Poststratification Into Many Categories Using Hierarchical Logistic Regression*

Andrew Gelman, Department of Statistics, Columbia University, New York, NY 10027
Thomas C. Little, Morgan Stanley Dean Witter, New York, NY

September 30, 1997

Abstract

A standard method for correcting for unequal sampling probabilities and nonresponse in sample surveys is poststratification: that is, dividing the population into several categories, estimating the distribution of responses in each category, and then counting each category in proportion to its size in the population. We consider poststratification as a general framework that includes many weighting schemes used in survey analysis (see Little, 1993). We construct a hierarchical logistic regression model for the mean of a binary response variable conditional on poststratification cells. The hierarchical model allows us to fit many more cells than is possible using classical methods, and thus to include much more population-level information, while at the same time including all the information used in standard survey sampling inferences. We are thus combining the modeling approach often used in small-area estimation with the population information used in poststratification.

We apply the method to a set of U.S. pre-election polls, poststratified by state as well as the usual demographic variables. We evaluate the models graphically by comparing to state-level election outcomes.

Keywords: Bayesian inference, election forecasting, nonresponse, opinion polls, sample surveys

1 Introduction

It is standard practice for weighting in opinion polls to be based entirely or primarily on poststratification, which we use generally to refer to any estimation scheme that adjusts to population totals. The basic approach is to divide the population into a number of categories, within each of which the survey is analyzed as simple random sampling. The poststratification step is to estimate population quantities by averaging estimates in the categories, counting each category in proportion to its size in the population. Poststratification categories are typically based on demographic characteristics (sex, age, etc.) as well as any variables used in stratification. Another level of complication, which we do not address here, would occur under cluster sampling.

There is a fundamental difficulty in setting up poststratification categories. It is desirable to divide the population into many small categories in order for the assumption of simple random

*To appear in *Survey Methodology*, December, 1997. We thank Xiao-Li Meng and several reviewers for helpful comments and the National Science Foundation for grant DMS-9404305 and Young Investigator Award DMS-9457824.

sampling within categories to be reasonable. But if the number of respondents per category is small, it is difficult to accurately estimate the average response within each category. For example, if we poststratify by sex, ethnicity, age, education, and region of the U.S., some cells may be empty in the sample, whereas others may have only one or two respondents.

A general solution to this problem is to model the responses conditional on the poststratification variables (see Little, 1993). For example, the standard approach to adjusting for several demographic variables is to rake across one-way or two-way margins (i.e., iterative proportional fitting, Deming and Stephan, 1940), which essentially corresponds to poststratification on the complete multi-way table, but with a model of the responses, conditional on the demographic variables, that sets higher-level interactions to zero. Methods based on smoothing weights can also be viewed as poststratification, with corresponding models on the responses (see Little, 1991). When the poststratification categories follow a hierarchical structure (for example, persons within states in the U.S.), one can improve efficiency of estimation by fitting a hierarchical model (e.g., Lazzeroni and Little, 1997). In the related context of regression estimation, Longford (1996) demonstrates the potential for hierarchical linear models to improve the precision of small area estimates based on sample survey data.

In this paper, we set up a hierarchical logistic regression model to be used for poststratification estimates for a binary variable. The advantage of the model, compared to standard poststratification, is that it allows for the use of many more categories, and thus much more detailed population information. The practical gains from this method are greatest for small subgroups of the population. We apply the method to the state-level results of a set of U.S. pre-election polls. This example has the nice feature that we can check our inferences externally by comparing to state-level election outcomes. Details appear in an appendix for computing the hierarchical model using an approximate EM algorithm.

2 Model

2.1 Sampling and poststratification information

Consider a partition of the population into R categorical variables, where the rth variable has J_r levels, for a total of $J = \prod_{r=1}^{R} J_r$ categories (cells), which we label $j = 1, \ldots, J$. Assume that N_j, the number of units in the population in category j, is known for all j. Let y be a binary response of interest; label the population mean response in each category j as π_j. Then the overall population mean is $\overline{Y} = \sum_j N_j \pi_j / \sum_j N_j$. Assume that the population is large enough that we can ignore all finite-population corrections.

A sample survey is now conducted in order to estimate $\overline{Y}$ (and perhaps some other combinations of the π_j's). For each j, let n_j be the number of units in category j in the sample. Conditional on the

R explanatory variables, assume that nonresponse is ignorable (Rubin, 1976). Thus, the R variables should include all information used to construct survey weights, as well as any other variables that might be informative about y.

For the example we shall consider in Section 3, we categorize the population of adults in the 48 contiguous U.S. states by $R = 5$ variables: state of residence, sex, ethnicity, age, and education, with $(J_1, \ldots, J_5) = (48, 2, 2, 4, 4)$. (Ethnicity, age, and education are discretized into 4 categories each, as described in Section 3.1.) The $J = 3072$ categories range from "Alabama, male, black, 18–29, not high school graduate" to "Wyoming, female, nonblack, 65+, college graduate," and, from the U.S. Census, we have good estimates of N_j in each of these categories. We shall consider population estimates (summing over all 3072 categories) and also estimates within individual states (separately summing over the 64 categories for each state). It is impossible for a reasonably-sized sample survey to allow independent estimates of the mean responses π_j for each category j (in fact, the vast majority of categories will be empty or contain just one respondent), and so it is necessary to model the π_j's in order to poststratify and thus make use of the known category sizes N_j. The (potential) advantage of poststratification is to correct for differential nonresponse rates among the categories.

2.2 Regression modeling in the context of poststratification

One can set up a logistic regression model for the probability π_j of a "yes" for respondents in category j:

$$\text{logit}(\pi_j) = X_j\beta, \tag{1}$$

where X is a matrix of indicator variables, and X_j is the jth row of X. If we were to assume a uniform prior distribution on β, then Bayesian inference, for different choices of X, under this model corresponds closely to various classical weighting schemes. These correspondences, which we present below, are general and rely on the linearity of the assumed model (that is, $X_j\beta$ in (1)). (In the case of binary data, which we are considering in this paper, the classical and uniform-prior-Bayesian estimates are not identical, because of the nonlinear logistic transformation in (1), but for large samples the differences are minor.)

The following models correspond to the most commonly-used classical poststratification estimates.

- Setting X to the $J \times J$ identity matrix corresponds to weighting each unit in cell j by N_j/n_j; that is, simple poststratification. This method is well known to work well only if the n_j's are reasonably large (and it will not work at all if $n_j = 0$ for any j).

- If we set X to the $J \times (\sum_{r=1}^{R} J_r)$ matrix of indicators for each individual variable, then the estimate of $\overline{Y}$ corresponds approximately to that obtained by raking across all R one-way margins.

- Including various interactions in X corresponds to including these same interactions in the raking. To put it most generally, assuming "structure" of any kind in X corresponds to pooling the poststratification across cells in some way.

- Including no explanatory variables in the model (that is letting X be simply a vector of 1's) leads to the sample mean estimate $\bar{y}$.

See Holt and Smith (1979) and Little (1993) for more discussion of the relation between weighting estimates and poststratification.

2.3 Hierarchical regression modeling for partial pooling

When the number of cells is large, none of the above options makes efficient use of the information provided by the categories (for example, simple poststratification gives estimates that are too variable, but if we exclude explanatory variables with many categories, we are discarding important information). Instead, we allow *partial* pooling across cells by setting up a mixed-effects model (see, e.g., Clayton, 1996). We write the vector β as $(\alpha, \gamma_1, \ldots, \gamma_L)$, where α is a subvector of unpooled coefficients and each γ_l, for $l = 1, \ldots, L$, is a subvector of coefficients (γ_{kl}) to which we fit a hierarchical model:

$$\gamma_{kl} \overset{\text{ind}}{\sim} \mathrm{N}(0, \tau_l^2), \;\; k = 1, \ldots, K_l$$

Setting τ_l to zero corresponds to excluding a set of variables; setting τ_l to ∞ corresponds to a noninformative prior distribution on the γ_{kl} parameters.

Given the responses y_i in categories j, we construct an $n \times J$ categorization matrix C, for which $C_{ij} = 1$ if respondent i is in cell j. Let $Z = CX$. The model (1) then can be written in the standard form of a hierarchical logistic regression model as

$$\begin{aligned} y_i &\sim \text{Bernoulli}(p_i) \\ \text{logit}(p_i) &= Z\beta \\ \beta &\sim \mathrm{N}(0, \Sigma_\beta), \end{aligned}$$

where Σ_β^{-1} is a diagonal matrix with 0 for each element of α, followed by τ_l^{-2} for each element of γ_l, for each l. We use the notation p_i, for the probability corresponding to the unit i, as distinguished from π_j, the aggregate probability corresponding to the category j. See Nordberg (1989) and Belin et al. (1993) for general discussions of hierarchical logistic regression models for survey data.

2.4 Inference under the model

To perform inferences about population quantities, we use the following empirical Bayes strategy: first, estimate the hyperparameters τ_l, given the data y; second, perform Bayesian inference for the regression coefficients β, given y and the estimated τ_l's; third, compute inferences for the vector of cell means $\pi = \text{logit}^{-1}(X\beta)$; fourth, compute inferences for population quantities by summing $N_j\pi_j$'s. We view this approach as an approximation to the full Bayesian analysis, which averages over the parameters τ_l. The two approaches will differ the most when components τ_l are imprecisely estimated or are indistinguishable from 0 (see for example, Gelman et al., 1995, Section 5.5). In the example we consider here, this is not a problem because the various components are clearly estimated to be different from 0. If this were not the case, it would probably be worth putting in the additional programming effort for a full Bayes analysis. The focus of this paper, however, is on the effectiveness of combining hierarchical modeling with poststratification, not on the relatively minor technical differences between Bayes and empirical Bayes analyses.

The shrinkage of the cell estimates comes in the second step, and the amount of shrinkage depends both on the sample sizes n_j and the data $\overline{y}_j$. More shrinkage occurs for smaller values of n_j and for values of $\overline{y}_j$ far from the predictions based on the logistic regression model. In addition, more shrinkage occurs if the parameters τ_l are small. A batch of coefficients γ_l with little predictive power will be shrunk toward zero in the estimation, because τ_l will be estimated to have a small value. This is how we can include a large number of coefficients in the hierarchical model without the estimates of population quantities becoming too variable.

3 Application: breaking down national surveys by state

3.1 Survey data

We apply the above methodology to state-by-state results from seven national opinion polls of registered voters conducted by the CBS television network during the two weeks immediately preceding the 1988 U.S. Presidential election. To follow our general notation, we assign $y_i = 1$ to supporters of Bush and $y_i = 0$ to supporters of Dukakis; we discard the respondents who expressed no opinion (about 15% of the total; we follow standard practice and count respondents who "lean" toward one of the candidates as full supporters). Since no data were collected from Hawaii and Alaska, only the 48 contiguous states are included in the model. Washington, D.C., although included in the surveys, was excluded from this analysis because its voting preferences are so different from the other states that a generalized linear model that fit the 48 states would not fit D.C. well, and as a result, the data from D.C. would unduly influence the results for the states. Since there are few observations for the smaller states and the between-poll variation in the estimated support for Bush is within binomial sampling variability (as measured by a χ^2 test of equality of the proportions of support for Bush in the seven polls), we combine the data from all the polls.

CBS creates survey weights by raking on the following variables, with default classifications for item nonresponse shown in brackets:

Census region:	Northeast, South, North Central, West
sex:	male, female
ethnicity:	black, [white/other]
age:	18-29, 30-44, [45-64], 65+
education:	not high school grad, [high school grad], some college, college grad

The raking includes all main effects plus the interactions of sex $\times$ ethnicity and age $\times$ education. We include all these variables as fixed effects in our logistic regression model, excluding from our analysis the relatively few respondents with nonresponse in any of the demographic variables. The CBS weights also correct for number of telephone lines and number of adults in household, which affect sampling probabilities; these have minor effects on estimates for Presidential preference (see Little, 1996, chapter 3), and we do not include them in our model. Further details of the CBS survey methodology and adjustment appear in Voss, Gelman, and King (1995).

Our model goes beyond the CBS analysis by including indicators for the 48 states as random effects, clustered into four batches corresponding to the four census regions. We check the performance of the model by comparing estimates for each state to the observed Presidential election. (Opinion polls just before the election are reliable indicators of the actual election outcome; see, e.g., Gelman and King, 1993.) We also compare the stability of estimates based on different polls over a short period of time.

3.2 Population data for poststratification

In order to poststratify on all the variables listed above, along with state, we need the joint population distribution of the demographic variables within each state: that is, population totals N_j for each of the $2 \times 2 \times 4 \times 48$ cells of sex $\times$ ethnicity $\times$ age $\times$ state. Since the target population is registered voters, we should use the population distribution of registered voters. As an approximation to that distribution we use the crosstabulations available in the Public Use Micro Survey (PUMS) data for all citizens of age 18 and over. The PUMS data contain records for 5% of the housing units in the U.S. and the persons in them, including over 12 million persons and over 5 million housing units. These data are a stratified sample of the approximately 15.9% of housing units that received long-form questionnaires in the 1990 Census. Persons in institutions and other group quarters are also included in the sample. Weights are given for both the housing unit and persons within the unit based on sampling probabilities and adjustment to Census totals for variables included in the short-form questionnaire. We use the weighted PUMS data to estimate N_j for each poststratification category and ignore sampling error in these numbers. The weighted PUMS numbers are very similar to the poststratification numbers used by CBS in their raking (see Little, 1996, chapter 3).

3.3 Results

We present results for four methods applied to the combined data from the seven surveys:

1. Classical estimate based on raking by demographic variables (region, sex, ethnicity, age, education, sex $\times$ ethnicity, and age $\times$ education). This is very close to the weighting method used by CBS. For estimates of results by states, we perform weighted averages within each state, using the weights obtained by the raking.

2. Regression estimate using the demographic variables and also indicators for the states, with no hierarchical model (i.e., "fixed-effects" regression). This is very similar to using iterative proportional fitting to rake on states as well as demographics. The state-by-state estimates from this model should improve upon those obtained by raking on demographics because the estimates of π_j's are weighted by the population numbers N_j rather than the sample numbers n_j within each state.

3. Regression estimate using only the demographic variables, with the state effects set to zero. This model allows the average responses within states to differ *only* because of demographic variation; to the extent that the demographics do not explain all the variation in opinion, the model should underestimate the variability between states.

4. Regression estimate using the demographic variables, with the 48 state effects estimated with a hierarchical model (in the notation of Section 2, $L = 4$ and $K_1, K_2, K_3, K_4 = 12, 13, 12, 11$). We expect this model to perform best, both because of the flexibility of the hierarchical regression model and because the poststratification uses the population numbers N_j.

We fit each of the regression models to the survey data, obtain posterior simulation draws for each coefficient (conditional on the estimated $\tau_1, \tau_2, \tau_3, \tau_4$), and reweight based on the PUMS data to obtain poststratified estimates for the proportion of registered voters in each state who support Bush for President.

Table 1 presents the raking estimate and the posterior medians and interquartile ranges for the three models, along with data on the survey responses and the actual election outcome. Table 2 gives the nationwide and mean absolute statewide prediction errors for the raking and the three models. The four methods give almost identical results at the national level; the real gain from the model-based estimates occurs in estimating the individual states. The reduction in mean absolute prediction error from about 6% to 5% can be attributed to using the poststratification information, with the further reduction to 3.5% attributable to the hierarchical modeling. In addition, the last two lines of Table 2 show that the uncertainty estimates from the hierarchical model are short and relatively well calibrated (slightly less than half of the true values fall inside the 50% intervals, which is reasonable since these intervals account only for sampling error and not for nonsampling errors and changes in opinion).

Figure 1 plots, by state, the actual election outcomes vs. the raking estimates and the posterior medians for the three models. As one would expect, the hierarchical model reduces variance, and thus estimation error, by shrinkage. Although the four methods correct the bias of the nationwide estimate by about the same amount, they act differently on the individual states, with the hierarchical model performing best. Figure 2 compares the prediction errors for the hierarchical and raking estimates for the states.

Interestingly, the hierarchical model does not seem to shrink the data enough to the nationwide mean: we can tell this because, in Figure 1d, the actual election outcome is higher than predicted for low-predicted values, and lower than predicted for high-predicted values. *Under*shrinkage means that the estimated parameters $\hat{\tau}_l$ are probably *higher* than their true values, which could be caused by a pattern of nonignorable nonresponse that varies between states so that observed variability in the state proportions is caused by varying nonresponse patterns as well as actual variation in average opinions (see Little and Gelman, 1996, for a discussion of this example and Krieger and Pfeffermann, 1992, for a more general treatment). The undershrinkage could be quantified by comparing the estimated to the optimal level of shrinkage, but this comparison can only be made after the true values are observed.

It is also possible to compare the models by fitting each separately to each survey and examining the stability of estimates over a short period of time. This would be a more reasonable way to study the models in the common situation that the true population means never become known. Figure 3 displays, for each of our seven surveys, the estimates from raking and from the hierarchical model. (When modeling the surveys individually, we fit a common hierarchical variance for all 48 states because there was not enough data to obtain reliable maximum likelihood estimates for the four regions separately from the data in each poll.) Results are shown for the entire United States and for three representative states: California (a large state), Washington (mid-sized), and Nevada (small). For convenience, the plot also show the estimates based on the seven surveys pooled and the actual election outcomes. For all the individual states, the hierarchical estimate is less variable over time than the raking estimate. The pattern is clearest in Nevada, where the sample size for the individual surveys was so low that the raking estimate degenerated to 0 or 1 in most cases, but the better performance of the hierarchical model is clear in the other states as well. For example, it was not reasonable to assign Bush only 46% of the support in California (in the poll 3 days before the election) or only 30% of the support in the state of Washington. For the United States as a whole, however, the two estimates are quite similar (in fact, when all seven polls are combined, the raking estimate performs very slightly better), indicating once again that the benefits from the modeling approach appear when studying subsets of the population.

The results for Washington have the surprising property that the regression estimate based on the combined surveys (shown at time "-1" on the graph) is lower than the seven estimates from the original surveys. This occurs because the data from the combined surveys show that the state of Washington supports Bush less than would be predicted merely by controlling for the demographic covariates (that prediction would be the estimate for Washington from the model with state effects set to zero, which from Table 1 is 0.58). But none of the individual surveys, taken alone, had enough data to make a convincing case that Washington was so far from the national mean, and so the Bayes estimate shrunk their estimates to a greater extent. This behavior, while it may seems strange at first, is in fact appropriate: with a smaller survey, there is less information about the individual poststratification categories, and the model-based estimate produces an estimate for each category that is closer to the sample mean. When all seven surveys are combined, more information is available, and the model relies more strongly on the data in each category. This is how the Bayes procedure essentially balances the concerns of poststratifying on too few or too many categories.

4 Discussion

Poststratification is the standard method of correcting for unequal probabilities of selection and for nonresponse in sample surveys. From the modeling perspective, raking or poststratification on a set of covariates is closely related to a regression model of responses conditional on those covariates, with population quantities estimated by summing over the known distribution of covariates in the population. Conditioning on more fully-observed covariates allows one to include more information in forming population estimates, but it is well known that raking on too large a set of covariates yields unacceptably variable inferences. We propose a method of poststratification on a large set of variables while fitting the resulting regression with a hierarchical model, thus harnessing the well-known strengths of Bayesian inference for models with large numbers of exchangeable parameters.

The Bayesian poststratification is most useful for estimation in subsets of the population (e.g., individual states in the U.S. polls) for which sample sizes are small. A related area in which modeling should be effective is in combining surveys conducted by different organizations, modeling conditional on all variables that might affect nonresponse in either survey. In addition, the methods in this paper can obviously be applied to continuous responses by replacing logistic regressions by other generalized linear models.

Our purpose in Bayesian modeling is not to fit a subjectively "true" model to the data or the underlying responses, but rather to estimate with reasonable accuracy the average response conditional on a large set of fully-observed covariates. More accurate models of the responses should allow more accurate inferences—but even the simple exchangeable mixed effects model we have fit, with hyperparameters estimated from the data, should perform better than the extremes of the fixed effects model or setting coefficients to zero. Ultimately, the goal of probability modeling and Bayesian inference in a sample survey context is to allow one to make use of abundant poststratification information (e.g., census data classified by sex, ethnicity, age, education, and state) to adjust a relatively small sample survey.

Difficulties with modeling approaches such as ours could arise in several ways. If one adjusts to a large number of categories using too weak a model (such as the model with unsmoothed state effects), the resulting estimates can be too variable. If the population distributions of the variables used in the poststratification are not available (for example, adjusting to a variable that is not measured or is measured inaccurately by the Census), then the N_j's must be modeled also, which requires additional work. Of course, such additional work would be required to rake on these variables as well. Since all of the methods, including raking and regression methods, assume ignorable models, they will yield incorrect inferences when unmeasured variables affect nonresponse and are correlated with the outcome of interest.

The methods described here are intended as an improvement upon raking-type poststratification adjustments and are not intended to, by themselves, correct for nonignorable nonresponse. However, by allowing one to adjust for more variables, the Bayesian poststratification should allow the use of models for which the ignorability assumption is more reasonable. Having a large number of poststratification categories (e.g., in 48 states) creates problems with classical weighting methods because many categories will have few or even no respondents. Interestingly, however, having many categories can make Bayesian modeling more reliable: more categories means more random effects in the regression, which can make it easier to estimate variance components.

Appendix: Computation

We use an EM-type algorithm to estimate the hyperparameters τ_l; given these, we sample from the posterior distribution of the coefficients β using a normal approximation to the logistic regression likelihood. We use this approximation for its simplicity and because it is reasonable for fairly large surveys, as in our application in Section 3; if desired, more exact computations can be performed using the Gibbs sampler and Metropolis algorithm (see Clayton, 1996), perhaps using the algorithm described here as a starting point.

When the data distribution is normal and the means are linear in the regression coefficients, the EM algorithm can be used to obtain estimates of the variance components (Dempster, Laird, and Rubin, 1977), treating the vector of coefficients β as "missing data." In this framework, the "complete-data" loglikelihood for τ_l is

$$L(\tau_l|\gamma_l) = \text{const} - K_l \log \tau_l - \frac{1}{2\tau_l^2} \sum_{k=1}^{K_l} \gamma_{kl}^2,$$

so the sufficient statistic for τ_l is $t(\gamma_l) = \sum_{k=1}^{K_l} \gamma_{kl}^2$. Given the current estimate τ^{old}, the expected sufficient statistic is

$$\mathrm{E}(t(\gamma_l)|y, \tau^{\text{old}}) = ||\mathrm{E}(\gamma_l|y, \tau^{\text{old}})||^2 + \text{ trace}(\text{var}(\gamma_l|y, \tau^{\text{old}})).$$

Since these two terms are not analytically tractable for our model, we use the following approximations which are easily obtained: (1) approximate $\mathrm{E}(\gamma_l|y, \tau^{\text{old}})$ with an estimate $\hat{\gamma}_l$, based on y and the estimate τ^{old}, and (2) approximate $\text{var}(\gamma_l|y, \tau^{\text{old}})$ from the curvature of the log-likelihood at the estimate, $\hat{V}_{\gamma_l} = (-L''(\hat{\gamma}_l))^{-1}$. We update these approximations iteratively for all $l = 1, \ldots, L$ simultaneously, converging to an approximate maximum likelihood estimate $(\hat{\tau}_1, \ldots, \hat{\tau}_L)$. Given an initial guess τ^{old}, the algorithm proceeds by iterating the following two steps to convergence.

Approximate E-step. Solve the likelihood equations iteratively, as described below. Use the estimate $\hat{\beta}$ to obtain an approximation to $\mathrm{E}(t(\gamma_l)|y, \tau^{\text{old}})$, for each $l = 1, \ldots, L$.

We solve the likelihood equations $\frac{d}{d\beta} L(\beta|y, \tau) = 0$ using iteratively weighted least squares, involving a normal approximation to the likelihood $p(y|\beta) = \prod_i p(y_i|\beta)$, based on locally approximating the logistic regression model by a linear regression model (see Gelman et al., 1995, p. 391). Let $\eta_i = (Z\beta)_i$ be the linear predictor for the ith observation. Starting with the current guess of $\hat{\beta}$, let $\hat{\eta} = Z\hat{\beta}$. Then a Taylor series expansion to $L(y_i|\eta_i)$ gives $z_i \approx N(\eta_i, \sigma_i^2)$, where

$$\begin{aligned} z_i &= \hat{\eta}_i + \frac{(1+\exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)} \left(y_i - \frac{\exp(\hat{\eta}_i)}{1+\exp(\hat{\eta}_i)} \right) \\ \sigma_i^2 &= \frac{(1+\exp(\hat{\eta}_i))^2}{\exp(\hat{\eta}_i)}. \end{aligned}$$

Let $\widehat{\Sigma}_\beta$ denote the value of Σ_β based on plugging in the current estimate $\hat{\tau}$, and let $\widehat{\Sigma}_z = \text{diag}(\sigma_i^2)$. Then we obtain an updated estimate and variance matrix using weighted least squares based on the normal prior distribution and the normal approximation to the logistic regression likelihood:

$$\hat{\beta} = (Z'\widehat{\Sigma}_z^{-1}Z + \widehat{\Sigma}_\beta^{-1})^{-1}Z'\widehat{\Sigma}_z^{-1}z \quad (2)$$

$$\widehat{V}_\beta = (Z'\widehat{\Sigma}_z^{-1}Z + \widehat{\Sigma}_\beta^{-1})^{-1}. \quad (3)$$

We iterate until convergence and then use $\hat{\beta}$ and the appropriate elements of $\widehat{V}_\beta$ to estimate $\mathrm{var}(\gamma_l|y,\tau^{\mathrm{old}})$.

M-step. Maximize over the parameters τ_l to obtain $\tau_l^{\mathrm{new}} = (\widehat{\mathrm{E}}(t(\gamma_l)|y,\tau^{\mathrm{old}})/K_l)^{1/2}$, for each $l = 1,\ldots,L$. Set τ^{old} to τ^{new} and return to the approximate E-step.

Once the approximate EM algorithm has converged to an estimate $\hat{\tau}$, we draw β from a normal approximation to the conditional posterior distribution $p(\beta|y,\hat{\tau})$, using the values from equations (2) and(3) at the last EM step as the mean and variance matrix in the normal approximation. For each draw of the vector parameter β, we compute the category means, $\pi = \mathrm{logit}^{-1}(X\beta)$, and any population totals of interest, counting each category j as N_j units in the population.

References

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. M. (1993). Hierarchical logistic regression models for imputation of unresolved enumeration status in undercount estimation (with discussion). *Journal of the American Statistical Association* **88**, 1149–1166.

Clayton, D. G. (1996). Generalized linear mixed models. In *Practical Markov Chain Monte Carlo*, ed. W. Gilks, S. Richardson, and D. Spiegelhalter, 275–301. New York: Chapman & Hall.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

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.

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

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

Lazzeroni, L. C., and Little, R. J. A. (1997). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, to appear.

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, 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. (1996). A model for differential nonresponse in sample surveys. Technical report.

Longford, N. T. (1996). Small-area estimation using adjustment by covariates. *Qüestió* **20**, to appear.

Nordberg, L. (1989). Generalized linear modeling of sample survey data. *Journal of Official Statistics* **5**, 223–239.

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.

State	Election result	Sample size	Unweighted mean	1: Raking est.	Poststratification estimates (and IQRs) 2: state effs unsmoothed	3: state effs set to 0	4: Hierarchical model
AL	0.60	134	0.72	0.67	0.63 (0.05)	0.56 (0.01)	0.62 (0.05)
AR	0.57	86	0.57	0.53	0.53 (0.06)	0.60 (0.01)	0.55 (0.06)
AZ	0.61	141	0.62	0.61	0.62 (0.05)	0.56 (0.02)	0.61 (0.05)
CA	0.52	1075	0.57	0.53	0.55 (0.02)	0.53 (0.01)	0.55 (0.02)
CO	0.54	126	0.59	0.59	0.58 (0.06)	0.57 (0.01)	0.57 (0.05)
CT	0.53	103	0.53	0.55	0.52 (0.06)	0.49 (0.02)	0.51 (0.06)
DE	0.56	30	0.40	0.37	0.42 (0.11)	0.60 (0.01)	0.52 (0.08)
FL	0.61	553	0.64	0.62	0.61 (0.03)	0.62 (0.01)	0.61 (0.03)
GA	0.60	211	0.62	0.58	0.56 (0.04)	0.56 (0.01)	0.56 (0.04)
IA	0.45	102	0.38	0.38	0.38 (0.06)	0.59 (0.01)	0.41 (0.06)
ID	0.63	31	0.52	0.58	0.52 (0.12)	0.59 (0.02)	0.55 (0.08)
IL	0.51	429	0.55	0.52	0.53 (0.03)	0.52 (0.01)	0.52 (0.03)
IN	0.60	215	0.75	0.73	0.74 (0.04)	0.56 (0.01)	0.72 (0.04)
KS	0.57	105	0.72	0.71	0.71 (0.06)	0.57 (0.01)	0.68 (0.05)
KY	0.56	146	0.57	0.53	0.56 (0.05)	0.64 (0.01)	0.57 (0.05)
LA	0.55	153	0.62	0.60	0.61 (0.05)	0.54 (0.01)	0.59 (0.04)
MA	0.46	277	0.47	0.41	0.46 (0.04)	0.50 (0.02)	0.47 (0.04)
MD	0.51	207	0.52	0.50	0.49 (0.04)	0.56 (0.01)	0.50 (0.04)
ME	0.56	44	0.52	0.52	0.55 (0.10)	0.52 (0.02)	0.54 (0.08)
MI	0.54	399	0.58	0.55	0.57 (0.03)	0.54 (0.01)	0.57 (0.03)
MN	0.46	210	0.54	0.53	0.53 (0.05)	0.59 (0.01)	0.53 (0.04)
MO	0.52	235	0.46	0.43	0.46 (0.04)	0.55 (0.01)	0.47 (0.04)
MS	0.61	170	0.69	0.70	0.65 (0.04)	0.53 (0.01)	0.63 (0.04)
MT	0.53	31	0.39	0.40	0.40 (0.12)	0.58 (0.02)	0.50 (0.09)
NC	0.58	239	0.59	0.60	0.55 (0.04)	0.58 (0.01)	0.55 (0.04)
ND	0.57	54	0.56	0.56	0.55 (0.09)	0.58 (0.01)	0.56 (0.08)
NE	0.61	90	0.58	0.60	0.56 (0.07)	0.58 (0.01)	0.56 (0.06)
NH	0.63	20	0.70	0.68	0.73 (0.13)	0.53 (0.02)	0.61 (0.10)
NJ	0.57	301	0.57	0.60	0.53 (0.04)	0.46 (0.01)	0.53 (0.03)
NM	0.53	87	0.55	0.54	0.57 (0.07)	0.54 (0.02)	0.56 (0.06)
NV	0.61	19	0.68	0.80	0.67 (0.13)	0.56 (0.02)	0.60 (0.09)
NY	0.48	639	0.42	0.37	0.41 (0.03)	0.45 (0.01)	0.41 (0.02)
OH	0.55	454	0.62	0.63	0.58 (0.03)	0.55 (0.01)	0.58 (0.03)
OK	0.58	93	0.57	0.62	0.59 (0.07)	0.63 (0.01)	0.60 (0.06)
OR	0.48	111	0.50	0.47	0.50 (0.06)	0.58 (0.02)	0.52 (0.06)
PA	0.51	431	0.54	0.54	0.52 (0.03)	0.48 (0.02)	0.52 (0.03)
RI	0.44	65	0.28	0.29	0.27 (0.07)	0.50 (0.02)	0.34 (0.06)
SC	0.62	151	0.70	0.67	0.66 (0.05)	0.55 (0.01)	0.64 (0.04)
SD	0.53	52	0.54	0.51	0.53 (0.09)	0.58 (0.01)	0.54 (0.08)
TN	0.58	252	0.68	0.69	0.66 (0.04)	0.60 (0.01)	0.65 (0.03)
TX	0.56	594	0.58	0.52	0.56 (0.03)	0.60 (0.01)	0.56 (0.02)
UT	0.67	61	0.80	0.85	0.79 (0.07)	0.60 (0.02)	0.72 (0.06)
VA	0.60	255	0.69	0.72	0.67 (0.04)	0.59 (0.01)	0.66 (0.03)
VT	0.52	12	0.54	0.58	0.60 (0.19)	0.53 (0.02)	0.55 (0.11)
WA	0.49	269	0.47	0.41	0.46 (0.04)	0.58 (0.01)	0.48 (0.04)
WI	0.48	264	0.49	0.53	0.48 (0.04)	0.57 (0.01)	0.49 (0.04)
WV	0.48	79	0.48	0.52	0.48 (0.07)	0.65 (0.01)	0.53 (0.06)
WY	0.61	13	0.50	0.36	0.59 (0.17)	0.59 (0.02)	0.59 (0.10)

Table 1: By state: election results (proportion of the two-party vote in 1988 received by Bush); survey data (unweighted mean and sample size) from the combined surveys; raking estimate using CBS variables; and posterior median (and interquartile range; that is, width of the central 50% uncertainty interval) of poststratified estimates based on state effects unsmoothed, set to zero, and fit by a hierarchical model. Estimates are labeled 1, 2, 3, 4 corresponding to the descriptions in Section 3.3.

Summary	Actual result	Unweighted mean	Raking est.	state effs unsmoothed	state effs set to 0	Hierarchical model
mean of national popular vote	0.539	0.568	0.549	0.548	0.547	0.550
mean absolute error of states	—	0.056	0.066	0.049	0.048	0.035
average width of 50% intervals	—	—	—	(0.069)	(0.016)	(0.057)
# states contained in 50% interval	—	—	—	18	3	20

Table 2: Summary statistics for raw mean of responses, raking estimate, and three poststratified estimates from the combined surveys. Summaries given are the estimated mean of the 48 state vote proportions weighted by state voter turnout (thus, estimated national popular vote proportion for Bush, excluding Alaska, Hawaii, and the District of Columbia); the mean absolute error of the 48 state estimates; the average width of the 50% intervals for the states; and the number of the 48 states whose true values fall within the 50% intervals.

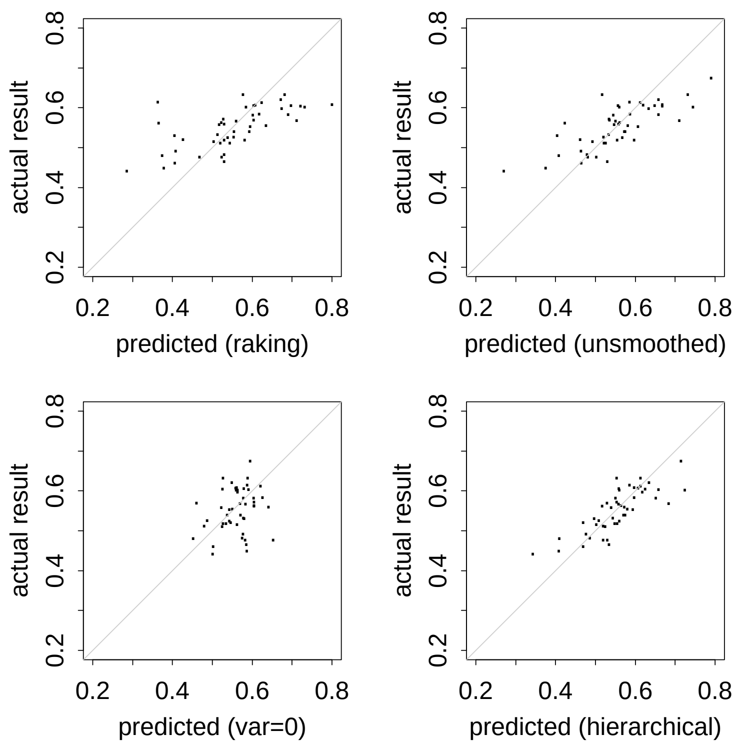

Figure 1: Election result by state, vs. posterior median estimate for (a) raking on demographics, (b) regression model including state indicators with no hierarchical model, (c) regression model setting state effects to zero, (d) regression model with hierarchical model for state effects.

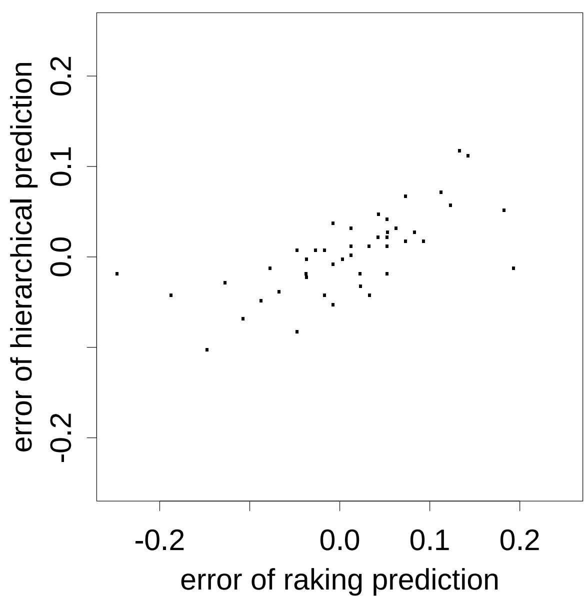

Figure 2: Scatterplot of prediction errors, by state, for the hierarchical model vs. the raking estimate. The errors of the hierarchical model are lower for most states.

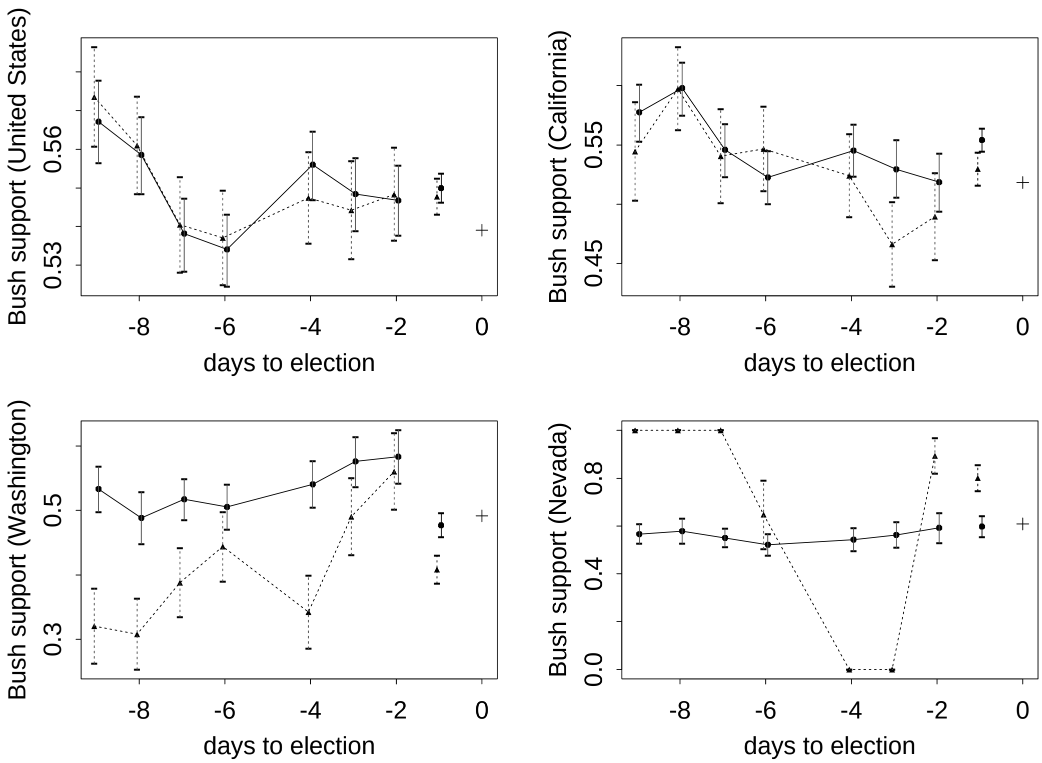

Figure 3: Estimated Bush support estimated separately from seven individual polls taken shortly before the election: for (a) the entire U.S. (excluding Alaska, Hawaii, and the District of Columbia), (b) a large state (California), (c) a medium-sized state (Washington), and (d) a small state (Nevada). Each plot shows the raking estimates as a dotted line and the estimates from hierarchical model as a solid line, with error bars indicating 50% confidence bounds for the raking and 50% posterior intervals for the model-based estimates. The polls were taken between nine and two days before the election. Estimates based on the combined surveys are shown at time "−1," and the actual election result is shown at time "0" on each plot.