범주형 자료분석
Models for Nominal Outcomes

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

May 19, 2025

진행 순서

1. 순서형 로짓 회귀분석
2. 회귀계수의 해석
3. 유의성 검정과 모형 적합도
4. 순서형 로짓 회귀모형의 대안
5. 비례 오즈 가정

순서형 로짓 회귀분석

순서형 로짓 회귀분석

종속변수가 순서형 척도라면 순서형 로짓 회귀모형 등을 활용할 수 있다.

- 종속변수가 설령 연속변수(continuous)가 아니더라도 명확한 순위(rank)만 있으면 서수적(ordinal)이고, 순서형(ordered or ordinal) 로짓 회귀모형을 사용할 수 있다.
- (고졸, 대졸 등) 교육수준(educational attainment)는 순서형 척도이고, 교육연수와는 달리 등간척도는 아니다. 범주형 소득 변수에 대해서도 생각해보자.
- 리커트 척도는 오늘날 사회조사에서 가장 흔하게 사용되는 척도일 것이다.

> ○○님은 현재 남한에 들어와 있는 탈북민을
> 대한민국 시민이라 생각하십니까?
>
> 1. 매우 그렇게 생각한다
> 2. 어느 정도 그렇게 생각한다
> 3. 별로 그렇게 생각하지 않는다
> 4. 전혀 그렇게 생각하지 않는다

순서형 로짓 회귀분석

다른 비선형 모형처럼 순서형 로짓 모형도 잠재변수 설정에서 출발한다.

- 관측변수(observed variable)와는 달리, 잠재변수(latent variable)란 (말 그대로) 보이지 않는 변수이다.
- 비선형 모형에서는 실제 관찰되는 관측변수 Y_i 그 자체를 모형화하는 것이 아니라, 잠재변수 Y_i^* 를 통해 그 잠재구조(latent structure)를 상수(constant) 없이 모형화한다.

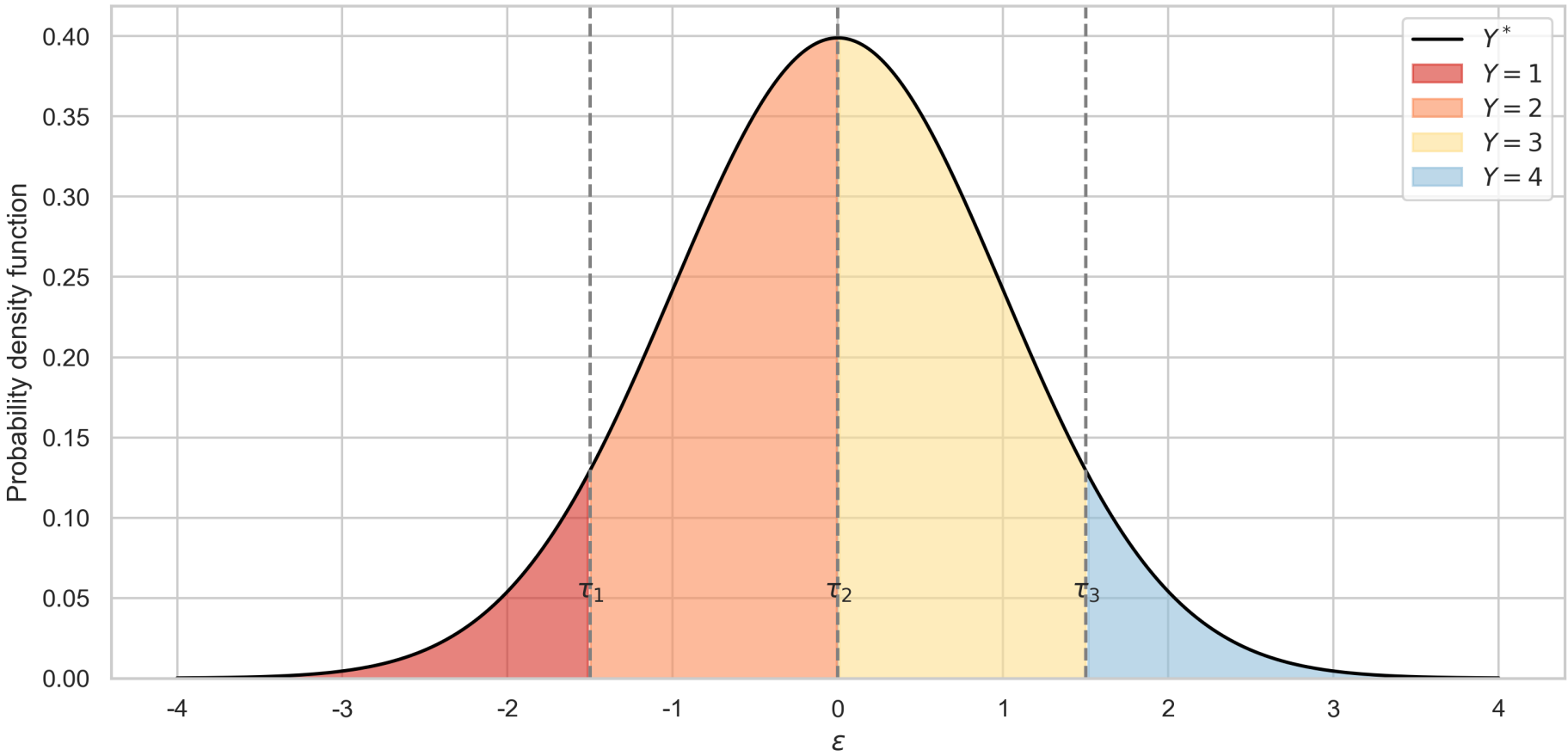
$$Y_i^* = \beta_1 X_i + \epsilon_i$$

- 특정 잠재변수 Y_i^* 가 분할점(cutpoint) τ_k 를 넘었는가에 따라, 실제 관찰되는 관측변수인 Y_i 의 $\{1, 2, 3, 4\}$에 대응시킨다.

$$Y_i = \begin{cases} 1 & (\text{if } -\infty < Y^* \leq \tau_1) \\ 2 & (\text{if } \tau_1 < Y^* \leq \tau_2) \\ 3 & (\text{if } \tau_2 < Y^* \leq \tau_3) \\ 4 & (\text{if } \tau_4 < Y^* \leq \infty) \end{cases}$$

순서형 로짓 회귀분석

- 잠재변수 Y^*와 관측변수 Y의 관계를 도식화해보자. 잠재변수 Y^*의 패턴은 부드러운 로지스틱 분포 곡선으로 표현된다.



순서형 로짓 회귀분석

순서형 로짓 확률모형을 구성해보자.

- 관측변수 Y의 모든 관찰값(1, 2, 3, 4)에 대해서 확률 구간을 설정할 수 있다.

$$
\begin{aligned}
P(Y=1) &= P(-\infty < Y^* \leq \tau_1) \\
P(Y=2) &= P(\tau_1 < Y^* \leq \tau_2) \\
P(Y=3) &= P(\tau_2 < Y^* \leq \tau_3) \\
P(Y=4) &= P(\tau_4 < Y^* \leq \infty)
\end{aligned}
$$

- 잠재변수 Y^*의 오차항 ϵ은 로지스틱 분포를 따른다고 가정된다. 일단 β가 추정된 뒤에는 자료의 X와 결합하여 $Y^* = \beta_1 X + \epsilon$ 전체가 로지스틱 분포를 따른다고 가정되는 셈이다.

순서형 로짓 회귀분석

- 만일 Y^*이 분할점 τ_1보다 작다면, $Y = 3$에 대응하는 관계를 상상해보자.

$$\begin{aligned} P(Y = 3) &= P(\tau_2 < Y^* \leq \tau_3) \\ &= P(\tau_2 - \beta_1 X < \epsilon \leq \tau_3 - \beta_1 X) \\ &= \int_{\tau_2 - \beta_1 X}^{\tau_3 - \beta_1 X} \lambda(\epsilon) d\epsilon \\ &= \Lambda(\tau_3 - \beta_1 X) - \Lambda(\tau_2 - \beta_1 X) \\ &= \frac{1}{1 + e^{-(\tau_3 - \beta_1 X)}} - \frac{1}{1 + e^{-(\tau_2 - \beta_1 X)}} \end{aligned}$$

- 이와 비슷하게 $P(Y = 1)$, $P(Y = 2)$, $P(Y = 4)$를 연습해보자.
- 순서형 로짓 회귀모형에는 (β_0가 따로 없고) 분할점 τ가 여러 개 있고 상수 역할을 한다.

순서형 로짓 회귀분석

- 아래처럼 살짝 다른 방식으로 표현할 수도 있다(Why?).

$$\begin{aligned} P(Y_i \leq j) &= P(Y^* < \tau_j) \\ &= P(\epsilon < \tau_j - \beta_1 X) \\ &= \int_{-\infty}^{\tau_j - \beta_1 X} \lambda(\epsilon) d\epsilon \\ &= \Lambda(\tau_j - \beta_1 X) \\ &= \frac{1}{1 + e^{-(\tau_j - \beta_1 X)}} \end{aligned}$$

- 물론 아래 또한 성립한다.

$$\begin{aligned} P(Y_i > j) &= 1 - \Lambda(\tau_j - \beta_1 X) \\ &= \frac{e^{-(\tau_j - \beta_1 X)}}{1 + e^{-(\tau_j - \beta_1 X)}} \end{aligned}$$

순서형 로짓 회귀분석

- 두 확률의 비율은 신기한 함의를 갖고 있다.

$$\begin{aligned}\frac{P(Y \le j)}{P(Y > j)} &= \frac{1}{1 + e^{-(\tau_j - \beta_1 X)}} \cdot \frac{1 + e^{-(\tau_j - \beta_1 X)}}{e^{-(\tau_j - \beta_1 X)}} \\ &= \frac{1}{e^{-(\tau_j - \beta_1 X)}} \\ &= e^{(\tau_j - \beta_1 X)}\end{aligned}$$

- 이제 양변에 자연대수를 취하면,

$$\ln \frac{P(Y \le j)}{P(Y > j)} = \tau_j - \beta_1 X$$

- 이것은 "순서 정보를 반영한" 로그 오즈를 의미한다. 가끔씩 누적 오즈(cumulative odds)라고도 불린다. 이것을 이항 로짓 모형과 비교해보자.

순서형 로짓 회귀분석

- 위의 리커트 4점 척도 문제를 생각해보면, 분할점 τ_j는 각각 다음과 같이 추정한다.

$$\ln \frac{P(Y \leq 1)}{P(Y > 1)} = \ln \frac{P(Y = 1)}{P(Y = 2) + P(Y = 3) + P(Y = 4)} = \tau_1 - \beta_1 X$$
$$\ln \frac{P(Y \leq 2)}{P(Y > 2)} = \ln \frac{P(Y = 1) + P(Y = 2)}{P(Y = 3) + P(Y = 4)} = \tau_2 - \beta_1 X$$
$$\ln \frac{P(Y \leq 3)}{P(Y > 3)} = \ln \frac{P(Y = 1) + P(Y = 2) + P(Y = 3)}{P(Y = 4)} = \tau_3 - \beta_1 X$$

- 추정되는 식은 3개 뿐이고, $\ln \frac{P(Y \leq 4)}{P(Y > 4)}$는 당연히 정의될 수 없다(Why?).
- 세 방정식에서 τ만 다를 뿐, β_1은 동일하게 값으로 추정된다!.

순서형 로짓 회귀분석

순서형 로짓 회귀분석에서도 최대우도 추정량을 사용한다.

- 예전에는 표본, 응답 범주, 독립변수 세 조합의 크기에 따라 계산 비용이 급격하게 높아진다는 것이 단점으로 꼽혔다(Winship and Mare 1984: 515). 지금은 아무도 신경쓰지 않는다.

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{N} \prod_{j=1}^{J} P(Y_i = j)^{I(Y_i = j)} \\
&= \prod_{i=1}^{N} \prod_{j=1}^{J} P(\tau_{j-1} < Y_i^* \leq \tau_j)^{I(Y_i = j)} \\
&= \prod_{i=1}^{N} \prod_{j=1}^{J} [\Lambda(\tau_j - X\beta) - \Lambda(\tau_{j-1} - X\beta)]^{I(Y_i = j)}
\end{aligned}
$$

- 로그 우도함수는 다음과 같다.

$$
\ln \mathcal{L} = \sum_{i=1}^{N} \sum_{j=1}^{J} I(Y_i = j) \ln P(Y_i = j)
$$

회귀계수의 해석

회귀계수의 해석

순서형 로짓 모형에서 회귀계수의 해석이 약간 혼란스러울 수 있다.

- nhanes2f.dta에서 주관적 건강상태(health)를 살펴보면 “poor (1)”에서 “excellent (5)”까지 리커트 5점 척도로 측정되어 있다.
- 이 자료를 순서형 로짓 회귀모형에 적합시켜, 회귀계수 β와 분할점 τ를 추정해보자.
- 앞서 우리가 유도한 누적오즈는 사실 일반적인 리커트 척도에서 할당된 점수와 반대이므로, 한 번 뒤집어 해석해야 한다(Why?).
- 종종 통계분석 소프트웨어는 (우리의 편의를 위해) 마치 아래처럼 알아서 뒤집어준다!

$$\frac{P(Y \geq j)}{P(Y < j)} = \beta_1 X + \tau_j$$

회귀계수의 해석

- 아래 순서형 로짓 모형을 잘 살펴보면서 몇 가지 해석법을 고민해보자.

$$\ln \frac{P(Y \leq j)}{P(Y > j)} = \tau_j - \beta_1 X$$

- 첫째, 로그 (누적)오즈를 그대로 해석할 수 있다.

  "(다른 변수의 영향력을 통제할 때) 나이(X)가 한 살 증가하면, 주관적 건강상태 악화의 로그 (누적)오즈가 0.041 만큼 증가한다."

- 그러나 통계분석 소프트웨어는 종종 이를 뒤집어 보여주므로,

  "(다른 변수의 영향력을 통제할 때) 나이(X)가 한 살 증가하면, 주관적 건강상태 증진의 로그 (누적)오즈가 0.041 만큼 감소한다."

회귀계수의 해석

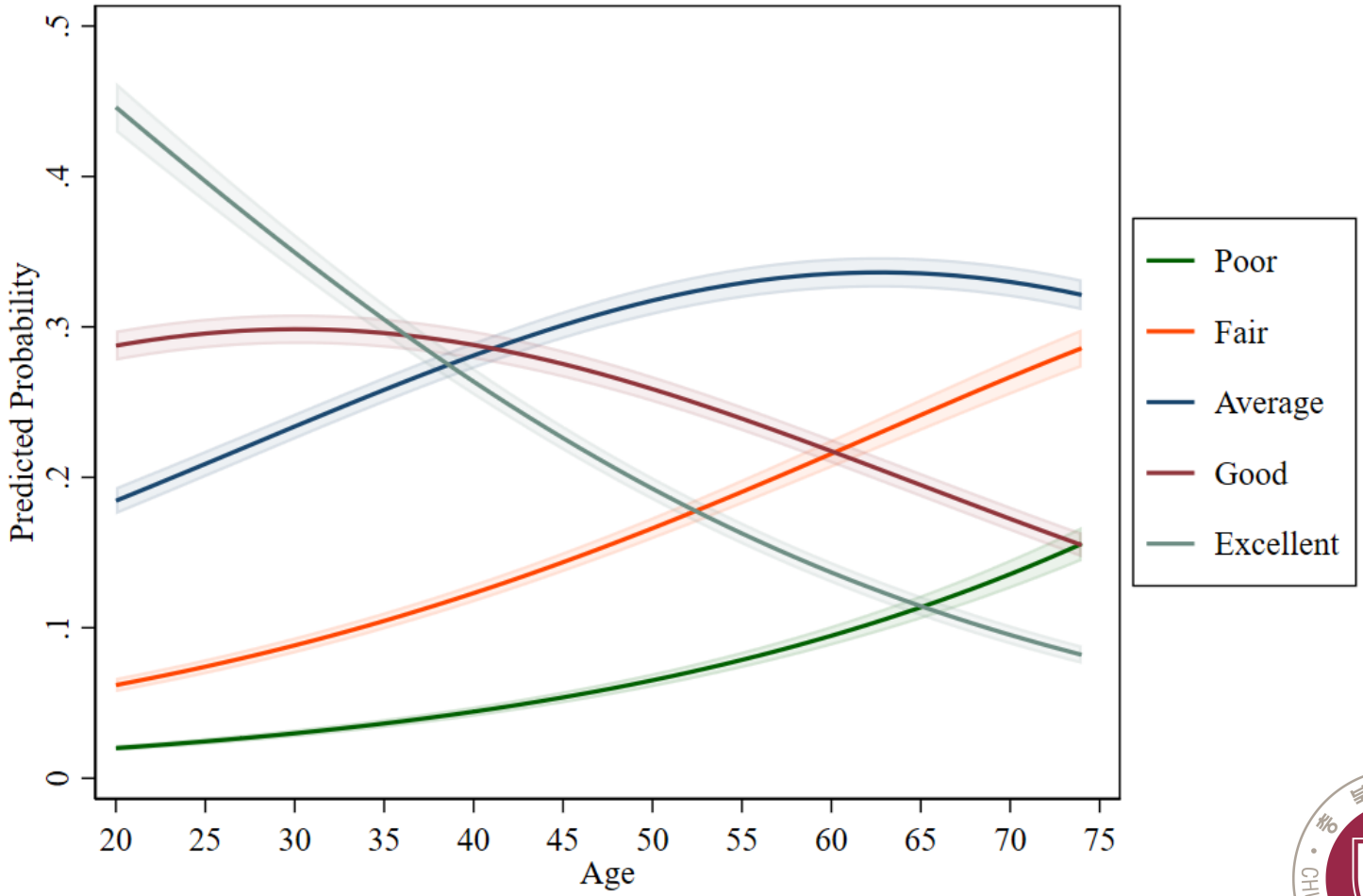
- 둘째, 이항 로짓 모형에서의 회귀계수처럼 (누적)오즈비 $e^{-\beta_1}$를 해석할 수 있다.

$$
\begin{aligned}
OR_{j=1} &= \frac{odds_{j=1|X+1}}{odds_{j=1|X}} \\
&= \frac{\frac{P(Y \leq j|X+1)}{P(Y > j|X+1)}}{\frac{P(Y \leq j|X)}{P(Y > j|X)}} \\
&= \frac{e^{\tau_1 - \beta_1(X+1)}}{e^{\tau_1 - \beta_1 X}} \\
&= e^{-\beta_1}
\end{aligned}
$$

- "(다른 변수의 영향력을 통제할 때) 나이를 한 살 더 먹을수록 주관적 건강상태 증진의 오즈가 4% 감소한다."

회귀계수의 해석

- 셋째, 예측 확률로 시각화할 수는 있지만 그다지 사용되지는 않는 것 같다.



회귀계수의 해석

- 넷째, 한계효과(marginal effect)로 해석할 수 있다.
- 이때 한계효과는 독립변수가 1 단위 증가할 때, 종속변수의 "범주별"로 확률이 얼마나 변하는지를 나타낸다. 다시 말해, 독립변수가 증가할 때, j번째 범주에 속할 확률의 변화량을 의미한다.
- 물론 순서형 회귀모형에서는 여러 범주(1=매우 아프다; $\cdots$; 5=매우 건강하다)가 존재하므로 각 범주별로 한계효과를 따로 계산해야 한다.
- 이런 측면에서 편의성이 다소 떨어진다.
- 아래와 같은 한계효과는 "소득이 10만 원 단위 증가하면 매우 건강하다(5)고 응답할 확률이 5% 증가한다"고 해석할 수 있다.

$$\frac{\partial \Pr(Y = 5)}{\partial \text{Income}} = 0.05$$

회귀계수의 해석

- 이항 로짓 회귀모형에서 한계효과는 다음과 같다. 이때 Λ와 λ는 각각 로지스틱 누적분포함수와 확률밀도함수를 뜻한다.

$$\frac{\partial \Pr(Y=1)}{\partial x_k} = \frac{\partial \Lambda(X\beta)}{\partial X\beta}\frac{\partial X\beta}{\partial x_k} = \lambda(X\beta)\cdot\beta_k$$

- 한편, 순서형 로짓 회귀모형에서 j번째 범주의 확률은 아래와 같은 누적분포함수로 표현된다.

$$P(Y=j) = \Lambda(\tau_j - X\beta) - \Lambda(\tau_{j-1} - X\beta)$$

- 그러므로 k번째 독립변수의 한계효과는 아래와 같다.

$$\begin{aligned}\frac{\partial P(Y=j)}{\partial X_k} &= \frac{\partial \Lambda(\tau_j - X\beta)}{\partial(\tau_j - X\beta)}\frac{\partial(\tau_j - X\beta)}{\partial X_k} - \frac{\partial \Lambda(\tau_{j-1} - X\beta)}{\partial(\tau_{j-1} - X\beta)}\frac{\partial(\tau_{j-1} - X\beta)}{\partial X_k} \\ &= \lambda(\tau_j - X\beta)\beta_k + \lambda(\tau_{j-1} - X\beta)\beta_k \\ &= \beta_k \cdot [\lambda(\tau_{j-1} - X\beta) - \lambda(\tau_j - X\beta)]\end{aligned}$$

회귀계수의 해석

- 평균한계효과(AME)는 한계효과를 각 관측치에 대해 계산한 후 평균을 취한 값이다.

$$AME = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial P(Y_i = j|X_i)}{\partial X_{ik}}$$

- 평균한계효과가 평균에서의 조건부 한계효과(conditional marginal effect at the means)보다 낫다.

$$\text{CME} = \left.\frac{\partial P(Y = j|X)}{\partial X_k}\right|_{X=\bar{X}}$$

회귀계수의 해석

모형간 회귀계수를 비교할 수 없다.

- 이미 수십 년 전부터 잘 알려졌듯, 모든 비선형 모형의 회귀계수는 모형 간에 비교할 수 없다(Winsihp and Mare 1984: 517).
- (오차항의 분산은 1 또는 $\pi^2/3$으로 가정되지만, 현실에서 이것을 충족시키기 어려우므로) 회귀계수는 스케일 인자(scale factor) σ_ϵ^2로 표준화된 값으로 추정된다.
- 모형마다 스케일 인자의 정확한 값을 알 수 없으므로 각 모형의 회귀계수는 동일한 독립변수라도 비교가 불가능하다.

유의성 검정과 모형 적합도

유의성 검정과 모형 적합도

유의성 검정과 모형적합도 역시 이항 로짓 회귀모형과 똑같다.

- 순서형 로짓 회귀모형에서도 회귀계수의 유의성 검정을 위해 왈드 검정을 사용한다.

$$\chi_1^2 = \left(\frac{\hat{b} - \beta}{\text{SE}_{\hat{b}}}\right)^2 = \left(\frac{\hat{b}}{\text{SE}_{\hat{b}}}\right)^2$$

- 자유도(degree of freedom)가 이항 로짓 회귀모형과 같고, 다항 로짓 회귀모형과는 다르다(Why?).
- χ_1^2 값이 충분히 크면 통계적으로 유의하게 귀무가설($H_0 : \beta = 0$)을 기각할 수 있다.
- Stata에서는 Z 검정을 한다. 표본 크기가 충분히 커지면 왈드 검정과 t 검정, Z 검정은 결국 수렴한다.

$$Z = \frac{\hat{b}}{\text{SE}_{\hat{b}}}$$

유의성 검정과 모형 적합도

- 모형적합도를 살펴보기 위해 (1) 우도비 검정(likelihood-ratio test), (2) 유사 결정계수(pseudo R^2), (3) 정보 기준(information criteria)을 주로 확인한다.
- 첫째, 우도비 검정의 검정통계량 G는 χ^2 분포를 따른다.

$$G = -2 \ln \frac{\mathcal{L}_{\text{null}}}{\mathcal{L}_{\text{model}}} = -2 \ln(\mathcal{L}_{\text{null}} - \mathcal{L}_{\text{model}}) \sim \chi^2_k$$

- 우도비 검정의 귀무가설은 다음과 같다. 이를 기각하지 못하면 "이 모형은 아무 짝에도 쓸모가 없다"라는 의미로 받아들여진다(Why?).

$$H_0 : \mathcal{L}_{\text{null}} = \mathcal{L}_{\text{model}}$$

유의성 검정과 모형 적합도

- 똑같은 원리를 사용하여 두 모형을 비교할 때도 우도비 검정을 사용할 수 있다.

$$G = -2 \ln \frac{\mathcal{L}_{\text{restricted}}}{\mathcal{L}_{\text{full}}} = -2(\ln \mathcal{L}_{\text{restricted}} - \ln \mathcal{L}_{\text{full}}) \sim \chi^2_{\Delta k}$$

- $\mathcal{L}_{\text{restricted}}$는 독립변수가 다소 적게 들어간 모형의 우도 함수값이고, $\mathcal{L}_{\text{full}}$은 그보다 독립변수가 좀 더 들어간 모형의 우도 함수값이다.
- 이때 제한모형(restricted model)은 완전모형(full model) 안에 내포되어(nested) 있어야 한다.
- χ^2로 검정하는 귀무가설은 아래와 같다. 이를 기각하지 못하면 "완전모형이 제한모형보다 나은 구석이 없다"라는 의미로 받아들여진다.

$$H_0 : \mathcal{L}_{\text{restricted}} = \mathcal{L}_{\text{full}}$$

유의성 검정과 모형 적합도

- 둘째, 유사결정계수는 선형회귀모형에서 사용되는 결정계수 R^2를 비선형모형에서 흉내낸 것이다.
- Stata에서는 (1) McFadden's R^2를 사용한다.

$$\text{McFadden's } R^2 = 1 - \frac{\ln \mathcal{L}_{\text{model}}}{\ln \mathcal{L}_{\text{null}}}$$

- 값이 클수록 모형의 설명력이 높다고 말할 수 있지만, 선형회귀모형의 R^2처럼 설명된 분산의 비율로 해석하지 않도록 주의해야 한다.

유의성 검정과 모형 적합도

- 셋째, 아카이케 정보기준(AIC)또는 베이즈 정보기준(BIC)을 보고할 수 있다.
- AIC는 독립변수의 수 k만 보지만, BIC는 표본 크기 n에도 주목한다.

$$AIC = -2\ln \mathcal{L}_{\text{model}} + 2k$$
$$BIC = -2\ln \mathcal{L}_{\text{model}} + k\ln(n)$$

- 당연히 AIC와 BIC 둘 다 그 값이 작을수록 좋다(Why?).
- 정보기준은 우도비 검정과는 달리 내포성(nestedness) 여부를 따지지 않는다. 그러나 유의성 검정 단계가 없으므로 모형의 개선 여부 판단이 약간 애매모호하다.

순서형 로짓 회귀모형의 대안

순서형 로짓 회귀모형의 대안

가정 위배와 무관하게 순서형 프로빗 회귀모형도 존재한다.

- 다행히 순서형 로짓 회귀모형과 추정 과정이 유사하다.
- 우선 Λ와 λ가 Φ와 ϕ로 대체되고, 오차항 ϵ이 로지스틱 분포가 아니라 표준정규분포를 따른다.
- 애초에 순서형 로짓 모형이 성립하지 않으므로, 아래처럼 로그(누적)오즈이나 (누적)오즈비로는 더이상 해석할 수 없다.

$$\ln \frac{P(Y \leq j)}{P(Y > j)} = \ln \frac{\frac{1}{1+e^{-(\tau_j - \beta_1 X)}}}{\frac{e^{-(\tau_j - \beta_1 X)}}{1+e^{-(\tau_j - \beta_1 X)}}} = \tau_j - \beta_1 X$$

- 예측확률 또는 평균한계효과가 훨씬 직관적이다.

순서형 로짓 회귀모형의 대안

다른 유력한 대안은 보통최소제곱 선형 회귀모형이다.

- 심리통계학(Psychometrics)에서 리커트 척도는 종종 대체로 등간(approximately interval) 자료인 것으로 여겨진다. 이를 받아들이면 그냥 보통최소제곱(OLS)에 기반한 선형 모형으로 추정될 수 있다.
- 5점 리커트 척도를 종속변수로 사용해도 정말 괜찮은 것일까?
- 기본적으로 (1) 회귀모형의 가정에 위배되는 점, (2) 상하한에 적절한 제약이 붙지 않는 점에서 문제가 될 수 있다(Why?).
- 그런데 OLS는 워낙 강건하기로 유명하고, Wu and Leung (2017)의 시뮬레이션 연구에 따르면 대체로 심각한 문제가 없다.

순서형 로짓 회귀모형의 대안

- 다른 한편, 순서형 로짓/프로빗도 각 순서형 범주의 분포에 대해 나름의 가정을 한다.
- 오차항이 로지스틱 또는 표준정규분포를 따른다고 가정하기 때문에 극단적인 값이 나타날 확률은 작고, 중간으로 모일 확률은 상대적으로 높다고 여겨진다(Why?).
- 이 가정은 현실적으로 나름 말이 되는 편이다. 또한 자료에 적합한지 통계적으로 검증될 수도 있다(정보기준(IC)이나 -2LL 값이 보다 작을수록 선호된다).
- Winship and Mare (1984: 514-515)에 따르면, 특히 (1) 특정 범주에 관측치가 크게 쏠려있거나, (2) 서로 다른 둘 이상의 집단들이 표본에 섞여있고 서로 다른 관측치 분포의 경향을 가졌을 때(양극화), 순서형 회귀모형이 선형 회귀모형보다 우수하다. 우선 분할점이 통계적으로 유의한지 살펴보자.

비례 오즈 가정

비례 오즈 가정

순서형 로짓 회귀분석은 성립을 장담하기 어려운 가정을 요구한다.

- 순서형 로짓 회귀모형은 다음과 같이 설정된다.

$$\ln \frac{P(Y \leq j)}{P(Y > j)} = \tau_j - X\beta$$

- 이때 이 식에서 β는 모든 j번째 선택에 대해 동일하다고 전제된다.
- 직접 모든 j에 대해 식을 풀어보면, τ만 다르고 β는 똑같다. 즉 범주 간 누적 오즈의 회귀계수(=기울기)는 동일하고, 오로지 분할점만 달라진다.

$$\ln \frac{P(Y \leq 1)}{P(Y > 1)} = \tau_1 - X\beta$$

$$\ln \frac{P(Y \leq 2)}{P(Y > 2)} = \tau_2 - X\beta$$

$$\vdots$$

비례 오즈 가정

- 각 누적 로짓 회귀식의 기울기 β가 동일하므로 로짓 공간에서 회귀선들은 평행한다(Why?).
- 이에 따라 평행회귀 가정(parallel regression assumption) 또는 평행선 가정(parallel line assumption)이 작동하고 있다.
- 로그 (누적) 오즈비 역시 모든 범주에 걸쳐 동일하다(Why?). 다시 말해, 범주 $Y \leq j$에 대한 누적 오즈 비는 X의 변화에 따라 항상 같은 비율로 변한다(Williams 2016: 8).

  "X가 한 단위 증가할 때, 누적 오즈비는 (범주 수준과 상관없이/모든 범주에 걸쳐) $e^{-\beta}$배 만큼 변화한다."
- 오즈비의 변화 크기가 항상 동일하다는 점에서 이를 비례 오즈 가정(proportional odds assumption)라고 부르기도 한다.

비례 오즈 가정

구체적인 사례와 함께 비례 오즈 가정의 의미를 생각해보자.

- 가령 종속변수는 "군대에 대해 신뢰한다(3), 중립이다(2), 불신한다(1)"이고, 독립변수 X는 교육 수준이다. 이 자료에 순서형 로짓 회귀모형을 적합시켜 $\beta = -0.1$를 얻었다고 하자.
- 교육 수준 X의 한 단위 변화가 $\frac{P(Y \leq 1)}{P(Y > 1)}$에 미치는 영향이나 $\frac{P(Y \leq 2)}{P(Y > 2)}$에 미치는 영향이나 똑같다.
- ($e^{-0.1} \approx .9$이므로) 교육 수준이 한 단위 증가할 때, "신뢰 대 중립+불신"의 오즈나 "신뢰+중립 대 불신"의 오즈나 10% 씩 감소하는 점에서 똑같다.
- 순서형 로짓 모형을 사용할 때, 우리는 독립변수의 변화가 모든 범주 구간에 걸쳐 동일한 영향을 주는 것으로 단순화하는 셈이다.

비례 오즈 가정

- 이 가정에 대해서는 브랜트-왈드 검정(Brant-Wald test)으로 확인해 볼 수 있다.
- 각 범주 별로 개별적인 β_j를 허용하는 비제한(unconstraint) 모형을 적합한다.

$$\ln \frac{P(Y \leq 1)}{P(Y > 1)} = \tau_1 - X\beta_1$$

$$\ln \frac{P(Y \leq 2)}{P(Y > 2)} = \tau_2 - X\beta_2$$

$$\vdots$$

- 이렇게 얻은 β_j들을 순서형 로짓 회귀모형의 공통된 β와 비교하는 것이 핵심이다.
- 귀무가설은 "비례오즈 가정이 성립한다"이고, 대립가설은 "적어도 하나 이상의 범주에서 β_j가 다르다"이다.
- 이때 비교 원리는 우도비 검정(likelihood-ratio test)이다.

비례 오즈 가정

일반화 순서형 로짓 회귀모형을 통해 가정을 완화해 볼 수 있다.

- 일반화 순서형 로짓(generalized ordered logit) 모형은 범주별로 β_j가 다른지 먼저 확인해보고, (만일 다르다면) 그것이 범주 별로 가변적일 수 있도록 허용한다.
- 이에 따라 비례 오즈 모형(proportional odds model)과 부분 비례 오즈 모형(partial proportional odds model)을 모두 포괄한다.

$$P(Y \leq j) = \frac{1}{1 + e^{-(\tau_j - X\beta_j)}}$$

- 부분 비례 오즈 모형은 일부 변수에서 β로 모든 범주에 걸쳐 동일하고, 일부 변수에서 범주 j 따라 다른 값을 가질 수 있도록 한다. 가령 X_1과 X_2가 비례 오즈 가정을 충족하고, X_3는 그렇지 않다면 다음과 같은 식을 얻는다.

$$P(Y \leq j) = \frac{1}{1 + e^{-(\tau_j - \beta_1 X_1 - \beta_2 X_2 - \beta_{3j} X_3)}}$$

비례 오즈 가정

- 비례 오즈 가정이 성립하지 않는 특정 변수에 대해서만 다르게 β를 추정할 수 있다는 점이 특징이다.
- 비례 오즈 가정을 테스트하기 위한 목적으로 혹은 강건성 확인(robustness check) 차원에서 수행할 수도 있다.
- 다만 해석이 복잡해지므로 이론적으로 의미 있는 변수에 대해서만 가정을 완화하는 편이 낫다.
- 일반화 순서형 로짓 회귀모형은 아무래도 무엇보다 강한(strong) 이론적 예측을 요구한다(Why?). 특히 Fullerton and Dixon (2010)의 비대칭 효과(asymmetrical effects) 같은 설명이 필요하다.
- 응용사회과학 연구의 관점에서 본다면, 순서형 로짓 회귀분석의 결과를 두고 일반화된 순서형 로짓 모형 그리고 다항 로짓(multinomial logit) 모형 결과와 조심스럽게 비교해야 한다(Why?).