Statistical Science
2007, Vol. 22, No. 2, 184–188
DOI: 10.1214/088342307000000203
Main article DOI: 10.1214/088342306000000691

Rejoinder: Struggles with Survey Weighting and Regression Modeling

Andrew Gelman

1. MOTIVATIONS

I was motivated to write this paper, with its controversial opening line, "Survey weighting is a mess," from various experiences as an applied statistician:

- Encountering this sort of statement in the documentation of opinion poll data we were analyzing in political science: "A weight is assigned to each sample record, and MUST be used for all tabulations." (This particular version was in the codebook for the 1988 CBS News/New York Times Poll; as you can see, this is a problem that has been bugging me for a long time.) Computing weighted averages is fine, but weighted regression is a little more tricky—I do not really know what a weighted logistic regression likelihood, for example, is supposed to represent.
- Constructing the weighting for the New York City Social Indicators Survey (SIS). It quickly became clear that we had to make many arbitrary choices about inclusion and smoothing of weighting variables, and we could not find any good general guidelines.
- We wanted to estimate state-level public opinion from national polls. If our surveys were simple random samples, this would be basic Bayes hierarchical modeling (with 50 groups, possibly linked using state-level predictors). Actually, though, the surveys suffer differential nonresponse (lower response by men, younger people, ethnic minorities, etc.) as signaled to the user (such as myself) via a vector of weights.

The weighting in others' surveys, as well as our own SIS, appeared to be a mess. In particular, different survey organizations weight on different variables, and use different smoothing of weights, even when using similar methodology to survey the same population (Voss, Gelman and King, 1995). The weights are clearly not the platonic inverse-selection probabilities envisioned in some of the classical statistical theory of sampling.

Having established that survey weighting is a mess, I should also acknowledge that, by this standard, regression modeling is also a mess, involving many arbitrary choices of variable selection, transformations and modeling of interaction. Nonetheless, regression modeling is a mess with which I am comfortable (Gelman and Hill, 2007) and, perhaps more relevant to the discussion, can be extended using multilevel models to get inference for small cross-classifications or small areas.

I was thus motivated to get the benefits of weighting—adjusting for expected or known differences between sample and population—in the familiar and expandable context of regression modeling. As indicated by the title of the paper, we are not there yet. I am thrilled to have my paper discussed by leaders in survey research who have made so many important contributions in the theory and practice of survey analysis, and I hope this discussion helps us move the field forward, both toward my ultimate goal of a unified design-based and model-based analysis, and toward the intermediate goal of identifying weak points of currently used weighting and poststratification adjustments.

2. SAMPLE SELECTION PROBABILITIES

Unequal probabilities of inclusion in a survey arise in three ways: stratification or multiple frames (so that units in different strata have different selection probabilities, perhaps unavoidably or perhaps by design), clustering and nonresponse. Unfortunately, surveys sometimes simply supply a weight without explaining where it came from. This can allow consistent estimates using weighted regression, but, as several discussants note, design knowledge is needed in order to correctly compute standard errors.

Nonresponse can contaminate selection probabilities that otherwise would be simple. For example, Table 1 shows nominal inverse-probability weights and actual poststratification weights for households of different sizes from two different national pre-election surveys

Andrew Gelman is Professor of Statistics and Professor of Political Science, Department of Statistics, Columbia University, New York, New York 10027, USA (e-mail: gelman@stat.columbia.edu).