eforensics: A Bayesian Implementation of A Positive Empirical Model of Election Frauds

Walter R. Mebane, Jr.

University of Michigan

Political Science 485, Oct 21, 2019

election forensics: a mixture model concept

- election forensics: use statistical methods to determine whether the results of an election accurately reflect the intentions of the electors
- Mebane (2016) is a likelihood implementation of the concept introduced by Klimek, Yegorov, Hanel and Thurner (2012) based on Normal distributions
- Ferrari, McAlister and Mebane (2018) and Mebane (2019) describe the Bayesian implementation in the **R** package `eforensics` of a formulation of a similar concept

core positive frauds model ideas

1. condition on the number of eligible voters at each observed aggregation unit, e.g., at each precinct or polling station
2. baseline assumption with no fraud: true vote distributions can be summarized by a conditional joint distribution for the number casting a valid vote and for the number voting for the "leader" at each unit
3. election fraud means that votes are added to the votes for the leader: some votes are manufactured from nonvoters and some votes are stolen from the "opposition"
4. the two kinds of election fraud refer to how many of the opposition votes and nonvoters counts are shifted
   - with "incremental fraud" moderate proportions of the votes are shifted
   - with "extreme fraud" almost all of the votes are shifted
5. frauds imply vote distributions are multimodal

observed data

- for aggregation units $i = 1, \ldots, n$ we observe counts
  - N_i: number of eligible voters at i
  - W_i: number of votes for the leader (sometimes "winner") at i
  - O_i: number of votes for opposition at i
  - $V_i = W_i + O_i$: number of valid votes at i
  - $A_i = N_i - V_i$: number of abstentions at i

observed data

- for aggregation units $i = 1, \ldots, n$ we observe counts
  - N_i: number of eligible voters at i
  - W_i: number of votes for the leader (sometimes "winner") at i
  - O_i: number of votes for opposition at i
  - $V_i = W_i + O_i$: number of valid votes at i
  - $A_i = N_i - V_i$: number of abstentions at i
- observed proportions
  - $t_i = V_i / N_i$: turnout proportion
  - $a_i = 1 - t_i$: proportion abstaining
  - $w_i = W_i / N_i$: leader proportion

unobserved data

- unobserved variables:
  - ν_i: true proportion of valid votes for the leader
  - τ_i: true turnout proportion
  - $Z_i \in \{1, 2, 3\}$: fraud type indicator
    - $Z_i = 1$: no fraud
    - $Z_i = 2$: incremental fraud
    - $Z_i = 3$: extreme fraud
  - ι_i^M, ι_i^S: proportion of votes manufactured from abstainers or stolen from opposition given incremental fraud
  - υ_i^M, υ_i^S: proportion of votes manufactured from abstainers or stolen from opposition given extreme fraud

general finite mixture model functional form

- conceptual formulation of data generating process:

$$E(a_i) \approx \begin{cases} 1-\tau_i, & \text{if } Z_i = 1 \\ (1-\tau_i)(1-\iota_i^M), & \text{if } Z_i = 2 \\ (1-\tau_i)(1-\upsilon_i^M), & \text{if } Z_i = 3 \end{cases} \tag{1a}$$

$$E(w_i) \approx \begin{cases} \nu_i\tau_i, & \text{if } Z_i = 1 \\ \nu_i\tau_i + \iota_i^M(1-\tau_i) + \iota_i^S\tau_i(1-\nu_i), & \text{if } Z_i = 2 \\ \nu_i\tau_i + \upsilon_i^M(1-\tau_i) + \upsilon_i^S\tau_i(1-\nu_i), & \text{if } Z_i = 3 \end{cases} \tag{1b}$$

  - a_i: (observed) proportion of N_i abstaining
  - w_i: (observed) leader proportion of N_i
  - ν_i: true proportion of valid votes for the leader
  - τ_i: true turnout proportion
  - $Z_i \in \{1,2,3\}$: {no fraud, incremental fraud, extreme fraud}
  - ι_i^M, ι_i^S: proportion manufactured or stolen | incremental fraud
  - υ_i^M, υ_i^S: proportion manufactured or stolen | extreme fraud

general finite mixture model functional form

- the model formulation is a finite mixture model: every aggregation unit is assumed to have all its counts from one of three conditions—no frauds, incremental fraud or extreme fraud

$$E\left(\frac{N_i - V_i}{N_i} = a_i\right) \approx \begin{cases} 1 - \tau_i, & \text{if } Z_i = 1 \\ (1 - \tau_i)(1 - \iota_i^M), & \text{if } Z_i = 2 \\ (1 - \tau_i)(1 - \upsilon_i^M), & \text{if } Z_i = 3 \end{cases}$$

$$E\left(\frac{W_i}{N_i} = w_i\right) \approx \begin{cases} \nu_i \tau_i, & \text{if } Z_i = 1 \\ \nu_i \tau_i + \iota_i^M (1 - \tau_i) + \iota_i^S \tau_i (1 - \nu_i), & \text{if } Z_i = 2 \\ \nu_i \tau_i + \upsilon_i^M (1 - \tau_i) + \upsilon_i^S \tau_i (1 - \nu_i), & \text{if } Z_i = 3 \end{cases}$$

- these comprise the components of the finite mixture model

qbl model: fraud probabiltiies

- the probabilties that there are frauds do not depend on conditioning factors
- we specify the Bayesian prior for the probabilities of no fraud (π_1), incremental fraud (π_2) and extreme fraud (π_3) so that π_1 is the largest probability

$$\tilde{\pi}_1 \sim U(0,1) \tag{2a}$$
$$\tilde{\pi}_2 \sim U(0,\tilde{\pi}_1) \tag{2b}$$
$$\tilde{\pi}_3 \sim U(0,\tilde{\pi}_1) \tag{2c}$$
$$\pi_j = \frac{\tilde{\pi}_j}{\tilde{\pi}_1 + \tilde{\pi}_2 + \tilde{\pi}_3}\,, \quad j \in \{1,2,3\} \tag{2d}$$

- the fraud type for each i has a single-draw multinomial prior

$$Z_i \sim \mathrm{Cat}(\boldsymbol{\pi}), \quad \boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3) \tag{3}$$

`qbl` model: logistic forms

- the likelihood for observed counts uses binomial distributions each having N_i “trials” and binomial probabilities given by (1a) and (1b); unknown proportions in (1a) and (1b) depend on covariates (at least intercepts) and random effects
- the unknown proportions are defined using logistic functions: for $k = .7$,

$$\nu_i = \frac{1}{1 + \exp[-(\beta^\top x_i^\nu + \kappa_i^\nu)]} \tag{4a}$$

$$\tau_i = \frac{1}{1 + \exp[-(\gamma^\top x_i^\tau + \kappa_i^\tau)]} \tag{4b}$$

$$\iota_i^l = \frac{k}{1 + \exp[-(\rho_l^\top x_i^\iota + \kappa_i^{\iota l})]}\, , l \in \{M, S\} \tag{4c}$$

$$\upsilon_i^l = k + \frac{1-k}{1 + \exp[-(\delta_l^\top x_i^\upsilon + \kappa_i^{\upsilon l})]}\, , l \in \{M, S\} \tag{4d}$$

`qbl` model: linear predictors

- each logistic function includes a linear predictor
- example: $\beta^\top x_i^\nu + \kappa_i^\nu$ is the linear predictor in

$$\nu_i = \frac{1}{1 + \exp[-(\beta^\top x_i^\nu + \kappa_i^\nu)]}$$

  - x_i^ν is a vector of observed covariates (including a constant term), and β is a vector of coefficients (Normal priors)
  - κ_i^ν is the realization of an unobserved random variable that for unknown mean $\mu^{\kappa\nu}$ and standard deviation $\sigma^{\kappa\nu}$ is assumed to have as prior the Normal distribution

$$\kappa_i^\nu \sim N(\mu^{\kappa\nu}, \sigma^{\kappa\nu})\,. \tag{5}$$

  Prior distributions for $\mu^{\kappa\nu}$ and $\sigma^{\kappa\nu}$ use standard Normal and exponential distributions:

$$\mu^{\kappa\nu} \sim N(0,1) \tag{6}$$
$$\sigma^{\kappa\nu} \sim \mathrm{Exp}(5) \tag{7}$$

`qbl` model: meaning of random effects

- in the true proportions of votes for the leader and the true turnout proportions, random effects capture overdispersion

$$\nu_i = \frac{1}{1+\exp[-(\beta^\top x_i^\nu + \kappa_i^\nu)]} \tag{8a}$$

$$\tau_i = \frac{1}{1+\exp[-(\gamma^\top x_i^\tau + \kappa_i^\tau)]} \tag{8b}$$

- in the fraud magnitude proportions, random effects capture additional variation in observation-level frauds: with $k = .7$

$$\iota_i^l = \frac{k}{1+\exp[-(\rho_l^\top x_i^\iota + \kappa_i^{\iota l})]}\,, l \in \{M,S\} \tag{9a}$$

$$\upsilon_i^l = k + \frac{1-k}{1+\exp[-(\delta_l^\top x_i^\upsilon + \kappa_i^{\upsilon l})]}\,, l \in \{M,S\} \tag{9b}$$

qbl model: estimation via MCMC

- fraud probabilities (π_1, π_2, π_3) are always positive
- estimation: Metropolis-Hastings (using JAGS) with MCMCSE stopping rules
  - the Metropolis-Hastings algorithm is a method for obtaining a sequence of random samples from a probability distribution: depending on the previous sample draw, a new draw is taken and then accepted or rejected with a probability that depends on the model and data
  - JAGS (Just Another Gibbs Sampler) is a software package for estimating models using MCMC (Markov Chain Monte Carlo) methods
  - MCMCSE (MCMC Standard Error) is a technique for deciding when the MCMC algorithm is drawing from the stationary distribution and so can be used to sample from the posterior distribution

qbl model: observation-level fraud estimates

- approach one: posterior mean only
- turnout and leader's vote proportions with incremental frauds

$$t_i^{\iota} = \tau_i + \iota_i^M (1 - \tau_i) \tag{10a}$$

$$w_i^{\iota} = \nu_i \frac{1 - \iota_i^S}{1 - \iota_i^M} \left(1 - \iota_i^M - \frac{A_i}{N_i}\right) + \frac{A_i}{N_i} \frac{\iota_i^M - \iota_i^S}{1 - \iota_i^M} + \iota_i^S \tag{10b}$$

and with extreme frauds

$$t_i^{\upsilon} = \tau_i + \upsilon_i^M (1 - \tau_i) \tag{11a}$$

$$w_i^{\upsilon} = \nu_i \frac{1 - \upsilon_i^S}{1 - \upsilon_i^M} \left(1 - \upsilon_i^M - \frac{A_i}{N_i}\right) + \frac{A_i}{N_i} \frac{\upsilon_i^M - \upsilon_i^S}{1 - \upsilon_i^M} + \upsilon_i^S \tag{11b}$$

using posterior mean estimates for τ_i, ν_i, ι_i^M, ι_i^S, υ_i^M and υ_i^S

qbl model: observation-level fraud estimates

- approach two: supports estimating posterior variability
- turnout and leader's vote proportions with incremental frauds

$$t_i^{\iota} = \tau_i + \iota_i^M(1-\tau_i) \tag{12a}$$
$$w_i^{\iota} = \tau_i\nu_i + \iota_i^M(1-\tau_i) + \iota_i^S\tau_i(1-\nu_i) \tag{12b}$$

and with extreme frauds

$$t_i^{\upsilon} = \tau_i + \upsilon_i^M(1-\tau_i) \tag{13a}$$
$$w_i^{\upsilon} = \tau_i\nu_i + \upsilon_i^M(1-\tau_i) + \upsilon_i^S\tau_i(1-\nu_i) \tag{13b}$$

using values from the MCMC chain for τ_i, ν_i, ι_i^M, ι_i^S, υ_i^M, υ_i^S

- supports computing credible intervals

qbl model: observation-level fraud estimates

- to compute posterior fraud proportions subtract the values that would occur if there were no frauds from the values that occur given that frauds occur:

$$p_{ti} = \begin{cases} 0, & \text{if } i \text{ is classified as no fraud} \\ t_i^{\iota} - \tau_i, & \text{if } i \text{ is classified as incremental fraud} \\ t_i^{\upsilon} - \tau_i, & \text{if } i \text{ is classified as extreme fraud} \end{cases}$$

$$p_{wi} = \begin{cases} 0, & \text{if } i \text{ is classified as no fraud} \\ w_i^{\iota} - \nu_i \tau_i, & \text{if } i \text{ is classified as incremental fraud} \\ w_i^{\upsilon} - \nu_i \tau_i, & \text{if } i \text{ is classified as extreme fraud}\,. \end{cases}$$

- numbers of fraudulent voters (turnout counts) and votes for the leading candidate at observation i are then $F_{ti} = p_{ti} N_i$ and $F_{wi} = p_{wi} N_i$

qbl model: fraud estimates variability

- values of unknown parameters occur in the stationary MCMC chain approximately as frequently as they are produced by the process that generated the data (as represented by the model we are using)
  - the posterior mean is estimated using the average of a quantity's values in the MCMC chain
- credible intervals: a range of values unknown parameters might have with specified probability
  - for $\alpha \in [0, 1]$, a credible interval for unknown parameter θ $(\theta_{\text{lower}}, \theta_{\text{upper}})$ is an interval of possible values of θ such that

$$\int_{\theta_{\text{lower}}}^{\theta_{\text{upper}}} \pi(\theta \mid \mathbf{x})\, d\theta = 1 - \alpha \tag{15}$$

  - the highest posterior density (HPD) credible region is defined by $\{\theta : \pi(\theta \mid \mathbf{x}) \geq c\}$ where c is chosen to solve

$$\int_{\{\theta:\pi(\theta|\mathbf{x})\geq c\}} \pi(\theta \mid \mathbf{x})\, d\theta = 1 - \alpha \tag{16}$$

References I

Ferrari, Diogo, Kevin McAlister and Walter R. Mebane, Jr. 2018. "Developments in Positive Empirical Models of Election Frauds: Dimensions and Decisions." Presented at the 2018 Summer Meeting of the Political Methodology Society, Provo, UT, July 16–18.

Klimek, Peter, Yuri Yegorov, Rudolf Hanel and Stefan Thurner. 2012. "Statistical Detection of Systematic Election Irregularities." *Proceedings of the National Academy of Sciences* 109(41):16469–16473.

Mebane, Jr., Walter R. 2016. "Election Forensics: Frauds Tests and Observation-level Frauds Probabilities." Paper presented at the 2016 Annual Meeting of the Midwest Political Science Association, Chicago, April 7–10, 2016.

Mebane, Jr., Walter R. 2019. "Notes Regarding Use of the **R** package `eforensics`: for POLSCI 485 in Fall 2019." working paper, class notes.