Hi everybody,

I am going to talk about statistical methods use to forecast elections.

So after introducing the topic, I will discuss a little bit about the main data used in election forcasting and then I will discuss some statistical models.

Election forecasting is a strategical task, especially for policy makers.

The challenge is being able to produce election forecasts timely and accurately.

I said election forecasts but I am focusing on US elections only.

Usually, two types of data are used to model election results, the so-called fundamental variables and trial-heat polls.

The fundamental indicators are related to the economic or the political dimensions. The assumption is that the voting process is influenced by the healtyness of the economy but also, obviously, by the political sphere.   
Many models have been developed during the years to forecast elections using only this kind of data, these are the so-called structural models.

Trial-heat polls, instead, are surveys in which at least one trial-heat question is asked. That is a preference question on future presidency.

This kind of data is nowadays very frequent since there are many pollesters agency, covering opinion in every state, but it suffers of some well-know problems, in particular this data shows high variability within the election year.

So now we are going to see how this data can be used to actually forecast elections.

First of all, there exists mainly three braod category of election models.

The structural models, econometrics models using only fundamental variables.

The trail-heat models, again econometrics models using polls data.

And the bayesian approach, that actually tries to combine the two.

For simplicity, let assume that the variable of interest is always the share of vote in one of the two major parties.

So first a structural model, and actually one of the most succesful. It is a national model that relates the share of vote to the GDP, the presidential approval rate and the so-called time-for-change variable.   
This model assumes that voters positively evaluate periodic government alternation which is captured by the time for change variable.

The drawback of this approach is that it relies only on fundamental variables, so it does not make use of current available information.

This is actually the purpose of trial-heat models, since most of the time they are just extensions of structural models that exploit current information using the available poll data.

For instance, Campbell proposed a simple national model that uses national poll data and GDP.

Gelman & King instead developed a more complex model based on both national and state variables. So in this case, it is a state model, that is we are estimating the share of vote in each state, but the poll data is still used on a national level.

The bayesian approach, instead, tries to combine the two previous methods actually using poll data to update forecasts produced by an historical model. This is the core idea of the following models.

The first is a model averaging weighting two different models. The hist equation represents structural model and the poll equation is the polling model, that is the share of vote is related to the percentage of survey respondents for that major party. Normality is assumed for the errors terms and the prediction is a weighted average where the weights are based on the proportion of the variances of the error terms. So that, the relative weight attached to each forecast is larger when its associated forecast error is smaller.

It is still a national model, but it updates the historical prediction every time new polls are available.

The second model is a Dirichlet-Multinomial conjugate model. It is assumed that election's outcomes are based on historical voting trends only, hence no structural variables are used.

The prior is based on past election results and the share of vote for each candidate is assumed to follow a Dirichlet distribution, while the likelihood is based on poll data and the sample proportions of a survey with n respondents is assumed to follow a Multinomial distribution. So that, the posterior is still a Dirichlet with updated parameters.

The calibration and the choice of parameters are based on historical election reasoning (expected normal votes for that election, average third party in previous elections, 3% undecided).

This model is actually the only one that incorporates third-parties and undecided.

Lock and Gelman, instead proposed a model with normal prior with the parameters estimated on past election results and a normal likelihood with parameters estimated on poll data. Normality is justified by the general lack of outliers in state election results and also by the large sample size of each poll. The predictions are draws from the posterior distribution which is a normal distribution where the mean is a weigthed average of the historical and poll estimates, weighted by their precisions. So higher prior precision places more weight on past election results, while higher likelihood precision places more weight on current poll data.

The last model, combines the historical forecasts based on structural variables with daily state-level poll data. The share of vote is based on two components, beta represents the historical voting preferences and delta is a national effect and both are assumed to be normally distributed.   
To obtain an estimate for those days when no polls are available, a reverse random walk strategy is used, that is starting on election day, the estimated voting preferences evolve randomly going back in time. The historical forecasts are obtained using the Time-for-Change model and are incorporated through a Normal prior for the final outcome and the higher the varibility the lower the weight attached to historical prediction over new polling data. In particular, when the election is soon estimates of the share of vote are based primarily on poll data, whereas if the election is farther ahead are driven by the historical prior.

To conclude, the bayesian approach allows to continuously update forecasts as new poll data is available.   
Usually using both structural variables and polls data outperform others method, but in general forecasts are accurate only within two month before the election day, which is not really an early forecast.

So there are still problems, it is difficult to produce timely and accurate forecast and also there is low accuracy and high uncertainty for states that are polled few and in those days with no polls at all.

Two suggestions to try to improve the early forecast accuracy and to solve the lack of data issue are:

using web data as another source of information to estimate public opinion on the share of vote. And also, none of the previous models take into account some possible correlation among voting preferences across states.