Seeking 50% of Seats, Needing More than 50% of Votes: Predicting the Seats-Votes Curve in the 2006 Elections*

Jonathan P. Kastellec† Andrew Gelman‡ Jamie P. Chandler§

October 14, 2006

Abstract

After their stunning loss of both houses of Congress in 1994, the Democrats have averaged over 50% of the vote in Congressional races in every year except 2002, yet they have not regained control of the House. The same is true with the Senate: in the last three elections (in which 100 senators were elected), Democratic candidates earned three million more votes than Republican candidates, yet they are outnumbered by Republicans in the Senate as well. 2006 is looking better for the Democrats, but our calculations show that they need to average at least 52% of the vote (which is more than either party has received since 1992) to have an even chance of taking control of the House of Representatives.

Why are things so tough? Looking at the 2004 election, the Democrats won their victories with an average of 69% of the vote, while the Republicans averaged 65% in their contests, thus "wasting" fewer votes. More formally, we estimated the seats-votes curve for 2006 by constructing a model to predict the 2006 election from 2004, and then validating the method by applying it to previous elections (predicting 2004 from 2002, and so forth). We predict that the Democrats will need 49% of the average vote to have a 10% chance, 52% of the vote to have an even chance, and 55% of the vote to have a 90% chance of winning the House. The Democrats might be able to do it, but it won't be easy.

*We thank Gary Jacobson for generously sharing his data, Robert Erikson, Eduardo Leoni, Robert Shapiro, and David Epstein for helpful comments, and the National Science Foundation for financial support.

†Department of Political Science, Columbia University, New York, jpk2004@columbia.edu

‡Department of Statistics and Department of Political Science, Columbia University, New York, gelman@stat.columbia.edu, www.stat.columbia.edu/∼gelman/

§Department of Political Science, City University of New York, Graduate School and University Center, jchandler@gc.cuny.edu

1 Introduction

With the 2006 midterm elections fast approaching, pollsters, scholars, and journalists are busy attempting to predict whether the Democrats will take back the House or Senate, or both. As with any election, most of these efforts focus on the production of a point prediction: the Democrats will pick up x seats, meaning they either will or will not control the House or Senate in the 110th Congress. Brandt and Brunell (2006), for example, predict that the Republicans will win 220 seats in the House, just two more than necessary to maintain the majority they have enjoyed since the 1994 elections. Similarly, Campbell (2006) predicts that the Democrats will pick up 13 seats, leaving them just short of a majority.[1] Congress scholars Ansolabehere, Erikson, and Jacobson have also registered their point predictions (Boston Review 2006).

While predicting seats is indeed an interesting endeavor, in this paper we ask a different question: how many votes do the Democrats need to have a realistic chance of winning the House? More specifically, we attempt to predict the seats-votes curve for the 2006 election, using a parsimonious model of elections to translate the average district vote for Democrats into the number of seats they will gain (or lose) in the House. Unlike macro-level studies that attempt to model the outcome in seats using such predictors as economic indicators and presidential approval, our approach has the comparative advantage of using information from every congressional district. This allows us to predict systematically the outcome in each district across a range of hypothetical swings in the average district vote from the 2004 House elections to the upcoming elections.

To get right to our conclusions, we predict that the Democrats will need slightly more than 52% of the average district vote to have a 50% probability of retaking the House. Obtaining about 49% of the vote will give them a small chance (10%) of gaining a majority, while obtaining 55% will all but ensure it (90%). To put these predictions into historical predictions, Figure 1 plots the time series of the average district vote in the post-war era, along with our 10%, 50% and 90% predictions (the shaded regions depict periods of Republican control). In 2004, the Democrats received 50.4% of the average district vote; according to our prediction, they would need about a two percent

[1] As of October 6, 2006, Republicans held 231 seats, Democrats 201 (including Bernard Sanders (VT), an independent who has organized with the Democrats and who is running for the Senate and will vacate his seat). In addition, there were three vacancies, two of which were created by Republican resignations (Tom Delay (TX-22) and Mark Foley (FL-16)) and one by a Democrat (Robert Menendez (NJ-13), who is also running for the Senate). Thus, assuming the Democrats retain Menendez's and Sanders's seats, they would need to pick up 15 seats to obtain a majority (Office of the Clerk 2006).

Figure 1: *Average district vote for Democrats, 1946–2004, along with prediction for seats-votes translation in 2006. The black line depicts the average district vote for Democrats in each House election from 1946 to 2004; the gray line is a reference for 50% of the vote. The lines connecting 2004 to 2006 show our estimates for what the Democrats need to have a 10%, 50% and 90% probability of retaking the House. The shaded regions indicate periods of Republican control of the House. In calculating the average district vote, uncontested seats are imputed to 25% and 75% of the district vote for Republican and Democratic winners, respectively (see the footnote on page 4 for a discussion of this imputation).*

national swing to have a 50% chance of the controlling the House, and about a four percent swing to have a 90% chance. Either of these vote shares would constitute the Democrats' largest since 1992, the last election in which they won a majority of House seats. But both fall well below the average vote shares the Democrats enjoyed during their four-decade control of the House.

The next section of this article discusses previous efforts to forecast House elections, and how our approach differs. We then present our method, which is based on a simple model of national swings and incumbency advantage. We validate our method by using it to predict previous national election outcomes. We then use the model to predict the seats-votes curve in 2006.

2 Predicting House Elections Nationally or by Districts

Predictions of presidential elections attempt to answer a simple question: which candidate will win? Likewise, most predictions of legislative elections attempt to answer the simple questions of which party will control the legislature, and by how much. But just as the outcome of a presidential election depends on 50 state elections, the outcome of a House election depends on 435 congressional elections.

For the most part, scholars who have attempted to predict House elections have nevertheless eschewed forecasts based on micro-level foundations, instead using macro-level variables to predict the aggregate vote total or the seat division itself. Tufte's (1975, 1978) famous referendum model, for instance, attempts to predict the national vote division using presidential approval and national economic indicators as predictors. Subsequent research directly models the seat division using similar predictors (see, for example, Hibbs 1982, Lewis-Beck and Rice 1984, Oppenheimer, Stimson and Waterman 1986, Marra and Ostrom 1989). Another line of inquiry has sought to use national polls to predict seat outcomes (see, for example, Abramowitz, Cover and Norpoth 1986, Erikson and Sigelman 1995).

Macro-level predictions have their advantages. The point predictions that result from them are intuitive and lead to a dichotomous answer—which party will control the House—that is of interest to scholars, politicians, the media, and the public alike. It is also much easier to collect data at a single level of analysis per election, allowing one to ignore idiosyncrasies at the district level.

The district level, however, incorporates a wealth of information that macro-level predictions throw away by definition. Variables such as the predisposition of voters for one party or the other, incumbency and whether a race is uncontested play crucial elements in determining the vote division in each race. This, in turn, determines who wins each race, which in turn affects the ultimate seat division. Thus, predicting the vote is true to the actual process by which seats are aggregated.[2]

A district-level forecast also allows us to focus on several interesting dynamics of House races that are relevant to predicting the 2006 vote. As Erikson (2006) notes, control of Congress is not simply a function of the preferences of the electorate; factors such as geography, incumbency advantage, and partisan bias in the seats-votes curve influence which party controls Congress, and for how long. The Democratic party will have to overcome a built-in incumbency advantage for the Republicans. And as we discuss below, they will also have to overcome a significant partisan bias that has emerged in favor of the Republicans since their takeover of the House in 1994. Democratic

[2]Lewis-Beck and Rice (1984, 476) argue as follows for their seat-level approach to election forecasting: "The models predict vote loss, not seat loss, which is the dependent variable of primary concern. Therefore, in order to estimate actual seat loss, it is necessary to apply the 'swing ratio of votes to seats given by a further regression equation.' Besides the extra calculation involved, this additional step produces more prediction error, for the fit in this votes-seats equation is less than perfect." Lewis-Beck and Rice are correct that predicting votes, then seats, is a two-step process. However, individual district vote proportions (along with incumbency information) give more information than are either in aggregate seats or votes. In this paper, we seek to determine what vote swing the Democrats need to have a reasonable probability of taking the House, and so we need to work with the seats-votes curve in any case.

candidates have been winning as many votes, on average, as their Republican counterparts (recall Figure 1), yet have come up short in seats. How many votes will they need to reverse this trend?

3 The Method

We draw our prediction of the seats-votes curve in 2006 from the a statistical model based on the framework for evaluating legislative elections advanced in Gelman and King (1994). The strength of this approach lies in its ability to forecast elections under a range of possible conditions through the use of simulation analysis that incorporates both a range of possible national swings and the predictive uncertainty inherent in any forecast. This incorporation of uncertainty allows us to make probabilistic predictions about the Democrats' chances of retaking the House. Finally, the approach is also quite parsimonious, employing as predictors only the vote in the prior election in each district, whether the incumbent is running for reelection and whether the race is contested or not.

3.1 Simulating Election Outcomes Under a Range of Hypothesized National Swings

Our method can be summarized informally as follows. (A formal description, along with details about variable coding, appears below.) We begin with the consideration that there is a range of possible national swings in the average district vote from one election year to the next. A swing in the direction of the Democratic party will help the Democrats pick up more seats; more specifically, every possible swing will have different implications in more closely contested elections and will lead to a different translation of seats for the Democrats. Our general strategy is to allow the swing to vary across a range of values, and then predict the outcome of every district at each interval in the range. Aggregating these results produces an estimated seats-votes curve.

The partisan swing, however, is unlikely to affect each district in the same manner. We account for district-level heterogeneity by incorporating incumbency and the 2004 district-level vote, which serves as a proxy for the district's overall predisposition to support the Democratic candidate.[3]

[3]For all analyses that appear in this paper, we impute the district vote for uncontested races to 0.75 for Democratic winners and 0.25 for Republican winners. These values are derived in King and Gelman (1991) and Gelman and King (1994), based on the vote shares received in the last election before a district became uncontested and the first election after a district became uncontested. The average of these values was about 0.75 for the incumbent party and

Within each interval of the possible national swing, we first use the results of the previous House election to estimate each district's baseline support for the Democratic candidate; we do so by reducing the observed vote total in each district in which an incumbent ran in the previous election by the estimated advantage of incumbency; this measure can be thought of the as "normal vote" for the previous election. We then shrink the normal vote toward 50% to represent that some of this vote arises from election-specific events, and we are interested in the part that continues on to the next election. The next step is to add back in the estimated incumbency advantage to the normal vote for incumbent candidates in the current election. This procedure results in a hypothetical proportion of the Democratic share of the two-party vote in each district. To account for uncertainty in the prediction, we simulated the outcome in each district 1000 times, adding independent normally-distributed errors whose standard deviation is based on the residual error from historical House election forecasts. Finally, for each simulation, we shift the entire election result (that is, add or subtract a constant from all the districts) so that the average district vote matches up to that of the previous election. The result is 1000 hypothetical election outcomes, each corresponding to a zero national swing (but various district-by-district differences).

The next task is to add the national swing to each set of district-level results. We do so in the form of adding a constant to the predicted value within each simulation; we set the constant to a value that ensures that the range of national swings is centered around the average district vote from the previous election.

For every interval in the national swing (in the entire range of average district vote), we thus have 1000 predicted outcomes in each congressional district. We summarize each simulation in two ways. First, we compute the average district vote; second, we compute the proportion of House seats won by the Democrats. Averaging these over all the simulations yields an expected seat share for the Democrats.[4] We then simply "connect the dots" of Democratic share in each interval, producing a predicted-seats votes curve.

Formally, we model the vote in each district as follows:

$$v_i^t = \alpha^t + \rho^t(v_i^{t-1} - \psi^t I_i^{t-1}) + \psi I_i^t + \delta^t + \epsilon_i, \quad \epsilon_i \sim \mathrm{N}(0, \sigma^2), \tag{1}$$

represents the average "effective support" for the party in uncontested races. For a measure based on the presidential vote in congressional districts, see Erikson (2006).

[4]Similar results are obtained by using the median instead of the mean to summarize predictive uncertainty.

where v = the Democratic share of the two-party vote (as a fraction), i indexes districts, t represents the current election (that in which the seats-votes curve is being predicted) and $t-1$ the previous election. The incumbency variable (defined in each district in each election year) is $I = 1$ if the incumbent in the district is a Democrat and seeking reelection, -1 if the incumbent is a Republican and seeking reelection, and 0 if the seat is open. For values of ψ (the estimated incumbency advantage), we use the trend of the values estimated in Gelman and Huang (in press), who find a steadily increasing incumbency advantage from the 1940s to 1980, and a constant value of around 8% since.[5] The coefficient ρ captures the autocorrelation of votes between elections.[6] The national swing δ is set to values ranging from -0.10 to 0.10 (that is, considering swings of up to 10% of the vote), with the constant term α set so that the national vote swing equals 0 when $\delta = 0$.[7]

For every value of δ^t, we simulate our model 1000 times. For each simulation, we record the share of seats predicted to be won by Democratic candidates (that is, $\frac{1}{n}\sum_{i=1}^{n} 1_{v_i^t > .5}$), where n equals the number of districts in a given election year.[8] We then take the average seat share measure to produce an expected seat proportion for the Democrats at the given value of δ^t. We repeat this procedure over a range of values of δ^t, allowing for a predicted translation between votes and seats at a potentially plausible range of national swings.

3.2 Validating by Fitting to Previous Elections

To demonstrate the validity of our method, we apply it to the 19 non-redistricting election years from 1958 to 2004. Figure 2 plots the predicted seats-votes curve for each year, along with the actual average district vote and seats share for the Democrats. The model performs fairly well, with a root mean-squared error of about 0.02, or about 9 seats out of 435. Two elections (1960 and 1980) contribute disproportionately to the predictive error, as the model is off by 26 and 19 seats, respectively, in those years. In 10 election years, the prediction error is less than 4 seats. The

[5] Specifically, we set $\psi = 0.08((\text{year} - 1946)/(1980 - 1946))$ if year < 1980, or $\psi = 0.08$ if year ≥ 1980.

[6] We obtain values of ρ by regressing v_i^t on v_i^{t-1}, along with incumbency and party indicators, for each non-redistricting election year (i.e., those not ending in "2") from 1946–2004. For each election year analyzed, we take the average of the coefficient on the lagged vote from the previous five years. The estimates of ρ have a mean of 0.68 with a standard deviation of 0.11.

[7] We set σ to 0.066, which is the mean of the residual standard deviations from the yearly regressions discussed in the previous footnote.

[8] In 1958 and 1960, respectively, 436 and 437 seats were contested due to the addition of Alaska and Hawaii as states. We also dropped the handful of races featuring third-party candidates (with the exception of Bernard Sanders—an independent who held Vermont's congressional seat from 1998–2004 who organized with the Democratic party—whom we classified as a Democrat).

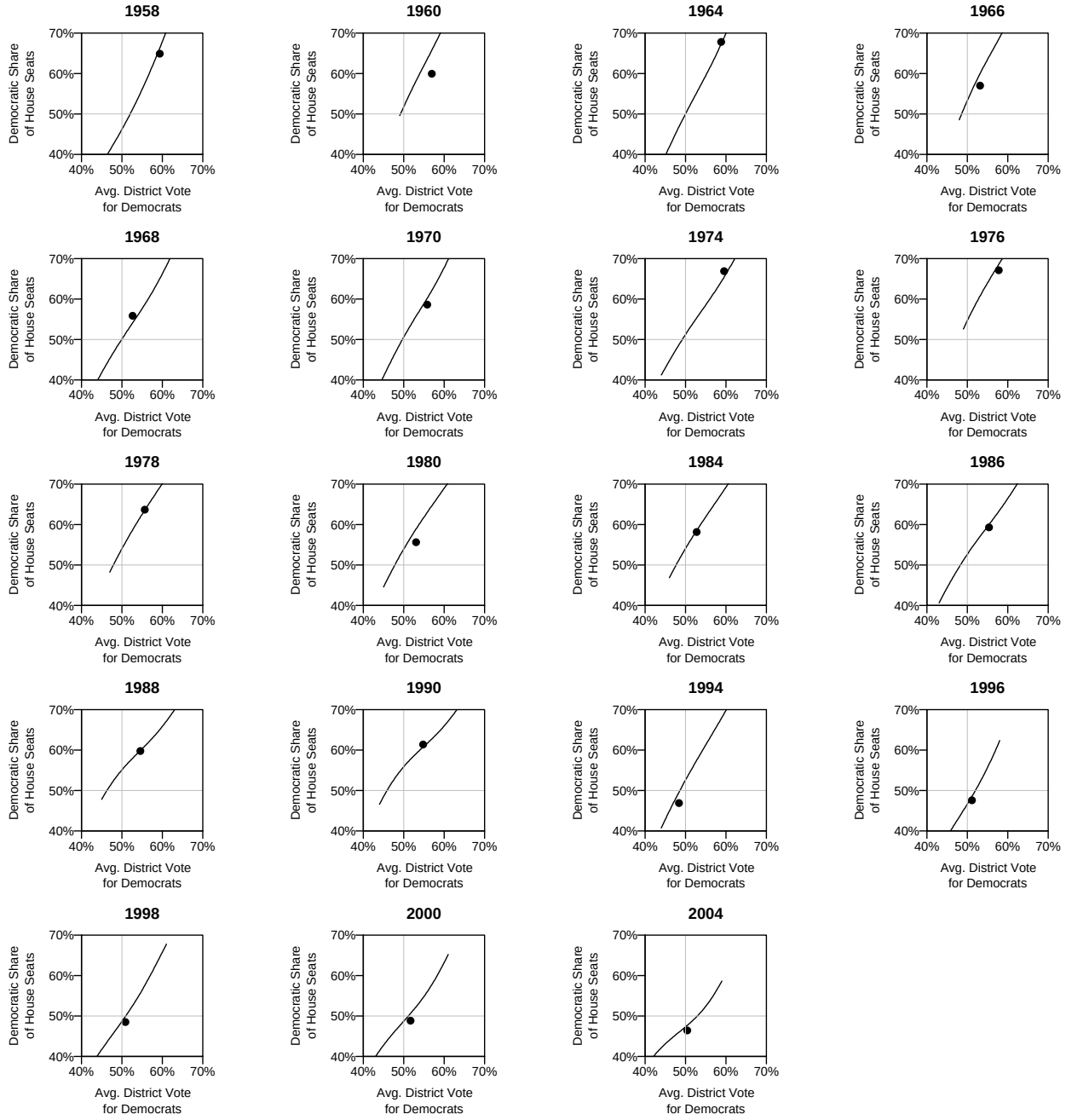

Figure 2: *Seats-votes curve prediction and actual election results, 1958–2004 (excluding redistricting years). For each year, the line depicts the estimated seats-votes curve, and the dot shows the actual election result. The light gray lines are references for 50% of the vote and 50% of the seats. As is well known (King and Gelman 1991), seats-votes curves have become less steep in recent decades—a given swing in votes will typically result in a lower swing in seats than before. As the plots show, the actual election typically conforms closely to the predicted seats-votes curve.*

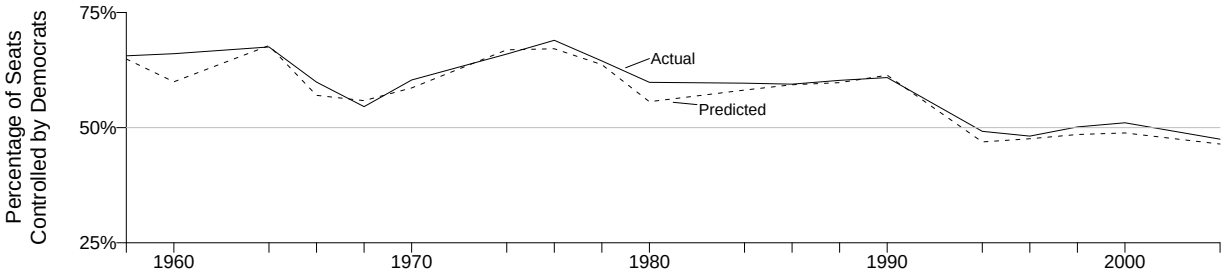

Figure 3: *Predicted and actual seats, 1958–2004. The solid line depicts the predicted percentage of seats controlled by Democrats, as obtained by our estimated seats-votes curve applied to the actual the average district vote for each year (excluding redistricting years). The dotted lines depicts the actual percentage of seats controlled by Democrats. The light gray line is a reference for 50% of the seats controlled by Democrats. The closeness of the two lines represents an accuracy in the prediction of seats given votes; we are not attempting to predict average district vote itself.*

root mean squared error of the predictive standard deviation of seats given votes is 0.01, or about 4 seats, implying that the rest of the predictive error is due to year-to-year variation unexplained by the model. Figure 3 plots the predicted Democratic percentage of the seats versus the actual percentage, showing how the two track closely. Thus, while the model is not perfect, we believe it demonstrates ample validity to serve as a useful tool for predicting the relationship between seats and votes for 2006.

3.3 Average District Vote and Total Vote

We have summarized election results using the average district vote—that is, the parties' share of the two-party vote, averaged over the 435 congressional districts: for the Democrats, this is $\frac{1}{n}\sum_{i=1}^{n} v_i$. This is the usual way that political scientists aggregate Congressional vote (King and Gelman 1991, Gelman and King 1994) and is a reasonable summary in that it is relatively immune to turnout variation and represents the national level of support for the party.

An alternative measure is the share of total vote that goes to the Democrats, which can be written as $\sum_{i=1}^{n} T_i v_i / \sum_{i=1}^{n} T_i$, where T_i is the turnout (the number of votes for the two major parties) in district i. Total vote differs from average district vote because of district-by-district variation in turnout, which is correlated with the parties' vote shares.

Figure 4 shows the time series of average district vote and total vote for the Democrats over the past sixty years. In the early period, the Democrats' average district vote was much higher than

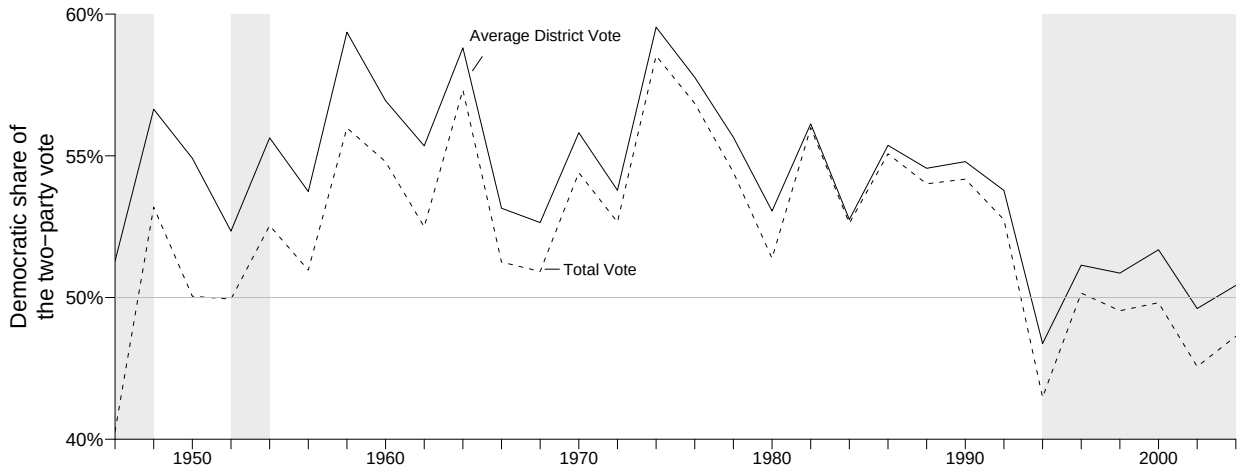

Figure 4: *Democrats' share of the two-party total vote and average district vote in House elections, 1946–2004. Shaded areas represent periods of Republican control of the House.*

their total vote share, which can be attributed to their large number of uncontested, low turnout districts in the South. In the past fifteen years, a gap has again opened up, this time of about three percentage points, because the Democrats do better in low-turnout districts (for example, in inner cities). Roughly speaking, if the Democrats need over 52% of the average district vote to have an even chance of winning the House, they need about 50% of the total vote to have this chance.

As noted above, we prefer average district vote as a summary measure of national support for the parties, and so that is what we have used in this paper. In any case, the Democrats are estimated to need a national vote swing of at least 2% to have an even chance of taking the House. Such a swing would represent, on average, slightly more than 52% of the vote within districts.

4 Results for the 2006 Election

We predict the seats-votes curve for 2006, along with the resulting probabilities of the Democrats taking control of the House given their average district vote, using the same model.[9] Figure 5 presents both the estimated seats-votes curve and probability versus average district vote. To reiterate, our model predicts that the Democrats need about 52% of the average district vote to have an even chance of winning a majority in the House, while 49% will give them a small chance, and 55% will nearly ensure it.

[9]Information on incumbency status and retiring members was obtained from `www.cq.com`. Information on uncontested candidates was obtained from Giroux (2006)—uncontested candidates are Democrats facing no Republican challenger and vice versa.

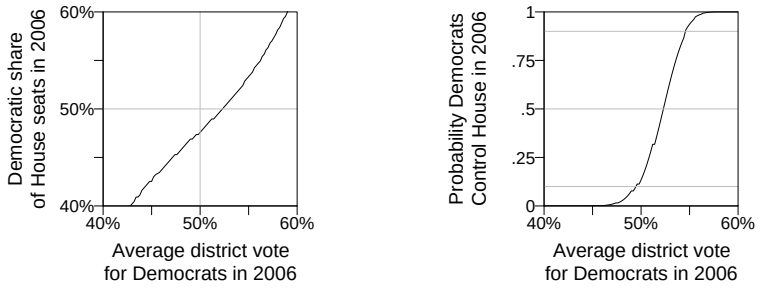

Figure 5: *(a) Predicted seats-votes curve for 2006; (b) predicted probability the Democrats will control the House given their average district vote. To have a reasonable chance of winning the House, the Democrats need to get between 50–55% of the vote.*

If no partisan bias existed in the seats-votes curve, then the Democrats would need only 50% of the vote to earn half of the seats. But as our predictions demonstrate, the electoral deck is stacked against the Democrats. As Erikson (2006) shows, since 1994 a clear partisan bias has existed in favor of the Republicans, a result that is consistent with the well-known tendency that the party in power tends to win a greater share of seats than votes (Tufte 1973).[10] Indeed, since the Republicans took control of the House in 1994, the Democrats have averaged more votes in four out of the five subsequent elections (2002 is the exception), yet have not managed to regain a majority of seats. Simply put, the Democrats are not receiving in seats what they are getting in vote shares.

How does this bias reveal itself in actual elections? Take 2004, for example. In races where Democratic candidates won, they averaged 69% of the vote; Republicans, by contrast, averaged only 65% in districts where they won, thereby "wasting" fewer votes. Many Democrats are in districts where they win overwhelmingly, while many Republicans are winning the close races—with the benefit of incumbency and, in some cases, favorable redistricting. This imbalance is depicted in Figure 6, which features histograms of the vote share in contested races won by the Democrats and Republicans, respectively. For Democratic winners, the distribution of the vote share is fairly symmetric, with about the same number of districts won with more than 70% of the vote than with less than 70%. By contrast, the distribution of vote share for Republican winners is skewed toward 50%; there are relatively few districts in which Republican candidates won with more than 70%. More concretely, the Republicans won 47 races with less than 60% of the vote; the Democrats only 28. Thus, Democrats in 2006 will need to win a greater share of these close races to have a chance

[10] From 1954 to 1994, a significant partisan bias existed in favor of the Democrats.

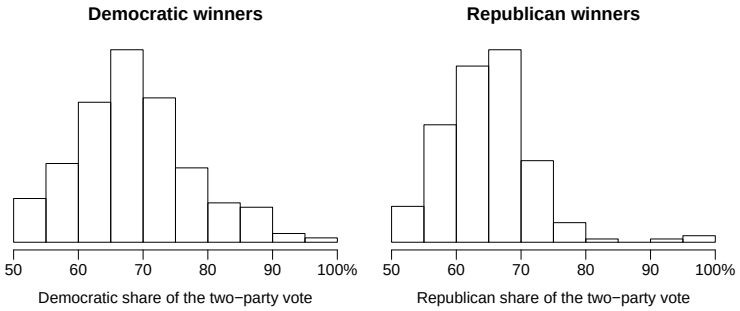

Figure 6: *Distribution of vote margins for Democrats and Republicans in winning races, 2004. The left histogram depicts the distribution of the Democratic share of the two-party vote in districts where Democrats won. The right histogram depicts the distribution of the Republican share of the two-party in districts won by Republicans. The distribution for the Democrats is symmetric around 68%, while the Republican wins are skewed towards 50%, illustrating that Republicans are winning more close elections than Democrats and are thus "wasting" fewer votes.*

of controlling in the House.

Incumbency will again favor the Republicans in 2006: of the 398 incumbents running for re-election, 211 of them are Republicans.[11] However, the Republicans' inherent advantage would be even larger than predicted if not for the large number of uncontested races in which the Democratic candidate is running unopposed—45, compared to only 10 unopposed races for the Republicans.[12]

5 Conclusion

We estimate that the Democrats need an average district vote of over 52%—an approximate 2% national swing compared to the last elections—to have an even chance of taking control of the House in 2006. Figure 1 shows the Democrats' share of the vote for the House of Representatives over the past few decades, along with what we estimate they need to have a 10%, 50%, and 90% chance of winning the crucial 218 seats in the House of Representatives. These calculations derive from a seats-votes curve estimated using a district-by-district extrapolation from the 2004 election, correcting for incumbency and election-to-election variation. (We validated the seats-votes curves on the previous half-century of House elections.) This statistical model is no substitute for experts' district-by-district analyses, but it does give a sense of how the parties' national vote shares might translate into seats, and thus how well they need to do nationally in order to have a chance at

[11] The Republicans enjoyed a nearly identical advantage in 2004.

[12] In 2004, Republicans ran unopposed in 37 races, Democrats 29.

winning the House.

References

Abramowitz, Alan I., Albert D. Cover and Helmut Norpoth. 1986. "The President's Party in Midterm Elections: Going from Bad to Worse." *American Journal of Political Science* 30:562–76.

Boston Review. 2006. "State of the Nation." *Boston Review* 31(5):6.

Brandt, Patrick T. and Thomas L. Brunell. 2006. "Forecasting the 2006 House Election." University of Texas, Dallas working paper.

Campbell, James E. 2006. "Forecasting the 2006 National Elections to the U.S. House of Representatives." *The Forum* 4(2).

Erikson, Robert E. 2006. "Party Control of the U.S. House of the Representatives." Columbia University Working Paper.

Erikson, Robert S. and Lee Sigelman. 1995. "Poll-Based Forecasts of Midterm Congressional Election Outcomes: Do the Pollsters Get it Right?" *Public Opinion Quarterly* 59:589–605.

Gelman, Andrew and Gary King. 1994. "A Unified Method of Evaluating Electoral Systems and Redistricting Plans." *American Journal of Political Science* 38(2):514–54.

Gelman, Andrew and Zaiyang Huang. in press. "Estimating Incumbency Advantage and its Variation, an Example of a Before-After Study." *Journal of the American Statistical Association*

Giroux, Greg. 2006. "For 55 House Incumbents, 'Safe' is an Understatement." *The New York Times* October 3.

Hibbs, Douglas A. 1982. "On the Demand for Economic Outcomes." *Journal of Politics* 44:426–62.

King, Gary and Andrew Gelman. 1991. "Systemic Consequences of Incumbency Advantage in U.S. House Elections." *American Journal of Political Science* 35(1):110–38.

Lewis-Beck, Michael S. and Tow W. Rice. 1984. "Forecasting U.S. House Elections." *Legislative Studies Quarterly* 9(3):475–86.

Marra, Robert F. and Charles W. Ostrom, Jr. 1989. "Explaining Seat Change in the U.S. House of Representatives." *American Journal of Political Science* 33(3):541–69.

Office of the Clerk, U.S. House of Representatives. 2006. "Vacanies and Successors." http://clerk.house.gov/members/electionInfo/vacancies.html. accessed 6 October 2006.

Oppenheimer, Bruce I., James Stimson and Richard W. Waterman. 1986. "Interpreting US Congressional Elections: The Exposure Thesis." *Legislative Studies Quarterly* 11(2).

Tufte, Edward R. 1973. "The Relationship Between Seats and Votes in Two-Party Systems." *American Political Science Review* 67(3):540–554.

Tufte, Edward R. 1975. "Determinents of the Outcomes of Midterm Congressional Elections." *American Political Science Review* 69:812–26.

Tufte, Edward R. 1978. *Political Control of the Economy.* Princeton, N.J.: Princeton University Press.