Latent Variables and Rolling Panels

A New Approach to Modeling Campaign Effects

Andrew Therriault, New York University

NEW YORK UNIVERSITY

Overview

What effects do political campaigns have on voters? Many researchers have argued that they have only "minimal effects", given the stability of voters' choices and the influence of factors such as partisanship and the economy. But in recent years, a growing body of evidence has suggested that campaigns indeed affect voters in a meaningful way. Exactly how campaigns' influences manifest, however, is still an unsettled debate.

I argue that diffculty of identifying campaign effects reflects the limitations of existing data and methods, *not* the absence of effects. In the paper from which this poster is drawn, I investigate campaign effects in the 2000 and 2004 US presidential elections by combining panel data from the National Annenberg Election Survey (NAES) with detailed records of campaign ads from the WiscAds/CMAG dataset. My paper is the first to employ both direct measurements of campaign messages and panel data with regard to specific issues, and the results point to a far greater role for campaigns than previously believed. I demonstrate that candidates can choose to clarify or obfuscate their positions effectively, that voters are persuaded by ads to change their policy preferences, and that these effects ultimately lead to substantial changes in vote choice.

While these findings are made possible through the use of over-time data on campaigns and voters, such data present a number of analytical challenges which threaten the ability of the analysis to credibly discern any causal relationships. As this is the point of using panel data in the first place—and existing models for panel data offer little guidance—I develop an original modeling approach and present it here. With election panels becoming more common, this method could prove increasingly valuable.

Data Structure and Challenges

The NAES studies are centered around a very large rolling cross-section (RCS), comprised of daily telephone interviews in the year before each election. The NAES then recontacts a subsample of respondents (approximately 10%) to conduct post-election interviews as well. The NAES general election panel combines both sets of responses for these individuals, allowing for the study of changes in variables repeated across waves.

This structure provides great potential for studying the effects of campaigns on voters, but the specific design of the NAES comes with disadvantages as well. In particular:

- The panel's sample size is an order of magnitude smaller than that of the RCS, which increases the risk of overfitting in very highly-specified models.
- Because of sample splits and variation in the questions asked over the course of the RCS, most questions asked in the reinterviews are missing a large proportion (often more than half) of pre-election responses.
- As the dates of pre-election interviews vary widely, changes in the distribution of responses over time could lead to biased results.

The last of these problems comes from the measurement error produced by translating latent variables to survey responses. As the distributions of these latent variables often change over time—for example, when voters learn about the candidates from media coverage—this error could be correlated with time. Since time is itself correlated with campaign exposure, this could lead to spurious results when analyzing campaign effects.

Modeling Campaign Effects

I look at three dependent variables to study the effects of campaigns: perceptions of candidate positions, policy preferences, and vote choice. To illustrate my general modeling strategy, I present below the model of how advertising affects voters' perceptions of candidates. The perception by individual i of the position of candidate c on issue j at time t—Z_{cijt}—is assumed to be oriented so that positive values are closer to the candidate's true position. This perception then leads to C_{cijt}, an indicator for whether the respondent correctly places the candidate in a survey:

$$C_{cijt} = 1 \text{ if } Z_{cijt} > 0, 0 \text{ otherwise.}$$

The perception Z_{cijt} at time $t = T$ is modeled as a linear function of ads by each candidate, on that issue in the respondents's market up until that date, and a set of control variables:

$$\begin{aligned} Z_{cijt} = {} & \beta_0 + \beta_1 KerryAds_{ij,t<T} + \beta_2 BushAds_{ij,t<T} + \beta_3 Demographics_i \\ & + \beta_4 MediaUse_i + \beta_5 PartyID_i + \beta_6 Ideology_i + \beta_7 Interest_i \\ & + \beta_8 Time_{it} + \epsilon_{it} \end{aligned}$$

Since the dependent variable in the applied models is the survey response C_{cijt}, a binary probit model can be used for the latent Z_{cijt}:

$$Pr(C_{cijt} = 1) = \Phi(Z_{cijt})$$

Because the only factors in this model which vary across panel waves are the ad measures and the time trend, post-election perceptions ($t = 1$) can be modeled as a function of these variables and pre-election perceptions ($t = 0$):

$$\begin{aligned} Z_{cij1} = {} & \beta_0 + \beta_1 KerryAds_{ij,0<t<1} + \beta_2 BushAds_{ij,0<t<1} \\ & + \beta_3 Time_{i0} + \beta_4 Z_{cij0} + \epsilon_{i1} \end{aligned}$$

If we observed Z_{cijt} directly, this model would be simple to implement. But because Z_{cijt} is both latent and continuous (while its observed counterpart, C_{cijt}, is categorical), the best approach is not so obvious. What is needed is a way to proxy for Z_{cijt}. Given the challenges of using the NAES panel described earlier, a suitable approach for implementing these models in the NAES data would need to meet three criteria:

1. It would avoid overfitting in a moderately-sized sample
2. Missing pre-election responses could be imputed and used in a straightforward manner
3. Bias from time-correlated measurement error would be minimized

Existing Methods

There are a variety of common techniques for analyzing categorical panel data, but none meets all three criteria:

Model	*Avoids Overfitting?*	*Imputation OK?*	*Minimizes Bias?*
Markov	No	No	No
First Differences	Yes	No	No
Lagged DV	Yes	Yes	No
LDV + Controls	No	Yes	Yes

The Markov model, in which separate analyses of post-election responses are run based on pre-election responses, can have sample-size problems when response distributionss are skewed, requires categorical initial responses (complicating MI), and does not address bias concerns. Using first differences solves the sample-size problem, but the others remain. Using a lagged dependent variable would meet the first two criteria, but still does nothing about bias. Finally, including the full set of controls (e.g., demographics, media consumption) in an LDV model would help to minimize bias (by taking weight off of the LDV), but risks overfitting the model.

A New Approach

As proxies for the lagged latent variables in models of post-election responses, I use a pair of variables: lagged survey responses and predictions of pre-election latent variables based upon models of cross-sectional responses (using the much-larger RCS sample, which allows for more precise estimates). Mathematically, the motivation behind using both variables is as follows. Let y_{it} be an individual's categorical survey response at time t. This response is the sum of the latent continuous variable y^*_{it} and a measurement error τ_{y*} (discussed at the end of the second section) produced by converting y^*_{it} into the categorical response y_{it}:

$$y_{it} = y^*_{it} + \tau_{y*}$$

The latent variable y^*_{it} is itself the sum of $x_{it}\beta$, the product of a vector of coefficents and a vector of observable independent variables; λ_i, which reflects the individual's unobserved heterogeneity; and ϵ_{it}, reflecting a random error term in the present period. Thus:

$$y_{it} = x_{it}\beta + \lambda_i + \tau_{y*} + \epsilon_{it}$$

Because y^*_{it} is never observed, the goal is to find the best proxy for y^*_{it} to use as a control variable in the follow period, $t+1$. By modeling y_{it} as a function of x_{it}, I can generate efficient, unbiased estimates of $x_{it}\beta$—denoted θ_{it}—which I assume to be the true values for this purpose. To proxy for y^*_{it}, I use the weighted average of θ_{it} and y_{it}:

$$\begin{aligned} \hat{y}^*{}_{it} &= \omega y_{it} + (1-\omega)\theta_{it} \\ \hat{y}^*{}_{it} &= \omega y_{it} + (1-\omega)x_{it}\beta \\ \hat{y}^*{}_{it} &= \omega(x_{it}\beta + \lambda_i + \tau_{y*} + \epsilon_{it}) + (1-\omega)x_{it}\beta \\ \hat{y}^*{}_{it} &= x_{it}\beta + \omega(\lambda_i + \tau_{y*} + \epsilon_{it}) \end{aligned}$$

The weight variable ω is not known *a priori*, but is instead determined empirically: the coefficients produced by the $t+1$ model reflect the weights on θ_{it} and y_{it}. The tradeoff is between increased precision from greater weight on λ_i and increased error from greater weight on ϵ_{it} and τ_{y*}. Determining these weights empirically allows the weight given to the individual and random error terms to be adjusted while keeping the weight on $x_{it}\beta$ constant. In expectation, if the coefficents on the time-invariant IVs are truly the same between waves, my approach will produce the same results as the LDV + Controls model discussed above, without the associated cost to the post-election model's degrees-of-freedom.

Comparing Models of Perceptions

To see how these approaches compare in practice, I use each to model perceptions of Bush's position on stem cell research (1 = correct placement, 0 = incorrect). The model fit (as expected proportional reduction in error—see Herron 1999), sample sizes, and estimated effects of ads (as the average change in probability of correctly placing Bush from $1 in per-household spending; SEs in parentheses) are listed below:

			Effect of $1 per Household of	
Model	*n*	*ePRE*	*Bush Ads*	*Kerry Ads*
Markov ($C_{t=0} = 0$)	182	7.34%	–0.47 (0.62)	+0.58 (0.41)
Markov ($C_{t=0} = 1$)	986	1.11%	–0.14 (0.13)	+0.20 (0.12)
First Differences	1168	0.57%	+0.09 (0.08)	–0.15 (0.06)
Lagged DV	4766	4.58%	–0.04 (0.04)	+0.04 (0.03)
LDV + Controls	4701	17.15%	–0.08 (0.04)	+0.08 (0.03)
LDV + Prediction	4756	11.73%	–0.08 (0.04)	+0.08 (0.03)

While most models show effects in the same direction (Bush's ads making voters' perceptions less accurate, Kerry's ads making them more so), the last two—lagged DVs with full controls, and my original approach—show the best fit and the most significant effects. Though there is little evidence of overfitting in the LDV + Controls results here (a potential problem noted above), models of policy preferences on the same issue tell another story.

Models of Policy Preferences

I next look to see whether voters are persuaded to change their policy preferences. On the issue of stem cell research, respondents place themselves on 5-pt scales (which I rescale to range from –1 = strongly oppose to 1 = strongly favor); though I treat these scales as continuous, they still carry potential measurement error. (Effect sizes are the estimated average change in self-placements from $1 in per-household ad spending, with SEs in parentheses.)

			Effect of$1 per Household of	
Model	*n*	R^2	*Bush Ads*	*Kerry Ads*
Markov ($Y_{t=0} = -1$)	211	0.1012	+0.17 (0.69)	–0.44 (0.46)
Markov ($Y_{t=0} = -0.5$)	105	0.0621	+0.86 (1.45)	–0.49 (0.90)
Markov ($Y_{t=0} = 0$)	77	0.1237	+1.22 (1.34)	+0.26 (0.90)
Markov ($Y_{t=0} = 0.5$)	179	0.0816	+0.19 (0.70)	–0.27 (0.50)
Markov ($Y_{t=0} = 1$)	545	0.0572	+0.18 (0.27)	–0.02 (0.25)
First Differences	1117	0.0094	+0.34 (0.27)	–0.28 (0.21)
Lagged DV	4766	0.2388	–0.10 (0.09)	+0.05 (0.06)
LDV + Controls	4766	0.3627	–0.13 (0.08)	+0.07 (0.06)
LDV + Prediction	4766	0.3360	–0.15 (0.08)	+0.10 (0.05)

Here the full set of controls (which includes many factor variables and interactions) offers little advantage over the prediction in terms of model fit, but the resulting coefficients on ads are of substantially-diminished magnitude and significance. Across all of the models, these results demonstrate how the use of ill-considered methods can lead to null or even perverse findings. With my approach, however, the results confirm what we would expect to see if persuasion were truly occurring: that Bush's ads made voters' preferences more conservative, while Kerry's made them more liberal.

Substantive Results

In my paper, I look at how each candidate's ads affect voters' candidate perceptions and policy preferences across 10 issues, then test whether such changes lead voters to update their vote preferences. I find that:

- Contrary to recent studies of campaign advertising, ads do *not* uniformly make voters better informed about candidates' positions. Rather, ads by competing candidates tend to have contrasting effects—when one candidate's ads clarify the candidates' positions for voters, the opponent's ads often serve to obfuscate them.
- Advertising can also have a persuasive effect on voters' policy preferences, *even after controlling for partisan cue-taking*.
- When these effects occur, individuals do update their vote choice, moving toward the candidates whose positions better fit their own policy preferences.

These results present a robust counterargument to those who dismiss campaigns as insignificant, and demonstrate that the effects of campaigns are *meaningful* rather than *minimal*.

Conclusion

The failure of previous studies to find campaign effects says less about the existence of these effects than it does about the inadequacy of common methodological approaches. By improving upon these approaches, I find stronger evidence for both informing and persuasion effects than any other study to date, and further show that such effects have a meaningful impact on votes. To be certain, this achievement was largely made possible by the advent of more precise data than was previously available. But as illustrated above, great data is not sufficient on its own—to be used effectively, it must be paired with a thoughtful modeling approach such as that presented here.

Andrew Therriault, New York University, therriault@nyu.edu