CHAPTER 12

Further Topics

12.1 Introduction

This chapter contains three major topics of interest in sample surveys. The first topic in Sections 12.2 through 12.5, presents the linearization, jackknife, bootstrap, and balanced repeated replication procedures. The last three methods in this group are also known as the **resampling procedures**. Estimating the variances and MSEs of nonlinear estimators, such as the ratio of two sample means and also more-complicated estimators, and finding confidence limits for the corresponding population quantities are two of the major purposes of these approaches. The linearization and jackknife procedures can also be used to reduce the bias of an estimator.

The second topic, presented in Section 12.6, describes the different procedures available for estimation of the population quantities for **small areas**, domains, and subpopulations of small sizes. The final section describes estimation methods for **complex surveys**.

12.2 Linearization

For the sake of illustration, consider estimation of the ratio $R = \bar{Y}/\bar{X}$ of two means, as in Chapter 9. As described in Appendix A9, the large sample variance of $\hat{R} = \bar{y}/\bar{x}$ in (9.1) was found by expanding $(\hat{R} - R)$ in a series and ignoring higher-order terms, that is, through the **linearization** of $\hat{R}$. The estimator $v(\hat{R})$ for this variance was also found using this procedure and was presented in (9.2).

Reducing the bias of $\hat{R}$ through linearization, the estimator $\hat{R}_T = \hat{R}[1 - (1 - f)(c_{xx} - c_{xy})/n]$ in (9.24) was obtained. Thus, linearization can be employed for finding an approximate expression for the variance or MSE of a nonlinear estimator, to estimate it from the sample and also to reduce its bias. The following example compares $\hat{R}$ and $\hat{R}_T$.

Example 12.1. College enrollments: For 1990, enrollment for private colleges (y_i) and the total enrollment for private and public colleges (x_i) appear in the *Statistical Abstracts of the United States* (1992), Table 264. For the $N = 48$ states, excluding the two largest states, California and New York, the ratio of the means of the private and total enrollment is $R = 0.2092$.

These enrollments for a sample of $n = 10$ from the 48 states are presented in the second and third columns of Table 12.1, along with their means and variances. From these figures, $\hat{R} = 52.8/198.7 = 0.2657$. Further, the sample covariance is $s_{xy} = 6949.4$. Now, $s_d^2 = s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}\, s_{xy} = 4508.6 + (0.2657)^2(17{,}628) - 2(0.2657)(6949.4) = 2060.14$. Thus, from (9.2), $v(\hat{R}) = [(48 - 10)/480](2060.14)/(198.7)^2 = 0.0041$.

Since the sample mean and variance of x are 198.7 and 17,628, $c_{xx} = 17{,}628/198.7^2 = 0.4465$. Since the sample standard deviations of x and y are 132.8 and 67.1, and the sample covariance is 6949.4, $c_{xy} = 6949.4/(198.7)(52.8) = 0.6624$.

From the above figures, $\hat{R}_T = 0.2705$, which does not differ much from $\hat{R} = 0.2657$. Since the sample is nearly 20% of the population, the reduction in the bias is not large.

In general, linearization conveniently provides expressions for the biases and variances of complicated estimators and it is routinely employed in theoretical and applied statistics. As another example, through this approach, Shao and Steel (1999) develop a variance estimator for the Horvitz–Thompson estimator in (7.34) obtained with imputed data. Woodruff (1971) notes that approximations to the variance of an estimator and its estimate can also be obtained through the *Taylor's series expansion*. Wolter (1984) describes linearization and other methods for variance estimation.

12.3 The jackknife

Quenouille (1956) proposed a method for reducing the bias of an estimator. Since it can also be used for estimating variances, finding confidence limits, and similar purposes, Tukey (1958) popularized it as the *jackknife*, a pocket- or penknife. The following subsections present this procedure for $\hat{R}$.

Bias reduction

For a large population, the expectation of $\hat{R}$ can be expressed as

$$E(\hat{R}) = R + c_1/n + c_2/n^2 + \ldots, \tag{12.1}$$

where $c_1, c_2, c_3, \ldots$ are the expectations of moments of increasing degree.

If the ith pair of the observations, $i = 1, 2, ..., n$, is deleted from the sample, the means become $\bar{y}_{n-1,i} = (n\bar{y} - y_i)/(n - 1)$ and $\bar{x}_{n-1,i} = (n\bar{x} - x_i)/(n - 1)$. As in the case of $\hat{R}$, the expectation of the ratio $\hat{R}_{n-1,i} = \bar{y}_{n-1,i}/\bar{x}_{n-1,i}$ can be expressed as

$$E(\hat{R}_{n-1,i}) = R + c_1/(n-1) + c_2/(n-1)^2 + ... \qquad (12.2)$$

Consider the *pseudo* values

$$\hat{R}_i^{/} = n\hat{R} - (n-1)\hat{R}_{n-1,i} \qquad (12.3)$$

for $i = 1, 2, ..., n$. From (12.1) and (12.2),

$$E(\hat{R}_i^{/}) = R - c_2/n(n-1) - (2n-1)c_3/n^2(n-1)^2 + -.... \qquad (12.4)$$

The biases of these pseudo values are approximately of order $1/n^2$. The jackknife estimator for R is given by their average,

$$\hat{R}_J = \sum \hat{R}_i^{/}/n = n\hat{R} - (n-1)\hat{R}_{n-1}, \qquad (12.5)$$

where $\hat{R}_{n-1} = \Sigma \hat{R}_{n-1,i}/n$, and from (12.4) its bias is also approximately of order $1/n^2$.

In the original development of this procedure, Quenouille divides the initial sample into g groups of sizes $m = n/g$ each, deletes one group at a time, constructs the pseudo values, and considers their average as the estimator. For example, when the sample is divided into two groups of size $m = n/2$ each, denote their means by $(\bar{x}_1, \bar{y}_1)$ and $(\bar{x}_2, \bar{y}_2)$, with the corresponding ratios $\hat{R}_1 = \bar{y}_1/\bar{x}_1$ and $\hat{R}_2 = \bar{y}_2/\bar{x}_2$ The pseudo values now are $2\hat{R} - \hat{R}_2$ and $2\hat{R} - \hat{R}_1$, and the jackknife estimator for R is given by $\hat{R}_J = 2\hat{R} - (\hat{R}_1 + \hat{R}_2)/2$.

Variance estimation

For sampling from a finite population, the jackknife estimator suggested for the MSE of $\hat{R}_J$ or $\hat{R}$ is obtained from

$$\begin{aligned} vj(\hat{R}) &= [(1-f)/n]\sum(\hat{R}_i^{/} - \hat{R}_J)^2/(n-1) \\ &= (1-f)(n-1)\sum(\hat{R}_{n-1,i} - \hat{R}_{n-1})^2/n \end{aligned} \qquad (12.6)$$

Table 12.1. Jackknife calculations for college enrollment.

Sample No.	Total Enrollment (x)	Private Enrollment (y)	$\bar{x}_{n-1,i}$	$\bar{y}_{n-1,i}$	$\hat{R}_{n-1,i}$
1	90	12	210.78	57.33	0.2720
2	227	26	195.56	55.78	0.2852
3	252	55	192.78	52.56	0.2726
4	171	53	201.78	52.78	0.2616
5	57	16	214.44	56.89	0.2653
6	419	234	174.22	32.67	0.1875
7	352	67	181.67	51.22	0.2820
8	34	8	217.00	57.78	0.2663
9	85	11	211.33	57.44	0.2718
10	300	46	187.44	53.56	0.2857
Average	198.7	52.8	198.7	52.8	0.2650
Variance	17,628.0	4508.6			0.000811

For the case of two groups, this estimator takes the form of $[(1 - f)/n](\hat{R}_1 - \hat{R}_2)^2/2$.

The estimator for the mean and its variance are obtained from $\bar{X}\hat{R}_J$ and $\bar{X}^2 v_J(\hat{R})$. For the college enrollments, the calculations needed for the jackknife are presented in Table 12.1. With these figures, from (12.5), $\hat{R}_J = 10(0.2657 - 9(0.265) = 0.272$. This estimate is not too far from $\hat{R} = 0.2657$ and $\hat{R}_T = 0.2705$. From (12.6), $v_J(\hat{R}) = (38/48)(9/10)(0.007299) = 0.0052$.

Durbin (1959), J.N.K. Rao and Webster (1966), Rao and Rao (1969), P.S.R.S. Rao (1969; 1974), Krewski and Chakrabarti (1981), Kreswski and J.N.K. Rao (1981), and others evaluated the merits of the linearization and jackknife methods for reducing the bias of the ratio estimator and estimating its variance, through the model in (9.10) and suitable assumptions. As g increased to n, the bias and MSE of $\hat{R}_J$ were found to decrease. However, the variance of $v_J(\hat{R})$ was found to be larger than that of $v(\hat{R})$ in (9.2) obtained from the linearization procedure; that is, $v_J(\hat{R})$ can be less **stable**. Since the S.E. for an estimator is obtained from its variance estimator, stability is a desirable property.

The jackknife procedure for ratio estimation with stratification is presented in Jones (1974). Following this approach, for a finite population, n and $(n - 1)$ in (12.5) should be replaced by $w = n(N - n + 1)/N$ and $(1 - w) = (n - 1)(N - n)/N$. Schucany et al. (1971) consider higher-order jackknife by the reapplication of the procedure. Since the ratio estimator for the mean $\hat{R}\bar{X}$ can be expressed as $\bar{y} + \hat{R}(\bar{X} - \bar{x})$,

P.S.R.S. Rao (1979) considered $\bar{y} + \hat{R}_J(\bar{X} - \bar{x})$ for reducing its bias, and compared it with $\hat{R}_J\bar{X}$.

For variance estimation, J.N.K. Rao and Shao (1992) examine the jackknife for the data from hot-deck imputation, J.N.K. Rao (1996) for the ratio and regression methods of imputation in single and multistage sampling, and Yung and J.N.K. Rao (1996) for stratified multistage sampling.

12.4 The bootstrap

This method proposed by Efron (1982) and also presented in Efron and Tibshirani (1993) can be used for estimating the variance or MSE of a nonlinear estimator and for finding confidence limits.

Different procedures of applying the bootstrap for variance estimation and other purposes were suggested by Gross (1980), McCarthy and Snowden (1985), J.N.K. Rao and Wu (1987), Kovar et al. (1988), Sitter (1992), among others. Booth et al. (1994) examine the coverage probabilities of the confidence limits obtained from applying the bootstrap procedure to the separate and combined ratio estimators. Efron (1994) describes the bootstrap for missing data and imputation. Shao and Tu (1995) present it along with the jackknife method, and Shao and Sitter (1996) examine it for the imputed data. For the variance of the two-phase regression estimator, Sitter (1997) compares it with the linearization and jackknife methods. C.R. Rao, et al. (1997) suggest a procedure for selecting the bootstrap samples.

The bootstrap method is illustrated below for estimating the variance of the mean of a sample selected from an infinite or finite population and for the estimation of $V(\hat{R})$.

Infinite population

Consider an infinite population for the characteristic y with mean $\mu = E(y)$ and variance $\sigma^2 = E(y - \mu)^2$. The mean $\bar{y} = \Sigma_1^n y_i/n$ and variance $s^2 = \Sigma_1^n(y_i - \bar{y}^2)/(n - 1)$ of a sample of observations, y_i, $i = 1, ..., n$, are unbiased for μ and σ^2, respectively. The variance of $\bar{y}$ and its unbiased estimator are $V(\bar{y}) = \sigma^2/n$ and $v(\bar{y}) = s^2/n$.

Let $\bar{y}_b = \Sigma_1^m y_i'/m$ and $s_b^2 = \Sigma_1^m(y_i - \bar{y})^2/(m - 1)$ denote the mean and variance of the observations y_i', $i = 1, 2, ..., m$ of a sample of size m selected *with* replacement from the above sample. With I denoting

the initial sample, $E(\bar{y}_b|I) = \bar{y}$ and

$$v(\bar{y}_b|I) = \frac{1}{m}\sum_1^n \frac{(y_i - \bar{y})^2}{n} = \frac{1}{m}\frac{(n-1)}{n}s^2. \tag{12.7}$$

If $m = n - 1$, this variance coincides with $v(\bar{y})$.

In the bootstrap method, the above procedure of selecting the m units is repeated independently a large number (B) of times and the means $\bar{y}_b$, $b = 1, ..., B$, are obtained. Let $\bar{y}_B = \Sigma_1^B \bar{y}_b/B$ and

$$v_B(\bar{y}) = \sum_1^B (\bar{y}_b - \bar{y})^2/B. \tag{12.8}$$

This variance is an estimate for $v(\bar{y}_b|I)$ and hence for $V(\bar{y})$. For an alternative estimator, $\bar{y}$ in this expression is replaced by $\bar{y}_B$ and the sum of squares is divided by $(B - 1)$.

Finite population

Consider a population of N units with mean $\bar{Y}$ and variance S^2. As seen in Chapter 2, the mean $\bar{y}$ of a sample of n units drawn *without* replacement from this population is unbiased for $\bar{Y}$ and has variance $V(\bar{y}) = (1 - f)S^2/n$, where $f = n/N$. The sample variance s^2 is unbiased for S^2, and an unbiased estimator for $V(\bar{y})$ is $v(\bar{y}) = (1 - f)\, s^2/n$. To estimate this variance, more than one method of adopting the bootstrap can be suggested. They differ in the method of selecting the bootstrap sample and its size. The selection *with* or *without* replacement is examined below.

Selection with replacement

If the bootstrap sample y_i', $i = 1, 2, ..., m$ is selected randomly *with* replacement from the original sample y_i, $i = 1, 2, ..., n$, its mean $\bar{y}_b = \Sigma_1^m y_i'/m$ is unbiased for $\bar{y}$ and has the variance $v(\bar{y}_b) = (n - 1)\, s^2/nm$. The expression in (12.8) is an approximation for this variance.

If $m = n - 1$, $v(\bar{y}_b)$ is the same as s^2/n. In this case, $v(\bar{y}) = (1 - f)\, s^2/n$ may be estimated from (12.8) by attaching $(1 - f)$ to its right side. On the other hand, $v(\bar{y}_b)$ will be the same as $v(\bar{y})$ if $m = (n - 1)/(1 - f)$, and hence $v(\bar{y})$ can be estimated from (12.8) with this value for m.

Selection without replacement

If the bootstrap sample of size m is selected *without* replacement, $\bar{y}_b$ is unbiased for $\bar{y}$ and $v(\bar{y}_b) = (n - m)s^2/nm$, which will be equal to $v(\bar{y})$ if $m = n/(2 - f)$.

Bootstrap for estimating the variance of the sample ratio

To estimate the variance of the ratio $\hat{R} = \bar{y}/\bar{x}$, the ratios $\hat{R}_b = \bar{y}_b/\bar{x}_b$ are obtained from the bootstrap samples of size m selected from (x_i, y_i), $i = 1, 2, \ldots, n$. The variance of $\hat{R}$ is obtained from

$$v_B(\hat{R}) = \sum_1^B (\hat{R}_b - \hat{R})^2/B. \tag{12.9}$$

As in the case of the mean, an alternative estimator for the variance is obtained by replacing $\hat{R}$ in this expression with $\hat{R}_B = \Sigma_1^B \hat{R}_b/B$.

To estimate the MSE of $\hat{\bar{Y}}_R = \hat{R}\bar{X}$, P.S.R.S Rao and Katzoff (1996) compared the relative merits of the linearization, jackknife, and bootstrap procedures. For the bootstrap procedure, the above approaches and other available methods were considered. In this investigation, all the procedures were found to underestimate the MSE of $\hat{\bar{Y}}_R$, but the underestimation was the least for the jackknife estimator. Among the bootstrap methods, sampling without replacement with the optimum size $n/(2 - f)$ was found to have relatively the smallest amount of underestimation. The differences in the stabilities of the different estimators were not significant.

12.5 Balanced repeated replication (BRR)

McCarthy (1966, 1969) presents this procedure for estimating the variance of a nonlinear estimator obtained from samples of size two selected *with* replacement from each of G strata. This method for the variance estimator of the population total is examined below.

With $n_g = 2$, an unbiased estimator for the total is obtained from

$$\hat{Y} = \sum_g N_g \bar{y}_g = \sum_g N_g(y_{g1} + y_{g2})/2 = \sum_g (e_{g1} + e_{g2}), \tag{12.10}$$

where $y_{gi}, i = (1, 2)$, are the sample observations and $e_{gi} = N_g y_{gi}/2$. Denoting the variance of y_{gi} by σ_g^2, the variance of this estimator

becomes $V(\hat{Y}) = \Sigma N_g^2 \sigma_g^2/2$. Since the sample variance of the gth stratum is given by $s_g^2 = \Sigma_i (y_{gi} - \bar{y}_g)^2 = (y_{g1} - y_{g2})^2/2$, an unbiased estimate of $V(\hat{Y})$ is obtained from

$$v(\hat{Y}) = \sum_g N_g^2 (s_g^2/2) = \frac{1}{4} \sum_g N_g^2 (y_{g1} - y_{g2})^2 = \sum (e_{g1} - e_{g2})^2. \tag{12.11}$$

Now, for the BRR method, one observation is selected from the two sample observations of each stratum. From these **half-samples** an estimator for the total is given by $K = \Sigma_g N_g y_{gi}$ where y_{gi} can be the first or the second unit in the sample. Note that $K - \hat{Y} = \Sigma_g \pm (e_{g1} - e_{g2})$. The sign in front of the parenthesis depends on whether the first or the second unit is chosen from the sample of the gth stratum. From this expression,

$$(K - \hat{Y})^2 = v(\hat{Y}) + \text{cross-product terms.} \tag{12.12}$$

McCarthy (1966) showed that a set of r, $(G + 1 \leq r \leq G + 4)$, half-samples can be selected such that for every pair of strata half the cross-product terms will have positive signs and the remaining half negative signs. These are known as the **balanced half-samples**. With such samples, from (12.12),

$$\frac{1}{r} \sum_1^r (K_j - \hat{Y})^2 = v(\hat{Y}). \tag{12.13}$$

With the sample units not included in K, another estimator for the total is given by $L = \Sigma_g N_g y_{gi}$. Using the balanced half-samples from this set, another estimator of variance is obtained by replacing K_j in (12.13) by L_j, $j = (1, \ldots, r)$. Further, since $\hat{Y} = (K + L)/2$, from (12.13) or otherwise,

$$\frac{1}{4r} \sum_1^r (K_j - L_j)^2 = v(\hat{Y}). \tag{12.14}$$

Any of these estimates or their averages can be used for estimating $V(\hat{Y})$. However, their biases and stabilities for estimating the MSEs of nonlinear estimators can be different.

Bean (1975) examined the BRR for ratio estimation with two-stage sampling and selection of units with unequal probabilities. Lemeshow and Levy (1978) compared it with the jackknife. Krewski and J.N.K. Rao (1981) compared the BRR with linearization and jackknife for ratio estimation, J.N.K. Rao and Wu (1985) with these methods for stratified sampling, and Valliant (1993) for poststratified estimation. All the three procedures were found to be satisfactory for finding the confidence limits. In some of the empirical studies, linearization seemed to provide slightly better stabilities for the variance estimators than the BRR and jackknife. Wolter (1985) presents this procedure along with the linearization and jackknife methods.

The BRR is closely related to the procedure of estimating the variance through the **interpenetrating subsamples** briefly described in Section 11.1. In this approach, the original sample is divided randomly into k groups of size $m = (n/k)$ each. Denoting their means by $\bar{y}_i$, $i = 1, \ldots, k$, $\Sigma(\bar{y}_i - \bar{y})^2/k(k-1)$ is an unbiased estimator for $V(\bar{y})$. Gurney and Jewett (1975) extend the BRR to the case of $n_g > 2$, with equal sample sizes in all the strata.

12.6 Small-area estimation

In several large-scale surveys, estimates for specified areas, regions, or subpopulations are needed. For example, the U.S. Department of Education requires estimates of the total number and proportion of children of 5 to 17 years age from low-income families in each of the school districts in more than 3000 counties. In 1997 to 1998, more than $7 billion was allocated for the educational programs of children in these types of families. The above estimates are obtained from the CPS of the U.S. Bureau of the Census. Suitable modifications of these estimates are presented in the report edited by Citro et al. (1998).

Sample estimates

In a sample survey conducted in a large area, the sample sizes observed for its small areas can be very small. Chapter 6 examined the estimation for subpopulations. The variance $V(\bar{y}_i) = (1-f)S_i^2/n_i$ in (6.7) for estimating the mean $\bar{Y}_i$ of the ith subpopulation or area clearly becomes large if n_i is small. Similarly, for small sample sizes, the variances in (6.11) and (6.17) for the estimates of the total and proportion and also

the variances of the estimates obtained in Sections 6.7 and 6.8 through stratification can be large.

If the mean of the supplementary characteristic $\bar{X}_i$ for the ith area is known, one can also consider the ratio estimator $(\bar{y}_i / \bar{x}_i)\bar{X}_i$ or the regression estimator $\bar{y}_i + b_i(\bar{X}_i - \bar{x}_i)$ for $\bar{Y}_i$, where $\bar{x}_i$, $\bar{y}_i$, and b_i are obtained from the units observed in that area. However, if the observed sample size n_i is small, these estimators can have large biases and MSEs.

Synthetic and composite estimators

In the synthetic estimation examined by Gonzalez (1973) and others, estimates for a small area are obtained by assuming that it is similar to the entire large area or population. For the mean of the ith area, the ratio estimator $(\bar{y} / \bar{x})\bar{X}_i$ or the regression estimator $\bar{y}_i + b(\bar{X}_i - \bar{x}_i)$, where $(\bar{y}/\bar{x})$ and b are obtained from all the n sample observations, are examples of the synthetic estimator. As can be seen, the biases and MSEs of these estimators will be large if the above assumption is not satisfied. Replacing $(\bar{y}/\bar{x})$ and b by the corresponding quantities obtained from the small areas similar to the ith area can reduce these biases and MSEs.

For composite estimation, a sample estimator is combined with the synthetic estimator. As an example, $W_i\bar{y}_i + (1 - W_i)(\bar{y}/\bar{x})\bar{X}_i$, where the weights W_i are suitably chosen, can be considered for the mean of the ith small area. Similar types of estimators were considered, for example, by Wolter (1979). Schaible et al. (1977) compare the synthetic and composite estimators. Drew et al. (1982) examine the composite estimator with data from the Canadian Labour Force survey. Särndal and Hidiroglou (1989) consider a composite estimator for estimating the wages and salaries for industries in census divisions, and suggest procedures for determining the weights. Longford (1999) presents procedures for multivariate estimation of small-area means and proportions.

Best linear unbiased predictor (BLUP)

For this procedure, the observations of the small areas are considered to constitute a random sample following the model

$$y_{ij} = \mu_i + \varepsilon_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \tag{12.15}$$

where $i = 1, 2, ..., k$ represents the areas and $j = 1, 2, ..., n_i$ represent the observations for the ith area. Note that the mean of the ith area

is μ_i, which is expressed as $\mu + \alpha_i$, where μ is the expected value of μ_i and α_i is the **random effect**. The random error ε_{ij} is assumed to have mean zero and variance σ_i^2. The random effect α_i, which is also known as the model error, is assumed to have mean zero and variance σ_α^2 and independent of ε_{ij}.

The BLUP for α_i is considered to be of the form $\hat{\alpha}_i = \Sigma c_{ij} y_{ij} + d$. The coefficients c_{ij} and d are obtained by minimizing $E(\hat{\alpha}_i - \alpha_i)^2$ with the unbiasedness condition $E(\hat{\alpha}_i - \alpha_i) = 0$. The BLUP for α_i can be expressed as $[\sigma_\alpha^2/(\sigma_\alpha^2 + v_i)](\bar{y}_i - \mu)$, where $v_i = V(\bar{y}_i) = \sigma_i^2/n_i$. As a result, the BLUP for μ_i is given by $\hat{\mu}_i = [\sigma_\alpha^2/(\sigma_\alpha^2 + v_i)]\ \bar{y}_i + [v_i/(\sigma_\alpha^2 + v_i)]\mu$. The error of this estimator is given by $E(\hat{\mu}_i - \mu_i)^2 = [\sigma_\alpha^2\ v_i/(\sigma_\alpha^2 + v_i)]$, which is smaller than v_i. For the BLUP, μ can be estimated from the WLS estimator $\Sigma W_i \bar{y}_i / W$, where $W_i = 1/(\sigma_\alpha^2 + v_i)$ and $W = \Sigma W_i$. P.S.R.S. Rao (1997), for instance, presents the derivation of the BLUP and different procedures for estimating σ_α^2 and v_i.

With the assumption that α_i and ε_{ij} follow independent normal distributions, the BLUP becomes the same as the Bayes estimator for the mean of the ith small area derived in Appendix 12A. The Bayes estimator and the BLUP obtained from the sample observations are known as the empirical Bayes estimator (EB) and the empirical best linear predictor (EBLUP). Ghosh and Meeden (1986), for example, describe the EB for the estimation in finite populations. Prasad and J.N.K. Rao (1990), Lahiri and J.N.K. Rao (1995), and Bell (1999) present procedures for estimating the MSE of the BLUP.

Applications, extensions, and evaluations

Estimation procedures for small areas using census data and additional information from administrative records or similar sources were described by Purcell and Kish (1980) and Bell (1996). Ghosh and J.N.K. Rao (1994) review the different procedures suggested for improved estimation in small areas.

In classical statistics, the model in (12.15) is known as the **One-Way Variance Components Model**, and it can be adapted as above to estimate the means and totals of small areas in finite populations. With supplementary variables or predictors x_1, x_2, ..., the mean μ can be replaced by $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots$. Suitable to the application, the intercept β_0 and some of the coefficients β_1, β_2, ..., are assumed to be random. The BLUP can be considered for the resulting model.

To predict the number of children from low-income families in the counties described above, a model of the type

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \alpha_i + \varepsilon_i \tag{12.16}$$

was employed. In this model, y_i is the logarithm of the 3-year weighted average of the number of children in the ith county. The predictors (x_1, x_2, x_3, x_4, x_5) included the logarithms of the child exemptions reported by families, population under 18 years of age, and other related variables. The sampling error of y_i is represented by ε_i and the model error by α_i. When the ith county is considered to be selected randomly from an infinite population, both these errors become random.

Models of the type in (12.15) and (12.16) were considered by Fay and Herriot (1979) for estimating per capita incomes in regions of the U.S. with fewer than 500 persons, by Ghosh and Lahiri (1987) for small-area estimation with stratified samples, by Dempster and Raghunathan (1987) to estimate wages paid by small business firms, by Battese et al. (1988) for the estimation of agricultural production in segmented areas, and by Stroud (1987) to estimate the average number of times students visit their homes during an academic year. Holt et al. (1979) derive the BLUP for the model of the type in (12.16). Ericksen (1974), Särndal (1984), and Sarndal and Hidiroglou (1985) use the regression-type models for small-area estimation.

The above types of models were also used by Stasny (1991) for a survey on crime, by Stasny et al. (1991) to estimate wheat production in counties, and by Fay and Train (1995) and Fisher and Siegel (1997) for estimating income in different states of the U.S.

Small-area estimation for census undercount was examined by Isaki et al. (1987) and others. Cressie (1989; 1992), Datta et al. (1992), and others consider the EB approach for this purpose. Zaslovsky (1993) uses the EB to estimate the population shares through census data and additional information. Wolter and Causey (1991) examine some of the procedures employed for the estimation of population in small areas. Estimation for small areas from the U.S. National Health Interview Survey was described by Marker (1993) and from the U.S. National Medical Expenditure Survey by Cohen and Braden (1993). Estimation for local areas and regions in Sweden was described by Lundstrom (1987) and Ansen et al. (1988) and in Finland by Lehtonen and Veijanen (1999). Pfeffermann and Burck (1990) and J.N.K. Rao and Yu (1992) present estimation for small areas from surveys conducted over periods of time.

12.7 Complex surveys

Many of large-scale surveys, especially nationwide, employ stratification and clustering of population units, and they are usually conducted in two or more stages. Samples at different stages are selected with equal or unequal probabilities, and the ratio or regression methods are employed for estimation of the population quantities. Nonresponse adjustments of the type described in Chapter 11 are also implemented in several cases. With all these factors, estimation of population quantities, finding standard errors of the estimators, and analysis of the data can become very **complex**.

Regression analysis for complex survey data was described by Holt et al. (1980) and Nathan (1988) among others. J.N.K. Rao and Scott (1981) describe the analysis of categorical data from complex surveys. For qualitative characteristics, Binder (1983) considers logistic regression and estimation of the standard errors. Through an empirical investigation, Katzoff et al. (1989) evaluate some of the procedures for analyzing complex survey data. Some of the methods of analyzing data from complex surveys is presented, for example, in the volume edited by Skinner et al. (1989) and in Lehtonen and Pahkinen (1995). Binder and Patak (1994) present procedures for point and interval estimation with complex survey data. J.N.K. Rao et al. (1992) review the bootstrap, jackknife, and BRR methods for variance estimation and confidence intervals with data from complex surveys. Skinner and J.N.K. Rao (1996) describe the estimation in complex surveys through dual frames. The linearization and jackknife procedures for the inference from complex surveys with multiple frames was described by Lohr and J.N.K. Rao (2000).

Exercises

12.1. (a) Since the sample mean is unbiased, show that jackknifing it will again result in the same estimator. (b) For the sample of the 1995 total enrollment in Table T4 in the Appendix, find the ten pseudo values and show that the jackknife estimator coincides with the sample mean.

12.2. The estimator $(n-1)s^2/n$ is biased for the population variance S^2. (a) Show that jackknifing this estimator results in the unbiased estimator s^2. (b) Verify this result from the sample of the 1995 total enrollments in Table T4 in the Appendix.

12.3. For a sample of ten states in the U.S., the 1995 expenditures for hospital care (y) and physician services (x) are presented in Table T7 in the Appendix. For estimating the ratio of the totals of y and x, compare $\hat{R}$, $\hat{R}_T$, and $\hat{R}_J$.

12.4. For estimation of the ratio in Exercise 12.3, compare the variance estimator in (9.2) and the jackknife estimator in (12.6).

12.5. (a) From the sample of the ten states in Table 12.1, find the regression coefficient b for the regression of the private enrollment (y) on the total enrollment (x). Denote by b_j the regression coefficient obtained by omitting one of the sample observations. The pseudo values are given by $nb - (n-1)b_j$. The jackknife estimator for the regression coefficient is $b_J = nb - (n-1)b'$, where $b' = \Sigma b_j/n$. (b) Find this estimate from the above sample, and compare it with the sample regression coefficient. (c) From the pseudo values, find the jackknife variance for b.

12.6. For the problem in Exercise 12.5, generate 100 bootstrap samples of size $m = (n-1)$ *with* replacement as described in Section 12.4, find the variance estimate of the first type in (12.8), and compare it with the jackknife estimator in Exercise 12.5(c).

12.7. The motivation for formulating the variance estimator in (12.6) is that the pseudo values in (12.3) are approximately uncorrelated. Show that this result is valid.

Appendix A12

The Bayesian approach

Consider a large group or an infinite population following a specified distribution, for example, normal with mean μ_i and variance σ_i^2. For given (μ_i and σ_i^2), the distribution of the mean $\bar{y}_i$ of a sample of size n_i from this population, which can be denoted by $f(\bar{y}_i \mid \mu_i, \sigma_i^2)$, follows the normal distribution with mean m and variance σ_i^2/n_i.

The mean μ_i is assumed to have a *prior distribution* $g(\mu_i)$, for example, normal with mean μ and variance σ_α^2. Now, the *joint distribution* of $\bar{y}_i$ and μ_i are given by $f(\bar{y}_i \mid \mu_i, \sigma_i^2)g(\mu_i)$. The *marginal distribution* of $\bar{y}_i$, denoted by $h(\bar{y}_i)$ is obtained from this joint distribution. The *posterior distribution*

$k(\mu_i \mid \bar{y}_i)$, which is the distribution of μ_i for given $\bar{y}_i$, is obtained by dividing $f(\bar{y}_i \mid \mu_i, \sigma_i^2)g(\mu_i)$ by $h(\bar{y}_i)$.

With the assumption for normality for $f(\bar{y}_i \mid \mu_i, \sigma_i^2)$ and $g(\mu_i)$ as above, $k(\mu_i \mid \bar{y}_i)$ becomes the normal distribution with mean $E(\mu_i \mid \bar{y}_i) = \mu + \beta(\bar{y}_i - \mu)$ and variance $V(\mu_i \mid \bar{y}_i) = \sigma_\alpha^2(\sigma_i^2/n_i)/(\sigma_\alpha^2 + \sigma_i^2/n_i)$. Notice that this mean can also be expressed as $E(\mu_i \mid \bar{y}_i) = a\mu + (1 - a)\bar{y}_i$, where $a = (\sigma_i^2/n_i)/(\sigma_\alpha^2 + \sigma_i^2/n_i)$. The weights for this posterior mean are inversely proportional to σ_α^2 and σ_i^2/n_i, and $V(\mu_i \mid \bar{y}_i)$ is smaller than both these variances. Further, it approaches $\bar{y}_i$ as n_i becomes large. For the EB approach, $E(\mu_i \mid \bar{y}_i)$ is obtained by estimating μ, σ_α^2, and σ_i^2 from the sample observations drawn from all the groups.

In general, $f(\bar{y}_i \mid \mu_i, \sigma_i^2)$ and $g(\mu_i)$ need not be normal. For such cases, $E(\mu_i \mid \bar{y}_i)$ may not be a simple linear combination of the prior mean μ and the sample mean $\bar{y}_i$. For Bayesian analysis, the entire posterior distribution or its mode are examined. For some applications, a prior distribution for σ_i^2 is also specified.