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:**
  - $\nu_i$: true proportion of valid votes for the leader
  - $\tau_i$: 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
  - $\iota_M$, $\iota_S$: proportion of votes manufactured from abstainers or stolen from opposition given incremental fraud
  - $\nu_i^M$, $\nu_i^S$: proportion of votes manufactured from abstainers or stolen from opposition given extreme fraud
Election Forensics

References

A Positive Empirical Model of Election Frauds

eforensics Model Quasi-Binomial (qbl) Specification

general finite mixture model functional form

- conceptual formulation of data generating process:

\[ E(a_i) \approx \begin{cases} 
1 - \tau_i, & \text{if } Z_i = 1 \\
(1 - \tau_i)(1 - \nu_i^M), & \text{if } Z_i = 2 \\
(1 - \tau_i)(1 - \nu_i^S), & \text{if } Z_i = 3 
\end{cases} \]  

(1a)

\[ E(w_i) \approx \begin{cases} 
\nu_i \tau_i, & \text{if } Z_i = 1 \\
\nu_i \tau_i + \nu_i^M (1 - \tau_i) + \nu_i^S \tau_i (1 - \nu_i), & \text{if } Z_i = 2 \\
\nu_i \tau_i + \nu_i^M (1 - \tau_i) + \nu_i^S \tau_i (1 - \nu_i), & \text{if } Z_i = 3 
\end{cases} \]  

(1b)

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

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

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

- these comprise the components of the finite mixture model
qbl model: fraud probabilities

- The probabilities that there are frauds do not depend on conditioning factors.
- We specify the Bayesian prior for the probabilities of no fraud ($\pi_1$), incremental fraud ($\pi_2$) and extreme fraud ($\pi_3$) so that $\pi_1$ is the largest probability.

\[
\begin{align*}
\tilde{\pi}_1 &\sim U(0, 1) \quad (2a) \\
\tilde{\pi}_2 &\sim U(0, \tilde{\pi}_1) \quad (2b) \\
\tilde{\pi}_3 &\sim U(0, \tilde{\pi}_1) \quad (2c)
\end{align*}
\]

\[
\pi_j = \frac{\tilde{\pi}_j}{\tilde{\pi}_1 + \tilde{\pi}_2 + \tilde{\pi}_3}, \quad j \in \{1, 2, 3\} \quad (2d)
\]

- The fraud type for each $i$ has a single-draw multinomial prior.

\[
Z_i \sim \text{Cat}(\pi), \quad \pi = (\pi_1, \pi_2, \pi_3) \quad (3)
\]
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 \( k = .7 \),

\[
\nu_i = \frac{1}{1 + \exp[-(\beta^T x_i^\nu + \kappa_i^\nu)]} \\
\tau_i = \frac{1}{1 + \exp[-(\gamma^T x_i^\tau + \kappa_i^\tau)]} \\
\nu_i^l = \frac{k}{1 + \exp[-(\rho_l^T x_i^l + \kappa_i^{\nu l})]} , l \in \{M, S\} \\
\nu_i^l = k + \frac{1 - k}{1 + \exp[-(\delta_l^T x_i^\nu + \kappa_i^{\nu l})]} , l \in \{M, S\}
\]

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

- $x_i^\nu$ is a vector of observed covariates (including a constant term), and $\beta$ is a vector of coefficients (Normal priors)
- $\kappa_i^\nu$ 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
  $$
  \kappa_i^\nu \sim N(\mu^{\kappa \nu}, \sigma^{\kappa \nu})
  $$

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

$$
\begin{align*}
  \mu^{\kappa \nu} &\sim N(0, 1) \quad (6) \\
  \sigma^{\kappa \nu} &\sim \text{Exp}(5)
\end{align*}
$$
qbl model: meaning of random effects

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

\[
\nu_i = \frac{1}{1 + \exp[-(\beta^T x_i^\nu + \kappa_i^\nu)]} \quad (8a)
\]
\[
\tau_i = \frac{1}{1 + \exp[-(\gamma^T x_i^\tau + \kappa_i^\tau)]} \quad (8b)
\]

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

\[
\nu_i^l = \frac{k}{1 + \exp[-(\rho_i^l x_i^\nu + \kappa_i^\nu)]}, \quad l \in \{M, S\} \quad (9a)
\]
\[
\nu_i^l = k + \frac{1 - k}{1 + \exp[-(\delta_i^l x_i^\nu + \kappa_i^\nu)]}, \quad l \in \{M, S\} \quad (9b)
\]
qbl model: estimation via MCMC

- Fraud probabilities ($\pi_1, \pi_2, \pi_3$) 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

\[ t^t_i = \tau_i + \nu_i^M (1 - \tau_i) \]  \hspace{1cm} (10a)

\[ w^t_i = \nu_i \frac{1 - \nu_i^S}{1 - \nu_i^M} \left( 1 - \nu_i^M - \frac{A_i}{N_i} \right) + \frac{A_i \nu_i^M - \nu_i^S}{N_i \left( 1 - \nu_i^M \right)} + \nu_i^S \]  \hspace{1cm} (10b)

and with extreme frauds

\[ t^v_i = \tau_i + \nu_i^M (1 - \tau_i) \]  \hspace{1cm} (11a)

\[ w^v_i = \nu_i \frac{1 - \nu_i^S}{1 - \nu_i^M} \left( 1 - \nu_i^M - \frac{A_i}{N_i} \right) + \frac{A_i \nu_i^M - \nu_i^S}{N_i \left( 1 - \nu_i^M \right)} + \nu_i^S \]  \hspace{1cm} (11b)

using posterior mean estimates for \( \tau_i, \nu_i, \nu_i^M, \nu_i^S, \nu_i^M \) and \( \nu_i^S \)
qbl model: observation-level fraud estimates

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

\[
\begin{align*}
t_i^k &= \tau_i + \nu_M^i (1 - \tau_i) \\
w_i^k &= \tau_i \nu_i + \nu_M^i (1 - \tau_i) + \nu_S^i \tau_i (1 - \nu_i)
\end{align*}
\]

and with extreme frauds

\[
\begin{align*}
t_i^\nu &= \tau_i + \nu_M^i (1 - \tau_i) \\
w_i^\nu &= \tau_i \nu_i + \nu_M^i (1 - \tau_i) + \nu_S^i \tau_i (1 - \nu_i)
\end{align*}
\]

using values from the MCMC chain for \( \tau_i, \nu_i, \nu_M^i, \nu_S^i \)

- 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:

\[
p_{ti} = \begin{cases} 
0, & \text{if } i \text{ is classified as no fraud} \\
(t_i^l - \tau_i), & \text{if } i \text{ is classified as incremental fraud} \\
(t_i^u - \tau_i), & \text{if } i \text{ is classified as extreme fraud}
\end{cases}
\]

\[
p_{wi} = \begin{cases} 
0, & \text{if } i \text{ is classified as no fraud} \\
(w_i^l - \nu_i \tau_i), & \text{if } i \text{ is classified as incremental fraud} \\
(w_i^u - \nu_i \tau_i), & \text{if } i \text{ is classified as extreme fraud}
\end{cases}
\]

- 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 $\theta$ ($\theta_{\text{lower}}, \theta_{\text{upper}}$) is an interval of possible values of $\theta$ such that
    \[ \int_{\theta_{\text{lower}}}^{\theta_{\text{upper}}} \pi(\theta \mid x) \, d\theta = 1 - \alpha \] \tag{15}
  - the highest posterior density (HPD) credible region is defined by \( \{ \theta : \pi(\theta \mid x) \geq c \} \) where $c$ is chosen to solve
    \[ \int_{\{ \theta : \pi(\theta \mid x) \geq c \}} \pi(\theta \mid x) \, d\theta = 1 - \alpha \] \tag{16}
References


