Proceedings of the “Seecond International Conceference on Establishment Surveys”, Buffalo, , June 17-21, 2000

MODELLING FIRM RESPONSE AND CONTACT PROBABILITIES IN WEB SURVEYS[1]

Silvia Biffignandi and Monica Pratesi, DMSIA
Silvia Biffignandi, Dipartimento di Matematica, Statistica, Informatica e Applicazioni (DMSIA)
Piazza Rosate, 2 I-24129 Bergamo
silvia@unibg.it

ABSTRACT

The framework, which is driving the contact and response in a Web survey, is briefly discussed, in order to set up the rules under which the contact event is identified. Given that framework and the rate measures computed within it, theory for modeling the participation in a Web survey and the respondents’ behavior during the compilation process is presented. It is proposed to detect the contact process through data collected on the server during the survey period. The probability of participation is then divided into two components, first the contact with the firm, then the decision to participate, and it is estimated separately by logistic regression. The respondents’ behavior is modeled using auxiliary Web process variables, too. Some results of the pilot survey of the CAWI project, which has been undertaken from the authors, are presented.

Key Words: **Web Survey, Contact, Non-response.**

1. INTRODUCTION

Surveys over Internet are now in a process of intensive development. Collecting data through Internet survey is useful either for marketing and other private research societies either for the National Statistical Institutes. The correct use of this data collection tool requires the development of a suitable survey methodology. One of the major topics is the exploration of the process of participation in Web surveys and of respondent’s behavior.

The main factors influencing the cooperation in surveys over Internet, as suggested from empirical studies (Batagelj, et al 1998; Dillman, et al 1998), could be grouped into:

- factors influencing the contact with the respondent (e.g. awareness of the survey, e-mail solicitation plan);
- individual factors (e.g. socioeconomic characteristics, knowledge of the subject, experience with Web surveys) that shape the decision to participate as the result of the interaction of the individual characteristics of the respondent with the invitation of the survey designer.

When the survey is directed toward a specific target population with access to Web, the first tentative contact with the respondent is often an e-mail message. Since e-mail message is not a guarantee of the contact event, before presenting the models, we propose and discuss concepts and definitions of various aspects of the non-contact versus contact event in a Web survey and suitable contact rates (section 2). The statistical models for participation in a survey given the contact and for the respondents’ behavior are specified in section 3. The main feature of the proposed models is the inclusion of auxiliary variables that can be chosen in the set of a priori, Web process and a posteriori variables (these sets are described in section 2). In section 4 and 5 we discuss the survey methodology of our empirical analysis and the results of the application of the model to the pilot survey data.

2. A FRAMEWORK FOR CONCEPTS AND DEFINITIONS IN A WEB SURVEY: CONTACT, NON-CONTACT, AND RESPONSE

The interviewing process may be affected by a number of events, and factors that originate them. There are factors that are in common with other kinds of survey, especially with mail surveys, and other factors that are specific of the contact process in Web surveys. Among the factors influencing the outcome of the survey there are the characteristic of the population, the difference between e-mail receiver and interviewed people and the questionnaire structure. In this paper we are focusing on the bias and errors connected to the contact-response process. Before presenting a methodology and our empirical analysis, we set up the framework of concepts and definitions which are needed to

[1] The research is carried out from the local unit of the University of Bergamo within the Murst 40% 1997 research grant on “ Strategie e tecniche di formazione di dati microeconomici per l'analisi del comportamento delle famiglie e delle imprese” (“Strategies and techniques for the construction of microeconomic data for households and enterprises behavioral analysis”).

identify the contact event and, therefore, to apply the proposed methodology for the analysis of the contact and response process.

Given an e-mail list of size N, the possible events in the survey process are non-contact (NC possible outcomes) and contact (C possible outcomes). The non-contact brings to no participation in the survey, whereas the contact opens the dichotomous possibility of participation or no participation (N=NC+N).

In Web surveys missed contacts (Table 1) can be caused by an error which explicitly appears in the e-mail address (WE), i.e. the e-mail is returned for unknown receiver, or by the absence of reaction of the firm (NoR). The list of e-mail addresses of the target population can contain not exact e-mail addresses that cause a missed contact (WE) and produce a *coverage error*. This error can be corrected with a check of the list before the survey, but the firm can change its e-mail address also after the design of the Web survey making it difficult to correct it during the survey period. Also when the e-mail address appears right but the firms do not react to the invitation in the survey period, the contact is missed (NoR). The *no reaction rate* plus the coverage rate compose the *non-contact rate*.

Table 1. Event: non-contact (NC)

EFFECT: NO PARTICIPATION IN THE SURVEY			
CAUSE	EVIDENCE	MEASURE	
Explicitly wrong e-mail address (WE)	Failed delivery message	Coverage error	Non-contact rate
Right e-mail address and no reaction (NoR)[2]	Nothing	No reaction error	

The contact event (C) is confirmed by various possible reactions after that the e-mail collaboration request has been sent. Some facts are registered and they state (directly, i.e. e-mail confirm or refusal letter, or indirectly, i.e. questionnaire entering) that the contact with the firm (C) has been successful. The questionnaire has been fulfilled and communication is given according to the compilation instructions. According to the survey procedures that we propose and adopt in our research, the contacts correspond to the following situation (see Table 2A and 2B). The firm receives the mail and completes the self-interview (CI) (also partial self-interview is possible (VI)). The firm sends only a return receipt and it does not enter the form for the self-interview (ORR), or it only enters the form without filling the items of the questionnaire (OE), or it sends an explicit refusal letter (ER).

Table 2. Event: contact (C)

TABLE 2A- **EFFECT: PARTICIPATION IN THE SURVEY**			
CAUSE	EVIDENCE	MEASURE	
Voluntary interruption of the self-interview: partial self-interview (VI)	Press of finish button	Response rate	Contact rate
Completed self-interview (CI)	E-mail message		

TABLE 2B - **EFFECT: NO PARTICIPATION IN THE SURVEY**			
CAUSE	EVIDENCE	MEASURE	
Only return receipt (ORR)	Return receipt	Non-response rate (which includes the Refusal rate, which refers only to explicit refusals)	Contact rate
Only entering the form (OE)	Log file		
Explicit refusal statement (ER)	E-mail message		

Only the contacted firms that respond and complete (CI) or partially fill the form (VI) for the self-interview participate in the survey. No participation for missed contact can be distinguished from no participation for refusal to cooperate.

The *contact rate* is the percentage of e-mails that corresponds to the previous situations. The *response rate* plus the *non-response rate* compose this rate. As regards the contact rate we propose the gross contact rate which refers to the total of the e-mail list and the net contact rate which removes from the total of the e-mail list the number of explicitly wrong e-mail addresses. Formally, we compute the rates as follows:

Non-contact rate $1-\bar{p}$ **:** $\frac{WE+NoR}{NC+C}$ **;**

[2] No reaction could be due to no interest at all in the survey, anyway we should have in mind that it could catch e-mail address error, too. At this proposal Swoboda et al. (1997) showed that the count of non-deliverable messages may underestimate technical difficulties in transmitting the instrument.

Contact rate $\bar{p}$ **:** **gross contact rate:** $\frac{VI + CI + ORR + OE + ER}{NC + C}$; **net contact rate:** $\frac{VI + CI + ORR + OE + ER}{NC + C - WE}$;

Response rate given the contact [3] $\bar{q}$ **:** $\frac{VI + CI}{C}$; **Non-response rate given the contact** $1 - \bar{q}$ **:** $\frac{ORR + OE + ER}{C}$ **.**

Having defined the framework of the contact and non-contact process and suitable measurement of various aspects of the participation/no participation in a Web survey, now some modeling of the process can be proposed.

It can be hypothesized that a Web survey with a high non-contact and low refusal rate should theoretically have a different set of correlates of overall non-response than one with a low non-contact rate and high refusal rate. Empirical support can be found in Vehovar (1998) using data from a comparison of a Web and a telephone survey. If we trust the previous distinction, the non-response can be seen as the result of a two-stage selection. The first stage, the contact process, produces the set of contacted firms (C). Then, only the contacted firms have a possibility of participating in the survey; it is the sub-set VI+CI. We propose to calculate the response rate conditionally on the set of contacts. As stated above, we can assume that contact and response may be correlated with different sets of explanatory variables. In Web surveys we can distinguish three kinds of auxiliary information according to the time when the information is collected and available:

- Information available before the survey: sets of possible explanatory variables often available in the file from which the e-mail list is obtained: e.g. denomination, size, legal form, economic activity of the firm (*a priori* variables);
- Information available during the survey: they are production process variables automatically collected by the CAWI system: e.g. number of return receipts, number and time/data of the accesses to the server where the form is resident, data and time of each e-mail message (invitation and following soliciting messages) (*Web process* variables);
- Information available after the survey: other characteristics of the firm collected with the self-interview: e.g. familiarity with the Web, use of e-commerce of the firm (*a posteriori* variables).

These sets of variables are candidates to correlate with the individual contact and response; particularly the *Web process* variables are candidate to explain the contact process and/or the respondent's behavior.

In order to estimate the individual probability to participate in a Web survey we consider the first stage of the data production process, the contact process, as the selection of a sort of Poisson sample *c* (Särndal et al, 1992) of the e-mail addresses. Then, another Poisson selection on the set *c* originates the set *r* of the respondents. Given the set of contacted firms, the sub-set of respondents can be selected; it allows for the investigation of the compilation behavior. A special interesting and quite uninvestigated topic is the identification of the most suitable survey period length and the number and periodicity of solicitations in a Web survey. We propose modeling the respondents' behavior by using Web process variables as auxiliary variables.

3. METHODOLOGY FOR CONTACT, RESPONSE AND PARTICIPATION PROBABILITY

3.1 Individual probability of contact

Let $\boldsymbol{C} = \{c : c \subseteq s\}$ be the set of all possible subsets of contacts from the e-mail list l and $p(c \mid l) \geq 0$ be the probability of each subset with $\sum_{c \in \boldsymbol{C}} p(c \mid l) = 1$. The individual probability of contact can be expressed by $p_i = \sum_{\boldsymbol{C}_i} p(c \mid l)$, where $\boldsymbol{C}_i = \{c : i \in c \subseteq l\}$ is the set composed of the whole of the subsets of contacts, which contain unit *i*. The set *c* has variable size $n_c \leq n$. The definition of p_i is similar to that of π_i, the probability of inclusion of the unit *i* in a sample *s* given the sample design *p*(*s*) (Cassel et al, 1983). Let C_i be the dichotomous variable with value 1 if *i* is contacted and 0 otherwise, the probability distribution $p(c \mid l)$ can be expressed by $p(c \mid l) = \prod_{i=1}^{n} p_i^{C_i} (1 - p_i)^{1 - C_i}$, under the hypothesis that the contact of *i* is independent on the contact of *j* with $i, j \in l$ and $\forall i, j \in l$.

[3] We propose to compute response and non-response rate based on the certification of the contact since our aim is to distinguish between access failure and respondent resistance and avoid the risk of getting low response rates which are largely due to delivery problems.

3.2 Individual probability of response

The contacts subset *c* is divided in three groups: 1) those who refuse the self-interview (*f*), 2) those who do not complete the interview in the survey period (*a*), 3) those who cooperate and complete the self-interview (*r*). Given $C_i = 1$ for the firm *i*, the possible outcome of the contact are indicated by the following variables: F_i, that assumes value 1 with probability f_i if the firm refuse to respond (0 otherwise); R_i, that is equal to 1 with probability q_i when the interview is completed (0 otherwise); A_i, that indicates the last outcome of the contact and is equal to 1 with probability a_i. The likelihood conditional on the contact, under the assumption of independent behavior of *i* and *j* after the contact, is then $q(f,a,r \mid c,s) = \prod_{i=1}^{n_c} \left[f_i^{F_i} \cdot a_i^{A_i} \cdot (1 - f_i - a_i)^{R_i} \right] \cdot p_i^{-1}$, where $q_i = 1 - f_i - a_i$ is the probability of cooperation.

3.3 The probability of participation in Web survey

Given the list *l*, the individual propensity of participating in Web survey $P(I_i \mid l)$ is expressed by the product of the contact probability p_i with the probability of cooperation given the contact q_i: $P(I_i \mid l) = p_i \cdot q_i$. Let us assume $\text{logit}(p_i) = \mathbf{x}_i' \beta$, where vector $\mathbf{x}_i$ contains the variables correlated with the contact of the firm *i*, p_i can be estimated maximizing on β the logarithm of the likelihood $p(c \mid s)$

$$l(\beta) = \sum_{i=1}^{n} \left[C_i \mathbf{x}_i' \beta + (1 - C_i) \ln(1 + \exp(\mathbf{x}_i' \beta)) \right] \qquad (1)$$

The result is $\hat{p}_i = \exp(\mathbf{x}_i' \beta) / (1 + \exp(\mathbf{x}_i' \beta)$. Further assuming that $\text{logit}(f_i / (1 - f_i - a_i)) = \mathbf{t}_i' \alpha_1$ and that: $\text{logit}(a_i / (1 - f_i - a_i)) = \mathbf{z}_i' \alpha_2$, where $\mathbf{t}_i$ e $\mathbf{z}_i$ are explanatory variables at individual level, the logarithm of the conditional likelihood $q(f,a,r \mid c)$ is:

$$l(\alpha_1, \alpha_2) = \sum_{i=1}^{n_c} \left[R_i \mathbf{t}_i' \alpha_1 + A_i \mathbf{z}_i' \alpha_2 - \ln(1 + \exp(\mathbf{t}_i' \alpha_1) + \exp(\mathbf{z}_i' \alpha_2)) \right] \qquad (2)$$

Maximum likelihood estimates of α_1 e α_2 allows for the estimation of f_i and a_i and then the estimation of $\hat{q}_i = 1 - \hat{f}_i - \hat{a}_i$ probability of cooperation given the contact, given the individual covariates $\mathbf{t}_i$, $\mathbf{z}_i$.

3.4 The behavior of the respondents

An important issue in managing a Web survey is the time when the firm completes the self-interview. The firms who participate in the survey can complete the form in a few days from the first e-mail of invitation or can complete the form after some solicitation messages at the end of the survey period. This behavior can be modeled again using a logit link, modeling the probability s_i of closing the self-interview in one, two or three weeks with reference to the firms who cooperate and complete the self-interview. If this set is composed by *r* firms and S_i is the indicator of the event of interest (i.e.: $S_i = 1$ when the firm close in two weeks, 0 otherwise) the probability s_i can be estimated assuming again that $\text{logit}(s_i) = \mathbf{x}_i' \gamma$, where the vector $\mathbf{x}_i$ contains the covariates of the time of response of the firm *i*. The probability s_i can be estimated by maximizing on γ the logarithm of the likelihood:

$$l(\gamma) = \sum_{i=1}^{r} \left[S_i \mathbf{x}_i' \gamma + (1 - S_i) \ln(1 + \exp(\mathbf{x}_i' \gamma)) \right] \qquad (3)$$

The behavior of the respondents during the survey period can be modeled using the auxiliary variables that we have indicated as *Web process* variable.

4. METHODOLOGY OF OUR WEB SURVEY

In this paper we present the research methodology of our CAWI survey and briefly describe the results of the pilot survey. The whole interviewing project is directed toward about 2000 firms of the provinces of Bergamo, Brescia, Lecco, Varese, and Mantova (each province is in the Lombardy region). The pilot survey was submitted to around 400 manufacturing firms of the Bergamo province. E-mail addresses were identified from a list of the Chamber of Commerce (as regards the Bergamo’s firms) and from a Unioncamere dabatase for the other provinces.

In our study, we do not enter the matter of the exhaustiveness of the population list; surely the firm list does not represent the whole population, it represents a segment with access to Internet. Firms with access to Internet may be rare at this moment in Italy, but their number is increasing[4]; moreover with reference to this segment the list is formed on a voluntary bases. Although no complete coverage of the list is supposed, we work under the hypothesis that the list is our target population. With regard to the objective of our analysis (the preliminary investigation of the contact, participation process and respondents behavior) hypotheses on the characteristics of the list should be ignored.

The data collection was carried out according to the following steps:

- The first invitation to participate in survey has been an e-mail sent (survey presentation letter) to each firm of the e-mail list at the beginning of the survey period. In order to encourage the participation incentives have been offered (survey report and other connected advantages). The discussion on the opportunity of promotion and incentives to cooperation in the first e-mail to reduce the non-contacts rate has been the topic of many contributions. Batagelj, et al (1998) has studied the effects of incentives on contacts.
- The access procedure to the questionnaire compilation has been made simple, by letting each firm straightforward receiving its own address for compilation; no identification code (id) and a password were required. It has been planned the recording of all the subsequent accesses of each firm and the day, hour t and duration of each access.
- During the survey period three soliciting e-mail messages have been sent. Soliciting e-mail messages have been sent weekly. They had a great impact in getting the questionnaire compilation, as we will note by analyzing the results.

The survey content has been on “Technological communication and links between enterprises”; the questionnaire has been kept simple[5] and required mainly qualitative answers and some percentage data[6]; the interviewed firms were asked for the use of e-commerce, the collaboration with other enterprises and/or their belonging to groups, the markets, the employment. We do not present in this paper the results on the content of the questionnaire. We show the rates, which describe the framework of our empirical analysis and the estimated response behavioral model.

5. SOME PRELIMINAR RESULTS

The e-mail list of the pilot survey corresponded to a population of 399 firms of Bergamo province. The pilot survey ended after four weeks. At the end of the survey period we obtained 118 completed self-interviews (CI) and 140 accesses for a visit of the form without filling any item of the questionnaire by 140 different IP (Internet Protocol) addresses. The questionnaire consisted of 8 pages (6 pages of 40 substantial items, 1 welcome page and 1 final page). The completed forms corresponded in average to 8.24 visited pages (c.v.=0.55, median=8) with a mean duration of each inspection of 0.408 minutes (c.v=2, median=0.1): the mean value greater than 8 indicates that some firm visited each page more than once during the same session. The visit of the form without any filling lasted 5.48 minutes in average (c.v=50.16, median=0.465) and the median value of visited pages is 3 (mean=4, c.v=0.78).

[4] The development of e-commerce has been a challenge to gain visibility on Internet, and many firms in several sectors of economic activity are now joining the net. The diffusion of Internet has been growing in the small medium size enterprises, too. An interesting segment of the Internet diffusion is regarding sub-contracting firms. In 1999 the sub-contracting database of the Banca Dati Subfornitura has registered the following penetration rates of Internet (Source: Banca dati Subornitura):

	Employees size class						
	-5	6-9	10-19	20-49	50-99	100 and over	Total
% sub-contracting firms with Internet access (1999)	25	19	31	58	68	77	405

[5] Dillman, et al (1998) noticed that the success of the contact (where success is the access to the server by the firm) is linked to the plain or fancy design of the questionnaire.

[6] Most questions in the questionnaire have been already tested by mail surveys. The questionnaire has been set up in an electronic format by using EZSurvey program.

In theory the 140 visits do not correspond to 140 firms which only enter the form (OE). The host identification number is not in a one to one correspondence with the firm. The providers can assign to different clients the same IP for different access in a day and the same client can have assigned different IP in different days or during a very busy day. Anyway, since the firms in the target population use many different providers (399 firms, 248 different IP addresses and 38 different providers), and since the mean number of accesses to visit the form has been 1.28 (c.v=0.5), it is likely that the same firm has the same IP number during the survey period. For this reason we considered the accesses of 140 different IP numbers as the accesses of 140 different firms, therefore as an estimate of OE. Thus, no participation in the survey due to non contacting reasons has consisted in 133 firms of which 62 assigned to explicitly wrong e-mail addresses (WE) and 71 detected through no reaction to the e-mail messages (Nor). The contacts have been $n_c = 266$. In particular, we obtained 8 explicit refusals (ER), 118 completed self-interviews (CI) and no partial self-interviews (VI=0), no only return receipt (ORR=0) and 140 only entering the form (OE).

At a global level, the contact and response process can be measured with reference to the rates given in section 2: non-contact rate $1-\overline{p} = 133/399 = 0.33$; gross contact rate= $266/399 = 0.67$; net contact rate: $266/337 = 0.798$; Response rate given the contact: $118/266 = 0.45$; Non response rate given the contact: $148/266 = 0.55$.

Our goal was to estimate $\hat{p}_i$ and $\hat{q}_i$ and the individual probability of participation $P(I_i \mid l) = p_i \cdot q_i$. With our data it is possible to estimate the probability of participation only at global level: $P(I_i \mid l) = P(I \mid l) = \overline{p} \cdot \overline{q} = 0.798 \cdot 0.45 = 0.3591$. The problem is that it is impossible to identify the 140 contacted firms, which entered the form and build the indicator $C_i = 1$. The behavior of the respondent who completed the self-interview can be studied at individual level with the models presented in section 3.3.1. The response variable is $S_i = 1$ when the firm closes the self-interview in two weeks (71 firms 60.2%), 0 otherwise (47 firms 39.8%).

Table 3:Interviews and accesses by soliciting messages

	Self-interviews		Accesses			Self-interviews		Accesses	
N.Soliciting Messages	freq	%	freq	%	**N.Survey Weeks**	freq	%	freq	%
0	29	24.6	114	81.4	**1**	35	29.6	69	49.2
1	24	20.3	17	12.1	**2**	36	30.5	32	22.9
2	35	29.7	9	6.4	**3**	31	26.3	29	27.9
3	30	25.4	-		**4**	16	13.6	-	-
Total	118	100	140	100	**Total**	118	100	140	100

Individual propensity to complete the self-interview in less than 2 weeks depends on the indicator of reactivity in 2 working days from the mail messages (invitation message or first, second or third soliciting message) and on the indicator of having been solicited exactly three times. The propensity seems to increase if the firm has a prompt reaction to the e-mail (odds ratio=3.12, p=0.0563) and seems to be reduced if the firm is difficult to be contacted (three soliciting messages) (odds ratio=0.042, p=0.0001; Chi-square for covariates=39.52 with 2 DF p=0.0001).

6. REFERENCES

Batagelj, Z., K. Lozar and V. Vehovar (1998), “Who are non respondents in Web Surveys?”, Paper presented at the 9t^{h} International Workshop on Household Survey Non-response, Bled, September 1998.

Cassel C.M, C.E Särndal.and J.H. Wretman (1983), “Some Uses of Statistical Models in Connection with the Non-response Problem”. In: *Incomplete Data in Sample Survey*, vol.3, Academic Press, New York.

Dillman, D. A., R. D. Tortora and D. Bowker (1998), “Influence of Plain vs Fancy Design on Response Rates for Web Surveys”, *Proceedings of Survey Methods Section*, 1998 ASA Meeting, Dallas, Texas.

Särndal C.E., B. Swensson.and J. H. Wretman (1992), *Model Assisted Survey Sampling*, Springer-Verlag: New York.

Swoboda W.J. et al. (1997), Internet Survey by Direct Mailing. *Social Science Computer Review*, **15** (3).

Vehover V, Z. Batagelj and K. Lozar (1999), Web Survey: can the weighting solve the problem?, Paper presented at the 1999 AAPOR Conference, St. Petersburg, Florida, May 13-16 1999.