THE UNIVERSITY OF WESTERN AUSTRALIA

Polling bias and undecided voter allocations US Presidential elections, 2004–2016

Joshua J. Bon with

T. Ballard & **B. Baffour**

27th of September, 2017

School of Mathematics and Statistics, University of Western Australia

Introduction

Motivation

- Something interesting to analyse
  - A “holiday” project that turned into a research paper
  - Trying to come to terms with the 2016 election
- Many aspects of the 2016 US president to analyse, however we had noticed:
  - Large undecided voter levels during the US election and other recent polls (e.g. Brexit)
  - Pervasive assumption of static/deterministic allocation of undecided voters by pollsters (e.g. poll modellers) and in election polling papers
- So we started to look...

US Presidential undecided voters, 2004 – 2016

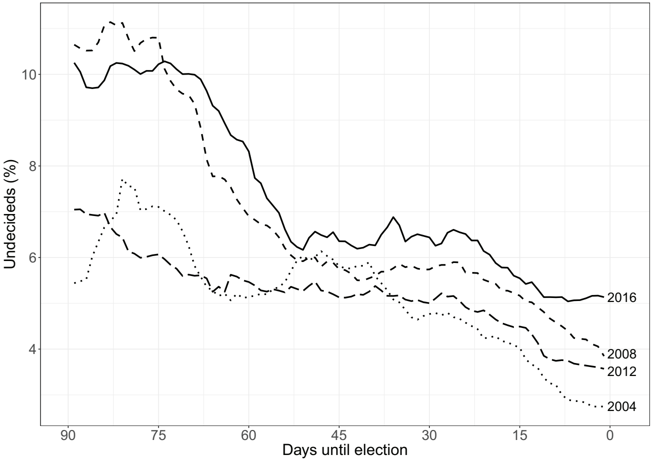

Figure 1: Mean level of undecided voters from US presidential elections. Weighted average from national polls that occur within a two-week window centred at x.

Measuring polling bias and variance

The total survey error approach[1]

- Survey/poll error = deviations of a survey response from true underlying value
- Error can occur through bias or variance
- Bias term captures systematic errors shared by all polls, e.g.
    - shared operational practices
    - sampling frames
- The variance term captures sampling variation, e.g.
    - different survey methodologies
    - different statistical models

[1] P. P. Biemer, *Public Opinion Quarterly* **74**, 817–848 (2010), R. M. Groves, L. Lyberg, *Public opinion quarterly* **74**, 849–879 (2010).

A Bayesian Model[2] (with no undecided voters)

$$p_i \sim \mathcal{N}\left(v_r + \alpha_r^p + t_i\beta_r^p, \sqrt{\frac{v_r(1-v_r)}{n_i}} + \tau_r^p\right) \tag{1}$$

- v_r is the actual poll result in state-year r for poll i
- $\alpha_r^p + t_i\beta_r^p$ time varying bias away from truth (election result)
- t_i is time until election day
- τ_r^p accounts for the excess deviation above a SRS
- ^{p} denotes these parameters relate to the polling model

[2] H. Shirani-Mehr *et al.*, *Disentangling Bias and Variance in Election Polls*, http://www.stat.columbia.edu/~gelman/research/unpublished/pollposition_v2.pdf, Accessed: 2016-11-15, 2016.

Pooling estimates

- We can't estimate bias and variance terms(s) for every poll
  - Otherwise p = 3n
  - Already ≈ 600 parameters
- Instead we pool the poll results within state and election year
- This way each parameter is shared in a state-year
  - e.g. California 2016 has 3 parameters, but many more polls prior to the election
- Each of these parameters are drawn from a shared hyper-prior
- Equivelant parameters can then "borrow strength" from each other

For example, the hierarchal structure of the election day bias is:

$$\alpha_r^p \sim \mathcal{N}(\mu_\alpha^p, \sigma_\alpha^p) \quad \mu_\alpha^p \sim \mathcal{N}(0, 0.05) \quad \sigma_\alpha^p \sim \mathcal{N}_+(0, 0.05) \tag{2}$$

Incorporating undecided voters

Standardising polls and undecided voters

A standard in the literature is to assume that

$$p_i = \frac{R_i}{R_i + D_i} \tag{3}$$

however you may include undecideds by letting

$$p'_i = \frac{R_i + \lambda U_i}{R_i + D_i + U_i} \tag{4}$$

where $0 \leq \lambda \leq 1$ “splits” the undecided voters. The measurements p_i and p'_i coincide under the assumption of static proportionate splitting:

$$\lambda = \frac{R_i}{R_i + D_i} \tag{5}$$

Incorporating undecided voters into the model

Assuming there is some bias away from proportionate splitting

$$\lambda = \frac{R_i}{R_i + D_i} + \theta_i \tag{6}$$

leads to the identity

$$p_i' = p_i + u_i\theta_i \tag{7}$$

which can be incorporated into the mean of the original model...

Incorporating undecided voters

...but, there several issues:

1. Undecided voter levels are time-varying
2. Undecided voters are not reported in $\approx 10\%$ of polls
3. Undecided voter levels are polled which has measurement error
4. θ_i is a parameter for every poll

Incorporating undecided voters

Let the undecided voters be modelled by:

$$u_i \sim \mathcal{N}\left(\rho_r + t_i\beta_r^u, \eta_y + \tau_r^u\right) \tag{8}$$

- ρ_r is the the election day mean for each state-year
- Polls that don't include u_i are covered since we estimate state-year parameters

Addresses time varying, missing data, and measurement error concerns by using state-year estimates of undecided voters on election day.

A Bayesian Model with undecided voters

$$p_i \sim \mathcal{N}\left(v_r + \alpha_r^p + t_i\beta_r^p - \rho_r\gamma_y, \sqrt{\frac{v_r(1-v_r)}{n_i} + \tau_r^p}\right) \tag{9}$$

- ρ_r is the the election day estimate of undecided voters for each state-year
- γ_y is election year parameter controlling biasing effect of undecided voters in each year

Data

Data sources

- State level polling data
  - 2012, 2016 from Pollster API[3]
  - 2004, 2008 from US Election Atlas[4]
- Polls up to 35 days prior to their respective election included
- 2,044 state-level polls total ($\approx$ 90% had undecideds reported)
- No 2000 or earlier polls with sufficient data on undecided voters were found.

[3] Huffington Post, *Pollster API V2*, http://elections.huffingtonpost.com/pollster/api/v2, Accessed: 2016-12-20, Huffington Post, 2016.

[4] D. Leip, *Atlas of US Presidential Elections*, http://uselectionatlas.org/, Accessed: 2016-12-20, 2008.

Results

Effect of undecided voters on polling bias

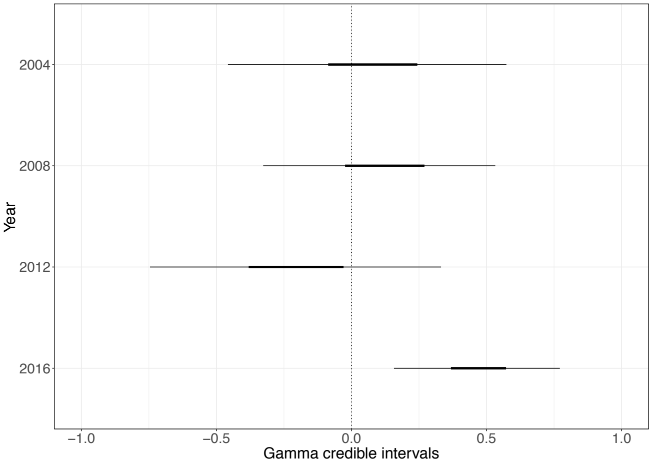

Figure 2: Credible intervals (95% and 50%) for undecided voter's effect on bias in the model (γ_y). A positive value indicates a bias away from proportional allocation in favour of the Republican candidate.

ALEX BRANDON/AP

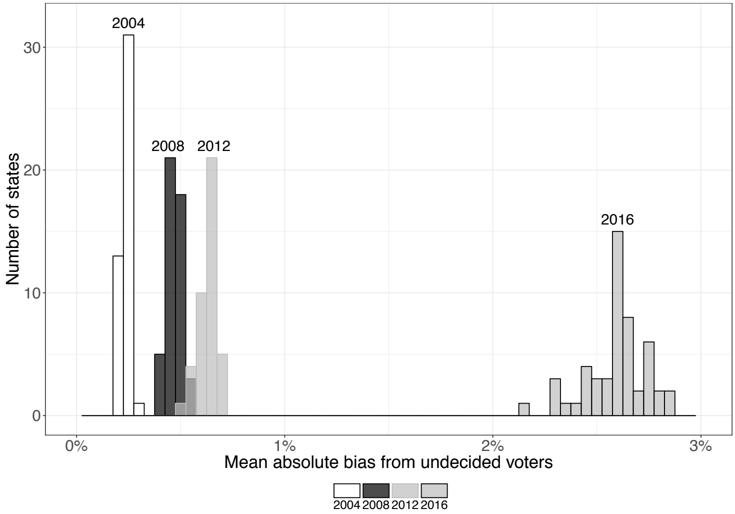

Figure 3: The bias from undecided voters is the quantity $\rho_r \gamma_y$ in the model. A positive value indicates a bias away from proportional allocation of undecided voters in favour of either candidate.

Concluding remarks

- In 2016, 5.5% of voters were undecided on election day, up from 3-4% in previous years
- Undecided voters biased polls in the 2016 US presidential election by 2-3% on average
- A static, proportionate split in undecided voters between leading candidates was a bad assumption in 2016, less so in previous years
- Pollsters and modellers should move towards stochastic allocation methods to allow uncertainty from undecided voters to propagate through models

Appendix

More concluding remarks

- No major changes after reanalysis assuming undecided voters have mean 50-50 split (i.e. $\lambda = 0.5 + \theta_i$)
- It should be noted that:
  - Only 4 years worth of data
  - Associative, no predictive testing or calibration done
  - Only estimated undecided voter's biasing effect at election year level

US Presidential undecided voters, 2004 – 2016

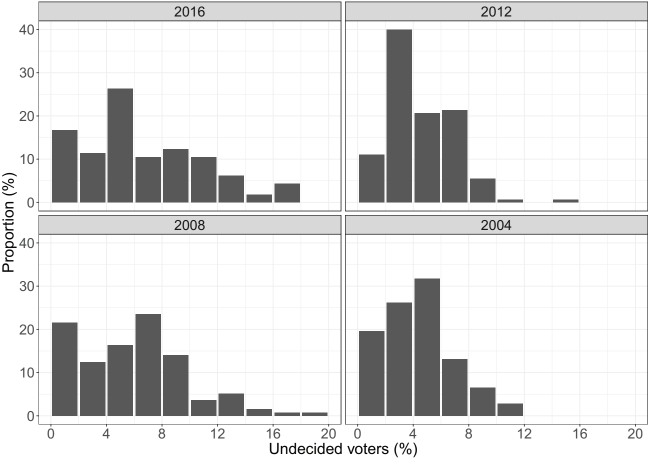

Figure 4: Histogram of undecided voters from national polls in the 3 months prior to the US presidential elections. Each bar is relative to the number of polls from that year.

Table 1: Priors used in models for analysis of state polls.

		Prior	*Hyper-prior*	
Model	Component		Mean	Variance
Polling	Mean	$\alpha_r^p \sim \mathcal{N}(\mu_\alpha^p, \sigma_\alpha^p)$	$\mu_\alpha^p \sim \mathcal{N}(0, 0.05)$	$\sigma_\alpha^p \sim \mathcal{N}_+(0, 0.05)$
		$\beta_r^p \sim \mathcal{N}(\mu_\beta^p, \sigma_\beta^p)$	$\mu_\beta^p \sim \mathcal{N}(0, 0.05)$	$\sigma_\beta^p \sim \mathcal{N}_+(0, 0.05)$
		$\gamma_y \sim \mathcal{N}(0, 0.5)$		
	Variance	$\tau_r^p \sim \mathcal{N}_+(0, \sigma_\tau^p)$		$\sigma_\tau^p \sim \mathcal{N}_+(0, 0.02)$
Undecided voters	Mean	$\alpha_r^u \sim \mathcal{N}(\mu_\alpha^u, \sigma_\alpha^u)$	$\mu_\alpha^u \sim \mathcal{N}(0, 0.05)$	$\sigma_\alpha^u \sim \mathcal{N}_+(0, 0.05)$
		$\beta_r^u \sim \mathcal{N}(\mu_\beta^u, \sigma_\beta^u)$	$\mu_\beta^u \sim \mathcal{N}(0, 0.05)$	$\sigma_\beta^u \sim \mathcal{N}_+(0, 0.05)$
		$\phi_y \sim \mathcal{N}(0.04, 0.02)$		
	Variance	$\tau_r^u \sim \mathcal{N}_+(0, \sigma_\tau^u)$		$\sigma_\tau^u \sim \mathcal{N}_+(0, 0.02)$
		$\eta_y \sim \mathcal{N}(0, 0.02)$		

$\mathcal{N}_+(\mu, \sigma)$ denotes half-normal distribution.