Forecasting elections with non-representative data

Literature / interesting precedents

Precedents on

The Xbox dataset (Wang et al., 2014)

- 750,148 interviews over the 45 days preceding the 2012 US presidential elections. 345,858 unique interviewees.

$$9\ \textit{questions}\ \left\{ \begin{array}{c} \text{Sex (2 categories)} \\ \text{Age (4 categories)} \\ \text{Education (4 categories)} \\ \text{Party ID (3 categories)} \\ \text{Ideology (3 categories)} \\ \text{2008 vote (3 categories)} \\ \text{Race (4 categories)} \\ \text{State (51 categories)} \\ \text{Voting intention (3 categories)} \end{array} \right\}$$

- Large but highly biased dataset.

Non-representativeness

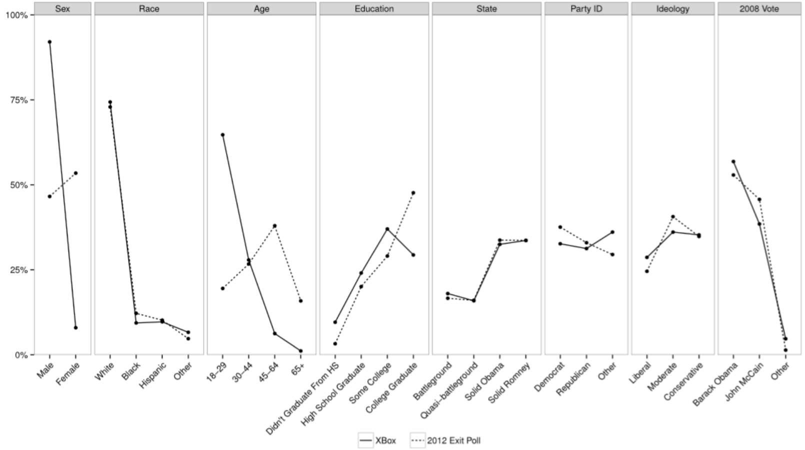

Figure 1:Xbox dataset vs adjustes exit 2012 exit poll

Methodology

Wang et al. try to correct for these bias and obtain accurate estimates in two steps

1. Estimating daily voting intent through Multilevel regression and poststratification (MRP).
2. Forecasting election results from estimates of daily voting intent and historical polling data through two nested regression models.

MRP (I)

- Poststratification: 176,256 cells. Too sparse to get reliable no-pooling estimates at a cell level.

$$\hat{y}^{PS} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j} \tag{1}$$

They use data from the 2008 exit poll to define N_j. Very useful approach, as it allows us to get estimates for any subset of interest from the population (in our case, provinces):

$$\hat{y}_s^{PS} = \frac{\sum_{j \in J_s} N_j \hat{y}_j}{\sum_{j \in J_s} N_j} \tag{2}$$

Is this technique equally effective when the biases come from less quantifiable variables? Age vs ideology.

MRP (II)

They estimate $\hat{y}_j$ in two steps:

1. $Pr(Y_i \in \{Obama, Romney\})$
   $= logit^{-1}(\alpha_0 + \alpha_1(state\ last\ vote\ share)$
   $+a^{state}_{j[i]} + a^{edu}_{j[i]} + a^{sex}_{j[i]} + a^{race}_{j[i]} + a^{partyID}_{j[i]} + b^{ideology}_{j[i]} + b^{lastvote}_{j[i]})$

with priors $a^{var}_{j[i]} \sim N(0, \sigma^2_{var})$ and $\sigma^2_{var} \sim inv - \chi^2(\nu, \sigma^2_0)$.

2. $Pr(Y_i = Obama | Y_i \in \{Obama, Romney\})$
   $= logit^{-1}(\beta_0 + \beta_1(state\ last\ vote\ share)$
   $+b^{state}_{j[i]} + b^{edu}_{j[i]} + b^{sex}_{j[i]} + b^{race}_{j[i]} + b^{partyID}_{j[i]} + b^{ideology}_{j[i]} + b^{lastvote}_{j[i]})$

with priors $a^{var}_{j[i]} \sim N(0, \eta^2_{var})$ and $\eta^2_{var} \sim inv - \chi^2(\mu, \eta^2_0)$.

MRP (III)

- Multilevel regression deals with sparsity.
- They run the model daily, using a four-day moving window.
- They use R package ‘lme4’, instead of full bayesian analysis, for computational convenience.

From daily estimates to election results (I)

- Need to go from daily estimates of voting intent to election day predictions.

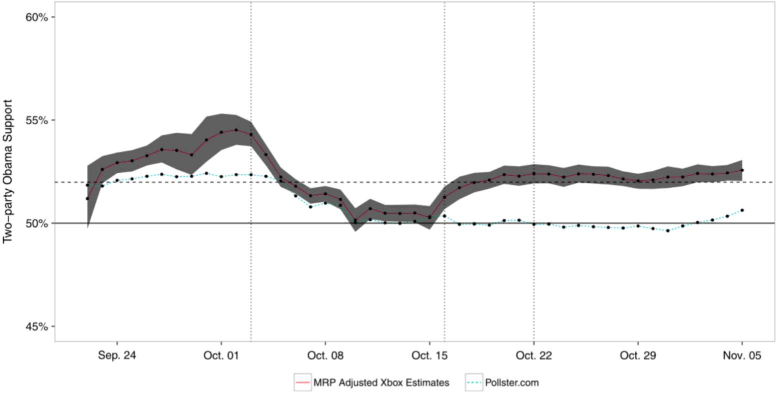

Figure 2:Adjusted MRP estimates vs Pollster.com

From daily estimates to election results (II)

- With historical data on polls from 2000, 2004 and 2008, they train two nested models

One at the national level

1. $y_e^{US} = a_0 + a_1 x_{t,e}^{US} + a_2 |x_{t,e}^{US}| x_{t,e}^{US} + a_3 t x_{t,e}^{US} + \eta(t, e)$

where $\eta \sim N(0, \sigma^2)$.

One at the state level

2. $y_{s,e}^{ST} = b_0 + b_1 x_{s,t,e}^{ST} + b_2 |x_{s,t,e}^{ST}| x_{s,t,e}^{ST} + b_3 t x_{s,t,e}^{ST} + \epsilon(s, t, e)$

where $Var(\epsilon(s, t, e)) = (t + a)^2$ and is allowed to be correlated across states.