Statistical Science
2007, Vol. 22, No. 2, 168–170
DOI: 10.1214/088342307000000195
Main article DOI: 10.1214/088342306000000691
© Institute of Mathematical Statistics, 2007

Comment: Struggles with Survey Weighting and Regression Modeling

F. Jay Breidt and Jean D. Opsomer

We congratulate the author on an informative and thought-provoking discussion on a topic of broad interest to the statistics community: the fitting of models to data collected through complex surveys. The number of papers written on this topic, whether from a model-based or design-based perspective, is substantial and goes back at least to Konijn (1962). This topic has led to some disagreements between those advocating that the design best be ignored when the primary interest is on the characteristics of the model, and those stating that the design cannot be ignored. More recently, both sides of this discussion have moved to something approaching a consensus, with those favoring a model-based approach acknowledging the need to account for nonignorable designs in the model fitting, while the traditional design-based view has been extended to explore certain circumstances under which it is appropriate to ignore the design.

The current article is an excellent example of those recent discussions of why the design needs to be accounted for in modeling, and how this can be done in practice. The importance of *fully* accounting for the design by incorporating all relevant interactions provides a good motivation for the discussion of the range of methods in the article. It also stresses other aspects of importance to people working with survey data, in particular the desirability of maintaining scale/location invariance and linearity of the model-based estimators. This ensures consistency of estimates for different variables in the survey, as well as additivity over domains within the population. (As an aside, the poststratified estimator arising from logistic regression in Section 3.2 can be modified to yield approximate weights by the method proposed in Wu and Sitter, 2001.)

The article mentions a number of disadvantages of design-based (weighted) model fitting and inference. Weights are viewed as complicated and mysterious, in the sense that the modeler often does not know how they were constructed and hence might not want to rely on them when it comes to model specification and estimation. Estimation, and especially variance estimation, are viewed as more cumbersome under the design-based paradigm compared to a model-based analysis. In what follows, we will argue that a weighted analysis offers some distinct advantages and might actually reduce the complexity of the analysis in many cases, at least from the perspective of a statistician interested in using previously collected and weighted survey data to fit a model.

A key feature of the design-based paradigm (broadly speaking) is that it makes it possible to separate design and postsample adjustments from data analysis. Individuals tasked with creating survey weights are typically within the organization collecting the data, and will be referred here as "the survey statisticians." They have knowledge of the sampling design and have access to detailed information on the nonresponse characteristics of the sample and to relevant auxiliary information. Based on these sources of information, they develop a set of survey weights (and sometimes also produce sets of replication weights for variance estimation). As noted in the article, these weights are often much more complicated than simple inverses of inclusion probabilities, and in fact reflect the best effort on the part of the survey statisticians creating the weights to account for nonresponse and incorporate potentially useful population-level information. These weights are appended to the dataset, which is then made available to individuals interested in analyzing those data. These individuals will be referred to as "the data analysts."

From the perspective of the data analysts, using these weights is convenient in the sense that they provide a simple way to account for the way the data were obtained, without requiring the data analysts to replicate many of the tasks of the survey statisticians. Overall, this "division of labor" allows both sets of statisticians to focus their efforts on the portion of the overall problem of most immediate interest to them, and for which

F. Jay Breidt is Professor and Chair, Department of Statistics, Colorado State University, Fort Collins, Colorado 80523, USA (e-mail: jbreidt@stat.colostate.edu). Jean D. Opsomer is Professor, Department of Statistics, Iowa State University, Ames, Iowa 50011, USA (e-mail: jopsomer@iastate.edu).