CHAPTER 17

Domain and Small Area Estimation

17.1 INTRODUCTION

Large-scale surveys cover wide geographical areas, and information on various items is included in the scope of the survey. In most situations, estimates from different sections of populations are required. For example, the "Botswana Aids Impact Survey II" (BAIS II) was conducted in 2003 to cover the whole Botswana to collect data on exposure to HIV infections as well as socioeconomic, demographic, and behavioral patterns, among others. The HIV infection status for the country as a whole was not the only important item required but information on different sections of populations (e.g., districts or counties) was also important for the management of HIV infections. A subpopulation of a population is called a domain. Here, district or counties may be considered as a domain (large domain) covering large geographical areas. In case sampling frames of different domains are available, one can select samples by employing a stratified sampling procedure and treating a domain as a stratum. But in reality, sampling frames of the domains under study would not be available at the planning stage of the survey, e.g., domain comprising HIV infection rates among business executives, middle class families, or child-headed households. Furthermore, if the number of domains increases, which is the case for a multicharacter survey (covering information on several characteristics), one cannot select samples from each of the domains directly. If the domain is large and a relatively large sample size is obtained, one may get reliable estimates of the parameters of interest. But reliable estimates of HIV infection rates of (for example) immigrants of different nationalities in different districts may not be obtained directly from the sample because the sample size belonging to such a subpopulation is not reasonably large enough. In this case, design-based estimates become very unstable. We call a section of population, whose representation in a sample is small or absent, a small area or small domain. We will consider the methods of estimation from domains and small areas separately. Theories of domain and small area estimation have been considered by Purcell and Kish (1979),

Survey Sampling Theory and Applications
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00017-0

Gonzalez (1973), Platek and Singh (1986), Platek et al. (1987), and Brackstone (1987), among others. Details have been given by Ghosh and Rao (1994), Chaudhuri and Stenger (1992), Rao (2003), and Särndal et al. (1992).

17.2 DOMAIN ESTIMATION

Consider a finite population U of N identifiable units from which a sample s of size n is selected with probability $p(s)$ such that inclusion probabilities for the ith unit π_i is positive for $i = 1,\ldots, N$. Let the population U be partitioned into mutually exclusive and exhaustive domains $U_1,\ldots, U_d,\ldots, U_D$; and y_i be the value of the ith unit of the variable of interest y and $Y_d = \sum_{i \in U_d} y_i$ be the dth domain total, $d = 1,\ldots, D$. Let

$$\widehat{Y} = \sum_{i \in s} b_{si} y_i \tag{17.2.1}$$

be an unbiased estimator of the population total $Y = \sum_{i \in U} y_i$, where b_{si}'s are constants that satisfy the unbiasedness condition $\sum_{s \supset i} b_{si} p(s) = 1$.

Let $s_d = s \cap U_d$ be the part of the sample s that belong to U_d of size n_d, which is not only positive but also reasonably large for every $d = 1,\ldots, D$. Then we find an unbiased estimator of Y_d as

$$\widehat{Y}_d = \sum_{i \in s} b_{si} y_i I_{di} = \sum_{i \in s_d} b_{si} y_i \tag{17.2.2}$$

where

$$I_{di} = 1 \quad \text{if} \quad i \in U_d \quad \text{and} \quad I_{di} = 0 \quad \text{if} \quad i \notin U_d \tag{17.2.3}$$

The variance of $\widehat{Y}_d$ is given by

$$\begin{aligned} V\left(\widehat{Y}_d\right) &= \sum_{i \in U} \alpha_i I_{di} y_i^2 + \sum_{i \neq} \sum_{j \in U} \alpha_{ij} I_{di} I_{dj} y_i y_j \\ &= \sum_{i \in U_d} \alpha_i y_i^2 + \sum_{i \neq} \sum_{j \in U_d} \alpha_{ij} y_i y_j \end{aligned} \tag{17.2.4}$$

where $\alpha_i = \sum_{s \supset i} b_{si}^2 p(s) - 1$ and $\alpha_{ij} = \sum_{s \supset i,j} b_{si} b_{sj} p(s) - 1$.

An unbiased estimator of $Var\left(\widehat{Y}_d\right)$ is given by

$$\begin{aligned} \widehat{V}\left(\widehat{Y}_d\right) &= \sum_{i \in s} c_{si} I_{di} y_i^2 + \sum_{i \neq} \sum_{j \in s} c_{sij} I_{di} I_{dj} y_i y_j \\ &= \sum_{i \in s_d} c_{si}\, y_i^2 + \sum_{i \neq} \sum_{j \in s_d} c_{sij}\, y_i y_j \end{aligned} \tag{17.2.5}$$

where $\sum_{s \supset i} c_{si} p(s) = \alpha_i$ and $\sum_{s \supset i,j} c_{sij} p(s) = \alpha_{ij}$.

The population mean $\overline{Y}_d = Y_d/N_d$ of the domain U_d of known size N_d may be estimated by

$$\widehat{\overline{Y}}_d = \widehat{Y}_d/N_d$$

If N_d is unknown, the following ratio estimator may be used.

$$\widehat{\overline{Y}}_{dR} = \frac{\sum\limits_{i\in s} b_{si} y_i I_{di}}{\sum\limits_{i\in s} b_{si} I_{di}} = \frac{\sum\limits_{i\in s_d} b_{si} y_i}{\sum\limits_{i\in s_d} b_{si}} \tag{17.2.6}$$

Approximate expressions of mean-square error (MSE) of $\widehat{\overline{Y}}_{Rd}$ and its unbiased estimators are obtained by using Theorem 8.2.2 as follows:

$$M\left(\widehat{\overline{Y}}_{dR}\right) \cong \sum_{i\in U_d} \alpha_i z_i^2 + \sum_{i\neq}\sum_{j\in U_d} \alpha_{ij} z_i z_j \tag{17.2.7}$$

$$\widehat{M}\left(\widehat{\overline{Y}}_{dR}\right) \cong \sum_{i\in s_d} \frac{\alpha_i}{\pi_i} z_i^2 + \sum_{i\neq}\sum_{j\in s_d} \frac{\alpha_{ij}}{\pi_{ij}} z_i z_j \tag{17.2.8}$$

where $z_i = y_i - \overline{Y}_d$ and $\widehat{z}_i = y_i - \widehat{\overline{Y}}_{dR}$.

17.2.1 Horvitz–Thomson Estimator

In particular, if $b_{si} = 1/\pi_i$, then the expressions (Eqs. 17.2.2, 17.2.4, and 17.2.5) become

$$\widehat{Y}_d = \widehat{Y}_d(ht) = \sum_{i\in s_d} y_i/\pi_i \tag{17.2.9}$$

$$V\left[\widehat{Y}_d(ht)\right] = \sum_{i\in U_d}\left(\frac{1}{\pi_i} - 1\right) y_i^2 + \sum_{i\neq}\sum_{j\in U_d}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right) y_i y_j, \tag{17.2.10}$$

and

$$\widehat{V}\left[\widehat{Y}_d(ht)\right] = \sum_{i\in s_d}\left(\frac{1}{\pi_i} - 1\right)\frac{y_i^2}{\pi_i} + \sum_{i\neq}\sum_{j\in s_d}\left(\frac{\pi_{ij}}{\pi_i\pi_j} - 1\right)\frac{y_i y_j}{\pi_{ij}}, \tag{17.2.11}$$

respectively.

In case the domain total N_d is known, one can modify the conventional estimator $\widehat{Y}_d(ht)$ as

$$\widehat{Y}_{dR}(ht) = \frac{\widehat{Y}_d(ht)}{\widehat{N}_d} N_d \tag{17.2.12}$$

where $\widehat{N}_d = \sum\limits_{i\in s_d} 1/\pi_i$.

Approximate expressions for the MSE of $\widehat{Y}_{dR}(ht)$ and its unbiased estimators are obtained using Eqs. (8.4.2) and (8.4.4) as follows:

$$M\left[\widehat{Y}_{dR}(ht)\right] \cong V\left(\sum_{i\in s_d}\frac{z_i}{\pi_i}\right)$$

$$= V\left[\widehat{Y}_d(ht)\right] \tag{17.2.13}$$

$$= \sum_{i\in U_d}\left(\frac{1}{\pi_i}-1\right)z_i^2 + \sum_{i\neq}\sum_{j\in U_d}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)z_iz_j$$

and

$$\widehat{M}\left[\widehat{Y}_{dR}(ht)\right] \cong \sum_{i\in s_d}\left(\frac{1}{\pi_i}-1\right)\frac{\widehat{z}_i^2}{\pi_i} + \sum_{i\neq}\sum_{j\in s_d}\left(\frac{\pi_{ij}}{\pi_i\pi_j}-1\right)\frac{\widehat{z}_i\widehat{z}_j}{\pi_{ij}} \tag{17.2.14}$$

where $z_i = y_i - \overline{Y}_d, \overline{Y}_d = Y_d/N_d, \widehat{z}_i = y_i - \widehat{\overline{Y}}_d$ and $\widehat{\overline{Y}}_d = \widehat{Y}_d(ht)/\sum_{i\in s_d} 1/\pi_i$.

For estimation of the population mean $\overline{Y}_d = Y_d/N_d$, we may use an unbiased estimator

$$\widehat{\overline{Y}}_d(ht) = \widehat{Y}_d(ht)/N_d \tag{17.2.15}$$

when N_d is known. However, a better estimator

$$\widehat{\overline{Y}}_d(ht) = \widehat{Y}_d(ht)/\sum_{i\in s} 1/\pi_i \tag{17.2.16}$$

may be used if strictly unbiasedness condition is not required.

17.3 SMALL AREA ESTIMATION

The terms "small area" or "small domain" are commonly used to denote small geographic areas or small subpopulations of a population. Sample sizes that belong to the small areas are generally very small because the overall sample size in a survey is determined with consideration of the accuracy of the estimates of the parameters for the entire population. This is why direct estimators based on the small areas very often yield a large variance. Small area estimation is given considerable importance because of the recent increasing demand for reliable estimates from government and public sector enterprises to determine, for example, birth rate, death rate, school enrollment, and revenue from different municipalities for counties. Various methods of small area estimation are available in the literature. Some of the important methods include symptomatic accounting technique (SAT), direct method, synthetic method, composite method, and various methods based on statistical modeling. Detailed discussions are available in Brackstone (1987), Ghosh and Rao (1994), and Lehtonen et al. (2003).

17.3.1 Symptomatic Accounting Technique

Demographers have proposed various methods of small area estimation, which are broadly known as SATs which use logical relationship in conjunction with statistical relationship based on previous data. SAT utilizes current data gathered from administrative records as well as related data from the latest census. This method was developed for estimation of birth rate, death rate, population projections, types of dwelling, and school enrollment based on diverse registration data, which are known as "symptomatic" variables. A detailed review has been given by Purcell and Kish (1980) and Ghosh and Rao (1994). SAT includes vital rates (VRs) method (Bogue, 1950), composite method (Bogue and Duncan, 1959), census component (CC) method, housing unit (HU) method (Smith and Lewis, 1980), ratio correlation, and difference correlation method, among others.

17.3.1.1 Vital Rates Method

The VRs technique was given by Bogue (1950). Let b_t and d_t be the estimated total number of births and deaths for a smaller area of interest at time t. The estimated total numbers of births and deaths at the larger area containing the smaller area are, respectively, B_t and D_t. Suppose from the last census period $t = 0$, the crude birth (death) rate for the smaller and larger areas are $g_0(h_0)$ and $G_0(H_0)$, respectively, while $G_t(H_t)$ denotes a reliable estimate of birth rate (death rate) at time t for the larger area. The birth rate g_t and death rate h_t for the period t are obtained assuming $g_t/g_0 = G_t/G_0$ and $h_t/h_0 = H_t/H_0$, respectively, and are given as follows:

$$g_t = g_0(G_t/G_0)$$
$$h_t = h_0(H_t/H_0)$$

Let P_t be the total population at time t for the small area of interest. Then $g_t \cong b_t/P_t$ and $h_t \cong d_t/P_t$. Hence P_t can be estimated by

$$\widehat{P}_t = \frac{1}{2}\left(\frac{b_t}{g_t} + \frac{d_t}{h_t}\right) \tag{17.3.1}$$

17.3.1.2 Composite Method

The composite method is an improvement on the VR method. The method was proposed by Bogue and Duncan (1959). In this method, local area population is divided into distinct age subgroup. From each of these subgroups, population estimates are obtained separately by VR method and then summing these estimates across the groups, the "composite" estimate of the current population is obtained. The composite method uses the group-specific birth and death counts for the local area as well as current population of each group for the larger area containing the local area.

17.3.1.3 Census Component Method

In the CC method the population at the local area at a particular time is estimated by using the formula

$$\widehat{P}_t = P_0 + b_t - d_t + m_t \tag{17.3.2}$$

where P_0 is the population in the local area at the last census period $t = 0$ and m_t is the net migration during the period 0 and t.

17.3.1.4 Housing Unit Method

Let $\widehat{U}_t$ and $\widehat{Q}_t$ be the estimated total number of occupied housing units (HUs) and group quarters at the local area at time t. In the HU method, the estimated population total at a time t is

$$\widehat{P}_t = \widehat{U}_t \overline{x}_t + \widehat{Q}_t \overline{x}_t^* \tag{17.3.3}$$

where $\overline{x}_t$ and $\overline{x}_t^*$ denote the estimated average number of persons per occupying HU and group quarters, respectively.

17.3.1.5 Ratio Correlation Method

Suppose we want to estimate the population count p_{ti} for the ith small area at the time t based on the recent two census population count data for the period 0 and 1 along with q symptomatic variables $x_1,\ldots, x_q$ for the periods 0, 1, and t. Let p_{ki} be the population count and x_{kij} be the value of the jth symptomatic variable for the ith area at the time k; $k = 0,\ 1,\ t$; $j = 1,\ldots, q$; $i = 1,\ldots, A$. In the ratio correlation method, the change in the population counts between the periods 0 and 1 of the ith area is measured as

$$r_i = \frac{p_{1i}/P_1}{p_{0i}/P_0} \quad \text{with} \quad P_k = \sum_{i \in A} p_{ki},\ k = 0, 1, t$$

Similarly, the change in the ratio of jth symptomatic variable is measured as

$$z_{ij} = \frac{x_{1ij}/X_{1j}}{x_{0ij}/X_{0j}} \quad \text{with } X_{kj} = \sum_{i \in A} x_{kij}, k = 0, 1, t$$

Consider a multiple regression of r_i on $z_{i1}, \ldots, z_{iq}$ as

$$r_i = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_j z_{ij} + \cdots + \beta_q z_{iq} + \in_i$$

where $\in_i$'s are usual independent identically distributed (iid) error components with mean zero and variance σ^2. Let the fitted multiple regression based on ordinary least squares (OLSs) method be

$$\widehat{r}_i = \widehat{\beta}_0 + \widehat{\beta}_1 z_{i1} + \cdots + \widehat{\beta}_j z_{ij} + \cdots + \widehat{\beta}_q z_{iq} \qquad (17.3.4)$$

with $\widehat{\beta}_j$ as the least squares estimator of β_j.

Let $r_i^* = \dfrac{p_{ti}/P_t}{p_{1i}/P_1}$ and $z_{ij}^* = \dfrac{x_{tij}/X_{tj}}{x_{1ij}/X_{1j}}$.

Then r_i^* can be estimated from the given values of $z_{i1}^*, \ldots, z_{iq}^*$ using the regression (Eq. 17.3.4) above, as

$$\widehat{r}_i^* = \widehat{\beta}_0 + \widehat{\beta}_1 z_{i1}^* + \cdots + \widehat{\beta}_j z_{ij}^* + \cdots + \widehat{\beta}_q z_{iq}^* \qquad (17.3.5)$$

Now noting $\widehat{r}_i^* = \dfrac{\hat{p}_{ti}/\hat{P}_t}{p_{1i}/P_1}$, we find an estimated value of the population count for the period t as

$$\widehat{p}_{ti} = \frac{p_{1i}}{P_1} \widehat{r}_i^* \widehat{P}_t \qquad (17.3.6)$$

where $\widehat{P}_t$ is the estimated reliable population count at time t assumed to be available for the larger area.

17.3.1.6 Difference Correlation Method

In the difference correlation method, the ratio of changes in population count and the jth symptomatic variable between the periods 0 and 1 are measured, respectively, by the differences

$$\widetilde{r}_i = p_{1i}/P_1 - p_{0i}/P_0 \quad \text{and} \quad \widetilde{z}_{ij} = x_{1ij}/X_{1j} - x_{0ij}/X_{0j}$$

Denoting the multiple regression of $\widetilde{r}_i$ on $\widetilde{z}_{i1}, \ldots, \widetilde{z}_{iq}$ as

$$\widehat{\widetilde{r}}_i = \widehat{\widetilde{\beta}}_0 + \widehat{\widetilde{\beta}}_1 \widetilde{z}_{i1} + \cdots + \widehat{\widetilde{\beta}}_j \widetilde{z}_{ij} + \cdots + \widehat{\beta}_q \widetilde{z}_{iq},$$

$\widetilde{r}_i^* = p_{ti}/P_t - p_{1i}/P_1$ and $\widetilde{z}_{ij}^* = x_{tij}/X_{tj} - x_{1ij}/X_{1j}$, we estimate $\widetilde{r}_i^*$ for given values $\widetilde{z}_{i1}^*, \ldots, \widetilde{z}_{iq}^*$ as

$$\widehat{\widetilde{r}}_i^* = \widehat{\widetilde{\beta}}_0 + \widehat{\widetilde{\beta}}_1 z_{i1}^* + \cdots + \widehat{\widetilde{\beta}}_j z_{ij}^* + \cdots + \widehat{\widetilde{\beta}}_q z_{iq}^*$$

Finally, noting $\widetilde{r}_i^* = p_{ti}/P_t - p_{1i}/P_1$, the population count p_{ti} is estimated by

$$\widehat{\widetilde{p}}_{ti} = \left(\widehat{\widetilde{r}}_i^* + p_{1i}/P_1\right)\widehat{P}_t \tag{17.3.7}$$

17.3.2 Direct Estimation

Suppose that we are interested in A small areas $\widetilde{U}_a$, $a = 1,\ldots, A$ with $\widetilde{U}_1 \cup \ldots \cup \widetilde{U}_A = \widetilde{U}$. Let the size of $\widetilde{U}_a$ and the part of the sample s that has intersection with the small area $\widetilde{U}_a$ be $\widetilde{N}_a$ and s_a, respectively. The direct estimators for the population total $\widetilde{Y}_a = \sum_{i \in \widetilde{U}_a} y_i$ and mean $\overline{\widetilde{Y}}_a = \sum_{i \in \widetilde{U}_a} y_i/\widetilde{N}_a$ are estimated by

$$\widehat{\widetilde{Y}}_a = \sum_{i \in s_a} y_i/\pi_i \quad \text{and} \quad \overline{\overline{Y}}_a = \frac{\sum_{i \in s_a} y_i/\pi_i}{\sum_{i \in s_a} 1/\pi_i} \tag{17.3.8}$$

The estimators given in Eq. (17.3.8) are generally unstable because they are often based on the small sample size n_a. So we consider the following alternative methods of estimation.

17.3.3 Synthetic Estimation

Synthetic estimation was proposed by Gonzalez (1973). This method is based on the assumption that the small areas have the same characteristics as larger area. Let $\mathcal{D}'$ be the set of larger domains that have the same characteristics as the smaller area $\widetilde{U}_a$ $\left(\widetilde{U}_a \subset \mathcal{D}'\right)$. Furthermore, let us assume that the auxiliary information x is available for the domains $\mathcal{D}'$ and the smaller area $\widetilde{U}_a$. Suppose $s_{ad} = s_a \cap U_d$ is the intersection of the small area sample s_a with the large domain $U_d (U_d \subset \mathcal{D}')$ with $s_a = \bigcup_{d \in \mathcal{D}'} s_{ad}$, and t_d is a reliable estimator for the domain total Y_d based on a larger sample $s_d = s \cap U_d$, then the synthetic estimator for the small area total $\widetilde{Y}_a$ is obtained as

$$\widehat{\widetilde{Y}}_a^S = \sum_{d \in \mathcal{D}'} \frac{X_{ad}}{X_d} t_d \tag{17.3.9}$$

where $X_{ad} = \sum_{i\in \widetilde{U}_a \cap U_d} x_i$ and $X_d = \sum_{i\in U_d} x_i$ are known totals and x_i be the value of the auxiliary variable for the ith unit.

In particular, if t_d is a ratio estimator of the form

$$t_d = \frac{\widehat{Y}_d}{\widehat{X}_d} X_d$$

with $\widehat{Y}_d = \sum_{i\in s_d} y_i/\pi_i$ and $\widehat{X}_d = \sum_{i\in s_d} x_i/\pi_i$, $\widehat{\widetilde{Y}}{}_a^S$ reduces to

$$\widehat{\widetilde{Y}}{}_a^S = \sum_{d\in\mathcal{D}'} \frac{\widehat{Y}_d}{\widehat{X}_d} X_{ad} \tag{17.3.10}$$

Corollary 17.3.1

For simple random sampling without replacement (SRSWOR) sampling $\pi_i = n/N$, $\widehat{X}_d = N\frac{n_d}{n}\overline{x}_{sd}$, $\widehat{Y}_d = N\frac{n_d}{n}\overline{y}_{sd}$, $\overline{x}_{sd} = \frac{1}{n_d}\sum_{i\in s_d} x_i$, and $\overline{y}_{sd} = \frac{1}{n_d}\sum_{i\in s_d} y_i$, we have

$$\widehat{\widetilde{Y}}{}_a^S = \sum_{d\in\mathcal{D}'} N_{ad}\frac{\overline{y}_{sd}}{\overline{x}_{sd}}\overline{X}_{ad} \tag{17.3.11}$$

where $\overline{X}_{ad} = X_{ad}/N_{ad}$ and $N_{ad} =$ size of the subpopulation $U_d \cap \widetilde{U}_a$.

Case 1: If only one domain is considered, i.e., $U_d = U$, $s_d = s$, $N_{ad} = \widetilde{N}_a$ and $\overline{X}_{ad} = \overline{\widetilde{X}}_a = \sum_{i\in\widetilde{U}_a} x_i/\widetilde{N}_a$, the estimator (Eq. 17.3.11) reduces to

$$\widehat{\widetilde{Y}}{}_a^S = N_a\frac{\overline{y}_s}{\overline{x}_s}\overline{\widetilde{X}}_a \tag{17.3.12}$$

Furthermore, if no auxiliary information is available and we take $x_i = 1$ for very i, Eq. (17.3.12) reduces to

$$\widehat{\widetilde{Y}}{}_a^S = N_a\overline{y}_s \tag{17.3.13}$$

Let $y_i = 1$ if, the ith individual belongs to certain group viz. HIV+ and $y_i = 0$ otherwise, then $\overline{y}_s = \widehat{\pi}_s =$ proportion of HIV+ persons in the sample s. The estimator of the total number of HIV+ persons in the small area "a" is

$$\widehat{\widetilde{Y}}{}_a^S = N_a\frac{n_+}{n}$$

where $n_+ =$ total number of HIV+ persons in the samples of size n.

17.3.4 Composite Estimation

The synthetic estimator has a potential bias whereas the direct estimator has large sampling variation. A composite estimator is obtained by taking the weighted average of the synthetic and direct estimator as

$$\widehat{\widetilde{Y}}_a^S = \phi_a \widehat{\widetilde{Y}}_a + (1 - \phi_a)\, \widehat{\widetilde{Y}}_a^S \tag{17.3.14}$$

Here $\widehat{\widetilde{Y}}_a$ is the direct unbiased estimator of the total $\widetilde{Y}_a$ and ϕ_a $(0 \le \phi_a \le 1)$ is a suitably chosen weight. The optimum value of ϕ_a that minimizes the MSE of $\widehat{\widetilde{Y}}_a^C$, assuming $Cov\left(\widehat{\widetilde{Y}}_a^S, \widehat{\widetilde{Y}}_a\right) \cong 0$, is given by

$$Opt(\phi_a) = \phi_{a0} = \text{MSE}\left(\widehat{\widetilde{Y}}_a^S\right) \Big/ \left[\text{MSE}\left(\widehat{\widetilde{Y}}_a^S\right) + Var\left(\widehat{\widetilde{Y}}_a\right)\right] \tag{17.3.15}$$

The MSE of $\widehat{\widetilde{Y}}_a^C$ with $\phi_a = \phi_{a0}$ is given by

$$\text{MSE}\left(\widehat{\widetilde{Y}}_a^C\right) = \left[\frac{1}{\text{MSE}\left(\widehat{\widetilde{Y}}_a^S\right)} + \frac{1}{Var\left(\widehat{\widetilde{Y}}_a\right)}\right]^{-1} \tag{17.3.16}$$

The optimum weight ϕ_{a0} cannot be used in practice because it involves unknown parameters $\text{MSE}\left(\widehat{\widetilde{Y}}_a^S\right)$ and $Var\left(\widehat{\widetilde{Y}}_a\right)$. One can use an estimate of the weight ϕ_{a0} as

$$\widehat{\phi}_{a0} = \widehat{M}\left(\widehat{\widetilde{Y}}_a^S\right) \Big/ \left[\widehat{M}\left(\widehat{\widetilde{Y}}_a^S\right) + \widehat{V}\left(\widehat{\widetilde{Y}}_a\right)\right] \tag{17.3.17}$$

where $\widehat{M}\left(\widehat{\widetilde{Y}}_a^S\right)$ = an estimator of $\text{MSE}\left(\widehat{\widetilde{Y}}_a^S\right)$ and $\widehat{V}\left(\widehat{\widetilde{Y}}_a\right)$ = an estimator of $V\left(\widehat{\widetilde{Y}}_a\right)$.

Purcell and Kish (1979) used a common weight ϕ for estimating $\widetilde{Y}_a$ for every $a = 1,\ldots, A$. The optimum value ϕ was obtained by minimizing $\frac{1}{A}\sum_{a=1}^{A} \text{MSE}\left(\widehat{\widetilde{Y}}_a^C\right)$ with respect to ϕ and it can be written as:

$$\phi_0 = 1 - \sum_{i=1}^{A} Var\left(\widehat{\widetilde{Y}}_a\right) \Big/ \sum_{i=1}^{A} Var\left(\widehat{\widetilde{Y}}_a^S - \widehat{\widetilde{Y}}_a\right)$$

An estimator of ϕ_0 was chosen as

$$\widehat{\phi}_0 = 1 - \sum_{i=1}^{A} \widehat{V}\left(\widehat{\widehat{Y}}_a\right) \Big/ \sum_{i=1}^{A} \left(\widehat{\widehat{Y}}_a^{S} - \widehat{\widehat{Y}}_a\right)^2 \tag{17.3.18}$$

Obviously, the common weight cannot provide an efficient estimator if the variation of $Var\left(\widehat{\widehat{Y}}_a\right)$ values is very large.

In case ϕ_{a0} depends on N_a, the population size of the area a, Drew et al. (1982) recommended the following weight:

$$\widehat{\phi}_a(1) = \begin{cases} 1 \quad \text{if } \widehat{N}_a \geq \delta N_a \\ \widehat{N}_a / \delta N_a, \quad \text{otherwise} \end{cases} \tag{17.3.19}$$

where $\widehat{N}_a$ is an unbiased estimator of N_a, and δ is a subjectivity factor to be determined by a statistician through his personal experience.

A similar recommendation was suggested by Särndal and Hidiroglou (1989) as

$$\widehat{\phi}_a(2) = \begin{cases} 1 \quad \text{if } \widehat{N}_a \geq \delta N_a \\ \left(\widehat{N}_a / N_a\right)^{k-1}, \quad \text{otherwise} \end{cases} \tag{17.3.20}$$

with a subjective factor k.

Clearly, $\widehat{\phi}_a(1) = \widehat{\phi}_a(2)$ if $\delta = 1$ and $k = 2$.

17.3.5 Borrowing Strength From Related Areas

Here we assume that the variables of interest y follow the following superpopulation model,

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{qi} + \varepsilon_i \tag{17.3.21}$$

where $x_{1i}, \ldots, x_{qi}$ are known auxiliary variables, $\beta_0, \beta_1, \ldots, \beta_q$ are unknown model parameters, ε_i's are iid random components with model expectation $E_m(\varepsilon_i) = 0$ and model variance $V_m(\varepsilon_i) = \sigma^2$.

Reliable estimates of the model parameters $\beta_0, \beta_1, \ldots, \beta_q$ cannot be obtained from the small sample s_a. This problem can be overcome if we suppose that the model (Eq. 17.3.21) holds not only for the area U_a under study but also for a larger area $\Im(\subset U)$, which contains the area U_a and the reliable estimators for the model parameters $\beta_0, \beta_1, \ldots, \beta_q$ based on the relatively larger sample obtained $s_{\Im} = s \cap \Im$ by applying least squares method. The technique of using information from related areas other than the small area to get reliable estimators of the model parameters is known as borrowing strength (or information). This technique was recommended by Ghosh and Rao (1994) and Pfeffermann (2002), among others. The vector of model parameter $\boldsymbol{\beta} = \left(\beta_0, \beta_1, \ldots, \beta_q\right)'$ may be estimated by using generalized least squares (GLSs) method (vide Lehtonen et al., 2003) as

$$\widehat{\boldsymbol{\beta}} = \left(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_q\right)' = \left(\sum_{i\in s_{\mathfrak{J}}} \mathbf{x}_i\mathbf{x}_i'/(c_i\pi_i)\right)^{-1}\left(\sum_{i\in s_{\mathfrak{J}}} \mathbf{x}_i y_i/(c_i\pi_i)\right) \tag{17.3.22}$$

where $\mathbf{x}_i' = (1, x_{1i}, \ldots, x_{qi})$ and c_i's are known positive weights.

Lehtonen et al. (2003) proposed the following types of estimators for $\widetilde{Y}_a$.

17.3.5.1 Synthetic Estimator

$$\widehat{\widetilde{Y}}_a^{*S} = \sum_{i\in U_a} \widehat{y}_i \tag{17.3.23}$$

where $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \ldots + \widehat{\beta}_q x_{qi}$ and $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_q$ are obtained from Eq. (17.3.22).

17.3.5.2 Generalized Regression Estimator

$$\begin{aligned}\widehat{\widetilde{Y}}_a^{*G} &= \sum_{i\in U_a} \widehat{y}_i + \sum_{i\in s_a} (y_i - \widehat{y}_i)/\pi_i \\ &= \widehat{\widetilde{Y}}_a^{*S} + \sum_{i\in s_a} (y_i - \widehat{y}_i)/\pi_i\end{aligned} \tag{17.3.24}$$

17.3.5.3 Composite Estimator

$$\begin{aligned}\widehat{\widetilde{Y}}_a^{*C} &= \phi^* \widehat{\widetilde{Y}}_a^{*S} + \left(1 - \phi^*\right) \widehat{\widetilde{Y}}_a^{*G} \\ &= \phi^* \widehat{\widetilde{Y}}_a^{*S} + \left(1 - \phi^*\right)\left(\widehat{\widetilde{Y}}_a^{*S} + \sum_{i\in s} \frac{y_i - \widehat{y}_i}{\pi_i}\right) \\ &= \widehat{\widetilde{Y}}_a^{*S} + \left(1 - \phi^*\right) \sum_{i\in s} \frac{y_i - \widehat{y}_i}{\pi_i}\end{aligned} \tag{17.3.25}$$

where $\phi^*(0 \le \phi^* \le 1)$ is a suitably chosen weight.

Example 17.3.1

The following table gives the estimated total number of births and deaths for the three cities Durban, Pietermaritzburg, and Richards Bay of the state KwaZulu-Natal (KZN) based on a household survey conducted in 2010 along with the birth and death rates for the last census year 2000. The estimated birth rate and death rate of KZN for the year 2010 were obtained 2.15% and 1.9%, respectively.

	2010 survey report		**2000 census report**	
	Births (b_t)	**Deaths (d_t)**	**Percentage birth rate ($g_0 \times 100$)**	**Percentage death rate ($h_0 \times 100$)**
Durban	2500	2000	2.25	1.75
Pietermaritzburg	3000	2500	2.00	1.25
Richards Bay	1000	6000	1.80	1.50
KZN	10000 (B_t)	7500 (D_t)	2.35 (G_0)	1.80 (H_0)

The estimated birth rate $g_t = g_0(G_t/G_0)$, death rate $h_t = h_0(H_t/H_0)$, and the estimated number of population $P_t = (b_t/g_t + d_t/h_t)/2$ for the three cities for the year 2010 are obtained by VR method as follows:

	Percentage birth rate ($g_t \times 100$)	**Percentage death rate ($h_t \times 100$)**	**Total population (P_t)**
Durban	2.058	1.847	114,858
Pietermaritzburg	1.830	1.319	176,713
Richards Bay	1.647	1.583	219,835

Example 17.3.2

A sample of 25 agricultural farms was selected from a list of 280 farms of a certain district by SRSWOR method and information of the production of a certain crop (y) and area of the farms (x) was obtained. The farms were classified into three categories large (1), medium (2) and small (3). The total number of farms and their average sizes were obtained from the last census. The data are given below:

Farm type	Production of crop (kg)	Farm size (acre)	Total number of farms	Mean area of farm (acre)
1	100	20	50	28.5
	180	27		
	150	25		
	175	28		
	160	30		
2	100	15	100	14.5
	87	12		
	60	10		
	75	12		
	65	14		
	70	14		
	80	15		
	85	17		
3	50	8	130	7.5
	60	8		
	60	7		
	50	6		
	45	5		
	45	5		
	50	8		
	40	7		
	30	5		
	20	6		
	20	5		
	30	6		

Let y_{ij} be the production of crop for the jth farm of the ith category, $j = 1,\ldots, n_i$; $i = 1, 2, 3$, and $n_1 = 5$, $n_2 = 8$ and $n_3 = 12$. The total numbers of large, medium, and small farms in the population and their average sizes are, respectively, $N_1 = 50$, $N_2 = 100$, $N_3 = 130$ and $\overline{X}_1 = 28.5$, $\overline{X}_2 = 14.5$, and $\overline{X}_3 = 7.5$ acres.

Direct Estimates

The direct estimates (sample mean) for mean production of three types of farms are $\overline{y}_1 = 153.00, \overline{y}_2 = 77.750$, and $\overline{y}_3 = 41.667$, respectively.

Synthetic and Composite Estimates

Case I: Sample Mean as an Estimator of Overall Mean

Synthetic Estimates

The estimated overall mean $\left(\overline{Y}\right)$ production is $\bar{y} = 75.48$.

The mean production of crop $\left(\overline{Y}_i\right)$ for the ith type of farm can be estimated by using synthetic method and have been given in the following table:

Farm type	**Number of farms** N_i	**Average farm size** $\overline{X}_i$	**Estimated mean production per farm** $\hat{\overline{Y}}_i^S = \frac{N_i\overline{X}_i}{N\overline{X}}\bar{y}$
Large	50	28.5	27.937
Medium	100	14.5	28.427
Small	130	7.5	19.115
Overall	280	13.75	

Composite Estimators

The composite estimator for $\overline{Y}_i$ is given by

$$\hat{\overline{Y}}_i^C = \phi_i\bar{y}_i + (1 - \phi_i)\hat{\overline{Y}}_i^S$$

Now using Drew et al. (1982) with $\delta = 1$, we find

$$\phi_i = \begin{cases} 1 & \text{if } \widehat{N}_i > N_i \\ \dfrac{\widehat{N}_i}{N_i} & \text{otherwise} \end{cases}$$

where $\widehat{N}_i = n_i \times N/n$. Since $\widehat{N}_1 = 56$, $\widehat{N}_2 = 89.5$ and $\widehat{N}_3 = 134.4$, we find $\phi_1 = 1$, $\phi_2 = 0.896$, and $\phi_3 = 1$. Hence we obtain the composite estimators as $\hat{\overline{Y}}_1^C = \bar{y}_1 = 153.0$, $\hat{\overline{Y}}_2^C = 0.896 \times \bar{y}_2 + (1 - 0.896) \times \hat{\overline{Y}}_2^S = 72.602$, and $\hat{\overline{Y}}_3^C = \bar{y}_3 = 41.667$.

Case II: Overall Mean Is Estimated by Ratio Estimator

The ratio estimator for the overall mean $\overline{Y}$ is given by

$$\widehat{\overline{Y}}_R = \frac{\overline{y}}{\overline{x}}\overline{X} = (75.458/12.6) \times 13.75 = 82.369.$$

The synthetic and composite estimators for the *i*th type farm is obtained by using the formula $\widehat{\overline{Y}}_i^S(R) = \dfrac{N_i\overline{X}_i}{N\overline{X}}\widehat{\overline{Y}}_R$ and $\widehat{\overline{Y}}_i^C(R) = \phi_i\overline{y}_i + (1-\phi_i)\widehat{\overline{Y}}_i^S(R)$, respectively. The estimates are given as follows:

			Estimated mean production per farm		
Farm type	**Number of farms** N_i	**Average farm size** $\overline{X}_i$	**Synthetic** $\widehat{\overline{Y}}_i^S(R)$	**Weight** ϕ_i	**Composite** $\widehat{\overline{Y}}_i^C(R)$
Large	50	28.50	30.487	1.000	153.00
Medium	100	14.50	31.022	0.896	72.890
Small	130	7.50	20.859	1.000	41.667
Overall	280	13.75			

Case III: Borrowing of Strength

Here we fit a linear regression $y_{ij} = \beta_0 + \beta_1 x_{ij} + \in_{ij}$ over 25 observations. We find estimates of β_0 and β_1 are $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x} = 4.316$ and $\widehat{\beta}_1 = \sum_{i\in s}(y_i - \overline{y})(x_i - \overline{x}) / \sum_{i\in s}(x_i - \overline{x})^2 = 5.648$, respectively, where s denotes the sample of 25 observations.

Synthetic Estimators

Synthetic estimates for the mean production of large, medium, and small farms are $\widehat{\overline{Y}}_1^{*S} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{X}_1 = 165.281$, $\widehat{\overline{Y}}_2^{*S} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{X}_2 = 86.211$, and $\widehat{\overline{Y}}_3^{*S} = \widehat{\beta}_0 + \widehat{\beta}_1\overline{X}_3 = 46.675$, respectively.

Generalized Regression Estimator

Let s_1, s_2, and s_3 be the sample corresponding to large, medium, and small farms with respective sizes $n_1 = 5$, $n_2 = 8$, and $n_3 = 12$. The generalized regression estimators for the mean production of large, medium, and small farms are obtained as follows:

$$\widehat{\overline{Y}}_1^{*G} = \frac{N}{nN_1}\sum_{i\in s_1}(y_i - \widehat{y}_i) + \widehat{\overline{Y}}_1^{*S} = \frac{280}{50\times 25}\times 9.189 + 165.281 = 167.340$$

$$\widehat{\overline{Y}}_2^{*G} = \frac{N}{nN_2}\sum_{i\in s_2}(y_i - \widehat{y}_i) + \widehat{\overline{Y}}_2^{*S} = \frac{280}{100\times 25}\times(-28.152) + 86.257 = 83.057$$

$$\widehat{\overline{Y}}_3^{*G} = \frac{N}{nN_3}\sum_{i\in s_3}(y_i - \widehat{y}_i) + \widehat{\overline{Y}}_3^{*S} = \frac{280}{130\times 25}\times 18.963 + 46.721 = 48.309$$

17.3.6 Use of Models

Indirect estimators such as synthetic and composite estimators are based on implicit or explicit models that connect small areas through supplementary data. In this section we will consider the general mixed effect model proposed by Henderson (1975). The random effects account for the area variation that cannot be explained by auxiliary variables. The general mixed effect model can be partitioned into area-level model, a unit-level model, and a hybrid. Area-level models relate small area direct estimators to area-specific auxiliary data. This model is useful if unit-level auxiliary information is not available. In unit-level model the study variable of each unit is related to a set of concomitant variables. Hybrid models are combination of unit-level and area-level models.

17.3.6.1 General Linear Mixed Model

Henderson (1975) considered the following general linear mixed model, where y_{ij}, the value of the study variable y of the jth unit of the ith area is related to q auxiliary variables through the following model

$$y_{ij} = x_{ij1}\beta_1 + \cdots + x_{ijt}\beta_t + \cdots + x_{ijq}\beta_q + z_{i1}v_1 + \cdots + z_{iA}v_A + \in_{ij};$$
$$j = 1, \ldots, N_i;\quad i = 1, \ldots, A \qquad (17.3.26)$$

where x_{ijt} is the value of the auxiliary variable x_t of jth unit of the ith small area is assumed to be known, z_{ij} is a known positive constant, and $v_1, \ldots, v_A$ are area-specific random effects that are assumed to be independently and identically distributed with

$$E_m(v_i) = 0 \text{ and } V_m(v_i) = \sigma_v^2;\quad i = 1, \ldots, A \qquad (17.3.27)$$

where E_m and V_m are expectation and variance operators with respect to the model (Eq. 17.3.26). The error components $\in_{ij}$'s are independently distributed with

$$E_m(\in_{ij}) = 0 \text{ and } V_m(\in_{ij}) = \sigma_e^2 \tag{17.3.28}$$

We further assume that $\in_{ij}$'s are independent of v_i's. Let a sample s of n units be selected from the population by some suitable sampling design, and let n_i be the number of units that falls in the ith small area $i = 1,\ldots, A$. Here we assume that the model (Eq. 17.3.26) is valid for the sampled data also, i.e.,

$$\begin{aligned} y_{ij} &= x_{ij1}\beta_1 + \cdots + x_{ijt}\beta_t + \cdots + x_{ijq}\beta_q + z_{i1}v_1 + \cdots + z_{iA}v_A + \in_{ij}; \\ j &= 1, \ldots, n_i; \quad i = 1, \ldots, A \end{aligned} \tag{17.3.29}$$

Battese et al. (1988) used the model (Eq. 17.3.29) to estimate county crop areas using satellite information as auxiliary variables, whereas Rao and Choudhry (1995) used this model to estimate total wages and salaries for Nova Scotia province using gross business as an auxiliary variable. Further applications of the model (Eq. 17.3.29) with real live data were provided by Kleffe and Rao (1992), Datta and Ghosh (1991), and Ghosh and Lahiri (1998), among others.

In matrix notation, Eq. (17.3.29) can be written as

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zv} + \mathbf{e} \tag{17.3.30}$$

where

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ \cdot \\ y_{1n_1} \\ \cdot \\ y_{A1} \\ \cdot \\ y_{An_A} \end{pmatrix}^{n\times 1}, \mathbf{X} = \begin{pmatrix} x_{111} & \cdot & x_{11q} \\ & \cdot & \\ x_{1n_1 1} & \cdot & x_{1n_1 q} \\ & \cdot & \\ x_{A11} & \cdot & x_{A1q} \\ & \cdot & \\ x_{An_A 1} & \cdot & x_{An_A q} \end{pmatrix}^{n\times q}, \mathbf{Z} = \begin{pmatrix} z_{11} & \cdot & z_{1A} \\ & \cdot & \\ z_{11} & \cdot & z_{1A} \\ & \cdot & \\ z_{A1} & \cdot & z_{AA} \\ & \cdot & \\ z_{A1} & \cdot & z_{AA} \end{pmatrix}^{n\times A},$$

$$\mathbf{e} = \begin{pmatrix} \in_{11} \\ \cdot \\ \in_{1n_1} \\ \cdot \\ \in_{A1} \\ \cdot \\ \in_{An_A} \end{pmatrix}^{n\times 1}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \cdot \\ \beta_q \end{pmatrix}^{q\times 1}, n = \sum n_i \text{ and } \mathbf{v}' = (v_1, \quad \cdot \quad, v_A). \tag{17.3.31}$$

Matrices $\mathbf{X}$ and $\mathbf{Z}$ are known and are of rank q and A, respectively; $\beta_1,\ldots,\ \beta_q$ are unknown regression coefficients (fixed effect) and $\mathbf{v}$ is an unknown vector of random effects and $\mathbf{e}$ is an unknown random error vector. The vectors $\mathbf{e}$ and $\mathbf{v}$ are distributed independently with mean $\mathbf{0}$ and variance—covariance matrices $\mathbf{R} = \sigma_e^2\mathbf{I}_n$ and $\mathbf{G} = \sigma_v^2\mathbf{I}_A$, respectively. $\mathbf{R}$ and $\mathbf{G}$ involve parameter vector $\boldsymbol{\tau} = \left(\sigma_v^2, \sigma_e^2\right)$, which is called the variance component vector. For the time being, we will assume that the vector $\boldsymbol{\tau}$ is known.

Here we are interested in the method of estimating (or predicting) a linear combination

$$\mu = \mathbf{l}'\boldsymbol{\beta} + \mathbf{m}'\mathbf{v} \tag{17.3.32}$$

where $\mathbf{l}'$ and $\mathbf{m}'$ are vectors of known constants.

We concentrate on the class C_l of linear unbiased estimators of μ, which consists of the estimators of the form

$$\widehat{\mu} = \mathbf{u}'\mathbf{y} + b \tag{17.3.33}$$

satisfying $E_m(\widehat{\mu}) = \mu$, where $\mathbf{u}'$ and b's are known constants. For a known parameter vector $\boldsymbol{\tau}$, we define the best linear unbiased prediction (BLUP) estimator of μ as one which minimizes $V_m(\widehat{\mu}) = E_m(\widehat{\mu} - \mu)^2$ for $\widehat{\mu} \in C_l$. Following Henderson (1975), we note that the BLUP estimator for μ is

$$\widehat{\mu}_h(\boldsymbol{\tau}) = \mathbf{l}'\widehat{\boldsymbol{\beta}} + \mathbf{m}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{17.3.34}$$

where $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$ is the GLS estimator of $\boldsymbol{\beta}$ and $\mathbf{V} = \mathbf{R} + \mathbf{Z}\mathbf{G}\mathbf{Z}'$ is the variance—covariance matrix of $\mathbf{y}$. The MSE of $\widehat{\mu}_h(\boldsymbol{\tau})$ was obtained by Henderson (1975) is as follows:

$$\text{MSE}\ \left[\widehat{\mu}_h(\boldsymbol{\tau})\right] = (\mathbf{l}', \mathbf{m}')\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{pmatrix}\begin{pmatrix} \mathbf{l} \\ \mathbf{m} \end{pmatrix} \tag{17.3.35}$$

where

$$\begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}' & \mathbf{C}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}^{-1}.$$

17.3.6.2 Nested Error Regression Model

Let y_{ij} be the value of the study variable for the unit j of the ith area, $j = 1,\ldots,\ N_i,\ i = 1,\ldots,\ A$. Assume for each unit $\mathbf{x_{ij}} = (x_{ij1}, \ldots, x_{ijq})'$, a

q-vector auxiliary information available. A unit-level model relates y_{ij} to the auxiliary variable through the following nested error regression model

$$y_{ij} = x_{ij1}\beta_1 + \cdots + x_{ijk}\beta_k + \cdots + x_{ijq}\beta_q + v_i + \in_{ij};\ j = 1, \ldots, N_i;$$
$$i = 1, \ldots, A \tag{17.3.36}$$

where $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_q\right)'$ is a q-vector regression parameter, v_i is the ith specific random effect, and $\in_{ij}$ is the error component. Battese et al. (1988) considered the model (Eq. 17.3.36) for estimating the mean acreage under crop for 36 counties in Iowa using satellite and survey data.

The population mean of the ith small area is given by

$$\overline{Y}_i = \sum_{j=1}^{N_i} y_{ij}/N_i = \overline{X}_{i\cdot 1}\beta_1 + \cdots + \overline{X}_{i\cdot k}\beta_k + \cdots + \overline{X}_{i\cdot q}\beta_q + v_i + \overline{E}_i$$

where $\overline{X}_{i\cdot k} = \sum\limits_{j=1}^{N_i} x_{ijk}/N_i$ and $\overline{E}_i = \sum\limits_{j=1}^{N_i} \in_{ij}/N_i$.

For large N_i we can write

$$\begin{aligned} \overline{Y}_i &= \overline{X}_{i\cdot 1}\beta_1 + \cdots + \overline{X}_{i\cdot j}\beta_j + \cdots + \overline{X}_{i\cdot q}\beta_q + v_i \\ &= \overline{\mathbf{X}}_i'\boldsymbol{\beta} + v_i \\ &= \mu_i(\text{say}) \end{aligned} \tag{17.3.37}$$

where $\overline{\mathbf{X}}_i' = \left(\overline{X}_{i\cdot 1}, \ldots, \overline{X}_{i\cdot q}\right)'$.

Suppose our objective is to estimate the mean μ_i from known $\overline{\mathbf{X}}_i'$ on the basis of the sampled data satisfying

$$y_{ij} = x_{ij1}\beta_1 + \cdots + x_{ijk}\beta_k + \cdots + x_{ijq}\beta_q + v_i + \in_{ij};\ j = 1, \ldots, n_i;$$
$$i = 1, \ldots, A$$

The BLUP estimator of μ_i is obtained from Eq. (17.3.34) by substituting $\mathbf{l}' = \overline{\mathbf{X}}_i' = \left(\overline{X}_{i\cdot 1}, \ldots, \overline{X}_{i\cdot q}\right)$, $\mathbf{m}' = (0, \ldots, 0, 1, 0, \ldots, 0)$ with 1 in the ith position and it is given by

$$\widehat{\mu}_h^{(i)}(\boldsymbol{\tau}) = \mathbf{l}'\widehat{\boldsymbol{\beta}} + \mathbf{m}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{17.3.38}$$

where $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\right)$, $\mathbf{V} = Diag(\mathbf{V}_1, \ldots, \mathbf{V_i}, \ldots, \mathbf{V_A})$, $\mathbf{V}_i = \sigma_e^2\mathbf{I}_i + \sigma_v^2\mathbf{E}_{i,i}$, $\mathbf{I}_i$ is a unit matrix of order n_i, $\mathbf{E}_{i,i}$ is a $n_i \times n_i$ matrix with each element is 1 and

$$\mathbf{Z}' = \begin{pmatrix} 1 & \cdot & 1 & \cdot & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & 0 & \cdot & 1 & \cdot & 1 \end{pmatrix}^{A\times n}$$

Furthermore, noting $\mathbf{V}_i^{-1} = \frac{1}{\sigma_e^2}\left(\mathbf{I}_i - \gamma_i\mathbf{E}_{i,i}/n_i\right)$ with $\gamma_i = n_i\sigma_v^2/\ \left(n_i\sigma_v^2 + \sigma_e^2\right)$, the BLUP estimator $\mu_i = \overline{Y}_i = \overline{\mathbf{X}}_i'\boldsymbol{\beta} + v_i$ is obtained by Prasad and Rao (1990) as

$$\begin{aligned}\widehat{\mu}_h^{(i)}(\boldsymbol{\tau}) &= \overline{X}_i'\widehat{\boldsymbol{\beta}} + \gamma_i\left(\overline{y}_i - \overline{\mathbf{x}}_{\mathbf{i}\cdot}'\,\widehat{\boldsymbol{\beta}}\right)\\ &= \gamma_i\left\{\overline{y}_i - \left(\overline{\mathbf{x}}_{i\cdot}' - \overline{\mathbf{X}}_i'\right)\widehat{\boldsymbol{\beta}}\right\} + (1-\gamma_i)\overline{\mathbf{X}}_i'\widehat{\boldsymbol{\beta}}\end{aligned} \qquad (17.3.39)$$

where $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$, $\overline{\mathbf{x}}_{i\cdot}' = \left(\overline{x}_{i\cdot 1}, \ldots, \overline{x}_{i\cdot q}\right)$ and $\overline{x}_{i\cdot k} = \sum_j x_{ijk}/n_i$.

The estimator (Eq. 17.3.39) is a weighted average of the sample regression estimator $\overline{y}_i - \left(\overline{\mathbf{x}}_{i\cdot}' - \overline{\mathbf{X}}_i'\right)\widehat{\boldsymbol{\beta}}$ and the synthetic estimator $\overline{\mathbf{X}}_i'\widehat{\boldsymbol{\beta}}$ of $\overline{Y}_i$ with weights γ_i and $1-\gamma_i$, respectively. For areas i with no samples, $\widehat{\mu}_h^{(i)}(\boldsymbol{\tau}) = \overline{X}_i'\widehat{\boldsymbol{\beta}}$. The coefficient γ_i is a "shrinkage factor" providing a compromise between the large variance of the regression predictor $\overline{y}_i - \left(\overline{\mathbf{x}}_{i\cdot}' - \overline{\mathbf{X}}_i'\right)\widehat{\boldsymbol{\beta}}$ and the bias of the synthetic estimator $\overline{X}_i'\widehat{\boldsymbol{\beta}}$. The estimator $\widehat{\mu}_h^{(i)}(\boldsymbol{\tau})$ reduces to the regression estimator $\overline{y}_i - \left(\overline{\mathbf{x}}_{i\cdot}' - \overline{\mathbf{X}}_i'\right)\widehat{\boldsymbol{\beta}}$ if the sample size n_i is very large or the model variance σ_v^2 is sufficiently large compared to σ_e^2. Conversely, if the model variance σ_v^2 is small compared to $\sigma_v^2 + \sigma_e^2/n_i$, then $\gamma_i \cong 0$ and the estimator $\widehat{\mu}_h^{(i)}(\boldsymbol{\tau})$ reduces to the synthetic estimator $\overline{\mathbf{X}}_{\mathbf{i}}'\widehat{\boldsymbol{\beta}}$.

The MSE of the BLUP estimator $\widehat{\mu}_h^{(i)}(\boldsymbol{\tau})$ was given by Prasad and Rao (1990) as

$$\begin{aligned}\text{MSE}\ \left[\widehat{\mu}_h^{(i)}(\boldsymbol{\tau})\right] &= E\left[\widehat{\mu}_h^{(i)}(\boldsymbol{\tau}) - \mu_i\right]^2\\ &= (1-\gamma_i)\sigma_v^2 + \left(\overline{\mathbf{X}}_i - \gamma_i\overline{\mathbf{x}}_{i\cdot}'\right)'\left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\left(\overline{\mathbf{X}}_i - \gamma_i\overline{\mathbf{x}}_{i\cdot}'\right)\end{aligned} \qquad (17.3.40)$$

17.3.6.3 Area-Level Model

Area-level random effect model is used when the auxiliary information is available at the area level. Here, direct survey estimates are models to area-specific auxiliary data. In an area-level model, we are often interested in estimating $\theta_i = g\left(\overline{Y}_i\right)$, a function of ith area population mean $\overline{Y}_i$ using the following linear model

$$\theta_i = x_{i1}\widetilde{\beta}_1 + \cdots + x_{ij}\widetilde{\beta}_j + \cdots + x_{iq}\widetilde{\beta}_q + z_i v_i;\ \ i = 1, \ldots, A \qquad (17.3.41)$$

where x_{ij}'s are known values of the jth auxiliary variable x_j for the ith area, z_i's are known positive constants, v_i's are area-specific random effects assumed to be independently and identically distributed with

$$E_m(v_i) = 0 \text{ and } V_m(v_i) = \widetilde{\sigma}_v^2;\ \ i = 1, \ldots, A \tag{17.3.42}$$

Numerous applications of the area-level model are available in the literature. Fay and Herriot (1979) used $\theta_i = \overline{Y}_i$, the true per capita income in "local government unit" i, while National Research Council, USA (2000) used $\theta_i = \log Y_i$ with Y_i as the poverty count for the ith area. Ericksen and Kadane (1985) took $\theta_i = (Y_i - C_i)/Y_i$, where Y_i and C_i denote respectively the true and census count for the ith area.

To make inference about θ_i, we assume

$$\widehat{\theta}_i = g\left(\widehat{\overline{Y}}_i\right) = \theta_i + \widetilde{\in}_i \tag{17.3.43}$$

where $\widehat{\overline{Y}}_i$ is a direct estimator based on the selected sample s using a suitable sampling design p and $\widetilde{\in}_i$'s are sampling errors distributed independently with

$$E_m(\widetilde{\in}_i|\theta_i) = 0 \text{ and } V_m(\widetilde{\in}_i|\theta_i) = \widetilde{\sigma}_i^2 \tag{17.3.44}$$

Here the sampling variance $\widetilde{\sigma}_i^2$ is assumed to be known for all areas $i = 1, \ldots, A$.

Eqs. (17.3.43) and (17.3.44) yield

$$\begin{aligned}\widehat{\theta}_i &= x_{i1}\widetilde{\beta}_1 + \cdots + x_{ij}\widetilde{\beta}_j + \cdots + x_{iq}\widetilde{\beta}_q + z_i v_i + \widetilde{\in}_i;\ \ i = 1, \ldots, A\\ &= \mathbf{x}_i'\widetilde{\boldsymbol{\beta}} + z_i v_i + \widetilde{\in}_i\end{aligned} \tag{17.3.45}$$

where $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iq})$, $\widetilde{\boldsymbol{\beta}} = \left(\widetilde{\beta}_1, \ldots, \widetilde{\beta}_j, \ldots, \widetilde{\beta}_q\right)'$, and z_i is a known positive constant.

Note that the model (Eq. 17.3.45) involves both the design-induced random variable $\widetilde{\in}_i$ as well the model-based random variable v_i. Now using Henderson's (1975) result given in Eq. (17.3.34), we find the BLUP estimator θ_i when the parametric function $\widetilde{\boldsymbol{\sigma}} = \left(\widetilde{\sigma}_v^2, \widetilde{\boldsymbol{\sigma}}_e^2\right)$ with $\widetilde{\boldsymbol{\sigma}}_e^2 = \left(\widetilde{\sigma}_1^2, \ldots, \widetilde{\sigma}_A^2\right)$ is known was obtained by Ghosh and Rao (1994) as

$$\widehat{\theta}_i^h(\widetilde{\boldsymbol{\tau}}) = \mathbf{x}_i'\widehat{\widetilde{\boldsymbol{\beta}}} + \widetilde{\gamma}_i\left(\widehat{\theta}_i - \mathbf{x}_i'\widehat{\widetilde{\boldsymbol{\beta}}}\right) \tag{17.3.46}$$

where $\widetilde{\gamma}_i = z_i^2\widetilde{\sigma}_v^2 / \left(z_i^2\widetilde{\sigma}_v^2 + \widetilde{\sigma}_i^2\right)$ and $\widehat{\widetilde{\boldsymbol{\beta}}} = \left(\sum_{i=1}^{A} \dfrac{\mathbf{x}_i\hat{\theta}_i}{\widetilde{\sigma}_i^2 + z_i^2\widetilde{\sigma}_v^2}\right) \Big/ \left(\sum_{i=1}^{A} \dfrac{\mathbf{x}_i\mathbf{x}_i'}{\widetilde{\sigma}_i^2 + z_i^2\widetilde{\sigma}_v^2}\right)$.

The expression (Eq. 17.3.46) can be written as a weighted average of the direct estimator $\widehat{\theta}_i$ and synthetic estimator $\mathbf{x}_i'\widehat{\widetilde{\boldsymbol{\beta}}}$ with weights $\widetilde{\gamma}_i$ and $1-\widetilde{\gamma}_i$ as follows:

$$\widehat{\theta}_i^h(\widetilde{\boldsymbol{\tau}}) = \widetilde{\gamma}_i\widehat{\theta}_i + (1-\widetilde{\gamma}_i)\mathbf{x}_i'\widehat{\widetilde{\boldsymbol{\beta}}} \tag{17.3.47}$$

The design bias and MSE of $\widehat{\theta}_i^h(\widetilde{\boldsymbol{\tau}})$ are given by Ghosh and Rao (1994) and Rao (2003) as follows:

$$\begin{aligned} B\left[\widehat{\theta}_i^h(\widetilde{\boldsymbol{\tau}})\right] &= E_p\left[\widehat{\theta}_i^h(\widetilde{\boldsymbol{\tau}}) - \theta_i\right] \\ &\cong (1-\widetilde{\gamma}_i)\left(\mathbf{x}_i'\widehat{\widetilde{\boldsymbol{\beta}}}^* - \theta_i\right) \end{aligned} \tag{17.3.48}$$

where $\widehat{\widetilde{\boldsymbol{\beta}}}^*$ is the conditional expectation of $\widehat{\widetilde{\boldsymbol{\beta}}}$ given $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$

$$\begin{aligned} \text{MSE } \left[\widehat{\theta}_i^h(\widetilde{\boldsymbol{\tau}})\right] &= E_p\left[\widehat{\theta}_i^h(\widetilde{\boldsymbol{\tau}}) - \theta_i\right]^2 \\ &= \widetilde{\gamma}_i\widetilde{\sigma}_i^2 + (1-\widetilde{\gamma}_i)^2\mathbf{x}_i'\left(\sum_{i=1}^{A}\frac{\mathbf{x}_i\mathbf{x}_i'}{\widetilde{\sigma}_i^2 + z_i^2\widetilde{\sigma}_v^2}\right)^{-1}\mathbf{x}_i \end{aligned} \tag{17.3.49}$$

In practice, the variances $\widetilde{\sigma}_i^2$ and $\widetilde{\sigma}_v^2$ are unknown and they are replaced by sample estimates.

17.3.6.4 Fay–Herriot Model

In the Fay and Herriot (1979) model, parameter of interest $\theta_i = \overline{Y}_i = \mu_i$, the population mean of the ith small area is related to the auxiliary information $\mathbf{x}_i' = (x_{i1}, \ldots, x_{ij}, \ldots, x_{iq})$ through the following model

$$\theta_i = x_{i1}\widetilde{\beta}_1 + \cdots + x_{ij}\widetilde{\beta}_j + \cdots + x_{iq}\widetilde{\beta}_p + v_i \quad \text{for } i = 1, \ldots, A \tag{17.3.50}$$

The sample mean of the ith area $\overline{y}_i\left(=\widehat{\theta}_i\right)$ is again related to $\overline{Y}_i(=\theta_i)$ as follows:

$$\overline{y}_i = \theta_i + \widetilde{\in}_i = \mathbf{x}_i'\widetilde{\boldsymbol{\beta}} + v_i + \widetilde{\in}_i \tag{17.3.51}$$

It is assumed that v_i's and $\widetilde{\in}_i$'s are independently distributed with $E_m(v_i) = 0$, $V_m(v_i) = \widetilde{\sigma}_v^2$, and $E_m(\widetilde{\in}_i|\theta_i) = 0$ and $V_m(\widetilde{\in}_i|\theta_i) = \widetilde{\sigma}_i^2$. The BLUP estimator for θ_i is given by

$$\begin{aligned} \widehat{\theta}_i^{fh}(\widetilde{\boldsymbol{\tau}}) &= \widehat{\mu}_i^{fh}(\widetilde{\boldsymbol{\tau}}) \\ &= w_i\overline{y}_i + (1-w_i)\mathbf{x}_i'\widehat{\widetilde{\boldsymbol{\beta}}} \end{aligned} \tag{17.3.52}$$

where $\hat{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}'\widetilde{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\left(\mathbf{X}'\widetilde{\mathbf{V}}^{-1}\bar{\mathbf{y}}\right) = \left(\sum_{i=1}^{A}\frac{\mathbf{x}_i\bar{y}_i}{\widetilde{\sigma}_i^2+\widetilde{\sigma}_v^2}\right)\Big/\left(\sum_{i=1}^{A}\frac{\mathbf{x}_i\mathbf{x}_i'}{\widetilde{\sigma}_i^2+\widetilde{\sigma}_v^2}\right)$,
$\bar{y}' = \left(\bar{y}_1,\ldots,\bar{y}_A\right)$, $\mathbf{V} = diag\left(\widetilde{\sigma}_1^2+\widetilde{\sigma}_v^2,\ldots,\widetilde{\sigma}_A^2+\widetilde{\sigma}_v^2\right)$ and

$$w_i = \widetilde{\sigma}_v^2\big/\left(\widetilde{\sigma}_v^2+\widetilde{\sigma}_i^2\right) \tag{17.3.53}$$

Furthermore, $\widehat{\theta}_i^{fh}(\widetilde{\boldsymbol{\tau}})$ reduces to the synthetic estimator $\mathbf{x}_i\hat{\hat{\boldsymbol{\beta}}}$ if $\widetilde{\sigma}_i^2$ is relatively large compared to $\widetilde{\sigma}_v^2$. Alternatively if $\widetilde{\sigma}_v^2$ is large compared to $\widetilde{\sigma}_i^2$, the estimator $\widehat{\mu}_i^{fh}$ becomes the sample mean $\bar{y}_i$.

17.3.7 Empirical Best Linear Unbiased Prediction, Empirical Bayes, and Hierarchical Bayes Methods

17.3.7.1 Empirical Best Linear Unbiased Prediction

The BLUP estimator $\widehat{\mu}_h(\boldsymbol{\tau})$ and $\widehat{\theta}_i^{h}(\widetilde{\boldsymbol{\tau}})$ given in Eqs. (17.3.34) and (17.3.46) involve model parameters $\boldsymbol{\tau}$ and $\widetilde{\boldsymbol{\tau}}$, respectively. The parameters $\boldsymbol{\tau}$ and $\widetilde{\boldsymbol{\tau}}$ are unknown in most practical situations. Hence, to use the BLUP estimator in practice, the parameters $\boldsymbol{\tau}$ and $\widetilde{\boldsymbol{\tau}}$ are estimated through a selected sample. Various methods of estimating the variance component parameters are available in literature. Popular among them are the method of moments, maximum likelihood (ML), and restricted ML (RML) methods (Rao, 2003). All these methods yield consistent estimators under general regularity conditions. Replacing the parameters $\boldsymbol{\tau}$ and $\widetilde{\boldsymbol{\tau}}$ by their suitable estimates $\widetilde{\boldsymbol{\tau}}$ and $\hat{\widetilde{\boldsymbol{\tau}}}$, in BLUP estimators, empirical best linear unbiased prediction (EBLUP) estimators are obtained. For details, readers are referred to Rao (2003). A few examples have been given below.

17.3.7.1.1 Onefold Nested Error Regression Model

For the nested error regression model (Eq. 17.3.36), the EBLUP estimator of $\overline{Y}_i = \overline{\mathbf{X}}_i'\boldsymbol{\beta} + v_i$ is obtained by replacing $\boldsymbol{\tau} = \left(\sigma_v^2, \sigma_e^2\right)$ by its unbiased estimator $\widehat{\boldsymbol{\tau}} = \left(\widehat{\sigma}_v^2, \widehat{\sigma}_e^2\right)$ in the expression of $\widehat{\mu}_h^{(i)}(\boldsymbol{\tau})$ of Eq. (17.3.38) and it will be denoted by

$$\widehat{\mu}_i^{\text{BLUP}} = \widehat{\mu}_i(\widetilde{\boldsymbol{\tau}}) \tag{17.3.54}$$

The unbiased estimators of $\widehat{\sigma}_e^2$ and $\widehat{\sigma}_v^2$ obtained by Prasad and Rao (1990) are

$$\widehat{\sigma}_e^2 = \sum_i\sum_j \widetilde{e}_{ij}^2\Big/(n-A-q) \text{ and } \widehat{\sigma}_v^2 = \frac{1}{n^*}\left[\sum_i\sum_j \widehat{u}_{ij}^2 - (n-q)\widehat{\sigma}_e^2\right] \tag{17.3.55}$$

where $n^* = n - \mathrm{tr}\left[(\mathbf{X}'\mathbf{X})^{-1}\sum_{i=1}^{A} n_i^2 \bar{\mathbf{x}}_{\mathbf{i}} \bar{\mathbf{x}}_{\mathbf{i}}'\right]$, $\bar{\boldsymbol{x}}_{\boldsymbol{i}}' = \left(\bar{x}_{i\cdot 1}, \ldots, \bar{x}_{i\cdot q}\right)$ and $\left\{\widehat{e}_{ij}\right\}$'s are the residuals from the OLS regression of $y_{ij} - \bar{y}_i$ on $\left(x_{ij1} - \bar{x}_{i\cdot 1}, \ldots, x_{ijq} - \bar{x}_{i\cdot q}\right)$ and $\left\{\widehat{u}_{ij}\right\}$'s are the residuals from the OLS regression of y_{ij} on (x_{ij1},..., x_{ijq}). The estimator $\widehat{\sigma}_v^2$ in Eq. (17.3.55) may turnout to be negative but the probability of the estimator taking a negative value becomes zero if $A \rightarrow \infty$. Hence we take the value of $\widehat{\sigma}_v^2$ as zero if it is negative.

17.3.7.1.2 Fay–Herriot Model

For Fay–Herriot model (Eq. 17.3.50), an unbiased estimator of $\widetilde{\sigma}_v^2$ assuming $\widetilde{\sigma}_i^2$ is known, is given by

$$\widehat{\widetilde{\sigma}}_v^2 = \left[\sum_{i=1}^{A} \widehat{u}_i^2 - \sum_{i=1}^{A} \widetilde{\sigma}_i^2 \left(1 - \mathbf{x}_{\mathbf{i}}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i\right)\right] \Big/ (A - q) \tag{17.3.56}$$

where $\widehat{u}_i = \bar{y}_i - \bar{\mathbf{x}}_{\mathbf{i}}'\widehat{\widetilde{\boldsymbol{\beta}}}$ and $\widehat{\widetilde{\boldsymbol{\beta}}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Here again, $\widehat{\widetilde{\sigma}}_v^2$ can take a negative value, and we set it equal to zero when it is negative. The EBLUP estimator for θ_i is obtained by placing $\widetilde{\sigma}_v^2 = \widehat{\widetilde{\sigma}}_v^2$ in the expression of Eq. (17.3.52) and it will be denoted by

$$\begin{aligned}\widehat{\theta}_i^{\mathrm{BLUP}} &= \widehat{\theta}_i^{fh}\left(\widehat{\widetilde{\boldsymbol{\tau}}}\right) \\ &= \widehat{w}_i \bar{y}_i + (1 - \widehat{w}_i)\bar{\mathbf{x}}_i'\widehat{\widetilde{\boldsymbol{\beta}}}\end{aligned} \tag{17.3.57}$$

where

$$\widehat{w}_i = \widehat{\widetilde{\sigma}}_v^2 \Big/ \left(\widehat{\widetilde{\sigma}}_v^2 + \widetilde{\sigma}_i^2\right)$$

17.3.7.2 Empirical Bayes Approach

Here we assume that the parameter of interest θ has some prior distribution $\xi(\theta|\boldsymbol{\lambda})$ with unknown parameter $\boldsymbol{\lambda}$. The posterior distribution of θ, given the data, is first obtained assuming that the parameter $\boldsymbol{\lambda}$ is known. The parameter $\boldsymbol{\lambda}$ is estimated through the marginal distribution of the data. Inference of θ is obtained from the estimated posterior distribution $\xi\left(\theta|\widehat{\boldsymbol{\lambda}}\right)$. Good details have been given by Morris (1983) and Ghosh and Rao (1994).

Let us assume that $\widehat{\theta}_i$'s given in Eq. (17.3.45) are independently normally distributed with unknown mean θ_i but known variance $\widetilde{\sigma}_i^2$. Let us further assume that the distribution of θ_i is normal with mean $\boldsymbol{\lambda} = \mathbf{x}_\mathbf{i}'\widetilde{\boldsymbol{\beta}}$ and variance $z_i^2\widetilde{\sigma}_\nu^2$, where $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}_\nu^2$ are unknown model parameters. The posterior distribution of θ_i, given $\widehat{\theta}_i$, $\widetilde{\boldsymbol{\beta}}$, and $\widetilde{\sigma}_\nu^2$, has the mean $\widehat{\theta}_i^B = \widetilde{\gamma}_i\widehat{\theta}_i + (1-\widetilde{\gamma}_i)\mathbf{x}_i'\widetilde{\boldsymbol{\beta}}$ and variance $\widetilde{\gamma}_i\widetilde{\sigma}_i^2$, where $\widetilde{\gamma}_i$ is given in Eq. (17.3.46). It can be shown that $\widehat{\theta}_i^B$ is the Bayes estimator of $\widehat{\theta}_i$ under squared error loss. This Bayes estimator $\widehat{\theta}_i^B$ involves unknown parameters $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}_\nu^2$. These unknown parameters can be estimated from the marginal distribution of $\widehat{\theta}_i$ which is normal with mean $\mathbf{x}_i'\widetilde{\boldsymbol{\beta}}$ and variance $z_i^2\widetilde{\sigma}_\nu^2 + \widetilde{\sigma}_i^2$. Let $\widehat{\widetilde{\boldsymbol{\beta}}}^{(ml)}$ and $\widehat{\widetilde{\sigma}}_\nu^{2(ml)}$ be the ML or RML estimates of $\widetilde{\boldsymbol{\beta}}$ and $\widetilde{\sigma}_\nu^2$, respectively, then the empirical Bayes estimator of θ_i is given by

$$\widehat{\theta}_i^{\mathrm{EB}} = \widehat{\widetilde{\gamma}}_i^{(ml)}\widehat{\theta}_i + \left(1 - \widehat{\widetilde{\gamma}}_i^{(ml)}\right)\mathbf{x}_\mathbf{i}'\widehat{\widetilde{\boldsymbol{\beta}}}^{(ml)} \tag{17.3.58}$$

with

$$\widehat{\widetilde{\gamma}}_i^{(ml)} = z_i^2\widehat{\widetilde{\sigma}}_\nu^{2(ml)} \Big/ \left(z_i^2\widetilde{\sigma}_\nu^{2(ml)} + \widetilde{\sigma}_i^2\right)$$

Further details have been given by Rao (2003).

17.3.7.3 Hierarchical Bayes Approach

In the hierarchical Bayes (HB) approach, we assume that the parameter of interest has some prior distribution with unknown parameters. The unknown parameters have again some prior distribution with unknown parameters and so on. At the ultimate stage, it is assumed that all the parameters of the prior distribution of the ultimate stage are known. The inference is based on the posterior distribution of the parameter of interest. In particular, under squared error loss, the parameter is estimated by the posterior mean, and its performance is measured by its posterior variance.

Consider the model (Eq. 17.3.45). Let us assume that the distribution of $\widehat{\theta}_i$ given $\theta_i, \widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}_\nu^2$ are independent normal with mean θ_i and variance $\widetilde{\sigma}_i^2$. Furthermore, the distribution of θ_i given $\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}_\nu^2$ are independent normal with mean $\overline{\mathbf{x}}_i'\widetilde{\boldsymbol{\beta}}$ and variance $z_i^2\widetilde{\sigma}_\nu^2$. Let us assume that the prior distribution of $\widetilde{\boldsymbol{\beta}}$ is uniform, and $\widetilde{\sigma}_v^2$ and $\widetilde{\sigma}_i^2$ are known. Then the posterior distribution of θ_i given $\widehat{\theta}_i$ and $\widetilde{\boldsymbol{\beta}}$ is normal with mean

$$E\left(\theta_i|\widehat{\theta}_i, \widetilde{\sigma}_v^2, \widetilde{\sigma}_i^2\right) = \widetilde{\gamma}_i\widehat{\theta}_i + (1-\widetilde{\gamma}_i)\mathbf{x}_i'\widehat{\widetilde{\boldsymbol{\beta}}} = \widehat{\theta}_i^{\mathrm{HB}} \tag{17.3.59}$$

and variance $\widetilde{\gamma}_i\widetilde{\sigma}_i^2$, where $\widetilde{\gamma}_i$ is given in Eq. (17.3.46). Hence the HB estimator for θ_i is $\widehat{\theta}_i^{\mathrm{HB}} = \widehat{\theta}_i^{h}(\widetilde{\boldsymbol{\tau}})$, the BLUP estimator given in Eq. (17.3.47).

To compare the performances of the estimators of the parameters of interest generated by different methods of estimation and estimation of confidence intervals based on these estimators, one needs to compute the MSEs of the estimators. Generally, exact expressions of MSEs are hard to obtain. However, Morris (1983), Kass and Steffey (1989), Prasad and Rao (1990), Singh et al. (1998), Butar and Lahiri (2001), Datta et al. (2002), among others, have provided approximate expressions of MSE under various small-area models. Ghosh and Lahiri (1987), Lahiri (1990), Ghosh (1992), Ghosh and Rao (1994), Chottopadhyaya et al. (1999), Prasad and Rao (1990), Kleffe and Rao (1992), Singh et al. (1994), Stukel and Rao (1999), Jiang et al. (2002), among others, presented simulation studies based on live data to compare efficiencies and biases of various small area estimators under different models.

17.4 EXERCISES

17.4.1 The total number of admission and the total number of dropout of secondary students of three countries of a certain district in the year 2008 are given below. The admission and dropout rates as recorded by the last census of schools in 2001 were 2.5% and 1.8%, respectively, whereas the estimated admission and dropout rates for the district in 2008 were 1.8 and 1.6, respectively. Use VR method and estimate the total number of students for the year 2008.

Counties	Admission 2008	Dropout 2008
1	525	480
2	625	550
3	300	250

17.4.2 A sample of 20 factories was selected at random from 250 factories of a certain district. The factories were classified into small, medium, and large according to the number of workers. The monthly output (in 000$) and number of workers in the sampled farms along with the total number of factories and their mean number of workers are given in the following table.

Factories	Monthly output (000$) *y*	Number of workers *x*	Total number of factories	Mean number of workers
Large	200	100	50	140
	180	150		
	175	125		
	280	120		
Medium	100	75	80	45
	120	70		
	90	50		
	75	40		
	65	40		
	70	30		
	80	30		
	60	40		
Small	18	10	120	15
	25	15		
	20	15		
	40	20		
	30	15		
	25	15		
	35	20		
	30	15		

Estimate the average monthly output of the different types of factories by synthetic and composite method taking direct estimator as (i) sample mean and (ii) ratio estimator. Suppose that a linear model $y_i = \beta x_i + \in_i$ is appropriate, estimate the mean output of the different types of factories by (i) synthetic and (ii) generalized regression method.

17.4.3 A sample of 35 agricultural farms was selected from a list of 1200 farms of a certain district at random. The farms were classified into three groups large (L), medium (M), and small (S) according to their size as well as irrigation facilities. The yield of wheat (y), farm size (x), and irrigation facilities are recorded from the sampled farms and is given in the following table.

Yield of wheat and farm sizes of the selected farms.

Irrigation facility	**Farm**	**Yield of wheat (00 kg) *y***	**Farm size (acre) *x***
Yes	L	100	20
		180	27
		150	25
	M	100	15
		87	12
		60	10
		75	12
		65	14
		70	14
		80	15
		85	17
	S	50	8
		60	8
		60	7
		50	6
		45	5
		45	5
		50	8
		40	7
		45	5
No	L	50	25
		70	30
		85	40
	M	30	15
		25	12
		30	10
		35	12
		45	15
	S	20	5
		15	5
		12	6
		15	6
		12	5
		15	5
		20	6

Estimate the average yield of wheat of the different types of farm by (i) direct, (ii) synthetic, and (iii) composite method (a) without assuming a model and (b) fitting a linear regression of yield of a farm on the area of the farm.