Estimation based on tendencies of non-response

Roland Szilágyi
Ph.D associate professor
University of Miskolc, Faculty of Economics

DESCRIPTION OF THE INITIAL DATA STRUCTURE

In this study I concentrate on one of the most important type of non-sampling error types, the non-response error, and within this, partial or item-level non-response. With the help of different analytical methods I investigated the impact of the substitution of refusals on the results of descriptive models. Non-responses of various degrees were generated using the database of the Household Budget Survey of 2005 provided by the Hungarian Central Statistical Office (KSH) and making an estimate of the average consumption expenses of households. Taking into consideration that people with higher income levels tend to be more reluctant to answer questions concerning income and consumption (as shown in the following empirical studies [4], [3], [2]), I have assumed that in this case, as well, they make up a significant proportion on non-respondents.

I took a sample consisting of 900 elements of the population of households according to the rules of uniform stratification where the stratification criterion was household income (monetary net annual income). This way, the number of strata was not naturally determined by the number of variants of the stratification criteria but was separated artificially, with the help of deciles. The procedure resulted in 10 strata of equal size in the population of households.

From the sample, as an experiment, in the case of certain elements I deleted the values related to consumption, artificially generating non-response based on different systems. (It should be noted that I solely concentrated on the analysis and impact of item non-response, without investigating the significant problem when unit non-response or failure erodes the database.)

This simulation was useful for various reasons. First of all because the original database provided the population expected value – the most effective and accurate estimation of which was the fundamental aim. Secondly, the differences between the complete sample and the sample truncated by failures , as well as the differences and biases between the estimation results derived from them became clearly visible.

During the simulations for the time being I investigated those cases when the non-respondents are from the "more affluent" strata of households. (Of course I ignored the possibility of random non-response, as in such cases it would not have an effect on the final result of the estimation.)

Relying on this presumption based on practical experience, I selected the non-respondents from the sample arranged in decreasing order according to consumption expenses, starting with a non-response ratio of 10%, which is considered quite favourable. According to this, the 10% non-response ratio meant that households which fall into the uppermost consumer decile did not respond to questions about their consumption expenses while they did respond to all the other questions.

DOI: 10.26649/musci.2015.093

Next I increased the ratio of non-respondents by 5 percent step by step (which means that I deleted the values for consumption expenses bit by bit every 5%) until a 50% non-response rate was reached. I did not investigate the rate of non-response beyond 50%.

EXAMINATION OF THE NON-RESPONSE TENDENCY

Among the steps taken to eliminate bias caused by non-response, the identification of tendencies has an important role. It is worth investigating the difference between the examined tendencies of respondents and non-respondents. Of course before all this it has to be determined whether there are any tendencies at all.

From the data of the selected sample, I attempted to decrease the bias caused by non-responses on the basis of the estimation of the average consumption expense with the help of the exploration of non-response tendencies. My aim is to decrease the consumption expenses' bias of estimation.

Based on my experience, the stratification criterion (income) is in a stochastic relation with the consumption data – this is what the value of the coefficient of correlation (r=0.719**) indicates. Thus it is presumable that as long as the reason for non-response is to conceal income data, then it can be related to the concealment of consumption data.

It is obvious that application of complementary information helps with reducing the degree of error. See the following studies [1], [5], [6]. Researchers do not always have the opportunity to use such information, so inner (within-sample) information should be utilised as much as possible. The tendency in responding groups (in the case of detailed grouping of sample items) can be projected to the entire sample, to the non-respondents. By modelling the tendencies we can reduce the bias caused by non-response.

Group formation

The fundamental aim of forming groups, therefore, is to decrease the effect of extremist values and to identify non-respondents as accurately as possible. To ensure the success of the procedure, appropriate groups need to be formed based on the criteria which are in a stochastic relation with the criterion investigated in the sample and which –generate non-response. To find such a criterion may be difficult but numerous methods can be of help with the identification of non-respondents. It is not necessary that the number of groups coincide with the number of stratification criterion variants.

The formation of groups – which occurs along a variable that is in stochastic relation with the potential and realised non-response – is arbitrary; the groups may be deciles or centiles but, of course, even the strata themselves. In the following example the formation of groups is not problematic, as I clearly consider non-response to be dependent on the income.

Mapping of tendencies

After the formation of groups I determined the average value of consumption expenses in each group (income category) and next I examined the tendencies in group averages. The groups are equal to the income deciles, so they move from the households with lower income to those with higher income. On this basis it can be presumed that consumption expenses of households that belong to the groups with higher incomes will be higher.

Therefore, if the group averages show some kind of tendency then it is describable with the help of some kind of mathematical function. In the case of the selected sample (with complete response), the changes in the average expenses of each household decile can be described with an exponential function, with an explanatory power of 96.5%. If the sampling is representative then the averages of the appropriate groups of the population also draw a similar (in this case, exponential) curve.

During my investigation I have worked with non-responses of different levels in such a way that the non-respondents were always from the upper deciles. In this case, the strata with lower income count as quasi-complete respondents, which means that the tendency uncovered in our data contains less bias than the inner tendencies of an imputated or weighted sample. Using this, I estimated the averages of the upper groups, extrapolating the tendencies found in the data of respondents. The main data of the exponential functions built on the data of respondents are given in Table 1.

Table 1: Parameters and explanatory power of the exponential functions built on the data of the respondents

level of non-response	b_j	a_j	R_j^2
upper 10% non-response	1.1659	664,262.24	0.952
upper 15% non-response	1.1660	664,215.84	0.952
upper 20% non-response	1.1743	648,690.46	0.944
upper 25% non-response	1.1774	643,453.10	0.948
upper 30% non-response	1.1839	632,936.05	0.932
upper 35% non-response	1.1908	623,265.03	0.941
upper 40% non-response	1.1981	613,166.01	0.917
upper 45% non-response	1.2020	608,514.30	0.922
upper 50% non-response	1.2132	595,488.81	0.886

As the non-response is systematic, the explanatory power of the descriptive functions describing the non-respondents' tendencies can be considered extremely good at every non-response level.

Of course the tendencies of different non-response levels differ from each other, sometimes under- or overestimating the population parameters. Figure 1 represents the data estimated at the different non-response levels.

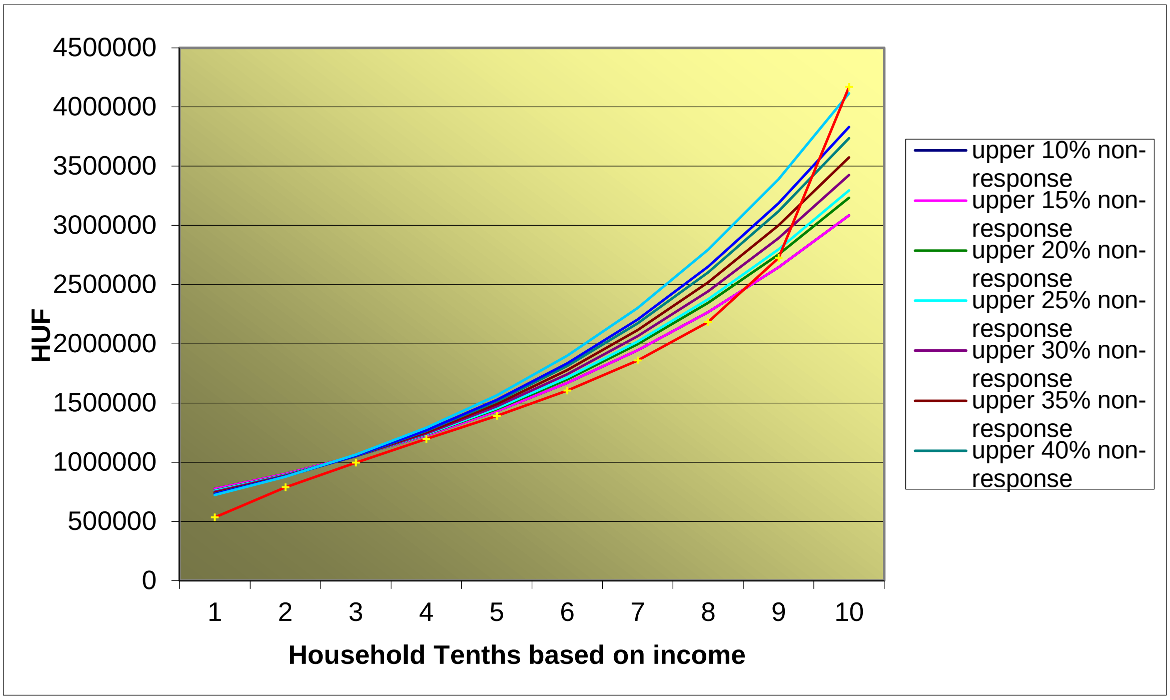

Figure 1: The estimated values of the average consumption expenses of households in the various income deciles, at different levels.

The exponential functions in the ninth and tenth deciles clearly underestimate the average of consumption expenses, while in the case of the other deciles, a minor overestimation can be seen.

When determining the final results, the overestimation bias of the first 5 groups should not be counted because it is likely that these groups represent a total response rate, so the estimation of these groups is not needed. This means that their actual data can be used.

THE ESTIMATE MODEL OF WEIGHTED TENDENCIES

Figure 1 shows that the function generated with a 50% non-response ratio overestimates the population value even in the 10th decile. However, it should be considered that this function has the weakest explanatory force. The explanatory force of the functions at lower levels of non-response are stronger; however, the functions significantly underestimate the actual values in the higher deciles.

Taking all these facts into account, the estimation of the expected values of each income deciles were defined as the average estimated value of the estimated values of the functions weighted by the proportion of the corrected explanatory force of the functions. Thus the functions that possess a weaker explanatory force receive a proportionally lower weight when defining the estimated value. This means that the functions generated at lower non-response levels – that were defined taking more strata into consideration and are thus more accurate – are represented by a greater weight when defining the final result.

The next step was to define the change of the average estimated values between each group, that is, the growth rate of the average estimated value while extending the income category. Within the case of the non-response of a given upper stratum an approximate, though not unbiased estimation of the average consumption expenditure can be given by expanding the respondent's data with the growth rate

The scale between the generated non-response measures can be changed with the condition that the scale must be proportional to the size of the non-responding groups. For example with a response ratio of 70% additional measures with a non-response ratio of 35%, 40%, 45% and 50% can be generated to ensure that the result of the estimations will be derived from weighing five different functions. Table 2 shows the process of the estimation and the application of the model at a non-response ratio of 30%.

Columns 3-7 of Table 2 show the average consumption expenditures at different rates of non-response estimated by the exponential functions. (The parameters of the functions can be found in Table 1). The adjusted R^2 values of the functions are provided in the row last but one (colored yellow), based on which the weights of the functions are calculated in the last row. In each income decile the average estimated consumption values in the 8th column can be calculated by using the estimated values in Columns 3-7 and the appropriate weights. From this data the relative rate of growth between the deciles can be easily calculated. (It must be noted that the rate of growth results in a perfectly fitting exponential function due to the weighting of the exponential functions.)

As 30% of the respondents did not answer and I assume that non-response is dependent on the income, it is clear that we should focus on the estimation of the last three deciles. Therefore the first seven rows of the last column (the lowest income deciles) containing the final estimated values are identical to the values in the second column calculated at a 100% response rate. By calculating with the actual values instead of the estimates in the seven lowest deciles the errors of the estimation method can be significantly reduced. The actual estimation in this case starts from the 7th decile by multiplying the decile's value by the respective growth rate. Thus at a non-response ratio of 30% the average value of the consumption expenditures – taking the tendencies of the non-respondents into consideration – is estimated to be 1,767,559 HUF. In order to analyse the data in Table 2 we should note that the population parameter I tried to estimate is known: 1,744,633 HUF.

Table 2: The estimate model of weighted tendencies at a real non-response level of 30%

Income deciles	***Consumption at overall response***	***30%NR***	***35%NR***	***40%NR***	***45%NR***	***50%NR***	***Average estimated function value***	***Estimated means of deciles***
1	*650,298*	749,355	742,179	734,640	731,450	722,461	736,179	*650,298*
2	*916,414*	887,189	883,780	880,178	879,221	876,507	881,437	*916,414*
3	*1,170,972*	1,050,374	1,052,398	1,054,549	1,056,846	1,063,400	1,055,428	*1,170,972*
4	*1,418,208*	1,243,575	1,253,186	1,263,464	1,270,355	1,290,142	1,263,851	*1,418,208*
5	*1,374,019*	1,472,313	1,492,283	1,513,767	1,526,999	1,565,231	1,513,536	*1,374,019*
6	*1,739,427*	1,743,123	1,776,998	1,813,658	1,835,492	1,898,976	1,812,674	*1,739,427*
7	*1,944,533*	2,063,746	2,116,034	2,172,959	2,206,308	2,303,883	2,171,083	*1,944,533*
8	*2,214,489*	2,443,342	2,519,755	2,603,441	2,652,038	2,795,126	2,600,538	2,329,175

9	*2,475,128*	2,892,759	3,000,503	3,119,205	3,187,817	3,391,114	3,115,157	2,790,094
10	*3,291,167*	3,424,840	3,572,973	3,737,147	3,831,836	4,114,181	3,731,872	3,342,456
mean	1,719,465	1,797,062	1,841,009	1,889,301	1,917,836	2,002,102	1,888,176	1,767,559
$\overline{R}^2$ (Rsq_adj)		0.9317	0.9408	0.9167	0.9218	0.8856		
weights of the functions	1	0.20269	0.20467	0.19943	0.20054	0.19266		

It can be stated that the model is becoming more biased with the increase in the realised non-response level. However, at a low rate of non-response it underestimates the population parameter.

The estimate model of weighted tendencies can be applied under the following conditions:

- there are criteria that determine non-response;
- based on these criteria the population can be grouped into groups (preferably of the same size);
- there exists a mathematical function that describes group tendencies significantly and reliably; and
- response ratio is larger than 50%.

If these conditions are met the model can provide a relatively good approximate value of average consumption expenses. Its greatest asset is that it can compensate for the significant underestimation which occurs when using the imputation and transweighing methods.

The disadvantage of the cold deck methods based on transweighting and imputation are that they are not suitable for estimation in cases when complete strata are left out as a result of non-response. In such cases – especially when estimating criteria with asymmetrical distribution – the information of the missing stratum is completely lost.

Although the model can result in overestimation, it is relatively small compared to its ability to deal with the bias arsing from non-response. Table 3 shows the effect of the model on reducing the bias caused by non-response.

Table 3. Comparison of results from non-response compensation methods by population parameter

level of non-response	***unweighted mean consumption expense***		***estimate model of weighted tendencies***	
	HUF	***percentage of expected value***	***HUF***	***percentage of expected value***
upper 10%	1,475,398	84.57%	1,684,073	96.53%
upper 15%	1,389,274	79.63%	1,719,168	98.54%
upper 20%	1,316,725	75.47%	1,721,414	98.67%
upper 25%	1,253,037	71.82%	1,762,698	101.04%
upper 30%	1,193,976	68.44%	1,767,559	101.31%
upper 35%	1,138,382	65.25%	1,850,261	106.05%
upper 40%	1,084,923	62.19%	1,858,925	106.55%

upper 45%	1,032,088	59.16%	1,806,398	103.54%
upper 50%	979,840	56.16%	1,826.144	104.67%

Average consumption expenses counted by weighting or omitting non-responses can show as much as a 40% bias at higher non-response levels. At the same time the estimate model of weighted tendencies shows only a 5% bias. Nevertheless, the combined application of different methods is advisable, as we must bear in mind that the given sample is only one variation of the sampling plan, and the examined criterion is a probability variable that can be influenced by several factors.

Analysis of different non-responses

For checking the operation of the model cases were examined in which non-response is not exclusively one-sided. According to the so-far simplified assumptions only the respondents with higher income were considered non-respondents and thus only cases where decrease occurs only at one side (from the maximal values) of the magnitude of the sample were examined. As mentioned before, this – according to research based on sampling – is not entirely realistic. Therefore it is necessary to examine cases where non-response occurs on both sides of the sample. Similarly to the previous problem, the non-response can take any value, but it is also assumed that its value does not exceed 50%. In order to present the calculations the levels of non-response was increased by a 10% scale at either side of the sample. Even in this case a large number of combinations are possible. In order to limit the number of possible combinations and still be able to produce illustrative results the non-response possibility of respondents with higher income is set at a maximum of 30% and of respondents with lower income at a maximum of 20%. Thus a 50% rate of non-response can be achieved, and two further possibilities of a 40% and 10% rate of non-response, and three possibilities of a 20% and 30% rate of non-response can be examined.

Table 4: Results of the estimate model of weighted tendencies in estimating the total consumption in different non-response cases

level of non-response	***estimation at non-response***		***estimate model of weighted tendencies***	
	estimated total consumption (HUF)	***percentage of expected value***	***estimated total consumption (HUF))***	***percentage of expected value***
nv_U10	1.544.832	88,55%	1,675,538	96.04%
nv_U20	1.428.545	81,88%	1,693,421	97.06%
nv_U30	1.316.267	75,45%	1,706,211	97.80%
nv_L10	1.838.262	105,37%	1,734,666	99.43%
nv_L20	1.953.493	111,97%	1,756,714	100.69%
nv_U10L10	1.656.649	94,96%	1,687,891	96.75%
nv_U10L20	1.762.396	101,02%	1,708,690	97.94%
nv_U20L10	1.539.723	88,25%	1,698,880	97.38%

nv_U20L20	1.643.608	94,21%	1,715,116	98.31%
nv_U30L10	1.427.262	81,81%	1,699,990	97.44%
nv_U30L20	1.529.432	87,66%	1,708,190	97.91%

The abbreviations used in Table 4 require some explanation. „U” stands for the non-response rate of the upper income deciles, while „L” stands for the non-response rate of the lower income deciles. According to this nv_U20L10 means that the upper 20% and the lower 10% of the sample did not respond thereby a non-response rate of 30% was realized.

The third column of the table shows the extent of bias at different levels of non-response. As non-response from both lower and upper levels was calculated, results quite close to the expected values can be achieved by finding (or not finding) certain rates. If the effect of non-response is not taken into account the results fluctuate between under- and overestimation, from 75% and 112%. However, the analysis cannot be left to chance, hoping that the missing high and low values will balance each other out.

The last column shows that by using the estimate model of weighted tendencies the estimated values approximate the expected value of the population, underestimating by only a few percentage points. By including non-response from both the upper and lower values the model becomes more balanced. Due to this the symmetric bias resulting from the increase of non-response level seen in the last column of Table 3 disappears. This confirms that the model can be flexibly applied to different levels of non-response and for different scales.

Acknowledgement: This research was carried out in the framework of the Center of Excellence of Sustainable Resource Management at the University of Miskolc.

BIBLIOGRAPHY

[1] ESTEVAO V. M. – SÄRNDAL C. E.: **The ten cases of auxiliary information for calibration in two-phase sampling**; Journal of Official Statistics, Vol. 18, No. 2, 2002, pp. 233–255.

[2] HAVASI Éva – SCHNELL Lászlóné: **Az 1996-os jövedelmi felvételre nem válaszoló háztartások – A megtagadások természete, a megtagadók sajátosságai**; Központi Statisztikai Hivatal. Budapest. 1996.

[3] HAVASI Éva: **Válaszmegtagadó háztartások**; Statisztikai Szemle 1997. 10 sz. pp. 831-843.

[4] KESZTHELYINÉ Rédei Mária: **A lakossági jövedelmek mérésének megbízhatóbb módszere**; Statisztikai Szemle, 2006. 84. évf. 5-6. szám pp. 518-551.

[5] Roy D. – SAFIQUZZAMAN Md.: **Variance estimation by Jackknife method under two-phase complex survey design**; Journal of Official Statistics, Vol. 22, No. 1, 2006, pp. 35–51.

[6] SÄRNDAL C. E. – LUNDSTRÖM S.: **Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator**; Journal of Official Statistics, Vol. 24, No. 2, 2008, pp. 167–191.