

Bayesian Multilevel Estimation with
Poststratification:
State-Level Estimates from National Polls¹

David K. Park ² Andrew Gelman ³ Joseph Bafumi ⁴

¹We thank the reviewers for their helpful comments and the National Science Foundation for their financial support.

²Assistant Professor, Political Science and Applied Statistics, Washington University, dpark@artsci.wustl.edu

³Professor, Statistics and Political Science, Columbia University

⁴Doctoral Candidate, Political Science, Columbia University

Abstract

We fit a multilevel logistic regression model for the mean of a binary response variable conditional on poststratification cells. This approach combines the modeling approach often used in small-area estimation with the population information used in poststratification (see Gelman and Little (1997)). To validate the method, we apply it to U.S pre-election polls for 1988 and 1992, poststratified by state, region, and the usual demographic variables. We evaluate the model by comparing it to state-level election outcomes. The multilevel model outperforms more commonly used models in political science. We envision the most important usage of this method, not in forecasting elections, but in estimating public opinion on a variety of issues at the state level.

1 Introduction

Hundreds of national opinion polls are conducted every year. These polls may, among other things, inform election prospects or reveal issue stances among the public. They are generally based on national random digit dialing with corrections for nonresponse based on demographic factors such as sex, ethnicity, age, and education. Often, it is desirable to estimate opinions at lower levels such as individual states or congressional districts. Certainly, scholars and politicians interested in public opinion (and political responses to opinion) would find such data enormously useful and interesting. While some state and district level polling is conducted, they tend to be few and exhibit high error variance due to small sample sizes. Further, their nonuniformity hinders comparisons between states.

One of the first projects to use national opinion polls to estimate opinions at the level of states was undertaken by Pool et al. (1965). They used national polls, voting and census data to construct 480 synthetic voter types based on a variety of socio-demographic factors.¹ They determined the percent of each type in each state and estimated state-level results. In the early 1970s, Weber et al. (1972-73) expanded the number of voter types from 480 to 960 categories and were able to estimate state-level opinions on a variety of issues.

We extend the work of Pool et al. (1965) and Weber et al. (1972-73) by expanding the number of synthetic voter types from 960 to 3264 and, more importantly, fitting a multilevel regression model while simultaneously correcting for nonresponse. In addition to allowing for more categories, our work moves beyond these precursors in political science by performing shrinkage estimation for individual categories and sets of categories (e.g., states).

We demonstrate the method by application to a set of 1988 and 1992 na-

¹The project was also the basis for a satirical novel by Eugene Burdick titled *The 480* (Burdick 1964).

tional pre-election polls. This is a useful testing ground for our method because polls immediately before the election can be externally validated by comparing to the vote outcomes. Thus, we apply our method to pre-election polls, *not* to forecast elections, but rather to provide an improved estimate for state-level opinions. Pre-election polling is simply a convenient and easily-validated example (see Jackman and Rivers (2001) for a recent example of state-level election forecasting via dynamic Bayesian hierarchical model.)

2 Overview

We construct a multilevel logistic regression model for the mean of a binary response variable conditional on poststratification cells. This approach combines the modeling approach often used in small-area estimation with the population information used in poststratification (Gelman and Little 1997). The procedure has two steps:

1. Fit a multilevel regression model² for the individual response y given demographics and state of residence. This model thus estimates an average response for each cross-classification j of demographics and state, π_j . In our example, we have sex (male or female), ethnicity (African-American or other), age (4 categories), education (4 categories), and 50 states; thus 3200 categories. If the District of Columbia is included, we have 3264 categories. The demographic categories we use are those used by national survey organizations in their weighting (see Voss et al. (1995))

²Many statistical applications involve multiple parameters that can be regarded as related in some way by the structure of the problem, implying that a joint probability model for these parameters should reflect the dependence among them. It is natural to model such a problem within a multilevel framework, with observable outcomes modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters. See Kreft and de Leeuw (1998), Snijders and Bosker (1999), Bryk and Raudenbush (2001), Gelman et al. (2003) for introductions to multilevel models from classical and Bayesian perspectives.

2. From the U.S. Census, we get the adult population N_j for each category j . The estimated population average of the response y in any states s is then

$$\theta_s = \frac{\sum_{j \in s} N_j \pi_j}{\sum_{j \in s} N_j},$$

with each summation over the 64 demographic categories in the state.

This procedure uses a large number of categories because (a) we are interested in separating out the response by state, and (b) nonresponse adjustments force us to include the demographics. As a result, any given survey will have few or no data in many categories. This is not a problem, however, if a multilevel model is fitted. Each factor or set of interactions in such a model is automatically given a variance component. This inferential procedure works well and outperforms standard survey weighted estimates when estimating state-level outcomes.

3 The Model

Multilevel Logistic Regression Model for Binary Data

We label the survey response y_i as 1 for supporters of the Republican candidate and 0 for supporters of the Democrat (with the undecideds excluded), with $Pr(y_i = 1) = \text{logit}^{-1}(X\beta_i)$. More generally, the model could apply to any yes/no survey response. The respondent-level design matrix \mathbf{X} is all 0's and 1's with indicators for the demographic variables in survey weighting: sex, ethnicity, age, education, interaction of sex and ethnicity, and age and education.³ We also include in \mathbf{X} indicators for the 50 states, plus the District of Columbia, and for the 5 regions of the country (Northeast, Midwest, South, West and D.C.

³These are the interactions used by polling organizations for survey weighting (see Voss et al. (1995)). The goal of our demographic adjustments are not to estimate demographic parameters, but to adjust for demographics during poststratification.

(considered as a separate region because of its distinctive voting patterns)). As part of a general approach for multilevel models, we give each batch of regression coefficients with greater than two groups an independent normal distribution centered at 0 and with standard deviation estimated from data. This allows us to estimate these parameters as varying effects, taking advantage of the multilevel structure of the data.

There is no gain to multilevel modeling for batches with $J < 3$ groups when prior distributions are noninformative (see e.g., Gelman et al. (2003) and Gelman (2004)), and so for simplicity we model sex and ethnicity as regression coefficients with no multilevel structure. The data model looks as follows:

$$\begin{aligned} Pr(y_i = 1) = & \text{logit}^{-1}(\beta^0 + \beta^{female} \cdot female_i + \beta^{black} \cdot black_i \\ & + \beta^{female.black} \cdot female_i \cdot black_i + \beta_{age(i)}^{age} \\ & + \beta_{edu(i)}^{edu} + \beta_{age(i),edu(i)}^{age,edu} + \beta_{state(i)}^{state}). \end{aligned} \quad (1)$$

We then set up a state-level model with region indicators and a measure of previous Republican vote share as predictors:

$$\beta_j^{state} \sim N(\beta_{region(j)}^{region} + \beta^{v.prev} \cdot v.prev_j, \sigma_{state}^2).$$

More precisely, $v.prev$ is the average Republican vote share in the three previous elections, adjusted for home-state and home-region effects in the previous election.

We assign normal distributions to the varying coefficients. These distributions have means 0 (no loss of generality given the inclusion of the constant term β^0 in the data model) and standard deviations $\sigma_{age}, \sigma_{edu}, \sigma_{age,edu}, \sigma_{region},$

estimated from data given noninformative uniform prior densities.⁴

Generalizations of the Model

There are certain generalizations that are particularly appropriate for the study of public opinion. Here we briefly outline how to extend the model in two ways: handling *leaners* and *undecided* voters, and restricting to subsets of the population such as *registered* or *likely* voters, or supporters of the two major parties. The model can be generalized in other ways, such as including effects for survey organizations by adding appropriate group-level predictors (Jackman and Rivers 2001).

First, *leaners* are traditionally included as full supporters of candidates, and voters who are undecided or express no opinion are typically treated as missing data or discarded. These data-analytic simplifications are generally reasonable in the context of U.S. politics (see, e.g., Gelman and King (1993)). In other opinion settings, however, it might be more informative to include the partial information provided by intermediate responses. Our model can be immediately generalized to an ordered probit to directly model intermediate opinions.

Second, this model can be made more complicated to account for turnout and third party candidates. For example, consider the 1992 presidential election. In that election, there was an unexpectedly large voter turnout rate that accompanied a particularly strong showing for a third party candidate. Thus, we estimate three separate models. In the first model, we label the survey response y_i as 1 for respondents registered or expected to vote and 0 for everyone else. The second model is restricted to the respondents who were registered or expected to vote; in this second model, $y_i = 1$ for all respondents who are

⁴Various noninformative prior distributions have been suggested for scale parameters in multilevel models. Gelman (2004) demonstrates that serious problems exist with the inverse-gamma family of noninformative prior distributions, and suggest the use of a uniform prior on the multilevel standard deviation parameters.

expected to vote for either the two major party candidates and 0 for everyone else. The third model is restricted to the respondents who were registered or expected to vote, and who expressed a preference for one of the two major-party candidates. Then, the proportion of registered or likely voters in any state s who support the Republican candidate, among those with a major-party preference, is

$$\theta_s^{1992} = \frac{\sum_{j \in s} N_j \pi_j^1 \pi_j^2 \pi_j^3}{\sum_{j \in s} N_j \pi_j^1 \pi_j^2}$$

where $\pi_j^1, \pi_j^2, \pi_j^3$ are the vectors of cell probabilities for each of the three models, and with each summation over the 64 demographic categories in the state.

4 Data

For each of 1988 and 1992, we use a CBS News/New York Times national poll conducted during the week before the U.S. Presidential election. In 1988, there were a total of 2,193 respondents in the sample; in 1992 there were 4,650 respondents. Following the standard convention, we place *leaners* among the full supporters. Respondents were excluded if any of the categories for sex, ethnicity, age or education were missing. Even though no data were included from Alaska and Hawaii, they are included in the model. The preferences in these states are estimated based on the demographic coefficients and previous Republican vote share.

The Census Bureau provides the joint adult population distributions of the demographic variables within each state.⁵ For example, we know from the Census that there were 66,177 adults who live in Alabama, were male, not Black, between the ages 18-29, and do not have a high school diploma.

⁵Census of Population and Housing 1990: Subject Summary Tape File (SSTF) 6, Education in the United States.

5 The Estimated Model

Fitting the Model

We fit the model using the Bayesian software WinBUGS (Spiegelhalter et al. 1999) as called from R (R Development Core Team 2003) using Gelman’s Bugs to R (Gelman 2003). We use a multiplicative and additive redundant parameterization to speed convergence. The Gibbs sampler can be slow to converge because of posterior dependence among parameters. Paradoxically, adding new parameters, thus performing the random walk in a larger space, can reduce dependence in the larger parameter set and improve the convergence of the Markov chain simulation.⁶

Approximate mixing of three parallel simulated chains was achieved after 5000 iterations.⁷ We construct summary plots of the multilevel model. With a binary outcome, we plot $Pr(y = 1) = E(y)$ as a function of the predictors. Because the model has many predictors, instead of plotting $E(y)$ as a function of each of the demographic inputs, we plot $E(y)$ as a function of the combined linear predictor:

$$\begin{aligned} u_i = & \beta^{female} \cdot female_i + \beta^{black} \cdot black_i + \beta^{female.black} \cdot female_i \cdot black_i \\ & + \beta_{age(i)}^{age} + \beta_{edu(i)}^{edu} + \beta_{age(i),edu(i)}^{age.edu} \end{aligned}$$

The estimate, 50% intervals, and 95% intervals⁸ for the demographic coef-

⁶For further discussion on efficient Gibbs samplers for multilevel models, see Gelman et al. (2003) sections 11.8 and 15.4.

⁷We independently simulate three sequences with starting points drawn from an overdispersed distribution. We monitored convergence by computing the potential scale reduction, \hat{R} , for all scalar estimands of interest and continued the simulation until \hat{R} was near 1 for all estimands of interest.

⁸A Bayesian (probability) interval for an unknown quantity of interest can be regarded as having a high probability of containing the unknown quantity, in contrast to a frequentist (confidence) interval, which may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice. Increased emphasis has been placed on interval estimation rather than hypothesis testing, and this provides a strong impetus to the Bayesian viewpoint, since it seems likely that most users of standard confidence intervals give

ficients are displayed in Figure 1. As can be seen from the graph, the demographic factors other than ethnicity are estimated to have little predictive power for this particular example. The multilevel model tends to shrink the coefficient estimates to zero. This is particularly true for the 16 age and education interactions. However, we follow the lead of the polling organization in keeping these predictors in the model since they can be important for some survey questions.

[Insert Fig. 1 Here]

The regression prediction can then be written as,

$$Pr(y_i = 1) = \text{logit}^{-1}(\beta^0 + \beta_{state(i)}^{state} + u_i).$$

Where u_i represents the combined linear predictor. We can plot this for each state. Figure 2 shows the result for a selection of 8 states. The solid lines display the estimated logistic regression in each state: thus, in any state, the probability of supporting Bush ranges from about 10% to 70% depending on the demographic variables, most importantly ethnicity. Roughly speaking, African-Americans have about a 10% probability of supporting Bush while others have about a 60% probability. Other demographic variables only slightly affect the predicted probability. The variation among states is fairly small but turns out to be important in allowing us to estimate average opinion by state. Changes of only a few percent in preferences can have large political impact.

[Insert Fig. 2 Here]

The gray lines on the graphs represent uncertainty in the state-level coefficients, β_j^{state} . Alaska has no data at all, but the inference there is still reasonably precise—its β_j^{state} is estimated from its previous election outcome, its demographic and regional predictors (Alaska is categorized as a Western state),
them a common-sense Bayesian interpretation (Gelman et al. 2003).

and from the distribution of the errors from the state-level regression. In general, the larger states such as California have more precise estimates than the smaller states such as Delaware, with more data in a state j , it is possible to estimate β_j^{state} more accurately.

[Insert Fig. 3 Here]

Figure 3 displays the estimated logistic coefficients for the 50 states, grouping them by region and, within each region, showing the multilevel regression on $v.prev$, the measure of Republican vote in the state in previous Presidential elections. Region and previous vote give good but not perfect predictions of the state-level coefficients in the public opinion model.

Using the Model Inferences to Get State-Level Estimates

The logistic regression model gives the probability that any adult will prefer Bush, given the person's sex, ethnicity, age, education, and state. We can now compute weighted averages of these probabilities to represent the proportion of Bush supporters in any specified subset of the population.

We first compute the expected response y^{pred} - the probability of supporting Bush for each of the categories $j = 1, \dots, J = 3264$ defined by the model. Since we have 1000 simulation draws, we compute a 1000 x 3264 matrix in R:

$$\begin{aligned}
 y^{pred} = & \text{logit}^{-1}(\beta^0 + \beta^{female} \cdot female_j + \beta^{black} \cdot black_j \\
 & + \beta^{female.black} \cdot female_j \cdot black_j + \beta_{age(j)}^{age} \\
 & + \beta_{edu(j)}^{edu} + \beta_{age(j),edu(j)}^{age.edu} + \beta_{state(j)}^{state}).
 \end{aligned} \tag{2}$$

for $j = 1$ to 3264. This is the same as (1) except with j in place of i , a notational change we make to emphasize that now that we have fit the model,

we are applying it to population categories j rather than survey respondents i .

We use the notation N_j for the number of adults as obtained from the 1990 Census and π_j to represent the probability of supporting Bush for each category j . For each state s , we are estimating the average response in the state,

$$y_s^{pred} = \frac{\sum_{j \in s} N_j \pi_j}{\sum_{j \in s} N_j}$$

summing over the 64 demographic categories within each state. The above calculations work because we have already prepared a 3264 x 6 matrix, from the 1990 Census, with columns N , indicators for female and ethnicity, and indexes for age, education, and state.

We can then summarize these 1000 simulations to get a point prediction and a uncertainty intervals for the proportion of adults in each state who supported Bush at the time of the surveys.

Comparing Public Opinion Estimates to Election Outcomes

The estimates of the model come from opinion polls taken just before the election, and they can be externally validated by comparing them to the actual outcomes. We can thus treat this as a sort of laboratory for testing the accuracy of multilevel models and any other methods that might be used to estimate state-level opinions from national polls.

[Insert Fig. 4 Here]

Figure 4 shows the actual outcomes for each state in 1988, compared to the model-based estimates of the proportion of Bush supporters. Validating by comparing to the actual election, the fit is pretty good, with no strong systematic bias and an average absolute error of only 4.0%; for 1992 it is 3.5%.

[Insert Fig. 5 Here]

By comparison, Figure 5 shows the predictive performance of the estimates based on complete pooling of states (estimating opinion solely based on demographics, thus setting $\beta_j^{state} = 0$ for all states) and no pooling (corresponding to completely separate estimates for each state, thus setting $\sigma_{state} = \sigma_{region} = \infty$) for 1988. The complete pooling model generally shrinks the state estimates too close towards the mean, whereas the no-pooling model does not shrink enough. To make a numerical comparison, the mean absolute error of the state estimates is 4.0% for the multilevel analysis, compared to 5.4% for complete pooling and 10.4% for no pooling. For 1992, Figure 6 compares the multilevel and no pooling models. Similar to 1988, the no pooling model does not shrink enough. The mean absolute error of the state estimates is 3.5% for the multilevel analysis, compared to 9.7% for no pooling model.⁹

[Insert Fig. 6 Here]

6 Discussion

We generated estimates of state-level vote choice opinions employing national data. The accuracy of these estimates was shown by comparison to actual vote outcomes. We use a Bayesian approach in this paper because of its generality and conceptual simplicity (Gelman et al. 2003), however, the key is really the placing of the model within a multilevel framework. Multilevel modeling is appropriate since the data have a hierarchical demographic and geographic structure (see Kreft and de Leeuw (1998); Snijders and Bosker (1999); Bryk and Raudenbush (2001)). This pattern of interlocking clustering is common in datasets employed by social scientists.

⁹We use the 3-step model to correct for nonregistrants and supporters of other candidates (mostly Perot). We do not bother to include the complete-pooling model for 1992 because, as seen in 1988, it generally shrinks the state estimates too close to the mean and therefore does not offer a credible alternative to the multilevel model.

As we have seen in this paper, multilevel modeling outperforms the simpler complete-pooling and no-pooling estimates, which is no surprise: not pooling ignores information and can give unacceptably variable inferences, and complete pooling suppresses variation that can be important or even the main objective of a study. These extreme alternatives can in fact be useful as preliminary estimates, but ultimately we prefer the partial pooling that comes out of a multilevel analysis.

We adopt multilevel modeling to the sample survey context via poststratifying, using Census data to weight our estimates by demographic characteristics per state. This allows us to use all the information in classical survey weights (in contrast to some modeling approaches that do not use information relevant to the data collection). In survey terminology, poststratification allows our estimates to correct for nonresponse bias, as well as would be done using classical weighting.

Moving forward, multilevel modeling and poststratification can be used to estimate state-level opinions on a variety of topics, such as ideology, partisanship, death penalty attitudes, spending on the poor, etc (Park 2004). Further, opinions can be estimated at levels other than the state. For example, researchers may be interested in estimating the opinions of populations in congressional districts or counties. It will also be interesting to study time trends in state and local opinions. We leave this to future research.

References

- Bryk, A. S. and Raudenbush, S. W. (2001), *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd edn, Sage Publications, Thousand Oaks, CA.
- Burdick, E. (1964), *The 480*, McGraw-Hill, New York.
- Gelman, A. (2003), *Bugs.R: Functions for calling Bugs from R*.
URL: <http://www.stat.columbia.edu/~gelman/bugsR>
- Gelman, A. (2004), ‘Prior distributions for variance parameters in hierarchical models’, *Bayesian Analysis* . Under Review.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2003), *Bayesian Data Analysis*, second edn, Chapman and Hall, London.
- 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* pp. 409–451.
- Gelman, A. and Little, T. C. (1997), ‘Postratification into many categories using hierarchical logistic regression’, *Survey Methodologist* **23**, 127–135.
- Jackman, S. and Rivers, D. (2001), ‘State level election forecasting during election 2000 via dynamic Bayesian hierarchical modeling’, *presented to the annual meeting of the American Political Science Association* .
- Kreft, I. G. G. and de Leeuw, J. (1998), *Introducing Multilevel Modeling*, Sage Publications, Thousand Oaks, CA.
- Park, D. (2004), Multilevel models of representation in the U.S. States. Doctoral Dissertation.

- Pool, I. d. S., Abelson, R. P. and Popkin, S. L. (1965), *Candidates, Issues, and Strategies*, M.I.T. Press, Cambridge, MA.
- R Development Core Team (2003), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3.
- URL:** <http://www.R-project.org>
- Snijders, T. A. B. and Bosker, R. (1999), *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*, Sage Publications, Thousand Oaks, CA.
- Spiegelhalter, D., Thomas, A. and Best, N. (1999), *WinBugs Version 1.4*, MRC Biostatistics Unit, Cambridge, UK.
- Voss, D., Gelman, A. and King, G. (1995), ‘Preelection survey methodology: Details from eight polling organizations, 1988 and 1992’, *Public Opinion Quarterly* **59**, 98–132.
- Weber, R. E., Hopkins, A. H., Mezey, M. L. and Munger, F. (1972-73), ‘Computer simulation of state electorates’, *Public Opinion Quarterly* **36**, 49–65.

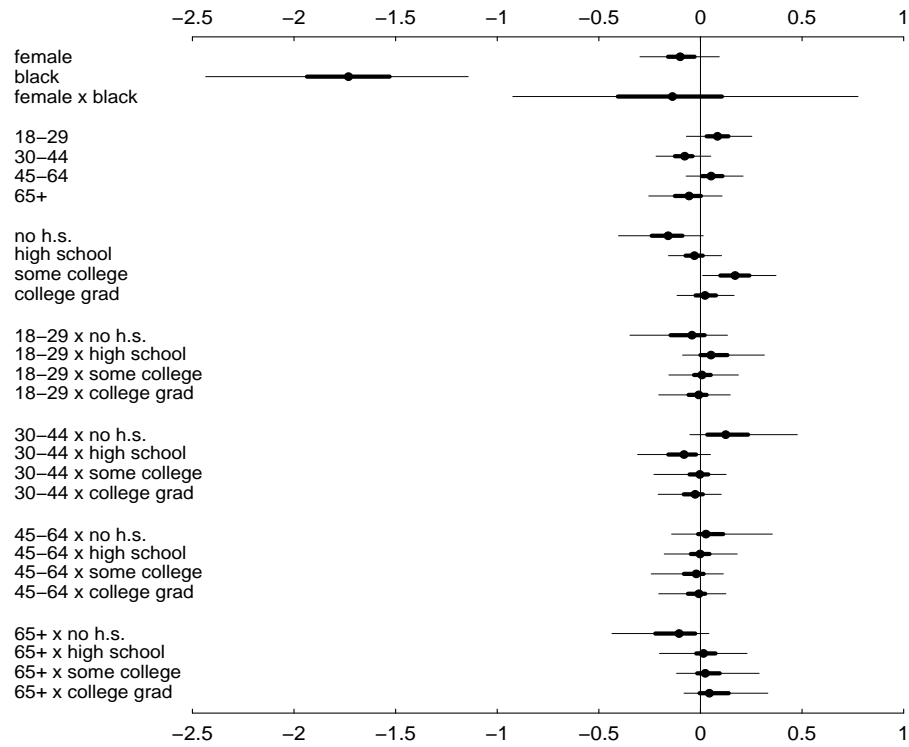


Figure 1: Estimates, 50% intervals, and 95% intervals for the demographic coefficients in the logistic regression of the probability of supporting George Bush in polls before the 1988 Presidential election. Recall that a change of x on the logistic scale corresponds to a change of at most $x/4$ on the probability scale. Thus, demographic factors other than ethnicity have very small estimated predictive effects on vote preference.

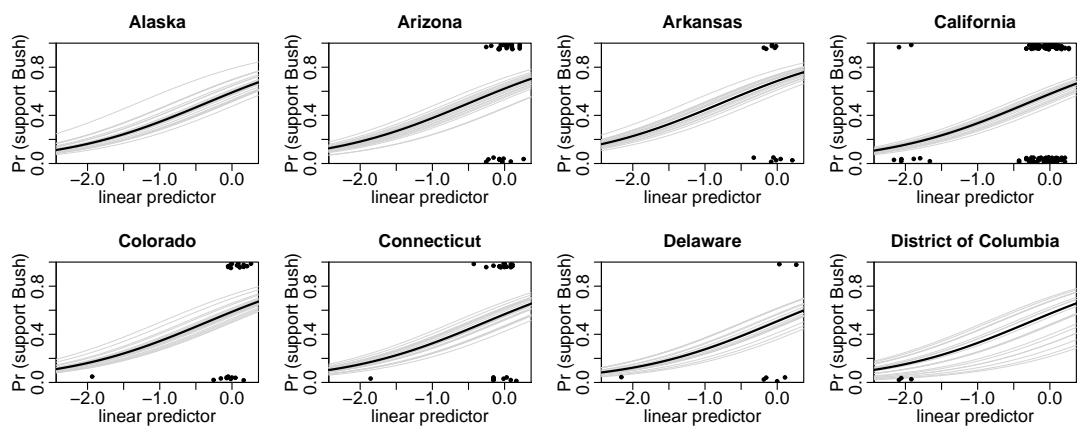


Figure 2: Estimated probability of a survey respondent supporting Bush for President in 1988, as a function of the linear predictor for demographics, in each state (displaying only a selection of 8 states to save space). Dots show the data (y -jittered for visibility), and the heavy and light lines show the median estimate and 20 random simulation draws from the estimated model.

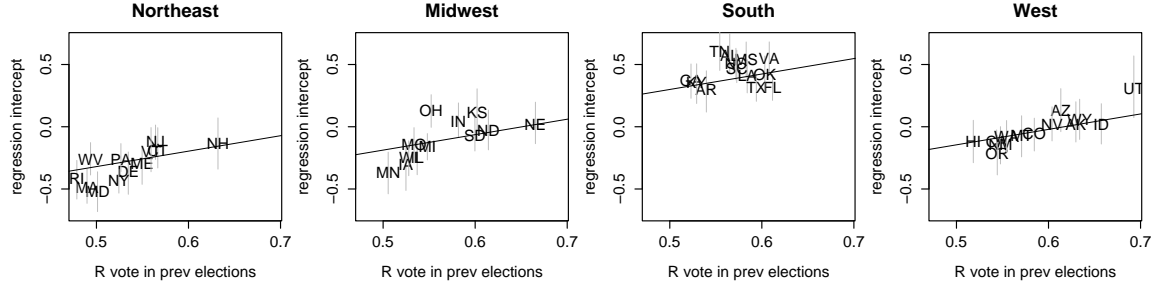


Figure 3: Estimates and 50% intervals for the state coefficients β_j^{state} , plotted versus previous state vote $v.\text{prev}_j$, in each of the four regions of the United States in 1988. The estimated group-level regression line, $\beta_j^{\text{state}} = \beta_k^{\text{region}} + \beta_j^{\text{v.prev}}$, is overlain on each plot (corresponding to the regions $k = 1, 2, 3, 4$).

1988

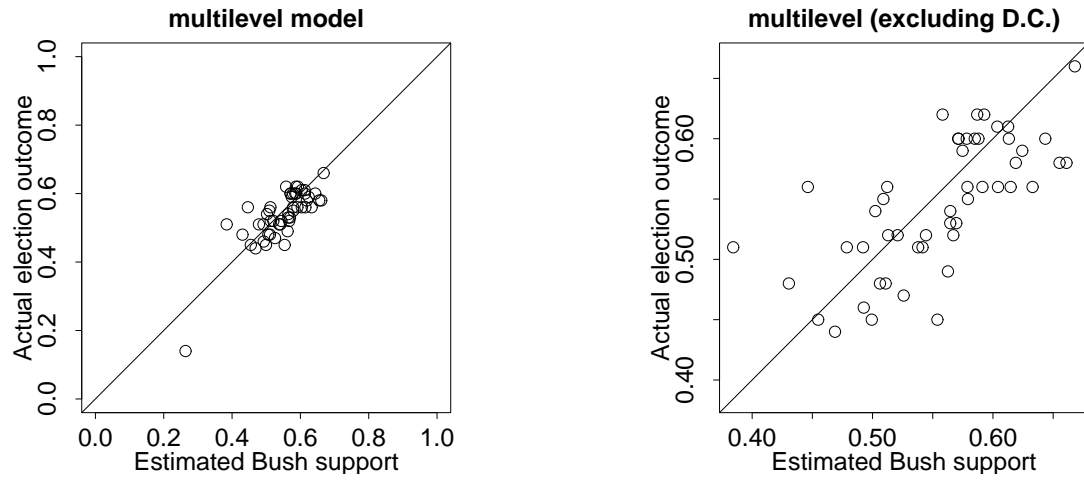


Figure 4: For each state, the proportion of the two-party vote received by George Bush in 1988, plotted vs. the support for Bush in the state, as estimated from a multilevel model applied to pre-election polls. The second plot excludes the District of Columbia in order to more clearly show the 50 states.

1988

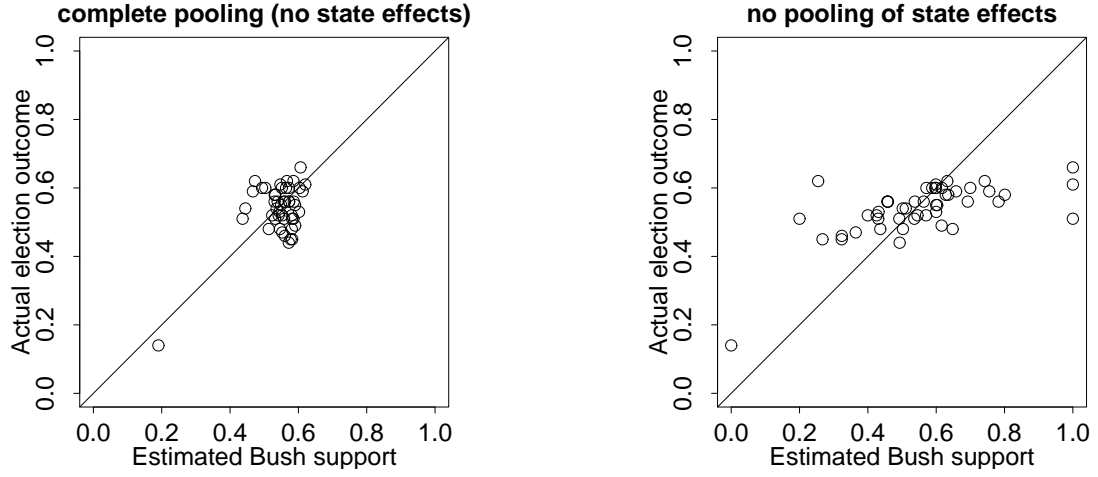


Figure 5: For each state, Bush's vote in 1988 plotted vs. his support in the polls, as estimated from (a) the complete-pooling model (using demographics alone with no state predictors), and (b) the no-pooling models (estimating each state separately). The three states in the no-pooling model with estimated 100 percent support for Bush were: Utah ($n=12$), Vermont ($n=2$), and Wyoming ($n=2$). The complete pooling and no pooling models correspond to $\sigma_{\text{state}} = \sigma_{\text{region}} = 0$ and ∞ , respectively. Compare to Figure 4a, which shows results from the multilevel model (with σ_{state} and σ_{region} estimated from data).

1992

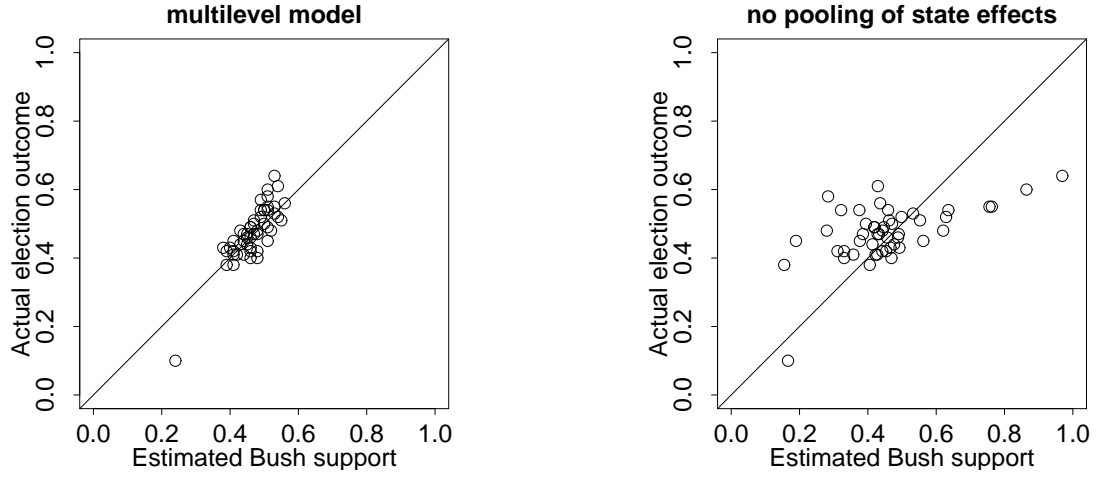


Figure 6: For each state, Bush's vote in 1992 plotted vs. his support in the polls, as estimated from (a) the multilevel model, and (b) the no-pooling model (estimating each state separately). The no pooling model correspond to $\sigma_{\text{state}} = \sigma_{\text{region}} = \infty$.