CHAPTER 11

Nonresponse and Remedies

11.1 Introduction

In almost every survey, some of the persons, households, and other types of units selected into the sample are not contacted. Persons away from home on business or vacation, wrong addresses and telephone numbers, households without telephones or with unlisted numbers, inability of the interviewers to reach households in remote places, and similar reasons contribute to the noncontacts. Even if they are contacted, some of the units may not respond to one or more characteristics of the survey. As noted in Section 3.7, estimators obtained from the responding units alone are biased, and increasing the sample size may reduce their variances but not the biases.

Public polls are also frequently affected by the **noncontacts** and nonresponse. In the Truman–Dewey presidential contest, although Truman emerged as the winner from the final count of the ballots, *The Literary Digest*, a newspaper in Chicago, initially declared Dewey the winner. A badly conducted poll and a large amount of nonresponse were blamed for this “fiasco.” Mosteller et al. (1949) analyzed the effects of nonresponse in this survey.

Cochran (1977, Chap. 12) presents some of the reasons for nonresponse and the procedures to counteract its effects. A review of progress in sample surveys including efforts to reduce the bias arising from nonresponse was provided by Cochran (1983). Godambe and Thompson (1986) describe theoretically optimum procedures of estimation in the presence of nonresponse. Little and Rubin (1987) present the general methods available for estimating missing observations in statistical designs of experiments, regression analysis, and other types of statistical procedures.

Surveys are also affected by errors of observation and measurement. Nonresponse together with these types of errors are known as **nonsampling errors**. Mahalanobis (1946) recommends the procedure of **interpenetrating subsamples**, in which independent samples are assigned to different interviewers. The above types of errors are assessed from the variation among the estimates of the interviewers. Lessler and Kalsbeek (1992) describe the effects of nonresponse and errors of measurements. Biemer et al. (1991) examine the effects of measurement errors.

Some of the exercises of previous chapters have examined simple procedures of adjusting estimates for nonresponse. In the following sections, the reasons for nonresponse and the procedures for reducing its effects are presented in detail.

11.2 Effects of survey topics and interviewing methods

Some of the units selected into the sample may not respond to the entire survey and some may provide answers only to selected items in the questionnaire. These two cases are classified as **unit nonresponse** and **item nonresponse**, respectively. A number of factors affect response rates, and some of the main reasons are described below.

Subject matter

To a large extent, response rates depend on the interest of the sampled units in the subject matter of the survey. Response rates of the public tend to be high on matters related to educational reforms, local improvements, political views, tax changes, and similar characteristics. Concerned citizens can be expected to respond readily to topics of national interest. Low response rates are frequently observed in surveys on personal characteristics, such as incomes and savings.

Types of interviewing

Surveys are conducted through the mail (post), e-mail, telephone, face-to-face personal interviews, or a combination of these procedures. Dillman (1978) describes surveys conducted through the mail and telephone. Nonresponse in telephone surveys is described in the book edited by Groves et al. (1988). Collins (1999) summarizes the response rates of the telephone interviews in the U. K. Sirken and Casady (1982) describe the nonresponse rates in the dual frame surveys in the U.S. in which samples are selected from the lists and telephone numbers of the population units. From the Canadian Labour Force Survey, Gover (1979) notes that response rates can differ with the interviewers.

In the **random digit-dialing** method, the telephone digits of people to be included in the sample are selected randomly; see Waksberg (1978), for example. This approach is practiced in some types of marketing surveys and for public polls, as an attempt to obtain responses from the persons with both listed and unlisted telephone numbers. Computer-aided telephone-interviewing (CATI) is being tested in both industrialized and developing nations; see Groves and Nicholls (1986), for example. Martin et al. (1993) describe Computer-aided personal interviewing (CAPI). Response rates have been found to be different for all these methods of conducting the surveys.

In some surveys, telephone reminders and follow-up letters are employed to increase response rates. In **callback** surveys, more than one attempt is made to contact and elicit responses from the sample units. Personal interviews are usually found to yield higher response rates than mail and telephone surveys.

Types of respondents

In some household surveys, responses to the survey questions are requested from **any adult** in the household. For some types of surveys, **specified** persons such as parents, heads of households, supervisors, and managers are asked to provide the required information. In some other types of surveys, the respondent is **randomly chosen** from all the persons eligible to provide answers to the items of the survey. Response rates are usually found to be relatively higher for the first type of respondent and lower for the last.

11.3 Response rates

A number of studies summarize the response rates of surveys. For example, Bailar and Lanphier (1978) found the response rates for 36 surveys conducted by the government and private organizations to be low; for three of these surveys, the rates were only 25, 46, and 50% in spite of several attempts to obtain the responses. From analyzing a large number of surveys, P.S.R.S. Rao (1983c) found that response rates for the *any*, *specified*, and *random adults* at the initial attempt were (69.4, 51.7, 33.1)% and (94.9, 89.1, 83.9)% after three or more attempts.

11.4 Bias and MSE

To examine the bias due to nonresponse, the population may be considered to consist of two strata or groups, with **respondents** and **nonrespondents**. Members of the first group provide answers to a question on the survey, if they were contacted. The second group does not provide answers even after repeated attempts.

In Section 3.7, the size of the responding group, their mean, and variance are denoted by N_1, $\bar{Y}_1$, and S_1^2, respectively. Similarly, $N_2 = N - N_1$, $\bar{Y}_2$, and S_2^2 denote the corresponding figures for the nonrespondents. The population mean now can be expressed as $\bar{Y} = W_1\bar{Y}_1 + W_2\bar{Y}_2$, where $W_1 = N_1/N$ and $W_2 = N_2/N$.

The number of respondents and nonrespondents in a simple random sample of n units from the population are denoted by n_1 and $n_2 = n - n_1$. The sample mean $\bar{y}_1$ is unbiased for $\bar{Y}_1$. As shown in Section 3.7, its bias, variance, and MSE for estimating $\bar{Y}$ are given by $B(\bar{y}_1) = \bar{Y}_1 - \bar{Y} = W_2(\bar{Y}_1 - \bar{Y}_2)$ and $V(\bar{y}_1) = (1 - f_1)S_1^2/n_1$ and MSE $(\bar{y}_1) = V(\bar{y}_1) + B^2(\bar{y}_1)$, where $f_1 = n_1/N_1$.

An unbiased estimator for the total $Y_1 = N_1\bar{Y}_1$ of the respondents is $\hat{Y}_1 = N_1\bar{y}_1$. Since N_1 is not known, an unbiased estimator can be obtained by replacing it with $N(n_1/n)$.

The sample ratio $\hat{R}_1 = \bar{y}_1/\bar{x}_1$ is approximately unbiased for $R_1 = \bar{Y}_1/\bar{X}_1$ of the respondents. Its bias in estimating $R = \bar{Y}/\bar{\mathrm{X}}$ will be small only if $R_2 = \bar{Y}_2/\bar{X}_2$ for the nonresponse group does not differ much from R_1. Similarly, the expected value of the regression estimator $\bar{y}_1 + b_1(\bar{X} - \bar{x}_1)$ is approximately equal to $\bar{Y}_1 + \beta_1(\bar{X} - \bar{X}_1)$, where β_1 and b_1 are the population and sample slopes for the respondents. Hence, its bias will be negligible only if the respondents and nonrespondents do not differ much in the means of x and y and the slopes.

In some situations, the n_1 responses may be considered to be a random sample from the n sampled units. This will be the case, for example, if the interviewer elicits responses through intensive efforts from a random selection of n_1 of the n sampled units. For such a situation, $\bar{y}_1$ is unbiased for $\bar{Y}$ and its variance is given by $V(\bar{y}_1) = (N - n_1)S^2/Nn_1$. An unbiased estimator of this variance is obtained by replacing S^2 by the sample variance of the respondents, $s_1^2 = \Sigma_1^{n_1}(y_i - \bar{y}_1)^2/(n_1 - 1)$.

11.5 Estimating proportions

As in Chapter 4, let C and $P = C/N$ denote the number and proportion of units in the population having a specified attribute. Similarly, let C_1 and C_2 denote the numbers of units having the attribute in the response and nonresponse groups. The corresponding proportions are denoted by $P_1 = C_1/N_1$ and $P_2 = C_2/N_2$.

Among the n_1 respondents of the sample, c_1 will be observed to have the attribute. The sample proportion $p_1 = c_1/n_1$ is unbiased for P_1, and its bias for estimating the population proportion P is $W_2(P_1 - P_2)$. The absolute value of this bias increases with the proportion of the nonrespondents W_2 and the difference between P_1 and P_2.

An unbiased estimator of the variance of p_1 is $v(p_1) = (N_1 - n_1)p_1 \times (1 - p_1)/N_1(n_1 - 1)$. Since N_1 is not known, it may be replaced by its estimator Nn_1/n. With the normal approximation, $(1 - \alpha)$% confidence limits for P_1 are obtained from

$$p_1 \pm Z\sqrt{v(p_1)}. \tag{11.1}$$

Example 11.1. Proportions: Consider a random sample of $n = 500$ units from a large population in which only $n_1 = 300$ units respond. If 120 of the respondents are observed to have the characteristic of interest, $p_1 =$ 120/300 or 40%. The variance of this estimate is $0.4(0.6)/299 = 0.0008$, and hence it has a S.E. of 0.0283. Approximate 95% confidence limits for P_1 are given by $0.4 \pm 1.96(0.0283)$; that is, (0.34, 0.46).

If one considers the n_1 responses to be a random sample from the n units of the initial sample, p_1 is unbiased for P. In this case, the variance of p_1 becomes $(N - n_1)\, PQ/(N - 1)n_1$, where $Q = 1 - P$. Now, the limits (0.34, 0.46) in the above example refer to P, the population proportion.

11.6 Subsampling the nonrespondents

Deming (1953) presented a model for studying the effectiveness of callbacks. Through this model and the data from practical surveys, P.S.R.S. Rao (1983c) found that for a variety of surveys on an average three calls are required to obtain high response rates. Since callbacks can be expensive, Hansen and Hurwitz (1946) suggest eliciting responses from a subsample of the nonrespondents. For the case of mail surveys at the initial attempt and personal interviews at the second stage, they present the following estimator for the mean and its variance. The procedure of subsampling, however, is applicable for any method of conducting a survey. A summary of this approach appears in P.S.R.S. Rao (1983b).

The procedure of Hansen and Hurwitz

As before, let N_1 and $N_2 = N - N_1$ denote the sizes of the response and nonresponse strata. In a simple random sample of n units from the population, n_1 responses and $n_2 = n - n_1$ nonresponses are obtained. A subsample of size $m = n_2/k$, with a predetermined value of k (>1), is drawn from the n_2 units and responses from all the m units are obtained.

Estimator and its variance

Let $\bar{y}_1$ and $\bar{y}_{2(m)}$ denote the means of the n_1 respondents at the first stage and the m subsampled units. An estimator for the population mean is given by

$$\bar{y}_H = w_1\bar{y}_1 + w_2\bar{y}_{2(m)}, \tag{11.2}$$

where $w_1 = n_1/n$ and $w_2 = n_2/n$. Note that for this estimator the mean of the m units at the second stage is inflated by w_2, and the mean $\bar{y}_2$ of the n_2 nonrespondents is not available.

For a given sample (s) consisting of n_1 respondents and n_2 nonrespondents $E(\bar{y}_H \mid s) = w_1\bar{y}_1 + w_2\bar{y}_2 = \bar{y}$, which is the mean of the n sampled units. Hence, $\bar{y}_H$ is unbiased for $\bar{Y}$, and from Appendix A11,

$$V(\bar{y}_H) = \frac{1-f}{n}S^2 + W_2\frac{k-1}{n}S_2^2. \tag{11.3}$$

The second term is the increase in the variance due to subsampling the nonrespondents. This increase will be large if the size of the second stratum or its variance is large. It can be reduced by increasing the subsampling fraction $1/k$.

Optimum sample sizes

For the above procedure, the cost of sampling is considered to be of the form

$$E' = e_0 n + e_1 n_1 + e_2 n_2, \tag{11.4}$$

where e_0 is the initial cost for arranging for the survey. The cost for obtaining the required information from a respondent at the first and second stages are denoted by e_1 and e_2, respectively. From (11.4), the expected cost becomes

$$E = (e_0 + e_1 W_1 + e_2 W_2/k)n. \tag{11.5}$$

Minimizing the variance in (11.3) for given cost in (11.5), or the cost for given variance V, is the same as minimizing $(V + S^2/N)E$. From this minimization, the optimum value of k is given by

$$k = \left[\frac{e_2(S^2 - W_2 S_2^2)}{(e_0 + e_1 W_1)S_2^2}\right]^{1/2}. \tag{11.6}$$

Thus, the size of the subsample will be large if e_2 is large relative to $(e_0 + e_1 W_1)$. Since additional effort is needed to elicit responses from a subsampled nonrespondent, e_2 is usually larger than e_1. For a specified V, from (11.3) and (11.6), the optimum value for the size of the initial sample is obtained from

$$\begin{aligned} n &= \frac{N[S^2 + (k-1)W_2 S_2^2]}{NV + S^2} \\ &= n_0\left[1 + \frac{(k-1)W_2 S_2^2}{S^2}\right], \end{aligned} \tag{11.7}$$

where $n_0 = NS^2/(NV + S^2)$ is the sample size required when $W_2 = 0$. For fixed E, the optimum sample size from (11.5) and (11.6) is given by

$$n = \frac{kE}{k(e_0 + e_1 W_1) + e_2 W_2}. \tag{11.8}$$

For finding the optimum sizes of the samples at the two stages, it is enough to know the relative value S_2^2/S^2 of the variances of the nonrespondents and the population. The value of W_2, however, should be known.

Example 11.2. Sample sizes and costs for a specified precision: Consider a population of $N = 30{,}000$ units with the standard deviation $S = 648$ for a characteristic of interest. If there is no nonresponse, for a sample of $n_0 = 2000$ units from this population, the variance of the sample mean is $(N - n_0)S^2/Nn_0 = 196$.

> One may require that when there is nonresponse, the variance in (11.3) should be the same, that is $V = 196$. For the costs, consider $e_0 = 0.5$, $e_1 = 1$, and $e_2 = 4$. Now, if $W_1 = 0.7$ and $S^2 = 1.8S_2^2$, from (11.6), $k = 2.23$ or $1/k = 0.45$; that is, 45% of the nonrespondents should be sampled at the second stage. From (11.7), the optimum size for the initial sample is $n = 2410$. From (11.5) this survey would cost on the average \$4189. Without nonresponse, it would cost $(e_0 + e_1)n_0$, that is, \$3000.

Since the subsampling fraction depends on W_2, Srinath (1971) suggests a modification, which was examined by P.S.R.S. Rao (1983a). Särndal and Swensson (1985) consider unequal probability sampling at both stages. The above procedure of subsampling nonrespondents can be extended to the case of stratification.

Variance estimation

For convenience, denote the mean $\bar{y}_{2(m)}$ and the variance of the m units $s_{2(m)}^2 = \Sigma^m(y_1 - \bar{y}_{2(m)})^2/(\mathrm{m}-1)$ by $\bar{y}_m$ and s_m^2. This variance is unbiased for s_2^2 of the $n_2 = n - n_1$ nonrespondents and hence for S_2^2. The sample variance s^2 of the n units is unbiased for S^2, but it is not available due to the nonresponse. An unbiased estimator for S^2 and $V(\bar{y}_H)$ are obtained in Appendix A11 from the derivations of Cochran (1977, p. 333) and J.N.K. Rao (1973). For large N, this estimator becomes

$$\begin{aligned} v(\bar{y}_H) = [&(n_1 - 1)s_1^2 + (n_2 - 1)ks_m^2 + n_1(\bar{y}_1 - \bar{y}_H)^2 \\ &+ n_2(\bar{y}_m - \bar{y}_H)^2]/n(n-1). \end{aligned} \tag{11.9}$$

Ratio and regression estimators

As in the case of $\bar{Y}$, an unbiased estimator for the mean $\bar{X}$ of an auxiliary characteristic is obtained from

$$\bar{x}_H = w_1\bar{x}_1 + w_2\bar{x}_{2(m)}. \tag{11.10}$$

In this expression, $\bar{x}_1$ and $\bar{x}_{2(m)}$ are the means of the n_1 respondents and the m subsampled units. Now, the ratio estimator for $\bar{Y}$ is given by

$$\bar{y}_{HR} = (\bar{y}_H/\bar{x}_H)\bar{X} = \hat{R}_H\bar{X}. \tag{11.11}$$

For large n and m, following the approach in Appendix A11, the variance of $\bar{y}_{HR}$ is obtained from

$$V(\bar{y}_{HR}) = \frac{(1-f)}{n}S_d^2 + W_2\frac{(k-1)}{n}S_{d2}^2, \tag{11.12}$$

where $S_d^2 = \Sigma_1^N(y_i - Rx_i)^2/(N-1)$ and $S_{d2}^2 = \Sigma_1^{N2}(y_{2i} - Rx_{2i})^2/(N_2 - 1)$. The estimator in (11.11) and the variance in (11.12) were suggested by Cochran (1977, p. 374).

For the regression estimator of $\bar{Y}$, let

$$A = \sum_1^{n1}(x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1) + (n_2 - 1)\sum_1^m(x_{2i} - \bar{x}_m)(y_{2i} - \bar{y}_m)/(m-1)$$

and

$$B = \sum_1^{n1}(x_{1i} - \bar{x}_1)^2 + (n_2 - 1)\sum_1^m(y_{2i} - \bar{y}_m)^2/(m-1). \tag{11.13}$$

Note that the subscript m refers to the subsample. An estimator for the slope is given $b = A/B$. For $\bar{Y}$, one can now consider

$$\bar{y}_{Hl} = \bar{y}_H + b(\bar{X} - \bar{x}_H). \tag{11.14}$$

For large n and m, the bias of this estimator is negligible and its variance approximately becomes

$$V(\bar{y}_{Hl}) = \frac{(1-f)}{n}S_e^2 + W_2\frac{(k-1)}{n}S_{e2}^2, \tag{11.15}$$

where $S_e^2 = \Sigma_1^N[(y_i - \bar{Y}) - \beta(x_i - \bar{X})]^2/(N-1)$ and $S_{e2}^2 = \Sigma_1^{N_2}[(y_{2i} - \bar{Y}_2) - \beta(x_{2i} - \bar{X}_2)]^2/(N_2 - 1)$.

There may not be any nonresponse for auxiliary variables such as family size and years of education. In such cases, $\bar{x} = \Sigma_1^n x_i/n$ will be available and an alternative ratio estimator for $\bar{Y}$ is given by $(\bar{y}_H/\bar{x})\bar{X}$, and a similar regression estimator can be found. P.S.R.S. Rao (1987, 1990) examines the biases and MSEs of these types of estimators.

11.7 Estimating the missing observations

Procedures developed for estimating the missing observations in statistical designs of experiments and regression analysis can also be utilized for predicting the observations of the nonrespondents.

Least squares estimation

For statistical analysis, Yates (1933) recommended estimation of the missing observations through the least squares principle. For the regression model in (10.5), if x is observed on all the n units, but y is observed on only n_1 units, the n_2 missing values of y along with the coefficients (α, β) can be estimated by minimizing

$$\delta = \sum_1^{n_1} (y_i - \alpha - \beta x_i)^2 + \sum_1^{n_2} (y_i - \alpha - \beta x_i)^2. \qquad (11.16)$$

This optimization results in

$$\hat{\beta} = b_1 = \frac{s_{xy1}}{s_{x1}^2}, \qquad \hat{\alpha} = \bar{y}_1 - b_1 \bar{x}_1$$

and

$$\hat{y}_i = \bar{y}_1 + b_1(x_i - \bar{x}_1), \qquad (11.17)$$

for $i = (1, ..., n_2)$. With these estimates,

$$\delta = \sum_1^{n_1} (y_i - \hat{\alpha} - \hat{\beta} x_i)^2 = \sum_1^{n_1} [(y_i - \bar{y}_1) - b_1(x_i - \bar{x}_1)]^2, \quad (11.18)$$

which is the same as the residual sum of squares computed from the n_1 completed observations on (x_i, y_i). Note that from (11.17), the average of the predictions for the n_2 missing observations becomes $\bar{y}_1 + b_1(\bar{x}_2 - \bar{x}_1)$.

For the regression through the origin, if $V(y_i \mid x_i)$ is proportional to x_i, following the model in (9.10),

$$\delta = \sum_1^{n_1} [(y_i - \beta x_i)^2 / x_i] + \sum_1^{n_2} [(y_i - \beta x_i)^2 / x_i] \qquad (11.19)$$

is minimized. The estimator of the slope β now is given by $b_1 = \bar{y}_1/\bar{x}_1$, and the observations of the n_2 nonrespondents are predicted from $b_1 x_i$. In this case, the average of the predictions for the n_2 missing observations becomes $(\bar{y}_1/\bar{x}_1)\bar{x}_2$.

> Example 11.3. Survey on families: Table 11.1 presents data on five variables from a simple random sample of 16 from 2000 families, along with the number of responses, means, and standard deviations. For income, mid-values of the ranges (in 1000s) 15 to 20, 20 to 25,... are considered.

Table 11.1. Survey of families.

No.	Family Size, x	Husband's Age, h	Wife's Age, w	Income Level, y	Television Time, t
1	3	35	32	62.5	15
2	4	45	45	59.5	12
3	5	45	—	48.5	15
4	3	28	26	22.5	12
5	5	48	—	72.5	18
6	5	50	—	49.5	12
7	2	26	26	47.5	10
8	3	28	25	37.5	12
9	4	36	32	52.5	—
10	2	27	23	22.5	—
11	4	44	—	62.5	—
12	2	29	29	32.5	—
13	4	46	38	—	14
14	3	32	28	—	12
15	5	44	—	—	15
16	6	45	35	—	14
Responses	16	16	11	12	12
Mean	3.75	38	30.82	47.50	13.42
S.D.	1.24	8.65	6.52	16.02	2.15

	Husband's Age			Wife's Age		
Source	d.f.	SS	F	d.f.	SS	F
Regression	1	1465.4	10.8	1	888.6	6.3
Residual	10	1358.6		6	849.3	
Total	11	2824		7	1737.9	

$\hat{y}_i =$	1.07	$+ 1.26h_i$	$\hat{y}_i =$	-6.0	$+ 1.62w_i$
S.E.	14.50	0.38	S.E.	19.7	0.65
t_{10}	0.07	3.28	t_6	−0.31	2.51
p-value	0.94	0.008	p-value	0.71	0.046

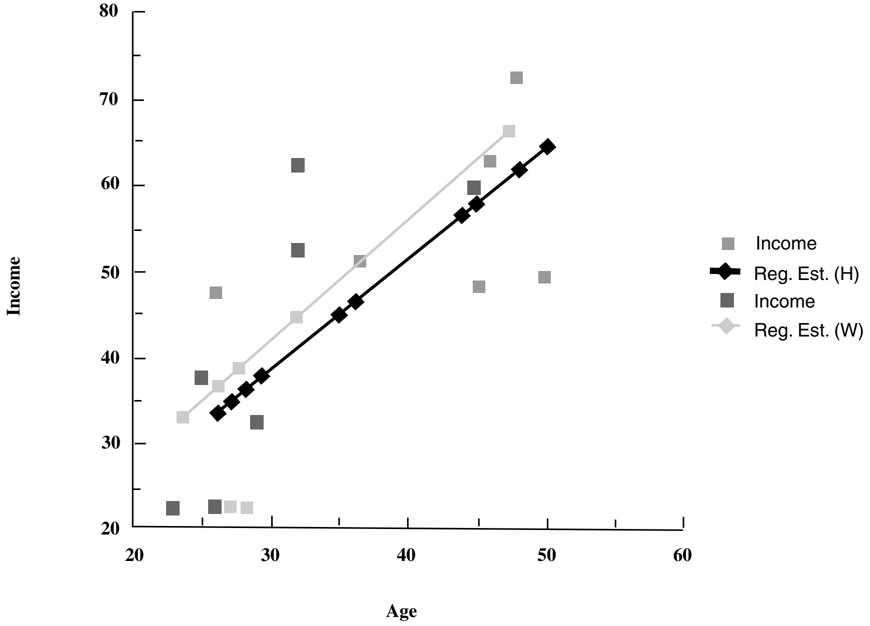

Figure 11.1. Regressions of income on husband's age and wife's age.

Regression of income on husband's age for the 12 pairs of completed observations and on wife's age for the eight pairs are presented in Figure 11.1, along with the summary figures needed for statistical analysis. For both the regressions, the slopes are significantly larger than zero but not the intercepts. If these regressions are still used from the first regression, prediction of the missing income at husband's age h = 46 is 1.07 + 1.26(46) = 59.03.

For the 12 pairs of responses on both husband's age and income, $\bar{h}_1 = 36.75$ and $\bar{y}_1 = 47.5$. For the four nonresponses on income, the mean for husband's age is $\bar{h}_2 = 41.75$. Since $b_1 = 1.26$, prediction of the average income $\bar{y}_2$ for the four nonrespondents from the regression on husband's age is 47.5 + 1.26(41.75 − 36.75) = 53.8.

One may now predict the missing observations on income from the regression on wife's age. At $w = 28$, prediction for the income is −6 + 1.62(28) = 39.36. Similarly, at wife's ages of 35 and 38, predictions for income are 50.7 and 55.56.

As described above, for the regression through the origin of income on husband's age, $b_1 = 47.5/36.75 = 1.29$. The prediction of income at h = 46 now is 1.29(46) = 59.34, which is slightly larger than 59.03 for the regression with the intercept.

Similarly, for the regression through the origin of income on wife's age with the eight pairs of responses, b_1= 42.125/29.75 = 1.42. Now, prediction for the income at w = 28 is 1.42(28) = 39.76. At w = 35 and 38, the predictions are 49.7 and 53.96.

One should note that for this type of predicting the missing observations, it is assumed that the slopes as well as the intercepts for the responding and nonresponding group are the same. Further, in the above illustration, multiple regression of income on both husband's and wife's age can be attempted, but it will be based on only eight completed sets of observations.

Alternative predictions

In the procedure suggested by Buck (1960), the missing observations on each variable are predicted by fitting its regression on the remaining variables, using the completed observations on all the variables. For example, in the case of two variables x and y, the observations on y are predicted from the regression of y on x. Similarly, the missing observations on x are predicted from its regression on y. Schafer (1997) describes statistical analysis for the case of missing observations on more than one variable.

Hendricks (1949) made one of the earliest suggestions for predicting the percentage for an attribute from the trend of the percentages at the successive attempts in a mail survey. As an illustration, suppose the response rates for three successive attempts are (60, 20, 10)%, with the corresponding percentages (54, 47, 52) favoring an attribute in the survey. For the regression of y_i = (54, 47, 52) on x_i = (1, 2, 3), the slope and intercept coefficients are $b = -0.5$ and $a = 52$. Hence, the equation for predicting the percentage at the successive attempts is $y_i = 52 - 0.5x_i$. From this expression, at $x_i = 4$, $y_i = 52 - 0.5(4) = 50$. Now, an estimate for the population percentage favoring the attribute is 54(0.60) + 47(0.20) + 52(0.10) + 50(0.10) = 52.

11.8 Ratio and regression estimation

Ratio estimation

If $\overline{X}$ is known, a ratio-type estimator for $\overline{Y}$ is $\bar{y}_R = (\bar{y}_1/\bar{x}_1)\overline{X} = \hat{R}_1\overline{X}$, where $(\bar{x}_1, \bar{y}_1)$ are the means of the n_1 respondents and $\hat{R}_1 = \bar{y}_1/\bar{x}_1$. If the respondents are considered to be a random sample from the n sample units, for large samples, $V(\bar{y}_R) = (N - n_1)(S_y^2 + R^2S_x^2 - 2RS_{xy})/Nn_1$.

If $\bar{X}$ is not known, it can be estimated by the mean $\bar{x}$ of the n units, provided there is complete response on the auxiliary characteristic. In this case, $\bar{Y}$ can be estimated from $\bar{y}_r = \hat{R}_1\bar{x}$. With the assumption of random responses, for large samples, the variance of this estimator is given by $(N-n)S_y^2/Nn + (n-n_1)(S_y^2 + R^2S_x^2 - 2RS_{xy})/nn_1$.

The large sample variances of both $\bar{y}_R$ and $\bar{y}_r$ are smaller than the variance $(N-n_1)S_y^2/Nn_1$ of $\bar{y}_1$, provided x and y are highly correlated. These variances can be estimated by replacing (S_x^2, S_y^2, S_{xy}) with $(s_{x1}^2, s_{y1}^2, s_{xy1})$ obtained from the n_1 responses and R by $\hat{R}_1$.

Notice that the sample mean can be expressed as $\bar{y} = (n_1/n)\bar{y}_1 + (n_2/n)\bar{y}_2$. When there is nonresponse on the n_2 units, if $\bar{y}_2$ is replaced by the predicted value $(\bar{y}_1/\bar{x}_1)\bar{x}_2$ obtained in Section 11.7, the estimator for the population mean now takes the form of $\bar{y}_r$. This estimator was also obtained by Jackson and P.S.R.S. Rao (1983) with suitable assumptions regarding the nonrespondents.

Example 11.4. Survey of families: The mean and variance of the 12 responses on income are $\bar{y}_1 = 47.5$ and $s_{y1}^2 = 256.73$. Hence, $v(\bar{y}_1) = (2000 - 12)(256.73)/2000(12) = 21.27$ and S.E.$(\bar{y}_1) = 4.61$. For the corresponding responses on husband's age, the mean and variance are $\bar{h}_1 = 36.75$ and $s_{h1}^2 = 83.48$, and its covariance with income is $s_{hy1} = 105.46$.

If the population mean for husband's age is 35, the ratio estimate for the average income is (47.5/36.75)(35) = 45.2. With the sample means and variances, an estimate of $V(\bar{y}_R)$ becomes $v(\bar{y}_R) = 10.24$, and hence S.E.$(\bar{y}_R) = 3.2$.

The mean of husband's age for the 16 sampled units is $\bar{h} = 38$. Now, the ratio estimator for the average income is $\bar{y}_r = (47.5/36.75)(38) = 49.1$. For this estimator, $v(\bar{y}_r) = 18.49$ and hence S.E.$(\bar{y}_r) = 4.3$.

In this illustration, for the average income, the sample mean 47.5 differs only a little from the ratio estimates 45.2 and 49.1. Also, the S.E. of $\bar{y}_R$ is only about two thirds of the S.E. of $\bar{y}_r$. Further, the decrease in the S.E. of $\bar{y}_r$ from $\bar{y}_1$ is not significant, since the population mean of husband's age is estimated from the sample.

Regression estimation

With the data from the n_1 respondents, a linear regression estimator for $\bar{Y}$ is given by $\bar{y}_L = \bar{y}_1 + b_1(\bar{X} - \bar{x}_1)$, where $b_1 = s_{xy1}/s_{x1}^2$. If the n_1 responses are considered to be random as before, for large samples, $V(\bar{y}_L) = (N-n_1)S_y^2(1-\rho^2)/Nn_1$ This variance can be estimated by replacing S_y^2 with s_{y1}^2 and $S_y^2(1-\rho^2)$ with $s_{e1}^2 = s_{y1}^2 - b_1^2 s_{x1}^2$ obtained from the n_1 observations.

If $\bar{X}$ is not known, the regression estimator becomes $\bar{y}_l = \bar{y}_1 + b_1 (\bar{x} - \bar{x}_1)$. The variance of this estimator is approximately given by $V(\bar{y}_l) = (N - n)S_y^2/Nn + (n - n_1)S_y^2(1 - \rho^2)/nn_1$. This variance can also be estimated as above.

11.9 Poststratification and weighting

Estimators for the mean and total

If the n units of the entire sample are poststratified into G strata, n_g of the units will be observed in the gth stratum. If the sizes N_g of the strata are known, an estimator for $\bar{Y}$ is

$$\hat{Y}_W = \sum_1^G W_g \bar{y}_{g1}, \tag{11.20}$$

where $\bar{y}_{g1}$ is the mean of the n_{g1} respondents in the gth stratum and $W_g = N_g/N$.

Since $\bar{y}_{g1}$ is unbiased for the mean $\bar{Y}_{g1}$ of the N_{g1} respondents of the gth stratum, the bias of the above estimator is

$$B(\hat{\bar{Y}}_w) = \sum_1^G W_g \bar{Y}_{g1} - \sum_1^G W_g \bar{Y}_g = \sum_1^G \frac{N_{g2}}{N}(\bar{Y}_{g1} - \bar{Y}_{g2}) \tag{11.21}$$

In this expression, $N_{g2} = N_g - N_{g1}$ is the size of the nonrespondents in the gth stratum and $\bar{Y}_{g2}$ is their mean. If (N_{g2}/N) is small and $\bar{Y}_{g2}$ does not differ much from $\bar{Y}_{g1}$, as expected, this bias will be small.

For a given n_{g1}, the variance of $\hat{\bar{Y}}_W$ is

$$V(\hat{\bar{Y}}_W) = \sum_g W_g^2 \left(\frac{1}{n_{g1}} - \frac{1}{N_{g1}}\right) S_{g1}^2 . \tag{11.22}$$

The estimator of this variance is obtained by replacing S_{g1}^2 by the sample variance s_{g1}^2 of the n_{g1} respondents and N_{g1} by its estimate $N_g(n_{g1}/n_g)$.

If it is assumed that the n_{g1} respondents are a random sample of the n_g units, N_{g1} and S_{g1}^2 in the above expression should be replaced by N_g and S_g^2.

The estimator for the total is

$$\hat{Y}_W = \sum_1^G N_g \bar{y}_{g1} = \sum_1^G \frac{N_g}{n_{g1}} \sum_1^{n_{g1}} y_{gi} , \tag{11.23}$$

and its variance is obtained by multiplying (11.22) by N^2.

For the estimator in (11.20), the means of the respondents are inflated with the strata weights W_g. If they are not known, one can replace them by $w_g = n_g/n$. The variance in (11.22) can now be estimated by replacing N_{g1} and S^2_{g1} by $N(n_{g1}/n)$ and s^2_{g1} respectively. For this case of replacing W_g by w_g, Oh and Scheuren (1983) derive an approximation to the variance of the estimator in (11.20) with the assumption that the n_{g1} units respond with probabilities Q_g.

If the strata or *adjustment cells* are formed through a row × column classification, an estimator for the mean is $\Sigma_i \Sigma_j \; N_{ij}\bar{y}_{ij1}/N$, where $i = (1,\ldots,r)$ and $j = (1,\ldots,c)$ represent the rows and columns, respectively. In this expression N_{ij} is the number of units in the ijth cell and $\bar{y}_{ij1}$ is the mean of the n_{ij1} respondents in that cell in the sample of n units.

When a sample is classified as above, the proportions (N_{ij}/N) can be estimated from the sample proportion (n_{ij}/n). If the totals of the numbers of observations in the rows and in the columns, the marginal totals, are known, improvements on these estimates can be made through the **raking** method described, for example, by Brackstone and J.N.K. Rao (1976). For the case of nonresponse, Oh and Scheuren (1987) consider a modification of this procedure. Binder and Theberge (1988) derive the variance of an estimator obtained through the raking method.

For the estimation of the population proportion P of an attribute, the means in (11.20) and (11.21) are replaced by the corresponding proportions. The following example illustrates the effects of poststratification.

Example 11.5. Bias reduction through poststratification: Consider three strata for classifying the responses of a sample selected from a population of $N = 2000$ units. The sizes (N_{g1}, N_{g2}) of the respondents and nonrespondents and the numbers $(C_{g1},\ C_{g2})$ of the units with an attribute of interest are presented in Table 11.2 for two compositions of the strata.

For the first type, the sizes of the three strata N_g are 900, 700, and 400. The numbers of units with the attribute of interest C_g are 510, 280, and

Table 11.2. Two types of stratification.

Sizes and Numbers		Type 1			Type 2	
N_{g1}	600	400	200	800	200	200
C_{g1}	360	160	20	200	180	160
N_{g2}	300	300	200	500	200	100
C_{g2}	150	120	10	120	80	80

30, respectively. Thus, the population proportion having the attribute is $P = (510 + 280 + 30)/2000 = 0.41$. For the second type, $N_g = (1300, 400, 300)$ and $C_g = (320, 260, 240)$.

The proportion for the 1200 respondents with the attribute is $P_1 = 540/1200 = 0.45$. Thus, if one does not poststratify the responses, the sample proportion p_1 of the n_1 responses has a bias of $0.45 - 0.41 = 0.04$.

For the respondents of the first type of stratification, the proportions with the attribute are $360/600 = 0.6$, $160/400 = 0.4$, and $20/200 = 0.10$. Thus, $\Sigma_1^G W_g P_{g1} = [9(0.6) + 7(0.4) + 4(0.1)]/20 = 0.43$. The bias of $\Sigma_1^G W_g p_{g1}$ for estimating the population proportion is $0.43 - 0.41 = 0.02$, which is only half the bias for p_1.

For the second type of stratification, the proportions with the attribute are $200/800 = 0.25$, $180/200 = 0.9$, and $160/200 = 0.8$, and hence $\Sigma_1^G W_g P_{g1} = [13(0.25) + 4(0.9) + 3(0.8)]/20 = 0.4625$. The bias of $\Sigma_1^G W_g p_{g1}$ now is $0.4625 - 0.41 = 0.0525$, which is larger than the bias of both p_1 and the estimator with the first type of stratification.

As seen in the above example, for a particular characteristic, one type of stratification can be more beneficial than the other. Proper poststratification followed by ratio or regression methods of estimation can be helpful in reducing the bias and MSE of the estimators based on the respondents.

Bailar et al. (1978) describe the procedures used for adjusting for the noninterviews in the Current Population Survey (CPS). Six cells based on three regions and two color categories were used for stratification. For estimating income-related variables in the CPS, Ernst (1978) examined the effects of weighting classes based on characteristics such as age, race, sex, educational level, occupational characteristics, and marital status of the head of the family.

The effects of several weighting procedures on the observations of a national telephone survey on smoking and health-related characteristics were evaluated by Boteman et al. (1982). The weights were based on the probability of selection of the units, nonresponse, telephone coverage, and poststratification. Jagers (1986) describes the procedures for reducing the nonresponse bias through poststratification.

Subpopulations

As in Section 6.7, let N_{gj}, Y_{gj}, $\overline{Y}_{gj}$, and S_{gj}^2, $g = (1,\ldots,G)$ and $j = (1,\ldots,k)$, denote the size, total, mean, and variance of the jth subpopulation in the gth stratum. As before, let the additional subscripts 1 and 2 denote the respondents and nonrespondents.

If N_{gj} is known, an estimator for the population total $Y_j = \Sigma_g Y_{gj}$ of the $N_j = \Sigma_g N_{gj}$ units of the subpopulation is

$$\hat{Y}_j = \sum_1^G \hat{Y}_{gj} = \sum_1^G N_{gj}\bar{y}_{gj1}, \tag{11.24}$$

where $\bar{y}_{gj1}$ is the mean of the n_{gj1} respondents among the n_{gj} units in the gth stratum.

Since $\bar{y}_{gj1}$ is unbiased for $\bar{Y}_{gj1}$, the bias of this estimator is

$$B(\hat{Y}_j) = \sum_1^G N_{gj}\bar{Y}_{gj1} - \sum_1^G N_{gj}\bar{Y}_{gj} = \sum_1^G N_{gj2}(\bar{Y}_{gj1} - \bar{Y}_{gj2}). \tag{11.25}$$

Its variance is given by

$$V(\hat{Y}_j) = \sum_1^G N_{gj}^2\left(\frac{1}{n_{gj1}} - \frac{1}{N_{gj1}}\right)S_{gj1}^2. \tag{11.26}$$

For estimating this variance, replace S_{gj1}^2 by the variance s_{gj1}^2 of the n_{gj1} respondents and N_{gj1} by its estimate $N_{gj}(n_{gj1}/n_{gj})$.

The estimator for the mean $\bar{Y}_j$ of the jth domain is given by $\hat{Y}_j/N_j$. If N_{gj} is not known, but N_g is known, for estimating the total and mean, replace it by its estimate $N_g(n_{gj}/n_g)$. If N_g is not known, it is estimated from $N(n_g/n)$ and hence N_{gj} is estimated from $N(n_{gj}/n)$. The following example will illustrate this procedure.

Example 11.6. Weighting for subpopulations: In a random sample of 400 from the 3600 undergraduate students of a university, 280 responded to a question on the number of hours of part-time work during a week. The results were poststratified into male and female groups. Figures for the freshman–sophomore (FS) and junior–senior (JS) classes are presented in Table 11.3.

With the assumption that all the four groups are of the same size, $\hat{N}_{gj}$ = 900. Now, an estimate of the average for the FS class is $\hat{\bar{Y}}_j$ = 0.5(8.5) +

Table 11.3. Part-time employment of students.

	Male		Female	
	FS	JS	FS	JS
Sample size	55	165	40	140
No. of responses	45	120	30	85
Mean	8.5	12.5	10.3	14.4
Variance	24	22	26	32

0.5(10.3) = 9.4. An estimate of the variance for this average becomes

$$v(\hat{\bar{Y}}_j) = 0.25\,\frac{(900-45)}{900(45)}\,(24) + 0.25\,\frac{(900-30)}{900(30)}\,(26) = 0.336$$

and hence it has a S.E. of 0.58.

If N_{gj} are estimated from $N(n_{gj}/n)$, the sizes for the male and female FS class are 3600(55/400) = 495 and 3600(40/400) = 360. Thus, they are in the proportion 495/855 = 0.58 and 0.42. The estimate for the average for the FS class now is given by $\hat{\bar{Y}}_j = 0.58(8.5) + 0.42(10.3) = 9.256$. The estimate of variance in this case is

$$v(\hat{\bar{Y}}_j) = (0.58)^2\,\frac{(495-45)}{495(45)}(24) + (0.42)^2\frac{(360-30)}{360(30)}(26) = 0.3032$$

and hence it has a S.E. of 0.55.

For the case of unknown domain sizes, Little (1986) compared the mean of the respondents $\Sigma_g n_{gj1}\bar{y}_{gj1}/\Sigma_g n_{gj1}$ with the *weighted in cell* mean $\Sigma_g w_{gj1}\bar{y}_{gj1}/\Sigma_g w_{gj1}$, where $w_{gj1} = (n_g/n_{g1})n_{gj1}$.

11.10 Response probabilities and weighting

As noted in Section 2.10, for the case of simple random sampling and complete response, the estimator $N\bar{y}$ for the total can be expressed as $\Sigma^n(y_i/\phi_i)$ where $\phi_i = (n/N)$ is the probability of selecting a unit into the sample. With stratification, the estimator for the population total $\hat{Y}_{\text{st}} = N\hat{\bar{Y}}_{\text{st}}$ studied in Chapter 5 can be expressed as $\Sigma_g(t_g/\phi_g)$, where $t_g = \Sigma^{n_g} y_{gi}$ is the sample total and $\phi_g = n_g/N_g$.

In the case of incomplete response, the sampled units can be considered to respond with certain probabilities, which can be estimated from the sample. Estimators for the population mean with weights based on the probabilities of responses are described below. The biases and MSEs of these estimators can be evaluated for suitable response probabilities.

For the case of a simple random sample of n units selected without replacement and n_1 responses, an estimate for $\bar{Y}$ can be expressed as

$$\hat{\bar{Y}}_Q = \frac{\sum^{n_1}(y_i/Q_i)}{\sum^{n_1}(1/Q_i)}, \tag{11.27}$$

where Q_i is the conditional probability of response of the ith unit which was selected into the sample. If $Q_i = n_1/n$, this estimator is the same as $\bar{y}_1$.

If the n sample units are selected with unequal probabilities π_i, the Horvitz–Thompson type estimator with the n_1 responses can be expressed as

$$\hat{\bar{Y}}_{\mathrm{HQ}} = \frac{\sum^{n_1}(y_i/\pi_i Q_i)}{\sum^{n_1}(1/\pi_i Q_i)}. \tag{11.28}$$

If $\pi_i = n/N$ and Q_i are all equal, this estimator again is the same as $\bar{y}_1$.

With nonresponse and poststratification, the estimator in (11.27) can be expressed as

$$\hat{\bar{Y}}_{\mathrm{SQ}} = \frac{\sum_g \sum_i (y_{gi}/\pi_{gi} Q_{gi})}{\sum_g \sum_i (1/\pi_{gi} Q_{gi})}, \tag{11.29}$$

where π_{gi} is the probability of selecting a unit into the gth stratum and Q_{gi} is its probability of response. In this expression, the second summation is carried over the number of responses n_{g1} in the gth stratum. Special cases of this estimator are as follows.

If $\pi_{gi} = n_g/N_g$ and $Q_{gi} = n_{g1}/n_g$, the above estimator becomes the same as the poststratified mean in (11.20). In this case, the response probabilities for the units of the gth stratum are estimated from n_{g1}/n_g.

If $\pi_{gi} = n/N$ and $Q_{gi} = Q_g$, (11.29) becomes

$$\hat{\bar{Y}}_{\mathrm{SQ}} = \frac{\sum n_{g1}\bar{y}_{g1}/Q_g}{\sum n_{g1}/Q_g}. \tag{11.29a}$$

The estimator $\bar{y}_H$ in (11.2) for subsampling the nonrespondents can be obtained from this expression by replacing (n_{11}, n_{21}) with (n_1, m), $(\bar{y}_{g1}, \bar{y}_{g2})$ with $(\bar{y}_1, \bar{y}_{2(m)})$, and noting that $Q_1 = 1$ and $Q_2 = m/n_2$.

A procedure for obtaining an estimator of the type in (11.29a) was considered by Politz and Simmons (1949) and Simmons (1954). In this method, interviews were assumed to be conducted during the six evenings from Monday through Saturday. The number of evenings the respondent was home during the previous five evenings was recorded at the time of the interview. If the respondent was home on g evenings, $g = (0,...,5)$, an estimate of the probability of his or her response is $(g + 1)/6$. With this estimate, the population mean is obtained from (11.29a).

If the nonresponse of a unit on a characteristic y_i depends on an auxiliary or supplementary variable x_i, Little and Rubin (1987) characterize it as the *ignorable nonresponse*. In this case, the probability of nonresponse can be estimated from the available information, and the above type of weighting procedures can be employed. If the nonresponse cannot be completely explained by x_i, it is recognized as *nonignorable nonresponse*. Suitable procedures are required in this case to adjust the estimators for the nonresponse. If neither of these two cases can explain the nonresponse, in some situations the responses and nonresponses may be considered to be obtained from random samples of the sampled units.

For the case of Q_i depending on an auxiliary characteristic, the merits of the resulting estimator relative to $\bar{y}_1$ were examined by Oh and Scheuren (1983) and Little (1986). Cassell et al. (1983) suggest the estimation of Q_i from the available auxiliary information.

11.11 Imputation

For some large-scale surveys, the effects of imputing suitable values for the missing units have been examined. If the imputation is preceded by grouping the responding units into homogeneous strata or cells, it can help reduce the nonresponse bias in some cases. Effects of some of the imputation schemes can be similar to the regression and weighting adjustments described in the previous sections. Another reason given for imputing the missing values is that the users of surveys may find it convenient to analyze the *clean* set of data obtained after imputation than the original data with observations missing on different characteristics for different units.

Mean imputation

In this procedure, the mean $\bar{y}_1$ of the n_1 respondents are duplicated for the $n_2 = n - n_1$ nonrespondents. The resulting estimator $\bar{y}_d$, with the subscript d denoting duplication, remains the same as $\bar{y}_1$, and its variance is the same as $(N_1 - n_1)S_1^2/n_1$. As noted earlier, the sample variance s_1^2 of the n_1 units is unbiased for S_1^2. With the completed observations, the sample variance becomes

$$s_d^2 = \frac{\sum_1^n (y_i - \bar{y}_d)^2}{n-1} = \frac{n_1 - 1}{n-1} s_1^2. \tag{11.30}$$

Hence, s_d^2 underestimates s_1^2. Further, the distribution of the n completed observations is concentrated around $\bar{y}_1$ and does not properly represent the frequency distribution of the n_1 responses.

Random duplication

For this method, n_2 values selected randomly without replacement from the n_1 respondents are imputed for the nonrespondents. In the completed sample, n_2 observations appear twice and the remaining $n - 2n_2 = n_1 - n_2$ observations appear once. Denote the duplicated observations by y_{i0}, $i = (1,...,n_2)$, and let $\bar{y}_0$ and s_0^2 denote their mean and variance.

With the completed observations, an estimator for the population mean is

$$\bar{y}_c = \frac{\sum^{n_1} y_i + \sum^{n_2} y_{i0}}{n} = w_1 \bar{y}_1 + w_2 \bar{y}_0, \tag{11.31}$$

where $w_1 = n_1/n$ and $w_2 = n_2/n$ as before. Let I denote the initial sample and the responses. Since $E(\bar{y}_0 | I) = \bar{y}_1$, $E(\bar{y}_c) = \bar{y}_1$. Hence, the bias of $\bar{y}_c$ remains the same as that of $\bar{y}_1$. Following Appendix A3,

$$\begin{aligned} V(\bar{y}_c) &= V(\bar{y}_1) + w_2^2 \left(\frac{n_1 - n_2}{n_1 n_2} \right) S_1^2 \\ &= \left[\frac{1}{n}(1 + 2w_2) - \frac{1}{N_1} \right] S_1^2. \end{aligned} \tag{11.32}$$

For large N_1, from (11.32), the proportional increase in the variance is

$$\frac{V(\bar{y}_c) - V(\bar{y}_1)}{V(\bar{y}_1)} = w_2^2 \frac{n_1 - n_2}{n_2} = w_2(1 - 2w_2). \tag{11.33}$$

This expression reaches its maximum at $w_2 = 1/4$. In this case, the relative increase in the variance due to duplication is 12.5%.

Variance estimation

An unbiased estimator of (11.32) is obtained by replacing S_1^2 with s_1^2. The variance s_c^2 of the completed observations can be expressed as

$$(n-1)s_c^2 = (n_1-1)s_1^2 + (n_2-1)s_0^2 + nw_1w_2(\bar{y}_0 - \bar{y}_1)^2. \quad (11.34)$$

Since the conditional expectations of s_0^2 and $(\bar{y}_0 - \bar{y}_1)^2$ are s_1^2 and $(n_1 - n_2)s_1^2/n_1n_2$, from the above expression,

$$E(s_c^2|I) = \frac{n(n-2)+(n_1-n_2)}{n(n-1)}s_1^2 = \left[1 - 2\frac{w_2}{n-1}\right]s_1^2 \quad (11.35)$$

Thus, s_c^2 underestimates s_1^2 and hence S_1^2. This underestimation, as expected, increases with the number of nonresponses but becomes negligible for large n. An unbiased estimator for S_1^2 is given by $(n-1)s_c^2/(n-1-2w_2)$.

Confidence limits

If the responses are considered to be a random sample from the n selected units, both $\bar{y}_1$ and $\bar{y}_c$ become unbiased for $\bar{Y}$. In this case, the variance of $\bar{y}_c$ is obtained by replacing N_1 and S_1^2 in (11.32) with N and S^2, respectively.

If the n_1 responses are considered to be a random sample from the n sampled units, for large N, the variance in (11.32) can be estimated from $(1 + 2w_2)s_1^2/n$. Now, $(1 - \alpha)\%$ confidence intervals for $\bar{Y}$ are obtained from

$$C_1 : \bar{y}_c \pm Z\left(\frac{1+2w_2}{n}\right)^{1/2} s_1 \quad (11.36)$$

Consider the alternative limits,

$$C_2 : \bar{y}_c \pm Zs_c/\sqrt{n}. \quad (11.37)$$

Note from (11.35) that the expected width of C_2 is smaller than the width of C_1.

Multiple imputation

Rubin (1978; 1979; 1986) suggests and illustrates this procedure for the estimation of means, totals, proportions, and other population quantities. One purpose of this procedure is to avoid underestimation of the variances and standard errors, noted in the above sections. Rubin (1987) presents the details of this approach, and its implementation is reviewed in Rubin (1996).

In this procedure, m sets of all the missing observations are randomly selected from a suitable *posterior distribution*. The derivation of a posterior distribution is briefly outlined in Appendix A12. With the selected observations of each set and the corresponding responses, estimates $\hat{T}_i$ for the population quantity and the variance v_i, $i = 1, 2, \ldots, m$ of the completed observations are obtained. The final estimate $\hat{T}$ is obtained from the average of the $\hat{T}_i$. The variance and the standard error of $\hat{T}$ are obtained by combining the variance among the $\hat{T}_i$ and all the within variances v_i.

This method is also illustrated in Herzog and Rubin (1983), Rubin and Schenker (1986), and Heitjan and Little (1988). Rubin et al. (1988) describe this approach for postenumeration surveys, and Glynn et al. (1993) for surveys with follow-ups. Gelman et al. (1995; 1998) describe multiple imputation and the Bayesian approach. Schafer and Schenker (2000) present a procedure for replacing the missing observations with predictions derived from a suitable model for imputation.

Hot- and cold-deck imputation

For large-scale surveys, these procedures were implemented by the U.S. Bureau of the Census and other organizations. For the *hot deck* method, observations are imputed from the current (*hot*) sample, with its units arranged in some order, for example, as a *deck* of computer cards. In contrast, for the *cold-deck* method, observations for imputation are selected from previous surveys, censuses, and similar sources.

Bailar et al. (1978) describe the hot-deck procedure used in the CPS to impute for nonresponse on the different variables; for complete nonresponse on the units, weighting procedures are used as described in Section 11.9. For imputation on the employment-related items, the sample units—about 50,000 households—were arranged in 20 cells. Five age groups, white and black or Hispanic classification, and male–female categories were used to form these cells. A sequential procedure of imputing observations from the most recently updated records was implemented. With this method, the same observation can be duplicated many times. For large sizes of the population and the sample, the variance of the imputed mean derived by the above authors, with the assumption that the responses constitute a random sample from the initial sample, is given by $(1 + 2n_2/n_1)S^2/n$.

11.12 Related topics

Efforts to increase response rates

For several types of surveys and polls, short and unambiguous questions were found to yield high response rates. They are usually low on personal questions, for example, on income and family finances. Response rates cannot be expected to be high on sensitive questions such as smoking habits or alcohol and drug addiction. To elicit responses on such items by assuring confidentiality, Warner (1965) and others suggest the **randomized response** method. Sirken (1983) describes **network sampling** to contact and obtain responses from the sample units. To enumerate the homeless and transient, Iachan and Dennis (1993) present the multiple frame methodology.

Imputation, estimation, and analysis

Bailar and Bailar (1978; 1983) compare the biases arising from the hot-deck and other types of imputation. Ernst (1980) derives the variances for the estimators of the mean obtained through different types of imputation. Chapman et al. (1986) summarize the imputation methods implemented by the U.S. Bureau of the Census. Kalton and Kish (1984) and Kalton and Kasprzyk (1986) describe the different procedures of imputation used in practice. Fellegi and Holt (1976), Platek and Gray (1983), Giles (1988), for example, describe the different types of imputation for the surveys conducted by Statistics Canada. Fay (1996) describes some of the procedures for imputing for the nonrespondents.

Analysis and estimation from imputed data are examined, for example, by Titterington and Sedransk (1986), Wang et al. (1992), and Kott (1994a). J.N.K. Rao and Shao (1992) and J.N.K. Rao (1996) present procedures for finding the S.E. of the estimates obtained from imputed data.

Models for prediction, poststratification, and weighting

Regression type of models for the prediction of observations for the nonrespondents or poststratification followed by weighting were suggested by Little (1982; 1995), Särndal (1986), Rancourt et al. (1994), Kott (1994b), and others. For longitudinal data, Stasny (1986; 1987), Duncan and Kalton (1987), Nordberg (1989), and Little (1995), for example, describe estimation for the case of nonresponse. Potthoff et al. (1993) describe weighting of the responses obtained from the callbacks.

Qualitative and categorical observations

When there is complete response, as in Chapter 4, denote by p the sample proportion having an attribute of interest, and $q = 1 - p$ its complement. The transformed variable $y = \log_e(p/q)$, where $\log_e$ stands for the logarithm with base e, is the *logit* of p. In some situations, this transformed variable is assumed to follow the regression model of the type in (10.5) and its extensions. To adjust for the nonresponse, this type of *logistic regression* is considered, for example, by Alho (1990).

If the characteristic is of the low–medium–high or similar categorical type, the proportions for the different classes can be transformed as above, and adjustments for the nonresponses can be considered. Fay (1986), Baker and Laird (1988), Binder (1991), Conaway (1992), Lipsitz et al. (1994), and others examine such procedures.

11.13 Bayesian procedures

As outlined briefly in Appendix A12, in these procedures, formalized prior beliefs, and information are combined with the sample information. Ericson (1969), Malec and Sedransk (1985), Calvin and Sedransk (1991), for example, describe this approach for finite population sampling. For the case of nonresponse, the Bayesian method is presented by Ericson (1967), J.N.K. Rao and Ganghurde (1972), Smouse (1982), Kadane (1993), and others. To adjust the nonresponse of categorical data, Bayesian procedures were described, for example, by Kaufmann and King (1973), Chiu and Sedransk (1986), and Raghunathan and Grizzle (1995).

11.14 Cesnsus undercount

Following the 1980 Decennial Census of the U.S., adjustments for the undercount of the population in several areas were examined through imputation, poststratification, weighting, and other procedures described in the above sections. Ericsen and Kadane (1985) and Ericsen et al. (1992) considered suitable models for this purpose. Cressie (1989; 1992) and others examine the Bayesian approach.

Fienberg (1992) provides bibliography for the capture–recapture method described in Section 4.12 and its application for the census undercount. Hogan (1993) and Ding and Fienberg (1994) describe the related *dual system* estimation for combining information from two or more sources, such as the census and the post-censal surveys.

Exercises

11.1. For the survey described in Example 11.1, consider $n_1 =$ 250 responses for another characteristic. If 150 of the respondents are observed to have the attribute of interest, find the 95% confidence limits for the percentage of the population having the attribute.

11.2. If $e_0 = 0.5$, $e_1 = 1$, $e_2 = 6$, and \$6000 are available for the survey in Example 11.2, find the optimum sizes for the samples at the two stages. If $S^2 = 2500$, find the S.E. of $\bar{y}$ that is obtained with these optimum sizes.

11.3. Consider the eight families in Table 11.1 with the completed observations on income, husband's age, and wife's age. Fit the multiple regressions of income on the ages of the couples through the origin and with the intercept, and test for the significance of the regressions. To predict the missing observations on income, examine whether these regressions can be preferred to the linear regressions on husband's and wife's age analyzed in Example 11.3.

11.4. From the responses of the families in Table 11.1, fit the linear regressions of television time on each of the remaining four variables, and examine which of the four regressions can be recommended for predicting the missing observations on television time.

11.5. In a sample of 150 from the 2000 population units of Example 11.5, only 95 responded. The observed sample sizes for the two types of stratification are (42, 35, 18) and (55, 26, 14). To estimate the population proportion, for both these cases, compare the variances and MSEs of the sample proportion and the poststratified estimator of the proportion (a) assuming that the responses are a random sample and (b) without this assumption.

11.6. Among the 95 respondents of the above example, 46 were observed to have the characteristic of interest. The numbers observed in the three strata of the first type were (22, 15, 9). (a) Compare the sample estimate of the proportion for the characteristic of interest with the poststratified estimate. (b) With the assumption of random response, find the sample standard errors of these estimates.

11.7. (a) Using the data in Table 11.3, find estimates for the average number of hours of part-time employment for the junior–senior class through the two procedures described in Example 11.6 and find their S.E. values.

11.8. If $n_1 < n_2$, the n_2 observations can be duplicated with replacement from the n_1 responses. (a) Show that the expectation of $\bar{y}_c$ for this case is also $\bar{y}_1$, but its variance is obtained by replacing $(n_1 - n_2)$ in the second term of (11.32) by $(n_1 - 1)$. (b) Show that the increase in the variance of $\bar{y}_c$ relative to $\bar{y}_1$ for this type of duplication approximately becomes $w_1 w_2$, and it reaches its maximum of 25% when the response rate is 50%.

Appendix A11

Variance of the Hansen–Hurwitz estimator

First note that $V[E(\bar{y}_H \mid s)] = V(\bar{y}) = (1 - f)S^2/n$, where $f = n/N$ and S^2 is the variance of the N population units. Next, from (11.2),

$$V(\bar{y}_H \mid s; n_2) = w_2^2 \frac{k-1}{n_2} s_2^2 = w_2 \frac{k-1}{n} s_2^2,$$

where s_2^2 is the variance of the n_2 units of the second stratum. Since $E(s_2^2) = S_2^2$, which is the variance of the N_2 nonresponding units, and $E(w_2) = W_2 = N_2/N$,

$$E[V(\bar{y}_H \mid s; n_2)] = W_2 \frac{k-1}{n} S_2^2.$$

The variance in (11.3) is obtained by combining these two expressions.

Variance estimator for the Hansen–Hurwitz estimator

The sample variance s^2 of the n units is unbiased for S^2, but it is not available due to the nonresponse. From the results of the samples at the two stages, an unbiased estimator of S^2 is given by

$$\begin{aligned}\hat{S}^2 &= \frac{n_1 - 1}{n - 1} s_1^2 + \frac{n_2 - k + w_2(k - 1)}{n - 1} s_m^2 \\ &\quad + \frac{n}{n - 1}[w_1(\bar{y}_1 - \bar{y}_H)^2 + w_2(\bar{y}_m - \bar{y}_H)^2].\end{aligned}$$

The term in the square brackets is the same as $w_1 w_2 (\bar{y}_1 - \bar{y}_m)^2$.

Replacing S^2 by the above estimator and S_2^2 by s_m^2, an unbiased estimator for the variance in (11.3) is obtained from

$$\begin{aligned}v(\bar{y}_H) &= \frac{(1-f)}{n(n-1)}[(n_1 - 1)s_1^2 + (n_2 - k)s_m^2 + n_1(\bar{y}_1 - \bar{y}_H)^2 \\ &\quad + n_2(\bar{y}_m - \bar{y}_H)^2] + \frac{(N-1)(k-1)w_2}{N(n-1)} s_m^2.\end{aligned}$$