Modeling Differential Nonresponse in Sample Surveys*

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

November 12, 1997

Abstract

The standard analysis of unit nonresponse in sample surveys is to assume missing at random—that is, that the probability a person responds is independent of their response to the question of interest, y, conditional on fully-observed covariates x or on sampling weights w. In this paper, we discuss weakening these assumptions without the use of additional covariates in the special case of a binary outcome variable, $y = 0$ or 1. We note frequentist confidence bounds that do not rely on strong assumptions about the response mechanism. From a Bayesian perspective, we discuss using prior distributions to average over uncertainty in the missing data mechanism. Surprisingly, a natural-looking "noninformative" prior distribution yields unappealing posterior inferences. We discuss methods of constructing informative prior distributions using hierarchical data structures.

We also show how to incorporate unequal sampling weights into the model using design-based sampling theory. This is important so that the nonresponse modeling can be an improvement upon rather than merely a replacement for standard weighted analysis of sample surveys.

We illustrate the hierarchical model by applying it to the state-level analysis of a series of national pre-election opinion polls. The use of a reasonable prior distribution for the relative response probabilities leads to substantial improvements in coverage of posterior intervals and prediction error of point estimates. We also consider the sensitivity to the prior distribution and the effect of including sampling weights in the analysis.

Keywords: Bayesian inference, hierarchical models, opinion polls, sampling weights

1 Introduction

1.1 Background

Response rates in the most carefully conducted academic and government surveys rarely exceed 70%, and response rates for surveys conducted by the large commercial polling organizations are generally much lower, in the 30% to 50% range (Brady and Orren, 1992). These high nonresponse rates leave the possibility of large biases in the estimates, so that sampling error accounts for only a small fraction of the total uncertainty. Weighting based on poststratification or sampling probabilities is often used as a corrective (see Little, 1991, 1993; Kish, 1992), but large potential biases remain (see Kish, 1965). Since these biases cannot be estimated from the data, they are generally ignored:

*To appear in *Sankhya*. We thank Henry Brady, Xiao-Li Meng, Donald Rubin, and a reviewer for helpful comments and the National Science Foundation for grant DMS-9404305 and Young Investigator Award DMS-9457824.

that is, it is assumed that data are missing completely at random or, when covariates are observed, missing at random, as defined in Rubin (1976).

In this paper, we set up a framework for including differential nonresponse rates for the case of a binary outcome variable y, where the parameter of interest is the population mean $\pi = \Pr(y = 1)$. We do not assume that the response mechanism is known, and so an identifiable parameterization of the likelihood which includes the population mean of Y does not exist. However, we derive confidence bounds for the population mean based on the sample mean and response rate.

From a Bayesian perspective, one can average over the prior distribution of the nonidentified parameters in the model. Important issues that arise in this context include: (1) parameterizing the model so that one can set up a reasonable class of prior distributions; (2) understanding the behavior of posterior inferences in the limit of large sample size; (3) methods of constructing informative prior distributions using external sources of data; and (4) understanding the sensitivity to the prior distribution of inferences for parameters of interest. To illustrate the methodology, we apply the model to the state-level analysis of a series of national pre-election opinion polls. We find that the use of a reasonable hierarchical model for the relative response probabilities can lead to substantial improvements in coverage of posterior intervals and prediction error of point estimates.

Rubin (1977) describes a similar Bayesian method to account for nonresponse in the normal case when covariates are available. However, in his paper, a prior distribution is specified for the parameters in the likelihood of the response variable Y conditional on whether an individual is a respondent or a nonrespondent. In contrast, here we specify a prior distribution for the relative response probabilities in the two groups characterized by $Y = 0$ and $Y = 1$ (see also the related work of Kaufman and King, 1973). Nordheim (1984) also allows for different classification probabilities for a binary variable, but does not use a prior distribution for these parameters. Section 2 of this paper reviews the basic results for bounding inferences given nonresponse rates, and Section 3 presents the Bayesian extension, revealing some poor behavior with a seemingly noninformative prior distribution. In Section 4, we relate the nonresponse models to the practial world of sample survey analysis by including in the model a design-based treatment of the nonresponse already accounted for by survey weights. Finally, in Section 5 we apply a hierarchical form of the Bayesian model to estimate state-by-state preferences in pre-election polls. Our work goes beyond the previous literature in this area in its criticism of the noninformative prior distribution, with its handling of unequal sampling weights, and with the hierarchical model for differential nonresponse.

1.2 Notation and model

For simplicity, we first set up the model in the context of simple random sampling; we generalize to unequal sampling probabilities in Section 4. Suppose that $y_1, \ldots, y_n$ are 0/1 responses, with $\pi = \Pr(y_i = 1)$ for each i. Consider the corresponding missing data indicator variables, $I_1, \ldots, I_n$, where unit i responds if $I_i = 1$ and does not respond if $I_i = 0$. In general, the probability of response can depend on the value of y_i. Label the conditional response probabilities as $\theta_0 = \Pr(I_i = 1|y_i = 0)$ and $\theta_1 = \Pr(I_i = 1|y_i = 1)$, and assume $I_i|y_i \overset{\text{ind}}{\sim} \text{Bernoulli}(\theta_{y_i})$. Let $m = \sum_{i=1}^n I_i$ denote the number of units for which y is observed, and let $m_0 = \sum_{i=1}^n I_i(1-y_i)$ and $m_1 = \sum_{i=1}^n I_i y_i$ denote the number of observed units for which $y = 0$ and $y = 1$, respectively. Then the distribution of $(m_0, m_1, n-m)$ is multinomial with density function,

$$p(m_0, m_1, n-m|n, \pi, \theta_0, \theta_1) = \binom{n}{m}\binom{m}{m_1}[(1-\pi)\theta_0]^{m_0}[\pi\theta_1]^{m_1}[(1-\pi)(1-\theta_0)+\pi(1-\theta_1)]^{n-m}. \quad (1)$$

The sample mean for the observed data is $\bar{y} = m_1/m$, and the population probability π is the estimand of primary interest. It can be seen from (1) that, conditional on n, the parameterization of the likelihood in terms of $(\pi, \theta_0, \theta_1)$ is unidentifiable, although the parameterization in terms of $((1-\pi)\theta_0, \pi\theta_1) = (\Pr(I = 1, y = 0), \Pr(I = 1, y = 1))$ is identifiable.

We will work with a parameterization of the model that separates identified and nonidentified parameters. Let $R = \theta_1/(\theta_0 + \theta_1)$; this is a measure of the relative response rates in the two groups characterized by $y = 0$ and $y = 1$. The missing-completely-at-random assumption corresponds to $\theta_0 = \theta_1$, or $R = 0.5$. We also define $\theta = (\theta_0 + \theta_1)/2$, so that the model can be parameterized in terms of (π, R, θ), with the parameters R and θ depending only on the nonresponse rates and not on the responses themselves.

We also will find it useful to work with an alternative parameterization in terms of the expectations of the proportion of respondents and the mean response. Let $p = m/n$ be the observed proportion of respondents; its expectation under the model is

$$\zeta_p = (1-\pi)\theta_0 + \pi\theta_1.$$

We also define

$$\zeta_{\bar{y}} = \frac{\pi\theta_1}{(1-\pi)\theta_0 + \pi\theta_1},$$

so that $\mathrm{E}(\bar{y}|n) = \zeta_{\bar{y}} + O_p(1/n)$.

Finally, we work at first under the assumption that the size of the original sample, n, is known, so that it is possible to construct conservative confidence bounds for π. We consider the case of unknown n in Section 3.4 and in the application in Section 5.

2 Frequentist analysis

We begin with a formal analysis of standard conservative bounds for the inferential errors caused by nonresponse; see Manski (1995) for a general discussion of inference for nonidentified parameters, for which this is a special case. We obtain conservative confidence bounds for π by considering separately the identified and nonidentified parts of the model. Let π_L, π_U be functions of the identifiable parameter vector $((1-\pi)\theta_0, \pi\theta_1)$ defined by $\pi_L = \pi\theta_1$ and $\pi_U = 1-(1-\pi)\theta_0$. Since $0 \le \theta_0, \theta_1 \le 1$ and $0 < \pi < 1$, it follows that $\pi_L \le \pi \le \pi_U$ for all π, θ_0, θ_1. In fact, there do not exist upper and lower boundaries on the parameter space of π which are functions of an identifiable parameter and uniformly closer to π. To see this, in the case of the lower boundary, let f be any function of the identifiable parameter $((1-\pi)\theta_0, \pi\theta_1)$ such that $\pi_L < f((1-\pi)\theta_0, \pi\theta_1) \le \pi$ for some π. This implies a contradiction for $\theta_1 = 1$ since $\pi_L = \pi$.

Lower confidence bounds for π_L and upper confidence bounds for π_U can be used to assign upper and lower confidence bounds for π. The distribution of m_1 under the model is $\mathrm{Bin}(n, \pi_L)$. When $m_1 \ge 1$, a level α lower confidence bound $\underline{\pi}_L$ for π_L, and therefore a conservative level α lower confidence bound for π, is the unique solution to the equation

$$\sum_{r=0}^{m_1} \binom{n}{k} \pi_L^k (1-\pi_L)^{n-k} = 1-\alpha.$$

If $m_1 = 0$ then $\underline{\pi}_L = 0$. Similarly, the distribution of $n - m_0$ is $\mathrm{Bin}(n, \pi_U)$. When $m_0 \ge 1$, a level α upper confidence bound $\overline{\pi}_U$ for π_U, and therefore a conservative level α upper confidence bound for π, is the unique solution to the equation

$$\sum_{k=n-m_0}^{n} \binom{n}{k} \pi_U^k (1-\pi_U)^{n-k} = 1-\alpha.$$

If $m_0 = 0$ then $\overline{\pi}_U = 1$.

If m_0 and m_1 are not too small, one can easily approximate the above bounds using the normal distribution, with the lower bound based on the assumption that $\theta_1 = 1$ (so that we could infer that $y = 0$ for all $n - m$ nonrespondents) and the upper bound based on the assumption that $\theta_1 = 0$ (so that $y = 1$ for all $n - m$ nonrespondents). Each bound is now based on binomial inference for a population of size n, and the normal approximation yields $\underline{\pi}_L = m_1/n - z_{1-\alpha}\sqrt{(m_1/n)(1-m_1/n)/n}$ and $\overline{\pi}_U = (1-m_0/n) + z_{1-\alpha}\sqrt{(m_0/n)(1-m_0/n)/n}$, where $z_{1-\alpha}$ is the appropriate standard normal quantile.

For example, Table 1 illustrates exact and approximate upper and lower conservative confidence bounds at level $\alpha = 0.05$ for π for data with observed mean $\bar{y} = 0.2$, response rate $p = 0.7$, and a range of sample sizes m. Although a response rate of 70% is relatively high for most types of sample surveys, confidence bounds for π remain fairly distant from 0.2 even for large n.

This example illustrates the well-known fact (see, e.g., Cochran, 1977) that model-free conservative confidence bounds for π tend to be so wide as to be often useless, and so it is important to understand how the inference for π depends on the relative response probability $R = \theta_1/(\theta_0 + \theta_1)$. Algebraic manipulation gives the following restrictions on the parameter space of $(\pi, R, \zeta_p, \zeta_{\bar{y}})$ for $(\zeta_p, \zeta_{\bar{y}}) \in (0, 1] \times [0, 1]$:

$$\frac{\zeta_p \zeta_{\bar{y}}}{1 - \zeta_p(1 - 2\zeta_{\bar{y}})} \leq R \leq \frac{1 - \zeta_p \zeta_{\bar{y}}}{1 + \zeta_p(1 - 2\zeta_{\bar{y}})}, \tag{2}$$

$$\zeta_{\bar{y}} - (1 - \zeta_p)\zeta_{\bar{y}} \leq \pi \leq \zeta_{\bar{y}} + (1 - \zeta_p)(1 - \zeta_{\bar{y}}). \tag{3}$$

Also, we can express π in terms of these parameters:

$$\pi = \frac{\zeta_{\bar{y}}(1 - R)}{\zeta_{\bar{y}}(1 - R) + (1 - \zeta_{\bar{y}})(R)}. \tag{4}$$

To understand the purpose of these transformations, first consider the limit $n \to \infty$, in which case there is no sampling variability, and all of the uncertainty about π comes from uncertainty about R. In this limit, ζ_p (the expected proportion of respondents) and $\zeta_{\bar{y}}$ (the expected value of y among all respondents) are known. The parameter $\zeta_{\bar{y}}$ determines the relation between π and R (equation (4)), and then ζ_p, in combination with $\zeta_{\bar{y}}$ gives us bounds on R (equation (2)) and thus π (equation (3)). These relations are illustrated in Figure 1; the curve segment shown in bold on the figure corresponds to the special case of $\zeta_p = 0.7$ and $\zeta_{\bar{y}} = 0.2$, the example considered Table 1.

If n is finite, ζ_p and $\zeta_{\bar{y}}$ are now estimated with binomial error, so that the relation between π and R becomes uncertain.

3 Bayesian analysis

When the proportion of respondents p is not close to 1, confidence bounds remain distant from π even for large n, as in the example summarized in Table 1. Closer bounds can be achieved by introducing assumptions about the relative response rate R. For example, under the missing-completely-at-random assumption $R = 0.5$ it follows that $\mathrm{E}(\bar{y}) = \pi$. This assumption is commonly made in practice, either implicitly or explicitly. Another possibility is to assume that R lies within some range; for example, Figure 1 illustrates that if we know that $R \in [0.4, 0.6]$, our inferences about π become relatively precise for any value of p. From a Bayesian viewpoint, uncertainty in (π, R) can be characterized by a probability distribution. A diffuse prior distribution for the relative response rate R results in increased posterior uncertainty for the parameter of interest π. For example, in the limit of $n \to \infty$, with $p = 0.7$ and $\bar{y} = 0.2$, the bold curve segment in Figure 1 is the support of the likelihood, which is overlain on the prior distribution for (π, R). If n is finite, the likelihood spreads above and below the bold segment and blurs at the endpoints.

A Bayesian analysis requires a prior distribution for all of the parameters in the likelihood. The first step, then, is to choose an appropriate parameterization of the likelihood. The probability π is the parameter of primary interest. If the parameter R is also included in the model, then the parameter $\theta = (\theta_0 + \theta_1)/2$ will complete the specification. From (1), the likelihood in this new parameterization is

$$p(m_0, m_1, n-m|n, \pi, R, \theta) = \binom{n}{m}\binom{m}{m_1}[2(1-\pi)(1-R)\theta]^{m_0}[2\pi R\theta]^{m_1}[2(1-\pi)R\theta+\pi(1-2R\theta)]^{n-m}.$$

This parameterization was chosen instead of $(\pi, \theta_0, \theta_1)$ because it may be more natural for specifying a prior distribution. The parameters θ_0 and θ_1 would probably be correlated; knowledge of the response probability in one group would affect the subjective estimation of the response probability in the other group. On the other hand, it may be reasonable to assume prior independence of π, R, and θ. For a particular specification of the prior $p(\pi, R, \theta)$, inference about π is based on the posterior distribution

$$\begin{aligned} p(\pi|n, m, y) &= \iint p(\pi, R, \theta|n, m, y)d\theta dR & (5) \\ &\propto \iint p(m, y|n, \pi, R, \theta)p(\pi, R, \theta)d\theta dR. & (6) \end{aligned}$$

3.1 Large-sample inference

In most Bayesian models encountered in practice, the likelihood dominates the prior for large n. This is not the case for unidentifiable models, however. Unless the fraction of missing data is small, the posterior distribution is sensitive to the specification of the prior, even for large n. We can see this by separating the posterior distribution into identified and unidentified parts. A prior distribution specified in terms of (π, R, θ) corresponds to a prior distribution for $(\pi, 2(1-\pi)(1-R)\theta, 2\pi R\theta)$, where the observed statistics follow the limits $m_0/n \to 2(1-\pi)(1-R)\theta$ and $m_1/n \to 2\pi R\theta$ in probability as $n \to \infty$. Since the observed data is independent of π given $((1-\pi)(1-R)\theta, \pi R\theta)$, the posterior distribution $p(\pi|n, m, y)$ tends to the conditional prior distribution evaluated at the observed values $p(\pi|2(1-\pi)(1-R)\theta = m_0/n, 2\pi R\theta = m_1/n)$. Then, in the limit as $n \to \infty$, the posterior density for π has the form of the prior density evaluated at particular values of the identified parameters.

Formally, let $h = (h_1, h_2, h_3)$ be the transformation defined on $(0,1)^3$ by $h(x_1, x_2, x_3) = (x_1, 2(1-x_1)(1-x_2)x_3, 2x_1x_2x_3)$, and let $J = x_1(1-x_1)x_3$ be the Jacobian of h. Then

$$\lim_{n\to\infty}\left|p(\pi|n, m, y) - \frac{J^{-1}p(\pi, R = h_2^{-1}(\pi, m_0/n, m_1/n), \theta = h_3^{-1}(\pi, m_0/n, m_1/n))}{\int J^{-1}p(\pi, R = h_2^{-1}(\pi, m_0/n, m_1/n), \theta = h_3^{-1}(\pi, m_0/n, m_1/n))d\pi}\right| = 0$$

in probability. Note that $J = \pi(1-\pi)\theta = \zeta_p((1-\zeta_{\bar{y}})\pi + \zeta_{\bar{y}}(1-\pi))$.

For example, if independent beta distributions are used in the prior so that

$$p(\pi, R, \theta) = \text{Beta}(\pi|a_1, b_1)\text{Beta}(R|a_2, b_2)\text{Beta}(\theta|a_3, b_3),$$

then the conditional prior distribution of π in terms of the other parameters is

$$\begin{aligned}
p(\pi|\zeta_p, \zeta_{\bar{y}}) \quad \propto \quad & ((1-\zeta_{\bar{y}})\pi + \zeta_{\bar{y}}(1-\pi))^{-1} \pi^{a_1-1}(1-\pi)^{b_1-1} \\
& \times \left(\frac{\zeta_{\bar{y}}\pi}{(1-\zeta_{\bar{y}})(1-\pi)+\zeta_{\bar{y}}\pi}\right)^{a_2-1} \left(\frac{(1-\zeta_{\bar{y}})(1-\pi)}{(1-\zeta_{\bar{y}})(1-\pi)+\zeta_{\bar{y}}\pi}\right)^{b_2-1} \\
& \times \left(\frac{(1-\zeta_{\bar{y}})\pi+\zeta_{\bar{y}}(1-\pi)}{\pi(1-\pi)}\right)^{a_3-1} \left(1-\frac{(1-\zeta_{\bar{y}})\pi+\zeta_{\bar{y}}(1-\pi)}{\pi(1-\pi)}\right)^{b_3-1} \\
& \text{for } \pi \in [\zeta_p\zeta_{\bar{y}}, 1-\zeta_p(1-\zeta_{\bar{y}})], \qquad (7)
\end{aligned}$$

and for all π,

$$|p(\pi|n,m,y) - p(\pi|\zeta_{\bar{y}} = \bar{y}, \zeta_p = p)| \to 0$$

in probability as $n \to \infty$.

3.2 Difficulties with a natural "noninformative" prior distribution

Consider the special case of independent uniform (i.e., Beta$(1,1)$) prior distributions on π, R, θ. In the limit, from (7), the posterior distribution for π is just

$$p(\pi|p,\bar{y}) \propto ((1-\bar{y})\pi + \bar{y}(1-\pi))^{-1}, \text{ for } \pi \in [p\bar{y}, 1-p(1-\bar{y})].$$

For example, if $p = 0.7$ and $\bar{y} = 0.2$, and $n \to \infty$, then π must lie within the range $[0.14, 0.44]$ and has density proportional to $1/(0.2+0.6\pi)$. The posterior mean of π is 0.278. The uniform prior distribution is thus not so "noninformative" as one might like, in that it shrinks π quite a ways from the raw estimate $\bar{y}$ toward 0.5. However, the uniform prior distribution is more reasonable if restricted to lie near $R = 0.5$; for instance, with a uniform prior distribution on (π, R, θ) but with R restricted to the range $[0.4, 0.6]$, the posterior mean for π becomes 0.205 in this example.

3.3 Constructing an informative prior distribution

In practice, our model for differential nonresponse rates is not particularly useful unless we have an informative prior distribution for the parameters in the model. As discussed above, it seems reasonable to set up prior distributions for π, R, and θ independently: (1) the distribution for π reflects substantive modeling of the responses in the population without any reference to the sampling mechanism; (2) $R = \theta_1/(\theta_1+\theta_2)$ is the relative rates of response in the two groups; and (3) $\theta = (\theta_1+\theta_2)/2$ reflects the level of response, averaging over the two groups. Models for π are widespread in the survey sampling literature, and we do not add anything to this topic here (see, e.g., Ericson, 1969, Scott and Smith, 1969, Rubin, 1987, Skinner, Holt, and Smith, 1989, Little, 1993, and Nadaram and Sedransk, 1993, for theoretical treatments, and Belin et al., 1993, Lazzeroni and Little, 1997, and Gelman and Little, 1997, for some recent examples). Models for θ are close to irrelevant for the problem of estimating π: what is relevant is the differential rates of nonresponse between the two groups corresponding to $y = 0$ and $y = 1$. Thus, the key part of our nonignorable nonresponse model is the prior distribution for R. The importance of the prior distribution is illustrated by Kadane (1993), who examines the sensitivity of inferences about π to different specified values of R in the context of a sample survey of jurors.

As with other Bayesian models, it is best to construct a prior distribution using some related data—in this case, this would mean other surveys in which the population proportions π were known, so that R could be estimated directly. We can generalize this idea by modeling our survey hierarchically. Suppose the population is divided into J groups, $j = 1, \ldots, J$, with known populations N_j, and it is known which respondents fall into which group. Then our data and model can be given a hierarchical structure: in each group j, we observe responses $y_{j\,1}, \ldots y_{j\,m_j}$, with parameters π_j, R_j, and θ_j. The mean response in the population is $\pi = \sum_j N_j \pi_j / \sum_j N_j$. We can set up a hierarchical model for the J sets of parameters (π_j, R_j, θ_j). Using a hierarchical model, we will be able to estimate some aspects of the prior distribution for these parameters. But, because we do not observe y for the nonrespondents, inferences for π will still depend on the prior distribution for the ensemble of R_j parameters, even in the limit of infinite sample size. The advantage of setting this up as a hierarchical model is that we can take advantage of any knowledge of the *distribution* of the R_j's, without having to accurately estimate any individual R_j ahead of time. We illustrate with an example in Section 5 of a U.S. opinion poll in which the groups j are individual states.

3.4 Proportion of missing data unknown

Sometimes the number of individuals in the original sample, n, and therefore the proportion of respondents $p = m/n$, are not known. This occurs, for instance, in a telephone poll: if no one answers the phone, the survey organization does not know whether no one is at home, or they are not answering the phone, or the phone is a non-residence (Brady and Orren, 1992, discuss the difficulty of estimating nonresponse rates in commercial telephone polls). If n is unknown, the likelihood (1) no longer applies. Instead inference is based on the conditional distribution

$$m_1 | m \sim \mathrm{Bin}(m, \zeta_{\bar{y}}). \tag{8}$$

Although confidence bounds can be found for ζ_π, it is easily seen that, for any n, any frequentist α-level upper and lower confidence bounds for π, the parameter of interest, are simply 0 and 1.

For a Bayesian analysis, we express the marginal posterior density for π as $p(\pi|m, m_1) \propto \int p(m_1|m, \pi, R) p(\pi, R) dR$, ignoring θ because it does not appear in the likelihood (8). Let $g =$

(g_1, g_2) be the transformation defined by

$$g(\pi, R) = \left(\pi, \frac{\pi R}{(1-\pi)(1-R) + \pi R}\right),$$

and let

$$J_g = \frac{(\pi + \zeta_{\bar{y}} - 2\pi\zeta_{\bar{y}})^2}{\pi(1-\pi)}$$

be the corresponding Jacobian. Then the conditional distribution of π is

$$p(\pi|\zeta_{\bar{y}}) \propto J_g^{-1} p(\pi, R = g_2^{-1}(\pi, \zeta_{\bar{y}})), \tag{9}$$

where $g_2^{-1}(\pi, \zeta_{\bar{y}}) = \zeta_{\bar{y}}(1-\pi)/(\zeta_{\bar{y}} + \pi - 2\zeta_{\bar{y}}\pi)$. As before, it follows that for all π,

$$|p(\pi|m, m_1) - p(\pi|\zeta_{\bar{y}} = \bar{y})| \to 0$$

in probability.

4 Accounting for sampling weights

It is standard for sample surveys to include some correction for nonresponse in the form of a *weight* w_i attached to each respondent i. Loosely speaking, w_i is proportional to the number of units in the population "represented" by this respondent. In the context of binary responses, the weights are set so that, assuming ignorable nonresponse, the weighted average $\sum_i w_i y_i / \sum_i w_i$ is intended to be a consistent estimate of the population proportion π. Weights are assigned as a function of measured covariates can be derived based on stratification, poststratification, sampling theory, or more elaborate modeling (see Kish, 1992, Little, 1991, and Pfeffermann, 1993, for recent reviews of these issues); here, we shall treat weights as inverse sampling probabilities, so that Pr(a unit with weight w is included in the set of respondents) $\propto 1/w$. We use this "design-based" perspective because it is standard in the practical analysis of sample surveys, and we want our nonresponse modeling to be an improvement upon rather than merely a replacement for standard weighted analysis of sample surveys.

Sampling weights affect our model of nonresponse because it is possible, and in fact generally occurs, that units with $y_i = 1$ have different weights, on average, than units with $y_i = 0$. That is, one often has direct evidence, from the survey weights themselves, that the response rates in the two groups differ. Obviously, we do not want to go to the trouble of setting up a nonignorable nonresponse model just for the purpose of finding out what we already know. Instead, we want to set up our model conditional on the weights, so that our parameter R represents differential response rates *after weighting.*

We develop a procedure to do this by formally defining, for each unit i in the population, its response Y_i and the weight W_i that would be assigned for that unit, based on the value of its covariates. We combine weighting and our differential nonresponse model as follows:

$$\Pr(\text{unit } i \text{ is included in the set of respondents}|Y_i, W_i) \propto \begin{cases} (1-R)/W_i & \text{if } Y_i = 0 \\ R/W_i & \text{if } Y_i = 1. \end{cases}$$

This reduces to the usual probability weights if $R = 0.5$ and to our earlier model if all weights are equal. Then the probability that a response is $y = 1$ is

$$\begin{aligned} \Pr(y_i = 1) &\propto \sum_i Y_i R/W_i \\ &\propto R \sum_i Y_i \frac{\sum_i Y_i/W_i}{\sum_i Y_i} \\ &\propto RN\pi \mathrm{E}(1/W_i|Y_i = 1). \end{aligned}$$

Similarly,

$$\Pr(y_i = 0) \propto (1-R)N(1-\pi)\mathrm{E}(1/W_i|Y_i = 0).$$

We now define

$$R^W = \frac{\mathrm{E}(1/W_i|Y_i = 1)}{\mathrm{E}(1/W_i|Y_i = 0) + \mathrm{E}(1/W_i|Y_i = 1)},$$

so that

$$\zeta_{\bar{y}} = \frac{\Pr(y_i = 1)}{\Pr(y_i = 0) + \Pr(y_i = 1)} = \frac{\pi R R^W}{(1-\pi)(1-R)(1-R^W) + \pi R R^W}. \tag{10}$$

The parameter R^W represents the differential nonresponse of the two groups *as explained by the weights*. Like R, the parameter R^W must lie in the range $[0, 1]$, and $R^W = 0.5$ corresponds to equal average weights among the two groups.

In general, we cannot know R^W, because it depends on the weights W_i in the population, whereas we only know the values of w_i in the sample. To estimate $\mathrm{E}(1/W_i|Y_i = 1)$ and $\mathrm{E}(1/W_i|Y_i = 0)$, and thus R^W, we use the fact that a consistent estimate of the population mean of any survey variable X is $\widehat{\mathrm{E}}(X) = \sum_i w_i x_i / \sum_i w_i$. Thus, we have

$$\begin{aligned} \widehat{\mathrm{E}}(1/W_i|Y_i = 0) &= \frac{\sum_i (1-Y_i) w_i (1/w_i)}{\sum_i (1-Y_i) w_i} = \frac{1}{\overline{w}_0} \\ \widehat{\mathrm{E}}(1/W_i|Y_i = 1) &= \frac{\sum_i Y_i w_i (1/w_i)}{\sum_i Y_i w_i} = \frac{1}{\overline{w}_1}, \end{aligned}$$

where $\overline{w}_1$ and $\overline{w}_0$ are the mean observed weights for the $y = 1$ and $y = 0$ respondents, respectively. A consistent estimator of R^W is then

$$R^w = \frac{\overline{w}_0}{\overline{w}_0 + \overline{w}_1}. \tag{11}$$

With a large enough sample size, one can simply use R^w in place of R^W; if the sample size is smaller, more sophisticated estimates can do better, as in the example in Section 5.

In either case, one can use expression (10) for $\zeta_{\bar{y}}$ in all formulas, so that the parameter R models only the differential nonresponse not already coded by the weights. We would expect this to pull R closer to 0.5 (since weights are generally explicitly included to make the response pattern closer to missing at random).

5 Application

The modeling described in the previous sections seems highly theoretical, and yet it can affect the analysis of survey data, beyond merely widening confidence intervals to take account of uncertainty about nonrespondents. We illustrate with an analysis of state-level data from nationwide opinion polls in the United States. In this application, it is possible to improve state-level inferences by using a hierarchical model for differential nonresponse that allows the parameters R to vary between states.

5.1 Problem and data

Nonresponse rates for high-quality professional political opinion polls can be in the 50–70% range (see Brady and Orren, 1992), and an obvious concern is differential nonresponse among supporters of two different candidates or positions. We illustrate with an analysis of data from seven national opinion polls conducted by CBS during the two weeks before the 1988 U.S. Presidential election. Figure 2 shows the unweighted and weighted means for Presidential preference at the national level. The weights are based on a combination of probability weighting and raking, performed separately for each survey, based on Census information about the population distribution of sex, race, age, and education. A general discussion of the use of raking to correct for nonresponse can be found in Oh and Scheuren (1983). A variation of random-digit dialing was used to select the sample. Details of the survey methodology and the adjustment appear in Voss, Gelman, and King (1995).

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). Figure 2 shows that the unweighted means are higher than the weighted, which indicates that, according to the weights, supporters of Bush were more likely to respond than supporters of Dukakis. That is, $R^w > 0.5$.

To illustrate our methodology, we fit several models of (π, R) to these data, first ignoring the weights and then including them. For all models, we allow separate parameters (π_j, R_j) for each of the 48 contiguous states j, and for the models that include weights, we allow separate values of R_j^W.

(Alaska and Hawaii were not included in the surveys. The District of Columbia, although included in the surveys, was excluded from analysis because its voting preferences are so different from the other states that it would be unduly influential in our model.) The number of nonrespondents is not known, so that inference is based on the likelihood (8). The target population is taken to be the set of registered voters. We validate the analysis by comparing our results with the November 4 election results, assuming that for each state the election result equals the true proportion of support for the candidate among registered voters.

Since there are few observations for the smaller states, and the between-poll variation displayed in Figure 2 is within binomial sampling variability, we combine the data from all seven polls. The first column in Table 2 gives the actual election results for the 48 contiguous states. Altering the notation for this example, let m_j denote the number of respondents in state j, and let y_j denote the number of those who say they will vote for Bush. The second and third columns in Table 2 give π_j and $\bar{y}_j = y_j/m_j$, respectively, for the 48 continental states. In the following discussion, let $y = (y_j), m = (m_j), \pi = (\pi_j), R = (R_j),\ j = 1, \ldots, 48$.

5.2 Models and estimation

To illustrate the methodology, we consider several models of varying complexity. In all of the models considered, the likelihood is given by

$$y_j | m_j, \zeta_{\bar{y}\,j} \overset{\text{ind}}{\sim} \text{Bin}(m_j, \zeta_{\bar{y}\,j}),$$

where

$$\zeta_{\bar{y}\,j} = \frac{\pi_j R_j R_j^W}{(1-\pi_j)(1-R_j)(1-R_j^W) + \pi_j R_j R_j^W}.$$

We consider prior distributions that are independent in (π, R) with the following form:

$$\begin{aligned} \pi_j &\overset{\text{ind}}{\sim} \text{Beta}(a_1, b_1) \\ R_j &\overset{\text{ind}}{\sim} \text{Beta}(a_2, b_2). \end{aligned}$$

The parameters R_j^W, which depend on the distribution of weights among the two groups of respondents, are assumed fixed in all the analyses.

The likelihood only tells us about the parameters $\zeta_{\bar{y}\,j}$, which depend on both π_j and R_j—so we can estimate the distribution of the π_j's (given a model for the R_j's) or the R_j's (given a model for the π_j's), but not both. The standard approach is to fix $R_j \equiv 0.5$ and estimate the π_j's. We do not do this, but we recognize that, in most surveys, the R_j's should vary less than the π_j's—it is hard to imagine the relative nonresponse probabilities for a question to vary more than the average response itself. In all the models we set up, (a_2, b_2) will be fixed (either set a priori or by using other data), so that only (a_1, b_1) will be estimated from the survey data.

5.2.1 Models for π

We consider two different models:

1. Independent uniform prior distributions; that is, $(a_1, b_1) = (1, 1)$. This is essentially equivalent to estimating each π_j using only the data from state j with no hierarchical model.

2. Hierarchical: (a_1, b_1) estimated from the survey data and the assumed distribution for R. The hyperparameters (a_1, b_1) can be estimated by maximum likelihood:

   (a) For the models with R fixed at 0.5, the marginal likelihood of the data given a_1, b_1 is $p(y|m, a_1, b_1) = \int p(y|m, \pi)p(\pi|a_1, b_1)d\pi = \prod_j p(y_j|m_j, a_1, b_1)$, where

   $$p(y_j|m_j, a_1, b_1) = \frac{\Gamma(m_j + 1)}{\Gamma(y_j + 1)\Gamma(m_j - y_j + 1)} \frac{\Gamma(a_1 + y_j)\Gamma(m_j + b_1 - y_j)}{\Gamma(a_1 + b_1 + m_j)} \frac{\Gamma(a_1 + b_1)}{\Gamma(a_1)\Gamma(b_1)}$$

   is beta-binomial. The marginal likelihood can easily be maximized over (a_1, b_1) numerically.

   (b) For the models in which R has a Beta(a_2, b_2) distribution, the marginal density of the data given all the hyperparameters is

   $$\begin{aligned} p(y|m, a_1, b_1, a_2, b_2) &= \iint p(y, \pi, R|m, a_1, b_1, a_2, b_2)d\pi dR \\ &= \iint p(y|m, \pi, R)p(\pi|a_1, b_1)p(R|a_2, b_2)d\pi dR \\ &= \prod_j \iint p(y_j|m_j, \pi_j, R_j)p(\pi_j|a_1, b_1)p(R_j|a_2, b_2)d\pi_j dR_j \quad (12) \end{aligned}$$

   The parameters (a_2, b_2) are assumed known (more on this in Section 5.2.2). For any (a_1, b_1), (12) can be evaluated by numerical integration. One can then use an optimization routine to find the (a_1, b_1) that maximizes the likelihood (12).

Given the estimated hyperparameters, the posterior distributions for the π_j's are independent; we sample posterior draws using rejection sampling applied to the product of the beta prior density and the likelihood.

5.2.2 Models for R

We consider several different prior distributions for the R_j's:

1. $R_j = 0.5$ for all j: this is the missing-at-random model and corresponds to $(a_2, b_2) = (\infty, \infty)$.

2. Hyperparameters (a_2, b_2) estimated from polls and election results. Although the election results were not available at the time of the surveys, for the purposes of illustration we use the

election results to estimate a prior distribution for R. This model can be expected to perform better than any estimated prior distribution specified at the time of the surveys. The marginal distribution of the data conditional on the election results π is

$$\begin{aligned} p(y|n,\pi,a_2,b_2) &= \int p(y,R|n,\pi,a_2,b_2)dR \\ &= \int p(y|n,\pi,R)p(R|a_2,b_2)dR \\ &= \prod_j \int p(y_j|n_j,\pi_j,R_j)p(R_j|a_2,b_2)dR_j. \end{aligned} \tag{13}$$

For any (a_2, b_2), we evaluate (13) by numerical integration. We use an optimization routine to find the (a_2, b_2) that maximizes the likelihood (13). At this point, we use these estimates as if they are the known values of the hyperparameters, and we make no more use of the election results π_j.

This model we have set up is a best possible model in the sense of using actual election results, but we emphasize that we are only using these to estimate the hyperparameters (a_2, b_2), *not* the individual R_j's.

3. Hyperparameters (a_2, b_2) fixed at other values. We consider $(50, 50)$ (mean 0.5, s.d. 0.07), $(20, 20)$ (mean 0.5, s.d. 0.11), and $(1, 1)$ (uniform on [0,1]). These prior distributions are more and more diffuse, but as we shall see, they do *not* give more and more diffuse inference for our parameters of interest, π_j.

5.2.3 Models for R^W

We do not complicate our analysis by estimating the parameters R_j^W simultaneously with π and R; rather, we estimate the R_j^W's first and then treat them as fixed in the subsequent analysis. We consider three different estimates of R_j^W:

1. Setting $R_j^W = 0.5$ for all j; that is, ignoring the weights. This has the effect of lumping all the relative nonresponse into the parameters R_j. Ignoring the weights would not be recommended for a serious analysis if weights are present, but we include this option because of its simplicity.

2. For each j, setting R_j^W to R_j^w (see equation (11)). This is the most direct way of including the weights in the data and should perform well for large states, for which the surveys have large sample sizes (see Table 2). For small states, however, the variation in the raw estimates R_j^w may be mostly sampling noise.

3. Smoothing the R_j^w's toward their common mean using a hierarchical model. This approach is based on the assumption that the true values of R_j^W probably do not vary much, at least

compared to the sampling variability of the R_j^w's in the smaller states. For simplicity, we deal with the boundedness of R^W by fitting a hierarchical normal model to the parameters $L_j = \text{logit}(R_j^W)$ based on the data $l_j = \text{logit}(R_j^w)$. Our model is

$$\begin{aligned} l_j &\stackrel{\text{ind}}{\sim} \text{N}(L_j, V_j) \\ L_j &\stackrel{\text{ind}}{\sim} \text{N}(\mu, \tau^2) \\ p(\mu, \tau) &\propto 1 \end{aligned} \tag{14}$$

We set the sampling variances V_j to fixed values as follows. First, for each j, we create a crude estimate of the sampling variance of l_j from elementary sample survey theory, based on the assumption of independent sampling of the respondents in state j:

$$\begin{aligned} l_j &= \text{logit}(R_j^w) = \log(\bar{w}_{1\,j}) - \log(\bar{w}_{0\,j}) \\ \text{var}(l_j) &= \text{var}(\log(\bar{w}_{1\,j})) + \text{var}(\log(\bar{w}_{0\,j})) \\ \widehat{\text{var}}(l_j) &= \frac{1}{y_{1j}} \frac{s^2_{w1\,j}}{\bar{w}_{1\,j}} + \frac{1}{y_{0j}} \frac{s^2_{w0\,j}}{\bar{w}_{0\,j}}, \end{aligned} \tag{15}$$

where y_{1j} and y_{0j} are the number of $y = 1$ and $y = 0$ respondents, respectively, and $s^2_{w1\,j}$ and $s^2_{w0\,j}$ are the sample variances of the weights for the $y = 1$ and $y = 0$ respondents, respectively, in state j. A plot of $\widehat{\text{var}}(l_j)$ versus the sample size $1/m_j$ (not shown here) shows approximate proportionality, as one would expect from simple theory. For each j, we set V_j to V/m_j, where V is the average value of $m_j\widehat{\text{var}}(l_j)$. We use V/m_j in the hierarchical analysis because $\widehat{\text{var}}(l_j)$ is extremely variable for small states. Given the V_j's, we estimate the parameters of the hierarchical model (14) Bayesianly, averaging over the hyperparameters μ and τ (see Rubin, 1981, and Gelman et al., 1995, chap. 5). We obtain the posterior medians of the parameters L_j using simulation and use the inverse-logits of these values as the fixed values of R_j^W in the subsequent analysis. The raw values R_j^w and the smoothed estimates $\widehat{R}_j^W$ for the pre-election polls appear as the last two columns of Table 2.

5.3 Results and assessing model fit

5.3.1 Estimates of hyperparameters

Table 3 gives the estimates of (a_1, b_1) and (a_2, b_2) corresponding to the various models. To explain this table, we shall first discuss the distributions of the R_j's (that is, the values of a_2 and b_2), then the estimated distributions of the π_j's (that is, the values of a_1 and b_1).

We consider several possibilities for the distribution of R_j's, ranging from fixed at 0.5 (the standard missing-at-random model, corresponding to $(a_2, b_2) = (\infty, \infty)$) to uniform on $[0, 1]$ (the "noninformative" prior distribution, corresponding to $(a_2, b_2) = (1, 1)$ that gives unappealing results, as discussed in Section 3.2). When we estimate (a_2, b_2) by comparing to the election data, we estimate the R_j's to have a mean of 0.534 and standard deviation of 0.044, which suggest that Bush supporters have a higher response rate than Dukakis supporters, and that differential nonresponse rates vary little from state to state. The latter observation explains why the Beta (1,1) model will not perform well. After correcting for the sampling weights (using either the raw or smoothed estimates), we estimate the R_j's to have a mean near 0.515; thus, the differential nonresponse is partially but not wholly explained by the sampling weights.

One surprising result is that including the weights in the analysis does not make our estimated distribution of R_j's less variable, as we might have expected.

For each of the models for R^W and R, we estimate the distribution of the π_j's from the survey data alone. The most consistent pattern here is that (a_1, b_1) become larger (that is, the π_j's are estimated to be *less* variable) as the R_j^W's become *more* variable (going from fixed at 0.5 to hierarchical estimates to raw estimates) and as the R_j's become *more* variable (going from fixed at 0.5 to the Beta (50,50) range to Beta (1,1)). This occurs because, with the hierarchical model, the variance of the π_j's is essentially being estimated from the variance of the $\overline{y}_j$'s, after subtracting (1) binomial sampling variability, (2) variability in the R_j's, and (3) variability in the R_j^W's. When one source of variability is raised, the others are estimated to be lower. At the most extreme case, when the R_j's are assigned a uniform prior distribution, there is not enough variance in the $\overline{y}_j$'s to explain this, and the π_j's are estimated to be all equal, that is, $(a_1, b_1) = (\infty, \infty)$. The other notable behavior of the estimated distribution of the π_j's is that the mean shifts after correcting for differential nonresponse, from about 0.568 with no correction, to about 0.560 after correcting for the R^W's, to about 0.545 after correcting for the R_j's.

5.3.2 Estimates of state results π_j

The test of the method is how well it estimates the individual state means, which in this case we can compare to the actual election results π_j^{actual} (under the assumption, reasonable in this case, that there is little opinion change in the last week of the election campaign). We are interested in prediction error of point estimates and also in coverage probability of posterior intervals.

To avoid an overwhelming display of results, we present inferences for a selection of the models that illustrate the behavior of the method under various assumptions:

1. Nonhierarchical model, no nonresponse adjustment: $(a_1, b_1) = (1, 1)$, $R_j = 0.5$ for all j, $R_j^W = 0.5$ for all j
2. Nonhierarchical model, nonresponse adjustment with diffuse prior distribution: $(a_1, b_1) = (1, 1)$, $(a_2, b_2) = (20, 20)$, $R_j^W = 0.5$ for all j

3. Nonhierarchical model, empirical nonresponse adjustment: $(a_1, b_1) = (1, 1)$, $(a_2, b_2) = (68.4, 62.2)$ (see Table 3), $R_j^W = 0.5$ for all j

4. Hierarchical model, no nonresponse adjustment: (a_1, b_1) estimated from polls, $R_j = 0.5$ for all j, $R_j^W = 0.5$ for all j

5. Hierarchical model, nonresponse adjustment with diffuse prior distribution: (a_1, b_1) estimated from polls, $(a_2, b_2) = (20, 20)$, $R_j^W = 0.5$ for all j

6. Hierarchical model, empirical nonresponse adjustment: (a_1, b_1) estimated from polls, $(a_2, b_2) = (68.4, 62.2)$, $R_j^W = 0.5$ for all j

7. Hierarchical model, empirical nonresponse adjustment, adjustment for raw weights: (a_1, b_1) estimated from polls, $(a_2, b_2) = (51.2, 48.2)$ (see Table 3), $R_j^W = R_j^w$ for all j

8. Hierarchical model, empirical nonresponse adjustment, adjustment for smoothed weights: (a_1, b_1) estimated from polls, $(a_2, b_2) = (65.6, 61.9)$ (see Table 3), $R_j^W = \widehat{R}_j^w$ for all j (see Section 5.2.3)

Table 4 presents, for each of the above models and for each of the 48 states j, the posterior median estimate of π_j and the (one-sided) p-value of the actual election result π_j^{actual} (that is, $\Pr(\pi_j \leq \pi_j^{\text{actual}} | \text{data, model})$). A p-value near 0 or 1 means that the actual election result was on the low or high end, respectively, of the posterior distribution for that state.

The summary statistics at the bottom of the table reveal that a large reduction in error comes simply from using a hierarchical model for the π_j's: models 1–3 have mean errors of about 0.05, whereas model 4 (estimating (a_1, b_1) from the data but making no correction for weights or differential nonresponse) has a mean error of about 0.4. Using appropriate values for (a_2, b_2) reduces the mean error to about 0.03, with the corrections for weights having little effect. These are the best possible results, in the sense that (a_2, b_2) are estimated using the election results themselves. However, setting (a_2, b_2) to reasonable approximate values such as $(50, 50)$ gives results of nearly the same accuracy.

Figure 3 displays the calibration of the error estimates with, for each model, a stem-and-leaf plot of the 48 p-values for the state forecasts. If the p-values for a model are clustered near 0 or 1, the forecasts are overconfident (that is, the standard errors are too small); if they are clustered near the middle of the range, the forecasts are underconfident and the standard errors too large. This behavior is summarized in the bottom rows of Table 4 by the scaled mean z-score and sum of squares of the normal-transformed p-values. The sum of squares can itself be compared to a χ^2 distribution, and this reveals that the models with (a_2, b_2) set to (∞, ∞) (that is, $R_j \equiv 0.5$

for all j) yield overconfident forecasts, whereas the models with (a_2, b_2) fit from actual election results yield underconfident forecasts. Both these results make sense: a model that assumes there is no differential nonresponse is ignoring a source of variability and thus should be expected to have standard errors that are too small, whereas a model that is fit using the best parameter values should have lower-than-expected errors.

To reveal what the models are doing, we plot in Figure 4 the actual vs. predicted results, by state, for each model. Models 1–3, which do not fit a hierarchical model to the π_j's, perform poorly, which is to be expected: the estimates are extremely variable, due to sampling variation, and no attempt is made to correct them. Model 4 achieves a great improvement by shrinking the estimates of π_j towards the grand mean. However, the model does not seem to shrink enough: for the states with low estimates, the actual result tends to be higher, and vice-versa.

As discussed earlier, the π_j's are shrunk more in the hierarchical model if the R_j's are allowed to vary. Model 5, which sets (a_2, b_2) to $(20, 20)$, shrinks the π_j's too much, as is shown by the fifth graph in Figure 4, because Beta (20,20) is too spread-out a distribution for the R_j's. Model 6, with a Beta (68.4,62.2) distribution for the R_j's, shrinks the π_j's about the right amount and thus improves the estimates for the 48 states. Models 7 and 8, which correct in different ways for the weights, perform similarly.

Figure 5 plots prediction error vs. sample size, by state, showing the expected inverse relation for all the models.

6 Discussion

We see the methods described in this article as a tool for survey analysis, to be used in addition to models for nonresponse based on observed covariates (Little, 1993). Fixing R at 0.5 is in some sense a default choice, corresponding to the ignorable model. In fact, in our election example, we find that the R_j's are quite close to 0.5. Allowing uncertainty in R should improve the calibration of error estimates, but it is important to keep that uncertainty realistic; allowing R to vary too much from 0.5 can lead to unreasonable inferences for the parameter of interest, π (see Section 3.2 and the performance of Model 5 in Figure 4).

In a hierarchical context, as achieved in our example by partitioning the population by state, accounting for variation in R_j's increases the accuracy of estimates of π_j's by allowing the model to shrink appropriately. The difficulty here is getting the population distribution for the R_j's for new problems.

In our example, correcting for sampling weights had little effect. In general, we prefer to correct for the weights provided with the survey so that our analysis can be viewed as an improvement upon, rather than an alternative to, the usual approach of weighted means. When the weighted and unweighted analysis give similar answers, we prefer the weighted method for its generality.

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.

Brady. H., and Orren, G. (1992). Polling pitfalls: sources of error in public opinion surveys. In *Media Polls in American Politics*, ed. T. Mann and G. Orren, 55–94. Washington, D.C.: Brookings Institution.

Cochran, W. G. (1977). *Sampling Techniques*, third edition. New York: Wiley.

Ericson, W. A. (1969). Subjective Bayesian models in sampling finite populations, I. *Journal of the Royal Statistical Society B* **31**, 195–234.

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

Gelman, A., and Little, T. C. (1997). Poststratification into many categories using hierarchical logistic regression. *Survey Methodology*, to appear.

Kadane, J. B. (1993). Subjective Bayesian analysis for surveys with missing data. *The Statistician* **42**, 415–426.

Kaufman, G. M., and King, B. (1973). A Bayesian analysis of nonresponse in dichotomous processes. *Journal of the American Statistical Association* **68**, 670–678.

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

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

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.

Manski, C. F. (1995). *Identification Problems in the Social Sciences*. Cambridge, Mass.: Harvard University Press.

Nadaram, B., and Sedransk, J. (1993). Bayesian predictive inference for a finite population proportion: two-stage cluster sampling. *Journal of the Royal Statistical Society B* **55**, 399–408.

Nordheim, E. (1984). Inference from nonrandomly missing categorical data: an example from a genetic study on Turner's syndrome. *Journal of the American Statistical Association* **79**, 772–780.

Oh, H., and Scheuren, F. (1983). Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*, vol. 2, ed. W. G. Madow, I. Olkin, and D. B. Rubin, 143–184. New York: Academic Press.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **3**, 581–592.

Rubin, D. B. (1977). Formalizing subjective notions about the effect of nonrespondents in sample surveys. *Journal of the American Statistical Association* **77**, 538–543.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6**, 377–401.

Rubin, D. B. (1987a). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Scott, A., and Smith, T. M. F. (1969). Estimation in multi-stage surveys. *Journal of the American Statistical Association* **64**, 830–840.

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

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.

n	50	100	200	400	1000	10000	∞
m	35	70	140	280	700	7000	∞
$\underline{\pi}_L$ (exact)	0.082	0.095	0.106	0.114	0.123	0.134	0.14
$\underline{\pi}_L$ (approximate)	0.059	0.083	0.100	0.111	0.122	0.134	0.14
$\overline{\pi}_U$ (exact)	0.546	0.517	0.496	0.480	0.465	0.448	0.44
$\overline{\pi}_U$ (approximate)	0.555	0.522	0.498	0.481	0.466	0.448	0.44

Table 1: A numerical example of conservative upper and lower confidence bounds for π at level $\alpha = 0.05$. For each n given it is assumed that the observed mean is $\bar{y} = 0.2$, and the observed response rate is $p = 0.7$. Exact bounds come from inverting the binomial distributions, approximated from the normal distribution.

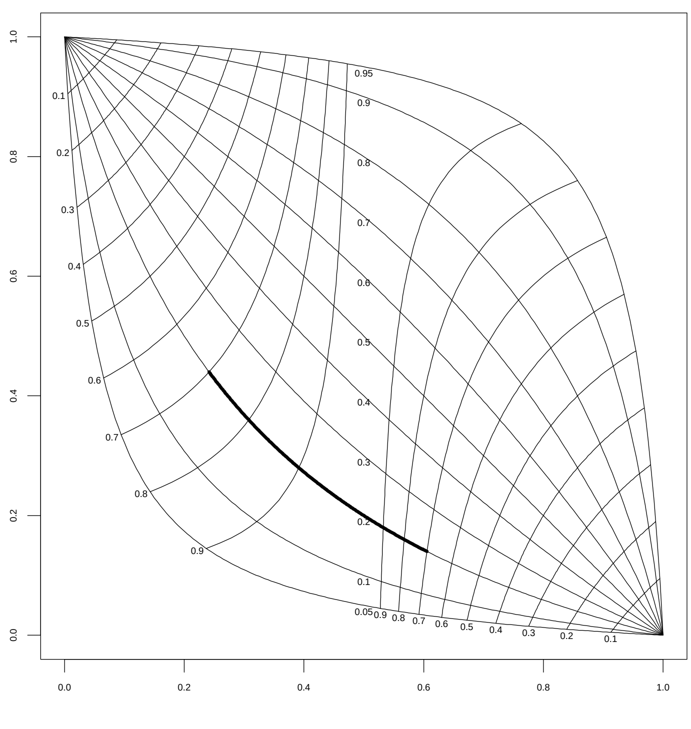

Figure 1: π as a function of $(R, \zeta_p, \zeta_{\bar{y}})$ for select values of $(\zeta_p, \zeta_{\bar{y}})$. (Note that ζ_p is estimated by $p = m/n$, and $\zeta_{\bar{y}}$ is estimated by $\bar{y}$.) Each curve that runs from upper-left to lower-right corresponds to a particular value of $\zeta_{\bar{y}}$ and gives π as a function of R conditional on $\zeta_{\bar{y}}$. Corresponding to each value of ζ_p is a set of two curves that run lower-left to upper-right. The two intersections of these curves with the curve for $\zeta_{\bar{y}}$ give the upper and lower bounds for R and π conditional on $\zeta_{\bar{y}}$ and ζ_p. For example, suppose $p = 0.7$, $\bar{y} = 0.2$, and n is large. Then $\zeta_p \approx 0.7$ and $\zeta_{\bar{y}} \approx 0.2$, and (π, R) must lie approximately on the curve segment labeled "0.2", between the two curves labeled "0.7"—this is shown in **bold** on the graph.

State, j	Election result	Unweighted mean, $\bar{y}_j$	Weighted mean, $\sum_i w_{ij} y_{ij} / \sum_i w_{ij}$	Sample size, m_j	R_j^w	$\widehat{R}_j^W$
AL	0.60	0.72	0.68	134	0.55	0.52
AR	0.57	0.57	0.51	86	0.56	0.52
AZ	0.61	0.62	0.61	141	0.51	0.51
CA	0.52	0.57	0.55	1088	0.52	0.52
CO	0.54	0.59	0.62	127	0.46	0.50
CT	0.53	0.53	0.55	103	0.48	0.50
DE	0.56	0.40	0.39	30	0.51	0.51
FL	0.61	0.63	0.62	565	0.51	0.51
GA	0.60	0.62	0.59	211	0.54	0.52
IA	0.45	0.38	0.30	102	0.58	0.53
ID	0.63	0.55	0.61	33	0.43	0.50
IL	0.51	0.54	0.52	439	0.53	0.52
IN	0.60	0.75	0.73	215	0.53	0.52
KS	0.57	0.72	0.67	105	0.56	0.52
KY	0.56	0.57	0.62	148	0.45	0.49
LA	0.55	0.62	0.57	153	0.55	0.52
MA	0.46	0.47	0.44	279	0.53	0.52
MD	0.52	0.52	0.50	207	0.52	0.52
ME	0.56	0.52	0.54	44	0.48	0.51
MI	0.54	0.57	0.56	403	0.51	0.51
MN	0.46	0.53	0.49	214	0.54	0.52
MO	0.52	0.46	0.43	235	0.52	0.51
MS	0.60	0.69	0.62	176	0.57	0.53
MT	0.53	0.39	0.45	31	0.44	0.50
NC	0.58	0.59	0.61	239	0.48	0.50
ND	0.57	0.56	0.57	54	0.48	0.51
NE	0.60	0.56	0.59	92	0.48	0.50
NH	0.63	0.70	0.68	20	0.52	0.51
NJ	0.57	0.56	0.55	306	0.50	0.51
NM	0.52	0.54	0.54	89	0.50	0.51
NV	0.61	0.62	0.62	21	0.50	0.51
NY	0.48	0.42	0.42	666	0.50	0.50
OH	0.56	0.62	0.64	459	0.47	0.49
OK	0.58	0.57	0.58	93	0.48	0.50
OR	0.48	0.50	0.46	113	0.55	0.52
PA	0.51	0.54	0.53	437	0.50	0.51
RI	0.44	0.27	0.27	67	0.50	0.51
SC	0.62	0.70	0.69	154	0.51	0.51
SD	0.53	0.54	0.55	52	0.48	0.51
TN	0.58	0.68	0.68	259	0.50	0.50
TX	0.56	0.58	0.57	601	0.52	0.51
UT	0.67	0.80	0.84	61	0.43	0.50
VA	0.60	0.69	0.72	257	0.46	0.49
VT	0.52	0.58	0.71	12	0.37	0.50
WA	0.49	0.47	0.44	274	0.54	0.52
WI	0.48	0.49	0.52	265	0.47	0.49
WV	0.48	0.49	0.50	80	0.48	0.50
WY	0.61	0.54	0.56	13	0.48	0.51

Table 2: By state: election results (proportion of the two-party vote in 1988 received by Bush); survey data (weighted mean, unweighted mean, and sample size) from the combined surveys; and estimated weighting adjustment (raw estimate and Bayes-shrunk estimate).

Model for R_j^W's	Model for R_j's	Distribution of R_j's (a_2, b_2)	mean (s.d.)	Distribution of π_j's (a_1, b_1)	mean (s.d.)
fixed at 0.5	fixed at 0.5	-	-	$(17.8, 13.6)$	0.567 (0.087)
fixed at 0.5	Beta (50,50)	-	-	$(25.4, 19.3)$	0.568 (0.073)
fixed at 0.5	Beta (20,20)	-	-	$(75.4, 57.1)$	0.569 (0.043)
fixed at 0.5	Beta (1,1)	-	-	(∞, ∞)	..... (.....)
fixed at 0.5	estimated	$(68.4, 62.2)$	0.524 (0.044)	$(21.9, 18.3)$	0.545 (0.078)
raw estimates	fixed at 0.5	-	-	$(15.2, 11.9)$	0.561 (0.094)
raw estimates	Beta (50,50)	-	-	$(20.4, 15.9)$	0.562 (0.081)
raw estimates	Beta (20,20)	-	-	$(44.3, 34.4)$	0.563 (0.056)
raw estimates	Beta (1,1)	-	-	(∞, ∞)	..... (.....)
raw estimates	estimated	$(51.2, 48.2)$	0.515 (0.050)	$(19.7, 16.3)$	0.547 (0.082)
hierarchical	fixed at 0.5	-	-	$(17.1, 13.5)$	0.559 (0.088)
hierarchical	Beta (50,50)	-	-	$(24.1, 19.0)$	0.559 (0.075)
hierarchical	Beta (20,20)	-	-	$(67.4, 52.9)$	0.560 (0.045)
hierarchical	Beta (1,1)	-	-	(∞, ∞)	..... (.....)
hierarchical	estimated	$(65.6, 61.9)$	0.514 (0.044)	$(21.5, 17.9)$	0.545 (0.078)

Table 3: Estimated values of hyperparameters under different models. Estimated distributions of R_j's are based on polls and election results; estimated distributions of π_j's are based on polls only, conditional on the distribution of R_j's.

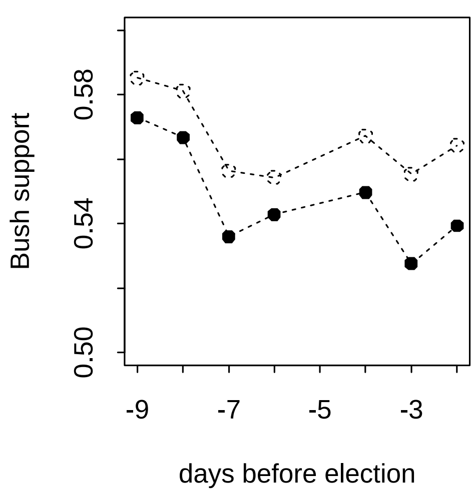

Figure 2: Support for Bush for President in 1988 in a series of CBS pre-election polls: raw means (open circles) and weighted means (solid circles). The actual election outcome was 53.9% of the two-party vote for Bush. The unweighted means are higher than the weighted, indicating that, according to the weights, supporters of Bush were more likely to respond than supporters of Dukakis.

Model:	1		2		3		4		5		6		7		8	
(a_1, b_1):	(1,1)		(1,1)		(1,1)		estimated		estimated		estimated		estimated		estimated	
(a_2, b_2):	(∞, ∞)		(20,20)		(68.4,62.2)		(∞, ∞)		(20,20)		(68.4,62.2)		(51.2,48.2)		(65.6,61.9)	
R_j^W:	0.5		0.5		0.5		0.5		0.5		0.5		raw		smoothed	
State	$\hat{\pi}$	P	$\hat{\pi}$	P	$\hat{\pi}$	P	$\hat{\pi}$	P	$\hat{\pi}$	P	$\hat{\pi}$	P	$\hat{\pi}$	P	$\hat{\pi}$	P
AL	0.71	0.00	0.71	0.09	0.69	0.05	0.69	0.01	0.60	0.49	0.64	0.18	0.62	0.34	0.64	0.21
AR	0.57	0.52	0.56	0.53	0.55	0.65	0.57	0.52	0.57	0.52	0.55	0.69	0.52	0.84	0.54	0.72
AZ	0.62	0.35	0.62	0.45	0.60	0.56	0.61	0.43	0.58	0.76	0.58	0.71	0.58	0.72	0.58	0.71
CA	0.57	0.00	0.57	0.26	0.55	0.27	0.57	0.00	0.57	0.09	0.55	0.24	0.54	0.31	0.54	0.28
CO	0.59	0.13	0.58	0.30	0.56	0.33	0.59	0.12	0.57	0.19	0.56	0.34	0.59	0.18	0.57	0.29
CT	0.53	0.44	0.53	0.48	0.51	0.60	0.54	0.35	0.56	0.17	0.52	0.51	0.54	0.38	0.53	0.48
DE	0.40	0.96	0.41	0.91	0.38	0.97	0.48	0.88	0.55	0.61	0.48	0.90	0.48	0.90	0.48	0.90
FL	0.63	0.22	0.62	0.45	0.60	0.57	0.62	0.26	0.58	0.78	0.59	0.71	0.59	0.71	0.59	0.71
GA	0.62	0.30	0.61	0.44	0.60	0.55	0.61	0.35	0.58	0.71	0.58	0.68	0.56	0.78	0.57	0.72
IA	0.38	0.91	0.39	0.74	0.36	0.91	0.43	0.70	0.54	0.01	0.43	0.61	0.39	0.85	0.42	0.68
ID	0.54	0.86	0.54	0.79	0.52	0.88	0.56	0.89	0.57	0.95	0.54	0.95	0.57	0.85	0.54	0.94
IL	0.54	0.08	0.54	0.35	0.52	0.42	0.55	0.06	0.56	0.08	0.53	0.34	0.52	0.45	0.52	0.40
IN	0.75	0.00	0.74	0.03	0.73	0.00	0.73	0.00	0.61	0.41	0.68	0.04	0.66	0.10	0.67	0.06
KS	0.72	0.00	0.71	0.04	0.70	0.02	0.69	0.00	0.60	0.21	0.64	0.07	0.61	0.19	0.64	0.09
KY	0.57	0.36	0.57	0.45	0.55	0.56	0.57	0.34	0.57	0.38	0.55	0.58	0.58	0.31	0.56	0.49
LA	0.62	0.04	0.61	0.23	0.60	0.22	0.61	0.05	0.58	0.24	0.58	0.28	0.55	0.50	0.57	0.34
MA	0.47	0.42	0.47	0.47	0.44	0.62	0.48	0.29	0.55	0.01	0.47	0.38	0.46	0.49	0.47	0.43
MD	0.52	0.42	0.52	0.48	0.50	0.62	0.53	0.34	0.56	0.12	0.51	0.51	0.50	0.58	0.51	0.54
ME	0.52	0.68	0.52	0.63	0.50	0.75	0.54	0.61	0.56	0.45	0.52	0.71	0.54	0.63	0.53	0.70
MI	0.57	0.11	0.57	0.36	0.55	0.45	0.57	0.10	0.57	0.21	0.55	0.44	0.54	0.47	0.55	0.45
MN	0.53	0.03	0.53	0.23	0.50	0.23	0.53	0.02	0.56	0.01	0.52	0.12	0.50	0.24	0.51	0.16
MO	0.46	0.98	0.46	0.77	0.43	0.95	0.47	0.95	0.55	0.25	0.47	0.88	0.46	0.88	0.46	0.89
MS	0.69	0.01	0.68	0.18	0.66	0.14	0.67	0.03	0.59	0.61	0.62	0.33	0.59	0.64	0.61	0.44
MT	0.39	0.94	0.40	0.88	0.37	0.95	0.48	0.79	0.55	0.33	0.47	0.81	0.50	0.66	0.48	0.80
NC	0.59	0.40	0.58	0.48	0.56	0.62	0.59	0.42	0.57	0.58	0.56	0.68	0.58	0.50	0.57	0.61
ND	0.56	0.57	0.55	0.56	0.53	0.67	0.56	0.54	0.57	0.48	0.54	0.68	0.55	0.60	0.54	0.67
NE	0.56	0.79	0.56	0.68	0.54	0.84	0.57	0.81	0.57	0.83	0.54	0.89	0.56	0.78	0.55	0.87
NH	0.69	0.30	0.68	0.34	0.67	0.38	0.62	0.57	0.58	0.89	0.59	0.76	0.59	0.74	0.59	0.75
NJ	0.56	0.65	0.56	0.57	0.53	0.74	0.56	0.64	0.57	0.52	0.54	0.76	0.54	0.71	0.54	0.74
NM	0.54	0.40	0.54	0.45	0.52	0.56	0.55	0.32	0.56	0.16	0.53	0.48	0.54	0.43	0.53	0.47
NV	0.61	0.49	0.61	0.50	0.59	0.57	0.59	0.61	0.57	0.79	0.56	0.76	0.57	0.72	0.56	0.75
NY	0.42	1.00	0.42	0.75	0.40	0.95	0.43	1.00	0.54	0.07	0.44	0.85	0.45	0.76	0.44	0.83
OH	0.62	0.00	0.61	0.24	0.59	0.22	0.61	0.00	0.58	0.25	0.58	0.27	0.60	0.13	0.60	0.16
OK	0.57	0.62	0.56	0.58	0.55	0.72	0.57	0.62	0.57	0.64	0.55	0.77	0.56	0.66	0.55	0.75
OR	0.50	0.27	0.50	0.38	0.48	0.47	0.52	0.16	0.56	0.02	0.51	0.28	0.48	0.45	0.50	0.31
PA	0.54	0.16	0.53	0.39	0.51	0.50	0.54	0.13	0.56	0.09	0.52	0.41	0.53	0.37	0.52	0.39
RI	0.27	1.00	0.28	0.96	0.26	1.00	0.36	0.94	0.52	0.02	0.38	0.85	0.39	0.81	0.39	0.84
SC	0.70	0.02	0.69	0.18	0.68	0.15	0.68	0.05	0.60	0.73	0.63	0.39	0.63	0.41	0.63	0.39
SD	0.54	0.47	0.54	0.49	0.51	0.58	0.55	0.37	0.56	0.20	0.53	0.51	0.54	0.43	0.53	0.50
TN	0.68	0.00	0.67	0.13	0.66	0.07	0.67	0.00	0.59	0.38	0.62	0.17	0.63	0.16	0.63	0.15
TX	0.58	0.20	0.58	0.44	0.56	0.56	0.58	0.20	0.57	0.41	0.55	0.59	0.55	0.62	0.55	0.61
UT	0.80	0.02	0.79	0.08	0.78	0.06	0.72	0.14	0.61	0.96	0.67	0.53	0.71	0.26	0.68	0.48
VA	0.69	0.00	0.68	0.16	0.66	0.11	0.68	0.01	0.60	0.58	0.63	0.27	0.66	0.13	0.64	0.18
VT	0.57	0.34	0.57	0.36	0.55	0.40	0.58	0.23	0.57	0.10	0.55	0.33	0.58	0.18	0.55	0.32
WA	0.48	0.72	0.48	0.58	0.45	0.77	0.48	0.61	0.55	0.07	0.48	0.61	0.46	0.71	0.47	0.67
WI	0.49	0.39	0.49	0.46	0.47	0.61	0.50	0.28	0.55	0.03	0.49	0.42	0.52	0.22	0.50	0.31
WV	0.49	0.42	0.49	0.45	0.46	0.57	0.51	0.24	0.56	0.02	0.50	0.33	0.51	0.25	0.50	0.31
WY	0.53	0.73	0.53	0.71	0.51	0.76	0.56	0.77	0.57	0.87	0.54	0.86	0.55	0.83	0.54	0.86
mean error	0.052		0.049		0.051		0.041		0.037		0.031		0.032		0.031	
mean Z	-3.839		-1.119		0.450		-4.288		-3.210		0.599		0.347		0.565	
χ^2	138.051		26.326		50.412		112.263		65.494		26.688		22.821		25.663	
P-value	0.000		0.995		0.378		0.000		0.047		0.995		0.999		0.997	

Table 4: Posterior medians and p-values ($\Pr(\pi_j \leq \pi^{\text{actual}} | \text{data,model})$) for each state for each of 8 models. At the bottom of the table are a measure of fit (mean absolute error of state estimates) and three measures of calibration (scaled mean Z-score of state p-values, sum of squares of the Z-scores, and p-value of the sum of squares compared to the χ^2_{48} distribution).

```
0 :
1 : 0336889
2 : 1456
3 : 11478
4 : 1335579
5 : 0089
6 : 13466
7 : 111224688
8 : 134558
9 : 0
```

```
0 : 111222377899
1 : 02679
2 : 011355
3 : 388
4 : 11589
5 : 2288
6 : 114
7 : 13689
8 : 379
9 : 56
```

```
0 : 02567
1 : 145
2 : 2237
3 : 38
4 : 0257
5 : 0566667778
6 : 0122257
7 : 24567
8 : 38
9 : 15557x
```

```
0 : 0000000112348
1 : 136
2 : 0279
3 : 04559
4 : 00222479
5 : 27
6 : 258
7 : 239
8 : 6
9 : 1468xx
```

```
0 : 69
1 : 5668
2 : 189
3 : 1112499
4 : 03457889
5 : 04
6 : 11778
7 : 011224555
8 : 034679
9 : 04
```

```
0 : 47
1 : 278
2 : 47778
3 : 3334489
4 : 1248
5 : 111389
6 : 118889
7 : 1116667
8 : 155689
9 : 05
```

```
0 : 000001123556
1 : 02346
2 : 034689
3 : 144557
4 : 23
5 : 247
6 : 11124
7 : 079
8 : 189
9 : 45x
```

```
0 : 3489
1 : 3688
2 : 3346
3 : 0456689
4 : 4455555677889
5 : 036788
6 : 38
7 : 14579
8 : 8
9 : 16
```

Figure 3: Stem-and-leaf plots of p-values of actual election outcome by state, compared to posterior distributions of π_j's under each of eight models. Each stem-and-leaf plot has 48 digits, one for each state. Ideally, the plots should follow Uniform(0,1) distributions. Bunching up of p-values near either or both ends would indicate posterior distributions that are too precise, whereas bunching up of p-values near the middle would indicate posterior distributions that are too conservative.

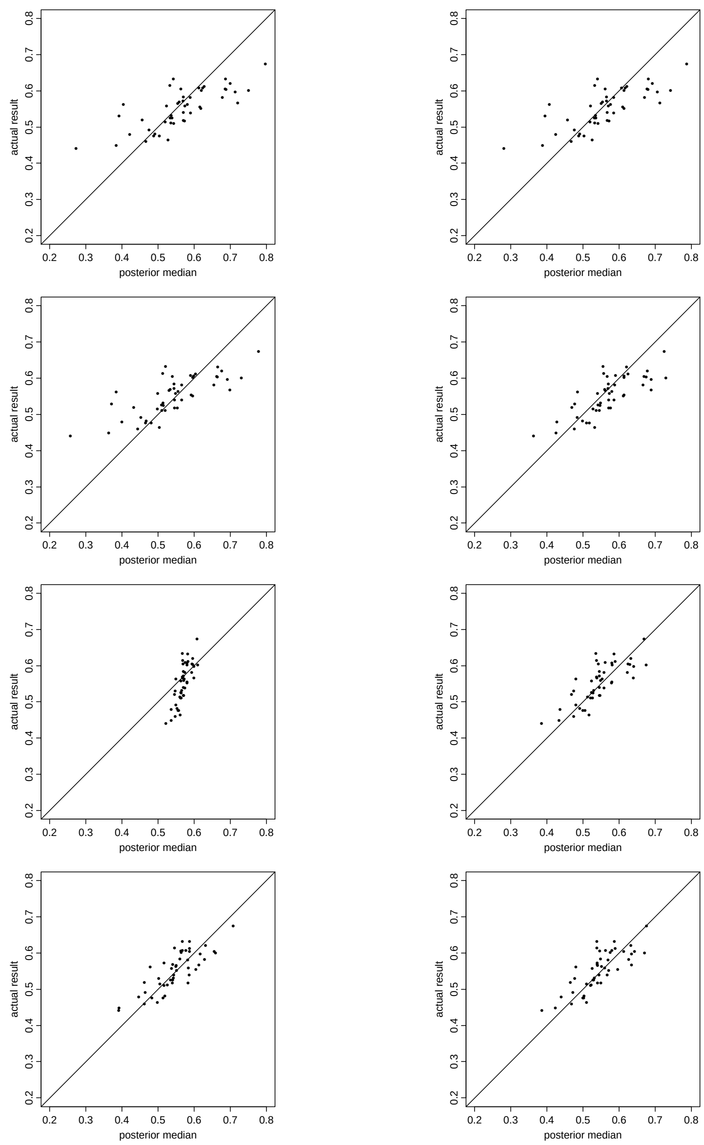

Figure 4: Plots of actual election results by state, π_j^{actual}, vs. posterior medians of π_j, for each of eight models. Diagonal lines indicate perfect estimates.

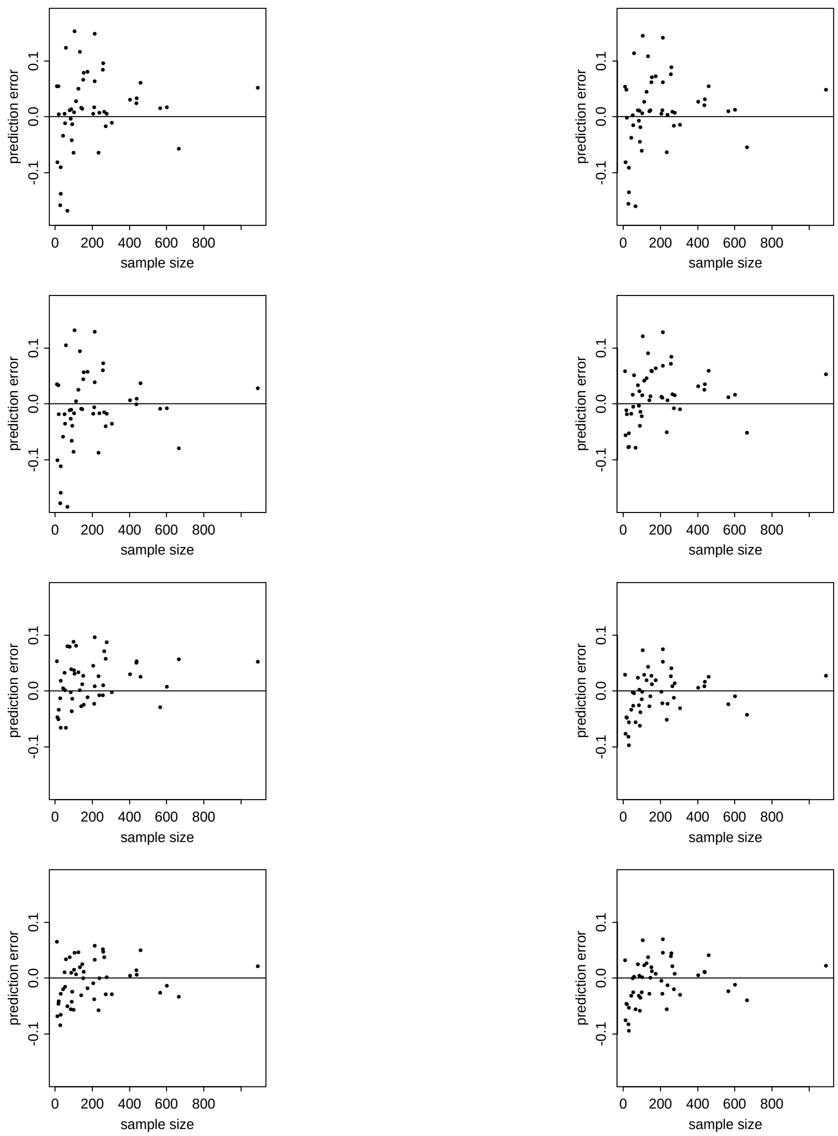

Figure 5: Prediction error for state j vs. sample size n_j, for each of the eight models. Horizontal line at zero indicates perfect estimates. Prediction errors for each model are actual vote minus the posterior median.