A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 255, 2011

Janusz L. Wywiał[*]

SIMULATION ANALYSIS OF ACCURACY ESTIMATION OF POPULATION MEAN ON THE BASIS OF REGRESSION TYPE STRATEGY DEPENDENT ON ORDER STATISTIC OF AUXILIARY VARIABLE

Abstract. The paper deals with an analysis of the accuracy of the strategies for estimating the mean as well as the total value of a variable under study in a fixed and finite population. They involve a positive valued auxiliary variable. A strategies called quantile-regression is compared with a simple sample mean and with other regression types strategies. Moreover, Horvitz-Thompson statistic is considered. A sampling design proportional to the difference of two quantiles is taken into account. Moreover, the well-known Singh-Srivastava's sampling design is considered, too. The comparison of the strategies' accuracy has been based on a computer simulation.

Key words: sampling design, order statistic, auxiliary variable, sampling scheme, estimation, strategy, accuracy comparison, relative efficiency, regression estimator, Horvitz-Thompson estimator

I. SAMPLING DESIGNS AND SCHEMES

Let $U=(1,2,...,N)$ be a fixed population of the size N. An observation of a variable under study (an auxiliary variable) attached to the i-th population element will be denoted by y_i $(x_i>0)$, $i=1,...,N$. The sample of size n, drawn without replacement from the population, will be denoted by s. The sampling design is denoted by $P(s)$ and inclusion probabilities of the first and second orders - by π_k, for k=1,....,N and $\pi_{k,t}$ for k≠t, k=1,..,N, t=1,..,N, respectively. Let $\boldsymbol{S}$ be the sample space of the samples of size n, drawn without replacement. We are going to consider the following sampling designs of simple samples drawn without replacement: $P_0(s)=\binom{N}{n}^{-1}$ for all $s\in\boldsymbol{S}$.

Singh and Srivastava (1980) proposed the following sampling design.

[*] Professor, Department of Statistics, Katowice University of Economics.

$$P_1(s) = \frac{v_{*s}(x)}{v_*(x)} \binom{N}{n}^{-1} \quad \text{for all } s \in \boldsymbol{S} \tag{1}$$

where $v_{*s}(x) = \frac{1}{n-1}\sum_{i\in s}(x_i - \bar{x})^2$, $v_*(x) = \frac{1}{N-1}\sum_{k\in U}(x_k - \bar{x})^2$. The sampling scheme implementing $P_1(s)$ is as follows. The first two element denoted by *(k,h)* of the sample are selected with the probability proportional to $(x_k - x_h)^2$, $k=1,...,N$, $k=1,...,N$, $k\neq h$. The next $(n$-2) elements are selected in the same way as the simple sample of size *(n-2)*, drawn without replacement.

Let $X_{(r)}$ be the r-th order statistic from a simple sample drawn without replacement. Let $\alpha\in(0;1)$ and $[n\alpha]$ is the integer part of the value $n\alpha$. The sample quantile of the order α is defined as $Q_{s,\alpha}$ = sampling $X_{(r)}$ where $r = [n\alpha]+1$ and $(r-1)/n \leq \alpha < r/n$. The sampling design depends on two order statistics $X_{(r)}$ and $X_{(z)}$. Let $U_1=(1,...,i-1)$ be a subpopulation of the population U and let s_1 be the simple sample of size (r-1), drawn without replacement from U_1. Similarly, let $U_2=(i+1,...,j-1)$ be a subpopulation of the population U and let s_2 be a simple sample of size (z-r-1), drawn without replacement from U_2. Finally, let s_3 be a simple sample of size *(n-z)*, drawn without replacement from $U_3=(j+1,...,N)$. Hence, $s = (s_1 \cup \{i\} \cup s_2 \cup \{j\} \cup s_3)$ be such a sample that the value of the r-th order statistic of an auxiliary variable - observed in the sample - equals x_i and the z-th order statistic of an auxiliary variable - observed in the sample - equals x_j. Wywiał (2009) proposed the sampling design:

$$P_{r,z}(s) = \frac{X_{(z)} - X_{(r)}}{h(r,z)} \quad \text{for } s \in \boldsymbol{S} \tag{2}$$

where

$$h(r,z) = \sum_{j=r}^{N-n+r} \sum_{j=i+z-r}^{N-n+z} g(r,z,i,j)(x_j - x_i),$$

$$g(r,z,i,j) = \binom{i-1}{r-1}\binom{j-i-1}{z-r-1}\binom{N-j}{n-z}.$$

Let us note that Wywiał (2006) straightforward generalized this sampling design into the following one.

$$P_{r,z}(s)=\frac{f\left(X_{(z)},X_{(r)}\right)}{h(r,z)}\text{ for } s\in\boldsymbol{S}$$

where f(.,.) is non-negative function of values of the order statistics and

$$h(r,z)=\sum_{j=r}^{N-n+r}\sum_{j=i+z-r}^{N-n+z}g(r,z,i,j)f\left(x_j,x_i\right).$$

In our case $f\left(X_{(z)},X_{(r)}\right)=X_{(z)}-X_{(r)}$.

Sampling scheme is as follows. Firstly, the *i*-th population element is selected according to the value of the following probability function of the statistic $X_{(r)}$:

$$p_{r,z}(i)=P(X_{(r)}=x_i)=\frac{1}{h(r,z)}\sum_{j=i+z-r}^{N-n+z}f(x_j,x_i)g(r,z,i,j)\text{ , } I=r,\ldots,N\text{-}n+r, \quad (3)$$

Next, the the *j*-th population element is selected according to the value of the following conditional probability function of the statistic $X_{(z)}$:

$$p_{r,z}(j\,|\,i)=P(X_{(z)}=x_j\,|\,X_{(r)}=x_i)=\frac{P(X_{(z)}=x_j,X_{(r)}=x_i)}{P(X_{(r)}=x_i)}$$

$$\text{for } j=i+1,\ldots,N\text{-}n+r \quad (8)$$

where

$$P(X_{(z)}=x_j,X_{(r)}=x_i)=\frac{g(r,z,i,j)}{h(r,z)}. \quad (9)$$

Finally, three simple sample (denoted by s_1, s_2 and s_3) are drawn without replacement. The sample s_1 of the size r-1 is selected from the sub-population $U_1=\{1,\ldots,i\text{-}1\}$ the sample s_2 of the size z-r-1 is selected from the sub-population $U_2=\{i+1,\ldots,j\text{-}1\}$ and the sample s_3 of the size n-z is selected from the sub-population $U_3=\{j+1,\ldots,N\}$. Finally, the sample is: $s=s_1\cup\{i\}\cup s_2\{j\}\cup s_3$.

II. ESTIMATORS AND STRATEGIES

The well known Horvitz-Thompson (1952) estimator is as follows.

$$t_{HTS} = \frac{1}{N}\sum_{k=1}^{N}\frac{a_k y_k}{\pi_k} \tag{10}$$

where $a_k = 1$ if the k-th population element was drawn to a sample. When $a_k = 0$, the k-th element was not drawn to the sample. It is well known that the strategy $(t_{HTS}, P(s))$ is unbiased for the population mean when all inclusion probabilities are positive. Moreover, the strategy $(t_{HTS}, P_0(s)) = (\bar{y}_S, P_0(s))$ is called a simple sample mean. In the next paragraph we are going to consider the strategy $(t_{HTS}, P_{r,z}(s))$.

The ordinary regression estimator is as follows

$$t_{eS} = \bar{y}_S + b_S(\bar{x} - \bar{x}_S) \tag{11}$$

where

$$b_S = \frac{c_{*S}(x, y)}{v_{*S}(x)}.$$

The strategy $(t_{eS}, P_1(s))$ is unbiased for the population mean. As it is well known, the strategy $(t_{eS}, P_0(s))$ is almost unbiased for population mean when the sample size is sufficiently large.

Wywiał (2004, 2009) considered the following estimators:

$$t_{r,z,S} = \bar{y}_{HTS} + b_{r,z,S}(\bar{x} - \bar{x}_{HTS}) \tag{12}$$

where

$$b_{r,z,S} = \frac{Y_u - Y_r}{X_{(u)} - X_{(r)}}.$$

The considered sampling strategies are: $(\bar{y}_S; P_0(s))$, $(t_{eS}; P_0(s))$, $(t_{eS}; P_1(s))$, $(t_{HTS}; P_{r,z}(s))$ and $(t_{r,zS}; P_{r,z}(s))$.

III. ACCURACY COMPARISON OF ESTIMATION STRATEGIES

The population in the demonstration consists of the municipalities in Sweden. The auxiliary variable x is: *1975 municipal population (in thousands)* and the variable under study y is: *1985 municipal taxation revenues (in millions of kronor).* Their observations have been published by Särndal, Swenson and Wretman (1992). The size of this population is 284 municipalities. There are three outlier observations of the variables, see Figure 1. Let $\bar{x}$, d, v_3 and $\beta_3 = v_3 / d^3$ be the mean, the standard deviation and the skewnees coefficient, respectively, of *municipal population*. In the case of data without outliers (size of the population N=281) $\bar{x}$ =24,263, d=24,153 and $\beta_3 = 0{,}043$. In the case of data with outliers (size of the population N=284) $\bar{x}$ =28.810, d=52,873 and $\beta_3 = 8{,}427$.

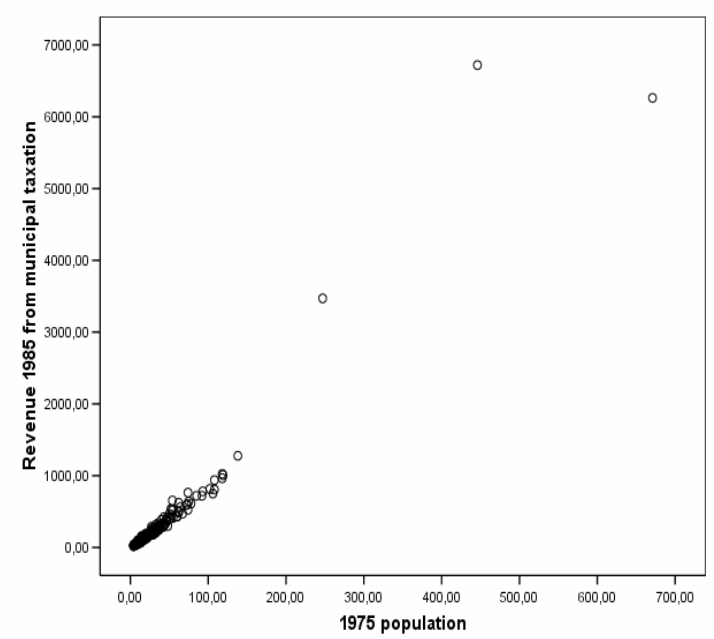

Figure 1. Scatterplot of the x and y variables; the population with three outliers.

The accuracy of the (t_S,$P(s)$) estimation strategy was measured by means of the relative efficiency - *deff*:

$$deff(t_S, P(s)) = \frac{MSE(t_S, P(s))}{D^2(\bar{y}_S, P_0(s))} 100\%$$

Let $deff1 = deff(t_{eS}; P_0(s))$, $deff2 = deff(t_{eS}; P_1(s))$, $deff3 = deff(t_{HTS}; P_1(s))$ $deff4 = deff(t_{r,zS}; P_{r,z}(s))$, $deff5 = deff(t_{HTS}; P_{r,z}(s))$.

Table 1. The relative efficiency coefficients for sampling design $P_{r,n-r+1}(s)$ for $r=1,\ldots,n/2$ and $n=12$

Data with outliers, N=284.						
r, z	1, 12	2, 11	3, 10	4, 9	5, 8	6, 7
m=z-r	11	9	7	5	3	1
deff4o	2,93	6,08	8,87	8,31	33,06	123,08
deff5o	16,76	42,75	55,80	55,73	63,71	90,41
Data without outliers, *N*=281.						
deff4	3,52	3,50	4,20	5,85	11,59	62,26
deff5	52,45	52,20	58,90	64,43	68,11	68,20

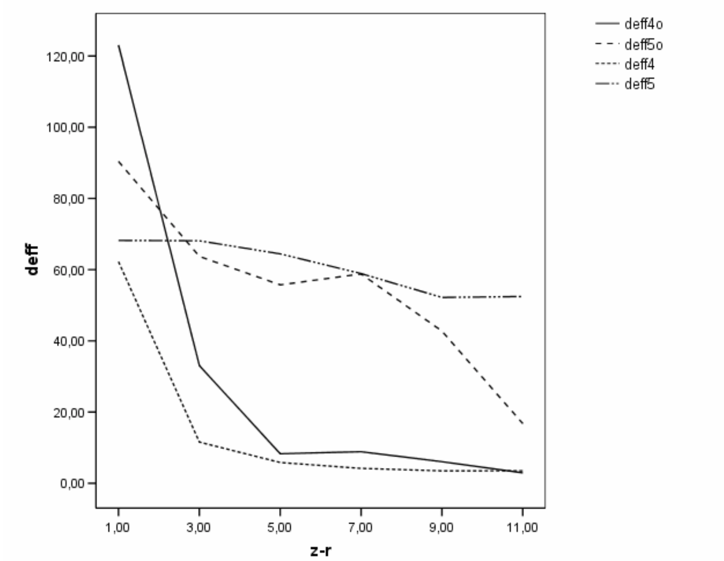

Figure 2. Presentation of the Table 1. Deff dependent on the difference of the ranks *r* and *z*

On the basis of the Table 1 and the Figure 2 we can infer that relative efficiencies of the strategies $(t_{r,n-r+1,S};P_{r,n-r+1}(s))$ and $(t_{HTS};P_{r,n-r+1}(s))$ decrease when the difference of the range *m=z-r* increases. In the case of existing outliers in the population the relative efficiencies of the strategies are better than in the case when the population without outliers. The strategy $(t_{r,n-r+1,S};P_{r,n-r+1,}(s))$ is more accurate than the strategy $(t_{HTS};P_{r,n-r+1,}(s))$. Hence, we can recommend the strategy $(t_{1,nS};P_{1,n}(s))$.

Table 2. Relative efficiency (%) of the strategies for N=284.
Data with autliers

	deff1	*Deff2*	*deff3*	*deff4*	*deff5*
2	4,96	5,79	10,03	1,38	2,67
3	1,65	5,29	9,28	1,44	5,03
4	1,76	4,04	8,56	1,67	6,73
5	1,45	3,69	9,40	1,71	8,94
6	1,67	3,68	9,63	1,80	10,07
8	1,89	3,71	11,25	2,23	12,95
10	2,05	3,47	11,73	2,67	15,65
12	2,30	3,61	13,25	2,93	16,76
15	2,43	3,76	14,00	3,41	20,71
20	2,89	3,64	17,15	4,14	23,44
25	3,74	4,43	20,59	4,38	26,26
30	4,08	4,33	22,98	4,85	27,68
40	5,22	5,04	25,21	5,08	31,25

Table 2 and Figure 3 deal with the case when the outliers exists in the population and lead to the following conclusions. The strategy $(t_{HTS};P_1(s))$ is the worst among the considered ones. In the case when $n \leq 3$ the strategy $(t_{1,n,S};P_{1,n}(s))$ is the best among the considered ones. In the case when n≥3 the accuracy of the strategies $(t_{eS};P_0(s))$, $(t_{1,n,S};P_{1,n}(s))$ and $(t_{eS};P_1(s))$ are significantly more efficient than the strategy $(t_{HTS};P_{1,n}(s))$. In general the strategies $(t_{eS};P_0(s))$ and $(t_{1,n,S};P_{1,n}(s))$ are similarly efficient and the best for rather small sample size *(n<10)*. For larger sample size *(n>9)* the strategies $(t_{eS};P_0(s))$, $(t_{1,n,S};P_{1,n}(s))$ *and* $(t_{eS};P_1(s))$ have similar efficiency and the are the best.

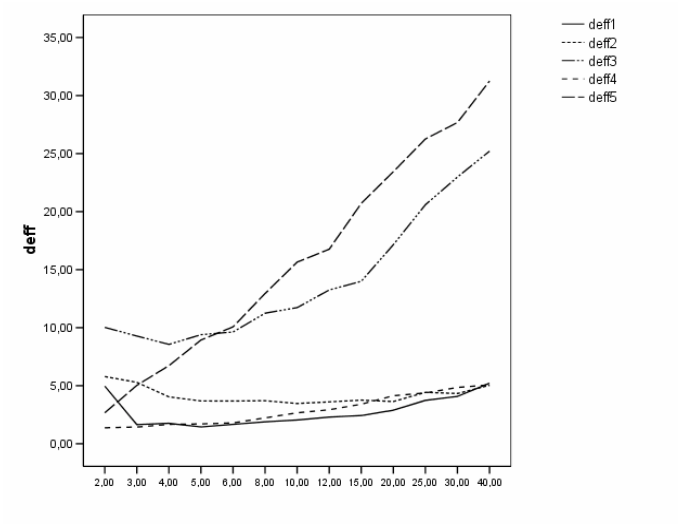

Figure 3. Deff for the data with outliers.

In the case when there are not outliers in the population on the basis of Table 3 and Figure 4 we infer that evidently the ordinary regression estimator from simple sample is the worst among the considered ones for *n=2*. For larger sample size (*n*>2) the Horvitz-Thompson's strategies $(t_{HTS};P_1(s))$ and $(t_{HTS};P_{1,n}(s))$ are significantly less efficient that the $(t_{eS};P_0(s))$, $(t_{eS};P_1(s))$ and $(t_{1,n,S};P_{r,z}(s))$ which accuracy is similar. The proposed strategy $(t_{1,n,S};P_{r,z}(s))$ is better than $(t_{eS};P_0(s))$ and it is better than $(t_{eS};P_1(s))$ for *n≤15*.

Table 3. Relative efficiency (%) of *the strategies for N=281.*
Data without autliers.

	deff1	*Deff2*	*deff3*	*deff4*	*deff5*
2	80,51	6,89	30,63	3,31	10,71
3	6,56	**2,82**	22,70	3,10	18,22
4	4,66	**2,55**	24,80	2,89	25,10
5	4,67	**2,52**	26,76	2,95	32,34
6	4,16	**2,71**	28,89	3,04	38,51
8	3,60	**2,72**	33,25	2,92	43,13
10	3,65	**2,53**	34,06	3,36	53,42
12	3,63	**2,83**	40,68	3,52	52,45
15	3,63	**2,63**	46,73	3,52	64,20
20	3,28	**2,96**	53,39	3,72	72,32
25	3,65	**2,64**	57,18	3,13	73,70
30	3,47	**2,74**	60,08	3,45	81,74
40	3,31	**2,75**	64,43	3,51	87,76

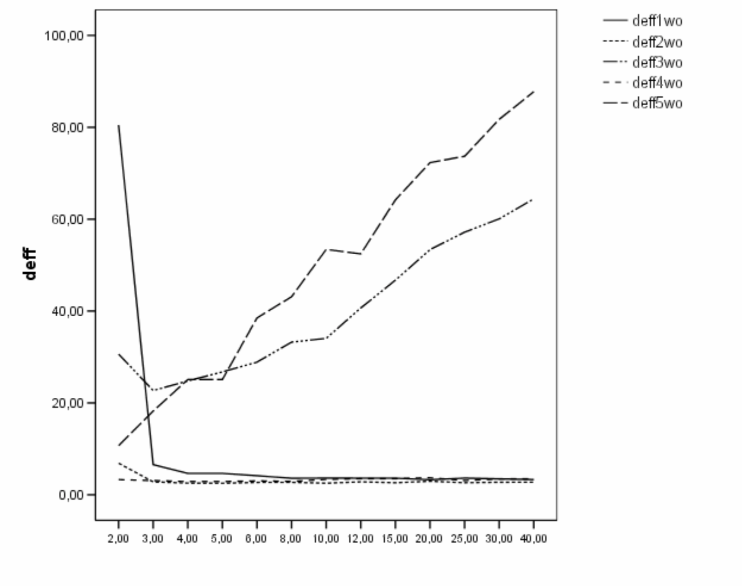

Figure 4. Deff for the data without outliers.

Generally, the regression type strategies are more accurate than the Horvitz-Thompson' ones. The accuracy of the regression estimators are comparable. But in the case of the population without outliers the Singh-Srivastava's regression strategy is slightly better than the others two regression ones. In the case of the population with outliers the quantil-regression strategy is better for sample sizes not greater than 4 but for the larger sample sizes the ordinary regression strategy is the best.

The received results are valid of course when we estimate the total value of a variable under study, In this case estimators are obtained through multiplying the considered ones by the size of a population,

REFERENCES

Horvitz D. G., Thompson D. J. (1952), A generalization of sampling without replacement from finite universe. *Journal of the American Statistical Association*, vol. 47, s. 663–685.

Särndal C. E., B. Swensson, J. Wretman: (1992): *Model Assisted Survey Sampling.* Springer Verlag, New York-Berlin-Heidelberg- London-Paris-Tokyo-Hong Kong- Barcelona-Budapest.

Singh P., Srivastava A.K. (1980), Sampling schemes providing unbiased regression estimators. *Biometrika*, vol. 67, 1, pp. 205–9.

Wywiał J. L. (2004), Quantile regression sampling strategy. In: *Metoda Reprezentacyjna w Badaniach Ekonomiczno-Społecznych* Editor: J. Wywiała, Prace Naukowe Akademii Ekonomicznej w Katowicach, str. 32–42.

Wywiał J. L. (2006), Plany losowania prób proporcjonalne do funkcji obserwacji zmiennej dodatkowej. Raport realizacji indywidualnego grantu nr 1 H02B 018 27 Ministerstwa Nauki i Szkolnictwa Wyższego, Akademia Ekonomiczna w Katowicach.
Wywiał J. L. (2009), Performing quantiles in regression sampling strategy. *Model Assisted Statistics and Applications* vol. 4, no. 2, pp. 131–142.

Acknowledgement
The research was supported by the grant number N N111 434137 from the Ministry of Science and Higher Education.

Janusz L. Wywiał

SYMULACYJNA ANALIZA DOKŁADNOŚCI ESTYMACJI ŚREDNIEJ W POPULACJI ZA POMOCĄ STRATEGII TYPU KWANTYLOWEGO

Problem dotyczy oceny wartości średnie (globalnej) zmiennej w populacji ustalonej I skończonej. Zakład się, że z góry są znane w populacji wartości dodatniej zmiennej pomocniczej. Do estymacji użyto strategia kwantylowej zależnej m.in. od planu losowania proporcjonalnego do nieujemnej funkcji kwantyla z próby zmiennej pomocniczej. Ponadto, brano pod uwagę estymator Horvitza-Thompsona oraz estymator ilorazowy. Porównanie dokładności przeprowadzono na podstawie symulacji komputerowej.