Statistical Science
2007, Vol. 22, No. 2, 165–167
DOI: 10.1214/088342307000000177
Main article DOI: 10.1214/088342306000000691

Comment: Struggles with Survey Weighting and Regression Modeling

Robert M. Bell and Michael L. Cohen

Andrew Gelman's article "Struggles with survey weighting and regression modeling" addresses the question of what approach analysts should use to produce estimates (and associated estimates of variability) based on sample survey data. Gelman starts by asserting that survey weighting is a "mess." While we agree that incorporation of the survey design for regression remains challenging, with important open questions, many recent contributions to the literature have greatly clarified the situation. Examples include relatively recent contributions by Pfeffermann and Sverchkov (1999), Graubard and Korn (2002) and Little (2004). Gelman's paper is a very welcome addition to that literature.

There are some understandable reasons for the current lack of resolution. First, U.S. federal statistical agencies have been historically limited by their mission statements to producing statistical summaries, primarily means, percentages, ratios and cross-classified tables of counts. This is one explanation for why Cochran (1977) and Kish (1965) devote the great majority of their classical texts to these estimates. As a result, the job of using regression and other more complex models to learn about any causal structure underlying these summary statistics was generally left to sister policy agencies and outside users.

However, things are changing. The federal statistical system (whether it likes it or not) is becoming more involved with complex modeling. This includes small-area estimation (e.g., unemployment estimates and census net undercoverage estimates) and research into models combining information from surveys with administrative data. (There will also likely be increased demands to use data mining procedures on federal statistical data.) This relatively new development has likely motivated several of the recent contributions on how to account for the sample design in complex models. Therefore, Gelman's article and the resulting discussion come at an important time.

Another reason for the failure to resolve this class of problems is that this general issue is not easy. Attempts to resolve this problem raise a number of clashing perspectives, including: (1) whether to be model-based or design-based in one's inference, (2) whether to take a Bayesian or a frequentist view, (3) whether one's inference should be conditional on (some of) the observed values of the design variables and other auxiliary data that one might have for the full population, (4) whether one evaluates a procedure based on its small-sample performance or its asymptotic properties, and (5) whether one wants an algorithm specific to a particular regression model or something more omnibus.

A variety of general schemes have been proposed to deal with this hard problem, and several of them can be expressed as members or mixtures of the following pure strategies: (1) use an unweighted analysis of the collected data, which is a pure model-based perspective assuming the model is correct for the entire (super) population, (2) use the inverses of the sample selection probabilities as weights, which derives from a pure design-based perspective and is therefore not dependent on model-based assumptions either, and (3) include the survey design in the model as predictors (Little, 2004). The last strategy, for instance, would make sense if it was obvious that separate models were needed for subgroups defined by the survey variables. Gelman's paper represents a mixture of strategies (2) and (3).

It is useful to take a closer look at the second example in Section 1.4 of Gelman's article, which addresses the bias of the race coefficient for predicting log income when the sample is unrepresentative of the population in terms of gender. Like Gelman, we are viewing the problem as one of estimating the "so-called" census regression coefficient, which in this case is the mean log income for whites minus the mean log income for nonwhites in the finite population. Some algebra shows that conditional on the population margins

Robert M. Bell is Member, Statistics Research Department, AT&T Labs–Research, 180 Park Avenue, Florham Park, New Jersey 07932, USA (e-mail: rbell@research.att.com). Michael L. Cohen is Study Director, Committee on National Statistics, National Academies, Room 1135 Keck Center, 500 5th St., N. W., Washington, District of Columbia 20001, USA (e-mail: mcohen@nas.edu).