범주형 자료분석
Models for Nominal Outcomes

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

April 28, 2025

진행 순서

1. 다항 로짓 회귀모형
2. 회귀계수의 해석
3. 유의성 검정과 모형 적합도

다항 로짓 회귀모형

다항 로짓 회귀모형

종속변수가 명목형이라면 다항 로짓 회귀모형을 활용할 수 있다.

- 앞서 우리는 종속변수가 가변수일 때, 이른바 이항(binary) 로짓 회귀모형을 사용하였다.
- 종속변수가 명목 척도(nominal scale)를 가진 범주형이라면 다항(multinomial) 로짓 회귀모형을 사용할 수 있다.
- 시간을 들여 흥미로운 범주형 종속변수(e.g., 지지하는 정당, 가입한 보험의 종류, 플로리다 악어가 선호하는 먹이 등)에 관해 상상해보자!
- 이항 로짓을 먼저 철저하게 이해하면 다른 로짓 회귀모형을 쉽게 이해할 수 있다!

다항 로짓 회귀모형

- Florida Game and Fresh Water Fish Commission는 Lake George에서 59 마리의 악어를 잡아 세 종류 중 하나의 먹이를 선택하게 하였다. 악어의 길이가 먹이 선택과 어떤 연관성을 갖는지 살펴보자.
- 플로리다 악어는 크기에 따라 어떤 먹이를 선호하는지 살펴보기 위해, 다음과 같이 상호배타적이고 전체포괄적인 먹이의 범주를 설정하였다.

$$\{\text{어류(fish; } F\text{), 연체류(invertebrates; } I\text{), 기타(other; } O\text{)}\}$$

- 먹이 선택을 잘 생각해보면 범주형 자료이므로 일반적인 선형회귀모형은 부적절할 뿐더러, 선형확률모형(LPM) 같은 대안도 없다(Why?).

다항 로짓 회귀모형

다항 로짓의 확률모형을 간단히 유도해보자.

- 개별 먹이 범주에 대응하는 선택 확률을 다음과 같이 정의할 수 있다.

$$\{P(Y = F), P(Y = I), P(Y = O)\}$$

- 먼저 $P(Y = F)$을 다음과 같이 동어반복으로 나타내보자.

$$P(Y = F) = \frac{P(Y = F)}{1} = \frac{P(Y = F)}{P(Y = F) + P(Y = I) + P(Y = O)}$$

- 범주형 변수를 독립변수로 사용할 때와 마찬가지로, 어느 하나(가령 $Y = O$)를 기준 대안(base outcome)으로 삼아 모형에서 빼고 변환해보자.

$$P(Y = F) = \frac{\frac{P(Y = F)}{P(Y = O)}}{\frac{P(Y = F)}{P(Y = O)} + \frac{P(Y = I)}{P(Y = O)} + 1}$$

다항 로짓 회귀모형

- 오힝 로짓 회귀식과 마찬가지로 다항 로짓 회귀식을 다음과 같이 구성할 수 있다.

$$\ln \frac{P(Y=F)}{P(Y=O)} = \beta_0^F + \beta_1^F X$$

$$\ln \frac{P(Y=I)}{P(Y=O)} = \beta_0^I + \beta_1^I X$$

- 이제 위의 확률모형에 이를 집어넣어 완성한다.

$$P(Y=F) = \frac{e^{\beta_0^F + \beta_1^F X}}{e^{\beta_0^F + \beta_1^F X} + e^{\beta_0^I + \beta_1^I X} + 1}$$

$$= \frac{e^{\beta_0^F + \beta_1^F X}}{1 + \sum e^{\beta_0^I + \beta_1^I X}}$$

- 이 유도 과정은 이해하기 쉬우므로 초보자에게 추천된다.

다항 로짓 회귀모형

사실 좀 더 이론적으로 탄탄한 유도 과정이 있다.

- 오차항 ϵ이 굼벨 분포(Gumbel distribution)를 따르는 임의효용모형(random utility model)으로 j번째 대안의 선택을 생각해보자.

$$U_j = V_j + \varepsilon_j = x\beta^{(j)} + \varepsilon_j$$

- k번째 대안이 아니라 j번째 대안을 선택한다는 것은 다음을 시사한다.

$$\begin{aligned} P(Y = j) &= P(U_j > U_k) \qquad (k \neq j) \\ &= P(\varepsilon_k < \varepsilon_j + V_j - V_k) \end{aligned}$$

- 대수 조작 과정을 거쳐 아래의 일반화된 확률모형이 도출될 수 있다(증명 생략).

$$P(Y = j) = \frac{e^{V_j}}{\sum_{k=1}^{m} e^{V_k}}$$

다항 로짓 회귀모형

- 이 확률모형은 수학적으로 중요한 강점을 갖는다.
- 첫째, k가 아닌 j번째 대안의 선택에 대한 점수 V_j의 값을 V_k 값에 비교하여 부드럽게 확률로 변환해 준다(Why?).
- 첫째, 미분가능하다(differentiable). 여러분은 당연히 한계효과(marginal effect)를 계산할 수 있다(Why?).
- 심지어 딥러닝(deep learning)에서 소프트맥스 함수(softmax function)로도 쓰인다.

다항 로짓 회귀모형

- 단점은 식별가능하지 않다(unidentifiable)는 점이다. 가변수를 포함하여 범주형 변수를 모두 다 모형에 집어넣을 수 없는 이유와 일맥상통한다.
- 식별가능하게 만들기 위한 가장 쉬운 방법은 어느 한 범주(즉 기준 대안)에 대해 $\beta^{(k)} = 0$을 설정하는 것이다.
- 그 결과 우리는 아래처럼 식별가능한 확률모형을 얻는다.

$$P(Y = j) = \frac{1}{1 + \sum_{k=2}^{J} e^{V_k}}$$

- 이것이 바로 우리가 앞서 사용한 다항 로짓 회귀모형의 예측 확률이다.

다항 로짓 회귀모형

- 여기서 식별성(identification) 문제는 상당히 까다로우므로, 직관적으로 살펴보기만 하자.
- 만약 $e^{X\beta^F} = 1$, $e^{X\beta^I} = 2$, $e^{X\beta^O} = 3$이라면, 우리는 아래의 확률값을 얻는다.

$$P(Y = F) = \frac{1}{6}, \quad P(Y = I) = \frac{2}{6}, \quad P(Y = O) = \frac{3}{6}$$

- 그런데 $e^{X\beta^F} = 0.1$, $e^{X\beta^I} = 0.2$, $e^{X\beta^O} = 0.3$이라도 동일한 확률값을 얻게 된다.
- 즉 다항 로짓의 일반화된 확률모형에서, 확률은 $\beta^{(j)}$의 절대값이 아니라 상대적인 차이에 의해 결정되므로 확률은 같으므로, 모든 $\beta^{(j)}$에 상수를 더하거나 빼도 확률은 같다. 즉 모형은 식별되지 않는다(unidentified).

$$P(Y = j) = \frac{e^{X\beta^j}}{\sum_{k=1}^{J} e^{X\beta^k}}$$

다항 로짓 회귀모형

- 모형이 식별가능하려면 어떻게 해야 할까? 널리 사용되는 방식은 기준 대안의 회귀계수를 $\beta = 0$로 고정시키는 것이다.

$$P(Y = F) = \frac{e^{X\beta^F}}{1 + e^{X\beta^F} + e^{X\beta^I}}$$

$$P(Y = I) = \frac{e^{X\beta^I}}{1 + e^{X\beta^F} + e^{X\beta^I}}$$

$$P(Y = O) = \frac{1}{1 + e^{X\beta^F} + e^{X\beta^I}}$$

- 이제 β^F와 β^I는 β^I에 대한 상대적 효과 또는 상대적 위험(relative risk)로 해석 가능하다(Why?).
- $e^{X\beta^F} = 1$, $e^{X\beta^I} = 2$일 때와 $e^{X\beta^F} = 0.1$, $e^{X\beta^I} = 0.2$일 때 확률값이 이제 확실히 다르다.

다항 로짓 회귀모형

- 이것이 우리가 이제부터 사용할 '식별가능한' 다항 로짓의 확률모형이다.

$$P(Y = j) = \frac{1}{1 + \sum_{k=2}^{J} e^{V_k}}$$

- 대안이 J개라면 회귀식은 $J - 1$개 추정되어야 한다. 가령 대안이 3개(F, I, O)였다면 추정할 회귀식은 2개가 된다.
- 대안이 여러 개일 때, 여러 개의 이항 로짓 모형으로 추정해 볼 수도 있을 것이다. 그에 비한다면 다항 로짓 모형에서는 이 식들을 동시에 추정하는 점에서 차이가 있다.

회귀계수의 해석

회귀계수의 해석

다항 로짓 회귀계수의 해석 역시 이항 로짓 회귀모형과 본질적으로 같다.

- F와 O의 선택에 관한 다항 로짓의 확률모형을 다시 생각해보자.

$$P(Y = F|X) = \frac{e^{\beta_0^F + \beta_1^F X}}{1 + \sum e^{\beta_0 + \beta_1 X}}$$

$$P(Y = O|X) = \frac{1}{1 + \sum e^{\beta_0 + \beta_1 X}}$$

- 이때 분모가 모두 같으므로, 다음과 같이 쓸 수 있다.

$$\frac{P(Y = F|X)}{P(Y = O|X)} = e^{\beta_0^F + \beta_1^F X}$$

- 다음과 같이 회귀계수의 의미를 명확히 이끌어 낼 수 있다.

$$\frac{\frac{P(Y = F|X+1)}{P(Y = O|X+1)}}{\frac{P(Y = F|X)}{P(Y = O|X)}} = \frac{e^{\beta_0^F + \beta_1^F (X+1)}}{e^{\beta_0^F + \beta_1^F X}} = e^{\beta_1}$$

회귀계수의 해석

- 이항 로짓과 비교해보면 사실상 똑같은 아이디어임을 확인할 수 있다.

$$\ln \frac{P(Y=1|X)}{1-P(Y=1|X)} = \beta_0 + \beta_1 X$$

- 독립변수 X가 한 단위 더 증가하면, 예전 상태와 비교하여 상대위험비(relative risk ratio; RRR)는 다음과 같이 정의된다.

$$RRR = e^{\beta_1}$$

- 즉 e^{β_1}는 X가 $X+1$로 한 단위 증가할 때 '상대위험이 달라지는 비율'을 보여준다 (Why?).
- 이때 오즈(odds)가 아니라 상대위험(relative risk)이라는 표현이 사용됨에 주의하자.

회귀계수의 해석

- 상대위험비는 정의상 비율(ratio)이므로, 가령 상대위험비가 1.5라면 분모의 상대위험보다 분자의 상대위험이 50% 크다는 것을 뜻한다.
- 그러므로 '상대위험이 달라지는 비율'을 다음과 같이 백분율로 나타낼 수 있다(Why?).

$$\Delta\% = 100 \cdot \left(e^{\beta_1} - 1\right)$$

- "X가 한 단위 증가하면, $Y = O$에 대신 $Y = F$를 선택할 상대위험이 $100 \times (e^{\beta_1} - 1)$ 퍼센트 증가한다."
- 다항 로짓 회귀계수의 상대위험비 해석은 혼동하기 쉬우므로 많이 연습해야 한다.

회귀계수의 해석

- 데이터를 통해 실제 추정해보면 상대위험비를 다음과 같이 해석할 수 있다.
- “악어의 길이가 1m 커질수록 기타(O) 먹이보다 연체류(I) 먹이를 선택할 로그상대위험비는 2.465만큼 감소한다.”
- “악어의 길이가 1m 커질수록 기타(O) 먹이보다 연체류(I) 먹이를 선택할 상대위험비는 91.5% (=$100 \times e^{2.465} - 1$)로 감소한다.”
- 상대위험비 해석에서 기준 대안이 무엇인가를 반드시 언급해야 한다(Why?).

회귀계수의 해석

만일 기준 대안을 연체류(I)로 바꾼다면 어떨까?

- 물론 기타(O) 말고 연체류(I)로 기준 대안으로 바꾸어 다시 한 번 다항 로짓 회귀식을 추정하는 것이 편하다.
- 하지만 수학적으로 다음의 관계가 성립함을 알 수 있다(Why?).

$$\ln \frac{P(Y=F)}{P(Y=O)} = \beta_{0,F} + \beta_{1,F} X$$

$$\ln \frac{P(Y=I)}{P(Y=O)} = \beta_{1,I} + \beta_{1,I} X$$

$$\begin{aligned} \ln \frac{P(Y=F)}{P(Y=I)} &= \ln \frac{P(Y=F)}{P(Y=O)} - \ln \frac{P(Y=I)}{P(Y=O)} \\ &= (\beta_{0,F} - \beta_{0,I}) + (\beta_{1,F} - \beta_{1,I}) X \end{aligned}$$

회귀계수의 해석

직관적인 해석을 위해서는 예측 확률로 전환하는 편이 낫다.

- j번째 대안을 선택할 예측 확률은 다음과 같다.

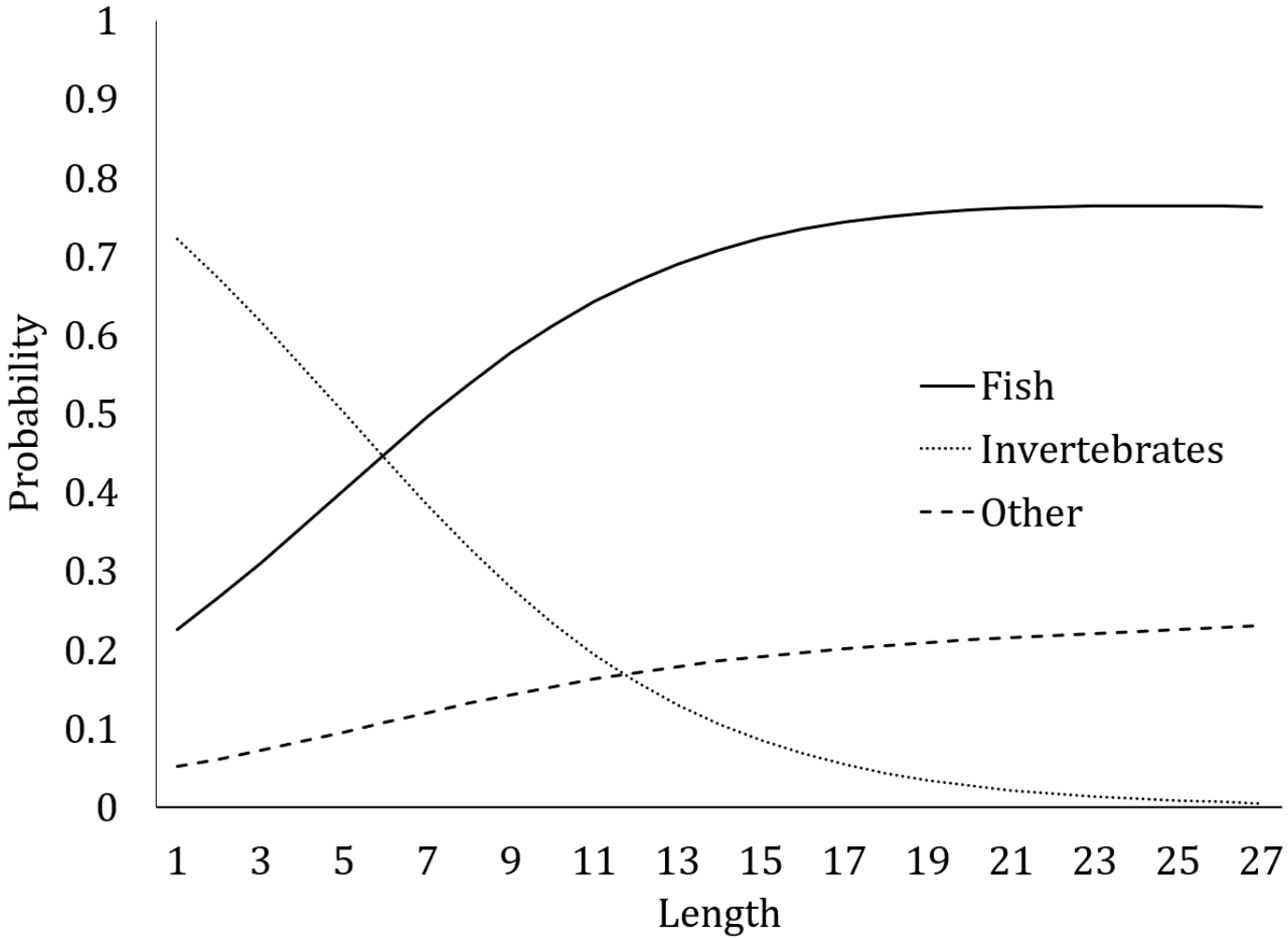
$$P(Y=j) = \frac{e^{\beta_0^{(j)} + \beta_1^{(j)} X}}{1 + \sum_{k=2}^{m} e^{\beta_0^{(k)} + \beta_1^{(k)} X}}$$

- 기준 대안(1)을 선택할 예측 확률은 다음과 같다.

$$P(Y=1) = \frac{1}{1 + \sum_{k=2}^{m} e^{\beta_0^{(k)} + \beta_1^{(k)} X}}$$

회귀계수의 해석

- 예측확률 해석은 다항 로짓에서 특히 빛을 발한다!



유의성 검정과 모형 적합도

유의성 검정과 모형 적합도

유의성 검정과 모형적합도 역시 이항 로짓 회귀모형과 똑같다.

- (이항 로짓과 마찬가지로) 다항 로짓 회귀모형에서도 회귀계수의 유의성 검정을 위해 왈드 검정을 사용한다.

$$Wald = \left(\frac{b-\beta}{SE_b}\right)^2 = \left(\frac{b}{SE_b}\right)^2 \sim \chi_1^2$$

- 이항 로짓 회귀모형과는 자유도(degree of freedom)가 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.

유의성 검정과 모형 적합도

- 모형적합도를 살펴보기 위해 (1) 우도비 검정(likelihood-ratio test), (2) 유사 결정계수(pseudo R^2), (3) 정보 기준(information criteria), (4) 분류표 (classification table)를 확인한다.
- 첫째, 우도비 검정의 검정통계량 G는 χ^2 분포를 따른다.

$$G = -2 \ln \frac{L_{null}}{L_{model}} = -2 \ln(L_{null} - L_{model}) \sim \chi^2_k$$

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

$$H_0 : \lambda_{null} = \lambda_{model}$$

- $\chi^2 = 16.8$이므로 99.9% 신뢰수준에서 통계적으로 유의하게 귀무가설을 기각한다 ($p < 0.0002$). 그러므로 '전반적으로' 악어의 길이와 먹이 선택에는 통계적 연관성이 존재한다!

유의성 검정과 모형 적합도

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

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

- $\lambda_{restricted}$는 독립변수가 다소 적게 들어간 모형의 우도 함수값이고, λ_{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.

$$H_0 : \lambda_{restricted} = \lambda_{full}$$

유의성 검정과 모형 적합도

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

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

- 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.

$$AIC = -2\ln\lambda_{model} + 2k$$
$$BIC = -2\ln\lambda_{model} + k\ln(n)$$

- 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.

유의성 검정과 모형 적합도

- 넷째, 분류표를 통해 로짓 회귀모형에 기반하여 예측된(predicted) $\widehat{Y}$와 실제 자료 Y를 행렬로 비교하여 나타낸다.
- 이것은 다항 로짓 회귀모형의 예측 정확성을 평가하는 도구이며 우도 함수는 사용하지 않는다.
- 당연히 실제 자료와 모형의 예측이 일치할수록, 즉 정확하게 분류된(correctly classified) 사례가 많을수록 모형이 우수하다고 볼 수 있다.
- 아쉽게도 Stata에서는 이를 간단히 구현할 수 있는 명령어가 아직 없다. 여러분이 스스로 만들 수 있을 것이다.