Statistical Science
2007, Vol. 22, No. 2, 175–178
DOI: 10.1214/088342307000000159
Main article DOI: 10.1214/088342306000000691

Comment: Struggles with Survey Weighting and Regression Modeling

Sharon L. Lohr

In the ideal samples of survey sampling textbooks, weights are the inverses of the inclusion probabilities for the units. But nonresponse and undercoverage occur, and survey statisticians try to compensate for the resulting bias by adjusting the sampling weights. There has been much debate about when and whether weights should be used in analyses, and how they should be constructed. Professor Gelman deserves thanks for clarifying the discussion about weights and for raising interesting issues and questions.

If we use weights in estimation, what would we like them to accomplish? Here are some desirable properties:

1. The mean squared error (MSE) of estimators is smaller if the weights are used than if the weights are not used.
2. Estimators produced using the weights are internally consistent. Thus, if $\hat{Y}_1$ is the estimated total medical expense for men in the population, $\hat{Y}_2$ is the estimated total medical expense for women in the population and $\hat{Y}_3$ is the estimated total medical expense for everyone in the population, then $\hat{Y}_1 + \hat{Y}_2 = \hat{Y}_3$.
3. We may have independent population counts from a census or administrative data source for sex, age, race/ethnicity and other variables. If we apply the weights to estimate these quantities, the estimates equal the true population counts. We refer to this as the calibration property.
4. The weight for unit i in the sample can be thought of as the number of population units represented by unit i.
5. The estimators have optimal properties under superpopulation models that are thought to fit the data.
6. The estimators are robust to misspecifications of the superpopulation models.
7. The procedure for constructing the weights is objective and transparent.

All of these are good properties. The problem is that one can only rarely construct a set of weights that satisfies all of them simultaneously.

In this discussion, we distinguish between design weights and weighting adjustments used for poststratification. The design weights are

$$d_i = \frac{1}{P(\text{unit } i \text{ included in sample})}.$$

The design weight d_i is a property of unit i; under design-based inference, it is a fixed constant. If two samples are drawn independently using the same probability sampling design and if each sample includes unit i, the weight d_i for the unit is the same in each sample. Poststratification weight adjustments, however, depend on the selected sample $\mathcal{S}$. In the simplest case of ratio adjustment, we multiply each sampling weight d_i by the factor $g_i(\mathcal{S}, x) = X/\hat{X}$, where X is the known population total of auxiliary variable x and $\hat{X} = \sum_{i\in\mathcal{S}} d_i x_i$. The resulting weight is $w_i(\mathcal{S}, x) = d_i g_i(\mathcal{S}, x)$; the weight depends on the sample selected and on the auxiliary variable x through the estimated total $\hat{X}$. The ratio estimator of the population total is then $\hat{Y}_r = (X/\hat{X})\sum_{i\in\mathcal{S}} d_i y_i = \sum_{i\in\mathcal{S}} w_i(\mathcal{S}, x) y_i$. Similarly, for generalized regression estimation,

$$g_i(\mathcal{S}, \mathbf{x}) = 1 + (\mathbf{X} - \hat{\mathbf{X}})^T \left(\sum_{j\in\mathcal{S}} d_j \mathbf{x}_j \mathbf{x}_j^T / c_j\right)^{-1} \mathbf{x}_i / c_i,$$

where the scaling constant c_i may depend on $\mathbf{x}$. For the special case of poststratification, $g_i(\mathcal{S}, \mathbf{x}) = N_c/\hat{N}_c$ for observation i in poststratification class c. Thus, for poststratification, the weight adjustments are positive; for general regression models, however, the weight adjustments are unrestricted. The weight $w_i(\mathcal{S}, x)$ varies from sample to sample. Since the weight adjustment depends only on x, though, and not on y, the weight $w_i(\mathcal{S}, x)$ will be the same for every response variable used in that sample.

The weights proposed by Gelman using hierarchical models add dependence on y to the mix. When a proper

Sharon L. Lohr is Thompson Industries Dean's Distinguished Professor of Statistics, Department of Mathematics and Statistics, Arizona State University, Tempe, Arizona 85287-1804, USA (e-mail: sharon.lohr@asu.edu).