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§

May 10, 2023

Abstract

Political scientists forecast elections, not primarily to satisfy public interest, but to validate statistical models used for estimating many quantities of scholarly interest. Although scholars have learned a great deal from these models, they can be embarrassingly overconfident: Events that should occur once in 10,000 elections occur almost every year, and even those that should occur once in a trillion-trillion elections are sometimes observed. We develop a novel generative statistical model of US congressional elections and validate it with extensive out-of-sample tests. The generatively accurate descriptive summaries provided by this model demonstrate that the 1950s was as partisan and differentiated as the current period, but with parties not based on ideological differences as they are today. The model also shows that even though the size of the incumbency advantage has varied tremendously over time, the risk of an in-party incumbent losing a midterm election contest has been high and essentially constant over at least the last two thirds of a century.

Words: 11,665

*Presented at the 39th annual meeting of the Society for Political Methodology, 21-23 July 2022. Our thanks for helpful comments to Neal Beck, Matt Blackwell, Patrick Brant, Devin Caughey, Aleksandra Conevska, Mo Fiorina, Andrew Gelman, Justin Grimmer, Bernie Grofman, Kosuke Imai, and Dave Mayhew.

†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 causal effects, counterfactual analyses, forecasts, and full generative models 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 produced an enviable record of reliable knowledge about the workings of this crucial democratic institution.

Yet, quite often, commonly used statistical models are spectacularly wrong. This is easiest to see in election prediction, where rigorous out-of-sample evaluations are unforgivingly obvious, and a major concern even when prediction is not the immediate goal. Although standard models do remarkably well much of the time, and have taught us a great deal, they are embarrassingly far off with regularity. These 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. But 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 we 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 prediction problems, do not inspire confidence when applied to other tasks more difficult to evaluate and of more interest to social scientists, such as causal inferences or generatively accurate descriptive summaries.

We aim to learn some fundamental characteristics of electoral democracy through a validated generative statistical model capable of estimating many of the diverse quantities political scientists find of interest. These include descriptive quantities — such as the probability of an incumbent losing, the odds of a competitive election, the expected vote of the median house seat, partisan bias, electoral responsiveness, among others — and, with appropriate additional assumptions, causal and other counterfactual inferences. Only a generative model can provide sufficient generality to estimate all these and other quantities, along with accurate uncertainty estimates, which is unlike approaches better for more specific purposes, such as via model-free, distribution-free, machine learning, or semi-parametric approaches. Such a model should also be capable of making election forecasts, but we (and most other political scientists) are not especially interested in forecasting in and of itself (except as citizens to participate in the fun and public interest leading up to an election). After all, from an academic perspective, the best method of forecasting is well known: just wait a bit. However, ensuring that we have a useful model requires that it be made vulnerable to being proven wrong in as many ways as possible, for which forecasting — along with leave-one-election-year-out cross-validation — is essential.

We thus build a new general purpose statistical model and validate it with extensive out-of-sample tests in 14,710 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 representations of empirical frequencies. Some of the generatively accurate descriptive summaries from this model reveal the rich complexity and dramatic changes in the landscape of US Congressional elections, including a reinterpretation the 1950s as very similar to the present day, except with parties then based on social-psychological groups rather than ideological distinctions. 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 that incumbents losing their seats has been quite high and essentially unchanged over our entire sample period. Of course, the same model can be used to estimate

numerous other quantities.

We describe the standard model and our proposed alternative in Section 2, perform many out-of-sample evaluations in Section 3, and give substantive findings and even suggest a broader theory of congressional elections consistent with these results in Section 4. Section 5 describes the broader methodological implications of generatively accurate descriptive summaries.

2 Statistical Models of District-Level Elections

We summarize the standard model used in the literature (Section 2.1) followed by our proposed alternative (Section 2.2). We construct our alternative approach to incorporate more substantive knowledge of elections, to simultaneously analyze more elections, and to attend to more of the known statistical issues than previously possible, all within a single Bayesian model. This led us to jointly estimate, integrate over, and represent the uncertainty of 3,567 parameters, including coefficients, missing cell values, uncontested districts, and random effects terms.

One of the reasons our approach has not been tried before is that it would have been computationally infeasible even a few years ago. With highly tuned computational algorithms we developed on a new server (with 20 cores and 128gb of RAM, and software tuned specially to this hardware), we are now able to complete one run of our model on a decade of congressional elections data in only about twenty minutes, although a full analysis of all our data with calibration and strictly out-of-sample evaluation requires about 48 hours of model run time generating about 44gb of output. Along with this paper, we are making available easy-to-use open source software that implements all our algorithms and methods.

2.1 Standard

The 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 con-

ducted 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 some by application but, to fix ideas and for the analyses below, we define X_{it} to include 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 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

We now build on the standard model to develop our proposed approach. We keep the same flexibility in choice of covariates within a fully Bayesian framework, but in three steps we describe our changes. First, in Section 2.2.1, we provide a qualitative description of model components we designed to reflect knowledge from the literature on elections that had been excluded from the standard approach. Second, in Section 2.2.2, we put together these components into a single Bayesian model, but for expository purposes focus only

[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.

on the simple special case where all elections are contested. Finally, in Section 2.2.3, we allow district elections to be either contested or uncontested.

See Appendix A for the full likelihood, Supplementary Appendix 4 for computational details, and Supplementary Appendix 2 for a set of "ablation" studies that, by sequentially removing each model component, demonstrates how all components are essential to the performance we achieve. Supplementary Appendix 5 considers alternative modeling assumptions.

2.2.1 Novel Model Components

Error terms in statistical models are designed to represent "known unknowns," features that reflect political scientists' knowledge of elections too difficult to code in the covariates. For example, the error term in Equation 1 allows for *district uniqueness* by adding a random term γ_i to model the persistence of this uniqueness for any one district i over time, beyond changes due to X. For example, Minnesota's 7th Congressional District has long been more Republican than the nation as a whole, favoring Donald Trump in 2016 and 2020 by about 30 percentage points. Yet, Democrat Colin Peterson 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 four other "known unknowns," modeling features that reflect valuable substantive political information well understood by students of elections or observable in the data 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 ten percentage points, with only rare sharp changes over time. 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 (a) modeling all elections within a "redistricting regime" (i.e., all elections for which the district geography remains unchanged) simultaneously rather than independently, and (b) assuming that each element β_{tk} of vector β_t (corresponding to covariate k, $k = 1, \ldots, K$ and time t) comes 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 and independently, as in

the standard approach, is equivalent to setting $\sigma_{\beta_k} \rightarrow \infty$. (The notation $\hat{\beta}$ is a shorthand reference to empirical Bayes, meaning that this distribution shrinks different covariate effects in the same redistricting regime toward the estimated mean without favoring one's a prior guess; 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 the estimation of the same parameter in other years, but without the rigidity and potential bias that would come from 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.[2]

Second, we allow for *positive cross-district covariances* by adding a random national swing term, η_t, that allows all districts in one election to be affected in roughly the same way by the same national event, over and above the information in X. For example, the 1994 Republican national congressional campaign strategy (known as the "Contract With America") seemed to be a successful heresthetical maneuver (Riker, 1990; Shepsle, 2003) that moved all the districts in the Republican direction by approximately the same amount. Although we cannot know ex ante what any one national swing will be, we can estimate the variation caused by the national swings, which we know occur regularly, and represent this uncertainty in the model with a common random effect for all districts. The result is the well known "approximate uniform partisan swing" pattern common across time periods, electoral systems, and even countries (Katz, King, and Rosenblatt, 2020).

Third, we model *district-level political surprises*, including intentional heresthetical maneuvers and unintentional exogenous political events that affect one district's vote at a point in time differently than others and are not included in X. Consider for example the election in Texas' 22nd district in 2006. Tom Delay was the popular Republican House majority leader from the district, regularly winning election by 35 or more percentage points. During the campaign, he was indicted and abruptly resigned. Worse for 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

[2]We could elaborate this assumption by allowing β_t to trend linearly, as a random walk, or as a function of other covariates, but we find no evidence for these alternative approaches in our data.

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 that can be used to model surprises, but a normal distribution indicates that deviations from a prediction this large would happen so infrequently that it would almost never be observed. Of course, as every election observer is aware, these surprises happen regularly, even if we do not know which ones will occur. As we explain below, we will therefore swap out the normal distribution for one that can more appropriately represent these political surprises, also keeping predictions within the [0,1] interval.

Finally, the normal distribution used in the standard approach turns out to be inadequate for two reasons. The first reason is that, although it often works well when the mean or average vote outcome is of interest, it fails miserably for most other aspects of the distribution, such as for uncertainty estimates or the probability of a close election or of one party winning. The second reason is that the normal tail implies that big surprises should almost never occur, meaning that it also gets the concentration around the mean wrong. To fix these problems, we use the additive logistic Student t (ALT) distribution, which, unlike the normal, constrains the vote proportion to the [0,1] interval and also has appropriately fatter tails to represent surprises. In addition, the ALT distribution has the simultaneous advantage of having more of its density concentrated near the mean, making the mean (and the covariates that account for its variation) more informative at the same time as it is accounting better for surprises.[3] The ALT distribution thus allows more informative predictions to coexist in the same model with the possibility of huge surprises.

2.2.2 The Model, with Fully Contested Elections

We now combine all the features described above in one model, reusing the notation (and redefining symbols) from Section 2.1. For expository simplicity, we imagine until the next section that all districts are contested. Thus, let

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

[3]Roughly, the ALT is the implied distribution on v (and so restricted to the [0,1] interval) when the t distribution is applied to the (unbounded) logistic transformation of the vote $\ln v_{it}/(1 - v_{it})$. For technical details, and extensive evaluations in multiparty elections, see Katz and King (1999).

$$\mu_{it} = X_{it}\beta_t + \gamma_i + \eta_t \tag{3}$$

where the variance is decomposed by the ALT for additional flexibility into scale ϕ and degrees of freedom ν_t parameters (as $\nu_t \to \infty$, the ALT approximates the additive logistic normal). The systematic component for the conditional expected value includes three independent random effect terms for covariate effects, district uniqueness, and national swing, respectively,

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

for $k = 1, \ldots, K$ covariates, $i = 1, \ldots, n$ observations, $t = 1, \ldots, T$ elections, and diffuse priors chosen for estimation convenience (see Appendix 4).

For intuition, we consider the voting data on the logistic scale by letting $y_{it} \equiv \ln[v_{it}/(1-v_{it})] = \mu_{it} + \omega_{it} = X_{it}\beta_t + \gamma_i + \eta_t + \omega_{it}$, with error term $\omega_{it} \equiv \ln[v_{it}/(1-v_{it})] - \mu_{it}$, which Equation 2 indicates is is t distributed. This enables us to see, first, that national swings induce a positive covariance between any two districts i and j ($i \neq j$) for each election year t: $\text{Cov}(y_{it}, y_{jt}|X_{it}, X_{jt}, \beta_t) = \sigma^2_\eta > 0$. This setup also makes clear that the random district uniqueness term γ_i induces a positive covariance for election outcomes in any one district i at two times t and t' (within the same redistricting decade), over and above differences due to X: $\text{Cov}(y_{it}, y_{it'}|X_{it}, X_{it'}, \beta_t, \beta_{t'}) = \sigma^2_\gamma > 0$.

As with the standard approach, some covariates one might put in this model vary over i and t (e.g., the lagged vote, v_{it}), some vary only over i (e.g., the confederate states indicator), and some vary only over t (e.g., presidential approval). A random effect can also be excluded, which can be useful when little information exists such as for covariates of the last type when T is small.

2.2.3 The Model, Allowing for Uncontested Elections

In the standard approach, the vote in uncontested elections is often recoded to fixed values such as $v_{it} = 0.25$ for Democrats running uncontested and $v_{it} = 0.75$ for Republicans running uncontested, or sometimes uncontested elections are deleted entirely. We instead formally distinguish between the observed vote v_{it} and the *effective vote* v^*_{it}, defined as the vote proportion that would be observed if the election had been contested (e.g., King

and Gelman, 1991). The effective vote is observed $v_{it}^* = v_{it}$ in contested elections but unobserved if one party runs unopposed. We then impute unobserved values (for uncontested elections) during Bayesian estimation simultaneous with the rest of the model. This approach includes all the information available and accounts for all uncertainty in the imputation.

To model v_{it}^* when unobserved, we replace the outcome variable v_{it} in Equation 2 with the effective vote, and add a "censoring assumption": candidates who run unopposed would have won even if the election were contested. This assumption is intuitive, probably accounts for why the district was uncontested in the first place, and is a special case of the assumption made by Katz and King (1999). We then replace Equation 2 with

$$v_{it}^* \sim \text{ALT}(\mu_{it}, \phi_t^2, \nu_t), \tag{4}$$

and write the likelihood function for an election district that is fully contested as $\text{ALT}(v_{it} \mid \mu_{it}, \phi_t^2, \nu_t)$, for a district where a Democrat runs uncontested as $\psi_{it} \equiv \int_0^{0.5} \text{ALT}(v^* \mid \mu_{it}, \phi_t^2, \nu_t) dv^*$, and for a district where a Republican runs uncontested as $1 - \psi_{it}$. The integral implements the censoring assumption.

The model contains one additional feature: When the lagged effective vote is used as a covariate, it too can be unobserved, which adds another level of modeling complexity. We describe this feature, along with the full likelihood, in Appendix A.

3 Evaluation

We now evaluate both the standard linear-normal approach and our proposed additive 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

standard modeling approach is not formally a limiting special case of the LogisTiCC although it can be thought of as an approximation in some situations. For one, as with all potentially misspecified models, point estimates from the linear-normal model 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 similar (and approximately 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 model we are putting forward that would seem to better represent the knowledge of election experts, and standard errors of β will be incorrect. 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.

As we demonstrate, a key problem with the linear-normal model is its incorrect independence assumptions, leading to substantial false precision in its uncertainty estimates (confidence intervals and standard errors that are too small). In contrast, the LogisTiCC allows for dependence among elections held in the same district at different times and among elections held in different districts on the same day. Correcting for this false precision leads to appropriately larger confidence intervals: the ratio of the nominal width of LogisTiCC-to-normal confidence intervals is about 1.4 for district-level predictions and about 5 for aggregate predictions such as the vote for the median house seat. (See Supplementary Appendix 1 for details.)

3.2 Rare Event Probabilities

We analyze 28 years of US Congressional elections from 1954 to 2020, including a total of 14,710 district-level contests, with forecasts limited to the 10,778 contests that exclude 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 calibration or

estimation). In each analysis, we use either a one-step-ahead or leave-one-out forecast, depending on context.

To begin, consider the probability of extraordinarily rare events under each model. For illustration, we use the notion of *moral certitude* from the Enlightenment, which is that events with probabilities smaller than 1 in 10,000 should be disregarded. (Because demographers of the time observed that the probability of a healthy person dying in the next day was smaller than 1 in 10,000 and does not seem to affect people's behavior in their daily lives, people act as if they are "morally certain" that these rare events will never occur; see Kavanagh 1990; Buffon 1777.) Updating this (quaint) idea, we make predictions for 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. If the interval is correct, we should observe about 1 in 10,000 outside the interval.

Figure 1 gives a count of these extraordinarily rare events (on the vertical axis) by election year (on the horizontal axis) and for the normal model (in gold) and the LogisTiCC (in black). As can be seen, the data dramatically violate the normal model's predictions in a disturbingly large number of elections. In the entire dataset of 10,778 elections, we would expect to see only about *one* 1-in-10,000 event, but this claim is wrong by a factor of more than sixty, in that surprise events the model is morally certain will not occur actually happened in 61 elections (and as many as 12 of the 435 elections in a single year, 1958) (see also Gelman, Carlin, et al., 1995, Ch. 8). The figure also annotates some of the points with the exact probability that we would expect to see these results under the model. These forecasts are stunningly bad. The late Richard McKelvey was fond of arguing that a fix for over-claiming in empirical work would be to require anyone reporting a p-value to take a bet with the implied odds (i.e., the reciprocal of the p-value to one) against someone finding evidence to the contrary. Using this logic, a one dollar bet against the linear-normal model's claimed level of certainty would give an equal chance of winning quadrillions of times more money than exists in circulation in all the world's currencies.

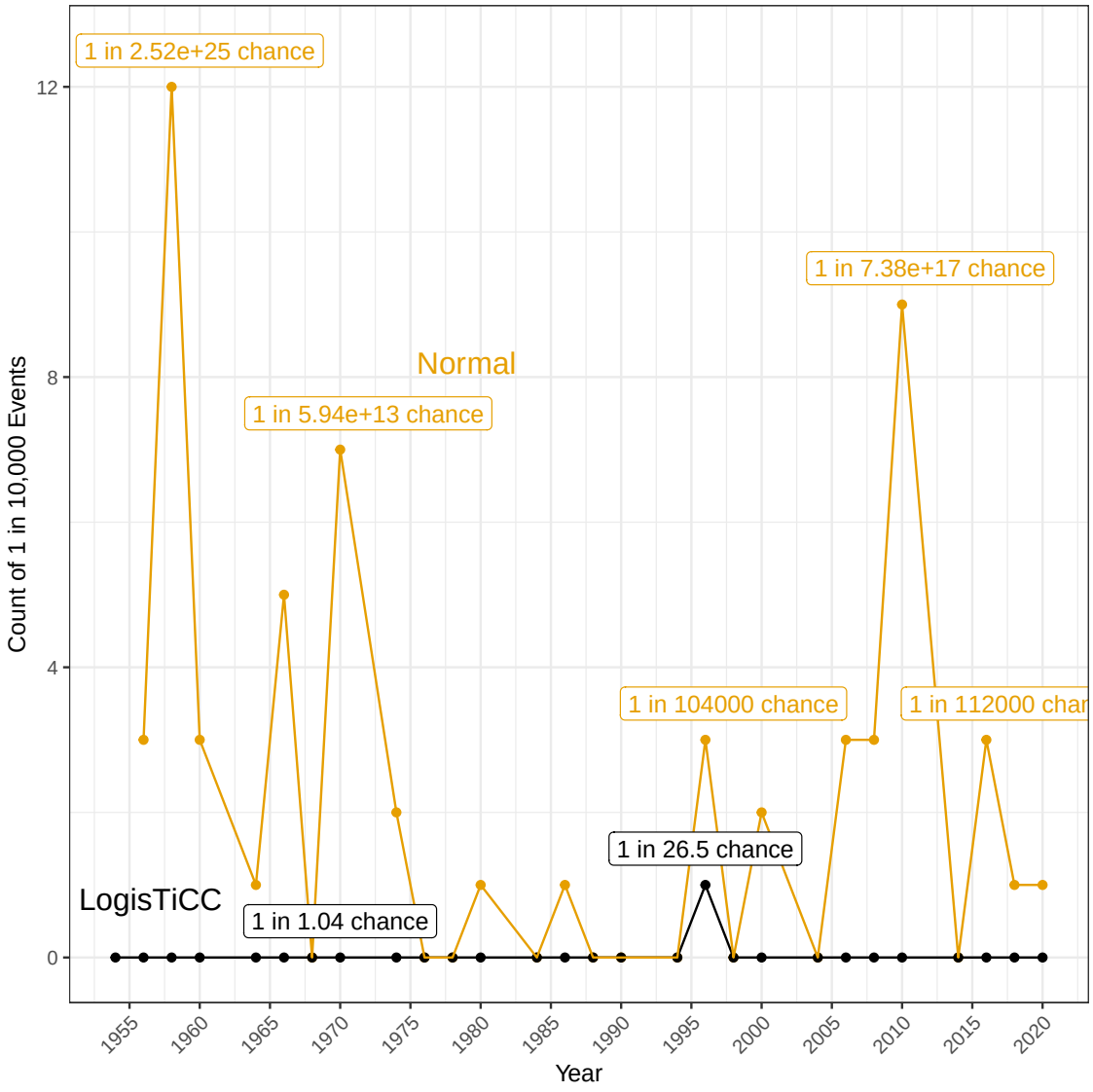


Figure 1: Moral Certitude: Count of elections outside a 99.99 credibility interval for each election year (with selected points labeled with the probability each model gives of seeing this many 1-in-10,000 events). Separate calculations appear for the normal model (in gold) and our proposed LogisTiCC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

In stark contrast, the black line in Figure 1 shows that only one of the 10,778 out-of-sample observed election results are much of a surprise to the proposed LogisTiCC model. All but one year has zero events and just one (in 1996) has one event with a modest probability of 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 30% 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 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.

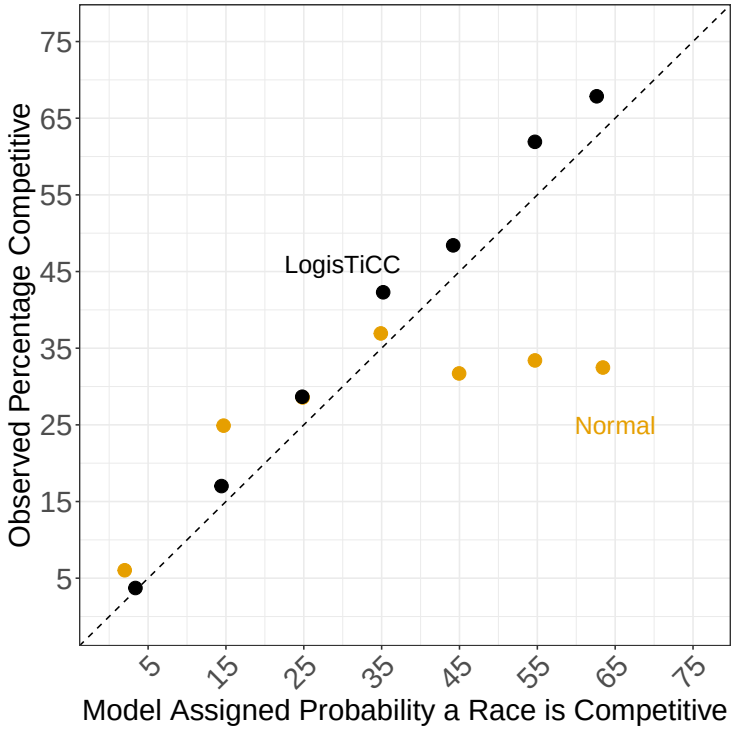


Figure 2: Calibration: Predicted out-of-sample probabilities (horizontally) by observed frequencies (vertically).

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 gold) substantially deviate from the 45 degree line of equality as the predicted 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 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 gold for the normal and black for the LogisTiCC. In addition to the LogisTiCC intervals being longer than for the normal because of the normal's false precision, the LogisTiCC intervals should be interpreted differently. First, recall that a t-based interval has both fatter tails to accommodate surprises and more concentration of density near the mean than the normal (making the mean prediction more informative). Second, the LogisTiCC intervals are accurate (See Figure 2) whereas the normal intervals are overconfident. This can be seen because in these out-of-sample tests, we would expect a well calibrated model to miss only about 1.4 elections, but the normal misses 20 of 27. In contrast, LogisTiCC's predictive confidence interval captures the observed outcome every time.

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 gold and LogisTiCC in black). A properly calibrated model should capture 95% of districts which, aside from estimation error, should be at 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 in all but a few years.

Finally, we evaluate our distributional assumption (a compound error term with 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

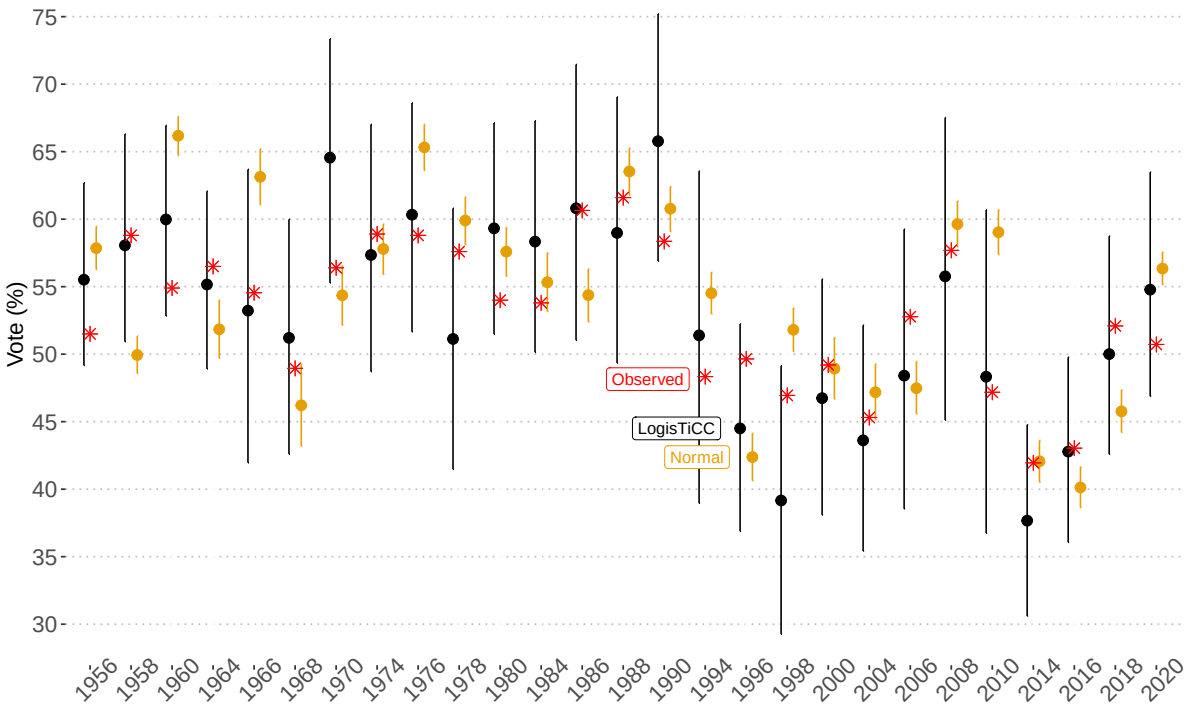


Figure 3: Expected Vote Share of the Median House Seat (95 Percent Credible Interval)

even under model misspecification, for any predictive model, and so we use it to check for misspecification in our model (Vovk, Gammerman, and Shafer, 2005). (Intuitively, the method works by computing confidence intervals based on errors from previous years' forecasts, assuming primarily that the data generation process is exchangeable conditional on the covariates.) In Figure 4 we add conformal confidence intervals (in red). We first confirm that the conformal intervals have accurate coverage, as designed, which we can see as the red line varies around the flat 95% line across the years. More relevant for our purposes is the comparison between the fit of the red and black lines to the 95% line. This comparison indicates that the LogisTiCC has approximately the same high quality coverage as these distribution-free intervals. These results thus provide evidence for the veracity of our distributional assumptions and for our Bayesian model as a generative model of US congressional elections data.

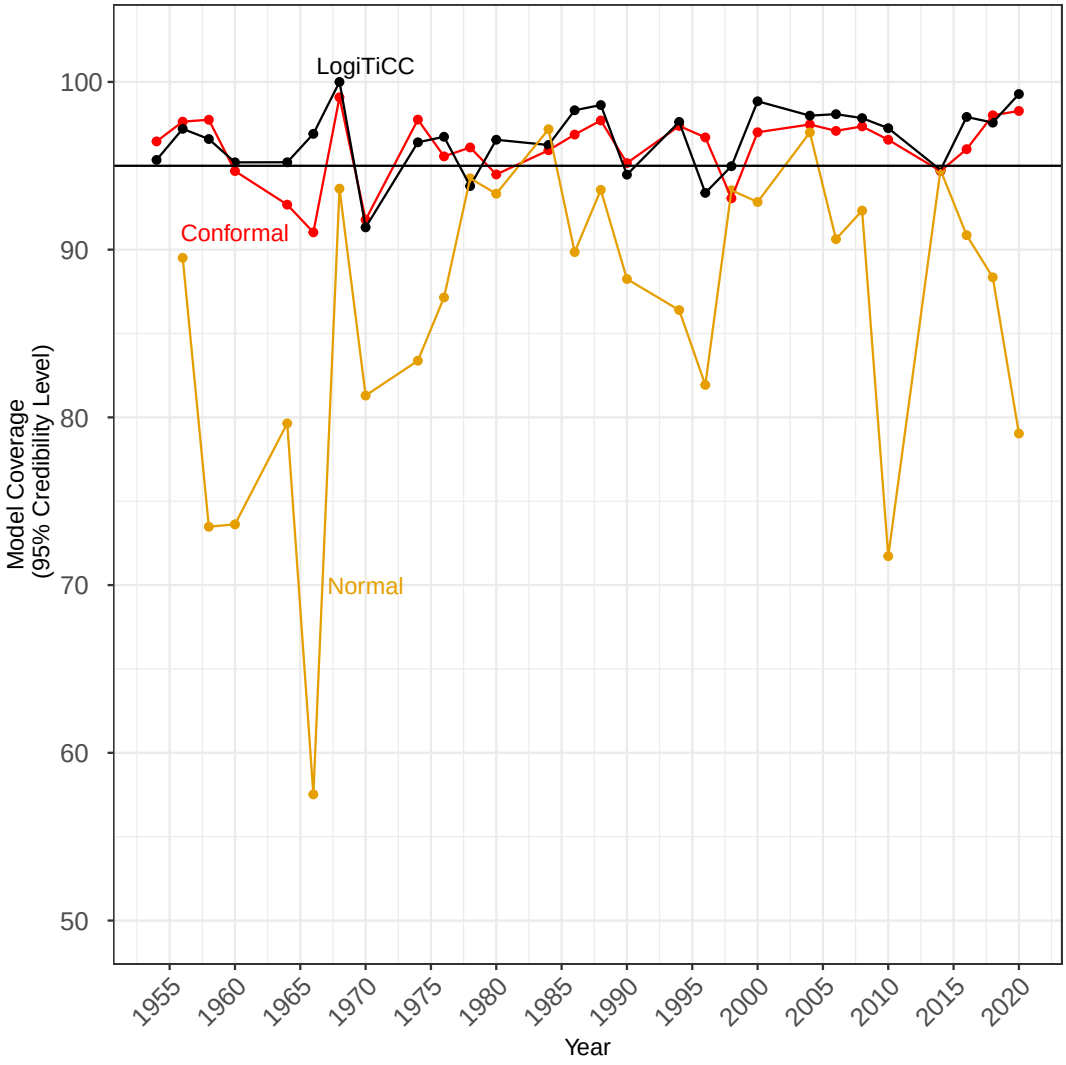


Figure 4: Coverage under Each Model at the 95 Percent Level

4 Electoral Implications

We use our model to compute generatively accurate descriptive summary statistics. First, in Section 4.1, we characterize election variation as falling into three regimes, at the start, middle, end of the 66 years of our study, and how elections throughout are powerfully driven mostly by national rather than local swings. Second, Section 4.2 builds on the first section with an empirical theory of congressional elections consistent with our empirical results and prior literature that tries to strip out several under-appreciated normative assumptions. Section 4.3 then focuses on a key feature of American democracy, the probability of an incumbent loss, and shows that it is essentially constant over time, despite

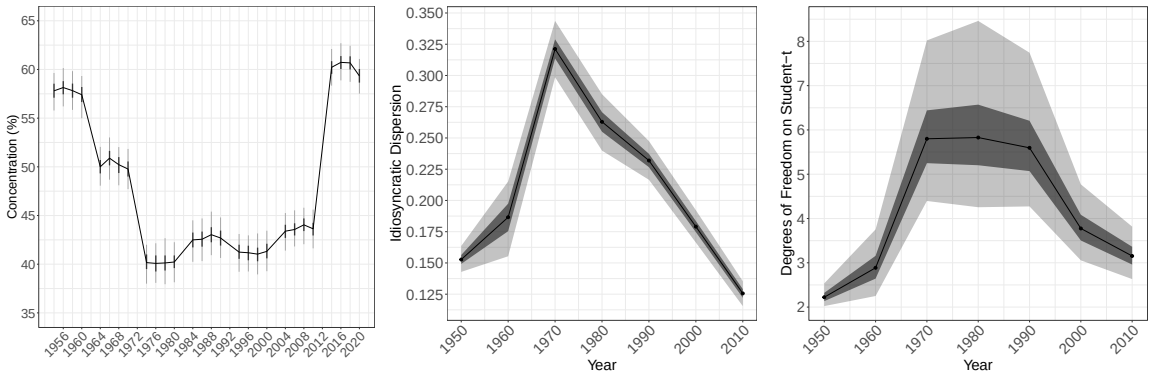
well known huge changes in the incumbent's expected vote advantage.

4.1 The Three Regimes of Election Prediction Variability

Our model decomposes election variability into district uniqueness, national swing, covariate effect stability, and political surprises, in addition to well known covariate effects. As Section 3 shows, these parts of the model provide far better fit to congressional elections data, making for accurate out-of-sample forecasts, uncertainty intervals, and calibrated probabilities. We now turn to the large scale patterns this modeling strategy reveals in congressional elections, leaving most of the substantive implications to the following sections.

First, we begin with an intuitive summary measure of the overall patterns in congressional elections data that we call vote *concentration*, the proportion of the vote probability mass in the interval [0.45,0.55], for mean predictions of 0.5. As Figure 5a shows, the early and late periods have high vote concentration, meaning that any one prediction conveys more certainty and more information, whereas the middle years have substantially lower concentration values, indicating that predictions in that period were of less (or more variable) value. These are not small differences: A prediction of $v = 0.5$ plus or minus five percentage points in the 1950s and the 2010s captures about 60% of likely voting outcomes, whereas in the 1970s-1990s the same interval only captures 40% of these outcomes.

Second, our results show that the national swing is far more important than the variation due to district uniqueness (even after accounting for the covariates), which is one reason for strong time series patterns in voter concentration. To see this, we compute the ratio of the standard deviation of the vote (on the logit scale) due to variations in national swing relative district uniqueness: $\sigma_\eta/\sigma_\gamma = 0.2/0.036 = 5.6$ (a ratio we find to be largely stable over time). Campaign observers have long known that exogenous events and heresthetical maneuvers by individual congressional candidates in their district campaigns can be important, but this result shows that exogenous national events and national-level heresthetical maneuvers are more than five times as consequential as the sum of all the individual district campaigns. All politics may well be local in its effect, but national level



(a) Concentration (vote probability mass within ± 5 points of 50% expected vote)

(b) Scale, ϕ

(c) Degrees of Freedom, ν

Figure 5: Model Features

political issues have a far bigger effect both nationally and locally than local issues (see also Hopkins 2018 and Caughey and Warshaw 2022: Sec. 3.3).

Finally, we decompose the vote concentration results from Figure 5a by noting that the ALT distribution partitions the overall variance into two parameters, the "scale" ϕ, which quantifies the amount of variation, and "degrees of freedom" ν, which controls the shape of the predictive distribution. Time series estimates of these parameters appear in Figures 5b and 5c, respectively. In both cases, we see a clear inverted U shape, revealing low variability in electoral outcomes at the start (1950s–60s) and the end of the series (2000s–2010s) and much higher variability in the middle years (1970s–1990s). The degrees of freedom parameter is similarly low at the start and end of the period, indicating sharper deviation from the normal with both longer tails and more concentration of density around the mean prediction, and higher values near the middle, indicating lower concentration.[4]

4.2 An Empirical Theory of American Democracy

The literature on American elections is increasingly scientific, but it has not always made its underlying normative assumptions transparent, which may have led to unrecognized biases and missed opportunities. We first clarify this point and then turn to a reevaluation

[4]See Supplementary Appendix 6 for additional empirical evidence of the three regimes. Note also that $\nu \approx 6$, the largest value in Figure 5c, still deviates substantially from an additive logistic normal, and both deviate from the normal.

of our empirical evidence.

Avoiding Normative Assumptions

Here we highlight the sometimes unrecognized philosophical assumptions in the literature. To do this, we begin with a simple characterization of American representative democracy as *a set of electoral rules that enables politicians to seek office by making public appeals and voters to choose among the politicians*. Importantly, the electoral rules constrain neither the arguments politicians make nor the calculus voters use in choosing candidates.

Political scientists and political philosophers have long layered on top of this simple definition various normative assumptions that they either consciously justify as important or effectively treat as facts. For example, scholars frequently ask whether voters pay attention to the important issues of the day, but they too often presumptuously define "importance" when in fact that's the voters' job. War, gun control, trade, unemployment, inflation, taxes, abortion, energy policy, and others, may sound important to political philosophers, but nowhere in American electoral rules do the normative preferences of a bunch of academics get to determine how voters make their decisions.

Similarly, when we impose our normative preferences for what counts as consistent positions across issues, voters may have a range of values of issue constraint from low to high. In fact, however, issue constraint is always "high" by definition, once we recognize again that the voters get to decide how much to count different issues in their voting decision. If voters decide only personality is important, or being pro-choice is consistent with support for the death penalty, no rule of American democracy is violated. Of course, philosophers can take normative positions, and political scientists can evaluate them systematically, but when we take on board normative views as if they are fixed features of the world, we can wind up with misleading conclusions.

These normative assumptions are so embedded in our empirical analyses that we can even miss that they are assumptions. The problem may be easier to see in older literature, on which much of our present empirical work is built. For example, consider the American Political Science Association's famous report, "Toward a more responsi-

ble two-party system" (APSA, 1950), which set the agenda for a generation of American politics researchers. The leading political scientists of the time wrote that when party positions and voter decision making are not based on the issues scholars deemed important, then "Party responsibility at the polls thus tends to vanish. This is a very serious matter, for it affects the very heartbeat of American democracy". They even clarified that "Those who suggest that elections should deal with personalities but not with programs suggest... that party membership should mean nothing at all" (APSA, 1950). (We should give the authors of this report a break, written as it was before most of the methodological developments in the social sciences, but, from a modern perspective, the report reads as breathtakingly reckless, with recommendations for numerous major reforms squeezed into single sentences, and all based on unevaluated normative assumptions and little systematic evidence.)

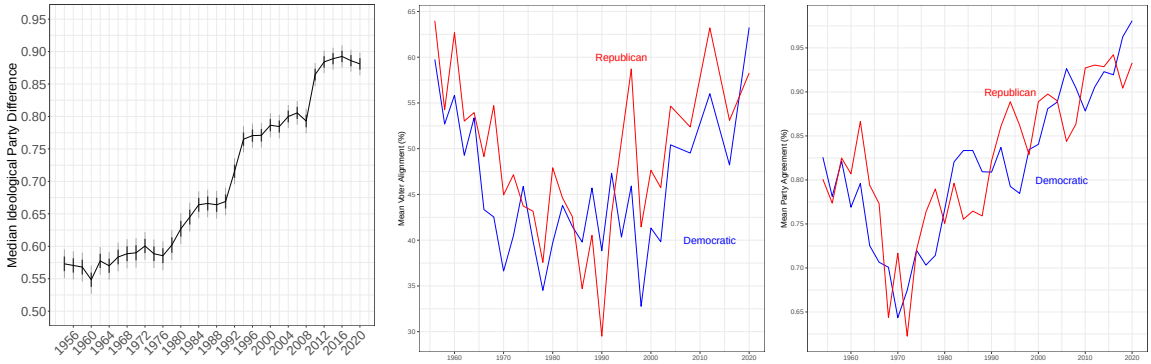
We might also ask whether these normative assumptions are merely reasonable viewpoints that no one would disagree with? After all, few have objected in the literature. For that matter, who would object to the claim that voters should cast ballots based on government programs rather than personality or temperament? Well, as it happens, we live in a representative democracy, not a direct democracy, and in most other situations where a person needs to be selected to do a job, temperament is a crucial factor. Personality evaluations are routinely made for job searches throughout the economy, choosing a romantic partner, picking an instructor, and in many other situations. Even if we could agree on the important issues of the day, new issues always arise after election day that cannot be the basis for voter decisions. In other words, one reasonable normative perspective is that voting should be based at least in part on subjects other than policy issues and programs. And whether we agree with this normative claim or not, it is perfectly consistent with the rules of American democracy: The decision makers are voters, not political philosophers.

Empirical Evidence

In Section 4.1, we described Figure 5a as showing that the distribution of vote predictions was just as concentrated around its mean in the 1950s as it is now, and much less concentrated for the years in between. From a casual reading of the literature, this result seems

awfully surprising: Where does it say that the political parties were as coherent, internally organized, and distinct from each other in the 1950s as they are today? The 1950 APSA report was designed to fix the lack of coherence in the parties, after all. How can it be that "the era of consensus," with Eisenhower as president and the parties in broad agreement over the cold war, economic prosperity, and support for international alliances like NATO, was as partisan as the 2010s and 2020s, with the gulf in ideological differences so large they seem impossible to span?

But wait, it's worse! Consider a direct measure of ideological polarization over time in Figure 6a, measured by a time series plot of the difference in DW-NOMINATE scores between the median Democratic and Republican members of the House (see McCarty, 2019). This figure shows a nearly monotonic increase in ideological polarization over the entire period, very low in the 1950s and very high in recent years. So why then would Figure 5a imply that the 1950s were highly partisan? The answer is that the 1950s were highly partisan, but the distinction between the parties was not based on the notions of ideology that political scientists and political philosophers happen to think are important. In fact, Figure 6a does not show that party polarization was at a low point in the 1950s; it highlights the failure of the political science concept of ideology to accurately describe this earlier period.



(a) Ideological Legislator Alignment

(b) Party ID–Vote Agreement

(c) Legislator-Leader Party Agreement

Figure 6: Ideological vs. Partisan Alignment

Scholars in the 1960s were aware of these patterns but they used them mostly to declare their dissatisfaction with how voters make their decisions. The leading empirical

book of the time, *The American Voter* (Campbell et al., 1980), showed empirically that voters were intensely partisan, but not very informed on the issues these political scientists decided were important. Of course, by definition, the voters were highly informed on the issues they chose to pay attention to, which can be seen by their highly predictable voting patterns. Voting decisions were based largely on partisan identification, which in turn was based on stable and measurable factors like group identities, such as race, religion, and union membership, and parental socialization.

We can also convey these basic empirical facts in simple time series plots. We do this in Figure 6b, by plotting percent agreement between party ID and the vote in ANES surveys, and Figure 6, for the percent agreement between members of the House and their party leaders (among roll call votes where leaders of one party oppose those of the other party). Both figures are characteristically U-shaped, mirroring our concentration graph in Figure 5a. (Note that the nadir of the time series comes earlier in Figure 6 than 6b, consistent with the idea that changes in voter behavior are mostly elite driven.)

Finally, note the asymmetry in the graphs we present here: for party differences, we give results among voters (Figure 6b) and legislators (Figure 6), but for ideological differences, we only present differences among legislators (Figure 6a). Why no graph for ideological differences among voters? The reason is that ideology is an idea invented by philosophers and used by political scientists; it was relatively unknown among voters until recently. In fact, questions about ideology were not even asked in the American National Election Survey until 1972 and even then prefaced with an explanation: "You may have recently heard a lot of talk about left/right…". Ideology is a normative idea that academics impose on voters, not necessarily one that voters chose to use themselves.

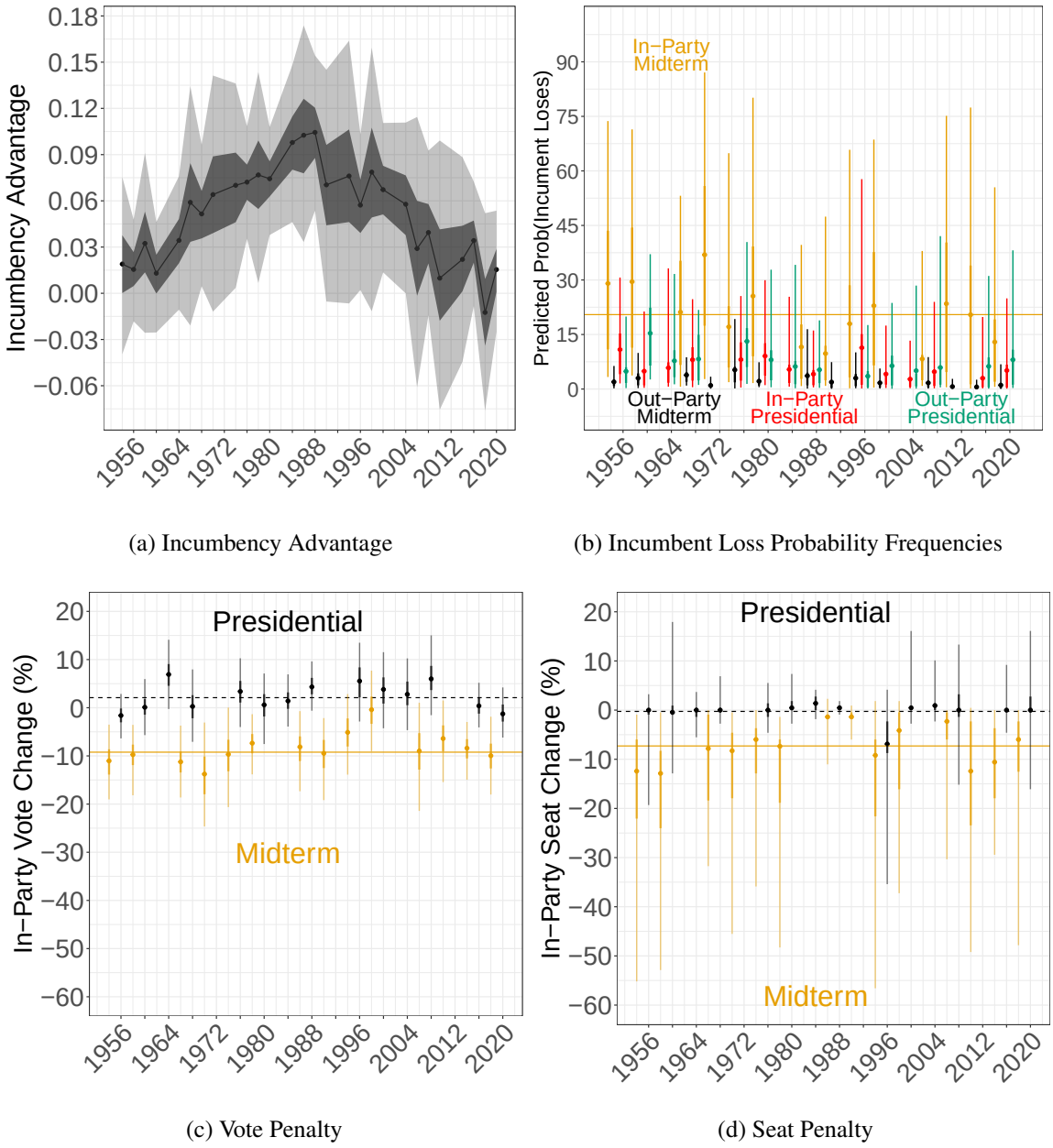
4.3 Changes in Incumbency Advantage, Stability in Incumbent Loss Probabilities

Section 4.2 shows the consequences, in terms understanding or misunderstanding empirical results, of substituting our own normative preferences for those of voters. In this section, we show the consequences of choosing a quantity of interest that we happen to find of interest and missing a related one of central importance for democracy. A funda-

mental question for any democracy is the responsiveness of its legislators to constituent preferences, and whether elections produce consequences for violating the voters' will. Mayhew (1974) famously noticed that this guarantee appeared to be breaking down in the 1970s given the decline in the number of competitive elections and what appeared to be an increase in estimates of the electoral value of incumbency (see also Abramowitz and Webster, 2016; Ferejohn, 1977). Studies of these "vanishing marginals," and corresponding increases in incumbency advantage (Gelman and King, 1990; Jacobson, 2015), were a major concern to generations of scholars. However, win margins and expected increases in incumbent votes, as important as they are in and of themselves, are only indirect indicators of the relevant quantity — *the probability that an incumbents will lose his or her job* in the next election. And it is the probability of losing one's job that is likely to be the motivating factor in keeping incumbents responsive to constituents and the whole democracy working. We show here that the broad regime changes in American politics described in Sections 4.1 and 4.2 counteract the expected advantages of incumbency, leading to long term stability in the risk of incumbents losing their seats. Moreover, this probability of loss is not only stable, it has been high over the last two-thirds of a century and across the three different electoral regimes we identify in Section 4.1, precisely because of the patterns identified there.

We begin with the familiar electoral advantage of incumbency, plotted over time in Figure 7a. For each year, the figure reports the expected vote for an incumbent minus that for a nonincumbent, with all else held constant. If we add appropriate identification assumptions, as in (Gelman and King, 1990), the vertical axis of this figure can be interpreted as an estimate of a causal effect, 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. This incumbency advantage was about two percent in the 1950s and 60s, increased to about ten percentage points in the 1980s, and then dropped back down again to around two percent by the third regime after 2000 (as noted by Jacobson, 2015).

Most of the information in incumbency advantage estimates comes from the difference



(a) Incumbency Advantage

(b) Incumbent Loss Probability Frequencies

(c) Vote Penalty

(d) Seat Penalty

Figure 7: Measures of Electoral Competitiveness

between the vote for incumbents and open seat candidates of the same parties. Each of these two components are strong functions of the national swing in any one year, which itself is of course closely related to the probability of an incumbent loss. This means that the value of the incumbency advantage, based on the difference, is mostly unrelated to the national swing. Thus, for clarity in Figure 7b, we give estimates from our model of the probability of incumbent loss for in-party members during midterm years, where there

is a well known large predictable negative national swing. The gold dots in this figure represent the average incumbent loss probability for all in-party midterm incumbents each year, with thick bars corresponding to the central 50% of the district loss probabilities and thin bars capturing 95% of them (i.e., these are not confidence intervals, representing uncertainty; they instead describe the distribution of district-level probabilities).

As Figure 7b reveals, the in-party midterm loss probabilities are large and variable, but do not trend over time. The average probability of an in-party incumbent loss during a midterm, represented by the horizontal gold line, is a substantial 20.6%. The vertical lines through the dots indicate that many incumbents have much higher probabilities of losing their jobs, which is indicated by the high end of the asymmetric intervals around the gold dots. The other three logical subsets have much lower average loss probabilities; these include in-party presidential in red, out-party midterm in black, and out-party presidential in green. Although these other three subsets have very small average loss probabilities, the competitiveness of the presidential election means that incumbents will sometimes wind up facing voters with a remarkable one-in-five chance of losing their jobs. Of course, nonincumbents in open seat races have much higher probabilities of losing and incumbent challengers's chances of losing are higher still. If you are or hope to be a tenured professor, think of how much more you might pay attention to the chair of your department, review committee, and students if every four or eight years one in five tenured professors where summarily fired. Your laurels wouldn't be very restful. Of course, this is excellent news for the incentives American democracy provides to its elected legislators to be responsive to their constituents.

Why, then, does incumbency advantage change so dramatically in Figure 7a even as the probability of incumbent loss remains so stable in Figure 7b? Indeed, these seemingly contradictory results are both computed from the same run of the same generative model. The answer comes from the results in Section 4.1: When incumbency advantage is low, near the beginning and end of our 66 year data set, variation is low and voter concentration is high, meaning that even a 2 percentage point incumbency advantage has some substantial value. When the expected advantage of incumbency rises to roughly 10 percentage

points in the middle of the period, the concentration and thus the value of that expected vote decreases by twice as much (from about 60% at the start and end, to only 40% in the middle, of Figure 5a). This increasing variability means that incumbents see little actual reduction in their probability of losing office. Getting a bonus of 10 percentage points (because you're an incumbent) may seem comforting, but if this "bonus" also comes with a much larger random component its value is degraded (see also Jacobson, 2015).

Finally, we summarize the consequences of these probabilities for the in-party's loss of votes in Figure 7c, and seats in Figure 7d. The average loss during midterm elections, represented by the gold flat lines, reflects about an 8.1 percentage point vote loss (± 2.3 points, a 95% CI) and 8.8 percentage point seat loss (± 1.7 points). These average effects are substantial, but do not miss the occasionally large and highly asymmetric confidence intervals in Figure 7d, meaning that we should also expect occasional extremely large in-party seat losses.

Consistent with Figure 7b, we also see little to no in-party vote or seat change during presidential years, which is reflected in the black dots and lines in Figures 7c and 7d.

5 Generatively Accurate Descriptive Summaries

We attempt in this paper to build *generatively accurate descriptive summaries* of our data, reducing the tremendous complexity of American politics and congressional elections to understandable summaries computed from a single internally consistent statistical model. While the cost of working with generative models is the modeling assumptions, the benefits include rigorous out-of-sample validation (see Section 3) and a far richer range of substantive political science questions that can be tackled, a topic we take up in this section.

Description is sometimes regarded as separate from inference and unaffected by the usual threats to proper statistical analysis (i.e., often as long as you say you're doing "mere" description, anything goes). In practice, however, the best descriptive summaries are those vulnerable to being proven wrong (and then ideally not actually wrong) and tailored to the many precise questions of substantive interest. In fact, descriptive summaries

are essential to addressing the breathtaking range of questions of interest to social scientists. Scholarship should not be limited to quantities that happen to be computationally or statistically convenient, or those in whatever methodological area happens to have made progress lately (such as causal inference in recent years; see Supplementary Appendix 7).

We outline in this section some of quantities that can be estimated from a generatively accurate model and explain how they can be used to enrich political science research. As inference is simply "using facts we know to learn about facts we do not know," we characterize the types of quantities we may wish to estimate by first detailing both the "unknowns" that may be of interest and then the "knowns" we have available to condition on. We characterize the unknowns in three ways. First, the *location* of an unknown is where the values are of the outcome variable that we want to know. This may involve a "forecast", i.e., into the future; "farcast," i.e., to an election in the present or past not in our dataset (such as for a different office or country); "nowcast", i.e., to unobserved features of elections in our dataset (such as the posterior distribution for a district vote, only one value of which is observed, as in posterior predictive checks); or even a "faroutcast," which refers to values of the outcome variable under counterfactual conditions (such as if no incumbents had run). Second is the *level* of aggregation of the quantity of interest, such as for district-, state-, regional-, or national-level statistics, or features of non-geographic groupings like all Democratic districts or all those without an incumbent. Finally, our quantities of interest involve a *concept*, such as partisan bias, electoral responsiveness, the probability that an incumbent will lose, the expected vote in a district, or the district vote of the median legislator.

Quantities of interest always condition on three types of features that are either known or, in the case of counterfactuals, assumed known. These include (1) the choice of covariates and their values for unknown quantities; (2) keeping, removing, or zeroing out random effects; and (3) keeping, removing, or adjusting surprises (such as to focus only on the expected value or other features of the posterior). We explain how to make decisions for choosing quantities of interest, such as those given in Section 4, and how to mix and match the location, level, and concept of the unknowns with the covariates, random

effects, or error term surprises to condition on.

Consider estimating the probability that a Democrat wins a particular district i in election year t. The simplest case is a "nowcast". Here we are interested in the ex ante probability that the Democrat would this district election, which in fact we have already observed (ex post). To do this, we set the values of X to their observed district values. For example, suppose the lagged vote for the Democrat incumbent is 74%. We might then consider setting $\beta_{tk} = \hat{\beta_{tk}}$ for all k covariates, and $\gamma_i = \hat{\gamma_i}$ and $\eta_t = \hat{\eta_t}$ for the random effects. Of course, we do not know any of these numbers for certain ex ante and, therefore, we would choose instead to include estimation uncertainty in our estimate of the probability that the Democrat would win this district. Thus, instead of fixing these parameters at their point estimates, we draw them from their posteriors, centered on the estimated values. Then, suppose we take 1,000 draws from this posterior; to generate our model-based "nowcast" of v_{it}, we then multiply each of these draws by the relevant X_{itk} and add the draw of the district effect and national swing. We then run this sum through the inverse logit function to get a hypothetical draw on the scale of votes. This generates 1,000 hypothetical draws of v_{it}. To estimate the probability a Democrat wins the district, we then simply count up the fraction of the draws greater than 0.5.

For forecasting, the choices about how to construct generatively accruate descriptive statistics is more flexible, and thus more complicated. Consider the simplest case. If all we want is a one-election-ahead forecast, we still have a number of important decisions to make. For example, how do we set the covariate values? If it is the next election, we have observed the lagged vote, so that is straightforward to use, but what about incumbency? Do we know yet if the current incumbent will run again? If so, we could use that value. But if we are making the forecast well before the election, and do not know this yet, what value should we use? We could assume they all run, or some randomly selected proportion run again. Perhaps, however, it is better to consider what would happen if the district were open? Regardless, the analyst must choose some value relevant to the question at hand, and must realize that this choice changes the question we are answering. In fact, the differences in forecasts across these assumptions may be of considerable substantive

value.

We also have important decisions about the district effect, γ_i. If this is were only one election ahead, and we do not think much else has changed, then we may want to fix $\gamma_i = \hat{\gamma}_i$ as we did in our "nowcast" above. However, we surely do not know β_{tk}. So instead we need to use draws of it from $\beta_{tk} \sim \mathcal{N}(\hat{\beta}_k, \sigma^2_{\beta_k})$. This will add additional uncertainty to our forecast, but is otherwise similar to our "nowcast". And we are unlikely to know the national swing and so must make a parallel choice about η_t. Given all these assumptions, we can then generate our hypothetical election draws and calculate the fraction of times the Democrat wins as our prediction as a probability.

Perhaps the most difficult set of choices comes from in making a "faroutcast", as for example, when we want to forecast what would happen in a new legislative map following the implementation of a proposed (or perhaps recently passed) redistricting plan. Here, the covariate choices are not obvious. First, we generally will not know where incumbents will be, but perhaps we can make some educated guesses. Alternatively, we might assume all seats are open to obtain a baseline probability that a Democratic candidate could win the seat. Harder yet, is what to do about lagged vote, which is giving the model a measure of the normal vote in the district. We could use precinct level returns for the previous election, subtract out the incumbency advantage and uncontestedness, aggregate into the new districts, and then add back in the incumbency advantage for districts where the decisions of incumbents and challengers is known. Or perhaps we could use presidential vote, or some average of statewide votes re-aggregated in the new district map. These constructed measures would be needed in the original model or in a separate model that imputes lagged vote from some statewide measures. Also, as with our forecast, we do do not know β_{tk}, γ_i, or η_t. And as before we could then generate our hypothetical election draws and calculate the fraction of times the Democrat won.

The flexibility of the model easily enables one to calculate even more sophisticated quantities of interest. For example, one of the largest sources of uncertainty in election predictions is the national swing. We can thus draw η_t directly from its posterior. Alternatively, we can model the parameters of the η_t prior with national-level covariates, such as

unemployment, presidential approval, or whether the country is at war. Yet another option is to fix it at the value of some previous election that seems similar to the current one.

By combining the location, level, and concept for a quantity of interest and fine tuning by making choices about the covariates, random effects, and surprises, accurate generative models like the one we describe here can reveal a vast amount about American elections, far richer than any one specific estimate or data analysis on its own.

6 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 observable implications of these models 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, and building on new statistical and computational technology not previously available. We validate our approach with extensive out-of-sample (and distribution-free) tests in 14,710 district-level elections. Our generative model is general in that it can be used, with the appropriate additional assumptions and covariates when necessary, 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 such as the incumbency advantage. We then build a more general model of American democracy consistent with these findings.

Further growth in computational power may one day enable feasible estimation of joint generative models that enable a richer substantive portrait of the electoral system, such as conducting modeling at the precinct-level to include redistricting periods, or encompassing other elections such as for the US senate, president, and state legislatures.

With a continual focus on rigorous out-of-sample validation, and larger generative models, it may even be possible, one day, to estimate these simultaneous with other sectors of society such as the economy, demography, public policy, and public health, or potentially data from other countries.

Appendix A Statistical Details

This appendix provides the full likelihood function for our model, including all the features described in Section 2.2, as well as situations where the effective vote is both included in the model as a lagged covariate and unobserved (because previous election was uncontested).

To write the full likelihood function, define an uncontestedness indicator U_{it} as 1 if the Democrat runs uncontested, 0 if contested, and -1 if the Republican runs uncontested in district i and time t. Then partition elections into four sets depending on whether the current election i, t and its lag $i, t-1$ are contested or uncontested. Denote CC as the set of all elections for which $U_{it} = 0$ and $U_{i,t-1} = 0$; UC as the set of elections for which $U_{it} \neq 0$ and $U_{i,t-1} = 0$; CU as the set of elections where $U_{i,t} = 0$ and $U_{i,t-1} \neq 0$; and UU as the set of elections for which $U_{it} \neq 0$ and $U_{i,t-1} \neq 0$. Then the likelihood function factors into four parts corresponding to these sets:

$$L = \left(\prod_{i,t \in \{\text{CC}\}} L_{it}^{\text{CC}} \right) \left(\prod_{i,t \in \{\text{UC}\}} L_{it}^{\text{UC}} \right) \left(\prod_{i,t \in \{\text{CU}\}} L_{it}^{\text{CU}} \right) \left(\prod_{i,t \in \{\text{UU}\}} L_{it}^{\text{UU}} \right) \tag{5}$$

each of which we now define.

The first component of the likelihood, for when election i, t and $i, t-1$ are both contested, is by far the most prevalent for the US congress. The likelihood for observation i, t is then simply

$$L_{it}^{\text{CC}} = \text{ALT}(v_{it} \mid \mu_{it}, \phi_t^2, \nu_t). \tag{6}$$

The second component of the likelihood accounts for which party is running uncontested at time t:

$$L_{it}^{\text{UC}} = \mathbf{1}(U_{it} = 1)\psi_{it} + \mathbf{1}(U_{it} = -1)(1 - \psi_{it}), \tag{7}$$

where our censoring assumption from Section 2.2.3 implies that $\psi_{it} \equiv \int_0^{0.5} \text{ALT}(v^* \mid \mu_{it}, \phi_t^2, \nu_t) dv^*$, given the indicator function defined as $\mathbf{1}(a) = 1$ if a is true and 0 otherwise, for any statement a.

To write the third component, where the lagged value of the effective vote is unobserved (because it is uncontested), we require a prior distribution for how this variable is distributed. The posterior will be computed from the entire model, but to begin we need an assumption about this prior. One option is to let $v_{i,t-1}^*$ be a censored ALT when unobserved (and equal to v_{it} when observed) but this creates a substantial computational burden with little substantive benefit. Instead, we find we can represent almost all relevant information by assuming that, when unobserved, $v_{i,t-1}^* \sim \mathcal{N}(Z_{i,t-1}\alpha_t, \sigma_v^2)$, with $Z_{i,t-1}$ a vector of covariates such as lagged presidential vote in a congressional district and incumbency status. Then this component of the likelihood is

$$L_{it}^{\text{CU}} = \int_{-\infty}^{\infty} \text{ALT}(v_{it} \mid \mu_{i,t}, \phi_t^2, \nu_t) \cdot \mathcal{N}(v^* \mid Z_{i,t-1}\alpha_t, \sigma_v^2) dv^*, \tag{8}$$

where the unobserved lagged effective vote v^* is included in X and so contributes to μ_{it}.

For the final component of the likelihood, we use features of all three previous components, so that

$$L_{it}^{\text{UU}} = \mathbf{1}(U_{it} = 1)\psi'_{it} + \mathbf{1}(U_{it} = -1)(1 - \psi'_{it}), \tag{9}$$

where

$$\psi' = \int_{-\infty}^{\infty} \int_0^{0.5} \text{ALT}(v \mid \mu_{i,t}, \phi_t^2, \nu_t) dv \cdot \mathcal{N}(v^* \mid Z_{i,t-1}\alpha_t, \sigma_v^2) dv^*.$$

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* 41, pp. 12–22.

APSA (1950). *Toward a More Responsible Two-Party System. A Report of the Committee on Political Parties of the American Political Science Association*.

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

Campbell, Angus, Philip E Converse, Warren E Miller, and Donald E Stokes (1980). *The american voter*. University of Chicago Press.

Caughey, Devin and Christopher Warshaw (2022). *Dynamic Democracy: Public Opinion, Elections, and Policymaking in the American States*. University of Chicago Press.

Ferejohn, John A (1977). "On the decline of competition in congressional elections". In: *American Political Science Review* 71.1, pp. 166–176.

Gelman, Andrew, J.B. Carlin, H.S. Stern, and D.B. Rubin (1995). *Bayesian Data Analysis*. Chapman and Hall.

Gelman, Andrew and Gary King (Nov. 1990). "Estimating Incumbency Advantage Without Bias". In: *American Journal of Political Science* 34.4, 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* 38.2, 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.

Hopkins, Daniel J (2018). *The increasingly United States: How and why American political behavior nationalized*. University of Chicago Press.

Jacobson, Gary C (2015). "It's nothing personal: The decline of the incumbency advantage in US House elections". In: *The Journal of Politics* 77.3, 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* 114.1, 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* 93.1, 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.

King, Gary and Andrew Gelman (Feb. 1991). "Systemic Consequences of Incumbency Advantage in the U.S. House". In: *American Journal of Political Science* 35.1, pp. 110–138. URL: bit.ly/SystCs.

Mayhew, David R (1974). "Congressional elections: The case of the vanishing marginals". In: *Polity* 6.3, pp. 295–317.

McCarty, Nolan (2019). *Polarization: What everyone needs to know®*. Oxford University Press.

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

Shepsle, Kenneth A (2003). "Losers in politics (and how they sometimes become winners): William Riker's heresthetic". In: *Perspectives on politics* 1.2, pp. 307–315.

Vovk, Vladimir, Alexander Gammerman, and Glenn Shafer (2005). *Algorithmic learning in a random world*. Springer Science & Business Media.

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