ROBUST SMALL AREA ESTIMATION FOR HOUSEHOLD CONSUMPTION EXPENDITURE QUANTILES USING M-QUANTILE APPROACH (CASE STUDY: POVERTY INDICATOR DATA IN BOGOR DISTRICT)

Kusman Sadik [1,a)], **Girinoto**[1,b)], **Indahwati**[1,c)]

[1]Department of Statistics, Faculty of Mathematics and Natural Science, Bogor Agricultural University, Bogor, Indonesia.

[a)] kusmansadik@gmail.com;. [b)] girinoto@gmail.com;[c)] indahwati_43@yahoo.co.id

Abstract. The consumption expenditure is one of important variable to estimate poverty in certain area. Although averages are widely used in many applications, relying only averages may not be very informative in providing complete picture of consumption expenditures. In fact, the distribution of expenditures commonly suffering from outlier that it may needs to be viewed with caution from misleading interpretation. To overcome such problem, employing quantile parameter or robust procedure can be viewed as an option. In Indonesia, The National Socio-Economic Survey (Susenas) samples are designed to produce estimates of parameters from planned domains (provinces and districts). The estimation of unplanned domains (sub-districts and villages) has its limitation to obtain reliable direct estimates. One of the possible solutions to overcome such problem is employing small area estimation techniques. In this paper, as an alternative approach for this purpose, we are using M-quantile regression for small area based on modeling quantile-like to estimate quantile of household consumption expenditure on unplanned domain. The aim of study is to give more complete insight for conditions the distribution consumption expenditure among sub-district in Bogor District-West Java in the perspective of quantiles and average estimator obtained from small area estimation framework.

Keywords: M-quantile regression, Podes, poverty, small area estimation, Susenas

JEL Classification : I32, C31, C83

1. INTRODUCTION

One of variable commonly used to describe the economic level in certain area is household consumption expenditure. However, it commonly suffering with the presence of outlier where it can be affect in matter of the parameter estimation. To remedy this problem, researcher have been developed a robust class methods finding estimator.

Indonesian government via Indonesia Statistics (*Badan Pusat Statistik, BPS*) published estimation household consumption expenditure is calculated by using National Socioeconomic Survey (Susenas) data. However, the estimation was available only at the level of province and districts. There has been rising demand of estimation of household consumption expenditure at sub-district and village levels as the support information of allocation of grants and fiscal transfer targeting in poverty reduction program by Indonesian Government. The problem arise when the result of direct estimation from the small domain/area yields estimates with the unacceptable level precision cause of matter a small sample size within small domain/area. The small area estimation technique can be employ to overcome the problem via models that borrow strength data from other area. The conventional method in small area estimation of model estimation is random effects models that include random area effects to account for between area variations beyond that explained by auxiliary variable (Rao 2003).

The application small area estimation techniques not merely applied to estimate the total and average. Recently, researchers in small area estimation have developed model to estimate other parameter such quantile/percentile. The popular model of application in small area estimation is dominated by a class of linear mixed model with rely on strong normality assumptions. In this paper we choose an approach base of small area estimation for using M-quantile estimator model proposed by Marchetti *et al.*(2012) to estimate the consumption expenditure quantiles and average. M-quantile regression models approaching small area estimation introduced by Chamber and Tzavidis (2006) for the first time. Unlike the conventional model of small area estimation techniques which rely on regression model that utilizing both covariates and random effects in order to explain variation between areas. M-

quantile small area estimation is based on modeling quantile-like parameters of conditional distribution target variable given covariates. The difference among area is characterized by area M-quantile coefficients.

In most practical application in estimating household consumption expenditure using unit level model of small area estimation technique based on survey data with "borrowing strength" by linking auxiliary information from census data. However, the major shortcoming is the availability of raw data of census, because of its restricted information protected by law and policy. We consider to utilizing the existence of village census Podes (The Villages Potential Statistics) data as the auxiliary information. We focus on estimation of household consumption expenditure at the level sub-districts in Bogor districts, West Java. The sub-district level is identified as small domain/unplanned domain part of National Socio-economic Survey project. We estimate the quantiles and average parameter of sub-districts level in Bogor district using small area estimation with the M-quantile regression approach. We also compare the results from small area estimation framework with the result from direct estimation as the conventional method to estimate quantiles and average parameters.

.

2. LITERATUR REVIEW

2.1. Robust Regression

In classical regression, the least squares method has been generally adopted, however there is presently awareness of the dangers posed by the occurrence of outliers, which may lead to a misleading result. In general, robust regression estimators aim to fit a model that describes the majority of a sample (Rousseeuw and Leroy 1987). Their robustness is achieved by such a calculation procedure of giving the data different weights, so that outlying data have relatively smaller influence. Comparatively, in least squares conventional regression all residuals are treated equally. M-estimator $\hat{\beta}$ minimize the objective function

$$\sum_{i=1}^{n}\rho(v^{-1}r_i)=\sum_{i=1}^{n}\rho(v^{-1}(y_i-\mathbf{x}_i\beta_i)) \quad (1)$$

where v as scale parameter. Define $\psi(u)=\rho'(u)$ as influence function in order to minimize (1), where $\hat{\beta}$ can be obtained by solving

$$\sum_{i=1}^{n}\psi(v^{-1}(y_i-\mathbf{x}_i\beta_i))\mathbf{x}_i=0 \quad (2)$$

The weight function of scaled residual define as $w(u)=\frac{\psi(u)}{u}$. There are several alternative of influence function. One of suggested is Huber function is defined as $\psi(u)=uI(-c\le u\le c)+c.\text{sgn}(u)I(|u|>c)$ where c is tuning constant. It gives different weights within the calculation of estimate depends on the value of residual u.

2.2. M-quantile Regression

In classical regression, a model is built is to describe relationship between dependent variable y and explanatory variable x with way of explaining behavior average value of y given x. Instead, quantile regression (Koenker and Basset 1987) is to describe the relationship with way of explaining behavior quantile value of y given x. In M-quantile regression (Breckling and Chambers 1988), the basic idea is to integrate the general concept of quantile regression and M-estimation to achieve a robust regression class based on influence function. M-quantile of order q ($0<q<1$) of conditional variable y given $\mathbf{x}$ is defined as solution $Q_q(\mathbf{x};\psi)$ that satisfied

$$\int\psi_q(y-Q_q(\mathbf{x};\psi))f(y\,|\,\mathbf{x})dy=0 \quad (3)$$

A linear model of M-quantile regression is defined as

$$Q_q(\mathbf{x};\psi) = \mathbf{x}^T \beta_\psi(q) \tag{4}$$

in contrast to standard regression, it allow a different set of regression parameters for each value of q. For specified q and influence function of ψ, an estimate $\hat{\beta}_\psi(q)$ of $\beta_\psi(q)$ can be obtained by solving

$$\sum_{i=1}^{n} \psi_q\left(v^{-1}(y_i - \mathbf{x}_i^T \hat{\beta}_\psi(q))\right)\mathbf{x}_i = 0 \tag{5}$$

where $\psi_q(u) = 2\psi(u)\{qI(u>0) + (1-q)I(t \le 0)\}$, $\psi(u)$ is influence function, in this paper we refer to Huber function with value of c is set at 1.345 scale estimator of v defined as Median Absolute Deviance (MAD). According to Chamber and Tzavidis (2006) the technique is no need distributional assumptions in contrary with classical regression.

2.3. M-quantile Modeling for Small Area Estimation

Suppose that a population U can be partitioned into d areas, indexed by j=1,..,d with area j containing N_j units. s_j is sample set containing n_j units is taken from population area j, with the $N_j - n_j$ as non-sample units denote by r_j. Following the work of Tzavidis *et al*. (2010) unit level model of small area estimation distribution function term of sample and non-sample elements

$$\hat{F}_j(t) = N_j^{-1}\left\{\sum_{i\in s_j} I(y_{ij} \le t) + \sum_{k\in r_j} n_j^{-1} \sum_{i\in s_j} I\{\hat{y}_{kj} + (y_{ij} - \hat{y}_{ij}) \le t\}\right\} \tag{6}$$

where the estimation $\hat{y}_{ij}$ of $\hat{y}_{kj}$ and based on linear M-quantile regression (4) are defined as $\hat{y}_{ij} = \mathbf{x}_{ij}^T \hat{\beta}_\psi(\hat{\theta}_j)$ and $\hat{y}_{kj} = \mathbf{x}_{kj}^T \hat{\beta}_\psi(\hat{\theta}_j)$ respectively. They are based on linear M-quantile regression for small area procedure Chamber and Tzavidis (2006). The idea is using non-sample covariates information (at population level) and sample of study variables. For unit i with value y_i and x_i there exist a value of M-quantile coefficient q_i such that $Q_{q_i}(\mathbf{x}_i;\psi) = y_i$. Models attempt to capture area effects by estimating an area specific q value $(\hat{\theta}_j)$ for each area of a hierarchical data set. Here the area specific $\hat{\theta}_j$ is estimated by average value q_{ij} in area.

2.4. Estimation Average and Quantile Parameter

Using (6), an estimator of average of y in small area j is the value obtained by

$$\hat{m} = \int_{-\infty}^{+\infty} y \, d\hat{F}_j(y) = N_j^{-1}\left[\sum_{i\in s_j} y_i + \sum_{i\in r_j} \hat{y}_i + (1-f_j)\sum_{i\in s_j} e_i\right] \tag{7}$$

where $f_j = n_j N_j^{-1}$ and e_i defined as residual model in the sample part. The parameter quantile $\phi \in (0,1)$ of the distribution function small area j value is then obtained by a numerical solution to the following estimating equation

$$\int_{-\infty}^{\hat{q}(j;\phi)} d\hat{F}(t) = \phi \tag{8}$$

3. RESEARCH METHOD

3.1. Data

Data used in this research were secondary data collected from Indonesia Statistics. Study variable is household per capita consumption expenditure in Bogor district (*Y*) where it taken from National Socio-economic Survey 2015. The Auxiliary variables were taken from PODES 2014. Bogor district has the largest area among districts in West Java, Indonesia. It comprises 434 Villages administration level clustered within 40 sub-districts. The selected covariates in auxiliary variables are known for each unit in the population and have resulted significant in the model expenditure described in Table 1.

TABLE 1 List of Variables

No	Variables	Explanation
1	Y	Household per capita consumption expenditure
2	X_1	Agricultural household (1=Agricultural hh; 0 = non Agricultural hh)
3	X_2	The number electricity connection
4	X_3	Total person received government aid in village
5	X_4	The number of traditional store in village

3.2. Methods of Data Analysis

In this section, we involve the direct estimate as the conventional methods obtaining the quantiles and average in order to compare the small area method. We only focus on quantiles parameter at value of 0.25, 0.5 0.75 as the form of quartile. The stages of data analysis in this research involved:

i. Preliminary Analysis
   Descriptive analysis was performed to explore the general description of data pattern in order to get the appropriate next analysis. In this step also we perform the linear mixed model as conventional method in small area estimation model.
ii. Direct estimate for Quantiles and Average
   We assume the survey design using simple random sampling (SRS) to get the direct estimate of quantiles and average from sample. In order obtaining the Mean Square Error (MSE) for the estimator we use bootstrap estimator for MSE.
iii. Small Area Estimation
   Small area estimation based on M-quantile regression modeling is the unit level type, where the auxiliary information covariates known for all the populations in the small area. As we utilizing the village census data for the auxiliary, we assume all households in the same village will have the same value or characteristics to construct the auxiliary population matrix. For more detail of the computing procedure can be found in Marchetti *et al.* (2012) work.

iv. Comparing the results

There are three aspects that we are going to compare from the results:

- Quartile range
- Average
- Root Mean Square Error (RMSE)

v. Conclusion

4. RESULT AND DISCUSSION

Figure 1 shows the distribution of household consumption expenditure per capita in sample. This figure indicated that the distribution of the household consumption expenditure have skewed distribution caused of several outliers. Figure 1 describes the distribution in the level of district and in the level sub-districts. In Figure 2 shows the result of residual from fitting linear mixed effect model are fail to satisfy the normality assumption. The preliminary analysis shows that it is plausible to use robust version in small area estimation.

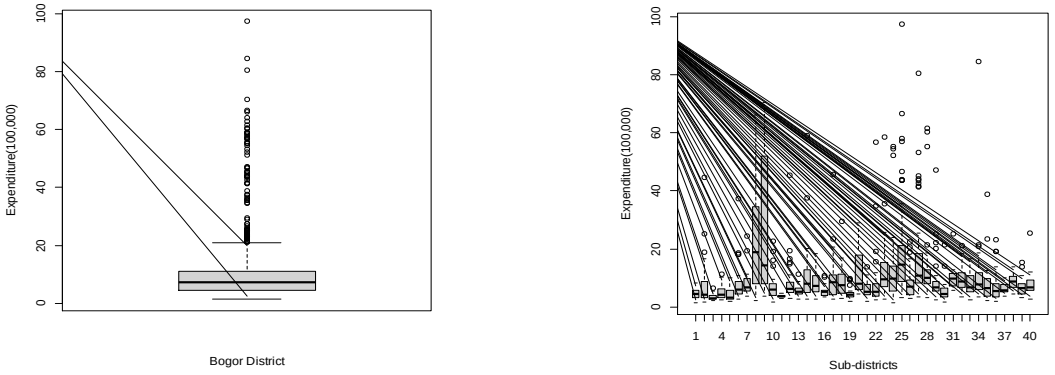

FIGURE 1 Distribution of household expenditure by district and sub-districts

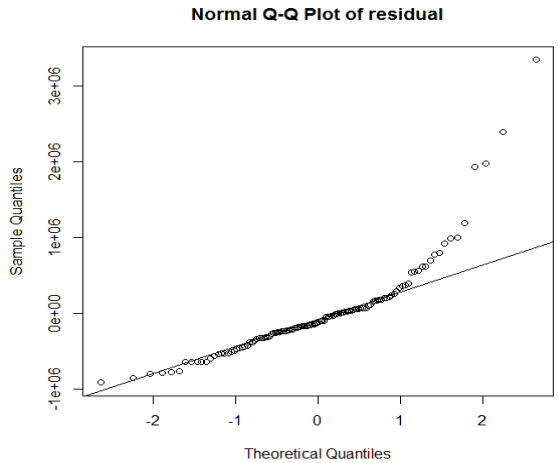

FIGURE 2 Normal plot of residual from fitting in linear mixed effect model

Direct estimate was performed in order to get more complete of picture about nature of data of household consumption expenditure. In fact, we already know at the previous preliminary analysis that data is suffered from outlier existences. Table 2 is the result of direct estimate of quantiles parameter for 0.25, 0.50 and 0.75. In the last column, we also have an average to give more insight about the central location of data in compare with median. As we can see the result of direct estimate in Table 2, the distribution of household consumption expenditure for most sub-districts tend to have different median (q0.50) from average.

TABLE 2 Household Consumption Expenditure Quantiles and Average Based On Direct Estimate From Sample

Subd	q(0.25)	q(0.50)	q(0.75)	mean	Subd	q(0.25)	q(0.50)	q(0.75)	mean
1	3.16	4.46	5.83	4.55	21	4.22	5.24	7.59	6.06
2	3.25	4.13	8.84	7.27	22	3.75	5.25	8.10	8.86
3	2.83	3.00	3.83	3.55	23	7.23	9.54	14.85	12.14
4	3.30	4.17	6.24	5.08	24	5.48	9.81	14.04	14.84
5	2.83	3.22	5.72	4.57	25	8.91	14.51	21.12	19.39
6	4.48	6.02	8.67	7.84	26	4.34	7.06	9.19	8.15
7	5.38	6.69	9.92	8.47	27	8.19	10.72	18.44	15.44
8	8.35	18.86	32.84	22.18	28	8.09	10.08	12.83	13.27
9	7.99	14.24	51.93	27.57	29	5.65	6.81	8.49	9.69
10	4.03	5.99	8.00	7.19	30	3.39	4.49	6.38	5.82
11	3.54	3.76	3.98	3.77	31	7.51	9.81	11.11	10.62
12	5.06	6.35	8.53	8.97	32	6.72	8.86	11.84	9.64
13	4.49	5.17	6.54	5.67	33	5.30	6.79	10.92	8.17
14	5.03	8.05	12.80	11.93	34	6.04	7.74	11.80	11.03
15	5.43	7.21	10.64	8.02	35	3.82	6.39	9.46	9.07
16	4.04	5.22	5.71	5.55	36	3.57	5.60	8.00	6.78
17	4.29	8.49	10.91	9.77	37	5.36	5.85	7.29	6.30
18	5.04	7.41	10.68	9.80	38	6.81	8.79	10.14	8.58
19	3.55	4.20	5.10	4.94	39	4.53	6.53	8.07	7.04
20	5.88	8.05	17.94	12.17	40	5.66	6.81	9.39	8.09

value x Rp100.000

TABLE 3 Household Consumption Expenditure Quantiles and Average Based On Small Area Estimation

Subd	q(0.25)	q(0.50)	q(0.75)	mean	Subd	q(0.25)	q(0.50)	q(0.75)	mean
1	2.67	4.88	6.39	4.55	21	4.14	5.19	7.81	6.02
2	2.60	4.77	7.44	6.48	22	2.71	4.44	7.50	7.84
3	1.94	2.52	3.57	2.92	23	5.35	7.93	10.37	10.45
4	2.83	3.59	5.82	4.66	24	5.20	9.02	10.21	13.91
5	4.21	5.57	7.93	6.12	25	6.52	11.10	12.71	16.39
6	3.75	5.54	7.99	7.22	26	4.30	6.06	7.88	7.10
7	4.62	6.66	8.34	8.15	27	5.65	8.80	10.77	12.95
8	6.64	12.24	12.88	20.63	28	5.82	7.33	10.15	10.79
9	8.18	13.24	14.06	26.38	29	4.57	5.97	8.33	8.88
10	3.46	5.55	8.00	6.75	30	2.76	3.84	5.31	5.10
11	3.16	3.42	3.85	3.48	31	7.14	9.20	11.44	10.63
12	4.69	6.44	9.64	8.95	32	5.59	8.03	9.76	8.74
13	3.88	4.86	6.28	5.27	33	4.91	6.85	9.31	7.93
14	4.98	8.28	10.02	11.40	34	4.72	6.71	9.11	10.04
15	3.45	5.58	8.54	6.64	35	3.21	4.33	8.23	7.86
16	3.23	4.40	5.96	4.89	36	3.03	4.99	7.54	6.23
17	4.38	7.24	9.79	8.72	37	4.11	5.34	6.68	5.57
18	5.22	8.23	11.21	10.75	38	6.40	8.66	11.02	8.67
19	2.49	3.25	4.35	4.00	39	3.73	6.49	8.33	6.42
20	5.86	7.65	10.29	11.82	40	3.83	6.20	8.16	7.00

value x Rp100.000

Based on Table 2 and Table 3, we will evaluate the results. The first objective is to evaluate the quantile range (q0.75 - q0.25) from both methods. The Figure 3 shows the range result from household consumption expenditure by direct estimate and M-quantile small area

estimation. As we can see there was a quite large of the difference at sub-district 8 and 9. It indicates that even the quantiles obtained from direct estimate is not effective to handle the tail of the distribution. The second, evaluating the average. Figure 4 shows the estimated value of averages from direct estimate and M-quantile small area estimation. As we can see, the model small area estimation gave a smaller estimation value for most districts. In the case of poverty application field, it will be such a correction quantity to enhance location grants and fiscal aid estimation.

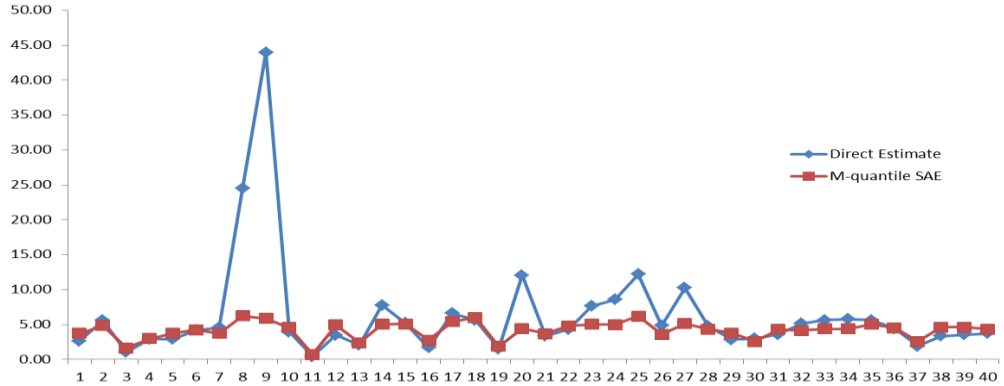

FIGURE 3 Plot of quantile range

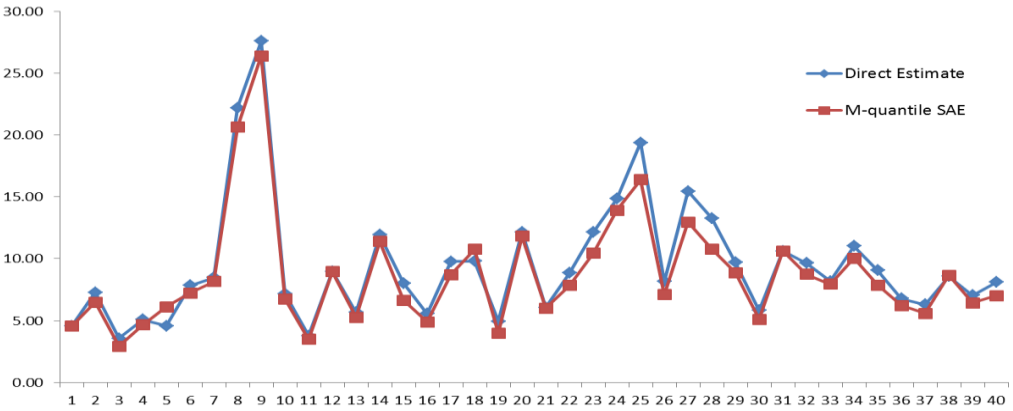

FIGURE 4 Plot of Average

In evaluating the quality aspect of the estimator, we focus on RMSE of average parameter from both methods. As the result in Figure 4, the pattern of RMSE from both methods are not significanly different. It shows that even the estimator obtained from robust small area estimation approach does not give any improvement in quality of estimation significantly. It might caused by such a trade of between robustness and precision of estimator.

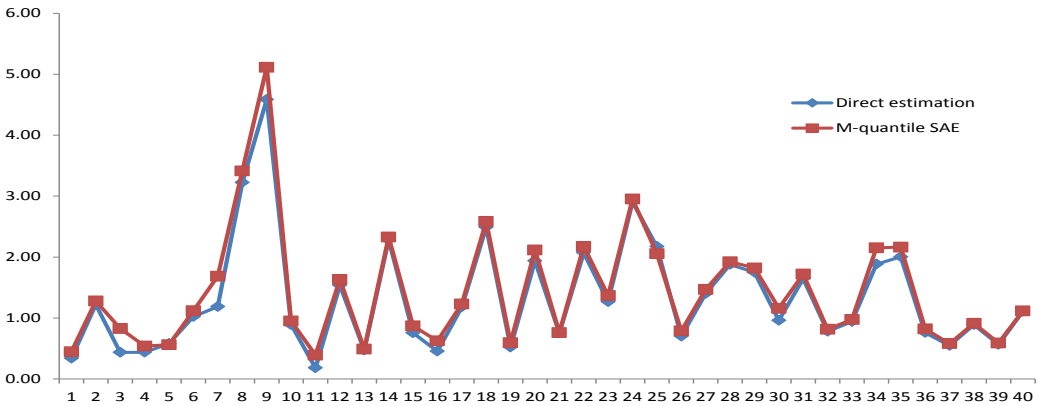

FIGURE 5 Plot of RMSE

5. CONCLUSIONS

From empiric evidence the household expenditure tends to have non-normal distribution, in application to describe the economic condition and welfare by its variable is need to be more wisely by avoiding the use of average from direct estimate methods. As suggested option for those purposes, using robust statistics such as quantile/quartile would be more reducing the influence of the outlier presence. We demonstrate how to utilize the available resource in practical situation such as the information of census in unit level is unavailable to build model of small area estimation framework. In this case study, even though from the quality of estimation aspect is not significantly different from direct estimate, considering the distribution of data of household expenditure tends to have non-normal distribution employing the robust approach based on small area estimation modeling become an advantage in estimating the parameter of quantiles and average.

REFERENCE

[1] Breckling J, Chambers RL. 1988. M-quantiles. *Biometrika* 75, 761–771
[2] Chambers RL, Tzavidis N. 2006. M-Quantile models for small area estimation. *Biometrika*. 93, 255-268
[3] Giusti C, Marchetti S, Pratesi M, Salvati N. 2012. Robust small area estimation and oversampling in the estimation of poverty indicators. *Survey Research Methods.* 6, 155-163
[4] J.N.K. Rao. *Small Area Estimation*. John Wiley and Sons. New York. 2003
[5] Koenker R, Basset G. 1978. Regression quantiles. *Econometrica* 46, 33–50
[6] Marchetti S, Tzavidis N, Pratesi M. 2012. Non-parametric bootstrap mean squared error estimation for m-quantile estimators of small area averages, quantiles and poverty indicators. *Computational Statistics & Data Analysis*. 56(10), 2889-2902
[7] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection*. John Wiley and Sons. New York. 1987, pp.65.
[8] Tzavidis N, Marchetti S, Chambers R. 2010. Robust prediction of small area means and distributions. *Australian and New Zealand Journal of Statistics*, 52, 167-186