If a Statistical Model Predicts That Common Events Should Occur Only Once in 10,000 Elections, Maybe it's the Wrong Model*

Danny Ebanks[†] Jonathan N. Katz[‡] Gary King[§]

July 13, 2022

Abstract

Election surprises are hardly surprising. Unexpected challengers, deaths, retirements, scandals, campaign strategies, real world events, and heuresthetical maneuvers all conspire to confuse the best models. Quantitative researchers usually model district-level elections with linear functions of measured covariates, to account for systematic variation, and normal error terms, to account for surprises. However, although these models work well in many situations they can be embarrassingly overconfident: Events that commonly used models indicate should occur once in 10,000 elections occur almost every year, and even those which the model indicates should occur once in a trillion-trillion elections are sometimes observed. We develop a new general purpose statistical model of district-level legislative elections, validated with extensive out-of-sample (and distribution-free) tests. As an illustration, we use this model to generate the first ever correctly calibrated probabilities of incumbent losses in US Congressional elections, one of the most important quantities for evaluating the functioning of a representative democracy. Analyses lead to an optimistic conclusion about American democracy: Even when marginals vanish, incumbency advantage grows, and dramatic changes occur, the risk of an incumbent losing an election has been high and essentially constant from the 1950s until the present day.

*Paper prepared for the 39th annual meeting of the Society for Political Methodology, 21-23 July 2022. Our thanks for helpful comments to Matt Blackwell, Devin Caughey, and Kosuke Imai.

[†]Ph.D. Candidate, California Institute of Technology; DEbanks@Caltech.edu, DannyEbanks.com.

[‡]Kay Sugahara Professor of Social Sciences and Statistics, California Institute of Technology; JKatz.Caltech.edu, JKatz@Caltech.edu

[§]Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, Harvard University; GaryKing.org, King@Harvard.edu.

1 Introduction

Political scientists have studied democratic elections over most of the history of our discipline, producing an extensive, high quality, and steadily improving scholarly literature with few equals across scholarly fields. Statistical studies of actual district-level election returns — including forecasts, causal inferences, and counterfactual analyses of numerous phenomena — supplemented by a wide variety of other approaches — such as intensive interviews, survey research, participant observation, archival work, and historical analyses — have generated an enviable record of reliable knowledge about the workings of this crucial democratic institution.

Yet, quite often, the models are spectacularly wrong. This problem is easiest to see in forecasting, where rigorous out-of-sample evaluations are unforgivingly obvious. Although standard models do remarkably well much of the time, and have taught us a great deal, they are embarrassingly far off with some regularity. These forecasting mistakes are not ordinary errors of ordinary magnitudes. Our best models indicate that certain events we see regularly should be rarely observed even if we had data from a trillion elections and some from even a trillion-trillion elections.

The intrepid political scientists who give media interviews after elections take one for our team trying to explain this to the public. Pretty much the best they can do is to say something like "Oops!... We Did It Again" and to explain that voters get to cast ballots for whomever they want. However, we all know (to paraphrase Britney Spears again) we're not that innocent. Errors of such magnitude are not merely mistakes. They are bugs in our logic, our models, our forecasts, our conclusions, our textbooks, our advice, and our public pronouncements — similar to what we would think if built a computer program to forecast the Democratic vote proportion, hit run, and it played a video of a galloping giraffe. This is not a missed forecast; it's the wrong model. And models that do so badly when they are vulnerable to being proven wrong, as in forecasting, do not inspire confidence when applied to other tasks more difficult to evaluate, such as causal or counterfactual inferences.

We build a new general purpose statistical model and validate it with extensive out-of-

sample (and distribution-free) tests in 10,687 district-level US Congressional elections, 1954-2020. We show that, unlike standard approaches, estimates from this model are correctly calibrated, meaning that its probability estimates are accurate. Our analyses from this model reveal the rich complexity and dramatic changes in the landscape of US Congressional elections. They also suggest an optimistic conclusion about a central feature of American democracy: Although, the marginals sometimes vanish and incumbency advantage sometimes soars, the probability of incumbents losing their seats has been quite high and essentially unchanged since the 1950s. In any one election, approximately 20% of have at least a 10% chance of losing, and 20% have at least a 10% chance.

We describe the standard model and our proposed alternative in Section 2, conduct several types of evaluations in Section 3, and give substantive findings about US Congressional elections in Section 4.

2 Statistical Models of District-Level Elections

We begin by summarizing the standard model used in the literature (Section 2.1) and then our proposed model (Section 2.2). (See Appendix A for estimation details.)

2.1 Standard

Our outcome variable for modeling US congressional elections is the Democratic proportion of the two-party vote, v_{it} for district i and election (time) t. The standard model is a linear-normal regression of v_{it} on a vector of K covariates X_{it}, with estimation conducted for each election year t run independently. For most applications in the last quarter century, an independent normal district-level random effect (constant over hypothetical or real elections but varying over districts) is added to the regression to model the political uniqueness of individual districts (Gelman and King, 1994, implemented in JudgeIt software).[1]

The specific content of the covariates varies with the application but, to fix ideas and

[1] Instead of directly estimating γ_i and modeling multiple elections together, which would have been computationally difficult in the 1990s, JudgeIt analyzes one election at a time, after a preprocessing step to estimate how much variation should be attributed to this random effect.

for the analyses below, define X_{it} as including a lagged vote share ($v_{i,t-1}$), incumbent party (the party that won the previous election, with 1 for Democrat and 0 for Republican), incumbency status (1 if the Democratic candidate in election t is an incumbent, 0 for open seat, and -1 for a Republican incumbent), uncontestedness (1 if a Democrat runs uncontested, 0 if contested, and -1 if Republican runs uncontested), an indicator for the old confederate states, and a presidential midterm penalty (coded 1 if t is a midterm year and the incumbent party in district i matches the president's party in that midterm and 0 otherwise).

We summarize this model as

$$\begin{aligned} v_{it} &\sim \mathcal{N}(\mu_{it}, \sigma^2) \\ \mu_{it} &= X_{it}\beta_t + \gamma_i \end{aligned} \tag{1}$$

where β_t is a vector of K linear regression effect parameters, $\gamma_i \sim \mathcal{N}(0, \sigma_\gamma^2)$ is an independent normal random effect with variance $\sigma_\gamma^2 > 0$, and σ^2 is the variance of the usual homoskedastic regression independent normal error term.

2.2 Proposed

Error terms in statistical models are designed to represent "known unknowns," features that reflect political scientists' knowledge of elections too difficult to code among the covariates. The error term in Equation 1 includes γ_i to model the *local uniqueness* of individual districts that often persist over time. For example, in recent years, Minnesota's 7th Congressional District has been much more Republican than the nation as a whole, favoring Donald Trump in 2016 and 2020 by about 30 percentage points. Yet, Democrat Collin Peterson has won this seat from 1991 to 2021 because of his personal brand and unusual political preferences, opposing abortion and supporting the border wall, but (perhaps accounting for how he wins the Democratic nomination) highly progressive economic views.

We now add to this model three other "known unknowns," modeling features that reflect valuable substantive political information well understood by students of elections but rarely modeled directly. First is *covariate effect stability*: β_t varies relatively little

over time. For example, the incumbency advantage might range between two and eight percentage points, with sharp changes over time quite rare. Similarly, the coefficient on the lagged vote is usually in the range of $[0.6, 0.8]$. We add this feature to the model by, first, modeling all elections within a redistricting regime simultaneously (i.e., all the elections for which the election districts remain unchanged), and explicitly allowing each element of β_{tk} (corresponding to covariate k, $k = 1, \ldots, K$ and time t) to come from the same distribution $\beta_{tk} \sim \mathcal{N}(\hat{\beta}, \sigma_{\beta_k})$, where $\sigma_{\beta_k} < \infty$; in contrast, estimating each equation separately as is usually done in the standard approach, is equivalent to setting $\sigma_{\beta_k} \to \infty$. (The notation $\hat{\beta}$ is shorthand to refer to empirical Bayes, meaning that this distribution shrinks different covariate effects in the same redistricting decade toward the same mean, without biasing the mean; this is equivalent to a fully Bayesian model, with the mean in the null space; see Girosi and King 2008.) The idea here is to borrow strength for the estimate of each parameter in each year from estimation of the same parameter in other years, but without the rigidity in a more informative prior. This will be especially valuable in smaller legislatures, such as many state assemblies and senates and the class up for election in the US Senate.

Second, we add a random *national swing*, η_t, allowing all districts in one election to be affected in the same way by the same national event. For example, the 1994 Republican national congressional campaign strategy (known as the "Contract With America") seemed to be a successful heresthetical maneuver that affected all the districts similarly. Although we cannot know ex ante what any this unpredictable national swing will be, we can estimate the distribution of national swings and include this in the model. The result of produces the well known "approximate uniform partisan swing" pattern common across time periods, electoral systems, and even countries (Katz, King, and Rosenblatt, 2020).

Finally, we model district-level *political surprises*, including intentional heresthetical maneuvers (Riker, 1990; Shepsle, 2003) and unintentional exogenous political events, that affect one district's vote differently than others. Consider for example the election in Texas' 22nd district in 2006. Tom Delay was the popular Republican House major-

ity leader from the district, regularly winning election by 35 or more percentage points. During the campaign, he was indicted and abruptly resigned. Worse for the his party, the deadline to field a candidate on the ballot line had passed and so his party could only field a write-in candidate, late in the campaign. The result was that this overwhelmingly Republican district elected a Democrat over the Republican write-in candidate by over 8 percentage points. Equation 1 already includes the usual normal error term, but a normal distribution predicts that extraordinary deviations from the prediction this large would happen so infrequently that it would almost never be observed. Of course, every election observer is aware that these surprises happen regularly.

The problem with the normal distribution is that the tails are not thick enough, which means that big surprises should almost never occur. We thus adopt the additive logistic Student t (ALT) distribution, which, unlike the normal, has appropriately fatter tails, is adjustable based on the data, and constrains the vote proportion to the [0,1] interval. This distribution also has the simultaneous advantage of having more of its density concentrated near the mean, making the mean more informative as well as accounting better for surprises.[2]

We now combine all these features in one model by reusing the notation (and redefining symbols) from Section 2.1. Thus, let

$$v_{i,t} \sim \text{ALT}(\mu_{it}, \sigma^2, \nu_t), \tag{2}$$

$$\mu_{i,t} = X_{i,t}\beta_t + \gamma_i + \eta_t \tag{3}$$

with independent error term components

$$\beta_{tk} \sim \mathcal{N}(\hat{\beta}_k, \sigma^2_{\beta_k}), \quad \gamma_i \sim \mathcal{N}(0, \sigma^2_\gamma), \quad \eta_t \sim \mathcal{N}(0, \sigma^2_\eta),$$

for $k = 1, \ldots, K$ covariates, $i = 1, \ldots, n$ observations, $t = 1, \ldots, T$ elections, and diffuse priors chosen for estimation convenience.[3]

This error structure represents local uniqueness via a positive contemporaneous correlation between any two districts i and j ($i \neq j$) in election year t. For simplicity, we study

[2]Roughly, the ALT is the t distribution applied to the (unbounded) logistic transformation of the vote $\ln v_{it}/(1 - v_{it})$, with everything transformed back to the [0,1] interval. For technical details, and extensive evaluations in multiparty elections, see Katz and King (1999).

[3]That is, $\sigma_\beta, \sigma_\omega, \sigma_{tk}, \sigma_i, \sim$ Exponential(0.2) and $\nu \sim \Gamma(3, 0.5)$.

this on the logistic scale and define $y_{it} \equiv \ln[v_{it}/(1-v_{it})] = X_{i,t}\beta_t + \gamma_i + \eta_t + \omega_{it}$, where Equation 2 implies that ω_{it} is t distributed. Then

$$\begin{aligned}\text{Cov}(y_{it}, y_{jt}|X_{it}, X_{jt}) &= \text{Cov}(X_{i,t}\beta_t + \gamma_i + \eta_t + \omega_{it}, \;\; X_{jt}\beta_t + \gamma_j + \eta_t + \omega_{jt}) \\ &= \Sigma_\beta^2(X_{it}'X_{jt}) + \sigma_\eta^2 > 0, \end{aligned} \tag{4}$$

where Σ_β is a diagonal matrix with element $\{\sigma_{\beta_k}^2\}$ on the diagonal.

The error structure also represents national swing via a positive correlation, for any one district, between any two time points (within a redistricting decade so district i is the same for all t): For any district i at times t and t',

$$\begin{aligned}\text{Cov}(y_{it}, y_{it'}|X_{it}, X_{it'}) &= \text{Cov}(X_{it}\beta_t + \gamma_i + \eta_t + \omega_{it}, \;\; X_{it'}\beta_{t'} + \gamma_i + \eta_{t'} + \omega_{it'}) \\ &= \Sigma_\beta^2(X_{it}'X_{it'}) + \sigma_\gamma^2 > 0. \end{aligned} \tag{5}$$

3 Evaluation

We now evaluate both the standard linear-normal approach and our proposed logistic t model with contemporaneous correlations, or LogisTiCC for short. We do this by summarizing the models' statistical properties (Section 3.1), comparing the probabilities of rare events from each approach to actual elections (Section 3.2), and studying the models' confidence interval coverage (Section 3.3).

3.1 Statistical Properties

As political scientists have long understood, the linear-normal model can reveal important information about elections, when its specification is correct or close to correct. The normal is not formally a limiting special case of the LogisTiCC although this is a reasonable way to think about the relationship heuristically. For one, as with all potentially misspecified models, point estimates from the normal will choose the distribution closest to the true data generation process (in the sense of the Kullback-Leibler information criterion; see White 1996) even if the data come from the LogisTiCC. In addition, if the linear specification is correct, both the normal and the LogisTiCC models will produce consistent estimates of (the same) β.

Unfortunately, given the covariance structure of the proposed model, estimates from the normal will be highly inefficient relative to the LogisTiCC, if data come from the the model we are putting forward that would seem to better represent the knowledge of election experts. Standard errors of β will be incorrect (but could be corrected with a robust version). However, most quantities of interest other than β, such as even the probability of a candidate winning an election, will be statistically inconsistent under the normal but consistent with the LogisTiCC.

3.2 Rare Events Probabilities

We analyze 28 years of US Congressional elections from 1954 to 2020, including a total of 10,687 district-level contests, usually limiting forecasts to all but the first year of each redistricting decade. This large dataset enables us to conduct numerous rigorous evaluations (cf. Grimmer, Knox, and Westwood, 2022), all of which we do out-of-sample (so that no data from the election being predicted is used during estimation). In each analysis, we use either a one-step-ahead or leave-one-out forecast, depending on context.

To begin, we consider the probability of extraordinarily rare events under each model. For illustration, we begin with the notion from the Enlightenment that events with probabilities smaller than 1 in 10,000 should be disregarded, a quantification of the idea of *moral certitude*. This idea is that we should be "morally certain" that these events will not occur, and so we should not expect to ever observe them (Buffon, 1777; Kavanagh, 1990). We thus now make predictions for 8,198 elections (all elections in our dataset except the first year in each redistricting decade) and count the number of elections for which the vote proportion observed out-of-sample appears outside a 99.99% (i.e., $1 - 1/10,000$) forecast credible interval.

Figure 1 gives a count of these extraordinarily rare events (on the vertical axis) by election year (on the horizontal axis). We present counts for the normal model (in salmon) and our LogisTiCC (in black) for each year. As can be seen, the data dramatically violate the normal model's predictions in a disturbingly large number of elections. In the entire dataset, we would expect a very small, probably single digit, number of such events. Yet, outcomes the normal model says we should be "morally certain" will not occur actually

occur in as many as 12 of 435 elections (in 2010). We also annotate some of the points with the exact probability that we would expect to see these results under the model. These forecasts are stunningly bad. Richard McKelvey, a prominent political scientist who died in 2002, was fond of arguing that a fix for much empirical work would be to require anyone reporting a p-value in a publication to take a bet with the reported odds against someone finding evidence to the contrary. This would mean, for a one dollar bet, that one would have an equal chance of winning quadrillions of times more money than exists in circulation of all the world's currencies.

In stark contrast, the black line in Figure 1 shows that only one of the out-of-sample observed election results are much of a surprise to our proposed (LogisTiCC) model. All but one year has zero events and just one (in 1996) has one event with a modest probability only 1 in 26.5, which is about what we would expect if the world generated all the data according to this model.

Thus, for this measure of extraordinarily unlikely events, the out-of-sample performance of our proposed model vastly exceeds that of the standard approach. We now show that this result is general in that the probabilities from our model, but not the normal, are well *calibrated*, meaning that for example when the model predicts that a certain event will occur with a 3% probability, that event actually occurs in about 3 of every 10 elections, and so on. We do this, for each election and model, by first computing the (out-of-sample) probability of a competitive outcome (which we define as $v_{it} \in [0.45, 0.55]$). We then sort these probabilities of a competitive outcome into bins, $[0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$,..., separately for each model and plot them in Figure 2, as follows. For each model, we plot a dot with a horizontal coordinate as the average of the estimated probabilities of elections in a bin and the vertical coordinate as the number of (out-of-sample) elections in the same bin that are in fact observed to be competitive. Dots for a perfectly calibrated model should fall approximately on the 45 degree line.

As Figure 2 demonstrates, the dots computed from the LogisTiCC bins (in black) are all close to the 45 degree line, and hence well calibrated. In contrast, those from the normal (in salmon) substantially deviate from the 45 degree line of equality as the predicted

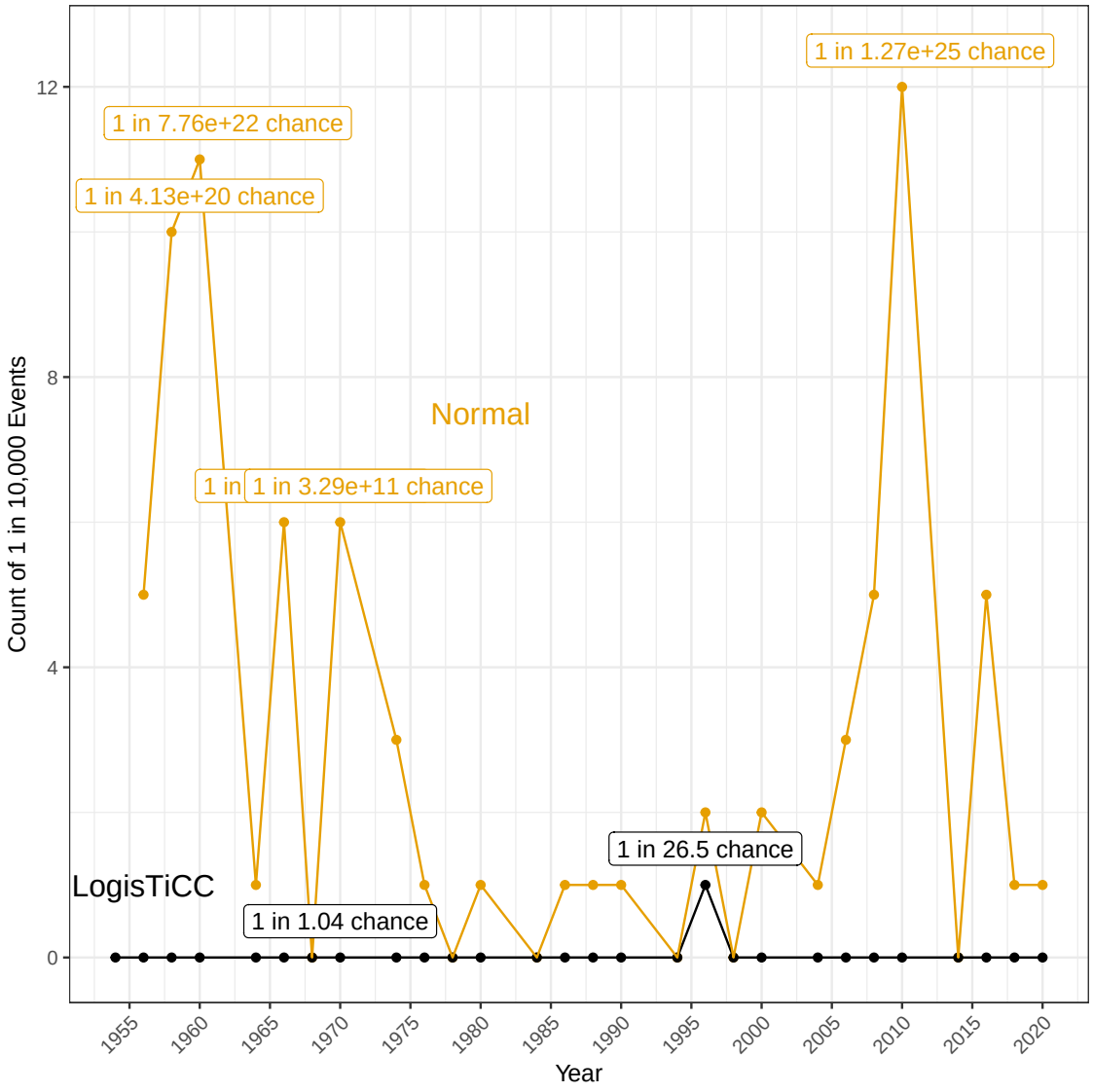


Figure 1: Moral Certitude: Count of elections outside a 99.99 credibility region for each election year (with selected points labeled with the probability each model gives of seeing this many "1 in 10,000" events). Separate calculations appear for the normal model (in salmon) and our proposed LogisTiCC model (in black).

probability of a competitive election gets higher. In other words, the normal model fails most dramatically in elections that are most politically important, the competitive ones.

3.3 Coverage

We now study, in three ways, the properties of credible intervals computed from the standard and proposed models. First, we plot in Figure 3 a time series of one of the most consequential quantities of interest in US politics — the Democratic proportion of the

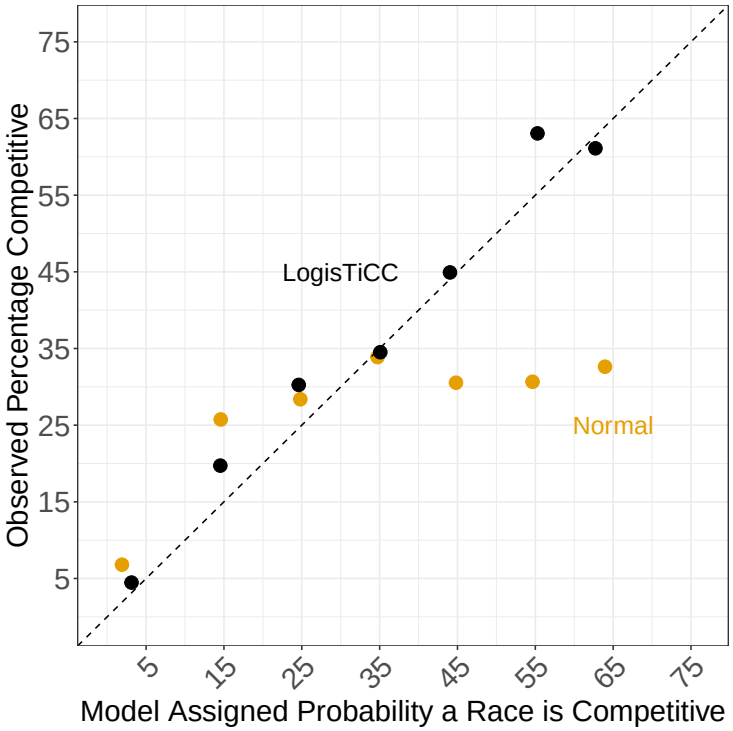


Figure 2: Calibration: Predicted out-of-sample probabilities (horizontally) by observed frequencies (vertically).

vote of the median seat in the House of Representatives (see the red stars). Then, for each year and model, we omit this year from the dataset and compute a point forecast and 95% out-of-sample credible interval around it. These appear in salmon for the normal and black for the LogisTiCC. The intervals for the LogisTiCC are longer than for the normal, but should be interpreted differently. First, recall that a t-based interval has more concentration of mass around the mean than the normal and slightly fatter tails to accommodate rare outliers, even if the variance is identical. Second, the LogisTiCC intervals are accurate (See Figure 2) whereas the normal intervals are hugely overconfident. This can be seen here because the LogisTiCC misses the (out of sample) true observed point in only 2 of 21 elections, whereas the normal misses 18 of 21.

Second, for each model, we compute a 95% out-of-sample credible interval around every individual district's vote share and tally up the percentage of districts that interval captures. Our results appear in Figure 4, with time on the horizontal axis and the percent coverage on the vertical axis (again with normal in salmon and LogisTiCC in black). A properly calibrated model should capture 95% of districts which, aside from some estima-

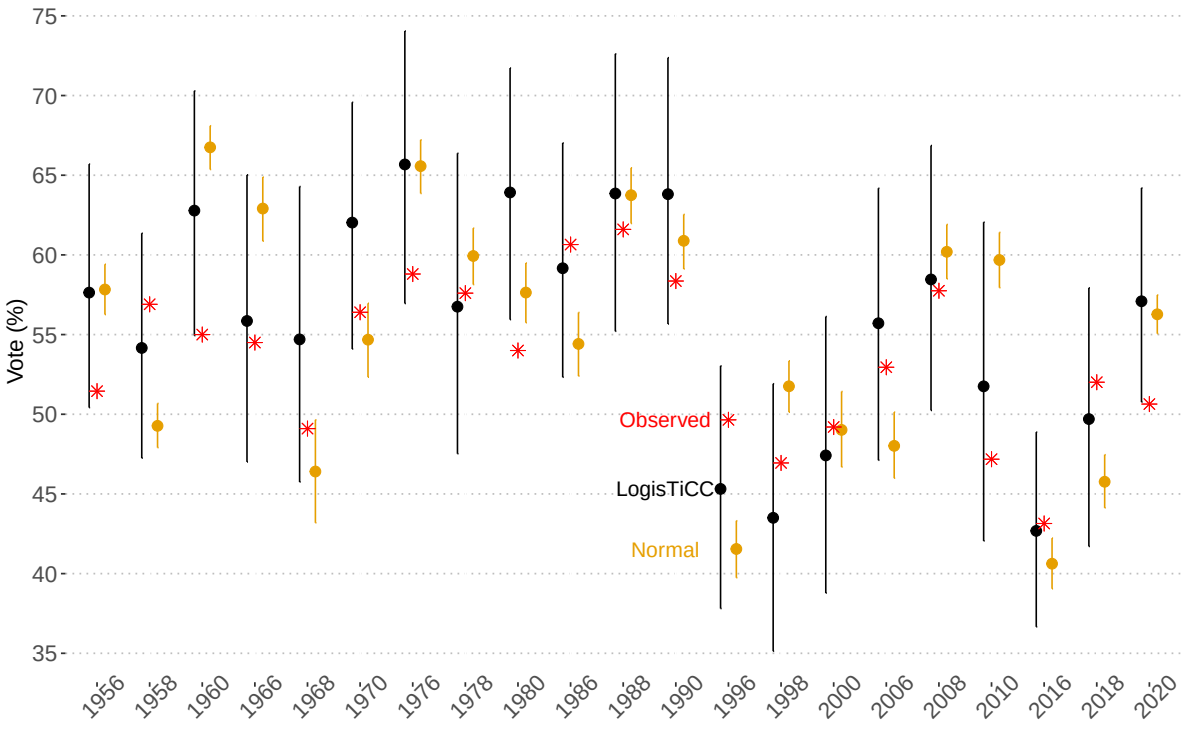


Figure 3: Expected Vote Share of the Median House Seat (95 Percent Credible Interval)

tion error, should be at approximately the flat black line near the top of the figure. This is the case for the LogisTiCC, which has well calibrated intervals. In contrast, the normal interval substantially deviates from capturing 95% of the elections. Importantly, the error of the normal model is always in the same direction, indicating massive overconfidence and reflecting what we see so often on election night. (Similar results hold for other coverage probabilities.)

Finally, we evaluate our distributional assumption (a compound error term with normal random effects and an additive logistic t distribution). To do this, we use methods of "conformal inference" that offer guarantees of accurate distribution-free finite sample coverage even under model misspecification, for any predictive model (Balasubramanian, Ho, and Vovk, 2014). Intuitively, the method works by computing confidence intervals based on errors from previous years' forecasts, assuming primarily that the data generation process is exchangeable (up to the covariates). Figure 4 also includes conformal confidence intervals (in red). The conformal intervals clearly have accurate coverage, as designed, which we can see as the red line hovers around the 95% line for all years.

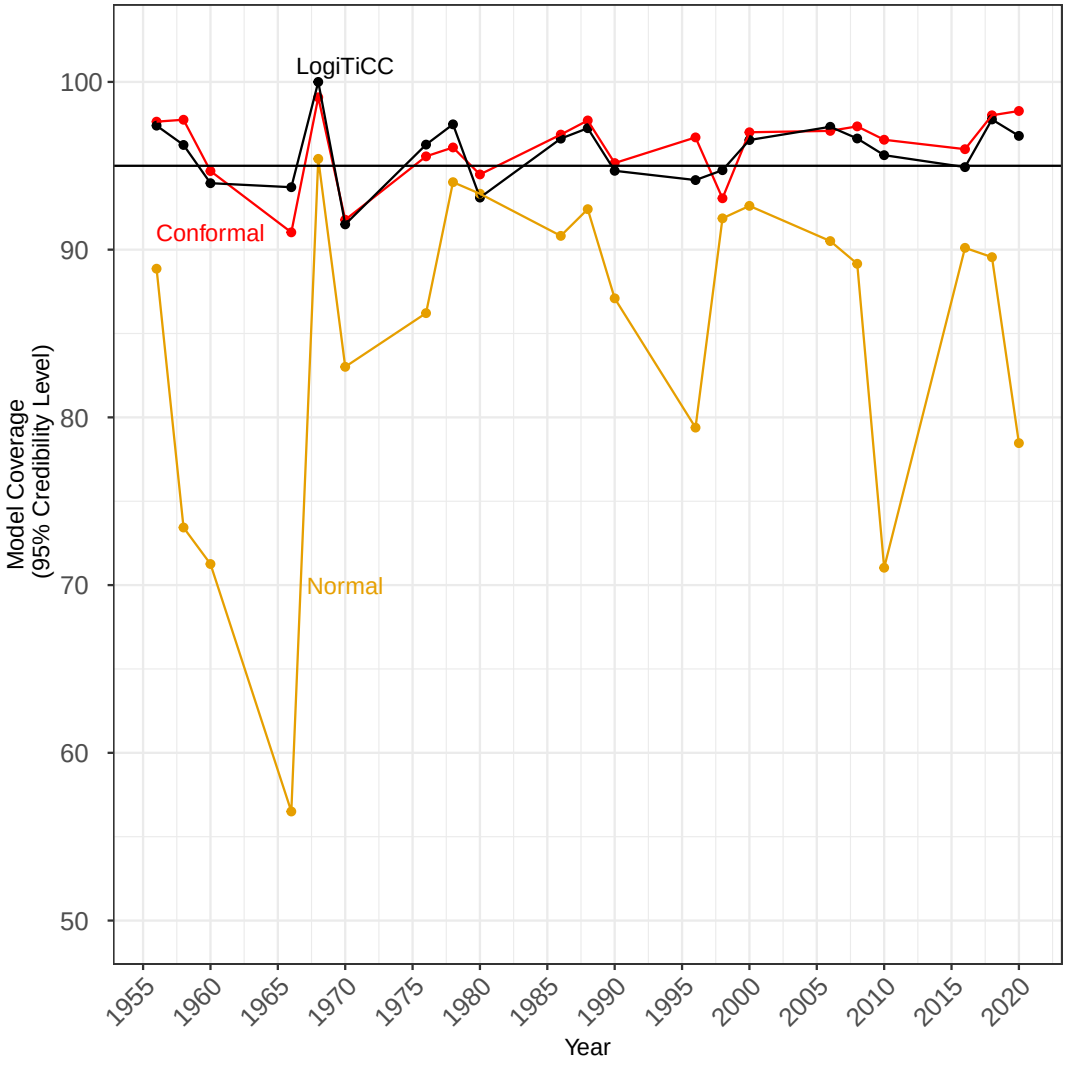


Figure 4: Coverage under Each Model at the 95 Percent Level

Remarkably, these distribution-free intervals have approximately the same high quality coverage as the LogisTiCC, which supports our fully Bayesian approach.

4 Change-Induced Stability in Congressional Elections

A key question for any democracy is the responsiveness of its legislators to constituent preferences. To motivate responsiveness, elections must produce a consequence for violating the voters' will. Abramowitz and Webster (2016), Ferejohn (1977), and Mayhew (1974) was the first to worry about this guarantee breaking down in the decline during the 1970s in the number of competitive elections, which he attributed to an increase in the

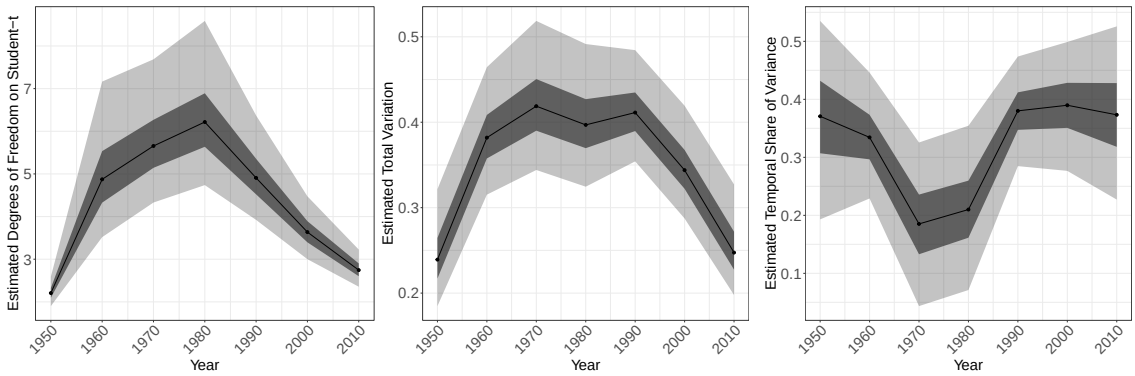
electoral value of incumbency. Studies of these "vanishing marginals," and corresponding incumbency advantages, were a major concern to a generations of scholars. Neither, however, was able to directly and accurately measure the probability that any one incumbent would lose the next election, and so neither necessarily translated directly into the likely responsiveness of legislative officeholders to their constituents.

We address these issues by first summarizing the titanic changes in US congressional elections over the last 66 years, as revealed by our model. We then show how these dramatic changes coexisted with remarkable stability over the entire period in the central quantity of interest with respect to the responsiveness of legislators — the risk of incumbents losing their seats.

4.1 Change

We summarize the key changes in congressional elections in three steps corresponding to the three panels in Figure 5. First, Figure 5a gives a time series plot of estimates of ν (the degrees of freedom parameter from our additive logistic t distribution). Recall that when ν is larger, as for example during the 1970s and 80s, the distribution is closer to normal, whereas when ν is small, such as at the beginning and end of this time series, the predictive distribution becomes more like a Student's t, with more density concentrated near the mean prediction and simultaneously fatter tails so that extraordinary events also happen more frequently. In the highly partisan 1950s-60s and 2000s-10s, we expect and see concentration of the vote around each party as different districts across the country swing more with each other. But the small value of ν also shows that extraordinary outliers are quite likely even when each party is speaking with one voice.

Second, in Figure 5b, we plot estimates of the total error variance not coded in the covariates ($\sigma_\gamma^2 + \sigma_\eta^2 + \sigma^2$). The results indicate that, in eras with less partisanship (1970s-80s), the vote is inherently less predictable. Since the parties are less coherently defined and individual members do not separate ideologically as much, there is more room for heresthetical maneuvers in the feasible policy space for candidates of both parties as they seek reelection. During the partisan eras at the start and end of the period, this variation declines (even though, recall, unexpected events can happen with large frequency due to



(a) Degrees of Freedom (b) Total Variation (c) Share of Variance due To Temporal Variation

Figure 5: Model Parameters

ν). See, relatedly Abramowitz and Webster (2016) and Carson, Sievert, and Williamson (2020)

Finally, we plot in Figure 5c the proportion of the total variance (from Figure 5b) that comes from variation over time. During the partisan eras, the variation for individual districts is small (concentrated around their party means) but the entire set of districts swing readily over time. In the less partisan era of the 1970s and 80s, the higher variance of individual districts leads to less coherence and less opportunity for districts to swing together over time.

4.2 Stability

Before moving to directly computing the probability of an election loss, we report a more familiar statistic, the electoral advantage of incumbency (Gelman and King, 1990). A time series plot of this quantity computed from our model is shown in Figure 6. This gives an estimate of the *expected* increase in the vote for a party that comes solely due to nominating the incumbent for reelection as compared to the best available nonincumbent willing to run. The advantage was about two percent in the 1950s and 60s, but increased to at least eight percentage points in the 1980s before it dropped back down again (as noted by Jacobson, 2015).

However, these large changes in incumbency advantage (and in marginal elections) are averages only. They do not reflect the different types of underlying variabilities conveyed

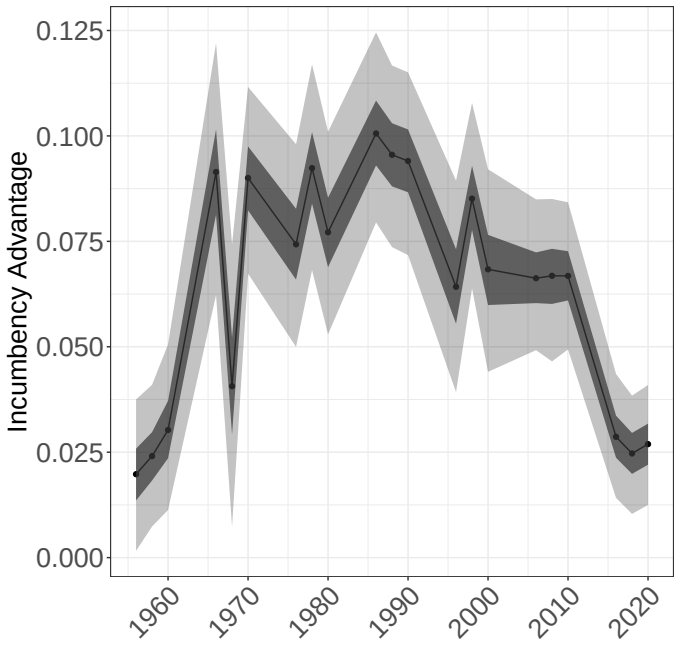


Figure 6: Incumbency Advantage over Time

in Figure 5 and so cannot, by itself, be used to convey the probability of an incumbent loss. Indeed, as we now show, using only the average as indicative of the probability of an incumbent electoral loss is a mistake. If the variation is high when the incumbency advantage is high, the risks of losing your seat may itself remain high. This indeed seemed to have happened. In the 1970s and 80s, partisanship was low and perhaps as a consequience incumbency advantage was high. However, as Figure 5b shows, this same period had very high variation around this average: Incumbents during this period appear to have had a higher *expected* vote outcome but also higher probability of losing (because the area below 0.5 was larger even when the mean prediction was much higher). In the partisan eras of the 1950s and 2010s, incumbency advantage was low but incumbents would be relatively protected by swinging along with their party; however the fat tails of this era meant they still had a substantial probability of losing their seat.

We summarize all these results by directly estimating from our model the probability of each incumbent losing a reelection contest. We then look for any possible trend in these probabilities, but find no evidence for anything but stability.[4] We summarize these results

[4]We regressed estimates of the probability of incumbent loss on a linear time trend. We then allowed for nonlinearities by using multiple types of splines. Every test we ran produced hypothesis tests far below

Figure 7: The Frequency of High Incumbent Loss Probabilities

in Figure 7 which shows that that the percent of incumbents with more than a 10% risk of losing is 20.3% with a relatively narrow credible interval and the percent of incumbents with more than a 20% risk of losing is about 9.7%, also with low levels of uncertainty.

Of course, nonincumbents in open seat races have much higher probabilities of losing and incumbent challengers's chances of losing are higher still. But even the numbers we report, for the most experienced and successful candidates, are substantial. If you are a tenured professor, think of how much more you might pay attention to the chair of your department or review committee if 10% of tenured professors like yourself had a 20% chance of losing their job every year, and 20% had a 10% chance of losing it annually. Your laurels wouldn't be very restful.

5 Concluding Remarks

Commonly used models of district-level election results have enabled political scientists to learn a wide variety of information about American legislative democracy. But the models also fail spectacularly quite often in ways that should almost never happen. We build on this existing approach by adding features of elections political scientists have learned over the years, with new statistical technology not available in previous decades. We validate our approach with extensive out-of-sample (and distribution-free) tests in

traditional levels of significance.

more than 10,000 district-level elections. Our model is general in that it can be used, with the right assumptions and covariates, to estimate almost any quantity of interest in the literature, and others, all with calibrated (i.e., accurate) probabilities and honest uncertainty intervals.

We apply the model to estimate one of the most central requirements of any representative democracy — the extent to which legislators have a serious chance of losing reelection. We reveal this number to be quite high and remarkably constant over more than half a century, a time period which we show has seen dramatic changes in many other important characteristics of electoral politics.

Appendix A Estimation

The usual normal model is usually estimated with a linear regression for forecasting (i.e., dropping γ_i) or, for other quantities of interest, via an approximate two-step procedure designed to avoid computational challenges difficult in the 1990s Gelman and King (1994).

Because of improvements in computation and Bayesian modeling, we estimate our LogisTiCC model via a fully Bayesian specification of Equation 2. We implement the model in "brms," which is open-source software that uses Hamiltonian Markov Chains sample from the posterior distribution of a mixed-effects model. In practice, we draw samples of the posterior distribution from the Bayesian mixed-effects representation. When lagged congressional vote share is a covariate, we drop the first election of each redistricting decade to fit the model.

Our Bayesian methods are computationally efficient, which enables us to analyze large legislatures, and does not require asymptotic assumptions, which is especially important for the US Senate class up for election in any one year, small legislatures, or many state upper and lower houses. These methods also able to draw from the full posterior distribution of the predicted values and parameters, which means we can easily calculate any relevant quantity of interest, along with honest, accurate, and calibrated uncertainty estimates.

References

Abramowitz, Alan I and Steven Webster (2016): "The rise of negative partisanship and the nationalization of US elections in the 21st century". In: *Electoral Studies*, vol. 41, pp. 12–22.

Balasubramanian, Vineeth, Shen-Shyang Ho, and Vladimir Vovk (2014): *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Buffon, George Louis Leclerc de (1777): "Essai d'arithmétique morale". In: *Euvres philosophiques*.

Carson, Jamie L, Joel Sievert, and Ryan D Williamson (2020): "Nationalization and the incumbency advantage". In: *Political research quarterly*, no. 1, vol. 73, pp. 156–168.

Ferejohn, John A (1977): "On the decline of competition in congressional elections". In: *American Political Science Review*, no. 1, vol. 71, pp. 166–176.

Gelman, Andrew and Gary King (Nov. 1990): "Estimating Incumbency Advantage Without Bias". In: *American Journal of Political Science*, no. 4, vol. 34, pp. 1142–1164. URL: tinyurl.com/yymdaj5r.

— (May 1994): "A Unified Method of Evaluating Electoral Systems and Redistricting Plans". In: *American Journal of Political Science*, no. 2, vol. 38, pp. 514–554. URL: j.mp/unifiedEc.

Girosi, Federico and Gary King (2008): *Demographic Forecasting*. Princeton: Princeton University Press. URL: j.mp/dsmooth.

Grimmer, Justin, Dean Knox, and Sean Westwood (2022): "Assessing the Reliability of Probabilistic US Presidential Election Forecasts May Take Decades". In.

Jacobson, Gary C (2015): "It's nothing personal: The decline of the incumbency advantage in US House elections". In: *The Journal of Politics*, no. 3, vol. 77, pp. 861–873.

Katz, Jonathan N, Gary King, and Elizabeth Rosenblatt (2020): "Theoretical foundations and empirical evaluations of partisan fairness in district-based democracies". In: *American Political Science Review*, no. 1, vol. 114, pp. 164–178. URL: GaryKing.org/symmetry.

Katz, Jonathan N. and Gary King (Mar. 1999): "A Statistical Model for Multiparty Electoral Data". In: *American Political Science Review*, no. 1, vol. 93, pp. 15–32. URL: bit.ly/mtypty.

Kavanagh, Thomas M (1990): "Chance and Probability in the Enlightenment". In: *French Forum*. Vol. 15. 1, pp. 5–24.

Mayhew, David R (1974): "Congressional elections: The case of the vanishing marginals". In: *Polity*, no. 3, vol. 6, pp. 295–317.

Riker, William H (1990): "Heresthetic and rhetoric in the spatial model". In: *Advances in the spatial theory of voting*, vol. 46, p. 50.

Shepsle, Kenneth A (2003): "Losers in politics (and how they sometimes become winners): William Riker's heresthetic". In: *Perspectives on politics*, no. 2, vol. 1, pp. 307–315.

White, Halbert (1996): *Estimation, Inference, and Specification Analysis*. New York: Cambridge University Press.