Statistical Science
2007, Vol. 22, No. 2, 179–183
DOI: 10.1214/088342307000000168
Main article DOI: 10.1214/088342306000000691

Comment: Struggles with Survey Weighting and Regression Modeling

Danny Pfeffermann

This is an intriguing paper that raises important questions, and I feel privileged for being invited to discuss it. The paper deals with a very basic problem of sample surveys: how to weight the survey data in order to estimate finite population quantities of interest like means, differences of means or regression coefficients.

The paper focuses for the most part on the common estimator of a population mean, $\bar{y}_w = \sum_{i=1}^n w_i y_i / \sum_{i=1}^n w_i$, and discusses different approaches to constructing the weights by use of linear regression models. These models vary in terms of the number and nature of the regressors in the model and in the assumptions regarding the regression coefficients, whether fixed or random with prespecified distributions. The idea behind regression weighting is to include in the regression model all the variables and interactions that are related to the outcome values and affect the sample selection and the response probabilities, such that the sampling and response mechanisms are ignorable in the sense that the model fitted to the observed data is the same as the population model before sampling. Assuming that all the important regressors affecting the sample selection and response are discrete, the set of all possible combinations of categories of these variables defines poststratification cells, which in turn define the dummy independent variables in the regression model. The target population parameter of interest can be written then as $\theta = \sum_{j=1}^J N_j \theta_j / \sum_{j=1}^J N_j$, where θ_j is the parameter for cell j (say the true cell mean, $\bar{Y}_j$), N_j is the cell size and J is the number of cells. The regression estimator has the general form $\hat{\theta}^{PS} = \sum_{j=1}^J N_j \hat{\theta}_j / \sum_{j=1}^J N_j$. For example, the estimator of the population mean is $\hat{\bar{Y}}^{PS} = \sum_{j=1}^J N_j \bar{y}_j / \sum_{j=1}^J N_j$, where $\bar{y}_j$ is the sample mean in cell j.

The discussion that follows is divided into two parts. In the first part I comment on the proposed weighting approach and some of the statements made in the article. In the second part I consider another approach for analyzing survey data that are subject to unequal sample selection probabilities and nonresponse, and compare it to the approach taken in this paper.

Danny Pfeffermann is Professor, Department of Statistics, Hebrew University of Jerusalem, Jerusalem 91905, Israel and Professor, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, United Kingdom (e-mail: msdanny@huji.ac.il).

1. REMARKS ON THE PAPER

• The first obvious remark, that is also made already in the Abstract, is that the number of poststratification cells can be extremely large, with inevitably very small or no samples in many of the cells. Having small or no samples in some or even most of the cells is theoretically not a problem under the model with random regression coefficients considered later, but it is not clear what should be done in such a case under the standard regression model with fixed coefficients. Note in particular the problems arising if the zero sample sizes are due to nonresponse. Deleting these cells from the regression model may violate the sample ignorability assumption. It is stated in Section 3.1 that it is not required to include in the model all the cells, but this raises the question of which cells to exclude and based on what criteria. It may imply also including different cells (interactions) for different outcome variables of interest.

• It is assumed that the cell sizes are known. This could be a strong assumption in a large-scale survey with many small cells. Also, it is often the case that the cell sizes are known to the person drawing the sample, but not necessarily to the person analyzing the data, who has limited access to the original files due to confidentiality restrictions or other reasons. The argument that the cell sizes can be estimated using iterative proportional fitting or other related procedures is well taken, but this raises questions of the effect of using estimated sizes on the performance of the estimators and how to estimate the variances, accounting for this source of variability.

• A third problem and in a way the most difficult one to handle is the implicit assumption that the analyst knows all the variables affecting the sample selection and nonresponse. Here again a distinction should be made between the person drawing the sample who