Vol. 20, n.° 2, pp. 125-138, 2021
Revista UIS Ingenierías
Página de la revista: revistas.uis.edu.co/index.php/revistauisingenierias

Item-nonresponse in a household sample survey in Colombia

No-respuesta al ítem en una encuesta de hogares en Colombia

Josefa Ramoni-Perazzi [1], Giampaolo Orlandoni-Merli [2], Surendra Prasad-Sinha [3]

[1] Grupo de investigación EMAR y GIDROT, Escuela de Economía y Administración, Universidad Industrial de Santander, Colombia. Orcid: 0000-0002-0493-1940. Email: jramonip@uis.edu.co
[2] Grupo de investigación CIBAS, Facultad de Ciencias Exactas, Naturales y Agropecuarias, Universidad de Santander, Colombia. Orcid: 0000-0002-0031-2659. Email: gorlandoni@udes.edu.co
[3] Instituto de Estadística Aplicada y Computación, Universidad de Los Andes, Venezuela. Email: sinha32@ula.ve

Received: 12 August 2020. Accepted: 17 November 2020. Final version: 8 February 2021.

Abstract

Item nonresponse occurs when sample units do not provide information on a particular variable, problem that may affect the representativeness of the sample and the reliability of the estimates. Efforts to reduce the item nonresponse rate do not necessarily improve the quality of the information. Besides the nonresponse rate, representativeness indicators can be used to measure the quality of the collected data. This paper analyzes the wage nonresponse mechanism of a household survey in Colombia and evaluates the quality of the wage information in two different periods of time (2008:4 and 2017:4). The results show a low but increasing nonresponse rate whose behavior does not seem to be associated with the set of observables considered. This result is of value since the selection of the adequate imputation method relies on the assumptions on the missing data mechanism.

Keywords: item nonresponse; representativeness; household survey; missing data; sample quality; sample weights; R indicator; MCAR; MAR; NMAR; hourly wages; Colombia.

Resumen

La no-respuesta a un ítem se produce cuando las unidades muestrales no proporcionan la información solicitada sobre una variable en particular, problema que puede afectar la representatividad de la muestra y la confiabilidad de las estimaciones. Los esfuerzos para reducir las tasas de no-respuesta no necesariamente mejoran la calidad de la información. Además de la tasa de no-respuesta, es posible utilizar indicadores de representatividad para medir la calidad de los datos recopilados. Este documento analiza el mecanismo de no-respuesta a salarios en una encuesta de hogares en Colombia y evalúa la calidad de la información sobre salarios en dos períodos diferentes (2008: 4 y 2017: 4). Los resultados muestran una tasa de no-respuesta baja pero creciente, cuyo comportamiento no parece estar asociado con el conjunto de observables consideradas. Este resultado es de interés ya que la selección del método de imputación adecuado depende de los supuestos en torno al comportamiento de los datos faltantes.

Palabras clave: no-respuesta al ítem; representatividad; encuesta de hogares; datos faltantes; calidad de la muestra; pesos muestrales; indicador R; MCAR; MAR; NMAR; salarios por hora; Colombia.

ISSN Printed: 1657 - 4583, ISSN Online: 2145 - 8456, **CC BY-ND 4.0**
How to cite: J. Ramoni-Perazzi, G. Orlandoni-Merli, S. Prasad-Sinha, "Item-nonresponse in a household sample survey in Colombia," *Rev. UIS Ing.*, vol. 20, no. 2, pp. 125-138, 2021, doi: 10.18273/revuin.v20n2-2021011

1. Introduction

A common way to reduce the costs of collecting information for large populations and alleviate the response burden is through probability sampling methods. Unfortunately, practical problems may arise in the collecting process, being nonresponse the most common one. In general, nonresponse often compromises surveys, or the information collected for a specific item. It occurs when eligible sample units in the survey do not provide the requested information on some or all items, or their answers are not suitable for the purpose of the study, which reduces the sample size, threatens the reliability of the sample selection mechanism and introduces potential selection bias.

Another consequence of the nonresponse problem is the potential lack of representativeness of the sample, with some groups ending up being under or overrepresented so that no reliable estimates of the population characteristics can be obtained, unless some corrective measures are taken. The response rate is considered an important but insufficient indicator of data quality. However, it is also necessary to determine whether the response can be assumed to be selective or at random. In the first case estimates may be biased, with that bias increasing with the nonresponse rate; in the second case, the precision of the estimates would not be affected.

There exists no standard definition for response rate. Particularly, [1] define it as the proportion of eligible units which provides the required information. By extension, an item nonresponse occurs when a sampled unit fails to provide any or a reliable answer to an item. In any case, the response rate (*RR*) is generally measured as:

$$RR = \frac{n_r}{n_e} \qquad (1)$$

with n_r and n_e indicating the number of respondents and the number of eligible units respectably. As said before, it is also important for the researcher to identify the missing data mechanism since it can help to determine the effect of nonresponse on the estimates:

• In the missing completely at random mechanism (MCAR), the nonresponse is totally independent of both the target variable (y) and all possible auxiliary variables (x) which are completely observed, so that $\Pr(r \mid y) = \Pr(\text{response})$ where response denotes an indicator variable which is equal to one for response and cero for nonresponse. In this case, the nonresponse is considered not selective so that no corrective measures are required since it does not generate biased estimates and only affects the efficiency.

• In the missing at random mechanism (MAR), the nonresponse is directly associated with x; however, since y and x are related to each other, there exists an indirect relationship between the observed y (y_{obs}) and nonresponse behavior, so that $\Pr(\text{response} \mid y) = \Pr(y_{obs})$ In this selective missing data mechanism, the causes of nonresponse are completely identified, so that it can be corrected based on the available information to avoid biased results. In other words, the probability that y is missing does not depend on the value of y but on a set of variables x.

• Finally, when the nonresponse behavior is strongly associated with y, indicative that there are observed and unobserved factors affecting it, this relationship cannot be accounted for the observed auxiliary variables x and therefore the bias cannot be corrected. This is the case of the not missing at random mechanism (NMAR).

Therefore, the attempt to identify the missing data mechanism that better explains the response behavior for the chosen target variable in the data set under study, implies to analyze its relationship with x. Suitable auxiliary variables must provide information about the distribution of individuals in the population for both respondents and nonrespondents.

In the case of population surveys, demographic factors such as age, gender, marital status, level of education, region, area, and household structure are among the most common. Also, social security or tax information can be used along with the living conditions observed by interviewers, if any. A set of representativeness indicators, the R-indicators developed by the Representativity Indicator for Survey Quality (RISQ) project can also be used to assess the quality of the collected data and understand its missing data mechanism.

The general literature emphasizes the unit nonresponse problem for which several authors have pointed out the declining survey response rates across countries over time. For example, [2] show several examples for The Netherlands where response rates have gone down below 50%. [3] summarizes the literature on nonresponse and nonresponse bias in surveys in the United States and Western Europe, describing the methods used to reduce refusal rates. Once again, the paper highlights the increasing nonresponse rate observed in US household surveys and the fact that bias is always present.

According to [4], renewed contact attempts can translate into bias reduction only if the effort targets sample units with a low probability of response. [5] use a simulated example to show how an adaptive survey design can improve the quality of the sample and the role of representativeness indicators in such a design, while [6] describe how these indicators can help to obtain more a representative response, using the monthly Dutch Survey of Consumer Confidence as a pilot.

In the Netherlands, [7] evaluates the effect of survey designs on nonresponse among minorities. The author analyzes the disposition to respond and estimates both nonresponse rates and representativeness indicators on the information of the Survey of the Integration of Minorities. [8] use the information provided by several samples in The Netherlands and the US to evaluate the effects of nonresponse adjustments compared to those of adaptive survey designs and find evidence in favor of introducing different treatments to different subgroups.

At the item level, the literature shows more concern for the accuracy of the reports provided by individuals, rather than the nonresponse to a particular item. In the case of sensitive questions such as those related to financial information, [9] conclude that respondents tend to adjust their answers based on what they consider as a desirable report and how uncomfortable they feel when providing the correct answer so that self-administered surveys can help reduce the problem.

[10] find evidence of a systematic income overreporting error among workers in Denmark attributable to social desirability, so that income should always be analyzed in a logarithmic transformation to avoid distorted conclusions.

[11] on the other hand, approach the problem of nonresponse to income in a local labor force survey in Italy. The authors resort to sensitivity analysis of deviations from the MAR assumption to apply a sequential regression multiple imputation method to deal with missing income amounts in rotating panel surveys. A similar analysis was previously conducted by [12] for a health survey in the United States.

[13] summarize the most important approaches to deal with item nonresponse using a German socioeconomic panel survey and highlight the effect of the missing data mechanism assumption on the robustness of the imputation methods. To our knowledge, there are no studies at this regard in Colombia.

Since a high proportion of the research is based on data already collected, it is important to try to understand the structural mechanism of nonresponse and how representative the sample is. Even though it is true that surveys help us to deal with general unit nonresponse by providing administrative sample weights to ensure representativeness, the researcher faces the problem of dealing with item nonresponse for a particular target variable, whose association with other variables and the way it is approached vary from case to case.

This is the nature of this paper, which attempts to evaluate the quality of the wage information obtained from the subsample of occupied workers from the Great Integrated Sample Survey (GIHS) in Colombia, evaluating whether the wage nonresponse mechanism is associated with some observables and comparing its behavior in two different periods of time (2008:4 and 2017:4).

2. Analyzing the representativeness of a sample

2.1. Methodology

The general literature recognizes that the response rate by itself is a poor indicator of the quality of the data [3]. To analyze the potential impact of nonresponse it is necessary to study the randomness of such nonresponse, to decide whether corrective measures are required. This implies to spread light on the wage nonresponse behavior.

For example, multiple correspondence analysis (MCA) can help to identify categories of the auxiliary variables associated with nonresponse. MCA is a factorial analysis useful to uncover the latent structures in a large set of variables, by measuring nonlinear relations among categories of qualitative variables. This descriptive technique allows us to analyze the data without imposing a priori restrictions on the expected association among categories of variables and generates a visual representation of its structure in a two-dimensional space. Although MCA can give us an idea about the potential randomness of the response behavior, yet a measure of representativeness of the sample is required.

A simple way to analyze the possible relation between nonresponse and a set of auxiliary variables is estimating the Cramér´s V statistic given by $V=\sqrt{\chi^2/Nx \min(r-1, c-1)}$, where r and c are the number of categories in a given variable $x \in X$ and in the nonresponse variable, respectively. The smaller the value, the lower the association between them. Unfortunately, this test considers the effect of one variable at the time, ignoring the impact of possible interactions on the response behavior.

A logit regression model can also be used to model the probability of response given the information provided by the set of variables X, so that:

$$\text{Log}\,(\text{P}/(1-\text{P})) = \text{X}'\beta \qquad (2)$$

with P as the vector of probabilities of response and (1-P) the probability of nonresponse; β is a k-vector of coefficients. This model is also useful to estimate the response propensity to be used to evaluate the data quality through representativeness indicators or R-indicators. These indicators are based on the idea that nonresponse leads to less accurate but still valid estimates of the population parameters, as long as there is no nonresponse bias, that is, on average there is no difference between respondents and nonrespondents to the target variable. To determine whether respondents resemble nonrespondents, [14] proposed the general R-indicator *R*, that measures whether a sample is representative based upon the standard deviation of response propensities, such that *R* in its population parametric form is given by:

$$R(\rho_x) = 1 - 2\,S(\rho_x) \qquad (3)$$

Where

$$S^2\,(\rho_X) \;=\; \frac{1}{N}\sum_{i=1}^{N}(\rho_x\,(x_i) - \bar{\rho})2 \qquad (4)$$

with *N* indicating the size of the population, ρ_{X_i} are the response propensities and $\bar{\rho}_x$ is the mean of the response propensities given by

$$\rho x = 1N\ i{=}1N\rho XI \qquad (5)$$

whose estimator is

$$\hat{\rho}_x = \frac{1}{n}\sum_{\text{i=1}}^{\text{n}}\hat{\rho}_{xi} \qquad (6)$$

This indicator can be estimated by $\hat{R}_{\rho_x} = 1 - 2\,\hat{S}\,(\hat{\rho}_x)$ where the population propensities ρ_x can be estimated based on a logistic regression model.

The theoretical properties of this type of indicators are analyzed in [5]. As indicated by [15], this R-indicator along with another one, based on auxiliary variables proposed by [16], were developed as part of the RISQ project to monitor the quality of the data at different stages of the collecting process. Of course, *R* somehow also depends upon the variables used to estimate the response probability.

The goodness of the *R* is that it is based on a Euclidean distance function so that the indicator can be normalized, and it is easy to interpret. Therefore, it takes values between 0 and 1, being 1 the most representative response (all individuals have the same ρ_i) and 0 the least representative response. This indicator is related to the Cramér´s *V* statistics since both indicators measure the lack of association between response behavior and other variables possibly affecting it. The meaning of representativeness in the context of this indicator as stated by [2] refers to the lack of observed selective forces, so that the weaker the association the more the respondent selection will resemble a simple random sample.

Keep in mind that survey topics may influence the probability of response, but such an influence cannot be measured; therefore, representativeness is based on a predefined set of observable variables X. This weak definition of representativeness implies that the missing data mechanism resembles an MCAR with respect to vector X, meaning that respondents are, on average, equal to nonrespondents. If this similarity holds only within a given subgroup, then the missing data mechanism is MAR; otherwise, the mechanism is NMAR [14].

When the measure of representativeness is limited to one auxiliary variable *z*, the indicator is a partial R-indicator. For categorical variables, the partial R-indicators can be defined for each category of z. Keep in mind that the general indicator *R* reflects the overall variation of the individuals´ response probabilities, while partial indicators separate this variation into components attributable to specific characteristics. There are two types of partial indicators:

• The unconditional partial indicator (*Pu*) measures the contribution of a single variable *z* or category *k* to the lack of representative response, where *z* can be or not an element of X. For categorical variables, it is given by

$$P_u\,(z{=}k,\,\rho_x) = \sqrt{\frac{N_k}{N}}\,(\bar{\rho}_{z_k} - \bar{\rho}_x) \qquad (7)$$

with N_k number of population units in category *k* and $\bar{\rho}_{x,k}$ as the weighted sample mean of the estimated response propensities in that category *k*. *Pu* takes values between -0.5 and +0.5. The larger the value of *Pu*(*z*=*k*, ρx), the greater the contribution of the category *k* of variable *z* to the lack of representativeness. Also, a positive (negative) value indicates that the category is over (under) represented. As [17] explain, this indicator measures the between variance of response propensities, while the within variance is accounted for by the conditional partial indicator.

• The conditional partial indicator (*Pc*) measures the contribution of a single variable $z \in X$ or a category k to the lack of representative response, considering other variables [18]. For the case of categorical variables, this indicator is given by

$$P_c\,(z=k,\ \rho_x) = \sqrt{\frac{1}{N-1}} \sum_{i=1}^{N} Z_k(\rho_{x_i} - \bar{\rho}_{z_k}) \qquad (8)$$

The *Pc* is expected to be smaller than the *Pu* for a given variable.

2.2. Data source

In this paper, all the above-mentioned approaches are used to evaluate de quality of the wage information from the GIHS subsample of occupied workers in the two periods considered. This information is collected based on the question "How much were you paid for this job last month? The interviewers are instructed to include the information regarding the monthly wage from the main job in the reference week. These wages are divided by the number of hours normally worked, to generate hourly wages. Notice that hourly wages are the result of the combination of two variables, both of which may suffer from nonresponse. However, the number of working hours is not as sensitive as wages and is generally reported.

The GIHS is a monthly face-to-face sample survey conducted by the National Administrative Statistics Department (DANE for its initials in Spanish) in Colombia since August 2006. It is the result of the combination of three other surveys (households, quality of life, and income and expenditures) aimed to collect information on social, economic, and demographic variables for a representative sample on individual and household levels. As in any other official survey, sample weights are provided to adjust for unit-nonresponse and resemble the original population. The DANE uses the SAS program Clan 97 v3.1. The dataset is used by the DANE to estimate relevant socioeconomic indicators such as unemployment, poverty, and informality rates. Its coverage has increased over time; despite this, the GIHS is still applied to 23 out of the 32 Departments in which Colombia is divided, plus the capital district. It is not a panel since each sample is independent of each other.

To analyze the item nonresponse and its evolution over time this study considers the data for the last quarter of the years 2008 and 2017. The reasons for using these years are as follow: First, we opt for the year 2008 to go back as far as possible in the life of the GIHS but leaving time for its consolidation after its beginning at the end of 2006. The last quarter of 2017 was the most recent database available by the time we started this study. Hence, we chose to work with the fourth quarter of 2008 as well, to avoid seasonal factors. Hourly wages for occupied paid workers aged 15 years and older are considered the target variable, while age, gender, level of education, marital status, region or department, sector and category of employment are assumed to be variables that can lead to differential nonresponse.

3. Result

The response to the wage variable is defined as a binary variable, named response, that takes value one if the individual reports his wage, zero otherwise. For the first period considered, the data include 74,590 workers; of them, 2.97 % are classified as unpaid workers; the nonresponse rate for the remaining workers is 5.62 %. This proportion increases to almost 13% in 2017 (see Table 1), behavior that is consistent with the declining item-response rate observed by [19] in the U.S. Current Population Survey.

Table 1. Wage nonresponse incidence

	2008		2017	
	Observations	**%**	**Observations**	**%**
Occupied	74,590	-	79,906	-
Unpaid work	2,218	2.97	2,535	3.17
Nonresponse	4,099	5.82	10,064	12.96

Source: Authors based on the GIHS.

3.1. Descriptive statistics

According to the data, men and married workers are slightly more likely to not provide information about their wage level. Nonresponse increases over time and appears to be more frequent among older workers, those with the highest level of education and those in the informal sector. Neither these results nor those for Cramer´s *V* test suggest a strong association between nonresponse and the auxiliary variables (see Table 2).

Notice also that the behavior of nonresponse across categories of variables is consistent over time, despite its higher incidence in 2017. A few words need to be said about region and category of worker. In the year 2008, workers in region 1 (Atlantic) are more likely to not report their labor income, with a very slow incidence of nonresponse in region 3 (rest of the country); however, by 2017 is region 2 (Andean) the one that shows a much smaller nonresponse rate. As for the category of worker, employers exhibit the highest and increasing nonresponse rate in both periods, followed by self-employed, especially in 2017.

Table 2. Wage response behavior by demographic characteristics

Variable	Group	2008			2017		
		Respondents	Nonrespondents	*V*	Respondents	Nonrespondents	*V*
Gender (gender)	Male (1)	94.31	5.69		86.16	13.84	
	Female (0)	94.48	5.52	0.004	88.06	11.94	0.028
Education (educ)	Basic (1)	95.06	4.94		88.47	11.53	
	Media (2)	95.11	4.89		88.11	11.89	
	Technical (3)	94.12	5.88		88.51	11.49	
	University (4)	90.99	9.01	0.063	80.32	19.68	0.088
Age (age)	15-25 (1)	94.65	5.35		89.47	10.53	
	26-35 (2)	95.42	4.58		88.18	11.82	
	36-45 (3)	94.77	5.23		87.25	12.75	
	>46 (4)	92.92	7.08	0.044	84.78	15.24	0.053
Civil status (civil)	Single (1)	94.31	5.69		87.68	12.32	
	Married (2)	94.34	5.66		86.62	13.38	
	Other (3)	94.68	5.32	0.006	87.39	12.61	0.014
Region (region)	Atlantic (1)	93.29	6.71		79.91	20.09	
	Andean (2)	94.56	5.44		91.81	8.19	
	Other (3)	97.59	2.41	0.052	81.72	18.28	0.169
Sector of Employment (formal)	Formal (1)	94.64	5.36		88.22	11.78	
	Informal (0)	94.23	5.77	0.009	86.16	13.84	0.030
Category of Worker (type)	Salaried (1)	95.73	4.27		91.28	8.72	
	Self-employed (2)	93.26	6.74		82.69	17.31	
	Employer (3)	88.29	11.71	0.076	76.89	23.11	0.136

Source: authors based on the GIHS.

Yet, the *V* test does not support the idea of a strong relationship between these two variables and the response behavior.

Concerned about the possibility of wage nonresponse to be affected by geographical factors, as suggested by the much smaller incidence of nonresponse in the Andean region and the higher value of the Cramer´s *V* test in the year 2017, we examine the nonresponse rate within the departments in these regions. Two findings are important to highlight: First, departments at the eastern side of the country is not included in the study since there are only part of the GIHS starting in 2012. Second, while departments surrounding the capital Bogotá (Cundinamarca, 7.7%) tend to have low nonresponse rates, the rates in some remote areas tend to be larger, starting by Chocó (54.8%) and followed by Bolivar (48%), Magdalena (43.5%), and Cauca (35.9%). In this case, the Cramer´s *V* test indicates a stronger association between nonresponse and geographical area, especially in the year 2017.

3.2. Multiple correspondence analysis

In MCA almost all the information contained in the database of n observations and m variables is collected in d dimensions, for $d < m$. Since this method only works with categorical variables, the variable age was recoded as shown in Table 2: 1 for ages between 15 and 25, 2 for ages between 26 and 35, 3 for ages between 36 and 45, and 4 for ages above 45. For each period, the MCA yields two dimensions which explain more than 70% of the variability of the variables. We use nonresponse (NR) as a supplementary variable (see Figure 1).

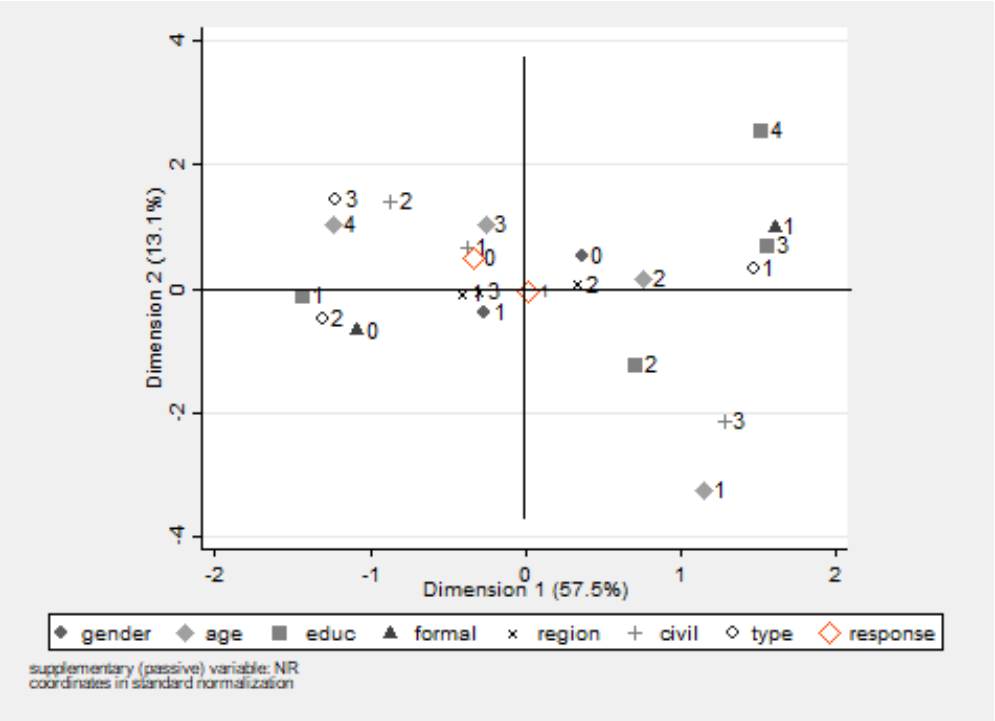
In the year 2008, the location of the response categories at the center of the plane indicates that this behavior does not contribute to the definition of the dimensions and, therefore its association to any particular category of the auxiliary variables is not statistically significant. In the year 2017, it can be observed a slight shift of the nonresponse option away from the center of the plane toward employerse, non-single workers, individuals aged 46 and over and those in the Atlantic region. Still, these results cannot be considered as indicative of any non-random behavior of nonresponse.

3.3. Logit regression model

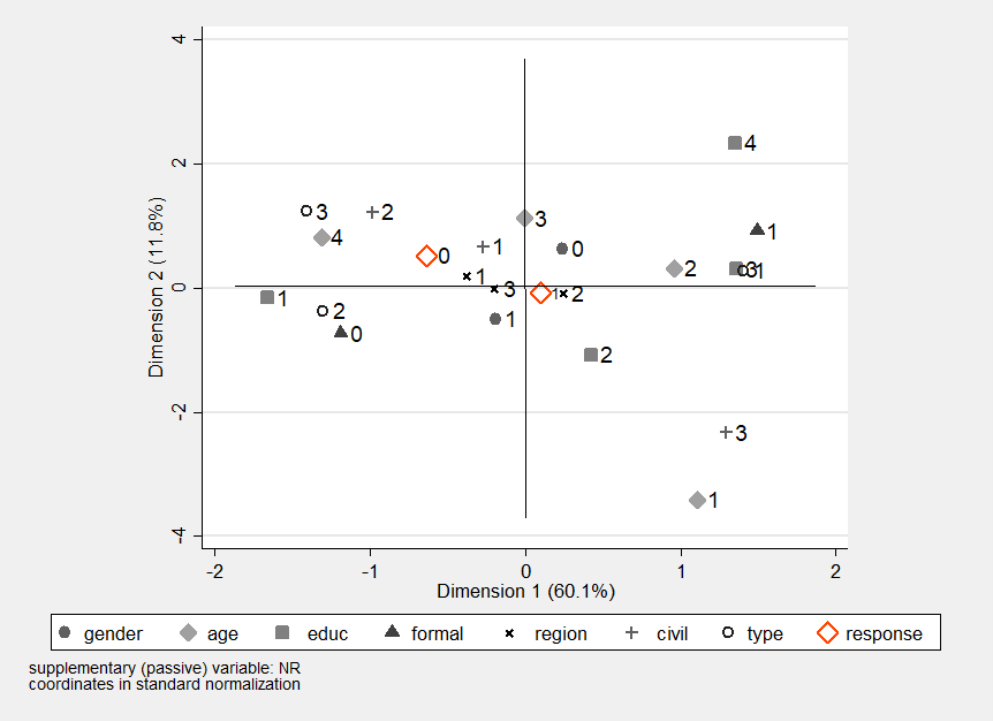
If every individual in the population has an unknown response propensity ρ_i, it is possible to estimate these scores using auxiliary variables. The first step is to fit a logistic regression model for both years as given by

$$\text{Response} = f(\text{gender, educ, age, marital status, region, formal, type}) + \varepsilon \quad (9)$$

with response taking value one for wage respondents, zero otherwise. As Table 3 summarizes, all variables are statistically significant as expected, due to the large data set, with some effects changing direction over time. Some interactions were considered but proved to be statistically not significant.



a. Year 2008



b. Year 2017

Figure 1. Multiple correspondence analysis.

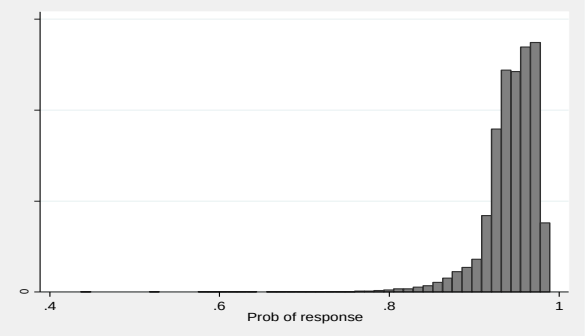
Table 3. Logit model estimates

Variable	2008	2017
Gender (male)	0.0626 *** (0.0337)	0.1135** (0.0051)
Age	0.0441** (0.0064)	0.0126** (0.0001)
Age^2	-0.0006 ** (0.0001)	-0.0003** (0.0294)
Single	-0.1508** (0.0421)	-0.1896** (0.0241)
Region 2 (Andean)	0.2189** (0.0342)	0.0996** (0.0343)
Region 3 (Other)	1.0956** (0.0833)	0.8172** (0.0255)
Self-employed	-0.6113** (0.0383)	-1.1762** (0.0523)
Entrepreneurs	-1.1275** (0.0634)	0.8574** (0.1015)
High School	-0.0169** (0.0427)	-0.2840 (0.0291)
Technological	-0.4141** (0.0644)	-0.3614 (0.0390)
University	-0.8628 ** (0.0464)	-0.9356** (0.0327)
Const	2.6126 (0.1472)	2.3181** (0.1109)
Pseudo R2	0.0357	0.0788
LR Chi2	1095.63*	4717.33*
Log likelihood	-14808.33	-27566.134

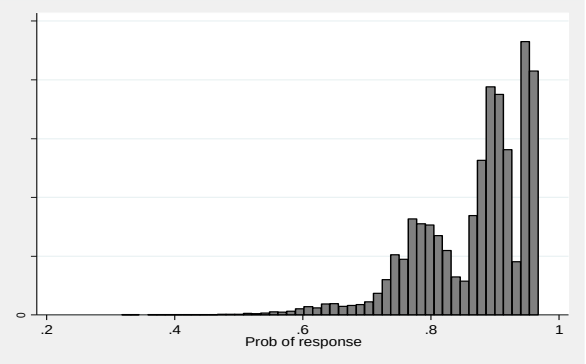
Note: standard errors in parenthesis. Variables significant at 1 (***) and 5% (**).

Source: authors based on the GIHS.

According to these results, the probability that a worker reports his wage decreases with the level of education and changes over regions. Employers are less likely to report their wages in 2008, while in 2017 are the self-employed the ones who are less likely to do so. In both cases, the likelihood ratio chi-square is statistically significant at 1%, indicative that both full models fit better than an empty model. The sector of employment ended up being not statistically significant in both years reason why it was removed. Following [20], propensity scores where obtained from these models; their distribution is shown in Figure 2. As the Figure show, in 2008 most of the scores fall between 0.8 and almost one, with a mean, median and mode of 0.9437, 0.9489, and 0.9586, respectively and a coefficient of variation equal to 0.0329. In 2017 there is a clear change in the pattern of behavior of these scores, with most of them ranging from 0.6 to a value closed to one. In this case, the mean, median and mode are, respectively 0.8725, 0.8961, and 0.9489, while the coefficient of variation goes up to 0.0991.



a. Year 2008



b. Year 2017

Figure 2. Histogram of the wage response propensities.

Despite these results, the main conclusion we can draw from them is that response propensities differ across individuals, with no clear evidence of their correlation with other variables. In the year 2017, the histogram for the response propensities suggests two different sample structures, which could be associated with the category of workers. In fact, 78.8% of the workers whose propensity falls below 0.85 are self-employed, while almost 65% of the workers with propensities above 0.85 are salaried.

The response propensity does not seem to differ between men and women, but rather varies across categories of worker, with an increasing dispersion over time. In the year 2008 the response behavior for salaried workers and self-employed looks similar; by the year 2017, similarities are still observed but between self-employed and employers (see Figure 3).

As for the level of education, only workers holding a university diploma seem to behave differently from the others in the year 2017 (see Figure 4).

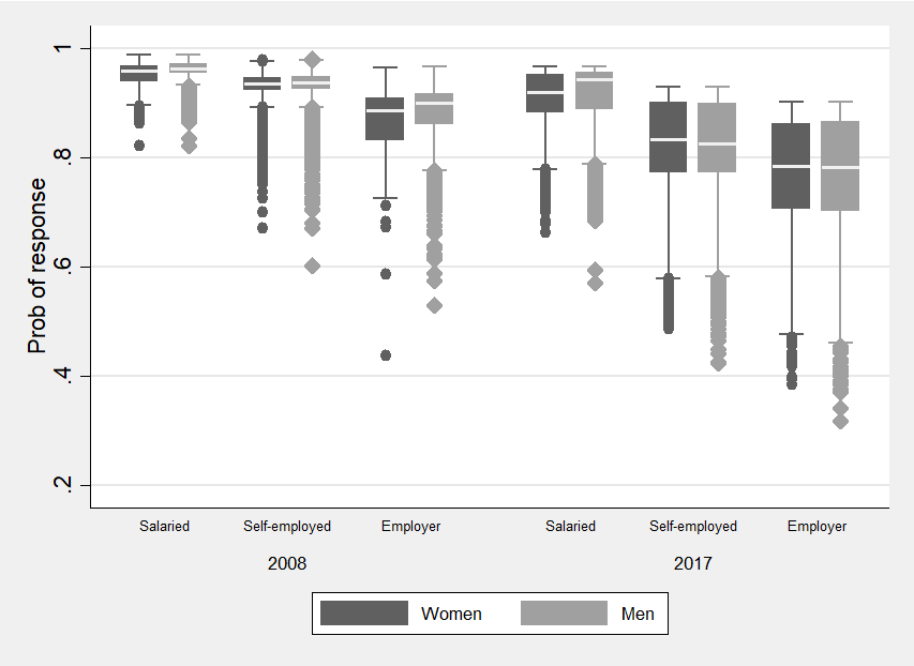


Figure 3. Box diagram of the wage response propensities by sex and category of employment.

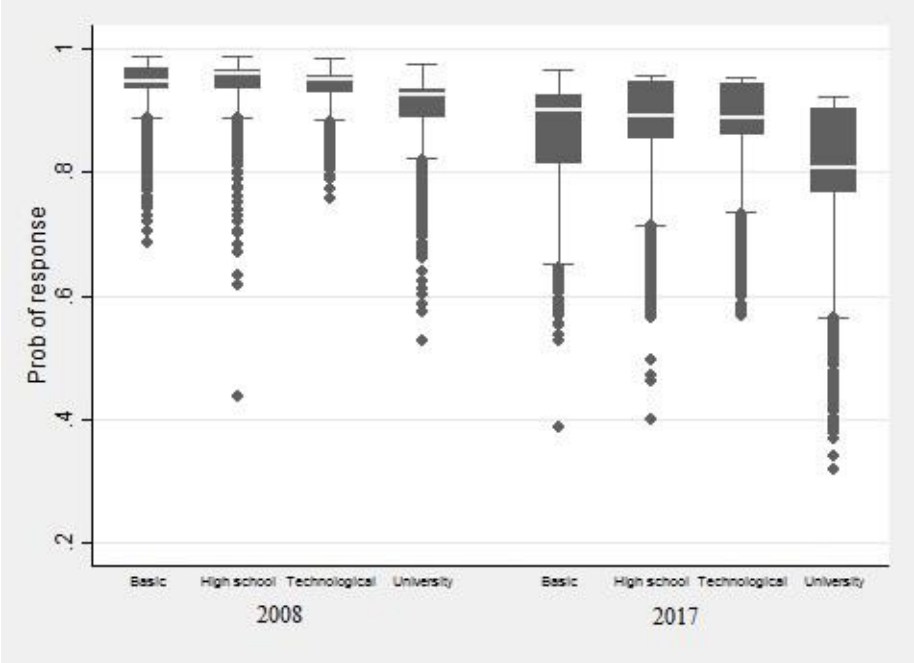


Figure 4. Box diagram of the wage response propensities by level of education.

In the same way, individuals aged 60 years and older show a different behavior in the response propensity in 2008, with variability increasing with age in the year 2017 (see Figure 5). The average response probabilities across regions were pretty much the same in 2008, with no significant differences across regions. Within regions, a very homogeneous pattern of behavior can be observed in all departments except Caquetá, Chocó, and Meta in the year 2008. In 2017 however, not only the dispersion within each department has increased, but also across all of them. Also, no significant differences are observed between formal and informal sector workers, regardless of the category of employment.

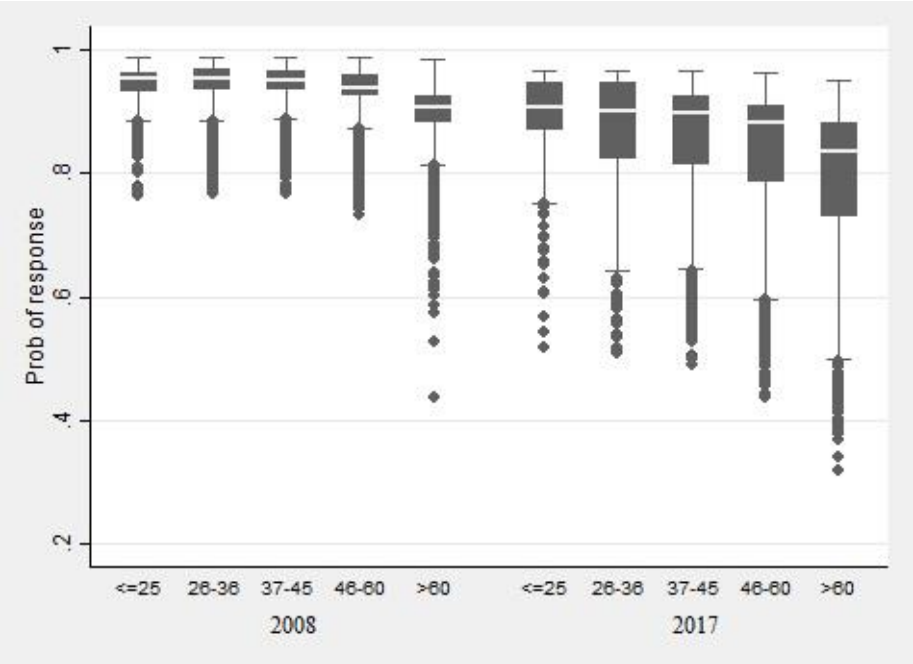


Figure 5. Box diagram of the wage response propensities by age.

3.4. The R-indicators

This section summarizes the results of the R-indicators based on the response propensities estimated from the previous logit models. As mentioned before, the general R-indicator is useful to evaluate the quality of response. As shown in Table 4, the wage response rate decreases over time without compromising its representativeness, given the high level of the indicator R. In fact, the response rate goes down from 94% to 87%, but R remains always above 0.94, suggesting a week association of the nonresponse with the auxiliary variables X.

The unconditional partial R-indicator *Pu* allows the comparison over time of the contribution of a given variable *z* to the lack of representative response. It measures the standard deviation of the response propensity for *z* in the population. In this case, we estimate the *Pu* by category level of the auxiliary variables. The larger the value of *Pu*, the more disperse the response for *z*, with negative (positive) values indicating under- (over-) representation. The small estimates of *Pu* suggest a low contribution of each category to the potential lack of representativeness of the sample. In 2008, the small but positive values indicate over-representation of each group; only in 2017, some unconditional partial indicators are negative, with the corresponding group being under-represented in the sample.

That is the case for employers, workers aged 60 years and over and all regions, especially in the following departments, some of which show the largest nonresponse rates: Atlántico, Bolívar, Caquetá, Cesar, Córdoba, Chocó, La Guajira, Magdalena, Meta, and Sucre (see Table 5).

It can be observed that the conditional categorical R-Indicators possess very low values for all the demographic variables studied, which indicates that these variables produce a very little conditional categorical impact on the wage response representativeness. However, there exists a slight increment in these Indicator values in the year 2017 in relation to 2008. Such a situation may be taken as an indication that the wage nonresponse is independent of all the survey variables considered and the estimators will not be biased.

4. Conclusions

Earnings are a variable which tends to show missing values, given its sensitive character for individuals. However, many studies require this information to carry out further analysis of working conditions, informality, poverty, and so on. When the individuals refuse to provide this information, it can put at risk the representativeness of the sample if such a nonresponse is affected by some factors and the researcher does not control for them.

Table 4. R-indicators and partial R-indicators

		2008		2017	
Response rate		94.18%		87.04%	
General R-indicator		0.9875		0.9480	
		Pu	***Pc***	***Pu***	***Pc***
Gender	Male	0.0572	0.0014	0.0443	0.0035
	Female	0.0462	0.0013	0.0392	0.0035
Level of Education	Basic	0.0512	0.0011	0.0395	0.0029
	High school	0.0474	0.0010	0.0420	0.0030
	Technological	0.0212	0.0011	0.0253	0.0030
	University	0.0153	0.0019	0.0037	0.0047
Category of employment	Salaried	0.0626	0.0008	0.0699	0.0021
	Self-employed	0.0422	0.0012	0.0143	0.0036
	Employer	0.0044	0.0022	-0.0038	0.0046
Region	Atlantic	0.0269	0.0018	-0.0151	0.0039
	Andean	0.0422	0.0011	-0.0143	0.0015
	Others	0.0214	0.0009	-0.0015	0.0056
Marital Status	Single	0.0386	0.0012	0.0330	0.0033
	Non-single	0.0626	0.0014	0.0492	0.0035
Age group	≤ 25	0.0344	0.0010	0.0346	0.0028
	26 – 36	0.0431	0.0012	0.0375	0.0032
	37 – 45	0.0370	0.0012	0.0284	0.0033
	46 – 60	0.0328	0.0014	0.0229	0.0036
	> 60	0.0069	0.0022	-0.0001	0.0044
Sector of employment	Informal	0.0538	-	0.0343	
	Formal	0.0503	-	0.0502	

Source: Authors based on the GIHS.

Table 5. Unconditional R-indicator by department

Departments	2008	2017
Antioquia	0.0277	0.0322
Atlántico	0.0148	-0.0070
Bogotá D.C.	0.0368	0.0417
Bolívar	0.0116	-0.0073
Boyacá	0.0119	0.0107
Caldas	0.0116	0.0105
Caquetá	0.0129	-0.0005
Cauca	0.0103	0.0113
Cesar	0.0071	-0.0054
Córdoba	0.0094	-0.0052
Cundinamarca	0.0164	0.0211
Chocó	0.0068	-0.0023
Huila	0.0059	0.0123
La Guajira	0.0076	-0.0044
Magdalena	0.0102	-0.0059
Meta	0.0157	-0.0005
Nariño	0.0108	0.0119
Norte de Santander	0.0117	0.0141
Quindío	0.0073	0.0120
Risaralda	0.0255	0.0131
Santander	0.0166	0.0172
Sucre	0.0064	-0.0047
Tolima	0.0130	0.0139
Valle del Cauca	0.0251	0.0285

Source: Authors based on the GIHS.

This work uses the information of GIHS of Colombia for two different periods of time (2008:4 and 2017:4) to evaluate whether the wage nonresponse among workers depends on a set of some observed factors. This information is important since nonresponse may affect the quality of the estimates and the methods used to deal with it may also affect the results of the empirical analysis. For example, ignoring observations with missing values may lead to substantial bias if such missingness is subject to some unaccounted but observed factors. The robustness of common imputation techniques depends on whether there are patterns in the missingness of the data or if it can be assumed at random.

The results show that even though the wage nonresponse rate has been tripled in 2017 with respect to 2008, this is still considerably low. Based on the results of Cramer's *V*, the willingness of the individuals to provide information about their wages does not appear to be associated with any variable. In the same way, the MCA was not able to detect any pattern of association between the wage response variable and the different demographic variables considered: age, gender, education level, marital status, department, sector, and employment category.

The logit models used to estimate the wage response propensities show that only the gender in the year 2008 is statistically significant at 1%; all the other factors are significant at 5% However, the low explanatory power of the models, below 8%, suggest that it is not possible to conclude that these variables may condition the willingness of the workers to respond.

Since the wage nonresponse rate is insufficient to determine the quality of a data, we resort to the R-indicators. The general R- indicator remains always above 0.94, which indicates the good quality of the sample and the low association of the nonresponse with the auxiliary variables considered in the estimation of the propensities.

The lower values of the unconditional partial R-indicator estimated by categories of the auxiliary variables put in evidence the low dispersion of the wage response propensities and the absence of an association between these and the categories considered. According to this indicator, several groups were overrepresented in 2008, while such a situation was observed in 2017 only for employers, advanced age workers and some departments like Atlántico, Bolívar, Caquetá, Cesar, Córdoba, Chocó, La Guajira, Magdalena, Meta, and Sucre.

The conditional categorical partial R-indicator measures the deviation from a representative response and the impact of a single variable conditional on the remaining variables. It was observed in our study that all the demographic variables considered conditionally showed very little impact in general and this again increased from the year 2008 to the year 2017.

All in all, the results suggest that the wage nonresponse behavior seems to be at with respect to the factors considered in this study. This assures confidence in the quality of the estimations obtained using the sampled information, without requiring additional adjustments to the sample weights in order to compensate for the nonresponse, process that can be cumbersome and not always necessary as shown by [21] for the same survey. Nevertheless, it is advised to keep under observation any additional increment that may occur in the wage nonresponse rate in the GIHS, following a worldwide tendency and, in the same way, the periodical inspection of the mechanism of generation of missing data.

References

[1] P. Lynn, R. Beerten, J. Laiho, J. Martin, "Towards standardization of survey outcome categories and response rate calculations," *Research in Official Statistics*, vol. 5, no. 1, pp. 63-86, 2002.

[2] J. G. Bethlehem, F. Cobben, B. Schouten, *Handbook of nonresponse in household surveys.* New Jersey, NJ, USA: John Wiley & Sons, 2011.

[3] R. M. Groves, E. Peytcheva, "The impact of nonresponse rates on nonresponse bias: a meta-analysis," *Public Opinion Quarterly*, vol. 72, no. 2, pp. 167-189, 2008. doi: 10.1093/poq/nfn011

[4] K. Beullens, G. Loosveldt, *R-indicators and fieldwork monitoring RISQ Project*. Oxford, Manchester, UK: Manchester University, 2010.

[5] N. Shlomo, B. Schouten, V. de Heij, "Designing adaptive survey design with R-indicators," in *Proc. of the New Techniques and Technologies for Statistics Conference*, Brussels, 2013.

[6] A. Luiten, W. Wetzels, "Differential survey strategies based on r-indicators," in, *European Conference on Quality in Official Statistics (Q2010), Helsinki, Finlandia,* 2010, pp. 4-6.

[7] J. Kappelhof, "The effect of different survey designs on nonresponse in surveys among non-western minorities in the Netherlands," *Survey Research Methods*, vol. 8, no. 2, pp. 81-98, 2013. doi: 10.18148/srm/2014.v8i2.5784

[8] B. Schouten, F. Cobben, P. Lundquist, J. Wagner, *Theoretical and empirical support for adjustment of nonresponse by design.* Statistics, Netherlands, 2014.

[9] R. Tourangeau, T. Yan, "Sensitive questions in surveys," *Psychological Bulletin*, vol. 133, pp. 859-883, 2007. doi: 10.1037/0033-2909.133.5.859

[10] J. G. Hariri, D. Dreyer, "Income and outcomes: social desirability bias distorts measurements of the relationship between income and political behavior," *Public Opinion Quarterly*, vol. 81, no. 2, pp. 564-576, 2017. doi: 10.1093/poq/nfw044

[11] C. Giusti, R.J. Little, "An analysis of nonignorable nonresponse to income in a survey with rotating panel design," Journal of Official Statistics, vol. 27, pp. 211-229, 2011.

[12] N. Schenker, T. E. Raghunathan, P. L. Chiu, D. M. Makuc, G. Zhang, A. J. Cohen, "Multiple imputation of missing income data in the national health interview survey," *Journal of the American Statistical Association*, vol. 101, pp. 924 -933, 2006. doi: 10.1198/016214505000001375

[13] S. Rässler, R. Riphahn, "Survey item nonresponse and its treatment," *Allgemeines Statistisches Archiv*, vol. 90, pp. 217-232, 2006. doi: 10.1007/s10182-006-0231-3

[14] B. Schouten, F. Cobben, J. G. Bethlehem, "Indicators for the representativeness of survey response," *Survey Methodology*, vol. 35, no. 1, pp. 101-113, 2009.

[15] C. Skinner, N. Shlomo, B. Schouten, L. Zhang, J. Bethlehem, "Measuring survey quality through representativeness indicators using sample and population base information," in *Proc. of the New Techniques and Technologies for Statistics*, Luxembourg, 2009, pp. 18-20.

[16] C.E. Särndal, S. Lundström, "Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator," *Journal of Official Statistics*, vol. 24, no. 2, pp. 167-191, 2008.

[17] B. Schouten, N. Shlomo, C. Skinner, "Indicators for monitoring and improving representativeness of response," *Journal of Official Statistics*, vol. 27, pp. 1-24, 2011.

[18] J. G. Bethlehem, F. Cobben, B. Schouten, "A new quality indicator for survey response," *AENORM*, vol. 19, pp. 24-28, 2011.

[19] B. K. Atrostic, C. Kalenkoski, "Item response rates: one indicator of how well we measure income," in *Proc. of the Joint Statistical Meetings, Section on Survey Research Methods*, Washington, 2002, pp. 94-99.

[20] G. Imbens, D. P. Rubin, *Causal inference for statistics, social, and biomedical sciences: an introduction*. New York, NY, USA: Cambridge University Press, 2015.

[21] S. P. Sinha, J. Ramoni-Perazzi, G. Orlandoni-Merli, E. Torres, "Weight adjustments after sub-sampling cross-sectional data," *Revista Estadística Española*, vol. 59, pp. 45-57, 2017.