정치 여론조사 자료에 대한 베이지안 메타 분석

Bayesian Meta-Analysis of Political Polling Data

박종희

서울대학교 정치외교학부

1 서론

정치 여론조사 자료는 일반적으로 다음과 같은 세 가지 단위를 기반으로 수집된다:

- i: 조사기관
- j: 조사 대상(예: 특정 후보나 정당에 대한 지지율)
- t: 조사 시점

정치 여론조사 자료를 이용한 메타분석의 주요 목적은 특정 시점까지의 다수 조사자료를 통합하여 여론의 추세(trend)를 분석하고, 이를 바탕으로 미래 여론 동향을 예측하는 것이다. 한국 정치 여론은 급변성이 크고 조사기관마다 방법론적 차이가 존재하기 때문에, 개별 조사 자료는 실제 여론 추세뿐 아니라 조사기관의 방법론적 특성과 응답자 구성의 편향까지 함께 반영한다고 볼 수 있다. 이러한 복합적 요소들을 효과적으로 모델링할 수 있는 통계적 방법으로 베이지안 다층 동적 선형 모형(Bayesian Multilevel Dynamic Linear Model, DLM)이 적합하다.

DLM은 자료 생성 과정을 초기 분포, 관측 방정식, 상태 방정식으로 분해하여 복잡한 시계열 자료의 구조를 체계적으로 재현한다. 그동안 여론M은 조사기관 효과를 고정된 값으로 가정하는 'Multilevel Dynamic Linear Model with Constant House Effects'를 통해 메타분석을 수행해왔다.

그러나 2024년 명태균 사건, 비상계엄 선포, 현직 대통령 체포, 그리고 헌법재판소의 탄핵 인용으로 이어지는 일련의 정치적 격변 과정에서 기존 메타분석 방법론의 한계가 드러났다. 특히 조사기관 효과를 고정적인 것으로 보는 가정이 현실을 정확히 반영하지 못한다는 점이 확인되었고, 이에 다양한 새로운 분석모형을 개발하고 테스트했다.

광범위한 검증 결과, 조사기관 효과의 시간적 변동을 반영하는 'Multilevel Dynamic Linear Model with Time-varying House Effects'가 가장 안정적인 성능과 높은 정확성을 보이는 것으로 확인되었다. 또한 대선 후보 구성의 차이가 여론조사 결과에 미치는 영향을 통제하기 위해 'Multilevel Dynamic Linear Model with Menu Effects'를 개발하여 대통령선거 기간 동안의 지지도 분석에 활용하고 있다.

이러한 발전된 모형들은 2025 MBC 여론M의 공식 분석 방법론으로 채택되어, 정당지지도는 조사기관 효과 변동 모형으로, 대통령선거 후보 지지도는 메뉴 효과 모형과 조사기관 효과 변동 모형을 결합한 복합 모형으로 추정되고 있다. 이러한 접근법은 조사기관의 특성과 시간적 변동성, 후보자 구성의 차이가 미치는 영향을 체계적으로 통제함으로써, 한국의 역동적인 정치 여론조사 환경에 최적화된 분석 프레임워크를 제공한다.

2 기본 모형 (Multilevel Dynamic Linear Model with Constant House Effects)

이 모형에서는 조사기관 효과가 시간에 따라 일정하다고 가정한다. 즉 각각의 조사기관은 고유의 조사기법과 방법으로 인해 다른 조사기관의 조사결과와 갖는 일정한 차이가 있다는 가정이다. 또는 그 조사기관의 조사방법으로 수집된 응답자들의 응답태도가 다른 조사기관 조사방법으로 수집된 응답자들의 응답태도와 일정한 차이를 갖는다는 가정이다. 이를 모형으로 표현하면 다음과 같다:

$$
\begin{aligned}
\alpha_{j1} &\sim \text{Uniform}(a_{01}, a_{02}) \\
\alpha_{jt} &= \alpha_{j,t-1} + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim \mathcal{N}(0, \sigma_\alpha^2) \\
y_{ijt} &= \text{house}_{ij[t]} + \alpha_{jt[i]} + \epsilon_{ijt}, \quad \epsilon_{ijt} \sim \mathcal{N}(0, \sigma_{jt}^2) \\
\sigma_{jt}^2 &= N_{jt}^{-1} \sum_i \frac{y_{ijt}(1 - y_{ijt})}{s_{ijt}} \\
\text{house}_{ij} &\sim \mathcal{N}(0, 1) \\
\sigma_\alpha &\sim \text{Uniform}(\sigma_{low}, \sigma_{high}).
\end{aligned}
$$

s_{ijt}는 개별 조사의 표본 크기이고 N_{jt}은 t 시점에 j에 관해 발표된 조사의 수이다.
주요 기호 설명:

- α_{jt}: 후보 j의 시점 t에서의 여론 추세
- house_{ij}: 조사기관 i의 후보 j에 대한 조사기관 효과

3 조사기관 효과 변동 모형 (Multilevel Dynamic Linear Model with Time-varying House Effects)

이 모형은 조사기관의 효과(house effect)가 시간에 따라 동적으로 변화할 수 있다는 가정에 기반한다. 이러한 접근은 조사 방법론의 변화, 응답자 구성의 변동, 또는 사회적 맥락의 변화 등 다양한 요인을 반영한다. 조사기관 효과의 시간적 변화를 모델링하는 것은 특히 사회적으로 중요한 사건 전후로 여론 조사 결과에 체계적인 편향이 발생할 수 있기 때문에 중요하다. 예를 들어 비상계엄 선포, 윤석열 대통령 체포, 서부지법 폭동, 탄핵인용 등의 사건 전후로 보수 및 진보 성향 응답자들의 조사 참여율이 차별적으로 변화할 수 있다. 예컨대, 특정 시점에서는 보수 진영 응답자가 자동응답시스템(ARS) 조사에 과다하게 포함되는 반면, 다른 시점에서는 진보 진영 응답자가 과다 표집되는 현상이 발생할 수 있다.

$$
\begin{aligned}
\alpha_{j1} &\sim \text{Uniform}(a_{01}, a_{02}) \\
\alpha_{jt} &= \alpha_{j,t-1} + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim \mathcal{N}(0, \sigma_\alpha^2) \\
y_{ijt} &= \text{house}_{ijt} + \alpha_{jt[i]} + \epsilon_{ijt}, \quad \epsilon_{ijt} \sim \mathcal{N}(0, \sigma_{jt}^2) \\
\sigma_{jt}^2 &= N_{jt}^{-1} \sum_i \frac{y_{ijt}(1 - y_{ijt})}{s_{ijt}} \\
\text{house}_{ij,1} &\sim \mathcal{N}(0, \sigma_{\text{initial}}^2) \\
\text{house}_{ij,t} &= \text{house}_{ij,t-1} + \eta_{ij,t}, \quad \eta_{ij,t} \sim \mathcal{N}(0, \sigma_{\text{house}}^2) \\
\sigma &\sim \text{Uniform}(\sigma_{low}, \sigma_{high}) \\
\sigma_{\text{house}} &\sim \text{Uniform}(0, \sigma_{\text{house max}}) \\
\sigma_{\text{initial}} &\sim \text{Uniform}(0, \sigma_{\text{initial max}})
\end{aligned}
$$

4 메뉴 효과 모형 (Multilevel Dynamic Linear Model with Menu Effects)

이 모형은 여론조사에서 후보자의 포함 여부(메뉴)에 따라 조사결과에 나타나는 일정한 효과를 통제한다. 여론조사에서 어떤 후보들이 선택지로 제시되는지에 따라 응답 결과가 체계적으로 달라지는 현상을 메뉴 효과라고 한다. 예를 들어, 대통령 후보군 조사에서 A, B, C 후보가 D후보와 함께 조사된 경우와 그렇지 않고 삼자로만 조사된 경우 D 후보의 유무로 인한 응답 분포의 차이가 발생한다. 특히 D후보가 포함되었을 때 다른 후보들의 지지율이 일관되게 감소하거나, 특정 후보의 지지율만 크게 영향을 받는 경우가 있다. 이러한 메뉴 효과는 여론조사 결과를 해석하고 비교할 때 중요한 고려사항이 된다. 서로 다른 후보 구성으로 진행된 여론조사 결과를 직접 비교하면 왜곡된 결론에 도달할 수 있기 때문이다. 아래의 모형은 후보의 부재가 다른 후보들의 지지율에 미치는 영향을 체계적으로 반영한다:

$$
\begin{aligned}
\alpha_{j1} &\sim \text{Uniform}(a_{01}, a_{02}) \\
\alpha_{jt} &= \alpha_{j,t-1} + \varepsilon_{jt}, \quad \varepsilon_{jt} \sim \mathcal{N}(0, \sigma_\alpha^2) \\
y_{ijt} &= \text{house}_{ijt} + \alpha_{jt[i]} + \mathbf{D}_{ijt}\boldsymbol{\beta}_j + \epsilon_{ijt}, \quad \epsilon_{ijt} \sim \mathcal{N}(0, \sigma_{jt}^2) \\
\sigma_{jt}^2 &= N_{jt}^{-1} \sum_i \frac{y_{ijt}(1 - y_{ijt})}{s_{ijt}} \\
\text{house}_{ij,1} &\sim \mathcal{N}(0, \sigma_{\text{initial}}^2) \\
\text{house}_{ij,t} &= \text{house}_{ij,t-1} + \eta_{ij,t}, \quad \eta_{ij,t} \sim \mathcal{N}(0, \sigma_{\text{house}}^2) \\
\boldsymbol{\beta}_j &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \text{Uniform}(\sigma_{low}, \sigma_{high}) \\
\sigma_{\text{house}} &\sim \text{Uniform}(0, \sigma_{\text{house max}}) \\
\sigma_{\text{initial}} &\sim \text{Uniform}(0, \sigma_{\text{initial max}})
\end{aligned}
$$

주요 기호 설명:

- $\mathbf{D}_{ijt}$: 누락된 후보를 나타내는 더미 변수 벡터
- $\boldsymbol{\beta}_j$: 다른 후보의 부재가 후보 j 지지율에 미치는 영향 계수

5 관심 있는 사후분포

사후분포는 베이지안 깁스 샘플링 기법을 통해 추정된다. 메타분석은 다음과 같은 사후분포들을 추정하는 것이 주요 목적이다:

- $p(\boldsymbol{\alpha}_j|\mathbf{y}_j, \mathbf{D}_j)$: 후보 j의 추세에 대한 사후분포
- $p(\alpha_{j,t+1}|\mathbf{y}_j, \mathbf{D}_j)$: 다음 시점 후보 j의 추세에 대한 예측분포
- $p(\alpha_{j,t+1}|\mathbf{y}_j, \mathbf{D}_j) - p(\alpha_{k,t+1}|\mathbf{y}_j, \mathbf{D}_j) > 0$: 다음 시점 후보 j가 후보 k보다 더 많이 득표할 가능성에 대한 예측분포
- $p(\mathbf{house}_{ij}|\mathbf{y}_j, \mathbf{D}_j)$: 조사기관 효과 사후분포
- $p(\boldsymbol{\beta}_j|\mathbf{y}_j, \mathbf{D}_j)$: 메뉴 효과 사후분포

아래 그림(Figure 1)은 특정 후보에 대한 조사기관 효과 추정 결과를 보여주고 있다. 시간별 변화를 가진 조사기관 효과 추정치를 평균해서 시각화한 것이다. (+) 값을 보인 조사기관은 해당 후보에 대해 비교적 우호적인 조사결과를, (−) 값을 보인 조사기관은 해당 후보에 대해 비교적 덜 우호적인 조사결과를 보인 것으로 판단할 수 있다. 이를 다른 말로 하면, (+) 값을 보인 조사기관의 조사에 이 후보에 대해 우호적인 의견을 가진 응답자들이 과다표집된 반면, (−) 값을 보인 조사기관의 조사에서는 해당 후보에 대해 비교적 덜 우호적인 조사결과를 보이는 응답자들이 과다표집되었다고 볼 수 있다. 이러한 비교의 기준이 되는 것은 조사기관 고정값이 0에 가까운 "평균적인 조사"이다.

결론

2025 MBC 여론M의 분석결과는 정당지지도와 대통령선거 후보 지지도를 서로 다른 통계 모형으로 추정하고 있다. 정당지지도는 조사기관 효과 변동 모형을 활용하여 분석하며, 대통령선거 후보 지지도는 메뉴 효과 모형과 조사기관 효과 변동 모형을 결합한 복합 모형(Multilevel Dynamic Linear Model with Time-varying House Effects and Menu Effects)을 통해 추정한다.

이러한 통계적 접근법은 다양한 조사기관에 의해 수집되고 급변하는 여론을 반영하는 한국 정치 여론조사 자료의 특성을 고려한 것으로, 현 시점에서 한국 정치 여론조사 분석에 적합한 프레임워크를 제공한다. 조사기관별 특성과 그 시간적 변동성, 그리고 후보자 구성의 차이가 조사결과에 미치는 영향을 체계적으로 통제함으로써, 표면적인 여론조사 결과 이면에 존재하는 실제 여론 동향을 더 정확하게 파악할 수 있다.

이 분석 방법론은 한국의 정치적 맥락과 여론조사 환경에 맞게 조정되었으며, 조사방법, 표본추출, 질문방식 등 다양한 요인에서 발생하는 체계적 편향을 효과적으로 보정한다. 여러 조건에서 수행된 여론조사 결과들을 통합 분석함으로써 정밀한 여론 추세 분석과 예측이 가능해져, 한국 정치 여론의 이해와 선거 예측의 정확도 향상에 기여한다.

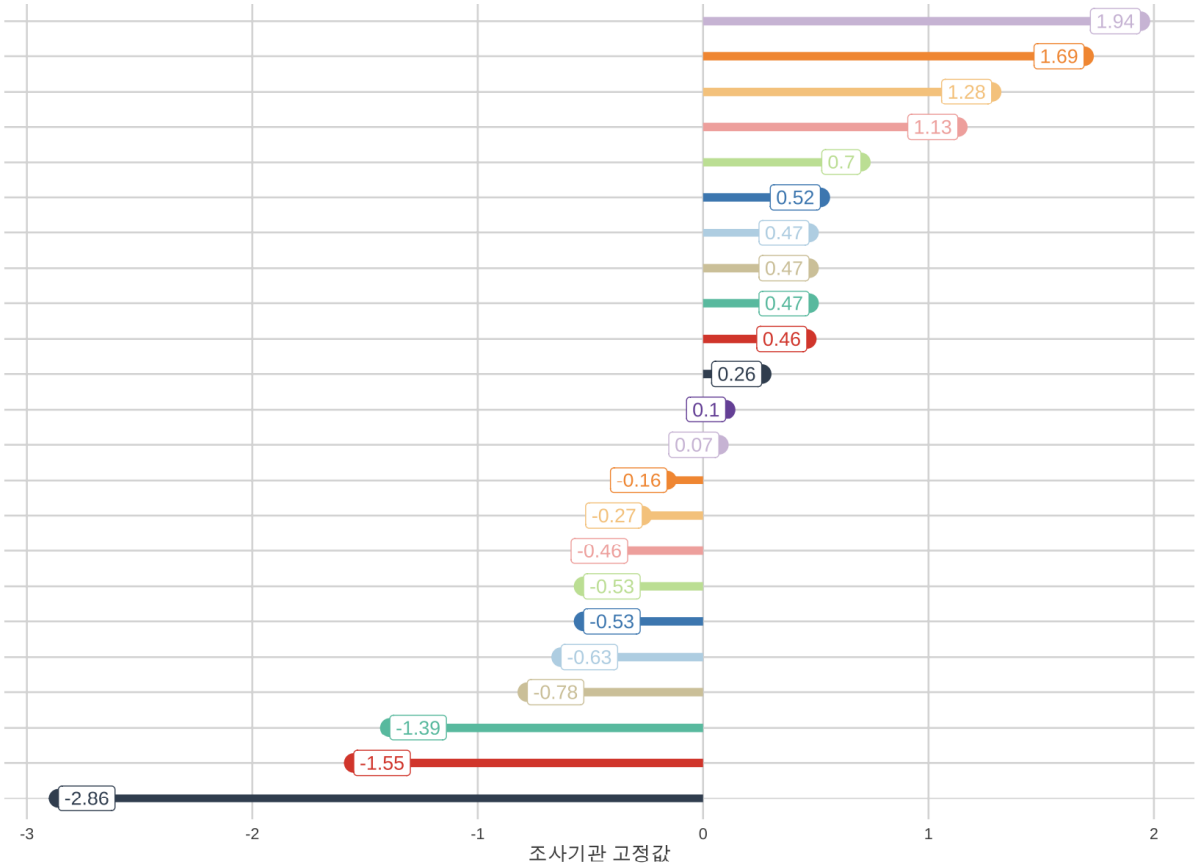

Figure 1: *조사기관 효과의 사후분포 (예시)*