Metrika (1996) 43:123–133

Metrika

Estimating Regression Coefficients from Survey Data by Asymptotic Design-Cum-Model Based Approach

ARIJIT CHAUDHURI

Indian Statistical Institute, 203, B. T. Road, Calcutta-700035, India

TAPABRATA MAITI

Kalyani University, Kalyani-741235, India

Abstract: Postulating a linear regression of a variable of interest on an auxiliary variable with values of the latter known for all units of a survey population, we consider appropriate ways of choosing a sample and estimating the regression parameters. Recalling Thomsen's (1978) results on non-existence of 'design-cum-model' based minimum variance unbiased estimators of regression coefficients we apply Brewer's (1979) 'asymptotic' analysis to derive 'asymptotic-design-cum-model' based optimal estimators assuming large population and sample sizes. A variance estimation procedure is also proposed.

Key Words and Phrases: Asymptotic analysis; model; optimal estimation; regression coefficients; survey sampling; variance estimation.

1 Introduction

Thomsen (1978) considered a survey population with a variable of interest and an auxiliary variable defined on it with values of the former unknown but those of the latter as known. Postulating a super-population linear regression model connecting them he showed that "minimum 'design-cum-model' variance" linear unbiased estimators do not exist for the regression coefficients. Assuming large sizes for the population and samples we apply Brewer's (1979) asymptotic analytical approach to show that among linear 'asymptotic-design-cum model unbiased' estimators of the regression coefficients it is possible to find those for which 'asymptotic-design-cum-model' mean square errors are the minimal. Introducing certain approximations variance estimators are also proposed.

2 Asymptotic Design-Cum-Model Based Optimal Estimators

Let $U = (1, \ldots, i, \ldots, N)$ denote a survey population of a known size N which is large. Let y be a real variable defined on it with unknown values y_i with a

total Y and x be a positive valued auxiliary variable with known values x_i and a total X. We postulate a linear super-population regression model $\underline{M}$ so as to write

$$y_i = \alpha + \beta x_i + \varepsilon_i \ , \qquad i \in U \ . \tag{1}$$

Here α and β are unknown regression coefficients and ε_i's are 'random' variables with model-based means, variances and covariances as

$$E_m(\varepsilon_i) = 0 \ , \qquad V_m(\varepsilon_i) = \sigma_i^2 \text{ and } C_m(\varepsilon_i, \varepsilon_j) = 0 \text{ for } i \neq j \ .$$

Thomsen (1978) considered choosing an appropriate design p with $p(s)$ as the probability of choosing a sample s of a given size n from U and basing on s and values of y_i for i in s linear estimators $\hat{\alpha}$ and $\hat{\beta}$, say, for α and β with the following properties. We shall write $\sum$ for sum over i in U, $\sum'$ for sum over i in s, E_p for the operator of design-based expectation and assume E_p to commute with E_m, denote by π_i the inclusion probability of i in a sample and note that $\sum \pi_i = n$.

Among $\hat{\alpha} = \sum' a_i(s) y_i$, $\hat{\beta} = \sum' b_i(s) y_i$ with $a_i(s)$, $b_i(s)$ as constants free of $\underline{Y} = (y_1, \ldots, y_i, \ldots, y_N)$, subject to

$$E_m E_p(\hat{\alpha}) = \alpha \text{ and } E_m E_p(\hat{\beta}) = \beta$$

his aim was to find $\hat{\alpha}_0$, $\hat{\beta}_0$ respectively of the forms $\hat{\alpha}$, $\hat{\beta}$ such that

$$E_m E_p(\hat{\alpha} - \alpha)^2 \geq E_m E_p(\hat{\alpha}_0 - \alpha)^2 \text{ and}$$

$$E_m E_p(\hat{\beta} - \beta)^2 \geq E_m E_p(\hat{\beta}_0 - \beta)^2 \ .$$

Thomsen (1978) showed that such $\hat{\alpha}_0$, $\hat{\beta}_0$ do not exist. So, we adopt the following asymptotic approach introduced by Brewer (1979). The population U along with $\underline{Y}$ and $\underline{X} = (x_1, \ldots, x_i, \ldots, x_N)$ is supposed to hypothetically re-appear $T(>1)$ times. Every time it re-appears a sample of size n of the type s is chosen from it employing the same design p 'independently' across re-appearances. The samples so drawn are amalgamated into a pooled sample s_T, say. The new design giving the selection probability for s_T is denoted by p_T and sum over units in s_T by $\sum_T'$. If for a linear parametric function $\theta(\underline{Y})$ of $\underline{Y}$ there exists an estimator e based on s, say $e(s)$, then when based on s_T, it i.e., $e(s_T)$ should estimate $T\theta(\underline{Y})$. Just as for any given s and y_i for i in s, one defines a real number $e(s)$ as the value of a function e taken as a point estimator for $\theta(\underline{Y})$, so is $e(s_T)$ defined as the value of the same function e when s_T and y_j for j in s_T

are given and purported to be taken as a point estimator for $T\theta(\underline{Y})$. So, it is desirable, allowing T to increase indefinitely that

$$\lim_{T\to\infty} \mathrm{E}_{p_T}\left(\frac{1}{T}e(s_T)\right),$$

written as $\lim E_p(e)$, in brief, should 'equal' or 'be close to' $\theta(\underline{Y})$. So, revising Thomsen's (1978) above-noted requirements we require that $\hat{\alpha}$, $\hat{\beta}$ above should be subject respectively to

$$\lim E_p E_m(\hat{\alpha}) = \alpha \tag{2}$$

$$\lim E_p E_m(\hat{\beta}) = \beta \tag{3}$$

and among such $\hat{\alpha}$, $\hat{\beta}$ we need $\hat{\alpha}_0$, $\hat{\beta}_0$ such that

$$\lim E_p E_m(\hat{\alpha} - \alpha)^2 \geq \lim E_p E_m(\hat{\alpha}_0 - \alpha)^2 \text{ and} \tag{4}$$

$$\lim E_p E_m(\hat{\beta} - \beta)^2 \geq \lim E_p E_m(\hat{\beta}_0 - \beta)^2 \ . \tag{5}$$

The requirement (2) leads to

$$I \qquad 1 = \lim_{T\to\infty} E_{p_T}\left(\frac{1}{T}\sum{}'_T a_i(s_T)\right) = \lim E_p\left(\sum{}' a_i(s)\right) \text{ and}$$

$$II \qquad 0 = \lim_{T\to\infty} E_{p_T}\left(\frac{1}{T}\sum{}'_T a_i(s_T)x_i\right) = \lim E_p\left(\sum{}' a_i(s)x_i\right) .$$

Similarly, (3) leads to

$$I' \qquad 0 = \lim_{T\to\infty} E_{p_T}\left(\frac{1}{T}\sum{}'_T b_i(s_T)\right) = \lim E_p\left(\sum{}' b_i(s)\right) \text{ and}$$

$$II' \qquad 1 = \lim_{T\to\infty} E_{p_T}\left(\frac{1}{T}\sum{}'_T b_i(s_T)x_i\right) = \lim E_p\left(\sum{}' b_i(s)x_i\right) .$$

Just as $a_i(s)$ is a real number free of y_i's defined uniquely for combination of a given s and i in s, so is $a_i(s_T)$ for a given s_T in place of s and i in s_T. Similarly for $b_i(s)$ and $b_i(s_T)$.

The main advantage of Brewer's (1979) approach is that we may achieve convenient simplifications by applying Slutzky's (vide Cramér (1946)) limit theorem on sequences of functions, some of which are illustrated below. Thus,

$$\begin{aligned}\lim E_p E_m(\hat{\alpha}-\alpha)^2 &= \lim E_p E_m(\hat{\alpha})^2 + \alpha^2 - 2\alpha \lim E_p E_m(\hat{\alpha}) \\ &= \lim E_p V_m(\hat{\alpha}) + (\lim E_p E_m(\hat{\alpha}))^2 - \alpha^2 \ , \\ &\quad \text{using (2) and Slutzky's theorem} \\ &= \lim E_p V_m(\hat{\alpha}) = \lim E_p\left(\sum{}' a_i^2(s)\sigma_i^2\right) \\ &= \lim_{T\to\infty} E_{p_T}\left(\frac{1}{T}\sum{}'_T a_i^2(s_T)\sigma_i^2\right) \\ &= F \ , \qquad \text{say} \ . \end{aligned} \tag{6}$$

So, to find $\hat{\alpha}_0$, we need to minimize F in (6) subject to I and II. Introducing Lagrangian multipliers $2\lambda_1$ and $2\lambda_2$ we need to solve, writing $\sum_{s_T}$ for sum over samples s_T, the equation:

$$\begin{aligned} 0 &= \frac{\delta}{\delta a_i(s_T)} \sum_{s_T} p_T(s_T)\frac{1}{T}[\textstyle\sum{}'_T a_i^2(s_T)\sigma_i^2 - 2\lambda_1 a_i(s_T) - 2\lambda_2 a_i(s_T)x_i] \\ &= \frac{2p_T(s_T)}{T}[a_i(s_T)\sigma_i^2 - \lambda_1 - \lambda_2 x_i] \ . \end{aligned}$$

This yields

$$a_i(s_T) = \frac{\lambda_1 + \lambda_2 x_i}{\sigma_i^2} \ , \qquad i \in s_T \ .$$

For this choice of $a_i(s_T)$, for every fixed T, the value of

$$E_{p_T}\left(\frac{1}{T}\sum{}'_T a_i^2(s_T)\sigma_i^2\right)$$

is minimized. By Slutzky's theorem, for the same choice, its limit, as $T \to \infty$, that is F itself is also minimized. Then by I and II we get, noting Brewer's

(1979) and Chaudhuri and Stenger's (1992) works,

$$1 = \lim_{T\to\infty} E_{p_T}\left[\frac{1}{T}\sum_T' \frac{\lambda_1 + \lambda_2 x_i}{\sigma_i^2}\right]$$

$$= \lambda_1 \sum \frac{\pi_i}{\sigma_i^2} + \lambda_2 \sum \frac{\pi_i x_i}{\sigma_i^2} \text{ and} \tag{8}$$

$$0 = \lim_{T\to\infty} E_{p_T}\left[\frac{1}{T}\sum_T' \frac{(\lambda_1 + \lambda_2 x_i)x_i}{\sigma_i^2}\right]$$

$$= \lambda_1 \sum \frac{\pi_i x_i}{\sigma_i^2} + \lambda_2 \sum \frac{\pi_i x_i^2}{\sigma_i^2} \tag{9}$$

Now writing

$$D = \sum \frac{\pi_i}{\sigma_i^2} \sum \frac{\pi_i x_i^2}{\sigma_i^2} - \left(\sum \frac{\pi_i x_i}{\sigma_i^2}\right)^2 , \tag{10}$$

from (7)–(9) we derive the optimal choice of $a_i(s)$, say, a_{i0} as

$$a_{i0} = \frac{\sum \frac{\pi_i x_i^2}{\sigma_i^2} - x_i \sum \frac{\pi_i x_i}{\sigma_i^2}}{D\sigma_i^2}$$

which involves only i no matter the sample it belongs to. So, the optimal choice of $\hat{\alpha}_0$ is

$$\hat{\alpha}_0 = \sum' a_{i0} y_i = \frac{\sum \frac{\pi_i x_i^2}{\sigma_i^2} \sum' \frac{y_i}{\sigma_i^2} - \sum' \frac{x_i y_i}{\sigma_i^2} \sum \frac{\pi_i x_i}{\sigma_i^2}}{D} . \tag{11}$$

By a similar analysis, minimizing $\lim E_p E_m(\hat{\beta} - \beta)^2$ subject to I' and II' we derive the optimal choice of $b_i(s)$ as, say, b_{i0} given by

$$b_{i0} = \frac{x_i \sum \frac{\pi_i}{\sigma_i^2} - \sum \frac{\pi_i x_i}{\sigma_i^2}}{D\sigma_i^2} ,$$

no matter the sample containing i, the resulting optimal choice of $\hat{\beta}$ is $\hat{\beta}_0$ given by

$$\hat{\beta}_0 = \sum' b_{i0} y_i = \frac{\sum \frac{\pi_i}{\sigma_i^2} \sum' \frac{y_i x_i}{\sigma_i^2} - \sum' \frac{y_i}{\sigma_i^2} \sum \frac{\pi_i x_i}{\sigma_i^2}}{D} . \tag{12}$$

So, we state the following

Theorem: Under model $\underline{M}$, we have

(i) *for* $\hat{\alpha} = \sum' a_i(s) y_i$ *subject to* $\lim E_p E_m(\hat{\alpha}) = \alpha$, $\lim E_p E_m(\hat{\alpha} - \alpha)^2 \geq \lim E_p E_m(\hat{\alpha}_0 - \alpha)^2$ *with* $\hat{\alpha}_0$ *given by* (11) *and*

(ii) *for* $\hat{\beta} = \sum' b_i(s) y_i$ *subject to* $\lim E_p E_m(\hat{\beta}) = \beta$, $\lim E_p E_m(\hat{\beta} - \beta)^2 \geq \lim E_p E_m(\hat{\beta}_0 - \beta)^2$ *with* $\hat{\beta}_0$ *given by* (12).

Remark I: If (y_i, x_i), $i \in U$ are available, the usual least squares estimators for β and α are respectively

$$\hat{\beta}_N = \frac{N \sum y_i x_i - \sum y_i \sum x_i}{N \sum x_i^2 - \left(\sum x_i \right)^2} \text{ and } \hat{\alpha}_N = \frac{\sum y_i}{N} - \hat{\beta}_N \frac{\sum x_i}{N} .$$

If (y_i, x_i) were available only for $i \in s$, then the estimators would respectively be

$$\hat{\beta}_s = \frac{\sum' w_i \sum' y_i x_i w_i - \sum' y_i w_i \sum' x_i w_i}{\sum' w_i \sum' x_i^2 w_i - \left(\sum' x_i w_i \right)^2}, \quad \hat{\alpha}_s = \frac{\sum' y_i w_i}{\sum' w_i} - \hat{\beta}_s \frac{\sum' x_i w_i}{\sum' w_i}; \quad w_i = \frac{1}{\sigma_i^2} .$$

Since x_i is available for $i \in U$ and y_i only for $i \in s$ in the present case, $\hat{\beta}_s$ may reasonably be replaced by

$$\beta_s^* = \frac{\sum w_i \pi_i \sum' y_i x_i w_i - \sum' y_i w_i \sum x_i w_i \pi_i}{\sum w_i \pi_i \sum x_i^2 w_i \pi_i - \left(\sum x_i w_i \pi_i \right)^2} ;$$

similarly for $\hat{\alpha}_s$. This β_s^* matches $\hat{\beta}_0$; so for $\hat{\alpha}_s$.

We shall take as a measure of error of $\hat{\beta}_0$ the following quantity denoted by V and get an estimator for V and treat that estimator of V, say, v as the variance estimator for $\hat{\beta}_0$. Let us write $V = \lim E_p E_m(\hat{\beta}_0 - \beta)^2$. Then we get

$$V = \lim E_p \sum{}' b_{i0}^2 \sigma_i^2$$

$$= \lim E_p \left[\sum{}' \frac{1}{\sigma_i^2} \left(x_i \sum \frac{\pi_i}{\sigma_i^2} - \sum \frac{\pi_i x_i}{\sigma_i^2} \right)^2 \right] \Big/ D^2$$

$$= \left[\left(\sum \frac{\pi_i}{\sigma_i^2} \right)^2 \sum \frac{\pi_i x_i^2}{\sigma_i^2} - \left(\sum \frac{\pi_i x_i}{\sigma_i^2} \right)^2 \sum \frac{\pi_i}{\sigma_i^2} \right] \Big/ D^2 \tag{13}$$

Remark II: Thomsen (1978) in his model considered the special case, namely,

$$\sigma_i^2 = \sigma^2 \ , \qquad \sigma(>0, \text{unknown}) \ , \tag{14}$$

for every i in U. Restricting to (14), we get

$$V = \frac{\sigma^2}{\sum \pi_i \left(x_i - \frac{\sum \pi_i x_i}{n} \right)^2} = \frac{\sigma^2}{B} \ , \qquad \text{say} \ , \tag{15}$$

writing

$$B = \sum \pi_i \left(x_i - \frac{\sum \pi_i x_i}{n} \right)^2 = \sum \pi_i x_i^2 - \frac{(\sum \pi_i x_i)^2}{n} \ . \tag{16}$$

Remark III: Assuming model $\underline{M}$ with restriction (14), the optimum design is one for which "π_i's are such that B is the maximum for a given $\underline{X} = (x_1, \ldots, x_i, \ldots, x_N)$", recalling that "$\sum \pi_i = n$". Since B is known, we need to estimate σ^2 in order to estimate V. For this we proceed as follows. Let us write $\hat{y}_i = \hat{\alpha}_0 + \hat{\beta}_0 x_i$, take $\hat{y}_i$ as a predictor for y_i and $\hat{Y} = \sum \hat{y}_i = N\hat{\alpha}_0 + \hat{\beta}_0 X$ as a predictor for Y. Let

$$e_i = y_i - \hat{y}_i = y_i - \hat{\alpha}_0 - \hat{\beta}_0 x_i \ ,$$

the residual in predicting y_i, $i \in U$. Let us also note that the following simplified forms under (14), for $\hat{\alpha}_0$, $\hat{\beta}_0$, namely,

$$\hat{\alpha}_0 = \frac{\bar{y}\sum \pi_i x_i^2 - \bar{z}\sum \pi_i x_i}{B} , \tag{17}$$

writing $z_i = x_i y_i$, $\bar{z}$, $\bar{y}$, $\bar{x}$, $\bar{\varepsilon}$ for the sample means of z_i, y_i, x_i, ε_i's and

$$\hat{\beta}_0 = \frac{n\bar{z} - \bar{y}\sum \pi_i x_i}{B} . \tag{18}$$

It is then easy to check that

$$\hat{\alpha}_0 = \bar{y} - \hat{\beta}_0 \frac{\sum \pi_i x_i}{n} . \tag{19}$$

So, we may write

$$e_i = (y_i - \bar{y}) - \hat{\beta}_0 \left(x_i - \frac{\sum \pi_i x_i}{n} \right) . \tag{20}$$

Using Brewer's (1979) approach and Slutzky's theorem let us observe, omitting the easy proofs, the following

Lemma:

$$(i) \lim E_p \left(\sum{}' x_i \right) = \sum \pi_i x_i , \qquad (ii) \lim E_p \sum{}' x_i^2 = \pi_i x_i^2 \text{ and}$$

$$(iii) \lim E_p \sum{}' (x_i - \bar{x})^2 = \sum \pi_i \left(x_i - \frac{\sum \pi_i x_i}{n} \right)^2 .$$

So, assuming a large sample size, we may approximate

$$(i)' \frac{1}{n} \sum \pi_i x_i \text{ by } \bar{x} , \qquad (ii)' \sum \pi_i \left(x_i - \frac{\sum \pi_i x_i}{n} \right)^2 \text{ by } \sum{}' (x_i - \bar{x})^2$$

$$(iii)' \hat{\beta}_0 \text{ by } \hat{\beta}_0' = \frac{n(\bar{z} - \bar{x}\bar{y})}{(n-1)s_x^2} , \tag{21}$$

the usual 'ordinary least squares' (OLS) estimator of β for simple random sampling,

(iv) $\hat{\alpha}_0$ by $\hat{\alpha}'_0 = \bar{y} - \dfrac{n\bar{x}(\bar{z} - \overline{\bar{x}\bar{y}})}{(n-1)s_x^2}$, using (19) and

(v) e_i by $\hat{e}_i = (y_i - \bar{y}) - \hat{\beta}'_0(x_i - \bar{x})$, using (20) .

Then using (1) we get

$$\hat{e}_i = (\varepsilon_i - \bar{\varepsilon}) - \frac{x_i - \bar{x}}{(n-1)s_x^2} \sum{}' (\varepsilon_i - \bar{\varepsilon})(x_i - \bar{x}) \ .$$

So,

$$E_m \frac{\sum{}' \hat{e}_i^2}{n-1} = \sigma^2 \ . \tag{22}$$

So, finally we propose the quantity

$$s_e^2 = \frac{\sum{}' \hat{e}_i^2}{n-1}$$

as an estimator of σ^2 and hence

$$v = \frac{s_e^2}{B} = \frac{s_e^2}{\sum \pi_i \left(x_i - \dfrac{\sum \pi_i x_i}{n} \right)^2} \ , \tag{23}$$

as an estimator for V.

Similarly, we may take as a measure of error of $\hat{\alpha}_0$ as an estimator of α the quantity

$$W = \lim E_p E_m (\hat{\alpha}_0 - \alpha)^2$$

which simplifies, using Slutzky's theorem, to

$$W = \lim E_p \sum{}' a_{i0}^2 \sigma_i^2$$

$$= \sum \frac{\pi_i x_i^2}{\sigma_i^2} \left[\sum \frac{\pi_i}{\sigma_i^2} \sum \frac{\pi_i x_i^2}{\sigma_i^2} - \left(\sum \frac{\pi_i x_i}{\sigma_i^2} \right)^2 \right] \Big/ D^2 \ .$$

Using restriction (14) it reduces to

$$W = \frac{\sigma^2 \sum \pi_i x_i^2}{n \sum \pi_i x_i^2 - \left(\sum \pi_i x_i\right)^2}$$

$$= \frac{\sigma^2 \sum \pi_i x_i^2}{n \sum \pi_i \left(x_i - \frac{\sum \pi_i x_i}{n}\right)^2} \tag{24}$$

So, we propose the estimator w for W i.e., w as a variance estimator for $\hat{\alpha}_0$ given by

$$w = \frac{s_e^2 \sum \pi_i x_i^2}{n \sum \pi_i \left(x_i - \frac{\sum \pi_i x_i}{n}\right)^2} . \tag{25}$$

If we retained the form $\sigma_i^2 = \sigma^2 f_i$, then we would get

$$\hat{\alpha}_0 = \frac{\sum' y_i/\pi_i}{\sum \pi_i/f_i} - \hat{\beta}_0 \frac{\sum \pi_i x_i/f_i}{\sum \pi_i/f_i} \text{ and}$$

$$\hat{\beta}_0 = \frac{\sum \frac{\pi_i}{f_i} \sum' \frac{y_i x_i}{f_i} - \sum \frac{\pi_i x_i}{f_i} \sum' \frac{y_i}{f_i}}{\sum \frac{\pi_i}{f_i} \sum \frac{\pi_i x_i^2}{f_i} - \left(\sum \frac{\pi_i x_i}{f_i}\right)^2} .$$

Using these it is possible to work out formulae for variance estimators of these $\hat{\alpha}_0$, $\hat{\beta}_0$ following the above approach, with or without approximations indicated above. But as the formulae look more cumbersome we do not present them for the sake of simplicity.

Acknowledgements: The authors are grateful to a referee for his helpful recommendations which led to an improvement upon our earlier draft. The work of the second author is supported by grant number 9/106(26)/91 – *EMR* – *I* of CSIR, India.

References

Brewer KRW (1979) A class of robust sampling designs for large-scale surveys. Jour Amer Stat Assoc 74:911–915

Chaudhuri A, Stenger H (1992) Survey sampling: Theory and methods. Marcel Dakker Inc NY

Cramér H (1946) Mathematical methods of statistics. Princeton Univ Press

Thomsen I (1978) Design and estimation problem when estimating a regression coefficient from survey data. Metrika 15:27–35

Received: 10.06.1994