Dealing with frame imperfections in a business survey: an empirical study using Without Replacement Bootstrap

Ana Cristina M. Costa

Instituto Superior de Estatística e Gestão de Informação, Univ. Nova de Lisboa, Campus de Campolide, 1070-312 Lisbon, Portugal

ccosta@isegi.unl.pt

Abstract

This paper approaches issues related to frame problems and nonresponse in surveys. These nonsampling errors affect the accuracy of the estimates whereas the estimators become biased and less precise. We analyse some estimation methods that deal with those problems and give an especial focus to post-stratification procedures. We address the bootstrap methodology for variance estimation. Some applications of reweighting procedures and the Without Replacement Bootstrap algorithm proposed by Sitter (1992) are presented using data from the 1997 Annual Business Survey, conducted by Portugal's National Statistics Institute. The precision of the analysed estimators is discussed and some recommendations are made regarding its applications.

Keywords: Post-stratification; frame problems; nonresponse; reweighting; adjustment methods; bootstrap.

1. Introduction

The target population definition is particularly important during the design stage of a survey. The target population is the finite set of identifiable elements which statistical data should refer to, according to the objectives of the survey. A perfect sampling frame is an up-to-date list of all elements in the target population. Such complete, perfect and up-to-date information is in general difficult to obtain, especially when applying for business surveys.

Four relevant types of frame errors can be distinguished during the estimation stage (Lessler and Kalsbeek 1992, p. 48-51): undercoverage (missing units), overcoverage (inclusion of nonpopulation units), duplicate or multiple listings and incorrect auxiliary information (size, activity, location, etc.).

The Annual Business Survey (ABS) is a major survey conducted by Portugal's National Statistics Institute. Like most business surveys the ABS suffers from more than one category of

frame imperfections. The ABS design uses stratified simple random sampling without replacement and the sampling frame is a statistical business register.

Survey answers suggest that statistical units don't belong to strata defined over the sampling frame implying that there isn't a perfect correspondence between population strata and frame strata. Furthermore, some design weights are out of date (Costa, 2001) and the ABS data also faces the problem of unit nonresponse. These nonsampling errors affect the accuracy of the estimates whereas the estimators become biased and less precise.

To deal with the problem of missing data two strategies are common in survey practice, namely *reweighting* and *imputation*. In the former approach missing or incomplete units in the sample are ignored and the inclusion weights (or design weights) for responding units are adjusted by dividing them by estimates of the probability of response.

Adjustment methods that perform the *reweighting* of the design weights are usually used when unit nonresponse occurs. Among them, post-stratification estimation is often pointed out as an adequate method to handle frame problems as well (Little, 1986; Lazzeroni and Little, 1998; Gelman and Carlin, 2002).

Reweighting procedures are an appealing methodology to handle those problems since they aim to correct for known differences between sample and target population, whether these discrepancies arise from frame errors, nonresponse, sampling fluctuations, or other sources (Gelman and Carlin, 2002; Kalton and Flores-Cervantes, 2003).

In this paper, several weighted estimators are analysed in the design-based perspective and the Without Replacement Bootstrap (BWO) algorithm, proposed by Sitter (1992), is addressed for variance estimation. Section 2 briefly describes the Annual Business Survey design and outlines the reweighting procedures that are subject to analysis under the empirical study.

Applications of several reweighting schemes and the BWO algorithm are presented, using data from 1997 Annual Business Survey. The precision of the analysed estimators is discussed and some recommendations are made regarding its applications (section 3).

We hope that the empirical study may help the survey practitioners to choose alternative estimators that might improve the accuracy of the estimates that are currently produced.

2. Methodological framework

The design of the Annual Business Survey uses stratified simple random sampling without replacement, with stratification by region, activity, number of workers classification and legal classification. Estimation uses the Horvitz-Thompson estimator for this sampling scheme (Horvitz and Thompson, 1952). Changes in activity or geographical classification are dealt with Horvitz-Thompson domain estimators resulting in a less efficient sample.

The population was divided in two non-overlapping sub-populations. All statistical units with 100 or more workers (*major businesses*) were included in the sample. Statistical units with less than 100 workers (*medium and small businesses*) were selected through the sample selection scheme.

The performance of several reweighting schemes was investigated using 1997 ABS data concerning the sub-population of *medium and small businesses* in Portugal mainland. The variables used in the study were: *mean number of workers* (*V1*), *total sales* (*V2*) and *total services rendered* (*V3*). Item nonresponse doesn't occur for these variables.

The analysed estimators were the adjustment cell estimator, the post-stratified estimator and the post-stratified estimator with adjustment cells. Point estimates for the population total and the population mean were also computed through the Horvitz-Thompson (*HT*) estimator, although known biased under this survey. The next sections discuss those estimators.

2.1. Adjustment cell estimator

In the adjustment cell procedure the obtained sample (including respondents and nonrespondents) is divided in *H* exhaustive non-overlapping sub-populations called *nonresponse adjustment cells* and the response rates are estimated within each cell.

For the 1997 ABS data the nonresponse adjustment cells are defined by initial strata. We assume that all units within the same stratum have similar values for the considered variables and equal response probabilities.

Let $s_h = s_{1h} \cup s_{0h}$ denote the set of sample units belonging to the *h*th (*h = 1,…, H*) nonresponse adjustment cell (with sample size n_h); where s_{1h} is the subset of s_h composed by respondent units (with n_{1h} elements) and s_{0h} is the subset composed by nonrespondent units (with n_{0h} elements). The subscript *1* (one) refers to respondents and subscript *0* (zero) refers to nonrespondents. Let w_{hk} denote the *k*th element design weight of the *h*th adjustment cell. The symbol τ denotes the population total.

For an arbitrary sampling design, the adjustment cell estimator (*AC*) of the population total is

$$\hat{\tau}_{AC} = \sum_{h=1}^{H} \sum_{k=1}^{n_{1h}} \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk} y_{hk} \qquad (1)$$

with y_{hk} the value of the study variable *y* for the *k*th element of the *h*th adjustment cell and

$$\hat{N}_h = \sum_{k=1}^{n_h} w_{hk} \qquad (2)$$

$$\hat{N}_{1h} = \sum_{k=1}^{n_{1h}} w_{hk} \qquad (3)$$

2.2. Post-stratified estimator

Post-stratification consists in stratifying the sample data set after the sample has been selected using auxiliary information, namely the population post-strata sizes which can be derived from administrative registers or can be present in the frame at the moment that estimation takes place.

Post-stratification techniques are often used to increase the precision of estimates, in particular when the sample is selected by simple random sampling without replacement, and have been examined from different points of view by several authors such as Williams (1962), Holt and Smith (1979), Rao (1985), Valliant (1993), Leonard *et al.* (1994) and Rao (1994).

The 1997 ABS sample was stratified according to three schemes of post-stratification. On the first one, strata were formed by a *number of workers classification* (*scheme 1*) leading to 5 post-strata; on the second one by a *total sales classification* (*scheme 2*) inducing 2 post-strata and on the last one by a *workers/sales classification* (*scheme 3*) which lead to 10 post-strata.

In this case, the nonresponse adjustment cells are defined by pos-strata and therefore we assume that all units within the same post-stratum have equal response probabilities. The population post-strata sizes were provided by Portugal's National Statistics Institute (Machado and Costa, 2001).

Let N_i denote the known population size for post-stratum i ($i = 1,\ldots, L$); n_{1i} denote the number of respondent units within post-stratum i and w_{ik} denote the kth element design weight of the ith post-stratum. For an arbitrary sampling design, the post-stratified estimator (*PS*) of the population total is

$$\hat{\tau}_{PS} = \sum_{i=1}^{L} \sum_{k=1}^{n_{1i}} \frac{N_i}{\hat{N}_{1i}} w_{ik} y_{ik} \tag{4}$$

with y_{ik} the value of the study variable y for the kth element of the ith post-stratum (nonresponse adjustment cell) and

$$\hat{N}_{1i} = \sum_{k=1}^{n_{1i}} w_{ik} \tag{5}$$

The post-stratified estimator is denoted by *PS_s1*, *PS_s2* and *PS_s3* when it refers to one of the three post-stratification schemes mentioned above, respectively.

2.3. Post-stratified estimator with adjustment cells

A widely used method to deal with unit nonresponse consists in reweighting the design weights by the adjustment cell procedure and adjusting them afterwards by post-stratification. This technique will be named post-stratification with adjustment cells procedure.

For the 1997 ABS data the nonresponse adjustment cells are again defined by initial strata and therefore we assume that all units within the same stratum have similar values for the considered variables and equal response probabilities.

In this case, the sample was stratified according to two schemes of post-stratification: *scheme 1* with 5 post-strata and *scheme 3* with 10 post-strata, previously mentioned.

To compute the final weights, the design weights must be first adjusted within every nonresponse adjustment cell *h* (*h = 1,…, H*) by the adjustment cell procedure:

$$w_{hk}^{(AC)} = \frac{\hat{N}_h}{\hat{N}_{1h}} w_{hk} \ , \ k \in s_h , \ h = 1,...,H \tag{6}$$

with $\hat{N}_h$ and $\hat{N}_{1h}$ given by (2) and (3), respectively; and w_{hk} denotes the *k*th element design weight of the *h*th initial stratum (nonresponse adjustment cell). On the next step these weights are adjusted by post-stratification (*L* post-strata cut across nonresponse adjustment cells):

$$w_{ik}^{(ACPS)} = \frac{N_i}{\hat{N}_{1i}^*} w_{ik}^{(AC)} \ , \ k \in s_i , \ i = 1,...,L \tag{7}$$

with $w_{ik}^{(AC)}$ the adjusted weight (6) of the *k*th element within the *i*th post-stratum; s_i the set of sample units of the *i*th post-stratum and

$$\hat{N}_{1i}^* = \sum_{k=1}^{n_{1i}} w_{ik}^{(AC)} \tag{8}$$

For an arbitrary sampling design, the post-stratified estimator with adjustment cells (*ACPS*) of the population total is

$$\hat{\tau}_{ACPS} = \sum_{i=1}^{L} \sum_{k=1}^{n_{1i}} w_{ik}^{(ACPS)} y_{ik} \tag{9}$$

with $w_{ik}^{(ACPS)}$ the final weight (7) and y_{ik} the value of the study variable *y* for the *k*th element of the *i*th post-stratum.

The post-stratified estimator with adjustment cells is denoted by *ACPS_s1* and *ACPS_s3* when referring to one of the two post-stratification schemes mentioned above, respectively.

2.4. Variance estimation

As stated before, application of adjustment methods allows for publication of better quality estimates. However, it is important to compute variances of estimates in order to judge the accuracy and usefulness of those estimates.

Although adjustment methods are commonly used, for complex sampling designs it seems difficult to investigate the properties of estimators and design-based variance estimates usually rely on resampling methods (see, e.g., Yung and Rao, 2000; Valliant, 2003).

For simple random sampling without replacement, the post-stratified estimator has smaller mean squared error than the adjustment cell estimator (Little, 1986; Lessler and Kalsbeek 1992, p. 195-196). It is expected that this property also holds for other sampling schemes.

The estimator proposed by Rao (1985) for an arbitrary sampling design was used for variance estimation of the post-stratified estimator (4) as it agrees with known conditionally correct results, in the absence of nonresponse, in the special case of simple random sampling without replacement. Thus it may have good properties as well, in a conditional approach, if units have similar values for the considered variables and equal response probabilities within every post-stratum.

The Without Replacement Bootstrap (BWO) algorithm (Sitter, 1992) was used for variance estimation of estimators (1), (4) and (9), although alternative techniques could be considered such as linearization or jackknife-type methods (Lu and Gelman, 2003). Variance estimates were determined with the Monte Carlo approximation and 1000 bootstrap samples were drawn from the pseudo-population created by means of the BWO algorithm.

The next section presents and discusses the results of the empirical study. The data analysis for this paper was generated with specific programs developed by the author using SAS software[1].

3. Results of the empirical study and discussion

The performance of the described reweighting schemes was investigated using 1997 ABS data and summary results are presented on Table 1, Table 2 and Table 3. Detailed results and a similar study for 1996 ABS data can be found in Machado and Costa (2001).

As previously discussed, one would expect that post-stratification estimators perform better than the adjustment cell estimator (*AC*) but that hasn't occurred for the *PS_s2* estimator since population mean estimates seem to be highly biased. However, this is a natural conclusion as for that post-stratification scheme only two post-strata were defined, hence the assumption of post-strata homogeneity is false. For that reason bootstrap estimates for the *PS_s2* estimator weren't computed. As expected, bootstrap estimates for the *AC* estimator also seem biased.

Both post-stratification estimators (*PS* and *ACPS*) reveal a similar performance when using the same post-stratification scheme (*scheme 1* or *scheme 3*). Recall from sections 2.3 and 2.4

[1] Copyright, SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

that the *PS* estimator assumes post-strata as the nonresponse adjustment cells and that the *ACPS* estimator assumes initial strata as those cells. The pointed similarity derives from the fact that the variable used on *scheme 1* was used both for initial stratification and post-stratification.

It wasn't possible to compute bias estimates for these estimators and therefore it is difficult to state exactly which one is the proper post-stratification scheme (using *scheme 1* or *scheme 3*). When one must choose between post-stratification schemes, the option should take into consideration post-strata homogeneity. For that reason we were expecting that post-stratification *scheme 3* would hold better results than the other considered schemes. However, a closer look at standard deviation and coefficient of variation bootstrap estimates shows that the post-stratification *scheme 1* performed better than *scheme 3* (see Table 1, Table 2 and Table 3).

If other proper post-stratification schemes were available (assuming that population post-strata sizes were known) it would be interesting to investigate the performance of both post-stratification estimators since the results could be somewhat different from the ones stated under this study.

The bootstrap estimates for the coefficients of variation indicate that the *PS_s1* estimator performs a little better than the *ACPS_s1* estimator, except for variable *total sales* (*V2*). These results suggest that the *PS_s1* estimator holds better results than the other adjustment methods.

Observed similarities between bootstrap variance estimates and those ones computed using the estimator proposed by Rao (1985) are due to the fact that the number of units within the intersection of initial strata and post-strata is very large. However, if this doesn't happen or if other variables were used (other study or post-stratification variables) this estimator could perform worse.

The above discussion implies that the post-stratified estimator using the post-stratification *scheme 1* (*PS_s1*) and the variance estimator proposed by Rao (1985) are appropriate techniques under this survey. Note that in this case the computational effort for computing bootstrap estimates can be avoided.

4. Concluding remarks

The application of adjustment methods to the 1997 ABS data was motivated by issues related to frame problems and unit nonresponse. As expected from theoretical evidences and according to the above discussion we conclude that post-stratification methods perform better than the other procedures considered here. Moreover, the post-stratified estimator using post-stratification by a *number of workers classification* (*PS_s1*) turn out to be the most appropriate under this survey.

For the considered variables, the bootstrap variance estimates of the post-stratified estimator and those ones computed using the variance estimator proposed by Rao (1985) are similar and therefore the computational effort for computing bootstrap estimates can be avoided, particularly when preliminary survey results are required. However, we must call attention to the fact that this estimator underestimates the true variance of the post-stratified estimator, thus it may not be appropriate in other situations.

As a final remark we would like to draw attention to the fact that when a substantial amount of auxiliary information is available (which was not the present situation) a variety of complex weighting adjustments can be used to handle nonresponse and frame imperfections. See, for example, Kalton and Flores-Cervantes (2003).

Acknowledgments

The author is grateful to Professor José Ferreira Machado of the New University of Lisbon (Universidade Nova de Lisboa) for helpful discussions and suggestions.

This study was supported by the "Mudança de Estratos no IEH" Research Project of the Statistics and Information Management Research Centre (CEGI – Centro de Estatística e Gestão de Informação) of the New University of Lisbon (Universidade Nova de Lisboa) and Portugal's National Statistics Institute (INE – Instituto Nacional de Estatística).

References

Costa, A.C. (2001). Pós-estratificação e Without Replacement Bootstrap: Aplicações ao Inquérito às Empresas/Harmonizado. *Rev. Estatística* 3(3), 125–148.

Gelman, A. and Carlin, J.B. (2002). *Poststratification and weighting adjustments*. In: R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little (Eds.), Survey Nonresponse, John Wiley & Sons, New York, 289–302.

Holt, D. and Smith, T.M.F. (1979). Post stratification. *J. R. Stat. Soc. Ser. A-Stat. Soc.* 142, 33–46.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *J. Am. Stat. Assoc.* 47, 663–685.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics* 19(2), 81–97.

Lazzeroni, L.C. and Little, R.J.A. (1998). Random-effects models for smoothing poststratification weights. *Journal of Official Statistics* 14(1), 61–78.

Leonard, K.A., An, A.B., Nusser, S.M. and Breidt, F.J. (1994). *Approximating the variance of the survey regression estimator using poststratification*. In: American Statistical Association (Ed.), Proceedings of the 1994 Joint Statistical Meetings, Survey Research Methods Section, Vol. I, 222–227.

Lessler, J.T. and Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. John Wiley & Sons, New York.

Little, R.J.A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review* 54(2), 139–157.

Lu, H. and Gelman, A. (2003). A method for estimating sampling variances for surveys with weighting, poststratification, and raking. *Journal of Official Statistics* 19(2), 133–151.

Machado, J.F. and Costa, A.C. (2001). *Mudanças de Estrato nos Inquéritos às Empresas / Harmonizados*. Tech. Report, Instituto Superior de Estatística e Gestão de Informação, Univ. Nova de Lisboa, June 2001.

Rao, J.N.K. (1985). Conditional inference in survey sampling. *Survey Methodology* 11(1), 15–31.

Rao, J.N.K. (1994). *Resampling methods for complex surveys*. In: American Statistical Association (Ed.), Proceedings of the 1994 Joint Statistical Meetings, Survey Research Methods Section, Vol. I, 35–41.

Sitter, R.R. (1992). Comparing three Bootstrap methods for survey data. *The Canadian Journal of Statistics* 20(2), 135–154.

Valliant, R. (1993). Poststratification and conditional variance estimation. *J. Am. Stat. Assoc.* 88(421), 89–96.

Valliant, R. (2003). The effect of multiple weighting steps on variance estimation. Survey Methodology Program, University of Michigan, Working Paper Series, Accepted for publication by *Journal of Official Statistics*, April 2003, 35 p.

Williams, W.H. (1962). The variance of an estimator with post-stratified weighting. *J. Am. Stat. Assoc.* 57, 622–627.

Yung, W. and Rao, J.N.K. (2000). Jackknife variance estimation under imputation for estimators using poststratification information. *J. Am. Stat. Assoc.* 95, 903–915.

Table 1
Estimates for the population mean for variable *V1*
** Estimates computed using the estimator proposed by Rao (1985)*

Estimator	**Mean**	**Std. deviation**	***Bootstrap estimates***	
			Std. deviation	**Coeff. of variation (%)**
HT	2.03	0.0173	-	-
AC	2.71	–	0.0173	1.09
PS_s1	2.88	0.0173*	0.0200	0.67
ACPS_s1	2.85	-	0.0224	0.74
PS_s2	8.94	0.0872*	-	-
PS_s3	4.50	0.0346*	0.0400	0.84
ACPS_s3	4.46	-	0.0436	0.91

Table 2
Estimates for the population mean for variable *V2* (in 1000 PTE[2])
** Estimates computed using the estimator proposed by Rao (1985)*

Estimator	**Mean**	**Std. deviation**	***Bootstrap estimates***	
			Std. deviation	**Coeff. of variation (%)**
HT	21060.76	377.65	-	-
AC	27868.98	-	283.65	1.84
PS_s1	30319.91	528.50*	626.12	1.94
ACPS_s1	29668.95	-	680.90	2.04
PS_s2	124205.97	2018.08*	-	-
PS_s3	79742.23	1630.53*	1928.42	2.29
ACPS_s3	79398.59	-	1918.28	2.36

Table 3
Estimates for the population mean for variable *V3* (in 1000 PTE)
** Estimates computed using the estimator proposed by Rao (1985)*

Estimator	**Mean**	**Std. deviation**	***Bootstrap estimates***	
			Std. deviation	**Coeff. of variation (%)**
HT	5735.40	165.17	-	-
AC	8007.16		104.10	2.56
PS_s1	8043.40	238.51*	243.30	2.87
ACPS_s1	8336.61	-	221.17	2.59
PS_s2	30209.56	906.20*	-	-
PS_s3	16772.26	828.23*	850.26	4.88
ACPS_s3	17668.88	-	661.02	4.09

[2] PTE stands for Portuguese Escudo (former currency).