UNIVERSITÀ
CATTOLICA
del Sacro Cuore

DIPARTIMENTO DI SCIENZE ECONOMICHE E SOCIALI

Forecasting the United States Election Through Primary Vote Results

Sebastiano Mars
Elena Calegari
Luca Bagnato

Quaderno n. 150/febbraio 2021

VP VITA E PENSIERO

Università Cattolica del Sacro Cuore

DIPARTIMENTO DI SCIENZE ECONOMICHE E SOCIALI

Forecasting the United States Election Through Primary Vote Results

Sebastiano Mars
Elena Calegari
Luca Bagnato

Quaderno n. 150/febbraio 2021

VITA E PENSIERO

Sebastiano Mars, Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, Piacenza.

Elena Calegari, Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, Piacenza.

Luca Bagnato, Dipartimento di Scienze Economiche e Sociali, Università Cattolica del Sacro Cuore, Piacenza (corresponding author).

✉ mars.sebastiano@gmail.com

✉ elena.calegari@unicatt.it

✉ luca.bagnato@unicatt.it

I quaderni possono essere richiesti a:
Dipartimento di Scienze Economiche e Sociali,
Università Cattolica del Sacro Cuore
Via Emilia Parmense 84 - 29122 Piacenza - Tel. 0523 599.342
http://dipartimenti.unicatt.it/dises

✉ dises-pc@unicatt.it

www.vitaepensiero.it

ISBN 978-88-343-4548-1

Abstract

In this work we present a model, named logistic primary model (LPM), which aims to describe the probability of a Democratic (Republican) victory at the U.S. presidential elections. The proposed model is based on a logistic regression with a unique regressor and exploits the primary results of the candidates to the White House. It follows the idea of the existing primary model (PM) proposed by Helmut Norpoth since 2004, which is a ARIMAX model for the two-party popular votes obtained by the Democratic Party. The LPM, applied to the U.S. election data 1912-2012, shows good performances both in terms of goodness-of-fit and forecasting. In addition, the paper presents an extensive review of the electoral forecasting models proposed in the literature.

Keywords: U.S. Electoral System, Time Series Analysis, Primary Model, Logistic Regression
JEL Classification: C53, C22, D7.

1. Introduction

The U.S. presidential election can be considered as a pivotal case for election forecasting models. Indeed, the U.S. represents one of the world's most powerful democracy and, on the other hand, the characteristics of the U.S. electoral system, such as the regular time span between elections, allow for methodological considerations (Linzer and Lewis-Beck, 2015). The roots of the literature on the forecast of the U.S. electoral outcome date back in the early 1980s (Lewis-Beck and Rice, 1992) and encompass several types of models that have been updated over time. The proposed models differ with respect to the analyzed outcome, the econometric specification, the forecast horizon and the included predictors. In particular, prior to the 2020, most of the models focused on predicting the percentage of votes for one of the two main parties at national level through linear models (for a review on popular vote models see Holbrook (2010)), whereas only rare examples predicted the electoral college votes by state employing different econometric strategies (Linzer, 2013). The forecast horizon widely varies among models being, for instance, of 78 days as median with a range of 60 to 246 days for the 2016 election (Jennings et al., 2020). Regarding the included predictors, a relevant subset of models includes variables describing the state of the economy and the approval rate of the incumbent president: see, for example, Cuzán (2012), Abramowitz (2016), Lewis-Beck and Tien (2016), Lockerbie (2016), Fair (2011), and Hibbs (2012). For a recent review see, among others, Cuzán (2020).

Despite their impressive record of successful forecasts, these models have been criticized because of some theoretical limitations: i) most of the models assume that voting is retrospective, i.e. it is based on how the economy and the incumbent president performed during the mandate, and pay only little attention to the electoral campaign; ii) most of the models are centered on the incumbent and on the judgment from the electorate, but do not explicitly consider how the opponent is perceived from the voters; iii) most models overlook measurement error affecting the variables used to describe the state of the economy and the misperception of citizens about the economic situation (Graefe, 2013).

Moreover, after the 2016 election, some general criticisms about U.S. presidential elections forecasting emerged. First, the necessity to account for the fact that, in the U.S. electoral system, the candidate that gains the majority of the popular vote may not become president and, secondly, it has been argued that in particular conditions the explanatory power of the economical predictors may be reduced (Abramowitz, 2020).

In trying to overcome some of the theoretical limitations, a stream of literature proposes to put the personality and leadership of the candidates, as well as their ability to focus on issues relevant to the public - such as economic and social conditions -, on the forefront (Graefe and Armstrong, 2012). In these models the variables that measure the strength of these subjective perceptions are based on opinion polls, registered as close as possible to the election day (Graefe et al, 2014). Following this intuition, Graefe (2013) proposes an "Issue and Leader Model", that includes only variables obtained by opinion polls, whereas other models propose to mix both economic and polls information. Among these, for example, Campbell (2016), Campbell et al. (2017), Holbrook (2016), and Erikson and Wlezien (2016). However, if it is true that the opinion polls can be a useful tool for electoral forecast models, it is also true that they introduce additional uncertainty in the predictions, arising some critics (Linzer and Lewis-Beck, 2015; Jennings and Wlezien, 2018; Shirani-Mehr, 2018). These and other existing electoral models had been extensively reviewed in Appendix A.

A pioneering model that accounts for the candidates' appeal during the electoral campaign without the employment of opinion polls, is the "Primary Model" (PM) proposed by Helmut Norpoth since 2004. The PM employs the outcome of the earliest primary elections in both the major parties as the fundamental ingredient to predict the share of the popular vote in the presidential election (Norpoth 2004, 2016). The intuition is that a strong win in the first primaries signals a leader with a personality appealing to the public and able to focus on issues that people consider important. The PM has three main merits: first, it can be used to forecast the outcome of the presidential elections many months in advance (early primaries take place usually in February), second, it considers both the incumbent and the opponent strength and, third, it is straightforward and uses predictors that are very easy to obtain. However, the PM has also some limitations: in forecasting the vote share of one of the two largest parties at the national level, it neglects the possibility that the candidate that attains the largest share of the votes is not the elected one, as in the case of the 2016 election.

The present paper aims at contributing to the literature on the forecast models for the U.S. presidential elections by proposing a modification of the PM, named Logistic Primary Model (LPM). The LPM borrows the strong points of the PM, such as the intuition regarding the predictive power of the results of the primary elections avoiding the drawbacks of the opinion polls, and the consequent possibility to forecast the outcome of the elections many months in advance. However, the LPM differs from the Norpoth (2016) proposal under three respects: i) it targets the election outcome and not the share of votes from one party; ii) it relies on a single predictor that reflects how stronger (weaker) the Democratic candidate is with respect to the Republican candidate, by evaluating their respective primary performances; iii) it is based on logistic instead of linear regression.

The proposed modifications have two main objectives. First, according to the principle of parsimony, they aim at simplifying the model, especially with respect to the construction of the predictor variables, and second, they aim at improving the forecasting ability, accounting for the possible discrepancy between the share of electoral vote and the actually elected candidate.

The paper is organized as follows: in Section 2 we recall the PM while in Section 3 we present our proposal. In Section 4 we evaluate, via an application to real data, the performance in terms of forecasting of the proposed model. We conclude the paper with a brief discussion in Section 5.

2. *The Primary Model*

The PM, developed by Norpoth (2016), is a five-variables and second-order autoregressive model, specified as follows:

$$DEM_t = c + \beta_1 DPS_t + \beta_2 RPS_t + \beta_3 DEM_{t-1} + \beta_4 DEM_{t-2} + \beta_5 P + \varepsilon, \quad (1)$$

where the considered electoral outcome DEM_t is the share of two-party popular votes obtained by the Democratic Party and the predictors are:

- DPS_t is the primary support received by the Democratic candidate. It is defined as the ratio between the votes that the winner obtains in the earliest democratic primary elections and the sum of votes, in the same elections, obtained by the first two most voted candidates;

- RPS_t is the primary support received by the Repubblican candidate, defined analogously to DPS_t;

- DEM_{t-1} and DEM_{t-2} are the shares of two-party popular votes obtained by the Democratic Party in the two previous presidential elections;

- P is the dummy variable reflecting the social changes that took place after the New Deal realignment, leading to a major shift in the electorate among the two main parties. It is coded 1 for years prior to 1940 and 0 thereafter.

The model has been estimated by Norpoth (2016) using data from the 1912 elections onwards. Up to the 1948 all the primaries were used in the computations of DPS_t and RPS_t. From 1952 to 2004 only the primaries in New Hampshire and from 2016 onwards the primaries in New Hampshire and South Carolina have been considered.

There are at least two major assumptions underlying the theoretical framework of the model. Firstly, the model embraces the idea that voters display a cyclical behavior: after a while people seek change, and therefore will turn their preferences despite their precedents. Because the U.S. voting system has just two main parties, predicting the elections can be compared to the prediction of the swings of a pendulum. Besides the "electoral cycle" theory (Norpoth, 2014), the second argument posed by Norpoth (2016) is that primaries are a leading predictor when it comes to electoral forecasts. The general idea is that the major portion of electorate already expressed its preferences during the primary contest: after that, only the minor undecided portion of people will shift their preferences toward the Democratic or Republican candidate.

These two theories concretize into two sets of variables. Two lagged values of the Democratic share of popular votes are used to detect the actual position of the electoral pendulum, while primary performances are encoded into two peculiar metrics. Finally, a dummy variable reflecting the changes across the historical social contest is included.

The PM uses very simple predictors that allow the model to be estimated on a large number of elections (since 1912); moreover, as based on the earliest primaries, that usually take place in February of each election year, it allows for a long-horizon forecast (approximately 9 months).

3. The Logistic Primary Model

Despite its merits, such as the idea of analyzing primaries as potential predictors, several improvements of the PM are possible.

In the first place, the choice of predicting the share of the democratic party vote is not the same as predicting the presidential election winner: because of the peculiarities of the U.S. electoral voting system, the candidate that attains the largest share of the votes can ultimately lose the race, as it happened in the 2016 elections when Hillary Clinton took more votes nationwide than the effective winner Donald Trump.

Secondly, linear models are probably not the best choice when describing percentages across time, because the forecast could be meaningless: for example, by predicting values lower than zero percent or higher than one hundred percent.

Thirdly, in the PM not all the included variables carry the same predictive power, as demonstrated by the fact we can obtain similar results by dropping all predictors except for primaries. The primary electoral contest alone has enough predictive power to be encoded in a single variable.

In this section we propose a modified PM in which we directly target the outcome of the presidential election using a logistic regression. Moreover, we reconsider the set of predictors that lead us to a simplification of the model.

3.1. The Model

The LPM aims to describe the probability of a Democratic victory at time t (DW_t), through a unique predictor, namely the Net Primary Score at time t (NPS_t) of the Democratic Candidate with respect to the Republican candidate. We define the variable NPS_t as the difference between the Democratic Primary Strength at time t (DPS_t) and the Republican Primary Strength at time t (RPS_t), that is:

$$NPS_t = DPS_t - RPS_t \, . \tag{2}$$

The calculation of the variables DPS_t and RPS_t is described in the following sections and is based on a measure of relative strength of each presidential candidate within his/her own party (Democratic or Republican). More specifically, the Candidate Primary Strength at time t (CPS_t), is defined as follow:

$$CPS_t = \frac{CNV_t}{CNV_t + CCNV_t} \, , \tag{3}$$

where CNV_t is number of votes in the primary contest in which the presidential candidate took part, being either the Democratic or Republican party primaries, and $CCNV_t$ is the number of votes, in the same primary contest, of the chief rival of the candidate.

The LPM is defined with a logistic specification (Hosmer, 2013), as follows:

$$logit(DW_t) = \beta_0 + \beta_1 NPS_t + \epsilon_t \quad , \tag{4}$$

where $logit(p) = \log\left(\frac{p}{1-p}\right)$. The idea behind this short formulation is simple: in order to predict the chances of winning for the Democratic Party, given that primary elections results are an efficient predictor of the presidential election, it is sufficient to measure the difference between presidential candidates in terms candidates' relative strength within the party they belong to.

The LPM formulation has one leading advantage over the original PM: every reliance on popular votes is abandoned. The model explains the probabilities for a candidate of being elected by the electoral college, regardless the popular votes obtained in the election. Another advantage is the reduction of uncertainty provided by the prediction intervals. The use of an ARIMAX model in the PM, and a time series consisting of less than 30 observations, make predictions enclosed in very large prediction intervals.

3.2. Model setting and data

The first empirical choice that characterizes the implementation of the LPM is the choice of which primary elections to consider in the calculation of the predictor variable. As described in Section 2, Norpoth (2016) includes all the primaries for electoral tournaments between 1912 and 1948, only the primaries in New Hampshire from 1952 to 2008 and from 2016 the primaries in New Hampshire and South Carolina.

After testing for correlation between primary scores and the popular votes obtained during the general election in different states, we endorse the choice of Norpoth (2016) to use New Hampshire and South Carolina primaries results in recent times. Norpoth (2016) motivates this choice with two main arguments. First, New Hampshire and South Carolina are among the earliest states where the primaries take place (early February and end of February, respectively) and, second, they display a very different demographical composition, proved to be a relevant determinant in the voting attitudes (Greenwald, 2009; Zeedan, 2019). Our empirical results thus confirm the theoretical choice.

The main difference that emerges with respect to the original version of the PM lies in the time periods adopted. Indeed, our results suggest including the South Carolina from 2008 rather than 2016. In particular, the primary results used in the LPM are the following: a) 1912-1948: all primaries are included in the model; b) 1952-2004: only primary results in New Hampshire are included; c) 2008-2016: both primary results in New Hampshire and South Carolina are included. According to the defined setting, the present model is based on 26 elections (from 1912 to 2012) with two leading candidates: Democrats and Republicans. As in the PM, third parties are not considered in the computation.

The second empirical choice in the implementation of the LPM is the definition of the dependent variable as a dichotomous variable given by the result of the general election (1 in case of a winning Democratic Party, 0 in case of a losing Democratic Party).

Finally, the last set of decisions regards the theoretical and empirical choices that lead to the definition and the computation of the predictor of the model, defined Net Primary Score (NPS), as shown in Section 3.1. According to the assumptions of the LPM, the predictor must reflect the strength of the Democratic presidential candidate with respect to the Republican one based on the relative strength that each candidate exhibits during the primary contest of his/her own party. Therefore, as first, it is necessary to define a metric to measure the relative strength of a candidate through the information on his/her primary performance in terms of number of votes. To do so, the LPM follows the metric suggested by Norpoth (2016), that defines the relative strength of a candidate as the ratio of the candidate's votes over the sum of his/her votes with his/her chief's rival votes, as shown in Eq. 3. The proposed metric has indeed shown good forecasting performances (Cuzán, 2020). At this scope, the series of Candidate Primary Strength (CPS) has been calculated, irrespective to the belonging party of the candidates. A further note is that, as the PM, the LPM, in defining the relative strength of both candidates, accounts for the fact that the incumbent candidate is a sitting president or not, as discussed in section 3.5.

Starting from this framework, the next sections illustrate the empirical validation of the assumptions of the LPM and the calculation of the NPS together with empirical strategy applied to address the related issues.

3.3. Preliminary analysis

To empirically validate the key assumption of the LPM, according to which there is a strong relationship between the CPS and the votes obtained in the presidential election, the CPS has been divided between Incumbent Primary Scores (IPS) and Opposition Primary Scores (OPS). The IPS are historic CPS obtained by incumbent candidates being them Democrats or Republicans. Incumbent candidates are all those candidates belonging to the current ruling party (Democratic or Republican). The OPS are historic CPS obtained by opposition candidates to the presidency, being them Democrats or Republicans either. Figure 1 shows the scatterplots between the IPS and the general election share of votes for each incumbent candidate, as well as the scatterplot between the OPS and the general election share of votes. What clearly emerges looking at Figure 1, is that the higher is the incumbent candidate primary relative strength, the more are the votes for the incumbent party at the presidential elections. On the contrary, the higher is the opposition candidate primary relative strength, the fewer are the votes for the incumbent party. This empirical relationship thus enhances the intuition that primaries can be a useful predictor for the general election.

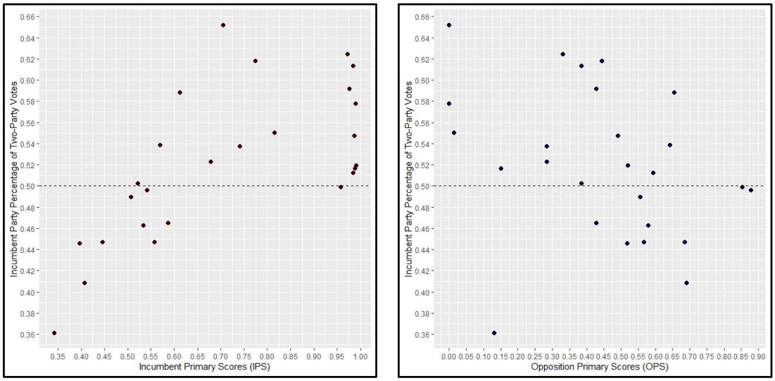

Figure 1: Scatter plots displaying the relationship between primary performances (for both the incumbent and the opposition party) and the incumbent party percentage of the two-party vote at the general election.

Considering data from 1912 onwards, Figure 1 also suggests two additional relevant features: on the one hand, that an incumbent candidate whose primary relative strength is at least 50% has good chances to win the general election, and, on the other hand, that the opposition candidate appears to be a serious threat for the incumbent party when he/she exhibits a primary strength at least greater than 40%.

A final consideration is that, looking at the scatterplot, another relevant trait seems to be manifested: after a certain primary score, the positive relationship between the incumbent candidate primary relative strength and the incumbent party votes weakens. A possible interpretation is that, on average, increments in primary consensus for less popular candidates (up until about 50% primary score) result in stronger increases in the general election performances with respect to what happens in case of very popular candidates (from 50% to around 80% primary score). However, considering popular candidates, there is not a great difference between a very popular candidate (80% primary score) and an extremely popular candidate (100% primary score) when it comes to the general election. Interestingly, also for the opposition candidates the extreme values of relative strength appear to have weaker relationship with the general election performance with respect to less extreme values. The empirical strategy adopted to face this issue in the calculation of the predictor of the LPM is discussed in the next section.

Finally, it is worth noting that most of the extremely popular candidates are in fact sitting presidents, which of course faced a far less competitive primary contest: this issue will be addressed in section 3.5.

3.4. Range of limit values

To model the relationship between the *IPS* and incumbent party votes at the general election accounting for the possible nonlinearity that emerges from the visual inspection of Figure 1, we adopt a range of limit values in the likes of the model through a winsoring strategy. This means choosing a minimum and a maximum primary *IPS* values: observations with *IPS* lower than the minimum are set equal to the minimum, while observations with *IPS* higher than the maximum are set equal to the maximum.

As first, the choice of the limit values is based on the approximation of the limits detected by observing the related plots. After running the inceptive model obtained with those limits, the limit values are then optimized using a BFGS algorithm (the R function "optim") (Byrd et al., 1995) in order to detect the range of limit values that minimizes the AIC of the model. The same range limit values have been adopted also for the *OPS*.

After the optimization procedure for the winsoring limits of *IPS* and *OPS*, the *CPS* series has been updated accordingly. However, since the obtained *CPS* does not allow for comparisons between the primary relative strength of the presidential candidates, the subsequent step to obtain the predictor of the LPM is to define a normalized measure. Indeed, intuitively, the higher the primary relative strength, the better the candidate's result in the general election; however, the benchmark that defines relatively strong and weak primary results may be different according to of the typology of candidates involved in each election.

3.5. Thresholds

The empirical strategy used to obtain a normalized value of the *CPS* follows three main steps: i) To define the possible category of presidential candidates according to the electoral setting; ii) For each category, to estimate the threshold of *CPS* that leads to an even presidential election result; iii) Finally, the normalized primary relative strength is defined as *CPS* of each candidate minus the threshold estimated for that category of candidate.

According to the different settings of each election, the candidates to the presidential elections have been distinguished in three categories: opposition candidates, incumbent candidates that are sitting presidents and "emerging" incumbents. The rationale behind this choice is straightforward: as noted above, the fact that the incumbent party candidate is a sitting president influences the level of primary consensus. Sitting presidents are current presidents seeking re-election: those candidates face a far less competitive primary contest, so their performance should be evaluated separately from other emerging incumbents.

The method applied to estimate the threshold of incumbent candidates, being them either sitting presidents or "emerging" incumbent, is the same: as first, the linear regression between incumbent party share of votes at the presidential election and the historic *CPS* for that category is estimated; secondly, the regression line is intersected with the 50-50 horizontal line of a theoretical even election; the *CPS* threshold value that leads to an even election is given by the intersection between the two lines. Considering the non-linearity of the relationship between *CPS* and the share of votes obtained in the general election, in the linear regression only value of CPS lower than 0.9 are included. A graphical example of the procedure is shown in Figure 2. For opposition candidates, given the weaker relationship between votes and primaries, we simply adopt the average opposition candidate performance as threshold.

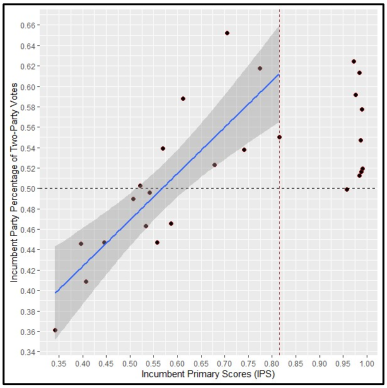

Figure 2: Detection procedure for the threshold value for incumbent candidates. Incumbents performing around 55% produce an even general election; the more positive the deviation from 55% the better the performance; the more negative the deviation from 55% the worse the performance. Note that this procedure is applied separately on incumbent candidates and sitting presidents.

Once computed the threshold that defines the benchmark between relatively strong and weak primary results for the three categories of candidates, a measure of distance to the threshold for each candidate *i* belonging to the category *q* is calculated as follows:

$$NCPS_{iq} = CPS_{iq} - T_q \,, \tag{5}$$

with *q* = *Opposition candidates, Sitting incumbent, Emerging incumbents*. *NCPS* is the Normalized Candidate Primary Strength and *T* the estimated threshold.

In the LPM setting, a further step is functional for the definition of the predictor of the LPM. Once the *NCPS* is calculated, it is defined separately for the candidates of the two party. Specifically, the *NCPS* is called Democratic Primary Strength (DPS) for Democratic candidates and Republican Primary Strength (RPS) for Republicans.

3.6. The Net Primary Score

The empirical definition of the predictor variable of the LPM finds its basis in the preliminary analyses on the relationships between incumbent party popular votes at the general elections and incumbent and opposition candidate primary relative strength, as described in previous sections.

The included predictor, defined as Net Primary Score (*NPS*) aims to explain the Democratic Party probabilities of winning the general election and is defined as the net advantage of the Democratic candidate over the Republican one. In this setting, as shown in Eq. 2, for each presidential election, the *NPS* is calculated as the difference between *DPS* and *RPS*. Recalling that *DPS* and *RPS* are normalized measures of primary relative strength, positive if the candidate primary performance is relatively strong and negative if the performance is relatively weak, the *NPS* is thus positive if the primary relative strength of the Democratic candidate is higher than the primary relative strength of the Republican one and negative otherwise. The higher (lower) the value of *NPS*, the stronger is, in terms of leadership, the Democratic (Republican) candidate.

4. Results

In this section the LPM is applied to electoral data and the main results, both in terms of goodness-of-fit and forecast, are reported.

4.1. Goodness-of-fit

The LPM analyses data for U.S. presidential election that involved a primary contest from 1912 to 2012 with the objective to predict the Democratic party probabilities of winning the elections, assuming the electoral system perfectly bipartisan. Following this assumption, the results can be interpreted as follow: i) an estimated probability greater than 0.5 suggests for Democratic Party victory; ii) an estimated probability lower than 0.5 suggests for Republican Party victory; iii) an estimated probability equal to 0.5 suggests for an even election between the two main parties.

The main results of the LPM are reported in Table 1. What clearly emerges is that the model can correctly explain almost the totality of the presidential elections of over one hundred years of elections (1912-2012). Overall, the model fails to explain the results of three elections over 26: the fifth (1928), the twentieth (1988) and the twenty-first election (1992). The Net Primary Score as unique predictor has shown therefore to be sufficient to explain almost all the analyzed U.S. elections, confirming the value of primary elections as leading indicator over the presidential election results.

The goodness-of-fit of the estimated LPM for historical data is shown in Figure 3.

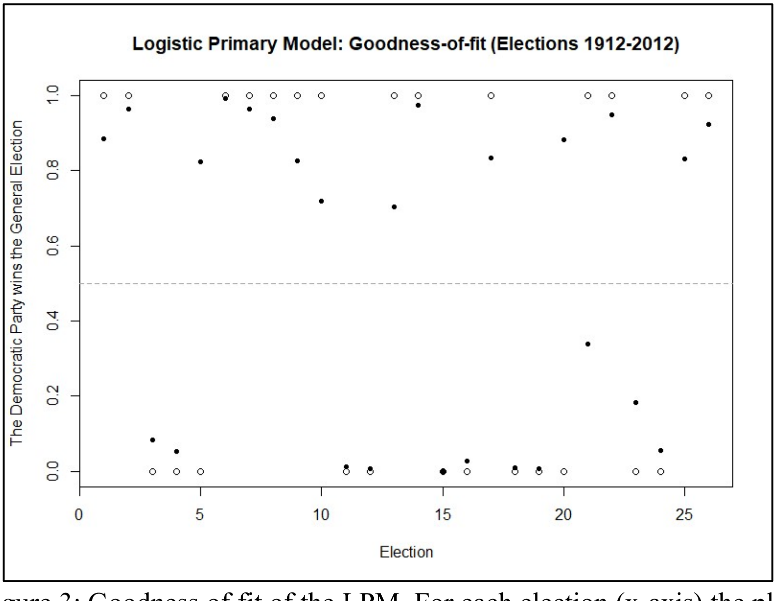

Figure 3: Goodness of fit of the LPM. For each election (x-axis) the plot displays (y-axis): the result (white dots) of whether the Democratic Party won (1.0) or lost (0.0) the election; the probabilities assigned for that historical event to be occurred (black dots). Probabilities favor the correct historic results in all but three cases.

Election	Year	Democratic party probability of winning	True Democratic party victory
1	1912	88.41%	1
2	1916	96.53%	1
3	1920	8.51%	0
4	1924	5.34%	0
5	1928	82.50%	0
6	1932	99.13%	1
7	1936	96.53%	1
8	1940	93.93%	1
9	1944	82.74%	1
10	1948	71.99%	1
11	1952	1.35%	0
12	1956	0.79%	0
13	1960	70.32%	1
14	1964	97.53%	1
15	1968	0.05%	0
16	1972	2.68%	0
17	1976	83.51%	1
18	1980	0.88%	0
19	1984	0.79%	0
20	1988	88.24%	0
21	1992	34.02%	1
22	1996	94.98%	1
23	2000	18.23%	0
24	2004	5.57%	0
25	2008	83.17%	1
26	2012	92.28%	1

Table 1: LPM estimated probability of Democratic victory and actual electoral results.

4.2. Forecasting performance

The model can be used for predictive purposes. In accordance with the Primary Model, after proper analysis oriented toward the search of potential improvements, the New Hampshire and South Carolina primaries are confirmed as the primaries better correlated with the general election in November. This allows for predictions 8-9 months before the election day.

The prediction of the 2016 outcome after the imputation of the 2016 primary results over the model built on historic data (1912-2012) leads to an estimated probability of victory equal to 0.0117 for the Democratic candidate Hillary Clinton and 0.9883 for the Republican candidate Donald Trump.

As a further check, a series of out-of-sample forecasts are conducted since 1996, for a total of six elections (1996, 2000, 2004, 2008, 2012 and then 2016). The electoral outcomes are predicted by the model with the data available at their respective time. The results are reported in Table 2 and confirm the forecasting abilities of the presented LPM.

Election	Year	Democratic party probability of winning	True Democratic party victory
22	1996	82.97%	1
23	2000	38.15%	0
24	2004	7.60%	0
25	2008	79.80%	1
26	2012	90.68%	1
27	2016	2.61%	0

Table 2: Out-of-sample forecasts 1996-2016.

5. *Conclusions*

From 1912 to 2016, presidential primaries had proven to be the leading predictor for the United States general election: this intuition, firstly embedded in the Norpoth's PM, is now improved in the LPM, at least for these specifics: i) contrarily to the PM, the LPM aims to predict the general election outcome and not the amount of popular votes received: this goal better reflects the nature of the American electoral system ii) logistic regression always ensures a meaningful prediction in terms of probabilities of winning the election, rather than a linear regression applied to percentage values.

The result is a model with several desired features, that aims at overcoming the limitations of some of the previous U.S. electoral forecasting models both from a theoretical and methodological perspective. From a theoretical point of view, as already recalled, it considers the characteristics of the U.S. electoral system, as invoked by recent literature (Dassoneville, 2020), moreover, it does not include economic variables, being considered as confounding factors especially in certain conditions (Erikson, 2020). Methodologically, it is estimated considering all the U.S. presidential election from 1912 onwards, and not only the more recent elections, as other electoral models. In addition, the LPM is a model with two desired properties, accuracy and lead (distance from the event) (Jennings, 2020).

Nevertheless, the choice of a unique predictor can be criticized as it might reduce the theoretical framework of the model. However, as pointed out by Campbell (2008) and Holbrook (2010), an electoral model must be accurate and interesting, that means based on theoretically plausible assumptions, as in the LPM case. Methodologically, the choice of a unique predictor is based on a parsimony principle, and it has been applied elsewhere in electoral forecasting models (Abramovitz, 2020).

Undoubtedly, electoral forecasting is at its earliest historical stages because of the short number of observations available. As to the actual state of the analysis, primary elections results appear as a leading indicator: they alone enable an accurate prediction, at almost a year before election date. This evidence poses the LPM in a unique class of electoral models to forecast the presidential election. Over the natural course of time, it will be possible to see if this observed relationship persists, or if additional predictors will emerge.

References

Abramowitz, A. I. (2016). Will time for change mean time for trump? *PS: Political Science & Politics 49 (4)*, 659–660.

Abramowitz, A. I. (2020). It's the Pandemic, Stupid! A Simplified Model for Forecasting the 2020 Presidential Election. *PS: Political Science & Politics*, 1-3.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on scientific computing*, 16(5), 1190-1208.

Campbell, J. E. (2008). Evaluating US presidential election forecasts and forecasting equations. *International Journal of Forecasting*, *24*(2), 259-271.

Campbell, J. E. (2016). The trial-heat and seats-in-trouble forecasts of the 2016 presidential and congressional elections. *PS: Political Science & Politics 49 (4)*, 664–668.

Campbell, J. E., H. Norpoth, A. I. Abramowitz, M. S. Lewis-Beck, C. Tien, R. S. Erikson, C. Wlezien, B. Lockerbie, T. M. Holbrook, B. Jerˆome, et al. (2017). A recap of the 2016 election forecasts. *PS: Political Science & Politics 50 (2)*, 331–338.

Cuzàn, A. G. (2012). Forecasting the 2012 presidential election with the fiscal model. *PS: Political Science & Politics 45 (4)*, 648–650.

Cuzán, A. G. (2020). The Campbell Collection of Presidential Election Forecasts, 1984–2016: A Review. *PS: Political Science & Politics*, 1-5.

Dassonneville, R., & Tien, C. (2020). Introduction to Forecasting the 2020 US Elections. *PS: Political Science & Politics*, 1-5.

Erikson, R. S. and C. Wlezien (2016). Forecasting the presidential vote with leading economic indicators and the polls. *PS: Political Science & Politics 49 (4)*, 669–672.

Erikson, R. S., & Wlezien, C. (2020). Forecasting the 2020 Presidential Election: Leading Economic Indicators, Polls, and the Vote. *PS: Political Science & Politics*, 1-4.

Fair, R. (2011). *Predicting presidential elections and other things*. Stanford University Press.

Gelman, A., S. Goel, D. Rivers, D. Rothschild, et al. (2016). The mythical swing voter. *Quarterly Journal of Political Science 11* (1), 103–130.

Graefe, A. (2013). Issue and leader voting in us presidential elections. *Electoral Studies 32* (4), 644–657.

Graefe, A., & Armstrong, J. S. (2013). Forecasting elections from voters' perceptions of candidates' ability to handle issues. *Journal of Behavioral Decision Making*, *26*(3), 295-303.

Graefe, A., Armstrong, J. S., Jones Jr, R. J., & Cuzán, A. G. (2014). Combining forecasts: An application to elections. *International Journal of Forecasting*, *30*(1), 43-54.

Greenwald, A. G., Smith, C. T., Sriram, N., Bar-Anan, Y., & Nosek, B. A. (2009). Implicit race attitudes predicted vote in the 2008 US presidential election. *Analyses of Social Issues and Public Policy*, *9*(1), 241-253.

Hibbs, D. A. (2012). Obama's reelection prospects under "bread and peace" voting in the 2012 us presidential election. *PS: Political Science & Politics 45* (4), 635–639.

Holbrook, T. (2010). Forecasting us presidential elections. In J. E. Leighley (Ed.), *The Oxford handbook of American elections and political behavior*, pp. 346–371. Oxford: Oxford University Press.

Holbrook, T. M. (2016). National conditions, trial-heat polls, and the 2016 election. *PS: Political Science & Politics 49* (4), 677–679.

Holbrook, T. M. and J. A. DeSart (1999). Using state polls to forecast presidential election outcomes in the american states. *International Journal of Forecasting 15* (2), 137–142.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Jennings, W., and Wlezien, C. (2018). Election polling errors across time and space. *Nature Human Behaviour*, *2*(4), 276-283

Jennings, W., Lewis-Beck, M., and Wlezien, C. (2020). Election forecasting: Too far out?. *International Journal of Forecasting*, *36* (2020), 949-962.

Jerôme, B. and V. Jerôme-Speziari (2016). State-level forecasts for the 2016 us presidential elections: Political economy model predicts hillary clinton victory. *PS: Political Science & Politics 49* (4), 680–686.

Lewis-Beck, M. S. and C. Tien (2016). The political economy model: 2016 us election forecasts. *PS: Political Science & Politics 49* (4), 661–663.

Lewis-Beck, M. S., and Rice, T. W. (1992). *Forecasting elections*. Washington D.C. : CQ Press.

Linzer, D. A. (2013). Dynamic Bayesian forecasting of presidential elections in the states. *Journal of the American Statistical Association*, *108* (501), 124-134.

Linzer, D. and Lewis-Beck, M. S. (2015). Forecasting US presidential elections: New approaches (an introduction). *International Journal of Forecasting*, *3*(31), 895-897.

Lockerbie, B. (2016). Economic pessimism and political punishment. *PS: Political Science & Politics 49* (4), 673–676.

Norpoth, H. (2004). From primary to general election: A forecast of the presidential vote. *PS: Political Science and Politics*, *37*(4), 737-740.

Norpoth, H. (2014). The electoral cycle. *PS, Political Science & Politics*, *47*(2), 332.

Norpoth, H. (2016). Primary model predicts trump victory. *PS: Political Science & Politics 49* (4), 655–658.

Shirani-Mehr, H., Rothschild, D., Goel, S., and Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association*, *113*(522), 607-614.

Zeedan, R. (2019). The 2016 US Presidential Elections: What Went Wrong in Pre-Election Polls? Demographics Help to Explain. *J—Multidisciplinary Scientific Journal*, *2*(1), 84-101.

Appendix

Review of U.S. presidential election models

A.1. The Issues and Leaders model by Andreas Graefe

$$V_t = P_t + bI_t + L_t + \epsilon\,.$$

The model covers all the elections from 1972 to 2016. It uses the following variables: the Incumbent's actual share of the two-party popular vote (V), the party identification (P), the measure of dealing with issues that matters (I) and the leadership score (L). The Issues and Leaders model casts its prediction 2-3 months before election day (Graefe, 2013).

A.2. The "Time for Chang" model by Alan Abramowitz

$$V_t = A + \beta_0 NETAPP_t + \beta_1 Q2GDP_t + \beta TERM1INC_t + \epsilon\,.$$

The model covers all the elections from 1988 to 2016. It uses the following variables: the Incumbent share of the two-party popular vote (V), the Incumbent president's net approval rating (approval less disapproval) in the final Gallup Poll in June ($NETAPP$), the Annualized growth rate of real GDP in the second quarter of the election year ($Q2GDP$), the presence (1) or absence (0) of a first-term incumbent in the race ($TERM1INC$). The Time for Change model casts its prediction 4-5 months before election day (Abramowitz, 2016).

A.3. The Trial-heat model by Jim Campbell

$$V_t = A + \beta_0 POLL_t + \beta_1 GDP_t + \epsilon\,.$$

The equation is estimated over the 16 elections from 1948 to 2012, with 2008 excluded. The model covers all the elections from

1992 to 2016. It uses the following variables: the Incumbent share of the two-party popular vote (V), the Incumbent party's candidate two-party support in early September Gallup preference poll ($POLL$) and the Annualized real GDP growth rate in the second quarter as indicated by the Bureau of Economic Analysis' second estimate released at the end of August (GDP). The Trial-heat model casts its prediction 60 days before election day (Campbell, 2016).

A.4. The Convention-bump model by Jim Campbell

$$V_t = A + \beta_0 PREC.POLL_t + \beta_1 NETC.BUMP_t + \beta_2 GDP_t + \epsilon \, .$$

The model covers all the elections from 2004 to 2016. It uses the following variables: the Incumbent share of the two-party popular vote (V), the Incumbent party's candidate two-party support in polls before the parties' first convention ($PREC.POLL$), the change in the incumbent party's candidate two-party support polls before the first convention to after the second convention ($NETC.BUMP$), the annualized real GDP growth rate in the second quarter as indicated by the Bureau of Economic Analysis' second estimate released at the end of August (GDP). The Convention-bump model casts its prediction 74 days before election day (Campbell et al., 2017).

A.5. The Political Economy model by Michael S. Lewis-Beck and Charles Tien

$$V_t = A + \beta_0 POPULARITY_t + \beta_1 ECONOMICGROWTH_t + \epsilon \, .$$

The model covers all the elections from 1948 to 2016. It uses the following variables: the Incumbent share of the two-party popular vote (V), the Gallup Presidential Approval measure in July of the election year ($POPULARITY$), the GNP growth in the first two quarters of the election year ($ECONOMICGROWTH$). The

Political Economy model casts its prediction 74 days before election day (Lewis-Beck and Tien, 2016).

A.6. The National Conditions and Trial Heat model by Thomas Holbrook

$$V_t = A + \beta_0 CONDITIONS_t + \beta_1 TRIAL + \epsilon\ .$$

The model covers all the elections from 1952 to 2016. It uses the following variables: the Incumbent share of the two-party popular vote (V), the Index of national conditions ($CONDITIONS$), the Measure of the incumbent party candidate's performance in pre-election trial-heat polls ($TRIAL$). The National Conditions and Trial Heat model casts its prediction 1-2 months before election day (Holbrook, 2016).

A.7. The DeSart and Holbrook model by Jay A. DeSart and Thomas Holbrook

$$\begin{aligned} V_{i,t} = \alpha + \beta_0 (POLL)_{i,t} + \beta_1 (PRIORVOTE)_{i,t} \\ + \beta_2 (NATIONALPOLLS)_{i,t} + \epsilon\ . \end{aligned}$$

The model covers all the elections from 2000 to 2016. It uses the following variables: the State-level Democratic popular vote (V), the average Democratic share of support among the major party candidates in all trial-heat polls taken in each state during the month of September ($POLLS$), the average Democratic share of the two-party popular vote across the four previous elections ($PRIORVOTE$), the Nation-level poll ($NATIONALPOLLS$). The DeSart and Holbrook model casts its prediction 2 months before election day (Holbrook and DeSart, 1999).

A.8. The Long-Rage DeSart and Holbrook model by Jay A. DeSart and Thomas Holbrook

$$V_{i,t} = \alpha + \beta_0(PREVIOUSRESULT)_{i,t} + \beta_1(PRIORVOTE)_{i,t} + \beta_2(PRIOROCTOBERPOLLS)_{i,t} + \epsilon\,.$$

The model covers all the elections from 2000 to 2016. It uses the following variables: the State-level Democratic popular vote (V), the Lagged state-level Democratic popu-lar vote ($PREVIOUSRESULT$), the average Democratic share of the two-party popular vote across the four previous elections ($PRIORVOTE$), the average national match-up polls between the two eventual nominees at October one-year prior to the election (PRIOROCTOBERPOLLS). The Long-Rage DeSart and Holbrook model by Jay A. De-Sart and Thomas Holbrook casts its prediction 1 year before the election day.

A.9. The Fiscal model by Alfred G. Cuzàn

$$VOTE2_t = \alpha + \beta_0 G_t + \beta_1 Z_t + \beta_2 D_t - \beta_3 P_t + \beta_4 F1_t + \beta_5 F3_t + \epsilon\,.$$

The model covers all the elections from 1916 to 2016. It uses the following variables: the Incumbent party share of two-party vote ($VOTE2$), the Growth rate of real per capita GDP in the first three quarters of the election year (G), the number of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2 percent at an annual rate (Z), a dummy variable coded 0 if the incumbent party has been in power for one term, 1 if the incumbent party has been in power for two consecutive terms, 1.25 if the incumbent party has been in power for three consecutive terms, 1.50 for four consecutive terms and so on (D), a dummy variable coded 1 if the Democrats occupy the White House and -1 if the Republicans are the incumbent (P), the Federal outlays to GDP in percentage (F), the differences in F across the actual and previous election year, $F1 = F_t - F_t - 1$ ($F1$), a dummy variable coded 1 if F1 > 1, -1 if F1 < 1 and 0 if F1

= 0 ($F3$). The Fiscal model casts its prediction 3 months before election day (Cuzàn, 2012).

A.10. The Lockerbie model by Brad Lockerbie

$$V_t = A + \beta_0 NYWORSE_t + \beta_1 LOGTWH_t + \epsilon\ .$$

The model covers all the elections from 1956 to 2016. It uses the following variables: the Incumbent share of the two-party popular vote (V), the people's economic expectations taken from a Survey of Consumer Attitudes and Behavior ($NYWORSE$), the length of White House control by the incumbent party ($LOGTWH$). The Lockerbie model casts its prediction 4 months before election day (Lockerbie, 2016).

A.11. The Leading indicators model by Robert Erikson and Christopher Wlezien

$$V_t = A + \beta_0 GROWTH_t + \beta_1 POLLS_t + \epsilon\ .$$

The model covers all the elections from 1952 to 2016. It uses the following variables: the Incumbent share of the two-party popular vote (V), the summed weighted growth in leading economic indicators through quarter 13 of the election cycle, with each quarter weighted 0.8 times the following quarter ($GROWTH$), the Incumbent party's candidate two-party support in polls ($POLLS$). The Leading indicators model casts its prediction 5 months before election day (Erikson and Wlezien, 2016).

A.12. The Fair model by Ray Fair

$$\begin{aligned} V_t = \beta_0(G_t * I_t) + \beta_1(P_t * I_t) + \beta_2(Z_t * I_t) + \beta_3(DPER_t) \\ + \beta_4(DUR_t) + \beta_5(I_t) + \beta_6(WAR_t) + \epsilon\ . \end{aligned}$$

The model covers all the elections from 1916 to 2016. It uses the following variables: the Democratic share of the two-party popular vote (V), the growth rate of real per capita GDP in the first three quarters of the on-term election year (G), the absolute value of the growth rate of the GDP deflator in the first 15 quarters of the administration except for 1920, 1944 and 1948 where the values are zero (P), the numbers of quarters in the first 15 quarters of the administration in which the growth rate of real per capita GDP is greater than 3.2 percent at an annual rate except for 1920, 1944 and 1948 where the values are zero (Z), a dummy variable coded 1 if there is a Democratic presidential incumbent and -1 if there is a Republican presidential incumbent (I), a dummy variable coded 1 if the Democratic incumbent is seeking reelection, -1 if it is Republican and 0 otherwise ($DPER$), a dummy variable coded 0 if either party has been in the White House for one term, 1[-1] if the Democratic [Republican] party has been in the White House for two consecutive terms, 1.25 [-1.25] if the Democratic [Republican] party has been in the White House for three consecutive terms, 1.50 [-1.50] if the Democratic [Republican] party has been in the White House for four consecutive terms, and so on (DUR), a dummy variable coded 1 for election years 1918, 1920, 1942, 1944, 1946 and 1948, and 0 otherwise (WAR). The Fair model casts its prediction 2 weeks before election day (Fair, 2011).

A.13. The Bread and Peace model by Douglas A. Hibbs

$$V_t = \propto + \beta_1 \left(\sum_{j=0}^{14} \lambda^j \Delta lnR_{t-j} \left({}^{1}\!\Big/_{\sum_{j=0}^{14} \lambda^j} \right) \right) + \beta_2 Fatalities_t + \varepsilon \, .$$

The model covers all the elections from 1952 to 2016. It uses the following variables: the Incumbent share of the two-party popular vote (V), the per capita disposable personal income deflated by the Consumer Price Index (R), the Quarter-to-quarter log-percentage change expressed at annual rates (ΔlnR), the Lag weight from 0 to 1 with different effects on the model (λ), the cumulative number of American military fatalities per millions of US population in Korea, Vietnam, Iraq and Afghanistan during the presidential terms preceding the 1952, 1964, 1968, 1976 and 2004, 2008 and 2012 elections ($Fatalities$). The Brad and Peace model casts its prediction 3 months before election day (Hibbs, 2012).

A.14. The State-by-state political economy model by Bruno and Véronique Jérôme

$$\begin{aligned} INCV_{i,t} = C + \Delta U_{i,t-n} + PJA_{t-n} + PPI_{i,t-n} \\ + Politics\ and\ Institutions_{i,t-n} \\ + \ President's\ local\ Strongholds_{i,t-n} \\ + Local\ Peculiarities_{i,t-n} + OPPVP + \varepsilon\ . \end{aligned}$$

The model covers all the elections from 1980 to 2016. It uses the following variables: the Incumbent share of two-party vote in the i-th state, 1 ≤ o ≤ 51 for the t-th time period 1980 ≤ t ≤ 2012 ($INCV_{i,t}$), the change in the state-level unemployment rate from the month after the president was elected to the month prior to the next presidential election (ΔU), the Gallup President's Job Approval, at national level, six months before the election (PJA), the Partisan domination variable, giving for each state over the 1952-2012 period the rate of success for each party when this rate exceeded 85 per-cent for the Republicans and 63 percent for the Democrats, zero otherwise (PPI), Some variables dealing with politics and institutions such as the electoral weight of independent candidates ($Politics\ and\ Institutions$), Variables accounting for the president's advantage in electoral strongholds ($President's\ local\ Strongholds$), Variables reflecting local peculiarities such as the scores usually deviating from the standard like Democrats in Washington DC ($Local\ Peculiarities$), the Opposition nominee's vote share during their party's primaries ($OPPVP$). The State-by-state political economy model casts its prediction 1 months before election day Jérôme, B. & Jérôme-Speziari (2016).

Finito di stampare
GieGi srl - Triuggio (MB)
Febbraio 2021

9788834345481