$\alpha_g \sim ii(0, \sigma_\alpha^2),$$

$$\epsilon_{gj} \sim ii(0, \sigma_\epsilon^2),$$

and assume α_g, ϵ_{rj}, and e_i to be independent for all g, r, j, and i. Then

$$\begin{aligned} V\{\bar{Y}_{..} - \bar{y}_N \mid \mathcal{F}_y\} &= V\{\bar{e}_{.} + \bar{\alpha}_{.} + \bar{\epsilon}_{..} \mid \mathcal{F}_y\} \\ &= (1 - f_n)n^{-1}S_e^2 + k^{-1}\sigma_\alpha^2 + n^{-1}\sigma_\epsilon^2, \end{aligned} \tag{5.6.8}$$

where

$$\begin{aligned} (\bar{Y}_{..}, \bar{e}_{.}, \bar{\epsilon}_{..}) &= n^{-1} \sum_{gj \in A} (Y_{gj}, e_j, \epsilon_{gj}), \\ \bar{\alpha}_{.} &= k^{-1} \sum_{g=1}^{k} \alpha_g, \end{aligned}$$

and $f_n = N^{-1}n$. Because there is one observation per person, the summation over gj is a summation over persons. For large interviewer assignments the term $k^{-1}\sigma_\alpha^2$ can be very important even when σ_α^2 is relatively small.

The usual estimator of variance is seriously biased. For a simple random sample,

$$E\{s_Y^2\} = \sigma_e^2 + \sigma_\epsilon^2 + n(n-1)^{-1}m^{-1}(m-1)\sigma_\alpha^2, \tag{5.6.9}$$

where

$$s_Y^2 = (n-1)^{-1} \sum_{gj \in A} (Y_{gj} - \bar{Y}_{..})^2$$

and $\sigma_e^2 = E\{S_e^2\}$.

Example 5.6.1. Let model (5.6.7) hold and assume that σ_α^2 is 2% of σ_e^2 and that σ_ϵ^2 is 15% of σ_e^2. Let a simple random sample of 1000 be selected, and let each of 20 interviewers be given a random assignment of 50 interviews. If the finite population correction can be ignored, the variance of the sample mean is

$$V\{\bar{Y}_{..} - \bar{y}_N\} = n^{-1}(\sigma_e^2 + 0.15\sigma_e^2) + 0.02k^{-1}\sigma_e^2 = 0.00215\sigma_e^2.$$

Although the variance of the interviewer effect is small, it makes a large contribution to the variance because each interviewer has a large number of interviews. By (5.6.9)

$$E\{s_Y^2\} = 1.1696\sigma_e^2$$

and the usual estimator of variance of $\bar{Y}_n$ has a bias of -45.6%.

By treating interviewer assignments as the first stage of a two-stage sample, it is possible to construct an unbiased estimator of the variance of $\bar{Y}_{..}$. Under the assumptions that the interviewer assignments are made at random and that the finite population correction can be ignored, an unbiased estimator of $V\{\bar{Y}_{..} - \bar{y}_N\}$ is

$$\hat{V}\{\bar{Y}_{..} - \bar{y}_N\} = (380)^{-1} \sum_{g=1}^{20} (\bar{Y}_{g.} - \bar{Y}_{..})^2,$$

where $\bar{Y}_{g.}$ is the mean for the gth interviewer. ■■

Many large-scale surveys are stratified multistage samples. In such surveys the interviewer assignments are often primary sampling units. If an interviewer is not assigned to more than one primary sampling unit, and if the finite population correction can be ignored, the usual variance estimator for a design linear estimator remains appropriate.

Situations that seem simple at first can be quite difficult in the presence of measurement error. The estimation of the distribution function is an example. Assume that

$$Y_i = y_i + e_i,$$
$$y_i \sim (\mu, \sigma_y^2),$$

where $e_i \sim NI(0, \sigma_e^2)$, independent of y_i. Then the mean of Y_i for a simple random sample is an unbiased estimator of μ, but the estimator

$$\hat{F}(y_o) = n^{-1} \sum_{i \in A} \delta_i(y_o),$$

where

$$\begin{aligned} \delta_i(y_o) &= 1 \qquad \text{if } Y_i \leq y_o \\ &= 0 \qquad \text{otherwise,} \end{aligned}$$

is, in general, biased for the probability that $y_i < y_o$. The mean of e_i is zero for y_i, but the mean of the measurement error for $\delta_i(y_o)$ is not zero.

If $e_i \sim NI(0, \sigma_e^2)$ and if $y_i \sim N(\mu, \sigma_y^2)$ then $Y_i \sim N(\mu, \sigma_y^2 + \sigma_e^2)$ and the parameters of the distribution function of y_i are easily estimated. Similarly, one can use likelihood methods to estimate the parameters of the distribution if one can specify the form of the distribution of y_i and the form of the error distribution.

Nonparametric or semiparametric estimation of the distribution function in the presence of measurement error is extremely difficult. See Stefanski and Carroll (1990), Cook and Stefanski (1994), Nusser et al. (1996), Cordy and Thomas (1997), Chen, Fuller, and Breidt (2002), and Delaigle, Hall, and Meister (2008). If σ_e^2 is known, the variable

$$z_i = \bar{Y} + [\hat{\sigma}_Y^{-2}(\hat{\sigma}_Y^2 - \sigma_e^2)]^{-1/2}(Y_i - \bar{Y})$$

has sample mean and variance equal to estimators of the mean and variance of y, where those estimators are $(\hat{\mu}_y, \hat{\sigma}_y^2) = (\bar{Y}, \hat{\sigma}_Y^2 - \sigma_e^2)$ and $(\bar{Y}, \hat{\sigma}_Y^2)$ is an estimator of the mean and variance of Y. Therefore, the sample distribution function of z is a first approximation to the distribution function of y that can be used to suggest parametric models for the distribution of y_i.

5.6.3 Complex estimators

As demonstrated in the preceding section, measurement error with zero mean increases the variance of linear estimators, but the estimators remain unbiased. Alternatively, the expectation of nonlinear estimators, such as regression coefficients, can be seriously affected by zero-mean measurement error. Consider the simple regression model,

$$y_i = \beta_0 + x_{1i}\beta_1 + e_i, \tag{5.6.10}$$

where $e_i \sim (0, \sigma_e^2)$ independent of x_{1i}. Assume that the observation on the explanatory variable is $X_{1i} = x_{1i} + u_i$, where $u_i \sim ind(0, \sigma_u^2)$ is the

measurement error as specified in (5.6.1). Consider the finite population to be a simple random sample from an infinite population where (y_i, x_{1i}) satisfies (5.6.10). Given a probability sample, the weighted estimator

$$\hat{\gamma} = (\mathbf{X}'\mathbf{D}_\pi^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_\pi^{-1}y, \tag{5.6.11}$$

where $\mathbf{D}_\pi = \text{diag}(\pi_1, \pi_2, \ldots, \pi_n)$, $\mathbf{X}' = (\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_n')$, and $\mathbf{X}_i = (1, X_{1i})$, was discussed in Chapter 2. We call the estimator $\hat{\gamma}$ because, in the presence of measurement error, we shall see that $\hat{\gamma}$ is a biased estimator of $\boldsymbol{\beta}$. The estimator of the coefficient for X_{1i} can be written

$$\hat{\gamma}_1 = \left(\sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})^2\pi_i^{-1}\right)^{-1} \sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})\pi_i^{-1}(y_i - \bar{y}_\pi). \tag{5.6.12}$$

Under the assumption that u_i is independent of (x_{1i}, π_i, e_i), and under the usual assumptions required for consistency of a sample mean,

$$E\left\{N^{-1}\sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})^2\pi_i^{-1} \mid \mathcal{F}_{(x,y),N}\right\} = S_{x,N}^2 + \sigma_u^2 + O_p(n^{-1})$$

and

$$E\left\{N^{-1}\sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})(y_i - \bar{y}_\pi)\pi_i^{-1} \mid \mathcal{F}_{(x,y),N}\right\} = S_{xy,N} + O_p(n^{-1}),$$

where $\mathcal{F}_{(x,y),N} = [(y_1, x_{1,1}), (y_2, x_{1,2}), \ldots, (y_N, x_{1,N})]$ is the finite population of values for the true x_1 and y. Then

$$\hat{\gamma}_1 = (S_{x,N}^2 + \sigma_u^2)^{-1}S_{xy,N} + O_p(n^{-1/2}) \tag{5.6.13}$$

and $\hat{\gamma}_1$ is a consistent estimator of $(\sigma_x^2 + \sigma_u^2)^{-1}\sigma_{xy} = \kappa_{xx}\beta_1$.

Consistent estimation of β_1 requires additional information beyond the set of (y_i, X_i) vectors. There are several forms for such information. If we know κ_{xx} or have an estimator of κ_{xx}, then

$$\tilde{\beta}_{1,\kappa} = \kappa_{xx}^{-1}\hat{\gamma}_1 \tag{5.6.14}$$

is a consistent estimator of β_1. For example, one could use the estimate of κ_{xx} from the U.S. census study if the explanatory variable is education. Similarly, if σ_u^2 is known, or estimated,

$$\tilde{\beta}_{1,\sigma} = \left(\sum_{i\in A}[(X_{1i} - \bar{X}_{1\pi})^2 - \sigma_u^2\,]\pi_i^{-1}\right)^{-1} \sum_{i\in A}(X_{1i} - \bar{X}_{1\pi})\pi_i^{-1}(y_i - \bar{y}_\pi) \tag{5.6.15}$$

is a consistent estimator of β_1. The matrix expression for the estimator of $\boldsymbol{\beta}$ is

$$\tilde{\boldsymbol{\beta}}_\sigma = (\mathbf{M}_{X\pi X} - \Sigma_{uu})^{-1}\mathbf{M}_{X\pi y},$$

where $\Sigma_{uu} = \text{diag}(0, \sigma_u^2)$, $\mathbf{X}_i = (1, X_{1i})$, and

$$(\mathbf{M}_{X\pi X}, \mathbf{M}_{X\pi y}) = N^{-1}\sum_{i\in A}\mathbf{X}_i'\pi_i^{-1}(\mathbf{X}_i, y_i).$$

Our usual Taylor approximation gives

$$\hat{V}\{\tilde{\boldsymbol{\beta}}_\sigma\} = (\mathbf{M}_{X\pi X} - \Sigma_{uu})^{-1}\hat{V}_{HT}\{\bar{\mathbf{b}}_{HT}\}(\mathbf{M}_{X\pi X} - \Sigma_{uu})^{-1},$$

where $\mathbf{b}_i = \mathbf{x}_i' a_i$,

$$\bar{\mathbf{b}}_{HT} = N^{-1}\sum_{i\in A}\pi_i^{-1}\mathbf{x}_i' a_i$$

and

$$\hat{V}_{HT}\{\bar{\mathbf{b}}_{HT}\} = \hat{V}_{HT}\left\{\left(\sum_{i\in A}\pi_i^{-1}\right)^{-1}\sum_{i\in A}\mathbf{X}_i'\pi_i^{-1}a_i\right\}$$

is computed using $\hat{a}_i = y_i - \bar{y}_{HT} - (X_{1i} - \bar{X}_{1HT})\hat{\boldsymbol{\beta}}_\sigma$. In most situations one is interested in $\boldsymbol{\beta}$ as a superpopulation parameter. If so, finite population corrections are not appropriate. See Chapter 6.

Establishing the relationship between two error-prone measures is an important application of measurement error models. In some situations there is a relatively inexpensive procedure appropriate for large-scale data collection and an expensive procedure believed to be unbiased for the characteristic of interest. Then a subsample may be used to study the relationship between the two measures. Similarly, if one measuring procedure is to be replaced by another, it is important to establish the relationship between the two measures. Example 5.6.2 is an illustration.

Example 5.6.2. The National Resources Inventory is described in Example 1.2.2. In 2004, the Natural Resources Conservation Service changed the way in which data were collected for the segments. In 2003 and in prior years, the data collectors outlined, on a transparent overlay placed on an aerial photograph, the areas designated as developed. Developed land includes urban areas, built-up areas, and roads. Beginning in 2004 a digital process was used in which roads and certain types of developed land, such as manufacturing plants and cemeteries, were outlined digitally, but for single- or double-unit residences, the location of the residence was entered as a simple "dot" location. A computer program was designed to convert the digital information to area information. To calibrate the computer program, a study was conducted

in which two data collectors, using the new procedure, made independent determinations on segments that had been observed in 2003. The determinations were treated as independent because only unmodified photographs were available to the data collector for each determination. See Yu and Legg (2009).

We analyze data collected in the calibration study for the western part of the United States. The data are observations where at least one of the three determinations is not zero. The computer program has parameters that can be changed to improve the agreement between the old and new procedures. Our analysis can be considered to be a check on parameters determined on a different data set.

Let (Y_{1i}, Y_{2i}, X_i) be the vector composed of the first determination by the new procedure, the second determination by the new procedure, and the determination by the old procedure, respectively. In all cases, the variable is the fraction of segment acres that are developed. The old procedure is assumed to be unbiased for the quantity of interest. Our analysis model is

$$\begin{aligned} y_{ji} &= \beta_0 + \beta_1 x_i, \\ (Y_{1i}, Y_{2i}, X_i) &= (y_i, y_i, x_i) + (e_{1i}, e_{2i}, u_i), \\ x_i &\sim ind(0, \sigma_x^2), \\ e_{ji} &\sim ind(0, \sigma_{ei}^2), \\ u_i &\sim ind(0, \sigma_{ui}^2), \end{aligned} \tag{5.6.16}$$

for $j = 1,\ 2$, and it is assumed that u_i, e_{jt}, and x_i are mutually independent. It seems reasonable that the measurement error has smaller variance for segments with a small fraction of developed land than for segments with a fraction near 50%. One could specify a model for the error variance, but we estimate the model estimating the average variance of u_i, denoted by $\sigma_{a,u}^2$, and the average variance of e_i, denoted by $\sigma_{a,e}^2$.

To estimate the parameters, it is convenient to define the vector

$$\mathbf{Z}_i = (Z_{1i}, Z_{2i}, Z_{3i}) = [X_i,\ 0.5(Y_{1i} + Y_{2i}),\ (0.5)^{0.5}(Y_{1i} - Y_{2i})] \tag{5.6.17}$$

and let

$$\mathbf{m} = (n-1)^{-1} \sum_{i \in A} (\mathbf{Z}_i - \bar{\mathbf{Z}})'(\mathbf{Z}_i - \bar{\mathbf{Z}}).$$

Then

$$E\{\mathbf{m}\} = \begin{bmatrix} \sigma_x^2 + \sigma_{a,u}^2 & \beta_1 \sigma_x^2 & 0 \\ \beta_1 \sigma_x^2 & \beta_1^2 \sigma_x^2 + 0.5\sigma_{a,e}^2 & 0 \\ 0 & 0 & \sigma_{a,e}^2 \end{bmatrix}. \tag{5.6.18}$$

If the elements of $\mathbf{Z}_i$ are normally distributed, m_{13} and m_{23} contain no information about the parameters because $E\{m_{13}, m_{23}\} = \mathbf{0}$ and the covariance between (m_{13}, m_{23}) and the other elements of $\mathbf{m}$ is the zero vector. If the distribution is not normal, it is possible that (m_{13}, m_{23}) contains information, but we ignore that possibility and work only with $(m_{11}, m_{12}, m_{22}, m_{33})$. By equating the sample moments to their expectations, we obtain the estimators

$$\begin{aligned}
\hat{\beta}_0 &= \bar{Z}_1 - \hat{\beta}_1 \bar{Z}_2, \\
\hat{\beta}_1 &= m_{12}^{-1}(m_{22} - 0.5m_{33}) = m_{12}^{-1}\hat{\sigma}_y^2, \\
\hat{\sigma}_x^2 &= (m_{22} - 0.5m_{33})^{-1} m_{12}^2 = \hat{\sigma}_y^{-2} m_{12}^2, \\
\hat{\sigma}_{a,e}^2 &= m_{33}, \\
\hat{\sigma}_{a,u}^2 &= m_{11} - (m_{22} - 0.5m_{33})^{-1} m_{12}^2 = m_{11} - \hat{\sigma}_x^2, \quad (5.6.19)
\end{aligned}$$

where $\hat{\sigma}_y^2 = m_{22} - 0.5m_{33}$.

For our sample of 382 segments,

$$\begin{aligned}
(\bar{Z}_1, \bar{Z}_2) &= \begin{matrix} (0.1433, & 0.1464), \\ (0.0073) & (0.0070) \end{matrix} \\
(m_{11}, m_{12}, m_{22}, m_{33}) &= \begin{matrix} (2.061, & 1.758, & 1.873, & 0.060) \times 10^{-2}, \\ (0.184) & (0.129) & (0.120) & (0.014) \end{matrix} \\
(\hat{\beta}_0, \hat{\beta}_1) &= \begin{matrix} (-0.0039, & 1.0482), \\ (0.0037) & (0.0374) \end{matrix}
\end{aligned}$$

and

$$\begin{aligned}
100(\hat{\sigma}_{a,e}^2, \hat{\sigma}_{a,u}^2, \hat{\sigma}_x^2) &= \begin{matrix} (0.0601, & 0.3837, & 1.6775), \\ (0.0139) & (0.0546) & (0.1583) \end{matrix}
\end{aligned}$$

where the standard errors were calculated with 382 delete-one jackknife replicates. The jackknife is appropriate under model (5.6.16) because all estimators are continuous differentiable functions of the moments. Also see Exercise 15.

The calibration sample was selected to have a much higher fraction of segments with developed land than the general population of segments. Because our analysis is conducted on unweighted data, $(\hat{\mu}_x, \hat{\sigma}_x^2)$ is not an estimate for the general population.

The estimated variance of $\hat{\sigma}_{a,e}^2$ is much larger than one would expect if $e_i \sim N(0, \sigma_e^2)$. There are two reasons: (1) the e_i have unequal variances, and (2) the distribution of the measurement errors has long tails.

The approximate F test for $H_0 : (\beta_0, \beta_1) = (0, 1)$ is

$$F(2, 380) = 0.5(\hat{\beta}_0, \hat{\beta}_1 - 1)[\hat{V}\{\hat{\beta}_0, \hat{\beta}_1\}]^{-1}(\hat{\beta}_0, \hat{\beta}_1 - 1)' = 0.83,$$

where the estimated covariance matrix for $(\hat{\beta}_0, \hat{\beta}_1)$ is

$$\hat{V}\{(\hat{\beta}_0, \hat{\beta}_1)\} = \begin{pmatrix} 0.1397 & -1.0746 \\ -1.0746 & 14.0068 \end{pmatrix} \times 10^{-4}.$$

Also, plots of the data give little indication of a nonlinear relationship between the old and new procedures. Therefore, the data are consistent with the hypothesis that the new measuring procedure produces values that are equal to the true value plus a zero-mean measurement error. ■■

5.7 REFERENCES

Section 5.1. Chang and Kott (2008), Fuller and An (1998), Fuller, Loughin, and Baker (1994), Kalton and Kasprzyk (1986), Kalton and Kish (1984), Kott (2006b), Little (1983b, 1988), Little and Rubin (2002), Meng (1994), Sande (1983), Särndal (1992).

Section 5.2. Chen and Shao (2000, 2001), Fay (1996, 1999), Fuller and Kim (2005), Kim and Fuller (2004), Kim, Fuller, and Bell (2009), Rancourt, Särndal, and Lee (1994), Rao and Shao (1992), Särndal (1992).

Section 5.3. Cressie (1991), Wolter (2007).

Section 5.4. Fuller (1991), Hidiroglou and Srinath (1981), Rivest (1994).

Section 5.5. Battese, Harter, and Fuller (1988), Fay and Herriot (1979), Ghosh and Rao (1994), Harville (1976), Kackar and Harville (1984), Mukhopadhyay (2006), Pfeffermann and Barnard (1991), Prasad and Rao (1990, 1999), Rao (2003), Robinson (1991), Wang and Fuller (2003), Wang, Fuller, and Qu (2009), You and Rao (2002).

Section 5.6. Biemer et al. (1991), Carroll, Ruppert, and Stephanski (1995), Fuller (1987b, 1991b, 1995). Hansen et al. (1951), Hansen, Hurwitz, and Pritzker (1964), Yu and Legg (2009).

5.8 EXERCISES

1. (a) (Section 5.2) Let a simple random sample of size n have m missing values and r respondents. Assume that response is independent of

y. Let a simple random replacement sample of M of the respondents be used as a set of donors for each missing value. Each donated value is given a weight of $n^{-1}M^{-1}$. Show that the variance of the imputed mean is

$$V\{\bar{y}_I \mid (\mathcal{F}, m)\} = (r^{-1} - N^{-1})S_y^2 + n^{-2}M^{-1}mr^{-1}(r-1)S_y^2.$$

(b) Assume that m values are missing and r are present in a simple random sample of size n and assume that the probability of response is independent of y. The data set is completed by imputing a single value for each missing value. Let $m = kr + t$, where k is the largest nonnegative integer such that $kr \leq m$. Consider a hot deck procedure in which $r - t$ respondents are used as donors k times, and t respondents are used as donors $k+1$ times. In human nonresponse, k is generally zero. The t respondents are chosen randomly from the r. Show that the variance of the imputed mean is

$$V\{\bar{y}_I \mid \mathcal{F}, m\} = \left(r^{-1} - N^{-1}\right) S_y^2 + n^{-2}tr^{-1}\left(r - t\right) S_y^2.$$

2. (Section 5.2) Using the data of Table 5.2, compute the estimated mean of y for each of the three x-categories. Using the replicates of Table 5.5, estimate the variance of your estimates.

3. (Section 5.2) Using the replicates of Table 5.7, estimate the variance of the estimated mean of x, where x is as given in Table 5.6. Is this a design-unbiased estimator?

4. (Section 5.2) Using the imputed data of Table 5.7, compute the weighted regression for y on x using the weights in the table. Using the replicates of Table 5.9, compute the estimated covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1)$, where $(\hat{\beta}_0, \hat{\beta}_1)$ is the vector of estimated coefficients and $\hat{y}_t = \hat{\beta}_0 + x_t\hat{\beta}_1$. In this simple example, do you think the regression coefficient based on the imputed estimator is a good estimator? Would your answer change if we had a sample of 100 observations with 30 y-values missing?

5. (Section 5.3) Assume that a one-per-stratum sample is selected from a population with stratum sizes N_h, $h = 1, 2, \ldots, H$. Let $\pi_{h,i}$, $i = 1, 2, \ldots, N$, be the selection probabilities, where the subscript h is redundant but useful. Assume that H is even and that the strata are collapsed to form $H/2$ strata. Let $y_{h,i}$ be the sample observation in the hth original stratum and let

$$\hat{V}\{\hat{T} \mid \mathcal{F}\} = 0.5 \sum_{h=1}^{H/2} (\pi_{2h,i}^{-1} y_{2h,i} - \pi_{2h-1,i}^{-1} y_{2h-1,i})^2,$$

where

$$\hat{T} = \sum_{h=1}^{H} \pi_{h,i}^{-1} y_{h,i}.$$

Assume that the finite population is a sample from a superpopulation with

$$y_{h,i} \sim ind(\mu_h,\ \sigma_h^2) \text{ for } (h,i) \in U_h.$$

What is the expected value of $\hat{V}\{\hat{T} \mid \mathcal{F}\}$?

6. (Section 5.6) Assume that it is known that the measurement error variance for a variable x is σ_ϵ^2. Let a stratified sample be selected, where the observations are $X_{hj} = x_{hj} + \epsilon_{hj}$. Let the usual estimator of variance of the estimated total be

$$\hat{V}\{\hat{T}_X \mid \mathcal{F}\} = \sum_{h=1}^{H} N_h^2 (1 - f_h) n_h^{-1} s_h^2,$$

where $f_h = N_h^{-1} N_h$, s_h^2 is the sample stratum variance of X, and N_h is the stratum size. Assume the model (5.6.1) holds for each h. Show that

$$\tilde{V}\{\hat{T}_X - T_x \mid \mathcal{F}_x\} = \sum_{h=1}^{H} N_h^2 n_h^{-1} [(1 - f_h) s_h^2 + f_h \sigma_\epsilon^2]$$

is an unbiased estimator of the variance of $\hat{T}_X - T_x$.

7. (Section 5.6) Let a simple random sample of size n be selected from a population of size N. Let duplicate measures be made on $m, m < n$ of the observations, where $X_{ij}, j = 1, 2$ are the measurements. Assume that

$$X_{ij} = x_i + u_{ij},$$
$$u_{ij} \sim ind(0, \sigma_u^2),$$

where u_{ij} is independent of x_t for all ij and t. Let

$$\hat{T}_X = N n^{-1} \left(\sum_{i \in A_r} \bar{X}_{i\cdot} + \sum_{i \in A_s} X_i \right),$$

where $\bar{X}_{i\cdot}$ is the mean of the two determinations on elements with two determinations, A_r is the set of indices for elements with two

determinations and A_s is the set of indices for elements with a single determination. What is the variance of $\hat{T}_X$ as an estimator of T_x? Give an unbiased estimator of $V\{\hat{T}_X - T_x \mid \mathcal{F}_x\}$. Include the finite population correction.

8. (Section 5.6) Let the predictor of u_g for known σ_e^2 and σ_{eg}^2 be given by (5.5.2). Show that $E\{(\hat{u}_g - u_g)^2\} = (\sigma_u^2 + \sigma_{eg}^2)^{-1}\sigma_u^2\sigma_{eg}^2$, where $\hat{u}_g$ is as defined in (5.5.2).

9. (Section 5.5) Prove that the estimator (5.5.6) is unbiased for θ_g in the sense that $E\{\hat{\theta}_g - \theta_g\} = 0$. Derive expression (5.5.7).

10. (Section 5.5) The y_g values for counties 201, 202, 203, and 204 were treated as unobserved in Table 5.13. Assume that the values are 0.754, 0.975, 1.065, and 0.874, with n_g values of 15, 10, 16, and 15, respectively. Are the predicted values in the table consistent with these observations? Using the estimated parameters given in Example 5.5.1, compute predicted values using the given observations.

11. (Section 5.5) The standard error for $\hat{\sigma}_u^2$ of Example 5.5.1 is 0.0062. Use a Taylor approximation to estimate the variance of $\hat{\gamma}_3$ for county 3 of Table 5.13, treating σ_e^2 as known. Use a Taylor expansion to find the leading term in the bias of $\hat{\gamma}_g$ under the assumption that $\hat{\sigma}_u^2$ is unbiased and that σ_e^2 is known.

12. (Section 5.5) Use the facts that $\mathbf{V} = \text{diag}(\sigma_u^2 + \sigma_{eg}^2)$ and that $\mathbf{X}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$, to show that (5.5.16) is sufficient for (5.5.15).

13. (Section 5.6) Assume the model

$$y_i = \beta_0 + x_i\beta_1 + e_i,$$
$$X_i = x_i + u_i,$$

where (e_i, u_i) is independent of x_i, and e_i is independent of u_i. It is desired to estimate β_1 using a simple random sample of n values of (y_i, X_i). How small must the true κ_{xx} be for the ordinary least squares estimator to have a MSE smaller than that of $\kappa_{xx}^{-1}\hat{\beta}_{1,ols}$? Assume that

$$\begin{pmatrix} y \\ X \end{pmatrix} \sim NI\left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} 1.4 & 1.0 \\ 1.0 & 1.0 \end{pmatrix}\right].$$

14. (Section 5.6) In Example 5.6.1 the finite population correction was ignored. Calculate the variance under the assumption that $Nn^{-1} = 0.6$. Construct an unbiased estimator of the variance of the sample mean

assuming it is known that $\sigma_\alpha^2 = 0.02$ and $\sigma_\epsilon^2 = 0.15\sigma_e^2$, but σ_e^2 is unknown.

15. (Section 5.6) The jackknife was used to estimate variances in Example 5.6.2. In this exercise we use Taylor methods. Let

$$\begin{aligned}\mathbf{b}_i &= [Z_{1i}, Z_{2i}, Z_{3i}^2, \psi_n(Z_{1i} - \bar{Z}_1)^2, \psi_n(Z_{1i} - \bar{Z}_1)(Z_{2i} - \bar{Z}_2),\\ &\quad \psi_n(Z_{2i} - \bar{Z}_2)^2, \psi_n(Z_{3i} - \bar{Z}_3)^2],\end{aligned}$$

where $\psi_n = [n(n-1)^{-1}]^{1/2}$. Then $\bar{\mathbf{b}} = (\bar{Z}_1, \bar{Z}_2, m_{11}, m_{12}, m_{22}, m_{33})$ and an estimator of the variance of $\bar{\mathbf{b}}$ is

$$\hat{V}\{\bar{\mathbf{b}}\} = n^{-1}(n-1)^{-1}\sum_{i\in A}(\mathbf{b}_i - \bar{\mathbf{b}})'(\mathbf{b}_i - \bar{\mathbf{b}}).$$

The variance estimator is biased for the sample covariances, but is judged an adequate approximation in large samples. Then the estimated covariance matrix of

$$\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}_x^2, \hat{\sigma}_{a,e}^2, \hat{\sigma}_{a,u}^2)'$$

is

$$\hat{V}\{\hat{\boldsymbol{\theta}}\} = \hat{\mathbf{H}}\hat{\mathbf{V}}\{\bar{\mathbf{b}}\}\hat{\mathbf{H}}',$$

where $\hat{\mathbf{H}}$ is the estimator of the partial derivative of $\hat{\boldsymbol{\theta}}$ with respect to $\bar{\mathbf{b}}$ evaluated at $\hat{\boldsymbol{\theta}}$. Show that the rows of $\hat{\mathbf{H}}$ are

$$\frac{\partial\hat{\theta}_1}{\partial\bar{\mathbf{b}}} = (1, -\hat{\beta}_1, 0, \bar{Z}_2 m_{12}\hat{\beta}_1, -\bar{Z}_2 m_{12}^{-1}, 0.5\bar{Z}_2 m_{12}^{-1}),$$

$$\frac{\partial\hat{\theta}_2}{\partial\bar{\mathbf{b}}} = (0, 0, 0, -m_{12}^{-1}\hat{\beta}_1, m_{12}^{-1}, -0.5m_{12}^{-1}),$$

$$\frac{\partial\hat{\theta}_3}{\partial\bar{\mathbf{b}}} = (0, 0, 0, 2m_{12}\hat{\sigma}_y^{-2}, -\hat{\sigma}_y^{-2}\hat{\sigma}_x^2, 0.5\hat{\sigma}_y^{-2}\hat{\sigma}_x^2),$$

$$\frac{\partial\hat{\theta}_4}{\partial\bar{\mathbf{b}}} = (0, 0, 0, 0, 0, 1),$$

$$\frac{\partial\hat{\theta}_5}{\partial\bar{\mathbf{b}}} = (0, 0, 1, -2m_{12}\hat{\sigma}_y^{-2}, \hat{\sigma}_y^{-2}\hat{\sigma}_x^2, -0.5\hat{\sigma}_y^2\hat{\sigma}_x^2),$$

where $\hat{\sigma}_y^2 = \hat{\beta}_1^2 \hat{\sigma}_x^2$. The estimated covariance matrix of $(m_{11}, m_{12}, m_{22}, m_{33})$ is

$$\begin{pmatrix} 3.4025 & 2.1599 & 1.4496 & 0 \\ 2.1599 & 1.6520 & 1.3556 & 0 \\ 1.4496 & 1.3556 & 1.4412 & 0 \\ 0 & 0 & 0 & 1.9254 \end{pmatrix} \times 10^{-6}$$

and the covariance matrix of $(\bar{Z}_1, \bar{Z}_2)$ is

$$\begin{pmatrix} 53.960 & 46.031 \\ 46.031 & 49.038 \end{pmatrix} \times 10^{-6}.$$

Compute the Taylor estimated variance of $\hat{\theta}$.

16. (Section 5.5) Let $a_g \sim NI(0, \sigma_u^2 + \sigma_{eg}^2)$ and let a sample $a_1, a_2, \ldots, a_m$ be given. Define an estimator of σ_u^2 to be the solution to the equation

$$m^{-1} \sum_{g=1}^{m} (\hat{\sigma}_u^2 + \sigma_{eg}^2)^{-1} a_g^2 = 1.00,$$

where the σ_{eg}^2, $0 < \sigma_{eg}^2 < C_\sigma$, $g = 1, 2, \ldots, m$, are known and C_σ is a positive constant. Let $\hat{\sigma}_u^2 = 0$ if $\Sigma_{g=1}^m \sigma_{eg}^{-2} a_g^2 < m$. Assume that $\sigma_u^2 > 0$ and that

$$\lim_{m \to \infty} m^{-1} \sum_{g=1}^{m} (\sigma_u^2 + \sigma_{eg}^2)^{-1} = \Phi,$$

where Φ is a positive constant. Obtain the limiting distribution of $\hat{\sigma}_u^2$ as $m \to \infty$.

Hint: The quantity $(\sigma_u^2 + \sigma_{eg}^2)^{-1} = \partial log(\sigma_u^2 + \sigma_{eg}^2)/\partial \sigma_u^2$. Also, $(\sigma_u^2 + \sigma_{eg}^2)^{-1}$ is monotone decreasing in σ_u^2 for positive σ_u^2. See Fuller (1996, Chapter 5) to prove that $\hat{\sigma}_u^2$ is consistent for σ_u^2.