DOUBLE SAMPLING IN MULTI-AUXILIARY REGRESSION ESTIMATION BASED ON CONDITIONAL SPECIFICATION

Grace O. Esimai and Chien-Pai Han
Iowa State University

1. Introduction

Consider a $(p + 1)$ random vector $\begin{pmatrix} Y \\ \underline{X} \end{pmatrix}$ which follows a multivariate normal distribution where Y is a scalar and $\underline{X}$ is a $p \times 1$ vector $(p \geq 1)$. In estimating the population mean μ_y of Y, it is well known that the precision of the estimator can be increased if auxiliary information is available. In this paper, we shall consider the linear regression estimator of μ_y with $\underline{X}$ as the auxiliary variable. To use the regression estimator we need to know the population mean $\underline{\mu}_X$ of $\underline{X}$. When $\underline{\mu}_X$ is unknown, we may take a preliminary sample to estimate it. This sampling procedure is the double sampling technique. In certain situations, an investigator may have partial information about $\underline{\mu}_X$ and suspects that $\underline{\mu}_X = \underline{\mu}_0$. In order to utilize this partial information, the investigator can perform a preliminary test about the hypothesis H_0: $\underline{\mu}_X = \underline{\mu}_0$ versus H_1: $\underline{\mu}_X \neq \underline{\mu}_0$. As an example, consider estimating the average yield per acre of a certain crop. It is known that the yield is highly correlated with the moisture and nitrogen content of the soil. Hence, the moisture and nitrogen content can be used as the auxiliary variable, $\underline{X}$. The experimenter usually does not know $\underline{\mu}_X$; but from the amount of rainfall reported by the weather bureau or other sources and from analysis by the soil science department, he believes that $\underline{\mu}_X$ should be $\underline{\mu}_0$. Once a preliminary sample is available, the investigator may test H_0. He then will use $\underline{\mu}_0$ in the regression estimator if H_0 is accepted; otherwise he uses the sample mean based on the preliminary sample. This estimator is usually known as the preliminary test estimator. If the investigator's prior information or experience is reliable, then the true mean $\underline{\mu}_X$ of $\underline{X}$ will be expected to be very close to $\underline{\mu}_0$. In this situation, the efficiency of the preliminary test estimator is high. Thus in practice, it is desirable to use the preliminary test estimator when partial information is available to the investigator.

Preliminary test estimator was first studied by Bancroft (1944). It belongs to the area of inference based on conditional specification. A bibliography on inference based on conditional specification was recently compiled by Bancroft and Han (1977).

Let $\begin{pmatrix} Y \\ \underline{X} \end{pmatrix} \sim N(\underline{\mu}, \underline{\Sigma})$; $\underline{\mu} = \begin{pmatrix} \mu \\ \underline{\mu}_X \end{pmatrix}$ and $\underline{\Sigma} = \begin{pmatrix} \sigma^2 & \underline{\Sigma}_{12} \\ \underline{\Sigma}_{21} & \underline{\Sigma}_{22} \end{pmatrix}$. We assume $\underline{X}$ is cheaply observed while Y is more expensive to observe. Let $(Y_i, X_{1i}, X_{2i}, \ldots, X_{pi})'$ $i = 1, \ldots, n_2$ be a random sample from $N(\underline{\mu}, \underline{\Sigma})$. This is supplemented by $n_1 - n_2$ $(n_1 > n_2)$ more independent observations on $\underline{X} = (X_1, \ldots, X_p)'$. In practice, the sample of n_2 observations is usually a subsample from the sample of n_1 observations. From all the observations, we define

$\bar{\underline{X}}_1 = (1/n_1)\left(\sum_{i=1}^{n_1} X_{1i}, \ldots, \sum_{i=1}^{n_1} X_{pi}\right)'$, and from the subsample, we define $\bar{y} = 1/n_2 \sum_{i=1}^{n_2} y$,

and $\bar{\underline{X}}_2 = (1/n_2)\left(\sum_{i=1}^{n_2} X_{1i}, \ldots, \sum_{i=1}^{n_2} X_{pi}\right)'$. If the vector $\underline{\mu}_X$ and $\underline{\Sigma}$ are known, then given $\underline{X}$ an unbiased estimator of μ_y is $\hat{\mu}_{y|\underline{X}}$

$= \bar{y} + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,(\underline{\mu}_X - \bar{\underline{X}}_2)$ with variance

$(1/n_2)\{\sigma^2 - \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}\}$. If $(1/n_2)\,\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}$ is considerably large, we have an appreciable gain in precision.

If $\underline{\mu}_X$ is unknown and partial information about $\underline{\mu}_X$ is available, without loss of generality we let $\underline{\mu}_0 = \underline{0}$, the linear regression preliminary test estimator is defined as

$$\hat{\mu}_{lr} = \begin{cases} \bar{y} - \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\bar{\underline{X}}_2 \\ \quad \text{if } n_1\,(\bar{\underline{X}}_1'\,\underline{\Sigma}_{22}^{-1}\,\bar{\underline{X}}_1) \leq \chi^2_{p,\alpha} \\ \bar{y} + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) \\ \quad \text{if } n_1(\bar{\underline{X}}'_1\,\underline{\Sigma}_{22}^{-1}\,\bar{\underline{X}}_1) > \chi^2_{p,\alpha} \end{cases} \qquad (1.1)$$

where $\chi^2_{p,\alpha}$ is the $100(1-\alpha)$ percent point of the Chi-squared distribution with p degrees of freedom and α is the level of significance of the preliminary test. Han (1973) studied the estimator $\hat{\mu}_{lr}$ when $p = 1$. This paper will consider the general case when $p \geq 1$. The bias,

mean squared error (MSE) and relative efficiency of $\hat{\mu}_{\ell r}$ are derived in Esimai (1977) and are given in Section 2. The optimal sample design is discussed in Section 3.

When Σ is unknown, the linear regression preliminary test estimator is

$$\tilde{\mu}_{\ell r} = \begin{cases} \bar{y} - \underline{S}_{12}\,\underline{S}_{22}^{-1}\,\bar{\underline{X}}_2 \\ \qquad \text{if } m_1 n_1 (\bar{\underline{X}}_1' \underline{S}_{22}^{-1} \bar{\underline{X}}_1) \le T_0^2 \\ \bar{y} + \underline{S}_{12}\,\underline{S}_{22}^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) \\ \qquad \text{if } m_1 n_1 (\bar{\underline{X}}_1' \underline{S}_{22}^{-1} \bar{\underline{X}}_1) > T_0^2 \end{cases} \tag{1.2}$$

where $m_1 = n_1 - 1$, T_0^2 is the $100(1-\alpha)$th percentile of the Hotelling's T^2 distribution with m_1 degrees of freedom. We define

$$\underline{S} = \begin{pmatrix} S_{11} & \underline{S}_{12} \\ \underline{S}_{21} & \underline{S}_{22} \end{pmatrix} \text{ where } S_{11} = \sum_{i=1}^{n_2} (y_i - \bar{y}),$$

$$\underline{S}_{12} = \sum_{i=1}^{n_2} (y_i - \bar{y})(\underline{X}_i - \bar{\underline{X}}_2)',$$

$$\underline{S}_{22} = \sum_{i=1}^{n_1} (\underline{X}_i - \bar{\underline{X}}_1)(\underline{X}_i - \bar{\underline{X}}_1)' \text{ and } \bar{y},\ \bar{\underline{X}}_1 \text{ and } \bar{\underline{X}}_2 \text{ are as defined above.}$$

2. Bias, MSE and Relative Efficiency of $\hat{\mu}_{\ell r}$

The joint distribution of $(\bar{\underline{X}}_1', \bar{\underline{X}}_2', \bar{y})$ is normal. Denote the acceptance region for the preliminary test by A and its complement by $\bar{A}$ and let $\chi^2_{p,\alpha} = b$.

$$\begin{aligned} E(\hat{\mu}_{\ell r}) &= E\{(\bar{y} - \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\bar{\underline{X}}_2) | A\}\, P(A) \\ &\quad + E\{[\bar{y} + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)] | \bar{A}\}\, P(\bar{A}) \\ &= \mu_y - \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\, E\{\bar{\underline{X}}_1 | \bar{A}\}\, P(\bar{A}) \\ &= \mu_y + B_1 \end{aligned} \tag{2.1}$$

B_1 is evaluated in Esimai (1977) and found to be

$$B_1 = -\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X\, H_{p+2}(b;\ \delta) \tag{2.2}$$

where $H_{p+2}(b;\ \delta)$ is the cumulative distribution function of the noncentral chi-squared distribution with $p+2$ degrees of freedom and noncentrality parameter $\delta = n\underline{\mu}_X'\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X$.

Without loss of generality, we let $\underline{\Sigma}_{22} = I$ and $\sigma^2 = 1$. Since B_1 changes sign with $\underline{\Sigma}_{12}$ and $\underline{\mu}_X$, we need only study the bias for $\mu_X > 0$ and $\rho > 0$ for $p = 1$. The bias was also studied by Han (1973) where the bias was expressed in terms of the cumulative distribution function of the standard normal distribution. The two expressions are equivalent as they should be. The general behavior of $-B_1$ is as follows. The bias is zero when $\mu_X = 0$ which is when the null hypothesis is true. It is an increasing function of ρ, but a decreasing function of α. For fixed n_1, α and ρ, the bias increases from zero and then decreases to zero as μ_X increases from zero to one. The values of $-B_1$ for $n_1 = 30$, $p = 2$ and certain values of $\underline{\Sigma}_{12}$, $\underline{\mu}_X$ and α are given in Table 1. The properties of the bias are found to be identical with those recorded for $p = 1$.

Table 1. Values of $-B_1$ for $p = 2$, $n_1 = 30$

$\underline{\mu}_X$ \ $\underline{\Sigma}_{12}'$	$\begin{pmatrix}.7\\0\end{pmatrix}$	$\begin{pmatrix}.7\\.7\end{pmatrix}$	$\begin{pmatrix}-.5\\.7\end{pmatrix}$
	$\alpha = .05$		
(0, 0)	0	0	0
(.5, 0)	.07	.07	-.05
(.5, .5)	.06	.12	.02
(1, 0)	0	0	0
	$\alpha = .10$		
(0, 0)	0	0	0
(.5, 0)	.04	.04	-.03
(.5, .5)	.04	.07	.01
(1, 0)	0	0	0

The MSE of $\hat{\mu}_{\ell r}$ was found to be $M_1 = MSE(\hat{\mu}_{\ell r}) = g_1 + h_1$ where

$$\begin{aligned} g_1 &= (1/n_2)\sigma^2 + (1/n_1)\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21} \\ &\quad -(1/n_2)\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}, \end{aligned} \tag{2.3}$$

$$\begin{aligned} h_1 &= \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X\,\underline{\mu}_X'\,\underline{\Sigma}_{22}\,\underline{\Sigma}_{21} \\ &\quad -(1/n_1)\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}\, H_{p+2}(b;\ \delta) \\ &\quad -2\,\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X\,\underline{\mu}_X'\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}[1 - H_{p+2}(b;\ \delta)] \\ &\quad +\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X\,\underline{\mu}_X'\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}[1 - H_{p+4}(b;\ \delta)]. \end{aligned}$$

Now we compare the performance of the preliminary test estimator, $\hat{\mu}_{\ell r}$ with the usual linear regression estimator, $\bar{y} + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$,

when the information of $\underline{\mu}_X$ is ignored. The relative efficiency of $\hat{\mu}_{\ell r}$ to the linear regression estimator is defined as

$$e_1 = \frac{MSE\ (\bar{y} + \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)}{MSE\ (\hat{\mu}_{\ell r})} = \frac{g_1}{g_1 + h_1} \quad (2.4)$$

Table 2. Values of e_1 for $p = 2$, $n_1 = 30$, $n_2 = 10$

$\underline{\Sigma}_{12}$ \ $\underline{\mu}_X$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}.5\\0\end{pmatrix}$	$\begin{pmatrix}.5\\.5\end{pmatrix}$	$\begin{pmatrix}1.0\\0\end{pmatrix}$
		$\alpha = .05$		
(.7, 0)	1.24	.69	.70	1.0
(.5, .5)	1.25	.83	.52	1.0
(.7, .7)	4.07	.56	.22	1.0
(-.5, .7)	1.64	.80	1.04	1.0
		$\alpha = .10$		
(.7, 0)	1.19	.78	.79	1.0
(.5, .5)	1.20	.88	.64	1.0
(.7, .7)	2.71	.67	.32	1.0
(-.5, .7)	1.48	.87	1.02	1.0
		$\alpha = .25$		
(.7, 0)	1.11	.91	.92	1.0
(.5, .5)	1.11	.95	.83	1.0
(.7, .7)	1.61	.85	.57	1.0
(-.5, .7)	1.24	.95	1.01	1.0

Without loss of generality we let $\underline{\Sigma}_{22} = I$ and $\sigma^2 = 1$. The values of e_1 for $p = 1$ are given in Han (1973) and will not be given here. The values of e_1 for $p = 2$, $n_1 = 30$, $n_2 = 10$ and certain values of $\underline{\Sigma}_{12}$, α and $\underline{\mu}_X$ are given in Table 2. It is seen that e_1 assumes maximum value at $\underline{\mu}_X = \underline{0}$. The maximum value of e_1 is an increasing function of ρ for fixed α, n_1 and n_2. The value of e_1 decreases to a minimum and then increases to unity as $\underline{\mu}_X'$ increases from $(0, 0)$.

The estimator $\tilde{\mu}_{\ell r}$ in (1.2) is given when $\underline{\Sigma}$ is unknown. The bias, B_2, and the mean square error, M_2, are derived in Esimai (1977) and are omitted here. The behavior of B_2 is the same as that of B_1 and the behavior of M_2 is similar to that of M_1.

3. The Optimal Sample Design

We shall now consider the problem of finding the optimum allocation of the sample sizes n_1 and n_2 for the estimator $\hat{\mu}_{\ell r}$ the cost function is

$$C = n_1 c_1 + n_2 c_2 \quad (3.1)$$

where c_1 is the cost of observing the vector $\underline{X}$ and c_2 is the cost of observing Y. The optimum values of n_1 and n_2 are obtained by minimizing M_1 subject to the constraint (3.1). We recall that in practice, under the supposition of a conditional specification, the experimenter has only partial information based on which he believes that $\underline{\mu}_X$ is close to $\underline{0}$. The relative efficiency of $\hat{\mu}_{\ell r}$ is the largest at $\underline{\mu}_X = \underline{0}$ and so it would be reasonable to consider the problem of optimum allocation under the optimum situation by letting $\underline{\mu}_X = \underline{0}$ in M_1. When $\underline{\mu}_X = \underline{0}$, M_1 becomes

$$M_1 = k_1/n_1 + k_2/n_2 \quad (3.2)$$

where

$$k_1 = \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ \underline{\Sigma}_{21}\ [1 - H_{p+2}\ (b;\ 0)]$$

$$k_2 = \sigma^2 - \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ \underline{\Sigma}_{21}$$

Minimizing (3.2) subject to (3.1) we find

$$n_1 = \frac{C\sqrt{k_1}}{\sqrt{k_2 c_1 c_2} + c_1\sqrt{k_1}},$$

$$n_2 = \frac{C\sqrt{k_2}}{\sqrt{k_1 c_1 c_2} + c_2\sqrt{k_2}} \quad (3.3)$$

and the optimum value of M_1 is

$$M_{1,\ opt} = \frac{(\sqrt{k_1 c_1} + \sqrt{k_2 c_2})^2}{C} \quad (3.4)$$

We now compare $M_{1,\ opt}$ with the optimum value of the MSE of $\bar{y} + \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$, the regression estimator under double sampling without using the preliminary test. If we denote the MSE of $\bar{y} + \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$ by M

$$M = k_1'/n_1 + k_2'/n_2 \quad (3.5)$$

where $k_1' = \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ \underline{\Sigma}_{21}$, $k_2' = \sigma^2 - \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ \underline{\Sigma}_{21}$ and the optimum value of M is

$$M_{opt} = \frac{(\sqrt{k_1' c_1} + \sqrt{k_2' c_2})^2}{C} \quad (3.6)$$

To compare (3.4) and (3.6) we note from (3.2) that $(1 - H_{p+2}\ (b;0))$ is a decreasing function of b with a maximum equal to unity at $b = 0$. Hence the numerator of $M_{1,\ opt}$ at most as large as that of M_{opt} and $M_{1,\ opt} \leq M_{opt}$ with equality holding for $b = 0$, i.e. when the two estimators coincide.

References

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. Ann. Math. Stat. 15, 190-204.

Bancroft, T. A. and Han, C. P. (1977). Inference based on conditional specification: A note and a bibliography. Accepted for publication in International Statistical Review.

Esimai, Grace O. (1977). Regression estimation for multivariate normal distributions. Unpublished Ph.D. thesis, Iowa State University, Ames, Iowa.

Han, C. P. (1973). Double sampling with partial information on auxiliary variables. J. Amer. Statist. Assoc. 68, 914-918.