ATLANTIS PRESS

Advances in Computer Science Research (ACSR), volume 85

Third Workshop on Computer Modelling in Decision Making (CMDM 2018)

Estimation of Fuzziness for Respondents' Quantitative Data on Income

Viacheslav P. Sirotin*[0000−0001−7783−8790], Ksenia V. Alexeeva†[0000−0003−4557−2937]

National Research University Higher School of Economics, Moscow, Russia

*vpsirotin@yandex.ru, †Ksenia.v.alexeeva@gmail.com

***Abstract*—The problem of impreciseness of data from different sources is in the focus of the research. While values of income based on diary notes may be considered as relatively exact measuring results, the values of income in other surveys founded on memory of respondents are rather ambiguous. The ambiguity reveals itself first of all in the tendency of rounding numbers. The usage of rounded values for identification of statistical and econometric models can make the estimates of the parameters biased. To prevent this effect, we propose using survey data in fuzzy format and show the way to estimate the measure of fuzziness. The main idea behind this estimation is in finding such measure of ambiguity (fuzziness) that provides the closest distribution of fuzzy data to the distribution of corresponding exact (crisp) data. As a source of crisp data, we use the household budget survey provided by the Federal State Statistics Service, and the Russian Longitudinal Monitoring Survey RLMS HSE supplies us with fuzzy data for corresponding period of time. It is shown that the algorithm presented in the paper allows us to improve data to a good level of conformity.**

***Index Terms*—estimation bias, household income, rounding numbers, crisp data, fuzzy data**

I. Introduction

With the widespread of the Big Data every researcher can access lots of data about different aspects of life, the vast majority of the data are polls and surveys conducted by research organizations. Thus, the answers of the respondents, especially in numerical scales can be specifically distorted. Using this data as exact may affect the result of the research. The alternative for the data from the respondent's answers is the polls carefully conducted based on diaries. Therefore, it is necessary to study the ways the data from the responses of surveys based on the respondents answers by memory (fuzzy data) and the data from diaries (crisp data) fit each other.

This paper studies the difference between the exactly recorded income per household and the income stated by memory. To study the effects of the errors the results of two surveys has been used and the random sample of 1000 responses has been selected from them. The first database is household budget survey by the Federal State Statistics Service, which presents the exact numbers (crisp data) of income of households all over Russia and the survey is constructed by using the consumption diaries where the respondents were asked to submit their actual income and spending every day. The second database is the Russia Longitudinal Monitoring Survey - Higher School of Economics (RLMS-HSE), which presents the remembered household income of the respondents all over Russia (soft, or fuzzy, data).

It is assumed the results of the actual and remembered household income to be different in two ways: the first one, the ordinary rounding may reflect the distortion, and the second one, the rounding by five will also have an effect on this difference.

It is also expected that rounding by five will be more influential, as it shows more uncertainty as does not refer to the new digit, which from the authors' point of view, adds accuracy to measurements.

Furthermore, it is supposed that with the correction the crisp and fuzzy datasets of households' income can be conformed.

II. The Problem of Distortion of Respondents' Monetary Statements in Scientific Literature

Lots of authors have studied the topic of the paper from different points of view. For example, there has been done a research by Borisov A. N. and his colleagues showing that people round numbers differently, the expert opinions have been collected from the borders of the numbers from range 1 to 99. The authors have used the data to develop the algorithm to construct the functions on the base of the expert assessments, the fuzzy sets concept was used. As a result, the algorithm of rounding is quite complicated [1].

A different approach was used in a series of surveys, the error was modeled using OLS and Quintile regression models by subtracting remembered income from real one including different groups of people separation. The examples of such research are made by Michael E. Borus 1970 [2], Martin David 1962 [3], Chang Hwan Kim and Christopher R. Tamborini 2014 [4]. The main idea of the latest research presented was to model the errors the following way:

$$u_i^{SR} = \delta y_i + \nu_i^{SR}, \tag{1}$$

where u_i^{SR} is the income stated by the memory, y_i is real income, gained from different sources, usually governmental organizations and income forms submitted, and ν_i^{SR} is the error between the two, then OLS and Quintile regression models were used. This research is based on more modern concepts of econometrics and has used the latest data and technology to analyze large samples.

Hence, the linear regression by itself does not give any interesting results, the authors add control variables to compare errors for different groups of people. For example, it was found out that female workers overreport at bottom 20% and underreport at top second 20% their earnings, people with

higher education give more accurate responses than those who do not have any education, at the same time higher earners tend to underreport earnings. As all of the researches presented are made in the USA at different times and the topic of nationality and skin color was hot, the attention is drawn to the topic in the researches, the following was found – black workers tend to underreport earnings to a greater degree than white ones. Another topic was to look at the frequency of changing jobs by the population, it was found out that job switchers tend to underreport earnings regardless other conditions.

Even though the results of the surveys and the methods applied show much patience and accuracy in calculation they do not include the rounding factors, and do not differentiate the origin of errors, which may be a minus of the research and therefore this missing valuation gives us the space for additional analysis. Another limitation characteristic of the surveys provided above is the fact that all information used is linked, which means that the authors know the remembered information of the person and the real information of the same person from the official sources. Thus, this may not be the case for many other countries, including Russia, therefore a more varied method should be provided to correct the calculations.

III. Modelling of Responses in Surveys

1) Identification of the Rounding Error.: To begin with a critical point is to show the importance of the research. The regression presented below is an example of how the income estimated with bias influences the conclusions drawn from the research. As spending is secondary to income, therefore income is one of the independent variables, some other descriptive variables of households can also be added to the regression model. The spending in the model can be any kind of spending including most often addressed currently – spending on information technologies including smart-phones, laptops, tablets and the Internet usage.

$$\text{spending} = \beta_0 + \beta_1 \text{income} + \\ + \sum_{j=2}^{l+1} \beta_j (\text{Descriptive variable}) + \varepsilon_i, \quad i = \overline{1,n}, \quad (2)$$

where l is a number of descriptive variables included, n is a number of observations in the sample.

Based on the fact proved in the previous research, income is measured with the uncertainty, therefore, there is a problem of stochastic regressors. Let's observe a simplified model for one of the cases of settlement type:

$$y = x\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma_\varepsilon^2), \quad (3)$$

where y is spending on the Internet, x is income.

Hence, income is stated with an error, it can also be modelled as follows:

$$x = x^* + u, u \sim N(0, \sigma_u^2). \quad (4)$$

For the case of the only regressor *income* the estimated coefficient will tend to the value:

$$p\lim \hat{\beta}_1 = \beta_1 \left(1 - \frac{\sigma_u^2}{\sigma_u^2 + \sigma_x^2}\right). \quad (5)$$

It can be clearly seen that the coefficient is estimated with bias towards zero, and, as a result, the regression model is not accurately predicting the examining value [5].

Speaking closely about income there may be several reasons for errors: one of them is calculation (when people cannot correctly summarize the earnings), the second one is unwillingness of respondents to present actual numbers, and finally, one more error is rounding numbers. At this part of the research we cannot estimate and correct the first two errors (as knowledge of the real earnings of the respondents is needed for this), thus, we can estimate the rounding errors. The further research will present the algorithm of how the income rounding error can be corrected, so that effects of stochastic regressors can be decreased. As we are not sure which direction the number is rounded, we will use the fuzzy variables theory to test the rounding hypothesis.

IV. Rounding Error Estimation

First of all, let us observe the level of conformity of the responses from representative household RLMS 2015 survey (fuzzy data) and the crisp data of the Federal State Statistics survey on household budgeting. The reason for this is that both datasets used include household income and are representative towards population of Russia, therefore, they should be conformed. The presented histogram shows that the samples poorly fit each other as there are lots of solid grey output and output with lines on white background, when the distributions are not the same (see Fig. 1), therefore, the modelling should be presented.

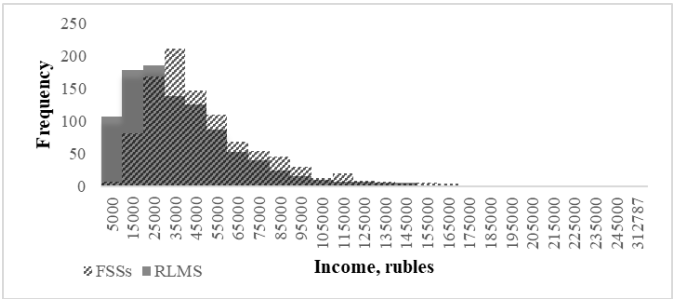

Fig. 1. Initial frequencies of the Federal State Statistics Survey and the RLMS household income samples

To do so the sample of size $(n = 1000)$ was derived from representative household RLMS 2015 survey, the only variable needed for the analysis is household income. Once again, as the stated number is not a crisp one, it is reasonable to present it as a "fuzzy number". The number of income is considered not to be a number, but some set of numbers which form some law of distribution $F(z)$. In accordance with the fuzzy theory, the stated number of income is a fuzzy set Z (following the theoretical base of fuzzy sets, Z is named the core of the fuzzy set, thus, here Z is the crisp number stated in the RLMS survey), defined on a number area X:

$$Z = (\mu_Z^*(x), x), \quad \forall x \in X, \quad (6)$$

where for each element of $\forall x \in X$, $\mu^*_Z(x)$ is set by a formula with $\mu^*_Z(x) \in [0, 1]$.

This function assigns to each value of x a certain number from the interval $[0, 1]$, which is called the membership step. Conversely, a crisp set is just a number. The difference between fuzzy set and a crisp set is presented in Fig. 2. It should be mentioned that the area under the crisp and fuzzy numbers is the same; for the crisp number it is assumed that the width is very small, therefore, the height is rather big (that is why the error in the graph points upwards).

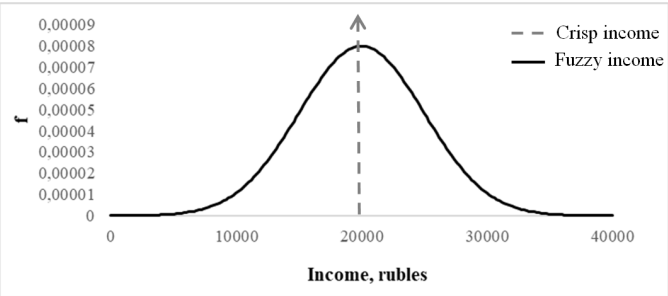

Fig. 2. Graphical representation of fuzzy and crisp sets in terms of households' monthly income of 200,000 rubles

There are several commonly used laws of distribution for modelling fuzzy sets like triangle, rectangle, trapezoid and Gaussian curve.

The Gaussian curve function used in the research looks the following way:

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-Z)^2}{2\sigma^2}}, \tag{7}$$

where $Z = E(x)$, $\sigma^2 = Var(x)$

In the modelling, the mean is the number stated in the survey income (crisp Z). The variance is calculated as follows.

Firstly, as it is assumed that people are rounding numbers the ordinal way, the number (m) of zeros at the end of each income response will be counted. As rounding by five is considered as a special case, these numbers will be remembered.

Secondly, as we are not sure to which degree people round their spending, the two parameters (p – fuzziness of ordinary rounding, p_5 – fuzziness of numbers rounded by five) are added, which allows changing the degree to which the number will be rounded. It was chosen to add separate parameters for numbers, for which ordinary rounding was presented and those ending on five, and in case the parameters are the same, then people do not differentiate in the rounding degree of these two cases. The idea presented above is graphically represented in Fig. 3 in the terms of probability theory.

The deviation value for each income statement will be calculated, following this formula:

$$d = p10^m, \text{ in case the number last non-zero digit is not five,}$$
$$\text{or } d_5 = p_5 10^m, \text{ if there five is the last non-zero digit.}$$

The calculated deviation can be graphically interpreted as follows (Fig. 4). By adding or subtracting the deviation from the stated fuzzy number, the interval of where the exact number lays can be found.

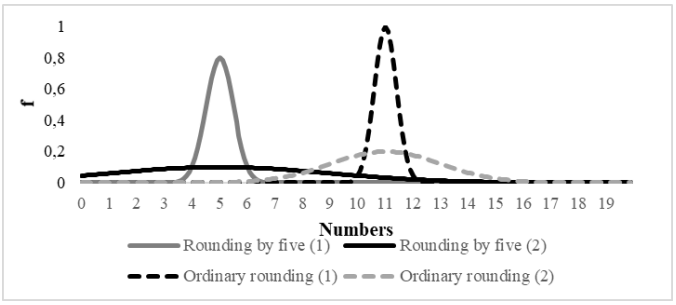

Fig. 3. The range of rounding using the Gaussian curves for tens and five rounding in terms of households' monthly income, where 1 – small effect of fuzziness, 2 – high effect of fuzziness

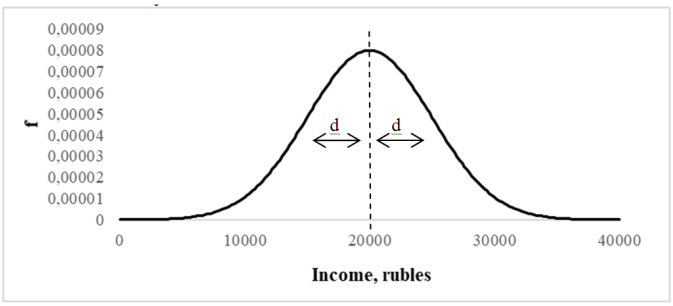

Fig. 4. Graphical representation of deviation modelling for fuzzy variables of households' income of 200,000 rubles

Now all the responses by households can be presented as normally distributed numbers. The next valuable part of the research is to optimize the parameters.

Therefore, the 25 intervals were chosen to find the frequencies and compare them with the step 10000 and the last interval has an open border. Let's look at how the frequencies of the RLMS 2015 responses are calculated. For i^{th} $(i = 1, 2 \dots 25)$ probability density function probability P_i to fit in each interval is calculated:

$$P_i = F(b_i - Z_j) - F(a_i - Z_j), \quad i = 1 \dots 24, \quad j = 1 \dots n$$
$$P_{25} = 1 - F(a_{25} - Z_j)$$

Then all probabilities are summarized for each interval, this values (w_i) are the frequencies for each interval.

Before calculating the frequency of the Federal State Statistics Survey, the shift of mean should be made to obtain the same location and to omit the differences between two samples. This is made due to the fact that the means can be different as for bias of under and misreporting income (e. g. shadow income), which may be corrected by shifting means of two datasets. To calculate the mean of RLMS 2015 the formula of weighted average is used:

$$\bar{Z} = \frac{\sum_{i=1}^{25} w_i \frac{a_i + b_i}{2}}{n}, \tag{8}$$

where w_i – frequencies (as summary of probabilities of fuzzy variables) for each interval.

The weighted average income for sample from RLMS 2015 households' responses is 44,507 rubles, while for the Federal State Statistics survey sample the ordinary formula for finding average income ($\widetilde{x}$) is used.

The average income for the Federal State Statistics Survey sample is 50,857.78 rubles and the difference between real income and income stated with an error is 5,786.78 rubles, therefore, this number should be substituted from each observation of the Federal State Statistics Survey. The next step is to compare the obtained frequencies of RLMS 2015 household responses (1000 random sample of representative dataset) and the Federal State Statistics Survey samples. To do so, the Smirnov criteria will be used. The null hypothesis to be tested is as follows:

H_0: the RLMS 2015 household responses sample's distribution with the correction is conformed to the Federal State Statistics Survey one. The Smirnov criterion is intended to test the hypothesis of the coincidence of the distribution laws of two or more general populations from the grouped samples extracted from these sets, and the division of the ranges of the investigated random variables into grouping intervals in all samples is carried out in the same way as the Pearson criteria.

Let's suppose there are s intervals common for all samples, ν_{ij} – number of elements of sample i in the interval j ($i = 1 \dots l$, $j = 1 \dots s$) and $n = \sum_{i=1}^{j} \sum_{j=1}^{s} v_{ij}$. For the case of two samples the following formula is used:

$$Y^{(n)} = \sum_{j=1}^{s} \frac{n_1 n_2 \left(\frac{\nu_{1i}}{n_1} - \frac{\nu_{2i}}{n_2}\right)^2}{\nu_{1j} + \nu_{2j}}. \quad (9)$$

Smirnov N. V. proved that with an unlimited increase in the volumes of all the samples and under the conditions of validity of the hypothesis being tested, the statistics described above tends on the distribution χ^2 with degrees of freedom equal to $(l-1, s-1)$. If $Y^{(n)} \leq \chi^2_{1-q}(l-1, s-1)$ the hypothesis is rejected at the chosen level of significance and vice versa for a different sign [6].

To make the procedure more robust, the Winsor approach was used for lowest and highest intervals respectively [7].

Returning to the parameters, by changing them the values of RLMS frequencies will also change, therefore, the χ^2 will change. The aim is to minimize χ^2 observed by the following restrictions: the $0 \leq p \leq 1$ as if there will be more than a unit the following might happen: a person states income as 10,000, and with the prediction of rounding used, the income of a person is between 0 and 20,000, which does not bring enough statistical accuracy; the $0 \leq p_5 \leq 5$ the upper limit for the parameter of fuzziness in rounding by five is higher, thus the logic is the same, it only matters that five is smaller than ten.

Using the optimizing procedure, the optimal parameters were found. They are $p = 0.025$, $p_5 = 2$ for income statements ending with zeros and fives respectively. If the number is rounded by 5, then the exact number may be within $\pm 2 \cdot 10^m$; if the number is rounded by 10, the exact number lays within $\pm 0.025 \cdot 10^m$ deviation. This shows that the rounding by five has greater effect than the ordinary rounding.

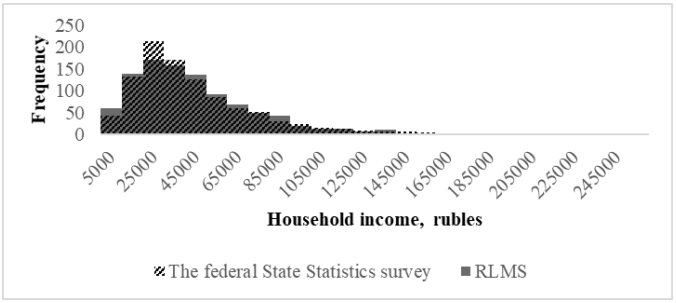

Fig. 5. Frequencies of the Federal State Statistics Survey 2015 and the RLMS 2015 household income samples.

With these parameters $p = 0.025, p_5 = 2$ the $\chi^2_{observed}$ is 12.9, and $\chi^2_{critical}$ is 26.216. The hypothesis is that the distributions conformity is not rejected at the significance level 0.05. Once again, the hypothesis tested was not that the distributions are the same, but that they conform. Now it is clearly seen that the changes made have significant influence on the way distributions look.

V. Histogram vs. Kernel Density Estimate

The histogram presented above is a not smooth non-parametric method which is used to evaluate probability density function of the samples. Thus, the method is rather inaccurate, as it displays several intervals which may hide the shifts of the distributions as well as it makes it difficult to differentiate the function as it is a discontinuous function. Furthermore, the histogram may present very different results by changing the step of the interval. Therefore, the better representation will be provided to obtain more accurate results. To do so, the Kernel density estimate method will be used. As for the Kernel density function the Gaussian curve is used:

$$K(u) = (2\pi)^{-i/2} exp\Big(\frac{-u^2}{2}\Big)$$

Before proceeding to the analysis, the samples of the Federal State Statistics survey and the RLMS one were standardized. The shifted sample of the Federal State Statistics survey was standardized the following way:

$$x_{st\ i} = \frac{x_i - \bar{x}}{\sigma}, \quad i = 1 \dots 1000,$$

where $\bar{x} = \frac{1}{1000} \sum_{j=1}^{1000} x_j$, $\sigma^2 = \frac{1}{1000} \sum_{j=1}^{1000} (x_j - \bar{x})^2$.

After the calculation, $\bar{x} = 45157.12$, $\sigma = 33167.89$ for the Federal State Statistics survey sample.

As the RLMS income responses are presented as fuzzy sets, therefore the variance should include the effect of the fuzziness (while the mean is calculated the same way). The overall variance is a sum of two different variances: between

and within ones. Intergroup variance is calculated by the ordinary formula using the sample of household income from representative household RLMS 2015 dataset as crisp data. The formulas for calculation are as follows:

$$\widetilde{\sigma}^2_{between} = \frac{1}{1000}\sum_{i=1}^{1000}(x_i - \bar{x})^2$$

$$\widetilde{\sigma}^2_{within} = \frac{1}{1000}\sum_{i=1}^{1000}\widetilde{\sigma}^2_i$$

Since the variance of each fuzzy set has been calculated previously, then the intragroup group variance is the average of those calculated ones (it should be mentioned that by this point, the parameters of estimation should be optimized, as they influence the intragroup variance). Therefore, the overall variance is:

$$\widetilde{\sigma}^2_{overall} = \widetilde{\sigma}^2_{between} + \widetilde{\sigma}^2_{within}$$

After the calculation, the following was found $\bar{x} = 45431.6$, $\widetilde{\sigma} = 33617.38$.

for the RLMS survey income responses. The Gaussian Kernel function was used for both samples, resulting in matrix $n \times n$ cells for each sample. Thus, as the matrices are very large, they will not be presented in the papers. The idea of calculation is the same for RLMS households' responses (soft data) as well.

The graph shows that the optimized responses of the RLMS sample are close to the ones of the Federal State Statistics survey. After constructing the matrix, the h value is optimized. For this case the optimal h value was found to be 0.1 to avoid both under smoothing and over smoothing.

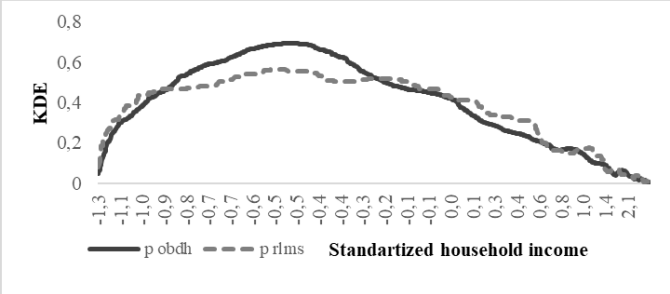

Fig. 6. The KDE for the optimized Federal State Statistics survey and RLMS survey income responses by households in 2015.

VI. Conclusion

The research presented in the paper has shown the existence of the errors in evaluation between crisp and fuzzy data. The shift between crisp and fuzzy datasets of household income was proved, based on the theory of rounding numbers by the respondents. Using the concept of fuzzy variables, the rounding by five and by ten was modelled. As expected, it was calculated that rounding by five is greater than rounding by ten. These findings were presented and the algorithm which was established in the research helped to improve the conformity level of two datasets, when not linked data is used, which is critical for further researches not only on this topic, but on any topic which required usage of fuzzy data in modelling and dealing with stochastic regressors.

The graph shows that there is quite good conformity of the kernel density estimates of the probability density functions, therefore, the changes made are significant for the research and should be considered in the further analysis.

Acknowledgements

This work was supported by a grant of Russian Foundation for Basic Research #18-010-00564 Modern Tendencies and Social and Economic Consequences of Digital Technologies Development in Russia "Russia Longitudinal Monitoring survey, RLMS-HSE", conducted by National Research University "Higher School of Economics" and OOO "Demoscope" together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology of the Federal Center of Theoretical and Applied Sociology of the Russian Academy of Sciences. (RLMS-HSE web sites: http://www.cpc.unc.edu/projects/rlms-hse, http://www.hse.ru/org/hse/rlms)

References

[1] A. N. Borisov, O. A. Krumberg, and I. P. Fedorov, "Decision-making based on fuzzy models," *Zinatne*, 1990.

[2] M. E. Borus, "Response error and questioning technique in surveys of earnings information." *Journal of the American Statistical Association*, vol. 65, no. 330, pp. 566–575, Jun. 1970.

[3] M. David, "The validity of income reported by a sample of families who received welfare assistance during 1959," *Journal of the American Statistical Association*, vol. 57, no. 299, pp. 688–685, Sep. 1962.

[4] C. Kim and C. R. Tamborini, "Response error in earnings: An analysis of the survey of income and program participation matched with administrative data," *Sociological Methods & Research*, vol. 43, no. 1, pp. 39–72, 2014.

[5] M. Verbeek, *A Guide to Modern Econometrics*, 5th ed. Wiley, 2017.

[6] S. A. Aivazian and V. S. Mkhitarian, *Probability Theory and Applied Statistics*, 2nd ed. Moscow: Unity, 2001, vol. 2.

[7] V. S. Mkhitaryan, M. Y. Arkhipova, T. A. Dubrova, Y. N. Mironkina, and V. P. Sirotin, *Data analysis*. Moscow: Yurayt Publishing House, 2018.