# CHAPTER 9

# Regression, Product, and Calibrated Methods of Estimation

## 9.1 INTRODUCTION

We have seen in Chapter 8 that the ratio estimator is optimal if the regression of $y$ on $x$ is linear and passing through the origin. The ratio estimator may not be efficient if the regression line does not pass through the origin. In this situation, one may use auxiliary information effectively through the regression estimator to improve precision of the conventional estimator. Like the ratio estimator, the regression estimator is not unbiased for the population mean or total. In this section we will study the difference, regression, product, calibrated, and related estimators along with their properties in detail.

## 9.2 DIFFERENCE ESTIMATOR

If the population total $X$ of the auxiliary variable $x$ is known, then the estimator

$$\widehat{Y}_\lambda = \widehat{Y} - \lambda\left(\widehat{X} - X\right) \tag{9.2.1}$$

is unbiased for the population total $Y$ for any known value of $\lambda$. The estimator (Eq. 9.2.1) is known as a difference estimator. To use the difference estimator one needs to find the optimum value of $\lambda$, which minimizes the variance of $\widehat{Y}_\lambda$.

Now minimizing the variance of $\widehat{Y}_\lambda$ with respect to $\lambda$, the optimum value of $\lambda$ comes out as

$$\lambda_{opt} = \frac{Cov\left(\widehat{X}, \widehat{Y}\right)}{V\left(\widehat{X}\right)} = \beta \quad \text{(say)} \tag{9.2.2}$$

Putting $\lambda = \beta$ in Eq. (9.2.1), we get the optimum estimator

$$\widehat{Y}_\beta = \widehat{Y} - \beta\left(\widehat{X} - X\right) \tag{9.2.3}$$

*Survey Sampling Theory and Applications*
ISBN 978-0-12-811848-1
http://dx.doi.org/10.1016/B978-0-12-811848-1.00009-1

The variance of the optimum estimator $\widehat{Y}_\beta$ is given by

$$V\left(\widehat{Y}_\beta\right) = \left(1 - \rho^2_{\hat{X},\hat{Y}}\right) V\left(\widehat{Y}\right) \tag{9.2.4}$$

where $\rho_{\hat{X},\hat{Y}}$ is the correlation between $\widehat{X}$ and $\widehat{Y}$.

Because $|\rho_{\hat{X},\hat{Y}}| \leq 1$, $\widehat{Y}_\beta$ possesses smaller variance than the convention estimator $\widehat{Y}$. The estimator $\widehat{Y}_\beta$ is seldom used in practice because the value of $\beta$ is unknown in most practical situations. Let $\beta_0$ be a guess value of $\beta$ obtained from a past experiment. Then substituting $\beta = \beta_0$ in Eq. (9.2.3), one gets the estimator

$$\widehat{Y}_{\beta_0} = \widehat{Y} - \beta_0\left(\widehat{X} - X\right) = \widehat{Y}_\beta + (\beta - \beta_0)\left(\widehat{X} - X\right) \tag{9.2.5}$$

The loss of efficiency of using $\widehat{Y}_{\beta_0}$ over the optimum estimator $\widehat{Y}_\beta$ comes out as

$$\begin{aligned} \frac{V\left(\widehat{Y}_{\beta_0}\right)}{V\left(\widehat{Y}_\beta\right)} - 1 &= \frac{(\beta - \beta_0)^2 V\left(\widehat{X}\right)}{V\left(\widehat{Y}_\beta\right)} \\ &= \frac{(\beta - \beta_0)^2 V\left(\widehat{X}\right)}{\left(1 - \rho^2_{\hat{X},\hat{Y}}\right) V\left(\widehat{Y}\right)} \\ &= \left(1 - \frac{\beta_0}{\beta}\right)^2 \frac{\rho^2_{\hat{X},\hat{Y}}}{1 - \rho^2_{\hat{X},\hat{Y}}} \end{aligned} \tag{9.2.6}$$

To keep the loss of efficiency to a certain level $\gamma$ (say), we must have

$$\left(1 - \frac{\beta_0}{\beta}\right)^2 < \gamma \frac{1 - \rho^2_{\hat{X},\hat{Y}}}{\rho^2_{\hat{X},\hat{Y}}} \tag{9.2.7}$$

For simple random sampling without replacement (SRSWOR) and simple random sampling with replacement (SRSWR), Eq. (9.2.7) reduces to

$$\left(1 - \frac{\beta_0}{\beta}\right)^2 < \gamma \frac{1 - \rho^2}{\rho^2} \tag{9.2.8}$$

where $\rho$ is the correlation coefficient between $x$ and $y$.

In particular if $\gamma = 0.25$, $\rho = 0.5$, Eq. (9.2.8) reduces to $0.134 < \dfrac{\beta_0}{\beta} < 1.866$.

For the value of $\lambda = 1$, the estimator $\widehat{Y}_\lambda$ reduces to a very simple estimator viz.

$$\widehat{Y}_1 = \widehat{Y} - \left(\widehat{X} - X\right) \tag{9.2.9}$$

This estimator $\widehat{Y}_1$ is unbiased for $Y$ and becomes more efficient than the conventional estimator $\widehat{Y}$ if $\rho_{\widehat{X},\widehat{Y}} > \frac{1}{2}\sqrt{\frac{V\left(\widehat{Y}\right)}{V\left(\widehat{X}\right)}}$. In particular if the variances of $\widehat{X}$ and $\widehat{Y}$ are equal, then the difference estimator given in Eq. (9.2.9) performs better than the conventional estimator $\widehat{Y}$ if $\rho_{\widehat{X},\widehat{Y}} > 1/2$. For SRSWOR and SRSWR sampling, $\widehat{X} = N\overline{x}_s$ and $\widehat{Y} = N\overline{y}_s$. Here the condition of superiority of the difference estimator over the conventional estimator reduces to $\rho > 1/2$ and it is realized in most practical situations.

## 9.3 REGRESSION ESTIMATOR

The value of $\beta$ in Eq. (9.2.2) is generally unknown and it is estimated by

$$\widehat{\beta} = \frac{C\widehat{o}v\left(\widehat{X}, \widehat{Y}\right)}{\widehat{V}\left(\widehat{X}\right)} \tag{9.3.1}$$

where $C\widehat{o}v\left(\widehat{X}, \widehat{Y}\right)$ and $\widehat{V}\left(\widehat{X}\right)$ are unbiased for $Cov\left(\widehat{X}, \widehat{Y}\right)$ and $V\left(\widehat{X}\right)$, respectively.

Hence, replacing $\beta$ by its estimator $\widehat{\beta}$ given in Eq. (9.3.1), we get the well known linear regression estimator (or simply regression estimator) for the total $Y$ as

$$\widehat{Y}_{reg} = \widehat{Y} - \widehat{\beta}\left(\widehat{X} - X\right) \tag{9.3.2}$$

Like the ratio estimator, the regression estimator is not unbiased for $Y$, but it is more efficient than the conventional estimator $\widehat{Y}$.

### 9.3.1 Exact Expression of Bias

The bias of the regression estimator $\widehat{Y}_{reg}$ is given by

$$\begin{aligned} B\left(\widehat{Y}_{reg}\right) &= E\left(\widehat{Y}_{reg} - Y\right) \\ &= -E\left\{\widehat{\beta}\left(\widehat{X} - X\right)\right\} \\ &= -Cov\left(\widehat{\beta}, \widehat{X}\right) \end{aligned} \tag{9.3.3}$$

Because in most situations, $Cov\left(\widehat{\beta}, \widehat{X}\right)$ decreases as the sample size increases, the regression estimator is approximately unbiased when the sample size is large.

### 9.3.2 Approximate Expression of Bias

Now writing $\varepsilon_{xy} = \dfrac{\widehat{Cov}(\widehat{X}, \widehat{Y})}{Cov(\widehat{X}, \widehat{Y})} - 1$ and $\varepsilon_{xx} = \dfrac{\widehat{V}(\widehat{X})}{V(\widehat{X})} - 1$, we get

$$\begin{aligned}\widehat{\beta} &= \frac{(1+\varepsilon_{xy})Cov(\widehat{X}, \widehat{Y})}{(1+\varepsilon_{xx})V(\widehat{X})} \\ &= \beta(1+\varepsilon_{xy})(1+\varepsilon_{xx})^{-1} \\ &= \beta(1+\varepsilon_{xy})(1-\varepsilon_{xx}+\varepsilon_{xx}^2-\cdots)\end{aligned} \tag{9.3.4}$$

The expression of bias of the regression estimator comes out as

$$\begin{aligned}B(\widehat{Y}_{reg}) &= E(\widehat{Y}_{reg} - Y) \\ &= -\beta E\left[(1+\varepsilon_{xy})(1-\varepsilon_{xx}+\varepsilon_{xx}^2-\cdots)(\widehat{X}-X)\right] \\ &= -\beta E\left[(\varepsilon_{xy}-\varepsilon_{xx}+\varepsilon_{xx}^2-\varepsilon_{xy}\varepsilon_{xx}+\cdots)(\widehat{X}-X)\right]\end{aligned}$$

Now neglecting the terms containing $E\left[\varepsilon_{xx}^i \varepsilon_{xy}^j (\widehat{X}-X)\right]$ for $i+j>1$ (which is expected to be small), we get an expression of bias up to the first order of approximation as

$$\begin{aligned}B(\widehat{Y}_{reg}) &\cong \beta E\left[(\varepsilon_{xx}-\varepsilon_{xy})(\widehat{X}-X)\right] \\ &= \beta\left[\frac{Cov\{\widehat{V}(\widehat{X}), \widehat{X}\}}{V(\widehat{X})} - \frac{Cov\{\widehat{Cov}(\widehat{X}, \widehat{Y}), \widehat{X}\}}{Cov(\widehat{X}, \widehat{Y})}\right]\end{aligned} \tag{9.3.5}$$

#### 9.3.2.1 Bias Under Simple Random Sampling Without Replacement

For an SRSWOR, $\widehat{X} = N\overline{x}_s$, $V(\widehat{X}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_x^2$, $\widehat{V}(\widehat{X}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_x^2$, $Cov(\widehat{X}, \widehat{Y}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_{xy}$, and $\widehat{Cov}(\widehat{X}, \widehat{Y}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_{xy}$. Hence the approximate bias of the regression estimator is obtained using Eq. (9.3.5) as follows:

$$\begin{aligned}B(\widehat{Y}_{reg}) &\cong \beta\left[\frac{Cov\left\{N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_x^2, N\overline{x}_s\right\}}{\left\{N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_x^2\right\}} - \frac{Cov\left\{N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_{xy}, N\overline{x}_s\right\}}{\left\{N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_{xy}\right\}}\right] \\ &= N\beta\left[\frac{Cov(s_x^2, \overline{x}_s)}{S_x^2} - \frac{Cov(s_{xy}, \overline{x}_s)}{S_{xy}}\right]\end{aligned}$$

Now using the following results of Sukhatme (1944) (detailed derivation is given in Appendix 9A), we find

$$Cov\left(s_x^2, \overline{x}_s\right) = \frac{N(N-n)}{(N-1)(N-2)} \frac{\mu_{30}}{n} \text{ and}$$

$$Cov\left(s_{xy}, \overline{x}_s\right) = \frac{N(N-n)}{(N-1)(N-2)} \frac{\mu_{21}}{n}$$

where $\mu_{ij} = \frac{1}{N} \sum_{i \in U} \left(x_i - \overline{X}\right)^i \left(\gamma_i - \overline{Y}\right)^j$.

The expression for the bias (Eq. 9.3.5) comes out as

$$B\left(\widehat{Y}_{reg}\right) = \frac{N(N-n)}{(N-2)} \frac{\beta}{n} \left(\frac{\mu_{30}}{\mu_{20}} - \frac{\mu_{21}}{\mu_{11}}\right)$$

In case $x$ and $\gamma$ follow bivariate normal distribution, we have $\mu_{30} = \mu_{21} = 0$ and the bias of the regression estimator $B\left(\widehat{Y}_{reg}\right)$ reduces to zero.

### 9.3.3 Approximate Expression of the Mean Square Error

The mean square error (MSE) of the regression estimator $\widehat{Y}_{reg}$ is given by

$$\begin{aligned}
M\left(\widehat{Y}_{reg}\right) &= E\left(\widehat{Y}_{reg} - Y\right)^2 \\
&= E\left[\left(\widehat{Y} - Y\right) - \widehat{\beta}\left(\widehat{X} - X\right)\right]^2 \\
&= E\left[\left(\widehat{Y} - Y\right) - \beta(1 + \varepsilon_{xy})\left(1 - \varepsilon_{xx} + \varepsilon_{xx}^2 - \cdots\right)\left(\widehat{X} - X\right)\right]^2 \\
&= E\left[\left\{\left(\widehat{Y} - Y\right) - \beta\left(\widehat{X} - X\right)\right\} - \beta\left(\widehat{X} - X\right)\right. \\
&\qquad \left. \times \left\{\varepsilon_{xy} - \varepsilon_{xx} + \varepsilon_{xx}^2 - \varepsilon_{xy}\varepsilon_{xx} + \cdots\right\}\right]^2
\end{aligned}$$

Neglecting terms $E\left(\widehat{X} - X\right)^2 \varepsilon_{xy}^i \varepsilon_{xx}^j$ for $i + j \geq 2$, the expression of MSE up to the first order of approximation is given by

$$\begin{aligned}
M\left(\widehat{Y}_{reg}\right) &\cong E\left[\left(\widehat{Y} - Y\right) - \beta\left(\widehat{X} - X\right)\right]^2 \\
&= V\left(\widehat{Y}\right) + \beta^2 V\left(\widehat{X}\right) - 2\beta Cov\left(\widehat{X}, \widehat{Y}\right) \qquad (9.3.6) \\
&= \left(1 - \rho_{\widehat{X},\widehat{Y}}^2\right) V\left(\widehat{Y}\right) \qquad (9.3.7)
\end{aligned}$$

Let $\widehat{X} = \sum_{i \in s} b_{si} x_i$ and $\widehat{Y} = \sum_{i \in s} b_{si} \gamma_i$ be linear homogeneous unbiased estimators of the population totals $X$ and $Y$, respectively, based on a sample

$s$ of size $n$ selected with probability $p(s)$. The constants $b_{si}$'s satisfy the unbiasedness condition $\sum_{s\supset i} b_{si}p(s) = 1$. Then we can write Eq. (9.3.6) as

$$\begin{aligned} M\left(\widehat{Y}_{reg}\right) &\cong E\left[\sum_{i\in s} b_{si}(\gamma_i - \beta x_i) - (Y - \beta X)\right]^2 \\ &= V\left[\sum_{i\in s} b_{si}(\gamma_i - \beta x_i)\right] \\ &= V\left(\sum_{i\in s} b_{si}E_i\right) \\ &= \sum_i (\alpha_i - 1)E_i^2 + \sum_{i\neq} \sum_j (\alpha_{ij} - 1)E_iE_j \\ &= V\left(\widehat{E}\right) \end{aligned} \tag{9.3.8}$$

where $E_i = \gamma_i - \beta x_i$, $\widehat{E} = \sum_{i\in s} b_{si}E_i$, $\alpha_i = \sum_{s\supset i} b_{si}^2 p(s)$, and $\alpha_{ij} = \sum_{s\supset i,j} b_{si}b_{sj}p(s)$.

From Eq. (9.3.8), we set an approximate unbiased estimator for $M\left(\widehat{Y}_{reg}\right)$ as

$$\begin{aligned} \widehat{M}\left(\widehat{Y}_{reg}\right) &= \sum_{i\in s} (\alpha_i - 1)\widehat{E}_i^2/\pi_i + \sum_{i\neq} \sum_j (\alpha_{ij} - 1)\widehat{E}_i\widehat{E}_j/\pi_{ij} \\ &= \widehat{V}\left(\widehat{E}\right) \end{aligned} \tag{9.3.9}$$

where $\widehat{E}_i = \gamma_i - \widehat{\beta}x_i$, $\pi_i =$ inclusion probability of $i$th unit, $\pi_{ij}$ $(>0)$ inclusion probability of the $i$th and $j(\neq i)$th unit and $\widehat{\beta}$ as in Eq. (9.3.1).

## 9.3.4 Mean Square Errors for Some Sampling Designs

### *9.3.4.1 Arbitrary Fixed Effective Sample Size Design*

Consider a fixed effective sample size $n$ design with $b_{si} = 1/\pi_i$. In this case

$$\widehat{X} = \sum_{i\in s} x_i/\pi_i,\ \widehat{Y} = \sum_{i\in s} \gamma_i/\pi_i,\ V\left(\widehat{X}\right) = \frac{1}{2}\sum_{i\neq} \sum_{j\in U} \Delta_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_i}{\pi_j}\right)^2,$$

$$Cov\left(\widehat{X}, \widehat{Y}\right) = \frac{1}{2}\sum_{i\neq} \sum_{j\in U} \Delta_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)\left(\frac{\gamma_i}{\pi_i} - \frac{\gamma_j}{\pi_j}\right),$$

$$\beta = \sum_{i\neq} \sum_{j\in U} \Delta_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)\left(\frac{\gamma_i}{\pi_i} - \frac{\gamma_j}{\pi_j}\right) \Bigg/ \sum_{i\neq} \sum_{j\in U} \Delta_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)^2$$

$$\widehat{V}(\widehat{X}) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j}\right)^2,$$

$$C\widehat{o}v(\widehat{X},\widehat{Y}) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_i}\right)\left(\frac{y_i}{\pi_i}-\frac{y_i}{\pi_j}\right)$$

$$\widehat{\beta} = \sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi_j}\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)\Big/\sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{x_i}{\pi_i}-\frac{x_j}{\pi j}\right)^2$$

where $\Delta_{ij} = \pi_i\pi_j - \pi_{ij}$.

The expression for the regression estimator $\widehat{Y}_{reg}$, its approximate MSE, and an approximate unbiased estimator of the MSE, respectively, come out as follows:

$$\widehat{Y}_{reg} = \sum_{i\in s}\frac{y_i}{\pi_i} - \widehat{\beta}\left(\sum_{i\in s}\frac{x_i}{\pi_i} - X\right) \tag{9.3.10}$$

$$M(\widehat{Y}_{reg}) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}\Delta_{ij}\left(\frac{E_i}{\pi_i}-\frac{E_j}{\pi_j}\right)^2 \tag{9.3.11}$$

$$\widehat{M}(\widehat{Y}_{reg}) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{\widehat{E}_i}{\pi_i}-\frac{\widehat{E}_j}{\pi_j}\right)^2 \tag{9.3.12}$$

#### 9.3.4.2 Simple Random Sampling Without Replacement

For an SRSWOR design $\pi_i = \frac{n}{N}$ and $\pi_{ij} = \frac{n(n-1)}{N(N-1)}$. In this case $\widehat{X} = N\overline{x}_s$, $\widehat{Y} = N\overline{y}_s$, $V(\widehat{X}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_x^2$, $\widehat{V}(\widehat{X}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_x^2$, $V(\widehat{Y}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_y^2$, $\widehat{V}(\widehat{Y}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_y^2$, $Cov(\widehat{X},\widehat{Y}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_{xy}$, $C\widehat{o}v(\widehat{X},\widehat{Y}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_{xy}$, $\beta = \frac{S_{xy}}{S_x^2}$, $\widehat{\beta} = \frac{s_{xy}}{s_x^2}$, and $\rho_{\widehat{X},\widehat{Y}} = \rho = \frac{S_{xy}}{S_xS_y}$, where $\overline{x}_s = \sum_{i\in s}x_i/n$, $\overline{y}_s = \sum_{i\in s}y_i/n$, $S_x^2 = \sum_{i\in U}(x_i-\overline{X})^2/(N-1)$, $s_x^2 = \sum_{i\in s}(x_i-\overline{x}_i)^2/(n-1)$, and $S_y^2 = \sum_{i\in U}(y_i-\overline{Y})^2/(N-1)$.

So, under SRSWOR

$$\widehat{Y}_{reg} = N\left[\overline{y}_s - \widehat{\beta}(\overline{x}_s - \overline{X})\right] \tag{9.3.13}$$

$$M(\widehat{Y}_{reg}) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)S_E^2 = N^2\left(\frac{1}{n}-\frac{1}{N}\right)(1-\rho^2)S_y^2 \tag{9.3.14}$$

$$\widehat{M}\left(\widehat{Y}_{reg}\right) = N^2\left(\frac{1}{n}-\frac{1}{N}\right)s_{\hat{E}}^2 = N^2\left(\frac{1}{n}-\frac{1}{N}\right)(1-\widehat{\rho}^2)s_y^2 \qquad (9.3.15)$$

where

$$S_E^2 = \sum_{i\in U}\left(E_i-\overline{E}\right)^2/(N-1),\ \ s_{\hat{E}}^2 = \sum_{i\in U}\left(\widehat{E}_i-\widehat{\overline{E}}_s\right)^2/(n-1),$$
$$\overline{E} = \sum_{i\in U}E_i/N,\ \ \widehat{\overline{E}}_s = \sum_{i\in s}\widehat{E}_i/n \text{ and } \widehat{\rho} = s_{xy}/(s_x s_y).$$

## 9.3.5 Efficiency of Regression Estimator

From Eq. (9.3.7) we note that the regression estimator $\widehat{Y}_{reg}$ is more efficient than the conventional estimator $\widehat{Y}$ because $V\left(\widehat{Y}\right) - M\left(\widehat{Y}_{reg}\right) = \rho_{\hat{X}\hat{Y}}^2 V\left(\widehat{Y}\right) \geq 0$.

### *9.3.5.1 Comparison With the Ratio Estimator*

Theorem 9.3.1

The regression estimator is more efficient than the ratio estimator.

Proof

The difference between the approximate expressions of the MSEs of the ratio and regression estimators obtained from the Theorem 8.3.1 and Eq. (9.3.7) is

$$\begin{aligned} M\left(\widehat{Y}_R\right) - M\left(\widehat{Y}_{reg}\right) &= V\left(\widehat{Y}\right) + R^2 V\left(\widehat{X}\right) - 2\rho_{\hat{X}\hat{Y}} R\sqrt{V\left(\widehat{Y}\right)V\left(\widehat{X}\right)} \\ &\quad - \left(1-\rho_{\hat{X}\hat{Y}}^2\right)V\left(\widehat{Y}\right) \\ &= \left[R\sqrt{V\left(\widehat{X}\right)} - \rho_{\hat{X}\hat{Y}}\sqrt{V\left(\widehat{Y}\right)}\right]^2 \\ &\geq 0 \end{aligned}$$

## 9.3.6 Optimality of the Regression Estimator

It is pointed out in Corollary 6.4.3 that the regression estimator $\widehat{Y}_{reg} = N\{\overline{y}_s - \widehat{\beta}\left(\overline{x}_s - \overline{X}\right)\}$ is the optimal in the class of linear $\xi$-unbiased predictors $C_{l\xi}$ of the population $Y$ under the superpopulation model $E_\xi(y_i) = \beta_0 + \beta x_i$; $V_\xi(y_i) = \sigma^2$ and $C_\xi(y_i, y_j) = 0$ for $i \neq j$.

Example 9.3.1

From the data given in Example 8.5.1, estimate total production of apples ($Y$) by the regression method of estimation. Estimate the standard error of

the regression estimator. Compare the relative efficiencies of the regression estimator with the ratio estimator and the expansion estimator for estimating the total $Y$.

From the given data, we get $N = 125$, $n = 15$, $\overline{X} = 40$, $\overline{x}(s) = 35$, $\overline{y}(s) = 1050$, $s_x^2 = 164.286$, $s_y^2 = 133942.9$, and $s_{xy} = 4617.857$. The sample regression coefficient $\widehat{\beta} = \dfrac{s_{xy}}{s_x^2} = 28.1087$ and $\widehat{\rho} =$ sample correlation coefficient $= r = 0.984$.

The regression estimator for the population total $Y$ is

$$\widehat{Y}_{reg} = N\left[\overline{y}_s - \widehat{\beta}\left(\overline{x}_s - \overline{X}\right)\right] = 125 \times \{1050 - 28.1087 \times (35 - 40)\}$$
$$= 148,817.9$$

An estimate of the MSE of the regression estimator is

$$\widehat{M}\left(\widehat{Y}_{reg}\right) = N^2(1-f)(1-\widehat{\rho}^2)s_y^2/n$$
$$= (125)^2(1 - 15/125)(1 - 0.9844^2) \times 13394.9/15$$
$$= 3975,839.8$$

Estimated standard error of $\widehat{Y}_{reg}$ is $S\widehat{E}\left(\widehat{Y}_{reg}\right) = \sqrt{\widehat{M}\left(\widehat{Y}_{reg}\right)} = 1948.29$.

From the Example 8.5.1, we get $\widehat{Y}_R = N\widehat{R}\overline{X} = 125 \times 30 \times 40 = 150{,}000$ and the estimated MSE is $\widehat{M}\left(\widehat{Y}_R\right) = N^2\widehat{M}\left(\widehat{\overline{Y}}_R\right) = 125^2 \times 277.389 = 4334203.13$.

An estimate of $Y$ based on the expansion estimator is $\widehat{Y} = N\overline{y}_s = 125 \times 1050 = 131{,}250$.

Estimated variance of $\widehat{Y}$ is $\widehat{V}\left(\widehat{Y}\right) = N^2(1-f)s_y^2/n = 125^2(1 - 15/125) \times 133,942.9/15 = 122{,}780{,}952.4$.

Estimated efficiency of the regression estimator compared to the ratio estimator is

$$\left\{\widehat{M}\left(\widehat{Y}_R\right)\Big/\widehat{M}\left(\widehat{Y}_{reg}\right)\right\} \times 100 = 114.18\%$$

Estimated efficiency of the regression estimator compared to the expansion estimator is

$$\left\{V\left(\widehat{Y}\right)\Big/\widehat{M}\left(\widehat{Y}_{reg}\right)\right\} \times 100 = 3234.61\%$$

## 9.3.7 Unbiased Regression Estimator

Singh and Srivastava (1980) proposed a sampling scheme for which the regression estimator (Eq. 9.3.13) is unbiased for the total $Y$. The sampling scheme is described as follows.

### *9.3.7.1 Singh and Srivastava Sampling Scheme*

Under the Singh and Srivastava (SS) sampling scheme, we first select two units $i$ and $j(i \neq j)$ with probability proportional to $(x_i - x_j)^2$, then select $(n-2)$ units from the remaining $(N-2)$ units by SRSWOR method. So, for the SS sampling scheme, probability of selection a sample $s$ of $n$ units is

$$p(s) \propto \sum_{i\neq} \sum_{j\in s} (x_i - x_j)^2 \tag{9.3.16}$$

$$\text{i.e.} \quad p(s) = \sum_{i\neq} \sum_{j\in s} (x_i - x_j)^2 \Big/ \sum_{s} \sum_{i\neq} \sum_{j\in s} (x_i - x_j)^2$$

Now noting $\sum_{i\neq} \sum_{j\in s} (x_i - x_j)^2 = 2n(n-1)s_x^2$ and $\sum_{s} \sum_{i\neq} \sum_{j\in s} (x_i - x_j)^2 = 2N(N-1)S_x^2 \binom{N-2}{n-2}$, we find

$$p(s) = s_x^2 \Big/ \left\{ \binom{N}{n} S_x^2 \right\} \tag{9.3.17}$$

Now noting $\widehat{\beta} = \left(s_{xy}/s_x^2\right)$, we find for the SS sampling scheme

$$E\left(\widehat{Y}_{reg}\right) = N \sum_{s} \left\{ \overline{y}_s - \left(s_{xy}/s_x^2\right)\left(\overline{x}_s - \overline{X}\right) \right\} s_x^2 \Big/ \left\{ s_x^2 \binom{N}{n} \right\}$$

$$= \frac{N}{S_x^2} \sum_{s} \left\{ \overline{Y} s_x^2 + \left(\overline{y}_s - \overline{Y}\right) s_x^2 - \left(s_{xy}\right)\left(\overline{x}_s - \overline{X}\right) \right\} \Big/ \binom{N}{n}$$

$$= N\left[ \overline{Y} E\left(s_x^2 \middle| \text{SRSWOR}\right) + Cov\left(\overline{y}_s, s_x^2 \middle| \text{SRSWOR}\right) - Cov\left(s_{xy}, \overline{x}_s \middle| \text{SRSWOR}\right) \right] / S_x^2$$

where

$$E\left(s_x^2 \middle| \text{SRSWOR}\right) = S_x^2, \tag{9.3.18}$$

$$Cov\left(\bar{y}_s, s_x^2 | \text{SRSWOR}\right) = \frac{1}{n}\frac{N(N-n)}{(N-1)(N-2)}\mu_{21} \tag{9.3.19}$$

$$Cov\left(s_{xy}, \bar{x}_s | \text{SRSWOR}\right) = \frac{1}{n}\frac{N(N-n)}{(N-1)(N-2)}\mu_{21} \tag{9.3.20}$$

and $\mu_{21} = \frac{1}{N}\sum_{i\in U}\left(x_i - \bar{X}\right)^2\left(y_i - \bar{Y}\right)$.

The results (Eqs. 9.3.19 and 9.3.20) were obtained by Sukhatme (1944) (details given in Appendix 9A). From Eqs. (9.3.18–9.3.20), we note that $\hat{Y}_{reg}$ is an unbiased estimator of the total $Y$.

The expression of the exact variance of $\hat{Y}_{reg}$ for this sampling scheme is very complex. However, an approximate expression of variance and its approximate unbiased estimator are stated in the following theorem without derivation.

Theorem 9.3.2

Under the SS sampling scheme

(i) $\hat{Y}_{reg} = N\left\{\bar{y}_s - \left(s_{xy}/s_x^2\right)\left(\bar{x}_s - \bar{X}\right)\right\}$ is unbiased for the $Y$

(ii) $V\left(\hat{Y}_{reg}\right) \cong N^2\left[\left(1-\rho^2\right)\left(\frac{1}{n}+\frac{1}{n^2}\right)\mu_{02} + \frac{1}{n^2\mu_{20}}\left(\frac{2\mu_{11}\mu_{31}}{\mu_{20}} - \frac{\mu_{11}^2\mu_{40}}{\mu_{20}^2} - \mu_{22}\right)\right]$

(iii) $\hat{V}\left(\hat{Y}_{reg}\right) \cong Y_{reg}^2 - \frac{NS_x^2}{ns_x^2}\left(\sum_{i\in s} y_i^2 + \frac{N-1}{n-1}\sum_{i\neq j}\sum_{\in s} y_i y_j\right)$

where $\mu_{jk} = \frac{1}{N}\sum_{i\in U}\left(x_i - \bar{X}\right)^j\left(y_i - \bar{Y}\right)^k$

## 9.3.8 Stratified Regression Estimator

Here we suppose that a population is stratified into $K$ strata. Let $y_{ij}$ and $x_{ij}$ be the value of the study ($y$) and auxiliary ($x$) variables for the $j$th unit of the $i$th stratum $U_i$, $i = 1,\ldots, K$. From each of the strata, samples are selected independently. Let the sample $s_i$ of size $n_i$ be selected from the $i$th stratum of size $N_i$ by some sampling procedure. Let $\hat{Y}_i$ and $\hat{X}_i$ be the unbiased estimators of the totals $Y_i$ and $X_i$ of the study and auxiliary variables of the $i$th stratum, respectively. Then we may construct the following two types of regression estimators, viz. separate and combined regression estimators for the total $Y$ analogous to separate and combined ratio estimators.

### *9.3.8.1 Separate Regression Estimator*

Here we assume that the linear regressions of $y$ and $x$ for the different strata are different and that the population totals of the auxiliary information $X_i$'s

of all the strata are known. Under this situation, we construct a regression estimator of the total $Y_i$ as

$$\widehat{Y}_{i(reg)} = \widehat{Y}_i - \widehat{\beta}_i\left(\widehat{X}_i - X_i\right) \tag{9.3.21}$$

where $\widehat{\beta}_i = \widehat{Cov}\left(\widehat{Y}_i, \widehat{X}_i\right)/\widehat{V}\left(\widehat{X}_i\right)$ is an estimate of the population regression coefficient $\beta_i = Cov\left(\widehat{Y}_i, \widehat{X}_i\right)/V\left(\widehat{X}_i\right)$ of the $i$th stratum.

The separate regression estimators for the population total $Y$ is given by

$$\widehat{Y}_{reg}(s) = \sum_{i=1}^{K} \widehat{Y}_{i(reg)} \tag{9.3.22}$$

The bias of $\widehat{Y}_{reg}(s)$ is

$$B(s) = \sum_{i=1}^{K} B\left(\widehat{Y}_{i(reg)}\right)$$

where $B\left(\widehat{Y}_{i(reg)}\right)$ is the bias of $\widehat{Y}_{i(reg)}$.

Eqs. (9.3.3) and (9.3.5) yield

$$\begin{aligned} B(s) &= -\sum_{i=1}^{K} Cov\left(\widehat{\beta}_i, \widehat{X}_i\right) \\ &\cong \sum_{i=1}^{K} \beta_i \left[\frac{Cov\left\{\widehat{V}\left(\widehat{X}_i\right), \widehat{X}_i\right\}}{V\left(\widehat{X}_i\right)} - \frac{Cov\left\{\widehat{Cov}\left(\widehat{X}_i, \widehat{Y}_i\right), \widehat{X}_i\right\}}{Cov\left(\widehat{X}_i, \widehat{Y}_i\right)}\right] \end{aligned} \tag{9.3.23}$$

An approximate expression for the MSE of $\widehat{Y}_{reg}(s)$ is obtained from Eqs. (9.3.6) and (9.3.7) as

$$M\left[\widehat{Y}_{reg}(s)\right] \cong \sum_{i=1}^{K} V\left\{\left(\widehat{Y}_i\right) - 2\beta_i Cov\left(\widehat{X}_i, \widehat{Y}_i\right) + \beta_i^2 V(X_i)\right\} \tag{9.3.24}$$

$$= \sum_{i=1}^{K} \left(1 - \rho^2_{\hat{X}_i, \hat{Y}_i}\right) V\left(\widehat{Y}_i\right) \tag{9.3.25}$$

where $\rho_{\hat{X}_i, \hat{Y}_i}$ is the correlation between $\widehat{X}_i$ and $\widehat{Y}_i$.

In case samples are selected from each of the strata by SRSWOR, we have

$$\widehat{Y}_{reg}(s) = \sum_{i=1}^{K} N_i\left\{\overline{y}_i - \widehat{\beta}_i\left(\overline{x}_i - \overline{X}_i\right)\right\} \tag{9.3.26}$$

$$B(s) = \sum_{i=1}^{K} \frac{N_i(N_i - n_i)}{(N_i - 2)} \frac{\beta_i}{n_i} \left(\frac{\mu_{30}^{(i)}}{\mu_{20}^{(i)}} - \frac{\mu_{21}^{(i)}}{\mu_{11}^{(i)}}\right) \tag{9.3.27}$$

$$
\begin{aligned}
M\left[\widehat{Y}_{reg}(s)\right] &\cong \sum_{i=1}^{K} N_i^2(1-f_i)\left(S_{yi}^2 - 2\beta_i S_{xyi} + \beta_i^2 S_{xi}^2\right)\Big/ n_i \\
&= \sum_{i=1}^{K} N_i^2(1-f_i)S_{qi}^2\Big/ n_i \\
&= \sum_{i=1}^{K} N_i^2(1-f_i)\left(1-\rho_i^2\right)S_{yi}^2\Big/ n_i
\end{aligned}
\tag{9.3.28}
$$

$$
\begin{aligned}
\widehat{M}\left[\widehat{Y}_{reg}(s)\right] &\cong \sum_{i=1}^{K} N_i^2(1-f_i)s_{qi}^2\Big/ n_i \\
&= \sum_{i=1}^{K} N_i^2(1-f_i)\left(1-\widehat{\rho}_i^2\right)s_{yi}^2\Big/ n_i
\end{aligned}
\tag{9.3.29}
$$

where $f_i = n_i/N_i$, $\overline{y}_i = \sum_{j\in s_i} y_{ij}\Big/ n_i$, $\overline{x}_i = \sum_{j\in s_i} x_{ij}\Big/ n_i$,

$\widehat{\beta}_i = \sum_{j\in s_i}\left(x_{ij}-\overline{x}_i\right)\left(y_{ij}-\overline{y}_i\right)\Big/ \sum_{j\in s_i}\left(x_{ij}-\overline{x}_i\right)^2, S_{yi}^2 = \sum_{j\in U_i}\left(y_{ij}-\overline{Y}_i\right)^2\Big/(N_i-1)$,

$S_{xi}^2 = \sum_{j\in U_i}\left(x_{ij}-\overline{X}_i\right)^2\Big/(N_i-1)$, $S_{xyi} = \sum_{j\in U_i}\left(y_{ij}-\overline{Y}_i\right)\left(x_{ij}-\overline{X}_i\right)\Big/(N_i-1)$,

$\overline{X}_i = \sum_{j\in U_i} x_{ij}\Big/ N_i$, $\overline{Y}_i = \sum_{j\in U_i} y_{ij}\Big/ N_i$, $\beta_i = S_{xyi}^2\Big/ S_{xi}^2$, $\rho_i = S_{xyi}\Big/(S_{xi}S_{yi})$,

$s_{xi}^2 = \sum_{j\in s_i}\left(x_{ij}-\overline{x}_i\right)^2\Big/(n_i-1)$, $s_{yi}^2 = \sum_{j\in s_i}\left(y_{ij}-\overline{y}_i\right)^2\Big/(n_i-1)$,

$s_{xyi} = \sum_{j\in s_i}\left(x_{ij}-\overline{x}_i\right)\left(y_{ij}-\overline{y}_i\right)\Big/(n_i-1)$, $\overline{x}_i = \sum_{j\in s_i} x_{ij}\Big/ n_i$ $\overline{y}_i = \sum_{j\in s_i} y_{ij}\Big/ n_i$,

$\widehat{\rho}_i = s_{xyi}\Big/(s_{xi}s_{yi})$, $S_{qi}^2 = \sum_{j\in U_i}\left(q_{ij}-\overline{Q}_i\right)^2\Big/(N_i-1)$, $q_{ij} = y_{ij} - \beta_i x_{ij}$,

$\overline{Q}_i = \sum_{j\in U_i} q_{ij}\Big/ N_i$, $s_{qi}^2 = \sum_{j\in s_i}\left(\widehat{q}_{ij}-\overline{q}_i\right)^2\Big/(n_i-1)$, $\widehat{q}_{ij} = y_{ij} - \widehat{\beta}_i x_{ij}$,

$\overline{q}_i = \sum_{j\in s_i}\widehat{q}_{ij}\Big/ n_i$, and $\mu_{kl}^{(i)} = \sum_{j\in U_i}\left(x_{ij}-\overline{X}_i\right)^k\left(y_{ij}-\overline{Y}_i\right)^l\Big/ N_i$.

#### 9.3.8.2 Combined Regression Estimator

Here we assume that a single regression line may be fitted over the entire population. The combined regression estimator for the total $Y$ is given by

$$
\widehat{Y}_{reg}(c) = \widehat{Y}_{st} - \widehat{\beta}_c\left(\widehat{X}_{st} - X\right) \tag{9.3.30}
$$

where $\widehat{X}_{st} = \sum_{i=1}^{K}\widehat{X}_i$, $\widehat{Y}_{st} = \sum_{i=1}^{K}\widehat{Y}_i$ and $\widehat{\beta}_c = \widehat{Cov}\left(\widehat{X}_{st}, \widehat{Y}_{st}\right)\Big/\widehat{V}\left(\widehat{X}_{st}\right)$ = an estimate of the combined regression coefficient $\beta_c = Cov\left(\widehat{X}_{st}, \widehat{Y}_{st}\right)\Big/ V\left(\widehat{X}_{st}\right)$.

The bias of $\widehat{Y}_{reg}(c)$ is obtained by using Eqs. (9.3.3) and (9.3.5) as

$$B(c) = -Cov\left(\widehat{\beta}_c, \widehat{X}_{st}\right) \cong \beta_c \left[\frac{Cov\left\{\widehat{V}\left(\widehat{X}_{st}\right), \widehat{X}_{st}\right\}}{V\left(\widehat{X}_{st}\right)} - \frac{Cov\left\{\widehat{C}ov\left(\widehat{X}_{st}, \widehat{Y}_{st}\right), \widehat{X}_{st}\right\}}{Cov\left(\widehat{X}_{st}, \widehat{Y}_{st}\right)}\right] \tag{9.3.31}$$

An approximate MSE of $\widehat{Y}_{reg}(c)$ upto first order of approximation is obtained from Eq. (9.3.6) as

$$\begin{aligned} M\left[\widehat{Y}_{reg}(c)\right] &\cong V\left(\widehat{Y}_{st} - \beta_c \widehat{X}_{st}\right) \\ &= \sum_{i=1}^{K}\left\{V\left(\widehat{Y}_i\right) - 2\beta_c \mathrm{cov}\left(\widehat{X}_i, \widehat{Y}_i\right) + \beta_c^2 V(X_i)\right\} \\ &= \sum_{i=1}^{K} V\left(\widehat{Q}_{ci}\right) \end{aligned} \tag{9.3.32}$$

where $\widehat{Q}_{ci} = \widehat{Y}_i - \beta_c \widehat{X}_i$.

An approximate unbiased estimator of $M\left[\widehat{Y}_{reg}(c)\right]$ is

$$\widehat{M}\left[\widehat{Y}_{reg}(c)\right] = \sum_{i=1}^{K}\left\{\widehat{V}\left(\widehat{Y}_i\right) - 2\widehat{\beta}_c \widehat{C}ov\left(\widehat{X}_i, \widehat{Y}_i\right) + \widehat{\beta}_c^2 \widehat{V}(X_i)\right\} \tag{9.3.33}$$

For SRSWOR sampling, $\widehat{Y}_{reg}(c)$ reduces to

$$\widehat{Y}_{reg}(c) = \left(\sum_{i=1}^{K} N_i \overline{y}_i\right) - \widehat{\beta}_c \left(\sum_{i=1}^{K} N_i\left(\overline{x}_i - \overline{X}\right)\right) \tag{9.3.34}$$

where $\widehat{\beta}_c = \dfrac{\sum_{i=1}^{K} N_i^2(1-f_i)s_{xyi}/n_i}{\sum_{i=1}^{K} N_i^2(1-f_i)s_{xi}^2/n_i}$.

The expressions for MSE and its approximate unbiased estimator of $\widehat{Y}_{reg}(c)$ under SRSWOR are, respectively, given by

$$\begin{aligned} M\left[\widehat{Y}_{reg}(c)\right] &\cong \sum_{i=1}^{K} N_i^2(1-f_i)\left\{S_{yi}^2 - 2\beta_c S_{xyi} + \beta_c^2 S_{xi}^2\right\}\Big/ n_i \\ &= \sum_{i=1}^{K} N_i^2(1-f_i)S_{gi}^2\Big/ n_i \end{aligned} \tag{9.3.35}$$

and

$$\begin{aligned} \widehat{M}\left[\widehat{Y}_{reg}(c)\right] &\cong \sum_{i=1}^{K} N_i^2(1-f_i)\left\{s_{yi}^2 - 2\widehat{\beta}_c s_{xyi} + \widehat{\beta}_c^2 s_{xi}^2\right\}\Big/ n_i \\ &= \sum_{i=1}^{K} N_i^2(1-f_i)s_{gi}^2\Big/ n_i \end{aligned} \tag{9.3.36}$$

where $S_{gi}^2 = \sum_{j \in U_i} \left(g_{ij} - \overline{G}_i\right)^2 / (N_i - 1)$, $g_{ij} = y_{ij} - \beta_c x_{ij}$, $\overline{G}_i = \sum_{j \in U_i} g_{ij} / N_i$, $s_{gi}^2 = \sum_{j \in s_i} \left(\widehat{g}_{ij} - \overline{g}_i\right)^2 / (n_i - 1)$, $\widehat{g}_{ij} = y_{ij} - \widehat{\beta}_c x_{ij}$, and $\overline{g}_i = \sum_{j \in s_i} \widehat{g}_{ij} / n_i$.

#### 9.3.8.3 Comparison Between Combined and Separate Regression Estimators

Approximate expressions of biases and MSEs for separate and combined regression estimators are obtained assuming sample sizes are large for each of the strata. The bias of a regression estimator decreases with the increase in sample size. The total sample size $n$ may be large, but the sample sizes $n_i$'s for each of the strata may not be large. Hence the bias of $\widehat{Y}_{i(reg)}$ may not be negligible compared to its standard error if the sample size $n_i$ is small. Hence the amount of bias of the separate regression estimator is expected to be much larger than the combined regression estimator.

It is obvious that the separate regression estimator $\widehat{Y}_{reg}(s)$ is more efficient than the combined regression estimator $\widehat{Y}_{reg}(c)$ because

$$M\left[\widehat{Y}_{reg}(c)\right] - M\left[\widehat{Y}_{reg}(s)\right] \cong \sum_{i=1}^{K} \left\{ V\left(\widehat{Y}_i - \beta_c \widehat{X}_i\right) - V\left(\widehat{Y}_i - \beta_i \widehat{X}_i\right) \right\} \geq 0$$

as $V\left(\widehat{Y}_i - \beta_c \widehat{X}_i\right)$ attains a minimum when $\beta_c = \beta_i$.

In practice, the values of $\beta_i$'s are unknown and they are estimated from the samples. In case the sample sizes are not large enough for each of the strata, the estimate $\widehat{\beta}_i$ becomes unreliable and hence the separate regression estimator becomes inefficient. But, if the regression coefficients are approximately equal, then $\widehat{\beta}_c$, the estimator of $\beta_c$ becomes more efficient than each of the separate estimators $\widehat{\beta}_i$ because the former is computed using a relatively large sample size. In this situation combined regression estimator certainly becomes more efficient than the separate regression estimators. As a rule of thumb, a combined regression estimator should be used in case regression lines for each of the strata are the same. Otherwise, separate regression estimator should be used.

Example 9.3.2

From the data given in Example 8.7.1, estimate the average wage of the factory workers by using (i) separate and (ii) combined regression estimators when average years of services of the three categories are known to be 15, 8, and 3 years, respectively. Estimate the relative efficiency of the separate regression estimator with respect to the combined regression estimator.

Here we prepare the following tables:

| Strata | $N_i$ | $\overline{X}_i$ | $N_i\,\overline{X}_i$ | $n_i$ | $\overline{x}(s_i)$ | $\overline{y}(s_i)$ | $N_i\,\overline{x}(s_i)$ | $N_i\,\overline{y}(s_i)$ | $s_{x_i}^2$ | $s_{y_i}^2$ | $s_{xy_i}$ | $\widehat{\beta}_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Skilled | 50 | 15 | 750 | 10 | 16.3 | 610 | 815 | 30,500 | 59.122 | 13,222.220 | 824.444 | 13.944 |
| Semiskilled | 75 | 8 | 600 | 12 | 9 | 425 | 675 | 31,875 | 6.727 | 6,590.909 | 150.000 | 22.297 |
| Unskilled | 100 | 3 | 300 | 15 | 3.8 | 260 | 380 | 26,000 | 2.171 | 12,571.430 | 134.285 | 61.842 |
| Total | 225 | | 1650 | | | | 1,870 | 88,375 | | | | |

| $\widehat{Y}_{i(reg)}$ | $\widehat{\rho}_i = r_i$ | $\dfrac{N_i^2(1-f_i)(1-r_i^2)s_{yi}^2}{n_i}$ | $\dfrac{N_i^2(1-f_i)s_{xyi}}{n_i}$ | $\dfrac{N_i^2(1-f_i)s_{xi}^2}{n_i}$ | $\dfrac{N_i^2(1-f_i)\left\{s_{yi}^2 - 2\widehat{\beta}_c s_{xyi} + \widehat{\beta}_c^2 s_{xi}^2\right\}}{n_i}$ |
|---|---|---|---|---|---|
| 29,593.591 | 0.932 | 345,110.568 | 164,888.889 | 11,824.440 | 660,175.087 |
| 30,202.702 | 0.712 | 1,278,236.333 | 59,062.500 | 2,648.864 | 1,305,202.464 |
| 21,052.631 | 0.813 | 2,417,919.799 | 76,095.238 | 1,230.476 | 4,665,162.568 |
| 80,848.925 | | 4,041,266.701 | 300,046.627 | 15,703.780 | 6,630,540.119 |

Separate regression estimate for the average wage $\overline{Y}$ is

$$\widehat{\overline{Y}}_{reg}(s) = \widehat{Y}_{reg}(s)/N = \sum_{i=1}^{K} \widehat{Y}_{i(reg)}/N = 80848.925/225 = 359.329$$

Estimated MSE of $\overline{Y}_{reg}(s) = \widehat{M}\left\{\widehat{Y}_{reg}(s)\right\}/N^2$

$$= \left\{\sum_{i=1}^{K} N_i^2(1-f_i)\left(1-\widehat{\rho}_i^2\right)s_{yi}^2/n_i\right\}\Big/N^2$$

$$= 4041266.701/(225)^2 = 79.827$$

An estimate of the combined regression $\beta_c$ is

$$\widehat{\beta}_c = \frac{\sum_{i=1}^{3} N_i^2(1-f_i)s_{xyi}/n_i}{\sum_{i=1}^{3} N_i^2(1-f_i)s_{xi}^2/n_i} = 300046.627/15703.780 = 19.107$$

Combined regression estimate of the average wage $\overline{Y}$ is

$$\widehat{\overline{Y}}_{reg}(c) = \widehat{Y}_{reg}(c)/N = \left(\sum_{i=1}^{K} N_i\overline{y}_i/N\right) - \widehat{\beta}_c\left(\sum_{i=1}^{K} N_i\overline{x}_i/N - \overline{X}\right)$$

$$= 88375/225 - 19.107 \times (1870/225 - 1650/225) = 374.096$$

Estimated MSE of $\widehat{M}\left[\widehat{\overline{Y}}_{reg}(c)\right]$ is

$$\widehat{M}\left[\widehat{\overline{Y}}_{reg}(c)\right] = \left[\sum_{i=1}^{K} N_i^2(1-f_i)\left(s_{yi}^2 - 2\widehat{\beta}_c s_{xyi} + \widehat{\beta}_c^2 s_{xi}^2\right)/n_i\right]\Big/N^2$$

$$= 6630540.119/(225)^2 = 130.974$$

Estimated relative efficiency of separate regression estimator with respect to the combined regression estimator for estimating $\overline{Y} = \widehat{M}\left[\widehat{\overline{Y}}_{reg}(c)\right] \times 100\Big/\widehat{M}\left[\widehat{\overline{Y}}_{reg}(s)\right] = (130.974/79.827) \times 100 = 164.071$ percent.

### 9.3.9 Regression Estimator for Several Auxiliary Variables

Suppose that we have $p$ auxiliary variables $x_1,\ldots, x_p$, each of which is well related to the study variable $y$. Let $x_{ij}$ be the value of $i$th unit of the jth auxiliary variable $x_j$. In this case, we may set the following two types of regression estimators for the total $Y$ when the totals of each of the auxiliary variables are known.

#### 9.3.9.1 Multivariate Regression Estimator

$$\widehat{Y}_{reg}(1) = \widehat{Y} - \lambda_1\left(\widehat{X}_1 - X_1\right) - \cdots - \lambda_j\left(\widehat{X}_j - X_j\right) - \cdots - \lambda_p\left(\widehat{X}_p - X_p\right) \tag{9.3.37}$$

where $\widehat{X}_j = \sum_{i\in s} b_{si}x_{ji}$ is an unbiased estimator of the total $X_j$ and $\lambda_1,\ldots, \lambda_p$ are known constants. The optimum values of $\lambda_1,\ldots, \lambda_p$ are obtained by minimizing the variance of $\widehat{Y}_{reg}(1)$. The variance of $\widehat{Y}_{reg}(1)$ is

$$V\left\{\widehat{Y}_{reg}(1)\right\} = V\left(\widehat{Y}\right) + \sum_{i=1}^{p}\lambda_i^2 V\left(\widehat{X}_i\right) + \sum_{i\neq}^{p}\sum_{j=p}^{p}\lambda_i\lambda_j Cov\left(\widehat{X}_i, \widehat{X}_j\right) - 2\sum_{i=1}^{p}\lambda_i Cov\left(\widehat{Y}, \widehat{X}_i\right) \tag{9.3.38}$$

Differentiating $V\left\{\widehat{Y}_{reg}(1)\right\}$ with respect to $\lambda_j$ and equating to zero for $j = 1,\ldots, p$, the optimum values of $\lambda_j = \beta_j$ (say) are obtained as

$$\begin{pmatrix} \beta_1 \\ \cdots \\ \beta_j \\ \cdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} V\left(\widehat{X}_1\right) & \ldots & Cov\left(\widehat{X}_1, \widehat{X}_j\right) & \ldots & Cov\left(\widehat{X}_1, \widehat{X}_p\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Cov\left(\widehat{X}_j, \widehat{X}_1\right) & \ldots & V\left(\widehat{X}_j\right) & \ldots & Cov\left(\widehat{X}_j, \widehat{X}_p\right) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ Cov\left(\widehat{X}_p, \widehat{X}_1\right) & \ldots & Cov\left(\widehat{X}_p, \widehat{X}_j\right) & \ldots & V\left(\widehat{X}_p\right) \end{pmatrix}^{-1} \times \begin{pmatrix} Cov\left(\widehat{Y}, \widehat{X}_1\right) \\ \cdots \\ Cov\left(\widehat{Y}, \widehat{X}_j\right) \\ \cdots \\ Cov\left(\widehat{Y}, \widehat{X}_p\right) \end{pmatrix} \tag{9.3.39}$$

The solution (Eq. 9.3.39) indicates that the optimum value $\beta_j$ is the partial regression coefficient of $\widehat{Y}$ on $\widehat{X}_j$.

In particular for SRSWOR sampling design $V\left(\widehat{X}_j\right) = N^2(1-f)S_{x_j}^2/n$, $Cov\left(\widehat{X}_j, \widehat{X}_k\right) = N^2(1-f)S_{x_j,x_k}^2/n$ and $Cov\left(\widehat{Y}, \widehat{X}_k\right) = N^2(1-f)S_{y,x_k}^2/n$ where $S_{x_j}^2 = \sum_{i\in U}\left(x_{ji} - \overline{x}_j\right)^2/(N-1)$,

$$S^2_{x_j,x_k} = \sum_{i\in U}\left(x_{ji}-\overline{x}_j\right)\left(x_{ki}-\overline{x}_k\right)/(N-1),$$
$$S_{y,x_k} = \sum_{i\in U}\left(y_i-\overline{y}\right)\left(x_{ki}-\overline{x}_k\right)/(N-1), \text{ and } \overline{x}_j = \sum_{i\in U} x_{ji}/N.$$

The value of $\widehat{Y}_{reg}(1)$ with the optimum value of $\lambda_i = \beta_i$ comes out as

$$V_{\min}\left(\widehat{Y}_{reg}(1)\right) = N^2(1-f)\left(1-R^2_{y|x_1,\ldots,x_k}\right)/n \tag{9.3.40}$$

where $R^2_{y|x_1,\ldots,x_k}$ = multiple correlation between $y$ and $x_1,\ldots, x_p$.

The optimum values $\beta_j$ are generally unknown. Hence regression estimator of $Y$ is obtained by replacing $\beta_j$ by its suitable estimate $\widehat{\beta}_j$ and it is given as

$$\widehat{Y}_{reg}(1) = \widehat{Y} - \widehat{\beta}_1\left(\widehat{X}_1 - X_1\right) - \cdots - \widehat{\beta}_j\left(\widehat{X}_j - X_j\right) - \cdots - \widehat{\beta}_p\left(\widehat{X}_p - X_p\right)$$

#### 9.3.9.2 Two Auxiliary Variables

For $p = 2$, Eq. (9.3.39) reduces to

$$\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} = \begin{pmatrix} S^2_{x_1} & \rho_{x_1,x_2}S_{x_1}S_{x_2}\\ \rho_{x_1,x_2}S_{x_1}S_{x_2} & S^2_{x_2}\end{pmatrix}^{-1}\begin{pmatrix}\rho_{yx_1}S_{x_1}S_y\\ \rho_{yx_2}S_{x_2}S_y\end{pmatrix}$$
$$= \begin{pmatrix}\dfrac{\rho_{yx_1}-\rho_{x_1x_2}\rho_{yx_2}}{1-\rho^2_{x_1,x_2}}\dfrac{S_y}{S_{x_1}}\\ \dfrac{\rho_{yx_2}-\rho_{x_1x_2}\rho_{yx_1}}{1-\rho^2_{x_1,x_2}}\dfrac{S_y}{S_{x_2}}\end{pmatrix}$$

where $\rho_{uv}$ is the correlation coefficient between $u$ and $v$.

Hence the variance of $\widehat{Y}_{reg}(1) = \widehat{Y} - \lambda_1\left(\widehat{X}_1 - X_1\right) - \lambda_2\left(\widehat{X}_2 - X_2\right)$ attains a minimum when $\lambda_1 = \beta_1$ and $\lambda_2 = \beta_2$. The corresponding minimum variance is

$$V_{\min}\left\{\widehat{Y}_{reg}(1)\right\} = N^2(1-f)\left(1-R^2_{y|x_1,x_2}\right)S^2_y/n$$

where $R^2_{y|x_1,x_2} = \dfrac{\rho^2_{yx_1}+\rho^2_{yx_2}-2\rho_{yx_1}\rho_{yx_2}\rho_{x_1x_2}}{1-\rho^2_{x_1x_2}}$

#### 9.3.9.3 Raj's Regression Estimator

Raj (1965a,b) computed $p$ difference estimators based on each of the auxiliary variables and then obtained a combined regression estimator by taking the weighted average of these separate estimators. Raj's estimator is given by

$$\widehat{Y}_{reg}(R) = \sum_{j=1}^{P} w_j\,\widehat{Y}_{reg}(j) \tag{9.3.41}$$

where $\widehat{Y}_{reg}(j) = \widehat{Y} - \lambda_j^*\left(\widehat{X}_j - X_j\right)$, $\lambda_j^*$ is a suitably chosen constant, and $\sum_{j=1}^{K} w_j = 1$.

The optimum value of $\lambda_j^*$ that minimizes the variance of $\widehat{Y}_{reg}(j)$ is $\beta_j^* = \rho_{yx_j} S_y \Big/ S_{x_j}$ and the corresponding minimum variance of $\widehat{Y}_{reg}(j)$ is $N^2(1-f)\left(1-\rho_{yx_j}^2\right) S_y^2 \Big/ n$. The optimum values of $w_i$'s that minimize the variance of $\widehat{Y}_{reg}(R)$ are obtained following Section 8.8. It is obvious that the minimum variance of Eq. (9.3.41) is higher than that of Eq. (9.3.40). Hence Raj's estimator is less efficient than the estimator given in Eq. (9.3.40). For $p = 2$, details are given by Sukhatme et al. (1984).

## 9.4 PRODUCT METHOD OF ESTIMATION

The ratio method of estimation is more efficient than the conventional estimator $\widehat{Y}$ when the study variable $y$ has a high positive correlation with the auxiliary variable $x$ and $R = Y/X$ is positive. Goodman (1960) and Murthy (1964) proposed the product estimator that is used when the study and the auxiliary variables are negatively related. The product estimator is given by

$$\widehat{Y}_p = \frac{\widehat{Y}\widehat{X}}{X} \tag{9.4.1}$$

The product estimator $\widehat{Y}_p$ is not unbiased. The bias is negligible for a large sample size. The expressions of bias and MSE of the product estimator are given as follows.

### 9.4.1 Bias of the Product Estimator

Writing $\varepsilon_x = \dfrac{\widehat{X}}{X} - 1$ and $\varepsilon_y = \dfrac{\widehat{Y}}{Y} - 1$, we get the bias of the product estimator as

$$\begin{aligned} B\left(\widehat{Y}_p\right) &= E\left[\frac{XY(1+\varepsilon_x)(1+\varepsilon_y)}{X} - Y\right] \\ &= YE(\varepsilon_x\varepsilon_y) \\ &= \frac{Cov\left(\widehat{X}, \widehat{Y}\right)}{X} \end{aligned} \tag{9.4.2}$$

An estimator of the bias $B\left(\widehat{Y}_p\right)$ is given by

$$\widehat{B}\left(\widehat{Y}_p\right) = \frac{C\,\widehat{ov}\left(\widehat{X}, \widehat{Y}\right)}{\widehat{X}} \tag{9.4.3}$$

### 9.4.2 Mean Square Error of the Product Estimator

The MSE of the product estimator is given by

$$M(\widehat{Y}_p) = E\left[\frac{XY(1+\varepsilon_x)(1+\varepsilon_y)}{X} - Y\right]^2 = Y^2 E(\varepsilon_x + \varepsilon_y + \varepsilon_x\varepsilon_y)^2 \tag{9.4.4}$$

Neglecting terms $E(\varepsilon_x^i \varepsilon_y^j)$ with $i+j>2$ in Eq. (9.4.4), we get the expression of the MSE up to the first order of approximation as

$$M(\widehat{Y}_p) = V(\widehat{Y}) + R^2 V(\widehat{X}) + 2RCov(\widehat{X}, \widehat{Y}) \tag{9.4.5}$$

An approximate unbiased estimator of $M(\widehat{Y}_p)$ is

$$\widehat{M}(\widehat{Y}_p) = \widehat{V}(\widehat{Y}) + \widehat{R}^2 \widehat{V}(\widehat{X}) + 2\widehat{R}C\widehat{o}v(\widehat{X}, \widehat{Y}) \tag{9.4.6}$$

### 9.4.3 Comparison With the Conventional Estimator

The product estimator $\widehat{Y}_p$ becomes superior to the conventional estimator $\widehat{Y}$ if

$$V(\widehat{Y}) - M(\widehat{Y}_p) \geq 0$$

$$\text{i.e. if } \rho_{\hat{X},\hat{Y}} \leq -\frac{1}{2}\frac{C_{\hat{X}}}{C_{\hat{Y}}} \quad \text{if } R>0$$

and

$$\rho_{\hat{X},\hat{Y}} \geq \frac{1}{2}\left|\frac{C_{\hat{X}}}{C_{\hat{Y}}}\right| \quad \text{if } R<0$$

where $C_{\hat{X}} = \sqrt{V(\widehat{X})}/X$ and $C_{\hat{Y}} = \sqrt{V(\widehat{Y})}/Y$ are the coefficient of variation of $\widehat{X}$ and $\widehat{Y}$, respectively.

### 9.4.4 Product Estimator for a Few Sampling Designs

#### *9.4.4.1 Fixed Effective Sample Size Design*

Consider a fixed effective sample size $n$ design and $b_{si} = 1/\pi_i$. In this case

$$\widehat{X} = \sum_{i\in s} x_i/\pi_i,\ \widehat{Y} = \sum_{i\in s} y_i/\pi_i,\ \widehat{Y}_p = \frac{\left(\sum_{i\in s} x_i/\pi_i\right)\left(\sum_{i\in s} y_i/\pi_i\right)}{X}$$

$$B(\widehat{Y}_p) = \frac{1}{X}\frac{1}{2}\sum_{i\neq}\sum_{j\in U}\Delta_{ij}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right),$$

$$\widehat{B}(\widehat{Y}_p) = \frac{1}{\widehat{X}}\frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{x_i}{\pi_i} - \frac{x_j}{\pi_j}\right)\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)$$

$$M(\widehat{Y}_p) = \frac{1}{2}\sum_{i\neq}\sum_{j\in U}\Delta_{ij}\left(\frac{F_i}{\pi_i}-\frac{F_j}{\pi_j}\right)^2, \widehat{M}(\widehat{Y}_p) = \frac{1}{2}\sum_{i\neq}\sum_{j\in s}\frac{\Delta_{ij}}{\pi_{ij}}\left(\frac{\widehat{F}_i}{\pi_i}-\frac{\widehat{F}_j}{\pi_j}\right)^2$$

where $F_i = y_i + Rx_i$ and $\widehat{F}_i = y_i + \widehat{R}x_i$.

#### 9.4.4.2 Simple Random Sampling Without Replacement

For SRSWOR, $\widehat{Y}_p = \frac{N\overline{y}_s\overline{x}_s}{\overline{X}}$, $B(\widehat{Y}_p) = \frac{N(1-f)}{n}\frac{S_{xy}}{\overline{X}}$, $\widehat{B}(\widehat{Y}_p) = \frac{N(1-f)}{n}\frac{s_{xy}}{\overline{x}_s}$, $M(\widehat{Y}_p) = N^2\frac{(1-f)}{n}\left(S_y^2 + R^2S_x^2 + 2RS_{xy}\right)$, and $\widehat{M}(\widehat{Y}_p) = N^2\frac{(1-f)}{n}\left(s_y^2 + \widehat{R}^2s_x^2 + 2\widehat{R}S_{xy}\right)$.

### 9.4.5 Unbiased Product Type Estimators

Chaudhuri and Arnab (1982a,b) proposed the unbiased product type estimator and it is constructed as follows:

Let $\widehat{X} = \sum_{i\in s} b_{si}x_i$ and $\widehat{Y} = \sum_{i\in s} b_{si}y_i$ be the unbiased estimators of the population totals $X$ and $Y$ as defined in Section 9.3.3. The sample $s$ of size $n$ is divided at random into $k$ groups so that the ith group $s_i$ consists of $n_i$ units $\left(\sum_{i=1}^{k} n_i = n\right)$. Let the product estimator of $Y$ based on the sample $s_j$ is denoted by

$$\widehat{Y}_p(j) = \widehat{X}_j\widehat{Y}_j/X \tag{9.4.7}$$

where $\widehat{Y}_j = n\sum_{i\in s_j}\xi_i/n_j$, $\widehat{X}_j = n\sum_{i\in s_j}\eta_i/n_j$, $\xi_i = b_{si}y_i$, and $\eta_i = b_{si}x_i$.

The proposed modified product estimator is defined as follows:

$$\widehat{Y}_p(m) = \sum_{i=1}^{k} w_i\widehat{Y}_p(j) + \left(1 - \sum_{i=1}^{k} w_i\right)\widehat{Y}_p \tag{9.4.8}$$

where $w_i$'s are suitably chosen weights.

Denoting $\varepsilon_{jx} = \{\widehat{X}_j - X\}/X$ and $\varepsilon_{jy} = (\widehat{Y}_j - Y)/Y$, we have the following theorem due to Chaudhri and Arnab (1982a,b).

Theorem 9.4.1

The modified product estimator $\widehat{Y}_p(m)$ is unbiased for $Y$ and has the same efficiency as $\widehat{Y}_p$ upto the first order of approximation if

$$(i)\sum_{i=1}^{k} c_iw_i + \left(1 - \sum_{i=1}^{k} w_i\right) = 0 \text{ and } (ii)\ w_i = \frac{n_i}{n}\left(\sum_{i=1}^{k} w_i\right)$$

where $c_i = E(\varepsilon_{ix}\varepsilon_{iy})/E(\varepsilon_x\varepsilon_y)$.

Proof

$$E\{\widehat{Y}_p(m)\} = \sum_{i=1}^{k} w_i E\{\widehat{Y}_p(j)\} + \left(1 - \sum_{i=1}^{k} w_i\right) E(\widehat{Y}_p)$$
$$= Y\left\{\sum_{i=1}^{k} w_i E(1 + \varepsilon_{ix} + \varepsilon_{iy} + \varepsilon_{ix}\varepsilon_{iy}) + \left(1 - \sum_{i=1}^{k} w_i\right) E(1 + \varepsilon_x + \varepsilon_y + \varepsilon_x\varepsilon_y)\right\}$$
$$= Y\left[\left\{\sum_{i=1}^{k} w_i c_i + \left(1 - \sum_{i=1}^{k} w_i\right)\right\} E(\varepsilon_x\varepsilon_y) + 1\right] \tag{9.4.9}$$

From Eq. (9.4.9), we note that $E\{\widehat{Y}_p(m)\}$ is unbiased if and only if

$$\sum_{i=1}^{k} c_i w_i + \left(1 - \sum_{i=1}^{k} w_i\right) = 0 \tag{9.4.10}$$

The MSE of $\widehat{Y}_p(m)$ is

$$M\{\widehat{Y}_p(m)\} = E\{\widehat{Y}_p(m) - Y\}^2$$
$$= Y^2\left[\sum_{i=1}^{k} w_i(1 + \varepsilon_{ix} + \varepsilon_{iy} + \varepsilon_{ix}\varepsilon_{iy}) + \left(1 - \sum_{i=1}^{k} w_i\right)(1 + \varepsilon_x + \varepsilon_y + \varepsilon_x\varepsilon_y) - 1\right]^2$$

Now neglecting terms with $E\left(\varepsilon_x^r \varepsilon_y^l\right)$ for $r + l > 2$, we get upto the first order of approximation:

$$M\{\widehat{Y}_p(m)\} \cong Y^2 E\left[\sum_{i=1}^{k} w_i(\varepsilon_{ix} + \varepsilon_{iy}) + \left(1 - \sum_{i=1}^{k} w_i\right)(\varepsilon_x + \varepsilon_y)\right]^2$$
$$= Y^2 E\left[E\left\{\sum_{i=1}^{k} w_i(\varepsilon_{ix} + \varepsilon_{iy}) + \left(1 - \sum_{i=1}^{k} w_i\right)(\varepsilon_x + \varepsilon_y)\right\}^2 \middle| s\right]$$
$$= Y^2 E\left[V\left\{\sum_{i=1}^{k} w_i(\varepsilon_{ix} + \varepsilon_{iy}) + \left(1 - \sum_{i=1}^{k} w_i\right)(\varepsilon_x + \varepsilon_y)|s\right\}\right]$$
$$+ Y^2 E\left[E\left\{\sum_{i=1}^{k} w_i(\varepsilon_{ix} + \varepsilon_{iy}) + \left(1 - \sum_{i=1}^{k} w_i\right)(\varepsilon_x + \varepsilon_y)|s\right\}\right]^2 \tag{9.4.11}$$

Because the first term of Eq. (9.4.11) is nonnegative and $E\left\{\sum_{i=1}^{k} w_i(\varepsilon_{ix}+\varepsilon_{iy})+\left(1-\sum_{i=1}^{k} w_i\right)(\varepsilon_x+\varepsilon_y)|s\right\}=\varepsilon_x+\varepsilon_y$, we find

$$M\{\widehat{Y}_p(m)\} \geq Y^2 E(\varepsilon_x+\varepsilon_y)^2 = M(\widehat{Y}_p) \tag{9.4.12}$$

Equality in Eq. (9.4.12) holds if and only if

$$\begin{aligned}&\sum_{i=1}^{k} w_i(\varepsilon_{ix}+\varepsilon_{iy})+\left(1-\sum_{i=1}^{k} w_i\right)(\varepsilon_x+\varepsilon_y)\\ &\quad = E\left\{\sum_{i=1}^{k} w_i(\varepsilon_{ix}+\varepsilon_{iy})+\left(1-\sum_{i=1}^{k} w_i\right)(\varepsilon_x+\varepsilon_y)|s\right\}=\varepsilon_x+\varepsilon_y\end{aligned} \tag{9.4.13}$$

for every $s$ with $p(s)>0$.

Now noting $\varepsilon_x+\varepsilon_y=\sum_{i=1}^{k}\frac{n_i}{n}(\varepsilon_{ix}+\varepsilon_{iy})$, the above condition (Eq. 9.4.13) reduces to

$$\sum_{i=1}^{k} w_i(\varepsilon_{ix}+\varepsilon_{iy}) = \left(\sum_{i=1}^{k} w_i\right)\sum_{i=1}^{k}\frac{n_i}{n}(\varepsilon_{ix}+\varepsilon_{iy}) \tag{9.4.14}$$

The condition (Eq. 9.4.14) is realized if we choose

$$w_i = \frac{n_i}{n}\left(\sum_{i=1}^{k} w_i\right) \quad \text{for } i=1,\ldots,k \tag{9.4.15}$$

The theorem follows from Eqs. (9.4.10) and (9.4.15).

#### 9.4.5.1 Simple Random Sampling Without Replacement

Let $b_{si}=N/n$, then $\widehat{X}=N\overline{x}(s)$, $\widehat{Y}=N\overline{y}(s)$, $\widehat{X}_j=N\sum_{i\in s_j} x_i/n_j$, $\widehat{Y}_j=N\sum_{i\in s_j} y_i/n_j$, and $c_j=1+\left(\frac{1}{n_j}-\frac{1}{n}\right)\left(\frac{1}{n}-\frac{1}{N}\right)^{-1}$. In this case Eq. (9.4.10) reduces to

$$\left(\frac{1}{n}-\frac{1}{N}\right)^{-1}\sum_{j=1}^{k} w_j\left(\frac{1}{n_j}-\frac{1}{n}\right)+1=0 \tag{9.4.16}$$

In particular if we choose $n_j=n/k$, Eqs. (9.4.15) and (9.4.16) yield a very simple solution of $w_j$ as

$$w_j = -\frac{1}{k(k-1)}\frac{N-n}{N} \quad \text{for } j = 1, \ldots, k \tag{9.4.17}$$

Shukla (1976) derived Eq. (9.4.17) for $k = 2$.

## 9.5 COMPARISON BETWEEN THE RATIO, REGRESSION, PRODUCT, AND CONVENTIONAL ESTIMATORS

The regression estimator is always superior to the ratio, product, and the conventional estimator provided the estimate $\widehat{\beta}$ becomes very close to the true value $\beta$. In practice, the reliability of the estimator of the regression coefficient $\beta$ is in question, especially if the sample size is small. Hence, in this situation, the regression estimator may not be efficient. Noting the difference $M(\widehat{Y}_p) - M(\widehat{Y}_R) = 4RCov(\widehat{X}, \widehat{Y})$, we conclude (i) for $R > 0$, the ratio estimator is superior to the product estimator if $\rho_{\hat{X},\hat{Y}} > 0$ and the product estimator is superior to the ratio estimator if $\rho_{\hat{X},\hat{Y}} < 0$; (ii) for $R < 0$, the ratio estimator is superior to the product estimator if $\rho_{\hat{X},\hat{Y}} < 0$ but for $\rho_{\hat{X},\hat{Y}} > 0$, the product estimator is superior. In case $\left|\rho_{\hat{X},\hat{Y}}\right| < \frac{1}{2}\frac{C_{\hat{X}}}{C_{\hat{Y}}}$, the conventional estimator $\widehat{Y}$ is always better than the product and ratio estimator. In particular, if $C_{\hat{X}} = C_{\hat{Y}}$, then the conventional estimator becomes more efficient than both the ratio and product estimators when $\left|\rho_{\hat{X},\hat{Y}}\right| < 1/2$.

## 9.6 DUAL TO RATIO ESTIMATOR

Srivenkataramana (1980) proposed the following product type estimator, known as a dual to ratio estimator

$$\widehat{Y}_{dR} = \frac{\widehat{Y}\widehat{X}_{\bar{s}}}{X} \tag{9.6.1}$$

where $X_{\bar{s}} = \sum_{i \in \bar{s}} \frac{x_i}{1 - \pi_i}$ is an estimator for the total $X$ based on the set $\bar{s}(= U - s)$ of $N - n$ units, which are not selected in the sample $s$. So the estimator $\widehat{Y}_{dR}$ is usable when the values of the auxiliary variable $x$ are known at unit level. For SRSWOR, $\pi_i = n/N$ and the estimator $\widehat{Y}_{dR}$ reduces to

$$\begin{aligned}\widehat{Y}_{dR} &= \frac{\widehat{Y}(X - n\overline{x}_s)N}{(N-n)X} \\ &= \frac{N\widehat{Y} - n\widehat{Y}_p}{N-n}\end{aligned} \tag{9.6.2}$$

where $\widehat{Y}_p$ is the product estimator defined in Eq. (9.4.1).

### 9.6.1 Bias of the Dual Estimator

Now writing

$$\widehat{Y}_{dR} - Y = \frac{N\left(\widehat{Y} - Y\right) - n\left(\widehat{Y}_p - Y\right)}{N-n} \tag{9.6.3}$$

we find the bias of $\widehat{Y}_{dR}$ as

$$B\left(\widehat{Y}_{dR}\right) = -\frac{n}{N-n}B\left(\widehat{Y}_p\right) \tag{9.6.4}$$

Furthermore, from Section 9.4.4.2 we note

$$B\left(\widehat{Y}_{dR}\right) = -\frac{S_{xy}}{\overline{X}} \tag{9.6.5}$$

The expression (Eq. 9.6.4) indicates that the absolute bias of $\widehat{Y}_{dR}$ is less than the absolute bias of the product estimator $\widehat{Y}_p$ if the sampling fraction $f = n/N < 1/2$. The expression (Eq. 9.6.5) indicates that $\widehat{Y}_{dR}$ is an inconsistent estimator for $Y$ because the bias does not decrease with the increase in the sample size.

### 9.6.2 Mean Square Error of the Dual Estimator

Writing $\varepsilon_x = \dfrac{\widehat{X} - X}{X}$ and $\varepsilon_y = \dfrac{\widehat{Y} - Y}{Y}$, we find

$$\widehat{Y}_{dR} - Y = Y\frac{(N-n)\varepsilon_y - n\varepsilon_x - n\varepsilon_x\varepsilon_y}{N-n} \tag{9.6.6}$$

The expression of the MSE of $\widehat{Y}_{dR}$ upto the first-order approximation comes out as

$$\begin{aligned}M\left(\widehat{Y}_{dR}\right) &= E\left(\widehat{Y}_{dR} - Y\right)^2 \\ &= N^2(1-f)\left(S_y^2 + g^2R^2S_x^2 - 2gRS_{xy}\right)\Big/n\end{aligned} \tag{9.6.7}$$

where $g = n/(N-n)$.

### 9.6.3 Comparison With Other Estimators

#### 9.6.3.1 Conventional Estimator

The $\widehat{Y}_{dR}$ becomes more efficient than the conventional estimator $\widehat{Y} = N\overline{y}(s)$ if

$$\text{(i) } \rho_{xy} \geq \frac{g}{2}\frac{C_x}{C_y} \quad \text{with } R > 0, \quad \text{or (ii) } \rho_{xy} \leq -\frac{g}{2}\left|\frac{C_x}{C_y}\right| \quad \text{with } R < 0.$$

#### 9.6.3.2 Ratio Estimator

$$M\left(\widehat{Y}_{dR}\right) - M\left(\widehat{Y}_R\right) = (g-1)\left[(g+1)R^2S_x^2 - 2R\rho_{xy}S_xS_y\right] \tag{9.6.8}$$

The estimator $\widehat{Y}_{dR}$ will be more efficient than the ratio estimator $\widehat{Y}_R$ if the difference (Eq. 9.6.8) is negative. Here we consider the following situations:

(i) $g - 1 > 0$, i.e., $f > 1/2$, in this case the difference (Eq. 9.6.8) is negative if either $\rho_{xy} > \frac{g+1}{2}\frac{C_x}{C_y}$ with $R > 0$ or $\rho_{xy} < -\frac{g+1}{2}\left|\frac{C_x}{C_y}\right|$ with $R < 0$. The situation $f > 1/2$ rarely happens in practice.

(ii) $g - 1 < 0$, i.e., $f < 1/2$, in this case (Eq. 9.6.8) is negative if either $\rho_{xy} < \frac{g+1}{2}\frac{C_x}{C_y}$ with $R > 0$ or $\rho_{xy} > -\frac{g+1}{2}\left|\frac{C_x}{C_y}\right|$ with $R < 0$. The situation $f < 1/2$ most likely happens in practice.

#### 9.6.3.3 Product Estimator

From Eq. (9.6.7) and Section 9.4.4.2, we note that

$$M\left(\widehat{Y}_{dR}\right) - M\left(\widehat{Y}_P\right) = (g+1)\left[(g-1)R^2S_x^2 - 2R\rho_{xy}S_xS_y\right] \tag{9.6.9}$$

The estimator $\widehat{Y}_{dR}$ will be more efficient than $\widehat{Y}_P$ if the difference (Eq. 9.6.9) is negative, i.e., if (i) $\rho_{xy} > \frac{g-1}{2}\frac{C_x}{C_y}$ with $R > 0$ or (ii) if $\rho_{xy} < -\frac{g-1}{2}\left|\frac{C_x}{C_y}\right|$ with $R < 0$.

## 9.7 CALIBRATION ESTIMATORS

The calibration estimator was proposed by Deville and Särndal (1992). In calibration estimators, the weights of the conventional unbiased estimators are modified by using auxiliary information. The weights of the

conventional estimator are calibrated in such a way that the average distances between the calibrated and original weights attain a minimum according to a certain distance measure and at the same time, the calibrated weights satisfy some constraints that are known as calibration equations. The calibrated estimators become asymptotically design consistent and more efficient than the corresponding conventional estimator.

Suppose our objective is to estimate the population total $Y$ on the basis of a sample $s$ selected with probability $p(s)$. Let the conventional estimator for $Y$ be the HTE

$$\widehat{Y}_{ht} = \sum_{i \in s} d_i y_i \tag{9.7.1}$$

where $d_i = 1/\pi_i$ and $\pi_i$ is the inclusion probability for the $i$th unit, assumed to be positive for every $i \in U$.

In case $X$, the population total of the auxiliary variable $x$ is known, we use this known total $X$ to form a calibrated estimator

$$\widehat{Y}_c = \sum_{i \in s} w_i y_i \tag{9.7.2}$$

where $w_i$'s are the weights that satisfy the following calibration constraint or calibration equation.

$$\sum_{i \in s} w_i x_i = X \tag{9.7.3}$$

Here it is important to note that in deriving the constraints (Eq. 9.7.3), no model between $x$ and $y$ has been assumed. To determine the modified weights $w_i$'s, we define the distance function $D(w, d)$, which satisfies the following conditions.

For every fixed $d > 0$, (i) $D(w, d) \geq 0$, differentiable with respect to $w$; (ii) strictly convex on an interval containing $d$, and such that $D(d, d) = 0$; and (iii) $D'(w, d) = \dfrac{\partial D(w, d)}{\partial w}$ is continuous.

Deville and Särndal (1992) proposed the following distance functions: (i) $D(w, d) = (w_i - d_i)^2/(d_i)$, (ii) $D(w, d) = w_i \log (w_i/d_i) - w_i + d_i$, (iii) $D(w, d) = 2\left(\sqrt{w_i} - \sqrt{d_i}\right)^2$, (iv) $-d_i \log (w_i/d_i) + w_i - d_i$, and (v) $D(w, d) = (w_i - d_i)^2/(2w_i)$

The most popular is the chi-square type distance function

$$D(w_i, d_i) = \frac{(w_i - d_i)^2}{d_i} \tag{9.7.4}$$

Here we minimize the average chi-square distance

$$\sum_{i\in s}\frac{(w_i-d_i)^2}{d_iq_i}\ \text{subject to}\ \sum_{i\in s}w_ix_i=X\ \text{ for every } s \tag{9.7.5}$$

where $1/q_i$ is a known positive weight.

For minimization consider

$$\phi=\frac{1}{2}\sum_{i\in s}\frac{(w_i-d_i)^2}{d_iq_i}-\lambda\Big(\sum_{i\in s}w_ix_i-X\Big) \tag{9.7.6}$$

with $\lambda$ as an undetermined Lagrange multiplier.

$$\begin{aligned}&\frac{\partial\phi}{\partial w_i}=0\ \text{ implies}\\ &w_i=d_i+\lambda d_iq_ix_i\end{aligned} \tag{9.7.7}$$

On eliminating $w_i$ from Eqs. (9.7.3) and (9.7.7), we get

$$\lambda=\frac{(X-\sum_{i\in s}d_ix_i)}{\sum_{i\in s}d_iq_ix_i^2} \tag{9.7.8}$$

Eqs. (9.7.7) and (9.7.8) yield

$$w_i=d_i+d_iq_ix_i\frac{(X-\sum_{i\in s}d_ix_i)}{\sum_{i\in s}d_iq_ix_i^2} \tag{9.7.9}$$

Finally substituting Eq. (9.7.9) in Eq. (9.7.2), we get the calibrated estimator as

$$\begin{aligned}\widehat{Y}_c&=\sum_{i\in s}d_iy_i-B_s\Big(\sum_{i\in s}d_ix_i-X\Big)\\ &=\widehat{Y}_{ht}-B_s\big(\widehat{X}_{ht}-X\big)\end{aligned} \tag{9.7.10}$$

where

$$\widehat{X}_{ht}=\sum_{i\in s}d_ix_i\ \text{ and }\ B_s=\frac{\sum_{i\in s}d_iq_ix_iy_i}{\sum_{i\in s}d_iq_ix_i^2} \tag{9.7.11}$$

The resulting calibrated estimator (Eq. 9.7.10) is known as generalized regression estimator (Cassel et al., 1976 and Särndal, 1982).

Remark 9.7.1

The calibrated weights $w_i$ may take negative values. Distance functions (ii), (iii), (iv) and (v) guarantee positive weights. Details of properties are given by Deville and Särndal (1992), Singh and Mohl (1996), and Anderson and Thorburn (2005), among others.

Example 9.7.1: Ratio Estimator

If we choose $q_i = 1/x_i$, then we get $B_s = \dfrac{\sum_{i\in s} d_i y_i}{\sum_{i\in s} d_i x_i}$ and the estimator $\widehat{Y}_c$ in Eq. (9.7.10) reduces to

$$\widehat{Y}_c = \frac{\widehat{Y}_{ht}}{\widehat{X}_{ht}} X = \widehat{Y}_R \tag{9.7.12}$$

the well known ratio estimator.

Example 9.7.2

If we choice of $q_i = 1$ and $d_i = N/n$, then the estimator (Eq. 9.7.10) reduces to

$$\widehat{Y}_c = N\left[\overline{y}(s) - \frac{\sum_{i\in s} y_i x_i}{\sum_{i\in s} x_i^2}\left(\overline{x}(s) - \overline{X}\right)\right] \tag{9.7.13}$$

The estimator (Eq. 9.7.13) is quite different from the regression estimator given in Eq. (9.7.12).

## 9.7.1 Efficiency of Calibrated Estimator

The calibrated estimator $\widehat{Y}_c$ is asymptotically design unbiased. The asymptotic variance (AV) of $\widehat{Y}_c$ given by Deville and Särndal (1992) is

$$AV\left(\widehat{Y}_c\right) = \sum_{i\neq} \sum_{j\in U} \Delta_{ij}\left(d_i E_i - d_j E_j\right)^2 \tag{9.7.14}$$

where $\Delta_{ij} = \pi_i\pi_j - \pi_{ij}$, $E_i = y_i - Bx_i$, and $B = \sum_{i\in U} x_i y_i q_i \Big/ \sum_{i\in U} x_i^2 q_i$.

From the expression (Eq. 9.7.14), it is evident that the estimator $\widehat{Y}_c$ is more efficient than $\widehat{Y}_{ht}$ in most situations since $AV\left(\widehat{Y}_c\right)$ is expected to be smaller than $V\left(\widehat{Y}_{ht}\right)$.

A design-consistent estimator of $AV\left(\widehat{Y}_c\right)$ is given by

$$\widehat{V}\left(\widehat{Y}_c\right) = \sum_{i\neq} \sum_{j\in s} \frac{\Delta_{ij}}{\pi_{ij}} \left(w_i \widehat{E}_i - w_j \widehat{E}_j\right)^2 \tag{9.7.15}$$

where $\widehat{E}_i = y_i - \widehat{B}_s x_i$ and $\widehat{B}_s = \left(\sum_{i\in s} w_i q_i x_i y_i\right) \Big/ \left(\sum_{i\in s} w_i q_i x_i^2\right)$.

## 9.7.2 Calibration Estimator for Several Auxiliary Variables

Suppose we have $p$ auxiliary variables $x_1,\ldots, x_p$ with known totals $X_1,\ldots, X_p$, respectively. In this situation the weights $w_i$ are obtained by minimizing

$$\sum_{i\in s} \frac{(w_i - d_i)^2}{d_i q_i}$$

Subject to

$$\sum_{i\in s} w_i x_{1i} = X_1, \ldots, \sum_{i\in s} w_i x_{ji} = X_j, \ldots, \sum_{i\in s} w_i x_{pi} = X_p$$

where $x_{ji}$ be the value of the $i$th unit of the $j$th auxiliary variable $x_j$, $j = 1,\ldots, p$.

For this minimization problem, we consider

$$\phi = \frac{1}{2} \sum_{i\in s} \frac{(w_i - d_i)^2}{d_i q_i} - \lambda_1 \left(\sum_{i\in s} w_i x_{1i} - X_1\right) - \cdots - \lambda_j \left(\sum_{i\in s} w_i x_{ji} - X_j\right) - \cdots - \lambda_p \left(\sum_{i\in s} w_i x_{pi} - X_p\right)$$

with $\lambda_1,\ldots, \lambda_p$ as Lagrange multipliers.

$\frac{\partial \phi}{\partial w_i} = 0$ implies

$$w_i = d_i + \lambda_1 d_i q_i x_{1i} + \cdots + \lambda_j d_i q_i x_{ji} + \cdots + \lambda_p d_i q_i x_{pi} \quad \text{for } i = 1, \ldots, n \tag{9.7.16}$$

On multiplying both sides of Eq. (9.7.16) by $x_{ki}$ then summing over $i \in s$ and writing $\sum_{i\in s} w_i x_{ki} = X_k$, for $k = 1,\ldots, p$, we get estimated value of $\boldsymbol{\lambda}' = (\lambda_1, \ldots, \lambda_p)$ as

$$\widehat{\boldsymbol{\lambda}} = \mathbf{A}_s^{-1}\mathbf{T}_s \tag{9.7.17}$$

where

$$\mathbf{T}_s = \begin{pmatrix} X_1 - \sum_{i\in s} d_i x_{1i} \\ \cdots \\ X_j - \sum_{i\in s} d_i x_{ji} \\ \cdots \\ X_p - \sum_{i\in s} d_i x_{pi} \end{pmatrix},$$

$$\mathbf{A}_s = \begin{pmatrix} \sum_{i\in s} d_i q_i x_{1i}^2 & \cdots & \sum_{i\in s} d_i q_i x_{ji} x_{1i} & \cdots & \sum_{i\in s} d_i q_i x_{pi} x_{1i} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i\in s} d_i q_i x_{1i} x_{ji} & \cdots & \sum_{i\in s} d_i q_i x_{ji}^2 & \cdots & \sum_{i\in s} d_i q_i x_{pi} x_{ji} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i\in s} d_i q_i x_{1i} x_{pi} & \cdots & \sum_{i\in s} d_i q_i x_{pi} x_{ji} & \cdots & \sum_{i\in s} d_i q_i x_{pi}^2 \end{pmatrix} \text{ and}$$

$$\widehat{\boldsymbol{\lambda}} = \begin{pmatrix} \widehat{\lambda}_1 \\ \cdots \\ \widehat{\lambda}_j \\ \cdots \\ \widehat{\lambda}_p \end{pmatrix}$$

Eq. (9.7.17) yields

$$\begin{aligned} \widehat{Y}_c &= \sum_{i\in s} w_i y_i \\ &= \widehat{Y}_{ht} + \widehat{\lambda}_1 \sum_{i\in s} d_i q_i x_{1i} y_i + \cdots + \widehat{\lambda}_j \sum_{i\in s} d_i q_i x_{ji} y_i + \cdots + \widehat{\lambda}_p \sum_{i\in s} d_i q_i x_{pi} \\ &= \widehat{Y}_{ht} - B_{1s}\left(\widehat{X}_{ht1} - X_1\right) - \cdots - B_{js}\left(\widehat{X}_{htj} - X_j\right) - \cdots B_{ps}\left(\widehat{X}_{htp} - X_p\right) \end{aligned} \tag{9.7.18}$$

where $\widehat{X}_{htj} = \sum_{i \in s} d_i x_{ji}$, $\mathbf{B}_s' \mathbf{A}_s = \left( \sum_{i \in s} d_i q_i x_{1i} y_i, \ldots, \sum_{i \in s} d_i q_i x_{ji} y_i, \ldots, \sum_{i \in s} d_i q_i x_{pi} y_i \right)$, and $\mathbf{B}_s' = (B_{1s}, \ldots, B_{js}, \ldots, B_{ps})$.

Following Deville and Särndal (1992), the a symptotic variance (AV) of $\widehat{Y}_c$ and its approximate unbiased estimators are obtained as

$$AV\left(\widehat{Y}_c\right) = \sum_{i \neq} \sum_{j \in U} \Delta_{ij} \left( d_i E_i^* - d_j E_j^* \right)^2 \tag{9.7.19}$$

and

$$\widehat{V}\left(\widehat{Y}_c\right) = \sum_{i \neq} \sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \left( d_i \widehat{E}_i^* - d_j \widehat{E}_j^* \right)^2 \tag{9.7.20}$$

where $E_i^* = y_i - B_1 x_{1i} - \ldots - B_j x_{ji} - \ldots - B_p x_{pi}$, $\widehat{E}_i^* = y_i - B_{1s} x_{1i} - \ldots - B_{js} x_{ji} - \ldots - B_{ps} x_{pi}$, and

$$\begin{pmatrix} B_1 \\ \cdots \\ B_j \\ \cdots \\ B_p \end{pmatrix} = \begin{pmatrix} \sum_{i \in U} q_i x_{1i}^2 & \cdots & \sum_{i \in U} q_i x_{ji} x_{1i} & \cdots & \sum_{i \in U} q_i x_{pi} x_{1i} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i \in U} q_i x_{1i} x_{ji} & \cdots & \sum_{i \in U} q_i x_{ji}^2 & \cdots & \sum_{i \in U} q_i x_{pi} x_{ji} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_{i \in U} q_i x_{1i} x_{pi} & \cdots & \sum_{i \in U} q_i x_{pi} x_{ji} & \cdots & \sum_{i \in U} q_i x_{pi}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i \in U} q_i x_{1i} y_i \\ \cdots \\ \sum_{i \in U} q_i x_{ji} y_i \\ \cdots \\ \sum_{i \in U} q_i x_{pi} y_i \end{pmatrix}$$

The variance estimator (Eq. 9.7.20) provides strictly design-consistent variance estimator. When design-based as well as model-based properties are considered, the following alternative variance estimator is proposed.

$$\widehat{V}_w\left(\widehat{Y}_c\right) = \sum_{i \neq} \sum_{j \in s} \frac{\Delta_{ij}}{\pi_{ij}} \left( w_i \widehat{E}_i^* - w_j \widehat{E}_j^* \right)^2$$

## 9.8 EXERCISES

**9.8.1** Describe the regression methods of estimation. Derive approximate expressions of bias and mean square errors of the regression estimator of the population mean. Show that the regression estimator is more efficient than the ratio estimator for estimating the population mean.

**9.8.2** Consider the sampling scheme where the first unit is selected with probability proportional to $\left(x_i - \overline{X}\right)^2$ and the rest $n - 1$ units are

selected from the remaining $(N-1)$ units by SRSWOR method. Show that for this sampling scheme

(i) probability of selection of an unordered sample$s$ of $n$ units is

$$p(s) = \frac{\sum_{i\in s}\left(x_i - \overline{X}\right)^2}{\binom{N-1}{n-1}\sum_{i\in U}\left(x_i - \overline{X}\right)^2} \text{ and}$$

(ii) $\frac{n(N-1)}{N(n-1)}\left[\overline{y}(s) - \widehat{\beta}^*\left\{\overline{x}(s) - \overline{X}\right\}\right]$ is an unbiased estimator $\overline{Y}$ where $\widehat{\beta}^* = \frac{\sum_{i\in s}\gamma_i\left(x_i - \overline{X}\right)}{\sum_{i\in s}\left(x_i - \overline{X}\right)^2}$ (Singh and Srivastava, 1980).

**9.8.3** Let $s$ be sample selected from a population by SRSWR method. Show that the regression estimator is more efficient than the ratio estimator under the following superpopulation model where $x$ is used as an auxiliary variable.

$$E_m(y|x_i) = \alpha + \beta x_i,\ V_m(y_i|x_i) = \sigma^2 x_i^g \text{ and } Cov(y_i, y_j|x_i, x_j) = 0 \quad \text{for } i \neq j$$

with $\sigma^2$, $g > 0$ (Raj, 1958).

**9.8.4** To determine the average height of 50 plants in a certain field, a sample of 10 plants was selected by SRSWOR method. The following table gives the exact height and eye estimate heights.

| Serial number of plants | Eye estimate height (in m) | Exact height (in m) |
|---|---|---|
| 1 | 15 | 16.8 |
| 2 | 10 | 12.2 |
| 3 | 20 | 18.5 |
| 4 | 8 | 9.0 |
| 5 | 5 | 4.8 |
| 6 | 10 | 10.5 |
| 7 | 15 | 16.2 |
| 8 | 10 | 10.0 |
| 9 | 15 | 16.3 |
| 10 | 14 | 12.2 |
| 11 | 12 | 12.0 |
| 12 | 8 | 9.5 |
| 13 | 12 | 12.8 |
| 14 | 9 | 8.5 |
| 15 | 6 | 5.8 |

Given that the average eye estimate of heights for the 50 plants is 16.5 m.

(i) Estimate the average height of the plant by ratio, regression, and product method of estimation.

(ii) Estimate the biases of the estimators used in (i).

(iii) Estimate the relative efficiencies of the above estimators with respect to the sample mean.

**9.8.5** A sample of 10 households is selected from a population of 50 households by PPSWR method of sampling using household size as a size variable. The information of household income, household expenditure, and amount of income tax paid were obtained and presented in the following table.

| **Sampled households** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| Size | 5 | 4 | 3 | 2 | 4 | 5 | 6 | 4 | 4 | 3 |
| Expenditure on food ($) | 2000 | 2500 | 1500 | 1000 | 2500 | 3000 | 4000 | 2000 | 1500 | 1200 |
| Income ($) | 5500 | 4500 | 5000 | 2500 | 7500 | 6000 | 8500 | 6000 | 2500 | 5000 |
| Income tax paid ($) | 1000 | 850 | 950 | 400 | 1400 | 1100 | 1450 | 1050 | 400 | 950 |

From the past survey it is known that the average family size and average income of the population are 4.2 and $5000, respectively. The total tax collected from the 50 households was $45,000. Use this information estimate the following:

(i) Average expenditure on food by using the regression estimator taking tax paid and income as auxiliary variables.

(ii) Give two separate estimates of the average expenditure of food by using the regression method using income and income tax paid as auxiliary variables. Give a pooled estimate of these two regression estimates.

(iii) Estimate relative efficiency of the regression estimators used in "i" with respect to the pooled estimator used in "ii."

**9.8.6** Using the data given in Exercise 9.8.5: (i) Calibrate the Hansen–Hurwitz estimator for the population mean consumption on food using total tax collection as $45,000. Estimate the standard error of the calibrated estimator you used. (ii) Obtain the calibrated estimator

for the average consumption on food using average income and total tax collection, which were \$5000 and \$45,000, respectively. (iii) Compare precision of the calibrate estimator used in "i" and "ii."

**9.8.7** To estimate the average production of milk per farm, the farms were classified into three strata: small, medium, and big with respect to cattle population. From each of the strata, samples were selected by SRSWOR method and information on production of milk and the number of cattle was collected.

| | | | Sample | | | | |
|---|---|---|---|---|---|---|---|
| Strata | Number of farms | Average number of cattle | Size | Mean no. of cattle | Mean yield of milk (kg) | SD of no. of cattle | SD of yield of milk (kg) |
| Small | 550 | 10 | 50 | 12 | 22,000 | 15 | 200 |
| Medium | 400 | 25 | 30 | 28 | 15,000 | 25 | 325 |
| Large | 250 | 45 | 20 | 50 | 10,000 | 10 | 250 |

(i) Estimate the average production of milk per farm using (a) separate and (b) combined regression methods taking cattle population as an auxiliary variable.

(ii) Estimate the relative efficiency of the separate regression estimation with respect to the combined regression estimator and comment on your findings.

## APPENDIX 9A

$$Cov\left(\overline{y}_s, s_x^2 \middle| \text{SRSWOR}\right) = Cov\left\{s_{xy}, \left(\overline{x}_s - \overline{X}\right)\right\} = \frac{1}{n}\frac{N(N-n)}{(N-1)(N-2)}\mu_{21}$$

**Proof**:

$$Cov\left(\overline{y}_s, s_x^2 \middle| \text{SRSWOR}\right)$$

$$= E\left[\overline{y}_s - \overline{Y}, s_x^2\right]$$
$$= A_0 \sum_s \left[\sum_{i\in s} h_i \left\{\sum_{i\in s} z_i^2 - \left(\sum_{i\in s} z_i\right)^2 \Big/ n\right\}\right], z_i = x_i - \overline{X}$$

$$\text{(where } A_0 = \left[ n(n-1)\binom{N}{n} \right]^{-1} \text{)}$$

$$= \frac{A_0}{n}\left[ \sum_s \left\{ (n-1)\sum_{i\in s} h_i \sum_{i\in s} z_i^2 - \left( \sum_{i\in s} h_i \right)\left( \sum_{i\neq} \sum_{j\in s} z_i z_j \right) \right\} \right]$$

$$= \frac{A_0}{n}\left[ \sum_s (n-1)\left\{ \sum_{i\in s} h_i z_i^2 + \sum_{i\neq} \sum_{j\in s} h_i z_j^2 \right\} - \left\{ 2\sum_{i\neq} \sum_{j\in s} h_i z_i z_j + \sum_{i\neq} \sum_{j\neq} \sum_{k\in s} h_i z_j z_k \right\} \right]$$

$$= \frac{A_0}{n}\left[ (n-1)\left\{ M_1 \sum_{i\in U} h_i z_i^2 + M_2 \sum_{i\neq} \sum_{j\in U} h_i z_j^2 \right\} - \left\{ 2M_2 \sum_{i\neq} \sum_{j\in U} h_i z_i z_j + M_3 \sum_{i\neq} \sum_{j\neq} \sum_{k\in U} h_i z_j z_k \right\} \right]; M_i = \binom{N-i}{n-i}$$

$$= \frac{A_0}{n}\left[ (n-1)(M_1 - M_2)\sum_{i\in U} h_i z_i^2 + 2\left\{ M_2 \sum_{i\in U} h_i z_i^2 - M_3 \sum_{i\in U} z_i^2 h_i \right\} \right]$$

$$= \frac{A_0}{n}\left[ \left\{ (n-1)(M_1 - M_2) + 2(M_2 - M_3) \right\} \right] \sum_{i\in U} h_i z_i^2$$

$$= \frac{A_0}{n}\left[ \left\{ (n-1)M_1(1 - M_2/M_1) + 2M_2(1 - M_3/M_2) \right\} \right] \sum_{i\in U} h_i z_i^2$$

$$= \frac{A_0}{n} M_1 \left[ (n-1)\frac{N-n}{N-1} + 2\frac{(n-1)}{N-1}\frac{N-n}{N-2} \right] \sum_{i\in U} h_i z_i^2$$

$$= \frac{1}{n^2}\frac{n}{N}\frac{(N-n)}{N-1}\frac{N}{N-2}\sum_{i\in U} h_i z_i^2$$

$$= \frac{1}{n}\frac{(N-n)}{N-1}\frac{1}{N-2}\sum_{i\in U} h_i z_i^2$$

$$= \frac{1}{n}\frac{(N-n)}{N-1}\frac{1}{N-2}\sum_{i\in U} \left( y_i - \overline{Y} \right)\left( x_i - \overline{X} \right)^2$$

$$= \frac{1}{n}\frac{N(N-n)}{(N-1)(N-2)}\mu_{21} \tag{A1}$$

Similarly,

$$
\begin{aligned}
&Cov\left\{s_{xy}, \left(\overline{x}_s - \overline{X}\right)\right\} \\
&= E\left[\left\{\frac{1}{n-1}\sum_{i\in s}\left(y_i - \overline{Y}\right)\left(x_i - \overline{x}_s\right)\right\}\left\{\frac{1}{n}\sum_{i\in s}\left(x_i - \overline{X}\right)\right\}\right] \\
&= E\left[\frac{1}{(n-1)n}\sum_{i\in s} h_i(z_i - \overline{z}_s)\sum_{i\in s} z_i\right] \\
&= \frac{1}{(n-1)n^2 M_o}\sum_{s}\left[\left\{(n-1)\sum_{i\in s} h_i z_i - \sum_{i\neq}\sum_{j\in s} h_i z_j\right\}\sum_{i\in s} z_i\right] \\
&= \frac{1}{(n-1)n^2 M_o}\left[(n-1)(M_1 - M_2) + 2(M_2 - M_3)\right]\sum_{i\in U} h_i z_i^2 \\
&= \frac{1}{n}\frac{N(N-n)}{(N-1)(N-2)}\mu_{21} \qquad \text{(A2)}
\end{aligned}
$$

From (A1) and (A2), we verify the result.