Alternative Weighting Options in Regression Analysis of Survey Data

Chris Skinner
London School of Economics and Political Science, London, United Kingdom - c.j.skinner@lse.ac.uk

Abstract

Survey weighting may be desirable when estimating regression models if sampling is informative. The paper reviews alternative approaches to weighting when fitting regression models to public use survey data. The focus will be on two main approaches to improving estimation efficiency whilst avoiding biasing effects of informative sampling: (i) stabilizing weights using functions of the explanatory variables; (ii) smoothing weights using functions of the dependent (as well as explanatory) variables.

Keywords: calibration; smoothed weight; stabilized weight; variance estimation.

1. Introduction

Survey weights are often used in regression analysis of survey data to ensure consistent estimation of regression coefficients when sampling is informative, that is when sample inclusion may be related to the outcome variable conditional on covariates (Fuller, 2009, Sect. 6.3). Thus, for a standard linear model

$$y_i = \mathbf{x}_i'\beta + e_i. \tag{1}$$

the survey weight w_i may be used in a weighted least squares estimator $\hat{\beta} = (\sum w_i \mathbf{x}_i \mathbf{x}_i')^{-1} \sum w_i \mathbf{x}_i y_i$ of β.

A number of approaches have been proposed to modify general-purpose survey weights for use in specific regression analyses, primarily with a view to improving efficiency. In this paper, we first illustrate this idea of weight modification by reference to calibration and then focus on two approaches: (i) stabilizing weights using functions of the explanatory variables; (ii) smoothing weights using functions of the dependent (as well as explanatory) variables.

We frame our discussion within a setting where a researcher has access to a public use dataset and wishes to use survey routines in standard statistical software to conduct regression analysis. We explain this context further in the next section.

2. Public Use Data

We assume that the researcher has access to a microdata file, which contains n records corresponding to responding sampled units in a set denoted $s = \{1, \ldots, n\}$. The record for unit $i \in s$ contains values $(y_i, \mathbf{x}_i)$ of the the outcome and explanatory variables, respectively, and a survey weight w_i, which may be used as above in the weighted least squares estimator $\hat{\beta}$ of β for model (1). Moreover, it is assumed that the file contains further identifiers or replicate weights, which enable valid variance estimation. These might consist of primary sampling unit and stratum identifiers, plus possibly finite population corrections and further (e.g. secondary) sampling unit identifiers, as are used in standard survey software to construct linearization variance estimators for standard stratified multistage designs, or they may consist of a series of replicate weights, $w_i^{(1)}, \ldots, w_i^{(B)}$, to enable construction of replication variance estimators.

We shall assume that the survey weights and identifiers/replicate weights enable consistent point and variance estimation for relevant regression parameters. In practice, unit non-response arises in most surveys and our notion of consistency here refers not only to repeated sampling under a probability design but also to

the non-response mechanism. We suppose that the survey weights w_i are designed to correct for bias from nonresponse as well as from sampling. Consistent variance estimation, in the presence of weighting for both nonresponse and sampling, can in fact be quite complex (e.g. Kim and Kim, 2007) but we shall, nevertheless, assume that standard variance estimators (as are used in standard survey software) using the identifiers in the public use data will provide consistent variance estimation.

We shall be interested in transformations of the survey weights w_i which can be undertaken with the public use file. For practical purposes, it is desirable not only that the weighted estimator $\hat{\beta}$ remains consistent for β under transformation of the w_i, but also that consistent variance estimation for $\hat{\beta}$ can still be achieved either, for linearization variance estimators by applying the same approach as for the original weighted estimator using the survey identifiers on the file or, for replication variance estimators, through some natural corresponding transformation of the replicate weights. The latter case typically involves applying the same transformation to the replicate weights as to the basic weight w_i and this is illustrated in the next section for calibration.

3. Calibrated Weights

Weight transformation may be illustrated by the case of calibration. Suppose the user of the data file has available the population totals $t_{\mathbf{z}}$ of a vector of variables $\mathbf{z}$ which are all included in the file and, for which, it may be assumed that $\hat{t}_{\mathbf{z}} = \sum_s w_i \mathbf{z}_i$ is consistent for $t_{\mathbf{z}}$. In this case, weight calibration may be of interest (Lumley and Scott, 2017). This is achieved by transforming the weights w_i into calibrated weights w_{ci} which minimise a measure of distance between the w_{ci} and the w_i such that the calibration constraint $\sum_s w_{ci}\mathbf{z}_i = t_{\mathbf{z}}$ holds. Following Deville and Särndal (1992), the user might take $w_{ci} = w_i F(\mathbf{z}_i'\lambda)$, where $F(.)$ is a specified function, such as $F(\mathbf{z}_i'\lambda) = 1 + \mathbf{z}_i'\lambda$ in the case of generalized regression estimation, and λ is determined by the calibration constraint. Calibrated replicate weights may then be constructed similarly as $w_{ci}^{(b)} = w_i^{(b)} F(\mathbf{z}_i'\lambda^{(b)})$, where $\lambda^{(b)}$ is determined by the calibration constraint $\sum_s w_{ci}^{(b)}\mathbf{z}_i = t_{\mathbf{z}}$. The replication variance estimator remains valid when the weights w_i and $w_i^{(b)}$ are replaced by w_{ci} and $w_{ci}^{(b)}$ (Rust and Rao, 1996).

Adaptation of the public use file and the calibration weights for linearization variance estimation seems somewhat less straightforward, unless calibration is explicitly handled in the survey software. Rao et al. (2002) discuss how this estimator needs to be modified as a result of calibration. The replacement of the weights w_i by the calibrated weights w_{ci} provides part of the modification needed but, more importantly, the regression residuals in model (1) need themselves to be regressed on $\mathbf{z}_i$ to construct modified residuals.

4. Stabilized Weights

Several authors have proposed modifying the weight w_i by a function of $\mathbf{x_i}$. Specifically, under the model in (1), where e_i has expectation zero and constant variance, Skinner and Mason (2012) propose to replace w_i in a survey weighted least squares estimator of β by $w_i\hat{q}_i$, where $q_i = E(w_i \mid \mathbf{x}_i)^{-1}$ and the estimate $\hat{q}_i$ of q_i is obtained from fitting an auxiliary weight model relating w_i to $\mathbf{x}_i$. Related approaches are discussed by Magee (1998), Pfeffermann and Sverchkov (1999), Fuller (2009, sect. 6.3) and Kim and Skinner(2013). The term 'stabilized weights' is recommended by Lumley and Scott (2017), following Robins et al. (2000).

Stablizing weights can offer major efficiency gains, especially for sampling designs where there are 'design' variables amongst the covariates $\mathbf{x_i}$, which strongly influence sample inclusion probabilities, for example such as in probability proportional to size sampling in business surveys, where the size of a business may be common covariate. Weight stabilization also has the attractive robustness property that the weighted estimator of β remains consistent even if the auxiliary weight model is misspecified.

It is clear that the fitting of the auxiliary model and hence weight stabilization can be implemented using a public use file. As discussed by the authors above, the error in estimating the auxiliary weight model can be ignored to first order when estimating the variance of the weighted least squares estimator of β with weights $w_i\hat{q}_i$. It follows that this variance can be estimated consistently using a linearization variance estimator

simply by modifying the weight w_i and without further modification of the file or the variance estimation procedure.

Similarly, the replication variance estimator should remain valid if, in addition, the replication weights are scaled in the same way as w_i by $\hat{q}_i$. This differs from calibration in that we do not propose here to fit the auxiliary model $E(w_i \mid \mathbf{x}_i)$ anew for each replicate, using a replicate weight in place of the dependent variable w_i. It does not appear that this would be helpful, in particular given the occurrence of zero weights for many replication methods.

5. Smoothed Weights

Beaumont (2008) proposed smoothing Horvitz-Thompson weights w_i when estimating the total of a variable y_i by smoothed weights $\tilde{w}_i = E(w_i \mid y_i, I_i)$, where I_i is the sample inclusion indicator and the auxiliary weight model $E(w_i \mid y_i, I_i)$ requires estimating, as for stabilized weights. Kim and Skinner (2013) proposed extending this approach for a regression model, such as in (1), by taking the smoothed weights as $\tilde{w}_i = E(w_i \mid \mathbf{x}_i, y_i, I_i)$. They proposed alternative auxiliary weight models, particularly of parametric form $E(w_i \mid \mathbf{x}_i, y_i, I_i) \equiv \tilde{w}(\mathbf{x}_i, y_i; \phi)$, and approaches to estimating such models.

Unlike for stabilized weights, the weighted estimator of β based on smoothed weights does not remain consistent in general under misspecification of the auxiliary weight model (Kim and Skinner, 2013) and this may be seen as a disadvantage. On the other hand, there may be significant efficiency gains of the smoothed estimator when the weights w_i exhibit major variation in ways that are unrelated to y_i given $\mathbf{x_i}$.

As for stabilized weights, it is clear that weight smoothing can be implemented using a public use file. Assuming a parametric model $\tilde{w}_i \equiv \tilde{w}(\mathbf{x}_i, y_i; \phi)$, the basic smoothed weights may be expressed as $\tilde{w}(\mathbf{x}_i, y_i; \hat{\phi})$, where $\hat{\phi}$ is obtained by regressing w_i on $\mathbf{x}_i$ and y_i using the sample data. Note that, given the conditioning on $I_i = 1$, it is not necessary to use survey weighting in the estimation of ϕ.

Variance estimation is more complex than with stabilized weights since errors in the estimation of the auxiliary weight model cannot be ignored. Kim and Skinner (2013) derive a linearization variance estimator which could, in principle, be used with a public use file. This variance estimator simplifies somewhat if the sampling fraction n/N is negligible, in which case the variance estimator could be obtained from standard software, using modified residuals, as for calibrated weights.

As to replication variance estimation, Beaumont (2008) does propose a bootstrap approach for the problem of estimating a population total and it is possible that this could be extended to the regression case. As for linearization, it seems likely that this will be more straightforward in the case when n/N is negligible, although in this case, it is still necessary to account for the error in estimating the auxiliary weight model. It is possible that this might be achieved by constructing smoothed replicate weights $w_i^{(b)}$ by regressing $w_i^{(b)}$ on $\mathbf{x}_i$ and y_i, but this requires further research.

References

Beaumont, J. F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, **95**, 539-53.

Deville, J. C., & Särndal, C. E. (1992). Calibration estimators in survey sampling. *J. Amer. Statist. Ass.*, **87**, 376-82.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ. Wiley.

Kim, J. K., & Kim, J. J. (2007). Nonresponse weighting adjustment using estimated response probability. *Canad. J. Statist.*, **35**, 501-14.

Kim, J. K., & Skinner, C. J. (2013). Weighting in survey analysis under informative sampling. *Biometrika*, **100**, 385-98

Lumley, T., & Scott, A. (2017). Fitting regression models to survey data. *Statistical Science*, **32**, 265-78.

Magee, L. (1998). Improving survey-weighted least squares regression. *J. Roy. Statist. Soc. B*, **60**, 115-26.

Pfeffermann, D., & Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya B*, **61**, 166-86.

Rao, J. N. K., Yung, W., & Hidiroglou, M. (2002). Estimating equations for the analysis of survey data using poststratification information. *Sankhyā, A*, **64**, 364-78.

Robins, J. M., Hernán, M., & Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, **11**, 550-60.

Rust, K. F., & Rao, J. N. K. (1996). Variance estimation for complex surveys using replication techniques. *Stat. Methods Med. Res.*, **5**, 283-310.

Skinner, C., & Mason, B. (2012). Weighting in regression analysis of surey data wiht a cross-national application. *Canad. J. Statist*, **40**, 697-711.