International Journal of Mathematics and
Computer Science, **17**(2022), no. 2, 635–646

A Regression Estimator in Path Sampling

Mena Lao[1], **Sudarat Nidsunkid**[1], **John J. Borkowski**[2]

[1]Department of Statistics
Faculty of Science
Kasetsart University
Bangkok, Thailand

[2]Department of Mathematical Sciences
Montana State University
Montana, USA

email: mena_patummasut@yahoo.com, sudarat.n@ku.th,
john.borkowski@montana.edu

(Received October 1, 2021, Accepted November 2, 2021)

Abstract

It is known that regression estimation utilizes known auxiliary information to develop efficient estimators. The objective of this paper is to propose a regression estimator in path sampling to improve the precision of estimators of the population mean by using generalized regression estimation with unequal probability sampling when the relationship between the variable of interest and auxiliary variable is linear with a straight line that does not pass through the origin. The mean square error of this proposed regression estimator and the estimator of this mean square error are obtained. The efficiency of the proposed regression estimator is investigated by simulation study on the real world data, and the estimated mean square error of the proposed regression estimator is compared to two original existing estimators, which are an unbiased estimator and ratio estimator. The simulation results show that the proposed regression estimator has very small estimated bias, and it is more efficient than those two original estimators.

Key words and phrases: Regression estimator, Auxiliary variable, Path sampling, Unequal probability sampling.
AMS (MOS) Subject Classifications: 62D05, 62F40, 94A20.
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

1 Introduction

It has been known that a regression estimator is more efficient than a ratio estimator especially when the regression line is a straight line that does not pass through the origin [2]. A ratio estimator can be as efficient as a regression estimator only when such regression line passes through the origin. To improve the estimation by applying the auxiliary information, a regression estimator is widely used in many sampling designs, such as simple random sampling, unequal probability sampling [12], double sampling [11], and ranked set sampling [9].

Path sampling is an economical sampling design for a spatial population where the sampling cost is the number of units traveled to observe the sampled units [7]. Two estimators, the unbiased estimator and the ratio estimator, were proposed in path sampling [7], [8]. It is found that the ratio estimator is more efficient than the unbiased estimator when a variable of interest y and auxiliary variable x have a linear relationship with a straight line passing through the origin. However, in some situations, the linear relationship between x and y might be a straight line that does not go through the origin. For this situation, the ratio estimator is no longer appropriate; therefore, a regression estimator is more suitable and more efficient [6], [12]. Thus, the objective of this paper is to propose a regression estimator in path sampling to improve the precision of estimators of the population mean.

In this paper, detail of path sampling is described; regression estimation with unequal probability sampling is explained; and the new estimator, regression type estimator, is derived. Finally, the simulation study for examining the efficiency of the new estimator is presented by applying path sampling in the real population data, the species Beilschmiedia pendula trees data in the tropical rainforest of Barro Colorado Island.

2 Path sampling

Consider a population region partitioned into I rows and J columns resulting in IJ quadrats or units as shown in Figure 1. A variable of interest is y taking values $y_{(i,j)}$ for $i = 1, 2, 3, , I$ and $j = 1, 2, 3, , J$, and the population mean is $\mu_y = \sum_{i=1}^{I}\sum_{j=1}^{J} y_{(i,j)}/IJ$. All possible I-1 paths, denoted, $R_1, R_2, R_3, ..., R_{I-1}$, are created in the population region where each path is a collection of connected units travelled from a starting unit $(1, \tilde{j})$ to ending unit $(1, \tilde{j}+1)$ as shown in Figure 2.

Under path sampling, p paths are sampled by simple random sampling without replacement from I-1 paths in the population, and the obtained path sample can be written as $r_s = \{r_1, r_2, r_3, ..., r_p\}$. Let s denote the set of distinct units in the path sample r_s. In path sampling, two estimators of the population mean were proposed. The first estimator is an unbiased estimator given by

$$\hat{\mu}_{ps} = \frac{1}{IJ} \sum_{(i,j)\in s} \frac{y_{(i,j)}}{\pi_{(i,j)}} \tag{2.1}$$

where $\pi_{(i,j)}$ is the inclusion probability of unit (i, j) [7]. The variance of $\hat{\mu}_{ps}$ is

$$\begin{aligned} \nu(\hat{\mu}_{ps}) &= \frac{1}{(IJ)^2} \left[\sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{1-\pi_{(i,j)}}{\pi_{(i,j)}} \right) y_{(i,j)}^2 \right. \\ &\quad \left. + \sum_{i=1}^{I} \sum_{i'\neq i} \sum_{j=1}^{J} \sum_{j'\neq j} \left(\frac{\pi_{(i,j),(i',j')} - \pi_{(i,j)}\pi_{(i',j')}}{\pi_{(i,j)}\pi_{(i',j')}} \right) y_{(i,j)} y_{(i',j')} \right] \end{aligned} \tag{2.2}$$

where $\pi_{(i,j),(i',j')}$ is the joint inclusion probability of unit (i, j) and (i', j') . The estimator of this variance is

$$\begin{aligned} \hat{\nu}(\hat{\mu}_{ps}) &= \frac{1}{(IJ)^2} \left[\sum_{(i,j)\in s} \left(\frac{1}{\pi_{(i,j)}^2} - \frac{1}{\pi_{(i,j)}} \right) y_{(i,j)}^2 \right. \\ &\quad \left. + \sum_{\substack{(i,j),(i',j')\in s \\ (i,j)\neq(i',j')}} \left(\frac{1}{\pi_{(i,j)}\pi_{(i',j')}} - \frac{1}{\pi_{(i,j),(i',j')}} \right) y_{(i,j)} y_{(i',j')} \right] \end{aligned} \tag{2.3}$$

The second estimator of the population mean proposed by [8] is a ratio estimator given by

$$\hat{\mu}_{psr} = \hat{R}_{ps}\mu_x \tag{2.4}$$

where $x_{(i,j)}$ is a value of auxiliary variable of unit (i, j) related to the variable of interest, and $\hat{R}_{ps} = \sum_{(i,j)\in s} \frac{y_{(i,j)}}{\pi_{(i,j)}} / \sum_{(i,j)\in s} \frac{x_{(i,j)}}{\pi_{(i,j)}}$ is an estimator of the population ratio $R = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} y_{(i,j)}}{\sum_{i=1}^{I}\sum_{j=1}^{J} x_{(i,j)}}$. Assume the bias of $\hat{\mu}_{psr}$ is approximately equal to zero.

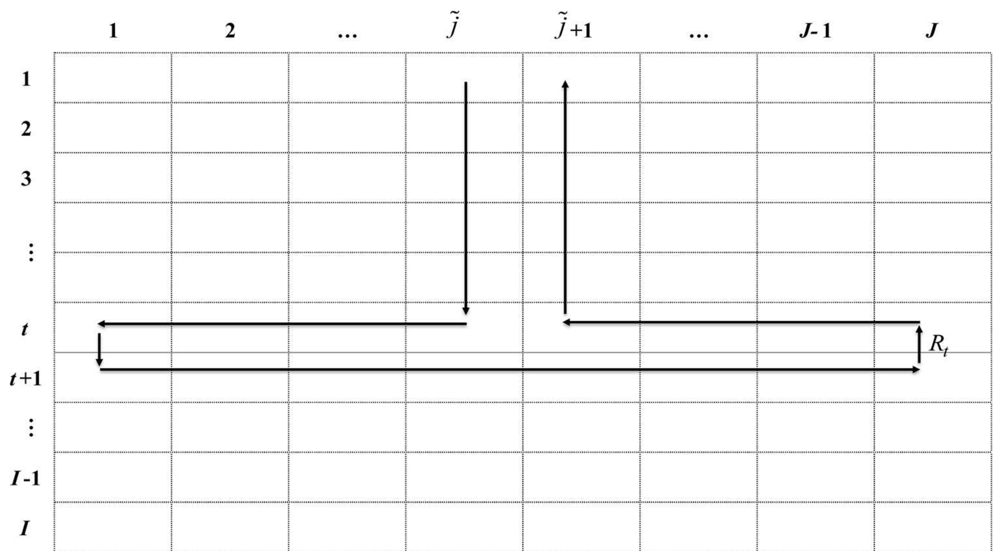

Figure 1: Path R_t with starting unit $(1, \tilde{j})$ in a population region partitioned into I rows and J columns.

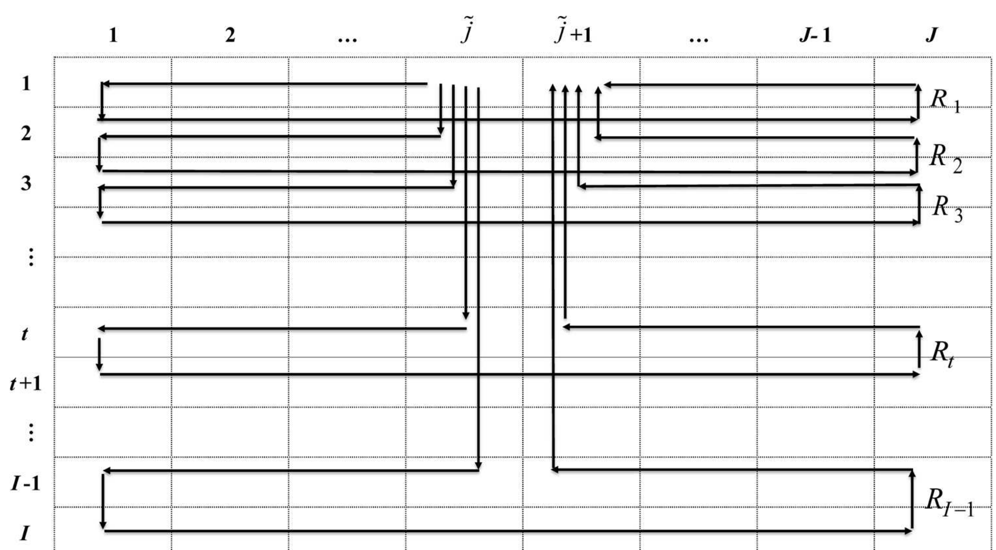

Figure 2: All possible I-1 paths with a starting unit $(1, \tilde{j})$ in a population region partitioned into I rows and J columns.

The approximation of the mean square error or the variance of $\hat{\mu}_{psr}$ is

$$\nu(\hat{\mu}_{psr}) = \frac{1}{(IJ)^2}\left[\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{1-\pi_{(i,j)}}{\pi_{(i,j)}}\right) z^2_{(i,j)} + \sum_{i=1}^{I}\sum_{i'\neq i}\sum_{j=1}^{J}\sum_{j'\neq j}\left(\frac{\pi_{(i,j),(i',j')}-\pi_{(i,j)}\pi_{(i',j')}}{\pi_{(i,j)}\pi_{(i',j')}}\right) z_{(i,j)}z_{(i',j')}\right] \quad (2.5)$$

where $z_{(i,j)} = y_{(i,j)} - Rx_{(i,j)}$. The estimator of this variance is

$$\hat{\nu}(\hat{\mu}_{psr}) = \frac{1}{(IJ)^2}\left[\sum_{(i,j)\in s}\left(\frac{1}{\pi^2_{(i,j)}} - \frac{1}{\pi_{(i,j)}}\right)\hat{z}^2_{(i,j)} + \sum_{\substack{(i,j),(i',j')\in s\\(i,j)\neq(i',j')}}\left(\frac{1}{\pi_{(i,j)}\pi_{(i',j')}} - \frac{1}{\pi_{(i,j),(i',j')}}\right)\hat{z}_{(i,j)}\hat{z}_{(i',j')}\right] \quad (2.6)$$

where $\hat{z}_{(i,j)} = y_{(i,j)} - \hat{R}_{ps}x_{(i,j)}$.

3 Regression Estimation with Unequal Probability Sampling

Let π_k, for $k = 1, , N$, be the inclusion probability for unit k for any sampling design, and let π_{kl} be the joint inclusion probability for unit k and l. For a sample of δ distinct units, a generalized regression estimator for the population mean is

$$\hat{\mu}_{rg} = \tilde{\mu}_y + \hat{B}(\mu_x - \tilde{\mu}_x) \quad (3.7)$$

where $\tilde{\mu}_y$ and $\tilde{\mu}_x$ are the generalized ratio estimator of the population mean of variables y and x, respectively [5]. $\hat{B}$ is a weighted regression slope estimator. The approximate expression of mean square error or variance of $\hat{\mu}_{rg}$ is

$$\nu(\hat{\mu}_{rg}) \approx \nu\left(\frac{1}{N}\sum_{k=1}^{\delta}\frac{(y_k - A - Bx_k)}{\pi_k}\right) \quad (3.8)$$

where A is a regression intercept and B is a regression slope.

Let $\hat{y}_i = y_i - \hat{A} - \hat{B}x_i$ where $\hat{A}$ is a weighted regression intercept estimator. The estimator of this variance is

$$\hat{\nu}(\hat{\mu}_{rg}) = \sum_{k=1}^{\delta}\left(\frac{1}{\pi_k^2} - \frac{1}{\pi_k}\right)\hat{y}_k^2 + \sum_{k=1}^{\delta}\sum_{k\neq l}\left(\frac{1}{\pi_k\pi_l} - \frac{1}{\pi_{kl}}\right)\hat{y}_k\hat{y}_l \tag{3.9}$$

for all $\pi_{kl} > 0$ [14].

4 A Proposed Regression Estimator in Path Sampling

This paper proposes a regression estimator in path sampling described as follows. Let $x_{(i,j)}$ be the value of the auxiliary variable of unit (i,j) associated with the value of the variable of interest. Suppose the population mean of x is known, denoted by μ_x. Each unit (i,j) has the inclusion probability $\pi_{(i,j)}$. Since the inclusion probability and joint inclusion probability can be obtained in path sampling, the regression estimator can be developed by applying a generalized regression estimator with unequal probability sampling. Based on the set of distinct units in the path sample, denoted s, the generalized ratio estimator of the population mean of the variable of interest y and the auxiliary variable x can be calculated from the formula

$$\hat{\mu}_y = \frac{\sum\limits_{(i,j)\in s}\frac{y_{(i,j)}}{\pi_{(i,j)}}}{\sum\limits_{(i,j)\in s}\frac{1}{\pi_{(i,j)}}} \text{ and } \hat{\mu}_x = \frac{\sum\limits_{(i,j)\in s}\frac{x_{(i,j)}}{\pi_{(i,j)}}}{\sum\limits_{(i,j)\in s}\frac{1}{\pi_{(i,j)}}} \tag{4.10}$$

respectively [12]. By applying a regression estimator for unequal probability sampling as shown in (3.7), a proposed regression estimator of the population mean of y in path sampling can be derived as

$$\hat{\mu}_{psg} = \hat{\mu}_y + \hat{m}(\mu_x - \hat{\mu}_x) = \hat{a} + \hat{m}\mu_x \tag{4.11}$$

where $\hat{m}$ is a weighted regression slope estimator given by

$$\hat{m} = \frac{\sum\limits_{(i,j)\in s}\frac{(x_{(i,j)}-\hat{\mu}_x)(y_{(i,j)}-\hat{\mu}_y)}{\pi_{(i,j)}}}{\sum\limits_{(i,j)\in s}\frac{(x_{(i,j)}-\hat{\mu}_x)^2}{\pi_{(i,j)}}} \tag{4.12}$$

and $\hat{a}$ is a weighted regression intercept estimator given by

$$\hat{a} = \frac{\sum_{(i,j)\in s} \frac{y_{(i,j)}}{\pi_{(i,j)}} - \hat{m}\frac{x_{(i,j)}}{\pi_{(i,j)}}}{\sum_{(i,j)\in s} \frac{1}{\pi_{(i,j)}}} = \hat{\mu}_y - \hat{m}\hat{\mu}_x \tag{4.13}$$

The mean square error of $\hat{\mu}_{psg}$ is

$$MSE(\hat{\mu}_{psg}) = E(\hat{\mu}_{psg} - \mu_y)^2 \tag{4.14}$$

[10].

By applying the approximate expression of mean square error or variance of the regression estimator for unequal probability sampling as shown in (3.8), the approximated formula of the mean square error or variance of $\hat{\mu}_{psg}$ can be derived as

$$\nu(\hat{\mu}_{psg}) \approx \nu\left(\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{y_{(i,j)} - a - mx_{(i,j)}}{\pi_{(i,j)}}\right)\right) \tag{4.15}$$

Now, let $\hat{y}_{(i,j)} = y_{(i,j)} - a - mx_{(i,j)}$, this variance can also be written in the form of Horvitz-Thompson [4] formula as

$$\begin{aligned} \nu(\hat{\mu}_{psg}) \approx & \sum_{i=1}^{I}\sum_{j=1}^{J}\left(\frac{1-\pi_{(i,j)}}{\pi_{(i,j)}}\right) y'^2_{(i,j)} + \\ & \sum_{i=1}^{I}\sum_{i'\neq i}\sum_{j=1}^{J}\sum_{j'\neq j}\left(\frac{\pi_{(i,j),(i',j')} - \pi_{(i,j)}\pi_{(i',j')}}{\pi_{(i,j)}\pi_{(i',j')}}\right) y'_{(i,j)}y'_{(i',j')} \end{aligned} \tag{4.16}$$

By applying the estimator of the variance of the regression estimator for unequal probability sampling as shown in (3.9), an estimator of $\nu(\hat{\mu}_{psg})$ can be derived using $\hat{y}_{(i,j)} = y_{(i,j)} - \hat{a} - \hat{m}x_{(i,j)}$ in the Horvitz-Thompson variance estimation formula. Thus, the resulting estimator of $\nu(\hat{\mu}_{psg})$ is

$$\begin{aligned} \hat{\nu}(\hat{\mu}_{psg}) = & \sum_{(i,j)\in s}\left(\frac{1}{\pi^2_{(i,j)}} - \frac{1}{\pi_{(i,j)}}\right)\hat{y}^2_{(i,j)} \\ & + \sum_{\substack{(i,j),(i',j')\in s \\ (i,j)\neq(i',j')}}\left(\frac{1}{\pi_{(i,j)}\pi_{(i',j')}} - \frac{1}{\pi_{(i,j),(i',j')}}\right)\hat{y}_{(i,j)}\hat{y}_{(i',j')} \end{aligned} \tag{4.17}$$

for all $\pi_{(i,j),(i',j')} > 0$.

16	12	16	10	8	17	21	16	55	35	13	24	8	12	8	9	14	14	19	7
12	8	15	4	5	10	7	14	28	18	14	23	11	17	14	13	6	33	9	34
9	2	5	10	9	13	0	7	23	18	18	23	25	16	8	10	7	13	22	24
8	2	4	13	29	7	9	0	3	18	7	14	10	5	5	9	15	22	7	25
11	4	3	1	4	1	2	2	0	0	0	4	4	5	10	14	11	13	5	13
8	13	1	2	5	2	2	3	0	0	0	0	0	0	20	46	40	26	123	16
5	12	4	10	6	9	2	0	0	0	0	0	0	53	50	13	6	10	82	30
2	2	7	4	3	2	0	0	0	0	0	0	0	1	0	1	6	6	10	4
8	4	2	2	1	3	0	0	1	0	1	0	0	7	10	6	10	9	7	0
12	14	11	0	0	2	1	0	0	0	2	1	7	6	12	21	15	7	7	6
10	31	17	4	1	3	0	0	2	0	5	8	9	5	4	6	8	6	2	6
57	24	16	22	5	9	1	0	1	0	3	2	3	1	1	6	8	1	6	2
40	37	28	27	17	4	1	0	2	2	6	1	1	3	1	1	2	6	0	1
25	19	36	14	14	0	1	2	0	1	2	0	1	1	1	0	0	0	0	0
10	15	34	24	62	15	4	3	3	0	0	3	3	1	1	0	1	0	1	1
4	7	9	8	10	5	6	1	2	2	9	0	3	4	5	8	2	9	4	5
7	6	12	10	20	15	8	11	51	21	12	12	5	25	3	7	12	20	10	31
0	8	8	6	3	4	2	3	10	13	11	25	7	16	12	9	11	23	9	18
0	0	0	0	0	2	0	12	4	5	2	6	5	15	14	21	14	22	1	5
0	0	0	0	2	1	0	0	0	0	0	0	0	0	3	7	18	15	5	1

Figure 3: Beilschmiedia pendula trees data in the tropical rainforest of Barro Colorado Island.

5 Simulation study

In order to examine the efficiency of the proposed regression estimator $\hat{\mu}_{psg}$, the simulation study on real-world data is performed to compare the estimated mean square error of the proposed regression estimator to the two original estimators, which are ratio estimator $\hat{\mu}_{psr}$ and unbiased estimator $\hat{\mu}_{ps}$. The studied population is a 1000×500 meter rectangular region in the tropical rainforest of Barro Colorado Island consisting of 3605 trees of the species Beilschmiedia pendula [3]. Under path sampling, this region is divided into 20 rows and 20 columns forming 400 units, 50×25 meter each, as shown in Figure 3.

Here, the starting and ending units are set to be unit (1, 10) and (1, 11), respectively. The variable of interest y is the number of Beilschmiedia pendula trees in a unit, and the population mean μ_y is 9.01. In this simulation study, two types of relationships between x and y are considered, resulting in the values of auxiliary variable x from simulation results of two models. The first model is $x_{(i,j)} = 8y_{(i,j)} + 4 - \epsilon_{(i,j)}$ where $\epsilon_{(i,j)} \sim N(0, y_{(i,j)})$ for a linear relationship with a straight line that does not pass through the origin, and the second model is $x_{(i,j)} = 8y_{(i,j)} - \epsilon_{(i,j)}$ for a linear relationship with a straight line passing through the origin.

In simulation, path samples of different number of paths, p, are selected from the population using R statistical software, and repeated 10,000 times. To compare the efficiency of the proposed regression estimator $\hat{\mu}_{psg}$ to the unbiased estimator $\hat{\mu}_{ps}$ and ratio estimator $\hat{\mu}_{psr}$, the estimated mean square error of each estimator is calculated by

$$mse(\hat{\mu}) = \frac{1}{10,000} \sum_{i=1}^{10,000} (\hat{\mu} - \mu_y)^2 \tag{5.18}$$

where $\hat{\mu}_i$ is the value of the relevant estimator for a sample i [1], [15].

Table 1 and Table 2 show simulation results of three estimators for each p with the estimated final sample size, denoted ω, which is the sum of all inclusion probabilities, that is, $\omega = \sum_{i=1}^{I}\sum_{j=1}^{J} \pi_{(i,j)}$ [13]. The relative efficiencies of the proposed regression estimator $\hat{\mu}_{psg}$ and ratio estimator $\hat{\mu}_{psr}$ are compared to $\hat{\nu}(\hat{\mu}_{ps})$, denoted,

$$re(\hat{\mu}_{psg}) = \frac{\hat{\nu}(\hat{\mu}_{ps})}{mse(\hat{\mu}_{psg})} \text{ and } re(\hat{\mu}_{psr}) = \frac{\hat{\nu}(\hat{\mu}_{ps})}{mse(\hat{\mu}_{psr})} \tag{5.19}$$

The estimated bias of the proposed regression estimator and ratio estimator are denoted by $\hat{B}(\hat{\mu}_{psg})$ and $\hat{B}(\hat{\mu}_{psr})$, respectively.

Table 1: The estimated bias and mean square error of the regression and ratio estimators and the relative efficiencies of the estimators for the Beilschmiedia pendula trees data with x-values from the model $x_{(i,j)} = 8y_{(i,j)} + 4 - \epsilon_{(i,j)}$.

p	ω	Estimated bias		Estimated MSE			Relative efficiency	
		$\hat{B}(\hat{\mu}_{psr})$	$\hat{B}(\hat{\mu}_{psg})$	$\hat{\nu}(\hat{\mu}_{ps})$	$mse(\hat{\mu}_{psr})$	$mse(\hat{\mu}_{psg})$	$re(\hat{\mu}_{psr})$	$re(\hat{\mu}_{psg})$
1	58.00	-0.07155	-0.01107	13.08982	0.05810	0.00867	225.29	1510.29
2	98.92	-0.03458	-0.00753	7.00936	0.02580	0.00158	271.71	4427.32
3	134.18	-0.02224	-0.00386	4.47563	0.01447	0.00083	309.37	5406.56
4	167.77	-0.01103	0.00339	3.64131	0.00890	0.00048	408.95	7590.47
5	195.59	-0.00947	-0.00321	2.48222	0.00571	0.00025	434.76	10061.82

From Table 1, for p =1, the estimated bias of the regression estimator $\hat{\mu}_{psg}$ and the ratio estimator $\hat{\mu}_{psr}$ are $\hat{B}(\hat{\mu}_{psg})$ = -0.01107 and $\hat{B}(\hat{\mu}_{psr})$ = -0.07155 respectively. This shows that the estimated bias of $\hat{\mu}_{psg}$ is smaller than that of $\hat{\mu}_{psr}$. The estimated mean square error of $\hat{\mu}_{ps}$, $\hat{\mu}_{psr}$ and $\hat{\mu}_{psg}$

are 13.08982, 0.05810 and 0.00867, respectively. Notice that the regression estimator $\hat{\mu}_{psg}$ gave the smallest estimated mean square error, thus $\hat{\mu}_{psg}$ is better than $\hat{\mu}_{ps}$ and $\hat{\mu}_{psr}$. The relative efficiency of $\hat{\mu}_{psr}$ and $\hat{\mu}_{psg}$ are $re(\hat{\mu}_{psr}) =$ 225.29 and $re(\hat{\mu}_{psg})$ =1510.29, respectively. This indicates that the regression estimator $\hat{\mu}_{psg}$ is more efficient than $\hat{\mu}_{ps}$ and $\hat{\mu}_{psr}$ because $\hat{\mu}_{psg}$ gave the largest relative efficiency. Similarly, for p = 2, 3, 4 and 5, the regression estimator $\hat{\mu}_{psg}$ is more efficient than all other estimators because it gave the largest relative efficiency for every p, and its estimated bias is smaller than that of the ratio estimator $\hat{\mu}_{psr}$. Therefore, the regression estimator $\hat{\mu}_{psg}$ is the most efficient for a linear relationship with a straight line that does not pass through the origin. Furthermore, it can be noticed that when p is increasing, the estimated bias and mean square error are smaller, and the relative efficiencies of estimators are larger.

Table 2: The estimated bias and mean square error of the regression and ratio estimators and the relative efficiencies of the estimators for the Beilschmiedia pendula trees data with x-values from the model $x_{(i,j)} = 8y_{(i,j)} - \epsilon_{(i,j)}$.

p	ω	Estimated bias		Estimated MSE			Relative efficiency	
		$\hat{B}(\hat{\mu}_{psr})$	$\hat{B}(\hat{\mu}_{psg})$	$\hat{\nu}(\hat{\mu}_{ps})$	$mse(\hat{\mu}_{psr})$	$mse(\hat{\mu}_{psg})$	$re(\hat{\mu}_{psr})$	$re(\hat{\mu}_{psg})$
1	58.00	0.01377	0.02332	12.38756	0.00430	0.00428	2882.93	2895.85
2	98.92	0.00629	-0.01327	6.36796	0.00209	0.00181	3049.06	3517.01
3	134.18	-0.00258	-0.00351	4.50898	0.00085	0.00083	5305.28	5428.44
4	167.77	-0.00176	0.00080	2.99906	0.00056	0.00054	5355.46	5553.81
5	195.59	0.00106	-0.00028	2.65283	0.00046	0.00045	5749.86	5871.41

From Table 2, for p =1, the estimated bias of the regression estimator $\hat{\mu}_{psg}$ and the ratio estimator $\hat{\mu}_{psr}$ are $\hat{B}(\hat{\mu}_{psg}) = 0.02332$ and $\hat{B}(\hat{\mu}_{psr}) =$ 0.01377. This illustrates that the estimated bias of the regression estimator $\hat{\mu}_{psg}$ and the ratio estimator$\hat{\mu}_{psr}$ are slightly different. The estimated mean square error of $\hat{\mu}_{ps}$, $\hat{\mu}_{psr}$ and $\hat{\mu}_{psg}$ are 12.38756, 0.00430 and 0.00428, respectively. This shows that the estimated mean square error of $\hat{\mu}_{psg}$ is slightly less than that of the ratio estimator $\hat{\mu}_{psr}$, therefore $\hat{\mu}_{psg}$ is slightly better than $\hat{\mu}_{psr}$. The estimated mean square error of both $\hat{\mu}_{psr}$ and $\hat{\mu}_{psg}$ are considerably less than that of $\hat{\mu}_{ps}$, indicating that both $\hat{\mu}_{psr}$ and $\hat{\mu}_{psg}$ are much better than $\hat{\mu}_{ps}$. Besides, the relative efficiency of $\hat{\mu}_{psr}$ and $\hat{\mu}_{psg}$ are $re(\hat{\mu}_{psr}) = 2882.93$ and $re(\hat{\mu}_{psg})$ =2895.85, respectively. Notice that $re(\hat{\mu}_{psg})$ is slightly more than

$re(\hat{\mu}_{psr})$, thus $\hat{\mu}_{psg}$ is slightly more efficient than $\hat{\mu}_{psr}$. Additionally, both $\hat{\mu}_{psg}$ and $\hat{\mu}_{psr}$ are much more efficient than $\hat{\mu}_{ps}$ because $re(\hat{\mu}_{psr})$ and $re(\hat{\mu}_{psg})$ are extremely greater than 1. Likewise, for $p = 2$, 3, 4 and 5, both $\hat{\mu}_{psg}$ and $\hat{\mu}_{psr}$ are more efficient than the unbiased estimator $\hat{\mu}_{ps}$ because they have larger relative efficiency for every p. The regression estimator $\hat{\mu}_{psg}$ is slightly more efficient than the ratio estimator $\hat{\mu}_{psr}$ since $\hat{\mu}_{psg}$ has slightly larger relative efficiency for every p. Notice that the estimated mean square errors of both estimators are quite similar, so it can be said that the regression estimator $\hat{\mu}_{psg}$ is as efficient as the ratio estimator $\hat{\mu}_{psr}$ when x and y have a linear relationship with a straight line passing through the origin. Moreover, it can be seen that when p is increasing, the relative efficiencies of estimators are larger, and the estimated bias and mean square error are smaller.

6 Conclusions

In path sampling, the proposed regression estimator, $\hat{\mu}_{psg}$, is very useful and efficient when a variable of interest is linearly related to the auxiliary variable with a straight line not passing through the origin, and in such situation, the ratio estimator is no longer appropriate. From the simulation study on the Beilschmiedia pendula trees data, the regression estimator is highly advantageous than all original estimators in path sampling when x and y have a linear relationship with a straight line that does not pass through the origin since it yields very larger values of relative efficiencies. However, the regression estimator is slightly more efficient than the ratio estimator when a linear relationship of x and y is a straight line passing through the origin.

Acknowledgment. This work was supported by a research grant from International SciKU Branding (ISB), Faculty of Science, Kasetsart University

References

[1] N. Chutiman, M. Chiangpradit, Ratio estimator in adaptive cluster sampling without replacement of networks, J. Probab. Stat., **2014**, (2014), 1-6.

[2] E. P. Clement, E. I. Enang, On the efficiency of ratio estimator over the regression estimator, Commun. Stat. Theory Methods, **46**, (2017), 5357-5367.

[3] R. Condit, S. P. Hubbell, R. B. Foster, Changes in tree species abundance in a neotropical forest: impact of climate change, J. Trop. Ecol., **12**, (1996), 231-256.

[4] D. G. Horvitz, M. E. Thompson, A generalization of sampling without replacement from a finite universe, J. Amer. Statist. Assoc., **47**, (1952), 663-685.

[5] C. T. Isaki, W. A. Fuller, Survey design under the regression super population model, J. Amer. Statist. Assoc., **77**, (1982), 89-96.

[6] S. L. Lohr, Sampling: Design and Analysis, 2nd ed., Brooks/Cole, Boston, 2010.

[7] M. Patummasut, A. L. Dryver, A new sampling design for a spatial population: Path sampling, J. Appl. Sci., **12**, (2012), 1355-1363.

[8] M. Patummasut, S. Nidsunkid, A ratio estimator in path sampling. Commun. Stat. Simul. Comput., **50**, (2021), 376-387.

[9] L. Philip, H. Yu, K. Lam, Regression estimator in ranked set sampling. Biometrics, **53**, (1997), 1070-1080.

[10] R. L. Scheaffer, W. III. Mendenhall, R. L. Ott, Elementary Survey Sampling, 6th ed., Duxbury Press, California, 2006.

[11] A. C. Tamhane, Inference based on regression estimator in double sampling, Biometrika, **65**, (1978), 419-427.

[12] S. K. Thompson, Sampling, 3rd ed., Wiley, New York, 2012.

[13] S. K. Thompson, G. A. F. Seber, Adaptive Sampling, Wiley, New York, 1996.

[14] R. L. Wright, Finite population sampling with multivariate auxiliary, J. Amer. Statist. Assoc., **78**, (1983), 879884.

[15] F. Younis, J. Shabbir, Estimators for population mean in adaptive cluster sampling, Thail. Stat., **15**, (2017), 105-110.