On Forecasting House Seats from Generic Congressional Polls— A Retrospective on 2006 (with a Glimpse at 2008)*

Joseph Bafumi
Dartmouth College (jbafumi@gmail.com)

Robert S. Erikson
Columbia University (rse14@columbia.edu)

Christopher Wlezien
Temple University (Wlezien@temple.edu)

* Prepared for presentation at the Annual Meeting of the American Association for Public Opinion Research, New Orleans, May, 2008.

On October 24, 2006, we completed an analysis to forecast the partisan division of seats in the US House of Representatives following the 2006 election. We launched our paper (Bafumi, Erikson, and Wlezien, 2008 [2006]) into cyberspace via the internet with the expectation that interested parties might read our forecast before the sell-by date of November 8, the day after the election.[1] The paper, titled "Forecasting House Seats from Generic Congressional Polls," received some attention at the time.[2]

In the forecast, we predicted a Democratic pickup of 32 seats (as an expectation). When taking into account the uncertainty around this expectation, we boldly proclaimed a Democratic takeover to be a virtual sure thing. This turned out to be a near bulls-eye as the Democrats did win 30 seats.

As an aspect of our bragging rights, we note that at the time of our forecast, the arbiters of the conventional wisdom were still debating whether the House would turn over, not whether the Democrats would exceed the necessary 15 seat gain. By Election Day, speculation of a Democratic blowout of a possible 30+ seat gain had become common. But two weeks in advance, this was not the case.

The success of our forecast breeds a follow-up question long after the electoral dust has settled. Put simply, was it that we were lucky or was it that we were good? The present paper addresses this question with a post-mortem on our forecast. We present our model and assumptions in detail, examining their validity versus other choices we could have made. Along the way, we look for possible change and stability in the "laws" that govern congressional elections, e.g., how is a "surge" election different? Also, we keep an eye out toward how we might predict the next congressional election, even far in advance.

We saw our forecasting task as consisting of two parts. First, we had to forecast the national vote for the House of Representatives, as a point estimate but with a variance around it. Second, we had to translate the projected vote outcome into a seat outcome (and win-probability). While both processes are interesting, the latter was the more challenging task by far.

Predicting the National Vote

We predict the vote from the results of the so-called generic polls—the questions pioneered by Gallup ascertaining whether the respondent plans to vote Republican or vote Democratic in the next congressional election, with no candidate names offered. The conventional wisdom is that generic polls are poor augers. Historically they

[1] The link to the original paper is: http://www.dartmouth.edu/~news/releases/2006/10/houseforecast.pdf.

[2] Andrew Gelman posted it on his well-read blog "Statistical Modeling, Causal Inference, and Social Science," along with comments. Three days later, Mark Blumenthal and Charles Franklin posted a condensed version of the original paper. By November 3, a still shorter version was posted on *The Huffington Post.*

overpredict the Democratic vote, and often by a wide mark. However, with the proper keys, the generic polls offer a remarkably accurate forecast for midterm elections, even well in advance.

-- Figure 1 about here --

Figure 1 displays the graph of the November midterm vote as a function of the vote using the pollsters' generic ballot question (coded as percent Democratic of Democratic and Republican votes). This particular graph is based on polls 240 to 300 days before the election, in other words, early in the midterm election year.

Note how poorly these early polls seem to predict. Almost without exception, they exaggerate the Democratic vote, as the observations systematically fall below (rather than above) the diagonal line representing perfect prediction. The 2006 election turns out to be no exception. If one plugs in the early generic poll Democratic showing for 2006 (5 or 55% of the two-party share), the prediction from the scatterplot would be near zero, as if the early polls gave no information about party control in 2006, especially when taking into account that the tie in the vote goes to the party holding the seats currently due to their incumbency advantage.

But are generic polls really useless information? If so, why would these particular polls perform so badly when other types of polls do not? Figure 2 transforms Figure 1 by simply taking into account the party of the president. Here, the diagonal lines represent parallel regression lines predicting the vote from the generic polls, with separate lines for each presidential party. As can be seen, the party holding the presidency is systematically penalized in the sense that the out-party gains strength between the time of the poll and Election Day. Our interpretation is that, over the campaign season, the electorate begins to focus on the upcoming election and increasingly takes the presidency into account with its decision, guided by a principle to vote for the opposition in order to offset presidential influence over policy. The result is as if the electorate increases its support for the out-party to balance or offset the president ideologically.[3]

-- Figure 2 about here --

For our purposes here as electoral forecasts, the interesting thing about Figure 2 is that it shows that the vote can be forecast from the generic polls. One merely must discount the size of the lead in the polls and adjust for the party in power.

Interestingly, had we made our forecast for 2006 from early 2006 data, it would have lead to virtually the same forecast as our final result. One did not need to foresee the further unraveling of the Iraq debacle or the Foley scandal to expect a Democratic win. Katrina,

[3] Many political scientists find this balancing argument implausible because they find it to be too cognitively-demanding of voters. However, notice that all it demands is awareness of the presidential party and some notion, perhaps not conscious or articulated, of the effect that the president deserves some checking by the opposition party. For more on ideological balancing in midterms in conjunction with the generic polls, see Bafumi, Erikson, and Wlezien, 2006.

the Harriet Miers nomination fiasco, and a host of smaller incidents accumulated sufficiently to auger Democratic success. The early generic polls pointed to a Democratic win. And in past midterms, the generic vote is stable until the election, except for moving further toward the out party as the campaign progresses.

For our vote predictions, we modeled the vote in past midterms as a function of the generic polls within the final month of the campaign plus the presidential party. Figure 3 presents the picture. By the end of the campaign, the presidential party variable is now largely factored in by the electorate, but still the tie goes to the out-party.

-- Figure 3 about here --

Note however that the experience of 2002 makes clear that there are exceptions to every rule. In that election, the midterm arrow of fortune turned in favor of the in-party Republicans, thanks presumably to 9/11 and the electorate's still positive evaluation of President Bush's performance. In forecasting 2006 from the past, could 2002 have posed an early warning sign that rules of thumb that worked in the past no longer are in force?

Our vote equation was:

$$\text{Dem Vote Share} = \underset{(0.63)}{24.38} + \underset{(0.10)}{0.51} * \text{Dem Poll Share} - \underset{(0.52)}{1.07} * \text{Presidential Party} \qquad \text{(Eq. 1)}$$

Adjusted R-squared = 0.75; Root MSE = 1.90,

where Presidential Party takes the value "1" under a Democratic President and "-1" under a Republican.

For the period between October 8 and October 23, we scored the average result from several generic ballot bolls as 57.7 percent Democratic of the two-party vote.[4] Translating via equation 1, we obtain a forecast of 55.0 percent Democratic in the actual vote, with a confidence interval from 51.1 to 58.7 percent Democratic. With the official 2004 tally at 48.6 percent Democratic, this represents a swing of 6.4 percentage points. So how close was this to the actual result?

The Clerk of the House of Representatives has produced the official count of the House of Representatives vote by party in 2006, and the national two-party vote is 54.1 percent Democratic, 45.9 percent Republican division. This represents a swing of 5.5 percentage

[4] For the period from October 8 through October 23, PollingReport.com reported results from five polls using the generic ballot, by CNN (2), ABC/Washington Post, Fox/Opinion Dynamics, and Newsweek. The average Democratic two-party share in these polls is 57.7%. We used the results among "likely voters" whenever possible. When only results for registered voters were reported, we adjusted the reported vote to reflect the expected "likely vote" result. Specifically, we subtracted 1.53 from the registered voter poll. This adjustment was derived from a regression predicting the generic polls during presidential election years by population and year indicators and using the coefficient of the population indicators. In this way, we avoided the possibility that the year-to-year variation in the selected universe would affect the result. See Bafumi et al., 2006.

points from 2004. For our purposes, we note that the vote is within 0.9 percent of our generic poll based forecast, and well within the margin of error.

We now exit from the discussion of forecasting the vote and turn to the hard question of predicting the seats from the votes. We had predicted a vote division that would be 55 percent Democratic producing a 32 seat Democratic gain. The true vote was 54.1 percent Democratic with a 30 seat gain.

Predicting the Seats from Votes

This section describes our procedure for estimating the seat division from the projected vote division from the previous section. As a template, we draw on information from the prior election, in 2004. The general assumption is that the same rules that govern the vote outcome for individual races in 2004 apply to 2006—except that the national vote will shift by some amount from 2004 to 2006.

We estimate two equations for the district level vote in 2004, one for incumbent races and one for open seats. In our incumbent equation, we predict the Democratic vote from the lagged (2002) Democratic vote plus a freshman variable. The freshman variable is a dummy variable coded -1 if a Republican freshman, 0 if not a freshman, and +1 if a Democratic freshman. This freshman dummy is necessary because of the sophomore surge, by which newly-minted incumbents increase their vote between their first election (as a nonincumbent) and their second (as a freshman seeking sophomore status).[5]

The incumbent equation is as follows:

$$\%\,\mathit{Dem}(2004) = \underset{(0.81)}{4.41} + \underset{(0.02)}{0.95} * \%\mathit{Dem}(2002) + \underset{(0.91)}{6.58} * \mathit{Frosh}, \qquad \text{(Eq. 2)}$$

where *%Dem*(2004) = the district's percent Democratic for the House in 2004,
%Dem(2002) = the district's percent Democratic for the House in 2002,
and *Frosh* = 1 if a Dem. Incumbent, -1 if a Rep. incumbent, otherwise 0.

Adjusted $R^2 = .940$ *RMSE* = 4.50 *N* = 271.

In our equation for open seats, we regress the 2004 vote not on the 2002 vote but rather the district's vote percentage for John Kerry, the Democratic candidate, in the 2004 presidential election. Presidential voting is an excellent predictor of Democratic partisanship across districts.

$$\%\mathit{Dem}(2004) = \underset{(5.70)}{5.99} + \underset{(0.12)}{0.89} * \%\text{Kerry}(2004), \qquad \text{(Eq. 3)}$$

[5] In the appendix to the original forecast, the incumbent seat equation is slightly misreported, as the estimated equation inadvertently includes Texas districts that had been redistricted between 2002 and 2004. The equation reported here as Equation 2 is in fact the one used to produce the forecasts in 2006. There is little difference except for the standard error.

where *%Dem*(2004) = the district's percent Democratic for the House in 2004.

%*Kerry*(2004) = the Kerry percent of the district's two-party presidential vote in 2004.

Adjusted $R^2 = 0.658$ $RMSE = 7.52$ $N = 29$.

We used these 2004 equations as input for simulations of the 2006 vote. Our assumptions were that the incumbent and open-seat equations (including their RMSE or estimated standard deviation of the error) would carry over from 2004 to 2006 except for the crucial matter of the equation constants (or intercepts), which vary as a function of the vote swing. An additional necessary assumption is that the swing of the national vote percentage (based on all Democratic and all Republican votes nationally) would be identical to the mean swing of the vote in districts with a contested vote in both elections (and no redistricting 2002-2004).

For any hypothetical vote swing, the open-seat simulation equation is readily adjusted by moving the intercept the amount of the vote swing. Borrowing from equation 3,

$$\textit{Simulated \%Dem}(2006) = 5.99 + 0.89 * \%\text{Kerry}(2004) + \text{SWING} + e + u_{\text{open}}. \qquad \text{(Eq. 4)}$$

For incumbent races, the intercept needs an adjustment based on some algebra and the frequencies of the two types of races. This is so the net vote swing across the two types of races equals the intended swing of the simulation:[6]

$$\textit{Simulated \% Dem}(2006) = 2.63 + 0.95 * \%\textit{Dem}(2004) + 6.58 * \textit{Frosh} + \text{SWING} + e + u_{\text{inc}} \qquad \text{(Eq. 5)}$$

The two sets of simulations include two sources of random error. The *e* term represents the RMSE of Equation 1, the vote equation. The *u* terms represent the error in Equations 2 and 3 predicting the district vote from its independent variables. The standard deviations of these terms are borrowed directly from the RMSEs of equations 1, 2, and 3.[7]

[6] For any fixed net vote swing across all races, the size of the vote shift in open seats constrains the mean vote shift in incumbent races. The 2004 open seat equation, based on the 2004 presidential vote, provides a baseline equation. With no net swing 2004-2006, equation 3 would apply unmodified for 2006. For incumbent races, some algebra must be applied. For incumbent races, the constant equals the equation 2 intercept (for 2004 incumbent races) *minus 1.72* plus the projected vote swing. The 1.72-point adjustment is necessary so that the projected mean district swing across all districts equals the targeted vote swing.. We start with the identity equation, N(I)*Incumbent Seat Vote Swing + N(O)* Open Seat Vote Swing = N(Total)*Vote Swing where N(I) is the number of incumbent seat cases and N(O) is the number of open seat cases. Given our projected swing of the national vote, we observe the resultant Open Seat Vote Swing and work the algebra to solve for the "unknown"—the Incumbent Seat Vote Swing. Given our methodology, all seats that were uncontested in 2004 must be assigned to the 2004 winner. Bernie Sanders's former seat as an Independent is assigned to the Democratic candidate.

[7] As noted two footnotes earlier, the incumbent race standard deviation in the original (on-line) version of the forecast is a misreport. The RMSE reported here (4.50) was used in practice.

The simulations were constructed as follows. For each possible generic ballot integer value from 50% Democratic to 60% Democratic, we computed 1,000 simulations of the 435 seat outcomes. Each simulation includes:

(a) a random draw from the density of the possible vote outcomes from our generic poll regression equation, based on the predict from the generic poll plus forecast error; and

(b) a set of 435 random draws of district level predictions conditional on the 2006 national shock (from [a]) plus district-level characteristics and shocks based on a regression model from the 2004 election.[8]

Evaluation

As we saw, our vote prediction based on the generic polls was on target, missing by only 0.9 percent. The test regarding the translation of votes to seats is more demanding. That test is, given that hindsight tells us the national vote, how accurate is our modeling of the seat division?

First, we must consider that our seat simulations are based on the assumption that the national swing of the total vote and the mean district-level vote swing are identical. When measured for all non-redistricted seats that were contested in 2004 and 2006, the mean vote swing is only 4.5, one point less than the 5.5 swing of the national vote. The discrepancy reflects the shifting balance of uncontested seats from Republican seats in 2004 to Democratic seats in the less GOP-friendly 2006 election.[9] In effect, by Democrats contesting in 2006 in Republican districts where they yielded in 2004 and the Republicans yielding in Democratic districts in 2006 where they had contested in 2004, the national vote moves an extra point just due to shifting concessions in advance of national party tides. For our purposes here, the relevant swing is 4.5 as an average district swing where the House race was contested in both elections. This is because our simulations of the 2006 district vote in our forecast were based on the mean district swing, not the national swing.[10]

Recall that our October forecast was a 32 seat gain from a 6.4 percent vote swing. If we record the actual swing as 5.5 instead of 6.4, overshooting the actual 30 seat swing by 2 might look pretty good. But the interesting test is, given a known 4.5 percent mean district vote swing, does the pre-election model predict something close to the actual thirty seat gain?

[8] Actually, the number of random draws is 331. The remaining 114 districts are assigned automatically to the current party holding the seat due to unopposed candidacies in 2004 or 2006 redistricting or to the uncontested 2006 candidate.

[9] Although the redistricted seats are in conservative Georgia and Texas, this does not appear to be a factor.

[10] We have also assembled a parallel analysis based on the 5.5 percent vote swing.

Using our hindsight about the vote, we would like to estimate what our model predicts (retrospectively) about the seat division, given the 4.5 percent vote swing. We ask, given this hindsight, what is the result of the simulations when the mean swing is fixed at 4.5 percent? And we can examine the internal parts. Are the 2006 equations equivalent to those from 2004's as we assume, and if not, what difference does it make? We also can ask whether the equations contain a well-behaved error structure with a homoscedastic normal distributions and if deviations distort predictions.

To find out, we ran 1000 simulations of the 2006 seat distribution based on a known 4.5 percent vote swing. From these simulations, the mean number of Democratic seats is 224.2 for a mean projected seat swing of 21 seats. This undershoots by 9 seats the correct answer of 233 Democratic seats. Observing that the standard deviation of our 1000 simulated outcomes is 3.69 seats, our point prediction is outside the conventional margin of error. To put meaning to the idea, if the vote divisions had been determined by lottery, using the error terms from Equations 4 and 5 for chance, a discrepancy from the mean expectation as large as we obtained (9 seats) could have occurred less than one time in 20. Somehow, the Democrats got more bang for their vote than our modeling predicts they could, given that they gained only 4.5 percentage points on average in 2006.

This leads us to investigate the internal parts to our model. How good were the simulation equations? In fact, our simulation equations based on modeling the district vote in 2004 almost perfectly match the vote equations directly modeling 2006. This presents some satisfaction, but also a puzzle.

For incumbent races, the actual 2006 equation was:

$$\underset{}{\%\,Dem(2006)} = \underset{(0.83)}{9.55}+ \underset{(0.02)}{0.90} * \%Dem(2004) + \underset{(0.86)}{5.71} * Frosh \qquad \text{(Eq. 6)}$$

Adjusted $R^2 = .925$ $RMSE = 4.66$ $N = 296$.

From equation 5, the simulation of the 2006 vote in incumbent races, based on our October model and plugging in a 4.5 percent swing is:

$$\%\,Dem(2004) = 7.13+ 0.95 * \%Dem(2002) + 6.58 * Frosh, \qquad \text{(Eq. 7)}$$

with 4.5 as the standard deviation of the error, compared to the observed 4.66 RMSE. For open seat races, the actual 2006 equation was:

$$\%Dem(2006) = \underset{(4.08)}{12.49} + \underset{(0.08)}{0.89} * \%\text{Kerry}(2004) \qquad \text{(Eq. 8)}$$

Adjusted $R^2 = 0.799$ $RMSE = 6.20$ $N = 30$.

From equation 4, the simulation of the 2006 vote in incumbent races, based on our October model and plugging in a 4.5 percent swing is:

$$Simulated\ \%Dem(2006) = 10.49 + 0.89 * \%\text{Kerry}(2004), \qquad \text{(Eq.9)}$$

with 7.52 as the standard deviation of the error, compared to the observed RMSE of 6.20. The presidential vote coefficient was quite stable. The unexplained variance in 2006 open seats was less than simulated, as if open seat outcomes were more tied to district partisanship in 2006 than 2004.

Our modeling is very successful in accounting for the net shift in the vote. That is, we target the observed 4.50 mean swing in twice contested races and obtain a simulated mean swing of 4.42, within 0.1 percentage points of the target. But we should also look at the swing separately for incumbent races and open seats. In 2006 incumbent races contested in both 2004 and 2006, the mean vote shifted from 50.1 percent Democratic to 54.1 percent Democratic, for a 4 point swing. The predicted swing from the model was 4.3 percent, so we were off by only a trivial two tenths of one percent.

With the swing in incumbent seats slightly less than the model prediction, it follows that the swing in the smaller number of open seats must have been greater than predicted. This is the case. In 2006 open seats contested in both 2004 and 2006, the mean vote shifted from 44.3 to 53.8 percent Democratic, for a whopping 9.5 percent shift. This was 2.5 percentage points more than the model predicts.

To illustrate our seeming success with incumbent races, Figure 4 displays four panels representing the 2006 vote as a function of the lagged vote in Republican-held incumbent races. Each highlights the Republican losses and displays freshman outcomes as distinct from those veteran Republican incumbents. Three of the four panels are simulations based on our model and the assumption of a net 4.5 percent vote swing. The fourth is the actual result. Can one tell the difference? Can one identify the true result as different?

-- Figure 4 about here --

The four scatterplots are similar in form, varying only in minor details. The actual result—as opposed to a simulation—is the bottom right panel. Whereas the three simulations show Republican losses of 15, 13, and 17, reality was less kind to the Republicans 21 incumbent losses. Thus, it can be said that the Democrats were lucky to do as well at sacking incumbents as they did, considering the limited 4.5 percent overall gain.[11]

Figure 5 displays a similar set of scatterplots for Republican-held open seats. Again, three are simulations and one is the actual data. They appear similar in form, with incumbent losses of 8, 4, 8 and 7. The panel with the actual data is the fourth.[12] With a loss rate similar to those of the simulations, the open-seat districts are not responsible for our model's underestimation of the seat swing given the 4.5 percent mean district vote

[11] The actual vote loss for Republican incumbents as a group was 5.0 points, which is also the approximate net loss in the three simulations shown.

[12] The reader who is counting will notice totals of 21 incumbent seats plus 7 open seats switching from Democratic to Republican. That is two short of the actual 30 switches. One of the extra switches involved a seat that had gone uncontested in 2004. The other was a redistricted seat.

swing. Although the open-seat vote swing was higher than our model anticipated, its impact was offset by the tighter fit of the open-seat vote around the prediction from the presidential vote than anticipated.

--Figure 5 about here—

So what is our final accounting of the 9 seat undercount, assuming a 4.5 point vote swing? Based on 1000 simulations of contested incumbent races, the Democrats won a surplus of 7 seats beyond what our simulations say they should have won, although it is not clear why. The Democrats won the correct number in opens seats. The other two were miscounted as follows. One was PA 7 where a Democrat defeated previously unopposed but newly scandal-plagued Curt Weldon. (The fine print shows that we conceded 2006 races to incumbents who were uncontested in 2004). The other was TX 23 where a Democrat upset Republican Henry Bonilla in a redistricted seat.

One further check is to see how our simulation model predicts if we not only know the mean vote swing in advance but also the exact form of the 2006 equations. For this exercise we plug in equations 6 and 8 based on the actual 2006 data. For 1000 simulations, we obtain a mean Democratic holding of 227.8 seats with a standard deviation of 3.56. Since the 233 seat outcome is well within 2 standard deviation of the mean estimate, we have reached a stopping point. Although the observed equations with their assumption of heteroscadiscity produce an expected undercount of 5 Democratic seats, this departure is within the margin of error. The Democrats were lucky in two respects. First, if equations 6 and 8 rather than our simulations represent the data generating process for determining vote outcomes with a stochastic term. Second, given equations 6 and 8 as the data generating processes, the Democrats were somewhat luckier in drawing winning outcomes than the laws of chance predict.

Uniform Swing?

Readers might ask why we went to all that trouble in our modeling exercise, when all we needed to do was extrapolate the 2006 seat division from the 2004 distribution of seats by assuming a uniform swing of the vote across districts. Assuming a uniform swing of the vote from one election to the next is a sophisticated—but flawed—way of predicting seat swings from vote swings. For a swing of *x* percent, all one needs to do is calculate the number of seats the advantaged party lost by *x* percent in the previous election. While the assumption of a strictly uniform swing is not realistic, one might assume the errors in the assumption are benignly random in some fashion. As we will see, that is not the case.

-- Figure 6 about here --

Figure 6 applies the logic of the uniform swing method to the 2006 election. It shows the cumulative distribution of seats as a function of the size of the 2004 Democratic vote. We immediately see the reason for the pre-election conventional wisdom that saw a high hurdle for the Democrats. The slope is extremely flat within the competitive 50 percent range as there were few seats at risk of changing partisan hands in 2004. The 2004 landscape was an exaggerated version of the classic bimodal distribution with Mayhew's "vanishing marginals" in the middle of the range.

To control the House, the Democrats needed the 218th least Democratic district to contain more Democratic than Republican votes. As Figure 6 shows, this 218th least Democratic district in 2004 was only 44.7 percent Democratic. Based on a uniform swing, the Democrats would have needed a massive 5.3 point gain just to achieve House control. Recognition of this daunting hurdle was one factor that kept informed observers hesitant to project Democratic control. With the 233rd least Democratic seat at 42.7 percent Democratic, by the uniform swing rule, the Democrats could have achieved their result of a thirty-seat gain only by means of an unprecedented 7.3 percentage point swing. Meanwhile, a vote swing as lowly as the 4.5 mean observed swing would have shifted a mere 9 seats in the Democrats' direction! With inferences such as these available, it was no wonder that sophisticated observers hesitated about envisioning the possibility of something like a Democratic landslide in 2006.

Where uniform swing fails is that it assumes a static set of campaign dynamics at each percentile range of the district vote. What our 2004-2006 uniform swing exercise is useful for is for simulating counterfactual 2004 (not 2006) scenarios. It reveals that if the Democrats received an extra unexpected across-the-board boost in votes in 2004, this stealth gain would have had little impact on seats. There would have had to be an extra unanticipated vote swing of at least 5.3 percent for the Democrats to win the 15 new seats necessary to take control. In 2006, the 5.3 percentage point gain would yield more than 15 new seats because the dynamics were different within the competitive range. Anticipating a Democratic gain, vulnerable Republicans retired at greater than usual rates, adding to the Democratic yield.[13] Moreover, Republican incumbents who barely won in 2004 could face stronger challenges in anticipation of a Democratic surge in 2004.

Figure 7 shows this by overlaying the uniform swing predictions on the actual cumulative seat distribution from 2006. This 2006 cumulative distribution presumably does represent what might have happened for various possible swings of the vote. We see that in retrospect, a surprisingly low 2.2 percentage point vote swing would have been sufficient to turn the House to Democratic control. Figure 7 also suggests that the Democrats actually gain 2 seats even if the vote swing were zero. This quite plausibly is due to strategic retirement with Republican retirements outnumbering Democratic retirements, giving them an initial boost in anticipation of the vote swing.

We see the contrast between counterfactual projections from the uniform swing of the cumulative 2004 vote, known in advance of the election and the actual 2006 cumulative vote, observed after the fact. We can also calculate the projected cumulative seats based on our simulations. That is, given a vote swing of a certain magnitude, we can project the

[13] Republican retirements outnumbered those on the Democratic side 21 to 9 in 2006. On the other hand, there is no visible evidence that retirement decisions among Republican incumbents were related to the degree of electoral vulnerability.

expected seat swing based on our simulations. We have seen that we underpredict at 4.5 percent swing (21 instead of 30 seats). We overpredict at zero swing (a net gain of 6 instead of 2 seats). Thus, while our simulations capture the Democratic gain from Republican strategic retirements and certainly outperform uniform swing, they err by not fully capturing the steepness of the slope representing seats as a function of votes.[14] *The "swing ratio" of 2006 turns out to be greater than our simple model could predict.*

-- Figure 7 about here --

Incumbency

An observer might ask why we did not include an incumbency dummy variable in our model. Consider the graph of Figure 8. If we add an incumbency variable to the 2004 incumbent equation, the slope for the lagged vote declines from 0.95 to 0.78 and the incumbency coefficient is 3.55 with a highly significant t-value of over 5. Adding this variable reduces the RMSE from 4.50 to 4.28.

-- Figure 8 about here --

We chose not to add the incumbency variable on the grounds that it misrepresents the likely outcome in close races. Where the Republican incumbent almost lost in 2004, we saw the Democrats putting up a stronger struggle in 2006. Where the Democrat incumbent almost lost in 2004, we saw the Republicans as conceding in 2006.

This turned out to be the correct decision. Figure 9 shows the slope of the 2006 vote on 2004 vote in the range of 40 to 60 percent Democratic. There is no visible regression discontinuity at 50 percent. If we add an incumbent party variable, it is an insignificant 0.60.

-- Figure 9 about here --

We were fortunate to have excluded the incumbency dummy in the incumbent race equation. If we had mistakenly included the incumbency dummy for our simulations (and assumed a 4.5 percent swing) we would have underestimated the seat shift by a further 6 seats, projecting a seat swing only half the size of the thirty seats the Democrats gained. The exclusion evidently made both theoretical and empirical sense.[15]

[14] With a fixed 1.5 percentage point Democratic vote gain, we simulate a gain of 10 seats; with a fixed 2.5 gain we simulate a 15 seat gain; not until the fixed seat swing exceeds 6 points do our simulations project a 30 seat gain. A key to these simulations is the word "fixed." Recall that our October forecast of a 6.4 point seat swing projected a 32 seat gain taking into account the uncertainty of the 6.4 estimate.

[15] It is tempting to see the presence and then absence of a sizeable incumbency coefficient as evidence about the incumbency advantage. The causes area more complicated than that. In 2006 the incumbency advantage survives in the form of a sophomore surge. That is, incumbents gained more than their veteran counterparts of their parties, as usual in 2006.

Lucky or Good?

At the outset we asked whether our accurate forecast of House seats in 2006 was lucky or good. Clearly we were lucky to be able to claim credit for nearly nailing the 30 seat Democratic gain on the nose. While we overestimated the vote swing by a slight and readily forgivable 0.9 points and overestimated the seat swing by 2, we erred in assuming that the mean vote swing and the national vote swing would be identical. Combined with our mild overshooting of the vote swing, this mistake was "lucky" in the sense that otherwise we could have underestimated the seat swing by a greater amount than our 2 seat overestimate. Put another way, we overpredicted the mean vote swing but underpredicted the 2006 swing ratio of seats gained per vote. We were lucky that these were offsetting rather than reinforcing errors.

Forecasting 2008

What about the 2008 election? We know that it certainly is possible to forecast the vote from the late generic polls using similar technology as we employed for 2006. The key difference is that the vote equation used to predict the 2008 congressional vote would need to be based on the presidential-year generic polls. Pollsters give less attention to forecasting the vote for Congress when there is an election for president also going on. The historic data on presidential-year generic polls is sparser (but see Erikson and Sigelman, 1996). To avoid serious missing data problems, we must lump the presidential-year generic polls in a broader time frame. As the Fall election approaches, we must use a 1-60 day frame rather than the 1-30 days we used for making our midterm predictions for 2006.

At this juncture (May 6, 2008), we can make a crude vote forecast based on past generic polls during the broad interval 61-180 days before the election. The prediction equation is:

$$\text{Dem Vote Share} = \underset{(0.54)}{0.95} + \underset{(0.07)}{0.31}\ \text{Dem Poll Share} \qquad \text{(Eq. 10)}$$

$$\text{Adjusted } R^2 = .54 \quad \text{RMSE} = 1.71 \quad N=16$$

The party of the president is not a significant predictor of the congressional vote share in presidential election years and therefore is omitted from our equation 10 prediction.

The two recent available polls (CBS/NYT and NBC/WSJ) show a 60-40 split among registered voters. Making our formulaic projection of registered voter polls to likely voter polls (in this case a 1.5 point drop in the Democratic vote) yields 58.5 as the estimate to plug into the right-hand side of the equation above.

The result is an estimate of 53.6 percent Democratic in the vote, only half a point below the Democrats' share in 2006. The estimate is admittedly crude, but also seemingly strong enough to assure Democratic control. The details about converting votes to seats must await the final tallies of retirements and be based on later polling. With reports of massive retirements by Republican incumbents from competitive districts, a national vote division similar to what we observed in 2006 should propel a further Democratic gain in seats.

References

Bafumi, Joseph, Robert S. Erikson and Christopher Wlezien. 2006. "Ideological Balancing, Generic Polls and Midterm Congressional Elections." Institute for Public Affairs Working Paper, Temple University. [Http://www.temple.edu/ipa/workingPapers/]

Bafumi, Joseph, Robert S. Erikson and Christopher Wlezien. 2008 [2006]. "Forecasting House Seats from Generic Congressional Polls." In Wendy Alvey and Fritz Scheuren (eds.), *Elections and Exit Polling*. New York: Wiley and Sons. [Http://www.dartmouth.edu/~news/releases/2006/10/houseforecast.pdf]

Erikson, Robert S. and Lee Sigelman. 1996. "Poll-Based Forecasts of the House Vote in Presidential Election Years." *American Politics Research* 24:520-531.

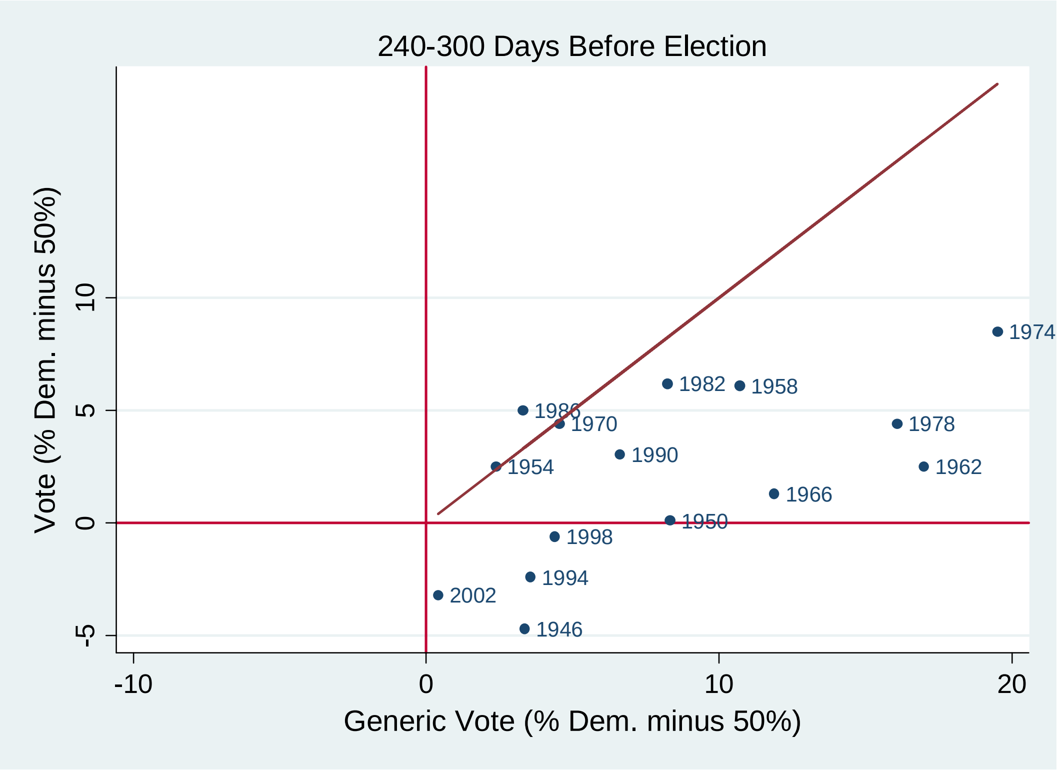

Figure 1. Midterm Vote by Vote in Generic Ballot, 240-300 Days before the Election, 1946-2002

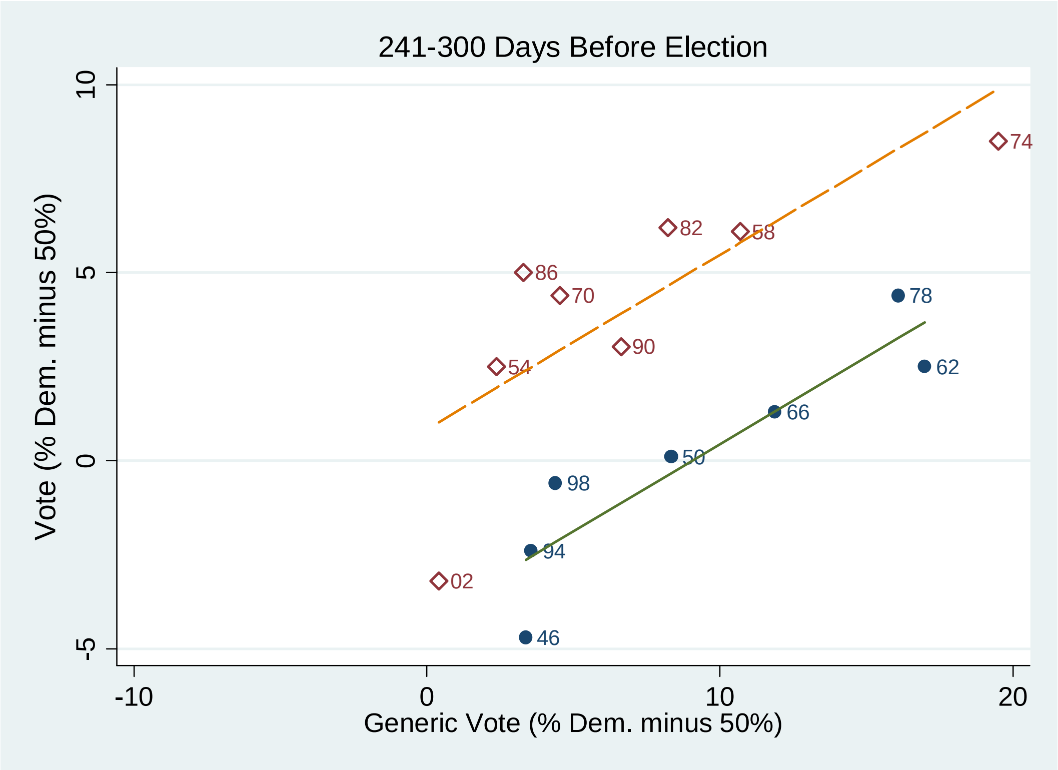

Figure 2. Midterm Vote by Vote in Generic Ballot, 240-300 Days before the Election, 1946-2002, with Party Control of Presidency Highlighted. Election years with Democratic presidents are blue; those with Republican presidents are red. The parallel lines represent the results of a regression equation predicting the vote from the generic vote plus a presidential party dummy.

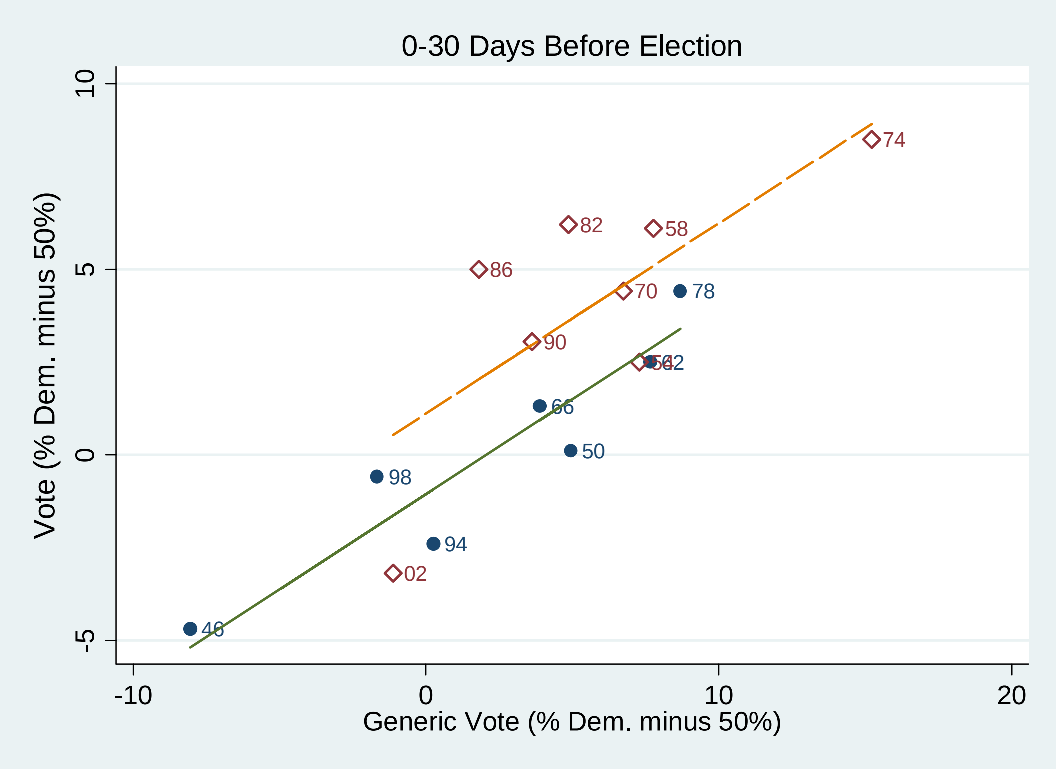

Figure 3. Midterm Vote by Vote in Generic Ballot, 1-30 Days before the Election, 1946-2002, with Party Control of Presidency Highlighted. Election years with Democratic presidents are blue; those with Republican presidents are red. The parallel lines represent the results of a regression equation predicting the vote from the generic vote plus a presidential party dummy.

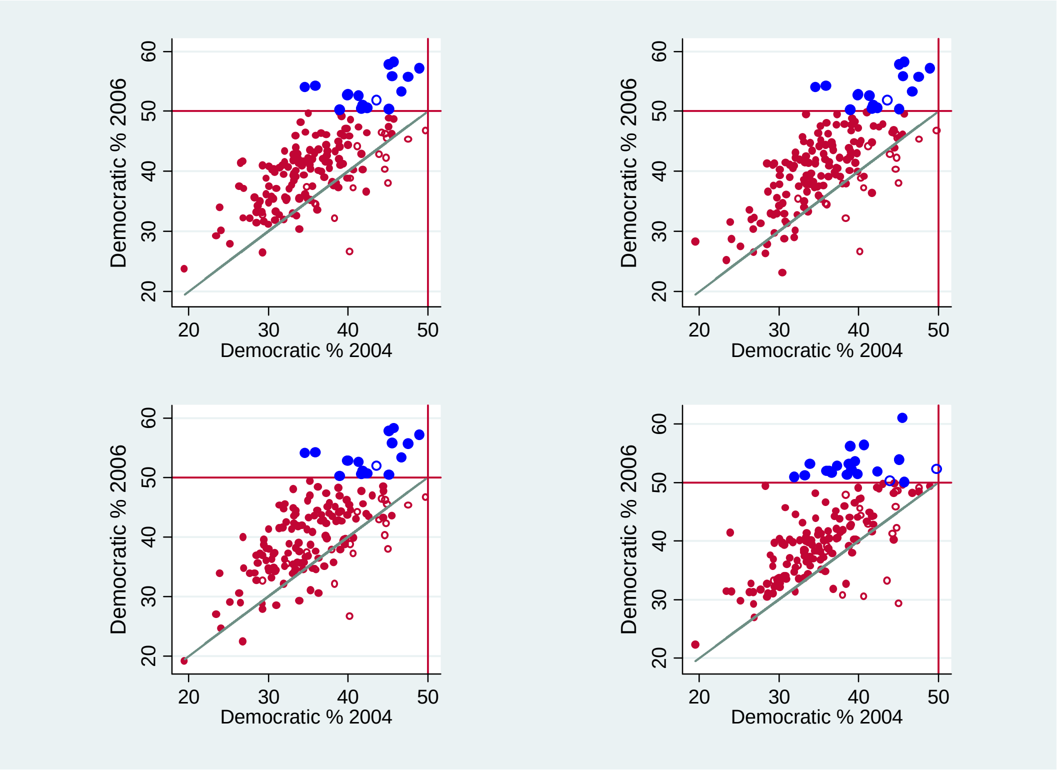

Figure 4. Four presentations of the 2006 Democratic Vote in Races with Republican Incumbents Running. Freshman races are highlighted by hollow dots. Three are simulations using a data generating function based on a 2004 equation and a 4.5 vote swing. The lower right panel represents the actual data.

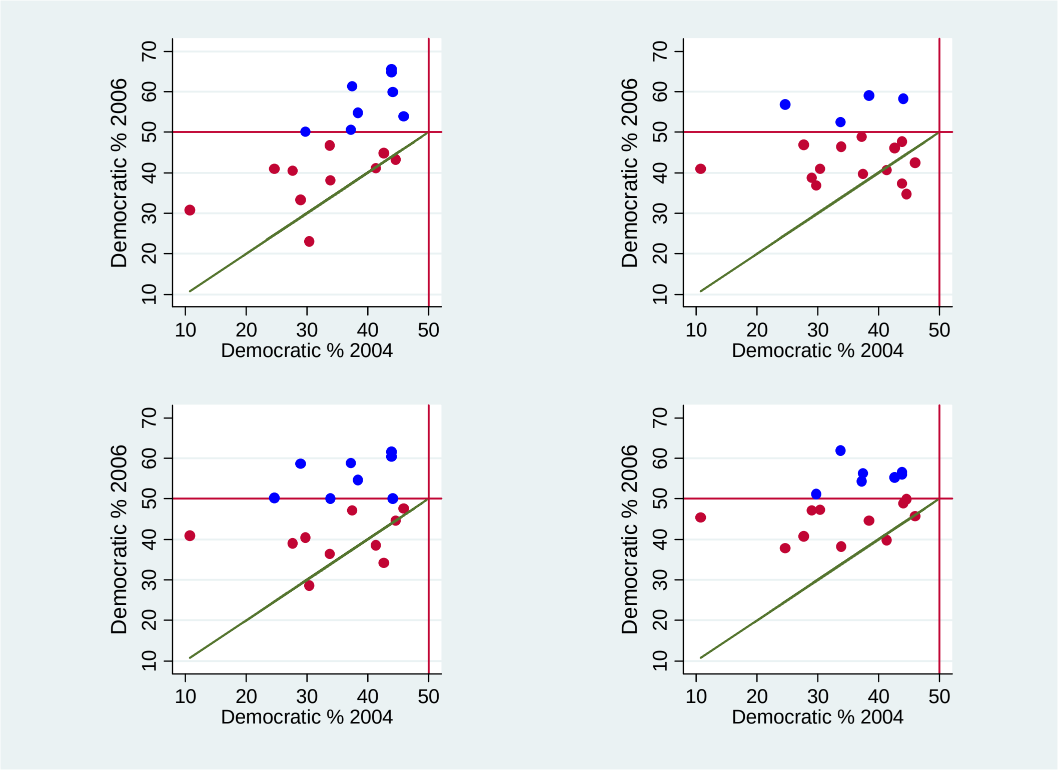

Figure 5. Four presentations of the 2006 Democratic Vote in Open Seats. Three are simulations using a data generating function based on a 2004 equation and a 4.5 vote swing. The lower right panel represents the actual data.

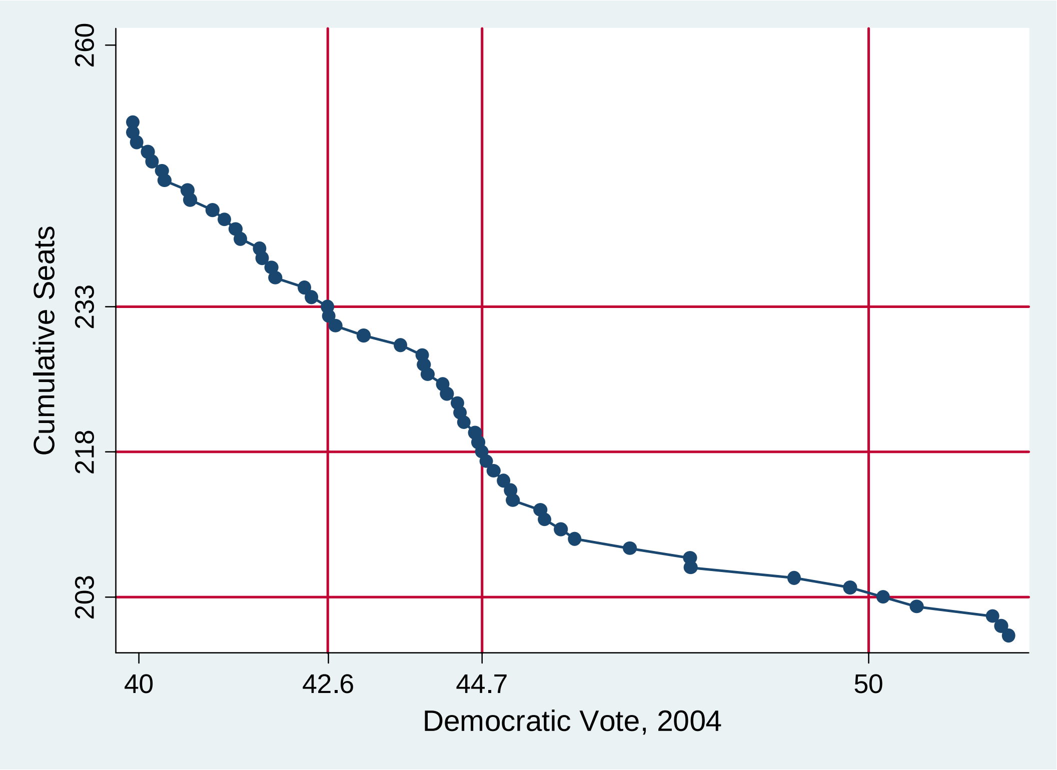

Figure 6. Extrapolating 2006 from the 2004 distribution of the vote. The projection is a 5.3 swing of the vote for a Democratic win and a 7.4 swing to achieve a 30 seat gain.

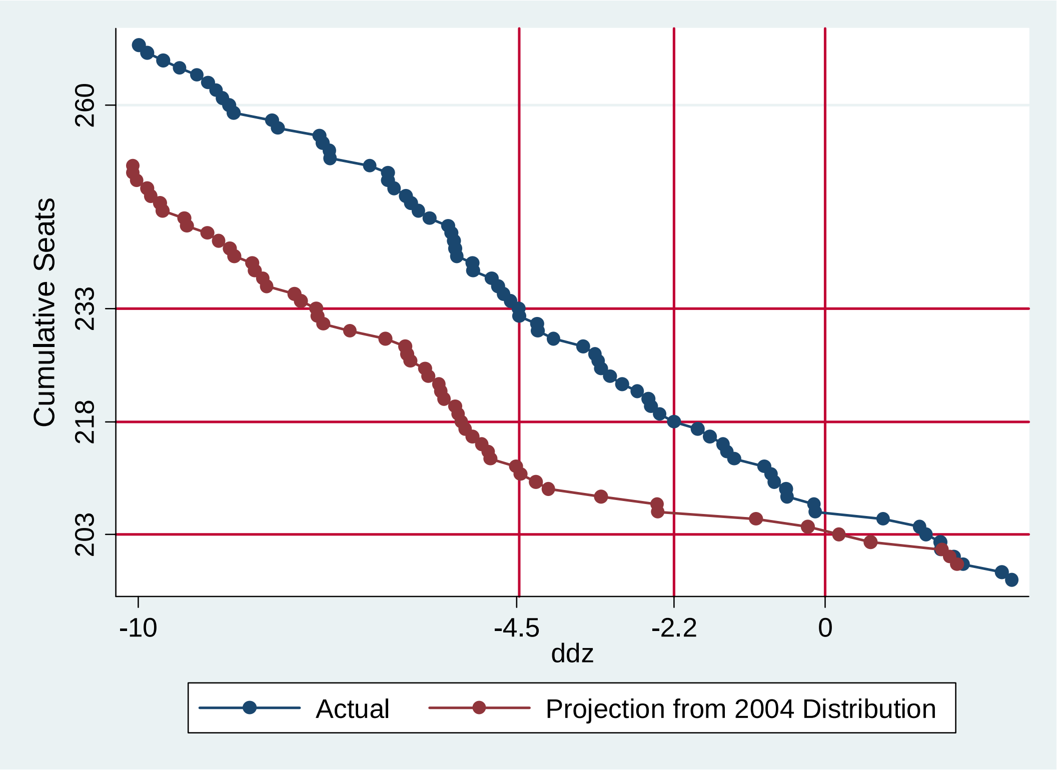

Figure 7. Comparing the actual 2006 cumulative seat distribution and the 2004 projection assuming a 4.5 percentage-point uniform swing. The 2004 observations are anchored by the point (0.203) representing 50 percent as the 2004 vote (no swing) and 203 seats (the status quo). The 2006 observations are anchored by the point (-4.5, 233), the new outcome.

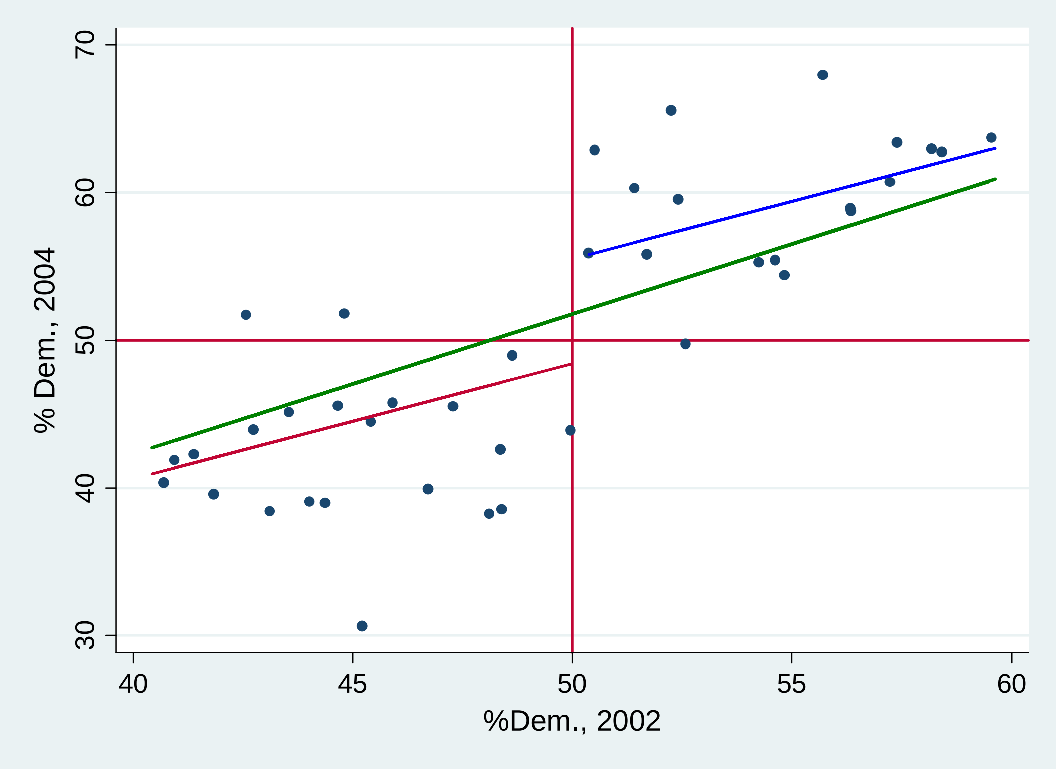

Figure 8. Veteran incumbent vote in 2004 by vote in 2002. The thick regression line is based on all districts. The thin lines with a break at 50% control for incumbent party. The graph is limited to the region where the vote is between 40 and 60 percent Democratic.

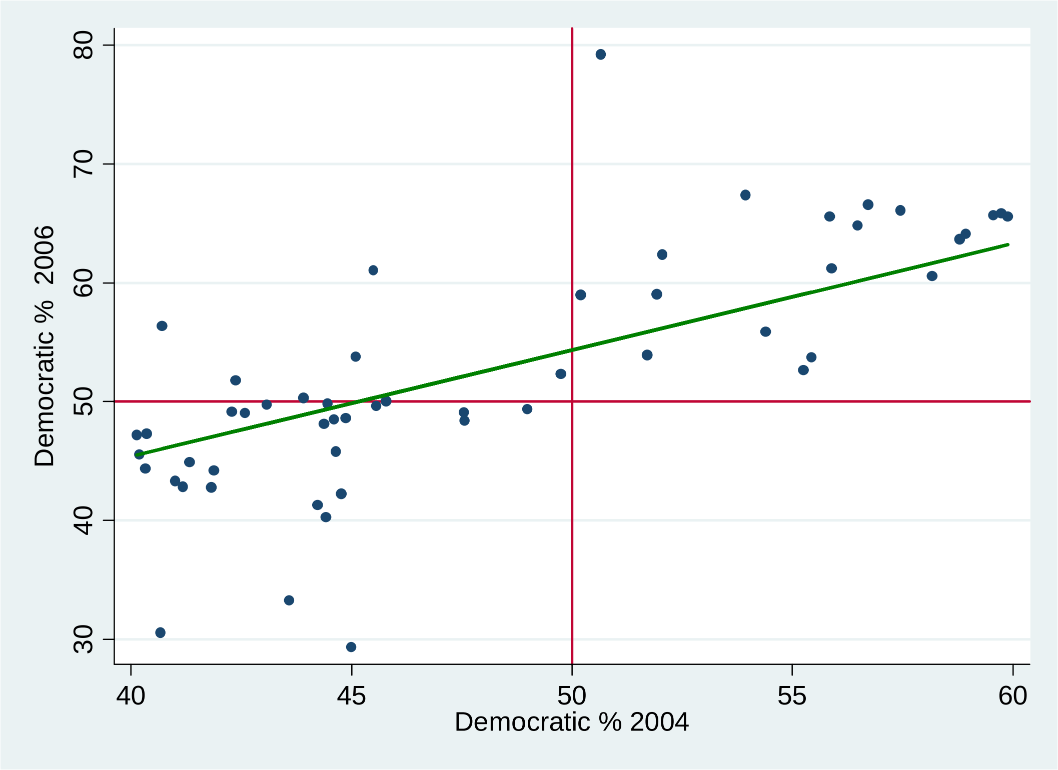

Figure 9. Veteran incumbent vote in 2006 by vote in 2004. The regression line is based on all districts. The graph is limited to the region where the vote is between 40 and 60 percent Democratic. Note that there is no visible break at 50 percent for 2006.