creative commons
COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

보건학 석사 학위논문

복합 표본자료 기반 인과효과 추정을 위한 역확률 가중치 방법

– 국민건강영양조사 자료를 중심으로 –

Inverse Probability Weighting Methods for Estimating Causal Effects using Complex Survey Design Data

2023년 2월

서울대학교 보건대학원

보건학과 보건학전공

하 나 영

복합 표본자료 기반 인과효과 추정을 위한 역확률 가중치 방법

– 국민건강영양조사 자료를 중심으로 –

지도 교수 이 우 주

이 논문을 보건학 석사 학위논문으로 제출함
2022년 11월

서울대학교 보건대학원
보건학과 보건학전공
하 나 영

하나영의 보건학 석사 학위논문을 인준함
2022년 12월

위 원 장 원 성 호 (인)

부위원장 권 순 선 (인)

위 원 이 우 주 (인)

초 록

보건 분야에서는 국민의 건강 증진을 위한 정책 설립의 근거자료로 활용하기 위해서 복합 표본조사 설계를 기반으로 한 대규모 표본조사가 널리 시행되고 있다. 본 연구는 이러한 복합 표본조사 자료를 활용하여 역확률 가중치(Inverse Probability Weighting) 방법에 표본 가중치를 적절하게 적용하는 방법을 검토하는 것을 목적으로 한다. 역확률 가중치 방법은 두 단계로 이루어진다. 첫번째는 성향점수를 추정하는 단계이고, 두번째는 추정한 성향점수를 이용하여 모집단에 대한 인과효과를 추정하는 단계이다. 첫번째 단계에서 선행연구들의 접근 방식에 따른 차이를 파악하고, 국민건강영양조사 자료 분석 시 고려사항과 가이드라인을 제공하고자 한다. 각 단계에서 표본 가중치 고려여부에 따라 4가지 조합의 시나리오를 구성할 수 있다. 이러한 시나리오 하에서 모집단 평균 처치효과(Population Average Treatment Effect)의 추정치를 비교하기 위해 모의시험과 국민건강영양조사 자료분석을 수행하였다. 모의실험 결과에 따르면 성향점수를 추정하는 첫번째 단계에서 표본 가중치를 고려하는 여부와 관계없이 두번째 단계에서 이를 고려할 때 좋은 성능을 보였다. 더 나아가, 두 단계 모두 표본 가중치를 고려하는 방법이 표본에 선택될 확률과 처치가 의존하는 정도와 관계없이 가장 강건함을 확인하였다. 국민건강영양조사 자료의 표본 가중치는 무응답 보정이 포함되어 있으므로 처치와 표본 선택이 항상 독립이라고 보기는 어렵다. 따라서 가능하면 역확률 가중치 방법의 두 단계 모두 표본 가중치를 고려하는 방법으로 분석하는 것을 권장한다. 만약 이들이 독립이라면, 성향점수 추정 모델에 표본 가중치를 고려한 것과 하지 않은 결과가 비슷하므로 이를 통해 민감도 분석을 진행해보는 것도 추천한다.

주요어 : 복합 표본조사 자료, 역확률 가중치 방법, 국민건강영양조사, 모집단 평균처치효과

학 번 : 2021-27619

목 차

제 1 장 서 론 1

제 1 절 연구의 배경 1

제 2 절 선행연구 고찰 2

제 3 절 연구의 목적 3

제 2 장 이론적 배경 4

제 1 절 평균 처치효과 4

1. 잠재결과 프레임 워크
2. 평균 처치효과
3. 식별가능한 가정

제 2 절 성향점수를 이용한 역확률 가중치 방법 6

1. 성향점수
2. 역확률 가중치 방법

제 3 절 국민건강영양조사 8

1. 국민건강영양조사 표본 설계
2. 국민건강영양조사 표본 가중치

제 3 장 복합 표본조사 자료 기반 역확률 가중치 방법 10

제 1 절 복합 표본조사 자료 기반 역확률 가중치 방법 10

제 2 절 모의실험 12

1. 모의실험 설계
2. 모의실험 결과

제 3 절 국민건강영양조사 자료분석 18

1. 자료원 및 분석 대상
2. 분석 방법
3. 분석 결과

제 4 장 결 론 27

제 1 절 연구 결과 고찰 27

제 2 절 연구의 한계 및 제언 29

부록 30

참고문헌 35

Abstract 37

표 목차

[표 1] IPW 방법 각 단계에서 표본 가중치 사용 여부 13
[표 2] ATE 추정치에 대한 모의실험 결과 16
[표 3] 국민건강영양조사 제5기 자료 변수 정의 및 설명 20
[표 4] 표본 성향점수 모델의 오즈비 25
[표 5] 모집단 성향점수 모델의 오즈비 26

그림 목차

[그림 1] 잠재결과에 대한 유사 모집단 추정 과정 11
[그림 2] 모의실험 성능 비교 17
[그림 3] 평균 처치 효과 및 95% 신뢰구간 추정 24

부록 목차

[부록 A] 95% 신뢰구간 길이에 대한 박스 플롯 30

제 1 장 서 론

제 1 절 연구의 배경

보건 분야에서는 국민의 건강 증진을 위한 정책의제 설정 및 학술 연구의 근거 자료로 활용하고자 대규모 건강조사를 실시한다. 이러한 건강조사는 모집단의 대표성과 조사의 비용효과성을 확보하기 위해 집락(cluster), 층(strata) 등 복합 표본 설계(Complex sampling design)가 고려된다. 대표적으로 보건 분야에서 많이 활용되는 국민건강영양조사의 경우 2단계 층화집락 표본설계를 이용하여 추출되었으며, 통계분석에서 층, 집락, 표본 가중치 등 복합표본 요소를 고려하도록 권고하고 있다[1]. 목표 모집단의 대표성 및 정확성을 높이기 위해서 분석과정에서 표본 가중치(이하 SW; Sampling Weight)를 고려해야만 표본 설계 시점과 조사 시점간 가구수 및 인구수 차이에 따른 포함오차, 불균등 추출률, 조사 미참여자의 무응답오차 등을 보정할 수 있다[1]. 하지만 자료의 특성과 분석방법에 대한 복잡성으로 인해 이를 고려하지 못하고 잘못된 통계적 방법을 사용하여 결론을 도출하는 문제가 제기되고 있다[2]. 이는 잘못된 정책으로 이어질 수 있으므로 이에 대한 연구가 시급하다.

표본 가중치를 고려하지 않으면 모집단 수준(population level)의 인과 추론이 불가능하고 모수 및 분산의 추정에 있어 편향(bias)이 발생할 것이다. 또한, 관찰연구에서 처치군과 통제군의 분포간 균형을 보정하지 않으면 발생되는 선택 편향(selection bias)을 피할 수가 없게 되므로 이러한 문제를 해결하기 위해 성향점수(이하 PS; Propensity Score)를 활용한 방법이 있다. 하지만 복합 표본조사 자료를 활용하여 인과관계를 추정 및 추론하는 방법을 명확하게 제안한 연구가 많지 않다. 몇 안되는 문헌을 살펴보았을 때, 각자의 접근 방식과 의견이나 해석이 미묘하게 다르므로 분석하는 입장에서 혼동을 갖거나 잘못된 결과를 도출할 가능성이 있다.

자료의 대표성을 확보하기 위한 표본 가중치, 처치군과 통제군의 사전분포 균형을 맞추어 선택편향을 최소화하기 위한 성향점수를 모두 고려하기 위해 활용가능한 방법이 필요하다. 성향점수 방법 중 성향점수의 역수를 가중치로 활용한 역확률 가중치 방법(이하 IPW 방법; Inverse Probability Weighting Method)으로 복합 표본설계

자료에서 목표 모집단에 대한 인과효과 추정을 하고자 한다. 본 연구에서 표본 가중치를 고려한 IPW 방법에 대해 선행연구를 고찰하고, 모의실험을 통해 이를 검토하고자 한다. 또한, 국민건강영양조사 자료에 적용한 통계적 분석과 함께 분석 가이드라인을 제안하고자 한다.

제 2 절 선행연구 고찰

대부분의 연구에서 복합 표본 설계된 대규모 조사 자료를 통해 모집단의 처치효과를 추정하는 것에 관심이 있다. 성향점수를 이용한 방법으로 이를 추정할 때 성향점수를 추정하기 위한 모델과 이를 이용하여 인과효과를 추정하는 최종 결과모델을 구축하는 두 단계로 이루어진다. 선행 연구를 살펴보면 최종 결과모델 단계에서 모집단 수준의 처치효과를 추정하기 위해서 표본 가중치를 고려해야하는 것이 중요하다는 의견은 일치하는 것을 알 수 있다. 하지만 성향점수를 추정하는 단계에서 표본 가중치를 어떻게 적용하는지에 대해 크게 두가지 의견이 존재한다.

먼저, 성향점수 추정단계에서 표본 가중치를 고려하지 않아도 된다는 의견이 있다. Zanutto는 성향점수 방법 중 표본 수준에서 처치군과 통제군의 균형을 맞추기 위해 성향점수 모델이 사용되므로 성향점수를 추정하는 단계에서 표본 가중치를 고려하지 않아도 된다고 제안한다[3, 4]. Dugoff는 Zanutto와 의견을 같이 하며 성향점수를 추정하는 단계에서 두 집단의 분포를 고르게 할 뿐 성향점수의 분산을 사용하지 않으므로 표본 가중치를 고려하지 않아도 된다고 덧붙였다[5]. 28개의 제한적인 비판적 고찰, 모의실험과 실제 자료분석을 통해 관심 estimand와 그에 따른 해석의 중요성을 강조하였다. 특히 PATE(Population Average Treatment Effect)에 관심이 있는 경우 층화(Subclassification), 가중(Weighting) 방법을 이용하고, PATT(Population Average Treatment on Treated)의 경우 앞선 두가지 방법과 매칭(Matching) 중에서 선택하는 것을 제안한다. 하지만 앞선 두 문헌은 수학적 근거를 제시하지 않았고, 처치가 표본 선택과 독립인 경우에만 성립된다.

또 다른 하나는 성향점수 추정단계에서 표본 가중치를 고려하지 않으면 모집단에 대한 평균 처치효과를 올바르게 추정할 수 없다는 의견이다. Ridgeway는 처치가 표본 선택에 의존하는 상황에서 이들의

접근방식이 문제를 야기시킬 수 있음을 수학적으로 유도하였다[6]. 모의실험을 통해 두 단계에서 모두 표본 가중치를 고려하는 것이 강건(robust)하고, 경쟁력 있는 방법임을 보였다. Austin과 Lenis는 성향점수 추정하는 단계에서 표본 가중치를 고려하는 여부와 상관없이 모집단의 처치효과 추정의 성능이 비슷한 것을 보였다[7, 8]. Dong은 앞선 문헌에서 특정 성향점수방법만을 적용하여 추론하였음을 지적하였고, 층화, 매칭, 가중 방법 모두에서 성향점수 모델에 표본가중치를 고려하는 것을 추천하였다[9].

제 3 절 연구의 목적

본 연구의 목적은 다음과 같다.

(1) 복합 표본조사 자료의 설계요소를 고려하여 인과효과를 추정한 기존 문헌을 고찰하고 접근방식에 따른 방법을 정리하고자 한다.
(2) 표본 가중치와 성향점수의 역가중치를 활용한 IPW 방법을 모의실험을 통해 검토하고자 한다.
(3) (2)에서 검토한 분석방법을 국민건강영양조사자료에 적용하여 추정 모수와 분산을 비교하고, 복합 표본조사 자료에서 표본 가중치를 적절하게 고려한 인과성 추론의 중요성을 강조하고자 한다.

제 2 장 이론적 배경

제 1 절 평균 처치 효과

1. 잠재결과 프레임 워크 (Potential Outcomes Framework)

개인의 결과(Outcome)에 대한 처치(treatment), 노출(exposure), 개입(intervention) 등의 인과 효과를 설명하기 위해 잠재결과 프레임워크라는 개념이 필요하다.

처치변수는 T_i로 개체 i가 처치를 할당 받으면 1 아니면 0이라 설정한다. 결과변수를 Y_i라고 할 때, 개체 i가 처치를 할당 받았을 때의 잠재적 결과를 $Y_i(T_i = 1)$ 또는 $Y_i(1)$ 으로 나타내고, 처치를 받지 않았거나 다른 처치를 받았을 때의 잠재적 결과를 $Y_i(T_i = 0)$ 또는 $Y_i(0)$로 나타낸다. 각 개체의 두 잠재적 결과의 차이를 인과효과(Causal Effect)라 정의한다 [10]. 하지만 현실에서 두 잠재적 결과 중 할당 받은 처치에 따라 하나만 관찰되므로 개별적 인과효과를 보는 것은 사실상 불가능하다. 그 대신, 모집단의 평균 인과 효과에 대한 추론은 가능하다.

2. 평균 처치효과 (ATE ; Average Treatment Effect)

평균 처치 효과는 목표 모집단에 대한 잠재적 결과의 평균의 차이로 다음과 같이 정의한다.

$$\Delta_{ATE} = E[Y(1)] - E[Y(0)]$$

이는 만약 모든 사람들이 처치를 받았을 때의 잠재적 결과와 모든 사람들이 처치를 받지 않았을 때 잠재적 결과의 평균 차이를 의미한다. 평균 처치 효과는 모집단과 표본 수준의 평균 처치효과로 나눌 수 있다.

N명의 모집단에서 n명의 표본을 추출하였다면, S는 표본에 포함된 n명의 집합이고, 모집단 단위는 $i = 1, \dots, N$ 이라고 하겠다. 모집단 평균 처치 효과(PATE; Population ATE)는 모집단 N에 대한 잠재결과의 평균의 차이로 정의한다.

$$\Delta_{PATE} = E[Y_i(1)] - E[Y_i(0)]$$
$$= \frac{1}{N}\{\sum_{i=1}^{N} Y_i(1) - \sum_{i=1}^{N} Y_i(0)\}$$

표본 평균 처치 효과(SATE; Sample ATE)는 다음과 같이 n개의 표본에 대한 잠재결과의 평균의 차이로 나타낼 수 있다.

$$\Delta_{SATE} = E[Y_i(1) \mid S = 1] - E[Y_i(0) \mid S = 1]$$
$$= \frac{1}{n}\{\sum_{i \in S} Y_i(1) - \sum_{i \in S} Y_i(0)\}$$

3. 식별 가능한 가정 (Identifiability Assumptions)

목표 모집단의 인과효과를 추정하기 위해 다음과 같이 세 가지 가정을 만족해야 한다.

(1) 일치성 (Consistency) : 개체들 간 간섭이 없고, 처치 수준이 고정되었을 때 잠재 결과와 동일하게 관찰된다. 즉, 관찰된 결과변수 $Y = I(T = 1)Y(1) + I(T = 0)Y(0)$ 이다. 이는 문헌에 따라 SUTVA(Stable Unit Treatment Value Assumptions)라고도 불린다[11].

(2) 조건부 교환가능성 (Conditional exchangeability) : 처치와 잠재 결과변수는 교란변수가 주어졌을 때 독립임을 가정한다. X가 교란변수일 때 이는 $\{Y(1).Y(0)\} \perp T \mid X$로 나타낼 수 있다.

(3) 양수성 (Positivity) : 관찰된 교란변수가 주어졌을 때 처치를 할당 받을 확률이 0과 1사이에 있다.

위의 세가지 가정이 모두 성립한다면 평균 처치효과를 다음과 같이 조건부 기대값의 차이를 통해 추론할 수 있다.

$$\tau = E[Y(1)|T = 1, X] - E[Y(1)|T = 0, X] \quad \text{(교환가능성)}$$
$$= E[Y|T = 1, X] - E[Y|T = 0, X] \quad \text{(일치성)}$$

제 2 절 성향점수를 이용한 역확률 가중치 방법

1. 성향점수 (Propensity Score)

성향점수는 교란변수 X가 주어졌을 때 처치를 할당 받을 조건부 확률로 다음과 같이 $e(X) = P(T = 1|X)$로 정의한다. 이때 처치변수 T는 할당을 받으면 1 아니면 0으로 나타낸다. 이는 관찰연구에서 교란변수 X의 정보를 요약하고, 처치군과 통제군의 사전 분포의 균형을 맞추어 선택 편향(selection bias)를 최소화하는 방법으로 사용된다[12].

Rosenbaum & Rubin은 앞서 1.1절에서 언급한 식별가능한 세가지 가정 하에서 교란변수 X 대신에 성향점수 $e(X)$ 가 주어지더라도 교환가능성 성질은 다음과 같이 유지된다는 것을 보였다.

$$\{Y(1), Y(0)\} \perp T \mid e(X) \quad , \quad X \perp T \mid e(X)$$

위의 수식은 동일한 성향점수를 가지는 개체들간 처치 할당이 무작위임을 의미한다. 이러한 성질을 이용하여 아래와 같이 관찰연구에서 인과효과를 추정할 수 있다.

$$E\{E[Y \mid T = 1, e(X)]\} = E\{E[Y(1) \mid e(X)]\} = E\{Y(1)\},$$
$$E\{E[Y \mid T = 0, e(X)]\} = E\{E[Y(0) \mid e(X)]\} = E\{Y(0)\}$$

성향점수는 고차원의 교란변수 X를 1차원으로 요약할 수 있는 장점이 있다. 이는 가중(weighting), 매칭(matching), 층화(stratification)등 다양한 방식으로 인과효과 추론에 사용된다.

2. 역확률 가중치 방법 (Inverse Probability Weighting)

역확률 가중치 방법은 성향점수의 역수를 가중치로 활용하여 처치군과 통제군의 공변량 X의 분포에 대해서 균형을 맞추도록 한다. 처치군에는 $e(X) = \mathrm{P(T = 1|X)}$ 의 역수를 통제군에는 $1 - \mathrm{e}(X)$ 를 가중치로 부여 하여 유사 모집단(pseudo−population)을 나타낸다[13, 14]. $E\left[\frac{T_i Y_i}{e(X_i)}\right] = E\left[\frac{T_i Y_i(1)}{e(X_i)}\right]$ 는 조건부 교환 가능성 가정 하에서 $E\left[\frac{T_i Y_i}{e(X_i)}\right] = E[Y_i(1)]$이 되므로 평균 처치효과를 다음과 같이 나타낼 수 있다.

$$\Delta_{ATE} = E\left[\frac{T_i Y_i}{e(X_i)}\right] - E\left[\frac{(1 - T_i)\, Y_i}{1 - e(X_i)}\right]$$

Δ_{ATE}를 추정하기 위해 추정된 성향점수의 역수를 가중치로 가중 평균하여 다음과 같이 구할 수 있다.

$$\hat{\Delta}_{ipw} = \frac{1}{N}\{\sum_{i=1}^{N} \frac{T_i Y_i}{\widehat{e(X_i)}} - \sum_{i=1}^{N} \frac{(1 - T_i)\, Y_i}{1 - \widehat{e(X_i)}}\}$$

제 3 절 국민건강영양조사

질병관리본부에서 실시한 제5기 국민건강영양조사(2010-2012년)는 전국 규모의 검진조사, 건강설문조사, 영양조사 등을 통해 국민의 건강수준, 건강행태, 식품 및 영양섭취 실태 등에 대한 전국 및 시도 단위의 통계를 산출하기 위한 목적으로 조사되었다[1].

1. 국민건강영양조사 표본 설계

제5기 국민건강영양조사 자료는 3년 동안 매년 2단계 층화집락추출 방법으로 표본조사를 실시한 순환표본조사이다. 192개 표본 조사구를 추출하여 약 3,800 가구의 만 1세이상 가구원 전체를 대상으로 12개월간 시행되었다. 일반조사구와 아파트조사구를 각각 2009년 주민등록인구와 2008년 아파트시세조사 자료로 구분하여 1차 추출단위로, 각 조사구 내 가구를 2차 단위로 활용하였다. 1차 층화는 시도 및 주택 유형으로 하였고, 2차 층화는 성별, 연령대별 인구비율 등을 기준으로 26개 층의 일반 지역과 단지별 평당가격, 평균 평수 등을 기준으로 24개 층의 아파트 지역으로 이루어진다.

2. 국민건강영양조사 표본 가중치[15]

목표 모집단의 인구 및 가구에 대해 정확하게 추정하기 위해서 표본추출법에 근거한 설계 가중치로 보정하는 것이 필요하다고 권고한다. 국민건강영양조사의 자료분석을 위한 가중치는 매년 조사된 자료를 분석하기 위해 각 년도 별 가중치와 여러 해에 걸쳐 수집된 자료 분석을 위한 통합 가중치로 분류할 수 있다. 각 년도 별 가중치는 설계 가중치, 무응답 조정, 사후층화, 극단가중치 처리 단계를 거쳐 산출된다.

설계 가중치는 1,2차 추출단위인 조사구 및 가구 추출률의 곱의 역수로 정의한다. 무응답 조정은 대상 가구 및 대상자에 대한 응답률 곱의 역수를 가중치로 반영하여 무응답 편향을 보정할 수 있다. 즉, 추출률과 응답률의 각각 역수의 곱을 기본 가중치로 정의한다. 각 년도 별 우리나라 전체 인구수 및 가구수와 같도록 사후층화 과정을 통해 기본 가중치를 보정한다. 사후보정 가중치는 시도별, 성별, 5세 단위

16개 연령대별로 사후 층으로 사용하고, 이에 대한 분포 통계량 중 사분위범위를 벗어나는 극단 가중치의 경우, 범위의 경계 값으로 대체한 후 다시 사후보정을 하여 최종 가중치를 산출한다. 국민건강영양조사는 순환표본조사 방법으로 매년 전국통계가 산출되고, 시도통계 또는 세부집단 분석을 위해서 여러 해에 걸쳐 수집된 자료를 통합하여 사용한다. 예를 들어, 제5기의 2010-2011년 기수내 자료를 통합하는 경우 조사구 수가 각각 192개 이므로 통합가중치는 2010년 1/2, 2011년 1/2로 계산한다.

제 3 장 복합 표본조사 자료 기반 역확률 가중치 방법

제 1 절 복합 표본조사 자료 기반 역확률 가중치 방법

복합 표본조사 자료는 모집단의 특성을 잘 대표하고, 대규모 조사를 효율적으로 수행하기 위해서 층, 집락 등 복합 표본설계를 고려하여 표본을 추출하게 된다. 이러한 관찰 자료로부터 모집단 수준의 인과성 추론을 하기 위해서는 복합표본 설계를 고려하여 목표 모집단으로 이동하고, 잠재 결과 프레임 워크 개념을 도입하여 잠재 결과변수에 대한 유사 모집단을 형성하는 과정이 필요하다.

Nattino는 관찰된 표본자료에서 모집단 수준의 인과효과를 추정하기 위해서 두가지 경로를 제안하였다[16]. **그림 1**에서 우측 하단의 하늘색 박스에서 두가지 경로를 통해 잠재결과에 대한 유사 모집단을 형성할 수 있다. 첫번째로 고려할 수 있는 경로는 잠재결과에 대한 유한 모집단 수준에서 처치를 먼저 할당한 후 표본을 추출한 것이다. 이때 모집단 수준에서 처치를 할당 받을 확률은 $e_P(x) = P(T = 1|X)$으로 모집단 성향점수를 의미한다.

처치를 할당 받은 후 표본에 선택될 확률은 $\pi_{T,X} = P(S = 1|T = t, X)$ 이고, 그 역수를 표본 가중치 SW_T 라고 정의한다. 두번째 경로는 잠재결과에 대한 유한 모집단에서 잠재결과에 대한 표본을 추출한 후 처치를 할당하는 것이다. 처치를 할당 받기 전 표본에 선택될 확률은 $\pi_X = P(S = 1|\, X)$ 이고, 그 역수를 SW 라고 정의하였다. 표본수준에서 처치를 할당 받을 확률은 $e_s(x) = P(T = 1|\, S = 1, X)$ 으로 표본 성향점수를 의미한다. 표본에 선택되는 것과 처치를 할당 받는 것이 서로 독립인 경우에 이러한 경로를 구분할 필요가 없다. 하지만 서로 의존하는 경우 각각의 확률을 혼동하면 목표 모집단과 유사한 모집단을 형성하지 못하고 왜곡된 결과를 초래하므로 다음과 같이 구분하여 추정해야 한다.

$$\hat{\Delta}_{PATE,TS} = \frac{1}{N}\{\sum_{i\in S}\frac{T_i}{\widehat{e_P(X)}\,\pi_{T,X}}Y_i - \frac{1-T_i}{(1-\widehat{e_P(X)})\,\pi_{T,X}}Y_i\},$$

$$\hat{\Delta}_{PATE,ST} = \frac{1}{N}\{\sum_{i\in S}\frac{T_i}{\widehat{e_s(X)}\,\pi_T}Y_i - \frac{1\ -\ T_i}{(1-\widehat{e_s(X)})\,\pi_T}Y_i\}$$

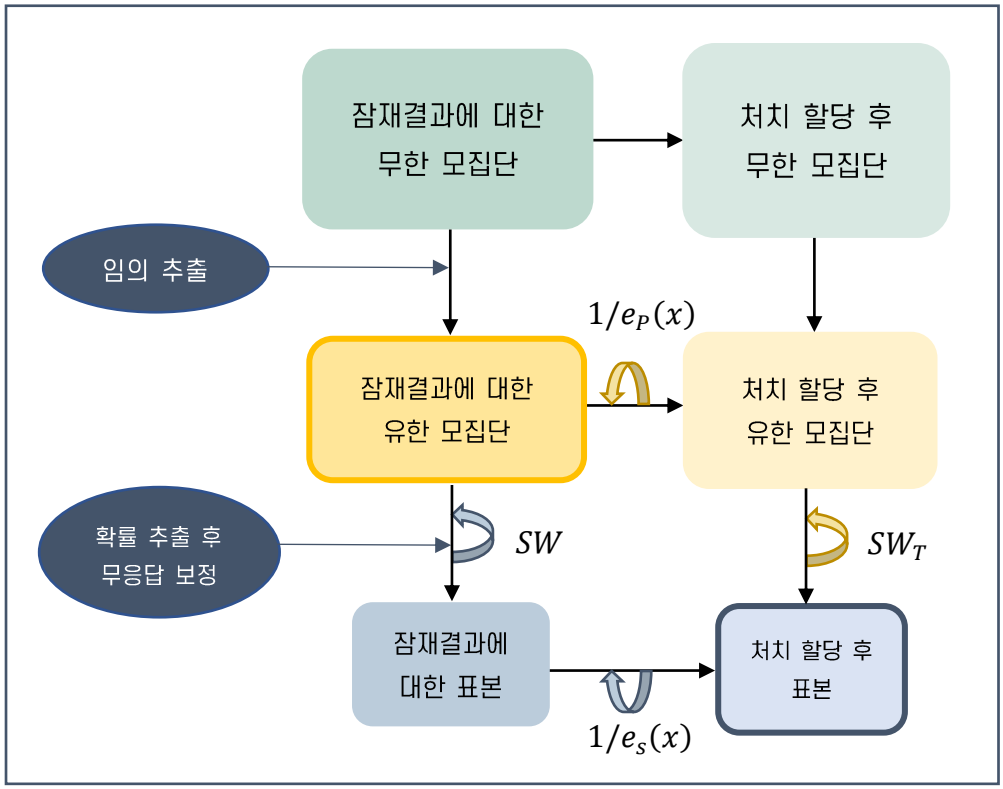

그림 1.관찰된 표본자료로부터 잠재결과에 대한 유사 모집단 추정 과정 : 우측 하단의 파란색 상자는 모집단으로부터 추출되고 처치 할당을 받은 후 최종적으로 관찰된 표본을 의미한다. 이를 시작으로 노란색 화살표 방향은 처치할당 후 표본에 선택된 확률 $\boldsymbol{\pi_{T,X}}$의 역수 인 $\boldsymbol{SW_T}$를 이용하여 처치 후 유한 모집단으로 이동하고, 모집단의 성향점수 $\boldsymbol{e_P(x) = P(T = 1|X)}$를 이용하여 잠재결과에 대한 유한 모집단에 도달하는 것을 의미한다. 파란색 화살표 방향은 표본 수준의 성향점수 $\boldsymbol{e_s(x) = P(T = 1|\, S = 1, X)}$ 를 이용하여 표본의 잠재결과 공간으로 이동하고, 표본 가중치 $\boldsymbol{SW = 1/\pi_X}$ 로 잠재결과에 대한 유한 모집단에 도달한다.

제 2 절 모의 실험

모의실험을 통해 복합 표본설계 하에서 모집단 수준의 평균 처치 효과(PATE)를 올바르게 추정하기 위한 IPW 방법을 검토하고자 한다. 모의실험 설계는 Dugoff와 Ridgeway의 문헌을 참고하였다[5, 6].

1. 모의실험 설계

본 모의실험 설계는 층화 추출과 표본설계 가중치를 포함하여 간단한 구조로 진행하였다. 공변량 X는 정규분포를 따르고, 처치변수 T 와 잠재결과변수 $Y(1), Y(0)$ 는 이산형이다. 9만 명의 모집단에서 동일한 크기의 층이나 선택변수 S가 1이 되는 확률 즉, 표본에 선택될 확률이 서로 다른 층 3개를 나누어 생성하였다. 또한, 표본 선택이 처치변수 T 에 의존하는 정도에 따라 각 방법의 결과가 어떻게 달라지는지 비교하고자 $S\sim(X,T)$ 에 대해 모델링 하였다. 이때 처치가 표본 선택에 영향을 주는 효과가 0인 경우 S는 T와 독립이고, X에 대해서만 의존한다. 자세한 내용은 아래 수식을 통해 설명하고자 한다.

단일 공변량 X는 각 층에 따라 정규분포 $X \sim \mathrm{N}\left(\mu_j, 1\right)$를 따르고, 각 층의 평균은 $\mu_j = \frac{j}{4} - \frac{1}{2}\ for\ j = 1,2,3$ 이다. 처치확률은 성향점수 모델 $logit\ P(t = 1|x) = -1 + log(4)\,X$ 을 사용하였고, 확률이 $\mathrm{e(x)} = P(t = 1|x)$ 인 베르누이 시행으로 처치변수 T 를 추출하였다. 선택확률은 $logitP(s = 1|x, t) = -2.8 - log(4)x + \delta\, t$ 이고, 델타 δ는 -1과 1 사이의 연속된 값을 가진다. 델타 δ가 0인 경우 선택변수 S가 처치변수 T와 독립이 되고, 0이 아닌 경우 처치가 표본 선택에 영향을 주는 크기가 변화함에 따른 각 방법의 결과를 비교하고자 한다. 이산형 잠재결과변수 $Y(1), Y(0)$는 확률 $p(1), p(0)$를 아래 수식과 같이 정의하고, 각 확률에 따른 베르누이 시행으로 추출된다. 또한, 처치를 할당 받았을 때 관찰된 결과변수 Y는 $Y = T\,Y(1) + (1 - T)Y(0)$로 정의한다.

$$logit\big(p(0)\big) = -0.2 + 0.2X$$
$$logit\big(p(1)\big) = -0.2 + 0.2X + log(2)$$

모의실험 시나리오는 **표 1**에 정리하였다. 1단계 성향점수 추정 모델과 이를 이용하여 인과효과를 추정하는 2단계 결과변수 모델에서 각각 표본 가중치를 고려하는 여부에 따라 4가지 조합으로 이루어진다. 첫번째 시나리오 SW_00은 1단계와 2단계 모두 표본 가중치를 사용하지 않으므로 표본 수준에서 추정한 성향점수 $P(T = t \mid S = 1, X)$만을 최종 가중치로 사용한다. 두번째 시나리오 SW_10은 1단계에만 표본 가중치를 사용한 것으로 모집단 수준의 성향점수 $P(T = t \mid X)$를 최종가중치로 사용한다. 세번째 시나리오 SW_01은 2단계에서만 표본 가중치를 사용한 것으로 1단계에서 표본 수준에서 추정한 성향점수와 표본 가중치의 곱인 $P(T = t \mid S = 1, X) * P(S = 1 \mid X)$을 최종 가중치로 사용한다. 네번째 시나리오 SW_11은 두 단계 모두 표본 가중치를 사용하는 것으로 모집단 수준의 성향점수와 표본 가중치의 곱인 $P(T = t \mid X) * P(S = 1 \mid X)$를 최종 가중치로 사용하였다. 각 시나리오의 성능은 Bias, MSE, 95% 신뢰구간의 coverage로 평가하고자 한다.

표 1. IPW 방법 각 단계에서 표본 가중치 사용 여부

Scenario	<Stage1> PS Model	<Stage2> Outcome Model
SW_00	None	None
SW_10	Weight	None
SW_01	None	Weight
SW_11	Weight	Weight

2. 모의실험 결과

앞서 설계한 4가지 시나리오 하에서 모의실험을 각 1,000번씩 수행하였다. ATE에 대한 추정치 결과를 T가 S에 주는 영향(δ)에 따라 **표 2**에 나타내었다. Mean은 1000번의 반복에 대한 추정치들의 평균, SE(Standard Error)는 1000번의 반복에 대한 추정치들의 표준오차의 평균이다. SD(Standard deviation)는 SE의 참값으로서 1000번의 반복에 대한 추정치들의 표본 표준편차이다. ATE의 참값은 0.171이다.

모의실험에 의한 결과는 다음과 같다. SW_00방법과 SW_10방법의 추정치들의 평균이 약 0.190이고, SW_01방법과 SW_11방법의 추정치들의 평균은 약 0.170으로 참값과 유사하다. SW_00방법과 SW_10방법은 전반적으로 SE값이 SD값보다 크므로 SE값이 과대 추정된 것임을 알 수 있다. 하지만 SW_10방법은 SE값이 SD값보다 작으나 앞선 두 방법에 비해 δ가 커짐에 따라 변동성(SE값과 SD값)이 커진다. SW_11방법은 δ에 관계없이 SE값이 SD값과 거의 비슷함을 보인다.

그림 2의 그래프 A와 B는 각각 편향(Bias)과 평균 제곱오차(MSE; Mean Squared Error)를 각 시나리오 방법에 따라 비교한 것이다. 두 평가지표 모두 0에서 0.1 범위 내에 존재하고, 이는 값이 작을수록 좋은 성능을 보인다. SW_00방법과 SW_01방법은 편향이 0보다 조금 큰 편이고, SW_01방법은 편향이 거의 0에 가깝지만 δ가 커질수록 조금 과소 추정되는 것을 알 수 있다. SW_11방법은 δ에 관계없이 대체로 0에 가깝다. MSE는 δ가 0일 때, 4가지 방법 모두 거의 차이가 없지만 δ가 커질수록 SW_10방법의 MSE가 증가한다. 그래프 C는 95% 신뢰구간 포함확률(CP; Coverage Probability of 95% Confidence Interval)로 0과 1 사이의 값을 가지고, 확률이 0.95에 가까울수록 성능이 좋다. SW_00방법과 SW_10방법은 95% 신뢰구간의 포함확률이 0.95에 거의 못 미치는 성능을 보인다. SW_01방법과 SW_11방법의 95% 신뢰구간의 포함확률이 0.95에 가깝다. δ가 변화함에 따라 4가지 방법의 95% 신뢰구간의 길이를 박스 플롯으로 나타낸 것을 **부록 A**(그림 A-1 ~ 그림 A-5)에 제시하였다.

모의실험의 결과를 종합하여 평가하면, 두 단계 모두 표본 가중치를 고려한 SW_11의 결과가 표본에 선택될 확률이 처치에 의존하는 정도와 관계없이 Bias, MSE, Coverage probability 관점에서 강건(robust)하다고 볼 수 있다. 만약 표본에 선택될 확률이 처치와

독립이라면 SW_01방법도 좋은 성능을 보인다. 즉, S와 T가 독립인 경우에는 성향점수 추정 단계에서 표본 가중치를 고려하는 여부와 관계없이 마지막 단계에서만 표본 가중치를 잘 고려해주면 된다. 하지만 S가 T에 의존하는 경우에는 SW_10과 같이 성향점수 추정 단계에서 표본 가중치를 고려하지 않으면 변동성이 크고 편향된 결과를 도출하게 된다.

표 2. ATE 추정치에 대한 모의실험 결과

	SW_00			SW_10			SW_01			SW_11		
δ	Mean	SD	SE	Mean	SD	SE	Mean	SD	SE	Mean	SD	SE
−1	0.199	0.031	0.035	0.198	0.031	0.036	0.176	0.030	0.028	0.171	0.035	0.034
−.5	0.199	0.025	0.029	0.198	0.025	0.030	0.175	0.029	0.027	0.172	0.031	0.030
0	0.196	0.020	0.025	0.196	0.020	0.025	0.169	0.033	0.030	0.170	0.029	0.028
.5	0.193	0.017	0.022	0.196	0.016	0.021	0.162	0.044	0.036	0.169	0.029	0.028
1	0.192	0.016	0.020	0.197	0.015	0.019	0.161	0.058	0.044	0.173	0.028	0.027

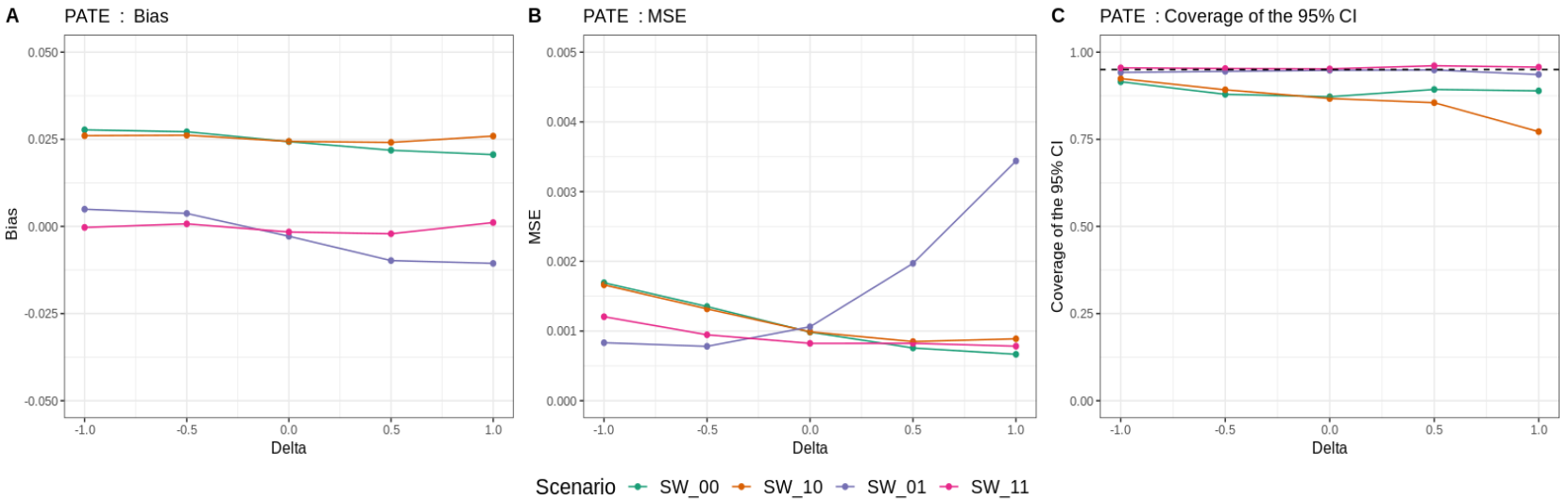

그림 2. 모의실험 성능 비교 : 좌측 그래프 A는 Bias을 추정한 것이고, 가운데 그래프 B는 MSE를 추정한 것이다. 우측 그래프 C는 95% 신뢰구간의 포함확률을 나타낸 것이다. 각 그래프의 x축은 선택변수 $\boldsymbol{S}$가 처치변수 $\boldsymbol{T}$에 영향을 받는 정도를 −1에서 1사이 의 5개 값으로 나타낸 것이다. y축은 각 시나리오의 성능 평가지표 값을 의미한다. 첫번째 시나리오 SW_00은 IPW 방법의 두 단계에서 모두 표본 가중치를 사용하지 않은 것이다. 두번째 시나리오 SW_10은 1단계에서만 표본 가중치를 사용한 것이다. 세번째 시나리오 SW_01은 2단계에서만 표본 가중치를 사용한 것이고, 네번째 시나리오 SW_11은 두 단계 모두 표본가중치를 사용한 것이다.

제 3 절 국민건강영양조사 자료 분석
:
50세 이상 중년 여성의 현재 골관절염 유병에 대한 과거 모유수유 경험의 효과

1. 분석 배경 및 목적

국민건강영양조사에서 제공하는 표본 가중치는 단순히 표본 추출률만 보정된 것이 아니라 무응답 보정도 포함되어 있다. 처치변수가 표본조사 대상자 추출에 직접적인 영향을 주는 것은 아니지만 무응답에는 영향을 줄 수 있다. 표본 가중치가 처치변수와 독립인 경우를 판단하기 위해서는 표본 추출 대상에 선택될 뿐만 아니라 무응답 여부까지도 고려해야한다. 독립인 사실이 분명한 상황에서는 제2절 모의실험에서 보인 것과 같이 마지막 단계인 결과 모델에서만 표본 가중치를 이용하여 목표모집단에 대한 추정을 하면 된다. 하지만 연구 설계에 따라 처치변수가 달라지고, 그때마다 처치변수가 표본 선택에 얼마나 영향을 주는지에 대해 판단하기가 어렵다. 따라서 일반적인 상황에서 국민건강영양조사 자료를 안전하게 분석하는 방법에 대해 앞서 제안한 시나리오 모델을 토대로 검토해보고자 한다.

국민건강영양조사는 단면 연구(Cross-sectional study)로 변수 간 시간적 선후관계를 판단하기 어렵다. 하지만 인과성 추론을 하기 위해서 노출(T)과 질병(Y)의 선후관계가 명확해야 하므로 국민건강영양조사 자료를 활용한 기존 문헌 중 이를 만족하는 문헌을 선택하여 재현하였다. Kim(2020)은 모유수유 경험과 관절통증 및 골관절염에 대한 연관성 분석을 하였다[17]. 모유수유는 대부분의 여성들이 2~30대에 경험한 것이고, 50세 이상 여성의 현재 골관절염은 이후에 일어난 일이므로 선후관계가 명확하다. 이 논문은 연관성을 분석한 것이라 인과효과를 추정한 결과와 비교하기는 어렵지만 두 변수간 연관성을 보이는 동일한 세팅에서 인과효과를 추정하는 것에 의의가 있다. 따라서 해당 논문을 재현하고, 동일한 세팅에서 제2절 모의실험에서 제시한 4가지 시나리오에 따른 IPW 방법을 적용하여 비교하고자 한다.

2. 분석 대상

제5기 국민건강영양조사자료 중 골밀도 검사가 시행된 2010년, 2011년 원시자료를 이용하여 분석을 진행하였다. 2010년 8,958명과 2011년 8,518명 중에서 50세 이상이 아닌 참여자 10,759명, 모유수유 정보가 없는 3,213명, 골관절 관련 정보가 없는 512명, 관절통증 정보가 없는 446명을 제외하면 3,454명이다. 건강설문조사 및 검진조사에 모두 참여한 사람 중 관련된 항목에 결측이 없는 3,093명을 최종 대상자로 선정하였다.

(1) 관심 결과변수인 골관절염 또는 관절통증 여부는 건강설문 이환 항목에서 무릎 또는 엉덩관절 통증에 대한 설문에 응답하였고, 방사선적 진단 Kellgren Lawrence grade 기준으로 2점 이상인 경우를 골관절염이 있다고 정의한다.

(2) 처치변수인 모유수유 여부에 대해서는 여성건강 항목에서 “최소한 한달 이상 모유수유 한 경험이 있습니까?”에 대한 질문에 “예”라고 응답하고, 총 모유수유한 기간이 1개월 이상인 경우를 모유수유를 받은 그룹을 처치 그룹으로 정의한다. 또한 모유수유를 하지 않았거나 그 기간이 1개월 미만인 경우를 통제 그룹이라 정의한다.

(3) 결과변수와 처치변수에 모두 영향을 줄 것으로 예상되는 교란변수는 연령, 체질량 지수, 가구 소득, 흡연여부, 음주 소비량, 신체활동, 과거 병력(당뇨, 고혈압, 이상지질혈증), 모유수유 자녀 수, 유산 횟수, 폐경 여부, 여성 호르몬제 복용여부 등이 있다. 교란변수를 포함하여 분석에서 사용한 변수 정의 및 설명은 **표 3**에 정리하였다.

표 3. 국민건강영양조사 제5기 자료 변수 정의 및 설명

변수명		원시자료 변수	변수 설명
처치변수 (Treatment)			
Breastfeeding	모유수유 여부	lw_br, lw_br_dur	최소 한달 이상 모유수유한 경험 여부
결과변수 (Outcome of interest)			
OA	골관절염 유병여부	oa_kh	골관절염 유병 여부 (방사선적 진단값)
교란변수 (Confounder)			
Age	연령	Age	만 나이
Income	가구소득 사분위수	ho_incm	가구 소득 1: 하, 2: 중하, 3: 중상, 4: 상
BMI	BMI (범주형)	he_bmi	체질량 지수 1 : 저체중(≤ 18.5) , 2 : 보통, 3 : 비만 (≥ 25)
Smok	흡연 여부	bs3_1	현재흡연 여부 1 : 현재흡연, 2 : 과거 흡연 3 : 비흡연
Alchol	음주 빈도	bd1_11	1 : 전혀 마시지 않음, 2 : 월1회미만, 3 : 월4회이하, 4 : 주2~4회이상

Daily_act	일상활동	be2	1: 안정 또는 가벼운 활동, 2 : 보통 활동, 3 : 심한 또는 격심한 활동
Diabetes	당뇨	he_dm,	공복혈당≥126 또는 당뇨병약 복용 또는 인슐린주사 투여 또는 의사진단 (건강설문 포함)/ 해당없는 경우 0으로 간주
Hypertension	고혈압	he_sbp, he_dbp, di1_2	수축기혈압≥140mmHg 또는 이완기혈압≥90mmHg 또는 고혈압약 복용
Dyslipid	이상지질혈증	di2_pr	이상지혈증 현재 유병 여부
Total_child	모유 수유 자녀 수	lw_br_dur	모유수유 자녀 수 (명)
Abortion	유산 횟수	lw_pr_2 (자연유산), lw_pr_3 (인공유산)	자연유산, 인공유산 횟수로 총 유산 횟수 산출
Menopausal	폐경 여부	lw_ms	1: 월경 여부가 “3.아니오” 또는 “9.모름”인 경우, 0 : 월경 여부가 “2.예”인 경우
Hormone	여성호르몬제 복용기간	lw_wh_dur	1: 복용안함, 2: 6개월미만, 3: 12개월미만, 4: 1년이상

3. 분석 방법

대한민국 50세 이상 중년 여성의 모유수유 경험과 현재 골관절염 유병 여부의 인과관계를 알아보기 위해서 모집단 평균 처치 효과(PATE)를 추정하였다.

1단계인 성향점수 추정 모델에서는 처치변수인 모유수유 여부가 종속변수가 되고, 앞서 정의한 교란변수가 예측변수로 사용된다. 이 단계에서 표본 가중치를 고려하지 않는 SW_00과 SW_01 방법은 종속변수가 이항변수일 때 일반적으로 사용하는 로지스틱 회귀모형을 사용하였다. SW_10과 SW_11 방법은 표본 가중치를 포함한 복합 표본설계 요소를 고려하기 위해 R version 4.2.0에서 제공하는 'survey' 패키지를 이용하여 분석하였다. 집락, 층, 표본 가중치를 svydesign 함수로 설계하고, svyglm 함수의 design 옵션에 포함하였다. 또한, family 옵션은 quasi-binomial likelihood로 설정하여 분석하였다.

2단계는 성향점수로 이루어진 가중치의 역수를 최종 가중치로 하여 목표 모집단에 대한 marginal effect를 추정하였다. 이때 결과변수를 골관절염 유병여부, 예측변수를 모유수유 여부로 사용하였다. SW_00과 SW_10 방법과 같이 표본 가중치를 고려하지 않는 경우, 앞서 추정한 성향점수의 역수를 최종 가중치로 사용하는 일반적인 IPW 방법과 같이 분석하였다. 하지만 표본 가중치를 고려하는 SW_01과 SW_11의 경우 성향점수와 표본가중치를 곱한 값의 역수를 최종 가중치로 사용하였다. 이렇게 구한 최종 가중치를 앞서 사용한 svyglm 함수에 반영하여 quasi-binomial likelihood로 분석하였다.

4. 분석 결과

분석에 포함된 50세 이상 여성 3093명 중 최소 한달 이상 모유수유를 경험하지 않은 270명은 통제군에 포함되고, 모유수유 경험이 있는 2823명은 처치군에 포함되었다.

그림 3은 50세 이상 여성의 골관절염 유병에 대한 최소 한달 이상 모유수유 경험의 모집단 평균 처치효과를 4가지 방법에 따라 추정한 것이다. SW_00은 두 단계 모두 표본 가중치를 고려하지 않은 것으로 모유수유를 경험한 처치군이 통제군 대비 골관절염 유병의 1.38배 더 높으나 유의한 결과는 아니다[95% CI 0.50−3.82]. 1단계에만 표본 가중치를 고려한 SW_10의 결과는 OR 1.33 [95% CI 0.44−4.05]이다. 2단계에만 표본가중치를 고려한 SW_01은 OR 1.96 [95% CI 0.83−4.61]이고, 두 단계 모두 표본 가중치를 고려한 SW_11은 OR 1.91 [95% CI 0.83−4.61]이다. 4가지 방법 모두 오즈비(Odds Ratio)의 신뢰구간이 1을 포함하고 있어 유의한 결과를 나타나지는 않았다.

하지만 SW_00과 SW_10의 결과가 비슷하고, SW_01과 SW_11의 결과가 서로 비슷한 패턴을 보임을 알 수 있다. 전자의 경우 성향점수를 이용하여 인과효과를 추정하는 2단계에서 표본 가중치를 고려하지 않는다는 공통점이 있다. 또한, SW_01과 SW_11은 2단계에서 표본 가중치를 고려하는 공통점이 있다. 즉, 각각의 두 방법들은 1단계에서 표본 가중치를 고려한 여부가 다름에도 그 결과가 서로 비슷하므로 앞 절의 모의실험 결과를 견주어 보았을 때 S와 T가 독립인 상황이라고 볼 수 있다. 이를 통해서 1단계에서 표본 가중치를 고려한 것과 고려하지 않는 결과를 비교하여 그 결과가 비슷하다면 S 와 T 가 독립인 것을 간접적으로 확인 할 수 있게 된다. 이러한 방식은 S와 T의 독립 여부가 모호한 경우 민감도 분석으로 고려해볼 수 있겠다.

표 4는 표본 가중치를 고려하지 않은 표본 성향점수 모델의 오즈비를 나타낸 것이고, 연령, 강도가 높거나 격한 활동을 하는 경우, 다자녀인 경우 모유수유 여부에 유의한 결과를 보인다. **표 5**는 표본 가중치를 고려하여 모집단에 대한 성향점수 모델의 오즈비를 나타낸 것으로 **표 4**과 비슷한 경향을 보이나 고소득 가구에 대한 유의한 결과가 추가되었다. 이는 표본 가중치에 가구소득이나 지역, 연령 등의 정보가 포함되어 있어서 이러한 변수가 처치변수나 결과변수에 영향을 주는 경우 성향점수 모델에 표본 가중치를 고려하지 않으면 편향된 결과를 가져올 수 있음을 시사한다.

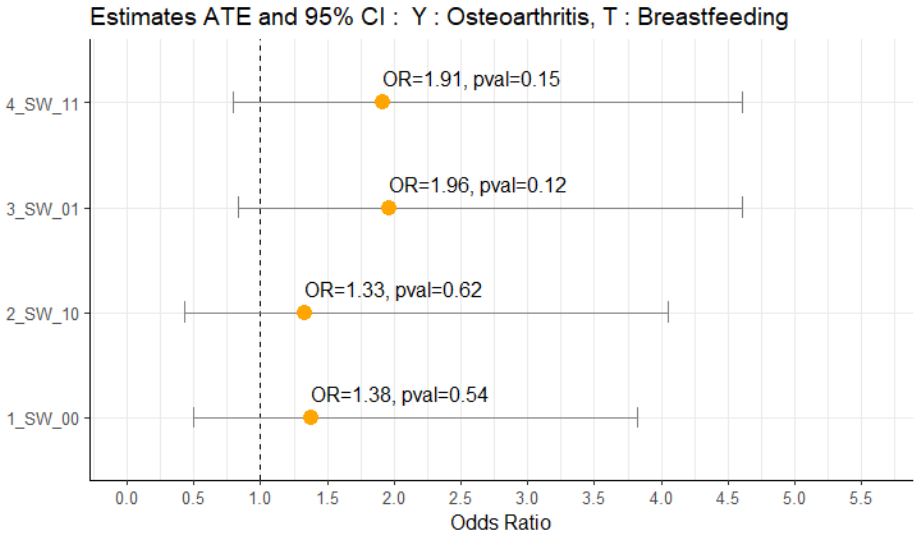

그림 3. 평균 처치 효과(ATE) 및 95% 신뢰구간 추정 : x축은 Odds Ratio(OR)로 통제군 대비 처치군의 골관절염 발생 오즈 비를 나타낸 것이다. y축은 각 방법에 따른 결과값을 나타낸 것으로 1_SW_0은 표본 가중치를 두 단계 모두 사용하지 않은 방법이고, 2_SW_10은 성향점수 추정하는 1단계에서 표본 가중치를 고려한 방법, 3_SW_01은 인과효과 추정하는 2단계에서 표본 가중치를 고려한 방법, 4_SW_11은 두 단계 모두 표본 가중치를 고려한 방법이다. 각 점은 OR을 나타낸 것이고, 점 위에 OR과 p값을 나타내었다. 각 선은 OR에 대한 95% 신뢰구간을 의미한다.

표 4. 표본 성향점수 모델의 오즈비

Confounder		OR (95% CI, p-value)
Age		**1.06 (1.03-1.09, p<0.001)**
Income	Low	-
	middle-high	0.79 (0.47-1.31, p=0.356)
	middle-low	0.86 (0.50-1.47, p=0.588)
	High	0.59 (0.35-0.98, p=0.043)
Smok	never	-
	ex-smoker	0.62 (0.27-1.70, p=0.314)
	current	0.59 (0.29-1.35, p=0.184)
Alchol	Never	-
	under1	1.20 (0.81-1.78, p=0.370)
	under5	1.53 (1.00-2.39, p=0.055)
	over5	1.49 (0.74-3.27, p=0.290)
BMI	Normal	-
	Under	0.42 (0.17-1.20, p=0.076)
	Obese	1.11 (0.78-1.59, p=0.566)
Daily_act	Light	-
	moderate	1.07 (0.77-1.49, p=0.670)
	intense	**4.57 (1.86-14.02, p=0.003)**
Diabetes	yes	1.09 (0.76-1.58, p=0.645)
Hypertension	yes	1.02 (0.72-1.45, p=0.902)
Dyslipid	yes	0.88 (0.56-1.42, p=0.581)
Total_child	none	-
	1_2	**144.20 (53.16-509.13, p<0.001)**
	3_4	**399.48 (144.56-1423.96, p<0.001)**
	over5	**1154.28 (308.59-5728.94, p<0.001)**
Abortion	None	-
	1_2	1.05 (0.71-1.53, p=0.818)
	over3	1.12 (0.69-1.82, p=0.647)
Menopausal	yes	0.99 (0.58-1.64, p=0.972)
Hormone	None	-
	Under6	1.89 (0.87-4.89, p=0.141)
	Under12	2.57 (0.49-37.46, p=0.373)
	Over12	1.41 (0.84-2.52, p=0.218)

표 5. 모집단 성향점수 모델의 오즈비

Confounder		OR (95% CI, p-value)
Age		**1.07 (1.03-1.11, p<0.001)**
Income	Low	-
	middle-high	0.93 (0.54-1.59, p=0.779)
	middle-low	0.80 (0.47-1.37, p=0.418)
	High	**0.57 (0.35-0.93, p=0.025)**
Smok	never	-
	ex-smoker	0.71 (0.30-1.68, p=0.438)
	current	0.92 (0.40-2.16, p=0.855)
Alchol	Never	-
	under1	1.25 (0.81-1.92, p=0.310)
	under5	1.45 (0.90-2.34, p=0.130)
	over5	1.29 (0.59-2.83, p=0.518)
BMI	Normal	-
	Under	0.49 (0.20-1.23, p=0.129)
	Obese	1.19 (0.81-1.74, p=0.376)
Daily_act	Light	-
	moderate	1.13 (0.78-1.63, p=0.528)
	intense	**3.54 (1.20-10.51, p=0.023)**
Diabetes	yes	1.20 (0.79-1.83, p=0.384)
Hypertension	yes	0.94 (0.63-1.39, p=0.743)
Dyslipid	yes	0.75 (0.42-1.33, p=0.320)
Total_child	none	-
	1_2	**131.57 (41.46-417.52, p<0.001)**
	3_4	**314.68 (100.08-989.42, p<0.001)**
	over5	**1241.05 (279.07-5518.97, p<0.001)**
Abortion	None	-
	1_2	1.18 (0.77-1.82, p=0.440)
	over3	1.07 (0.67-1.71, p=0.792)
Menopausal	yes	0.89 (0.50-1.57, p=0.680)
Hormone	None	-
	Under6	1.75 (0.70-4.38, p=0.233)
	Under12	3.30 (0.73-14.84, p=0.119)
	Over12	1.59 (0.90-2.81, p=0.113)

제 4 장 결 론

제 1 절 연구 결과 고찰

본 연구를 통해서 복합 표본조사 자료를 활용하여 모집단의 인과효과를 추정하는 IPW 방법에 대해 검토하였다. 선행연구 고찰을 통해 성향점수 추정하는 첫번째 단계에서 표본 가중치 고려여부에 대한 의견이 상이한 것을 알 수 있었다. 하지만 모집단 수준의 인과효과를 추정하는 두번째 단계에서 성향점수의 역수와 표본 가중치의 곱을 최종 가중치로 사용하는 것에 대해서는 이견이 없었다.

본 연구에서 관찰된 표본자료를 잠재결과에 대한 목표 모집단 수준으로 추정할 때 제 3장의 그림 1과 같이 두 가지 경로가 존재하고, 이러한 이유로 선행연구에서 이견이 발생한 것을 알 수 있었다. 두 경로 중 하나는 처치 할당을 받은 표본 자료를 모집단 수준으로 이동하기 위해 표본 가중치를 사용하고, 인과효과를 추정하기 위해서 성향점수의 역수로 잠재결과에 대한 유사 모집단을 형성한다. 이 경우에는 Ridgeway et al., Dong et al.에서 언급된 바와 같이 첫번째 단계에서 표본 가중치를 고려하여 모집단 수준의 성향점수를 추정해야 한다. 이때 표본 가중치는 처치에 따라 표본에 선택될 확률이 다를 수 있으므로 표본 선택변수 S와 처치변수 T가 서로 의존한다고 볼 수 있다. 또 다른 경로는 표본 수준에서 성향점수를 추정하여 잠재결과에 대한 표본을 형성하고, 이에 표본 가중치를 곱하여 잠재결과에 대한 모집단에 도달한다. 이 경우는 표본 자료로부터 성향점수를 추정하기 때문에 Dugoff et al.에서 언급된 내용과 같이 첫번째 단계에서는 표본 가중치를 고려하지 않아도 된다. 또한, 처치 할당을 받기 이전에 표본을 선택하였으므로 T와 S가 독립이라고 볼 수 있다. 이처럼 표본이 어떠한 과정으로 추출되었는지에 따라 성향점수 추정하는 첫번째 단계에서 표본 가중치를 고려하는 여부가 달라진다.

본 연구에서는 4가지 시나리오 하에서 진행된 모의실험을 통해 두 단계 모두 표본 가중치를 고려한 SW_11 방법이 T가 S에 미치는 영향과 관계없이 추정치에 대한 Bias, MSE, Coverage probability 관점에서 가장 우수한 성능을 보였다. T와 S가 독립인 경우, 두번째 단계에서만 표본 가중치를 고려한 SW_01 방법도 SW_11방법과 유사하게 좋은 성능을 보였다.

국민건강영양조사 자료를 활용하여 모집단 수준의 분석을 할 때, 대부분 성향점수 추정단계에서 표본 가중치를 고려하지 않는 SW_01방법을 많이 사용한다. 국민건강영양조사 자료는 표본조사 이전에 처치 할당이 이루어진 단면연구 자료로써 처치변수에 따라 표본에 선택되는 대상이 달라진다고 보기는 어렵다. 하지만 자료에 주어진 표본 가중치는 표본 추출률뿐만 아니라 무응답보정 등 사후보정을 하여 제공된 값이므로 처치변수에 따라 무응답율에는 영향을 줄 수 있을 것이다. 처치변수는 연구설계에 따라 달라지므로 S와 T가 항상 독립이라고 보기는 어렵다. 현실적으로 S와 T의 관계를 명확하게 아는 것은 어려우므로 두 단계 모두 표본 가중치를 고려하는 SW_11방법으로 모집단 수준의 인과효과를 추정하는 것이 안전하다. 더 나아가 민감도 분석으로 두번째 단계에서만 표본 가중치를 고려한 SW_01방법을 앞서 제시한 SW_11방법과 분석하여 비교해보는 것을 추천한다. 만약 S와 T가 독립이라면 두 방법의 결과가 비슷하게 나올 것이다. 하지만 두 결과가 다르게 나온다면 두 변수가 서로 독립이 아니거나 성향점수 추정 모델이 틀렸다는 근거가 될 수 있다.

제 2 절 연구의 한계 및 제언

본 연구 사용한 IPW 방법은 복잡한 자료 구조로 확장이 용이하여 복합 표본설계자료를 활용하여 목표 모집단과 유사한 수준에서 인과효과를 추정하기에 쉽게 고려할 만한 방법이다. 하지만 성향점수의 역수와 표본 가중치의 곱을 최종 가중치로 사용할 때 극단적인 값이 나오는 경우를 피할 수 없으므로 이를 절단하거나 보정하는 과정이 필요할 것이다. 본 연구에서 성향점수 모델에 표본 가중치를 고려하는 여부에 초점을 두었는데 만약 모델 자체가 틀린 경우 표본 가중치를 적절하게 고려하더라도 편향된 결과를 도출할 수 있는 한계가 있다. 이에 대해서는 앞서 제시한 민감도 분석을 통해서 모델이 틀린 근거를 검토해보는 것이 좋다. 또한, T와 S에 모두 영향을 주는 X뿐만 아니라 S에만 영향을 주는 변수를 고려하지 않고 보정을 하거나 T와 S가 실제로 독립인데 의존한다고 가정하여 분석하게 되면 충돌변수(collider) 역할을 하여 편향된 결과를 초래하게 된다. 추후에 국민건강영양조사 표본설계와 더 유사한 복잡한 구조에 대해 검토할 필요가 있다.

국민건강영양조사 자료의 특성상 표본 가중치 계산에 사용된 변수와 처치 변수 간의 관계를 명확하게 파악하는 것이 쉽지 않을 것이다. 또한, 성향점수 모델에 표본 가중치를 고려한 분석이 크게 복잡하지 않으므로 두 변수 간의 관계와 무관하게 강건한 결과를 제공하는 두 단계 모두 표본 가중치를 고려하는 방법을 권장한다. 본 연구를 통해서 보건 분야에서 많이 활용되는 국민건강영양조사 자료로 목표 모집단의 평균 처치효과를 추정하기 위해 편향을 최소화하는 분석 방법과 분석 시 고려사항을 제공함으로 국민의 건강을 위한 정책 및 의사결정에 기여할 수 있다.

부록 A

95% 신뢰구간 길이에 대한 박스플롯 : S에 T가 미치는 정도(Delta)에 따라 4가지(SW_00, SW_10, SW_01, SW_11) 시나리오 하에서 1000번의 모의실험 결과에 따른 95% 신뢰구간 길이를 박스플롯으로 나타낸 것이다.

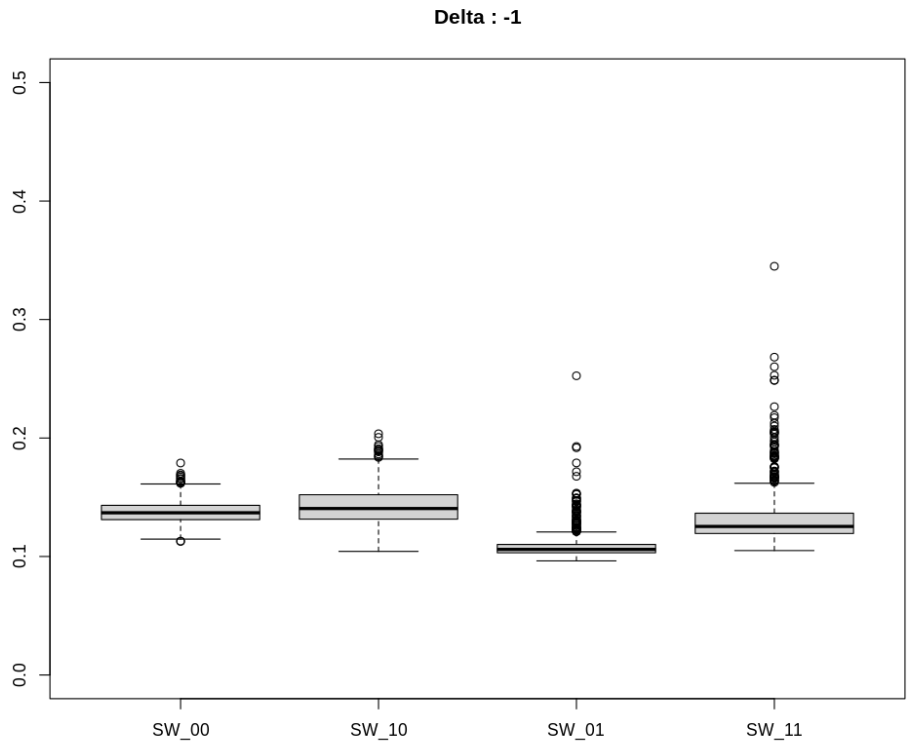

그림 A-1. Delta가 -1인 경우의 95% 신뢰구간 길이에 대한 박스플롯

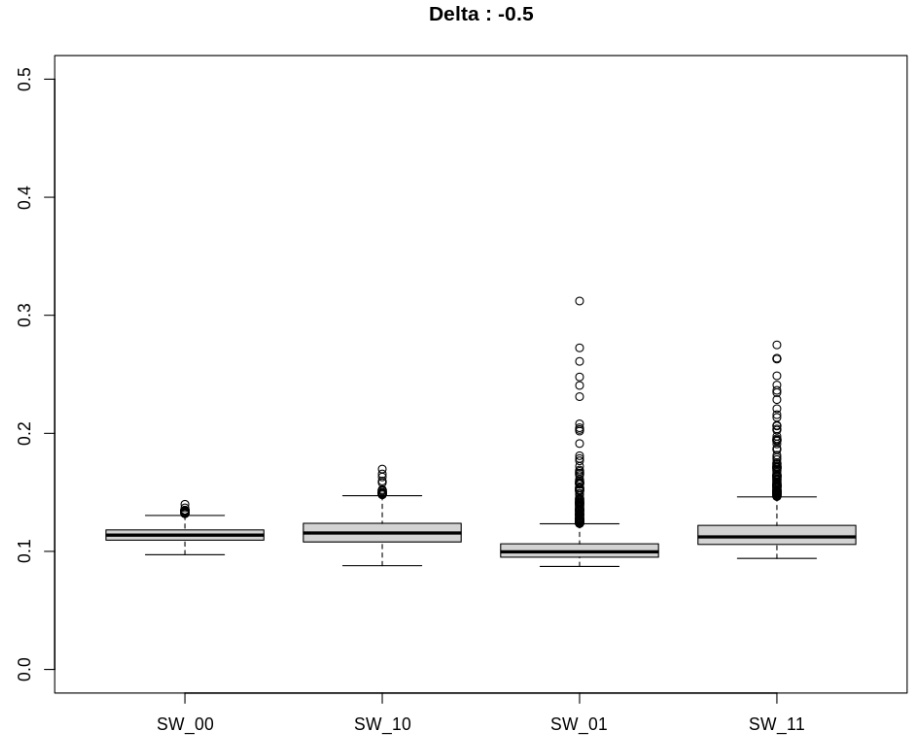

그림 A-2. Delta가 -.5 인 경우의 95% 신뢰구간 길이에 대한 박스플롯

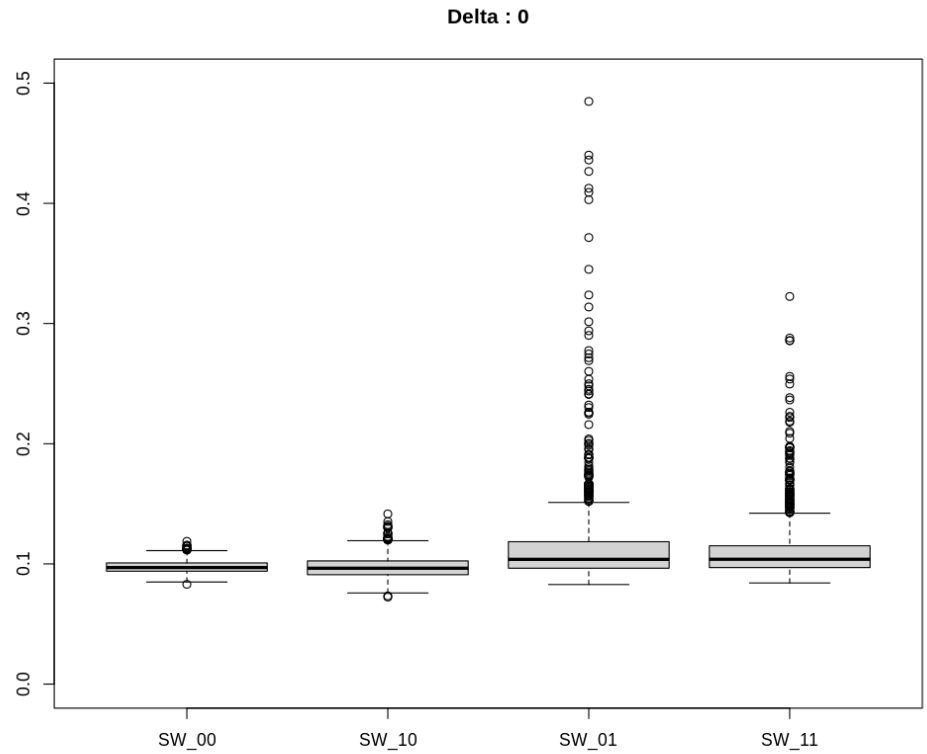

그림 A-3. Delta가 0인 경우의 95% 신뢰구간 길이에 대한 박스플롯

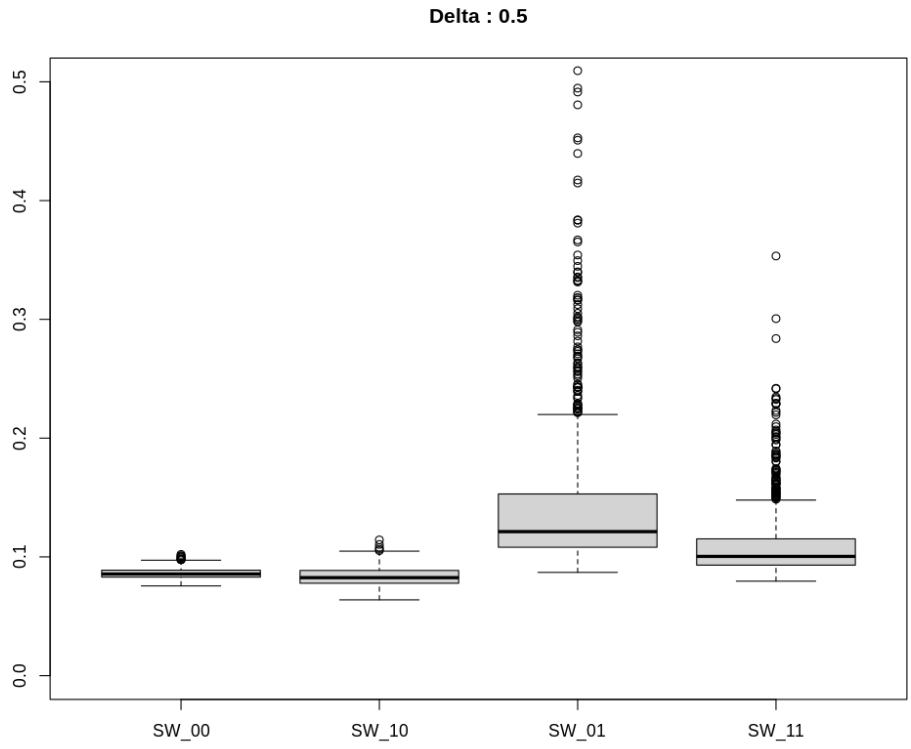

그림 A-4. Delta가 .5인 경우의 95% 신뢰구간 길이에 대한 박스플롯

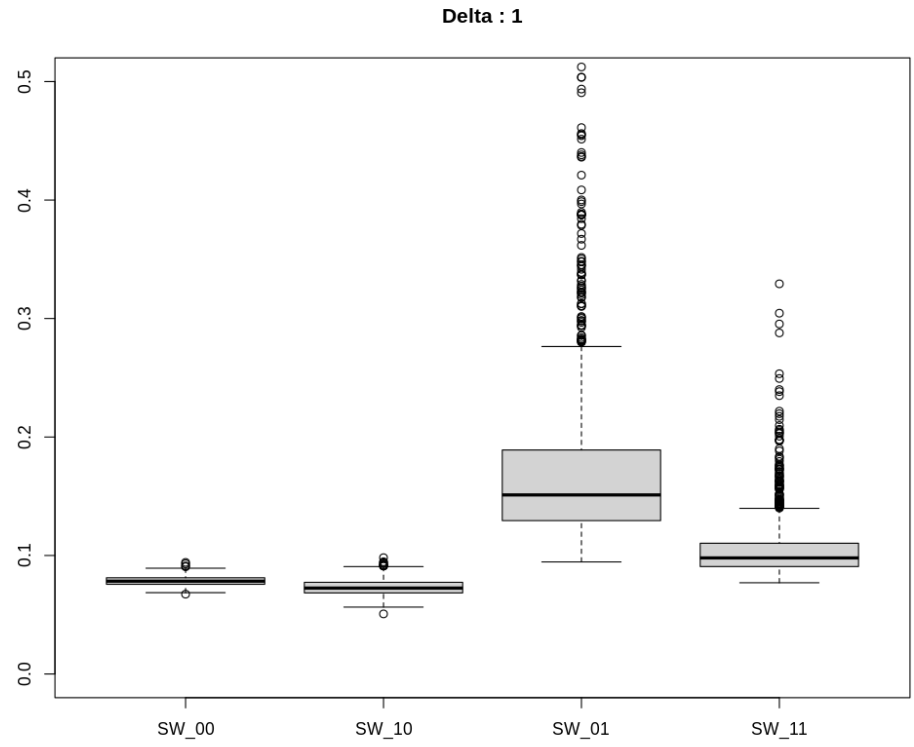

그림 A-5. Delta가 1인 경우의 95% 신뢰구간 길이에 대한 박스플롯

참고 문헌

1. *제5기(2010-2012) 원시자료 이용지침서.*
2. 정진은, *국민건강영양조사 자료의 복합표본설계효과와 통계적 추론.* 한국영양학회지, 2012. **45**(6): p. 600-612.
3. Yanovitzky, I., E. Zanutto, and R. Hornik, *Estimating causal effects of public health education campaigns using propensity score methodology.* Evaluation and program planning, 2005. **28**(2): p. 209-220.
4. Zanutto, E.L., *A comparison of propensity score and linear regression analysis of complex survey data.* Journal of data Science, 2006. **4**(1): p. 67-91.
5. DuGoff, E.H., M. Schuler, and E.A. Stuart, *Generalizing observational study results: applying propensity score methods to complex surveys.* Health services research, 2014. **49**(1): p. 284-303.
6. Ridgeway, G., et al., *Propensity score analysis with survey weighted data.* Journal of causal inference, 2015. **3**(2): p. 237-249.
7. Austin, P.C., N. Jembere, and M. Chiu, *Propensity score matching and complex surveys.* Statistical Methods in Medical Research, 2018. **27**(4): p. 1240-1257.
8. Lenis, D., et al., *It's all about balance: propensity score matching in the context of complex survey data.* Biostatistics, 2019. **20**(1): p. 147-163.
9. Dong, N., et al., *Using propensity score analysis of survey data to estimate population average treatment effects: A case study comparing different methods.* Evaluation review, 2020. **44**(1): p. 84-108.
10. Holland, P.W., *Statistics and causal inference.* Journal of the American statistical Association, 1986. **81**(396): p. 945-960.
11. Rubin, D.B., *Randomization analysis of experimental data: The Fisher randomization test comment.* Journal of the American statistical association, 1980. **75**(371): p. 591-593.
12. Rosenbaum, P.R. and D.B. Rubin, *The central role of the propensity score in observational studies for causal effects.* Biometrika, 1983. **70**(1): p. 41-55.
13. Austin, P.C., *An introduction to propensity score methods for reducing the effects of confounding in observational studies.* Multivariate behavioral research, 2011. **46**(3): p. 399-424.
14. Hirano, K. and G.W. Imbens, *Estimation of causal effects using propensity score weighting: An application to data on right heart catheterization.* Health Services and Outcomes research methodology, 2001. **2**(3): p. 259-278.
15. *국민건강영양조사 제5기 가중치산출 최종결과보고서*
16. Nattino, G., *Causal Inference in Observational Studies with Complex Design: Multiple Arms, Complex Sampling and Intervention Effects*. 2019: The Ohio State University.
17. Kim, M.-Y., et al., *Relationship of breastfeeding duration with joint pain and knee osteoarthritis in middle-aged Korean women: a cross-*

sectional study using the Korea National Health and Nutrition Examination Survey. BMC women's health, 2020. **20**(1): p. 1-10.

Abstract

Inverse Probability Weighting Methods for Estimating Causal Effects using Complex Survey Design Data

Nayoung Ha
Department of Biostatistics
The Graduate School of Public Health
Seoul National University

In the field of public health, large-scale surveys based on complex survey designs are widely conducted to be used as fundamental data to formulate health policies. This study aims to investigate how sample weights might be incorporated into Inverse Probability Weighting (IPW) methods using data from complex survey design. IPW method consists of two stages. The first stage involves to estimate the propensity score (Stage 1) and the second stage uses the propensity score to estimate the causal effect on the population (Stage 2). We attempt to identify the differences between the various approaches of precious studies in the first stage. We also try to propose considerations and recommendations when analyzing data from Korean National Health and Nutrition Examination Survey (KNHANES). There are four possible scenarios depending on whether sample weights are considered at each stage. To compare the Population Average Treatment Effect (PATE) estimates under

such scenarios, simulation studies and data analysis using KNHANES data were conducted. According to the simulation studies under the scenarios, sample weights included in Stage 2 performed well regardless of whether they were considered in Stage 1. Furthermore, it is robust to consider sample weights in both Stages 1 and 2 regardless of how much the probability of selection for the sample depends on the treatment. Treatment and sample selection are not necessarily independent since non-response corrections are considered when generating sample weights from the KNHANES data. As a result, it is recommended to use a method that incorporates the sample weights in both stages. If the treatment and selection are independent, results are similar whether or not the sample weights are considered in Stage 1. For that reason, it is suggested to conduct a sensitivity analysis.

Keywords : Complex Survey Data, Inverse Probability Weighting Methods, Population Average Treatment Effect (PATE), KNHANES

Student Number : 2021-27619