Struggles with survey weighting and regression modeling[*]

Andrew Gelman[†]

June 22, 2005

Abstract

The general principles of Bayesian data analysis imply that models for survey responses should be constructed conditional on all variables that affect the probability of inclusion and nonresponse, which are also the variables used in survey weighting and clustering. However, such models can quickly become very complicated, with potentially thousands of post-stratification cells. It is then a challenge to develop general families of multilevel probability models that yield reasonable Bayesian inferences. We discuss in the context of several ongoing public health and social surveys. This work is currently open-ended, and we conclude with thoughts on how research could proceed to solve these problems.

Keywords: multilevel modeling, poststratification, sampling weights, shrinkage

1 Background

Survey weighting is a mess. It is not always clear how to use weights in estimating anything more complicated than a simple mean or ratios, and standard errors are tricky even with simple weighted means. Contrary to what is assumed by many theoretical statisticians, survey weights are *not* in general equal to inverse probabilities of selection but rather are constructed based on a combination of probabilities and nonresponse adjustments.

Regression modeling is a potentially attractive alternative to weighting. In practice, however, the potential for large numbers of interactions can make regression adjustments highly variable. This paper reviews the motivation for hierarchical regression, combined with poststratification, as a strategy for correcting for differences between sample and population. We sketch some directions toward a practical solution, which unfortunately has not yet been reached.

1.1 Estimating population quantities from a sample

Our goal is to use sample survey data to estimate a population average or the coefficients of a regression model. The regression framework also includes small-area estimation, since that is simply a regression on a discrete variable corresponding to indicators for the small areas.

Figure 1: The proportion of adults surveyed who answered yes in the Gallup Poll to the question, "Are you in favor of the death penalty for a person convicted of murder?", among those who expressed an opinion on the question. It would be interesting to estimate these trends in individual states.

We shall consider two running examples: a series of CBS/New York Times national polls from the 1988 election campaign, and the New York City Social Indicators Survey, a biennial survey of families that is conducted by Columbia University's School of Social Work (Garfinkel and Meyers, 1999, Meyers and Teitler, 2001, Garfinkel et al., 2003). Both sets of surveys use random digit dialing.

For the pre-election polls, our quantity of primary interest is the proportion of people[1] who support the Republican candidate for President in the country or in each state. We would also like to use series of national polls to estimate state-by-state time trends, for example in the support for the death penalty over the past few decades. (See Figure 1 for the national trends.)

For the Social Indicators Survey, we are interested in population average responses to questions such as, "Do you rate the schools as poor, fair, good, or very good?", average responses in subpopulations (for example, the view of the schools among parents of school-age children), and so-called "analytical" studies that can be expressed in terms of regressions (for example, predicting total satisfaction given demographics and specific attitudes about health care, safety, etc.). In this paper, we focus on trends from 1999 to 2001, as measured by changes in two successive Social Indicators Surveys, on a somewhat arbitrary selection of questions chosen to illustrate the general concerns of the survey.

Table 1 shows the questions, the estimated average responses in each year, and the estimated differences and standard errors as obtained using two different methods of inference. This paper is centered on the puzzle of how these two estimation methods differ. We shall get back to this

---

[1]Or the proportion of *voters* who support the Republican candidate, which is a ratio: the proportion of people who will vote and support the Republican, divided by the proportion who will support the Republican. It is straightforward to move from estimating a population mean to estimating this ratio, as discussed in the context of this example by Park, Gelman, and Bafumi (2004).

Question	weighted averages 1999 2001	(a) time change in percent	(b) linear regression coefficient of time	(a) time change on logit scale	(b) logistic regression coefficient of time
Adult in good/excellent health	75% 78%	3.4% (2.4%)	6.6% (1.4%)	0.19 (0.13)	0.48 (0.10)
Child in good/excellent health	82% 84%	1.7% (1.5%)	1.2% (1.3%)	0.24 (0.21)	0.18 (0.20)
Neighborhood is safe/very safe	77% 81%	4.5% (2.3%)	4.1% (1.5%)	0.27 (0.14)	0.27 (0.10)

Table 1: Estimates for some responses from two consecutive waves of the New York City Social Indicators Survey, and estimated changes, with standard errors in parentheses. Changes are estimated in percentages and on the logit scale. In each scale, two estimates are presented: (a) simple differences in weighted means, (b) regression controlling for the variables used in the weighting. Approaches (a) and (b) can give similar results but sometimes are much different.

question in a moment after reviewing some basic ideas in survey sampling inference.

1.2 Poststratification and weighting

Naive promulgators of Bayesian inference—or the modeling approach to inference in general—used to say that the method of data collection was irrelevant to estimation from survey data. All that matters, from this slightly misguided perspective, is the likelihood, or the model of how the data came to be. However, as has been pointed out by Rubin (1976), the usual Bayesian or likelihood analysis implicitly assumes the design is "ignorable," which in a sampling context roughly means that the analysis include all variables that affect the probability of a person being included in the survey (see chapter 7 of Gelman et al., 2003, for a review).

In a regression context, the analysis should include, as "X variables," everything that affects sample selection or nonresponse. Or, to be realistic, all variables should be included that have an important effect on sampling or nonresponse, if they also are potentially predictive of the outcome of interest. In a public survey such as the CBS polls, a good starting point is the set of variables used in their weighting scheme: number of adults and number of telephone lines in the sampled household; region of the country; and sex, ethnicity, age, and education level of the respondent (see Voss, Gelman, and King, 1995). For the Social Indicators Survey, we did our own weighting (Becker, 1998) using similar information: number of telephone lines (counted as 1/2 for families with intermittent phone service), number of adults and children in the family, and ethnicity, age, and education of the head of household. Weights for each survey are constructed by multiplying a series of factors.

With this as a reference, we quickly review the unified notation for poststratification and survey weighting of Little (1991, 1993) and Gelman and Carlin (2001). Here we use the notation y, z for variables that are observed in the sample only, and X for variables that are observed in the sample and known in the population.

Poststratification. The purpose of poststratification is to correct for known differences between sample and population. In the basic formulation, we have variables X whose joint distribution in the population is known, and an outcome y, whose population distribution we are interested in estimating. We shall assume X is discrete, and label the possible categories of X as *poststratification cells* j, with population sizes N_j and sample sizes n_j. In this notation, the total population size is $N = \sum_{j=1}^{J} N_j$ and the sample size is $n = \sum_{j=1}^{J} n_j$. The implicit model of poststratification is that the data are collected by simple random sample within each of the J poststrata. The assignment of sample sizes to poststrata is irrelevant. In fact, classical stratification (in which the sampling really *is* performed within strata) is a special case of poststratification as we formulate it. We assume the population size N_j of each category j is known. These categories include all the cross-classifications of the predictors X.[2]

The population mean of any survey response can be written as a sum over poststrata:

$$\text{definition of population mean:} \quad \theta = \frac{\sum_{j=1}^{J} N_j \theta_j}{\sum_{j=1}^{J} W_j},$$

with corresponding estimate

$$\text{poststratified estimate:} \quad \hat{\theta}^{PS} = \frac{\sum_{j=1}^{J} N_j \hat{\theta}_j}{\sum_{j=1}^{J} N_j}.$$

We use the general notation θ_j rather than $\bar{Y}_j$ to allow for immediate generalization to other estimands such as regression coefficients.

Weighting. When you look at sample survey data from a public-use dataset, the "survey weight" looks like a unit-level characteristic—just one more column in the data—and it's easy to think of it almost as a survey response, w_i. In this context it seems natural to use weighted averages of the form $\bar{y} = \sum_{i=1}^{n} (w_i y_i) / \sum_{i=1}^{n} w_i$.

But survey weights are *not* attributes of individual units—they are constructions based on an entire survey. Within any poststratification cell, all units have the same weight.[3] We shall refer to

---

[2]In some cases the cell populations are unknown and must be estimated. For example, in the Social Indicators Survey, we adjust to estimated demographics from the Current Population Survey, which includes about 2000 New York City residents each year. This is enough to give reliable estimates of one-way and two-way margins (for example, the proportion of city residents who are white females, white males, black females, black males, etc.), but the counts are too sparse to directly estimate deep interactions (for example, the proportion who are white females, 30–45, married, with less than a high school education, etc.). The usual practical solution in this case is to poststratify on the margins (that is, raking; see, for example, Deville, Sarndal, and Sautory, 1993). If the whole table of population counts is required, it can be estimated using iterative proportional fitting (Deming and Stephan, 1940) which sets interactions to be as small as possible while being consistent with the available population data. For this paper, we shall ignore this difficulty and treat the full vector of N_j's as known.

[3]In theory, continuously-varying survey weights could arise from a survey with a continuous range of sampling probabilities. For example, one could imagine a survey of college-bound students where the probability of selection is a continuous function of background variables (e.g., $\text{Pr(selection)} = \text{logit}^{-1}(a + b \cdot \text{SAT})$). Or one could model nonresponse as a continuous function of predictors such as age and previous health status in a medical surveys. These continuous weights do not come up much the sorts of social surveys under consideration in this paper, but they are interesting research directions that are potentially important in other areas of application.

```
       respid     org year survey  y state edu age female black adults phones weight
11352   6140 cbsnyt    7   9158 NA     7   3   1      1     0      2      1    923
11353   6141 cbsnyt    7   9158  1    39   4   2      1     0      2      1    558
11354   6142 cbsnyt    7   9158  0    31   2   4      1     0      1      1    448
11355   6143 cbsnyt    7   9158  0     7   3   1      1     0      2      1    923
11356   6144 cbsnyt    7   9158  1    33   2   2      1     0      1      1    403
```

Table 2: Data from the first five respondents of a CBS pre-election poll. The weights are listed as just another survey variable, but they are actually constructed after the survey has been conducted, so as to match sample with known population information.

unit weights w_i, $i = 1, \ldots, n$, and *cell weights* $W_j = n_j w_i$ for units i within cell j.

$$\text{weighted average:} \quad \bar{y} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{j=1}^{J} W_j \bar{y}_j}{\sum_{j=1}^{J} W_j}. \tag{1}$$

Survey weights in general depend on the actual data collected as well as on the design of the survey. For example, consider the seven CBS polls conducted during the week before the 1988 Presidential election. These surveys had identical designs and targeted the same population. However, the weighting factor assigned to men (compared to a factor of 1 for women) varies from as low as 1.10 to 1.27 among the seven surveys. The different samples happened to contain different ratios of men to women and hence needed different adjustments.

1.3 Competing methods of estimation: weighted averages, weighted regression, and unweighted regression controlling for X

Many researchers have noted the challenge of using survey weights in regression models (as reviewed, for example, by DuMouchel and Duncan, 1983, Kish, 1992, and Pfefferman, 1993). For the goal of estimating a population mean, it is standard to use the weighted average (1), but it is not so clear what to do in more complicated analyses. For example, when estimating a regression of y on z, one recommended approach is to use weighted least squares, and another option is to perform unweighted regression of y on z, also controlling for the variables X that are used in the weighting.

Computing standard errors is not trivial for weighted estimates, whether means or regressions, because the weights themselves are random variables that depend on the data. In particular, correct classical standard errors cannot simply be obtained from the data and the weights; one also needs to know the procedure used to create the weights. Table 3 illustrates problems with some variance estimates that do not account for the weighting design. Similarly, with regressions, simple weighted regression procedures do not in general give correct standard errors.

| | True | Different standard error estimates | | | |
Opinion of NYC	standard error	assuming SRS	conditioning on weights	assuming inv-prob	design-based
Become a better place	2.2%	1.2%	2.5%	1.9%	2.1%
Remained the same	2.0%	1.2%	2.3%	1.6%	1.9%
Gotten worse	2.0%	1.2%	2.4%	1.7%	2.0%

Table 3: From simulation study: true standard error and four different standard error estimates for a question on the Social Indicators Survey. Ignoring the weighting or treating the weights as constant underestimates uncertainty, whereas uncertainty is overestimated by treating the weights as inverse probabilities. Accurate standard errors can be obtained using a jackknife-like procedure that explicitly takes account of the design of the weighting procedure. From Lu and Gelman (2003).

1.4 The crucial role of interactions

Consider a regression of y on z, estimated in some way from a survey where inclusion probabilities depend on X. In general, y can depend on both X and z, in which case the appropriate way to estimate the regression of y on z is to regress y on X, z and then average over the population distribution of X. In general, estimating the regression of y on z requires estimation of the relation between z and X as well (Graubard and Korn, 2002).

Because of the potential dependence of z and X, it can be important to include interactions between these predictors in the model for y, even if the ultimate goal is simply to estimate the relation between y and z.

2 The challenge

2.1 Estimating simple averages and trends

We now return to the example of Table 1. The goal is to estimate $\bar{Y}^{2001} - \bar{Y}^{1999}$, the change in population average response between two waves of the Social Indicators Survey. This can be formulated as the coefficient β_1 in a regression of y on time: $y = \beta_0 + \beta_1 z + \text{error}$, where the data from the two surveys are combined, $z = 0$ and 1 for respondents of the 1999 and 2001 surveys, respectively. Or, $y = \beta_0 + \beta_1 z + \beta_2 X + \text{error}$, where β_2 is a vector of coefficients for the variables X used in the weighting. Now the quantity of interest is $\beta_1 + \beta_2(\overline{X}^{2001} - \overline{X}^{1999})$, to account for demographic changes between the two years. For New York City these changes were minor, and we would be happy to simply estimate β_2.

This brings us to the puzzle of Table 1. For each of three binary outcomes y, we compute the weighted mean for each year, $\bar{y}_w^{1999}$ and $\bar{y}_w^{2001}$, and two estimates of the change:

- Our first estimate is the simple difference, $\bar{y}_w^{2001} - \bar{y}_w^{1999}$, with standard error $\sqrt{\text{var}(\bar{y}_w^{2001}) + \text{var}(\bar{y}_w^{1999})}$, where the sampling variances are computed using the design of the weights (as in the right-

most column in Table 3).

- Our other estimate is obtained by linear regression. We combine the data from the two surveys into a single vector, $y = (y^{1999}, y^{2001})$ and create an associated indicator vector z that equals 0 for the data from 1999 and 1 for the data from 2001. We fit a linear regression of y on z, also controlling for the variables X used in the weighting.[4] To estimate the change from 1999 to 2001, we use the coefficient of z, with standard error automatically coming from the (unweighted) regression.

As indicated in the third and fourth column of Table 1, the change in weighed averages and the regression coefficient tend to have the same sign, but the two estimates sometimes differ quite a bit in magnitude. (Similar results are obtained if we work on the logit scale, as can be seen from the final two columns of the table.)

What should we believe? For this particular example, the direct analysis of weighted averages seems more believable to us, since we specifically created the weighting procedure for the goal of estimating these citywide averages. More generally, however, using weighted averages is awkward and we would prefer to use the more general techniques of regression and poststratification.

Where do we go from here? We would like an approach to statistical analysis of survey data that gives the right answers for simple averages and comparisons, and can be smoothly generalized to more complicated estimands.

2.2 Deep poststratification

One of the difficulties of survey weighting is that the number of poststratification cells can quickly become large, even exceeding the number of respondents. This leads naturally to multilevel modeling to obtain stable estimates in all the poststratification cells, even those with zero or one respondent. Choices must then be made in the modeling of interactions.

For example, in our time-trend estimation problem, we could model $y = \beta_0 + \beta_1 z + \beta_2 X + \beta_3 X z + \text{error}$, where β_3 is a vector of coefficients for the interaction of X and time. We would then be interested in $\beta_1 + \beta_2(\overline{X}^{2001} - \overline{X}^{1999}) + \beta_3 \overline{X}^{2001}$. Where should the interaction modeling stop? The practical problem of adjusting for survey nonresponse leads to general questions of inference under multi-way interactions, an issue that becomes even more relevant in small-area estimation.

Gelman and Carlin (2001) and Park, Gelman, and Bafumi (2004) discuss the estimation of state-by-state opinions from national polls, using a hierarchical logistic regression with demographics and state effects, followed by poststratification on Census population totals for 64 demographic categories in each of the 50 states. The method worked well, but it is not clear how it would perform if the model included interactions of demographic and state effects.

---

[4] These X variables are number of adults in household, number of children in family, number of telephone lines, marital status, and sex, age, ethnicity, and education, and ethnicity $\times$ education for head of household.

3 Using regression modeling to connect weighting and poststratification

When cell means are estimated using certain linear regression models, poststratified estimates can be interpreted as weighted averages (Little, 1991, 1993). The idea is to work with the poststratified estimate (1.2)—an average over cell estimates $\hat{\theta}_j$, with the regression model providing the $\hat{\theta}_j$'s based on characteristics of the cells j. Under certain conditions, the poststratified estimate can be reinterpret this as a weighted average of the form (1), and then we can solve for the cell weights W_j and the unit weights w_i.

3.1 Classical models

Full poststratification. The simplest case is full poststratification of raw data, in which case the cell estimates are the cell means, $\hat{\theta}_j = \bar{y}_j$, and (1.2) becomes

$$\hat{\theta}^{PS} = \frac{\sum_{j=1}^{J} N_j \bar{y}_j}{\sum_{j=1}^{J} N_j},$$

which is equivalent to (1) with cell weights $W_j \propto N_j$ or unit weights $w_i \propto N_{j(i)}/n_{j(i)}$, where $j(i)$ is the poststratification cell to which unit i belongs.

This estimate can also be viewed as a classical regression including indicators for all J poststratification cells.

No weighting. The other extreme is no weighting, that is, unit weights $w_i = 1$ for all i, which is equivalent to poststratification if the cell estimates $\hat{\theta}_j$ are all equal to the sample mean $\bar{y}$, which in turn corresponds to classical regression including only a constant term.

Classical regression on cell characteristics. Intermediate cases of weighting can be obtained by regression models that include information about the poststratification cells without going to the extreme of fitting a least-squares predictor to each cell. For example, in the CBS/New York Times pre-election surveys, one could regress y on indicators for sex, ethnicity, age, education, and region, without necessarily including all their interactions.

Suppose the regression model is $y \sim \mathrm{N}(X\beta, \sigma_y^2 I)$. We shall use X to represent the $n \times k$ matrix of predictors in the data, and X^{pop} to represent the $J \times k$ matrix of predictors for the J poststratification cells. We also label the vector of poststratum populations as $N^{\mathrm{pop}} = (N_1, \ldots, N_J)$, with a sum of $N = \sum_{j=1}^{J} N_j$.

The estimated vector of regression coefficients is then $\hat{\beta} = (X^t X)^{-1} X^t y$, and the estimated cell

means are $X^{\mathrm{pop}}\hat{\beta}$. The poststratified estimate of the population mean is then,

$$
\begin{aligned}
\hat{\theta}^{PS} &= \frac{1}{N}\sum_{j=1}^{J}N_j(X^{\mathrm{pop}}\hat{\beta})_j \\
&= \frac{1}{N}(N^{\mathrm{pop}})^t X^{\mathrm{pop}}(X^t X)^{-1}X^t y, \tag{2}
\end{aligned}
$$

which can be written as $\hat{\theta}^{PS} = \frac{1}{n}\sum_{i=1}^{n}w_i y_i$, with a vector of unit weights,

$$
w = \left(\frac{n}{N}(N^{\mathrm{pop}})^t X^{\mathrm{pop}}(X^t X)^{-1}X^t\right)^t. \tag{3}
$$

For convenience, we have renormalized these weights to sum to n (see below). In 3, w is a vector of length n that takes on at most J distinct values. The vector of J possible unit weights (corresponding to units in each of the J poststrata) is,

$$
w^{\mathrm{pop}} = \left(\frac{n}{N}(N^{\mathrm{pop}})^t X^{\mathrm{pop}}(X^t X)^{-1}(X^{\mathrm{pop}})^t\right)^t, \tag{4}
$$

and the poststratified estimate can also be expressed as,

$$
\hat{\theta}^{PS} = \frac{1}{n}\sum_{j=1}^{J}w_j^{\mathrm{pop}}\bar{y}_j.
$$

The key result that makes the above computations possible—that allows $\hat{\theta}^{PS}$ to be interpreted as a weighted average of data—is that the derived unit weights w in (3) sum to n. The identity $\sum_{j=1}^{n}w_j = n$ can be proved using matrix algebra but is more easily derived from an invariance in the classical regression model. With a least-squares regression, if a constant is added to all the data, that same constant will be added to the intercept, with the other coefficients not changing at all. The expression (2) includes exactly 1 of the intercept, and so when $\hat{\theta}^{PS}$ is expressed as $\hat{\theta}^{PS} = \frac{1}{n}\sum_{i=1}^{n}w_i y_i$, these w_i/n's must sum to 1.

The left panel of Figure 2 shows the unit weights obtained by fitting a sequence of classical regression models to the CBS/New York Times survey data. As more factors and interactions are included, the weights become more variable.

3.2 Hierarchical models

Hierarchical regression. The results above can be immediately generalized to multilevel regression models in which some of the coefficients are batches of indicator variables. We shall generalize the regression model to $y \sim \mathrm{N}(X\beta, \Sigma_y)$ with a prior distribution on β of the form $\beta \sim \mathrm{N}(0, \Sigma_\beta)$. For simplicity, we assume independence of the components of β in the prior distribution, conditional on hyperparameters for the variance components.

The prior precision matrix Σ_β^{-1} is then diagonal, with zeroes for non-hierarchical regression coefficients (including the constant term in the regression). For example, consider a regression for the CBS/New York Times polls, with the following predictors:
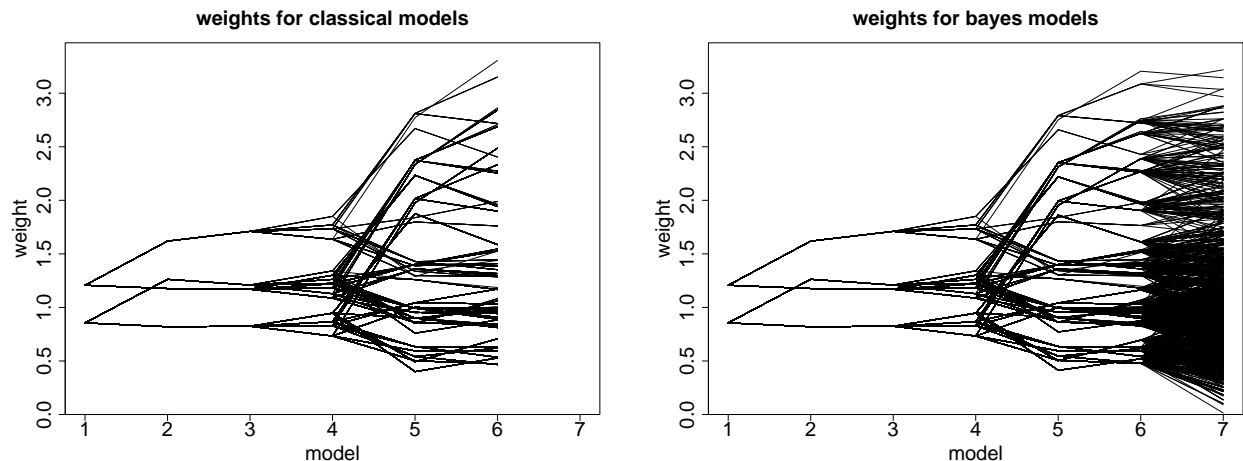
Figure 2: Implicit weights w_i for one of the CBS/New York Times surveys, based on a series of models fit first using classical regression and then using Bayesian hierarchical regression. The models are nested, controlling for (1) male/female, (2) also black/white, (3) also male/female $\times$ black/white, (4) also four age categories, (5) also four education categories, (6) also age $\times$ education, (7) also state indicators. Each model includes more factors and thus has more possible weights, which are renormalized to average to 1 for each model. For the Bayes models, the indicators for age, education, age $\times$ education, and state are given independent batches of varying coefficients. For the classical weights, model (7) is not included because of collinearity.

The lines in each graph connect the weights for individual respondents, which are divided into successively more categories as predictors are added to the models.

- A constant term

- An indicator for sex (1 if female, 0 if male)

- An indicator for ethnicity (1 if black, 0 otherwise)

- Sex $\times$ ethnicity

- 4 indicators for age categories

- 4 indicators for education categories

- 16 age $\times$ education indicators.

The classical regression has 1+1+1+1+3+3+9=19 predictors (avoiding collinearity by excluding the baseline age and education categories). The hierarchical regression has 1+1+1+1+4+4+16=28 predictors, and its prior precision matrix has the form,

$\Sigma_\beta^{-1} = \mathrm{Diag}(0, 0, 0, 0, \sigma_{\mathrm{age}}^{-2}, \sigma_{\mathrm{age}}^{-2}, \sigma_{\mathrm{age}}^{-2}, \sigma_{\mathrm{age}}^{-2}, \sigma_{\mathrm{edu}}^{-2}, \sigma_{\mathrm{edu}}^{-2}, \sigma_{\mathrm{edu}}^{-2}, \sigma_{\mathrm{edu}}^{-2}, \sigma_{\mathrm{age.edu}}^{-2}, \dots, \sigma_{\mathrm{age.edu}}^{-2})$, with the parameters $\sigma_{\mathrm{age}}, \sigma_{\mathrm{edu}}, \sigma_{\mathrm{age.edu}}$ estimated from data.

The estimated vector of regression coefficients is then $\hat{\beta} = (X^t \Sigma_y^{-1} X + \Sigma_\beta^{-1})^{-1} X^t \Sigma_y^{-1} y$, and expressions (2)–(4) become,

$$\begin{aligned}
\hat{\theta}^{PS} &= \frac{1}{N}(N^{\text{pop}})^t X^{\text{pop}} (X^t \Sigma_y^{-1} X + \Sigma_\beta^{-1})^{-1} X^t \sigma_y^{-1} y \\
w &= \left(\frac{n}{N}(N^{\text{pop}})^t X^{\text{pop}} (X^t \Sigma_y^{-1} X + \Sigma_\beta^{-1})^{-1} X^t \Sigma_y^{-1} \right)^t \\
w^{\text{pop}} &= \left(\frac{n}{N}(N^{\text{pop}})^t X^{\text{pop}} (X^t \Sigma_y^{-1} X + \Sigma_\beta^{-1})^{-1} (X^{\text{pop}})^t \Sigma_y^{-1} \right)^t .
\end{aligned} \tag{5}$$

Conditional on the variance parameters in Σ_y and Σ_{beta}, then, estimates from this model correspond to weighted averages.

The right panel of Figure 2 shows the unit weights obtained by fitting a sequence of Bayesian models to the CBS/New York Times poll. The first three models are actually identical to the classical versions, since we assign noninformative uniform prior distributions to the coefficients for sex, ethnicity, and their interactions. Models 4 and 5 are similar to the classical fits because age and education have only four categories, so there is little information available for partial pooling of these effects (see Gelman, 2005). The weights in model 6, with age × education interactions included, are smoothed somewhat compared to the corresponding classical model. Finally, introducing state effects leasds to a downweighting of some of respondents in states that happen to be overrepresented in the survey, and an upweighting for respondents in the undersampled states. There is no corresponding classical model here because the survey does not actually include data from all 50 states.

Exchangeable normal model. To understand these formulas better, we consider the special case of an exchangeable normal model for the J cell means (see also Lazzeroni and Little, 1998, and Elliott and Little, 2000). This model can be expressed in terms of the cell means:

$$\begin{aligned}
\bar{y}_j &\sim \mathrm{N}(\theta_j, \sigma^2/n_j) \\
\theta_j &\sim \mathrm{N}(\mu, \sigma_\theta^2).
\end{aligned}$$

This is a special case of the hierarchical regression model discussed above, so we already know that the poststratified estimate, conditional on the (estimated) variance parameters σ_y, σ_θ, can be expressed as a weighted average of the cell means, $\bar{y}_j$, or equivalently as a weighted average of the data points y_i.

In this simple example, however, we can gain some understanding by deriving algebraic expressions for the weights. Our goal is to express them in terms of the completely-smoothed weights, $w_j = 1$, and the weights from full poststratification, $w_j = \frac{N_k/N}{n_j/n}$.

We start with the posterior means (conditional on the variance parameters) of the cell means. We write these as $\hat{\theta}_k, k = 1, \ldots, J$ (using k as a subscript rather than k because this results in more

11

convenient notation later):

$$\hat{\theta}_k = \frac{\frac{n_k}{\sigma_y^2} y_k + \frac{1}{\sigma_\theta^2} \hat{\mu}}{\frac{n_k}{\sigma_y^2} + \frac{1}{\sigma_\theta^2}}, \tag{6}$$

where

$$\hat{\mu} = \frac{\sum_{k=1}^{J} \frac{1}{\sigma_y^2/n_k + \sigma_\theta^2} y_k}{\sum_{k=1}^{J} \frac{1}{\sigma_y^2/n_k + \sigma_\theta^2}}. \tag{7}$$

We can combine (6) and (7) to express each $\hat{\theta}_j$ as a linear combination of the cell means $\bar{y}_k$:

$$\hat{\theta}_k = \sum_{j=1}^{J} c_{kj} \bar{y}_j.$$

After some algebra, we can write these coefficients as

$$c_{kj} = \begin{cases} \frac{\sigma_y^2}{n_k} A_k A_j / A & \text{for } j \neq k \\ \sigma_\theta^2 A_k + \frac{\sigma_y^2}{n_k} A_k^2 & \text{for } j = k, \end{cases}$$

where

$$A_k = \sum_{k=1}^{J} \frac{1}{\sigma_y^2/n_k + \sigma_\theta^2}.$$

The payoff now comes in computing the poststratified estimate,

$$\begin{aligned} \hat{\theta}^{PS} &= \sum_{k=1}^{J} N_k \hat{\theta}/N \\ &= \sum_{k=1}^{J} \sum_{j=1}^{J} \frac{N_k}{N} c_{kj} \bar{y}_j \end{aligned}$$

and equating this to $\sum_{j=1}^{j} W_j \bar{y}_j$, thus deriving the cell weights,

$$\begin{aligned} W_j &= \sum_{k=1}^{J} \frac{N_k}{N} c_{kj} \\ &= A_j \left[\frac{N_j}{N} \sigma_\theta^2 + \sum_{k=1}^{J} \frac{N_k}{N} \frac{A_k}{A} \frac{\sigma_y^2}{n_k} \right]. \end{aligned}$$

The implicit unit weights are then $w_j^{\text{pop}} = (n/n_j) W_j$, or

$$\begin{aligned} w_j^{\text{pop}} &= A_j \frac{n}{n_j} \left[\frac{N_j}{N} \sigma_\theta^2 + \frac{\sigma_y^2}{AN} \sum_{k=1}^{J} A_k \frac{N_k}{n_k} \right] \\ &= \frac{n}{\sigma_y^2 + n_j \sigma_\theta^2} \left[\frac{N_j}{N} \sigma_\theta^2 + \frac{\sigma_y^2}{N} \frac{\sum_{k=1}^{J} \frac{1}{\sigma_y^2 + n_k \sigma_\theta^2} N_k}{\sum_{k=1}^{J} \frac{1}{\sigma_y^2 + n_k \sigma_\theta^2} n_k} \right] \end{aligned} \tag{8}$$

The ratio of sums in (8) is a constant (given the fitted model) that does not the depend on j. Let us approximate it by N/n (which is appropriate if the sample proportions n_k/N_k are independent of the group sizes N_k. Under this approximation, the unit weights can be written as,

$$\text{approximate } w_j^{\text{pop}} = \frac{\frac{n_j}{\sigma_y^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\theta^2}} \cdot \frac{N_k/N}{n_k/n} + \frac{\frac{1}{\sigma_\theta^2}}{\frac{n_j}{\sigma_y^2} + \frac{1}{\sigma_\theta^2}} \cdot 1, \tag{9}$$

which is a weighted average of the full poststratification unit weight, $\frac{N_k/N}{n_k/n}$, and the completely-smoothed weight of 1. Hierarchical poststratification is thus approximately equivalent to a shrinkage of weights by the same factors as in the shrinkage of the parameter estimates (6).

Thus, as with hierarchical regression models in general, the amount of shrinkage of the weights depends on the between and within-stratum variance in the outcome of interest, y.

Other hierarchical models. Lazzeroni and Little (1998, and Elliott and Little (2000) discuss various hierarchical linear regression models, including combinations of the two models described above (that is, a hierarchical regression a cell-level variance component) and models with correlations between adjacent cell categories for ordered predictors.

Another natural generalization is to use logistic regression for binary inputs. Unfortunately, when we move away from linear regression, we abandon the translation-invariance of the parameter estimates (that is, the property that adding a constant to all the data affects only the constant term and none of the other regression coefficients). As a result, for logistic regression, the poststratified estimate $\hat{\theta}^{PS}$ is no longer a weighted average of the data, even after controlling for the variance parameters in the model. However, we suspect that the model could be linearized, yielding approximate weights.

3.3 Properties of the model-based poststratified estimates

Standard errors. The variance of the poststratified estimate, ignoring sampling variation in X, can be expressed using various formulas:

$$\text{var}(\hat{\theta}^{PS}) = \frac{1}{n^2} \sum_{i=1}^{n} w_i^2 \sigma_y^2 = \frac{1}{n} \sum_{j=1}^{J} (w_j^{\text{pop}})^2 n_j \sigma_y^2 = \frac{1}{nN} \sum_{j=1}^{J} w_j^{\text{pop}} N_j^{\text{pop}} \sigma_y^2.$$

Any of these equivalent expressions can be viewed as the posterior variance of θ given a noninformative prior distribution on the regression coefficients, and ignoring posterior uncertainty in σ_y (Holt and Smith, 1979, Little, 1993).

Dependence of implicit weights on the outcome variable. Classical survey weights depend only on the n_j's and the N_j's, as well as the design matrix X (used, for example, to define the margins used in raking), but do not formally depend on y. (There is an informal dependence on y

in the sense that there is no urgency to weight on variables X that do not help predict outcomes y of interest.) Similarly, the implicit weights (3) obtained from a classical regression model depend only on n, N, and X, not on y.

However, the implicit weights (5) from hierarchical regression *do* depend on the data, implicitly, through the hyperparameters in Σ_y and Σ_β, which are estimated from the data. Thus, the appropriate weights could differ for different survey responses.

4   Where to go next

There are currently two standard approaches to adjusting for known differences between sample and population in survey data: weighting and regression modeling.

Practical limitations of weighting.   The weighting approach has the advantage of giving simple estimates for population averages but has several disadvantages. First, it is not generally clear how to apply the weighting to more complicated estimands such as regression coefficients. There has been some work on weighted regression for surveys (e.g., DuMouchel and Duncan, 1983, and Pfefferman, 1993) but these procedures are not very flexible, which is one reason why the modeling approach is more popular for problems such as small-area estimation. A second problem with weighted estimates is that standard errors are more difficult to evaluate (recall Table 3). Finally, weighting may be "dirty" but it is not always "quick": actually constructing the weighting for a survey is more difficult than you might think. Creating practical weights requires arbitrary choices about inclusion of weighting factors and interactions, pooling of weighting cells, and truncation of weights.[5]   The resulting vector of weights is in general a complicated and not-fully-specified function of data and prior knowledge. Subjective choices arise in virtually all statistical methods, of course, but good advice on creating weights tends to be much vaguer than for other methods in the statistical literature (see, e.g., Lohr, 1999).

Practical limitations of modeling.   Regression modeling is easy to do—even hierarchical regression is becoming increasingly easy in Bugs, Stata, and other software packages—but for analysis of survey data it has the disadvantage that, to combine with population information, the regression must theoretically condition on all the poststratification cells, which can lead to very complicated models—more complicated than we are comfortable with in current statistical practice—even in surveys of moderate size (see Section 2.2). When a model is too complicated, it becomes difficult

---

[5]For example, in the Social Indicators Survey, we decided to weight on some interactions and not others in order to control variability of the weights. While setting up the weightin procedure, we repeatedly compared weighted estimates to Census values for various outcomes that we thought could be "canaries in the coal mine" if the survey estimate did not fit the population. These "canary" variables included percentage of New York City residents who are U.S. citizens, the percent who own their own home, and income quintiles.

to interpret or use the results, leading to awkward situations such as in Table 1, where we simply can't trust the regression coefficients for time trends in the Social Indicators Survey.

It's a delicate point, because sometimes we do have confidence in regression coefficients, even with complicated hierarchical models with many parameters. For example, as discussed in Gelman and Carlin (2001) and Park, Gelman, and Bafumi (2004), hierarchical regression combined with poststratification performs excellently at estimating state-level opinions from the national CBS/New York Times polls. So it's not just the number of parameters that's important, but rather some connection between the model and the quantities of interest, which is somehow more difficult to establish in the models whose results are shown in Table 1.

Putting it together using hierarchical models and poststratification. Our ideal procedure should be as easy to use as hierarchical modeling, with population information included using poststratification as in (1.2). The procedure should feature a smooth transition from classical weighting so that when different estimation methods give different results, it is possible to understand this difference as a result of interactions in the model (as discussed by Graubard and Korn, 2002).

How do we get there? One place to start is to focus on examples such as in Table 1 where different methods give different answers, and try to figure out which, if either, of the two estimates make sense. Our goal is to come up with a general modeling procedure which gives believable estimates for time trends and as a byproduct produces a good set of weights that can be used for simple estimands. Given the difficulties with current methods for weighting and modeling, we believe this approach is of both practical and theoretical interest.

References

Becker, D. E. (1998). The New York City Social Indicators Survey: an analysis of the weighting procedure. Technical report, School of Social Work, Columbia University.

Binder, D. A. (1983). On the variance of asymptotically normal estimators from sample surveys. *International Statistical Review* **51**, 279–292.

Deming, W. E., and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics* **11**, 427–444.

Deville, J., Sarndal, C., and Sautory, O. (1993). Generalizing raking procedures in survey sampling. *Journal of the American Statistical Association* **88**, 1013–1020.

DuMouchel, W. H., and Duncan, G. J. (1983). Using sample survey weights in multiple regression analysis of stratified samples. *Journal of the American Statistical Association* **78**, 535–543.

Elliott, M. R., and Little, R. J. A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics* **16**, 191–209.

Fay, R. E., and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* **74**, 269–277.

Garfinkel, I., Kaushal, N., Teitler, J., and Garcia, S. (2003). Vulnerability and resilience: New Yorkers respond to 9/11. School of Social Work, Columbia University

Garfinkel, I., and Meyers, M. K. (1999). New York City social indicators 1997: a tale of many cities. School of Social Work, Columbia University.

Gelman, A. (2005). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, to appear.

Gelman, A., and Carlin, J. B. (2001). Poststratification and weighting adjustments. In *Survey Nonresponse*, ed. R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little. New York: Wiley.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2003). *Bayesian Data Analysis*, second edition. London: CRC Press.

Graubard, B. I., and Korn E. L. (2002). Inference for superpopulation parameters using sample surveys. *Statistical Science* **17**, 73–96.

Holt, D., and Smith, T. M. F. (1979). Post stratification. *Journal of the Royal Statistical Society A* **142**, 33–46.

Kish, L., (1992). Weighting for unequal P_i. *Journal of Official Statistics* **8**, 183–200.

Lazzeroni, L. C., and Little, R. J. A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics* **14**, 61–78.

Little, R. J. A. (1991). Inference with survey weights. *Journal of Official Statistics* **7**, 405–424.

Little, R. J. A. (1993). Post-stratification: a modeler's perspective. *Journal of the American Statistical Association* **88**, 1001–1012.

Lohr, S. (1999). *Sampling: Design and Analysis*. Pacific Grove, Calif.: Duxbury.

Lu, H., and Gelman, A. (2003). Sampling variances for surveys with weighting, poststratification, and raking. *Journal of Official Statistics*.

Meyers, M. K., and Teitler, J. O. (2001). New York City social indicators 1999: pulling ahead, falling behind. School of Social Work, Columbia University.

Park, D. K., Gelman, A., and Bafumi, J. (2004). Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis*.

Pfefferman, D. (1993). The role of sampling weights when modeling survey data. *International*

Statistical Review **61**, 317–337.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Voss, D. S., Gelman, A., and King, G. (1995). Pre-election survey methodology: details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly* **59**, 98–132.

Yung, W., and Rao, J. N. K. (1996). Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology* **22**, 23–31.