ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

Janusz Wywiał[*]

PERFORMING QUANTILES IN MULTIPLE REGRESSION SAMPLING STRATEGY

ABSTRACT. Estimation of the population average in a finite population by means of sampling strategy dependent on the sample quantile of an auxiliary variables is considered. The sampling design is proportionate to the determinant of the matrix dependent on some quantiles of an auxiliary variables. The sampling scheme implementing the sampling design is proposed. The derived inclusion probabilities are applied to estimation the population mean using the well known Horvitz-Thompson estimator. Moreover, the regression estimator is defined as the function of the coefficient dependent on the quantiles of the auxiliary variables. The properties of this estimator under the above defined sampling design are studied. The considerations are supported by empirical examples.

Key words: sampling design, order statistic, sample quantile, auxiliary variable, Horvitz-Thompson statistic, inclussion probabilities, sampling scheme, regression estimator.

I. INTRODUCTION

We are going to consider a finite and fixed population of the size N. A variable under study will be denoted by *y*, an auxiliary variable by x. Let $((y_1,x_1), (y_2,x_2),\dots, (y_i,x_i),\dots, (y_n,x_n))$ be the observations of the variable (y, x). The sample means of the variables *y* and *x* will be denoted by $\bar{y}$ and $\bar{x}$, respectively. Let $s_R=\{i:x_i\le\bar{x}\}$, $s_L=\{i:x_i>\bar{x}\}$, $n_R=Card\{s_R\}$, $n_L=n-s_R$ and

$$\bar{x}_R=\frac{1}{n_R}\sum_{i\in s_R}x_i\ ,\ \bar{x}_L=\frac{1}{n_L}\sum_{i\in s_L}x_i,\ \bar{y}_R=\frac{1}{n_R}\sum_{i\in s_R}y_i,\ \bar{y}_L=\frac{1}{n_L}\sum_{i\in s_L}y_i.$$

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

So, the $\overline{x}_R$ and $\overline{x}_L$, statistics are the sample means of the variable x from the right truncated and left truncated sample, respectively, in the point $\overline{x}$.

The estimator of the slope coefficient of the linear regression of the variable y on x considered by Wald (1940), Kendall and Stuart (1961), pp. 399-400 and Hellwig (1956, 1963), pp.138-155 is as follows

$$b_s = \frac{\overline{y}_R - \overline{y}_L}{\overline{x}_R - \overline{x}_L}.$$

In this paper we are going to generalize those results into the multidimensional case when the linear regression function depends on at least two explanatory variables.

II. BASIC DEFINITIONS AND NOTATIONS

We are going to consider the finite population U of size $N>k$.The observation of a variable under study and an auxiliary variables are denoted by y_i and $x_{i,j}$, respectively i=1,…,N, j=1,…,k. Let $[y\ x]$ where $x=[x_1\ x_2\ ...\ x_n]$ be the vector which values are coordinates of a point in a *(k+1)* -dimensional Euclidean space. Elements of the vector $\mathbf{y}^T=[y_1\ y_2\ \dots\ y_N]$ be values of a variable under study observed in the population. The vector $\mathbf{y}_k$ of dimensions k×1 consists of k-elements of the vector $\mathbf{y}$, $k \le N$. The observed in the population values of k-auxiliary variables are the elements of the matrix $\mathbf{x}=[x_{ij}]$ of dimensions $N \times k$ and $i=1,2,...,N$, $j=1,2,...,k$. Moreover, $\mathbf{x}=[\mathbf{x}_{\bullet 1}\ \mathbf{x}_{\bullet 2}\ \dots\ \mathbf{x}_{\bullet k}]$ where $(\mathbf{x}_{\bullet j})^T=[x_{1j}\ x_{2j}\ \dots\ x_{kj}]$ and $\mathbf{x}=\begin{bmatrix}\mathbf{x}_{1\bullet}\\ \dots\\ \mathbf{x}_{N\bullet}\end{bmatrix}$ where $\mathbf{x}_{i\bullet}=[x_{i1}\ x_{i2}\ \dots\ x_{ik}]$. Let $\mathbf{x}_k$ ($k<N$) be the submatrix of the matrix $\mathbf{x}$. The matrix $\mathbf{x}_k$ is obtained after removing (N-k) rows from the matrix $\mathbf{x}$. The column vector consisted of *k-th* values each equal to one will be denoted by $\mathbf{J}_k$. The column vector consisted of *k-th* values each equal to zero will be denoted by $\mathbf{0}_k$.

The well known equation of a k dimensional hiper-plain spanned on *(k+1)* points $\mathbf{z}_i=[y_i\ \mathbf{x}_{i\bullet}]$, (i=0,1,2,...,k) in the *(k+1)*-dimensional Euclidean space is as follows (see. e.g. Borsuk (1969)):

$$\begin{vmatrix} 1 & y & x \\ 1 & \mathrm{y}_0 & \mathbf{x}_0 \\ \mathbf{J}_k & \mathbf{y}_\mathrm{k} & \mathbf{x}_k \end{vmatrix} = 0,$$

If we subtract the second row of the above matrix from the remainder rows we get:

$$\begin{vmatrix} 0 & y-\mathrm{y}_0 & x-\mathbf{x}_0 \\ 1 & \mathrm{y}_0 & \mathbf{x}_0 \\ \mathbf{0}_k & \mathrm{d}_\mathrm{y}(\mathrm{y}_0) & \mathbf{d}(\mathbf{x}_k,\mathbf{x}_0) \end{vmatrix} = 0 \qquad (1)$$

or

$$\begin{vmatrix} y-\mathrm{y}_0 & x-\mathbf{x}_0 \\ \mathbf{d}_\mathrm{y}(\mathrm{y}_0) & \mathbf{d}(\mathbf{x}_k,\mathbf{x}_0) \end{vmatrix} = 0\,, \qquad (2)$$

where

$$\mathbf{c}(\mathrm{y}_0)=\mathbf{y}_\mathrm{k} - \mathrm{y}_0\mathbf{J}_\mathrm{k}, \qquad (3)$$

$$\mathbf{d}(\mathbf{x},\mathrm{x}_0)=\mathbf{x}_\mathrm{k}- \mathbf{J}_\mathrm{k}\mathbf{x}_0. \qquad (4)$$

Let $\mathbf{d}^{(.,j)}=\mathbf{d}^{(.,j)}(\mathbf{x}_\mathrm{k},\mathrm{x}_0)$ be the matrix obtained through removing a *j-th* column of the matrix $\mathbf{d}(\mathbf{x}_\mathrm{k},\mathrm{x}_0)$, j=1,…,k. Similarly, let $\mathbf{d}^{(i,.)}=\mathbf{d}^{(i,.)}(\mathbf{x}_\mathrm{k},\mathrm{x}_0)$ be the matrix obtained through removing an *i-th* row of the matrix $\mathbf{d}(\mathbf{x}_\mathrm{k},\mathrm{x}_0)$, i=1,…,k. Finally, let $\mathbf{d}^{(i,j)}=\mathbf{d}^{(i,j)}(\mathbf{x}_\mathrm{k},\mathrm{x}_0)$ be the matrix obtained through removing an i-th row and a *j-th* column of the matrix $\mathbf{d}(\mathbf{x}_\mathrm{k},\mathrm{x}_0)$, i,j=1,…,k. Moreover, let $\mathbf{D}(\mathbf{x}_\mathrm{k},\mathrm{x}_0)=[\mathbf{d}^{(i,j)}(\mathbf{x}_\mathrm{k},\mathrm{x}_0)]=[\mathbf{d}^{(i,j)}]$. This notation let us rewrite the determinant (1) or (2) in the following ways.

$$(y-\mathrm{y}_0)\left|\mathbf{d}(\mathbf{x}_k,\mathbf{x}_0)\right|+\sum_{j=1}^{k}(-1)^j\left(\mathrm{x}_j-\mathrm{x}_{0j}\right)\left|\mathrm{c}(\mathrm{y}_0)\ \mathbf{d}^{(.,j)}(\mathbf{x}_k,\mathbf{x}_0)\right|=0$$

or

$$(y-\mathrm{y}_0)\left|\mathbf{d}(\mathbf{x}_k,\mathbf{x}_0)\right|+\sum_{i=1}^{k}(-1)^i(\mathrm{y}_\mathrm{i}-\mathrm{y}_0)\left|\mathbf{d}^{(i,.)}(\mathbf{x}_k,\mathbf{x}_0)\right|=0$$

or

$$(y-\mathrm{y}_0)\left|\mathbf{d}(\mathbf{x}_k,\mathbf{x}_0)\right|+\sum_{i=1}^{k}\sum_{j=1}^{k}(-1)^{i+j}\left(\mathrm{x}_j-\mathrm{x}_{0j}\right)\left(\mathrm{y}_\mathrm{i}-\mathrm{y}_0\right)\left|\mathbf{d}^{(i,j)}(\mathbf{x}_k,\mathbf{x}_0)\right|=0$$

or

$$\left(y-\mathrm{y}_0\right)\left|\mathbf{d}^{-1}(\mathbf{x}_k,\mathbf{x}_0)\right|-(x-\mathrm{x}_0)\mathbf{D}(\mathbf{x}_k,\mathbf{x}_0)\mathbf{c}(y_0)=0$$

or

$$y-\mathrm{y}_0-(x-\mathrm{x}_0)\mathbf{d}^{-1}\left(\mathbf{x}_k,\mathbf{x}_0\right)\mathbf{c}(y_0)=0.$$

These equations are equivalent and can be rewritten in the following ways:

$$y=\mathrm{y}_0-\frac{1}{\left|\mathbf{d}\left(\mathbf{x}_k,\mathbf{x}_0\right)\right|}\sum_{j=1}^{k}(-1)^j\left(\mathrm{x}_j-\mathrm{x}_{0j}\right)\left|\mathrm{c}(\mathrm{y}_0)\ \mathbf{d}^{(\cdot,j)}\left(\mathbf{x}_k,\mathbf{x}_0\right)\right|$$

or

$$y=\mathrm{y}_0-\frac{1}{\left|\mathbf{d}\left(\mathbf{x}_k,\mathbf{x}_0\right)\right|}\sum_{i=1}^{k}(-1)^i\left(\mathrm{y}_\mathrm{i}-\mathrm{y}_0\right)\left|\mathbf{d}^{(i,\cdot)}\left(\mathbf{x}_k,\mathbf{x}_0\right)\right|$$

or

$$y=\mathrm{y}_0-\frac{1}{\left|\mathbf{d}\left(\mathbf{x}_k,\mathbf{x}_0\right)\right|}\sum_{i=1}^{k}\sum_{j=1}^{k}(-1)^{i+j}\left(\mathrm{x}_j-\mathrm{x}_{0j}\right)\left(\mathrm{y}_\mathrm{i}-\mathrm{y}_0\right)\left|\mathbf{d}^{(i,j)}\left(\mathbf{x}_k,\mathbf{x}_0\right)\right|$$

or

$$y=\mathrm{y}_0+\frac{1}{\left|\mathbf{d}\left(\mathbf{x}_k,\mathbf{x}_0\right)\right|}(x-\mathbf{x}_0)\mathbf{D}\left(\mathbf{x}_k,\mathbf{x}_0\right)\mathbf{c}(y_0)$$

or

$$y=\mathrm{y}_0+(x-\mathbf{x}_0)\,\mathbf{d}^{-1}\left(\mathbf{x}_k,\mathbf{x}_0\right)\mathbf{c}(y_0). \qquad (5)$$

Let U be a fixed population of size *N*. Moreover, let $\mathrm{x}_{i,1}<\mathrm{x}_{i+1,1}$, i=1,...,N-1. Our problem is estimation of the population average $\bar{y}=\sum_{i=`1}^{N}y_i/N$. Let *s* be a simple sample of the fixed size *n>k+1*. The sample mean of the variable under study is defined by $\bar{y}_s=\sum_{i\in s}y_i/n$. The population mean vector of auxiliary variables we denote by $\bar{\mathbf{x}}=\left[\bar{x}_1\quad\bar{x}_2\quad\ldots\quad\bar{x}_k\right]$ where $\bar{x}_j=\sum_{i=1}^{N}x_{i,j}/N$ and the sample mean vector of auxiliary variables will be denoted by

$\overline{\mathbf{x}}_s = [\overline{x}_{1s} \quad \overline{x}_{2s} \quad \ldots \quad \overline{x}_{ks}]$ where, $j=1,\ldots,k$. Let $\mathbf{x}_n$ (n<N) be such a submatrix of the matrix **x** that the rows of the matrix $\mathbf{x}_n$ are observations of the auxiliary variables in the sample s of size n.

Let us consider the estimator which we obtain through changing $\overline{y}_s$ for y_0 and $(\overline{\mathbf{x}} - \mathbf{x}_0)$ for $(x - \mathbf{x}_0)$ only at the first row of the matrix given by the expression (2) and $\overline{\mathbf{x}}$ for $\mathbf{x}_k$. This leads to the following estimator of the population mean:

$$\tilde{\mathrm{y}} = \overline{\mathrm{y}}_s + (\overline{\mathbf{x}} - \mathbf{x}_0)\, \mathbf{d}^{-1}(\mathbf{x}_k, \mathbf{x}_0)\, \mathbf{c}(y_0) \tag{6}$$

or

$$\tilde{\mathrm{y}} = \overline{\mathrm{y}}_s + \frac{1}{|\mathbf{d}(\mathbf{x}_k, \mathbf{x}_0)|}(\overline{\mathbf{x}} - \mathbf{x}_0)\, \mathbf{D}(\mathbf{x}_k, \mathbf{x}_0)\, \mathbf{c}(y_0). \tag{7}$$

Particularly, if k=1 then

$$\tilde{y}_s = \overline{\mathrm{y}}_s + \frac{\mathrm{y}_1 - \mathrm{y}_0}{\mathrm{x}_1 - \mathrm{x}_0}(\overline{\mathrm{x}} - \overline{\mathrm{x}}_s).$$

The next estimator can be constructed in the following way. At the right side of the equation (5) let us change: the mean $\overline{y}_s$ for y_0, $\overline{\mathbf{x}}$ for **x** and $\overline{\mathbf{x}}_s$ for $\mathbf{x}_0$. This leads to the following estimator of the population mean:

$$\hat{\mathrm{y}} = \overline{\mathrm{y}}_s + (\overline{\mathbf{x}} - \overline{\mathbf{x}}_s)\, \mathbf{d}^{-1}(\mathbf{x}_k, \overline{\mathbf{x}}_s)\, \mathbf{c}(\overline{y}_s). \tag{8}$$

Let us note that particularly, if $k=1$, then

$$\hat{y}_s = \overline{\mathrm{y}}_s + \frac{\mathrm{y}_1 - \overline{\mathrm{y}}_s}{\mathrm{x}_1 - \overline{\mathrm{x}}_s}(\overline{\mathrm{x}} - \overline{\mathrm{x}}_s).$$

Our problem is how to determine the observation of the auxiliary variables consisting the rows of the matrix **x**. Firstly, let us note that they should be evaluate in such a way that $\mathbf{d}(\mathbf{x}_k, \overline{\mathbf{x}}_s) \neq 0$ or $\mathbf{d}(\mathbf{x}_k, \mathbf{x}_0) \neq 0$. The several ways of the matrix $\mathbf{x}_k$ determining can be proposed. The elements of $\mathbf{x}_k$ can be truncated means or quantiles. Some proposition of determining the elements of the matrix will be presented in the next sections.

III. QUANTILES OF AUXILIARY VARIABLES

The sample space of the samples s we denote by $\boldsymbol{S}$. The sample is of the fixed effective size $1<n<N$. The sampling design is denoted by $P(s)$. We assume that $P(s)>0$ for all $s\in\boldsymbol{S}$ and $\sum_{s\in\mathbf{S}} P(s)=1$.

Let us assume that observations of the first auxiliary variable increase when their indexes increase. Let $(X_{(j)1})=(X_{(1)1}, X_{(2)1},\ldots, X_{(n)1})$ be the sequence of the order statistics of observations of the first auxiliary variable in the sample s. The sample quantile of order α is defined, see e.g. Fisz (1963), as follows:

$$Q_{s,\alpha}=X_{(r)1}, \tag{9}$$

where $r=[n\alpha]+1$, the function $[n\alpha]$ means the integer part of the value $n\alpha$, $r=1,2,\ldots,n$. Let us note that $X_{(r)1}=Q_{s,\alpha}$ for $\frac{r-1}{n}\leq\alpha\leq\frac{r}{n}$. In this paper it will be more conveniently to consider the order statistic than the quantile.

Let $G(r_i,t_i,i=1,\ldots,k+1)=\{s:X_{(r_i)1}=x_{j_i,1},\ i=1,\ldots k+1\}$ be the set of all samples which r_i *-th* order statistics $(i=1,\ldots,k+1)$ of the first auxiliary variable are equal to $x_{t_i,1}$, respectively where $r_1\leq t_1<t_2<\ldots<t_{k+1}\leq N\text{-}n+r_k$. The size of the set $G(r_i,t_i,\ i=1,\ldots,k+1)$ is denoted by $g(r_i,t_i,\ i=1,\ldots,k+1)$ and

$$g(r_i,t_i,\ i=1,\ldots,k+1)=\prod_{i=1}^{k+2}\binom{t_i-t_{i-1}-1}{r_i-r_{i-1}-1}, \tag{10}$$

where $r_0=0$, $t_0=0$, $r_{k+2}=n+1$, $t_{k+2}=N+1$.

The sets $G(r_i,t_i,\ i=1,\ldots,k+1)$ and $G(r_e,t_e,\ e=1,\ldots,k+1)$ are disjoint for $i\neq e$. This and the expression (10) lead to the following ones:

$$\bigcup_{t_1=1}^{N-n+r_1}\ \bigcup_{t_2=t_1+r_2-r_1}^{N-n+r_2}\ \ldots\ \bigcup_{t_k=t_{k-1}+r_k-r_{k-1}}^{N-n+r_k} G(r_i,t_i,\ i=1,\ldots,k+1)=\mathbf{S},$$

$$\sum_{t_1=1}^{N-n+r_1}\ \sum_{t_2=t_1+r_2-r_1}^{N-n+r_2} \cdots \sum_{t_k=t_{k-1}+r_k-r_{k-1}}^{N-n+r_k} g(r_i,t_i,\ i=1,\ldots,k+1)=\binom{N}{n}. \qquad (11)$$

Hence, we have the following probability distribution of the order statistics from the simple sample.

$$P\left(X_{(r_1)1}=x_{t_1,1},\ X_{(r_2)1}=x_{t_2,1},\ldots,X_{(r_{k+1})1}=x_{t_{ik+1},1}\right)=\frac{g(r_i,t_i,i=1,\ldots,k+1)}{\binom{N}{n}}.(12)$$

In order to simplifying the notation let $\mathbf{X}_{(\mathrm{r})1}^{\mathrm{T}}=\left[X_{(r_1)1}\ X_{(r_2)1}\ \cdots\ X_{(r_{k+1})1}\right]$ and $\mathbf{x}_{(\mathrm{t})1}^{\mathrm{T}}=\left[x_{t_1,1}\ x_{t_2,1}\ \cdots\ x_{t_{ik+1},1}\right]$. Now we have: $G\left(\mathbf{X}_{(\mathrm{r})1}=\mathbf{x}_{(\mathrm{t})1}\right)=G(r_i,t_i,i=1,\ldots,k+1)$ and $g\left(\mathbf{X}_{(\mathrm{r})1}=\mathbf{x}_{(\mathrm{t})1}\right)=g(r_i,t_i,i=1,\ldots,k+1)$. Moreover, the equation (12) take the following shorter form: $P\left(\mathbf{X}_{(\mathrm{r})1}=\mathbf{x}_{(\mathrm{t})1}\right)=\dfrac{g\left(\mathbf{X}_{(\mathrm{r})1}=\mathbf{x}_{(\mathrm{t})1}\right)}{\binom{N}{n}}$.

III. SAMPLING DESIGN AND ESTIMATION STRATEGY

Let $\begin{bmatrix}\mathbf{x}_{(t)}\\ \mathbf{x}_{t_{k+1}*}\end{bmatrix}$ where $\mathbf{x}_{t_{k+1}*}=\left[x_{t_{k+1},1}\quad x_{t_{k+1},2}\quad \cdots\quad x_{t_{k+1},k}\right]$,

$$\mathbf{x}_{(t)}=\begin{bmatrix} x_{t_1,1} & x_{t_1,2} & \cdots & x_{t_1,k}\\ x_{t_2,1} & x_{t_2,2} & \cdots & x_{t_2,k}\\ \cdots & \cdots & \cdots & \cdots\\ x_{t_k,1} & x_{t_k,2} & \cdots & x_{t_k,k}\end{bmatrix}.$$

The first column $\mathbf{x}_{(t)1}$ of the matrix $\begin{bmatrix}\mathbf{x}_{(t)}\\ \mathbf{x}_{t_{k+1}*}\end{bmatrix}$ is the observation of the vector of the order statistic $\mathbf{X}_{(r)1}$. Now let us determine the matrix $\mathbf{d}\left(\mathbf{x}_{(t)},\mathbf{x}_{t_{k+1}*}\right)$ in the following way

$$\mathbf{d}\left(\mathbf{x}_{(t)},\mathbf{x}_{t_{k+1}*}\right)=\mathbf{x}_{(t)}-\mathbf{J}_k\mathbf{x}_{t_{k+1}*}. \tag{13}$$

Moreover,

$$\mathbf{c}(\mathrm{y}_0)=\mathbf{y}_{\mathrm{k}}\text{- }\mathrm{y}_0\mathbf{J}_{\mathrm{k}}, \tag{14}$$

Let us define the following expression.

$$z\left(\mathbf{X}_{(r)1},\mathbf{x}_{t_{k+1}*}\right)=\sum_{t_1=1}^{N-n+r_1}\sum_{t_1=t_2+r_2-r_1}^{N-n+r_2}\cdots\sum_{t_k=t_{k-1}+r_k-r_{k-1}}^{N-n+r_k}\left|\mathbf{d}\left(\mathbf{x}_{(t)},\mathbf{x}_{t_{k+1}*}\right)\right|g\left(\mathbf{X}_{(r)1}=\mathbf{x}_{(t)1}\right). \tag{15}$$

Definition 1. The sampling design proportional to the determinant $\left|\mathbf{d}\left(\mathbf{x}_{(t)1},\mathbf{x}_{j_{k+1}*}\right)\right|$ is as follows.

$$P\left(s\left|\mathbf{X}_{(r)1},\mathbf{x}_{t_{k+1}*}\right.\right)=\frac{\left|\mathbf{d}\left(\mathbf{x}_{(t)},\mathbf{x}_{t_{k+1}*}\right)\right|}{z\left(\mathbf{X}_{(r)1},\mathbf{x}_{t_{k+1}*}\right)} \tag{16}$$

for $s\in G\left(\mathbf{X}_{(r)1}=\mathbf{x}_{(t)1}\right)=G(r_i,t_i,i=1,...,k+1)$.

On the basis of the expression (7) we construct the following regression estimator:

$$\tilde{\mathrm{y}}=\overline{\mathrm{y}}_s+\frac{1}{\left|\mathbf{d}\left(\mathbf{x}_{(t)},\mathbf{x}_{t_{k+1}*}\right)\right|}\left(\overline{\mathbf{x}}-\mathbf{x}_{t_{k+1}*}\right)\mathbf{D}\left(\mathbf{x}_{(t)},\mathbf{x}_{t_{k+1}*}\right)\mathbf{c}(y_{k+1}) \tag{17}$$

or

$$\tilde{\mathrm{y}}=\overline{\mathrm{y}}_s+\left(\overline{\mathbf{x}}-\mathbf{x}_{t_{k+1}*}\right)\mathbf{d}^{-1}\left(\mathbf{x}_{(t)},\mathbf{x}_{t_{k+1}*}\right)\mathbf{c}(y_{k+1}) \tag{18}$$

Particularly, Wywial (2004) considered the case if k=1:

$$\widetilde{y}_s = \overline{\mathrm{y}}_s + \frac{\mathrm{y}_2 - \mathrm{y}_1}{\mathrm{x}_{(r_2)} - \mathrm{x}_{(r_1)}}\left(\overline{\mathrm{x}} - \overline{\mathrm{x}}_s\right). \tag{19}$$

He show that the strategy $\left(\widetilde{y}_s, P\left(s \middle| \mathbf{X}_{(r)1}, \mathbf{x}_{t_{k+1}*}\right)\right)$ is not unbiased for population mean $\overline{\mathrm{y}}$. So, in the considered case of multidimensional auxiliary variable the strategy is biased, too. From the other point of view the strategy can be useful when some outliers exists or the some observations of variable under study are censored.

IV. SUPERPOPULATION APPROACH

Let us consider the following regression model:

$$\mathbf{Y}=\mathbf{x}\mathbf{b}+\xi, \tag{19}$$

where $\mathbf{x}$ is the matrix of dimension N×k of non-random auxiliary variables observations, the observations of the random variables $\mathbf{Y}^T=[Y_1 \ldots Y_N]$ are values of the variable under study, $\mathbf{b}$ is the column vector of non-random parameters and $\xi^T=[\xi_1 \ldots \xi_N]$ is the vector of residuals and $E(\xi)=\mathbf{0}$, variance covariance matrix $\mathbf{V}(\mathbf{e})=\mathbf{I}_N\sigma^2$, where $\mathbf{I}_N$ is the unit matrix of the degree N. Hence, $Y_i=\mathbf{x}_{i*}\mathbf{b}+\xi_i$ and $E(Y_i)=\mathbf{x}_{i*}\mathbf{b}$.

On the basis of the expression (8) we construct the following predictor of the mean value $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$:

$$\hat{\mathrm{Y}} = \overline{\mathrm{Y}}_s + \left(\overline{\mathbf{x}} - \overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k, \overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{Y}_s). \tag{20}$$

The similar operations lead to the following result:

$$E_\xi\left(\hat{\mathrm{Y}}_{\mathrm{s}}\right) = \overline{\mathbf{x}}\mathbf{b} \tag{21}$$

So, the predictor $\hat{\mathrm{Y}}_{\mathrm{s}}$ is ξ-unbiased. Its variance is as follows:

$$\begin{aligned}
&D_\xi^2\left(\hat{\mathrm{Y}}\right)=D_\xi^2\left(\overline{\mathrm{Y}}_s\right)+D_\xi^2\left(\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{Y}_s)\right)+\\
&+2Cov_\xi^2\left(\overline{\mathrm{Y}}_s,\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{Y}_s)\right)=\\
&\quad=D_\xi^2\left(\overline{\xi}_s\right)+D_\xi^2\left(\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{\xi}_s)\right)+\\
&\quad+\_\,Cov_\xi^2\left(\mathrm{Y}\overline{\xi}_s,\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{\xi}_s)\right)=\\
&=\frac{\sigma^2}{n}+E_\xi\left(\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{\xi}_s)\right)^2+E_\xi\left(\overline{\xi}_s\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{\xi}_s)\right),
\end{aligned}\tag{22}$$

$$\begin{aligned}
&E_\xi\left(\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{\xi}_s)\right)^2=\\
&=\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)E_\xi\left(\mathbf{c}(\overline{\xi}_s)\mathbf{c}^T(\overline{\xi}_s)\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)^T\\
&E_\xi\left(\mathbf{c}(\overline{\xi}_s)\mathbf{c}^T(\overline{\xi}_s)\right)=E_\xi\left(\xi_k-\overline{\xi}_s\mathbf{J}_k\right)\left(\xi_k^T-\overline{\xi}_s\mathbf{J}_k^T\right)=\sigma^2\left(\mathbf{I}_k-\frac{2}{n}\mathbf{J}_k\mathbf{J}_k^T\right)
\end{aligned}$$

$$\begin{aligned}
&E_\xi\left(\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{\xi}_s)\right)^2=\\
&=\sigma^2\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{M}_k\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)^T,
\end{aligned}\tag{23}$$

where

$$\mathbf{M}_k=\mathbf{I}_k-\frac{1}{n}\mathbf{J}_k\mathbf{J}_k,\tag{24}$$

$$E_\xi\left(\overline{\xi}_s\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{c}(\overline{\xi}_s)\right)=\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)E_\xi\left(\mathbf{c}(\overline{\xi}_s)\overline{\xi}_s\right)=0\tag{25}$$

because

$$E_\xi\left(\mathbf{c}(\overline{\xi}_s)\overline{\xi}_s\right)=E_\xi\left(\xi_k-\overline{\xi}_s\mathbf{J}_k\right)=\frac{\sigma^2}{n}\left(\mathbf{J}_k-\mathbf{J}_k\right)=\mathbf{0}_k.$$

The expressions (12)-(25) lead to the following one

$$D_\xi^2\left(\hat{\mathrm{Y}}\right)=\frac{\sigma^2}{n}\left(1+\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{M}_k\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)^T\right)\tag{26}$$

Particularly, if k=1 then:

$$\hat{Y}_s = \bar{Y}_s + \frac{Y_1 - \bar{Y}_s}{x_1 - \bar{x}_s}\left(\bar{x} - \bar{x}_s\right).$$

We can show that $E_m\left(\hat{Y}_s\right) = E_m\left(\bar{Y}\right)$ and

$$D_\xi^2\left(\hat{Y}\right) = \sigma^2\left(\frac{1}{n} + \left(1 - \frac{1}{n}\right)\frac{\left(\bar{x} - \bar{x}_s\right)^2}{\left(x_1 - \bar{x}_s\right)^2}\right) \geq \frac{\sigma^2}{n}.$$

So, when a purposive sample s_1 is such a one that $\bar{x} = \bar{x}_{s_1}$ the above variance takes minimal value. The sample s_1 is called the balanced one.

The next particular case of the predictor given by the expression (20) is as follows

$$\hat{Y}_s = \bar{Y}_s + \frac{Y_2 - Y_1}{x_{(r_2)} - x_{(r_1)}}\left(\bar{x} - \bar{x}_s\right).$$

We can show that $E_\xi\left(\hat{Y}_s\right) = E_\xi\left(\bar{Y}\right)$ and

$$D_\xi^2\left(\hat{Y}\right) = \sigma^2\left(\frac{1}{n} + \frac{2\left(\bar{x} - \bar{x}_s\right)^2}{\left(x_{(r_2)} - x_{(r_1)}\right)^2}\right) \geq \frac{\sigma^2}{n}.$$

Similarly like in the previous case when a purposive sample s_1 is such a one that $\bar{x} - \bar{x}_{s_1} = 0$ the above variance takes minimal value.

The more general predictor is as follows:

$$\breve{Y} = Y_0 + \left(\bar{\mathbf{x}} - \mathbf{x}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k, \mathbf{x}_s\right)\mathbf{c}(Y_0) \qquad (27)$$

where Y_0 is an observed value of variable under study and it is not the element of the vector $\mathbf{Y}_k$, so $Cov(Y_0, Y_i) = 0$ for $Y_i \in \mathbf{Y}_k$. We can show that

$$E_\xi\left(\breve{\mathrm{Y}}_{\mathrm{s}}\right)=\overline{\mathbf{x}}\mathbf{b}. \tag{28}$$

So, the predictor $\breve{\mathrm{Y}}_{\mathrm{s}}$ is ξ-unbiased. Its variance can be derived similarly as the parameter $D_\xi^2\left(\hat{\mathrm{Y}}\right)$ and is as follows:

$$D_\xi^2\left(\breve{\mathrm{Y}}\right)=\sigma^2\left(1+\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\mathbf{L}_k\mathbf{d}^{-1}\left(\mathbf{x}_k,\overline{\mathbf{x}}_s\right)\left(\overline{\mathbf{x}}-\overline{\mathbf{x}}_s\right)^T\right), \tag{29}$$

where

$$\mathbf{L}_k=\mathbf{I}_k+\mathbf{J}_k\mathbf{J}_k. \tag{30}$$

Hence, the statistic $\breve{\mathrm{Y}}_{\mathrm{s}}$ is not consistent predictor of $\overline{\mathrm{Y}}$.

Finally let us consider the following predictor:

$$\tilde{\mathrm{Y}}_{\mathrm{s}}=\overline{\mathrm{Y}}_s+\left(\overline{\mathbf{x}}-\mathbf{x}_0\right)\mathbf{d}^{-1}\left(\mathbf{x}_{(t)},\mathbf{x}_0\right)\mathbf{c}(Y_0), \tag{31}$$

where $\overline{Y}_s=\frac{1}{N}\sum_{i\in s}Y_i$, $\mathbf{c}(\mathrm{Y}_0)=\mathbf{Y}_k\text{-}\mathrm{Y}_0\mathbf{J}_k$ and Y_0 is an observed value of variable under study and it is not the element of the vector $\mathbf{Y}_{k;}$ Under these assumptions we evaluate the ξ-expected value:

$$\begin{aligned}E_\xi\left(\mathbf{c}(Y_0)=\left[E_\xi\left(Y_i\right)-E_\xi\left(Y_0\right)\right]=\left[\mathbf{x}_i\mathbf{b}-\mathbf{x}_0\mathbf{b}\right]=\left(\mathbf{x}_{(t)}-\mathbf{x}_0\mathbf{J}_k\right)\mathbf{b}=\mathbf{d}\left(\mathbf{x}_{(t)},\mathbf{x}_0\right)\mathbf{b}\\ =\overline{\mathbf{x}}_s\mathbf{b}+\left(\overline{\mathbf{x}}-\mathbf{x}_0\right)\mathbf{d}^{-1}\left(\mathbf{x}_{(j)},\mathbf{x}_0\right)\mathbf{d}\left(\mathbf{x}_{(j)},\mathbf{x}_0\right)\mathbf{b}\end{aligned}$$

$$E_\xi\left(\tilde{\mathrm{Y}}_{\mathrm{s}}\right)=\overline{\mathbf{x}}_s\mathbf{b}+\left(\overline{\mathbf{x}}-\mathbf{x}_0\right)\mathbf{b}. \tag{32}$$

So, the predictor $\widetilde{\mathrm{Y}}_{\mathrm{s}}$ is ξ-biased. Its variance is derived in the following way:

$$\begin{aligned}&D_\xi^2\left(\tilde{\mathrm{Y}}\right)=D_\xi^2\left(\overline{\mathrm{Y}}_s\right)+D_\xi^2\left(\left(\overline{\mathbf{x}}-\mathbf{x}_0\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\mathbf{x}_0\right)\mathbf{c}(Y_0)\right)+\\ &+2Cov_\xi\left(\overline{\mathrm{Y}}_s,\left(\overline{\mathbf{x}}-\mathbf{x}_0\right)\mathbf{d}^{-1}\left(\mathbf{x}_k,\mathbf{x}_0\right)\mathbf{c}(Y_0)\right)=\end{aligned}$$

$$= D_\xi^2\left(\overline{\mathrm{Y}}_s\right) + D_\xi^2\left(\left(\overline{\mathbf{x}} - \mathbf{x}_0\right)\mathbf{d}^{-1}\left(\mathbf{x}_k, \mathbf{x}_0\right)\mathbf{c}(Y_0)\right) +$$
$$+2Cov_\xi\left(\overline{\mathrm{Y}}_s, \left(\overline{\mathbf{x}} - \overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k, \mathbf{x}_0\right)\mathbf{c}(Y_0)\right) =$$
$$= D_\xi^2\left(\overline{\mathrm{Y}}_s\right) + D_\xi^2\left(\left(\overline{\mathbf{x}} - \mathbf{x}_0\right)\mathbf{d}^{-1}\left(\mathbf{x}_k, \mathbf{x}_0\right)\mathbf{c}(Y_0)\right) + 0$$
$$D_\xi^2\left(\tilde{\mathrm{Y}}_\mathrm{s}\right) = \sigma^2\left(\frac{1}{n} + \left(\overline{\mathbf{x}} - \overline{\mathbf{x}}_s\right)\mathbf{d}^{-1}\left(\mathbf{x}_k, \overline{\mathbf{x}}_s\right)\mathbf{L}_k\mathbf{d}^{-1}\left(\mathbf{x}_k, \overline{\mathbf{x}}_s\right)\left(\overline{\mathbf{x}} - \overline{\mathbf{x}}_s\right)^T\right) \tag{33}$$

Hence, the statistic $\widetilde{\mathrm{Y}}_\mathrm{s}$ is not consistent predictor of $\overline{\mathrm{Y}}$.

The particular case of the above predictor can be as follows:

$$\widetilde{Y}_s = \overline{\mathrm{Y}}_s + \frac{\mathrm{Y}_2 - \mathrm{Y}_1}{\mathrm{x}_{(r_2)} - \mathrm{x}_{(r_1)}}\left(\overline{\mathrm{x}} - \mathrm{x}_0\right).$$

We can show that $E_m\left(\widetilde{Y}_s\right) = E_m\left(\overline{Y}\right)$ and

$$D_\xi^2\left(\widetilde{Y}_s\right) = \sigma^2\left(\frac{1}{n} + \frac{2\left(\overline{x} - x_0\right)^2}{\left(x_{(r_2)} - x_{(r_1)}\right)^2}\right) \geq \frac{\sigma^2}{n}.$$

Let us remind the bias of the predictor is $E_\xi\left(\tilde{Y}_s\right) = \overline{x}_s b + \left(\overline{x} - x_0\right)b$. So, when a purposive sample s$_2$ is such a one that $\overline{x} = x_0$ the above variance takes minimal value. For instance s$_2$ can be such a balanced sample that sample median x$_0$ is equal to the population mean $\overline{x}$.

Acknowledgement

The research was supported by the grant number 1 H02B 018 27 from the Ministry of Science and Higher Education

REFERENCES

Borsuk K. (1969): *Multidimensional Analytic Geometry*. PWN, Warsaw.

Fisz, M. (1963). *Probability Theory and Mathematical Statistics*. Wiley and Sons Inc., New York.

Hellwig Z. (1963). *Linear Regression and Its Application in Economy (in Polish)*. PWN, Warszawa.

Kendall, M. G., Stuart, A. (1961). *The Advanced Theory of Statistics. Vol. II: Inference and Relationship*. Charles Griffin & Company Limited, London.

Wald, A. (1940): The fitting the stright lines if both variables are subject to errors. *Annals of Mathematical Statistics*, 11, pp. 284

Wywiał. J. (2004). Quqantile regression sampling strategy. In: *Metoda Reprezentacyjna w Badaniach Ekonomiczno-Społecznych (Survey Sampling in Economical and Social Research)*. (Edited by J. Wywiał) Katowice University of Economics, Katowice.

Janusz Wywiał

OCENA WARTOŚCI PRZECIĘTNEJ ZA POMOCĄ REGRESYJNEJ STRATEGII LOSOWANIA WYKORZYSTUJĄCEJ KWANTYLE ZMIENNEJ POMOCNICZEJ

Problem oceny wartości średniej z wykorzystaniem danych o wszystkich wartościach cech pomocniczych jest rozważany. W tym celu znany estymator regresyjny zależny od wielu zmiennych pomocniczych jest wykorzystywany. W odróżnieniu od zwykłego podejścia znanego w metodzie reprezentacyjnej do oceny parametrów regresji są wykorzystywane kwantyle jednej ze zmiennych dodatkowych. Otrzymane na tym polu wyniki są adoptowane do konstrukcji predytorów wartości średniej w nadpopulacji. Wyprowadzono również wariancje różnych odmian proponowanych predykatorów.