A FIRST APPROACH TO NONRESPONSE BIAS IN THE INDUSTRIAL CLIMATE SURVEY

Abstract. We try to assess whether firms in a bad situation are more prone to stop responding to the survey.

1. INTRODUCTION

Nonresponse is one of the nonsampling errors that affect almost every survey. Its consequences range from relatively harmless (moderate variance increase) to very damaging (large bias). This depends both on the nonresponse rate and on whether the distribution of the variables in the nonresponding units is the same that in the rest of the population. When this happens, we say that the data are Missing At Random (MAR)[1]. In this situation, nonresponse does not cause bias, but otherwise it does.

These biases are much more serious a problem that variance increase, because they may easily go undetected. The smaller size of the effective sample caused by nonresponse produces an increase of the variance that can be accounted for by the same methods we use to estimate variance, but the biases caused by NMAR (Missing Not At Random) nonresponse cannot be estimated directly.

We want to check the hypothesis that nonresponse is MNAR. More specifically, we suspect that it might be happening that firms in a bad economic situation have a larger probability of not responding.

Since clearly, we cannot observe the answers of the firms that do not respond, we will use as a proxy the closest thing we have, their questionnaires from adjacent periods. Therefore, we will focus on firms that answer in month *t-1* but do not in month *t*. For these firms, we make a comparison between their answers in month t and the answers of the other firms in the same activity branch.

[1] The distinction between MAR and MCAR (*Missing Completely At Random*) is irrelevant if we consider complete nonresponse, as it is the case. See Rubin (1976).

In section 2, we argue the validity of using the answers from the previous month to detect MNAR. In section 3, we describe the model we will use to estimate the differences between distributions. In section 4 we show the results of the experiment and we conclude with some remarks in section 5.

2. The answers at t-1 as a proxy

To check that the answers at *t-1* are a good approximation to the ones they would have given in *t* in case they responded to the survey, we will estimate a simple Markov Chain model (see for example Meyn and Tweedie, 2012). For that, we need just to estimate, for each question, the probability that a firm gives answer *k* at period *t*, conditional to that it gave answer *k* al time t-1.

For example, in the question about production level, we have the conditional probabilities reported in table 1.

Table 1. Transition probabilities for Q10

	increase at t-1	**stability at t-1**	**decrease en t-1**
increase at t	0.9503	0.0064	0.0067
stability at t	0.0366	0.9802	0.0304
Decrease at t	0.0131	0.0134	0.9628

In table 1, we can check that the answers given by a specific firm are considerably stable. The overall probability of not changing their answer is 0.97.

Moreover, even if this probability were not so close to unity, it seems unlikely that the differences in the distribution at t vanished completely going a month back.

3. Model

We consider the interval from 2000 to 2014. In this period, we find 2 686 situations as described above, that is, when a firm answers at *t-1* but not at *t*. However, if we count cases by NACE division (two digits), we find too few to make comparisons.

Hence, we need a way to compare the distribution that combines the information of all divisions, but taking their differences into account.

If we were dealing with continuous variable, it would be relatively easy to make the comparison. For example, one can calculate the average of the difference

$$x_{hit} - \bar{x}_{ht}$$

or for positive variables, the average of the ratio

$$\frac{x_{hit}}{\bar{x}_{ht}}$$

where the subscript *h* indicates the division, *i* the firm and *t* the month. The average would be calculated across all the combinations *(h,i,t)* that satisfy the condition that there are data at *t* and nonresponse at *t+1.*

In our case, we are dealing with binary variables. We indicate by $x_{hit} = 1$ the event that firm *i* has given answer *k* to question *j* (for simplicity, we omit subscripts *j* and *k* in the remainder). To model this variable, we will use a logit or probit model with the form

$$\mathrm{P}[\mathrm{x}_{\mathrm{hit}} = 1] = \mathrm{L}(\mathrm{L}^{-1}(\bar{\mathrm{x}}_{\mathrm{ht}}) + \mathrm{b})$$

where *L* is the distribution function of t standard normal for the probit case or the logistic function for the logit case (see for example, Cameron and Trivedi, 2005).

That identity entails that if the probability of firm *i* giving answer *k* to question *j* is the same as the one in its NACE division, then coefficient *b* should be zero. If *b* were positive, that would mean that it is more probable that firm *i* would give that answer.

We fit this model with all the questionnaire data we mentioned (that is, n=2 686). The result of the estimation would tell us whether there are significant differences between the firms at are not going to respond the following month and the others.

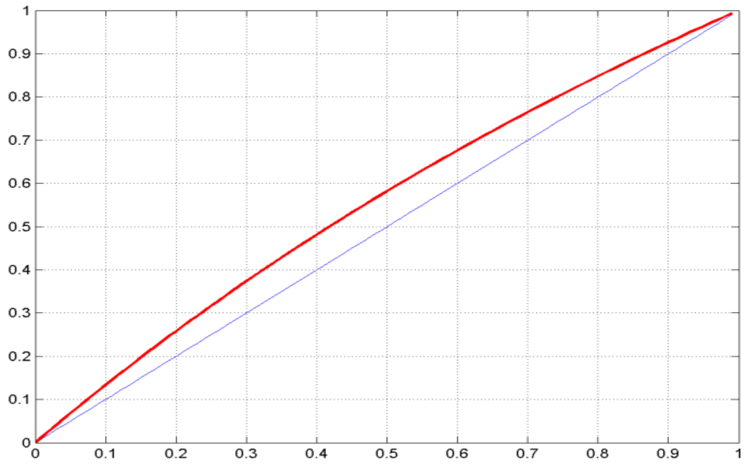
Coefficient *b* is estimated by maximum likelihood using the function glmfit of MATLAB.

4. RESULTS

Results with probit and logit models are very similar. In tables 2 and 3 we report the results for each of the questions and possible answers. It can be seen that the results are generally consistent with the idea that nonresponse is MAR.

To interpret the magnitude of the coefficients, we present an example of how a 0.3 coefficient would affect to the answer probability. In picture 1, we represent a curve in which the X axis represents the probability without the effect and the Y axis represents the probability with the effect. Therefore, for probabilities around 0.5, there is an increase of about 0.1.

Picture 1. Increase of probability



This large value of the coefficient is chosen to convey how this affects the results, but in tables 2 and 3 no values so large are observed. Moreover, the signs of the coefficients do not seem to be consistent with a reasonable interpretation, that is, that firms that are in a 'bad' situation are less likely to answer.

7. FINAL REMARKS

The conclusion of the exercise is not final, because it is possible that there are nonresponse biases that cannot be detected by our model. However, as a first-order approximation to the problem, we can take this as a sign that this kind of bias may not be a very pressing concern.

REFERENCES

Rubin, Donald B. (1976): "Inference and missing data", *Biometrika* 63.3, 581-592.

Meyn, Sean P. and Richard L. Tweedie (2012): *Markov chains and stochastic stability*, Springer Science & Business Media.

Cameron, A. Colin, and Pravin K. Trivedi (2005): *Microeconometrics: methods and applications*, Cambridge university press.

APPENDIX: TABLES

Table 2. Coefficientes of the logit models

		1	2	3
1	Total order books level	High	Normal	Weak
		-0.03 (0.02)	0.00 (0.01)	0.00 (0.01)
2	Domestic order books level	High	Normal	Weak
		-0.04 (0.02)	0.02 (0.01)	-0.02 (0.01)
3	Foreign order books level	High	Normal	Weak
		0.02 (0.02)	-0.03 (0.01)	-0.03 (0.01)
4	Order books tendency	Increase	Stable	Decrease
		0.02 (0.01)	-0.05 (0.01)	0.07 (0.01)
5	Domestic order books tendency	Increase	Stable	Decrease
		0.00 (0.02)	-0.05 (0.01)	0.07 (0.01)
6	Foreign order books tendency	Increase	Stable	Decrease
		0.06 (0.02)	-0.08 (0.01)	0.02 (0.02)
7	Finished products stock level	Excessive	Adequate	Insufficient
		0.00 (0.01)	-0.03 (0.01)	-0.02 (0.02)
8	Finished products stock tendency	Excessive	Adequate	Insufficient
		-0.01 (0.02)	-0.08 (0.01)	0.12 (0.02)
9	Raw materials stock level	Excessive	Adequate	Insufficient
		-0.20 (0.03)	-0.08 (0.02)	-0.37 (0.04)
10	Production level	Aumento	Estabilidad	Descenso
		0.04 (0.02)	-0.07 (0.01)	0.06 (0.01)
11	Production tendency	Increase	Stable	Decrease
		0.05 (0.01)	-0.09 (0.01)	0.09 (0.01)

Table 3. Coefficients of the probit models

		1	2	3
1	Total order books level	High	Normal	Weak
		-0.02 (0.01)	0.00 (0.01)	0.00 (0.01)
2	Domestic order books level	High	Normal	Weak
		-0.02 (0.01)	0.01 (0.01)	-0.01 (0.01)
3	Foreign order books level	High	Normal	Weak
		0.00 (0.01)	-0.02 (0.01)	-0.02 (0.01)
4	Order books tendency	Increase	Stable	Decrease
		0.01 (0.01)	-0.03 (0.01)	0.03 (0.01)
5	Domestic order books tendency	Increase	Stable	Decrease
		-0.00 (0.01)	-0.03 (0.01)	0.04 (0.01)
6	Foreign order books tendency	Increase	Stable	Decrease
		0.03 (0.01)	-0.05 (0.01)	0.01 (0.01)
7	Finished products stock level	Excessive	Adequate	Insufficient
		0.00 (0.01)	-0.02 (0.01)	-0.02 (0.01)
8	Finished products stock tendency	Excessive	Adequate	Insufficient
		-0.01 (0.01)	-0.05 (0.01)	0.06 (0.01)
9	Raw materials stock level	Excessive	Adequate	Insufficient
		-0.12 (0.01)	-0.07 (0.01)	-0.19 (0.02)
10	Production level	Aumento	Estabilidad	Descenso
		0.02 (0.01)	-0.04 (0.01)	0.03 (0.01)
11	Production tendency	Increase	Stable	Decrease
		0.03 (0.01)	-0.05 (0.01)	0.05 (0.01)