Using state polls to forecast presidential election outcomes in the American states

Thomas M. Holbrook[a],*, Jay A. DeSart[b]

[a]*Department of Political Science, University of Wisconsin-Milwaukee, Milwaukee, WI 53201, USA*
[b]*Department of Political Science, Florida Atlantic University, Boca Raton, FL, USA*

Abstract

This paper presents a simple forecasting model for state-level presidential outcomes, based on statewide preference polls and a lagged vote variable. The analysis illustrates two important points. First, the candidate who is leading in a state in September usually goes on to win that state in the November election. Second, the combination of pre-election preference polls and a lagged dependent variable generates highly accurate estimates of presidential election outcomes in the states. The limits of using statewide preference polls are also discussed. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: State polls; Presidential elections; Forecasting; America

1. Introduction

While the media and campaign organizations have been using scientific polling for several decades, there has been a real explosion in the number of public opinion polls used by the media in the last several years (Ladd and Benson, 1992). This increase in the use of public opinion polls is most noticeable during presidential campaigns, as media organizations tend to focus on the ''horse race'' aspect of the campaign. While most of the media's attention is on the national race and national poll results, the use of state polls has also become increasingly common during presidential campaigns. Although these state polls are most often used to give residents a glimpse of how the race is unfolding within the state, they also present a real opportunity to scholars who are interested in forecasting election

outcomes. Specifically, state-level presidential preference polls may provide a useful instrument for forecasting the outcomes of the presidential contests in the states.

The enterprise of forecasting elections has been around since the early work of Bean (1948) and found renewed contemporary interest beginning with the work of Kramer (1971) and Tufte (1978). Although election forecasting is not without its critics (Beck, 1992; Greene, 1993), it has gained fairly wide acceptance as a legitimate academic endeavor.[1] With few exceptions, forecasting studies have focused on the national presidential outcome. Campbell (1992) and Rosenstone (1983), however, developed forecasting models for presidential out-

---

[1]See the October 1996 issue of *American Politics Quarterly* for a recent collection of essays on forecasting the 1996 presidential election.

*Corresponding author.

comes in the states. Both of these studies found that a combination of long term trends in party support and political ideology, along with short-term factors, such as the state of the state economy and the home state advantage, help to explain (and forecast) state-level presidential outcomes. In a recent article, Cohen (1998) examined the relationship between state poll results (trial heats and candidate image measures) and state outcomes in the 1996 election and found a fairly strong relationship between the poll results and election outcomes. While Cohen's findings are informative, his study focused less explicitly on developing a forecasting model for state outcomes than does the current analysis. The goal of this analysis is to contribute to the body of forecasting literature by providing a parsimonious, public opinion-based forecasting model for state-level outcomes. By examining the extent to which pre-election polls can be used to forecast state-level election outcomes, it is hoped that this paper can contribute to and improve upon the literature on forecasting elections in the states.

It is perhaps useful to begin with an articulation of what it is that constitutes a ''good'' forecasting model. First, a good forecasting model must provide accurate statistical estimates of the dependent phenomenon. Second, a good model should also allow forecasts to be made in time significantly prior to the dependent phenomenon. If variables used in the model do not occur or cannot be measured prior to the event being forecast, true forecasting is not possible. And, finally, a good forecasting model should be as simple and parsimonious as possible. This is especially true for national election models, when the number of cases is relatively small, but is also true in the case of forecasting state-level outcomes, when the number of cases is larger. In both cases, it is important to keep the model simple so that forecasts can be made and understood with relative ease. The model developed here is a nice combination of accuracy, lead time, and parsimony.

2. State polls and election outcomes

In this analysis, results from the state polls taken in September are used to predict the November

outcomes in the 1992 and 1996 presidential races.[2] September polls are used because they are taken after the national party conventions, when the campaign has settled into its fall mode, and because they are taken long enough before the election to make interesting and useful predictions. Using polls from August or earlier as the benchmark would give us better lead time but would also entail dropping most states (especially the smaller ones) from the analysis due to a lack of polling data.[3] Using September polls also grants this analysis some real practical applications. For instance, when allocating resources for October, campaigns are likely to be interested in whether their standing in the polls in September provides them with useful information concerning their likely position on election day. In the event that they provide accurate forecasts, the September poll results could inform strategists concerning which states are ''safe'', which states are out of reach, and which states are up for grabs.

[2]Several sources were used to gather the poll data. For the 1992 data, we relied primarily on the November/December (1992) issue of *The Public Perspective*. For the 1996 data we relied on the now defunct POLITICSNOW web page, as well as data provided to us by one of the pollsters, Mason Dixon Political Media Research. In each year the polls were administered by dozens of different organizations, frequently from within the state. There were a couple of organizations that polled in virtually every state: the American Research Group and Mason Dixon Political Media Research in 1992, and Mason Dixon Political Media Research in 1996. Unfortunately, we do not have information on question wording and, in some cases, also do not have information regarding sample size.

[3]Campbell (1996) has found in his forecasting studies that the predictive ability of the national polls settles into a stable pattern by the post-convention period of the campaign. In both years of this analysis the September polls offer a significant predictive improvement over the August polls, and the October polls offer some improvement over September, although just marginally in 1996:

Correlation Between Poll Results and Election Outcome

	August	September	October
1992	0.49	0.81	0.94
1996	0.82	0.91	0.95

No doubt, the poorer performance of earlier polls in 1992 is due to Ross Perot's renewed candidacy in October, which occurred after the August and September polls.

Table 1
September polls as predictors of state-level presidential election outcomes

Actual outcome	September poll standings		
	Bush ahead in polls	Toss-up states	Clinton ahead in polls
1992[a]			
Bush victory	5	10	3
Clinton victory	0	6	26
1996[b]			
Dole Victory	7	11	1
Clinton Victory	0	5	26

Note: Cell entries are the number of states in each cross-classification. The September poll standings are based on which candidate had a significant lead in the polls conducted in each state during the month of September. Toss-up states are those states for which the candidate lead is not significantly different from zero, using a 95% confidence interval.
[a] $X^2 = 22.05$ ($P < 0.0000$) Tau-c $= 0.64$.
[b] $X^2 = 31.32$ ($P < 0.0000$) Tau-c $= 0.80$.

The specific measure of poll standing used in this analysis is the state-level average of Clinton's percent of the two-party vote in state-level trial heat polls conducted during the month of September.[4] This figure is calculated separately for each state in 1992 and 1996. Table 1 offers an initial, if somewhat simple glimpse of the potential ability of state polls to predict state outcomes. This table presents the cross-tabulation of a blunt but useful classification of state-level poll standing and state outcomes. States are categorized, based on the September poll results, into three groups: those in which the Republican candidate was ahead in the polls, those in which the Clinton was ahead in the polls, and those in which the outcome was within the margin of error of the polls.[5] The states are also classified on the dependent variable according to which candidate won the state. Both the cell contents and the measures of associa-

tion point to the potential usefulness of state polls as indicators of eventual outcomes. In 1992, the candidates won 31 out of the 34 states (91%) in which they held a significant lead in the polls in September. Bush won all 5 states in which he was ahead and Clinton won 26 of the 29 states in which he was ahead. Of the three states Clinton lost, two (Alaska and Arizona) were states in which Perot garnered more than his national average (28% and 24%, respectively), which no doubt had a deleterious impact on Clinton's standing,[6] and one (North Carolina) was lost by less than a percentage point. In 1996, the candidates won 33 of the 34 states (97%) in which they held a significant lead in the September polls. The only observation on the discordant diagonal in 1996 is Colorado, which Clinton lost by less than a percentage point.

Of course, there are still a number of states left up in the air because neither candidate held a significant lead in the polls. However, if one wanted to throw caution to the wind and make forecasts without regard to the margin of error in the polls and simply rely on point estimates, there would still be a fairly high level of accuracy: when all states are classified according to which candidate is ahead, regardless of

[4] Polls that were taken largely in September but finished on October 1 are included in this analysis. Also, in South Dakota in 1996, the only September poll available started on September 30 and ended on October 2.

[5] The margin of error for individual polls (based on a 95% confidence interval) ranged from 2.17 to 5.51 percentage points in 1992, and from 3.09 to 4.89 percentage points in 1996. The average margin of error was 4.24 in 1992 and 3.66 in 1996. In 1996, there were a number of observations for which we could not determine sample size and, therefore, could not calculate the margin of error. Rather than exclude these cases from the analysis, we substituted the average error (3.66) for those observations for which we did not have information on sample size.

[6] There is strong negative correlation at the state level between Perot support and Clinton support (−0.66) and virtually no relationship between Perot support and Bush support (−0.05).

the margin, 39 states are predicted correctly in 1992 and 44 states are predicted correctly in 1996 (tau-$b_{1992} = 0.51$, tau-$b_{1996} = 75$. The weaker performance of the polls in 1992 is no doubt due to Perot's renewed candidacy during the month of October. Since the polls were all conducted in September, they had no way of taking into account the impact of Perot's candidacy on the allocation of votes to the major party candidates.

3. A forecasting model

A somewhat more sophisticated method for evaluating the predictive capacity of September polls is presented in Eq. (1) and tested Table 2.

$$\text{VOTE}_i = \alpha + \beta_1 \, (\text{POLL})_i + \beta_2 (\text{PRIOR VOTE})_i \tag{1}$$

Where VOTE is Clinton's share of the two-party statewide vote, POLL is Clinton's average share of the two-party vote in polls taken in September, and PRIOR VOTE is the mean percentage of the two-party vote going to the Democratic candidate in the previous two presidential elections. The use of the lagged vote variable is necessary to control for trends in party support that may not be reflected in the September polls. In addition to controlling for longer term voting trends, this variable also serves other useful functions. First, it provides a benchmark against which the impact of the September polls can

be judged. Second, this variable allows us to control for the possibility that voters are returning to their base between the time of the polls and election day. To the extent that the prior vote variable is significant when it is in the same equation as the September poll result, it could indicate that the party faithful, who may have flirted with other candidates early in the fall, return home by election day.

The OLS regression analysis of the model, presented in Table 2, reveal a number of interesting findings. First, both prior vote and the September poll results are strongly correlated with the level of candidate support on election day. Second, both variables are more strongly related to candidate support in 1996 than in 1992, once again perhaps indicating the impact of the Perot candidacy in 1992. Third, while each variable provides fairly accurate estimates by itself, a model including both variables offers significant improvement over the bivariate models. The full model generates higher R^2 values and lower standard errors (based on out-of-sample predictions) than either of the bivariate models. Based on the coefficients from the full model, 90% of all cases were predicted correctly in 1992 and 92% were predicted correctly in 1996.[7] The full model also indicates that the September polls were more strongly related to election outcomes in 1996 than in 1992.

One of the drawbacks of the analysis in Table 2 is

---

[7]These predictions are made on the basis of point estimates and do not take confidence intervals into account.

Table 2
Forecasting Model of Presidential Votes in the States

OLS Regression	Prior vote model		September poll model		Full model	
	1992	1996	1992	1996	1992	1996
Constant	12.78 (3.70)*	−5.28 (4.74)	11.05 (4.33)*	−0.31 (3.47)	4.40 (3.42)	−7.10 (3.47)*
Prior Vote	0.93 (0.09)*	1.21 (0.10)*	–	–	0.59 (0.10)*	0.51 (0.13)*
September Polls	–	–	0.74 (0.08)*	0.98 (0.06)*	0.41 (0.08)*	0.65 (0.10)*
N	50	50	50	50	50	50
R^2	0.70	0.76	0.65	0.83	0.81	0.88
SE (Prediction)	3.53	3.93	3.81	3.31	2.96	2.96
% Correct	90%	86%	86%	92%	90%	92%

Note: The dependent variable is Clinton's share of the two-party state outcome, Prior Vote is the average Democratic share of the state presidential vote in the previous two elections, and September Polls is the average of all polls taken in the state during September of the election year. Coefficients are unstandardized, and the standard error of the coefficient is in parentheses. The standard error of the model is based on the error in out-of-sample predictions.
*$P < 0.01$.

that there are slight differences in the coefficients for the two years. This is not terribly important for understanding the general impact of the model on the dependent variable but it does have important implications for using the model for forecasting outcomes. Suppose, for instance, that one wanted to use these results to forecast the state-level outcomes for the 2000 election. With two different sets of coefficients it would not be possible to apply this model to 2000. Even applying the most recent results (1996) would be unwise, since it is clear that the coefficients vary from year to year.

The analysis in Table 3 represents an effort to provide a more general model that could be applied to future elections. In this table, data from the two years are pooled together to provide a single set of coefficients. One model adds a year dummy variable ($0 = 1992$, $1 = 1996$), intended to capture the additive influence of the year of the election, to the original model (Eq. (1)). This dummy variable is significant and indicates that Clinton did worse in 1996 than he should have, based on the prior vote and September poll standings. This could represent the impact of the campaign in October, or perhaps the revelations concerning fund-raising improprieties late in the campaign. The problem with this variable, however, is that it provides a means for identifying only what is unique about 1996. In forecasting future elections,

it is not possible to know what is unique about a given year until after the election. For that reason, we feel it is preferable to incur slightly more error and use the more general model in the second column of Table 3, which excludes the dummy variable. The general model indicates a strong influence from both independent variables, although the September poll variable is somewhat more important than prior vote. Both variables, together, once again provide highly accurate projections of the eventual outcome, although the level of accuracy is somewhat stronger in 1996 than in 1992.

4. Conclusions

This analysis has presented a simple and useful model for forecasting statewide presidential election outcomes. The model has all of the features of a good forecasting model, according to the criteria discussed earlier: it is parsimonious, provides accurate forecasts, and includes a reasonable amount of lead time. But the model is not without drawbacks. For instance, the extent to which the results of this analysis can provide an accurate forecast of state outcomes in 2000 remains to be seen. Obviously, we would feel more comfortable if the forecasts were based on more years of data. But statewide polling data have only been plentiful in the last two election cycles, so we are somewhat limited in terms of data availability. Of course, it is expected that the model will be strengthened by the addition of more data over time. But the important point that should not be lost is that state polls taken early in the fall campaign can be very strong predictors of the eventual outcome. Even in the simple analysis in Table 1, which looked only at whether the candidates were ahead in the polls, it was clear that candidates who are leading in a state in September generally go on to win that state on election day. Ultimately, this is the type of knowledge we hope to glean from a election forecasting models.

Table 3
A general forecasting model for state-level presidential outcomes

	OLS Regression	
	Year-specific model	General model
Constant	−1.03 (2.39)	−0.52 (2.47)
Prior vote	0.58 (0.08)*	0.43 (0.06)*
September Poll	0.52 (0.06)*	0.61 (0.05)*
Year	−2.14 (0.77)*	–
N	100	100
R^2	0.84	0.83
SE (Prediction)	2.94	3.03
% Correct Overall	91%	90%
% Correct 1992	90%	88%
% Correct 1996	92%	92%

Note: All variables are the same as those used in Table 2, with the exception of the year dummy variable ($1992 = 0$, $1996 = 1$). Coefficients are unstandardized, and the standard error of the coefficient is in parentheses. The standard error of the model is based on the error in out-of-sample predictions.
*$P < 0.01$.

References

Beck, N. (1992). Forecasting the 1992 presidential election: the message is in the confidence interval. *The Public Perspective*, *3*, 32–34.

Bean, L. (1948). *How to Predict Elections*. A.A. Knopf, New York.

Campbell, J. E. (1992). Forecasting the presidential vote in the states. *American Journal of Political Science*, *36*, 386–407.

Campbell, J. E. (1996). Polls and votes: the trial-heat presidential forecasting models, certainty, and political campaigns. *American Politics Quarterly*, *24*, 408–433.

Cohen, J. (1998). State-level public opinion polls as predictors of presidential election results: the 1996 race. *American Politics Quarterly 26*, (to be provided).

Greene, J. P. (1993). Forewarned before forecast: presidential election forecasting models and the 1992 election. *PS*, *26*, 17–21.

Kramer, G. H. (1971). Short-term fluctuations in U.S. voting behavior, 1896–1964. *American Political Science Review*., *65*, 131–143.

Ladd, E. C., Benson, J. (1992). The growth of news polls in American politics. In: Mann, T. E., Orren, G. R. (Eds.), *Media Polls in American Politics*. Brookings Institute, Washington.

Rosenstone, S. (1983). *Forecasting Presidential Elections*. Yale University, New Haven CT.

Tufte, E. R. (1978). *Political Control of the Economy*. Princeton University, Princeton.