The Korean Journal of Applied Statistics (2020) DOI: https://doi.org/10.5351/KJAS.2020.33.4.439
33(4), 439–449

Pairwise pseudolikelihood approach for adjusting selection bias in meta-analysis

Sunghee Kuk[a] · Woojoo Lee[b,1]

[a]Department of Statistics, Inha University; [b]Department of Public Health Science, Graduate School of Public Health Seoul National University

(Received June 3, 2020; Revised June 19, 2020; Accepted June 26, 2020)

Abstract

Meta-analysis provides a way of integrating several independent studies of interest. Since small studies with statistically significant results are more likely to be published, publication bias, which is a special case of selection bias, often occurs in meta analysis. Conditional likelihood and weighted estimating equation have been proposed to deal with publication bias, but they require to specify a correct selection probability model. In contrast, the pairwise pseudolikelihood approach can correct publication bias without fully specifying the correct selection probability model, but its performance in meta-analysis was not investigated. In this paper, we perform a numerical study about whether the pairwise pseudolikelihood approach is effective for solving publication bias arising from typical meta-analysis settings.

Keywords: meta-analysis, pairwise pseudolikelihood, publication bias, selection bias

1. 서론

메타 분석은 사회 과학, 의학 그리고 경제학 등 많은 연구 분야에서 합리적인 의사결정 및 판단을 위해 다양한 연구 결과들을 통합시켜주는 분석 방법들 중 하나이다 (Egger 등, 1997a). 체계적이고 계량적인 분석을 하기위해서는 메타 분석에서 사용 될 연구 결과들이 담겨있는 문헌들에 대한 체계적인 검토가 선행되어야 한다. 이 때 수집되는 연구 문헌들은 소규모 연구인 경우 통계적으로 유의한 결과를 보이는 연구가 출간될 확률이 높기 때문에, 출판 편향(publication bias)이 발생할 수 있다 (Stern과 Simes, 1997; Egger 등, 1997b). 출판 편향의 문제를 고려하지 않는 경우, 관심있어하는 모수의 추정량은 무시하기 어려운 편향을 갖게 된다. 그렇기 때문에 메타 분석에서 출판 편향은 중요한 주제로 인식되고 있으며 출판 편향을 탐지하고 보정하기 위한 다양한 방법들이 검토되고 있다 (Sutton 등, 2000; Thornton과 Lee, 2000; Egger 등, 1997; Copas와 Shi, 2000).

통계적으로 유의한 연구 결과의 문헌이 수집될 확률이 더 높기 때문에 발생되는 출판 편향은 선택 편향(selection bias)의 특수한 경우로 간주될 수 있다. 여기서 선택 편향이란 연구자가 분석하려는 자료

This work was supported by a Grant from the Next-Generation BioGreen 21 program (Project N0. PJ01337701), Rural Development Administration, Republic of Korea.
[1]Corresponding author: Department of Public Health Science, Graduate School of Public Health Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea. E-mail: lwj221@gmail.com

가 모집단에서 랜덤 표본으로 얻어지는 것이 아닌 자료 자체의 특정 요인에 의존하여 수집될 때 발생되는 편향을 의미한다. 선택 편향을 보정하는 분석 방법으로는 조건부 가능도(conditional likelihood)를 이용한 방법과 관측 자료들이 수집되는 선택 확률을 가중치로 사용하는 가중 추정 방정식(weighted estimating equation)이 있다 (Robins 등, 1994, 1995). 그러나 이 방법들은 모두 관측 자료에 대한 선택 확률 모형을 사용하고 있기 때문에 관측 자료에 대한 실제 선택 확률 모형을 정확히 알지 못하는 일반적인 상황에서 활용하기 쉽지 않다. 이 문제에 대한 대안으로 본 연구에서는 Liang과 Qin (2000)의 쌍별 유사가능도 접근법(pairwise pseudolikelihood approach)을 메타분석에서 고려해 보고자한다. Liang과 Qin (2000)이 제안한 쌍별 유사가능도 접근법은 조건부 가능도나 가중 추정 방정식과는 달리 선택 확률 모형을 완전히 결정하지 않으면서도 관심있는 모수의 추정이 가능하다. 그러나 Liang과 Qin (2000)의 연구에서는 메타분석에서 쌍별 유사가능도 접근법이 고려가 가능함을 언급하였으나, 실제 관련 연구는 진행되지 않았다. 이에 따라, 본 논문에서는 메타 분석에서 쌍별 유사가능도 접근법이 출판 편향을 보정해 주는 것에 대해 확인하고, 다양한 선택 확률 모형과 표본의 크기에 따른 쌍별 유사가능도 접근법의 성능과 문제점을 수치적으로 연구한다.

본 논문의 순서는 다음과 같다. 2절에서는 선택 편향과 분석 방법론들에 대해 소개하고 3절에서는 쌍별 유사가능도 접근법을 메타 분석에 직접 적용한 내용을 다룬다. 그리고 4절에서는 이를 위한 모의실험의 설계 및 결과에 대해 제시하며, 마지막으로 5절에서는 결론에 대해 논의할 것이다.

2. 선택 편향 문제와 분석 방법

선택 편향된 자료의 추정값을 보정하기 위해 활용되는 분석 방법은 대부분 분석대상이 되기위해 관측자료가 선택되어지는 확률을 요구한다. 그러나 정확한 선택 확률 모형이 주어지는 경우는 매우 희귀하기 때문에 선택 편향을 보정하는 문제는 쉽지 않다. 이를 해결하는 방안으로 Liang과 Qin (2000)은 선택 확률 모형을 완전히 알지는 못 하더라도 편향을 보정할 수 있는 방법인 쌍별 유사가능도 접근법을 제안하였다. 이 절에서는 선택 확률 모형을 가정한 기존의 분석 방법들과 쌍별 유사가능도 방법을 검토하고, 쌍별 유사가능도 방법의 효율성(efficiency)을 올리기 위한 새로운 변형을 소개한다.

2.1. 조건부 가능도

먼저 $f(y \mid x;\beta)$는 x가 주어졌을 때 y의 조건부 밀도함수를 나타내고, 연관된 모수로 β가 있음을 의미한다. δ_i는 i번째 자료가 선택되어 분석대상에 포함되는 경우는 1, 그렇지 않은 경우는 0의 값을 갖는 변수이다. 이 때, 주어진 n개의 자료에 대한 조건부 가능도 함수는 다음과 같다.

$$\begin{aligned}\prod_{i=1}^{n} f(y_i \mid x_i, \delta_i = 1) &= \prod_{i=1}^{n} \frac{\Pr(\delta_i = 1 \mid x_i, y_i) f(y_i \mid x_i;\beta)}{\Pr(\delta_i = 1 \mid x_i)} \\ &= \prod_{i=1}^{n} \frac{\Pr(\delta_i = 1 \mid x_i, y_i) f(y_i \mid x_i;\beta)}{\int \Pr(\delta_i = 1 \mid x_i, y) f(y \mid x_i;\beta) dy}, \end{aligned} \tag{2.1}$$

여기서 $\Pr(\delta_i = 1 \mid x_i)$은 설명 변수 x의 i번째 값이 주어져 있을 때의 관측 자료가 수집될 확률이며 $\Pr(\delta_i = 1 \mid x_i, y_i)$는 설명 변수 x와 반응 변수 y의 i번째 값이 주어졌을 때의 관측 자료의 수집 확률이다. 그런데 만일 관측 자료가 수집되는 확률이 반응 변수 y에 의존하지 않고 모두 동일하다면 $\Pr(\delta_i = 1 \mid x_i, y_i) = \Pr(\delta_i = 1 \mid x_i)$이 되어 제거되기 때문에 선택 확률의 정보를 고려하지 않더라도 β에 대해 올바른 추정을 할 수 있다. 그러나 관측 자료가 y에 따라 다른 확률로 수집되는 경우는 $\Pr(\delta_i = 1 \mid x_i, y_i) \neq \Pr(\delta_i = 1 \mid x_i)$이 되므로 β에 대한 올바른 추정을 하기 위해서는 관측자료 마다의

정확한 $\Pr(\delta_i = 1 \mid x_i, y_i)$이 요구된다. 이 방법은 쌍별 유사가능도 접근법과 비교를 위해 사용할 것이며 본 논문에서는 조건부 가능도 추정법이라고 언급하도록 하겠다.

2.2. 가중 추정 방정식

조건부 가능도 추정법 외에 선택 편향을 보정하기 위한 또 다른 방법에는 가중 추정 방정식이 있다 (Robins 등, 1994, 1995). 가중 추정 방정식은 다음과 같이 표현된다.

$$S(\beta) = \sum_{i=1}^{n} \left\{ \frac{1}{\pi_i} \frac{\partial[log\{f(y_i \mid x_i; \beta)\}]}{\partial \beta} \right\} = 0, \tag{2.2}$$

여기서 π_i는 $\Pr(\delta_i = 1 \mid x_i, y_i)$를 의미한다. 가중 추정 방정식은 정확한 $\Pr(\delta_i = 1 \mid x_i, y_i)$을 가정하는 경우 β에 대해 일치추정량을 얻는다는 장점을 갖는다. 그러나 잘못된 $\Pr(\delta_i = 1 \mid x_i, y_i)$이 가정될 때 계산되는 추정량은 일치추정량이라고 할 수 없기 때문에 조건부 가능도 추정법처럼 관측 자료 마다의 정확한 $\Pr(\delta_i = 1 \mid x_i, y_i)$이 요구된다. 이 방법 역시 쌍별 유사가능도 접근법과 비교를 위해 사용될 것이며 본 논문에서는 가중 추정 방정식이라고 언급하도록 하겠다.

2.3. 쌍별 유사가능도 접근법

Liang과 Qin (2000)은 선택 확률 $\Pr(\delta_i = 1 \mid x_i, y_i)$을 사용하지 않더라도 β를 추정하는 방법으로 쌍별 유사가능도 접근법을 제안하였다. 제안된 쌍별 유사가능도 함수는 다음과 같다.

$$L_P(\beta) = \prod_{i<k} \frac{f(y_i \mid x_i; \beta) f(y_k \mid x_k; \beta)}{f(y_i \mid x_i; \beta) f(y_k \mid x_k; \beta) + f(y_i \mid x_k; \beta) f(y_k \mid x_i; \beta)}. \tag{2.3}$$

위의 식의 분모와 분자를 모두 분자로 나누어 주면,

$$L_P(\beta) = \prod_{i<k} \frac{1}{1 + R(y_i, x_i; y_k, x_k)}$$

이 되고, 여기서

$$R(y_i, x_i; y_k, x_k) = \frac{f(y_i \mid x_k) f(y_k \mid x_i)}{f(y_i \mid x_i) f(y_k \mid x_k)}$$

으로 주어진다. Liang과 Qin (2000)은 $R(y_i, x_i; y_k, x_k)$을 일반화 오즈비라고 소개하였으며, $L_P(\beta)$를 최대화 하는 문제는 결국 일반화 오즈비에 의해 결정됨을 이야기하였다. Liang과 Qin (2000)은 $\Pr(\delta_i = 1 \mid x_i, y_i) = \Pr(\delta_i = 1 \mid y_i)$이면, $L_p(\beta)$를 최대화하는 추정량은 일치성을 갖는다는 것을 보였다. 또한

$$f(y \mid x; \beta) = \exp\{\frac{\theta(x^T\beta)y - b(x^T\beta)}{a(\phi)} + c(y, \phi)\}, \tag{2.4}$$

인 경우, 즉 Nelder과 Wedderburn (1972)의 일반화 선형 모형의 지수족이 가정된 경우,

$$R(y_i, x_i; y_k, x_k) = \exp\left[\frac{(y_i - y_k)(\theta(x_k^T\beta) - \theta(x_i^T\beta))}{a(\phi)}\right] \tag{2.5}$$

으로 정리된다. 식 (2.4)에서 $\theta(x^T\beta)$는 자연모수(natural parameter)를 의미하며, $a(\phi)$는 산포 모수(dispersion parameter)에 대한 함수를 의미한다. 예를들어 정규분포에서 항등연결함수가 사용된 경우에는 $\theta(x^T\beta) = x^T\beta$, $a(\phi) = \sigma^2$이 된다. Liang과 Qin (2000)에 따르면 이 쌍별 유사가능도 접근법은 데이터의 일부 정보만을 사용하기 때문에 분산이 증가하므로 실제 사용에서의 걸림돌이 될 수 있음을 지적하였다.

2.4. 확장된 유사가능도 접근법

2.3절에서는 Liang과 Qin (2000)은 두 개의 조합들로 이루어진 쌍별 유사가능도이기 때문에, 이 절에서는 더 많은 정보를 활용하고자 하는 목적으로 세 개의 조합들과, 네 개의 조합들로 확장시킨 유사가능도 접근법을 새롭게 고려해보고자 한다. 특히 이 후 시뮬레이션 연구에서 이 방법이 쌍별 유사가능도에서 얻어진 추정량의 분산을 얼마나 감소시킬 수 있는지를 살펴보고자 한다. 먼저, 세 개의 조합들을 사용한 유사가능도 접근법은 다음과 같이 정의될 수 있다.

$$L_{P2}(\beta) = \prod_{i<j<k} \frac{f(y_i|x_i;\beta)f(y_j|x_j;\beta)f(y_k|x_k;\beta)}{\left[\begin{array}{l} f(y_i|x_i;\beta)f(y_j|x_j;\beta)f(y_k|x_k;\beta) + f(y_i|x_i;\beta)f(y_j|x_k;\beta)f(y_k|x_j;\beta) \\ +f(y_i|x_j;\beta)f(y_j|x_i;\beta)f(y_k|x_k;\beta) + f(y_i|x_j;\beta)f(y_j|x_k;\beta)f(y_k|x_i;\beta) \\ +f(y_i|x_k;\beta)f(y_j|x_i;\beta)f(y_k|x_j;\beta) + f(y_i|x_k;\beta)f(y_j|x_j;\beta)f(y_k|x_i;\beta) \end{array}\right]}. \quad (2.6)$$

위에 정의한 유사가능도 접근법은 $f(y_i \mid x_i;\beta)f(y_j \mid x_j;\beta)f(y_k \mid x_k;\beta)$를 분모 분자에 각각 나눠준 경우 아래와 같이 정의 될 수 있다.

$$L_{P2}(\beta) = \prod_{i<j<k} \frac{1}{\left[\begin{array}{l} 1 + R(y_j,x_j;y_k,x_k) + R(y_i,x_i;y_j,x_j) + R(y_i,x_i;y_k,x_k) \\ +W1(y_i,x_i;y_j,x_j;y_k,x_k) + W2(y_i,x_i;y_j,x_j;y_k,x_k) \end{array}\right]}, \quad (2.7)$$

여기서 $R(y_j,x_j;y_k,x_k)$, $R(y_i,x_i;y_j,x_j)$과 $R(y_i,x_i;y_k,x_k)$은 식 (2.5)의 정의를 따르며, $W1(y_i,x_i;y_j,x_j;y_k,x_k)$과 $W2(y_i,x_i;y_j,x_j;y_k,x_k)$은 아래와 같이 정의된다.

$$W1(y_i,x_i;y_j,x_j;y_k,x_k) = \exp\left[\frac{\left[y_i\left(x_j^T - x_i^T\right) + y_j\left(x_k^T - x_j^T\right) + y_k\left(x_i^T - x_k^T\right)\right]\beta}{a(\phi)}\right],$$

$$W2(y_i,x_i;y_j,x_j;y_k,x_k) = \exp\left[\frac{\left[y_i\left(x_k^T - x_i^T\right) + y_j\left(x_i^T - x_j^T\right) + y_k\left(x_j^T - x_k^T\right)\right]\beta}{a(\phi)}\right].$$

네 개의 조합들을 사용한 유사가능도 접근법도 위에서 살펴본 것과 동일한 방법으로 정의가 가능하다. 만약 네 개의 조합들을 사용하는 경우 식 (2.6)의 분모에는 총 24개의 조합이 나열될 수 있다. 조합의 수를 확장시킨 유사가능도 접근법은 더 많은 정보를 활용하기 때문에 분산을 줄어들 수 있을 것으로 기대할 수 있으나 추정량을 계산하기 위해 계산량이 빠르게 증가함을 알 수 있다. 본 논문에서는 분산이 감소되는 경향을 확인하기 위해 위의 확장된 유사가능도 접근법을 적용할 것이며 두 개의 조합들을 사용하는($l=2$) 경우를 쌍별 유사가능도 접근법, 세 개의 조합들을 사용하는($l=3$) 경우를 삼중 유사가능도 접근법, 네 개의 조합들을 사용하는($l=4$) 경우를 사중 유사가능도 접근법이라고 언급하도록 하겠다.

3. 유사가능도를 이용한 메타 분석

3.1. 메타 분석

편의상, y_i와 σ_i를 i번째 연구에서의 치료 효과(treatment effect)와 표준 오차(standard error)라 하자. 수집된 n개의 문헌들에 대한 메타 분석은 다음의 고정 효과 모형을 가정한다.

$$y_i = b + \epsilon_i \quad (i = 1,\ldots,n), \quad (3.1)$$

이때 b는 치료 효과에 대한 전체 평균(global mean)을 의미하며, ϵ_i은 $N(0,\sigma_i^2)$을 따른다. 메타 분석에서 σ_i^2은 알려진 값으로 사용된다. 그리고 메타 분석에 사용되는 문헌의 수 n은 일반적으로 수 십 정도

이다. 메타 분석에서 치료 효과에 대한 전체 평균 b를 추정하는 대표적인 방법으로 다음과 같은 가중 평균 방법이 제안되어왔다 (Cochran, 1954).

$$\hat{b} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}$$

이 때, w_i는 i번째 연구에 대한 가중치을 의미한다. 이 가중치를 무엇으로 선정하는 것이 가장 이상적인지에 대해 많은 연구들이 수행되어 왔으나 그 중에서도 연구들의 정밀한 정도에 따라 가중치를 선정하는 방법인 역 분산 가중 평균(inverse variance-weighted average) 방법이 대표적으로 사용된다 (Cochran, 1954). 그러므로 본 논문에서도 분산의 역수($1/\sigma_i^2$)를 가중치로 사용하는 역 분산 가중 평균을 사용할 것이다.

3.2. 쌍별 유사가능도를 이용한 메타 분석

메타 분석에서 발생하는 선택 편향은 분석에 사용하는 자료가 유의한 검정 결과를 보이는 소규모 연구일 때, 더 자주 수집되기 때문에 발생한다. 이로 인해 Hedges (1992)와 Dear와 Begg (1992)은 검정통계량(y_i/σ_i)에 대한 가중 함수를 이용한 보정 방법을 제시하기도 하였다. 다시 말해서, $z_i = y_i/\sigma_i$라고 한다면 선택 확률은 $\Pr(\delta_i = 1|z_i)$이라고 할 수 있기 때문에 선택 편향을 고려한 식 (3.1)의 확률 밀도 함수는 다음과 같이 정리 될 수 있다.

$$\begin{aligned} f(z_i \mid \delta_i = 1) &= \frac{\Pr(\delta_i = 1 \mid z_i) f(z_i; b)}{\Pr(\delta_i = 1)} \\ &= \frac{\Pr(\delta_i = 1 \mid z_i)}{\Pr(\delta_i = 1)} \frac{1}{\sqrt{(2\pi)}} \exp\left\{ -\frac{(z_i - b/\sigma_i)^2}{2} \right\}. \end{aligned} \tag{3.2}$$

정리된 식 (3.2)에서 $x_i = 1/\sigma_i$, $\beta = b$로 고려하여 2.3절에서 설명한 쌍별 유사가능도 접근법에 적용 시키는 경우 $\Pr(\delta_i = 1 \mid z_i)/\Pr(\delta_i = 1)$는 서로 상쇄되어 사라지기 때문에 선택 확률 모형을 지정하지 않더라도 b를 추정할 수 있다. b를 추정하기 위한 일반화 오즈비는

$$R\left(z_i, \frac{1}{\sigma_i}; z_k, \frac{1}{\sigma_k}\right) = \exp\left[\frac{b(z_i - z_k)(\sigma_i - \sigma_k)}{(\sigma_i \sigma_k)}\right] \tag{3.3}$$

과 같이 정의될 수 있다.

3.3. 삼중과 사중 유사가능도를 이용한 메타 분석

메타 분석에 쌍별 유사가능도 접근법을 확장한 삼중 유사가능도 접근법과 사중 유사가능도 접근법을 적용하였을 때 식 (3.2)에 대한 $L_{P2}(b)$의 식은 다음과 같이 정리된다.

$$L_{P2}(b) = \prod_{i<j<k} \frac{1}{\left[\begin{array}{l} 1 + R\left(z_j, \frac{1}{\sigma_j}; z_k, \frac{1}{\sigma_k}\right) + R\left(z_i, \frac{1}{\sigma_i}; z_j, \frac{1}{\sigma_j}\right) + R\left(z_i, \frac{1}{\sigma_i}; z_k, \frac{1}{\sigma_k}\right) \\ +W1\left(z_i, \frac{1}{\sigma_i}; z_j, \frac{1}{\sigma_j}; z_k, \frac{1}{\sigma_k}\right) + W2\left(z_i, \frac{1}{\sigma_i}; z_j, \frac{1}{\sigma_j}; z_k, \frac{1}{\sigma_k}\right) \end{array}\right]}, \tag{3.4}$$

여기서 $R(z_j, 1/\sigma_j; z_k, 1/\sigma_k)$, $R(z_i, 1/\sigma_i; z_j, 1/\sigma_j)$, 그리고 $R(z_i, 1/\sigma_i; z_k, 1/\sigma_k)$는 식 (3.3)의 정의를 따르며, $W1(z_i, 1/\sigma_i; z_j, 1/\sigma_j; z_k, 1/\sigma_k)$과 $W2(z_i, 1/\sigma_i; z_j, 1/\sigma_j; z_k, 1/\sigma_k)$은 $W1$, $W2$ 정의에 따라 메타

Table 4.1. Selection probability model

경우	올바른 선택 확률 모형 사용 여부	실제 선택 확률 모형	적합시 사용된 선택 확률 모형
1	O	$\exp(0.5\|z\|)/\{1+\exp(0.5\|z\|)\}$	$\exp(0.5\|z\|)/\{1+\exp(0.5\|z\|)\}$
2	O	$\exp(2\|z\|)/\{1+\exp(2\|z\|)\}$	$\exp(2\|z\|)/\{1+\exp(2\|z\|)\}$
3	O	$\exp(0.5z)/\{1+\exp(0.5z)\}$	$\exp(0.5z)/\{1+\exp(0.5z)\}$
4	O	$\exp(2z)/\{1+\exp(2z)\}$	$\exp(2z)/\{1+\exp(2z)\}$
5	X	$\exp(2z)/\{1+\exp(2z)\}$	$\exp(0.5z)/\{1+\exp(0.5z)\}$
6	X	$\exp(0.5z)/\{1+\exp(0.5z)\}$	$\exp(2z)/\{1+\exp(2z)\}$
7	X	$\exp(2z+z^2)/\{1+\exp(2z+z^2)\}$	$\exp(2z)/\{1+\exp(2z)\}$
8	X	$\exp(-0.5\|z\|)$	$\exp(-1.5\|z\|)$

분석에서 다음과 같이 계산된다.

$$W1\left(z_i,\frac{1}{\sigma_i};z_j,\frac{1}{\sigma_j};z_k,\frac{1}{\sigma_k}\right)=\exp\left[b\left\{\frac{z_i(\sigma_i-\sigma_j)}{\sigma_i\sigma_j}+\frac{z_j(\sigma_j-\sigma_k)}{\sigma_j\sigma_k}+\frac{z_k(\sigma_k-\sigma_i)}{\sigma_k\sigma_i}\right\}\right],$$
$$W2\left(z_i,\frac{1}{\sigma_i};z_j,\frac{1}{\sigma_j};z_k,\frac{1}{\sigma_k}\right)=\exp\left[b\left\{\frac{z_i(\sigma_i-\sigma_k)}{\sigma_i\sigma_k}+\frac{z_j(\sigma_j-\sigma_i)}{\sigma_j\sigma_i}+\frac{z_k(\sigma_k-\sigma_j)}{\sigma_k\sigma_j}\right\}\right].$$

사중 유사가능도 접근법도 위와 동일한 방식으로 정의되어질 수 있으며 본 논문의 모의실험에서는 삼중 유사가능도와 사중 유사가능도 접근법을 모두 고려하여 쌍별 유사가능도의 성능과 비교할 것이다.

4. 모의실험

4.1. 모의실험 설계

본 논문에서의 모의실험은 크게 두 가지의 목적을 갖는다. 첫 번째는 메타 분석에서 쌍별 유사가능도 접근법의 성능을 확인하기 위함이며, 두 번째는 $l=2$인 경우의 쌍별 유사가능도 접근법에서 $l=3$과 $l=4$에 해당하는 확장된 유사가능도 접근법을 사용하였을 때의 분산 감소의 정도를 파악하기 위함이다.

첫 번째 목적을 위한 모의실험에서는 다양한 선택 확률 모형과 표본의 크기($n=10,20,50,100,200,500$)에서 조건부 가능도 추정법, 역분산 가중평균법, 가중 추정 방정식, 쌍별 유사가능도 접근법에서 얻어진 b의 추정치의 정확도를 비교한다. 먼저 균일 모집단 $U(1,3)$으로부터 1,000개의 분산(σ_i^2)을 생성하고, 치료 효과(y_i)는 정규분포 $N(1,\sigma_i^2)$에서 생성하였다. 그리고 Table 4.1에서의 선택 확률 모형에 따라 베르누이 변수를 생성하여 n개의 표본을 선택하고 b를 추정한다. 이와 같은 작업을 100번 반복하는 모의실험을 진행하였다. Table 4.1에서의 z는 3.2절에서 언급한 y/σ를 의미한다.

Table 4.1에서는 실제 데이터를 생성할 때 사용한 선택 확률 모형과 분석에서 모형 적합에서 사용된 선택 확률 모형을 제시하고 있는데, 1번 경우부터 4번 경우까지는 선택 확률 모형을 정확히 아는 상황으로 설계하였고, 5번 경우부터 8번 경우까지는 잘못된 선택 확률 모형을 가정하는 상황으로 설계하였다. 조건부 가능도 추정법, 가중 추정 방정식에서는 적합시 사용된 선택확률 모형에 따라 가능도 함수의 형태가 바뀌어지는 반면에 역분산 가중 평균법과 쌍별 유사가능도 방법에서는 선택확률 모형이 직접적으로는 사용되지 않음에 주목해야 한다.

두 번째 목적의 모의실험 연구에서는 쌍별 유사가능도 접근법을 확장하여 세 개의 조합, 네 개의 조합에 기반한 확장된 유사 가능도 방법(삼중 유사가능도, 사중 유사가능도 접근법)이 실질적인 추정치의 분산 감소 효과를 가져오는지 여부를 확인한다. 이를 위해, 확장 정도($l=2,3,4$)와 다양한 표본의 크

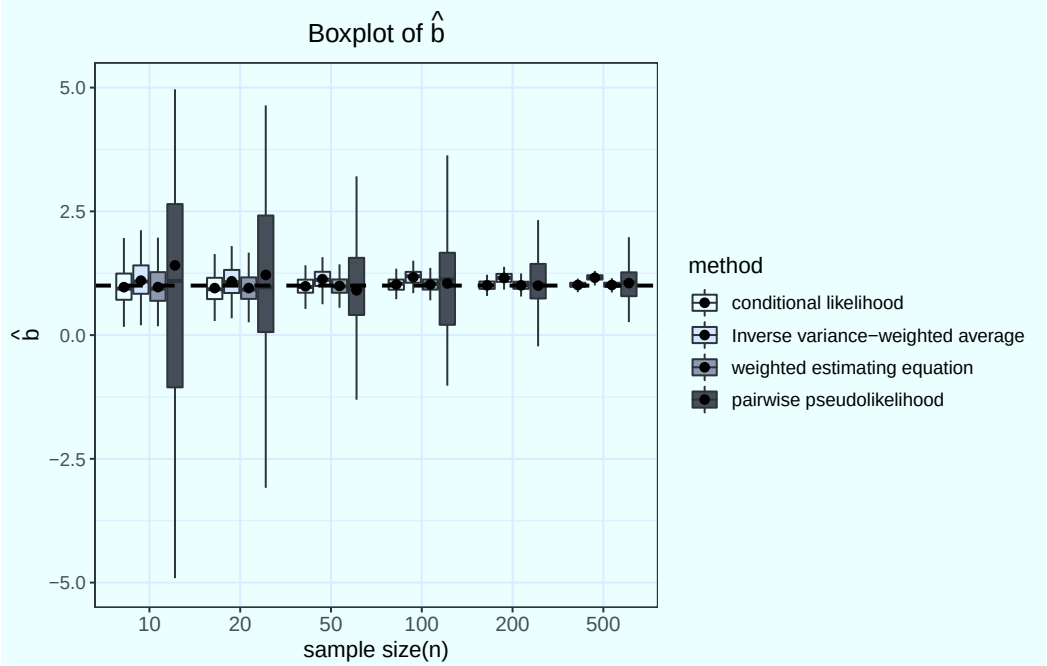

Figure 4.1. Boxplots of $\hat{b}$ from the first simulation result: the true selection probability model and the assumed selection probability model are same as $\exp(2|z|)/\{1+\exp(2|z|)\}$.

기($n = 10, 20, 50, 100, 200$)에 따라 첫 번째 모의실험에서 사용된 선택 확률 모형을 고려하여 수치연구를 진행하였다.

4.2. 모의실험 결과

첫 번째 모의실험의 결과에서 Figure 4.1은 실제 선택 확률 모형과, 모형 적합에 사용된 선택 확률 모형이 일치하는 2번 경우에 대한 결과를 나타내고 있다. 이 그림은 표본의 수를 점차 증가시키며 치료 효과들의 전체 평균(b)에 대해 네 가지 분석 방법별로 추정한 결과를 상자그림으로 나타냈다. 이 그림을 보면 역분산 가중 평균의 경우 지속적으로 편향이 발생하고 있으며 표본의 크기가 500일 때 편향의 크기가 0.1598(표준오차 0.0604) 정도가 되는 것으로 나타났다. 반면, 조건부 가능도 추정법과 가중 추정 방정식 방법이 편향과 분산의 측면에서 우수한 결과를 보이는것을 알 수 있다. 그러나 이는 선택 확률 모형을 알고있다는 가정 하에서 얻어진 결과로 실제 상황에서는 얻을 수 없는 비교를 위한 이상적인 결과로써 이해되어야 한다. 반면, 쌍별 유사가능도 접근법의 경우 표본의 수가 적을 때 상대적으로 분산이 매우 크게 나타났다. 따라서 쌍별 유사가능도 접근법은 소표본에서의 메타 분석에서 사용하는 것은 매우 주의를 필요로 해 보인다. 그러나 실제 선택 확률 모형과 모형 적합시 사용된 선택 확률 모형이 일치하지 않는 8번 경우에 대한 결과를 나타낸 Figure 4.2를 보면, 앞선 결과와 마찬가지로 역분산 가중평균의 경우 여전히 편향이 발생하고 있으며, 표본이 500인 경우 편향의 크기는 -0.3013(표준오차 0.0539) 정도로 나타났다. 또한, 쌍별 유사가능도 접근법은 여전히 표본의 수가 매우 적은 상황에서 다른 방법들에 비해 매우 큰 분산을 보이고 있으나 편향의 정도는 상대적으로 작고, 표본의 수가 증가할 수록 분산이 점차 줄어들기 때문에 표본의 수가 큰 상황에서는 조건부 가능도 추정법을 포함한 나머지 방법에 비

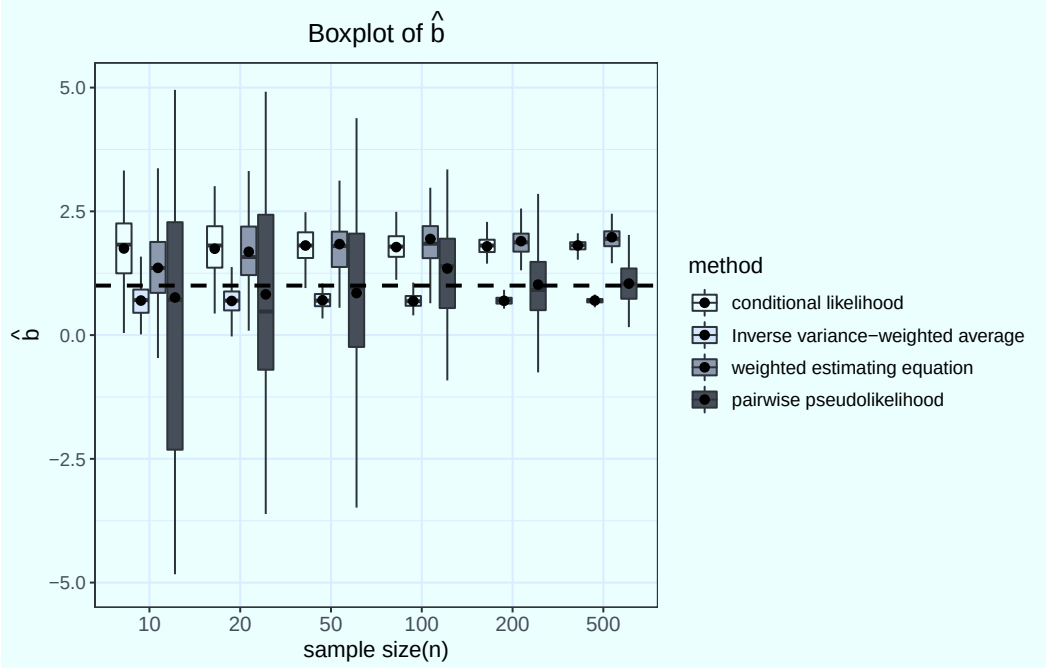

Figure 4.2. Boxplots of $\hat{b}$ from the first simulation result: the true selection probability model is $\exp(-0.5|z|)$ and the assumed selection probability model is $\exp(-1.5|z|)$.

해 월등히 좋은 성능을 보이는 것을 알 수 있다. 이처럼 선택 확률 모형이 잘못 지정된 경우 쌍별 유사가능도 접근법을 제외한 나머지 방법들은 상대적으로 분산은 작을 수 있으나 큰 편향을 보이고 있기 때문에 선택 확률 모형을 알지 못하는 일반적인 상황에서 사용하기에는 바람직하지 못하다는 것을 알 수 있다. 모의실험의 설계에서 언급된 나머지 선택 확률 모형의 경우에 대해서도 같은 결론을 보였기 때문에 본 논문에서 그림은 생략하였다.

두 번째 모의 실험의 결과인 Figure 4.3은 쌍별 유사가능도 접근법($l = 2$), 삼중 유사가능도 접근법($l = 3$) 그리고 사중 유사가능도 접근법($l = 4$)을 선택 확률 모형이 일치하는 2번 경우에서 적용한 결과이다. 이 그림을 보면 l값이 증가함에 따라 분산의 크기는 거의 변동이 없다는 것을 확인할 수 있었다. 이러한 결과는 가정한 선택 확률 모형이 일치하지 않는 8번 경우에 대한 결과인 Figure 4.4에서도 동일한 경향을 보였다. 이 모의실험 결과를 볼 때, 분산의 감소 효과는 l이 커짐에 따라 실질적이지 않은 것으로 판단된다.

5. 결론

본 논문에서는 선택 편향의 특수한 경우인 출판 편향의 문제가 빈번히 발생하는 메타 분석에서 쌍별 유사가능도 접근법을 적용하여 이 방법의 성능과 문제점에 대해 확인하고자 하였다. 메타 분석은 체계적인 문헌의 검토 과정이 선행되어야 하는 분석이기 때문에 표본의 수가 10개에서 20개정도의 소표본이 사용되는 경우가 종종 있다. 그래서 이 연구에서는 메타 분석에서의 쌍별 유사가능도 접근법이 표본에 따라 어떠한 성능의 차이를 보이는 지를 파악하기 위해 모의실험을 통해 표본의 수가 매우 적은 경우부터 큰 경우까지 다양한 선택 확률 모형들에 따라 결과를 확인하였다. 그 결과, 선택 확률 모형을 정확

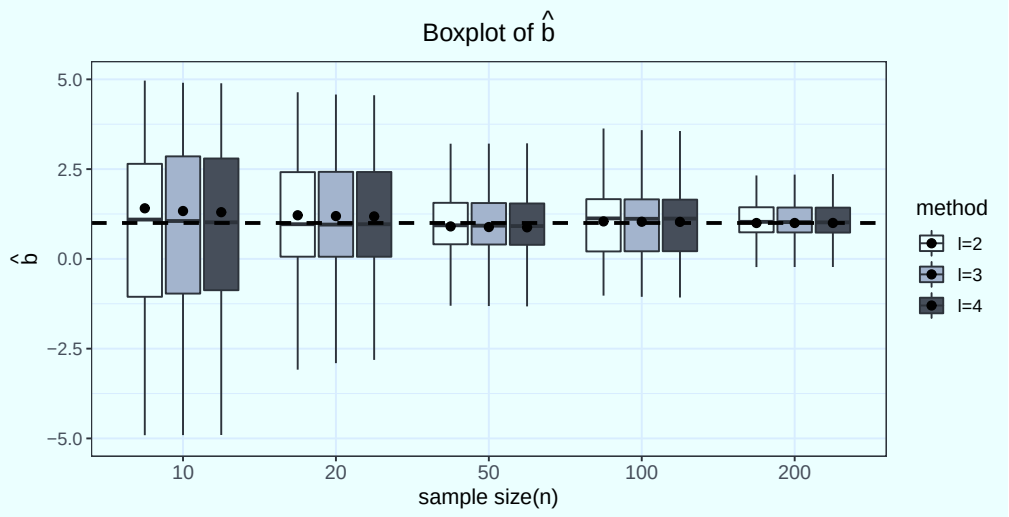

Figure 4.3. Boxplots of $\hat{b}$ from the second simulation result (comparison of different l): the true selection probability model and the assumed selection probability model are same as $\exp(2|z|)/\{1+\exp(2|z|)\}$.

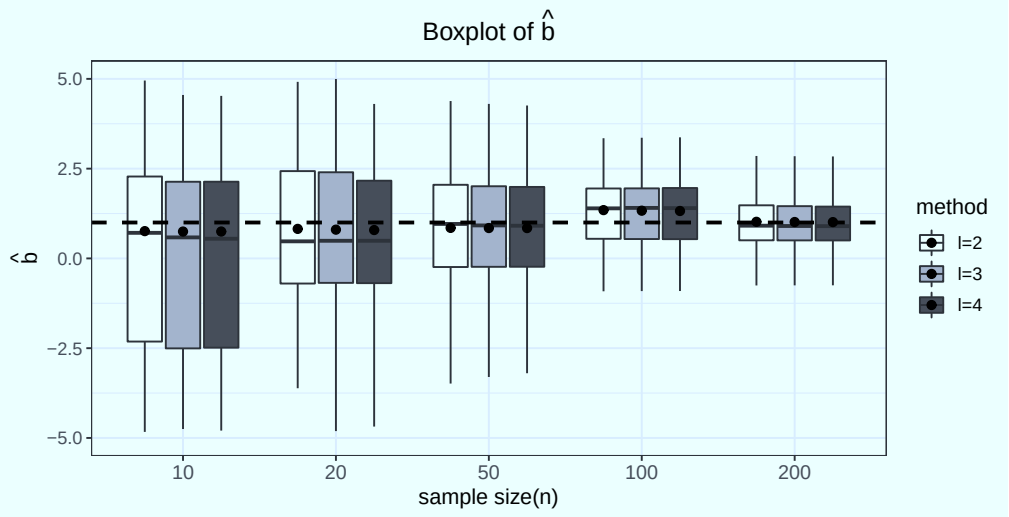

Figure 4.4. Boxplots of $\hat{b}$ from the second simulation result (comparison of different l): the true selection probability model is $\exp(-0.5|z|)$ and the assumed selection probability model is $\exp(-1.5|z|)$.

히 알고 있다고 가정하는 경우에는 조건부 가능도 추정법과 가중 추정 방정식이 가장 좋은 성능을 보이는 것을 알 수 있었으나, 선택 확률 모형이 잘못 지정되어 있는 경우에는 조건부 가능도 추정법과 가중 추정 방정식은 편향의 문제가 심각함을 알 수 있었다. 쌍별 유사가능도 접근법의 경우 표본의 수가 증가함에 따라 성능이 점차 좋아지는 것을 알 수 있었고, 표본의 수를 500개까지 증가시켰을 때는 다른 추정 방법들에 비하여 월등히 좋은 성능을 보이는 것을 알 수 있었다. 또한 쌍별 유사가능도 접근법의 분산을 감소시키기 위해 쌍별 유사가능도 접근법을 확장시킨 삼중 유사가능도 접근법과 사중 유사가능도 접근법을 적용하여 분산의 차이를 확인하였으나 큰 변화를 보이지는 않았다. 이러한 결과들을 종합해 볼 때,

쌍별 유사가능도 접근법을 확장시키는 것은 실질적인 개선의 효과를 기대하기 어려워 보이며, 선택 확률 모형을 모르는 일반적인 상황에서 표본의 수가 충분히 큰 경우에 제한적으로 쌍별 유사가능도 접근법을 사용하는 것이 좋은 선택이 될 수 있을 것으로 보인다. 그러나 표본의 수가 충분히 크지 않은 소표본을 분석하는 상황에서는 쌍별 유사가능도 접근법의 사용은 매우 조심스럽게 고려되어야 하며, 소표본에서의 분산을 감소시키기 위한 별도의 전략을 개발하는 것은 앞으로 소표본에서도 신뢰성 있는 결론을 얻을 수 있는 좋은 연구주제가 될 것으로 생각된다.

References

Cochran, W. G. (1954). The combination of estimates from different experiments, *Biometrics*, **10**, 101–129.

Copas, J. and Shi, J. Q. (2000). Meta-analysis, funnel plot and sensitivity analysis, *Biostatistics*, **1**, 247–262.

Dear, B. G. and Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta analysis, *Statistical Science*, **7**, 237–245.

Egger, M., Smith, G. D., and Phillips, A. N. (1997a). Meta-analysis: principles and procedures, *British Medical Journal*, **315**, 1533–1537.

Egger, M., Zellweger-Zahner, T., Schneider, M., Junker, C., Lengeler, C., and Antes, G. (1997b). Language bias in randomised controlled trials published in English and German, *Lancet*, **350**, 326–329.

Hedges, L. V. (1992). Modeling publication selection effect in meta-analysis, *Statistical Science*, **7**, 246–255.

Liang, K. Y. and Qin, J. (2000). Regression analysis under non-standard situations: a pairwise pseudolikelihood approach, *Journal of Royal Statistical Society. Series B*, **62**, 773–786.

Nelder J. A. and Wedderburn R. W. M. (1972) Generalized linear models, *Journal of the Royal Statistical Society. Series A*, **135**, 370–384.

Robins J. M, Rotmitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed, *Journal of the American Statistical Association*, **89**, 846–866.

Robins, J. M., Rotmitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the American Statistical Association*, **90**, 106–121.

Stern, J. M. and Simes, R. J.(1997). Publication bias: evidence of delayed publication in a cohort study of clinical research projects, *British Medical Journal*, **315**, 640–645.

Sutton, A. J., Song, F., Gilbody, S. M., and Abrams, K. R. (2000). Modelling publication bias in meta-analysis: a review, *Statistical Methods in Medical Research*, **5**, 421–445.

Thornton, A. and Lee, P. (2000). Publication bias in meta-analysisits causes and consequences, *Journal of Clinical Epidemiology*, **53**, 207–216.

메타분석의 선택 편향 보정을 위한 쌍별 유사가능도 접근법

국성희[a] · 이우주[b,1]

[a]인하대학교 통계학과, [b]서울대학교 보건대학원

(2020년 6월 3일 접수, 2020년 6월 19일 수정, 2020년 6월 26일 채택)

요 약

메타 분석은 여러 연구 결과들을 종합시켜주는 분석 방법 중 하나이다. 이 때 수집되는 연구 문헌들은 소규모 연구인 경우 통계적으로 유의한 결과를 보이는 연구가 출간될 확률이 높기 때문에, 선택 편향의 특수한 경우인 출판 편향이 종종 발생한다. 선택 편향을 보정하는 방법에는 조건부 가능도와 가중 추정 방정식이 있는데 이 방법들은 실제 얻기 힘든 정확한 선택 확률 모형을 필요로한다. 반면 쌍별 유사가능도 접근법은 선택 확률 모형을 정확히 알 수 없는 경우에도 선택 편향을 보정할 수 있는 방법으로 제안되었다. 본 논문은 메타분석에서 쌍별 유사가능도 접근법의 성능과 문제점을 수치적으로 연구한다.

주요용어: 메타 분석, 선택 편향, 쌍별 유사가능도, 출판 편향

이 논문은 농촌진흥청 차세대 바이오그린21사업(PJ01337701)의 지원에 의해 수행되었습니다.

[1]교신저자: (08826) 서울특별시 관악구 관악로1, 서울대학교 보건대학원. E-mail: lwj221@gmail.com