범주형 자료분석

Cautionary Notes about Nonlinear Models

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

April 14, 2025

진행 순서

1. 로짓 회귀모형에서 상호작용항
2. 로짓 회귀모형에서 누락변수 편의

로짓 회귀모형에서 상호작용항

로짓/프로빗 회귀모형에서 상호작용항

비선형 모형에서 상호작용항의 계수는 상호작용 효과와 다르다.

- 상호작용효과(interaction effect)는 교차편미분(cross-partial derivative) 또는 교차 차이(cross-difference)로 정의된다.
- Stata 같은 컴퓨터 프로그램에서 보고하는 상호작용항의 계수는 상호작용 효과를 제대로 보여주지 않는다.
- 또한 이는 상호작용 효과와 일반적으로 같지 않다.

$$\frac{\partial^2 F(X\beta)}{\partial x_1 \partial x_2} \neq \frac{\partial F(X\beta)}{\partial (x_1 x_2)}$$

로짓/프로빗 회귀모형에서 상호작용항

(선형모형과 달리) 비선형모형에서 상호작용효과는 관측치마다 다르다.

- 프로빗 회귀모형이라면 $F(X\beta) = \Phi(x\beta)$이고 상호작용 효과는 다음과 같다(Why?).

$$\frac{\partial^2 \Phi(x\beta)}{\partial x_1 \partial x_2} = (\beta_1 + \beta_{12}x_2)(\beta_2 + \beta_{12}x_1)\phi'(x\beta) + \beta_{12}\phi(x\beta)$$

- 컴퓨터 프로그램에서 보고된 상호작용의 계수는 당연히 다음과 같다.

$$\frac{\partial \Phi(X\beta)}{\partial(x_1 x_2)} = \beta_{12}$$

- 결정적인 차이에 주목해야 한다. 그것은 $\phi(x\beta)$와 $\phi'(x\beta)$가 공변량 x의 값에 따라 달라진다는 점이다.
- 즉 동일한 β 값을 가지더라도 관측치마다 상호작용효과의 크기와 방향이 달라진다 (Why?).

로짓/프로빗 회귀모형에서 상호작용항

심지어 β_{12}의 부호와 실제 상호작용 효과의 부호가 반대일 수도 있다.

- 가령 $\beta_{12} < 0$이지만, $\frac{\partial^2 \Phi(x\beta)}{\partial x_1 \partial x_2} > 0$일 수도 있다(Why?).
- 상호작용 효과는 β_1, β_2 및 x의 값, $\phi(x\beta)$, $\phi'(x\beta)$ 등에 모두 영향을 받기 때문이다.
- 물론 β_{12}가 음수라도 “특정 조건 하에서” 상호작용은 양(+)의 방향일 수도 있다 (Why?).
- 그러므로 β_{12}의 부호로 상호작용의 방향마저 단순히 해석하는 것은 위험하다.

로짓/프로빗 회귀모형에서 상호작용항

통계적 유의성마저 신중하게 해석해야 한다.

- Ai and Norton (2003)의 지적대로, 설령 β_{12}가 유의하더라도 상호작용 효과는 관측치마다 다르므로 일률적인 유의성 판단은 불가능하다(Why?).
- 다시 말해, $\beta_{12} = 0$이더라도 상호작용 효과는 0이 아닐 수 있다.
- 통계적 유의성은 단순한 t-검정이나 Wald 검정이 아닌, 델타 방법(Delta method을 통해 계산해야 한다.

로짓/프로빗 회귀모형에서 상호작용항

비선형 모형에서는 상호작용항이 없어도 상호작용 효과가 나타날 수 있다.

- 가령 프로빗 회귀모형에서 상호작용항 없이 추정해보자.

$$E(Y|x_1, x_2) = \Phi(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$$

- $A = \beta_0 + \beta_1 x_1 + \beta_2 x_2$라고 하자. 여기에는 $x_1 x_2$ 항이 없지만,

$$\frac{\partial^2 E(Y|x_1, x_2)}{\partial x_1 \partial x_2} = \beta_1 \beta_2 \cdot \Phi''(A) = -\beta_1 \beta_2 A \cdot \phi(A) \neq 0$$

- 이 효과는 모형의 비선형 구조 때문에 생긴 것일 뿐, 변수 간의 실제 상호작용 때문은 아니다. 즉 상호작용 효과는 때로 모형의 함수 형태의 산물일 수도 있다.

로짓/프로빗 회귀모형에서 상호작용항

그래서 어떻게 하면 좋을까?

- 좀 더 단순하고 기본적인 대안을 먼저 생각해보자.
- (1) 선형확률모형(linear probability model; LPM)은 생각보다 나쁘지 않으니 거기에서 추정된 상호작용 효과와 비교한다.
- (2) 2×2 교차표 등 간략한 표를 만들고 오즈비(OR)를 계산하여 상호작용 효과를 파악한다. 가령 Ganzach, Saporta, and Weber (2000)와 Kuha and Mills (2020)를 참고할 것.
- 이 방식들은 결국 로짓/프로빗 회귀모형에서 상호작용항의 직접적인 사용을 포기한다는 점에서 다소 아쉬움을 남긴다.

로짓/프로빗 회귀모형에서 상호작용항

- (3) $\frac{\partial F(X\beta)}{\partial(x_1 x_2)}$가 아니라 $\frac{\partial^2 F(X\beta)}{\partial x_1 \partial x_2}$을 똑바로 계산한다.
- Karaca-Mandic, Norton, and Dowd (2012)은 이 방식을 지지했다.
- Stata에서 inteff, margins, predictnl을 사용하면 $\frac{\partial^2 F(X\beta)}{\partial x_1 \partial x_2}$를 옳게 계산할 수 있다.
- 여기까지는 큰 의미가 있다. 그러나 $\frac{\partial^2 F(X\beta)}{\partial x_1 \partial x_2}$ 자체를 읽는다고 해도 어차피 의미가 와닿지 않는다. 원래 상호작용 효과는 항상 시각화로 해석했었다!
- Ai and Norton (2003)과 Norton, Wang, Ai (2004)은 독특한 방식의 시각화를 제안했다. Greene (2010)도 지적했지만, 그 시각화 결과물은 도무지 전달이 어렵다. 이 시각화 방법은 추천할 수 없다.

로짓/프로빗 회귀모형에서 상호작용항

- (4) 예측확률 곡선(predicted probability curve)을 여러 개 그리는 방식으로 상호작용 효과를 설명할 수 있다.
- Allison (mimeo)와 Greene (2010)은 이 방식을 지지했다.
- 특히 Greene (2010: 295)이 성별과 나이의 상호작용이 어떻게 나타냈는지를 확인해보자(Fig 4). 이때 델타 방법으로 추정된 신뢰구간(confidence interval)을 음영으로 표현하였음에 주의하자.
- 모형 추정 단계에서 상호작용항을 넣고 추정한다(Why?). 그것이 통계적으로 유의하든 유의하지 않든, 계수 부호가 (+)이든 (-)이든 당장 해석하지 않는다(Why?). 그림을 그리고 “구간에 따라” 해석한다(Why?).

로짓/프로빗 회귀모형에서 상호작용항

- (5) $\dfrac{\partial F(X\beta)}{\partial x_1}$를 y축으로 하고, x_2를 x축으로 하여 시각화한다.
- Uberti (2022)는 이 방식을 지지했다.
- 이 방식이 (3)과 결합하면 가장 수학적으로 명확하다. 즉 유의성 검정은 (3)의 방식대로 하고, 실제 시각화는 이 방식대로 하는 셈이다.
- 특히 Uberti (2022: 70–75)는 (좀 더 세밀하게) 다른 통제변수는 물론이고 아내 소득을 제외한 가구소득(nwifeinc)에까지 특정 값을 부여하고, 그 근처에서 $P(Y=1|X)$의 변화량을 시각화하여, 여러가지 시나리오를 표현하고 있다.

로짓 회귀모형에서 누락변수 편의

로짓 회귀모형에서 누락변수 편의

일반적으로 회귀계수는 누락변수 편의에 취약하다.

- 선형 회귀모형에서 x_2의 누락으로 인한 x_1의 편의(bias) 정도는 x_1과 x_2가 상관되어 (correlated) 있는가 여부에 달려 있다.
- 지금까지 비선형 회귀모형도 비슷할 것으로 추측되었다.
- 그러나 Mood (2010)에 따르면 로짓 회귀모형에서는 x_1과 x_2가 상관성이 있든 없든 편의가 발생한다.
- 즉 로짓 회귀모형은 누락변수 편의(omitted variable bias), 즉 관찰되지 않은 이질성(unobserved heterogeneity)에 아무래도 더 취약하다.

로짓 회귀모형에서 누락변수 편의

잠재변수 모형에 관해 먼저 이해할 필요가 있다.

- 잠재변수(latent variable) y^*를 종속변수로 한 프로빗(probit) 모형을 다음과 같이 생각해보자.

$$Y^* = X\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1)$$

- 이때 y^*는 실제로 관측되지 않는 가상의 연속 변수이고(Why?), 실제 관찰된 (observed) y와 다음과 같은 관계를 갖는다.

$$Y = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

- 위의 이론적 모형은 아래처럼 바뀌어 추정된다. 이때 $\mathbb{E}(Y|X) \neq Y^*$임에 주의하자.

$$\mathbb{E}(Y|X) = P(Y = 1|X) = P(Y^* > 0) = \Phi(X\beta)$$

로짓 회귀모형에서 누락변수 편의

- 잠재변수(latent variable) y^*를 종속변수로 한 비선형 모형을 다음과 같이 생각해보자.

$$y^* = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- 이때 로짓 회귀모형은 $\mathrm{Var}(\epsilon) = \frac{\pi^2}{3} = 3.29$이라고 가정된다(증명 생략).
- 오차 분산이 자유롭다고 잠시 가정하고, 왜곡없는(unbiased) 회귀계수 β^*를 구할 수 있다(Why?).

$$y^* = \alpha + \beta_1^* x_1 + \beta_2^* x_2 + \sigma\epsilon$$

- 이때 σ는 (가정된 오차 분산을) 진정한 오차 분산으로 조정해주는 역할을 수행하는 스칼라(scalar)일 뿐, 실제 오차에 곱해지는 것은 아니다.
- 그러나 로짓 회귀모형을 사용하는 이상 우리는 하는 수 없이 가정을 따라야만 한다. 그 결과 우리는 왜곡된 회귀계수와 왜곡없는 오차 ϵ^*를 얻는다.

$$\frac{y^*}{\sigma} = \frac{\alpha}{\sigma} + \frac{\beta_1}{\sigma} x_1 + \frac{\beta_2}{\sigma} x_2 + \epsilon^*$$

로짓 회귀모형에서 누락변수 편의

- 만일 $x_2 = \gamma_0 + \gamma_1 x_1 + \nu$이라면, x_2가 위 모형에서 누락될 때, 이렇게 다시 정리할 수 있다(Mood 2010: 69).

$$y^* = (\alpha + \beta_2\gamma_0) + (\beta_1 + \beta_2\gamma_1)x_1 + (\beta_2\nu + \epsilon)$$

- 만일 오차 분산에 대한 가정이 없었다면, 진정한(true) 오차 분산은 $\beta_2^2 \mathrm{Var}(\nu) + \mathrm{Var}(\epsilon)$이다.

$$\frac{y^*}{\sigma} = \frac{\alpha}{\sigma} + \frac{\beta_1}{\sigma}x_1 + \epsilon^*$$

- 그러나 로짓 회귀모형대로 추정한다면 가정된(assumed) 오차 분산은 무조건 $\mathrm{Var}(\epsilon^*) = 3.29$이다.

로짓 회귀모형에서 누락변수 편의

- 이를 아까처럼 $\sigma\epsilon$으로 표현한다면, σ는 아래처럼 나타낼 수 있다(Why?).

$$\sigma = \sqrt{\frac{\text{True residual variance}}{\text{Assumed residual variance}}} = \sqrt{\frac{\beta_2^2 \mathrm{Var}(\nu) + 3.29}{3.29}}$$

- 즉 아래 회귀모형은 (가정을 무시하고 추정해낸) 진정한 β_1을 표현하고 있다.

$$y^* = \alpha + \beta_1^* x_1 + \sigma\epsilon$$

- 물론 로짓 회귀모형에서는 $\sigma\epsilon$를 허락해주지 않으므로, 결국 식별된 것은 β_1이 아니라 $\frac{\beta_1}{\sigma}$이다.
- 우리는 이제 x_2가 누락되면, (1) 오차 분산 Var(ϵ)이 커지고, (2) β_1는 σ가 커지면 하향 편의(downwardly biased)됨을 확인하였다(Why?).

로짓 회귀모형에서 누락변수 편의

로짓 회귀계수는 효과 크기 뿐 아니라 이질성에도 영향을 받는다.

- 먼저 x_1과 x_2가 상관된(correlated) 경우에 $\frac{\beta_1}{\sigma}$는 다음과 같다.

$$\frac{\beta_1}{\sigma} = (\beta_1 + \beta_2\gamma_1) \cdot \frac{\sqrt{3.29}}{\sqrt{3.29 + \beta_2^2 \cdot \mathrm{Var}(\nu)}}$$

- 다음으로 x_1과 x_2가 상관없는(uncorrelated) 경우에 $\frac{\beta_1}{\sigma}$는 다음과 같다.

$$\frac{\beta_1}{\sigma} = \beta_1 \cdot \frac{\sqrt{3.29}}{\sqrt{3.29 + \beta_2^2 \cdot \mathrm{Var}(\nu)}}$$

- 우리는 이제 (상관일 때는 물론이고) 비상관일 때조차도 오차 분산이 커짐에 따라 계수가 작아짐을 확인하였다(Why?). 그러므로 선형 회귀모형보다 이중으로 취약함을 알 수 있다.

로짓 회귀모형에서 누락변수 편의

우리는 무엇을 배웠을까?

- 불완전하게 설정된 모형이라면, 즉 $\epsilon = 3.29$를 완벽하게 맞추지 못했다면, $\sigma \neq 1$이 되므로, 실제 식별한 회귀계수는 β_1/σ가 되어 왜곡된다.
- 좀 더 구체적으로, 설명 안된 변동성(unobserved heterogeneity)인 오차 ν가 클수록 똑같은 β라도 로짓 회귀계수는 작아진다.
- 이 모든 문제는 로짓/프로빗 회귀모형에서 (선형회귀모형과는 달리) 오차 분산을 고정하기 때문에 생기는 특수한 현상이다.

로짓 회귀모형에서 누락변수 편의

- 그러므로 표본 간(across samples), 표본내 그룹간(across groups within samples), 혹은 시간 경과에 따른(over time) 로짓 회귀계수 크기 비교는 곤란하다 (Mood 2010: 73).
- 우리에게 A그룹과 B그룹이 있다고 할 때, $\mathrm{Var}(\epsilon) = 3.29$이므로 다음과 같이 모형을 쓸 수 있다.

$$y^* = \alpha^A + \beta^A x_1 + \sigma^A \epsilon$$
$$y^* = \alpha^B + \beta^B x_1 + \sigma^B \epsilon$$

- 이때 우리는 β^A와 β^B를 직접 비교하지 못하고, $\frac{\beta^A}{\sigma^B}$와 $\frac{\beta^B}{\sigma^B}$를 비교하게 된다.
- 비교를 꼭 하려면 설명 안된 변동성이 모두 같다는 가정($\sigma^A = \sigma^B$)이 필요하다. 물론 이는 검증 가능하지 않다.

로짓 회귀모형에서 누락변수 편의

그래서 어떻게 하면 좋을까?

- Kuha and Mills (2020)에 따르면, (1) 연구자의 관심이 실제 관측된(observed) Y에 있을 때, 비선형 회귀모형에서 계수 비교는 원칙적으로 가능하다. 그리고 대부분의 사회과학 연구는 잠재변수 Y^*에 이론적 관심이 없고, 단지 관측된 Y에 관심이 있을 뿐이므로, 로짓 계수의 비교 가능성에 대한 우려는 실제 사회과학 연구에서 중요하지 않다.
- 이들은 또한 (2) 관찰되지 않은 이질성(unobserved heterogeneity)은 본질적으로 존재하며, 계수 비교가 발생하는 현실 그 자체를 반영한다고 본다. 즉 서로 다른 그룹이 이질성(heterogeneity)을 가진다는 건 회귀모형의 오류가 아니라, 그룹 간 효과의 정의가 원래 그렇다는 것이다. 따라서 계수 비교 자체가 잘못된 것은 아니라고 주장한다.

로짓 회귀모형에서 누락변수 편의

- 나는 동의하기 어렵다.
- Kuha and Mills (2020)가 상정한 이변량(bivariate) 처리효과(treatment effect)를 보는 그런 단순한 상황이면 모를까, 사회과학 연구에서 다변량(multivariate) 모형을 구축하다 보면 Y^*에 의존하지 않을 수 없다(Why?).
- 다변량 로짓 모형에서 각 독립변수의 계수는 공변량이 일정할 때의 조건부 변화율, 즉 기울기이고, 이는 잠재연속변수 Y^*를 전제로 한 해석이다.
- 게다가 이질성이 불가피한 현실 그 자체라는 설명은 다소 납득하기 어려운 현실도피처럼 보인다.

로짓 회귀모형에서 누락변수 편의

- (1) 매개 효과(mediation effect)를 최대한 설명에서 배제한다. 반드시 매개 효과를 써야할 상황이면 연속변수를 종속변수로 쓰도록 노력한다.
- (2) 로짓/프로빗 말고 선형확률모형(LPM)을 통해 회귀계수를 비교한다. 가령 정인관(2022)을 참고할 것.
- 선형확률모형에서 그냥 비교할 수도 있고, (정 마음에 걸리면) 로짓/프로빗에서 비교한 결과가 크게 다르지 않음을 보이면 된다.
- (3) 가정의 위배 가능성을 분명히 고지한다. 즉 "관찰되지 않은 이질성이 모든 그룹에 걸쳐 균일하게 유지된 것"으로 가정하였음을 밝힌다(Why?). 그 위험을 고지한 상태에서 독자에게 이 결과를 믿을 것인지 맡긴다.

로짓 회귀모형에서 누락변수 편의

- (4) 이질선택모형(Heterogeneous choice model)이나 KHB 모형 (Karlson–Holm–Breen Model) 등 좀 더 고급 모형을 사용한다(Allison 1999; Karlson, Holt, and Breen 2011; Williams 2009, 2010).
- 아쉽게도 이것은 훨씬 어렵다. 비선형 모형에서 식별성(identification) 문제를 깊이 있게 이해하고 있어야 한다.
- 다행인지 불행인지 사회과학 분야에서 이것을 사용한 실증연구는 매우 적다. 나중에 기회가 닿는다면 이 내용을 다루어보자.