조사연구(2025. 11) 26권 4호: 141－172

DOI http://dx.doi.org/10.20997/SR.26.4.6

연구논문

로짓 또는 프로빗 회귀모형에서 상호작용효과의 유도와 확률 해석*

김현우**

이 연구는 로짓과 프로빗 등 비선형 회귀모형에서 상호작용효과의 올바른 유도와 한계효과로서의 해석 방법을 제시한다. 선형 모형과 달리 비선형 모형에서는 상호작용항의 회귀계수가 상호작용효과와 일치하지 않을 수 있으며, 특히 그 한계효과는 공변량의 구체적인 값에 따라 크기와 방향이 달라질 수 있다. 이 연구는 수리적 검토를 통해 비선형 모형에서 상호작용의 한계효과는 교차편미분으로 계산되어야 하며, 이는 상호작용항 외에도 주요항의 회귀계수와 확률밀도함수의 도함수로 구성되어 있음을 보였다. 다양한 시나리오를 설정한 시뮬레이션 결과, 상호작용항의 계수와 상호작용효과가 반대로 나타나는 구간이 존재할 뿐 아니라, 상호작용항이 통계적으로 유의하지 않은 경우에도 그러한 한계효과는 나타날 수 있음을 확인하였다. 이러한 문제를 해결하기 위해 다섯 가지 실용적 대안을 제시하였다. 이 연구는 사회학 연구에서 비선형 모형의 상호작용효과 해석의 정확성을 높이는 데 중요한 방법론적 함의를 제공할 것이다.

주제어: 로짓, 프로빗, 비선형 모형, 상호작용효과, 상호작용항, 한계효과, 범주형 자료분석

* 이 논문은 2021년도 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구입니다 (NRF-2021S1A5C2A02088321). 소중한 코멘트를 주신 익명의 네 분 심사자들께 진심으로 감사드립니다.

** 충북대학교 사회학과 조교수(hxk271@cbnu.ac.kr).

Ⅰ. 서론

이 연구는 로짓(logit)과 프로빗(probit)과 같은 비선형 모형(nonlinear models)에서 상호작용효과(interaction effect)를, 특히 한계효과(marginal effect)로서 명확하게 도출하고 해석하는 방법에 관한 것이다. 어떤 효과는 종종 다른 조절요인(moderator)의 효과에 의존한다. 이 아이디어는 종종 회귀모형에서 상호작용항(interaction term)을 통해 구현된다. 임금에 대한 교육의 효과가 성별에 따라 변화하는지, 또는 어떤 치료법의 효과가 환자의 사회경제적 지위에 따라 달라지는지 등 사회(과)학 분야의 수많은 연구 질문은 상호작용효과를 살펴보아야 적절한 답을 찾을 수 있다. 다만 비선형 모형의 독특한 성질로 인해 연구자가 상호작용효과를 해석할 때, 해석상 오류를 저지르기 쉽고, 그로 인해 지속적으로 부정확한 지식을 생산할 위험이 크다. 따라서 비선형 모형에서 상호작용효과를 둘러싼 함정을 이해하는 것은 단순히 방법론적 엄밀성을 위한 것이 아니라, 사회 현상의 맥락 의존성을 더 깊이 있게 이해하기에 필수적이다.

Ai & Norton(2003)의 연구가 발표된 지 20년이 지났음에도, 비선형 회귀모형에서 상호작용항의 계수는 여전히 실제 상호작용효과와 동일한 것으로 종종 오해되고 있다. 그러나 이 두 개념은 수학적으로 전혀 다르다. 가령 로짓 회귀모형에서 상호작용효과는 교차편미분(cross-partial derivative)으로 정의되며, 이는 상호작용항의 계수와 일반적으로 같지 않다. 게다가 상호작용항의 계수 부호와 실제 상호작용효과의 부호가 정반대일 수도 있다. 통계적 유의성 판단도 마찬가지로 복잡하다. 상호작용항의 계수가 통계적으로 유의하더라도, 실제 상호작용효과는 관측치(observations)마다 다를 수 있다. 심지어 상호작용항의 계수가 0이더라도 상호작용효과는 0이 아닐 수도 있다. 그러므로 비선형 회귀분석에서 컴퓨터 프로그램이 보여주는 상호작용항의 통계적 유의성의 유무(*)에 일희일비하는 것은 섣부르다.

이 연구의 목적은 널리 알려진 비선형 모형인 로짓과 프로빗 회귀모형을 중심으로 비선형 모형에서 상호작용효과가 왜 잘못 해석되기 쉬운지 해명하고, 이를 토대로 상호작용효과를 올바르게 유도하는 원리에 관해 설명하는 것이다. 우선 선형 모형에서 상호작용항의 회귀계수는 곧바로 상호작용효과를 의미하지만, 비선형 모형

에서는 그렇지 않음을 보인다. 뒤이어, 다양한 상황을 가정한 시뮬레이션을 통해 회귀계수들의 방향과 크기에 따라 상호작용효과가 얼마나 극적으로 달라질 수 있는지 확인한다.

이 연구의 또 다른 목표는 단순히 문제를 지적하는 데 그치지 않고, 연구자들이 직접 활용할 수 있는 실용적인 대안을 제시하는 것이다. 다양한 계산법과 시각화 방법을 실제 자료 및 코드와 함께 제공하여, 연구자들이 향후 연구에 있어 혼동과 오류를 줄일 수 있도록 돕는다. 다양한 비선형 모형의 상호작용효과를 올바르게 분석함으로써, 궁극적으로 사회 현상 간의 복잡한 상호의존적 성격에 대한 더 정확하고 신뢰할 수 있는 이해를 얻을 수 있을 것이다.

Ⅱ. 이론적 검토

이 장에서는 우선 회귀분석 프레임워크에서 상호작용효과는 수학적으로 조건부 기대값의 교차편미분을 의미함을 살펴본다. 선형 모형에서는 상호작용항이 곧 상호작용효과와 일치하지만, 비선형 모형에서 이 둘은 일반적으로 다름을 증명하는 것이 그 핵심이다.

1. 선형 모형에서 상호작용효과

아래와 같은 선형 모형에서 공변량 X_1과 X_2 그리고 그 곱(product)인 X_1X_2를 투입하였을 때, X_1X_2의 회귀계수인 b_{12}를 상호작용항이라고 지칭하며 이는 두 공변량의 결합적 영향력을 반영한다고 이해된다.

$$E(Y|X_{1,}X_2) = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2 \quad (1)$$

한편 상호작용효과란 가령 (1) X_2의 변화가 종속변수 Y의 조건부 기댓값의 변화에 미치는 영향이 (2) X_1의 변화에 따라 달라지는 정도를 의미하며, 바로 이러한 직관을 아래와 같은 교차편미분으로 표현할 수 있다.

$$\frac{\partial^2 E(Y|X_1, X_2)}{\partial X_1 \partial X_2} = \frac{\partial}{\partial X_1}\left(\frac{\partial E(Y|X_1, X_2)}{\partial X_2}\right)$$

위 식에서 우변의 두 번째 부분(괄호 안 부분)은 X_2의 변화가 종속변수 Y의 조건부 기댓값의 변화에 미치는 영향을 나타내며, 그 앞부분은 X_1의 변화가 다시 괄호 안 변화에 미치는 영향력을 나타낸다. 또한 식 (1)에서 제시한 선형 모형에서 상호작용항의 계수 b_{12}는 이 교차편미분과 정확히 일치함을 보일 수 있다.

$$\frac{\partial^2 E(Y|X_1, X_2)}{\partial X_1 \partial X_2} = \frac{\partial}{\partial X_1}\left(\frac{\partial E(Y|X_1, X_2)}{\partial X_2}\right) = \frac{\partial}{\partial X_1}(b_2 + b_{12}X_1) = b_{12}$$

그러므로 선형 모형의 교차편미분 값인 상호작용항의 계수 b_{12}는 곧장 (한계효과로서 표현한) 상호작용효과로 파악할 수 있다. 다시 말해, 선형 회귀모형에서 상호작용항과 상호작용효과의 동일성은 보장된다.

2. 비선형 모형에서 상호작용효과

한편, 비선형 회귀모형에서는 종속변수인 조건부 기대값 $E(Y|X_{1,}X_2)$이 우변과 선형 회귀모형처럼 항등함수(identity function)로 연결되지 않고, 모종의 비선형적인 누적분포함수 $F(\cdot)$를 연결함수(link function)로 사용한다. 가령 널리 사용되는 로짓 모형에서는 식 (2)처럼 누적 로지스틱 분포함수 Λ를 사용하며, 회귀모형은 잠재변수(latent variable) Y^*로 설정된다.

$$E(Y|X_{1,}X_2) = \Lambda(Y^* = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2) \quad (2)$$

$$\text{이때 } \Lambda(Y^*) = \frac{1}{1+\exp(-Y^*)}.$$

이제 상호작용효과를 유도하기 위해서는 식 (2)에서 조건부 기대값의 X_1과 X_2에 대한 교차편미분을 계산해야 한다. 이때 연쇄법칙(chain rule)과 합성함수의 미분법 등을 활용하면 다음의 결과를 얻는다.

$$\begin{aligned}\frac{\partial^2 \Lambda(Y^*)}{\partial X_1 \partial X_2} &= \frac{\partial}{\partial X_1}\left(\frac{\partial \Lambda(Y^*)}{\partial Y^*}\frac{\partial Y^*}{\partial X_2}\right) \\ &= \frac{\partial}{\partial X_1}\left(\lambda(Y^*)(b_2 + b_{12}X_1)\right) \\ &= \frac{\partial \lambda(Y^*)}{\partial X_1}(b_2 + b_{12}X_1) + \lambda(Y^*)b_{12} \\ &= \frac{\partial \lambda(Y^*)}{\partial Y^*}\frac{\partial Y^*}{\partial X_1}(b_2 + b_{12}X_1) + \lambda(Y^*)b_{12} \\ &= \lambda'(Y^*)(b_1 + b_{12}X_2)(b_2 + b_{12}X_1) + \lambda(Y^*)b_{12}\end{aligned} \tag{3}$$

이때 λ는 로지스틱 분포의 확률밀도함수이고, 정의상 Λ의 도함수이다. 이 유도의 결과, b_{12}은 상호작용항이자 상호작용효과가 그 자체가 아니라, 단지 그 일부에 지나지 않음에 주목해야 한다.

다음으로, 로짓 모형보다 덜 인기 있지만 더 오래된 프로빗 모형은 다음과 같이 설정될 수 있다.

$$E(Y|X_1,X_2) = \Phi(Y^* = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2)$$

$$\Phi(Y^*) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{Y^*}\exp\left(-\frac{\epsilon^2}{2}\right)d\epsilon$$

여기에서 Φ는 물론 누적표준 정규분포함수이며, 프로빗 모형에서 다음과 같이 교차편미분을 유도할 수 있다.

$$\begin{aligned}\frac{\partial^2 \Phi(Y^*)}{\partial X_1 \partial X_2} &= \frac{\partial}{\partial X_1}\left(\frac{\partial \Phi(Y^*)}{\partial Y^*}\frac{\partial Y^*}{\partial X_2}\right) \\ &= \frac{\partial}{\partial X_1}\left(\phi(Y^*)(b_2 + b_{12}X_1)\right) \\ &= \frac{\partial \phi(Y^*)}{\partial X_1}(b_2 + b_{12}X_1) + \phi(Y^*)b_{12} \\ &= \frac{\partial \phi(Y^*)}{\partial Y^*}\frac{\partial Y^*}{\partial X_1}(b_2 + b_{12}X_1) + \phi(Y^*)b_{12} \\ &= \phi'(Y^*)(b_1 + b_{12}X_2)(b_2 + b_{12}X_1) + \phi(Y^*)b_{12}\end{aligned} \tag{4}$$

이때 ϕ는 표준정규분포의 확률밀도함수이고, 정의상 Φ의 도함수이다. 만일 두 공변량 X_1과 X_2가 이산변수(discrete variable)라면, 교차편차분(cross-partial difference)으로 다루는 것이 좀 더 엄밀하다. 이렇게 계산된 교차편차분은 아래와 같으며, 여기에서도 b_{12}는 명백히 상호작용효과의 (전부가 아니라) 그 일부에 지나지 않는다(Greene 2010; Karaca-Mandic, Norton, & Dowd 2012).

$$\begin{aligned}\frac{\Delta^2 \Lambda(Y^*)}{\Delta X_1 \Delta X_2} &= \frac{\Delta}{\Delta X_1}\left(\frac{\Delta \Lambda(b_0 + b_1 X_1 + b_2 X_2 + b_{12} X_1 X_2)}{\Delta X_2}\right) \\ &= \frac{\Delta}{\Delta X_1}\left(\Lambda(b_0 + b_1 X_1 + b_2 + b_{12} X_1) - \Lambda(b_0 + b_1 X_1)\right) \\ &= \left(\Lambda(b_0 + b_1 + b_2 + b_{12}) - \Lambda(b_0 + b_{1)})\right) - \left(\Lambda(b_0 + b_2) - \Lambda(b_0)\right)\end{aligned}$$

다만 Puhani (2012)는 비선형 이중차분(nonlinear difference-in-difference) 프레임워크에서 처리집단에 대한 평균처리효과(average treatment effect on the treated) 계산이 위에서 계산된 교차편차분과는 다름을 보였다. 이 논문의 직접적인 주제는 아니지만, 비실험적 인과추론을 연구하는 사회학자는 이 주제에 관해서도 특별한 주의를 기울일 필요가 있다(Athey & Imbens 2006). 이에 관심 있는 연구자는 저자의 코드를 참고할 수 있다.

한편, 사회학 분야에서 연구자들이 흔히 사용하는 SPSS, R, Stata, SAS 등 여러 통계분석 소프트웨어는 로짓이나 프로빗 등 비선형 모형의 분석 결과에서 (교차편미분이 아니라) 단지 주요항(main terms)의 곱인 $X_1 X_2$ 그 자체에 대한 편미분 값 b_{12}를 다음과 같이 계산하여 보고할 뿐이다.

$$\frac{\partial \Lambda(Y^*)}{\partial(X_1 X_2)} = \lambda(Y^*) b_{12} \quad \text{그리고} \quad \frac{\partial \Phi(Y^*)}{\partial(X_1 X_2)} = \phi(Y^*) b_{12}$$

위 식들은 식 (3) 또는 식 (4)에서 전개한 상호작용효과의 일부일 뿐이므로, 통계분석 소프트웨어에서 보여주는 상호작용항만을 보고 상호작용효과의 유무나 크기를 올바르게 판단할 수 없다. (선형 모형과는 달리) 비선형 모형에서 상호작용효과와 상호작용항의 회귀계수 b_{12}의 동일성은 보장되지 않기 때문이다. 이때 연구자는 비선형 모형에서 아래와 같음을 살펴볼 필요가 있다.

$$\frac{\partial^2 \Lambda(Y^*)}{\partial X_1 \partial X_2} \neq \frac{\partial \Lambda(Y^*)}{\partial (X_1 X_2)} \quad \text{그리고} \quad \frac{\partial^2 \Phi(Y^*)}{\partial X_1 \partial X_2} \neq \frac{\partial \Phi(Y^*)}{\partial (X_1 X_2)}$$

3. 비선형 모형에서 상호작용효과의 구성요소

지금까지 검토하였듯, 로짓과 프로빗 등 비선형 모형의 교차편미분인 상호작용효과는 상호작용항인 b_{12} 이외에도 여러 요소로 구성되어 있다. 가령 로짓 모형이라면 식 (3)이 보여주듯, b_{12} 이외에도, (a) 주요항인 X_1과 X_2, (b) 그들의 회귀계수인 b_1과 b_2, 그리고 (c) $\lambda'(Y^*)$가 개입하여 상호작용효과를 만들어낸다. 이렇게 b_{12} 이외의 구성요소가 있다는 사실은 다음과 같은 내용을 시사한다.

첫째, 선형 모형에서는 상호작용효과가 b_{12}에 의해서만 결정되므로 곧 표본 전체에 걸친 전역적 효과(global effect)인 반면, 비선형 모형에서는 주요항의 구체적인 값 X_1과 X_2에 따라 상호작용효과가 달라진다. 구체적인 상호작용효과는 구간(local)에 따라 다르므로 분석에 사용된 표본 안에서 직접 계산해 보기 전에는 구간별 차이가 얼마나 큰지 사전에 알 수 없다.

둘째, 비선형 모형에서 상호작용항의 회귀계수 b_{12}가 가령 양(+)의 값으로 추정되었다고 할지라도, 실제로 양의 상호작용효과가 있다고 말할 수 없다. 식 (3)과 (4)에서 유도된 교차편미분에서 확인할 수 있는 확률밀도함수 $\lambda(Y^*)$와 $\phi(Y^*)$는 반드시 양수임이 자명하나, 그 도함수인 $\lambda'(Y^*)$와 $\phi'(Y^*)$는 부호가 어느 쪽으로도 변할 수 있으므로, 직접 계산해 보기 전에는 양의 상호작용효과가 있는지 확신할 수 없기 때문이다(Online Appendix 참고).

지금까지 첫 번째와 두 번째 문제들은 비선형 모형 자체의 내재적 성질에서 기인한다. 다시 말해, 비선형 모형이라면 설령 상호작용항을 갖고 있지 않더라도 이 문제를 피할 수 없으므로, 주효과(main effect)의 한계효과를 해석할 때 역시 주의를 기울여야 한다.

셋째, 상호작용항의 회귀계수가 설령 귀무가설($H_0 : \beta_{12} = 0$)을 통계적으로 유의하게 기각하지 못했더라도, 나머지 구성요소에 의해 여전히 0보다 큰 상호작용효과의 절대값을 얻을 수 있다. 식 (3)에 $b_{12} = 0$을 대입하면 아래의 식을 얻을 수 있는데, 이 경우조차 값이 0이라는 보장은 여전히 없기 때문이다.

$$\frac{\partial^2 \Lambda(Y^*)}{\partial X_1 \partial X_2} = \lambda'(Y^*) b_1 b_2$$

조금 더 확장해서 생각해 본다면, (설령 이론적으로 관심이 없었다던가 등의 이유로) 애초에 상호작용항을 모형에 추가하지 않은 경우조차 상호작용효과가 존재할 수 있음을 염두에 둔 접근이 필요하다(Berry, DeMeritt, & Esarey 2010).

결론적으로, 통계분석 소프트웨어를 통해 추정된 비선형 모형의 상호작용항 회귀계수 b_{12}가 어떠한 계수 크기, 방향, 그리고 통계적 유의성을 가졌는가와 무관하게, 표본으로부터 X_1과 X_2, b_1과 b_2, 그리고 $\lambda'(Y^*)$ 또는 $\phi'(Y^*)$를 통해 상호작용효과를 연구자가 직접 계산해야 할 필요가 있다. 이렇게 개별적으로 추정된 상호작용효과의 통계적 유의성은 일반적인 t 검정이나 Z 검정 등이 아니라, 델타 방법(delta method) 또는 MCMC (Markov Chain Monte Carlo) 시뮬레이션 등을 거쳐 얻을 수 있다(Ai & Norton 2003; Greene 2010).

Ⅲ. 시뮬레이션

지금까지 비선형 모형에서는 단순히 상호작용항 해석을 통해 상호작용효과를 유추할 수 없음을 수리적으로 살펴보았다. 그러나 위에서 유도한 올바른 비선형 상호작용효과는 상호작용항에 비해 훨씬 복잡하므로, 만일 문제의 심각성이 크지 않다면(즉 상호작용항과 상호작용효과가 대체로 일치하는 방향과 크기라면), 그냥 상호작용항에 준하여 상호작용효과를 해석하는 것도 연구자에게는 하나의 현실적인 대안이 될 수 있다. 상호작용효과는 상호작용항과 얼마나 괴리가 큰가? 이 질문에 대해 답하기 위해 이 장에서는 몇 가지 시나리오에 따라 설정된 비선형 모형에서 상호작용효과가 얼마나 크게 달라질 수 있는가를 시뮬레이트한다.

시뮬레이션 결과에 따르면, 상호작용항과 무관하게 상호작용효과가 구간에 따라 상당히 크거나 작고, 심지어 계수의 방향과 정반대로 나타난다. 또 상호작용항의 회귀계수가 통계적으로 유의하지 않은 경우($\beta_{12} = 0$)조차도 (특정 상황이 아니라면) 상호작용효과가 다양하게 나타날 수 있다.

1. 모형 설정

여기서 사용되는 시뮬레이션은 식 (2)와 같은 로짓 모형을 기초로 한다. 우선 두 개의 공변량 X_1와 X_2를 각각 $[-2,2]$ 구간에서 일정한 간격으로 분할된 격자(grid) 위에서 정의하고, 이들의 크로넥커 곱(Kronecker product)으로 구성된 2차원 공변량 공간(covariate space)을 구현한다. (아래에서 구체적으로 설명하듯) 세 가지 시나리오에 따른 회귀계수를 공변량 공간 위에 적용하여 개별적인 상호작용효과를 모두 컴퓨터가 계산한다.

그런데 실제 계산 과정의 어려움이 하나 남아있다. 교차편미분을 실제로 계산하여 구체적인 값을 얻어내려면 식 (3)에 그치지 않고, 그 안의 구체적인 $\lambda'(Y^*)$를 얻어야 실제 자료의 값을 대입할 수 있다. 다행히 통계학의 로지스틱 우도함수 헷시안(Hessian) 계산이나 딥러닝의 경사하강법(gradient descent) 등 각종 최적화 기법 등을 통해 다음이 널리 알려져 있다.

$$\begin{aligned}\lambda'(Y^*) &= \frac{\partial}{\partial Y^*}[1-\Lambda(Y^*)]\Lambda(Y^*) \\ &= \frac{\partial[1-\Lambda(Y^*)]}{\partial Y^*}\Lambda(Y^*)+[1-\Lambda(Y^*)]\frac{\partial\Lambda(Y^*)}{\partial Y^*} \\ &= -\lambda(Y^*)\Lambda(Y^*)+[1-\Lambda(Y^*)]\lambda(Y^*) \\ &= \lambda(Y^*)[1-2\Lambda(Y^*)]\end{aligned}$$

이때,

$$\begin{aligned}\lambda(Y^*) &= \frac{\exp(-Y^*)}{[1+\exp(-Y^*)]^2} = \frac{1}{[1+\exp(-Y^*)]}\cdot\frac{\exp(-Y^*)}{[1+\exp(-Y^*)]} \\ &= \Lambda(Y^*)\cdot[1-\Lambda(Y^*)]\end{aligned}$$

그러므로 실제 시뮬레이션에서 구체적으로 사용되는 상호작용효과로서의 교차편미분 값은 식 (3)과 살짝 달리 아래와 같음에 주의해야 한다.

$$\begin{aligned}\frac{\partial^2\Lambda(Y^*)}{\partial X_1\partial X_2} &= \lambda(Y^*)[1-2\Lambda(Y^*)](b_1+b_{12}X_2)(b_2+b_{12}X_1)+\lambda(Y^*)b_{12} \\ &= \lambda(Y^*)\left[[1-2\Lambda(Y^*)](b_1+b_{12}X_2)(b_2+b_{12}X_1)+b_{12}\right]\end{aligned}$$

시뮬레이션의 첫 번째 시나리오는 $b_1 = 1$, $b_2 = 1$, $b_{12} \in \{-0.5, -1, -2\}$이다. 다른 요소는 고정되고 b_{12}만을 -0.5, -1, -2로 각각 변화시켰을 때, 상호작용효과가 어떻게 달라지는가를 관찰하는 것이 목적이다. 두 번째 시나리오는 $b_1 = 1.5$, $b_2 = 0.5$, $b_{12} \in \{2, 1, 0.5\}$이다. b_{12}의 계수 방향을 뒤집고, b_1과 b_2를 각각 0.5씩 늘리거나 줄였다. 첫 번째 시나리오에 비해, b_{12}의 효과는 음수이고, b_1과 b_2는 다소 비대칭적으로 작용할 때, 그 결과가 어떻게 달라지는가를 관찰한다. 세 번째 시나리오는 $b_1 \in \{-1, 0, 1\}$, $b_2 = 1$, $b_{12} = 0$이다. 여기서는 상호작용항을 아예 모형에 투입하지 않거나, 통계적으로 유의하지 않았을 때($\beta_{12} = 0$), 상호작용효과 역시 정말 0인지 여부를 확인하기 위함이다. 모든 시나리오에서 계산 편의상 상수를 0으로 설정한다. 세 가지 시나리오가 모든 가능성을 포괄하는 것은 아니지만, 다른 경우는 대체로 이로부터 유추할 수 있다. 가령 $b_1 = b_2 = 1$의 경우, 첫 번째 시나리오의 결과를 뒤집은 것과 똑같이 나오기 때문에 구태여 다시 수행하지 않았다.

한편 비선형 모형에서는 추정된 회귀계수가 진정한 회귀계수에 스케일 팩터(scale factor)에 의해 표준화되어 얻어지므로, 계수 비교가 곤란하다는 사실이 이미 널리 알려져 있다(Breen, Karlson, & Holm 2018; Mood 2010). 즉 집단별로 종속변수가 다른 모형뿐 아니라, 심지어 같은 종속변수를 공유한 내포된 모형(nested model) 안에서조차 비선형 모형의 회귀계수는 일반적으로 불가능하며, 이것이 가능해지려면 비교하고자 하는 모형들에 걸쳐 스케일 팩터가 동등하다(equivalent)는 가정이 요구된다. 이 시뮬레이션의 주요 목적은 계수 비교가 아닌 상호작용효과를 비교하는 것이고, 시뮬레이션의 특성상 인위적으로 그와 같은 패러미터를 조정하여 일치시킬 수 있으므로 이 문제와는 무관하다.

각각의 시나리오는 다음과 같으며 모두 열지도(heatmap) 및 3D 표면도(3D surface plot)로 시각화한다. 수치 해석과 시각화를 위해 Python을 사용하였다. 모든 그래프와 코드는 저자의 GitHub (https://github.com/hxk271/nl_inteff)에서 확인할 수 있다.

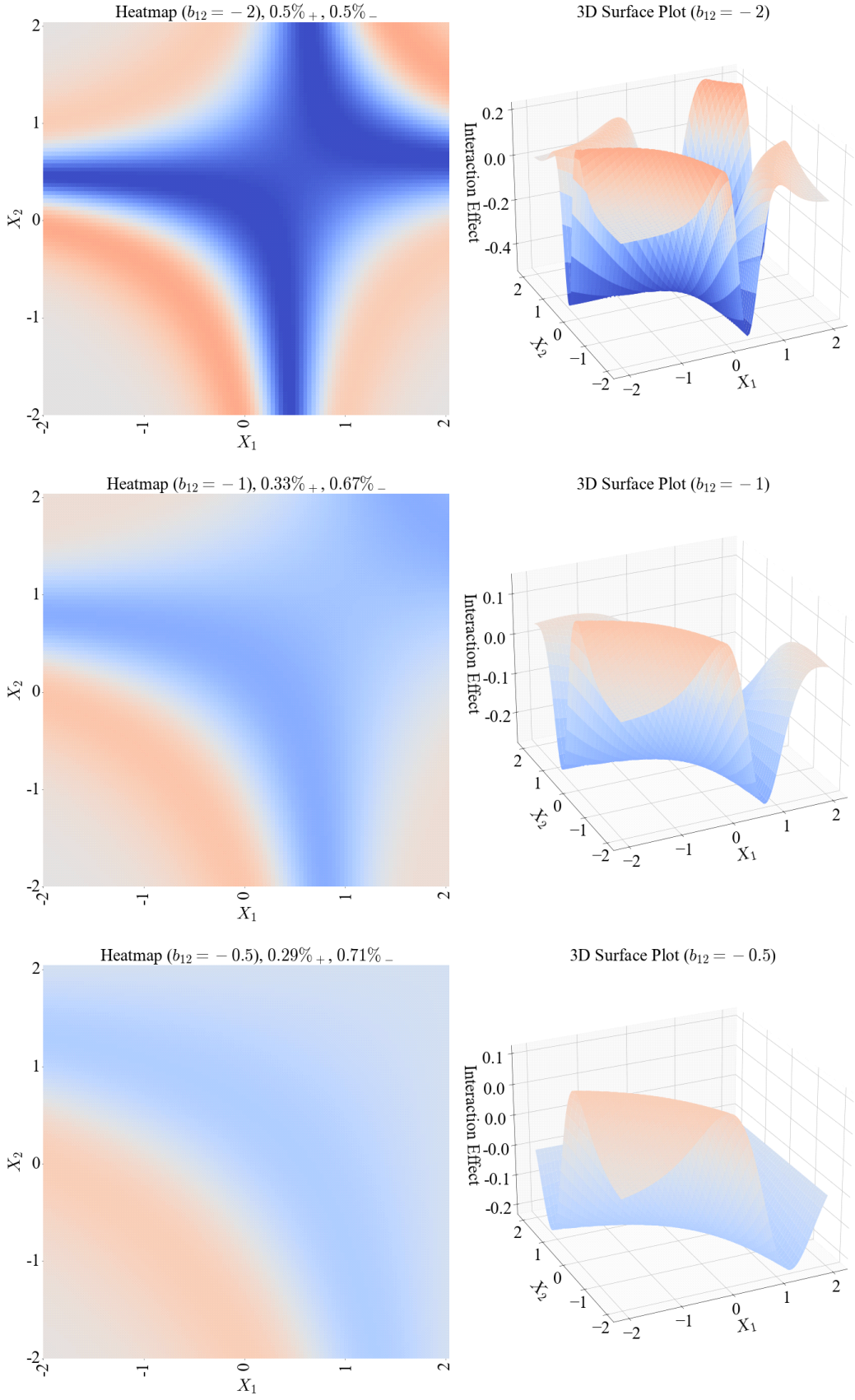
<표 1> 시뮬레이션 시나리오별 패러미터 설정

	b_1	b_2	b_{12}
1번 시나리오	1	1	{-2, -1, -0.5}
2번 시나리오	1.5	0.5	{2, 1, 0.5}
3번 시나리오	{-1, 0, 1}	1	0

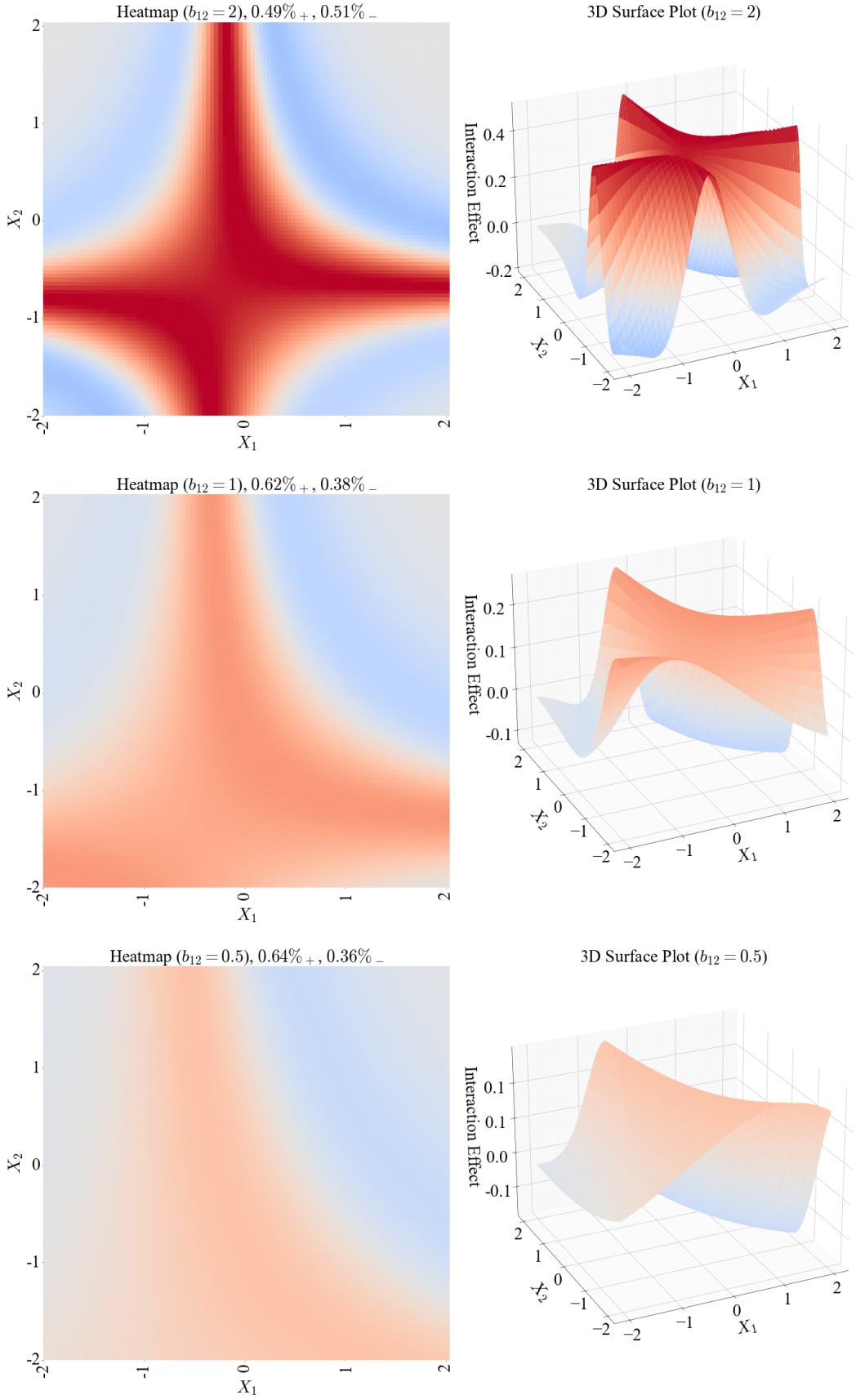
2. 분석결과

<그림 1>과 <그림 2>는 각각 첫 번째와 두 번째 시나리오 결과를 시각화한 것이다. 그림의 첫 번째 열(column)은 열지도를, 두 번째 열은 3D 표면도를 나타낸다. 붉은색이 강할수록(짙을수록) 양(+)의 방향으로 상호작용효과 $\partial^2 \Lambda(Y^*)/\partial X_1 \partial X_2$가 크게 나타나며, 푸른색이 강할수록 상호작용효과가 음(-)의 방향으로 크게 나타난다. 첫 번째 행(row)은 $b_{12} = -2$, 두 번째 행은 $b_{12} = -1$, 세 번째 행은 $b_{12} = -0.5$인 경우의 상호작용효과를 각각 시각화한다. 모든 그림에서 x축과 y축은 각각 공변량 X_1와 X_2을 나타낸다. 오른쪽의 3D 표면도에 한하여 상호작용효과의 크기를 별도의 z축으로 표현하고 있다. 이 시뮬레이션의 공변량 격자에서는 X_1과 X_2가 균등 분포하는 상황을 전제하였으나, 실제 연구 상황에서는 공변량의 분포 형태와 분산-공분산 행렬(variance-covariance matrix) 등 패러미터에 따라 그 결과가 달라질 수 있다.

<그림 1>에서 첫 번째 행은 상호작용항의 회귀계수가 명백히 음수($b_{12} = -2$)임에도 불구하고, 상호작용효과가 다양하게 나타나고 있음을 보여준다. 특히 $X_1, X_2 \in [-2, 2]$ 전체 범위의 공변량 공간에서 50% 정도의 상호작용효과는 양(+)의 방향으로 나타났음에 주목할 필요가 있다(실제 비율은 공변량 분포에 따라 상이하며, 가령 다변량 정규 분포라면 주어진 분산-공분산 행렬에 따라 그 비율이 달라질 수 있다). 이때 상호작용항의 회귀계수가 -2에서 -1을 거쳐 -0.5까지 감소하면, 양(+)의 상호작용효과가 나타나는 구간은 오히려 줄어드는 것을 확인할 수 있다. 그러므로 상호작용항이 상대적으로 크다고 해서, 표본에서 나타날 수 있는 상호작용효과의 대세까지 예단할 수 없다. 여러 행들을 대조하여 살펴보면 알 수 있듯, 상호작용항의 절대값이 클 때 상호작용효과의 크기가 비교적 뚜렷한 구간이 생겨난다. 그러나 그 구간을 벗어나면 상호작용항과 정반대의 효과를 보여주는 구간이 나타나므로, 상호작용항의 해석에는 많은 주의가 필요하다.



<그림 1> 상호작용효과 시뮬레이션 (1번 시나리오)



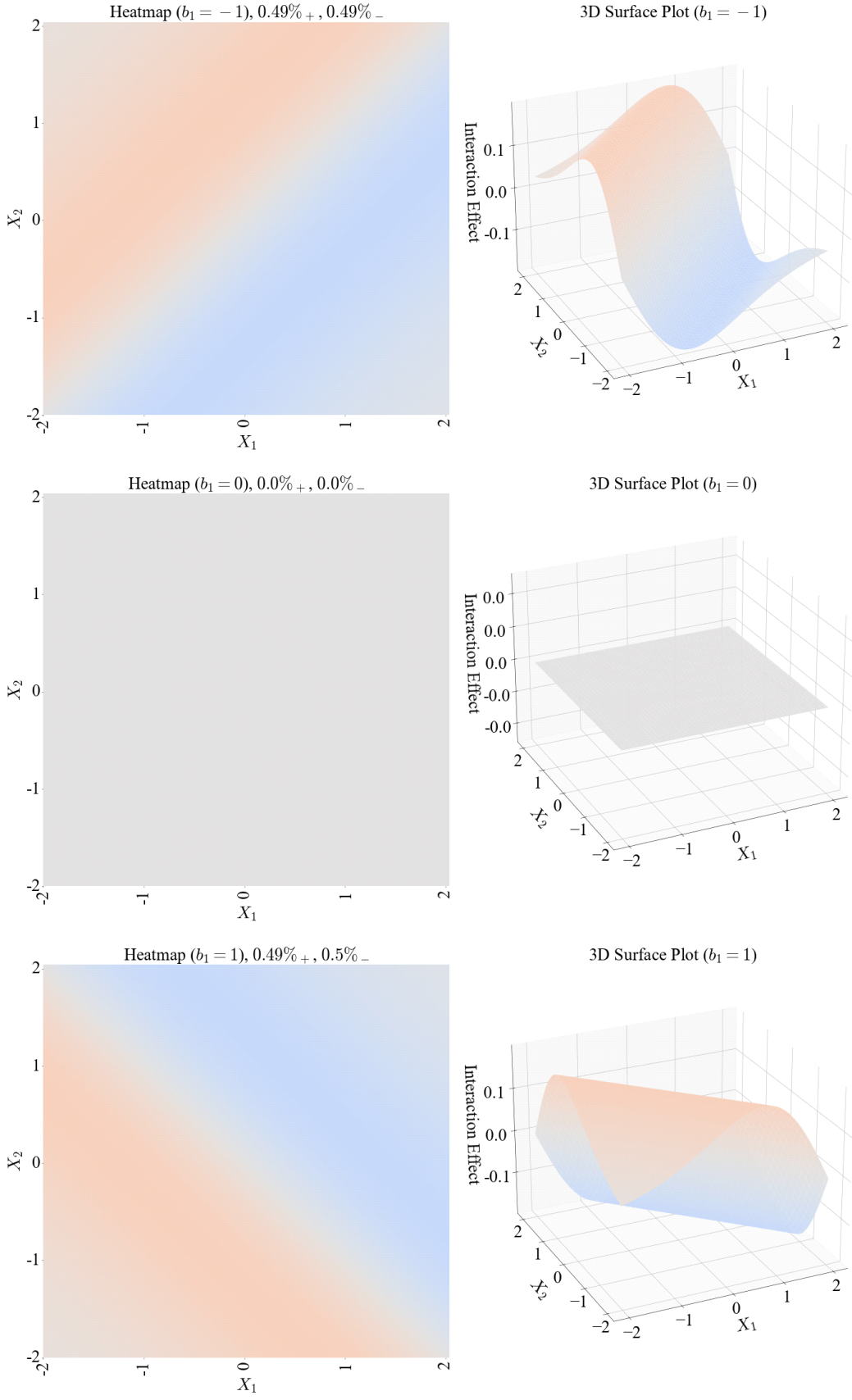
<그림 2> 상호작용효과 시뮬레이션 (2번 시나리오)

오른쪽 열의 3D 표면도는 $b_{12}=-2$일 때 종이학 꼴을 닮아있으나, 회귀계수의 절대값이 점차 감소함에 따라 점차 느슨하게 풀려 주어진 권역 안에서 음(-)의 상호작용효과가 나타나는 구간이 넓어짐을 보여주고 있다. 상호작용항의 계수 방향대로 음(-)의 상호작용효과를 얻는 구간이 점차 커지지만, 그러한 변화는 상호작용항의 계수가 음수로 커지는 방향이 아니라 0에 가까워지는 방향일 때 나타난다.

이와 유사하게, <그림 2>에서도 첫 번째 행은 상호작용항의 회귀계수가 양수($b_{12}=2$)임에도 불구하고, 공변량 X_1과 X_2의 조합에 따라 약 50% 가량의 공변량 공간에서 음(-)의 상호작용효과가 나타남을 확인할 수 있다. 오른쪽 열의 3D 표면도는 <그림 1>의 3D 표면도를 뒤집은 것과 유사하다.

<그림 1>과 <그림 2>에서 공통적으로 나타나는 특징 가운데 하나는 상호작용효과가 양(+)에서 음(-)으로 또는 음(-)에서 양(+)으로 뒤바뀌는 하얀색 구간을 경계로 십자 모양이 나타나는 점이다. 하얀색 구간은 상호작용항이 0이 아님에도 불구하고 상호작용효과가 0 또는 그에 매우 가깝게 근접하는 부분이다(즉 양에서 음으로 또는 음에서 양으로 효과의 방향이 바뀌는 부분이다). 식 (5)의 구성요소에 따라 나누어 살펴보면 그 이유를 대략 짐작할 수 있다. 누적분포함수의 정의상 $1 \geq \Lambda(Y^*) \geq 0$이므로 $1 \geq [1-2\Lambda(Y^*)] \geq -1$ 이기 때문에, 회귀계수와 공변량의 조합에 따라 십자 모양이 나타나는 구간에서 $(b_1+b_{12}X_2)(b_2+b_{12}X_1)$ 부분이 0에 근접하며 하얗게 나타난다.

한편 <그림 1>과 <그림 2>에서 두 시뮬레이션 모형은 상호작용항의 회귀계수가 정반대라는 것 이외에도, 주요항의 회귀계수 또한 달리 설정되어 있다. 그럼에도 불구하고, 전반적인 열지도와 3D 표면도는 상당히 유사하다. 주목할 만한 차이는 $X_1, X_2 \in [-2,2]$ 권역 안에서 특정 방향으로의 상호작용효과가 차지하는 구간이 달라지는 점이다. 보다 구체적으로, <그림 1>에서 회귀계수 -2, -1, -0.5일 때 양(+)의 상호작용효과 구간은 각각 50%, 33%, 29%이지만 <그림 2>에서 회귀계수 2, 1, 0.5일 때 음(-)의 상호작용효과 구간은 각각 51%, 38%, 36%로 나타난다.



<그림 3> 상호작용효과 시뮬레이션 (3번 시나리오)

지금까지 연구자가 가진 자료와 회귀계수 추정량의 특성에 따라, 엇비슷한 상호작용항 회귀계수에도 불구하고, 특정 방향으로의 상호작용효과가 더 혹은 덜 관찰될 수 있음을 확인하였다. 마지막으로 <그림 3>은 상호작용항의 회귀계수가 0인 경우를 시각화하고 있다. 비선형 모형에서 상호작용항을 아예 넣지 않았거나, 유의성 검정에서 귀무가설을 기각하지 못하고 $\beta_{12} = 0$인 상황이 이에 대응한다. 이 경우조차 명백히 상호작용효과는 0보다 크거나 작을 수 있다. 그러므로 비선형 모형에서는 상호작용항의 유무 또는 상호작용항의 유의성 검정 결과와 무관하게 상호작용효과는 나타날 수 있음을 재차 확인할 수 있다.

다만 두 번째 행을 통해 상호작용항의 회귀계수가 0일 때, 상호작용효과마저 정말로 사라지는 특수한 조건도 동시에 확인할 수 있다. 상호작용항의 회귀계수뿐 아니라, 주요항 중 하나 이상의 회귀계수도 0인 경우(두 번째 행의 $b_1 = 0$ 시나리오)가 이에 해당한다. 식 (5)를 다시 한번 살펴보면, $\lambda'(Y^*) = \lambda(Y^*)[1 - 2\Lambda(Y^*)]$와 상관없이, 이 시나리오에서는 우측의 모든 항이 0이 되므로, 이를 곱한 상호작용효과도 $X_1, X_2 \in [-2, 2]$의 모든 구간에 걸쳐 0이다.

전반적으로 시뮬레이션 결과는 상호작용효과가 (상호작용항의 크기를 통해) 전역적(global)으로 해석될 수 없고, 공변량 X_1과 X_2에 따라 지역적(local)으로 해석되어야 함을 시사한다. 상호작용항의 회귀계수 부호와 실제 상호작용효과의 부호는 불일치하는 구간이 상당히 많고, 계수 해석만으로는 진정한 효과의 방향을 알 수 없음을 시각적으로 확인할 수 있다.

Ⅳ. 대안

시뮬레이션 연구는 얻어진 추정량에 대응하여 주어진 공변량 공간에 따라 상호작용효과가 어떻게 분포될 수 있는가를 보여준다. 그러나 이 방식으로 사회학도가 자신의 연구에서 나타날 수 있는 상호작용효과를 하나하나 검토한다는 것은 그다지 실용적이지 못하다. 통계분석 소프트웨어에서 즉각 이용 가능한 보다 현실적인 대안이 필요하다. 이 장에서는 비선형 모형에서 상호작용효과를 확인하기 위해 연구자가 어떤 실용적 방법을 택할 수 있는지에 관해 논의한다. 각각의 대안을 이론적

으로 설명하면서, 또한 실제 분석 상황에 어떻게 그것들이 활용할 수 있는지 구체적인 예를 제시하기 위해 가공의 자료를 생성하였다.

자료생성 구조(data-generating process)는 다음과 같다: 표본 크기는 1,000으로 한다. 공변량 X_1은 평균이 0, 표준편차가 1인 정규분포를 따르는 연속변수이다. 공변량 X_2는 평균이 1, 표준편차가 2인 정규분포를 따르는 확률변수가 만일 0.5보다 작으면 1, 크면 0의 값을 갖는 가변수(dummy variable)이다. 진정한 상수와 회귀계수는 다음과 같이 설정한다: $\beta_0 = 0$, $\beta_1 = -1$, $\beta_2 = 1$, $\beta_{12} = 0.5$. 이렇게 만들어진 잠재변수 $Y^* = XB$를 사용하여 누적분포함수 $\Lambda(Y^*)$를 얻었다. 마지막으로 확률적 반응변수(response variable)를 생성하기 위해 일양 분포(uniform distribution)를 따르는 0과 1 사이의 임의의 수를 각 사례에 부여하고, 그 값이 위의 $\Lambda(Y^*)$보다 작으면 1, 크면 0의 값을 갖는 종속변수 Y를 만들었다.

이 자료생성 구조는 사회학 분야 전반에서 일반적인 양적 연구 상황과 유사하게 설계되었다. 가령 분석단위가 개인인 자료에 대해, 종속변수가 노조원 여부(1 또는 0)라면, X_1은 Z값으로 표준화된 인지된 노사관계 점수(클수록 우호적), X_2은 종사상 공공부문(1) 또는 사기업 부문(0)을 지시하는 가변수를 상상할 수 있다. 이 경우, 노사관계가 우호적일수록 노조원이 될 확률은 낮아질 것으로 예측할 수 있고, 다른 한편, 공공부문 노동자보다 사기업 부문 노동자가 노조원이 될 확률이 높을 것으로 예측할 수 있다. 동시에 인지된 노사관계 점수가 노조원 지위에 미치는 영향이 공사 부문에 따라 차등적일 것으로 예측한다면 X_1와 X_2의 상호작용항을 투입하게 된다.

이하에서는 크게 다음의 다섯 가지 방식을 검토한다. 첫째, 비선형 모형 대신 선형확률모형으로 상호작용효과를 확인하는 방법이 있다. 둘째, 예측 확률(predicted probabilities)을 시각화할 수 있다. 셋째, 다른 한 공변량의 변화에 따라 달라지는 평균한계효과(average marginal effect)를 시각화할 수 있다. 넷째, 평균 상호작용효과(average interaction effect; AIE)의 추정량과 표준오차를 수치 계산하는 방법이 있다. 마지막으로 한계 오즈비(marginal odds ratio)를 활용한 해석에 관해서도 검토한다. 각각은 고유한 특성과 장·단점이 있으므로 이를 잘 이해하고 연구자의 목적에 따라 적절히 활용하는 편이 바람직하다. 구체적인 자료와 대안을 실천하는 코드는 공개되어 있으므로 연구자가 각자 목적에 맞추어 수정하여 사용할 수 있다. 앞서 시뮬레이션에서는 연속변수와 연속변수의 상호작용효과를 검토하였고, 아래 대

안은 연속변수와 이산변수 간의 상호작용효과를 중심으로 설명하고 있다. 이 논문에서 더 다루어지지 못한 이산변수와 이산변수의 상호작용효과와 관련하여 중요한 비선형 이중차분 기법은 저자의 GitHub에서 코드와 시각화 자료를 참고할 수 있다.

1. 선형확률모형

1) 성격과 장·단점

첫 번째로 고려할 만한 대안은 처음부터 비선형 모형을 피하고 선형확률모형(linear probability model; LPM)을 통해 상호작용효과를 추정하는 것이다. 선형확률모형은 이분형 종속변수에 대해서 관찰된 0과 1의 값을 각각 $P(Y=0|X)$와 $P(Y=1|X)$인 것으로 의제한 뒤, (로짓 모형 대신 선형 모형을 세우고) 보통최소제곱(ordinary least square)에 따라 평이하게 회귀계수를 추정하는 방법이다.

이 방법은 연구자 입장에서 분석의 수행과 해석이 간편하다는 점이 가장 큰 장점이다. 이 연구에서 제기하고 있는 문제를 돌이켜보면, 선형확률모형은 비선형 모형이 아니므로, 적어도 상호작용효과의 해석 문제는 발생하지 않는다. 다차항(polynomial terms)처럼 복잡한 모형이라면 선형확률모형의 이러한 간편성이 연구자에게 있어 외면하기 어려운 중요한 강점이 된다. 추정된 예측 확률이 매우 낮거나 크지 않으면 로짓/프로빗 모형과 제법 유사한 결과를 제공한다는 점, 그리고 가정 위배로 인한 파급효과가 우려스럽다면 이분산성을 반영한(heteroskedasticity-consistent) 표준오차를 대신 보고할 수 있다는 점 등은 선형확률모형의 추가적인 장점이다. 그러나 이 방식으로 예측된 값(predicted values)은 예측 확률이 아니므로 $[0,1]$ 바깥의 값을 얻을 수 있다는 점, 이분산성(heteroskedasticity) 문제로 인해 더 이상 최우 불편추정량(best linear unbiased estimates)임을 보장할 수 없다는 점 등에서는 명확한 한계가 있다(Long 1997; 35-40).

선형확률모형은 비교적 간단하게 상호작용효과가 통계적으로 유의한가 여부를 진단할 수 있는 교차검증 도구로서 여전히 가치를 갖는다. 만약 로짓/프로빗 모형의 결과와 선형확률모형의 결과가 심각하게 다르다면, 보다 시간과 노력을 기울여 상호작용효과의 크기와 범위를 진지하게 검토할 수 있다.

2) 실제 활용

앞서 설명한 대로 생성한 가공의 자료에 상호작용항을 포함한 로짓 모형과 선형

확률모형을 적합시킨 뒤, 각각 회귀계수와 상수를 추정한 결과는 <표 2>와 같다. b_1과 b_2는 각각 연속변수와 가변수의 회귀계수이고, b_{12}는 상호작용항의 회귀계수이며, b_0는 상수이다. 로짓 회귀분석과 선형확률모형의 적합성은 각각 Pseudo R^2와 R^2로 보고하였다. 두 모형에 따라 1로 예측된 확률 $P(\widehat{Y=1}|X)$이 0.5보다 크고 실제 자료에서도 1이거나, $P(\widehat{Y=1}|X)$가 0.5보다 작고 실제 자료에서도 0인 경우 정확하게 분류한 것으로 평가하는 방식도 적합도 지수로 제시하였다(% correctly classified).

주목할 만한 부분은 로짓 모형에서 상호작용항이 통계적으로 유의하지 않으나 ($p < 0.286$), 선형확률모형에서는 95% 신뢰수준에서 통계적으로 유의하다는 점이다($p < 0.011$). 일반적인 표준오차를 사용하여 t값을 계산하더라도 이 값은 여전히 통계적으로 유의하다.

상반된 결과 중 어느 쪽을 믿을 수 있을까? 아직 알 수 없다. 그러나 선형확률모형에서 유의한 결과를 그냥 지나치거나, 로짓 모형에서 상호작용항이 통계적으로 유의하지 않기 때문에 상호작용효과가 없을 것이라고 섣불리 포기하기보다, (적어도 일정 구간 내) 상호작용효과를 좀 더 진지하게 살펴볼 필요가 있을 것이다.

<표 2> 로짓 추정치와 선형확률모형 추정치의 비교

Coefficients	Logit estimates		LPM estimates	
	Est.	SE	Est.	SE
b_1	-0.761***	(0.096)	-0.168***	(0.016)
b_2	1.113***	(0.151)	0.235***	(0.028)
b_{12}	0.170	(0.160)	0.066*	(0.026)
b_0	0.072	(0.087)	0.516***	(0.019)
Pseudo R^2	0.117			
R^2			0.148	
% correctly classified	67.6%		67.5%	
Number of observations	1,000			

Note: Heteroskedasticity-consistent standard errors for LPM estimates in parentheses.
*$p < 0.05$, ***$p < 0.001$.

2. 예측 확률의 시각화

1) 성격과 장·단점

선형확률모형을 벗어나 일단 비선형 모형을 사용하기로 한다면, 변화하는 예측 확률을 시각화하는 방법이 편리하다. 우선 로짓/프로빗 회귀모형 등 비선형 모형을 통해 회귀계수를 추정한 뒤, 간단한 대수 조작을 통해 다음의 예측 확률을 계산한다.

$$\widehat{P(Y=1|X)} = \Lambda(Y^*) = \frac{1}{1+\exp\left[-(b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2)\right]}$$

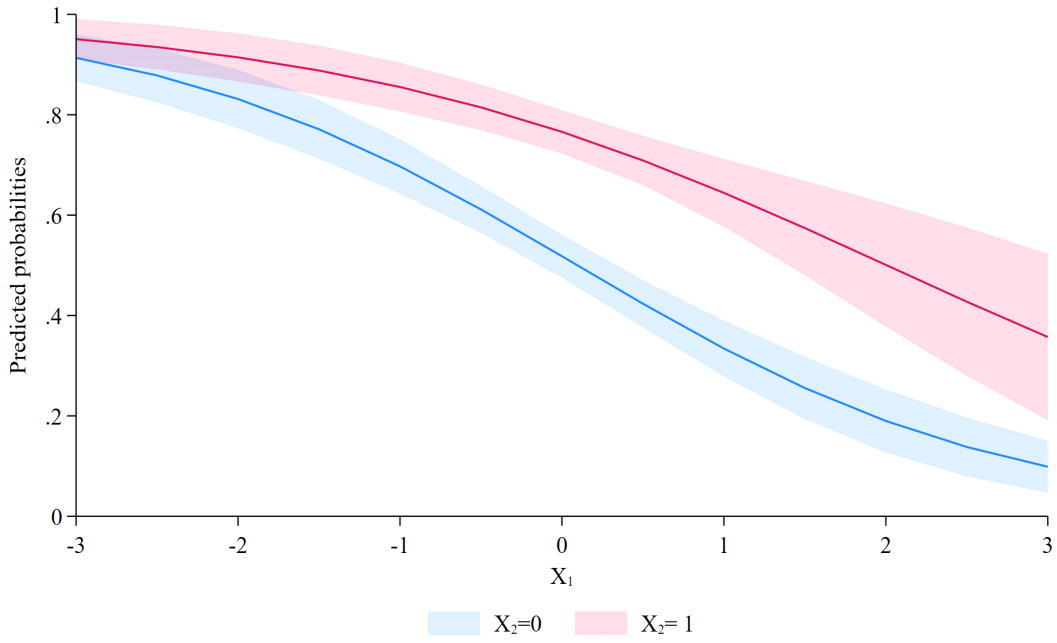
그리고 주어진 자료에 대응하는 X_1과 X_2를 대입하여 예측 확률을 계산한 뒤, X_1의 변화에 따라 평균적인 $\widehat{P(Y=1|X)}$가 변화하는 모습을 X_2 구간별로 각각 시각화한다(Greene 2010). 이때 핵심은 X_1와 평균적인 $\widehat{P(Y=1|X)}$의 연관성을 선그래프로 나타내고, X_2에 따라 그 연관성이 제법 달라져 분기하는(diverging) 양태를 드러내야 한다는 점이다. 만일 X_2가 가변수라면 두 개의 선을 그릴 수 있으며, 두 선이 얼마나 서로 다른 각도로 뻗어나가는가를 통해 상호작용효과를 표현할 수 있다.

다만 예측 확률을 시각화할 때 각 지점에 걸쳐 반드시 신뢰구간을 함께 그려야 한다. 신뢰구간은 계산하기 까다롭고 그래프를 다소 지저분하게 보이게 할 수 있지만, 상호작용효과의 통계적 유의성을 판별할 때 필수적이기 때문이다. 하나의 신뢰구간은 (연구자가 감수할 용의가 있는) 귀무가설이 참이지만 이를 기각할 오류(Type I error)의 크기를 반영하며, 연구자에 따라 99%, 95%, 90% 등의 기준을 선택할 수 있다. 만일 두 예측 확률 곡선 사이에 신뢰구간이 겹쳐있지 않다면, 이는 적어도 그 구간(local)에서 X_1의 $\widehat{P(Y=1|X)}$에 대한 효과가 X_2에 따라 상이함을 의미한다(즉 상호작용효과가 존재한다). 만일 두 예측 확률 곡선의 신뢰구간이 서로 겹쳐있어도, 해당 구간에서 상호작용효과가 통계적으로 유의하지 않다고 섣불리 단정할 필요는 없다(Schenker & Gentleman 2001). 두 신뢰구간이 겹쳐있다면 연구자가 직접 왈드 검정(Wald test) 등을 수행하여 해당 구간에서 상호작용효과가 통계적

으로 유의한지 재차 확인할 필요가 있다.

2) 실제 활용

<그림 4>는 앞서 설명한 가공의 자료에서 조건부 예측 확률을 계산하고, 이를 X_1의 변화에 따라 시각화하되, 상호작용효과를 표현하기 위해 X_2 값에 따라 두 개의 다른 궤적으로 나타내고 있다. 앞서 설명하였듯, 로짓 모형에서 상호작용항은 통계적으로 유의하지 않았으나, 대부분의 구간에 걸쳐 두 예측 확률 곡선은 통계적으로 유의하게 다른 것으로 나타난다(이 점에서 오히려 선형확률모형에서의 통계적 유의성과 더 일치하는 해석을 제공한다).



<그림 4> X_1에 따른 예측 확률과 95% 신뢰구간

이때 X_1이 $[-3,-1.5]$ 사이의 구간에서 두 예측 확률 곡선의 95% 신뢰구간은 서로 중첩되어 있음에 주목할 필요가 있다. 시뮬레이션을 통해 밝혔듯, 공변량 공간에 따라 상호작용효과는 서로 다르게 나타날 수 있다. 만일 연구 목적상 구간 상호작용효과(local interaction effects)를 좀 더 꼼꼼하게 살펴볼 필요가 있다면, 구간별로 왈드 검정을 수행해 볼 수 있다.

<표 3>은 X_1의 $[-3,3]$ 구간에서 0.5의 간격(interval)으로 상호작용효과를 살펴보고 있다. 실제로 <그림 4>처럼 −3과 −2.5의 두 구간에서는 상호작용효과가 통계적으로 유의하지 않음을 확인할 수 있다. 다만 −2 지점에서는 비록 시각적으로 두 신뢰구간이 중첩되어 있으나, 왈드 검정에 따르면 95% 신뢰수준에서 통계적으로 유의하게 구간 상호작용효과가 0과 다름을 시사하고 있다. 앞서 설명하였듯 두 신뢰구간의 중첩 여부로 통계적 유의성 판정은 지나치게 보수적이므로 연구자의 주의가 필요하다(Online Appendix 참고).

<표 3> X_1에 따른 상호작용효과

	$X_2 = 1$ versus $X_2 = 0$		
X_1	Est.	SE	χ^2
-3	0.037	0.031	1.41
-2.5	0.057	0.036	2.52
-2	0.083	0.039	4.63*
-1.5	0.117	0.039	8.89**
-1	0.158	0.037	18.05***
-0.5	0.203	0.033	37.18***
0	0.248	0.031	63.62***
0.5	0.285	0.035	67.09***
1	0.310	0.045	47.99***
1.5	0.319	0.058	30.65***
2	0.311	0.070	19.49***
2.5	0.289	0.081	12.67***
3	0.258	0.089	8.50***
Joint test			72.56***

Note: SE stands for delta-method standard errors.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

3. 조건부 한계효과의 시각화

1) 성격과 장·단점

한편, 평균 상호작용효과에 그치지 않고, 구간별로 이를 살펴보려는 시각화 기법이 일찌감치 제안된 바 있다(Ai & Norton 2003; Norton, Wang, & Ai 2004). 이들의 제안에 따르면, 일단 로짓/프로빗 회귀모형 등 비선형 모형을 통해 회귀계수를 추정하고, 이로부터 계산된 예측 확률을 x축으로 하고, 개별적인 상호작용효과 및 t값을 y축으로 하는 산점도(scatterplot)를 두 개 제시할 수 있다. 이 그래프들은 다양한 구간에 걸친 상이한 상호작용효과를 살펴볼 수 있도록 고안된 것이다. 그러나 막상 예측 확률의 증가에 따라 상호작용효과가 어떻게 달라지는가보다는, 차라리 특정 공변량(가령 X_1)의 변화에 따라 다른 쪽 공변량의 평균한계효과가 어떻게 달라지는가를 시각화하는 편이 훨씬 직관적이다. 즉 (예측 확률이 아니라) X_1을 x축에, 평균한계효과 $\sum\left(\partial^2 \Lambda(Y_i^*)/\partial X_{i2}\right)$를 y축에 둔 적합선(fitted curve)과 그 신뢰구간을 시각화할 수 있다(Berry et al. 2010; Uberti 2022).

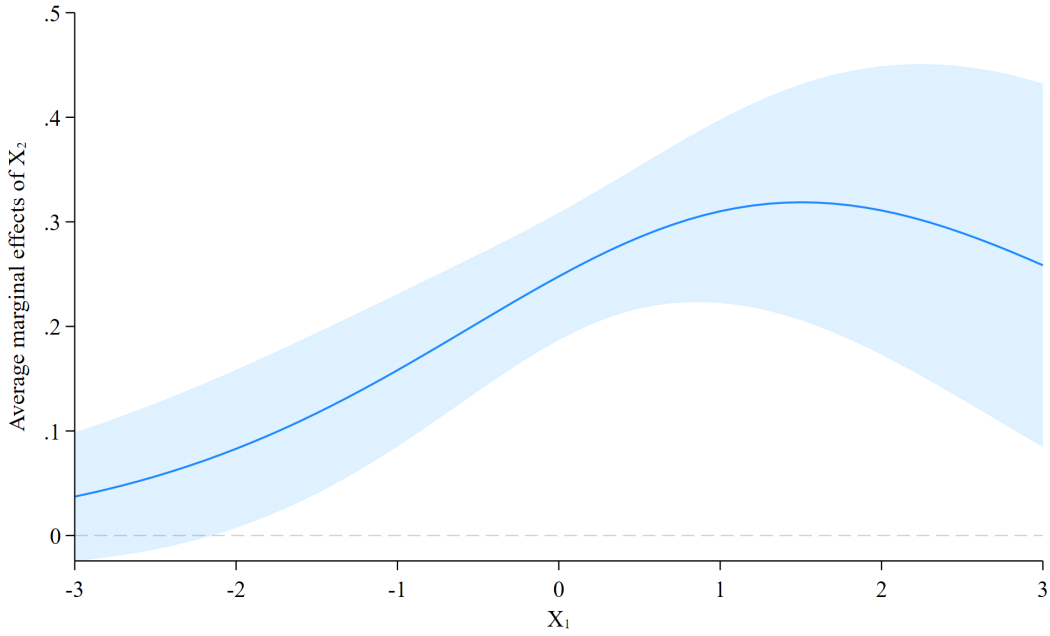
이 방법은 앞서 설명한 예측 확률의 시각화와 본질적으로 크게 다르지 않은 것임을 다소간 유의할 필요가 있다. 가령 <그림 5>의 조건부 한계효과 시각화 자료는 <그림 4>에 나타난 두 집단 간 예측 확률의 차이를 반영하고 있다. 또한 <그림 5>가 <표 3>에서 제시된 추정치를 나타내고 있다는 점에서도 계산 원리상 차이가 있다기보다는 단지 시각화의 방식이 다를 뿐이다.

한편 이 방법의 장점은 상호작용효과의 계산 원리인 교차편미분 개념을 비교적 명확히 반영하고 있다는 것이다. 앞서 설명하였듯, 상호작용효과란 가령 X_2의 변화가 종속변수 Y의 조건부 기대값에 미치는 영향이 X_1의 변화에 따라 달라지는 정도를 의미하며, 이는 교차편미분으로 표현된다. 이 시각화 기법은 교차편미분 우변의 첫째 부분(괄호 안)을 x축으로, 둘째 부분을 y축으로 삼아 상호작용효과를 표현한다.

$$\frac{\partial^2 E(Y|X_1, X_2)}{\partial X_1 \partial X_2} = \frac{\partial}{\partial X_1}\left(\frac{\partial E(Y|X_1, X_2)}{\partial X_2}\right)$$

2) 실제 활용

앞서 설명한 가공의 자료로부터 X_2의 평균한계효과가 X_1에 따라 어떻게 달라지는가를 시각화한 결과는 <그림 5>와 같다. 이 그래프를 통해 X_2의 평균한계효과 $\sum\left(\partial^2 \Lambda(Y_i^*)/\partial X_{i2}\right)$는 전반적으로 X_1이 커질수록 증가함을 확인할 수 있다. 한편 $X_1 < -2$의 구간에서 X_2의 평균한계효과 95% 신뢰구간이 0을 포함하고 있으므로, 그 구간부터 상호작용효과는 통계적으로 유의하지 않음을 확인할 수 있다. 또한 $X_1 > 1$ 구간부터는 상호작용효과가 점차 감소하는데, 이를 통해 X_1과 X_2이 단조증가하는(monotonically increasing) 상호작용효과는 아님을 확인할 수 있다.



<그림 5> X_1 조건부 X_2의 상호작용효과

4. 평균 상호작용효과 계산 및 시각화

1) 성격과 장·단점

네 번째 대안은 모든 개별적 상호작용효과의 평균을 계산하여 이른바 평균 상호작용효과(AIE)를 보고하는 방식이다. 이 값은 개별적인 상호작용효과 $\partial^2 \Lambda(Y_i^*)/\partial X_{i1}\partial X_{i2}$를 표본 안 모든 관측치에 대해 각각 계산한 뒤, 그것들의 표본평균을 계

산한다는 점에서 평균한계효과(average marginal effects)의 원리와 같다.

$$AIE = \sum_{i}^{N} \left(\frac{\partial^2 \Lambda(Y_i^*)}{\partial X_{i1} \partial X_{i2}} \right)$$

이 방식의 가장 중요한 장점은 교차편미분 값을 명확하게 계산하여 비선형 모형에서 상호작용효과를 둘러싼 문제를 정면에서 해결한다는 것이다. 바로 그 때문에 이 방식이 어쩌면 가장 폭넓게 검토되어 지지받고 있는 것 같다(Ai & Norton 2003; Karaca-Mandic, Norton, & Dowd 2012; Norton, Wang, & Ai 2004; Radean 2023). 이 방식과 관련하여 주의해야 할 단점도 있다. 우선 이 연구에서도 시뮬레이션을 통해 검토하였듯, 실제 상호작용효과는 공변량에 따라 국소적으로 다를 수 있는데, 그 수치 계산이 전역적(global)으로 이루어지는 평균 상호작용효과의 특성상 이를 고려하지 않는다는 것이다(물론 때에 따라 이런 부분이 단순함이라는 강점으로 여겨질 수도 있다). 다음으로, AIE 추정량과 시각화 모두 결과물이 연구자에게 실질적인 해석에 결정적인 도움을 주지 못한다는 점이다. 가령 Radean (2023)은 평균 상호작용효과의 계산 원리를 검토하고, 그 추정량과 신뢰구간을 상자-수염 도표(box-whiskers plot)처럼 시각화하는 방안을 제시하였다. 그러나 하나의 스칼라로 제시된 평균 상호작용효과 값만으로 연구자의 상호작용효과에 관한 이론적 예측을 어떤 식으로 지지하는지 해석하기 어렵다. 설령 그 추정량을 신뢰구간과 함께 시각화하더라도 해석에 추가적인 도움을 주는지는 다소 모호하다.

2) 실제 활용

앞서 설명한 가공의 자료로부터 평균 상호작용효과를 계산한 결과는 <표 4>와 같다.

<표 4> 평균 상호작용효과

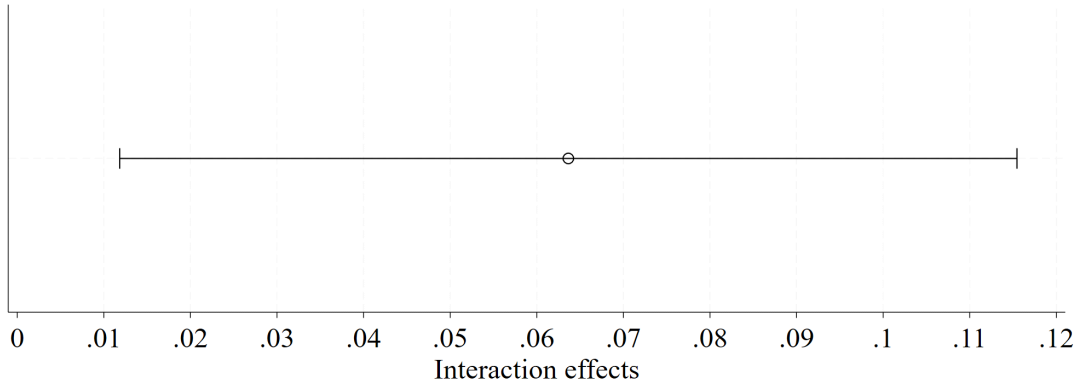
	$X_2 = 1$ versus $X_2 = 0$		
AIE	Est.	SE	χ^2
	0.064	0.026	5.8**

Note: Delta-method standard errors.

**$p < 0.01$.

Est.는 개별적인 상호작용효과 $\partial^2 \Lambda(Y_i^*)/\partial X_{i1}\partial X_{i2}$를 모든 관측치에 대해 계산한 뒤 얻은 표본평균이고, SE는 그 표준오차이다. 표본 내 평균 교차편미분 값은 0.064으로 이는 99% 신뢰수준에서 통계적으로 유의하게 0과 다름을 확인하였다.

다음으로 <그림 6>은 Radean (2023)이 제안한 방식대로 그린 것이다. 가운데 점은 평균 상호작용효과를 나타내며, 수염의 끄트머리는 95% 신뢰구간의 상하한을 표현한다. 여기서 이 신뢰구간이 0을 포함하지 않았기에 95% 신뢰수준에서 통계적으로 유의하게 상호작용효과가 존재한다고 말할 수 있다.



<그림 6> 평균 상호작용효과와 95% 신뢰구간

이상의 결과를 통해 연구자는 무엇을 알 수 있을까? 우선 <표 2>의 로짓 회귀분석 결과는 달리 상호작용효과가 통계적으로 유의하다는 결론을 지지한다. 이는 그 자체로 큰 장점일 수 있다. 그러나 이 방식은 상호작용효과의 본질인 교차편미분에 가장 부합한 접근임에도 불구하고, 사회학 연구에서 상호작용효과를 이론적으로 해석해 나갈 때는 이 이상 도움을 주지 못한다.

5. 한계 오즈비 해석

1) 성격과 장·단점

마지막으로 로짓 회귀모형에서 흔히 사용되는 해석법인 오즈비(odds ratio)를 살펴본다. (한계효과와는 달리) 오즈비는 특정 공변량을 상대적인 효과 크기로서 해석할 수 있고, 종속변수의 한계분표(marginal distribution)에 대해 강건하다는 점 등의 장점을 갖기 때문에(Karlson & Jann 2023), 상대적 불평등을 연구할 때 특히 인기

있는 해석 방법으로 여겨져 왔다. 그런데 주요항에 대한 오즈비 해석을 넘어, 상호작용항에 대해서도 오즈비 해석을 통해 상호작용효과를 도출할 수 있을까?

앞서 식 (2)의 우변을 아래와 같이 선형화하여 살펴보면, b_{12}는 $\Lambda(Y^*)$ 자체가 아니라 $\log(\Lambda(Y^*)/[1-\Lambda(Y^*)])$, 즉 로그 오즈(log odds)의 상호작용항이자 상호작용효과임을 알 수 있다.

$$\Lambda(Y^*) = \frac{1}{1+\exp(-Y^*)}$$
$$\log\left(\frac{\Lambda(Y^*)}{1-\Lambda(Y^*)}\right) = b_0 + b_1X_1 + b_2X_2 + b_{12}X_1X_2$$

따라서 로그 오즈를 있는 그대로 사용한다면, 마치 선형 모형에서처럼 비선형 모형에서도 상호작용항을 상호작용항효과 자체로 해석할 수 있다. 비록 이것은 중요한 장점이나, 다시 로그 오즈는 직관에 부합하지 않는 척도라는 것은 단점이다.

그렇다면 $\exp(b_{12})$를 계산하여 주요항 계수의 해석처럼 로그 오즈 또는 오즈비로 상호작용효과를 해석할 수는 없을까? 가령 X_1이 한 단위 증가(+1)하고, X_2는 0에서 1로 변화하는 순간의 상호작용효과를 오즈비로 직접 계산해 보면, $\exp(b_{12})$는 아래처럼 오즈비의 비(ratio of odds ratios)에 가까운 해석을 얻게 된다.

$$\exp(b_{12}) = \frac{\log\left(\frac{\Lambda(Y^*|X_1+1,X_2=1)}{1-\Lambda(Y^*|X_1+1,X_2=1)}\right)}{\log\left(\frac{\Lambda(Y^*|X_1,X_2=0)}{1-\Lambda(Y^*|X_1,X_2=0)}\right)}}{\frac{\log\left(\frac{\Lambda(Y^*|X_1+1,X_2=0)}{1-\Lambda(Y^*|X_1+1,X_2=0)}\right)}{\log\left(\frac{\Lambda(Y^*|X_1,X_2=0)}{1-\Lambda(Y^*|X_1,X_2=0)}\right)} \cdot \frac{\log\left(\frac{\Lambda(Y^*|X_1,X_2=1)}{1-\Lambda(Y^*|X_1,X_2=1)}\right)}{\log\left(\frac{\Lambda(Y^*|X_1,X_2=0)}{1-\Lambda(Y^*|X_1,X_2=0)}\right)}}$$

$$= \frac{\log\left(\frac{\Lambda(Y^*|X_1+1,X_2=1)}{1-\Lambda(Y^*|X_1+1,X_2=1)}\right)}{\log\left(\frac{\Lambda(Y^*|X_1+1,X_2=0)}{1-\Lambda(Y^*|X_1+1,X_2=0)}\right) \cdot \log\left(\frac{\Lambda(Y^*|X_1,X_2=1)}{1-\Lambda(Y^*|X_1,X_2=1)}\right)}$$

이에 따라 몇몇 연구자들은 $\exp(b_{12})$를 상호작용효과의 오즈비로 해석하는 것은 해석이 지나치게 난해하므로 피할 것을 권고하였다(Karaca-Mandic, Norton, & Dowd 2012). 주요항에 비한다면 상호작용항의 오즈비 해석이 까다롭고 직관적으로 이해하기 어려운 것은 사실이다. 그럼에도 (1) 로그 오즈로 상호작용항을 상호작용효과로써 해석하거나, (2) $\exp(b_{12})$를 상호작용효과의 오즈비로 해석하는 것에는 수학적으로 아무런 흠결이 없으므로, 연구 목적에 따라서 충분히 이 방식을 사용할 수 있다.

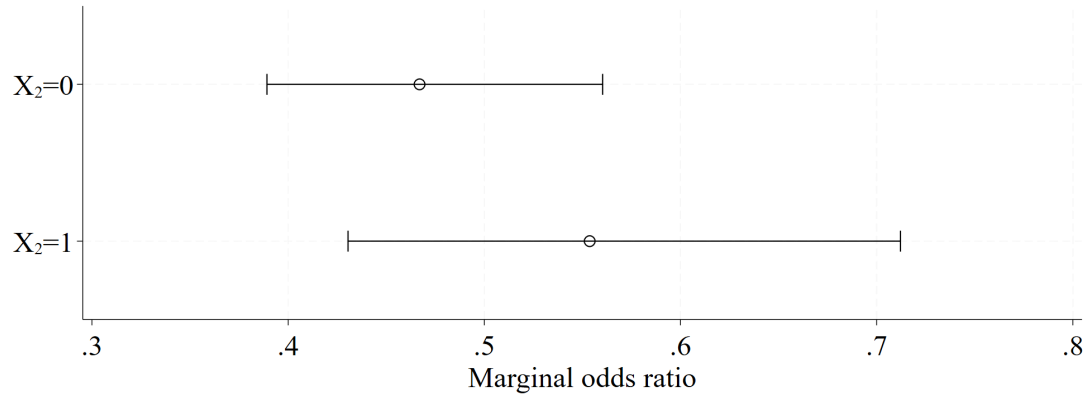
2) 실제 활용

한편 보다 최근에는 다음과 같이 로짓 모형의 회귀계수를 이른바 한계 오즈비 개념으로 해석할 수 있다는 논의가 제시되었다(Karlson & Jann 2023).

$$MOR = \frac{\left(\frac{E_X[P(Y_1 = 1|X = x)]}{1 - E_X[P(Y_1 = 1|X = x)]} \right)}{\left(\frac{E_X[P(Y_0 = 1|X = x)]}{1 - E_X[P(Y_0 = 1|X = x)]} \right)}$$

한계 오즈비 개념은 일반적인 오즈비 해석에 비해 몇 가지 차별점을 갖는다. 먼저 Donald Rubin이 정식화한 잠재적 결과 프레임워크(potential outcome framework)에 따라, Y_1과 Y_0는 각각 처리집단인 경우와 통제집단인 경우에 대응하며, $P(Y_1 = 1)$와 $P(Y_0 = 1)$은 만일 이들이 처리를 받았을 때 얻게 될 반응 확률을 의미한다. 다음으로, 이들의 연구에서 제시되는 한계 오즈비는 모집단-평균 추정량(population-average estimates)인데, 이는 일반적인 오즈비와는 달리 데이터 안에 주어진 공변량 X에 대해 Y_1과 Y_0의 기대값 $E_X[P(\cdot)]$을 사용하여 (적어도 데이터에서 관찰된) 이질성(heterogeneity)을 평균화(average-out)하기 때문이다.

Karlson & Jann (2023)은 통계분석 패키지인 Stata에서 사용할 수 있는 명령어 lnmor를 공개하였으므로, 연구자들은 이제 손쉽게 한계 오즈비를 추정할 수 있다. 가령 아래와 같은 <그림 7>을 통해 X_2가 각각 0인 경우와 1인 경우에 대응한 한계 오즈비를 시각화할 수도 있다. 한계 오즈비의 중첩된 신뢰구간을 해석할 때 역시 앞서 설명한 주의 사항을 따라야 하는 것은 물론이다.



<그림 7> 한계 오즈비 추정치와 95% 신뢰구간

V. 결론

이 연구는 로짓/프로빗 등 비선형 회귀모형에서 상호작용효과를 계산하는 원리가 선형 회귀모형과 근본적으로 다름을 확인하고, 올바른 상호작용효과의 계산법을 유도하여, 상호작용항에만 집중한 기존의 해석이 잘못되었음을 이론적으로 검토하였다. 또한 시뮬레이션을 통해 그 한계효과의 통계적 유의성이 전역적이라기보다 구간에 따라 상이할 수 있음을 보이고, 다양한 시나리오에서 상호작용항의 유무 및 계수 크기/방향과 다르게 나타나는 상호작용효과의 이질성을 재차 확인하였다.

이 연구의 핵심적인 시사점은 다음과 같다. 첫째, 비선형 회귀모형에서 상호작용효과를 해석할 때, 상호작용항에만 집중해서는 곤란하다. 상호작용항의 계수 부호와 실제 상호작용효과와 불일치하는 경우가 매우 흔하고, 계수 해석만으로는 진정한 효과의 방향조차 알 수 없기 때문이다. 특히 상호작용항이 존재하지 않거나 통계적으로 유의하지 않은 경우에도 상호작용효과가 나타날 수 있다.

둘째, 이에 따라 다섯 가지 대안의 장·단점을 충분히 이해하고, 연구 목적에 맞추어 적절히 활용해야 한다. 이 연구에서 작성된 모든 자료와 코드는 공개되어 있다(https://github.com/hxk271/nl_inteff).

각 대안들은 고유한 장·단점을 갖고 있으므로 어느 방식 하나만을 추천하기 어렵다. 그러나 대체로 사회학도가 비선형 회귀모형에서 상호작용효과를 분석할 때,

다음의 원칙을 고려할 필요가 있다: (1) 상호작용항의 계수에만 의존하지 않고 반드시 교차편미분을 직접 계산하고, (2) 필요시 전역적 해석에 그치지 않고 구간별 효과 변화를 살펴보며, (3) 적절한 시각화를 통해 상호작용효과의 이질성을 확인해야 한다.

이 연구는 주로 이분형 종속변수에 초점을 맞추었으나, 여기서 검토한 상호작용항 유도와 해석 문제는 모든 종류의 비선형 모형으로 일반화될 수 있다. 사회학 연구에서도 종종 사용되는 순서형(ordered) 또는 다항(multinomial) 로짓/프로빗 모형, 다층모형(multilevel models), 일반화 추정 방정식(generalized estimating equations), 구조 방정식 모형(structural equations models) 등 모형의 구조를 막론하고, 비선형 모형이라면 반드시 선형 모형에서의 상호작용항 해석과는 달리 접근해야 한다.

다만 기존의 통계분석 소프트웨어가 새롭게 고안되거나 보다 고급의 비선형 모형에 대해서도 예측 확률, 평균한계효과를 계산하고 시각화 기능까지 지원하는지는 별개의 문제이다. 이런 경우 연구자는 예측 확률이나 델타 방법 추정치를 직접 계산해야 하는 부담이 뒤따르게 되므로, 선형확률모형을 우선적인 대안으로 고려할 수 있다.

이 연구에서는 로짓과 프로빗 이외의 추가적인 비선형 모형에 관한 적절한 해석과 시각화는 모두 다루지 못하였으므로, 이에 관해 다른 연구자가 참고할 만한 추가 연구가 필요하다. 머신러닝과 딥러닝 프레임워크의 확산과 함께 비선형적인 모형들이 사회과학에 도입되고 있는 상황에서, 높은 해석 가능성(interpretability)을 확보하면서도 정확한 상호작용효과를 파악할 수 있는 방법론과 지침의 개발이 중요한 과제가 될 것이다.

참고문헌

Ai, Chunrong and Edward Norton. 2003. “Interaction Terms in Logit and Probit Models.” *Economics Letters* 80(1): 123-129.

Athey, Susan and Guido Imbens. 2006. “Identification and Inference in Nonlinear Difference-in-differences Models.” *Econometrica* 74: 431–497.

Berry, William D., Jacqueline H.R. DeMeritt, and Justin Esarey. 2010. “Testing for Interaction in Binary Logit and Probit Models: Is a Product Term Essential?”

American Journal of Political Science 54(1): 248-266.

Breen, Richard, Kristian Bernt Karlson, and Anders Holm. 2018. "Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models." *Annual Review of Sociology* 44: 39-54.

Cramer, J.S. 2007. "Robustness of Logit Analysis: Unobserved Heterogeneity and Mis-specified Disturbances." *Oxford Bulletin of Economics and Statistics* 69(4): 545-555.

Greene, William. 2010. "Testing Hypotheses about Interaction Terms in Nonlinear Models." *Economics Letters* 107(2): 291-296.

Karlson, Kristian Bernt, and Benn Jann. 2023. "Marginal Odds Ratios: What They Are, How to Compute Them, and Why Sociologists Might Want to Use Them." *Sociological Science* 10: 332-347.

Long, J. Scott. 1997. *Regression Models for Categorical Dependent Variables*. Thousand Oaks, CA: Sage.

Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It." *European Sociological Review* 26(1): 67-82.

Norton, Edward C., Hua Wang, and Chunrong Ai. 2004. "Computing Interaction Effects and Standard Errors in Logit and Probit Models." *Stata Journal* 4(2): 154-167.

Puhani, Patrick A. 2012. "The Treatment Effect, the Cross Difference, and the Interaction Term in Nonlinear 'Difference-in-differences' Models." *Economics Letters* 115: 85-87.

Radean, Marius. 2023. "Ginteff: A Generalized Command for Computing Interaction Effects." *Stata Journal* 23(2): 301-335.

Schenker, Nathaniel and Jane F Gentleman. 2001. "On Judging the Significance of Differences by Examining the Overlap between Confidence Intervals." *American Statistician* 55(3): 182-186.

Uberti, Luca J. 2022. "Interpreting logit models." *Stata Journal* 22(1): 60-76.

<접수 2025.10.14; 수정 2025.11.09; 게재확정 2025.11.24>

On Derivations and Probability Interpretations of Interaction Effects in Logit and Probit Regression Models

Hyun Woo Kim
(Chungbuk National University)

This article presents proper derivation and interpretation methods for interaction effects as marginal effects in nonlinear models including logit and probit regression models. Unlike linear models, interaction terms in nonlinear models do not directly correspond to interaction effects, and the magnitude and direction of interaction marginal effects can vary depending on the specific values of covariates. I demonstrate that interaction marginal effects in nonlinear models should be computed as cross-partial derivatives, consisting not only of interaction terms but also main term coefficients and derivatives of probability density functions. Simulation results under various scenarios reveal that substantial proportions of regions show interaction term coefficients show opposite signs to actual interaction marginal effects, and such marginal effects can emerge even when interaction terms are not statistically significant. To address these issues, five practical alternatives were proposed to be considered. This research contributes to improving the accuracy of interaction effect interpretation in social science research using nonlinear models, providing important methodological implications for the broader sociological research community.

Key words: logit, probit, nonlinear models, interaction effect, interaction term, marginal effect, categorical data analysis