Semi-Parametric Identification and Estimation of Ballot Stuffing

Anastasia Burkovskaya

School of Economics, University of Sydney

January 5, 2019

Motivation

- Elections and referenda help observing preferences of population that might be used for building a future political agenda.
- However, in countries where voting is not compulsory, many people decide not to vote and due to selection bias, the preferences of the entire population stay unknown.
  - For young British voters the turnout was very low but the government still has to negotiate the agreements related to student visas for those who would like to study in the EU.
  - Exceptional turnout of a specific socioeconomic group might play a large role, e.g., an unusual spike in African-American turnout in the 2008 U.S. presidential elections due to the Obama candidacy.
- How to estimate the preferences of the entire population from the available electoral data?
- How to identify the electoral preferences in the presence of ballot stuffing?

Results

- Offer a structural model that allows for derivation of joint distribution of turnout and voter share from unobservable preferences and costs of voting
  - The model assumes that preferences are exogenous in order to avoid assumptions about ideology, valence, etc.
  - Single elections
- Identification and semi-parametric estimation of the model with the normal asymptotics
- Application: ballot stuffing - illegal addition of extra ballots into the urn
  - Clean subsample allows to identify distribution of the fraud given observables at each polling station
- Empirical illustration on the 2011 Russian parliamentary dataset
  - Parametric vs. semi-parametric approach

Literature

- Electoral fraud:
  - Statistical irregularities: Benford's law (Mebane (2006), Breunig and Goerres (2011)), unusual kurtosis of the distribution of electoral data (Klimek et al. (2012)), etc.
    - ★ Test the presence of fraud, but do not evaluate its amount.
  - Natural experiments (Cantu (2014)) and randomized assignment of independent observers (Enikolopov et al (2013)).
    - ★ Evaluate the amount of fraud on average between stations with observers vs. stations without them. Without the structural model researchers cannot infer the amount of fraud at a polling station level.
  - Parametric model: Levin et al. (2009)
    - ★ Assumptions about underlying distributions of voter preferences

Electoral Model: Preferences

- Candidates, A and B are running for office
- A voter i in a polling station j in region K prefers A to B if

$$\sigma_A^{ijK} + \delta_A^{jK} + \mu_A^K > \sigma_B^{ijK} + \delta_B^{jK} + \mu_B^K$$

  - σ^{ijK} is a parameter of individual "pure" preferences towards the candidates
  - δ^{jK} is popularity of a candidate in the area of the polling station and is the same for one polling station j
  - μ^K is a regional effect in popularity of each candidate and it is a function of some observable characteristics of the region X_K, such as average income, level of education, share of old population, etc.

Electoral Model: Reduced Preferences

- A voter i in a polling station j in region K prefers A to B if

$$\sigma^{ijK} + \delta^{jK} + h(X_K) < 0$$

  - $\sigma^{ijK} = \sigma_B^{ijK} - \sigma_A^{ijK}$ is a parameter of individual "pure" preferences towards the candidates
  - $\delta^{jK} = \delta_B^{jK} - \delta_A^{jK}$ is popularity of a candidate in the area of the polling station and is the same for one polling station j
  - $\mu^K = \mu_B^K - \mu_A^K \equiv h(X_K)$ is a regional effect in popularity of candidates and it is a function of some observable characteristics of the region X_K

Electoral Model: Turnout

- A voter chooses to participate in elections if the difference in her preferences from different candidates is higher than costs of participation:

$$|\sigma^{ijK} + \delta^{jK} + h(X_K)| \geq c^{jK}$$

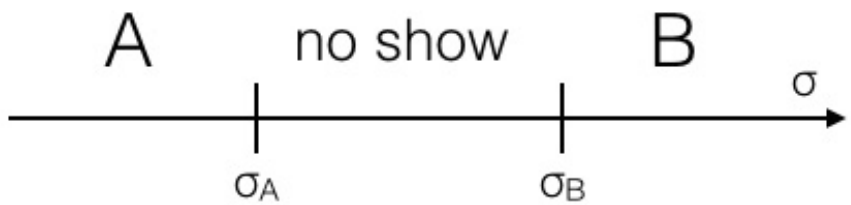
  - c^{jK} are random and the same for all voters in the same polling station, but different across different polling stations
  - Costs might represent the length of line to vote, the weather, difficulty of obtaining a voter card, etc.
  - If the preferences of a voter are close to indifference between the candidates, then she does not attend elections

Electoral Model: Swing Voters

- "Swing voters", σ_A^{jK} and σ_B^{jK}, in every polling station j of region K, who are indifferent between participating and not participating in elections:

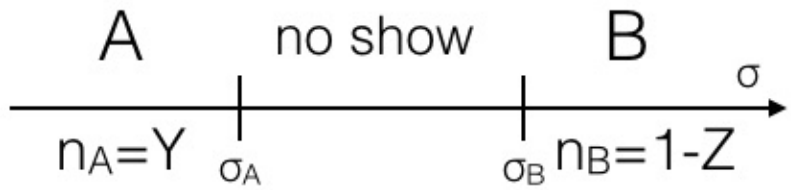
$$\sigma_A^{jK} = -\delta^{jK} - \mu^K - c^{jK}$$
$$\sigma_B^{jK} = \sigma_A^{jK} + 2c^{jK}$$

  - People with "pure" preferences $\sigma^{ijK} < \sigma_A^{jK}$ will vote for the candidate A
  - People with $\sigma^{ijK} > \sigma_B^{jK}$ will choose the candidate B
  - Everybody in between the swing voters σ_A^{jK} and σ_B^{jK} will abstain from elections



Electoral Model: Observables

- The number of people who vote for A in a polling station j in region K is $n_A^{jK} = \int_{-\infty}^{\sigma_A^{jK}} dG(x) = G(\sigma_A^{jK})$
- The same number for candidate B is $n_B^{jK} = \int_{\sigma_B^{jK}}^{+\infty} dG(x) = 1 - G(\sigma_B^{jK})$
- Turnout in the polling station is $\tau^{jK} = 1 - G(\sigma_B^{jK}) + G(\sigma_A^{jK})$
- A's share of votes is $\pi_A^{jK} = \frac{n_A^{jK}}{n_A^{jK}+n_B^{jK}} = \frac{G(\sigma_A^{jK})}{1-G(\sigma_B^{jK})+G(\sigma_A^{jK})}$



- Define the electoral variables:

$$Y^{jK} = G(\sigma_A^{jK}) = \pi_A^{jK}\tau^{jK}$$

$$Z^{jK} = G(\sigma_B^{jK}) = 1 - \tau^{jK} + G(\sigma_A^{jK}) = 1 - \tau^{jK} + \pi_A^{jK}\tau^{jK}$$

Electoral Model: Assumptions

Assumption 1.

Personal preferences σ is independent on X, δ and c, its support in $\mathbb{R}$ is compact, and it has continuously differentiable density $g(\cdot)$, strictly increasing on the support cumulative distribution function $G(\cdot)$, and $E\sigma = 0$.

Assumption 2.

Local preferences δ and costs of voting c have continuously differentiable joint density $f_{\delta,c}(\cdot,\cdot)$, cumulative distribution function $F_{\delta,c}(\cdot,\cdot)$, and they are independent on regional characteristics X.

Ballot Stuffing

- Ballot-stuffing: a number q of unused ballots are filled for the "right" candidate by a polling station official
- Assumption: only candidate A has access to unused ballots
- Total number of votes for A: $Y \Longrightarrow \bar{Y} = Y + q$
- Total number of votes for B: Z stays the same

Theorem 1.

If Assumptions 1 and 2 hold and $h(X) = \beta' X$, then $g(\cdot)$ and coefficients β are identified.

Estimation: β and $g(G^{-1}(\cdot))$

Identification strategy suggests the following estimator:

$$\widehat{G^{-1}(z)} = \hat{c}\hat{F}_Z(z),$$

where $\hat{c} = \int \frac{1}{f_Z(z)} dz$.
In addition,

$$Z_i = G\left(-\delta_i + c_i - \beta' X_i\right)$$

$\Rightarrow \hat{\beta}$ is an OLS estimator of $\widehat{G^{-1}(Z)}$ on X.

Empirical Illustration

- Dataset – the 2011 Russia parliamentary elections (94,795 obs.)
- Region – TIK for non-urban areas and city for urban areas (2,483 regions)
- Regional characteristics X are the total number of voters and the average size of a polling station normalized to 1
- $Y =$ the number of votes for the United Russia
- $Z = 1$ – the number of pro-opposition votes

Parametric Empirical Illustration: Assumptions

Assumption 3.

Suppose that $\sigma \sim U[0,1]$.

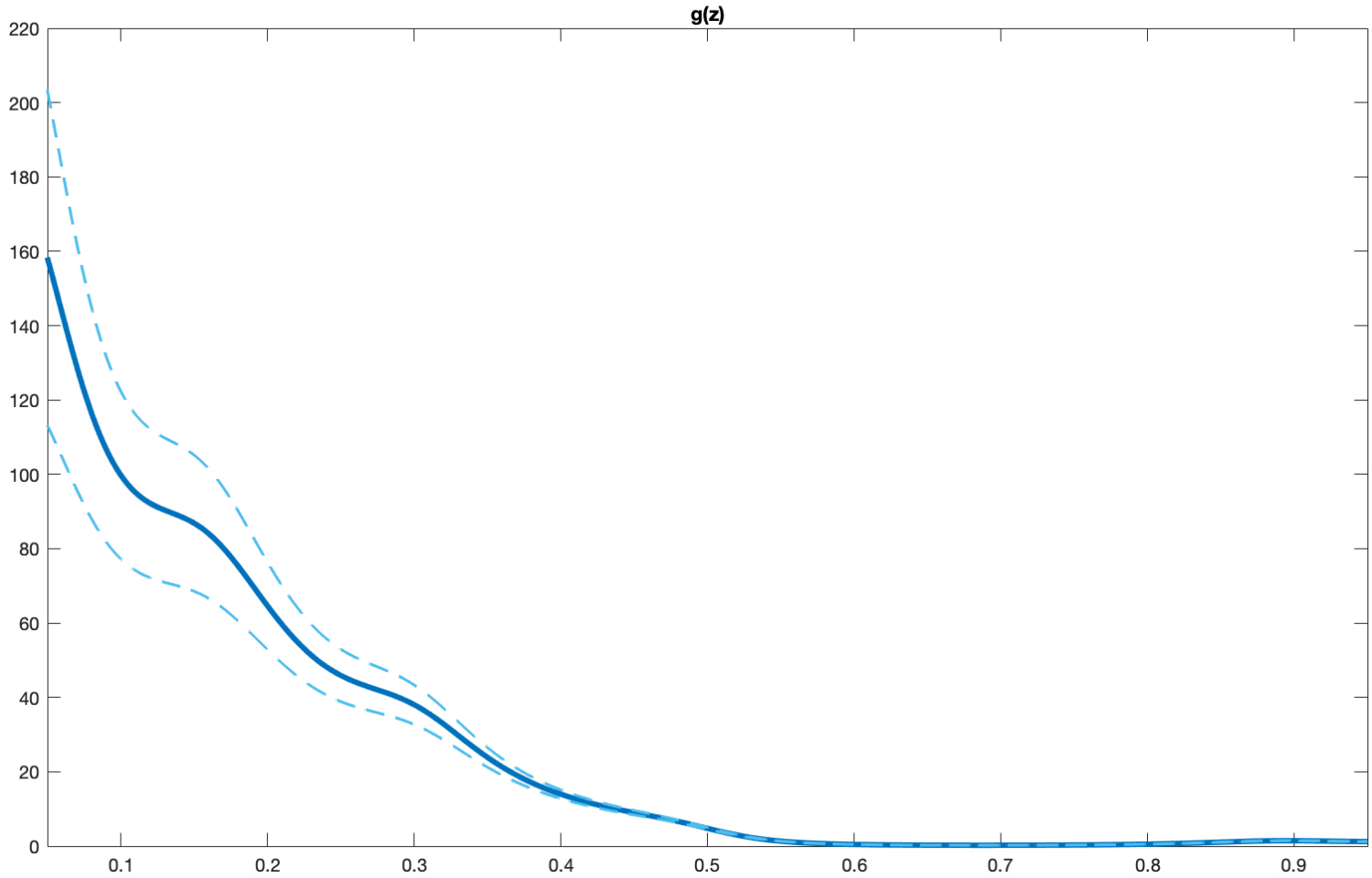
Assumption 4.

Suppose that $(\delta, c)' \sim N(\mu, \Sigma)$.

Assumptions 3 and 4 provide linearity of observables in regional characteristics:

$$Z = -\delta + c - X\beta$$

Non-parametric estimation: $g(\cdot)$



Note the difference in the shape with the uniform density used in the parametric illustration.

Parametric vs. Semi-parametric Estimation

	const	β_1	β_2
parametric coefficient	0.7550	0.0221	1.5986
standard error	0.0008	0.0021	0.0297
semi-parametric coefficient	0.6008	0.0677	3.9765
standard error	0.0082	0.0119	0.3682

- X_1 – the total number of voters
- X_2 – the average size of a polling station

Ballot Stuffing: "clean" subsample

Theorem 2.

The joint distribution of local preferences and costs $f_{\delta,c}(\cdot,\cdot)$ can be identified if $f_{Y,Z|X=X^k}$ can be recovered from a "clean" subsample in a region with characteristics X^k. Moreover, if $q \perp\!\!\!\perp c|X,Z$, then the distribution of the amount of fraud q, $f_{q|X,\bar{Y},Z}(\cdot)$, is identified at every polling station.

Estimation of Ballot Stuffing

- Estimate $\hat{f}_{Y,Z|X^k}(y,z)$ in the clean region
- Estimate $f_{-\delta-c,-\delta+c}(\cdot)$:

$$\hat{f}_{-\delta-c,-\delta+c}\left(\hat{h}(X^k)+\widehat{G^{-1}(y)},\hat{h}(X^k)+\widehat{G^{-1}(z)}\right)=\hat{f}_{Y,Z|X^k}(y,z)\hat{g}(\widehat{G^{-1}}(y))\hat{g}(\widehat{G^{-1}}(z))$$

- Obtain $f_{Y,Z|X}(\cdot,\cdot)$ for any region X by using the same formula.
- Evaluate q from $\bar{Y}=Y+q$:

$$\hat{f}_{q|X^j,\bar{Y},Z}(q)=\hat{f}_{Y|X^j,Z}(\bar{Y}-q)$$

Empirical Illustration: Ballot Stuffing

- Clean subsample – data from independent observers in Moscow who did not report violations (75 polling stations)
- We assume joint normality of Y and Z in Moscow and fit the distribution

	$\hat{\mu}_{Y\|\tilde{X}}$	$\hat{\mu}_{Z\|\tilde{X}}$	$\hat{\sigma}^2_{Y\|\tilde{X}}$	$\hat{\sigma}^2_{Z\|\tilde{X}}$	$\hat{\sigma}_{YZ\|\tilde{X}}$
parameter	0.1258	0.6270	0.0015	0.0023	-0.0001
standard error	0.0044	0.0056	0.0002	0.0004	0.0002

Ballot Stuffing: Parametric vs. Semi-parametric Estimation

Kostroma – UIK 213 and UIK 299

	UIK 213^n	UIK 299^n	UIK 213^p	UIK 299^p
$\bar{Y}$	0.1499	0.1964	0.1499	0.1964
average ballot stuffing	0.0624	0.0802	-0.0046	0.0404
standard error	0.0030	0.0034	0.0047	0.0058
reported voter share of UR	29.30%	33.68%	29.30%	33.68%
expected voter share of UR	19.47%	23.11%	29.93%	28.74%

Conclusion

- Offer a structural model that allows for derivation of joint distribution of turnout and voter share from unobservable preferences and costs of voting
- Identification and semi-parametric estimation of the model
- Identification and semi-parametric estimation of ballot stuffing from clean subsample of the electoral data
- Parametric vs. semi-parametric empirical illustration based on the 2011 Russia parliamentary data