Poststratification and weighting adjustments*

Andrew Gelman[†]and John B. Carlin[‡]

February 3, 2000

> "A weight is assigned to each sample record, and MUST be used for all tabulations."
>
> — codebook for the CBS News / New York Times Poll, 1988

1 Introduction

1.1 Overview

Poststratification and weighting are used to adjust for known or expected discrepancies between sample and population. In this chapter, we aim to review current methods for using these techniques in survey analysis, and to critically examine the methods in the context of new ideas for extending model-based (Bayesian) methods to handle some of the more difficult problems that arise in practice. In particular, we distinguish among several different types of weights that are commonly used and clarify the relationship between poststratification and weighting. Difficulties that arise with these concepts motivate further development of the model-based poststratification approach (Holt and Smith, 1979; Little, 1991, 1993), which is usefully linked to the more traditional approaches via what we call the basic poststratification identity. Some progress is illustrated with examples, and the need for further development of these ideas is emphasized.

We focus most of our discussion on the problem of estimating the population mean of a univariate survey response in a one-stage sampling design. Section 4 briefly considers more complex estimators (ratios and regression coefficients) and multistage designs, and Section 5 illustrates the potential advantages of a model-based poststratification approach with an example. We conclude in Section 6.

The goal of this chapter is not to make recommendations but rather to lay out the key choices and assumptions that must be made when using weighting and poststratification to correct for non-

*For the International Conference on Survey Nonresponse 1999. To appear in *Survey Nonresponse*, ed. R. Groves, D. Dillman, J. Eltinge, and R. Little. We thank Roderick Little and John Eltinge for helpful comments and the National Science Foundation for support through grant SBR-9708424 and Young Investigator Award DMS-9796129. The second author is grateful to Dr. C. Hendricks Brown and the Department of Epidemiology and Biostatistics, University of South Florida, for sabbatical support during this work.

[†]Department of Statistics, Columbia University, New York, USA

[‡]Clinical Epidemiology and Biostatistics Unit, Royal Children's Hospital and University of Melbourne, Australia

response. Similarly, we do not attempt a thorough literature review; more comprehensive references on weighting and sample survey analysis appear in books such as Lohr (1999), and a companion chapter in this volume (Bethlehem, 2000) provides a broad review of the design-based approach to these issues.

2 Weighting and poststratification: current practices

2.1 Weights

The essential idea of weights in sample survey estimation is that weighted averages over the sample should provide good estimates of the corresponding averages in the target population. The usual way of explaining this is that the weighted estimator will be unbiased for the population mean under repeated sampling that uses the same sampling plan (although poststratification weights cannot be justified in quite this way). The intuitive appeal of weighting seems to be based on a more fundamental notion of creating estimates that correct for differences between sample and population, whether these discrepancies arise from sampling fluctuation, nonresponse, frame errors, or other sources.

Weighted estimation is not something that many non-survey statisticians are familiar with. In fact, most mainstream statistics packages do not provide for inferences (that is, standard errors as well as point estimates) using so-called sampling or "probability" weights. One exception is Stata (StataCorp, 1999) which carefully distinguishes between these sampling weights and so-called "analytic" weights. The latter are weights used in standard regression estimation where the data values themselves each have different (known) variances; in contrast, sampling weights are used for estimating a finite population quantity about which auxiliary information is known. Some similar capabilities are available in the latest release of SAS (An and Watts, 1998).

2.1.1 Where do the weights come from?

Different survey organizations use different weighting schemes, even when using similar methods, asking similar questions to the same populations. The general principle is clearly to do enough weighting to correct for any dramatic discrepancies between sample and population, but just how much is "enough" is not easy to define. For example, Voss, Gelman, and King (1985) report on weights for national political polls by news organizations in the 1988 U.S. general election campaign. At one extreme, some of the ABC News / Washington Post polls weighted only for sex (and in fact these weights were fairly minor, for example, 1.04 for men and 0.96 for women). In contrast, the CBS News / New York Times polls included weights proportional to number of adults in household (see Section 3.3.3) divided by the number of telephone lines, then used ratio weights to match the sample to the population for sex × ethnicity and age × education.

As pointed out by Voss, Gelman, and King (1995), the weighted average estimates for population quantities of interest turned out to be similar for the different survey organizations, which is no surprise since each survey organization used the weighting it deemed necessary to match sample to population.

The CBS example may be used to illustrate an important distinction between two types of weights: inverse-probability and poststratification. The basic difference is that the former are known at the time the survey is designed whereas the latter can only be estimated after the data have been collected. A further distinction among types of inverse-probability weights is that sometimes these are created by the survey designer, for example using probability-proportional-to-size sampling schemes, and sometimes they are a byproduct of a multistage structure, as with the household size weights in the CBS polls.

Poststratification weights are calculated after the data are collected, with the weight (multiplier) for each stratum proportional to the number of units in the stratum in the population, divided by the number of units in the sample in this stratum. In the CBS polls, the final weight for each individual was the product of four factors: two approximate inverse-probability weights and two poststratification weights. Although both kinds of weights have the same intuitive interpretation, they have a different statistical standing, with potential implications for the estimation of standard errors, whether design or model-based; see Section 3.3.

2.1.2 How should the weights be used?

In using sampling weights, it is widely agreed that for obtaining point estimates of population means and ratios, weighted averages are appropriate. The weighted estimate of a mean $\overline{Y}$ is $\sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i$, and the weighted estimate of a ratio $\overline{Y}/\overline{X}$ is $\sum_{i=1}^{n} w_i y_i / \sum_{i=1}^{n} w_i x_i$.

If the aim of analysis is to estimate something more complicated, such as regression coefficients, or if standard errors are required, current practices and textbook recommendations vary. For estimating the coefficients of the regression of y on X, the most commonly used option among most survey analysts is probably to run a weighted regression using the survey weights. The basic weighted average notion is applied to the estimation of regression coefficients by applying the weights to "estimating equations" or pseudo-score statistics, which also take the form of a (weighted) average over the sample (Binder, 1983; Carlin et al., 1999).

Alternatively, statisticians with a stronger model-based persuasion might decide to ignore the weights, on the basis that regression relationships should be validly estimable even from a non-equally weighted sample as long as the model is adequate. Concerns about robustness to model assumptions tend to lead many away from this approach, and an alternative is to try to capture the information in the weighting by including in the model a further set of covariates that contain all the information used in the survey weighting. One then performs an *unweighted* regression of y on the augmented X matrix, and interprets the regression coefficients in the context of the larger model (see DuMouchel and Duncan, 1983, and Pfeffermann, 1993). In a more complicated scenario, interest lies in the regression of y on a subset of the variables in X; a problem we briefly discuss at the end of Section 4.1.

2.2 Poststratification

Poststratification may be defined simply as the use of stratified sample estimators for unstratified designs. Its use is traditionally motivated by the considerable gains in precision that can be made by using information about population structure that is predictive of the survey outcome. However, a more important reason to use poststratification is often as a means of correcting for differential nonresponse between cells. For example, the well-educated are much more likely than the poorly-educated to respond to national telephone opinion polls (see, e.g., Little, 1996).

We use a general definition of poststratification that includes all methods of adjusting the sample to fit known aspects of the population. For example, one can examine the sample averages of some demographic variables and compare them to the population averages (e.g., estimated from the Census or Current Population Survey). If the sample and population differ dramatically on some variables, one can reweight the sample to match the population. Considered from the poststratification perspective, one can consider the survey as giving separate estimates for each demographic category or poststratum, and then these estimates are combined using population totals. We elaborate on this perspective below.

When adjusting for many variables, a standard approach, called *raking*, is to use ratio weights, adjusting for one variable or set of variables at a time. In *iterative proportional fitting*, this weighting procedure is followed several times, looping thorough all of the variables until the weights stabilize (Deming and Stephan, 1940).

3 A unifying framework

3.1 Notation for weighting and poststratification

We have found it useful to develop a unified notation, derived from Little (1991, 1993), for weighting and poststratification of sample surveys. We shall follow standard practice and focus on a single survey response at a time, labeling the values on unit i in the population as Y_i, $i = 1, \dots, N$, and in the sample as y_i, $i = 1, \ldots, n$. To start with, we assume the goal is to estimate the population mean $\theta = \overline{Y} = \sum_{i=1}^{N} Y_i/N$.

We suppose a population is divided into J stratification/poststratification cells, with population N_j and sample size n_j in each cell $j = 1, \ldots, J$, with $N = \sum_{j=1}^{J} N_j$ and $n = \sum_{j=1}^{J} n_j$. For example, if the population of U.S. adults is classified by sex, ethnicity (white or nonwhite), 4 categories of education, 4 categories of age, and 50 states, then $J = 2 \times 2 \times 4 \times 4 \times 50 = 3200$, and the cell populations N_j would be (approximately) known from the public-use subset of the long form of the U.S. Census.

We define π_j as the probability that a unit in cell j in the population will be included in the sample. For some designs, π_j is known but, in general, when nonresponse is present, it can only be estimated. The ratio n_j/N_j is an obvious estimate but this does not take account of unequal-probability sampling designs. Moreover, smoothed estimates can perform better if cell sample sizes are small.

We label the population mean within cell j as $\theta_j = \overline{Y}_j$ and the sample mean within cell j as $\bar{y}_j$. The overall mean in the population is then

$$\theta = \overline{Y} = \frac{\sum_{j=1}^{J} N_j \theta_j}{N}, \tag{1}$$

which we refer to as the basic poststratification identity. We focus on weighted estimates of the form

$$\hat{\theta} = \sum_{j=1}^{J} W_j \hat{\theta}_j, \tag{2}$$

where the *cell weights* W_j sum to 1. So far, equation (2) has no restrictions: the W_j's and the $\hat{\theta}_j$'s can depend in any way on the design and the data.

We use (1) and (2) as a way of unifying a variety of existing estimation procedures. Classical weighting methods generally avoid any modeling of the responses and restrict themselves to unsmoothed estimates $\hat{\theta}_j = \bar{y}_j$ and weights W_j that depend only on the n_j's and N_j's (as well as inverse-probability weights, where present), but not on the y_j's, thus yielding population estimates of the form,

$$\begin{aligned} \hat{\theta}_{\mathrm{w}} &= \sum_{j=1}^{J} W_j \bar{y}_j \\ &= \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \end{aligned} \tag{3}$$

where $w_i = W_{j(i)}/n_{j(i)}$ is the *unit weight* of the items i in cell j. Strictly speaking, the denominator in (3) is unnecessary since, as we have defined them, the w_i's sum to 1, but the general ratio formula is useful when considering arbitrary unnormalized unit weights. The usual challenge for design-based methods is for the unit weights w_i to capture the unequal sampling fractions in the different cells without being so variable as to lead to an unstable estimate of θ.

"Model-based" estimates tackle (2) from a different direction by setting the cell weights to the population proportions—that is, $W_j = N_j/N$ for each j—and using a probability model or smoothing procedure to construct the estimates $\hat{\theta}_j$ from the sample means $\bar{y}_j$ and sample sizes n_j. Thus,

$$\hat{\theta}_{\rm m} = \sum_{j=1}^{J} \frac{N_j}{N} \hat{\theta}_j. \tag{4}$$

The implicit model underlying all these procedures, both design- and model-based, is of equal probability of inclusion in the sample within cells—where the probability encompasses both design and nonresponse issues. This is why $\bar{y}_j$ is considered a reasonable estimate for θ_j. It is also why, in the presence of nonresponse, it is desirable to poststratify as finely as possible, so that the implicit assumption of equal probability of inclusion is reasonable within each poststratification cell (with these probabilities being allowed to vary *between* poststrata).

3.2 Standard errors and inference

An essential part of survey estimation is going beyond point estimates to provide credible measures of uncertainty. We shall follow standard practice here and assume sample sizes are large enough that normal-theory inferences are acceptable, so that we can base inferences on point estimates and standard errors.

For standard survey problems, classical design-based and Bayesian model-based calculations tend to give similar inferences, as long as (a) sample sizes are large enough that sampling distributions of estimands of interest are approximately normal, (b) the inferences take into account design features such as stratification and clustering, and (c) the model uses noninformative prior distributions (see, e.g., Gelman et al., 1995, chapters 4 and 7). This similarity allows one to use model-based calculations to get reasonable repeated-sampling inference or, conversely, to use design-based standard errors to make probability statements about unknown population quantities.

With complex weighting schemes, the design-based perspective can be used to derive variances of weighted/poststratified estimates by accounting for the design factors by using the form (2) and recognizing the sampling variability of the weights W_j. To start with, in simple poststratification, the weights W_j are fixed ($W_j = N_j/N$) and the cell estimates are simply $\hat{\theta}_j = \bar{y}_j$, and so we can use the simple variance formula for a stratified estimate:

$$\text{for simple poststratification:} \quad \text{var}(\hat{\theta}) = \sum_{j=1}^{J} W_j^2 \sigma_j^2 / n_j, \tag{5}$$

where σ_j^2 can be estimated from the within-stratum sample variance. (For expression (5), we ignore the generally very minor correction arising from the randomness of the $1/n_j$ factors.)

With inverse-probability weights, raking, or iterative proportional fitting, the cell weights W_j depend on the vector of data sample sizes $n = (n_1, \ldots, n_J)$. As a result, the sampling variance can be decomposed as

$$\mathrm{var}(\hat{\theta}) = \mathrm{E}(\mathrm{var}(\hat{\theta})|n) + \mathrm{var}(\mathrm{E}(\hat{\theta}|n)), \tag{6}$$

the first term of which is essentially identical to (5) and the second term of which accounts for the randomness in the cell weights. For a complex survey design, (6) can be estimated using linearization or jackknife-type methods (Binder, 1983, Lu and Gelman, 2000).

Design and model-based inferences begin to differ when sample sizes become small or models become more complicated, both of which happen when poststratification is applied with many cells. In this case the sample size in each cell becomes small and design-based approaches may suffer problems of excess variance. For such problems, model-based inferences may provide an attractive alternative, with the assumption of informative, structurally based hierarchical prior distributions (see Sections 3.4 and 3.5 for general discussion and Section 5 for an example). Standard errors for model-based inferences such as $\hat{\theta}$ in (4) come directly from the posterior distribution of the corresponding quantities of interest, such as θ in (2), which would be computed from posterior simulations of the parameter vector θ in a Bayesian analysis (e.g., Gelman et al., 1995).

3.3 Three simple examples illustrating the distinction between inverse-probability weights and poststratification weights

In classical sampling theory, unit weights can be defined in two ways. *Inverse-probability weights* (from Horvitz and Thompson, 1952) are defined as $w_i \propto 1/\pi_{j(i)}$ and *poststratification weights* are defined for unit i in cell j as $w_i \propto N_{j(i)}/n_{j(i)}$; in either case, the classical estimator is the weighted ratio (3). Because the two kinds of weights use the same estimation formula, they are often confused. However, the distinction between them is important (within the design-based perspective), especially when considering more complex weighing adjustments. Here, we illustrate the differences between inverse-probability and poststratification weights using three simple examples.

3.3.1 Unequal response probabilities for men and women: 1

For our first example, we consider a simple random sample (with no nonresponse) of adults with the population divided into two poststrata—men and women—with equal numbers in the population, $N_1 = N_2 =$ 500,000, and a sample of $n = 200$ with $n_1 = 90$ men and $n_2 = 110$ women. For this survey, we are assuming simple random sampling, so the inverse-probability weights are equal for all the units, and the corresponding Horvitz-Thompson estimator, ignoring the poststratification information, is $\hat{\theta}^{\text{H-T}} = \bar{y}$ (see, e.g., Lohr, 1999). The poststratification weights, however, are proportional to $1/90$ for the men and $1/110$ for the women, and so the poststratified estimate is $\hat{\theta}^{\text{PS}} = \frac{1}{2}\bar{y}_1 + \frac{1}{2}\bar{y}_2$.

The two estimates also differ in their standard errors. The Horvitz-Thompson estimate in this case is simply $\bar{y}$, so its standard error is $\sigma/\sqrt{n}$, where σ is the standard deviation of the Y_i values in the population. The poststratified estimate has an approximate standard error of $\sqrt{(N_1/N)^2\sigma_1^2/n_1 + (N_2/N)^2\sigma_2^2/n_2}$, where σ_1 and σ_2 are the within-stratum standard deviations in the population, which are typically smaller than σ (since we tend to poststratify on variables that are relevant for the survey responses of interest). For example, if $\sigma = 30$ and $\sigma_1 = \sigma_2 = 20$, then the Horvitz-Thompson estimate has standard error 2.1 and the poststratification estimate has standard error of approximately[1] 1.4. These are design-based standard errors; normal-theory model-based inferences would give essentially the same results.

Given the population proportions, poststratification is the standard weighting approach in this sort of problem, and it is *not* an inverse-probability weighting in this example; as noted above, the two approaches give different estimates and different standard errors.

3.3.2 Unequal response probabilities for men and women: 2

Conversely, consider the same example—simple random sampling, $n_1 = 90$ and $n_2 = 110$—but this time with $N_1 = 450{,}000$ and $N_2 = 550{,}000$. In this case, the inverse-probability weights are still equal and so the Horvitz-Thompson estimate is still $\bar{y}$, but now the poststratification weights are also equal (because $450{,}000/90 = 550{,}000/110$) and so its corresponding estimate is also $\bar{y}$.

However, in this case, even though the two estimates are the same for this particular sample, they have different properties in repeated sampling. In particular, the poststratification estimate has a lower standard error. In fact, the design-based (or normal-theory model-based) standard errors of the two estimators are as given in Section 3.3.1.

The increased precision of the poststratification estimator in this simple example is due to the conditioning on poststrata, not to any difference between the weighting (since of course there was none). Although in this case the point estimate was the same from the two approaches, in general the poststratification method will provide a better "fit" to the population. In practical terms, an approximate poststratification inference would be obtained in standard survey analysis software by specifying both the weights and the poststrata (as if they were in fact strata), whereas the Horvitz-

[1] This is only the approximate sampling standard error because it conditions on n_1, n_2 rather than treating them as random variables. In this case, however, a simple simulation calculation shows this approximation to be correct to two significant figures.

Thompson estimate would only specify the weights.

This example illustrates the weakness—from a design- or model-based perspective—of trying to obtain standard errors using only weights, without including the information used in constructing the weights.

3.3.3 Adjusting for household size in a survey of individuals

Our second example comes from a real problem in household surveys described in Gelman and Little (1998). In a survey in which households are sampled at random, and then a single individual is sampled from each sampled household, individuals in larger households have a smaller probability of being selected. If individuals within a household are selected with equal probability and there is no nonresponse, then the probability of an individual being included in the survey is inversely proportional to the size of the household. However, composition of the sample is also affected by nonresponse. One source of nonresponse is nonavailability—no one answers the phone, or no one receives the message on the answering machine. It seems reasonable to suppose that in a larger household it is more likely that someone will be home to receive the phone call. Another source of nonresponse is refusal to participate in the survey.

Gelman and Little (1998) compared the distribution of household sizes in the U.S. Census to three series of national opinion polls (two sets of pre-election telephone polls conducted by CBS News and the in-person National Election Study conducted by the University of Michigan) in order to compute the poststratification weights for adults in different household sizes. Table 1 compares these to the inverse-probability weights, which are simply proportional to the number of adults in the household. The poststratification weights for the higher categories are lower than the inverse-probability weights because the realized sample overrepresented the larger households, presumably because adults in smaller households are harder to reach. The discrepancy between the two kinds of weights is smallest with the in-person NES poll, which makes sense since it had the highest response rate of all these surveys.

A simple use of inverse-probability weights in this example will give inferences that overly weight the adults in larger households. Interestingly, Table 1 reveals that the poststratification weights are also less variable than the inverse-probability weights in this example.

A key assumption of the poststratification weighting here is that the Census numbers represent the target population of the survey. In general, this is not exactly the case; for example, the Census includes people who are apathetic about politics, and one might argue that a political poll should represent the people who plan to vote in the election. For this particular example, however, there is no substantial correlation between number of adults in a household and the likelihood of voting. The

Number of adults	Inverse-probability	Poststratification weights for three surveys		
in household	weights	early CBS	late CBS	NES
1	1	1.00	1.00	1.00
2	2	1.32	1.38	2.00
3	3	1.35	1.53	2.30
4+	4.25	0.95	1.20	2.55

Table 1: Inverse-probability weights and poststratification weights for late CBS polls, early CBS polls, and the National Election Study, all scaled so that the weight is 1 for respondents from households with 1 adult. (The inverse-probability weight in the last row is not exactly 4 because that poststratification category includes all households with 4 or more adults.) Systematic discrepancies between the two kinds of weights imply different nonresponse rates among the cells.

discrepancy between the sample and population is more plausibly explained by nonresponse, caused primarily by the difficulty of reaching anyone in a small household (and, as is shown by Gelman and Little, 1998, this differential nonresponse remains after adjusting for the demographic variables of sex, ethnicity, age, and education). If an analyst wishes to adjust the survey for household size, it seems to us much more reasonable and in line with other survey practice to poststratify rather than trying to adjust for unequal sampling probabilities while ignoring nonresponse.

3.4 Difficulties with classical weighted estimates

Estimate (3) is unbiased under the sampling design if the cell weights W_j are set to N_j/N, which corresponds to unit weights $w_i \propto N_{j(i)}/n_{j(i)}$ for units i in cell j. As the weighting (3) indicates, if these unit weights are too variable, then $\hat{\theta}$ will itself have an unacceptably high variance, and this will occur if the n_j's are small. There is thus a tension between two competing alternatives: (a) keeping the number of weighting cells small, so that the individual n_j's will be reasonably large and the weighted estimate not too variable; and (b) increasing the number of cells, which may make the implicit assumption of equal probability of inclusion within cells more plausible (in the presence of nonresponse). A commonly-used compromise is to keep a large number of weighting cells but to "smooth the weights": that is, to set the unit weights so that they are less variable than would arise from simply setting $w_i \propto N_{j(i)}/n_{j(i)}$. In practice, this means that units from cells with small sample sizes receive smaller weights than they would under the unbiased estimate.

As discussed by Elliott and Little (1999), stable estimates of θ in (2) can be obtained in two ways: by smoothing the weights W_j, or by estimating the cell means θ_j using a hierarchical model. The difference between these approaches is that smoothing of weights is usually done without reference to the responses y_i, whereas the amount of smoothing in a fitted hierarchical model depends on the variance of y_i between and within strata.

An extreme version of the instability problem occurs with non-structural zero cells: that is, cells for which $n_j = 0$ but $N_j \neq 0$. This can obviously happen; for example, if $J = 3200$ as in the second paragraph of Section 3.1 and n is 1500, say, which is typical in national polls, then by necessity most of the cells will be empty. In this case the classical solution is to adjust based on margins using the method of raking discussed at the end of Section 2.2, or to pool some weighting cells. The choice of which margins to adjust for or which cells to pool is somewhat arbitrary and contradicts the goal of including in the analysis all variables that affect the probability of inclusion, which is a basic principle in both classical and Bayesian sampling inference.

3.5 Difficulties with model-based estimates

Unfortunately, existing model-based estimates also have drawbacks. Most importantly, there is an understandable resistance on the part of survey sampling practitioners to the use of models for survey response, since models do not seem necessary when using standard design-based methods with moderate or large sample sizes. From this point of view, modeling assumptions can appear somewhat arbitrary and concerns arise as to the possible sensitivity of inferences to alternative model specifications.

At a minimum, model-based methods should be able to give similar answers to classical methods in settings where the classical methods make sense. This means that models for survey outcome variables must at least include all the information currently used in weighting estimates. This appears to be possible in principle for poststratification variables (see Section 5 for some examples), where including indicator variables in a regression model may arguably achieve the purpose, but it is less clear how information used in unequal-probability sampling schemes should be handled.

Once we have resolved to include in our model all the variables affecting the probability of selection, the problem arises that the resulting model becomes quite complicated and requires many assumptions. In the notation of Section 3.1, we must model the θ_j's conditional on all crossclassification variables—for example age, sex, ethnicity, education, age, and state—and all their interactions. The challenge is to construct a class of models for this problem for which the resulting inferences based on (4) are reasonable.

4 More complicated settings

We briefly discuss how the theoretical framework of the previous section can be applied to multistage designs and estimates more complicated than sample means.

4.1 Estimating ratios and regression coefficients

So far we have focused on estimating the population mean (1) or subgroup means. In general, however, one may be interested in more complex estimands, most notably ratios and regression estimates.

Ratios arise in various ways, perhaps the most common being means of subgroups with unknown population proportions. For example, suppose we are interested in θ, the average income of supporters of the Republican candidate for President. If we let Y_i be the income response in the population and U_i be the indicator for supporting the Republican, then

$$\theta = \frac{\sum_{i=1}^N U_i Y_i}{\sum_{i=1}^N U_i} = \frac{\overline{V}}{\overline{U}},$$

where $V_i = Y_i U_i$. Classically, the bias correction and standard error of a ratio estimate are estimated using Taylor expansion; in the Bayesian context, one would need to model Y and U jointly (perhaps by modeling U given X, then Y given (U, X), where X represents the variables that determine the poststratification categories described in Section 3.1).

Regression estimates commonly arise in analytical studies of sample survey responses that attempt to understand what variables U are predictive of an outcome of interest Y. This can be directly incorporated into our framework by including the variables in U as predictors, thus regressing Y on (U, X), where X represents the variables used in any weighting and poststratification. The implicit assumption underlying all weighting and poststratification is equal probability of inclusion in the sample conditional on these X variables, and so any analysis that conditions on X (in this case, a regression of Y on (U, X)) would yield valid inferences without any need for weighting in the estimates; see DuMouchel and Duncan, 1983).

For many problems, this result will be satisfactory. For example, Gelman and King (1993) model vote preferences as a function of party identification and political ideology (these variables represent U in our notation) as well as demographics (the variables X used in the weighting) using an unweighted regression (see also the rejoinder in Gelman, King, and Liu, 1998).

But what if one is ultimately interested in the regression of Y on U without conditioning on X? We can derive this "marginal regression" from our joint regression of Y on (U, X) by averaging over X, which means averaging over the distribution of X in the population, and then the weights come back in, to adjust for differences between sample and population. Unfortunately, the weighting is complicated by the presence of the additional predictors U. The marginal regression can be written as,

$$\mathrm{E}(Y|U) = \mathrm{E}(\mathrm{E}(Y|U, X)) = \sum_X p(X|U)\mathrm{E}(Y|U, X),$$

where the left side represents the predictive relation of interest, and the summation on the right side requires an additional modeling of the joint distribution of (U, X) required to estimate the conditional distribution $p(X|U)$. (If U were not there, this would simply be $p(X)$ which corresponds to the cell populations N_j used in poststratification.) Modeling and estimating $p(X|U)$ seems like a reasonable task but we are not aware of any examples in the literature.

4.2 Cluster sampling and unequal sampling probabilities

In a cluster sampling design, the population or subset of the population is partitioned into clusters, only some of which are sampled. This fits naturally into a hierarchical model that includes a parameter for each cluster. The key difference from stratification or poststratification is the need to generalize to the unsampled clusters. In the hierarchical model, this corresponds to additional parameters drawn from the estimated common distribution. In addition, depending on the design, it may also be necessary to estimate the population sizes of the clusters (the N_j's in (1)). This problem becomes more elaborate with unequal probability sampling designs such as probability proportional to size, where it is possible for sampling probabilities π_j to be known even though population sizes N_j are not, which adds another difficulty to inferences based on the poststratification identity (1). There are various reasons why one might want to define poststratification cells for which the population totals N_j are unknown, and one example is described in the next section.

Challenges arise when clustering is combined with weighting or poststratification. For example, consider a national survey of personal interviews that is clustered geographically (for example, with 20 persons interviewed in each of 50 counties selected at random with probability proportional to size) and then poststratified by demographics (for example, age, sex, ethnicity, etc.). A full modeling analysis requires inference based on (4) requires estimates $\hat{\theta}_j$ for cells j defined by all the counties in the U.S. crossclassified by all the demographic categories used in the weighting/poststratification. This should be possible but experience is limited with this sort of modeling, and further research is needed to see what sort of relatively simple models could work reliably here.

For completeness, we review how the standard weighting methods for cluster sampling can be understood in terms of unit weights as in (3). The first stage of weighting based on sample design: the weight w_i for each of the units in sampled cluster k is set to $N_k/(n_k p_k)$, where N_k is the number of units in the cluster, n_k is the number of units in the cluster included in the sample, and p_k is the probability of selection of cluster k (see, e.g., Lohr, 1999). This weighting expression is general enough to include sampling with probability proportional to size, or approximate measure-of-size, and the n_k in the denominator automatically corrects for unequal nonresponse between clusters (implicitly assuming, as is standard with these weighting or modeling procedures, that nonresponse is uncorrelated with the outcome under study). Once these weights are obtained, one can follow up with standard ratio weighting to poststratify on demographics, as discussed in Section 2.1.1.

As always, the standard error of the weighted estimate (3) should be computed based on the sampling design, not simply using the weights. A generally reasonable approximation is to compute standard errors conditional on the observed cluster means (see Kish, 1965). That is, if clusters $k = 1, \ldots, K$ have been sampled, to express estimate (3) as a weighted average of cluster means:

$$\hat{\theta}_w = \frac{\sum_{k=1}^{K} \omega_k z_k}{\sum_{k=1}^{K} \omega_k}, \tag{7}$$

where $\omega_k = \sum_{i \in k} w_i$ and $z_k = \sum_{i \in k} w_i y_i / \sum_{i \in k} w_i$. The variance of $\hat{\theta}_w$ in (7) can be computed using standard ratio estimation formulas or the jackknife (see, e.g., Lohr, 1999).

4.3 Item nonresponse

Weighting and poststratification are designed to correct for missing data at the unit level, whether the missingness arises by design (i.e., a survey is not a census, so many if not most of the units in the population are missing from the sample) or by nonavailability or nonresponse. As discussed in Section 3.3.3, poststratification can be used to correct for nonresponse in the context of unequal sampling probabilities.

Once we are working in a modeling framework with parameters θ_j for mean response within poststrata, it is natural to consider modeling individual responses, and then to go the next step and model individual responses to the set of survey questions as a multivariate outcome. Such a multivariate model can be used to impute missing items. It is in fact already becoming common to use model-based imputation methods for item nonresponse (e.g., Rubin, 1996, Schafer, 1997), but unit nonresponse is still usually handled by weighting methods. A full multivariate analysis of survey responses could in principle be used to model both kinds of nonresponse.

5 Poststratification in model-based inference: examples

The model-based approach can allow improved estimation when *partial* information is available on a variable that is predictive of the survey response of interest, in other words when there is partial but not complete information on the poststratification cell totals N_j. For example, Reilly and Gelman (1999) analyze a series of national opinion polls, focusing on the question of how strongly the respondent approves of the President's job performance. A highly effective predictor of Presidential approval is the respondent's "party identification," which can be Democrat, Republican, or neither. Using party identification as a poststratifier would seem to be hopeless since it is itself known only from opinion polls. However, polls in a closely-spaced time series are available, and party identification is known (and observed) to change only slowly over time. It was thus possible to fit a time-series model to party identification, estimate the N_j's for the cells corresponding to the three categories, and use these to poststratify and get lower-variance estimates of average Presidential approval at each time point. In this example, weekly snapshots of public opinion were analyzed, with sample sizes of about 40 to 60 in each survey. The estimates formed by the model-based poststratification reduced the estimation variances by factors of about 1.3 relative to the simple unadjusted estimates. This is an example of the modeling of population distributions of poststratifiers that we believe warrants further development.

A more detailed example will illustrate use of the model-based poststratification approach in combination with model-based small-area estimation methods (e.g., Fay and Herriot, 1979, Dempster and Raghunathan, 1987) to get inferences about subpopulations of interest. In effect, the model-based approach allows us to estimate population averages in a large number of poststrata, in settings where the poststratum means are too variable to be directly useful.

We illustrate the potential power of this approach with an example from Gelman and Little (1997). The goal in that paper was to get separate estimates for each state from a series of national pre-election polls. Two natural approaches to this problem are (1) the classical method of assigning unit-level design-based weights and then computing weighted means for each state, and (2) the naive Bayesian method of shrinking the mean of each state toward the national average, with the amount of shrinkage determined by the variance of the binomial distribution for the sample mean in each state. Both these approaches have problems, however: the classical method yields highly variable estimates for all but the largest states, and the naive Bayesian method ignores design information and thus fails to correct for known sampling biases (for example, that women and more educated persons are more likely than men and less educated persons to be reached and respond to a telephone survey).

For this problem, we focused on the binary response y_i equal to 1 if the respondent supported or leaned toward supporting the Republican candidate for President and 0 if the respondent supported or leaned toward the Democrat. (Respondents who supported other candidates or had no opinion were excluded from our analysis.) Our approach was to fit a model of the form $n_j\bar{y}_j \sim \text{Bin}(n_j, \theta_j)$ and $\text{logit}(\theta_j) = (X\beta)_j$, with X including indicators for each state and for all the demographic variables used by the survey organization's own weighting: sex × ethnicity, age × education, and region of the country. We modeled the state indicators as random effects, so that the Bayesian inference shrinks the state differences toward zero *after adjusting for the variables that affect nonresponse.* We view this Bayesian model as *design-based* in that it uses the design information that had been recognized

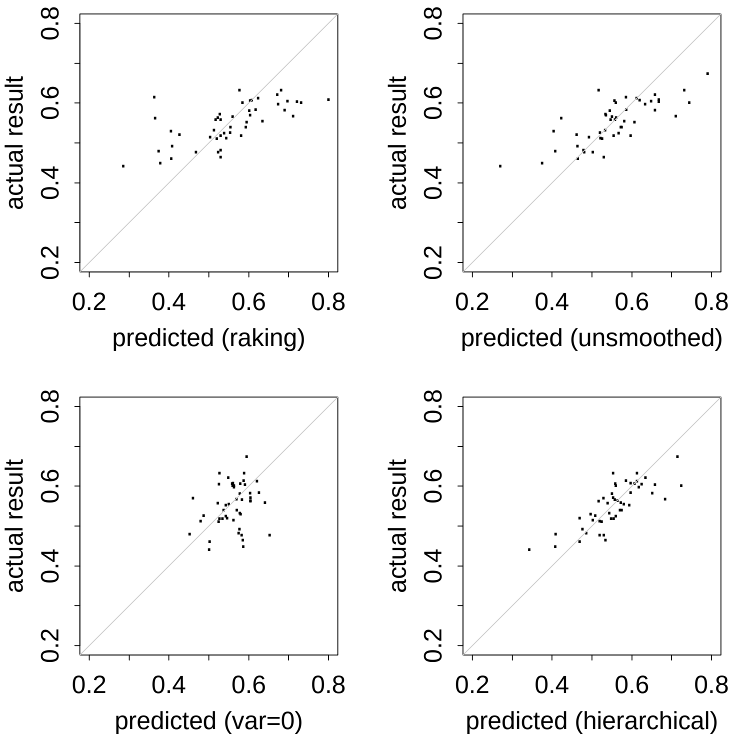

Figure 1: Election result by state, vs. posterior median estimate based on (a) raking on demographics, (b) regression model including state indicators with no hierarchical model, (c) regression model setting state effects to zero, (d) regression model with hierarchical model for state effects.

as relevant by the survey organization.

Once the model has been fitted and inferences obtained for all θ_j's, the key poststratification step is performed, computing estimates of the population mean within each state k as $\theta_k = \sum_{j\in k} N_j\theta_j / \sum_{j\in k} N_j$, where the summations are over all poststratification cells within state k, and the cell sizes N_j are given from the Census. Our Bayesian computation yields 1000 posterior simulation draws of the vector β; from each simulated vector β is computed the vector of cell means θ_j, which are summed to yield the vector of state means θ_k. For each θ_k, we can take the 1000 simulation draws and compute a point estimate as the median of the draws and 50% or 95% intervals from the appropriate quantiles (see, e.g., Gelman et al., 1995).

This approach of smoothing and poststratification performs quite well, as we can see by comparing our inferences, which were based on polls immediately preceding the presidential election, to the state-by-state outcomes of the election itself. Figure 1 displays result vs. prediction, by state, for four estimation methods: classical weighting ("raking"), Bayesian estimate with hierarchical variance set to infinity ("unsmoothed," which does no shrinkage and is thus very similar to the classical estimate,

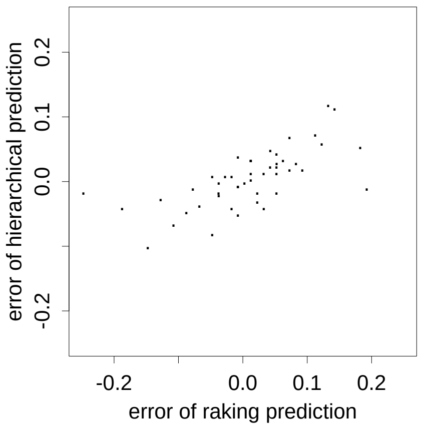

Figure 2: Scatterplot of prediction errors, by state, for the hierarchical model vs. the classical raking estimate. The errors of the hierarchical model are lower for most states.

as expected), Bayesian estimate with hierarchical variance set to 0 ("var=0," which overshrinks by assuming that states are identical after demographic adjustments), and finally hierarchical Bayes, which has the lowest prediction errors. This is a fair test of the model: the actual election results were *not* used in any way in the estimation procedure.

In addition, Figure 2 shows in a state-by-state comparison that the poststratified hierarchical estimates had lower errors than the classical weighting in 41 out of the 48 states (Alaska and Hawaii were not included in these surveys). In this example, the Bayes estimate worked well because it used all the information that was used in classical weighting, but in a model-based context.

An important feature of the model-based approach is its direct computation of posterior uncertainties. The average width of the 50% intervals for the 48 state estimates is 0.57, and 20 out of the 48 intervals contain the actual result for that state. (By comparison, the model-based 50% intervals for the raking estimates have an average width of 0.69, and only 18 of these intervals contain the actual results.)

Finally, Figure 1d shows that the hierarchical model does not seem to shrink the data enough toward the nationwide mean. As discussed by Gelman and Little (1997) and Little and Gelman (1996), this extra variation in the predictions could be caused by a pattern of nonignorable nonresponse that varies between states; see also Krieger and Pfeffermann (1992).

6 Conclusion

> "... it is the structure of the population, rather than the sample design, which an estimator should reflect." — Holt & Smith, 1979

This quotation reflects a belief that we find reasonable, despite the emphasis of many sample survey texts on sample design as the primary basis for deriving estimates from sample surveys. The use of weights, whether inverse-probability or based on poststratification, is traditionally supported with the concepts of unbiasedness and efficiency from the design-based approach to survey inference. We believe that it may be helpful to shift the emphasis somewhat, toward regarding weights as a tool for ensuring that inferences reflect as well as possible the structure of the target population. Extending this notion suggests that other efforts to capture population structure as part of the survey analysis task will be fruitful, and we have described examples where this was achieved through appropriate modeling.

The pre-election polls example in Section 5 illustrates how one can attack the problem of large numbers of poststrata, which challenges traditional "design-based" methods. This example also shows how a successful "model-based" approach works by conditioning on variables relevant in the design and nonresponse and then using population information on these variables to estimate population averages of interest. (The short example that begins Section 5 illustrates how this model-based poststratification approach can be used when the population stratum sizes are missing.)

We have attempted to clarify some aspects of existing practices and to suggest areas where existing methods may be open to improvement by greater investment in modeling technology. In particular, the goal of conditioning on all variables that might affect nonresponse leads to a large number of potential poststratification cells and thus many parameters θ_j in (1); Section 5 illustrates how hierarchical models can be used to estimate all these parameters simultaneously. Further work is needed, however, to define ways in which the model-based approach can successfully incorporate adjustments that are currently made in practice with operationally straightforward techniques such as inverse-probability weighting and raking of poststratification weights (see Little and Wu, 1991). Our hope is to see a unified approach to survey estimation that combines the benefits of modeling population structure while remaining "backwards compatible" with the more traditional *ad hoc* adjustment techniques.

References

An, A., and Watts, D. (1998). New SAS procedures for analysis of sample survey data. In *SUGI Proceedings.* Cary, N.C.: SAS Institute.

Bethelehem, J. G. (2000). Weighting adjustments for ignorable nonresponse. In this volume.

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

Carlin, J. B., Wolfe, R., Coffey, C., and Patton, G. C. (1999). Analysis of binary outcomes in longitudinal studies using weighted estimating equations and discrete-time survival methods (Tutorial in Biostatistics). *Statistics in Medicine* **18**, 2655–2679.

Dempster, A. P., and Raghunathan, T. E. (1987). Using a covariate for small area estimation: a common sense Bayesian approach. In *Small Area Statistics: An International Symposium*, ed. R. Platek, J. N. K. Rao, C. E. Sarndal, and M. P. Singh, 77–90. New York: Wiley.

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

Elliott, M. R., and Little, R. J. A. (1999). Model-based alternatives to trimming survey weights. Technical report, Department of Biostatistics, University of Michigan, Ann Arbor.

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.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Gelman, A., and King, G. (1993). Why are American Presidential election campaign polls so variable when votes are so predictable? *British Journal of Political Science* **23**, 409–451.

Gelman, A., King, G., and Liu, C. (1998). Not asked and not answered: multiple imputation for multiple surveys (with discussion and rejoinder). *Journal of the American Statistical Association* **93**, 846–874.

Gelman, A., and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology* **23**, 127–135.

Gelman, A., and Little, T. C. (1998). Improving upon probability weighting for household size. *Public Opinion Quarterly* **62**, 398–404.

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

Horvitz, D. G., and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47**, 663–685.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Krieger, A. M., and Pfeffermann, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology* **18**, 225–239.

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.

Little, R. J. A., and Wu, M. M. (1991). Models for contingency tables with known margins when target and sample populations differ. *Journal of the American Statistical Association* **86**, 87–95.

Little, T. C. (1996). Models for nonresponse adjustment in sample surveys. Ph.D. thesis, Department of Statistics, University of California, Berkeley.

Little, T. C., and Gelman, A. (1998). Modeling differential nonresponse in sample surveys. *Sankhya* **60**, 101–126.

Lohr, S. L. (1999). *Sampling: Design and Analysis.* Pacific Grove, Ca.: Brooks-Cole.

Lu, H., and Gelman, A. (2000). Sampling variances for classical survey weighting and poststratification. Technical report, Department of Statistics, Columbia University.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review* **61**, 317–337.

Reilly, C., and Gelman, A. (1999). Post-stratification without population level information on the post-stratifying variable, with application to political polling. Submitted to *Journal of the American Statistical Association.*

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics* **12**, 1151–1172.

Rubin, D. B. (1996). Multiple imputation after 18+ years (with discussion) *Journal of the American Statistical Association* **91**, 473–520.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data.* London: Chapman and Hall.

Skinner, C. J., Holt, D. and Smith, T. M. F. (1989). *Analysis of Complex Surveys.* New York: Wiley.

Smith, T. M. F. (1994). Sample surveys 1975–1990; an age of reconciliation? (with discussion). *International Statistical Review* **62**, 5–34.

StataCorp (1999). *Stata Statistical Software: Release 6.0*, College Station, Texas: Stata Corporation.

Thompson, M. E. (1997). *Theory of Sample Surveys.* London: Chapman and Hall.

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.