

A Mixture Model for the Identification of Fraudulent Interviewers in the Italian Multipurpose Surveys

Marco Fortini, Adriano Pareto
Istituto Nazionale di Statistica, Roma
e-mail: fortini@istat.it, pareto@istat.it

Abstract: In this paper we present a mixture model for the identification of cheating interviewers in the social surveys carried out by Istat. An EM algorithm is used to estimate the parameters of the model, starting from reinterviews results and from a set of variables describing the interviewer characteristics and performance. The above method is applied to data collected in the Italian Multipurpose Survey 2000. Non-parametric jack-knife is used to obtain variance estimates for the model parameters. The proposed model can be a useful tool for survey quality improvement.

Keywords: survey data quality, mixture model, EM algorithm, classification

1. Introduction

Interviewer cheating is an ubiquitous problem in survey organizations, especially with respect to personal interviews (Schäfer *et al.*, 2004).

The method used for detecting interviewer cheating in the Italian Multipurpose Surveys is based on telephone check to survey respondents (reinterview): for each interviewer, one respondent is recontacted in order to verify that the interview was conducted as required. If a dishonest interviewer is detected through the first check, all its interviews are checked and its proportion of invalid interviews is computed. Of course, a single telephone check for each interviewer does not allow to detect all the cheaters and the lower is the average number of invalid interviews carried out by cheaters, the higher is the expected number of undetected cheaters.

Here we propose a mixture modelling approach based on the idea that interviewers belong to one out of two unobserved sets: the ‘compliant’ interviewers and those who behave incorrectly with a certain probability. The model estimates the probability of interviewers to belong to the second set even when the initial check did not detect their dishonest behaviour. Moreover this probability is explained as a function of the interviewer characteristics and performance indicators.

2. The model

The data structure is showed in table 1. Being i the label assigned to the i -th interviewer, we name:

- T_i the indicator variable assuming value 1 if the contacted respondent gives evidence of interviewer cheating and 0 otherwise;
- M_i the number of interviews completed by the interviewer;

- K_i the number of interviews cheated by the interviewer coming from a full control on the remaining M_i-1 completed interviews ($K_i = 0$ by definition when $T_i = 0$);
- Φ_i the unobserved indicator variable assuming value 1 if the interviewer belongs to the group of cheaters and 0 otherwise.

Table 1: *Data matrix structure*

i	T_i	M_i	K_i	Φ_i
1	0	10	0	0
2	0	15	0	1
3	1	18	0	1
4	1	12	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	1	11	10	1

Given the definitions above, we can assume

$$\begin{cases} \pi = P(\Phi_i = 1) \\ p = P(T_i = 1 | \Phi_i = 1) \\ P(T_i = 0 | \Phi_i = 0) = 1 \end{cases}$$

The resulting Likelihood can be factorised into a couple of components where the first of them depends only on π whereas the second one depends only on p :

$$\begin{aligned} L(p, \pi; \Phi, T, K, M) &= \prod_i \pi^{\Phi_i} (1-\pi)^{(1-\Phi_i)} \prod_i p^{T_i(\Phi_i + K_i)} (1-p)^{[(1-T_i)\Phi_i + T_i(M_i - 1 - K_i)]} \\ &= L(\pi; \Phi) L(p; \Phi, T, K, M) \end{aligned}$$

A dependence of π from explanatory variables can be further assumed through the link function $\text{logit}(\pi) = \alpha + \beta \mathbf{x} + \varepsilon$.

The parameter estimation is achieved by EM algorithm where:

$$E\text{-step} \quad \hat{\Phi}_i = \begin{cases} \frac{(1-p)\pi_i}{(1-p)\pi_i + (1-\pi_i)} & \text{if } T_i = 0 \\ 1 & \text{if } T_i = 1 \end{cases}; \quad M\text{-step} \quad \begin{cases} \hat{\pi}_i = \frac{\exp(\hat{\alpha} + \hat{\beta} \mathbf{x}_i)}{1 + \exp(\hat{\alpha} + \hat{\beta} \mathbf{x}_i)} \\ \hat{p} = \frac{\sum_i T_i(\Phi_i + K_i)}{\sum_i \Phi_i + (M_i - 1)T_i} \end{cases}$$

If π is not modelled by explanatory variables, MLE solution is straightforward and $\hat{\pi} = \frac{1}{n} \sum_i \hat{\Phi}_i$. Differently, when a model structure is imposed on π , a logistic model solution has to be considered for the step of maximization.

3. An application to real data

The proposed method is applied to data from the Italian Multipurpose Survey 2000. The data set concerns 1,684 interviewers for whom the reinterview results were completely available (Mero and Pareto, 2005). The adopted model can be written as:

$$\text{logit}(\pi) = \beta_0 + \sum_{j=1}^7 \beta_j X_j + \varepsilon$$

where X_1 = number of completed interviews, X_2 = refusal rate, X_3 = non-contact rate, X_4 = rate of non-response for other reasons, X_5 = age ('0' = up to 30; '1' = over 30), X_6 = years of experience ('0' = up to 8; '1' = over 8) and X_7 = type of municipality ('0' = small / up to 50,000 inhabitants; '1' = large / over 50,000 inhabitants). No interactions were considered for simplicity.

Even though several methods can be used for estimating the standard error of $\hat{\beta}_j$, for practical reasons we propose the jack-knife method. Following this approach, each sample member is removed in turn from the data, thereby generating n separate samples each of size $n-1$. Let $\hat{\beta}_{j(i)}$ be the value of $\hat{\beta}_j$ obtained from the i th of these samples and let $\bar{\beta}_j$ denote the average of the n values $\hat{\beta}_{j(i)}$. Then the jack-knife estimate of the

standard error of $\hat{\beta}_j$ is given by $SE(\hat{\beta}_j) = \sqrt{\frac{n-1}{n} \sum_{i=1}^n (\hat{\beta}_{j(i)} - \bar{\beta}_j)^2}$.

The results are shown in Table 2.

Table 2: Parameter estimates, standard errors (SE) and odds ratios (OR)

Parameter	Estimate	SE	OR (95% CI)
Intercept	-5.248	0.817	-
Number of completed interviews	0.291	0.100	1.337 (1.099-1.627)
Refusals (%)	-0.056	0.019	0.946 (0.910-0.982)
Non-contacts (%)	0.022	0.018	1.022 (0.987-1.059)
Other non-responses (%)	0.264	0.096	1.302 (1.078-1.573)
Age	-2.192	0.973	0.112 (0.017-0.752)
Years of experience	2.904	0.933	18.249 (2.934-113.5)
Type of municipality	1.000	0.456	2.719 (1.113-6.642)

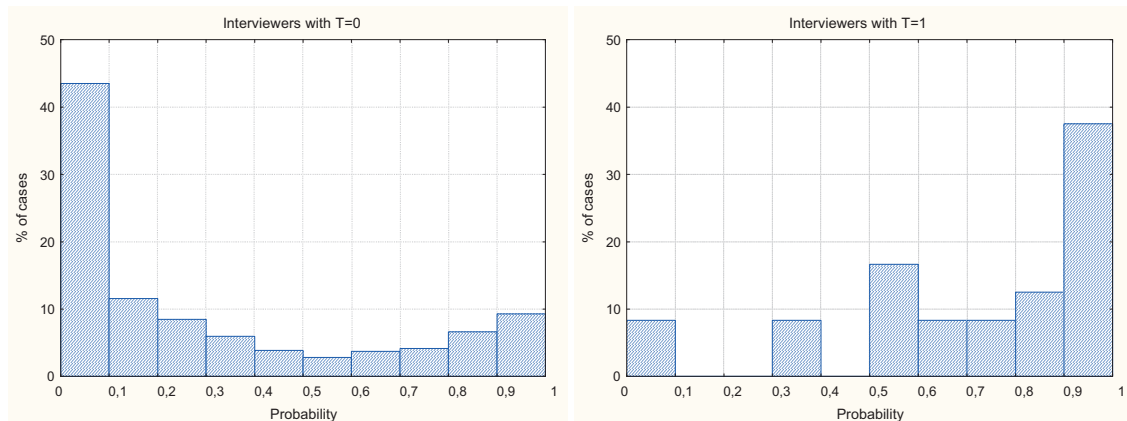
The number of completed interviews shows a positive effect on the probability that an interviewer is fraudulent (OR=1.337), i.e. the higher is the number of completed interviews, the larger becomes the probability of interviewer cheating. On the contrary, increasing refusal rate (OR=0.946) results in a less probability that an interviewer is dishonest. The non-contact rate can be considered not significant, since the 95% confidence interval (CI) of odds ratio includes 1. Moreover, age has a negative effect (lower probability for older interviewers), whereas interviewer experience has a strong

positive effect (much higher probability over 8 years of experience).

Regarding municipality effect, its corresponding odds ratio is 2.719, so implying a risk increase of 3 times for interviewers working in towns with more than 50,000 inhabitants when compared to small towns.

Figure 1 displays the distribution of the probability π that an interviewer is fraudulent by reinterview result ($T=0$ or $T=1$). Let us note that since T is not among explanatory variables, values of $\pi < 1$ are allowed also for interviewers showing $T=1$.

Figure 1: *Distribution of π by T*



The results show that about 45% of the interviewers where the initial reinterview did not detect a dishonest behaviour ($T=0$) have a value of $\pi < 0.1$, while almost 10% of them shows a probability to be fraudulent greater than 0.9 and should be further checked. Finally, more than 80% of the detected dishonest interviewers ($T=1$) obtain a predicted probability $\pi > 0.5$, with un peak of 38% greater than 0.9.

4. Conclusions

In this work we propose a method to improve the identification of cheating interviewers by a mixture model in which the probability that an interviewer is fraudulent is a function of a set of variables describing the interviewer characteristics and performance. A better detection protocol of interviewer charting behaviour based on this method could reduce the costs of questionnaire falsification in terms of biased data and improve survey quality.

References

- Mero C., Pareto A. (2005) Analisi e sintesi degli indicatori di qualità dell'attività di rilevazione nelle indagini campionarie sulle famiglie, *Contributi Istat*, 8/2005.
- Schäfer C., Schräpler J.P., Müller K.R. (2004) Identification, Characteristics and Impact of Faked and Fraudulent Interviews in Surveys, *European Conference on Quality and Methodology in Official Statistics*, Mainz.