정치 여론조사 자료에 대한 베이지안 메타 분석

서울대학교 정치외교학부
박종희

모형

선거후보에 대한 여론조사의 경우 우리가 관측하는 자료는 다음 세 가지 관측단위를 기본으로 갖는다고 볼 수 있다:

- i: 조사기관
- j: 조사대상(예: 후보의 지지율)
- t: 조사일

조사결과는 조사기관, 조사대상, 그리고 조사일에 따라 변화하는데 우리의 목적은 특정 시점까지의 자료를 기반으로 해서 다음 시점에서의 조사결과를 예측(forecast)하는 것이다. 특정한 패턴을 보이지 않는 시계열자료의 단기 예측에서 가장 작은 평균제곱오차(mean squared error)를 보이는 것은 동적선형모형(dynamic linear model)이다. 동적선형모형은 관측자료의 생성과정을 초기분포, 관측방정식(observation equation)과 상태방정식(state equation)으로 분해해서 재현한다.

여기서 상태방정식은 추세(γ_{jt})의 변화를 임의보행(random walk) 과정으로 가정한다. 관측방정식은 추세의 움직임에 조사기관 효과(α_{ij})와 메뉴효과($\mathbf{D}_{ijt}\boldsymbol{\beta}$), 그리고 조사오차에서 발생하는 노이즈($\epsilon_{ijt}$)를 결합하여 개별 관측치가 결정된다고 가정한다.

모형의 구성요소들은 다음과 같다:

- γ_{j0}: 추세의 초기값. 초기분포로부터 무작위로 추출
- y_{ijt}: i조사기관의 j후보에 대한 t시점에서의 개별 조사 관측치
- γ_{jt}: t시점에서 j후보에 대한 여론의 추세
- γ_{jt-1}: $t-1$시점에서 j후보에 대한 여론의 추세
- $\mathbf{D}_{ijt}\boldsymbol{\beta}_j$: 메뉴효과. j후보를 제외한 나머지 후보가 설문조사에서 빠짐에 따라 j후보의 지지율이 변하는 정도. $\mathbf{D}_{ijt} = (d_{1,ijt}, \ldots, d_{k,ijt})$이며 만약 i조사기관의 t시점에서의 조사가 k번째

후보를 설문에서 빠트렸으면 $d_{k,ijt} = 1$.

- ϵ_{ijt}: y_{ijt}의 조사오차에서 발생하는 랜덤 노이즈
- ε_{jt}: γ_{jt}의 변화 과정에서 동반되는 랜덤 노이즈
- σ^2_ε: γ_{jt}의 변화 과정의 변동폭을 반영하는 분산모수
- α_{ij}: i조사기관의 j후보에 대한 조사기관 효과
- $\sigma^2_{\alpha,j}$: j후보에 대한 조사기관 효과의 분산

이 구성요소들을 모형 안에 집어 넣어서 작성하면,

$$\begin{cases} \gamma_{j0} \sim \mathcal{N}(\mu_0, V_0) & \text{(초기분포)} \\ y_{ijt} = \alpha_{ij} + \gamma_{jt} + \mathbf{D}_{ijt}\boldsymbol{\beta}_j + \epsilon_{ijt}, & \epsilon_{ijt} \sim \mathcal{N}(0, \sigma^2_\epsilon) \quad \text{(관측방정식)} \\ \gamma_{jt} = \gamma_{jt-1} + \varepsilon_{jt}, & \varepsilon_{jt} \sim \mathcal{N}(0, \sigma^2_\varepsilon) \quad \text{(상태방정식)} \\ \alpha_{ij} \sim \mathcal{N}(0, \sigma^2_{\alpha,j}) & \text{(조사기관효과)}. \end{cases} \tag{1}$$

위 모형의 모수에 사전분포를 결합하면, 베이지안 추론방법에 의해 모수의 사후분포를 추출할 수 있다. 단 σ^2_ϵ는 표본사이즈의 다양성을 반영하기 위해 조사기관에 의해 보고된 표본오차 분포를 직접 반영한다.

우리가 관심을 갖는 사후분포는

- $p(\boldsymbol{\gamma}_j|\mathbf{y}_j, \mathbf{D}_j)$: 여기서 $\boldsymbol{\gamma}_j = (\gamma_{j1}, \ldots, \gamma_{jT})$이고 $\mathbf{y}_j, \mathbf{D}_j$는 j에 대해 관측된 자료 전체를 나타냄.
- $p(\gamma_{j,t+1}|\mathbf{y}_j, \mathbf{D}_j)$: 우리가 마지막으로 관측한 자료 바로 다음 시점$(t+1)$에서 j에 대한 추세의 예측치 분포. 베이즈정리를 이용하면 이는 다음과 같이 구할 수 있다:

$$p(\gamma_{j,t+1}|\mathbf{y}_j, \mathbf{D}_j) = \int p(\gamma_{j,t+1}|\mathbf{y}_j, \mathbf{D}_j, \boldsymbol{\theta}_j) p(\boldsymbol{\theta}_j|\mathbf{y}_j, \mathbf{D}_j, \boldsymbol{\gamma}_j) d\boldsymbol{\theta}_j. \tag{2}$$

  여기서 $\boldsymbol{\theta}_j = (\boldsymbol{\gamma}_j, \sigma^2_{\alpha,j}, \sigma^2_\varepsilon, \boldsymbol{\beta}_j)$.

- $p(\alpha_{ij}|\mathbf{y}_j, \mathbf{D}_j)$: 조사기관 효과의 분포. 조사기관 효과의 평균은 0으로 설정되며 개별 조사기관에 대한 조사기관 효과와 분산의 크기는 자료로부터 직접 추정된다.