Thailand Statistician
January 2013; 11(1): 31-43
http://statassoc.or.th
Contributed paper

The Effect of Sampling Methods for Linear Regression Estimation of Population Means of Dependent Variables

Natcha Buasiri and Tidadeaw Mayureesawan*

Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand.

*Author for correspondence; E-mail**:** tidadeaw@yahoo.com

Received: 19 October 2012

Accepted: 31 January 2013

Abstract

In this paper, we compare an efficient population mean estimator of dependent variable on simple linear regression model in three sampling methods, namely simple random sampling (SRS), ranked set sampling (RSS) and L ranked set sampling (LRSS) when the population mean of the independent variable is known. The independent variable and the dependent variable jointly followed a bivariate t-distribution with the different degree of freedom (df) levels, correlation coefficient (ρ) between ones with the different levels, set size and number of sample units allocated to each set (m) number of cycles (r) and LRSS coefficient (k) with the different levels. The efficiency of population mean estimator of all sampling methods are considered by the average of variance of population mean estimator (AVPME) values. The results indicate that for very low level of df (df=1), for all levels of ρ that less than 1, all levels of m and for low and moderate levels of r (r=5, 10), the RSS method is the most efficient but for high levels of r (r=20), the SRS method is the most efficient. For the other levels of df, for all levels of ρ that less than 1 and for all levels of r, the RSS method is the most efficient when m is in low levels (m=8 and 10), the LRSS method is the most efficient when m is in moderate levels (m=15, 20 and 30) at k=1, for large level of m (m=40) at k=2 and for very large level of m (m=50) at k=3. For very high level of ρ (ρ=1) in all levels of df, m and r, the performance of the three sampling methods are nearly the same. The efficiency of population mean estimator of dependent variable from each sampling methods increase and nearly the similar when m, r and ρ increases.

Keywords: Simple random sampling, ranked set sampling, L ranked set sampling, population mean estimator.

1. Introduction

Simple linear regression analysis is a conceptually simple method for investigating relationship between two variables that has a straight line relationship. The simple true relationship can be approximated by the regression model $Y = \alpha + \beta X + \varepsilon$ where ε is assumed to be random error, α and β are unknown regression parameters to be estimated from the data. Sampling is the process of obtaining a sample that is a good representative of the population for reliably reference. Sampling is divided into two categories, namely nonprobability sampling and probability sampling. Nonprobability sampling is the process of sampling that regardless the sample unit is much less chance of being selected so this sampling cannot refer to the population. In the another category of sampling, probability sampling is the process of sampling that can set the chance of being selected in each sample unit so this sampling can refer to the population. Generally Simple Random Sampling (SRS) is most commonly used that is a kind of probability sampling. The advantage of SRS is easy to use but the disadvantage is samples may not be distributed that does not the best representative population.

Ranked set sampling (RSS) is a kind of probability sampling that is two steps process and provide a more precise estimator of the population mean of the variable of interest (dependent variable) which is either difficult to measure so that the ranking is done on the basis of an independent variable that associate with dependent variable. The first step, select m random samples, each of size m units from the population and ranks the units within each sample with respect to an independent variable by a visual inspection or by any other method. The second step, each sample is ranked. Only one unit is measured. RSS is divided into two categories, namely perfect ranked set sampling and imperfect ranked set sampling. Perfect ranked set sampling is ranked set sampling when the value of correlation coefficient between independent and dependent variable is -1 or 1 that is independent and dependent variables are completely correlated and this one has no error in ranking. For another category, imperfect ranked set sampling is ranked set sampling when the value of correlation coefficient between independent and dependent variable is in range -1 to 1 that is independent and dependent variable are correlated and this one has error in ranking.

In 1952 McIntyre [1] first suggested a sampling method, namely RSS for estimating pasture yields that was more efficient than SRS, while assuming perfect ranking of its units. In 1977 Stokes [2] considered the case when the dependent variable is difficult to measure and order, but there is an independent variable that is correlated with dependent variable and can be used to judge the order of the dependent variable. In 1995 Muttlak [3] used RSS to estimate the parameters of the simple linear regression model treating the independent variable as a constant. In 1997 Yu and Lam [4] proposed a regression estimator based on RSS when independent and dependent variable jointly follow a bivariate normal distribution. They demonstrated that this estimator is always more efficient than SRS and RSS estimator without using the regression model unless the correlation coefficient is low ($|\rho|<0.4$). Moreover, it is always superior to the regression estimators under SRS for all ρ. In 2000 Demir and Çıngı [5] applied the estimator proposed by Yu and Lam with original Turkish data for estimating the population mean. They demonstrated that this estimator is superior to the SRS regression estimator, even for asymmetric distributions. In 2007 Al-Nasser [6] proposed L ranked set sampling (LRSS). LRSS is a random sampling process that has two steps like RSS but in the second step there is a k that is a value of the setting position in the sample to be measured or LRSS coefficient which k = [mp] such that $0 \le p < 0.5$ and k is the largest integer value less than or equal to mp. For each of first (k+1) samples, select the unit with rank k+1 and measure the dependent variable (Y) value that corresponding to $X_{(k+1)j}$ and denote it by $Y_{[k+1]j}$. For j=k+2, …, m-k-1, the unit with rank j in the j^{th} ranked sample is selected and measures the y value that corresponds then select the $(m-k)^{th}$ unit and measure the correspond y value for each of the last $(m-k)^{th}$ samples. In addition, Al-Nasser also compared the efficiency of the estimators for the population mean on SRS, RSS and LRSS methods by studying the symmetric and asymmetric distributions. It is shown that the LRSS method gives unbiased estimator for the population mean with minimum variance under symmetric distribution. In 2008 Al-Nasser and Radaideh [7] compared the efficiency of the SRS, RSS, ERSS and LRSS regression estimators when the distribution of independent variable is standard normal. Extreme ranked set sampling (ERSS) is a random sampling process that has two steps like RSS but in the second step there is an m that is the set size and number of sample units allocated to each set is a value of the setting position in the sample to be measured. It is shown that the regression estimator based on LRSS is superior to the SRS, ERSS and RSS.

From the researches are mentioned above, the researchers compared the efficiency of the population mean estimator of simple linear regression model when the distribution of independent or dependent variable is symmetric as normal and asymmetric as exponential but they have not studied in case of independent or dependent variable is symmetric in heavy tail class. The main objective of this paper is to study the efficiency of the population mean estimator of simple linear regression model when the population mean of an independent variable is in between three sampling methods include SRS, RSS and LRSS when independent and dependent variable are t-distribution. This will know the efficiency of the population mean estimator and the effect of the value of df, ρ, m, r and k to the efficiency of the population mean estimator of simple linear regression model.

2. Material and Method

The operation for study the efficiency of population mean estimator on simple linear regression model $Y = \alpha + \beta X + \varepsilon$ in three sampling methods, namely SRS, RSS and LRSS when the population mean of an independent variable is known using R program has the following steps.

Step 1 Generate the big population data of independent and dependent variable which have jointly followed a bivariate t-distribution at the degree of freedom levels df = 1, 5, 13, 15, 20 and 25 and correlation coefficient ρ = 0.5, 0.6, 0.7, 0.8, 0.9 and 1.

Step 2 For RSS and LRSS, the set size and number of sample units allocated to each set m = 8, 10, 15, 20, 30, 40, and 50 the number of cycles r = 5, 10, and 20 and LRSS coefficient k = 1, 2 and 3 only for to be available in all m parameters. For SRS, the sample size n = mr.

Step 3 Estimate the simple linear regression coefficient ($\hat{\beta}$) in each situations from three sampling methods.

Demir and Çıngı [5] showed (1) and (2)

SRS $$\hat{\beta} = \frac{\sum_{i=1}^{n} y_i x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}} \qquad (1)$$

RSS $$\hat{\beta} = \frac{\sum_{j=1}^{r}\sum_{i=1}^{m}\left(X_{(i)j}-\bar{x}_{rss}\right)\left(Y_{[i]j}-\bar{y}_{rss}\right)}{\sum_{j=1}^{r}\sum_{i=1}^{m}\left(X_{(i)j}-\bar{x}_{rss}\right)^2} \quad (2)$$

and Al-Nasser and Radaideh [7] showed (3)

LRSS $$\hat{\beta} = \frac{\sum_{j=1}^{r}\sum_{i=1}^{m}\left(d_{(i)j}^{x}-d^{\bar{x}}\right)\left(d_{[i]j}^{y}-d^{\bar{y}}\right)}{\sum_{j=1}^{r}\sum_{i=1}^{m}\left(d_{(i)j}^{x}-d^{\bar{x}}\right)^2}. \quad (3)$$

Step 4 Estimate the population mean of dependent variable on simple linear regression model.

Yu and Lam [4] defined (4) and (5)

SRS $$\bar{y}_{srs.reg} = \bar{y} + \hat{\beta}(\mu_X - \bar{x}) \quad (4)$$

RSS $$\bar{y}_{rss.reg} = \bar{y}_{rss} + \hat{\beta}(\mu_X - \bar{x}_{rss}) \quad (5)$$

and following Yu and Lam [4] to obtain (6)

LRSS $$\bar{y}_{lrss.reg} = d^{\bar{y}} + \hat{\beta}(\mu_X - d^{\bar{x}}) \quad (6)$$

where $d^{\bar{y}} = \frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} d_{[i]j}^{y}$ and $d^{\bar{x}} = \frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} d_{(i)j}^{x}$

$$d_{[i]j}^{y} = \begin{cases} Y_{[k+1]j} & i \le k \\ Y_{[i]j} & k+1 \le i \le m-k \quad ; j = 1,2, \dots, r \\ Y_{[m-k]j} & m-k+1 \le i \le m \end{cases}$$

$$d_{(i)j}^{x} = \begin{cases} X_{(k+1)j} & i \le k \\ X_{(i)j} & k+1 \le i \le m-k \quad ; j = 1,2, \dots, r \\ X_{(m-k)j} & m-k+1 \le i \le m. \end{cases}$$

Step 5 Calculate the variance of population mean estimator of dependent variable on simple linear regression model

Yu and Lam [4] showed (7) and (8)

SRS $$V\left(\bar{y}_{srs.reg}\right) = \frac{\sigma_Y^2}{n}(1-\rho^2)\left[1+\left(\frac{1}{n-3}\right)\right] \quad (7)$$

RSS $$V(\bar{y}_{rss.reg}) = \frac{\sigma_Y^2}{mr}(1-\rho^2)\left[1+E\left(\frac{\bar{z}_{rss}^2}{S_Z^2}\right)\right] \quad (8)$$

where $\bar{Z}_{rss} = \frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} Z_{(i)j}$, $Z_{(i)j} = \frac{X_{(i)j}-\mu_X}{\sigma_X}$

$$\text{and} \quad S_Z^2 = \frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m}\left(Z_{(i)j}-\bar{Z}_{rss}\right)^2$$

and following Yu and Lam [4] to obtain (9)

LRSS $$V(\bar{y}_{lrss.reg}) = \frac{\sigma_Y^2}{mr}(1-\rho^2)\left[1+E\left(\frac{\bar{z}_{lrss}^2}{S_Z^2}\right)\right] \quad (9)$$

where $\bar{Z}_{lrss} = \frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m} Z_{(i)j}$, $Z_{(i)j} = \frac{d_{(i)j}^{X}-\mu_X}{\sigma_X}$

$$\text{and} \quad S_Z^2 = \frac{1}{mr}\sum_{j=1}^{r}\sum_{i=1}^{m}\left(Z_{(i)j}-\bar{Z}_{lrss}\right)^2.$$

Step 6 In each situation, compute the AVPME of dependent variable on simple linear regression model with values obtained through extensive simulations.

Step 7 Compare the efficiency of population mean estimator of dependent variable on simple linear regression model in three sampling methods by comparison of the AVPME values.

3. Results

The AVPME of dependent variable we show for some level of ρ. Table 1 summarizes results for df=1, ρ=0.5, Table 2 for df=15, ρ=0.5 and Table 3 for df=25, ρ=0.5. From Table 1 indicates that for all m levels the RSS method gives minimum AVPME of dependent variable when r is small and medium (r=5 and 10) but for the large level of r (r=20) the SRS method gives minimum AVPME value and gives the same result as ρ=0.6, 0.7, 0.8 and 0.9.

Table 2 shows that for df=15, ρ=0.5 for all r levels the RSS method gives minimum AVPME value when m is small (m=8 and 10), for the medium level of m (m=15, 20 and 30) the LRSS method gives minimum AVPME value when k=1, for the large level of m (m=40) the LRSS method gives minimum AVPME value when k=2 and for the large level of m (m=50) the LRSS method gives minimum AVPME value when k=3 and gives the same result as ρ=0.6, 0.7, 0.8 and 0.9.

In Table 3 shows the AVPME values for df=25, ρ=0.5, we found that the result of minimum AVPME values are similar at the level of df=15, ρ=0.5 in Table 2 and for all levels of df which not shown here but the AVPME values decrease and gives the same result as ρ=0.6, 0.7, 0.8 and 0.9.

In the case of ρ is large (ρ=1), the AVPME of dependent variable equal to 0 for all sets m, r and df.

The effects of m, ρ, r and df values are shown in Figures 1 and 2. Figure 1(A) depicts to the effect of the value of m at df=1, ρ=0.5. It is shown that for all sampling methods give the same results that for all r levels the AVPME values decrease when the value of m is increased and the difference of the AVPME values between sampling methods are relatively small when the value of m is increased and gives the same result for all levels of df when ρ=0.6, 0.7, 0.8 and 0.9. Figure 1(B) depicts to the effect of the value of ρ at df=25, m=50. It is shown that for all sampling methods give the same results that for all r levels when the value of ρ is increased, the AVPME values decrease and the AVPME values between sampling methods are nearly the similar and gives the same result for all levels of df and m. Figure 2(A) depicts to the effect of the value of r at df=25, ρ=0.5. We show that for all sampling methods give the same results that for all m levels the AVPME values decrease when the value of r is increased and the difference of the AVPME values between sampling methods are relatively small when the value of r is increased and gives the same result for all levels of df when ρ=0.6, 0.7, 0.8 and 0.9. Figure 2(B) depicts to the effect of the value of df at r=5, ρ=0.5. It is shown that for all sampling methods give the same results that for all m levels the AVPME values are relatively large when df changes from 1 to 5 but the AVPME values are slightly different as df increases from 5 to 25 and gives the same result for all levels of r when ρ=0.6, 0.7, 0.8 and 0.9.

Table 1. The AVPME values of SRS, RSS and LRSS method for df=1 and ρ=0.5 .

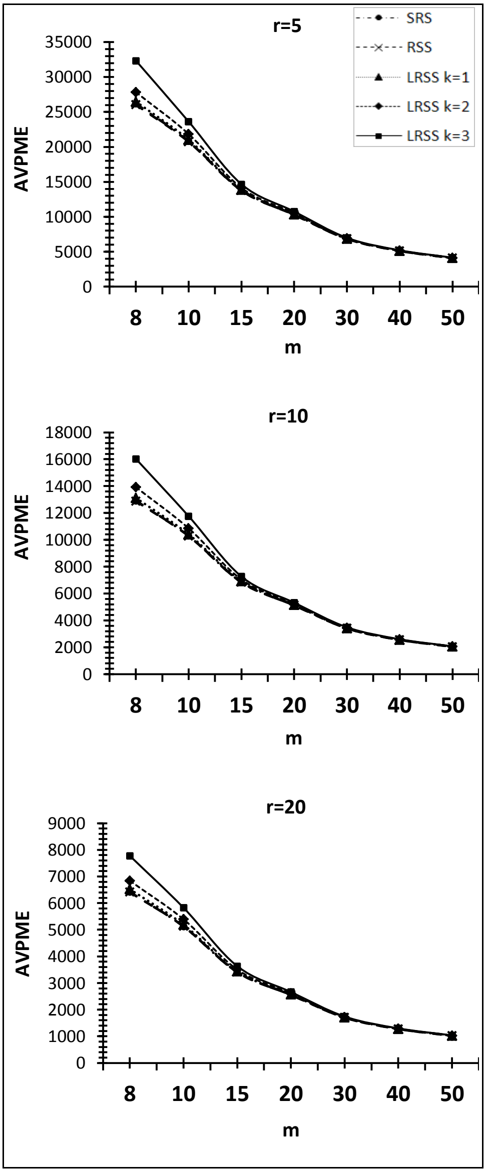
r	m	SRS	RSS	LRSS		
				k=1	k=2	k=3
5	8	26,176.070	**26,043.870**	26,478.030	27,830.680	32,345.570
	10	20,823.600	**20,736.710**	21,019.430	21,834.220	23,613.980
	15	13,781.980	**13,755.650**	13,867.080	14,128.800	14,660.820
	20	10,299.990	**10,285.240**	10,346.000	10,515.230	10,749.920
	30	6,842.828	**6,835.645**	6,856.582	6,911.934	6,989.160
	40	5,123.320	**5,119.925**	5,134.803	5,171.578	5,205.761
	50	4,094.466	**4,091.496**	4,099.450	4,115.884	4,144.168
10	8	12,909.110	**12,887.420**	13,121.180	13,921.950	16,029.320
	10	10,299.990	**10,282.570**	10,407.000	10,855.130	11,758.130
	15	6,842.828	**6,836.378**	6,911.847	7,041.200	7,266.741
	20	5,123.320	**5,122.955**	5,147.428	5,213.792	5,322.008
	30	3,409.739	**3,408.815**	3,418.215	3,447.311	3,489.043
	40	2,555.142	**2,553.457**	2,563.393	2,576.171	2,599.550
	50	2,043.080	**2,042.603**	2,046.181	2,057.678	2,069.876
20	8	**6,412.391**	6,412.563	6,521.255	6,837.054	7,773.288
	10	**5,123.320**	5,125.444	5,189.319	5,391.144	5,825.042
	15	**3,409.739**	3,409.923	3,434.592	3,504.158	3,632.010
	20	**2,555.142**	2,555.860	2,567.766	2,609.983	2,662.130
	30	**1,701.994**	1,702.527	1,707.667	1,722.817	1,741.841
	40	**1,275.960**	1,276.008	1,278.928	1,286.480	1,297.801
	50	**1,020.512**	1,020.529	1,022.251	1,027.733	1,034.715

Table 2. The AVPME values of SRS, RSS and LRSS method for df=15 and ρ=0.5 .

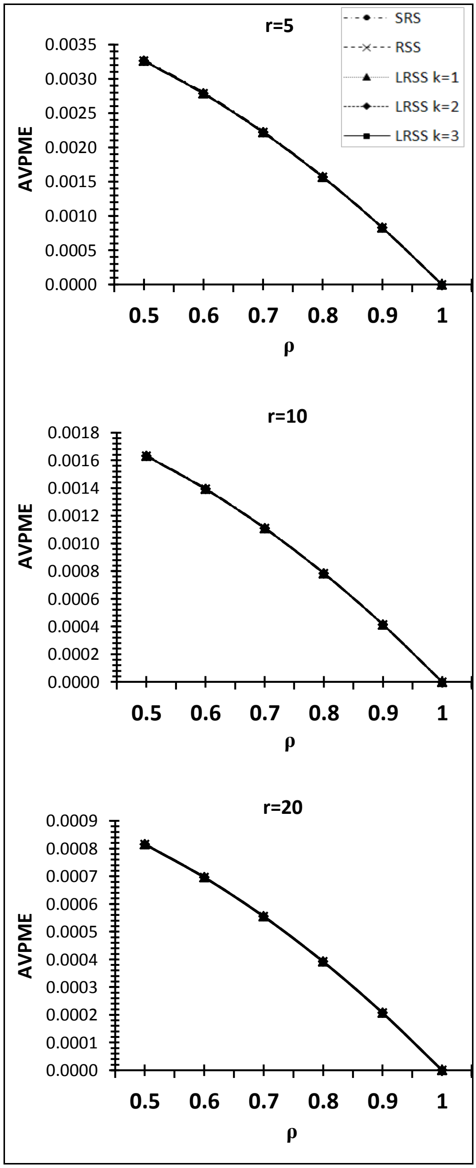
r	m	SRS	RSS	LRSS		
				k=1	k=2	k=3
5	8	0.022231840	**0.021765480**	0.021814490	0.021948000	0.022197110
	10	0.017685890	**0.017409190**	0.017414670	0.017435470	0.017462670
	15	0.011705300	0.011567610	**0.011566290**	0.011570370	0.011590990
	20	0.008747980	0.008672619	**0.008671218**	0.008671712	0.008675899
	30	0.005811745	0.005776524	**0.005775960**	0.005776532	0.005776917
	40	0.004351334	0.004330952	0.004330489	**0.004330318**	0.004330644
	50	0.003477508	0.003464334	0.003464143	0.003464189	**0.003464139**
10	8	0.010963960	**0.010859750**	0.010875840	0.010906040	0.010967920
	10	0.008747980	**0.008675114**	0.008681682	0.008691150	0.008709967
	15	0.005811745	0.005778810	**0.005778476**	0.005781758	0.005783455
	20	0.004351334	0.004332627	**0.004332413**	0.004332468	0.004333166
	30	0.002895956	0.002887123	**0.002887117**	0.002887147	0.002887183
	40	0.002170131	0.002165147	0.002165058	**0.002165023**	0.002165162
	50	0.001735227	0.001731927	0.001731927	0.001731903	**0.001731894**
20	8	0.005446166	**0.005418907**	0.005421820	0.005429585	0.005438953
	10	0.004351334	**0.004334233**	0.004334639	0.004337551	0.004339350
	15	0.002895956	0.002887629	**0.002887516**	0.002888476	0.002888898
	20	0.002170131	0.002165467	**0.002165355**	0.002165611	0.002165616
	30	0.001445536	0.001443343	**0.001443286**	0.001443324	0.001443424
	40	0.001083697	0.001082443	0.001082433	**0.001082431**	0.001082438
	50	0.000866740	0.000865923	0.000865922	0.000865923	**0.000865918**

Table 3. The AVPME values of SRS, RSS and LRSS method for df=25 and ρ=0.5 .

r	m	SRS	RSS	LRSS		
				k=1	k=2	k=3
5	8	0.020937360	**0.020482330**	0.020587020	0.020705520	0.020952910
	10	0.016656100	**0.016372360**	0.016385200	0.016444360	0.016486700
	15	0.011023740	0.010896570	**0.010895870**	0.010896900	0.010900720
	20	0.008238617	0.008163836	**0.008162657**	0.008166600	0.008168680
	30	0.005473348	0.005439359	**0.005439083**	0.005439774	0.005440492
	40	0.004097971	0.004078790	0.004078568	**0.004078536**	0.004078855
	50	0.003275025	0.003262722	0.003262664	0.003262489	**0.003262476**
10	8	0.010325570	**0.010223770**	0.010242120	0.010263350	0.010310170
	10	0.008238617	**0.008172605**	0.008175034	0.008184956	0.008201508
	15	0.005473348	0.005442475	**0.005442392**	0.005443654	0.005447752
	20	0.004097971	0.004079567	**0.004079421**	0.004080016	0.004080180
	30	0.002727335	0.002718911	**0.002718886**	0.002719120	0.002719167
	40	0.002043772	0.002039053	0.002039009	**0.002038995**	0.002039064
	50	0.001634191	0.001631115	0.001631069	0.001631090	**0.001631064**
20	8	0.005129056	**0.005102287**	0.005106700	0.005118887	0.005122110
	10	0.004097971	**0.004080385**	0.004082860	0.004084325	0.004092029
	15	0.002727335	0.002720079	**0.002719726**	0.002720205	0.002720499
	20	0.002043772	0.002039367	**0.002039297**	0.002039360	0.002039532
	30	0.001361368	0.001359319	**0.001359302**	0.001359313	0.001359338
	40	0.001020598	0.001019400	0.001019397	**0.001019392**	0.001019414
	50	0.000816273	0.000815497	0.000815495	0.000815495	**0.000815489**

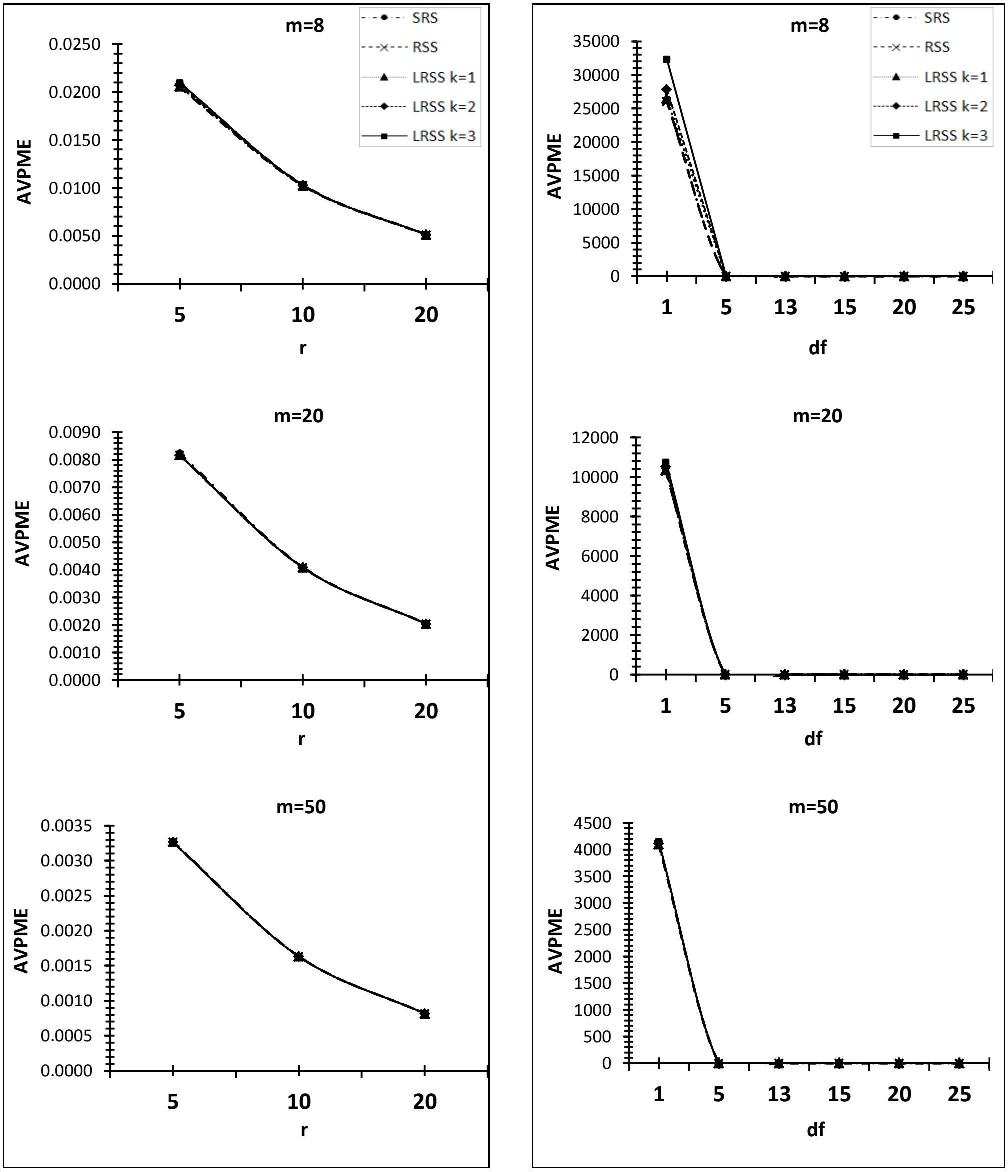


(A)



(B)

Figure 1. The AVPME values of SRS, RSS and LRSS method (A) for df=1, ρ=0.5 and (B) for df=25, m=50.



(A) (B)

Figure 2. The AVPME values of SRS, RSS and LRSS method (A) for df=25, ρ=0.5 and (B) for r=5, ρ=0.5.

4. Discussion and Conclusions

In this paper, we compare an efficient population mean estimator of dependent variable on simple linear regression model in three sampling methods, namely SRS, RSS and LRSS when the population mean of the independent variable is known by independent variable and the dependent variable jointly followed a bivariate t-distribution. We vary the different degrees of freedom (df) levels, correlation coefficient (ρ), set size and number of sample units allocated to each set (m) number of cycles (r) and LRSS coefficient (k) levels. The efficiency of population mean estimator of dependent variable is considered by the AVPME values. The results of the comparisons are summarized in Table 4 which gives a list of preferred sampling methods for a range of values of df, ρ, m and r. The results indicate that for df=1, ρ=0.5, 0.6, 0.7, 0.8 and 0.9 and for all levels of m, the RSS method is the most efficient when r=5 and 10 and SRS method is the most efficient when r=20. For df=5, 13, 15, 20 and 25, ρ=0.5, 0.6, 0.7, 0.8 and 0.9 and for all levels of r, the RSS method is the most efficient when m=8 and 10, the LRSS method is the most efficient at k=1 when m=15, 20 and 30, the LRSS method is the most efficient at k=2 when m=40 and the LRSS method is the most efficient at k=3 when m=50. For ρ=1 in all levels of df, m and r, the performance of the three sampling methods are nearly the same. Also we found that the parameters df, m and r are effect to the efficiency. When m, r and ρ are increased the efficiency of population mean estimator from each sampling methods is increased and nearly the similar. When df changes from 1 to 5, the efficiency of all sampling methods are relatively large but the efficiency of all samplings methods are slightly different when df changes from 5 to 25.

Table 4. Summary of preferred sampling method.

Parameters				Preferred sampling methods
df	**ρ**	**m**	**r**	
1	0.5, 0.6, 0.7, 0.8 and 0.9	All levels	5, 10	RSS
			20	SRS
5, 13, 15, 20 and 25	0.5, 0.6, 0.7, 0.8 and 0.9	8, 10	All levels	RSS
		15, 20, 30		LRSS when k=1
		40		LRSS when k=2
		50		LRSS when k=3
All levels	1	All levels	All levels	SRS, RSS and LRSS for all k levels

5. Acknowledgements

The authors would like to thank the Graduate College, King Mongkut's University of Technology North Bangkok for the financial support of this study.

References

[1] McIntyre, G.A., A method of unbiased selective sampling using ranked sets, Australian Journal of Agricultural Research, 1952; 59: 385-390.

[2] Stokes, S.L., Ranked set sampling with concomitant variables, Communication in Statistics: Theory and Methods, 1977; 6: 1207-1211.

[3] Muttlak, H.A., Parameters Estimation in a Simple Linear Regression Using Rank Set Sampling, Biometrical Journal, 1995; 37: 799-810.

[4] Yu, P.L.H. and Lam, K., Regression estimator in ranked set sampling, Biometrics, 1997; 53: 1070-1080.

[5] Demir, S. and Çıngı, H., An Application of the Regression Estimator in Ranked Set Sampling, Hacettepe Bulletin of Natural Sciences and Engineering Series B, 2000; 29: 93-101.

[6] Al-Nasser, A.D., L Ranked set sampling, A generalization procedure for robust visual sampling, Communications in Statistics: Simulation and Computation, 2007; 36: 33 – 43.

[7] Al-Nasser, A.D. and Radaideh, A., Estimation of Simple Linear Regression Model Using L Ranked Set Sampling, International Journal of Open Problems in Computer Science and Mathematics, 2008; 1: 18-33.