Advances in Methodology, Data Analysis, and Statistics
Anuška Ferligoj (Editor), Metodološki zvezki, 14
Ljubljana: FDV, 1998

Non-response Patterns from the Matched Data

Vasja Vehovar[1] and Metka Zaletel[2]

Abstract

The estimates based on survey samples can be biased due to various reasons. Besides observational errors (interviewers, respondents, mode of interview, measurement) there also exist errors of non-observation: under-coverage, sampling and non-interviews. The paper deals with non-response bias in official household surveys in Slovenia (Labour Force Survey, Family Budget Survey). The survey data were matched with the administrative records. The non-response patterns were analysed and compared with results from other studies. There were many similarities found in non-response behaviour, but also some features that were country specific. The target variable in Labour Force Survey (unemployment) had almost no non-response bias, but the estimate of the target variable in Family Budget Survey (income) was found to be strongly biased due to non-response.

1 Introduction

Non-response patterns have been studied in many different countries. Our starting question is whether the experience from other countries (Groves, 1993; Foster, Bushnell, 1994) can be generalised to the case of Slovenia. We are, of course, also interested in demographic characteristics of non-respondents and in the consistency of results across different surveys.

Similar matching projects were performed in the USA (Groves et al., 1993), Great Britain (Foster et al., 1994) and Canada (Michaud et al., 1995). In the first two studies, Census data at the level of address were used to define characteristics of non-respondents. The third study used data from longitudinal survey. Other research on this topic generally use the data from sampling frames that are used for comparing respondents and non-respondents (e.g., Bros et al. 1995).

[1] Vasja Vehovar, Faculty of Social Sciences, University of Ljubljana, Kardeljeva ploščad 5, 1000 Ljubljana, Slovenia; e-mail: Vasja.Vehovar@uni-lj.si

[2] Metka Zaletel, Statistical Office of the Republic of Slovenia, Vožarski pot 12, 1000 Ljubljana, Slovenia; e-mail: Metka.Zaletel@stat.sigov.mail.si

Major socio-demographic correlates reviewed in the literature on survey non-response are gender, age, household size and composition, race and ethnicity, socio-economic status (income, education, occupation), tenure, housing structure and urbanity.

Age, household size and urbanity are well-known and strong correlates of non-response (Bros et al. 1995). Highly educated people and people with higher occupational status generally tend to respond better (Groves, 1989; Foster, 1993; Barnes, 1993). On the other hand, results for income are relatively inconsistent (Couper and Groves, 1993). Of course, some of the above mentioned correlates of non-response are inapplicable for Slovenia (e.g.; race and tenure).

In the paper, census and register data were combined with survey data. Since Slovenian official statistics is register oriented we were able to match the survey records directly to the administrative and Census records of the same persons.

The paper starts with the description of the methodology (Section 2), and continues with bivariate and then multivariate data analysis (Section 3). A logit model was also constructed for the impact of the demographic characteristics on the co-operation in the survey and some interactions were discussed. We further examined the impact of individual variables on the decision to co-operate in the survey. At the end (Section 4),. conclusions are made.

2 Description of the surveys

2.1 The methodology

We analysed the following surveys:

- Labour Force Survey 1994 (LFS94);
- Family Budget Survey 1993 (FBS93).

Below we are describing some important issues about analysed surveys.

2.1.1 Sampling procedures

The households were selected through persons from the Central Register of Population of the Republic of Slovenia (CRP). Of course, larger household had larger probability of being selected and afterwards this effect was neutralised by weighting.

Sampling designs for both surveys were similar: they were stratified two (FBS93) or three (LFS94) stage surveys of households with enumeration areas as the final stage of the clusters. Stratification was performed by region and type of enumeration area (rural/urban). The sample size of LFS94 was 3,482 households and the sample size of HBS93 was 4,566 households.

2.1.2 Fieldwork procedures

Fieldwork strategies were similar for both surveys: personal advance letters were sent to selected persons, five follow-ups were performed in both cases, and proxy respondents were allowed.

Fieldwork of LFS94 lasted one month with beginning in May 1994. The average length of an interview was about 20 minutes. The survey mode was face-to-face paper-and-pencil interview. About 130 interviewers and 13 field supervisors were involved in the data collection.

Fieldwork of FBS93 lasted three weeks with beginning in December 1993. The average length of an interview was about 80 minutes. The survey mode was the same as for LFS94. There were 109 interviewers and 30 field supervisors involved.

2.1.3 Non-response rates

The unit non-response rate is defined (Groves, 1989) as a ratio of number of non-respondents and number of eligible units. The refusal rate is defined as a ratio of refusals and contacted units. Other non-response rates (e.g. non-contact rates) were also computed, but no analysis was performed separately for refusals or non-contacted units due to small number of cases.

In surveys analysed, either substitute units (FBS93) or weighting of respondents within the groups (LFS94) were used. We do not discuss these issues here. Some results can be found in Vehovar (1993). In LFS94, the non-response rate was 8.9%, the refusal rate 5.9% and the completion rate 88.1%. For FBS93, we have records about reasons of non-response for 4,293 households included in the first phase of the survey in May 1993. So we can only analyse the non-response rate (24.6%) and the refusal rate (11.9%) based on first wave non-respondents. The completion rate was 71.6%.

2.2 Matching

As external data sources we used the following three data sources:

- 1991 Census of Population, Dwellings, Households and Farm Economies, conducted by the Statistical Office of the Republic of Slovenia in April 1991;
- income tax records file for 1993,
- Register of Unemployed Persons in the Republic of Slovenia (May 1994).

The matching procedure was performed at the level of person selected, so we matched the outcome of the survey (response or non-response) for a certain person with the administrative records of the same person. The matching procedures in other countries generally used the data attached to the address (Foster, 1993), but in our case, we were also able to follow persons that had moved. The key variable that allowed us to combine the samples with the above data bases was the personal identification number (PIN). At the beginning (already attached with the survey data) we had all PINs only for persons included in LFS94. In order to combine the data from FBS93 we first had to combine the survey data with the Central Register of Population (CRP), so that we attached the PINs. Due to many technical problems[3] we successfully combined 77% of persons in FBS93, so the number of eligible persons for matching decreased to 3,484 (from 4,566) for FBS93.

After obtaining PINs we first matched them with Census '91 data. We lost about 2.5% of each sample during the matching with the Census data. Most of them were people who immigrated from abroad and were not included in the Census. We included in the analysis only persons, who had the PIN and for whom the match with Census '91 data was successfully performed. In Table 1 we can observe the loss of data.

Table 1: Loss of data during the matching process

	LFS94	FBS93
loss of data due to ineligibility	118	228
loss of data due to incorrect PINs	244	854
loss of data due to unsuccessful match with Census '91 data	73	88
total loss of data	435	1170

Non-response rates were calculated separately for matched and unmatched cases and no significant differences were detected in any subgroup. This result is reasonable since the main source for unsuccessful matching were the administrative errors in recording the numbers. Such errors cannot be correlated with any characteristic of the respondent. This is especially important for the FBS93 sample where we lost a large portion of data. We therefore assume that the process of losing the un-matched records does not influence the results of our analysis on non-response.

A specific problem appeared in combining the selected persons with the income tax records data. There, a large number of persons have no income of their own and are therefore without any income tax records. For persons who were not in the income tax database (1,022 units in LFS94 and 1,057 in HBS93) we do not know whether they do have income - but the match was not performed properly - or they do not have it at all. In any case we classify them into the lowest income category. In

[3] The problem appeared because only PINs were recorded correctly but not the name and the address, which were the key data in combining with CRP and might have changed for certain persons.

analysing the influence of the height of income on response or refusal we analysed both the data with and without those cases. The analysis was run also on cases without income data.

2.3 Variables analysed

We included the following variables into our analysis:

- **from Census data**:
  - *age of the person (younger than 70 years; 70 years and over);*
  - *education (less than 12 years of education; 12 years of education or more);*
  - *size of the household in which the person lives (single household; larger household);*
  - *family type (family with children; family without children);*
  - *urbanity of settlement in which the person lives (rural areas; urban areas)*
  - *type of housing (individual housing; multiunit building);*

- **from income tax records**:
  - *gross income in 1993 (1st and 2nd class on the tax scale, 3rd - 5th class on the tax scale).*

- **from the Register of Unemployed Persons**:
  - *registered unemployment in May 1994.*

Besides the variables mentioned above we had some other variables at our disposal (e.g., gender of selected persons), but initial analysis showed that they have no impact on non-response rates.

Only after extensive analysis we decided on the categorisation of variables described in brackets above. These are final classifications of the values of variables that are the most appropriate for interpretation. The analysis with more complex scales showed there is no loss in the explanatory power.

It is worth mentioning that the last two variables are extremely important: unemployment is the target variable in LFS94 and income is the target variable in FBS93. We should also notice that some of the above variables refer to persons and others to the household of the selected persons.

3 Analysis

We have analysed the following questions:

- whether the non-response rate is a function of explanatory variables;
- whether there are some interactions of two variables influencing the response rate;
- whether the refusal rate is a function of the observed variables.

3.1 Non-response rates across categories of analysed variables

Let us observe non-response rates by categories. The percentage of refusals in total non-response was calculated as a ratio of the number of refusals and the number of all non-respondents.

We split variables into two sets: variables describing households (Table 2) and variables describing selected persons (Table 3). In both tables, non-response and refusal rates are calculated at the level of the household.

Table 2: Non-response rates in categories of variables describing households

	LFS94			FBS93		
	non-response rate	refusal rate	% of refusals in non-response	non-response rate	refusal rate	% of refusals in non-response
Total	12.0	8.1	68.4	23.7	11.9	59.9
Household size						
single	23.3	9.33	34.15	34.6	17.02	48.19
larger	11.1	8.35	71.53	22.3	16.12	72.40
Housing						
individual housing unit	9.9	7.61	73.26	19.9	14.65	73.82
multiunit building	15.9	10.16	58.45	29.2	19.02	64.65
Urbanity						
rural	8.2	5.91	68.81	16.3	11.87	71.43
urban	15.6	10.76	65.91	28.6	19.52	68.97
Type of family						
family without children	15.2	8.13	72.60	26.8	16.13	74.85
family with children	10.6	9.07	57.02	21.7	16.28	60.30

First of all, one should distinguish between non-response rates in Table 2 and those in Table 3. In Table 2, we are describing the characteristics of a household or a family. On the other hand, in Table 3, we are describing the characteristics of a selected person in the household. It is necessary to repeat that non-response occurred at the level of the household. For example, the non-response rate 11.5% among persons under 70 years in Table 3 means that 11.5% of households with a person[4] under 70 years did not respond.

Comparing non-response and refusal rates in both surveys, we can notice that both rates are much higher in FBS93 than in LFS94. This is a general result found also in other European countries (Heer, 1992). This is caused by a higher respondent

[4] And this was the person included in the sample.

burden in Family Budget Surveys. But comparing the percentage of refusals in the whole non-response count, we notice the same pattern in LFS94 and FBS93. There are a few exceptions to this rule: the percentage of refusals among single person households is 34% in LFS94 and 48% in FBS93.

Table 3: Non-response rates in categories of variables describing selected persons

	LFS94			FBS93		
	non-response rate	refusal rate	% of non-response	non-response rate	refusal rate	% of non-response
Age						
under 70 years	11.5	8.05	67.74	23.4	16.74	71.78
70 years and over	17.4	11.44	62.00	25.7	14.11	54.90
Education						
less than high school	11.9	8.37	67.43	21.5	14.16	68.95
high school or more	12.3	8.48	65.77	30.7	20.64	73.43
Registered unemployment in May 1994						
unemployed	11.6	6.61	55.56	23.8	15.28	61.11
other	12.1	8.57	67.88	23.7	16.51	70.23
Income						
not known	12.1	7.77	60.90	27.6	16.47	59.47
1st and 2nd class	11.8	8.85	72.81	20.3	15.70	77.35
3rd -5th class	14.4	10.40	69.23	33.9	22.81	67.53

We have already mentioned that income and unemployment are the target variables in each survey. We can observe that unemployment has no impact on the total non-response rate, but it is worth mentioning that unemployed persons refuse to co-operate less often than other persons do, but are harder to contact.

On the other hand, we can observe a strong impact of income on non-response and refusal rates in FBS93. So income in LFS94 has almost no influence on refusal or non-contacts of households, but it is very important in FBS93

The largest differences in the non-response rate can be noticed in the level of urbanisation, type of housing and household size. Differences are statistically significant in both surveys. The differences in other variables, e.g., family type, education and age are statistically significant in only one survey or not at all, so we cannot generalise them to other surveys. Below we describe the differences between non-response rates and refusal rates for each of the independent variables:

- The refusal rate in **single households** is almost equal to the refusal rate in larger households (9.3% vs. 8.3% in LFS94 and 17.0% vs. 16.1% in FBS93). However, non-response rates are much higher in single households. It is clear that single households are significantly more absent. With the increase in the number of visits of interviewers (from 5 to e.g., 8 visits) we may significantly decrease non-response.

- Households in **multiunit buildings** have a higher refusal rate than households in individual houses (10.1% vs. 7.6% in LFS94; 14.6% vs. 19.0% in FBS93) and a higher non-response rate (15.9% vs. 9.9% in LFS94; 29.2% vs. 19.9% in FBS93). The differences are smaller than in the case of household size.
- The non-response rate of households living in **urban areas** is almost twice as large as the non-response rate of households living in rural areas (15.6% vs. 8.2% in LFS94, 28.6% vs. 16.3% in FBS93). We can observe approximately the same relation in refusal rates.
- In the case persons of **over 70 years of age,** we can not generalise results from one survey to the other. The result of FBS93 tells us that households with elderly are more absent than other, but the result of LFS94 can not justify it.
- Observing education, there is no significant difference in non-response rates in LFS94. The difference in FBS93 is larger (30.7% vs. 21.5%). Again, we can not generalise the results.

To summarise the above findings, we can state that with the increase in the number of visits of interviewers we may significantly decrease the level of non-response at least in certain categories, e.g. in single households and households living in multiunit buildings.

3.2 Interactions

a) It is shown in further analysis that there is an interaction between urbanity, education and response in the LFS94 database. It is the only interaction discovered that is statistically significant in LFS94 data.

Table 4: Non-response rates according to urbanity and education in LFS94

	EDUCATION	
URBANICITY	less than 12 years	12 years and over
Rural areas	9.5%	4.5%
Urban areas	15%	16.0%

In urban areas there is no difference between non-response rates with respect to education but in rural areas the education makes a large difference: educated persons responded more often than less educated ones. This interaction has not been reported in other countries. It has either not been observed or it is specific for Slovenia.

b) The only statistically significant interaction in FBS93 data was the interaction between single households and urbanity. In the bivariate table of associations (Table 2) we observed that single households have lower response rates. However, in Table

5 we see a distinctly higher non-response rate among single households in urban settlements. In rural settlements household size makes no difference to the non-response rate. Since both variables - urbanity and size of household - are proved to be very strong correlates of non-response (Groves, 1993; Bros et al., 1995), a similar interaction is to be expected also in other countries.

Table 5: Non-response rates according to urbanity and household size in FBS93

	HOUSEHOLD HOUSE	
URBANICITY	Single	larger
Rural areas	19.0%	16.7%
Urban areas	41.7%	27.3%

3.3 Logit models of predicting response

We constructed standard logit models for predicting response for both surveys:

$$g(y_i)^{(s)} = \beta_0^{(s)} + \sum\nolimits_{j=1}^{8} \beta_j^{(s)} x_{ij}^{(s)}.$$

We use the following notation:

- s=1 for LFS94, *s*=2 for FBS93;
- dependent variable y = response (0=non-response, 1=response),
- *g(y)* is logit-link function,
- independent variables are:
  - $x_1 = 1$ *for persons in urban areas, otherwise 0;*
  - $x_2 = 1$ *for persons with 12 years or more of education (finished secondary school or more), otherwise 0;*
  - $x_3 = 1$ *for persons over 70 years of age, otherwise 0;*
  - $x_4 = 1$ *for persons living in single households, otherwise 0;*
  - $x_5 = 1$ *for persons living in multiunit buildings, otherwise 0;*
  - $x_6 = 1$ *for persons belonging to the 3rd, 4th or 5th category of the tax scale, otherwise 0;*
  - $x_7 = 1$ *for persons in urban areas with 12 years or more of education, otherwise 0 - this variable represents interaction between the urbanity and education;*
  - $x_8 = 1$ *for persons living in single households in urban areas, otherwise 0 - this variable represents interaction between the urbanity and single households.*

Variables that are not statistically significant for individual surveys have empty cells in the table of coefficients for corresponding surveys.

Table 6: Logit models for response in a survey (dependent variable): logit coefficients and their standard errors in parenthesis.

	LFS94		FBS93	
	logit coeff.	stand.error	logit coeff.	stand.error
Intercept	2.3722	(0.1131)	1.7780	(0.0892)
Urban areas	-0.3672	(0.1662)	-0.5927	(0.1111)
Education	0.7394	(0.3048)		
Age over 70 years	-0.3699	(0.1807)	-0.4645	(0.1952)
Single households	-0.6241	(0.1998)		
Multiunit building	-0.2394	(0.1423)		
Income			-0.6098	(0.1521)
Interaction education/urban	-0.8731	(0.3383)		
Interaction single/urban			-0.7072	(0.2575)

The highest logit coefficient in LFS94 is the coefficient for the interaction education/urban. The coefficient (-0.8731) means that persons with higher education living in urban areas have probability 82% of response. Persons living in rural areas in individual houses with higher education of age under 70 years and not living in single households have the highest probability of response: i.e. 96%.

The results for FBS93 are somewhat different from those obtained for LFS94: people living in urban areas, younger than 70 years, with lower income and not living in single person households have the highest probabilities of response (85%). We can find the lowest probability of response (38%) for people living in urban areas, in single households, older than 70 years and with higher income.

Some of the variables that were proved to be very strong correlates of non-response in other countries (Groves, 1993) were important also in our analysis of both surveys: age and urbanity. Other variables are - at least in Slovenia - topic specific and the results cannot be generalised.

3.4 Relative bias of estimates

For variables that we had at our disposal we were able to calculate the relative non-response bias:

$$B = \frac{n_{nr}}{n} \cdot \frac{(y_r - y_{nr})}{y} \cdot 100,$$

where n_{nr} is the number of non-respondents in the sample, n is the total sample size, y_r is the value of the estimator y for respondents, y_{nr} is the value of the estimator y for non-respondents, and y is the value of the estimator in the total sample. The relative bias (B) can be interpreted as the percentage of error when the estimate is based only on respondents. We can observe the results in Table 7.

Table 7: Relative non-response bias in LFS94 and FBS93

		Total	Non-respondents	Respondents	R Bias
% of single households	LFS94	5.9	11.6	5.2	-13.2
	FBS93	7.5	11.2	6.4	-15.1
% of people living in multiunit building	LFS94	32.3	43.2	30.8	-4.6
	FBS93	35.2	44.3	32.5	-8.0
% of people living in urban areas	LFS94	52.2	67.0	50.2	-3.9
	FBS93	56.8	69.7	52.9	-7.0
% of families with children	LFS94	72.1	64.3	73.2	1.5
	FBS93	69.3	64.8	70.7	2.0
% of people older than 70 years	LFS94	9.6	13.8	9.0	-6.1
	FBS93	11.9	12.9	11.6	-2.6
% of people with 12 years or more of education	LFS94	33.1	33.7	33.0	-0.3
	FBS93	15.2	20.4	13.7	-10.5
average household size	LFS94	3.7	3.4	3.8	1.4
	FBS93	3.6	3.3	3.6	2.0
average apartment size	LFS94	75.2	73.1	75.5	0.4
	FBS93	73.3	70.8	74.1	1.1
average number of rooms in the apartment	LFS94	2.9	2.8	3.0	0.7
	FBS93	2.9	2.8	2.9	0.7
% of unemployed persons	LFS94	8.3	8.0	8.4	0.5
	FBS93	6.5	6.6	6.5	-0.1
average income (in 1000 Slovenian Tolars)	LFS94	619	638	617	-0.4
	FBS93	671	774	642	-4.7

Again, let us repeat the results for the target variables - unemployment rate and average income. The estimate of registered unemployment has no non-response bias in LFS94, but the estimate of income is strongly biased in FBS93. The estimates in FBS93 connected to income are underestimated by 5% because of non-response. Bias in the estimate of income in LFS94 is less than 0.5%, and thus almost negligible.

Usual assumption in post-survey non-response adjustment is that the MAR (missing at random) assumption holds and that we will be able to find the covariate that can explain the missing data mechanism. Typical this will result in a construction of weighting classes where we safely presume that non-respondents are similar to the respondents. But here we have an example of a missing data process, which is not

MAR (Little, Rubin, 1988). Asking about income (FBS93) makes people with a certain (high) income more uncooperative, but not asking them (LFS94) keeps them co-operative. In this case, simple adjustment procedures for non-response cannot be used without any precautions anymore.

Among demographic variables the estimate of the percentage of single households is especially biased. The estimates in FBS93 and LFS94 both underestimate the true value for about 15%. Also underestimated is education - we underestimated the percentage of persons with high school (or more) by 10%.

4 Conclusions

We can conclude the following facts about non-response behaviour:

- Certain characteristics of non-respondents in Slovenia are the same as in other countries, e.g., non-respondents are more likely to be older, living in single households and in urban areas (Groves, 1993).
- However, some of the characteristics are country specific, e.g., interaction of education, urbanity and response.
- Relative non-response bias was found to be very high (Rbias=5%) for the income variable, but surprisingly low for the unemployment rate.

We should remind of some methodological problems with the interpretation:

- Some variables describe selected persons (e.g., age) and other the households (e.g., family type).
- There is a certain time lag between census and our surveys and some information about selected persons are out of date. Fortunately, there is no time lag in both key variables we are most interested in - income and unemployment.
- Also, we must be aware of a relatively large loss of data due to no PINs in the case of FBS93 1993, which can create – despite the fact there was no difference detected in non-response rates among matched and lost records - certain problems with the interpretation.

However, we believe that the main conclusions are robust with respect to the above-described limitations.

With demographic characteristics of respondents and non-respondents we can predict the response in household surveys relatively well (70% in logit analysis). Knowing these predictors we can now be more efficient in using the optimal approach for reducing non-response.

It is obvious that a very interesting question was not yet addressed here; i.e. whether weighting and imputation can remove the non-response bias. The extent of such an improvement is definitely the issue of future work on this matched data.

References

[1] Allard, B. and Dufour J. (1994): *Characteristics of Respondents and Non-respondents in the Canadian Labour Force Survey.*

[2] Brehm, J. (1993): *The Phantom Respondent: Opinion Surveys and Political Representation.* Ann Arbor, Michigan: The University of Michigan Press.

[3] Bros, L. et al. (1995): *Non-respondents in a Mail Survey: Who are They? International Perspectives on Non-response.* RR 219. Statistics Finland: Helsinki.

[4] Cochran.W. (1978): *Sampling Techniques.* New York: Wiley.

[5] Couper, M. P. and Groves, R. M. (1993): *Household-Level Effects on Survey Participation.* Paper presented on 4th International Workshop on Household Survey Non-response, Bath 1993.

[6] Fingleton, B. (1984): *Models of Category Count.* Cambridge: Cambridge University Press.

[7] Foster, K. and Bushnell, D. (1994): *Non-response Bias on Government Surveys in Great Britain.* OPCS, Great Britain.

[8] Gilbert N. (1993): *Analysing Tabular Data.* London: UCL Press.

[9] Goyder, J. (1988): *The Silent Minority.* New York: Wiley.

[10] Groves, R. M. (1989): *Survey Errors and Survey Costs.* New York: Wiley.

[11] Groves, R. M. and Couper, M. P. (1993): *Correlates of Non-response in Personal Visit Survey.*

[12] Heer, W.F. et al. (1992): *Response Trends in Europe.* 52th ASA Conference.

[13] Kish, L. (1965): *Survey Sampling.* New York: Wiley.

[14] Little, R.J.A and Rubin, D.B. (1987): *Statistical Analysis with Missing Data.* New York: Wiley.

[15] Vehovar, V. (1993): Field substitutions in Slovene Public Opinion survey. In A. Ferligoj and A. Kramberger (Eds.): *Contributions to Methodology and Statistics.* Metodološki zvezki, **10**, Ljubljana: FDV, 39-66.