연구보고서

직업병 인과추론 가이드라인 및 통계분석법 개발 (3)

-인과추론과 복합노출 가이드라인의 활용 및 통계분석법 개발-

예신희, 이상길, 이유진, 마성원, 박성균, 김용진, 이우주

산업재해예방
안전보건공단
산업안전보건연구원

요 약 문

● **연 구 기 간** 2023년 02월 ~ 2023년 11월
● **핵 심 단 어** g-formula, Bayesian kernel machine regression, 통계분석법 개발, 가이드라인 개발
● **연구과제명** 직업병 인과추론 가이드라인 및 통계분석법 개발 (3)
-인과추론과 복합노출 가이드라인의 활용 및 통계분석법 개발-

1. 연구배경

인과추론(예: g methods)과 복합노출 건강영향 평가(예: BKMR) 각각에 초점을 맞춘 통계분석법은 이미 개발되어 있으나, 각각의 방법론은 몇 가지 제한점들을 가지고 있다. 작업환경에서의 유해물질 복합노출로 인해 발생하는 새로운 직업병을 발굴하기 위해서는 이러한 통계방법론의 제한점들을 개선하고, 국내 산업보건 역학 연구자들이 이러한 통계방법론을 쉽게 활용할 수 있게 하는 가이드라인 개발이 필수적이다.

g-formula은 반복 측정된 자료에서 발생 가능한 치료-교란 요인 되먹임의 존재를 모형에 반영하여 분석이 가능하며, 건강근로자 생존 편향과 같은 산업보건 역학 연구에서 흔히 나타날 수 있는 선택 편향을 효과적으로 통제할 수 있으며, marginal causal effect에 대응하는 인과 효과 추정치를 다양한 위험지표(risk measure)를 통해 산출이 가능하다.

예신희 등(2022)은 산업 보건 역학 연구에서 시간에 따라 변하는 복합 노출의 건강 영향을 평가할 수 있는 통계방법으로 g-formula와 BKMR을 소개함과 동시에 g-formula와 BKMR의 장점과 단점을 검토하였다. 이 연구에서 g-formula의

단점 및 제한점으로 용량-반응 곡선(dose-response relationship or curve)과 유해물질 사이의 인과적 교호 작용(causal interaction)을 시각적으로 제공해 주는 함수를 R 패키지 ‘gfoRmula’(Lin VL 등 (2019))에서 제공하지 않아 유해물질과 건강 결과 사이의 관계를 시각적, 직관적으로 확인하기 어렵다는 점이 지적되었다. g-formula로 분석한 결과가 모형의 일부 오지정(misspecification)에 의해 인과 효과 추정치가 얼마나 크게 변화하는지 등 안정성을 체계적으로 검토하는 방법이 없다는 것이 또한 큰 제한점으로 지적되었다.

g-formula과 Bayesian kernerl machine regression(BKMR)은 모두 이론적으로 노출 변수의 개수와 무관하게 적용이 가능하나 g-formula의 경우, 노출 변수 사이의 관계를 모형에 반영해야 하는 반면, BKMR은 노출 변수 사이의 상관성을 커널 행렬(kernel matrix)을 통해 모형에 반영하기 때문에 노출 변수 사이의 관계 구축의 어려움 없이 결과 변수와 유해물질에 대한 모형 적합이 가능하다. 또한, 노출 변수 사이의 교호 작용을 평가하고, 이를 시각적으로 표현하는 함수와 각 노출 변수의 사후포함확률(posterior inclusion probability; PIP)을 R 패키지 ‘bkmr’(Bobb JF 등 (2018))에서 제공하기 때문에 건강 결과와 복합 유해물질 사이의 관계에 대한 이해를 직관적으로 할 수 있다. 나아가, 모수적 모형(parametric model)을 사용하는 g-formula와 비교하여 BKMR은 복합 노출의 다양한 고차원 항(higher-order term) 또는 교호 작용 항을 반영하기 위해 커널 행렬을 이용한 혼합 효과 모형(mixed effect model)을 사용하기 때문에 복합유해물질과 건강결과 사이의 관계를 g-formula보다 유연하게 기술할 수 있다.

한편, BKMR은 반복 측정된 자료를 다룰 때 표본 수가 커짐에 따라 계산량이 매우 빠르게 증가하며, 수천-수만 정도의 대상자 수가 있는 경우 분석결과를 얻는데 2주~1달 정도의 시간이 소요되는 경우가 발생한다. 이는 표본의 수에 대응하는 차원을 가지는 커널 행렬과 마코프 체인 몬테-카를로(Markov chain Monte-Carlo; MCMC) 기반 베이지안 기법의 사용으로 표본 수에 따라 분석 시간이 급속도로 증가하기 때문이다. 또한 이분형 결과

변수에 대하여 BKMR은 프로빗(probit) 회귀 모형만 적합이 가능하기 때문에 보건, 의료 문제에 널리 사용되는 오즈 비(odds ratio)를 통해 복합 물질이 건강 결과에 미치는 영향을 해석하기 어렵다는 단점이 있다.

각 방법에 대해 위에서 언급된 제한점들이 존재하였고, 산업 보건 역학 연구자들이 g-formula와 BKMR 방법을 수월하게 산업 보건 역학 연구에 적용하기 위해서는 한계점들을 개선해야 한다. 그리고 개선된 결과를 국내 산업 안전 보건 역학 연구자들이 실제 현장에서 사용할 수 있도록 가이드라인을 제공하고자 한다.

2. 주요 연구내용

1) 인과추론 및 복합노출 국문 가이드라인 무료 배포용 책자 개발

- **합성 데이터를 활용한 g-formula 분석 가이드라인 개발**: 2021년과 2022년에 진행하였던 직업병 인과추론 가이드라인 및 통계분석법 개발(1, 2)를 요약하고, 특수건강진단 자료를 기반으로 'synthpop' R 패키지를 사용하여 실제 특수건강진단 자료와 유사한 예제용 합성 데이터를 만들었다.
- **인과추론 교과서 번역본에 대한 후속 과제 기획**: 인과추론에 대한 국내 연구진들의 이해를 높이고자, 후속 과제로 미국 하버드 대학의 Miguel A. Hernán 및 James M. Robins 교수가 발간한 인과추론 교과서 'Causal Inference: What If' 번역본을 만들어 산업안전보건연구원 홈페이지를 통해 무료배포 하고자 한다.

2) 인과추론 및 복합노출 국문 가이드라인의 활용

- **국문 가이드라인 활용 세미나 진행**: 전공의 3인과 산업위생 전문가 1인을 대상으로 특수건강진단 자료를 활용한 g-formula 분석에 대한 세미나를 진행하였다(이론 세미나 1회, g-formula 예제 분석 세미나 1회, 특수건강진단자료 데이터 클리닝 세미나 3회, g-formula 특수건강진단 자료 분석 세미나 1회).
- **질의 응답 정리**: 세미나 중 질의한 내용에 대한 응답을 정리하였다.

3) g-formula의 통계분석법 개선

- **용량-반응 곡선**: 근로자 종적 자료에서 단일 유해물질에 대한 노출의 건강 영향을 평가할 때, 노출 농도에 따른 건강 영향을 직관적으로 전달하기 위해 시각적인 그림을 제공하는 경우가 많다. 따라서 본 연구에서는 단일 유해물질의 노출 정도에 따른 건강 영향을 직관적으로 표현할 수 있게 하는 시각화 코드(DoseResponsePlot)를 개발하였다.
- **등고선 그림**: 단일 유해물질의 경우, 농도에 따른 건강 영향을 2차원 그래프로 쉽게 표현이 가능하지만 두 유해물질의 농도에 따른 건강 영향의 경우 3차원 그래프를 통해 표현해야하므로, 위의 코드를 곧바로 적용하기 어렵다. 그러한 이유로 근로자 종적 자료에서 두 개의 유해물질에 대한 복합 노출의 건강 영향을 직관적으로 표현하기 위한 시각화 코드(ContourPlot)를 개발하였다.
- **교호 작용 그림**: 2개 이상의 유해물질로 인한 복합 노출의 건강 영향을 평가할 때, 유해물질 간 교호 작용 효과(또는 시너지 효과)를 파악하여 근로자의 건강을 악화시키는 유해물질 사이의 조합을 확인할 수 있다. 그 효과는 additive interaction, multiplicative interaction 그리고 relative excess risk due to interaction (RERI)를 통해 계산이 가능하며,

계산된 additive interaction 값을 직관적으로 표현하기 위한 시각화 코드(InteractionPlot)를 개발하였다.

4) g-formula의 분석 결과의 안정성을 평가하는 방법

- **자료의 결측치**: 특수건강진단 자료와 같이 반복 측정된 자료에서 시간에 따라 변하는 노출 변수 또는 교란 변수에 결측 치가 존재하는 경우, 그 결측 치를 채우는 방법으로 last observation carried forward(LOCF)와 multiple imputation(MI) 방법을 고려할 수 있다. 본 연구는 두 방법을 적용하였을 때, 교란 변수의 결측 비율에 따라 인과효과 추정치에 발생하는 편향의 절댓값, 표준 오차 비 그리고 95% 신뢰구간의 포함률에 대해 수치적으로 조사하였다. 본 연구에서 수행한 모의실험 자료에서 결측 비율이 증가할 때, LOCF 방법과 비교하여 MI 방법이 상대적으로 작은 편향, 작은 표준 오차 그리고 높은 신뢰구간 포함률을 보였다. 특히, 결측 비율이 30% 이상인 경우, LOCF 방법에서 모두 편향 및 표준 오차 비의 급격한 증가 그리고 신뢰구간 포함률의 급격한 감소가 관찰되었다.
- **근로자의 불규칙한 특수건강진단 문제**: 업무 전환 조치 등의 적절한 사유로 자료에서 일부 근로자는 매년 특수건강진단을 받는 것으로 나타나지 않고, 불규칙적으로 검진을 받는 것으로 나타났다. 이러한 이유로 이전 과제에서는 검진 연도를 기준으로 분석을 시행한 것이 아닌 검진 순서에 따라 분석을 시행하였다. 검진 순서의 경우, 검진 순서 사이의 시간에 대한 고려가 어렵기 때문에 검진 연도를 기준으로 분석을 시행하고자 하는 경우, 이러한 근로자의 불규칙한 검진 패턴을 고려한 분석이 수행되어야 한다. 검진 받지 않은 연도에 해당하는 자료를 결측 자료로 생각하여 기존 종적 자료의 구조를 확장하고, MI 방법을 통해 결측치를 채운 후 g-formula를 적용하여 분석 결과의 안정성을 추정치의 편향의 절댓값, 표준 오차 비 그리고 95% 신뢰구간의 포함률을 사용하여 조사하였다. 불규칙한 자료의 비율이

증가할수록 g-formula가 제공하는 추정치의 편향의 절댓값과 표준 오차의 크기가 증가하였고, 95% 신뢰구간의 포함률은 감소하였다. 이러한 결과로부터 검진 받지 않은 연도에 해당하는 자료를 결측된 자료로 생각하여 종적 자료가 가지고 있는 불규칙한 자료의 구조를 규칙적인 자료의 구조로 확장하고 검진 받지 않은 연도에서 발생하는 결측치를 채워 g-formula를 적용하는 것이 한계점을 가지는 접근이라는 것을 확인할 수 있었다. 자료의 불규칙적 관측을 반영하는 새로운 방법의 개발이 필요하다고 판단하였다.

- **노출 변수 또는 교란 변수에 대한 모형 지정 검토**: g-formula는 사용되는 모든 모형(교란 변수, 노출 변수 그리고 결과 변수에 대한 모형)이 올바르게 지정(correct specification)되어야 편향 없는 추정치를 제공한다. 하지만 모형이 올바르게 지정되었는지 확인하기 위해 현재 gfoRmula R 패키지에서 제공하는 것은 자연 경과 조건에서 예측되는 노출 변수 및 교란 변수의 이력과 실제 관측 값을 비교하는 그래프이며, 각 모형이 잘 적합하였는지 그래프로 확인하는 것은 주관적이기 때문에 어려운 작업이다. 따라서 본 연구는 g-formula가 교란 변수에 대한 모형, 노출 변수에 대한 모형 그리고 결과 변수에 대한 모형 중 어느 모형에 크게 의존하는지 각 모형을 잘못 지정한 후, 얻어지는 g-formula의 인과 효과 추정치를 편향의 절댓값, 95% 신뢰구간의 포함률을 사용하여 조사하였다. 그 결과, 노출 변수에 대한 모형과 관계없이 매 시점 일정하게 개입(intervention)하는 경우, 교란 변수에 대한 모형 또는 결과 변수에 대한 모형을 잘못 지정하였을 때, 편향의 절댓값의 크기가 증가하였고, 신뢰구간의 포함률은 결과 변수 모형이 잘못 지정되었을 때 감소하는 결과가 나타났다. 또한, 결과 변수에 대한 모형뿐만 아니라 교란 변수에 대한 모형도 같이 잘못 지정되었을 경우, 그 크기는 더 감소하였다(편향에서는 그 크기가 더 증가하였음).

5) BKMR의 통계분석법 개선

- **BKMR의 분석 속도 개선**: 기존 BKMR 방법에서 제안한 변수 선택 사전 분포가 아닌 horseshoe 축소 사전 분포 및 변분 근사 알고리즘을 사용하여 사후 분포를 근사하여 계산 속도를 대폭 개선하였다.
- **반복 측정된 자료에서 기울기에 랜덤 효과 적용**: 기존 BKMR 방법에서 랜덤 절편만 허용이 가능하였지만, 성김 구조의 촐레스키 요인을 가정하여 사후 분포를 근사하고, mini-batch 확률적 경사법을 확장하여 랜덤 효과에 대한 분포의 공분산의 행렬식과 역행렬 계산에 필요한 계산량을 줄여 랜덤 기울기를 허용하는 BKMR 방법을 구축하였다.
- **BKMR의 로지스틱 회귀 모델로의 확장**: 기존 BKMR 방법의 경우, 이항 자료 (binary data)를 다루기 위해 프로빗 회귀 모델을 사용하였지만, 본 연구에서는 중요도 추출 방법을 통해 중요도 함수를 추정 및 변분 분포를 정의하고 쿨백-라이블러 발산을 기준으로 확률적 경사 알고리즘을 적용하여 주변 가능도 함수를 근사하는 확대 사후 분포를 구성하여 BKMR에서 역학 연구자 및 의료 분야 연구자들이 많이 사용하는 로지스틱 회귀 모델을 사용할 수 있도록 하였다.

6) 개선된 통계방법론에 대한 활용 가이드라인 작성

- 본 연구에서 개발한 g-formula를 이용한 용량-반응 곡선 및 교호 작용을 표현하는 시각화 코드 함수 및 BKMR의 분석 속도를 개선하고, 반복 측정된 자료에서 기울기에 랜덤 효과를 허용한 방법 나아가 BKMR에 로지스틱 모형을 적용한 방법을 산업보건 역학 연구자가 용이하게 사용할 수 있도록 함수의 사용법에 대한 활용 가이드라인을 작성하였다.

3. 연구 활용방안

- 다양한 산업 보건 역학 연구에서 알고자 하는 주된 관심사인 건강 결과와 유해물질 사이의 용량-반응 곡선 및 유해물질 사이의 교호 작용을 평가하고 이를 시각적으로 표현할 수 있는 그래프를 제공하여 g-formula로 추정한 여러 복합물질에 대한 위험을 다양한 지표 및 관점에서 평가할 수 있다. g-formula 분석 결과에 대한 안정성을 제고하는 근거를 제시함으로써 g-formula를 국내 산업 안전 보건 역학 연구의 결과를 신뢰성을 확보할 수 있다.
- 표본의 수가 큰 자료에서 수행하기 어려웠던 BKMR을 본 과제의 결과물을 통해 수행할 수 있게 되었으며, 분석 소요시간으로 인해 진행하지 못한 근로자의 반복 측정된 자료에 대한 분석 연구의 발전을 기대할 수 있다. 또한, 표본 수가 큰 빅 데이터를 통해 보다 정확한 인과 효과 추정치와 신뢰구간의 제공이 가능하다.
- 기존 BKMR에서 사용되던 프로빗 모형의 사용으로 인한 결과 해석의 어려움을 역학 연구 및 의학 연구에서 주로 사용되는 로지스틱 회귀 모델을 기반으로 한 BKMR의 개발을 통해 해소할 수 있다. 이를 통해 작업장에서 발생하는 다양하고 새로운 복합물질에 대한 노출과 건강 결과 사이의 이해를 확고히 하고, 나아가 새로운 유해물질 노출에 대한 기준 등의 정책 마련의 출발점이 될 수 있다.
- 위와 같은 평가를 통해 위험성이 높은 물질에 대해 노출 제한 용량 또는 기준을 재정비하고, 직업성 질환에 대한 누적 발생률 또는 사망률 감소시켜 근로자 개인으로서 삶의 질이 향상될 뿐만 아니라 국가적으로 근로자의 건강관리를 도울 수 있으며, 국가사업의 사회적 기여도를 높일 수 있다.

- 본 과제를 통하여 산업 보건 연구를 진행하는 연구자들이 복합노출에 대한 건강 영향 평가 분석방법인 g-formula와 BKMR을 자료에 적용하는데 필요한 시간을 단축시키며, 올바르게 사용하도록 하여 근로자의 사망 또는 건강 지표에 대한 복합노출의 효과를 추정, 산출할 수 있도록 한다.

4. 연락처

- 연구책임자: 산업안전보건연구원 중부권역학조사팀 팀장 예신희
  - ☎ 032) 510. 0754
  - E-mail shinheeye@kosha.or.kr

목 차

목 차

표 목차

그림목차

그림목차

Ⅰ. 서 론

Ⅰ. 서 론

1. 연구 목적 및 필요성

인과추론(예: g methods)과 복합노출 건강영향 평가(예: BKMR) 각각에 초점을 맞춘 통계분석법은 이미 개발되어 있으나, 각각의 방법론은 몇 가지 제한점들을 가지고 있다. 작업환경에서의 유해물질 복합노출로 인해 발생하는 새로운 직업병을 발굴하기 위해서는 이러한 통계방법론의 제한점들을 개선하고, 국내 산업보건 역학 연구자들이 이러한 통계방법론을 쉽게 활용할 수 있게 하는 가이드라인 개발이 필수적이다. 따라서, 본 연구에서는 산업보건 역학전문가와 통계전문가의 협업을 통해 다음과 같은 목적으로 연구를 진행하였다.

〈표 Ⅰ-1〉 연차 별 연구 목적

연차	연구 목적
2021년 과제: 직업병 인과추론 가이드라인 및 통계분석법 개발(1)	• 종적자료에 적용하는 인과추론 통계방법인 g-formula를 검토 • g methods 국문 가이드라인 개발
2022년 과제: 직업병 인과추론 가이드라인 및 통계분석법 개발(2)	• 인과추론 통계방법에 결합하기 적절한 복합 노출의 건강영향을 평가하는 통계방법(g-formula와 BKMR)을 검토 • 복합 노출의 건강영향 평가 국문 가이드라인 개발
2023년 과제: 직업병 인과추론 가이드라인 및 통계분석법 개발(3)	• 복합노출과 건강영향 간의 인과관계를 평가할 수 있는 통계분석법의 제한점을 개선한 통계분석법 개발

특정 단일 시점(time-fixed)에서 노출이 이루어진 관찰연구 자료로부터 인과추론을 하는 방법들은 standardization(Sjolander A(2018)), inverse probability of treatment weighting(Horvitz DG.(1952), Rosenbaum PR 등(1983), Robins JM 등(2000)), augmented inverse probability weighted estimator(Robins JM 등(1994)) 그리고 targeted minimum loss-based estimator(Van der laan MJ 등(2006), Van der laan MJ 등(2011)) 등 다양하게 개발되어 있으며, 널리 사용되고 있다. 또한, R 프로그램에서 이러한 방법들을 구현하기 위한 패키지가 잘 구현되어 있다(Sjolander A(2018), Van der wal WM 등(2011), Gruber S 등(2012), Zhong Y 등(2021), Zetterqvist J 등(2015)). 하지만 시간에 따라 값이 변하는 노출 변수(time-varying exposure)로부터 영향을 받는 교란 변수(confounder)가 존재하는 경우 표준적인 회귀분석을 기반으로 방법(traditional methods)들은 인과 효과 추정치에 편향(bias)을 발생시킨다는 것이 잘 알려져 있다(Hernan MA 등(2020)).

특히, 근로자에게 노출된 작업장 내 산재한 복합물질의 양을 시간에 따라 변하는 노출 변수로 본다면 근로자들의 고용 상태는 과거 노출 변수에 의해 영향을 받으면서 다음 기간의 노출 변수에 영향을 주는 시간에 따라 바뀌는 교란 변수의 역할을 하게 된다. 따라서 근로자의 고용 상태를 적절하게 고려하지 않고 표준적인 회귀모형만을 이용하여 분석한다면 노출 변수의 인과효과 추정치에는 편향이 발생하게 된다. 고용상태와 같이 과거 노출 변수에 의해 영향을 받고 다음 기간의 노출 변수에 영향을 주는 교란 변수를 치료-교란 변수 되먹임(treatment-confounder feedback)이라 일컫는다.

이러한 편향을 제거하기 위해 하버드대학교의 James M. Robins 교수는 시간에 따라 변하는 노출 변수가 있는 복잡한 관찰연구 자료로부터 이러한 치료-교란 변수 되먹임의 존재를 반영하여 인과효과를 추론하는 방법인 g-method (g-method에는 g-computation formula(줄여서 g-formula)(Robins JM 등(1986)), inverse probability weighting을 사용한 marginal structural model(Robins JM 등(2000)), g-estimation(Robins JM 등(1989)) 세 가지

방법이 포함됨)를 개발하였다. 이는 건강근로자 생존 편향(healthy worker survivor bias)과 같은 반복 측정된 자료를 분석하는 산업보건 연구에서 발생할 수 있는 선택 편향(selection bias) 문제를 효과적으로 다룰 수 있는 방법이다. 하지만 역학 연구자들은 g-method에 대한 개념과 기술적인 세부내용에 대한 이해 부족으로 g-method 사용에 어려움을 겪고 있다(Naimi AI 등(2017)). 국내 산업 보건 역학연구에서도 g-method 사용의 필요성은 남정모 등(2002) 및 이경무 등(2011)에서 언급한 바 있으나, 현재 국내 산업 보건 영역에서 많이 사용되고 있지 않다.

유해물질에 대한 복합 노출과 관련한 문제는 (ⅰ) 복합노출 원인 물질 각각과 근로자의 건강 지표 사이의 관계가 비선형(non-linear)이며 비가산(non-additive) 관계일 것이며, (ⅱ) 원인 물질들 사이의 교호 작용(interaction)이 건강 지표에 작용할 것이며, (ⅲ) 원인 물질 사이의 연관성이 매우 크다는 점이 있다. 하버드 대학교의 Jennifer F. Bobb, Linda Valeri 교수는 기존 kernel machine regression(KMR) 방법을 도입하여 이러한 문제들을 해결하고자 하였다(Bobb JF 등(2015)).

산업 보건 역학 연구에서 복합 물질에 대한 노출을 다루기 위해 사용하는 방법에는 Elasticnet regression(Forns J 등(2016)), Partial least square (Agier L 등(2016)), Weighted quantile sum regression(WQS)(Carrico C 등(2015)), Bayesian kernel machine regression(BKMR)(Bobb JF 등(2015)) 등이 있다. 특히 다중 노출(multiple cause)을 유연하게 다룰 수 있는 BKMR이 여러 환경역학 연구에서 주목을 받고 있다(이슬비 등(2019)).

예신희 등(2022)은 산업 보건 역학 연구에서 시간에 따라 변하는 복합 노출의 건강 영향을 평가할 수 있는 통계방법으로 g-formula와 BKMR을 소개함과 동시에 g-formula와 BKMR의 장점과 단점을 검토하였다.

g-formula은 반복 측정된 자료에서 발생 가능한 치료-교란 요인 되먹임의 존재를 모형에 반영하여 분석이 가능하며, 건강근로자 생존 편향과 같은 산업

보건 역학 연구에서 흔히 나타날 수 있는 선택 편향을 효과적으로 통제할 수 있으며, marginal causal effect에 대응하는 인과 효과 추정치를 다양한 위험 지표(risk measure)를 통해 산출이 가능하다.

반면, 용량-반응 곡선(dose-response curve)과 유해물질 사이의 인과적 교호작용(causal interaction)을 시각적으로 제공해주는 함수를 R 패키지 'gfoRmula' (Lin VL 등(2019))에서 제공하지 않아 유해물질과 건강 결과 사이의 관계를 시각적, 직관적으로 확인하기 어렵다는 점이 g-formula의 단점 및 제한점으로 지적되었다. 또한, g-formula로 분석한 결과가 모형의 일부 오지정(misspecification)에 의해 인과 효과 추정치가 얼마나 크게 변화하는지 등 안정성을 체계적으로 검토하는 방법이 없다는 것이 큰 제한점이라고 할 수 있다.

g-formula과 BKMR은 모두 이론적으로 노출 변수의 개수와 무관하게 적용이 가능하나 g-formula의 경우, 노출 변수 사이의 관계를 모형에 반영해야 하는 반면, BKMR은 노출 변수 사이의 상관성을 모형에서 커널 행렬(kernel matrix)을 통해 반영하기 때문에 노출 변수 사이의 관계 구축의 어려움 없이 결과 변수와 유해물질에 대한 모형 적합이 가능하다. 또한, 노출 변수 사이의 교호 작용을 평가하고, 이를 시각적으로 표현하는 함수와 각 노출 변수의 사후포함확률(posterior inclusion probability; PIP)을 R 패키지 'bkmr'(Bobb JF 등(2018))에서 제공하기 때문에 건강 결과와 복합 유해물질 사이의 관계에 대한 이해를 직관적으로 할 수 있다. 나아가, 모수적 모형(parametric model)을 사용하는 g-formula와 비교하여 BKMR은 복합노출의 다양한 고차원 항(higher-order term) 또는 교호 작용 항을 반영하기 위해 커널 행렬을 이용한 혼합 효과 모형(mixed effect model)을 사용하기 때문에 복합유해물질과 건강결과 사이의 관계를 g-formula보다 유연하게 기술할 수 있다.

한편, BKMR은 반복 측정된 자료를 다룰 때 표본 수의 증가에 따라 계산량이 매우 빠르게 증가하며, 수천-수만 정도의 대상자 수가 있는 경우 분석결과를 얻는데 2주~1달 정도의 시간이 소요되는 경우가 발생한다. 이는 표본의 수에 대응하는 차원을 가지는 커널 행렬과 마코프 체인 몬테-카를로(Markov chain

Monte-Carlo; MCMC) 기반 베이지안 기법의 사용으로 표본 수에 따라 분석 시간이 급속도로 증가하기 때문이다. 또한 이분형 결과 변수에 대하여 BKMR은 프로빗(probit) 회귀 모형만 적합이 가능하기 때문에 보건, 의료 문제에 널리 사용되는 오즈 비(odds ratio)를 통해 복합 물질이 건강 결과에 미치는 영향을 해석하기 어렵다는 단점이 있다.

위에서 언급된 결과와 같이 각 방법에 대해 제한점들이 존재하였고, 산업 보건 역학 연구자들이 g-formula와 BKMR 방법을 수월하게 산업 보건 역학 연구에 적용하기 위해서는 한계점들을 개선해야 한다. 그리고 개선된 결과를 국내 산업 안전 보건 역학 연구자들이 실제 현장에서 사용할 수 있도록 가이드라인을 제공하고자 한다.

2. 관련 선행 연구에 대한 분석

Lin VL 등(2019)의 연구에서 경쟁 위험 유무에 따른 생존 여부, 연속형 변수 또는 이분형 변수를 결과 변수와 시간에 따라 변하는 연속형 또는 이분형 노출 변수를 가지는 자료에서 parametric g-formula를 적합하여 분석할 수 있도록 하는 R 프로그램 패키지를 개발하였다.

Bobb JF 등(2018)의 연구에서 유해 물질 사이의 교호 작용 그리고 결과 변수와의 비선형성 용량-반응 관계를 모형에 반영하는 모형인 BKMR을 적합하여 복합 노출 자료를 분석할 수 있도록 R 프로그램 패키지를 개발하였다.

예신희 등(2022)은 반복 측정된 자료에서 복합 노출을 다룰 수 있는 방법 중 시간에 따라 변하는 결과 변수, 노출 변수 그리고 교란 변수를 다룰 수 있는 g-formula와 비모수적으로 복합 노출문제를 다룰 수 있는 BKMR의 가이드라인을 작성하고, 장점과 단점을 검토하였다.

Bobb JF 등(2015)은 연속형 결과 변수뿐만 아니라 이분형 결과 변수도 BKMR이 적용 가능하도록 R 패키지를 개발하였지만, 이분형 변수에 대해 보건,

의료 분야에서 빈도가 적게 사용되는 프로빗 회귀 모형만 개발하였다.

VanderWeele T 등(2014)은 단일 시점에서 주로 사용되는 인과적 교호 작용 지표인 additive interaction, multiplicative interaction 또는 relative excess risk due to interaction(RERI)에 대한 설명을 제공하였을 뿐만 아니라 통계적 교호 작용과 인과적 교호 작용 사이의 관계에 대해서도 기술하였다. 그 중 RERI의 경우, Sjolander A(2018)에서 standardization을 적용하여 RERI를 추정할 수 있는 R 패키지 stdReg를 개발하고 배포하였다.

Park 등(2021)의 연구는 사후 분포(posterior distribution)의 계산이 어려운 경우, 계산이 간편한 분포를 통해 사후 분포를 근사하는 변분 베이즈 방법(variational Bayes method) 중 하나인 자동 미분 변분 추론(automatic differentiation variational inference)을 사용하여 약물동태학 모형(pharmacokinetic model)에 적용하여 실제 현장에서 Monte Carlo Markov Chain(MCMC)의 대안으로 사용이 가능한지 검토하였다.

MHV Ong 등(2017)의 연구는 단일변수 비모수적 요소(univariate nonparametric component)가 가우시간 확률 과정 사전 분포(gaussian process prior distribution)의 spectral analysis에 기반한 코사인 수열로 표현된다면, 모수적 요소(parametric component)와 비모수적 요소(nonparametric component)로 구성된 준모수적 모형(semiparametric model)에 변분 베이즈 방법을 사용하여 MCMC 사용 대비 모형 적합에 소요되는 시간을 줄일 수 있음을 보였다.

3. 연구 목표

1) 인과추론과 복합노출에 대한 가이드라인 활용 및 무료 배포용 책자로 작성(자체 과제)

예신희 등(2021)은 시간에 따라 변화하는 단일 유해물질 노출로 인한 건강영향을 g-formula로 분석하는 방법에 대한 가이드라인을 작성하였다. 예신희 등(2022)은 2개 이상의 유해물질 노출로 인한 건강영향을 인과적으로 해석하기 위한 통계 가이드라인을 작성하였다.

위 가이드라인을 활용하여, 인과추론과 복합노출 통계분석 경험이 없는 전공의 2~3인과 산업위생 전문가 1인과 특수건강진단 자료를 분석해보며, 가이드라인을 수정하고, 이 때 발생하는 질의 응답을 정리하였다.

2) 현재 통계분석법(g-formula, BKMR)의 제한점을 개선한 통개분석법 개발(부분위탁 과제)

본 연구를 통해 복합물질에 대한 노출의 건강 영향 평가 방법인 g-formula를 다양한 유해물질의 농도에서 적용하고, 복합물질의 농도에 따른 건강 결과의 그래프(용량-반응 곡선)를 시각적으로 표현하는 코드를 개발, 제공하여 유해물질과 건강결과 사이의 관계를 국내 산업 보건 역학 연구자가 직관적으로 이해할 수 있게 하고자 한다.

두 가지 이상 유해물질이 복합적으로 영향을 주는 건강결과에 대하여 g-formula 방법을 통해 두 물질 사이의 인과적 교호 작용을 additive interaction, multiplicative interaction(또는 interaction on additive-, multiplicative-scale) 그리고 relative excess risk due to interaction (RERI)을 통해 평가하고, 교호 작용을 시각적으로 표현하는 코드를 개발, 제공하여

국내 산업 보건 역학 연구자가 유해물질과 건강결과 사이의 관계에 대한 이해를 용이하게 하고자 한다.

g-formula을 적용하여 분석한 결과의 안정성을 저해할 수 있는 요소(자료의 결측 치 문제, 근로자의 불규칙한 특수건강검진 문제 그리고 노출 변수 또는 교란 변수에 대한 분포 가정 문제)들을 평가하고 적절한 분석 가이드라인을 제공하고자 한다. 이러한 평가 내용을 토대로 g-formula에 대한 국내 산업 보건 역학 연구자의 이해를 증진시키고, 분석 결과의 안정성을 제고할 수 있다.

복합물질에 대한 노출의 건강 영향을 평가하는 방법인 BKMR의 분석 속도를 개선하여 표본 수가 큰 자료에도 BKMR을 적용할 수 있도록 하고자 한다. 이러한 속도 개선을 통해 일부 근로자에게만 노출되는 복합물질뿐만 아니라 다수의 근로자에게 노출되는 복합물질에 대해서도 건강 영향 평가를 수행할 수 있다.

프로빗 회귀 모형에만 적합이 가능하였던 BKMR 방법론을 로지스틱 회귀 모델로 확장하여 보건 의료 분야에서 사용되고 있는 위험 지표인 오즈 비를 사용하여 복합물질에 대한 건강 영향을 평가할 수 있다.

반복 측정된 자료에서 각 근로자에 대해 절편만 랜덤 효과를 적용할 수 있었던 것을 각 공변량에 대한 기울기까지 확장하여 개별 근로자에 대한 각 공변량의 효과를 제공할 수 있다.

Ⅱ. 연구 방법

Ⅱ. 연구 방법

1. 연구 내용

1) 인과추론과 복합노출에 대한 가이드라인 활용 및 무료 배포용 책자로 작성(자체 과제)

(1) 인과추론 및 복합노출 국문 가이드라인 무료 배포용 책자 개발

예신희 등(2021)와 예신희 등(2022)이 가이드라인을 요약하고, 특수건강진단 자료를 기반으로 'synthpop' R 패키지를 사용하여 실제 특수건강진단 자료와 유사한 예제용 합성 데이터를 만들었다.

인과추론에 대한 국내 연구진들의 이해를 높이고자, 후속 과제로 미국 하버드 대학의 Miguel A. Hernán 및 James M. Robins 교수가 발간한 인과추론 교과서 'Causal Inference: What If' 번역본을 만들어 산업안전보건연구원 홈페이지를 통해 무료배포하고자 하며, 이를 후속과제로 기획하였다.

(2) 인과추론 및 복합노출 국문 가이드라인의 활용

예신희 등(2021)와 예신희 등(2022)이 가이드라인을 활용하여, 인과추론과 복합노출 통계분석 경험이 없는 전공의 23인과 산업위생 전문가 1인과 특수건강진단 자료를 분석해보며, 가이드라인을 수정하고, 이 때 발생하는 질의응답을 정리하였다.

2) 현재 통계분석법(g-formula, BKMR)의 제한점을 개선한 통개분석법 개발(부분위탁 과제)

(1) g-formula의 통계분석법 개발

용량-반응 곡선과 교호 작용을 표현하는 시각화 코드를 개발하고자 한다. (1) 산업 보건 역학 연구자가 유해물질 별 평가하고자 하는 농도의 범위를 지정한 후, 농도의 조합 별 건강 결과에 대한 복합유해물질의 인과 효과를 추정하여 용량-반응 곡선을 꺾은선 그래프 또는 등고선 그림으로 제공하고자 한다. (2) 유해물질이 두 가지인 경우, 시간이 고정된 관찰연구 자료에서 인과적 교호 작용을 측정할 때 사용하는 지표인 additive interaction, multiplicative interaction 그리고 RERI을 반복 측정된 자료로 확장하여 계산하고, 시각적으로 표현하고자 한다.

분석결과의 안정성 평가 방법을 개발하고자 한다. g-formula를 사용하여 분석결과의 안정성을 저해하는 요소인 자료의 결측치 문제, 근로자의 불규칙한 특수건강검진 문제 그리고 노출 변수 또는 교란 변수에 대한 분포 가정 문제 각각이 발생시키는 편향 및 신뢰 구간의 coverage rate를 수치 실험을 통해 검토함으로써 g-formula로 분석한 결과의 안정성을 평가하고자 한다.

(2) BKMR의 통계분석법 개발

horseshoe 연속형 축소 사전 분포 및 변분 근사 알고리즘을 이용하여 BKMR의 분석 속도를 향상하고, 반복 측정된 자료에서 절편에만 랜덤 효과를 적용될 수 있었던 점을 기울기까지 랜덤 효과를 허용함으로써 BKMR의 성능을 개선하고자 한다.

프로빗 회귀 모형만 적용이 가능한 BKMR을 로지스틱 회귀 모델도 허용이 가능하도록 확장하여 이분형 결과 변수에 대하여 보건, 의료 문제에서 널리 사용되는 지표인 오즈 비에 대한 해석이 가능하도록 하고자 한다.

(3) 개선된 통계방법론에 대한 활용 가이드라인 작성

국내 산업 보건 역학 연구자가 역학 연구에 적용하기 용이하도록 용량-반응 곡선과 교호 작용을 표현하는 시각화 코드 및 개선된 BKMR의 사용법과 분석 결과의 안정성 평가 방법(자료의 결측치 문제, 근로자의 불규칙한 특수건강검진 문제, 노출 변수 또는 교란 변수의 분포 가정 문제)을 설명하는 가이드라인을 작성하고, 제공 및 배포하고자 한다.

2. 연구 방법

1) 인과추론 및 복합노출 국문 가이드라인 무료 배포용 책자 개발

(1) 합성 데이터를 활용한 g-formula 분석 가이드라인 개발

2021년과 2022년에 진행하였던 직업병 인과추론 가이드라인 및 통계분석법 개발(1, 2)를 요약하고, 특수건강진단 자료를 기반으로 'synthpop' R 패키지를 사용하여 실제 특수건강진단 자료와 유사한 예제용 합성 데이터를 만들었다.

(2) 인과추론 교과서 번역본에 대한 후속 과제 기획

인과추론에 대한 국내 연구진들의 이해를 높이고자, 후속 과제로 미국 하버드 대학의 Miguel A. Hernán 및 James M. Robins 교수가 발간한 인과추론 교과서 'Causal Inference: What If' 번역본을 만들어 산업안전보건연구원 홈페이지를 통해 무료배포하고자 한다.

2) 인과추론 및 복합노출 국문 가이드라인의 활용

(1) g-formula 이론 및 예제 분석 세미나

전공의 3인과 산업위생 전문가 1인을 대상으로 인과추론 및 복합노출 국문 가이드라인을 활용하여 g-formula 이론과 예제 분석에 대한 세미나를 진행한다.

(2) 가이드라인 실제 활용 시 질의 응답 정리

가이드라인 활용 시 발생하는 질의 응답을 정리한다.

3) g-formula의 통계분석법 개발

(1) 용량-반응 곡선과 교호 작용을 표현하는 시각화 코드 개발

가) 용량-반응 곡선

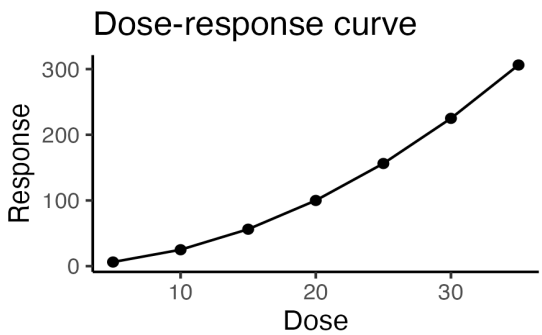
산업 보건 역학 연구자가 유해물질 별 농도의 최솟값, 최댓값 그리고 농도의 증가분(계산을 하는 간격의 크기)를 지정하여 평가하고자 하는 농도의 범위를 설정한다. 농도의 조합 별 건강 결과에 대한 복합유해물질의 인과 효과를 추정하여 용량-반응 곡선을 꺾은선 그래프(유해물질이 한 가지인 경우) 또는 등고선 그림(유해물질이 두 가지인 경우)으로 제공하고자 한다. 구체적인 예시는 아래와 같다.

① 유해물질 한 가지인 경우 용량-반응 곡선

[Step A1] 연구자가 평가하고자 하는 유해물질의 농도의 최솟값, 최댓값 그리고 농도의 증가분을 지정한다. 예를 들어, 납 농도의 최솟값(0$\mu g/ml$), 최댓값(40$\mu g/ml$) 그리고 농도의 증가분(5$\mu g/ml$)를 지정하여 평가하고자 하는 농도를 지정한다(0, 5, 10, 15, 20, 25, 30, 35, 40; 단위: $\mu g/ml$).

[Step A2] 연구 대상 집단에서 모든 근로자의 혈중 납 농도가 지정된 각 농도일 때의 건강 영향을 g-formula를 통해 추정한다.

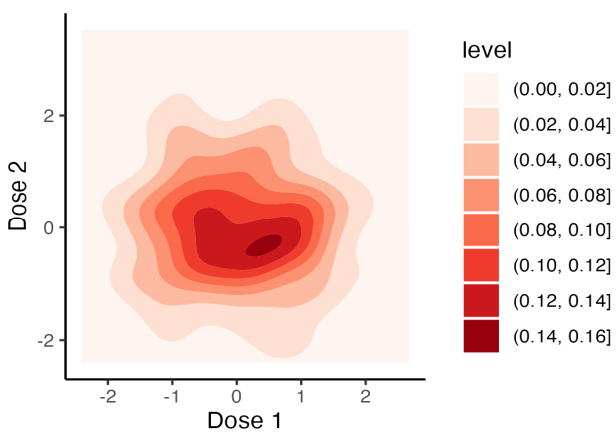
[Step A3] Step A2에서 추정된 각 농도 별 건강 영향을 그래프로 표현하면 [그림 Ⅱ-1]과 같다. 본 과제에서는 Step A1~Step A3을 포함하여 [그림 Ⅱ-1]을 표현하는 코드를 개발하고, 제공하고자 한다.



[그림 Ⅱ-1] 용량-반응 곡선의 예시 그림

② 유해물질이 두 가지인 경우 용량-반응 곡선

유해 물질이 한 가지였던 Step A1을 두 가지의 유해물질에 대해서 수행하고, Step A2에서 두 유해물질의 농도에 대한 모든 조합에 대하여 g-formula를 통해 건강 영향을 추정한다. 마지막으로 Step A2에서 추정된 각 농도 조합 별 건강 영향을 그래프로 표현하면 그림 [Ⅱ-2]와 같다.



[그림 Ⅱ-2] 등고선 그림의 예시 그림

나) 교호 작용

시간이 고정된 관찰연구 자료에서 결과 변수에 대한 이분형 노출 변수 사이의 인과적 교호 작용(causal interaction)의 크기를 측정할 때 사용하는 지표인 additive interaction, multiplicative interaction 그리고 RERI을 반복 측정된 자료 및 연속형 노출 변수로 확장하여 계산하고, 시각적으로 표현하고자 한다. 시간이 고정된 관찰연구에서 결과 변수에 대한 두 이분형 노출 변수 사이의 additive interaction 식은 아래와 같다.

Additive interaction =

$$P(Y(1,1)=1)-P(Y(1,0)=1)-P(Y(0,1)=1)+P(Y(0,0)=1)$$

위의 식에서 $Y(a,b), a, b \in \{0,1\}$은 두 이분형 노출 변수가 a, b 값일 때의 반사실적 또는 잠재적 결과 변수(counterfactual or potential outcome)를 의미한다. 이를 연속형 노출 변수 및 반복 측정된 자료로 확장하면 아래와 같이 변형이 가능하다.

Additive interaction =

$$P(Y(\bar{a},\bar{b})=1)-P(Y(\bar{a},\bar{0})=1)-P(Y(\bar{0},\bar{b})=1)+P(Y(\bar{0},\bar{0})=1)$$

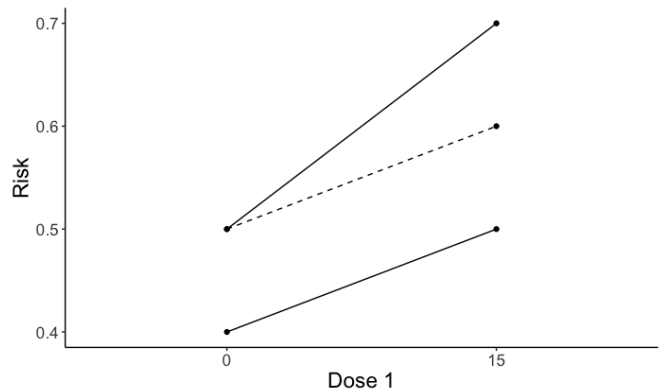
위의 식에서 $\bar{a}$은 반복 측정된 자료에서 분석기간동안 노출 변수의 값이 a의 값으로 유지되는 것을 의미하며, $\bar{0}$은 분석기간동안 노출 변수의 값이 0의 값으로 유지되는 것을 의미한다. 예를 들어, 특수건강진단 자료에서 분석기간이 8년이고, 노출 변수가 혈중 납 농도라면 $\overline{15}$은 분석기간에 해당하는 8년 동안 모든 근로자의 혈중 납 농도가 $15\mu g/ml$로 유지되는 개입(intervention)을 의미한다. 추가적으로 다른 노출 변수인 혈중 카드뮴 농도가 존재하여 잠재적 결과 변수 $Y(\overline{15},\overline{10})$은 모든 근로자가 8년동안 혈중 납 농도는 $15\mu g/ml$, 혈중 카드뮴 농도는 $10\mu g/ml$으로

유지되었을 때 나타나는 건강 결과를 의미한다. 따라서 반복 측정된 자료로 확장된 additive interaction을 계산하고, 시각화하는 과정은 아래와 같다.

[Step B1] 교호 작용이 있는지 평가하고자 하는 두 유해물질의 농도를 설정한다. 예를 들어, 납과 카드뮴 각각 15㎍/㎖, 10㎍/㎖에서 교호 작용이 있는지 평가하기 위해 납과 카드뮴의 농도를 각각 15㎍/㎖, 10㎍/㎖으로 설정한다.

[Step B2] 연구 대상 집단에서 모든 근로자에 대해 각 유해물질의 농도가 Step B1에서 지정된 농도일 때의 건강 영향을 g-formula를 통해 추정한다. 예를 들어, 연구 대상 집단에서 모든 근로자의 혈중 납 그리고 혈중 카드뮴 농도가 15㎍/㎖, 10㎍/㎖인 경우, 15㎍/㎖, 0㎍/㎖인 경우, 0㎍/㎖, 10㎍/㎖인 경우, 마지막으로 0㎍/㎖, 0㎍/㎖인 경우에서 g-formula를 통해 건강 영향을 추정한다.

[Step B3] Step B2에서 추정된 4가지 건강 영향을 토대로 RERI를 계산하고, 그 값을 시각화하면 [그림 Ⅱ-3]과 같다. 본 과제에서는 Step B1~Step B3을 포함하여 [그림 Ⅱ-3]을 표현하는 코드를 개발하고, 제공하고자 한다.



[그림 Ⅱ-3] 앞선 설명에서와 같이 모든 근로자의 혈중 납 농도가 15㎍/㎖, 혈중 카드뮴 농도가 10㎍/㎖으로 8년 동안 유지되었을 때의 두 유해물질 사이의 인과적 교호 작용을 시각적으로 보이기 위한 예시 그래프. Dose 1은 혈중 납 농도를 의미함. 위쪽 실선은 혈중 카드뮴 농도가 10㎍/㎖으로

8년 동안 유지되었을 때의 결과이고, 아래쪽 실선은 혈중 카드뮴 농도가 0$\mu g/ml$으로 8년 동안 유지되었을 때의 결과임. 점선은 아래쪽 실선을 평행이동 시켰다고 생각한 경우에 대응됨.

(2) 분석결과의 안정성을 평가하는 방법

가) 자료의 결측치 문제(missing data problem)

특수건강진단 자료와 같이 반복 측정된 자료에서 시간에 따라 변하는 노출변수 또는 교란 변수에 결측치(missing value)가 존재하는 경우, 그 결측치를 채우는 방법으로 last observation carried forward(LOCF)와 imputation 방법이 있다. LOCF 방법은 시간에 따라 변하는 변수에 대하여 t 시점에서의 결측치를 t-1 시점에서 관측된 값으로 채우는 방법을 말하며, imputation 방법은 관측된 자료에서 결측치가 없는 변수들을 사용하여 일부 결측치가 있는 변수를 예측하는 모형을 구축하고, 그 모형을 통해 예측 값을 추정하여 결측치의 값으로 채우는 방법을 의미한다. Imputation은 통계 프로그램 R에서 R 패키지 'mice'를 통해 수행할 수 있다(S Van Buuren 등(1999), S Van Buuren 등(2011)). 본 과제에서는 특수건강진단과 같은 반복 측정된 자료에 LOCF와 imputation 방법을 적용하여 결측치를 채웠을 때, 인과효과 추정치에 발생하는 편향 및 신뢰구간의 coverage rate를 조사하고, 두 방법 중 어떠한 방법이 국내 산업안전 보건 역학 연구자들에게 권장할 수 있는 방법인지 수치적으로 검토하고자 한다.

나) 근로자의 불규칙한 특수건강진단 문제(irregular visit process problem)

산업안전보건법 제 130조에 따르면 특수 직종에 근무하는 근로자는 작업장에서 쉽게 노출되는 중금속 등 유해물질로부터 건강이 위협받는지 확인 및 건강을 관리하기 위해 정기적으로 특수건강진단을 받도록 되어 있다. 하지만 업무 전환 조치 등의 적절한 사유로 자료에서 일부 근로자는 매년 특수건강진단을 받는 것으로 나타나지 않고, 불규칙적으로 검진을 받는 것으로 나타났다. 이러한 이유로

이전 과제에서는 검진 연도를 기준으로 분석을 시행한 것이 아닌 검진 순서에 따라 분석을 시행하였다. 본 과제에서는 검진 받지 않은 연도에 해당하는 자료를 결측치가 포함된 자료로 보고, 위에서 언급한 LOCF 와 imputation 방법을 적용하여 결측치를 채운 후, g-formula를 적용하여 검진 연도를 기준으로 특수건강진단 자료를 분석 및 이전 과제와의 결과를 비교하여 분석 결과의 안정성을 제공하고자 한다. 기존 자료의 구조와 확장된 자료의 구조의 차이는 아래의 예시를 통해 설명하고자 한다.

〈표 Ⅱ-1〉 기존 자료의 구조

근로자 ID	검진 연도	검진 순서	결과 변수	노출 변수
1	2012	1	0	0
1	2015	2	0	1
2	2011	1	0	0
2	2013	2	0	1
2	2016	3	1	1
3	2010	1	0	0

〈표 Ⅱ-2〉 확장된 자료의 구조
결과 변수와 노출 변수 열에서 NA는 결측치를 의미함.

근로자 ID	검진 연도	검진 순서	결과 변수	노출 변수
1	2012	1	0	0
1	2013	2	NA	NA
1	2014	3	NA	NA
1	2015	4	0	1
2	2011	1	0	0
2	2012	2	NA	NA
2	2013	3	0	1
2	2014	4	NA	NA
2	2015	5	NA	NA
2	2016	6	1	1
3	2010	1	0	0

ID가 1인 근로자는 2012년도와 2015년에 특수건강진단을 받았음〈표 Ⅱ-1〉. 하지만 이 근로자는 2013년, 2014년도에 특수건강진단을 받지 않았기 때문에 확장된 자료에서는 결과 변수와 노출 변수가 NA로 존재한다〈표 Ⅱ-2〉. 본 과제에서는 이러한 결과 변수와 노출 변수에 대하여 LOCF와 imputation를 적용하여 결측치를 채운 후, g-formula를 적용하고자 한다.

다) 노출 변수 또는 교란 변수에 대한 분포 가정(distributional assumption for exposure and confounder variables) 검토

반복 측정된 자료를 분석하기 위해 g-formula를 적용할 때, 결과 변수에 대한 모형뿐만 아니라 시간에 따라 변하는 노출 변수 및 교란 변수에 대해서도 모형을 구축하여야 하며, g-formula는 사용되는 모든 모형(결과 변수에 대한 모형과 노출 변수 및 교란 변수에 대한 모형)이 올바르게 지정(correct specification)되어야 편향이 없는 추정치를 제공한다. 하지만 현재 R 패키지 'gfoRmula'에서 제공하는 것은 자연 경과 조건에서 예측되는 노출 변수 및 교란 변수의 이력과 실제 관측 값을 비교하는 그래프로, 이를 통해 올바른 지정을 확인하기에는 한계가 있다. 본 과제에서는 g-formula를 통해 획득한 인과 효과 추정치가 결과 변수에 대한 모형, 노출 변수에 대한 모형 그리고 교란 변수에 대한 모형 중 어느 모형에 크게 의존하는가를 살펴보기 위해 각 모형을 잘못 지정하여 얻어지는 인과 효과 추정치의 편향 및 신뢰 구간의 포함률을 검토하고자 하였다.

4) BKMR의 통계분석법 개발

(1) BKMR 분석시간 단축 및 반복 측정된 자료에서 기울기에 랜덤 효과 적용

가) Horseshoe 연속형 축소 사전 분포(continuous shrinkage prior distribution)를 활용한 BKMR

본 과제에서는 BKMR에서 커널의 초모수(hyper-parameter)인 length-scale 모수에 대해 horseshoe 축소 사전 분포를 사용하는 BKMR을 제안하고, 자료의 크기가 큰 상황에서도 빠르게 사후 분포(posterior)를 계산할 수 있는 변분 근사 알고리즘을 개발하고자 한다.

분석 시간이 단축된 BKMR: 본 과제에서 개발하고자 하는 BKMR을 설명하기 위해 i번째 관측치의 결과 변수를 y_i , 반응 변수와 선형적인 관계에 있는 p차원의 설명 변수를 $x_i = (x_{i1}, \dots, x_{ip})^\top$, 결과 변수와 비선형 관계에 있는 M차원의 설명 변수를 $z_i = (z_{i1}, \cdots, z_{iM})^\top$로 표현하였다. 본 과제에서는 정규 분포를 따르는 오차를 갖는 다음의 BKMR을 다루고자 한다.

$$y_i = x_i^\top \beta + f(z_i) + \epsilon_i, \epsilon_i \sim N(0, \sigma^2), i = 1, \dots, n, \tag{1}$$

$$f = (f_1, \dots, f_n)^\top \sim N(0, \tau_f^2 K), K = (K_{jk})_{j,k=1}^n, \tag{2}$$
$$K_{jk} = \exp\left\{-\sum_{m=1}^{M} \rho_m (z_{jm} - z_{km})^2\right\}, \rho_m > 0.$$

여기서 $\beta = (\beta_1, \cdots, \beta_p)^\top$은 p차원의 회귀계수이고, $\sigma^2 > 0$은 분산을 나타낸다. 베이지안 추론(Bayesian inference)을 위해 본 연구에서는 먼저 회귀계수와 분산에 대해 가장 널리 사용되는 정규 분포와 역감마 분포(inverse-gamma distribution)

$$\beta \sim N(0, \tau_\beta^2 I_p), \ \sigma^2 \sim IGa(a_\sigma, b_\sigma),$$

를 사전 분포로 활용하고, 커널의 모수 length-scale에 대해서는 변수 선택의 효과를 주기 위해 아래와 같이 연속형 축소 사전 분포인 horseshoe 사전 분포를 사용하고자 한다. 이때, 분산 모수에 대해서는 역감마 분포를 가정한다.

$$\rho_m|\lambda_m,\tau_\rho \sim N^+(0,\lambda_m\tau_\rho^2),\ \lambda_m \sim C^+(0,1),\ \tau_\rho \sim C^+(0,1), \quad (3)$$
$$\tau_f^2 \sim IGa(a_f,b_f),$$

여기서 $N^+(\mu,s^2)$는 평균이 μ, 분산이 s^2인 양의 실수 공간에서 정의된 정규 분포를 나타내고, $C^+(0,1)$은 위치 모수가 0, 척도 모수가 1인 양의 실수 공간에서 코시 분포(Cauchy distribution)를 나타낸다.

제안된 사전 분포는 기존에 Bobb 등(2015)에 제안된 BKMR의 변수 선택 사전 분포(variable selection prior distribution) $\rho_m \sim (1-\omega_m)\delta_0 + \omega_m Unif(0,\infty)$에 비해 사후 분포 계산에 있어서 매우 효율적인 모형이 될 것이다.

나) 변분 근사 알고리즘(variational algorithm for approximate bayesian inference)

위에서 제안한 모형의 사후 분포를 계산하기 위해 먼저, 모형 (1)에서 가우시안 확률 과정을 가정한 비선형 함수 $f(\cdot)$을 적분하는 것이 필요하다. 그 결과로 모수 $\theta = (\beta, \log\rho_1, \ldots, \log\rho_M, \log\lambda_1, \ldots, \lambda_M, \log\tau_f^2, \log\sigma^2)^\top$에 대해 다음과 같이 주변 가능도 함수(marginal likelihood)가 주어지고

$$\log p(y|\theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_e| - \frac{1}{2}(y-X\beta)^\top \Sigma_e^{-1}(y-X\beta),$$

변분 하한(variational lower bound)은 아래와 같이 계산된다.

$$L(\phi) = \int q_\phi(\theta)\log\frac{p(y|\theta)p(\theta)}{q_\phi(\theta)}d\theta,$$

여기서 $q_\phi(\cdot)$은 변분 분포(variational distribution)로 다변량 정규 분포(multivariate normal distribution) 혹은 서로 독립인 정규 분포의 곱으로 가정한다. 즉, $q_\phi(\theta) = MVN(\mu, T^{-\top}T^{-1})$, $\phi = (\mu, T)$이다. 그 다음 단계로, 위의 변분 하한은 최대사후추정량(Maximum a posteriori, MAP) 계산에 제안된 mini-batch 확률적 경사법(stochastic gradient method)을 확장하여 변분 모수 ϕ을 최적화함(Chen등(2022)). 여기서 mini-batch를 사용하는 가장 큰 이유는 표본의 크기가 크면 $n \times n$차원의 공분산 Σ_e에 대한 행렬식과 역행렬에 대한 계산양이 많기 때문에 이를 효율적으로 계산하기 위함이다. 그리기 위해 mini-batch방법을 사용할 때, 자료의 부분집합은 전체 자료에서 균등하게 추출하는 uniform sampling 방법 혹은 서로 가까운 표본을 동시에 추출하는 nearby sampling 방법을 고려할 예정이다.

랜덤 효과를 고려할 수 있는 모형: 위의 (1)에서 제안된 모형을 다음과 같이 임의효과를 고려하는 모형으로 확장하고자 한다.

$$y_{ij} = x_{ij}^\top \beta + f_i(z_{ij}) + u_{ij}^\top b_i + \epsilon_{ij}, i = 1, \dots, n, j = 1, \dots, n_j, \qquad (4)$$

여기서 u_{ij}은 q차원의 설명 변수이고, b_i은 정규 분포를 따르는 랜덤 효과를 나타낸다.

$$b_i \sim MVN_q(0, \Sigma_b), i = 1, \dots, n.$$

랜덤 효과를 포함하는 모형 (4)의 사후 분포를 계산하기 위해 본 과제에서는 아래와 같은 성김 구조의 촐레스키(Cholesky) 요인 혹은 대각행렬로 가정하고,

$$T = \begin{pmatrix} T_{11} & 0 & \cdots & 0 & 0 \\ 0 & T_{22} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & T_{NN} & 0 \\ T_{N+1,1} & T_{N+1,2} & \cdots & T_{N+1,N} & T_{N+1,N+1} \end{pmatrix}$$

사후 분포 근사를 실시하고자 한다. 이때, 위에서 제안된 mini-batch 확률적 경사법을 랜덤 효과(지역적 모수)와 전역적 모수에 각각 적용하고, 복잡한 모형의 안정적인 최적화를 위해 피셔의 정보량(Fisher information)을 이용하는 natural gradient 방법을 고려하고자 한다. 참고로, 변분 하한의 natural gradient는 다음과 같이 주어진다.

$$\widetilde{\nabla}_{\phi} L = I_F(\phi)^{-1} \nabla_{\phi} L,\ I_F(\phi) = E_q\left[\{\nabla_{\phi} \log q_{\phi}(\theta)\}\{\nabla_{\phi} \log q_{\phi}(\theta)\}^{\top}\right]$$

(2) 로지스틱 회귀 모델로 확장하여 분석

본 과제에서는 기존 BKMR을 로지스틱 회귀 모델에서 사용 가능하도록 BKMR을 확장하고(이하 로지스틱 BKMR), 확장된 회귀 모델에서의 효율적인 사후 분포 계산방법에 대해 연구를 진행하고자 한다.

가) 확률적 경사 변분 근사 알고리즘

BKMR에서 이항 분포를 가능도 함수로 하는 로지스틱 회귀 모델로 확장에서 가장 어려운 점은 다음과 같이 주변 가능도 함수를 알려진 형태로 계산할 수 없다는 점이다.

$$p(y|\theta) = \int \left\{ \prod_{i=1}^{n} p(y_i|f_i, \theta) \right\} p(f|\theta) df,$$
$$f = (f_1, \ldots, f_n)^{\top} \sim MVN(0, \sigma_f^2 K).$$

따라서 본 과제에서는 이러한 문제를 해결하기 위해 먼저 중요도 추출(importance sampling) 방법을 활용하여 다음과 같이 불편 추정량(unbiased estimator)을 계산한다.

$$\hat{p}_N(y|\theta) = \frac{1}{N}\sum_{j=1}^{N} w(f^{(j)},\theta),$$
$$w(f^{(j)},\theta) = \frac{p(y|f^{(j)},\theta)p(f^{(j)}|\theta)}{h(f^{(j)}|y,\theta)},\ f^{(j)} \sim h(f|y,\theta),$$

여기서, $h(\cdot)$은 f에 대한 중요도 밀도 함수(importance density)를 나타내고, N은 중요도 밀도함수로부터 추출된 표본(particle)의 개수다. 그 다음 단계에서는 추정된 중요도 함수를 이용하여, 다음과 같이 주변분포가 근사하고자 하는 사후분포 $\pi(\theta|y)$가 되도록 확대 사후 분포(augmented posterior)를 구성한다.

$$\pi_N(\theta,z|y) = \frac{p(\theta)p(y|\theta)e^z g_N(z|\theta)}{p(y)} = \pi(\theta|y)e^z g_N(z|\theta), \qquad (5)$$

위 식에서 $z = \log\hat{p}_N(y|\theta) - \log p(y|\theta)$이고, $g_N(z|\theta)$은 z의 밀도함수다. 그리고 마지막 단계에서는 변분 분포를 $q_{\phi,N}(\theta,z) = q_\phi(\theta)g_N(z|\theta)$로 정의하고, 아래에 주어진 쿨백-라이블러 발산(Kullback-Leibler divergence)을 기준으로 하여 확률적 경사 알고리즘을 바탕으로 확대 사후 분포를 근사한다.

$$KL(\phi) = \int q_\phi(\theta)g_N(z|\theta)\log\frac{q_\phi(\theta)q_N(z|\theta)}{\pi_N(\theta,z)}dzd\theta.$$

다음은 연구를 통해 제안하는 확률적 경사 알고리즘을 나타낸다.

① $\phi^{(0)}$ 초기 값 설정

② $\phi^{(t+1)} = \phi^{(t)} - \alpha_t\widehat{\nabla}_\phi KL(\phi^{(t)}),\ t = 0,1,\ldots,$

위의 알고리즘 ②에서 α_t은 학습률(learning rate)을 나타내고, $\widehat{\nabla}_\phi KL(\phi^{(t)})$은 쿨백-라이블러 발산의 기울기(gradient)에 대한 불편 추정량을 나타냄. 참고로, 쿨백-라이블러 발산의 기울기는 아래와 같이 주어지고,

$$\nabla_\phi KL(\phi) = E_{\theta \sim q_\phi(\theta),\, z \sim g_N(z|\theta)}\big(\nabla_\phi[\log q_\phi(\theta)]\big(\log q_\phi(\theta) - \log p(\theta)\hat{p}_N(y|\theta, z)\big)\big),$$

실제 적용에 있어서는 다음과 같이 변량 조절(control variate)방법을 적용하여, 기울기에 대한 몬테-카를로 추정량(Monte-Carlo estimator)의 분산을 줄인다.

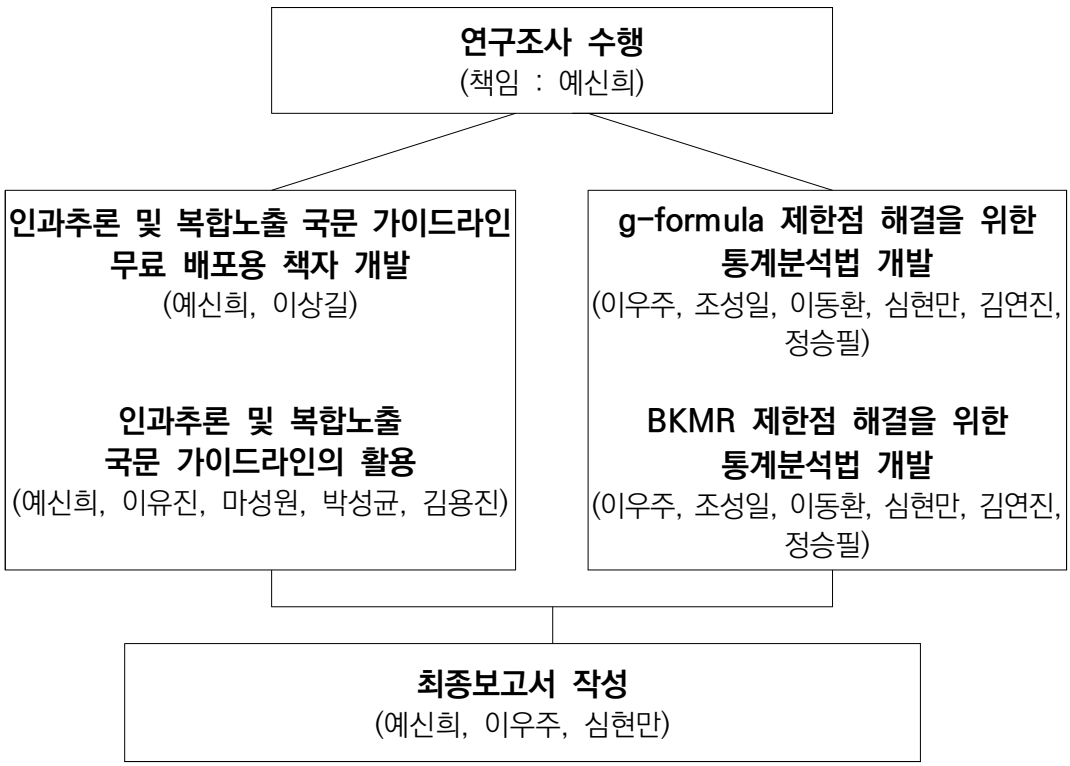
$$\widehat{\nabla_\phi KL}(\phi) = \frac{1}{S}\sum_{s=1}^{S}\nabla_\phi[\log q_\phi(\theta_s)](\log q_\phi(\theta_s) - h(\widehat{\theta_s}, z_s) - c),$$
$$c = \frac{Cov\big(\nabla_\phi[\log q_\phi(\theta)]\big(\log q_\phi(\theta) - \hat{h}(\theta, z)\big),\ \nabla_\phi[\log q_\phi(\theta)]\big)}{V\big(\nabla_\phi[\log q_\phi(\theta)]\big)}.$$

참고로, 위의 식에서 변량 조절 변수 c은 이전 단계에서 추출된 사후표본을 바탕으로 몬테-카를로 방법을 활용하여 계산한다.

5) 개선된 통계방법론에 대한 활용 가이드라인 작성

지금까지 언급한 기존의 g-formula의 한계점으로 지적된 용량-반응 곡선과 교호 작용을 표현하는 시각화 코드의 결여, g-formula 분석 결과의 안정성(자료의 결측치 문제, 근로자의 불규칙한 특수건강진단 문제, 노출 변수 또는 교란 변수에 대한 분포 가정 검토 그리고 g-formula와 같은 양을 구하는 서로 다른 두 통계 방법론) 평가 방법, 개선된 BKMR(분석 시간 단축, 로지스틱 BKMR로의 확장 그리고 기울기에 랜덤 효과를 적용)을 국내 산업 안전 보건 역학 연구자가 역학 연구에 용이하게 사용할 수 있도록 활용 가이드라인을 작성하고 제공, 배포하고자 한다.

3. 연구 추진 체계



4. 연구 윤리

본 조사를 위하여 2023년 산업안전보건연구원 기관생명윤리위원회의 심의(institutional review board, IRB)를 통과하였다(승인번호: OSHRI-202303-HR-010).

Ⅲ. 연구 결과

Ⅲ. 연구 결과

1. 인과추론 및 복합노출 국문 가이드라인 무료 배포용 책자 개발

1) 합성 데이터를 활용한 g-formula 분석 가이드라인 개발

2021년과 2022년에 진행하였던 직업병 인과추론 가이드라인 및 통계분석법 개발(1, 2)를 요약하고, 특수건강진단 자료를 기반으로 'synthpop' R 패키지를 사용하여 실제 특수건강진단 자료와 유사한 예제용 합성 데이터를 만들었다.

실제 특수건강진단 자료와 유사한 예제용 합성 데이터를 활용한 가이드라인은 부록 1에 작성하였고, 합성 데이터는 본 연구의 최종보고서와 함께 산업안전보건연구원 웹사이트에 업로드할 예정이다.

2) 인과추론 교과서 번역본에 대한 후속 과제 기획

인과추론에 대한 국내 연구진들의 이해를 높이고자, 후속 과제로 미국 하버드 대학의 Miguel A. Hernán 및 James M. Robins 교수가 발간한 인과추론 교과서'Causal Inference: What If' 번역본을 만들어 산업안전보건연구원 홈페이지를 통해 무료배포 하고자 한다. 구체적인 후속 과제 계획은 다음과 같다.

예신희 등(2021-2023)은 복합노출과 인과추론에 활용하는 통계방법인 g-formula와 BKMR의 국문 가이드라인을 개발하고, 기존 g-formula와 BKMR의 제한점을 개선한 통계 방법(R package 'vbayesGP'등)을 개발하였다. 또한, 국문 가이드라인을 활용하여 집단 역학조사를 수행하고 있고, 외부강의(2022년 한국보건정보통계학회 연수교육 등)를 7회 이상 진행하였으며, 연구결과를 바탕으로 SCI급 논문을 9건 이상 작성하고 있다.

다만 예신희 등(2021-2023)은 작업환경에서 유해물질 장기 복합노출로 인해 발생하는 새로운 직업병을 발굴하기 위해서는 g-formula와 같은 인과추론 통계방법을 적용하는 것이 적절하다고 결론 내렸는데, 연구결과의 신뢰성을 높이기 위해서는 g-formula 외에도 Marginal Structural Model(MSM), g-estimation, 그리고 Targeted Maximum Likelihood Estimation(TMLE)과 같은 통계방법을 추가로 활용하여 결과의 재현성을 확인해야 한다. 따라서, 국내 산업보건 역학 연구자들이 이러한 방법론을 쉽게 활용할 수 있도록 국문 가이드라인을 개발하고, 필요시 기존 방법론의 제한점 개선을 위한 통계 방법을 개발하고자 한다.

아울러, 미국 하버드 대학의 Miguel A. Hernán 및 James M. Robins 교수가 발간한 인과추론 교과서(Causal Inference: What If)를 한글로 번역하여 인과추론에 대한 국내 연구진들의 이해를 높이고자 한다.

따라서, 후속 과제서는 산업보건 역학전문가와 통계전문가의 협업을 통해 아래와 같은 목표로 3년간 연구를 수행하고자 한다.

〈표 Ⅲ-1〉 후속 과제의 연구 목표

년도	2024년		2025년		2026년	
1) 복합노출 인과추론 가이드라인 개발	주제: MSM		주제: g-estimation		주제: TMLE	
	예제 데이터 기반 단일노출 MSM 분석가이드라인 작성	복합노출분석으로 확장하여 수정 및 보완	예제 데이터 기반 단일노출 g-estimation 분석 가이드라인 작성	복합노출분석으로 확장하여 수정 및 보완	예제 데이터 기반 단일노출 TMLE 분석 가이드라인 작성	복합노출분석으로 확장하여 수정 및 보완
	MSM 장단점 평가		g-estimation 장단점 평가		TMLE 장단점 평가	
2) 인과추론 교과서 번역본 개발	교과서 1/2 해석	해석 감수	교과서 1/2해석	해석 감수	온라인 배포	g-methods 한계점 개선을 위한 후속 연구 기획

2. 인과추론 및 복합노출 국문 가이드라인의 활용

전공의 3인과 산업위생 전문가 1인을 대상으로 특수건강진단 자료를 활용한 g-formula 분석에 대한 세미나를 진행하였다(이론 세미나 1회, g-formula 예제 분석 세미나 1회, 특수건강진단자료 데이터 클리닝 세미나 3회, g-formula 특수건강진단 자료 분석 세미나 1회).

g-formula 분석에 대한 세미나 중 질의한 내용에 대한 응답은 다음과 같으며, 실제 특수건강진단 자료와 유사한 예제용 합성 데이터를 활용한 g-formula 가이드라인인 부록 2에 질의 응답 내용을 추가하였다.

질문 1: 연구 대상자의 수가 10,000명보다 큰 경우, nsimul은 기본 값으로 연구 대상자의 수로 설정되어 있는데, 더 줄이면 안 되나요?

답변: gfoRmula R package는 sample 함수를 사용하여 자료를 생성할 연구 대상자를 선정하기 때문에 연구 대상자보다 작은 값을 nsimul의 값으로 설정하게 되면 일부 분석 대상자만 선정하는 것과 동일합니다. 분석 시간을 단축하기 위한 목적이라면 nsimul을 작게 설정하시는 것보다 병렬 계산(parallel computing)을 권장드립니다(관련 인수는 parallel, ncores임).
병렬계산과 관련된 가이드라인의 내용은 아래와 같습니다.
gformula 함수는 계산 량이 많이 요구되는 몬테카를로 시뮬레이션과 붓스트랩을 모두 사용하기 때문에 결과를 제공하기까지 오랜 시간이 소요됩니다. 하지만 이러한 문제는 병렬 계산을 통하여 다소 해결할 수 있으며, 병렬계산을 사용할 수 있도록 gformula 함수는 parallel과 ncores를 제공합니다. parallel의 값을 TRUE로 설정하고, 이때 사용할 CPU core의 개수를 ncores에 입력하면 됩니다. 다만 ncores를 입력하는 경우에는 gformula 함수 이전에 아래의 예시와 같은 준비 코드가 필요합니다.

```
ncores <- parallel::detectCores() - 1
gformula(…,
  parallel = TRUE,
  ncores = ncores
)
```

첫 번째 줄에서 -1를 하는 이유는 데스크탑 또는 노트북의 CPU가 gformula 함수 외에 다른 프로그램을 처리할 수 있도록 여유 CPU를 설정하려는 목적으로 입력한 임의의 숫자이며, 연구자의 데스크탑이 보유하고 있는 CPU를 모두 gformula 함수를 처리하는데 사용하려는 연구자는 -1가 아닌 0을 또는 여유분을 더 남기려는 연구자는 -3, -4 등의 숫자를 사용하시면 됩니다.

질문 2: g-formula에서 보정 변수들 간의 교호작용 항(interaction term)을 가정하는 법과 노출 변수 간의 교호작용을 평가하는 방법이 있을까요?

답변: 모델에서 보정 변수들 사이의 교호작용을 포함하고 싶으시다면 모델을 기술할 때, 예를 들어 L1 * L2와 같이 입력하시면 됩니다. 다만 이때, 모형에 L1 * L2와 같이 입력한 교호작용은 통계적 교호작용(statistical interaction)으로 인과적 교호작용(causal interaction)과는 다른 개념이라는 점을 주의하셔야 합니다. 노출 변수들 간의 인과적 교호작용을 보고자 하실 때에는 가이드라인에서 기술한 것과 같이 각 노출 변수의 값을 바꿔가며 인과 효과 추정치를 산출한 후, 그림을 그려서 등고선 그림과 같이 시각적으로 확인하는 방법도 있고, additive interaction 또는 multiplicative interaction을 구하여 수치적으로 확인하는 방법도 있습니다.

질문 3: 예제에 있는 as.factor(t0)는 꼭 포함해야하는 걸까요?

답변: t0은 연구 등록으로부터의 시간이기 때문에 t0은 시간과 선형적인 관계가 있습니다. 이러한 관계를 모형에 포함된 시간에 따라 변하는 변수가

이미 가지고 있다면 t0를 모형에 포함하는 것은 시간에 대해 중복 보정하는 결과를 낳게 됩니다. 그러므로 시간에 선형적인 관계를 가지는 변수가 모형에 없을 경우에는 포함하는 것이 모형의 구축에 도움을 줄 수 있으나 이미 시간과 선형적인 관계를 포함하는 변수가 모형에 포함되어 있을 경우에는 포함하지 않는 것이 바람직하다고 생각됩니다. 예를 들어, 나이(age) 변수가 시간에 따라 증가하도록 코딩이 되어 있을 경우, 나이는 시간에 선형적으로 증가하기 때문에 t0을 모형에 포함하는 것은 나이를 중복 보정하는 것으로 볼 수 있습니다(예를 들어, 1년 단위의 시간을 가지는 자료의 경우, 1번 근로자에 대한 2012년, 2013년, 2014년 자료의 나이 변수의 값이 20, 21, 22와 같이 코딩되어 있을 경우). 나이를 연구 등록 시점의 나이로 코딩을 하셨다면 t0를 포함하셔도 될 것으로 생각되고(위의 예제를 기준으로 1번 근로자의 나이는 20으로 시간에 따라 변하지 않는 값으로 코딩되어 있을 경우), 나이를 시간에 따라 증가하도록 코딩을 하셨을 경우에는 포함하지 않는 것이 바람직하다고 생각됩니다.

질문 4: gfoRmula R package 실습 시 시간에 따라 변하는 변수의 형태를 지정하여 주는데, 연속형 변수의 경우, 분석 전 분포를 확인해야 하나요?

답변: 잘못된 분포를 지정하여 생기는 g-formula 추정치의 편향을 줄일 수 있기때문에 분석 전 분포를 확인하는 것을 권장드립니다.

질문 5: 연속형 변수에서 한쪽으로 치우친 skewed data의 경우, log-transformation이 필요한가요?

답변: 로그 변환하여 정규 분포에 가까워진다면 변환 후 변수의 분포를 정규 분포로 지정하신 후, g-formula를 적용하여 분석하시면 됩니다.
노출 변수를 로그 변환한 경우 interventions에서도 아래와 같이 로그 변환한 노출 값을 지정해주면 됩니다.

```
interventions <- list(list(c(static, rep(log(1.6), 7))),
                      list(c(static, rep(log(30), 7))))
```

질문 6: gfoRmula R package에서 reference 변경하는 법이 가이드 라인에 나와 있지 않은데 어떻게 바꿀 수 있나요?

답변: gfoRmula R package의 gformula 함수는 intervention으로 natural course를 기본적으로 제공하고 있기 때문에 어떠한 intervention을 수행하더라도 gformula 함수에서 natural course가 0번 intervention으로 나타나며, 기본 reference로 사용하고 있습니다(아래의 코드 기준으로 ref_int=0). 따라서 g-formula R package에서 두 가지 intervention(never treat: 1, always treat: 2)이 적용된 아래의 예시 코드에서 reference를 natural course에서 never treat으로 변경하시려면 <u>ref_int=1</u>을 코드에 추가하여 주시면 됩니다.

```
gform_basic <- gformula_survival(
obs_data = basicdata_nocomp,
id = id,
time_points = time_points,
time_name = time_name,
covnames = covnames,
outcome_name = outcome_name,
covtypes = covtypes,
covparams = covparams,
ymodel = ymodel,
intvars = intvars,
interventions = interventions,
int_descript = int_descript,
histories = histories,
histvars = histvars,
```

```
basecovs        = c('L3'),
nsimul          = nsimul,
ref_int         = 1,
seed            = 1234
)
```

질문 7: gfoRmula R package에서 covtypes를 사용하여 내생교란요인의 분포를 지정할 때, ordinal 분포를 사용할 수 있나요?

답변: 현재 gfoRmula R package에서 교란 변수의 형태로서 ordinal은 미리 만들어 둔 함수의 형태로 지원하고 있지 않습니다. 다만 ordinal 형태를 반영하고 싶으신 경우에는 VGAM 등의 패키지에서 제공하는 함수를 활용하여 gformula 함수에 입력 가능한 형태로 교란 변수에 적합할 함수를 생성하고, 교란 변수의 타입을 custom으로 지정한 뒤, g-formula를 사용하시면 됩니다.

질문 8: gfoRmula R package를 돌리다보면 에러가 너무 많이 뜹니다. 이러한 경우에는 어떻게 해야 할까요 ?

답변: 여러 연구자들이 이 패키지를 사용하면서 발생하는 에러들에 대해 gfoRmula R package를 만든 저자가 운영하는 깃허브 사이트(https://github.com/CausalInference/gfoRmula/issues) 에 질문을 올리면 저자가 직접 답변을 해주고 있습니다. 생성된 에러 중 깃허브 사이트에 있는 에러의 경우, 해결이 가능하며, 이외의 에러의 경우, 구글(google) 또는 ChatGPT를 이용하여 일부 해소가 가능할 것입니다.

3. g-formula의 통계분석법 개발

1) 용량-반응 곡선과 교호 작용을 표현하는 시각화 코드 개발

g-formula는 종적 자료에서 치료-교란 변수 되먹임의 구조를 반영하여 근로자가 근무하는 작업장에서 발생하는 유해물질에 대한 복합 노출의 건강 영향을 추정하기 위해 사용되는 인과추론 통계 방법론 중 하나이며, 또한 2개 이상의 복합노출로 인한 건강 영향을 평가하기 위해 사용될 수 있다.

유해물질에 대한 복합노출의 건강 영향을 평가할 때, 그 영향을 직관적으로 살펴보기 위해 결과를 다양하게 시각화할 수 있으며, 그러한 시각화 그림 중에서 단일 유해물질의 경우, 용량-반응 곡선, 두 가지 유해물질의 경우, 등고선 그림이 많이 사용된다. 본 연구에서는 종적 자료에서 g-formula를 적용하여 용량-반응 곡선과 등고선 그림을 그릴 수 있도록 시각화 코드를 개발하여 제공하였다. 또한, 개발한 함수들(DoseResponsePlot, ContourPlot 그리고 InteractionPlot)은 g-formula의 적합 결과를 그대로 입력이 가능하다. 또한, 개발한 함수를 산업보건학 연구자가 수월하게 사용 및 결과를 재현할 수 있게 하기 위해 R 패키지 'gfoRmula'에 내장된 자료 basicdata_nocomp를 사용하였다.

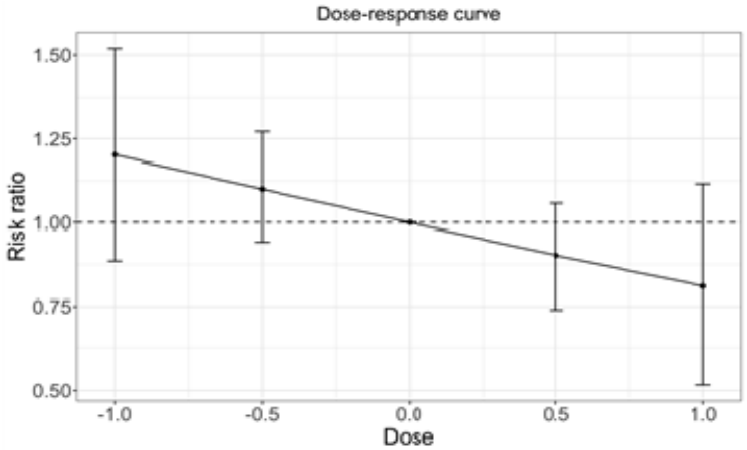
용량-반응 곡선에 대하여 먼저 설명하고자 한다. 개발한 함수 DoseResponsePlot를 통해 용량-반응 곡선을 그리기 위해서는 먼저 g-formula를 적합하여야 하며, 적합한 결과를 개발한 함수의 object 인수에 입력해야 한다.

```
DoseResponsePlot(
  object = gformRes_DoseResponse,
  Dose = (-2:2) * 0.5,
  xlab = "Dose",
  ylab = "Risk ratio",
  main = "Dose-response curve",
```

```
width = 0.03,
pointsize = 2,
lab_title_size = 17,
lab_text_size = 15,
main_size = 15)
```

개발한 함수 DoseResponsePlot을 통해 그림을 그리기 위해서는 object 인수를 제외하고 8개의 인수(Dose, xlab, ylab, main, width, lab_title_size, lab_text_size, main_size)가 추가적으로 필요하다. 용량-반응 곡선에서 개입의 범위를 Dose 인수를 통해 지정이 가능하다. xlab, ylab 그리고 main은 용량-반응 곡선에서 사용할 x-축, y-축 그리고 그림 제목을 지정할 수 있는 인수이며, lab_title_size, lab_text_size, main_size를 통해 글자 크기를 지정할 수 있다. width 인수는 신뢰구간의 상한과 하한을 표현하는 막대의 길이를, pointsize로 점의 크기를 조정할 수 있다.

아래의 그림은 basicdata_nocomp 자료에 개발한 DoseResponsePlot 함수를 적용하여 용량-반응 곡선을 시각화한 그림이다. 그림을 그리기 위해 사용한 자료와 함수의 사용법은 부록 1에서 보다 자세히 설명하고자 한다.



[그림 Ⅲ-1] 변수 'L2'의 값의 변화에 따른 용량-반응 곡선

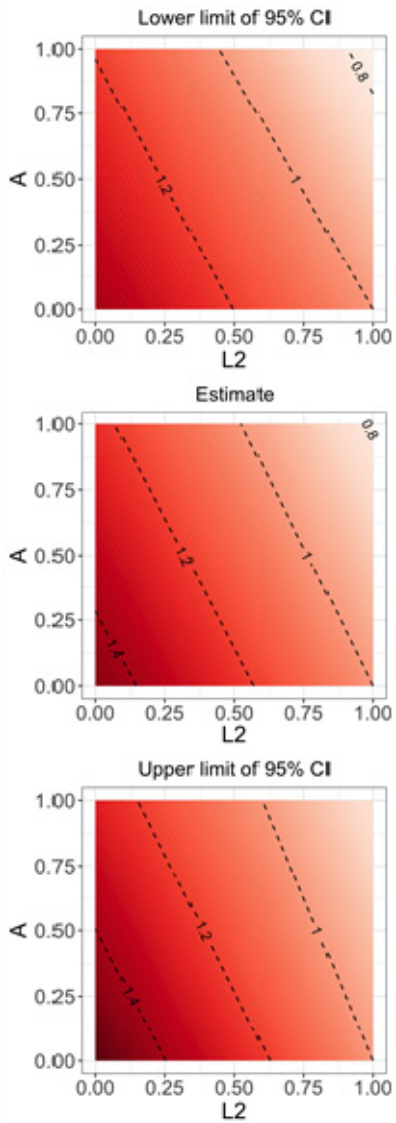
다음으로 등고선 그림에 대하여 설명하고자 한다. 등고선 그림을 그리는 함수 Contourplot 또한 용량-반응 곡선을 그리는 함수 DoseResponsePlot와 사용법이 비슷하나, 등고선 그림의 경우 개입의 대상이 되는 변수가 2개이기 때문에 지정해야할 변수가 1개 더 추가되어야 한다는 점에서 차이가 있다. 다음은 등고선 그림을 그려주는 개발한 함수 ContourPlot이다.

```
Dose1 <- Dose2 <- (-2:2) * 0.5

ContourPlot(
  object = gformRes_Contour,
  Dose = list(Dose1, Dose2),
  xlab = "L2",
  ylab = "A",
  vertical = FALSE,
  lab_title_size = 17,
  lab_text_size = 15,
  main_size = 15,
  textNum = 100,
  textwidth = 0.2)
```

개발한 함수 ContourPlot은 그림을 그리기 위해 추가적으로 9개의 인수 (Dose, xlab, ylab, vertical, lab_title_size, lab_text_size, main_size, textNum, textwidth)가 필요하다. Dose 인수를 통해 등고선 그림을 그리고자 하는 변수의 범위를 설정할 수 있으며, list의 형태로 입력되어야 한다. xlab과 ylab 인수를 통해 등고선 그림의 x-축과 y-축에 개입의 대상이 되는 변수의 이름을 입력할 수 있다. vertical 인수를 통해 등고선 그림을 세로로 표현할지 정할 수 있으며, 등고선 그림에서의 x-축, y-축 그리고 제목의 글자 크기는 lab_title_size, lab_text_size, main_size 인수를 통해 조정할 수 있다. textNum 인수를 통해 결과 값의 변화에 따른 색 변화의 정도를 조절할 수 있으며,

textwidth 인수는 등고선 그림 위 결과 값을 어느 정도의 단위로 표시하고자 하는지 결정하는 인수이다.

아래의 그림은 basicdata_nocomp 자료를 일부 수정한 후, 개발한 함수 ContourPlot 함수를 적용하여 그린 등고선 그림이다. 그림을 그리기 위해 사용한 자료와 함수의 사용법은 부록 1에서 보다 자세히 설명하고자 한다.



[그림 Ⅲ-2] 변수 'L2'과 변수 'A'의 농도에 따른 등고선 그림

산업보건 역학연구에서 2개 이상의 복합노출에 대한 건강 영향을 평가할 때, 유해물질 간 건강 영향에 대해 교호 작용(또는 시너지 효과)이 있는지 확인하는 것 또한 주된 관심사다. 교호 작용은 additive interaction인 RERI를 통해 계산이 가능하며, 그 효과를 직관적으로 확인하기 위해서는 종적 자료에 g-formula를 적용하여 얻어지는 결과를 시각화할 수 있는 코드가 필요하다. 본 연구에서는 용량-반응 곡선과 등고선 그림뿐만 아니라 교호 작용의 크기를 시각화할 수 있는 시각화 코드 또한 개발하여 제공하였다.

교호 작용의 크기를 표현하는 함수 InteractionPlot은 다음과 같다.

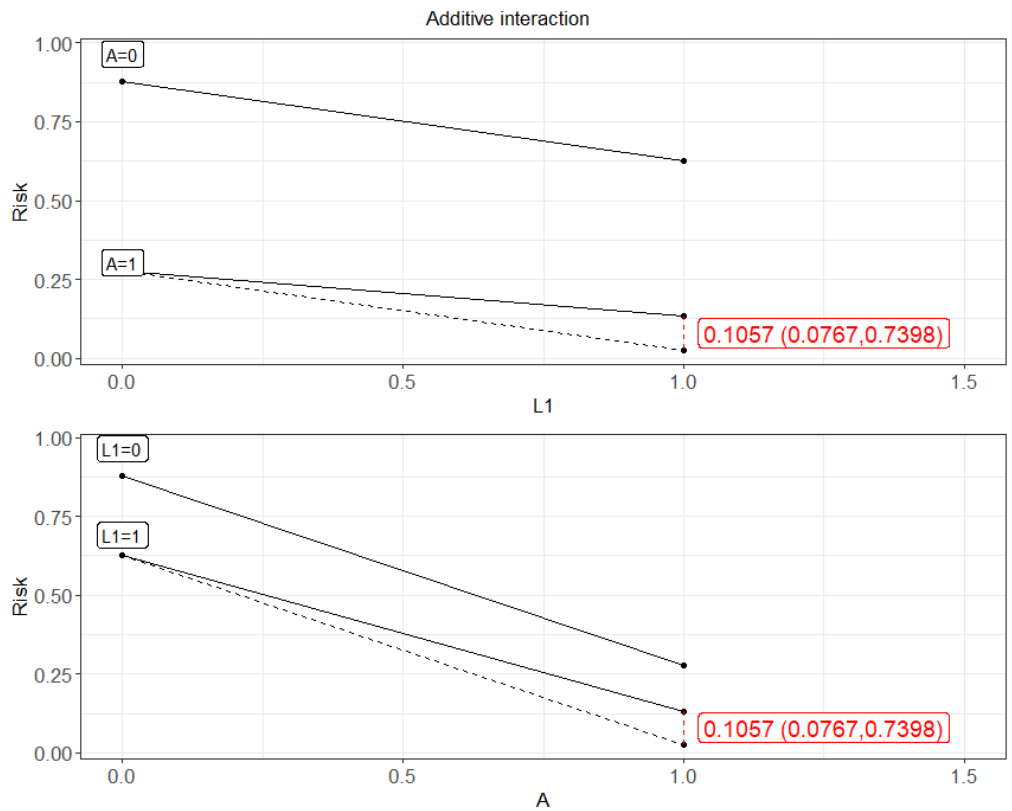
```
Dose3 <- Dose4 <- 0:1

InteractionPlot(
  obs_data = basicdata_nocomp,
  id = "id",
  time_points = timepoints3,
  time_name = "t0",
  covnames = c('L1', 'L2', 'A'),
  covtypes = c('binary', 'bounded normal', 'binary'),
  covparams = list(covmodels = c(L1 ~ lag1_A + lag_cumavg1_L1 + lag_
cumavg1_L2 + + L3 + t0,
L2 ~ lag1_A + L1 + lag_cumavg1_L1 + lag_cumavg1_L2 + L3 + t0,
A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 + lag_cumavg1_L2 + L3 + t0)),
  histories = c(lagged, lagavg),
  histvars = list(c('A', 'L1', 'L2'), c('L1', 'L2')),
  basecovs = "L3",
  outcome_name = "Y",
  outcome_type = "survival",
  ymodel  = Y ~ A + L1 + L2 + L3 + lag1_A + lag1_L1 + lag1_L2 + t0,
```

```
intvars = intvars3,
int_descript = int_descript3,
interventions = interventions3,
ref_int = 1,
baselags = TRUE,
alpha = 0.05,
nboots = 10,
seed= 12345678,
Dose = list(Dose3,Dose4),
xlab = c("L1","A"),
vertical = TRUE,
main_size = 13,
lab_title_size = 13,
lab_text_size = 13,
label_size = 5)
```

개발한 함수 InteractionPlot를 통해 교호 작용의 크기를 산출하고, 그림을 그릴 때, g-formula를 적합하기 위해 필요한 인수들이 동일하게 사용되며, 추가적으로 DoseCombination, alpha, nboots 인수를 필수적으로 입력하여야 한다. DoseCombination은 교호 작용의 크기를 구하기 위해 어떤 값이 변수의 개입 값으로 사용되었는지 입력하는 인수이며, data.frame의 형태로 입력이 되어야 하며, alpha와 nboots는 신뢰구간을 추정할 때 사용되는 인수다. 그림을 꾸미기 위해서는 다음의 인수들(xlab, vertical, main_size, lab_title_size, lab_text_size, label_size)을 지정하는 것이 필요하다. xlab은 교호 작용 그림에서 x-축에 표시할 변수 명을 의미한다. 교호작용의 경우 관심 있는 변수가 2개이기 때문에 길이가 2인 벡터가 입력되어야 한다. vertical은 각 변수에 대한 교호 작용 그림의 배치 형태를 지정할 수 있다(vertical = TRUE는 세로 형태).

lab_title_size, lab_text_size, label_size을 통해 축 제목의 크기, 축 글자 크기 그리고 교호작용의 크기를 표현한 글자의 크기를 조절할 수 있다.

아래의 그림은 basicdata_nocomp 자료에 InteractionPlot 함수를 적용하여 교호 작용의 크기를 시각화한 그림이며, 추정치와 95% 신뢰 구간을 그림 내 붉은 글씨로 표현하였다, 또한, InteractionPlot 함수는 multiplicative interaction과 RERI에 대한 추정치 및 신뢰구간을 제공한다. 그림을 그리기 위해 사용한 자료와 함수의 사용법은 부록 1에서 보다 자세히 설명하고자 한다.



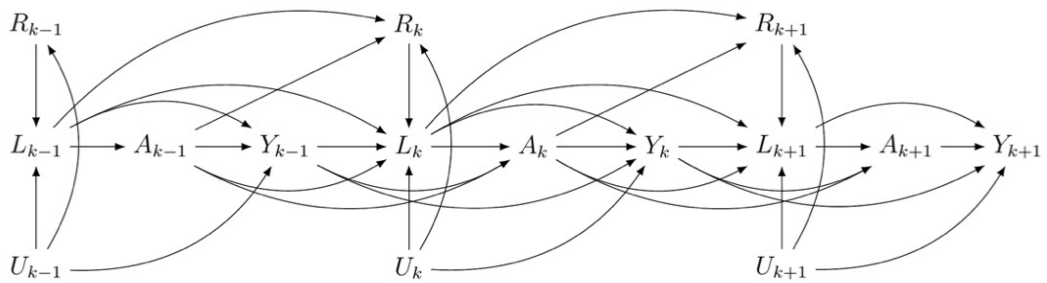
[그림 Ⅲ-3] 변수 'L1'을 기준으로 그린 교호 작용 그래프(왼쪽) 그리고 변수 'A'를 기준으로 그린 교호 작용 그래프(오른쪽)

2) 분석 결과의 안정성을 평가하는 방법

(1) 자료의 결측치 문제

특수건강진단 자료와 같은 종적 자료에서 시간에 따라 변하는 노출 변수 또는 교란 변수에 존재하는 결측치를 채우는 방식에 따라 추정치에 발생하는 편향의 크기가 달라질 수 있다. 따라서 본 연구는 교란 변수의 결측 비율에 따라 모의실험 자료를 생성하고, LOCF와 imputation 방법으로 결측치를 채운 후 얻어지는 추정치를 편향의 절댓값, 95% 신뢰구간의 참 값 포함률(coverage rate of 95% confidence interval) 그리고 평균 제곱근 오차(root mean sqaure error)를 통해 살펴보고자 한다. Imputation 방법으로 결측치를 채울 때는 R에서 제공하는 mice package의 mice 함수를 기본 값으로 사용하였다.

아래의 그림은 모의실험 자료를 생성하기 위해 사용한 방향성 비순환 그래프(directed acyclic graphs, DAGs)다.



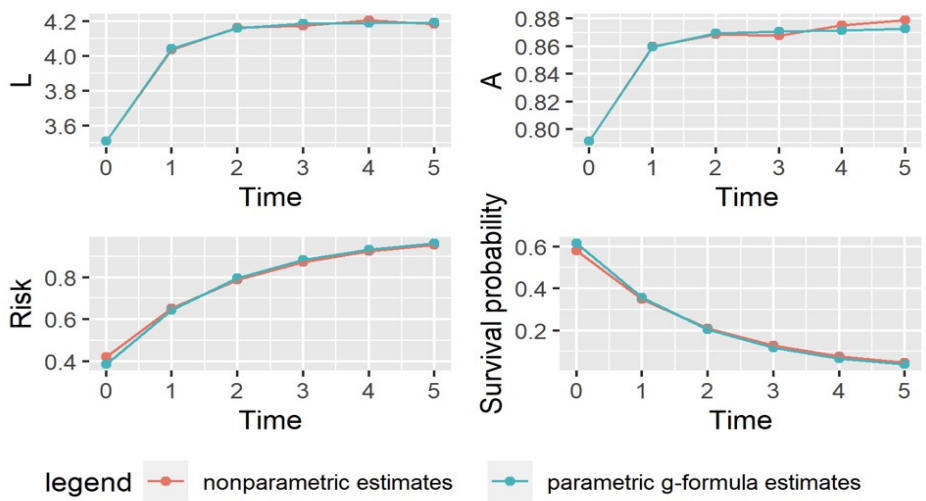
[그림 Ⅲ-4] 모의실험 자료의 생성에 사용된 방향성 비순환 그래프

위 방향성 비순환 그래프를 기반으로 하여 표본 수는 1,000, 최대 추적 관찰기간은 5, 측정되지 않은 연속형 교란 변수 U, 측정된 연속형 교란 변수 L, 이항 노출 변수 A, 생존 변수 Y 그리고 결측 여부 R로 이루어진 모의실험 자료를 500개 생성하였고, 결측 여부 R이 0, 즉 R=0인 관측치의 경우, 교란 변수 L의 값에 NA를 입력하여 결측치를 생성하였다. 본 모의 실험에서 인과 효과 측도(causal effect measure) 중 하나인 위험 비(risk ratio)를 목표

모수(target parameter)로 사용하였으며, 참 값(0.9539)은 표본 수가 100,000인 자료를 기반으로 하여 모든 대상자가 A=0일 때의 위험 대비 모든 대상자가 A=1일 때의 위험의 값으로 계산하였다. 또한, 95% 신뢰구간은 percentile bootstrap 방법을 통해 계산되었으며, bootstrap은 500번 시행되었다.

모의실험 자료에서 측정되지 않은 연속형 교란 변수 U는 시간에 관계없이 평균이 3, 분산이 1인 정규 분포(normal distribution)에서, 측정된 연속형 교란 변수 L은 이전 시점의 교란 변수 L, 이전 시점의 노출 변수 A 그리고 같은 시점의 교란 변수 U을 통해 산출된 평균, 분산이 1인 정규 분포에서, 이항 노출 변수 A는 이전 시점의 노출 변수 A 그리고 같은 시점의 교란 변수 L을 통해 산출된 확률 값을 가지는 베르누이 분포(Bernoulli distribution)에서, 결측 여부 R은 이전 시점의 교란 변수 L과 노출 변수 A 그리고 같은 시점의 교란 변수 U을 통해 산출된 확률 값을 가지는 베르누이 분포에서 그리고 생존 변수 Y는 같은 시점의 이항 노출 변수 A, 교란 변수 L 그리고 교란 변수 U를 통해 산출된 확률 값을 가지는 베르누이 분포에서 생성되었다.

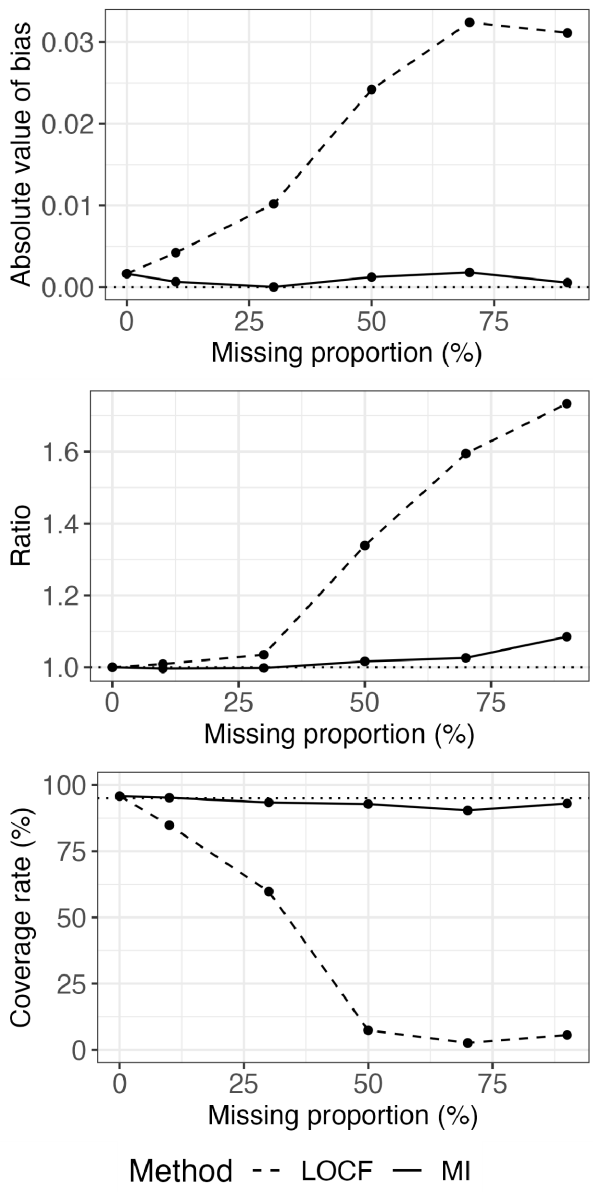
아래의 그림은 참 값을 구하기 위해 사용된 자료에 g-formula를 적용한 후 구축한 모형이 자료를 충분히 설명하고 있는지 확인한 그래프다.



[그림 Ⅲ-5] 참 값을 구하기 위해 사용된 자료에 대한 모형 적합 진단 그래프

분홍 선은 자료를 통해 그린 선이며, 초록 선은 적합한 모형을 통해 자료를 재생성하여 그린 선이다. 이 두 선이 비슷할수록 g-formula에 사용된 모형이 각 변수를 실제 자료에 근접하게 생성하고 있다는 것을 의미한다. 이 그래프를 통해 참 값을 구하기 위해 사용된 모형들이 실제 자료를 충분히 설명하고 있다는 것을 알 수 있다.

아래의 그림은 LOCF와 imputation 방법을 통해 산출한 추정치의 편향의 절댓값(상단), 표준 오차 비(중간) 그리고 95% 신뢰구간의 포함률(하단)을 자료의 결측 비율(0, 10, 30, 50, 70, 90%)에 따라 표현한 그래프임(표준 오차 비는 결측 비율이 0%일 때의 표준 오차 대비 각 결측 비율에서의 표준 오차로 정의하였음). [그림 Ⅲ-6]의 상단 그림에서 결측 비율이 높아지더라도 편향의 절댓값이 0 근처에 있던 imputation 방법과는 달리 LOCF 방법의 경우 자료 내 결측 비율이 높아질수록 추정치의 편향의 절댓값이 증가하는 것을 확인할 수 있다. [그림 Ⅲ-6]의 중간 그림에서 결측 비율이 증가함에 따라 두 방법 모두 결측 비율이 0%일 때의 표준 오차와 비교하여 표준 오차가 증가하는 경향을 보였으며, 특히 결측 비율이 30%보다 커질 때 그 변화 폭이 커지는 것이 관찰되었다. 마지막으로 [그림 Ⅲ-6]의 하단 그림에서 결측 비율이 증가할수록 LOCF 방법의 경우, 포함률이 급격히 감소하는 것을 확인할 수 있으며, 결측 비율이 30%에서 50%로 증가할 때, 더 큰 감소 폭을 보였다. Imputation 방법의 경우, 결측 비율이 증가함에 따라 포함률이 소폭 감소하는 것이 관찰되었다.

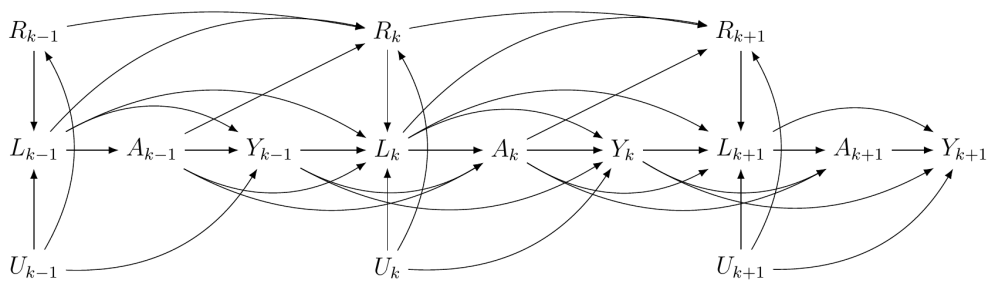


[그림 Ⅲ-6] 자료의 결측 비율에 따른 LOCF 방법과 imputation 방법으로 산출한 추정치의 편향의 절댓값(상단), 표준 오차 비(중간) 그리고 95% 신뢰구간의 포함률(하단) 그래프. 상단과 중간 그림에서 점선은 y=0 그리고 y=1인 직선을 의미하며, 하단 그림에서 점선은 y=95(%)인 직선을 나타냄. 세 그림에서 파선은 LOCF 방법, 실선은 imputation 방법을 나타냄.

(2) 근로자의 불규칙한 특수건강진단 문제

특수건강진단 자료와 같은 종적 자료에서 불규칙적으로 검진을 받는 일부 근로자들로 인하여 종적 자료 내 공백이 발생한다. 이러한 공백으로 인한 분석의 어려움을 검진 순서에 따라 분석을 수행하는 것으로 대체하는 경우도 있지만, 이러한 분석은 검진 순서 사이의 간격을 고려하기 어렵다. 따라서 본 연구에서는 검진 받지 않은 연도의 자료를 결측 자료로 생각하여 주어진 자료의 구조를 확장하고 결측치를 imputation 방법으로 채운 후 g-formula를 적용하였다. 또한, 모의실험을 통해 검진 받지 않은 연도에 해당하는 자료의 비율에 따라 g-formula가 제공하는 추정치의 성능(편향의 절댓값, 표준 오차 비 그리고 95% 신뢰구간의 포함률)을 조사하여 분석 결과의 안정성을 살펴보고자 한다(이때, 검진 받지 않은 연도에 해당하는 자료의 비율이란 처음 검진을 받은 연도부터 가장 마지막 검진을 받은 연도까지의 기간 중 검진을 받지 않은 연도의 비율을 의미한다. 예를 들어, 한 근로자가 2011년(첫 검진), 2013년, 2014년 그리고 2020년(마지막 검진)에 검진을 받았다면 이 근로자의 검진을 받지 않은 연도의 비율은 0.6 이다.

아래의 그림은 모의실험 자료를 생성하기 위해 사용한 방향성 비순환 그래프다.



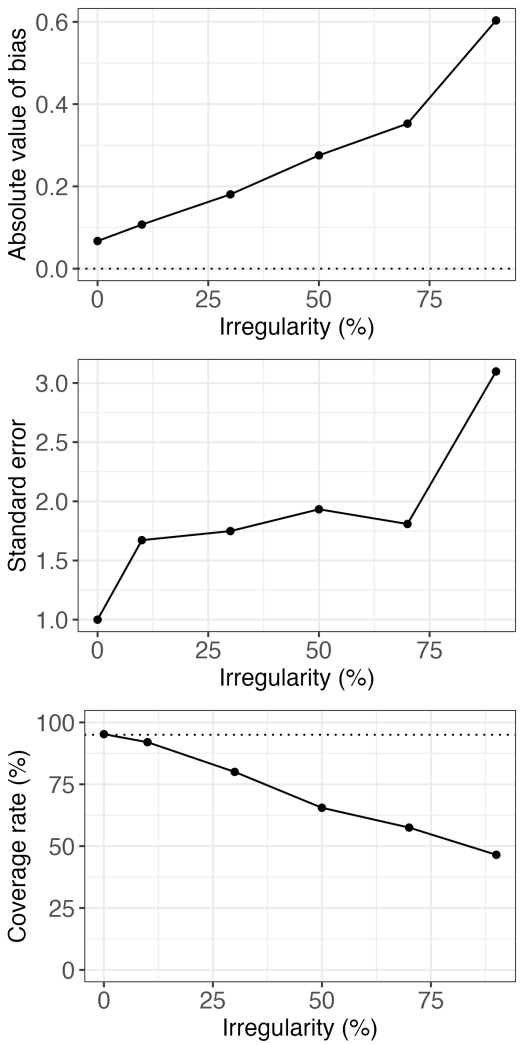
[그림 Ⅲ-7] 모의실험 자료의 생성에 사용된 방향성 비순환 그래프

위 방향성 비순환 그래프를 기반으로 하여 표본 수는 1,000, 최대 추적 관찰기간은 5, 측정되지 않은 연속형 교란 변수 U, 측정된 연속형 교란 변수 L, 이항 노출 변수 A, 생존 변수 Y 그리고 결측 여부 R로 이루어진 모의실험 자료를 500개 생성하였고, 결측 여부 R이 0, 즉 R=0인 관측치의 경우, 교란 변수 L, 노출 변수 A 그리고 생존 변수 Y의 값에 NA를 입력하여 결측 자료를 생성하였다. 본 모의 실험에서 위험 비를 목표 모수로 사용하였으며, 참값(0.5570)은 표본 수가 100,000인 자료를 기반으로 하여 모든 대상자가 A=0일 때의 위험 대비 모든 대상자가 A=1일 때의 위험의 값으로 계산하였다. 또한, 95% 신뢰구간은 percentile bootstrap 방법을 통해 계산되었으며, bootstrap은 500번 시행되었다.

모의실험 자료에서 측정되지 않은 연속형 교란 변수 U는 시간에 관계없이 평균이 3, 분산이 1인 정규 분포에서, 측정된 연속형 교란 변수 L은 이전 시점의 교란 변수 L, 이전 시점의 노출 변수 A 그리고 같은 시점의 교란 변수 U를 통해 산출된 평균 그리고 분산이 1인 정규 분포에서, 이항 노출 변수 A는 이전 시점의 노출 변수 A 그리고 같은 시점의 교란 변수 L을 통해 산출된 확률 값을 가지는 베르누이 분포에서, 결측 여부 R은 이전 시점의 결측 여부 R, 교란 변수 L과 노출 변수 A 그리고 같은 시점의 교란 변수 U을 통해 산출된 확률 값을 가지는 베르누이 분포에서 그리고 생존 변수 Y는 같은 시점의 이항 노출 변수 A, 교란 변수 L 그리고 교란 변수 U를 통해 산출된 확률 값을 가지는 베르누이 분포에서 생성되었다.

아래의 그림은 불규칙한 자료의 비율(0, 10, 30, 50, 70, 90%)에 따라 추정치의 편향의 절댓값(상단), 표준 오차 비(중간) 그리고 95% 신뢰구간의 포함률(하단)을 표현한 그래프다. [그림 Ⅲ-8]의 상단 그림에서 불규칙한 자료의 비율이 높아짐에 따라 추정치의 편향의 절댓값이 증가하는 것을 확인할 수 있다. [그림 Ⅲ-8]의 중간 그림을 보면 불규칙한 자료의 비율이 10%부터 표준 오차 비가 1.6보다 커지는 것을 확인할 수 있으며, 이는 곧 신뢰구간이 불규칙한 자료의 비율 0%일 때의 신뢰구간과 비교하여 불규칙한 자료의 비율이 10%일 때의

신뢰구간의 길이가 1.6배 되었음을 의미한다. 마지막으로 [그림 Ⅲ-8]의 하단 그림에서 불규칙한 자료의 비율이 증가할수록 95% 신뢰구간의 포함률이 감소하는 것을 확인할 수 있다.



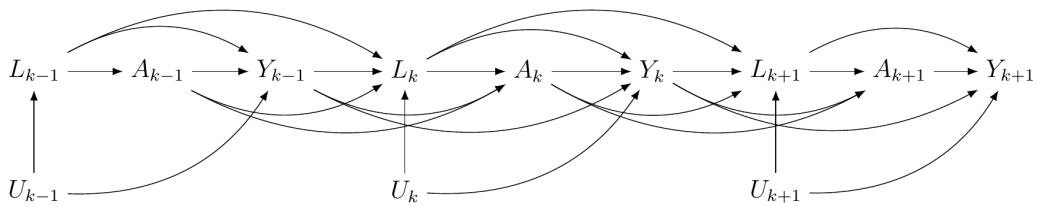
[그림 Ⅲ-8] 불규칙한 자료의 비율에 따른 g-formula가 제공하는 추정치의 편향의 절댓값(상단), 표준 오차(중간) 그리고 95% 신뢰구간의 포함률(하단) 그래프. 상단과 하단 그림에서 점선은 y=0인 직선을 의미하며, 중간 그림에서 점선은 y=95(%)인 직선을 나타냄.

(3) 노출 변수 또는 교란 변수에 대한 모형 지정 검토

g-formula는 시간에 따라 변하는 노출 변수, 교란 변수 그리고 결과 변수에 대한 모형을 모두 구축하여야 하며, 특히 구축된 모형이 모두 올바르게 지정되어야 편향이 없는 추정치를 제공한다. 실제 자료 분석 과정에서 연구자가 g-formula를 적합하기 위해 구축한 모형 중 어떠한 모형이 잘못 지정되었는지 확인하는 것은 쉽지 않은 과제다. 현재 gfoRmula package에서는 적합한 g-formula로 추정한 자연 경과(natural course)에서의 시간에 따라 변하는 변수들의 추세가 실제 자료에서 나타나는 시간에 따라 변하는 변수들의 추세와 일치하는지 비교하는 그림을 제공하고 있다. 하지만 이러한 주관적 비교를 통해 모형이 적합한지 판단하는 것은 어려운 작업이다.

그림이 아닌 수치적으로는 모의실험을 통해 모형이 잘못 지정되었을 때 g-formula 추정치의 성능을 검토하여 추정치에 대한 각 모형의 의존도를 살펴봄으로써 실제 자료 분석에서 어떠한 모형을 면밀히 검토해야 하는지 간접적으로 확인할 수 있다. 따라서 본 연구에서는 각 모형을 잘못 지정하였을 때, g-formula가 제공하는 추정치가 어떠한 모형에 크게 의존하는지 편향의 절댓값, 95% 신뢰구간의 포함률 그리고 평균 제곱근 오차를 통해 살펴보고자 한다.

아래의 그림은 모의실험 자료를 생성하기 위해 사용한 방향성 비순환 그래프다.



[그림 Ⅲ-9] 모의실험 자료의 생성에 사용된 방향성 비순환 그래프

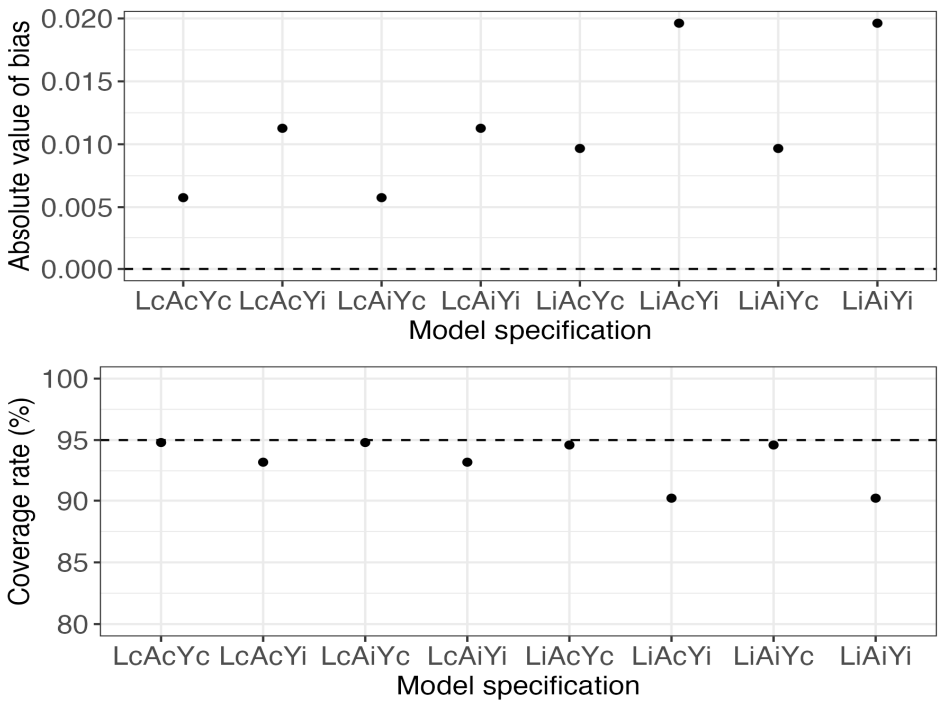
위 방향성 비순환 그래프를 기반으로 하여 표본 수는 1,000, 최대 추적관찰 기간은 5, 측정되지 않은 연속형 교란 변수 U, 측정된 연속형 교란 변수 L,

이항 노출 변수 A, 생존 변수 Y로 이루어진 모의실험 자료를 500개 생성하였다. 본 모의 실험에서 위험 비를 모수로 사용하였으며, 참 값(0.8901)은 표본 수가 100,000인 자료를 기반으로 하여 모든 대상자가 A=0일 때의 위험 대비 모든 대상자가 A=1일 때의 위험의 값으로 계산하였다. 또한, 95% 신뢰구간은 percentile bootstrap 방법을 통해 계산되었으며, bootstrap은 500번 시행되었다.

모의실험 자료에서 측정되지 않은 연속형 교란 변수 U는 시간에 관계없이 평균이 3, 분산이 1인 정규 분포에서, 측정된 연속형 교란 변수 L은 이전 시점의 교란 변수 L과 노출 변수 A 그리고 두 변수의 교호 작용 항(interaction term), 그리고 같은 시점의 교란 변수 U를 통해 산출된 평균 그리고 분산이 1인 정규 분포에서, 이항 노출 변수 A는 이전 시점의 노출 변수 A 그리고 같은 시점의 교란 변수 L 그리고 이 두 변수의 교호 작용 항을 통해 산출된 확률 값을 가지는 베르누이 분포에서 그리고 생존 변수 Y는 같은 시점의 이항 노출 변수 A, 교란 변수 L, 이 두 변수의 교호 작용 항 그리고 교란 변수 U를 통해 산출된 확률 값을 가지는 베르누이 분포에서 생성되었다. 또한, g-formula 적합 시 각 모형의 교호 작용 항을 포함하지 않는 것으로 모형을 잘못 지정하였다.

아래의 그림은 시간에 따라 변하는 교란 변수, 노출 변수 그리고 결과 변수에 대한 모형을 올바르게 지정 또는 잘못 지정하여 얻어지는 총 8개의 조합에 대해 g-formula가 제공하는 추정치의 편향의 절댓값(상단), 95% 신뢰구간의 포함률(하단)을 표현한 그래프임. [그림 Ⅲ-10]의 x-축에서 L, A 그리고 Y는 각각 교란 변수에 대한 모형, 노출 변수에 대한 모형 그리고 결과 변수에 대한 모형을 의미하며, 세 알파벳 옆 c, i에 대해 c는 모형을 올바르게 지정한 경우, i는 모형을 오지정한 경우를 나타낸다. 예를 들어, LcAcYc는 교란 변수, 노출 변수 그리고 결과 변수에 대한 모형이 모두 올바르게 지정된 경우를 의미하며, LiAcYi는 노출 변수에 대한 모형만 올바르게 지정되고, 나머지 두 모형, 즉 교란 변수에 대한 모형 그리고 결과 변수에 대한 모형은 모두 잘못 지정된 경우를 의미한다. [그림 Ⅲ-10]의 상단 그림에서 모든 모형이 올바르게 지정되었을

경우(LcAcYc)와 비교하여 교란 변수에 대한 모형 또는 결과 변수에 대한 모형이 잘못 지정되었을 때 편향이 생기는 것을 확인할 수 있으며, 그 크기는 결과 변수에 대한 모형을 잘못 지정하였을 때 더 크게 나타났다. 또한, 교란 변수와 결과 변수에 대한 모형을 모두 잘못 지정하였을 경우(LiAcYi 또는 LiAiYi), 편향이 가장 크게 나타났다. [그림 Ⅲ-10]의 하단 그림에서 교란 변수 또는 노출 변수에 대한 모형만 올바르게 지정되지 않았을 경우, 95% 신뢰구간의 포함률이 감소하는 문제가 발생하지 않았지만, 결과 변수에 대한 모형만 잘못 지정되었을 경우에는 포함률이 소폭 감소하였다. 더불어, 결과 변수에 대한 모형이 잘못 지정되었을 때, 교란 변수에 대한 모형이 추가적으로 잘못 지정되면 포함률이 더 크게 감소하는 것으로 나타났다.

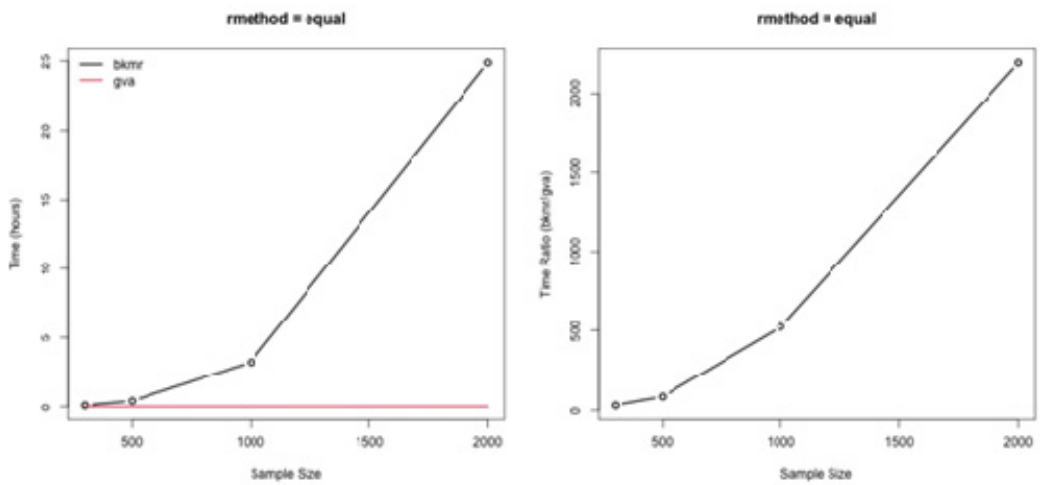


[그림 Ⅲ-10] 모형을 잘못 지정한 조합에 따른 g-formula가 제공하는 추정치의 편향의 절댓값(상단)과 95% 신뢰구간의 포함률(하단) 그래프. 상단 그림에서 점선은 y=0인 직선을 의미하며, 하단 그림에서 점선은 y=95(%)인 직선을 나타냄.

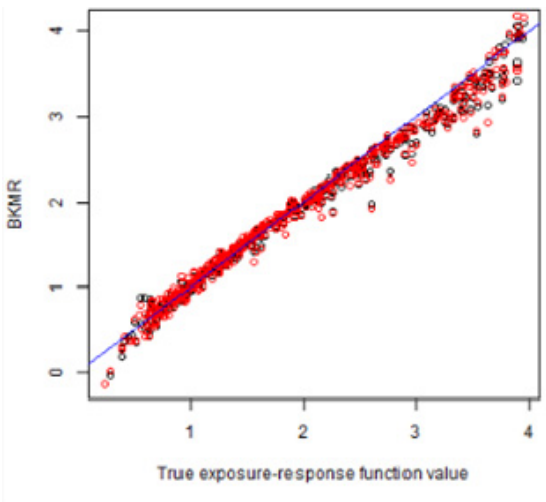
4. BKMR의 통계분석법 개발

1) BKMR 분석시간 단축 및 반복 측정된 자료에서 기울기에 랜덤 효과 적용

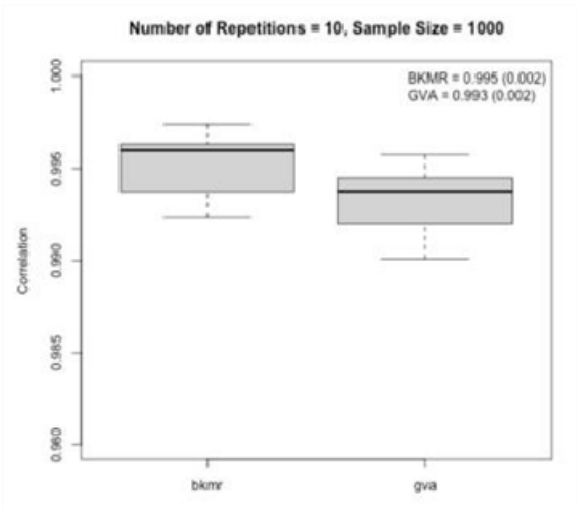
커널의 초모수(hyperparameter)인 length-scale 모수에 대해 기존 BKMR 모형에서 제안한 변수 선택 사전 분포가 아닌 horseshoe 축소 사전 분포 및 변분 근사 알고리즘을 적용하여 사후 분포의 계산 속도를 대폭 개선하였다. 이러한 개선점이 반영된 본 제안 방법을 통해 BKMR을 적용하여 얻은 복합노출에 대한 건강 영향 추정치의 근사 값을 기존 BKMR과 비교하여 매우 빠른 시간 내에 얻을 수 있다. 아래의 그림을 보면 개발한 BKMR 모형은 표본 수의 증가에 따라 분석에 소요되는 시간이 완만하게 증가하는 반면, 기존 BKMR 모형은 분석에 소요되는 시간이 매우 급격하게 증가하는 것을 확인할 수 있다.



[그림 Ⅲ-11] 표본 수에 따라 소요되는 분석 시간을 나타내는 그래프. 검은 선은 기존 BKMR 모형을, 빨간 선은 개발한 BKMR 모형을 나타냄.



[그림 Ⅲ-12] 기존 BKMR 모형과 개발된 방법을 적용하여 얻은 추정치와 참 값을 비교하는 그래프. 검은 점은 기존 BKMR 모형을 적합하여 얻은 추정치, 빨간 점은 개발된 BKMR 모형을 적용하여 얻은 추정치이며, 파란 선은 y=x임.



[그림 Ⅲ-13] 기존 BKMR 모형과 개발된 BKMR 모형에서 참 값과의 상관 계수 비교 그래프. bkmr은 기존 bkmr 모형을 나타내며, gva는 개발된 BKMR 모형을 의미함.

기존 BKMR 모형과 개발된 BKMR 모형이 참 값에 얼마나 근접하는지 확인하였다[그림 Ⅲ-13]. [그림 Ⅲ-13]에서 개발된 모형을 통해 얻어진 추정치가 기존 BKMR 모형으로부터 얻어진 추정치에 가깝게 근사하고 있음을 알 수 있으며, 나아가 참값(true exposure-response function value)와 비슷하게 나타나고 있음을 알 수 있다. 또한, 근로자 종적 자료와 같이 반복 측정된 자료의 분석에서 주로 사용되는 랜덤 효과 모형(random effect model)에서 허용되는 랜덤 절편과 랜덤 기울기에 대해 기존 BKMR 모형은 절편에만 랜덤 효과가 허용되었다. 하지만 본 연구에서는 성김 구조의 촐레스키(Cholesky) 요인을 가정하여 기울기에 랜덤 효과를 허용하는 BKMR 모형의 사후 분포를 근사하였고, mini-batch 확률적 경사법을 확장하여 랜덤 효과에 대한 분포의 공분산의 행렬식과 역행렬 계산에 필요한 계산량을 줄여 BKMR에서 기울기에 랜덤 효과를 허용하는 모형을 구축하였다.

분석 속도가 개선된 BKMR과 기울기에 랜덤 효과를 허용하는 BKMR 모형은 아래의 R 패키지 ‘vbayesGP’에 내장된 gvagpr 함수를 통해 사용이 가능하다.

```
priors <- list(lengthscale="horseshoe")

gvagpr(
y = y,
X = X,
Z = Z,
id = NULL,
random.slope = NULL,
priors = priors,
covstr = 'diagonal')
```

gvagpr 함수에서 인수 y는 결과 변수을 나타내는 벡터, X는 공변량을 포함하는 행렬, Z는 노출 변수를 포함하는 행렬, id는 반복 측정된 자료가 같은 근로자로부터 나왔음을 알려주는 index 그리고 random.slope는 공변량 행렬 X에서 랜덤 효과를 적용하고자 하는 열 번호를 의미한다. 또한, priors를 통해 커널 행렬에 사용되는 기여도에 사용하려는 사전 분포에 대한 정보를 입력할 수 있으며, covstr를 통해 공분산 행렬의 구조를 지정할 수 있다. 사전 분포(priors)는 list의 형태로 lengthscale이라는 인수를 입력받아 지정할 수 있다. 기본적으로 gvagpr 함수는 기존 BKMR R 함수 kmbayes와 비교하여 분석 속도가 향상된 함수이며, id만 입력하였을 때는 기존 BKMR과 동일하게 절편에만 랜덤 효과가 적용된 모형을 적합할 수 있지만, random.slope까지 입력하게 되면 기울기에도 랜덤 효과를 적용할 수 있다.

다음은 모의 실험자료에 gvagpr 함수를 적용한 결과이다.

```
summary(foutmbu.diag)

Fitted object of class 'gpr'
Outcome family: gaussian
Covariance Strucutre: diagonal
Lengthscale Parameter: varying and shrinkage
Minibatch-Epoch: 100,000
Running time: 1326.89 secs
Model fit on: 2023-11-04 23:31:31

Parameter estimates:
                mean       sd    q_2.5      q_50    q_97.5
beta         2.01770 0.01727  1.98601   2.01783   2.05056
sigsq.eps    0.55374 0.03502  0.48750   0.55374   0.62242
lambda      10.69106 1.55571  8.00172  10.61979  14.14561
```

```
r1          0.02699 0.00664 0.01632  0.02617  0.04137
r2          0.03304 0.00873 0.01946  0.03178  0.05209
r3          0.00000 0.00000 0.00000  0.00000  0.00000
r4          0.00000 0.00000 0.00000  0.00000  0.00000
r5          0.00000 0.00000 0.00000  0.00000  0.00000
Pseudo posterior inclusion probabilities:
  variable  PPIP
1       z1 0.956
2       z2 0.999
3       z3 0.000
4       z4 0.000
5       z5 0.000
```

R 패키지 'vbayesGP'는 결과물을 출력하기 위해 R 기본 함수인 summary 함수를 이용하는 것은 동일하지만, 참 노출-반응 함수 값 또는 계수의 수렴 여부를 확인하기 위해 R 패키지 'bkmr'는 자체 개발 함수인 bkmr::ComputePostmeanHnew 또는 TracePlot를 사용해야 하지만, R 패키지 'vbayesGP'는 R 사용자에게 친숙한 fitted 함수 또는 plot 함수를 통해 확인이 가능하다. 보다 자세한 함수의 사용법은 부록 1에서 자세히 설명하고자 한다.

2) BKMR의 로지스틱 회귀 모델로의 확장

많은 역학 연구자들이 이항 자료를 분석할 때, 프로빗(probit) 회귀 모델보다 로지스틱 회귀 모델을 주로 사용하고 있지만, 기존 BKMR 방법은 로지스틱 모델이 아닌 프로빗 모델을 통해 이항 자료를 다루고 있다. 따라서 본 연구는 BKMR 방법에서 로지스틱 모델을 통해 이항 자료를 다룰 수 있도록 확장하고자 한다. 이 때, BKMR 방법에서 이항분포를 가능도 함수로 하는 로지스틱 회귀 모델로 확장할 때 어려운 점은 주변 가능도 함수를 알려진 형태로 계산하기 어렵다는

것이다. 본 연구에서는 이러한 문제를 해결하기 위해 중요도 추출(importance sampling)방법을 활용하여 중요도 함수를 추정하고, 이를 통해 주변 분포가 근사하고자 하는 사후 분포가 되도록 확대 사후 분포(augmented posterior distribution)을 구성하였다. 이후 변분 분포를 정의하고 쿨백-라이블러 발산(Kullback-Leibler divergence)을 기준으로 하여 확률적 경사 알고리즘을 적용하여 확대 사후 분포를 근사하였다.

로지스틱 모델을 통해 이항 변수를 다루는 BKMR 방법은 아래의 R 패키지 'vbayesGP'에 내장된 gvaggpr 함수를 통해 사용이 가능하다.

```
gvaggpr(y,
        X,
        Z,
        id = NULL,
        random.slope = NULL,
        family = binomial,
        priors = list(),
        covstr = c("diagonal", "blockdiag"),
        control = list(),
        verbose = TRUE,
        seed = sample.int(.Machine$integer.max, 1))
```

앞서 설명한 gvagpr 함수에서와 동일하게 gvaggpr 함수에서 인수 y는 결과 변수를 나타내는 벡터, X는 공변량을 포함하는 행렬, Z는 노출 변수를 포함하는 행렬, id는 반복 측정된 자료가 같은 근로자로부터 나왔음을 알려주는 index 그리고 random.slope는 공변량 행렬 X에서 랜덤 효과를 적용하고자 하는 열 번호를 의미한다. 또한, priors를 통해 커널 행렬에 사용되는 기여도에 사용하려는 사전 분포에 대한 정보를 입력할 수 있으며, covstr를 통해 공분산 행렬의 구조를 지정할 수 있다. 사전 분포는 list의 형태로 lengthscale이라는 인수를 입력받아

지정할 수 있다. gvagpr 함수와 다른 점은 인수 family가 추가되었다는 것이며, 로지스틱 회귀 모델을 사용하고자 하는 경우에는 family="binomial"을 입력하면 된다.

다음은 모의실험 자료에 gvaggpr 함수를 적용한 결과이며, 함수 gvaggpr 또한 결과물을 출력하기 위해 R 기본 함수인 summary 함수를 이용하는 것과 참 노출-반응 함수 값 또는 계수의 수렴 여부를 확인하기 위해 plot 함수를 이용하는 것 모두 gvagpr 함수와 동일하다. 자세한 함수의 사용법은 부록 1에서 자세히 설명하고자 한다(모의실험 자료를 생성하는 코드는 부록 1에 수록되어 있음).

```
fit_logisticBKMR <- vbayesGP::gvaggpr(
y = datp$y,
Z = datp$Z,
X = datp$X,
priors = list(lengthscale = 'normal'),
family = 'binomial')

summary(fit_logisticBKMR)

Fitted object of class 'gpr'
Outcome family: binomial ( logit )
Covariance Strucutre: diagonal
Lengthscale Parameter: equal
Iterations: 3947
Running time: 16.5 secs
Model fit on: 2023-11-03 10:23:05
```

```
Parameter estimates:
             mean        sd      q_2.5      q_50    q_97.5
beta     0.03880 0.05122 -0.06392 0.03757 0.14231
lambda 2.21210 2.32495  0.31328 1.53661 8.30152
r        1.91242 1.49731  0.40541 1.53330 5.71578
```

5. 개선된 통계방법론에 대한 활용 가이드라인 작성

산업보건 역학연구에 g-formula를 적용한 후 얻어지는 결과를 사용하여 유해물질의 복합노출의 정도에 따른 건강 영향을 직관적으로 확인하기 위한 용량-반응 곡선과 교호 작용을 표현하는 시각화 코드를 산업보건 역학 연구자들이 용이하게 사용할 수 있도록 함수의 사용법에 대한 활용 가이드라인을 작성하였다. 또한, 표본 수에 대응하는 차원을 가지는 커널 행렬과 마코프 체인 몬테-카를로 기반 베이지안 기법의 사용으로 인한 BKMR의 분석 속도를 개선하였으며, 개선된 방법을 사용할 수 있는 함수의 사용법 또한 산업보건 역학 연구자들이 용이하게 사용할 수 있도록 활용 가이드라인을 작성하였다. 현재까지 개발된 방법에 대해 함수의 사용법은 부록 1에 기술하였다.

Ⅳ. 고찰

Ⅳ. 고찰

1. 주요 연구 결과

1) 인과추론 및 복합노출 국문 가이드라인 무료 배포용 책자 개발

- **합성 데이터를 활용한 g-formula 분석 가이드라인 개발**: 2021년과 2022년에 진행하였던 직업병 인과추론 가이드라인 및 통계분석법 개발(1, 2)를 요약하고, 특수건강진단 자료를 기반으로 'synthpop' R 패키지를 사용하여 실제 특수건강진단 자료와 유사한 예제용 합성 데이터를 만들었다.
- **인과추론 교과서 번역본에 대한 후속 과제 기획:** 인과추론에 대한 국내 연구진들의 이해를 높이고자, 후속 과제로 미국 하버드 대학의 Miguel A. Hernán 및 James M. Robins 교수가 발간한 인과추론 교과서'Causal Inference: What If' 번역본을 만들어 산업안전보건연구원 홈페이지를 통해 무료배포 하고자 한다.

2) 인과추론 및 복합노출 국문 가이드라인의 활용

- **국문 가이드라인 활용 세미나 진행**: 전공의 3인과 산업위생 전문가 1인을 대상으로 특수건강진단 자료를 활용한 g-formula 분석에 대한 세미나를 진행하였다(이론 세미나 1회, g-formula 예제 분석 세미나 1회, 특수건강진단자료 데이터 클리닝 세미나 3회, g-formula 특수건강진단 자료 분석 세미나 1회).
- **질의 응답 정리**: 세미나 중 질의한 내용에 대한 응답을 정리하였다.

3) g-formula를 이용한 용량-반응 곡선 및 교호 작용을 표현하는 시각화 코드 개발

- **용량-반응 곡선**: 근로자 종적 자료에서 단일 유해물질에 대한 노출의 건강 영향을 평가할 때, 노출 농도에 따른 건강 영향을 직관적으로 전달하기 위해 시각적인 그림을 제공하는 경우가 많다. 따라서 본 연구에서는 단일 유해물질의 노출 정도에 따른 건강 영향을 직관적으로 표현할 수 있게 하는 시각화 코드(DoseResponsePlot)를 개발하였다.
- **등고선 그림**: 단일 유해물질의 경우, 농도에 따른 건강 영향을 2차원 그래프로 쉽게 표현이 가능하지만 두 유해물질의 농도에 따른 건강 영향의 경우 3차원 그래프를 통해 표현해야하므로, 위의 코드를 곧바로 적용하기 어렵다. 그러한 이유로 근로자 종적 자료에서 두 개의 유해물질에 대한 복합 노출의 건강 영향을 직관적으로 표현하기 위한 시각화 코드(ContourPlot)를 개발하였다.
- **교호 작용 그림:** 2개 이상의 유해물질로 인한 복합 노출의 건강 영향을 평가할 때, 유해물질 간 교호 작용 효과(또는 시너지 효과)를 파악하여 근로자의 건강을 악화시키는 유해물질 사이의 조합을 확인할 수 있다. 그 효과는 additive interaction, multiplicative interaction 그리고 RERI를 통해 계산이 가능하며, 계산된 additive interaction 값을 직관적으로 표현하기 위한 시각화 코드(InteractionPlot)를 개발하였다.

4) g-formula의 분석 결과의 안정성을 평가하는 방법

- **자료의 결측치**: 특수건강진단 자료와 같이 반복 측정된 자료에서 시간에 따라 변하는 노출 변수 또는 교란 변수에 결측 치가 존재하는 경우, 그 결측치를 채우는 방법으로 LOCF와 imputation 방법을 고려할 수 있다. 본 연구는 두 방법을 적용하였을 때, 교란 변수의 결측 비율에 따라 인과효과

추정치에 발생하는 편향의 절댓값, 표준 오차 비 그리고 95% 신뢰구간의 포함률에 대해 수치적으로 조사하였다. 본 연구에서 수행한 모의실험 자료에서 결측 비율이 증가할 때, LOCF 방법과 비교하여 imputation 방법이 상대적으로 작은 편향, 작은 표준 오차 그리고 높은 신뢰구간 포함률을 보였다. 특히, 결측 비율이 30% 이상인 경우, LOCF 방법에서 모두 편향 및 표준 오차 비의 급격한 증가 그리고 신뢰구간 포함률의 급격한 감소가 관찰되었다.

- **근로자의 불규칙한 특수건강진단 문제**: 업무 전환 조치 등의 적절한 사유로 자료에서 일부 근로자는 매년 특수건강진단을 받는 것으로 나타나지 않고, 불규칙적으로 검진을 받는 것으로 나타났다. 이러한 이유로 이전 과제에서는 검진 연도를 기준으로 분석을 시행한 것이 아닌 검진 순서에 따라 분석을 시행하였다. 검진 순서의 경우, 검진 순서 사이의 시간에 대한 고려가 어렵기 때문에 검진 연도를 기준으로 분석을 시행하고자 하는 경우, 이러한 근로자의 불규칙한 검진 패턴을 고려한 분석이 수행되어야 한다. 검진 받지 않은 연도에 해당하는 자료를 결측 자료로 보고, imputation 방법을 통해 결측치를 채운 후 g-formula를 적용하여 분석 결과의 안정성을 추정치의 편향의 절댓값, 표준 오차 비 그리고 95% 신뢰구간의 포함률을 사용하여 조사하였다. 불규칙한 자료의 비율이 증가할수록 g-formula가 제공하는 추정치의 편향의 절댓값과 표준 오차의 크기가 증가하였고, 95% 신뢰구간의 포함률은 감소하였다. 이러한 결과로부터 검진 받지 않은 연도에 해당하는 자료를 결측된 자료로 생각하여 종적 자료가 가지고 있는 불규칙한 자료의 구조를 규칙적인 자료의 구조로 확장하고 검진 받지 않은 연도에서 발생하는 결측치를 채워 g-formula를 적용하는 것이 한계점을 가지는 접근이라는 것을 확인할 수 있었다. 자료의 불규칙적 관측을 반영하는 새로운 방법의 개발이 필요할 것으로 생각된다.
- **노출 변수 또는 교란 변수에 대한 모형 지정 검토**: g-formula는 사용되는 모든 모형(교란 변수, 노출 변수 그리고 결과 변수에 대한 모형)이 올바르게

지정(correct specification)되어야 편향 없는 추정치를 제공한다. 하지만 모형이 올바르게 지정되었는지 확인하기 위해 현재 gfoRmula R 패키지에서 제공하는 것은 자연 경과 조건에서 예측되는 노출 변수 및 교란 변수의 이력과 실제 관측 값을 비교하는 그래프이며, 각 모형이 잘 적합하였는지 그래프로 확인하는 것은 주관적이기 때문에 어려운 작업이다. 따라서 본 연구는 g-formula가 교란 변수에 대한 모형, 노출 변수에 대한 모형 그리고 결과 변수에 대한 모형 중 어느 모형에 크게 의존하는지 각 모형을 잘못 지정한 후, 얻어지는 g-formula의 인과 효과 추정치를 편향의 절댓값, 95% 신뢰구간의 포함률을 사용하여 조사하였다. 그 결과, 노출 변수에 대한 모형과 관계없이 매 시점 일정하게 개입(intervention)하는 경우, 교란 변수에 대한 모형 또는 결과 변수에 대한 모형을 잘못 지정하였을 때, 편향의 절댓값의 크기가 증가하였고, 신뢰구간의 포함률은 결과 변수 모형이 잘못 지정되었을 때 감소하는 결과가 나타났다. 또한, 결과 변수에 대한 모형뿐만 아니라 교란 변수에 대한 모형도 같이 잘못 지정되었을 경우, 그 크기는 더 감소하였다(편향에서는 그 크기가 더 증가하였음).

5) BKMR 분석시간 단축 및 반복 측정된 자료에서 기울기에 랜덤 효과 적용

- **BKMR의 분석 속도 개선**: 기존 BKMR 방법에서 제안한 변수 선택 사전 분포가 아닌 horseshoe 축소 사전 분포 및 변분 근사 알고리즘을 사용하여 사후 분포를 근사하여 계산 속도를 대폭 개선하였다.
- **반복 측정된 자료에서 기울기에 랜덤 효과 적용**: 기존 BKMR 방법에서는 랜덤 절편만 허용이 가능하였지만, 성김 구조의 촐레스키 요인을 가정하여 사후 분포를 근사하고, mini-batch 확률적 경사법을 확장하여 랜덤 효과에 대한 분포의 공분산의 행렬식과 역행렬 계산에 필요한 계산량을 줄여 랜덤 기울기를 허용하는 BKMR 방법을 구축하였다.

6) BKMR의 로지스틱 회귀 모델로의 확장

- 기존 BKMR 방법의 경우, 이항 자료를 다루기 위해 프로빗 회귀 모델을 사용하였지만, 본 연구에서는 중요도 추출 방법을 통해 중요도 함수를 추정 및 변분 분포를 정의하고 쿨백-라이블러 발산을 기준으로 확률적 경사 알고리즘을 적용하여 주변 가능도 함수를 근사하는 확대 사후 분포를 구성하여 BKMR에서 역학 연구자 및 의료 분야 연구자들이 많이 사용하는 로지스틱 회귀 모델을 사용할 수 있도록 하였다.

7) 개선된 통계방법론에 대한 활용 가이드라인 작성

- 본 연구에서 개발한 g-formula를 이용한 용량-반응 곡선 및 교호 작용을 표현하는 시각화 코드 함수 및 BKMR의 분석 속도를 개선하고, 반복 측정된 자료에서 기울기에 랜덤 효과를 허용한 방법 나아가 BKMR에 로지스틱 모형을 적용한 방법을 산업보건 역학 연구자가 용이하게 사용할 수 있도록 함수의 사용법에 대한 활용 가이드라인을 작성하였다.

2. 연구 활용방안

- 다양한 산업 보건 역학 연구에서 알고자 하는 주된 관심사인 건강 결과와 유해물질 사이의 용량-반응 곡선 및 유해물질 사이의 교호 작용을 평가하고 이를 시각적으로 표현할 수 있는 그래프를 제공하여 g-formula로 추정한 여러 복합물질에 대한 위험을 다양한 지표를 통해 평가할 수 있다. g-formula 분석 결과에 대한 안정성을 제고하는 근거를 제시함으로써 g-formula를 국내 산업 안전 보건 역학 연구의 결과를 신뢰성을 확보할 수 있다.

- 표본의 수가 큰 자료에서 수행하기 어려웠던 BKMR을 본 과제의 결과물을 통해 수행할 수 있게 되었으며, 분석 소요시간으로 인해 진행하지 못한 근로자의 반복 측정된 자료에 대한 분석 연구의 발전을 기대할 수 있다. 또한, 표본 수가 큰 빅 데이터를 통해 보다 정확한 인과 효과 추정치와 신뢰구간의 제공이 가능하다.
- 기존 BKMR에서 사용되던 프로빗 모형의 사용으로 인한 결과 해석의 어려움을 역학 연구 및 의학 연구에서 주로 사용되는 로지스틱 회귀 모델을 기반으로 한 BKMR의 개발을 통해 해소할 수 있다. 이를 통해 작업장에서 발생하는 다양하고 새로운 복합물질에 대한 노출과 건강 결과 사이의 이해를 확고히 하고, 나아가 새로운 유해물질 노출에 대한 기준 등의 정책 마련의 출발점이 될 수 있다.
- 위와 같은 평가를 통해 위험성이 높은 물질에 대해 노출 제한 용량 또는 기준을 재정비하고, 직업성 질환에 대한 누적 발생률 또는 사망률 감소시켜 근로자 개인으로서 삶의 질이 향상될 뿐만 아니라 국가적으로 근로자의 건강관리를 도울 수 있으며, 국가사업의 사회적 기여도를 높일 수 있다.
- 본 과제를 통하여 산업 보건 연구를 진행하는 연구자들이 복합노출에 대한 건강 영향 평가 분석방법인 g-formula와 BKMR을 자료에 적용하는 데 필요한 시간을 단축시키며, 올바르게 사용하도록 하여 근로자의 사망 또는 건강 지표에 대한 복합노출의 효과를 추정, 산출할 수 있도록 한다.

참고문헌

남정모, 김진흠, 강대룡, 안연순, 이후연, 이대희. Intermediate 변수의 영향을 통제하는 통계적 방법론에 대한 연구 : 건강근로자효과를 통제하기 위한 새로운 접근. Korean Journal of Epidemiology(한국역학회지). 2002; 24(1):7-16.

예신희, 이경은, 성정민, 박동준, 이우주. 직업병 인과추론 가이드라인 및 통계분석법 개발(1): -g methods 국문 가이드라인 개발. 산업안전보건연구원. 2021.

예신희, 이경은, 윤민주, 박동준, 마성원, 이영신, 이우주. 직업병 인과추론 가이드라인 및 통계분석법 개발(2): 복합노출의 건강 영향평가 국문 가이드라인 개발. 산업안전보건연구원. 2022.

이경무, 전재범, 박동욱, 이원진. 건강근로자효과의 최소화 방안과 보정 방법. 한국환경보건학회지(구 한국환경위생학회지). 2011; 37(5):342-347.

이슬비. 임신 중 복합 환경유해물질 노출이 6 개월 영유아 아토피 피부염 발생에 미치는 영향. 2019.

Agier L, Portengen L, Chadeau-Hyam M, Basagaña X, Giorgis-Allemand L, Siroux V, Robinson O, Vlaanderen J, González JR, Nieuwenhuijsen MJ, Vineis P, Vrijheid M, Slama R, Vermeulen R. A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations. Environ Health Perspect. 2016;124(12):1848-1856.

Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. Bayesian kernel machine regression for

estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015;16(3):493-508.

Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. Environ Health. 2018;17(1):67.

Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. J Agric Biol Environ Stat. 2015;20 (1):100-120.

Forns J, Mandal S, Iszatt N, Polder A, Thomsen C, Lyche JL, Stigum H, Vermeulen R, Eggesbø M. Novel application of statistical methods for analysis of multiple toxicants identifies DDT as a risk factor for early child behavioral problems. Environ Res. 2016;151: 91-100.

Gruber S, Van der laan MJ. tmle: An R Package for Targeted Maximum Likelihood Estimation. Journal of Statistical Software. 2012;51(13): 1-35.

Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc. 1952;47:663-685.

Hernán MA, Robins JM. Causal Inference: What If. Boca Raton: Chapman & Hall/CRC. 2020.

Keil AP, Buckley JP, O'Brien KM, Ferguson KK, Zhao S, White AJ. A Quantile-Based g-Computation Approach to Addressing the

Effects of Exposure Mixtures. Environ Health Perspect. 2020;128(4): 47004.

Lin V, McGrath S, Zhang Z, Petito LC, Logan RW, Hernan MA, Young JG. GfoRmula: an R package for estimating effects of general time-varying treatment interventions via the parametric g-formula. arXiv:1908.07072. 2019.

Naimi AI, Cole SR, Kennedy EH. An introduction to g methods. Int J Epidemiol. 2017;46(2):756-762.

Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1):41-55.

Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period: application to the healthy worker survivor effect. Mathematical Modelling. 1986;7:1393-1512. [Errata(1987) in Computers and Mathematics with Applications 14, 917-921. Addendum(1987) in Computers and Mathematics with Applications 14, 923-945. Errata(1987) to addendum in Computers and Mathematics with Applications 18, 477.]

Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. Health Service Research Methodology: A Focus on AIDS. 1989; 113-159. Eds: Sechrest L., Freeman H., Mulley A. Washington, D.C.: U.S. Public Health Service, National Center for Health Services Research.,pp. 113-159. Please see also the following link for the Errata to the article: ERRATA for The analysis of

randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal data.

Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11(5):550-560.

Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. J Am Stat Assoc. 1994;89(427):846-866.

Sjölander A. Regression standardization with the R package stdReg. Eur J Epidemiol. 2016 Jun;31(6):563-74.

Sun Park, Seongil Jo and Woojoo Lee. A variational Bayes method for pharmacokinetic model. The Korean Journal of Applied Statistics. 2021;34(1):53-67.

van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med. 1999;18(6): 681-694.

Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011; 45(3):1-67.

Van der laan MJ, Rubin DB. Targeted maximum likelihood learning. Int J Biostat. 2006;2(1):Article 11.

Van der Laan MJ, Rose S. Targeted learning: causal inference for observational and experimental data (Vol. 4). New York: Springer. 2011.

VanderWeele T, Knol M. A Tutorial on Interaction. Epidemiologic Methods. 2014;3(1):33-72.

MHV Ong, Mensah DK, Seongil Jo, Nott DJ, Beomjo Park and Taeryon Choi. A variational approach to Bayesian shape restricted regression with Gaussian process priors. Electronic Journal of Statistics, 2017;11(2):4258-4296.

Van der wal WM, Geskus RB. ipw: An R Package for Inverse Probability Weighting. Journal of Statistical Software. 2011;43(13):1-23.

Zhong Y, Kennedy EH, Bodnar LM, Naimi AI. AIPW: An R Package for Augmented Inverse Probability-Weighted Estimation of Average Causal Effects. Am J Epidemiol. 2021;190(12):2690-2699.

Zetterqvist J, Sjölander A. Doubly Robust Estimation with the R Package drgee. Epidemiologic Methods. 2015;4(1):69-86.

Abstract

Causal Inference and Statistical Methods for Environmental Mixtures in Occupational Studies

Objectives: We aim to enhance the interpretation of analysis results by visualizing the dose-response curve, contour plot and relative excess risk due to interaction(RERI) as one of measures of causal interaction. We also aim to develop a novel computational method to reduce the computational cost of implementing Bayesian kernel machine regression(BKMR) and to allow random slope effects in BKMR.

Method: We consider how to effectively visualize the analysis results from the g-formula. For reducing the computation burden, we apply variational bayes, Cholesky factorization for sparse structure, mini-batch stochastic gradient method for approximating the posterior.

Results: We provide R functions to draw plots for the dose-response curve, contour plot and RERI. In addition, we substantially reduce the computational burden of the existing BKMR method and enable researchers to handle the random slope effects in BKMR. We also develop a logistic BKMR. We provide the R function to implement the newly proposed methods.

Key words: environmental mixture, g-formula, Bayesian kernel machine regression, variational Bayes, causal interaction

부록 1.

개선된 통계방법론에 대한 활용 가이드라인

Ⅰ. g-formula의 통계분석법 개발

g-formula에 대하여 2021년 연구과제(예신희 등(2021))에서는 단일 유해물질에 대한 노출로 발생하는 건강 영향을 추론하는 국문 가이드라인을, 2022년 연구과제(예신희 등(2022))에서 여러 유해물질에 의한 복합 노출로 발생하는 건강 영향을 추론하는 국문 가이드라인을 작성하였다. 하지만 2023년 본 과제의 부록에서는 단일 또는 두 유해물질에 대한 노출로 인한 건강 영향을 직관적으로 해석할 수 있도록 하는 시각화 코드를 개발 및 제공하여 산업보건 역학 연구자들이 쉽게 사용할 수 있도록 하는 활용 가이드라인을 작성하고자 한다. 본 부록에서 설명할 3개의 함수(DoseResponsePlot, ContourPlot, InteractionPlot)들은 R 패키지 'gfoRmula'에 내장된 gformula()를 기반으로 하고 있으며, 이 함수에 대한 자세한 설명은 2020년에 발표된 S MCGrath 등(2020), 예신희 등(2021), 예신희 등(2022)을 참고하기 바란다.

```
library(dplyr)
```

1. 용량-반응 곡선과 교호 작용을 표현하는 시각화 코드 개발

산업보건 역학 연구자가 새로 개발한 함수를 수월하게 사용할 수 있도록 g-formula에 내장된 자료를 기반으로 함수의 사용법을 설명하고자 한다. basicdata_nocomp는 경쟁 위험이 없는 2,500명의 환자를 7년 간 추적 관찰(follow-up)한 모의실험 자료로 아래의 코드를 통해 R 패키지 'gfoRmula'를 실행시킨 후 확인이 가능하다.

```
library(gfoRmula)
data("basicdata_nocomp")
```

S McGrath 등(2020)의 설명에서는 basicdata_nocomp 자료에서 변수 'A'를 binary treatment로 사용하고 있으나, 본 부록에서는 시간에 따라 변하는 변수(time-varying variable)와 결과 변수(outcome) 사이의 용량-반응 관계를 설명하기 위해 자료에서 time-varying confounder로 사용된 연속형 변수 'L2'를 treatment로 사용하고자 한다. 용량-반응 관계를 살펴보기 위해 변수 'L2'의 범위를 확인한 결과 -1부터 1까지 0.5 단위로 용량-반응 관계를 살펴보고자 한다.

```
summary(basicdata_nocomp$L2)

Min. 1st Qu.  Median     Mean 3rd Qu.     Max.
-1.3782 -1.0183 -0.8912 -0.1674  0.9738  1.3485
```

결과 변수 'Y'와 변수 'L2' 사이의 용량-반응 곡선을 그리기 전에 용량-반응 곡선을 그리고자 하는 용량에 대해 g-formula를 적합하였다. 아래의 코드는 위에서 설정한 용량(-1, -0.5, 0, 0.5, 1)에 대해 g-formula를 적용하는 코드이다. 결과까지의 시간을 단축시키기 위해 bootstrap은 20번만 수행하였다(nsample = 20).

```
# 개입하고자 하는 값 지정
Dose <- (-2:2) * 0.5

# 개입의 대상이 되는 변수와 개입 전략 지정
intvars1 <- interventions1 <- list()
for(i in 1:length(Dose)){
  intvars1[[i]] <- c("L2","L2")
  interventions1[[i]] <- list(c(static, rep(x = Dose[i], times = 7)))
}; remove(i)
```

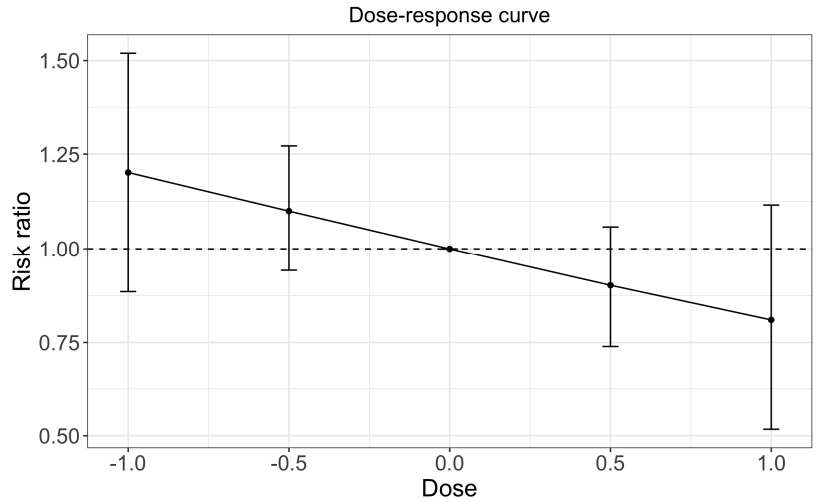
```
# 개입 전략 명 지정
int_descript1 <- c()
for(i in 1:length(Dose)){
  int_descript1[i] <- paste0("L2"," [Dose: ",Dose[i],"]")
}; remove(i)
```

```
# g-formula 적합
gformRes_DoseResponse <- gformula(
  obs_data = basicdata_nocomp,
  id = "id",
  time_points = 7,
  time_name = "t0",
  covnames = c('L1', 'L2', 'A'),
  covtypes = c('binary', 'bounded normal', 'binary'),
  covparams = list(covmodels = c(L1 ~ lag1_A + lag_cumavg1_L1 + lag_
cumavg1_L2 + + L3 + t0,
L2 ~ lag1_A + L1 + lag_cumavg1_L1 + lag_cumavg1_L2 + L3 + t0,
A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 + lag_cumavg1_L2 + L3 + t0)),
  histories = c(lagged, lagavg),
  histvars = list(c('A', 'L1', 'L2'), c('L1', 'L2')),
  basecovs = "L3",
  outcome_name = "Y",
  outcome_type = "survival",
  ymodel = Y ~ A + L1 + L2 + L3 + lag1_A + lag1_L1 + lag1_L2 + t0,
  intvars = intvars1,
  int_descript = int_descript1,
  interventions = interventions1,
  ref_int = 3,
  nsamples = 20,
  ci_method = "percentile",
  seed = 12345678)
```

다음은 위에서 적합한 g-formula의 결과를 기초로 하여 용량-반응 곡선을 그려주는 코드이다. 개발한 함수 DoseResponsePlot은 gformula 적합 결과를 그대로 함수에 입력하여 사용이 가능하며, 그림을 꾸미기 위해서는 다음의 인수들(xlab, ylab, main, width, lab_title_size, lab_text_size, main_size)을 추가적으로 정해주어야 한다. 용량-반응 곡선에서 x축의 범위는 Dose 인수를 통해 지정되며, 본 예제에서는 g-formula에서 개입의 범위를 Dose 벡터를 이용하여 지정하였으므로 Dose 벡터를 개발한 함수에서도 동일하게 사용하였다. xlab, ylab 그리고 main은 용량-반응 곡선에서 사용할 x-축, y-축 그리고 그림 제목을 지정할 수 있는 인수이며, lab_title_size, lab_text_size, main_size를 통해 글자 크기를 지정할 수 있다. width 인수는 신뢰구간의 상한과 하한을 표현하는 막대의 길이를, pointsize로 점의 크기를 조정할 수 있다.

함수의 결과물을 통해 용량-반응 곡선을 얻을 수 있다. 아래의 용량-반응 곡선을 통해 변수 'L2'의 값이 증가함에 따라 위험이 감소함을 알 수 있다.

```
# 개발한 함수를 이용하여 용량-반응 곡선
DoseResponse_Plot <- DoseResponsePlot(
  object = gformRes_DoseResponse,
  Dose = Dose,
  xlab = "Dose",
  ylab = "Risk ratio",
  main = "Dose-response curve",
  width = 0.03,
  pointsize = 2,
  lab_title_size = 17,
  lab_text_size = 15,
  main_size = 15)

# 용량-반응 곡선 그림 출력
DoseResponse_Plot
```



단일 유해물질에 대한 노출의 경우, 위의 그림과 같이 용량-반응 관계를 표현할 수 있지만, 2개의 유해물질에 대한 복합 노출의 경우, 두 유해물질의 농도가 같이 변하기 때문에 x-축 하나로 두 유해물질의 용량 조합을 표현하기 어렵기 때문에 용량-반응 관계를 표현하기 위해서는 등고선 그림이 필요하다. 등고선 그림을 그리는 함수의 사용법을 설명하기 위해 R 패키지 'gfoRmula'에 내장된 자료를 일부 수정하여 사용하였다. basicdata_nocomp 자료는 시간에 따라 변하는 변수로 두 이항 변수(L1, A)와 하나의 연속형 변수(L2)를 포함하고 있으므로 이항 변수 중 하나의 변수를 연속형 변수로 변환을 해서 등고선 그림을 그리는 것으로 하였다. 아래의 코드는 이항 변수 'A'를 연속형 변수로 변환하는 코드이다.

```
set.seed(seed = 20230816)
basicdata_contour <- basicdata_nocomp %>%
  rename(A_origin = A) %>%
  mutate(A = ifelse(test = A_origin == 1,
                    yes  = rnorm(n = 1, mean = 2, sd = 1),
                    no   = rnorm(n = 1, mean = -2, sd = 1)))
```

앞선 용량-반응 곡선과 같이 등고선 그림을 그리기 위해 변수 L2와 변수 A의 범위를 확인한 후, 변수 L2와 변수 A의 범위를 모두 -1부터 1까지 0.5 단위로 설정하였다.

```
summary(basicdata_contour$L2)

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.3782 -1.0183 -0.8912 -0.1674  0.9738  1.3485

summary(basicdata_contour$A)

Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.4260 -1.4260  1.2877  0.4046  1.2877  1.2877
```

등고선 그림을 그리기 전에 두 변수의 개입의 값으로 설정한(-1, -0.5, 0, 0.5, 1)에 대해 가능한 조합을 생성하고, 각 조합마다 g-formula를 적합하였다. 결과까지의 시간을 단축시키기 위해 bootstrap은 20번만 수행하였다(nsample = 20).

```
# 개입하고자 하는 값 지정
Dose1 <- Dose2 <- (-2:2) * 0.5

# 가능한 조합 생성
DoseCombination <- expand.grid(
  Dose2 = Dose2, Dose1 = Dose1,
  KEEP.OUT.ATTRS = FALSE, stringsAsFactors = FALSE)[,2:1]

# 개입의 대상이 되는 변수와 개입 전략 지정
intvars2 <- interventions2 <- list()
```

```
for(i in 1:nrow(DoseCombination)){
  intvars2[[i]] <- c("L2","A")
 interventions2[[i]] <- list(
  c(static, rep(x = DoseCombination$Dose1[i], times = 7)),
  c(static, rep(x = DoseCombination$Dose2[i], times = 7)))
}; remove(i)

# 개입 전략 명 지정
int_descript2 <- c()
for(i in 1:nrow(DoseCombination)){
  int_descript2[i] <- paste0("L2 / A"," [",DoseCombination$Dose1[i],",",Do
seCombination$Dose2[i],"]")
}; remove(i)
```

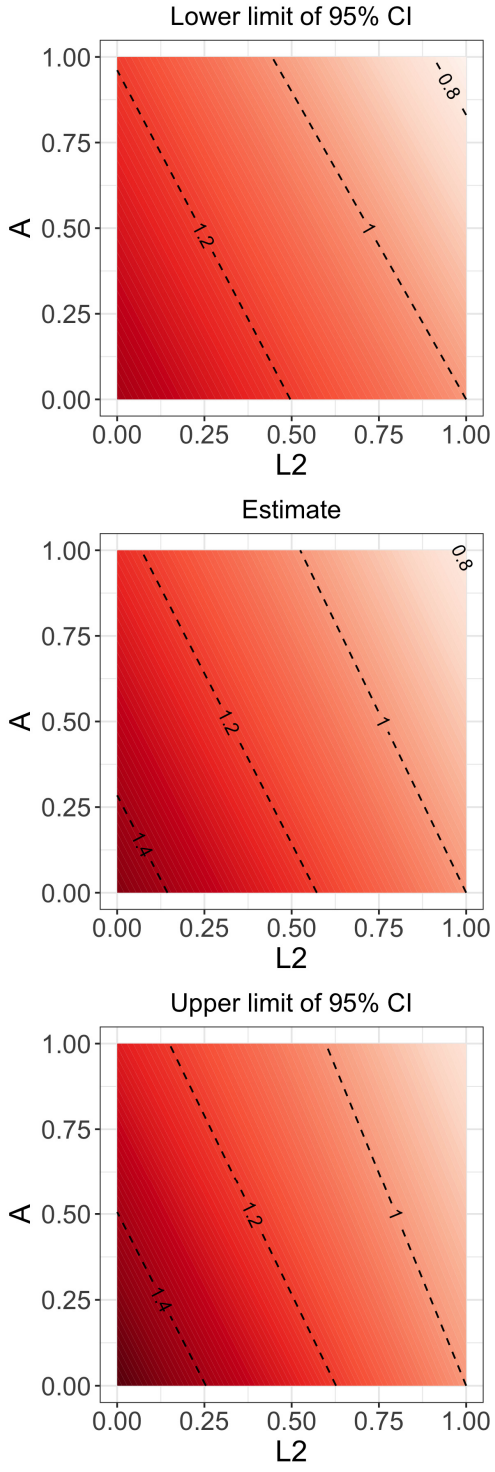
```
# g-formula 적합
gformRes_Contour <- gformula(
  obs_data = basicdata_contour,
  id = "id",
  time_points = 7,
  time_name = "t0",
  covnames = c('L1', 'L2', 'A'),
  covtypes = c('binary', 'bounded normal', 'normal'),
  covparams = list(covmodels = c(L1 ~ lag1_A + lag_cumavg1_L1 + lag_
cumavg1_L2 + + L3 + t0,
L2 ~ lag1_A + L1 + lag_cumavg1_L1 + lag_cumavg1_L2 + L3 + t0,
A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 + lag_cumavg1_L2 + L3 + t0)),
  histories = c(lagged, lagavg),
  histvars = list(c('A', 'L1', 'L2'), c('L1', 'L2')),
```

```
basecovs = "L3",
outcome_name = "Y",
outcome_type = "survival",
ymodel = Y ~ A + L1 + L2 + L3 + lag1_A + lag1_L1 + lag1_L2 + t0,
intvars = intvars2,
int_descript = int_descript2,
interventions = interventions2,
ref_int = 3,
nsamples      = 20,
ci_method     = "percentile",
seed          = 12345678)
```

다음은 위에서 적합한 g-formula의 결과를 기초로 하여 등고선 그림을 그려주는 코드이다. 개발한 함수 ContourPlot은 g-formula의 적합 결과를 그대로 함수에 입력하여 사용이 가능하며, 그림을 꾸미기 위해 다음의 인수들(xlab, ylab, vertical, lab_title_size, lab_text_size, main_size, textNum, textwidth)을 지정하는 것이 필요하다. Dose는 list의 형태로 등고선 그림을 그리기 위한 두 변수의 범위를 지정할 수 있다. xlab과 ylab 인수를 통해 등고선 그림의 x-축과 y-축에 개입의 대상이 되는 변수의 이름을 입력할 수 있으며, 등고선 그림에서의 x-축, y-축 그리고 제목의 글자 크기는 lab_title_size, lab_text_size, main_size 인수를 통해 조정할 수 있다. textNum 인수를 통해 결과 값의 변화에 따른 색 변화의 정도를 조절할 수 있으며, textwidth 인수는 등고선 그림 위 결과 값을 어느 정도의 단위로 표시하고자 하는지 결정하는 인수이다. 마지막으로 vertical 인수를 통해 등고선 그림을 세로로 표현할지 정할 수 있다. 아래의 등고선 그림에서는 textwidth를 0.2로 설정하였기 때문에 결과 값이 0.2 단위로 표시되는 것을 확인할 수 있다.

```
# 개발한 함수를 이용하여 등고선 그림
Contour_Plot <- ContourPlot(
  object = gformRes_Contour,
  Dose = list(Dose1, Dose2),
  xlab = "L2",
  ylab = "A",
  vertical = FALSE,
  lab_title_size = 17,
  lab_text_size = 15,
  main_size = 15,
  textNum = 100,
  textwidth = 0.2)

# 등고선 그림 출력
Contour_Plot$Plot; Contour_Plot$ColorLegend
```

함수의 결과물을 통해 등고선 그림과 색의 변화에 대응되는 반응 값의 범위(ColorLegend)를 얻을 수 있다. 아래의 등고선 그림을 통해 변수 'L2'와 변수 'A'의 값이 증가함에 따라 위험이 감소함을 알 수 있다.


Lower limit of 95% CI
1.00
0.75
0.50
0.25
0.00
A
0.8
1
1.2
0.00
0.25
0.50
0.75
1.00
L2
Estimate
1.00
0.75
0.50
0.25
0.00
A
0.8
1
1.2
1.4
0.00
0.25
0.50
0.75
1.00
L2
Upper limit of 95% CI
1.00
0.75
0.50
0.25
0.00
A
1
1.2
1.4
0.00
0.25
0.50
0.75
1.00
L2

마지막으로 교호 작용을 표현하는 함수에 대해서 설명하고자 한다. 용량-반응 곡선을 그리는 함수의 사용법을 기술할 때, 사용하였던 예시 자료 basicdata_nocomp를 사용하여 교호작용을 표현하는 함수의 사용법을 설명하고자 한다. 교호 작용을 간단히 설명하기 위해 시간에 따라 변하는 변수는 모두 이항 변수인 변수 'L1'과 변수 'A'에 대해 개입을 하고자 한다. 이항 변수를 다루고 있기 때문에 additive interaction는 두 변수 모두 1로 개입되었을 때의 위험의 크기(R11)와 두 변수 모두 0로 개입되었을 때의 위험의 크기(R00)의 합에서 변수 'L1'은 1로 개입, 변수 'A'는 0으로 개입되었을 때의 위험의 크기(R10)와 변수 'L1'은 0으로 개입, 변수 'A'는 1로 개입되었을 때의 위험의 크기(R01)의 합을 빼는 것으로 정의되며, multiplicative interaction은 R11과 R00의 곱을 R10과 R01의 곱으로 나눈 값으로 정의되고, 마지막으로 RERI는 additive interaction을 R00으로 나눈 값으로 정의된다. 개발한 함수를 통해 다양한 교호 작용의 효과를 추정하고, 그림으로 표현이 수월한 additive interaction의 경우 그림으로 표현하고자 한다.

교호 작용의 크기를 g-formula를 적용하여 산출하기 전 필요한 입력사항(교호 작용을 보고자 하는 변수의 값, 개입 전략의 형태 및 개입 전략의 이름)을 먼저 기술하여야 한다.

```
# 개입하고자 하는 값 지정 및 가능한 조합 생성
Dose3 <- Dose4 <- 0:1
DoseCombination3 <- expand.grid(
  Dose2 = Dose4, Dose1 = Dose3,
  KEEP.OUT.ATTRS = FALSE, stringsAsFactors = FALSE)[,2:1]

# 개입의 대상이 되는 변수 지정 및 개입 전략 지정
intvars3 <- interventions3 <- list()
for(i in 1:nrow(DoseCombination3)){
  intvars3[[i]] <- c("L1","A")
```

```
  interventions3[[i]] <- list(
  c(static, rep(x = DoseCombination3$Dose1[i], times = 7)),
  c(static, rep(x = DoseCombination3$Dose2[i], times = 7)))
}; remove(i)

# 개입 전략 명 지정
int_descript3 <- c()
for(i in 1:nrow(DoseCombination3)){
  int_descript3[i] <- paste0("L1 / A"," [",DoseCombination3$Dose1[i],",",
DoseCombination3$Dose2[i],"]")
}; remove(i)
```

이 사항들을 기술한 후, InteractionPlot을 사용하여 additive interaction, multiplicative interaction 그리고 RERI를 추정할 수 있다.

개발한 함수 InteractionPlot를 통해 교호 작용의 크기를 산출하고, 그림을 그릴 때, g-formula를 적합하기 위해 필요한 인수들이 동일하게 사용되며, 추가적으로 DoseCombination, alpha, nboots 인수를 필수적으로 입력하여야 한다. DoseCombination은 교호 작용의 크기를 구하기 위해 어떤 값이 변수의 개입 값으로 사용되었는지 입력하는 인수이며, data.frame의 형태로 입력이 되어야 하며, 유의 수준(alpha)와 붓스트랩 수(nboots)는 신뢰 구간을 추정할 때 사용되는 인수이다. 그림을 꾸미기 위해서는 다음의 인수들(xlab, vertical, main_size, lab_title_size, lab_text_size, label_size)에 대해 지정해주는 것이 필요하다. xlab은 교호 작용 그림에서 x-축에 표시할 변수 명을 의미하며, 교호작용의 경우 관심 있는 변수가 2개이기 때문에 길이가 2인 벡터가 입력되어야 한다. vertical은 각 변수에 대한 교호 작용 그림의 배치 형태를 지정할 수 있다(vertical = TRUE는 세로 형태). lab_title_size, lab_text_size, label_size을 통해 축 제목의 크기, 축 글자 크기 그리고 교호작용의 크기를 표현한 글자의 크기를 조절할 수 있다.

```
# 개발한 함수를 이용하여 교호 작용의 크기 추정 및 표현하기 하는 그림
Interaction_Plot <- InteractionPlot(
  obs_data = basicdata_nocomp,
  id = "id",
  time_points = timepoints3,
  time_name = "t0",
  covnames = c('L1', 'L2', 'A'),
  covtypes = c('binary', 'bounded normal', 'binary'),
```

```
  covparams = list(covmodels = c(L1 ~ lag1_A + lag_cumavg1_L1 + lag_
cumavg1_L2 + + L3 + t0,
L2 ~ lag1_A + L1 + lag_cumavg1_L1 + lag_cumavg1_L2 + L3 + t0,
A ~ lag1_A + L1 + L2 + lag_cumavg1_L1 + lag_cumavg1_L2 + L3 + t0)),
histories = c(lagged, lagavg),
  histvars = list(c('A', 'L1', 'L2'), c('L1', 'L2')),
  basecovs = "L3",
  outcome_name = "Y",
  outcome_type = "survival",
  ymodel = Y ~ A + L1 + L2 + L3 + lag1_A + lag1_L1 + lag1_L2 + t0,
  intvars = intvars3,
  int_descript = int_descript3,
  interventions = interventions3,
  ref_int = 1,
  baselags = TRUE,
  alpha = 0.05,
  nboots = 10,
  seed = 12345678,
  DoseCombination = DoseCombination3,
  xlab = c("L1","A"),
```

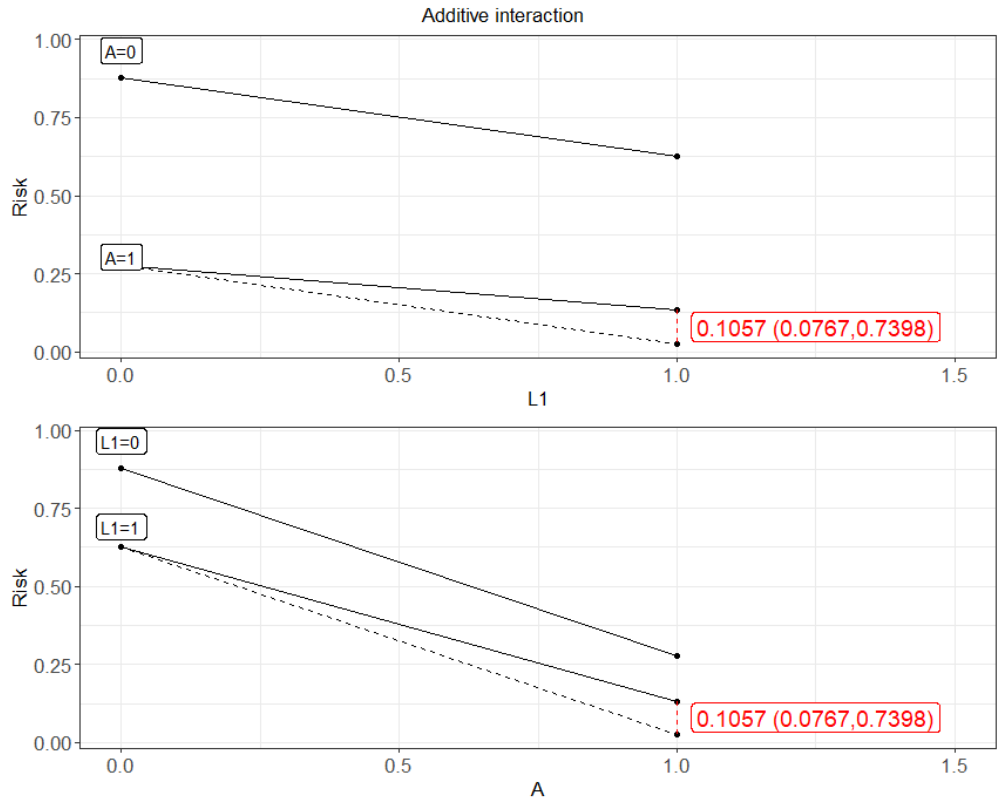
```
  vertical = TRUE,
  main_size = 13,
  lab_title_size = 13,
```

```
  lab_text_size = 13,
  label_size = 5)

# 다양한 교호 작용의 크기 및 additive interaction을 표현하는 그림 출력
Interaction_Plot$summary; Interaction_Plot$Plots
```

함수의 결과물을 통해 다양한 교호작용의 크기(Interaction_Plot$summary)와 additive interaction의 그림(Interaction_Plot$Plots)을 얻을 수 있다. 자료 basicdata_nocomp에서 결과 변수에 대한 이항 변수 L1과 이항 변수 A의 additive interaction, multiplicative interaction 그리고 RERI의 크기는 0.1057, 0.6710 그리고 0.1206으로 나타났으며, 각각의 95% 신뢰 구간(alpha=0.05)은 (0.0767,0.1163), (0.6294, 0.7398) 그리고 (0.0864, 0.1300)으로 나타났다(분석 시간을 단축하기 위해 현재 예시 자료에서는 붓스트랩의 수를 10으로 설정하였음(nboots=10)).

```
Interaction_Plot$summary
                                Est LowerCI UpperCI
Additive interaction         0.1057  0.0767  0.1163
Multiplicative interaction 0.6710  0.6294  0.7398
RERI                              0.1206  0.0864  0.1300
```

각 물질을 기준으로 additive interaction에 대한 교호 작용 그림은 아래와 같다.



위쪽 그림에서는 'A'의 수준 별(A=0, A=1)로 변수 'L1' 값에 따라 나타나는 위험의 크기를 보여주고 있다. 점선은 하나의 실선 (A=0)을 다른 쪽(A=1)으로 평행하게 옮겼을 때의 결과로 생각할 수 있고, 이 두 선이 일치하지 않으므로 변수 'A'의 수준에 따라 변수 'L1'이 미치는 위험의 크기가 변한다는 것을 알 수 있다. 또한, 붉은 글씨로 강조된 교호작용의 크기 및 신뢰 구간은 0.1057, 95% 신뢰 구간은 (0.0767,0.1163) 임을 알 수 있다. 두 번째 물질을 기준으로 한 아래쪽 그림에서도 비슷한 해석이 가능하다.

Ⅱ. BKMR의 통계분석법 개발

본 부록에서는 분석 속도가 대폭 개선된 BKMR 방법과 반복 측정된 자료에서 기울기에 랜덤 효과가 허용된 BKMR 방법을 자료에 적합할 수 있도록 하는 함수의 사용법에 대해서 설명하고자 한다. BKMR의 기본 이론과 R 패키지에 대한 설명은 Bobb JF 등(2015), Bobb JF 등(2018)의 논문, 예신희 등(2022)이 작성한 산업안전보건연구원의 연구보고서와 Bobb JF이 2017년에 GitHub에 작성한 'bkmr' R 패키지 소개를 참고하기 바란다. 두 모형의 적합에 필요한 R 패키지는 https://sites.google.com/view/lwj221에서 다운 받아 설치가 가능하며, 연구에서 사용된 R 버전은 4.0.2이며, CPU는 AMD Ryzen 3700X이다.

1. 분석 속도가 대폭 개선된 BKMR 방법

R 프로그램의 'vbayesGP' 패키지에는 기존 BKMR 방법보다 분석 속도가 대폭 개선된 BKMR 방법을 적합할 수 있는 함수, 즉 gvagpr() 함수가 포함되어 있으며 모형 적합(fitting)에 있어 가장 중요한 함수이다. 또한, 적합된 결과를 요약하고, 진단하며, 직관적으로 해석할 수 있도록 돕는 함수들이 내장되어 있다. 다음은 R 프로그램의 'vbayesGP' 패키지를 설치하고 불러오는 코드이다.

```
install.packages("vbayesGP")
library(vbayesGP)

#'vbayesGP' R 패키지가 CRAN에 업로드 되어 있지 않아 다운 받아 설치하는 경우에는 다음과 같이 설치하기
install.packages("D://CI example//vbayesGP_0.2.0.zip", repos = NULL, type="source")
library(vbayesGP)
```

모의실험 자료를 생성하여 R 패키지 'vbayesGP'에서 제공하는 함수들의 사용법에 대하여 설명하고자 한다. 모의실험 자료를 생성하는 함수 SimData()는 기존 BKMR 방법의 적합이 가능한 R 패키지 'bkmr'에서 사용이 가능하다. SimData() 함수에서 n는 표본 수(sample size), M은 유해물질의 수, sigsq.true은 정규 분포를 가정한 경우에서 에러의 분산, beta.true은 공변량의 계수, hfun은 참 노출-반응 함수(true exposure-response function)의 함수 형태(1: 첫 번째 유해물질에 대한 비선형 함수, 2: 처음 두 유해물질에 대한 선형 함수 및 두 유해물질의 곱을 포함하는 함수, 3: 처음 두 유해물질에 대한 비선형 그리고 비가법 함수), Zgen은 유해물질을 생성하는 방법, ind은 참 노출-반응 함수에 포함될 유해물질의 번호, family은 에러 변수의 분포를 의미한다. 또한, 결과의 재현성을 위해 난수 번호는 20230816을 사용하였다. 생성된 모의실험 자료 dat에서 주요하게 사용되는 것은 결과 변수를 의미하는 벡터 'y', 유해물질을 의미하는 행렬 'Z', 공변량(covariate)을 의미하는 행렬 'X' 그리고 참 노출-반응 함수 값인 벡터 'h'이다.

```
set.seed(seed = 20230816)
dat <- bkmr::SimData(n = 500, M = 5, sigsq.true = 0.5, beta.true = 2,
hfun = 3, Zgen = 'norm', ind = 1:2, family = 'gaussian')

y <- dat$y
Z <- dat$Z
M <- ncol(Z)
X <- dat$X
true.h <- dat$h
```

개발된 방법에 대한 소개 및 비교를 하기에 앞서 기존 제안된 BKMR 방법을 먼저 적합하고, 수행에 필요한 시간을 측정하고자 한다. 또한, 기존 BKMR은 모수를 추정할 때, 마코프 체인 몬테 카를로(Markov chain Monte Carlo;

MCMC) 알고리즘을 사용하며, 모수 추정치의 수렴을 위해 알고리즘을 얼마나 반복시킬지 횟수를 지정하는 것이 필요하다. MCMC 알고리즘을 실행시킬 때 필요한 반복 횟수는 iter 인수(argument)를 사용하여 10,000번의 반복을 지정하였다.

```
bkmr.time <- system.time({
  fout.bkmr <- bkmr::kmbayes(y = y, Z = Z, X = X, iter = 10000, rmeth
od = 'varying', varsel = TRUE, verbose = FALSE)
})[3]

summary(fout.bkmr)
bkmr.h <- bkmr::ComputePostmeanHnew(fout.bkmr)$postmean
```

R 패키지 'bkmr'에서 BKMR 방법의 적합을 위해 필요한 kmbayes 함수를 모의실험 자료에 적용하여 아래의 같은 결과를 얻었으며, 결과를 제공받기까지 약 14분이 소요되었다.

```
Fitted object of class 'bkmrfit'
Iterations: 10000
Outcome family: gaussian
Model fit on: 2023-08-03 12:17:14.814957
Running time:  13.94919 mins
```

```
Parameter estimates (based on iterations 5001-10000):
   param      mean       sd    q_2.5    q_97.5
1   beta   2.02194  0.01562  1.99161   2.05307
2 sigsq.eps  0.49240  0.03164  0.43489   0.55913
3     r1   0.05360  0.02878  0.01562   0.12200
4     r2   0.04207  0.02459  0.01272   0.10624
```

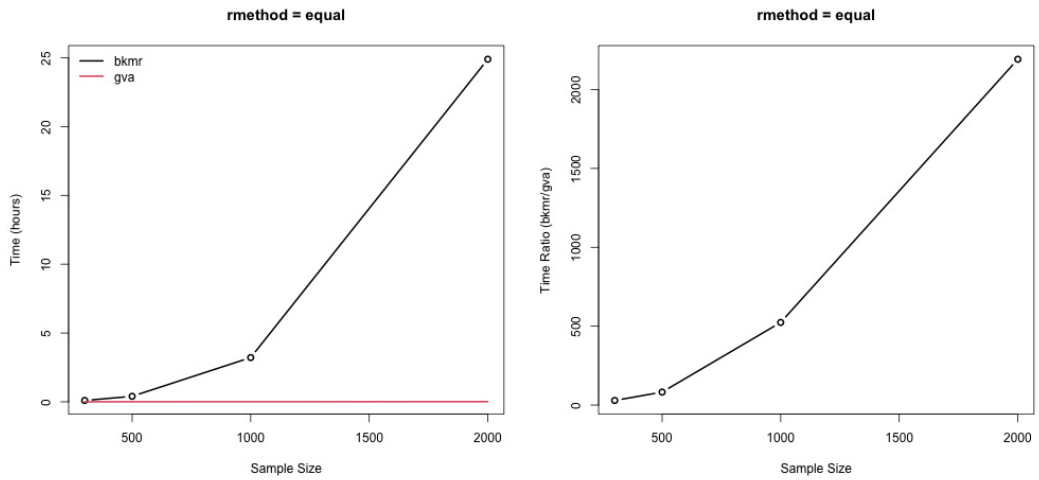
```
5          r3  0.00122 0.00398 0.00000  0.01386
6          r4  0.00000 0.00025 0.00000  0.00000
7          r5  0.00003 0.00063 0.00000  0.00000
8      lambda 12.65394 9.13997 3.42875 37.78383
```

BKMR의 적합에 필요한 분석 시간의 단축을 위해 새로 개발한 BKMR 방법은 R 패키지 'vbayesGP'를 통해 적합이 가능하다. 분석 시간을 단축시키기 위한 방법, 즉 사후 분포를 근사하기 위해 사용한 방법으로 (1) 기존 BKMR 방법에서 사용한 변수 선택 사전 분포 대신 horseshoe 축소 사전 분포의 사용, (2) 성김 구조의 촐레스키 요인의 가정 그리고 (3) mini-batch 확률적 경사법을 사용하였으며, 이 방법들을 함수에 입력하기 위한 코드는 다음과 같다. priors에서 lengthscale 인수를 통해 horseshoe 축소 사전 분포의 지정이 가능하며, 그 외의 인수 asig, bsig, alam, blam, lam, tau는 축소 사전 분포의 지정을 위해 필요한 초모수(hyperparameter)이다.

```
priors <- list(lengthscale='horseshoe', asig=0.001, bsig=0.001, alam=10,
blam=1, lam0=1, tau0=1)
control <- list(nbatch = 3)
```

```
mvbmbu.time <- system.time({
  foutmbu.diag <- vbayesGP::gvagpr(y, X, Z, priors = priors, covstr = 'di
agonal', control = control, minibatch = TRUE)
})[3]

summary(foutmbu.diag)
mvbmb.h <- fitted(foutmbu.diag)$fmean
```

BKMR을 적합하기 위해 'vbayesGP' 패키지를 사용하였을 때, 소요된 시간은 약 25초가 소요되었으며, 이는 'bkmr' 패키지를 사용하여 BKMR을 적합하였을 때 소요된 시간의 약 1/33에 해당하는 시간, 즉 33배 빠른 속도이다. 다음은 표본 수에 따른 기존 BKMR 방법과 새로 개발한 BKMR 방법의 분석 시간을 비교한 그림이다. 표본 수가 적을 경우, 두 방법간의 차이는 극적이지는 않지만, 표본 수가 2,000인 경우를 보면 기존 BKMR 방법과 비교하여 약 2,000배 빠른 것을 확인할 수 있다. 이러한 결과를 통해 분석 시간으로 인하여 기존 BKMR 방법의 적용이 어려웠던 표본의 수가 큰 자료에서 개발한 함수를 통해 BKMR의 적용하여 제한된 시간 내 산업 보건 역학연구가 가능하다.



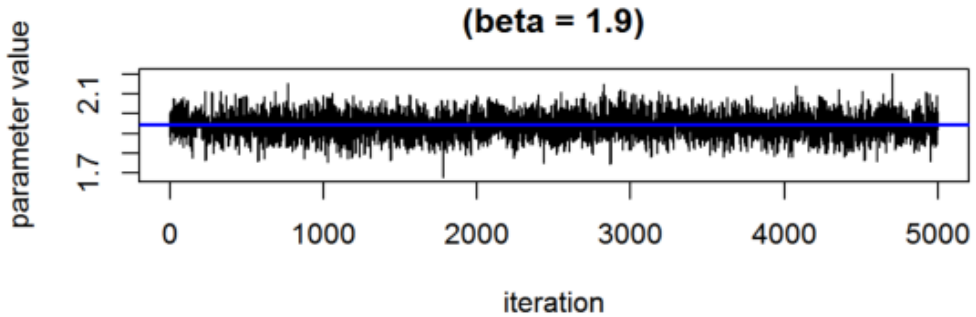
또한, 참 노출-반응 함수 값의 값을 추론하기 위해 기존 'bkmr' 패키지에서는 ComputePostmeanHnew()을 사용하여야 했으나, 본 연구에서는 산업보건 역학 연구자들이 선형 회귀 분석에서 익숙하였던 함수 fitted()를 적용할 수 있도록 하였다. 'vbayesGP' 패키지를 사용하여 적합한 결과는 아래와 같다.

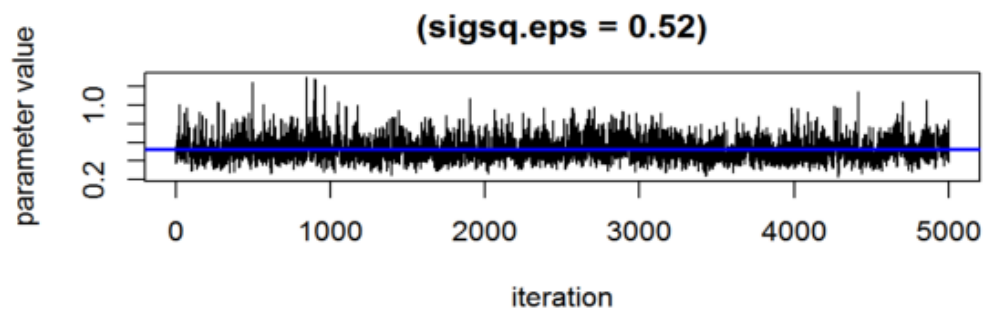
```
Fitted object of class 'gpr'
Outcome family: gaussian
Covariance Strucutre: diagonal
Lengthscale Parameter: varying and shrinkage
Minibatch-Epoch: 100,000
Running time: 1326.89 secs
Model fit on: 2023-11-04 23:31:31
```

```
Parameter estimates:
                mean        sd     q_2.5       q_50     q_97.5
beta         2.01770 0.01727  1.98601   2.01783   2.05056
sigsq.eps    0.55374 0.03502  0.48750   0.55374   0.62242
lambda      10.69106 1.55571  8.00172  10.61979  14.14561
r1           0.02699 0.00664  0.01632   0.02617   0.04137
r2           0.03304 0.00873  0.01946   0.03178   0.05209
r3           0.00000 0.00000  0.00000   0.00000   0.00000
r4           0.00000 0.00000  0.00000   0.00000   0.00000
r5           0.00000 0.00000  0.00000   0.00000   0.00000

Pseudo posterior inclusion probabilities:
  variable  PPIP
1       z1 0.956
2       z2 0.999
3       z3 0.000
4       z4 0.000
5       z5 0.000
```

기존 'bkmr' 패키지의 경우, 모수의 추정치가 안정적으로 수렴하는지 TracePlot()을 사용하여 확인이 가능하며, par 인수를 통해 어떤 모수에 대하여 trace plot을 그릴지 지정할 수 있다. 아래의 코드는 모의실험 자료를 통해 적합한 기존 BKMR 방법의 모수(beta, sigma square)들이 안정적으로 수렴하는지 확인하는 코드이다.

```
TracePlot(fit = fout.bkmr, par = "beta")
TracePlot(fit = fout.bkmr, par = "sigsq.eps")
```

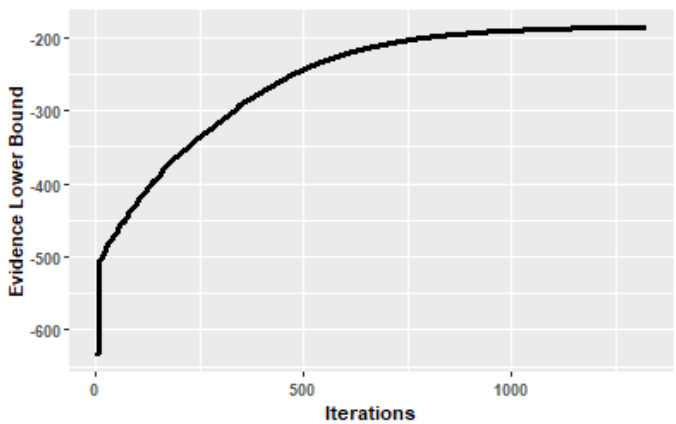




분석 속도를 개선한 본 연구에서의 BKMR 방법의 경우, 사후 분포를 변분 근사 알고리즘을 사용하여 근사하였기 때문에 MCMC를 사용한 기존 BKMR 방법에서 사용한 trace plot와 달리 expected lower bound(ELBO)를 통해 모수 추정치의 수렴 여부를 점검한다. ELBO 값은 extractELBO() 함수를 통해 추출이 가능하며, plot() 함수를 통해 시각적으로 확인이 가능하다. 또한, 기존

BKMR의 trace plot은 모수의 추정치가 수렴했는지 그림을 통해 확인을 해야 하며, 수렴 여부와 무관하게 함수에 지정한 반복횟수를 모두 수행하기 때문에 추정치가 수렴하였을 경우, 수렴 이후 불필요한 계산이 수반된다. 하지만 개발된 모형의 경우, 추정치의 차이가 함수에서 제공하는 적절히 작은 값 이내로 들어오면 수렴 알고리즘을 멈추도록 설계되어있어 지정한 반복횟수 이전에 추정치가 수렴하였을 경우, 더 이상 계산이 진행되지 않아 분석 시간의 단축시킬 수 있다. 기존 BKMR 방법의 경우, 현재 지정한 10,000번이 모두 수행되었으며, 개발된 모형의 경우, 2,158번의 반복 이후 추정치가 수렴하였기 때문에 그 이후에는 계산이 진행되지 않았다.

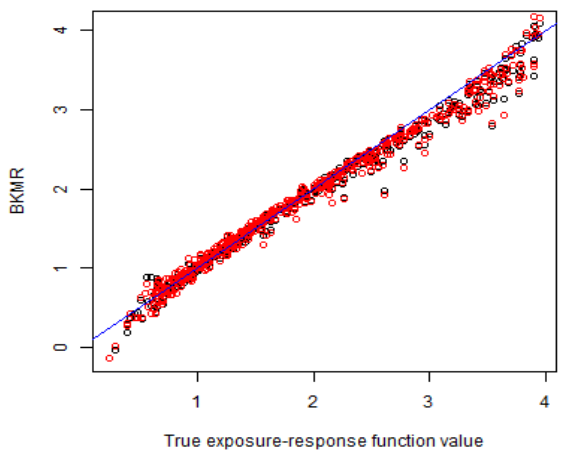
```
extractELBO(foutmbu.diag)
plot(foutmbu.diag)
```

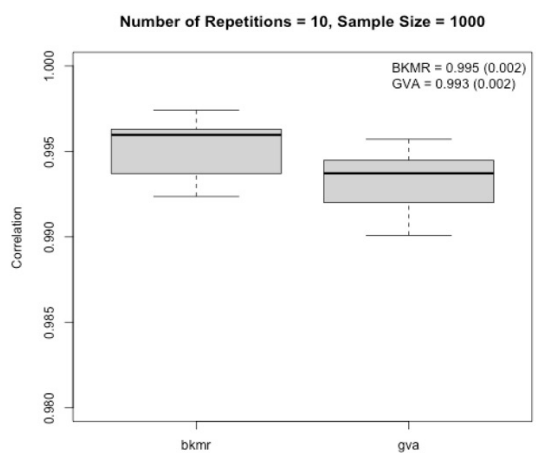


기존 BKMR 방법과 분석 속도가 개선된 BKMR 방법으로부터 얻어진 노출-반응 함수 값을 참 값과 비교한 그림(검은 점: 기존 BKMR 방법, 빨간 점: 개발된 BKMR 방법, 파란 선: y = x)은 아래와 같다. 그림을 보면 새로 개발된 BKMR 방법에서 얻어진 노출-반응 함수 값의 추정치(빨간 점)가 y=x 함수 위에 대부분 위치하고 있는 것을 확인할 수 있으며, 이는 개발된 방법의 추정치가 높은 정확도를 가지고 있음을 알 수 있다. 또한, 두 방법의 추정치의

정확도를 참 값과의 correlation을 비교하여서도 확인이 가능하다(bkmr은 기존 BKMR 방법, gva는 새로 개발한 BKMR 방법을 의미함). 이로부터 사후 분포를 근사하는 본 연구의 방법이 기존 BKMR 방법과 매우 유사한 결과를 주고 있음을 확인할 수 있다.

```
# 두 모형의 추정치와 참 노출-반응 함수 값을 비교한 그림
plot(x = true.h, y = bkmr.h, xlab = "True dose-response function value",
ylab = "BKMR")
points(x = true.h, y = mvb.h, col = "red")
abline(a = 0, b = 1, col = "blue")
```





2. 반복 측정된 자료에서 기울기에 랜덤 효과가 허용된 BKMR 방법

R 프로그램의 'vbayesGP' 패키지에는 앞서 설명한 기존 BKMR 방법과 비교하여 분석 속도가 대폭 개선된 BKMR 방법뿐만 반복 측정된 자료에서 기울기에 랜덤 효과를 허용하는 BKMR 방법 또한 gvagpr() 함수를 통해 적합이 가능하다. 랜덤 기울기를 허용하는 BKMR 방법에 대해서도 모의실험 자료를 생성하여 함수의 사용법에 대해 설명하고자 한다.

```
Hfun3 <- function (z, ind = 1:2) {
  4 * plogis(1/4 * (z[,ind[1]] + z[,ind[2]] + 1/2 * z[,ind[1]] * z[,ind[2]]), 0,
0.3)
}

N <- 100; R <- 3; D <- N * R; M <- 5; p <- 1; q <- 2

beta.true <- 2
sigsq.true <- 0.5
SIGMAb.true <- matrix(c(0.5, 0.1, 0.1, 0.3), q, q)

set.seed(1)
b.true <- MASS::mvrnorm(n=N, mu = rep(0, q), Sigma = SIGMAb.true)
bvec <- as.vector(t(b.true))
ID <- rep(1:N,each=R)
```

```
Z <- matrix(rnorm(D * M), D, M);colnames(Z) <- paste0("z", 1:M)
X <- matrix(rnorm(D * p), D, p)
U <- matrix(0, D, N*q)
for (i in 1:N) {
  U[((i-1)*R+1):(i*R),((i-1)*q+1):(i*q)] <- cbind(1, X[((i-1)*R+1):(i*R)])
}

true.h <- Hfun3(Z)
y <- X %*% beta.true + true.h + U %*% bvec + rnorm(D, sd = sqrt(sigs
q.true))
```

위와 같이 생성된 모의실험 자료에 기울기에 랜덤 효과를 허용하는 BKMR 방법을 다음과 같이 적합할 수 있다. gvagpr() 함수 또한 기존 kmbayes() 함수와 동일하게 반복 측정된 자료가 같은 근로자로부터 측정되었다는 것을 지정하기 위해 id 인수에 해당 정보를 입력하도록 개발되었다. 또한, 개발된 BKMR 방법의 계수의 수렴 여부를 확인하기 위해 사용하였던 함수 plot() 또한 기울기에 랜덤 효과를 허락하는 BKMR 방법에서도 사용 가능하다.

```
priors <- list(lengthscale='horseshoe', asig=0.001, bsig=0.001, alam=10,
blam=1, lam0=1, tau0=1, vtaub=100)

mvb.time <- system.time({
  fout.diag <- vbayesGP::gvagpr(y, X, Z, id = ID, random.slope = 1, prio
rs = priors, covstr = 'diagonal')
})[3]
```

```
summary(fout.diag)
mvb.h <- fitted(fout.diag)$fmean
```

아래의 결과는 기울기에 랜덤 효과를 허락한 BKMR 방법의 적합 결과이다.

```
Fitted object of class 'gpr'
Covariance Strucutre: diagonal
Lengthscale Parameter: varying and shrinkage
Random Effects: random slope
Iterations: 6030
Running time: 103.5 secs
Model fit on: 2023-08-05 16:10:58.545834

Parameter estimates:
              mean       sd    q_2.5    q_97.5
beta       1.96565  0.04383  1.87368   2.05473
sigsq.eps  0.59236  0.04652  0.50974   0.68862
lambda     8.68243  2.41543  4.80764  14.58592
r1         0.04750  0.01816  0.02151   0.09126
r2         0.03207  0.01541  0.01161   0.07134
r3         0.00277  0.00417  0.00025   0.01130
r4         0.00280  0.00314  0.00034   0.01133
r5         0.00190  0.00241  0.00016   0.00908
```

```
Variance estimates (random effects):
   [,1]    [,2]
[1,] 0.71982 0.12223
[2,] 0.12223 0.36235
```

3. BKMR의 로지스틱 회귀 모델로의 확장

새로 개발한 'vbayesGP' 패키지에는 기존 R 패키지 'bkmr'에서는 다룰 수 없었던 로지스틱 BKMR 방법을 계산해주는 gvaggpr() 함수가 포함되어 있으며 모형 적합에 있어 가장 중요한 함수이다. 또한, 적합된 결과를 요약하고, 진단하며, 직관적으로 해석할 수 있도록 돕는 함수들이 내장되어 있다.

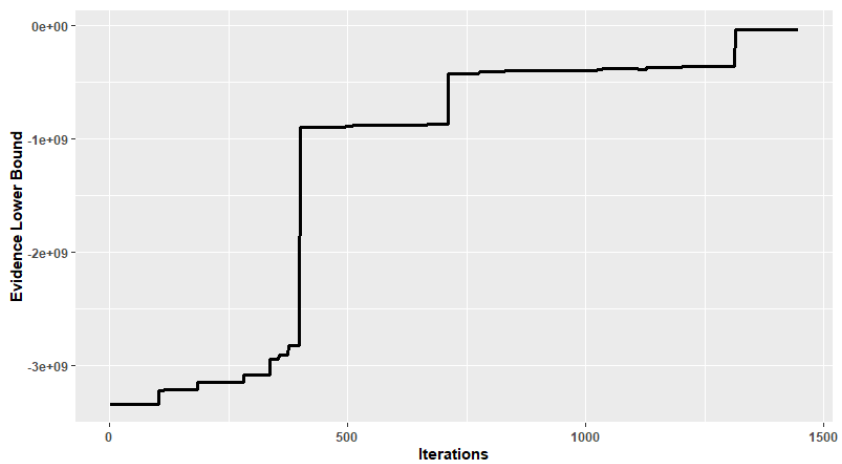
모의실험 자료를 생성하여 함수의 사용법에 대해 설명하고자 한다.

```
# 모의실험 자료 생성
seed <- 111
set.seed(seed)
n <- 200; M <- 4
beta.true <- 0.1
Z <- matrix(runif(n * M, -1, 1), n, M)
X <- as.matrix(3*cos(Z[, 1]) + 2*rnorm(n))
hfun0 <- function(z) (2*z + 0.5)^2
hfun <- function(zvec) hfun0(zvec[1]) + hfun0(zvec[2]) - hfun0(zvec[3]) -
hfun0(zvec[4]) + zvec[3]*zvec[4]
h <- apply(Z, 1, hfun) ## only depends on z1, z2, z3, z4
ystar <- X %*% beta.true + h
prob <- 1 / (1 + exp(-ystar))
y <- rbinom(n = n, size = 1, prob = prob)
datp <- list(n = n, M = M, beta.true = beta.true, Z = Z, h = h, X = cbind
(X), y = y, ystar = ystar)
```

```
# 로지스틱 BKMR의 적용
fit_logisticBKMR <- vbayesGP::gvaggpr(
y = datp$y,
Z = datp$Z,
X = datp$X,
priors = list(lengthscale = 'normal'),
family = 'binomial')
```

위의 코드와 같이 생성된 모의실험 자료를 통해 이항 자료에 대해 BKMR 방법을 프로빗 모델이 아닌 로지스틱 모델을 통해 적합할 수 있다. 결과 변수 y의 형태를 지정하는 인수인 family를 제외한 나머지 모든 인수는 gvagpr() 함수와 동일하다. 결과 변수가 이항 자료인 반복 측정된 자료 또한 본 연구에서 개발된 gvaggpr 함수를 통해 분석이 가능하며, 개발된 로지스틱 BKMR 방법에서 계수의 수렴 여부를 확인하기 위해 사용하였던 함수 fitted()와 plot(), 기존 BKMR에서 적합된 결과를 확인하기 위해 사용하였던 함수 summary() 모두 로지스틱 BKMR 방법에서도 사용 가능하다.

```
plot(fit_logisticBKMR)
```



gvagpr 함수에서와 같은 방식으로 위의 ELBO 그림을 통해 로지스틱 BKMR에서 모수 추정치들이 잘 수렴했음을 확인할 수 있다. 아래의 결과는 summary 함수를 적용하였을 때 출력되는 로지스틱 BKMR의 결과물이다.

```
summary(fit_logisticBKMR)

Fitted object of class 'gpr'
Outcome family: binomial ( logit )
Covariance Strucutre: diagonal
Lengthscale Parameter: equal
Iterations: 3947
Running time: 16.5 secs
Model fit on: 2023-11-03 10:23:05
```

```
Parameter estimates:
             mean       sd     q_2.5     q_50   q_97.5
beta    0.03880 0.05122 -0.06392 0.03757 0.14231
lambda 2.21210 2.32495  0.31328 1.53661 8.30152
r       1.91242 1.49731  0.40541 1.53330 5.71578
```

위의 summary 결과를 통해 결과 변수는 이항 변수이며, link 함수로 logit을 사용하고 있음을 확인할 수 있고, 모수 추정치들이 3765번째에서 수렴하였으며, 계산 시간이 16.5초가 소요되었다는 것을 확인할 수 있다(표본 수는 200, 다중 노출 수는 4개일 때). 아래의 코드는 동일한 모의실험 자료로 프로빗 BKMR을 수행하는 코드이다.

```
Fit_bkmr <- kmbayes(y = datp$y, Z = datp$Z, X = datp$X,
                    iter = 10000, verbose = FALSE,
                    varsel = TRUE, family = "binomial",
                    control.params = list(r.jump2 = 0.5))
```

프로빗 BKMR에 summary 함수를 적용한 결과이다.

```
summary(fit_bkmr)
```

아래의 결과를 보면 이항 변수에 프로빗 연결 함수를 사용하였으며, 프로빗 BKMR을 수행하는 데 약 5분이 소요된 것을 확인할 수 있다.

```
Fitted object of class 'bkmrfit'
Iterations: 10000
Outcome family: binomial (probit link)
Model fit on: 2023-11-03 10:18:33
Running time:  5.21231 mins

Acceptance rates for Metropolis-Hastings algorithm:
              param      rate
1            lambda 0.4813481
2 r/delta (overall) 0.1230123
3 r/delta (move 1) 0.0000000
4 r/delta (move 2) 0.2478839

Parameter estimates (based on iterations 5001-10000):
        param      mean        sd    q_2.5    q_97.5
1        beta   0.13274   0.06652  0.01054   0.26588
2   sigsq.eps   1.00000   0.00000  1.00000   1.00000
3          r1   0.25736   0.12351  0.08547   0.57846
4          r2   0.26290   0.13072  0.07021   0.50007
5          r3   0.22047   0.13068  0.06148   0.60009
6          r4   0.25971   0.14033  0.08489   0.74028
7      lambda 19.52689  11.01608  5.16946  47.40877
```

```
8       ystar1   1.38290   0.85294 0.08577   3.23668
9       ystar2   1.18105   0.87089 0.04909   3.26225
10   ystar200   2.77865   1.17694 0.57938   5.13504
```

```
Posterior inclusion probabilities:
  variable PIP
1       z1   1
2       z2   1
3       z3   1
4       z4   1
```

summary 함수 외에도 R 패키지 'bkmr'에서 사용 가능하였던 함수 predictor ResponseBivar(), predictorResponseBivarLevels(), predictorResponseB ivarPair() 그리고 predictorResponseUnivar() 등은 R 패키지 'vbayesGP'에서도 동일하게 동일한 목적으로 사용이 가능하다.

부록 1의 참고문헌

이슬비. 임신 중 복합 환경유해물질 노출이 6 개월 영유아 아토피 피부염 발생에 미치는 영향. 2019.

예신희, 이상길, 이지혜, 이경은, 성정민, 김민수. 저농도 복합유해물질 노출과 혈액검사 이상 관련성 탐색 연구. 산업안전보건연구원. 2020.

예신희, 이경은, 성정민, 박동준, 이우주. 직업병 인과추론 가이드라인 및 통계분석법 개발(1): -g methods 국문 가이드라인 개발. 산업안전보건연구원. 2021.

예신희, 이경은, 윤민주, 박동준, 마성원, 이영신, 이우주. 직업병 인과추론 가이드라인 및 통계분석법 개발(2): 복합노출의 건강 영향평가 국문 가이드라인 개발. 산업안전보건연구원. 2022.

Bobb JF. “Introduction to Bayesian kernel machine regression and the bkmr R package”, GitHub, 2017년 3월 24일 작성. 2022년 7월 10일 접속. URL https://jenfb.github.io/bkmr/overview.html#estimated_posterior_inclusion_probabilities.

Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, Godleski JJ, Coull BA. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015 Jul;16(3):493-508.

Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. Environ Health. 2018 Aug 20;17(1):67.

Keil AP, Richardson DB. Reassessing the Link between Airborne Arsenic Exposure among Anaconda Copper Smelter Workers and Multiple Causes of Death Using the Parametric g-Formula. Environ Health Perspect. 2017 Apr;125(4):608-614. doi: 10.1289/EHP438. Epub 2016 Aug 19. PMID: 27539918; PMCID: PMC5381993.

Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics. 2007 Dec;63(4):1079-88. doi: 10.1111/j.1541-0420.2007.00799.x. PMID: 18078480; PMCID: PMC2665800.

Neophytou AM, Costello S, Picciotto S, Brown DM, Attfield MD, Blair A, Lubin JH, Stewart PA, Vermeulen R, Silverman DT, Eisen EA. Diesel Exhaust, Respirable Dust, and Ischemic Heart Disease: An Application of the Parametric g-formula. Epidemiology. 2019 Mar;30(2):177-185.

Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. Int J Epidemiol. 2009 Dec;38(6):1599-611.

Valeri L, Mazumdar MM, Bobb JF, Claus Henn B, Rodrigues E, Sharif OI, Kile ML, Quamruzzaman Q, Afroz S, Golam M, Amarasiriwardena C. The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20–40 months of age: evidence from rural Bangladesh. Environmental health perspectives. 2017 Jun 26;125(6):067015.

부록 2.

특수건강진단 자료 합성 데이터를 활용한 g-formula 가이드라인

Ⅰ. 예제 자료에 대한 설명

1. 예제 자료 생성 방법

분석에 사용된 예제 자료는 산업안전보건연구원에서 보유하고 있는 2013-2019년 특수건강진단 자료이다. 특수건강진단 자료 중 혈중 납과 혈중 카드뮴, 혈중 헤모글로빈 값을 기반으로 정의한 빈혈 유무, 현재 음주 유무, 현재 흡연 유무, 체질량지수, 나이, 성별, 사업장 규모 변수를 분석용으로 사용하였으며, 1인당 1년에 1회 검진한 자료로 정제하였다.

synthpop R 패키지를 활용하여 정제한 특수건강진단 원자료와 유사한 가상의 합성 데이터를 예제 데이터로 만들었다. 예제로 활용할 합성 데이터의 특성은 다음과 같다.

2. 연구대상자 수

총 17,310명의 데이터를 합성하였고, 최대 7회까지 반복 측정된 데이터이기 때문에 총 42,846행의 데이터가 있다.

추적관찰 시점	추적관찰 된 자료 수
(회)	N(%)
1	17310(40.4)
2	10763(25.1)
3	6336(14.8)
4	3823(8.9)
5	2347(5.5)
6	1421(3.3)
7	846(2)

3. 예제 데이터의 변수 설명

시간에 따라 변하는 결과에 해당하는 요인은 빈혈 여부, 시간에 따라 변하는 노출은 혈중 납 농도와 혈중 카드뮴 농도, 시간에 따라 변하는 내생 교란 요인은 흡연 여부, 음주 여부, 비만도, 사후관리조치 결과, 시간에 따라 변하는 외생 교란 요인은 나이, 사업장 규모 그리고 시간에 따라 변하지 않는 교란 요인으로 성별이 있다.

역할	요인	설명
-	아이디	자료 내에서 같은 근로자임을 확인할 수 있도록 해주는 비식별화된 수
-	연도	특수건강진단을 받은 연도
시간에 따라 변하는 결과 변수	빈혈 여부	근로자가 빈혈을 앓고 있는지 여부
시간에 따라 변하는 교란 요인 및 노출 변수	음주 여부	근로자의 음주 유무
	흡연 여부	근로자의 흡연력 유무
	비만도	근로자의 체질량지수(kg/m^2)
	혈중 납 농도	근로자의 혈중 납 농도 수치($\mu g/dL$)
	혈중 카드뮴 농도	근로자의 혈중 카드뮴 농도 수치($\mu g/L$)
	사후관리조치 결과	판정결과에 따라 결정되는 사후관리조치 내용
시간에 따라 변하지 않는 교란 요인	나이	특수건강진단을 받았을 때의 근로자의 나이
	성별	근로자의 성별
	사업장 규모	근로자가 근무하는 사업장의 규모

1) 연속형 변수

나이와 혈중 납, 혈중 카드뮴은 연속형 변수로 설정하였다. 추적관찰 시점이 0인 baseline에서 연구대상자의 나이와 혈중 납, 혈중 카드뮴의 값은 다음과 같다.

변수명(단위)	N	평균±표준편차	최솟값	최댓값
나이(세)	17,310	38.7±10.8	18	65
혈중 납(μg/dL)	17,310	3.3±2.8	0.5	44.5
혈중 카드뮴(μg/L)	17,310	1±0.7	0.1	13.3

2) 범주형 변수

성별, 사업장 규모, 음주 유무, 흡연 유무, 사후관리, 빈혈 유무는 범주형 변수로 설정하였다. 추적관찰 시점이 0인 baseline에서 범주형 변수의 값은 다음과 같다.

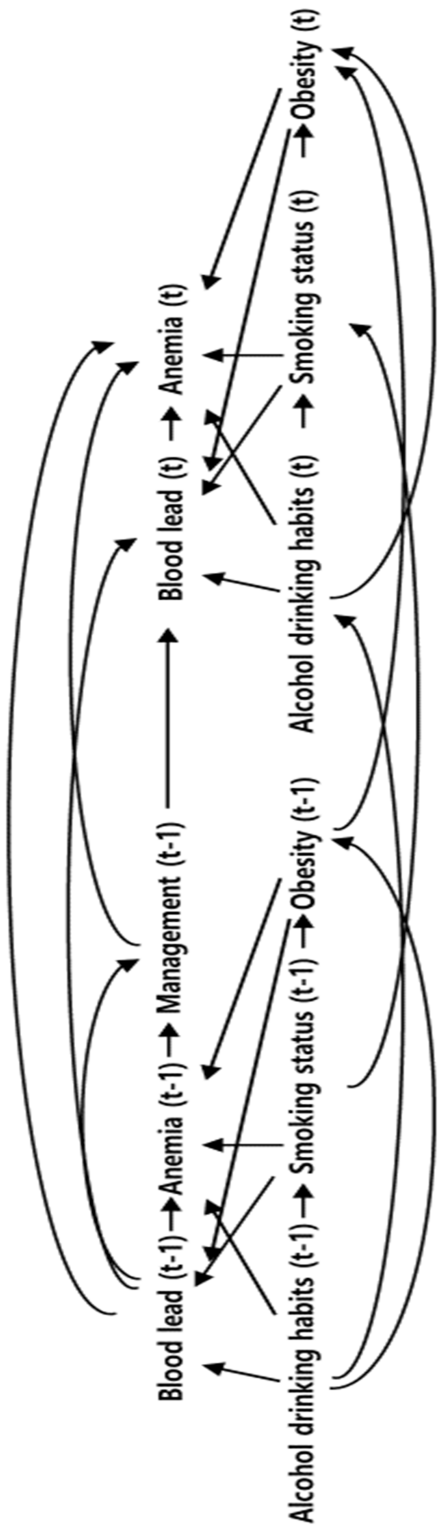
변수명	N(%)	변수명	N(%)
성별		**체질량지수(kg/m^2)**	
남성	14,343(82.9)	〈18.5	550(3.2)
여성	2,967(17.1)	18.5-22.9	5,918(34.2)
사업장규모		**23.0-24.9**	**4,369(25.2)**
〈50	3,803(22)	25.0-29.9	5,327(30.8)
50-299	6,886(39.8)	≥30.0	1,146(6.6)
≥300	6,621(38.3)	사후관리	
음주 유무		**필요없음**	**17,143(99)**
비음주자	1,022(5.9)	추적관찰 등 (근로시간 변화 없음)	149(0.9)
음주자	16,288(94.1)	작업전환 등 (근로시간 단축)	18(0.1)
흡연 유무		**빈혈 유무**	
비흡연자	7,684(44.4)	빈혈 없음	17,154(99.1)
흡연자(과거흡연 포함)	9,626(55.6)	빈혈 있음	156(0.9)

Ⅱ. 단일 노출 예제

1. 가설 설정

가설은 '장기적인 납 노출이 빈혈 발생의 위험도를 높인다'이며, 연구진 회의를 통해 다음 그림과 같이 인과 그래프로 표현하였다. 특수건강진단 자료를 활용하여 근로자의 종적 자료 안에 존재하는 요인들의 관계를 선험적 지식을 바탕으로 하여 작성한 인과 그래프이다.

구체적인 연구 가설은 작업장에서 시간에 따라 다르게 납에 노출된 근로자의 혈중 납 농도가 근로자의 빈혈의 발생률에 미치는 영향의 크기, 즉 효과를 구하고자 하는 것이며, 그 효과는 추적관찰 기간인 7년 동안 모든 근로자의 혈중 납 농도가 30μg/dL였을 때의 빈혈의 평균 발생률과 모든 근로자 혈중 납 농도가 일반 인구집단의 평균 혈중 납 농도 수준인 1.6μg/dL였을 때의 빈혈의 평균 발생률을 비교하여 측정하고자 한다.


Blood lead (t-1)
Anemia (t-1)
Management (t-1)
Blood lead (t)
Anemia (t)
Alcohol drinking habits (t-1)
Smoking status (t-1)
Obesity (t-1)
Alcohol drinking habits (t)
Smoking status (t)
Obesity (t)

2. R 패키지 분석

연구 가설을 분석하기 위한 R 코드는 다음과 같다.

```
#예제 데이터 불러오기
data <- read.csv("D://CI example//example_data.csv")

# 혈중 납 농도 log transformation 하기
hist(data$pb_result_n, main="Histogram of pb_result_n", xlab="pb_result_
n", col="blue")
data$log_pb <- log(data$pb_result_n)
hist(data$log_pb, main="Histogram of log_pb", xlab="log_pb", col="blue")

# data.table로 데이터 형태 바꾸기
install.packages("data.table")
library(data.table)
data <- as.data.table(data)

#g formula 분석하기
install.packages("gfoRmula")
library("gfoRmula")
id <- 'id'
time_points <- 7
time_name <- 't0'
covnames <- c('alcohol_sta','smoking_sta','obesity','log_pb', 'management
')
outcome_name <- 'anemia'
covtypes <- c('binary', 'binary', 'categorical', 'bounded normal','categorical
```

```
')
histories<-c(lagged)
histvars<-list(c('log_pb', 'management', 'smoking_sta','alcohol_sta', 'obesit
y'))
covparams<-list(covmodels=c(alcohol_sta~lag1_alcohol_sta+age+sex+scal
e,
                            smoking_sta~alcohol_sta+lag1_smoking_st
a+age+sex+scale,
                            obesity~alcohol_sta+smoking_sta+lag1_ob
esity+age+sex+scale,
                            log_pb~alcohol_sta+smoking_sta+obesity+l
ag1_log_pb+lag1_management+age+sex+scale,
                            management~log_pb+age+sex+scale))
ymodel<-anemia~log_pb+smoking_sta+alcohol_sta+obesity+lag1_log_pb+l
ag1_management+age+sex+scale
intvars=c('log_pb','log_pb')
interventions <- list(list(c(static, rep(log(1.6), 7))),
                      list(c(static, rep(log(30), 7))))
int_descript<-c('Never treat', 'Always treat')
nsimul<-42846
ncores<-parallel::detectCores()-1
gform_basic<-gformula_survival(obs_data=data, id=id,
                               time_points=time_points,
                               time_name=time_name, covnames=cov
names,
                               outcome_name=outcome_name,
                               covtypes=covtypes,
                               covparams=covparams,
                               ymodel=ymodel,
```

```
                                intvars=intvars,
                                interventions=interventions,
                                int_descript=int_descript,
                                histories=histories,
                                histvars=histvars,
                                basecovs=c('age','sex','scale'),
                                nsimul=nsimul,
                                ci_method='percentile',
                                nsamples=100,
                                ref_int=1,
                                parallel=TRUE,
                                ncores=ncores,
                                seed=1234)

#분석 결과 확인하기
gform_basic

#모델 적합도 확인하기
plot(gform_basic, survival = TRUE)
```

3. 결과 해석

시간에 따라 변하는 내생 교란 요인, 노출 변수, 그리고 결과 변수에 대하여 위에서 상정한 모형을 적용하여 g-formula를 분석한 결과는 다음과 같다표에서 Interv.은 노출 전략을 의미하며 Interv.의 값에서 0은 자연 경과를, 1은 모든 근로자의 혈중 납 농도를 7년 동안 1.6μg/dL로 유지시키는 전략을, 2는 모든 근로자의 혈중 납 농도를 7년 동안 30μg/dL로 유지시키는 전략을 의미한다.

현재 분석에서는 혈중 납 농도가 1.6μg/dL인 경우 대비 다른 노출 전략에서의 결과를 비교하고자 하였기 때문에 reference를 Interv.가 1인 경우로 설정하였다. Measure는 위험도를 계산하는 방법을 나타내며, NP risk, g-form risk, risk ratio 그리고 risk difference가 있다. NP risk는 비모수적 방법으로 risk를 추정한 값을 의미하며, 비모수적인 방법은 관찰된 자료에서만 계산할 수 있기 때문에 노출 전략이 자연 경과인 경우에만 계산이 가능하다. g-form risk는 g-formula를 적용하여 추정한 risk의 값을, risk ratio는 노출 전략 1을 reference로 구한 위험 비를, 그리고 risk difference는 노출 전략 1을 reference로 구한 위험 차이를 의미한다. 표에서 노출 전략이 1인 경우는 다른 노출 전략의 reference가 되기 때문에 Risk ratio와 Risk difference 각각에 관련된 값들이 모두 1 또는 0의 값을 갖는다. Estimates는 각 measure에 대한 점 추정치를 제공한다. lower 95% CI와 upper 95% CI는 추정치(estimates)의 95% 신뢰구간의 왼쪽과 오른쪽 값을 나타낸다.

Interv.	Measure	Estimates	Lower 95% CI	Upper 95% CI
0	NP risk	0.212	-	-
	g-form risk	0.306	0.203	0.454
	Risk ratio	1.118	1.058	1.191
	Risk difference	0.032	0.016	0.062
1	g-form risk	0.274	0.181	0.404
	Risk ratio	1.000	1.000	1.000
	Risk difference	0.000	0.000	0.000
2	g-form risk	0.616	0.433	0.813
	Risk ratio	2.247	1.741	2.826
	Risk difference	0.342	0.2	0.46

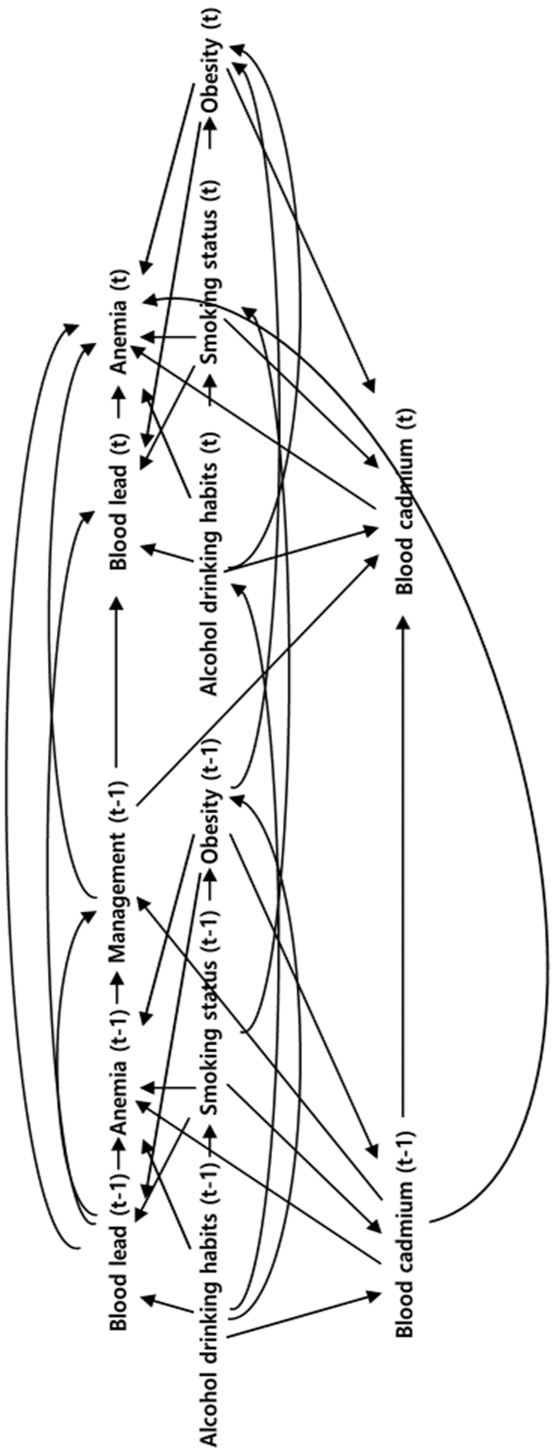
Ⅲ. 복합 노출 예제

1. 가설 설정

가설은 '납과 카드뮴에 대한 장기적인 복합노출이 빈혈 발생의 위험을 높인다.'이며 근로자의 건강 관련 정보(빈혈 여부, 혈중 납 농도, 혈중 카드뮴 농도, 나이, 성별, 사업장 규모, 음주 여부, 흡연 상태, 비만도, 건강진단 결과에 대한 사후관리조치 결과)를 활용하여 인과 그래프로 표현하였다.

실무지침 내 노출 기준치에 대해 납의 경우, 직업병 요관찰자에 해당하는 혈중 납 농도 기준치는 30μg/dL이며, 직업병 유소견자에 해당하는 혈중 납 농도 기준치는 40μg/dL이며, 카드뮴의 경우, 5μg/L이다.

구체적인 연구 가설은 작업장에서 시간에 따라 다르게 납에 노출된 근로자의 혈중 납 농도와 혈중 카드뮴의 복합노출이 근로자의 빈혈의 발생률에 미치는 영향의 크기, 즉 효과를 구하고자 하는 것이며, 그 효과는 추적관찰 기간인 7년 동안 모든 근로자의 혈중 납 농도가 40μg/dL이고 혈중 카드뮴 농도가 5μg/L였을 때의 빈혈의 평균 발생률과 모든 근로자 혈중 납 농도와 혈중 카드뮴 농도가 일반 인구집단의 평균 혈중 납 농도 수준인 1.6μg/dL와 일반 인구집단의 평균 혈중 카드뮴 농도 수준인 0.9187μg/L 이었을 때의 빈혈의 평균 발생률을 비교하여 측정하고자 한다.


Blood lead (t-1)
Anemia (t-1)
Management (t-1)
Obesity (t-1)
Smoking status (t-1)
Alcohol drinking habits (t-1)
Blood cadmium (t-1)
Blood lead (t)
Anemia (t)
Obesity (t)
Smoking status (t)
Alcohol drinking habits (t)
Blood cadmium (t)

2. R 패키지 분석

연구 가설을 분석하기 위한 R 코드는 다음과 같다.

```
#예제 데이터 불러오기
data <- read.csv("D://CI example//example_data.csv")

# 혈중 납 농도와 혈중 카드뮴 농도 log transformation 하기
hist(data$pb_result_n, main="Histogram of pb_result_n", xlab="pb_result_
n", col="blue")
data$log_pb <- log(data$pb_result_n)
hist(data$log_pb, main="Histogram of log_pb", xlab="log_pb", col="blue")
hist(data$cd_result_n, main="Histogram of cd_result_n", xlab="pb_result_
n", col="blue")
data$log_cd <- log(data$cd_result_n)
hist(data$log_cd, main="Histogram of log_cd", xlab="log_cd", col="blue")
data$obesity = as.factor(as.character(data$obesity))
data$management = as.factor(as.character(data$management))

# data.table로 데이터 형태 바꾸기
install.packages("data.table")
library(data.table)
data <- as.data.table(data)

#g formula 분석하기
#최근 버전의 "gfoRmula" R 패키지로 분석하였을 때 에러가 생겨 분석이 안될
경우, https://cran.r-project.org/src/contrib/Archive/gfoRmula/ 에서 과거
버전을 다운 받아 다음과 같이 실행한 뒤 분석하기
```

```
remove.packages("gfoRmula")
install.packages("D://CI example//gfoRmula_1.0.0.tar.gz", repos = NULL, type="source")
library("gfoRmula")
packageVersion("gfoRmula")

id <- 'id'
time_points <- 7
time_name <- 't0'
covnames <- c('alcohol_sta', 'smoking_sta', 'obesity', 'log_pb', 'log_cd', 'management')
outcome_name <- 'anemia'
covtypes <- c('binary', 'binary', 'categorical', 'bounded normal','bounded normal','categorical')
histories<-c(lagged)
histvars<-list(c('log_pb', 'log_cd', 'management', 'smoking_sta','alcohol_sta', 'obesity'))
covparams<-list(covmodels=c(alcohol_sta~lag1_alcohol_sta+age+sex+scale,
                            smoking_sta~alcohol_sta+lag1_smoking_sta+age+sex+scale,
                            obesity~alcohol_sta+smoking_sta+lag1_obesity+age+sex+scale,
                            log_pb~alcohol_sta+smoking_sta+obesity+lag1_log_pb+lag1_management+age+sex+scale,
                            log_cd~alcohol_sta+smoking_sta+obesity+lag1_log_cd+lag1_management+age+sex+scale,
                            management~log_pb+log_cd+age+sex+scale))
```

```
ymodel<-anemia~log_pb+log_cd+smoking_sta+alcohol_sta+obesity+lag1_l
og_pb+lag1_log_cd+lag1_management+age+sex+scale
intvars=list(c('log_pb','log_cd'), c('log_pb', 'log_cd'))
interventions <- list(list(c(static, rep(log(1.6), 7)),
                           c(static, rep(log(0.9187), 7))),
                      list(c(static, rep(log(40), 7)),
                           c(static, rep(log(5), 7))))
int_descript<-c('Never treat', 'Always treat')
nsimul<-42846
ncores<-parallel::detectCores()-4
gform_basic<-gformula_survival(obs_data=data, id=id,
                               time_points=time_points,
                               time_name=time_name, covnames=cov
names,
                               outcome_name=outcome_name,
                               covtypes=covtypes,
                               covparams=covparams,
                               ymodel=ymodel,
                               intvars=intvars,
                               interventions=interventions,
                               int_descript=int_descript,
                               histories=histories,
                               histvars=histvars,
                               basecovs=c('age','sex','scale'),
                               nsimul=nsimul,
                               ci_method='percentile',
                               nsamples=100,
                               ref_int=1,
                               parallel=TRUE,
```

```
                                    ncores=ncores,
                                    seed=1234)

#분석 결과 확인하기
gform_basic

#모델 적합도 확인하기
plot(gform_basic, survival = TRUE)
```

3. 결과 해석

시간에 따라 변하는 내생 교란 요인, 노출 변수, 그리고 결과 변수에 대하여 위에서 상정한 모형을 적용하여 g-formula를 분석한 결과는 다음과 같다. 표에서 Interv.은 노출 전략을 의미하며 Interv.의 값에서 0은 자연 경과를, 1은 모든 근로자의 혈중 납 농도와 혈중 카드뮴 농도를 7년 동안 1.6μg/dL와 0.9187μg/L로 유지시키는 전략을, 2는 모든 근로자의 혈중 납 농도와 혈중 카드뮴 농도를 7년 동안 40μg/dL과 5g/dL로 유지시키는 전략을 의미한다. 현재 분석에서는 혈중 납 농도와 혈중 카드뮴 농도가 1.6μg/dL과 0.9187μg/L인 경우 대비 다른 노출 전략에서의 결과를 비교하고자 하였기 때문에 reference를 Interv.가 1인 경우로 설정하였다. Measure는 위험도를 계산하는 방법을 나타내며, NP risk, g-form risk, risk ratio 그리고 risk difference가 있다. NP risk는 비모수적 방법으로 risk를 추정한 값을 의미하며, 비모수적인 방법은 관찰된 자료에서만 계산할 수 있기 때문에 노출 전략이 자연 경과인 경우에만 계산이 가능하다. g-form risk는 g-formula를 적용하여 추정한 risk의 값을, risk ratio는 노출 전략 1을 reference로 구한 위험 비를, 그리고 risk difference는 노출 전략 1을 reference로 구한 위험 차이를 의미한다. 표에서

노출 전략이 1인 경우는 다른 노출 전략의 reference가 되기 때문에 Risk ratio와 Risk difference 각각에 관련된 값들이 모두 1 또는 0의 값을 갖는다. Estimates는 각 measure에 대한 점 추정치를 제공한다. lower 95% CI와 upper 95% CI는 추정치(estimates)의 95% 신뢰구간의 왼쪽과 오른쪽 값을 나타낸다.

Interv.	Measure	Estimates	Lower 95% CI	Upper 95% CI
0	NP risk	0.212	-	-
	g-form risk	0.117	0.111	0.145
	Risk ratio	1.169	1.118	1.381
	Risk difference	0.017	0.012	0.04
1	g-form risk	0.100	0.091	0.109
	Risk ratio	1.000	1.000	1.000
	Risk difference	0.000	0.000	0.000
2	g-form risk	0.432	0.339	0.577
	Risk ratio	4.308	3.229	5.747
	Risk difference	0.332	0.232	0.475

Ⅳ. 질의 응답

특수건강진단 자료를 사용하여 g-formula R 패키지를 분석하면서 발생한 질의 응답을 아래와 같이 정리하였다.

질문 1: 연구 대상자의 수가 10,000명보다 큰 경우, nsimul은 기본 값으로 연구 대상자의 수로 설정되어 있는데, 더 줄이면 안 되나요?

답변: gfoRmula R package는 sample 함수를 사용하여 자료를 생성할 연구 대상자를 선정하기 때문에 연구 대상자보다 작은 값을 nsimul의 값으로 설정하게 되면 일부 분석 대상자만 선정하는 것과 동일합니다. 분석 시간을 단축하기 위한 목적이라면 nsimul을 작게 설정하시는 것보다 병렬 계산(parallel computing)을 권장드립니다(관련 인수는 parallel, ncores임). 병렬계산과 관련된 가이드라인의 내용은 아래와 같습니다.

gformula 함수는 계산 량이 많이 요구되는 몬테카를로 시뮬레이션과 붓스트랩을 모두 사용하기 때문에 결과를 제공하기까지 오랜 시간이 소요됩니다. 하지만 이러한 문제는 병렬 계산을 통하여 다소 해결할 수 있으며, 병렬계산을 사용할 수 있도록 gformula 함수는 parallel과 ncores를 제공합니다. parallel의 값을 TRUE로 설정하고, 이때 사용할 CPU core의 개수를 ncores에 입력하면 됩니다. 다만 ncores를 입력하는 경우에는 gformula 함수 이전에 아래의 예시와 같은 준비 코드가 필요합니다.

```
ncores <- parallel::detectCores() - 1
gformula(…,
 parallel = TRUE,
 ncores = ncores
)
```

첫 번째 줄에서 -1를 하는 이유는 데스크탑 또는 노트북의 CPU가 gformula 함수 외에 다른 프로그램을 처리할 수 있도록 여유 CPU를 설정하려는 목적으로 입력한 임의의 숫자이며, 연구자의 데스크탑이 보유하고 있는 CPU를 모두 gformula 함수를 처리하는데 사용하려는 연구자는 -1가 아닌 0을 또는 여유분을 더 남기려는 연구자는 -3, -4 등의 숫자를 사용하시면 됩니다.

질문 2: g-formula에서 보정 변수들 간의 교호작용 항(interaction term)을 가정하는 법과 노출 변수 간의 교호작용을 평가하는 방법이 있을까요?

답변: 모델에서 보정 변수들 사이의 교호작용을 포함하고 싶으시다면 모델을 기술할 때, 예를 들어 L1 * L2와 같이 입력하시면 됩니다. 다만 이때, 모형에 L1 * L2와 같이 입력한 교호작용은 통계적 교호작용(statistical interaction)으로 인과적 교호작용(causal interaction)과는 다른 개념이라는 점을 주의하셔야 합니다. 노출 변수들 간의 인과적 교호작용을 보고자 하실 때에는 가이드라인에서 기술한 것과 같이 각 노출 변수의 값을 바꿔가며 인과 효과 추정치를 산출한 후, 그림을 그려서 등고선 그림과 같이 시각적으로 확인하는 방법도 있고, additive interaction 또는 multiplicative interaction을 구하여 수치적으로 확인하는 방법도 있습니다.

질문 3: 예제에 있는 as.factor(t0)는 꼭 포함해야하는 걸까요?

답변: t0은 연구 등록으로부터의 시간이기 때문에 t0은 시간과 선형적인 관계가 있습니다. 이러한 관계를 모형에 포함된 시간에 따라 변하는 변수가 이미 가지고 있다면 t0를 모형에 포함하는 것은 시간에 대해 중복 보정하는 결과를 낳게 됩니다. 그러므로 시간에 선형적인 관계를 가지는 변수가 모형에 없을 경우에는 포함하는 것이 모형의 구축에 도움을 줄 수 있으나 이미

시간과 선형적인 관계를 포함하는 변수가 모형에 포함되어 있을 경우에는 포함하지 않는 것이 바람직하다고 생각됩니다. 예를 들어, 나이(age) 변수가 시간에 따라 증가하도록 코딩이 되어 있을 경우, 나이는 시간에 선형적으로 증가하기 때문에 t0을 모형에 포함하는 것은 나이를 중복 보정하는 것으로 볼 수 있습니다(예를 들어, 1년 단위의 시간을 가지는 자료의 경우, 1번 근로자에 대한 2012년, 2013년, 2014년 자료의 나이 변수의 값이 20, 21, 22와 같이 코딩되어 있을 경우). 나이를 연구 등록 시점의 나이로 코딩을 하셨다면 t0를 포함하셔도 될 것으로 생각되고(위의 예제를 기준으로 1번 근로자의 나이는 20으로 시간에 따라 변하지 않는 값으로 코딩되어 있을 경우), 나이를 시간에 따라 증가하도록 코딩을 하셨을 경우에는 포함하지 않는 것이 바람직하다고 생각됩니다.

질문 4: gfoRmula R package 실습 시 시간에 따라 변하는 변수의 형태를 지정하여 주는데, 연속형 변수의 경우, 분석 전 분포를 확인해야 하나요?

답변: 잘못된 분포를 지정하여 생기는 g-formula 추정치의 편향을 줄일 수 있기 때문에 분석 전 분포를 확인하는 것을 권장드립니다.

질문 5: 연속형 변수에서 한쪽으로 치우친 skewed data의 경우, log-transformation이 필요한가요?

답변: 로그 변환하여 정규 분포에 가까워진다면 변환 후 변수의 분포를 정규 분포로 지정하신 후, g-formula를 적용하여 분석하시면 됩니다. 노출 변수를 로그 변환한 경우 interventions에서도 아래와 같이 로그 변환한 노출 값을 지정해주면 됩니다.

```
interventions <- list(list(c(static, rep(log(1.6), 7))),
                      list(c(static, rep(log(30), 7))))
```

질문 6: gfoRmula R package에서 reference 변경하는 법이 가이드 라인에 나와 있지 않은데 어떻게 바꿀 수 있나요?

답변: gfoRmula R package의 gformula 함수는 intervention으로 natural course를 기본적으로 제공하고 있기 때문에 어떠한 intervention을 수행하더라도 gformula 함수에서 natural course가 0번 intervention으로 나타나며, 기본 reference로 사용하고 있습니다(아래의 코드 기준으로 ref_int=0). 따라서 g-formula R package에서 두 가지 intervention (never treat: 1, always treat: 2)이 적용된 아래의 예시 코드에서 reference를 natural course에서 never treat으로 변경하시려면 ref_int=1을 코드에 추가하여 주시면 됩니다.

```
gform_basic <- gformula_survival(
obs_data = basicdata_nocomp,
id = id,
time_points = time_points,
time_name = time_name,
covnames = covnames,
outcome_name = outcome_name,
covtypes = covtypes,
covparams = covparams,
ymodel = ymodel,
intvars = intvars,
interventions = interventions,
int_descript = int_descript,
histories = histories,
histvars = histvars,
basecovs = c('L3'),
```

```
nsimul          = nsimul,
ref_int         = 1,
seed            = 1234
)
```

질문 7: gfoRmula R package에서 covtypes를 사용하여 내생교란요인의 분포를 지정할 때, ordinal 분포를 사용할 수 있나요?

답변: 현재 gfoRmula R package에서 교란 변수의 형태로서 ordinal은 미리 만들어 둔 함수의 형태로 지원하고 있지 않습니다. 다만 ordinal 형태를 반영하고 싶으신 경우에는 VGAM 등의 패키지에서 제공하는 함수를 활용하여 gformula 함수에 입력 가능한 형태로 교란 변수에 적합할 함수를 생성하고, 교란 변수의 타입을 custom으로 지정한 뒤, g-formula를 사용하시면 됩니다.

질문 8: gfoRmula R package를 돌리다보면 에러가 너무 많이 뜹니다. 이러한 경우에는 어떻게 해야 할까요?

답변: 여러 연구자들이 이 패키지를 사용하면서 발생하는 에러들에 대해 gfoRmula R package를 만든 저자가 운영하는 깃허브 사이트(https://github.com/CausalInference/gfoRmula/issues) 에 질문을 올리면 저자가 직접 답변을 해주고 있습니다. 생성된 에러 중 깃허브 사이트에 있는 에러의 경우, 해결이 가능하며, 이외의 에러의 경우, 구글(google) 또는 ChatGPT를 이용하여 일부 해소가 가능할 것입니다.

연구진

연 구 기 관 : **산업안전보건연구원**

연구책임자 : 예신희 (팀장, 중부권역학조사팀)
연 구 원 : 이상길 (실장, 직업건강연구실)
연 구 원 : 이유진 (연구위원, 직업건강연구실)
연 구 원 : 마성원 (전공의, 역학조사부)
연 구 원 : 박성균 (전공의, 역학조사부)
연 구 원 : 김용진 (전공의, 역학조사부)

부분위탁

연구책임자 : 이우주 (부교수, 서울대학교 보건대학원)
연 구 원 : 조성일 (부교수, 인하대학교 통계학과)
연 구 원 : 이동환 (부교수, 이화여자대학교 통계학과)
연구책임자 : 심현만 (박사과정, 서울대학교 보건대학원)
연구책임자 : 김연진 (박사과정, 서울대학교 보건대학원)
연구책임자 : 정승필 (박사과정, 서울대학교 보건대학원)

연구기간

2023. 02.~ 2023. 11.

본 연구보고서의 내용은 연구책임자의 개인적 견해이며,
우리 연구원의 공식견해와 다를 수도 있음을 알려드립니다.

산업안전보건연구원장

직업병 인과추론 가이드라인 및 통계분석법 개발 (3)

(2023-산업안전보건연구원-744)

발 행 일 : 2023년 11월 30일
발 행 인 : 산업안전보건연구원 원장 김은아
연구책임자 : 산업안전보건연구원 팀장 예신희
발 행 처 : 안전보건공단 산업안전보건연구원
주 소 : (44429) 울산광역시 중구 종가로 400
전 화 : 032-510-0754
팩 스 : 032-510-0759
Homepage : http://oshri.kosha.or.kr
I S B N : 979-11-93642-17-7
공공안심글꼴 : 무료글꼴, 한국출판인회의, Kopub바탕체/돋움체