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
4. 무관한 대안의 독립성
5. 다항 프로빗 회귀모형

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.

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

- 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^F 와 β^I 는 β^I 에 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.

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

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?).

$$
\begin{aligned}
\ln \frac{P(Y=F)}{P(Y=O)} &= \beta_0^F + \beta_1^F X \\
\ln \frac{P(Y=I)}{P(Y=O)} &= \beta_1^I + \beta_1^I X \\
\ln \frac{P(Y=F)}{P(Y=I)} &= \ln \frac{P(Y=F)}{P(Y=O)} - \ln \frac{P(Y=I)}{P(Y=O)} \\
&= (\beta_0^F - \beta_0^I) + (\beta_1^F - \beta_1^I) X
\end{aligned}
$$

회귀계수의 해석

직관적인 해석을 위해서는 예측 확률로 전환하는 편이 낫다.

- j번째 대안을 선택할 예측 확률은 다음과 같다.

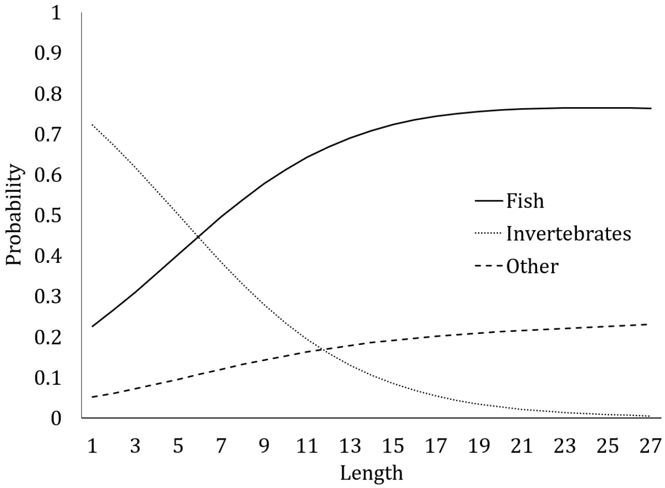
$$P(Y=j) = \frac{e^{\beta_0^j + \beta_1^j X}}{1 + \sum_{k=2}^{m} e^{\beta_0^k + \beta_1^k X}}$$

- 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!



회귀계수의 해석

좀 더 어렵지만 평균한계효과가 가장 훌륭한 해석을 제공한다.

- i번째 응답자의 한계효과(ME)는 어떤 변수 x가 한 단위 변화할 때, j번째 대안의 선택확률 $P(Y = j)$에 미치는 영향을 말한다.

$$\mathrm{ME}_{ij} = \frac{\partial P(Y_i = j)}{\partial x_i}$$

- 다항 로짓 회귀모형에서는 ME의 닫힌 해(closed-form solution)를 계산할 수 있다 (증명 생략).

$$\frac{\partial P(Y_i = j)}{\partial x_{ir}} = P(Y_i = j) \left(\beta_{jr} - \sum_{k=1}^{J} P(Y_i = k) \beta_{kr} \right)$$

- 이때 x_r는 r번째 독립변수이고, $P(Y_i = j) = \dfrac{\exp(X_i \beta_j)}{\sum \exp(X_i \beta_k)}$ 이다.

회귀계수의 해석

- 평균한계효과(AME)는 주어진 표본에서 개별적인 한계효과의 평균을 계산한 것이다.

$$\text{AME}_j = \frac{1}{N}\sum_{i=1}^{N} \text{ME}_{ij}$$

- ME와는 달리 AME에서는 하첨자 i가 붙어있지 않음에 주목하자.
- 물론 다항 로짓 회귀모형에서 AME를 계산할 때는 $j = 1, \ldots, J$에 대해 각각 계산해야 한다.

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!

- 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.

무관한 대안의 독립성

무관한 대안의 독립성

다항 로짓 회귀모형에도 몇 가지 가정이 있다.

- 가령 (1) 관측치의 모든 대안에 대한 오차항은 상호 독립적이고 동일한 분포를 가지고 (independent and identically distributed; i.i.d), (2) 오차항은 굼벨 분포를 따르며, (3) 효용 극대화(utility maximization) 등이 있다.
- 오늘 주목할 부분은 무관한 대안의 독립성(Independence of Irrelevant Alternatives; IIA)이다.
- 이것은 두 대안 중 어느 하나를 선택할 상대위험(relative risk)이 다른 대안(들)에 의해 영향받지 않는다는 가정이다. 제3의 무관한 대안(irrelevant alternative)을 더하거나 빼도, 다른 상대위험들에는 변화가 없어야 한다.

$$\frac{P(y=i)}{P(y=j)} = \frac{e^{V_i}}{e^{V_j}} = k_0$$

무관한 대안의 독립성

- 경우에 따라 무관한 대안의 독립성은 아주 쉽게 깨진다.
- 교통 수단 선택을 사례로 생각해보자. 버스(A)를 이용하는 것과 승용차(B)를 이용하는 것 사이에서 무차별한 경우의 상대위험을 계산해보자.

$$P(A) = 0.5, \quad P(B) = 0.5, \quad \frac{P(A)}{P(B)} = \frac{0.5}{0.5} = 1$$

- 여기에 지하철(C)을 추가해보자. 만약 IIA가 유지된다면 반드시 아래가 성립해야 한다(Why?).

$$P(A) = \frac{1}{3}, \quad P(B) = \frac{1}{3}, \quad P(C) = \frac{1}{3}, \quad \frac{P(A)}{P(B)} = \frac{0.5}{0.5} = 1$$

무관한 대안의 독립성

- 하지만 현실에서 같은 대중교통인 지하철(C)은 아무래도 버스(A)와 경쟁하기 때문에, 선택할 확률을 나눠 가지게 되고 그 결과 기존 상대위험이 변화한다.

$$P(A) = 0.3, \quad P(B) = 0.5, \quad P(C) = 0.2, \quad \frac{P(A)}{P(B)} = \frac{0.3}{0.5} = 0.6$$

- 이런 교통 수단처럼 복수의 대안이 서로 유사하거나 특정 대안이 다른 대안에 내포(nested)되어 있다면 IIA 가정은 쉽게 위배될 수 있다(e.g, 분당한 정당들, 일반 콜라와 다이어트 콜라 등).

무관한 대안의 독립성

IIA 가정의 성립 여부를 판정하기 위해 Hausman 검정을 사용할 수 있다.

- 다항 로짓 회귀모형에서 회귀계수의 유의성 검정에는 왈드 검정을 사용했다.

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

- 이 원리를 확장하여 Hausman-McFadden 설정 검정(specification test)은 다음과 같은 통계량을 사용한다.

$$(b_c - b_e)^T \left[\mathrm{Var}(b_c) - \mathrm{Var}(b_e)\right]^{-1} (b_c - b_e) \sim \chi_k^2$$

- 여기서 b_c는 전체 모형에서의 얻은 일치(consistent) 추정량이고, b_e는 대안이 제거된 축소 모형에서의 효율(efficient) 추정량이다.
- IIA가 성립한다면 두 추정량은 통계적으로 유의하게 다르지 않아야 한다(Why?).

무관한 대안의 독립성

- (분자 부분을 보면 알 수 있듯) 핵심 아이디어는 전체 대안을 포함한 모형과 일부 대안을 제거한 축소 모형에서의 회귀계수 추정값이 일치해야 한다는 것이다.
- 만일 두 추정량 간의 차이가 유의하다면, 대안의 제거가 다른 대안 간의 관계에 영향을 미친 것으로 간주되므로 IIA 가정이 위배되었다고 판단한다.
- Hausman 검정의 귀무가설은 b_c와 b_e 간에 차이가 없다는 것이다.
- 따라서 이 검정은 “대안 하나를 제거해도 나머지 회귀계수가 안정적인가?”를 확인하는 셈이다.
- 안타깝게도 Hausman 검정이 제대로 계산되지 않는 경우가 종종 있다. 특히 (각 대안에 속하는 관측치가 적거나 하면) 분모 부분을 구하지 못할 때가 있다.

무관한 대안의 독립성

IIA 가정이 성립하지 않았다면 다른 수단을 강구할 수 있다.

- 혼합 다항 로짓(mixed multinomial logit) 회귀모형은 임의효과(random effect) 항을 모형 안에 투입하여 IIA 문제에 대응한다. 이것은 확률선택모형(stochastic choice model) 중에서도 제법 수준이 높으므로 우리 수업의 범위를 벗어난다.
- 근본적으로 IIA 가정 위배로 인한 문제가 왜 생기는지 생각해보자.
- 대안들 사이에 어떤 상관관계가 있는데 이것이 적절히 모델링되어 있지 않다면, 관찰되지 않은 이질성(unobserved utilities)으로 인해 IIA가 위배된다.
- 다항 로짓 모형의 오차항은 i.i.d인 굼벨 분포를 따른다는 가정되므로, IIA 문제를 피할 수 없다.

$$\mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) = 0$$

다항 프로빗 회귀모형

다항 프로빗 회귀모형

그럼 다항 프로빗 회귀모형이라면 어떨까?

- 다항 프로빗(multinomial probit) 모형에서도 i번째 관찰값은 J개의 대안 중 하나를 다음과 같은 효용함수에 따라 선택한다.

$$
\begin{aligned}
U_{ij} &= V_{ij} + \varepsilon_{ij}, \quad (j = 1, \cdots, J) \\
V_{ij} &= X_{ij}\beta_j
\end{aligned}
$$

- 다만 오차항 $\varepsilon_{ij} = (\varepsilon_{i1}, \ldots, \varepsilon_{iJ})$이 다변량 정규(multivariate normal) 분포를 따른다고 가정한다.

$$\varepsilon_i \sim \mathcal{N}(0, \Sigma)$$

- 오차항들 사이의 공분산 구조(covariance structure)를 명시적으로 모형화할 수 있으므로, 제대로 모형 설정했다면 이론상 이 문제를 피할 수 있다(Dow and Endersby 2004).

다항 프로빗 회귀모형

- 다항 프로빗 모형에서 k번째 대안 대신 j번째 대안을 선택할 확률은 다음과 같이 쓸 수 있다($\forall k \neq j$).

$$\begin{aligned} P(y_i = j) &= P(U_{ij} > U_{ik}) \\ &= P(\varepsilon_{ij} - \varepsilon_{ik} < V_{ij} - V_{ik}) \end{aligned}$$

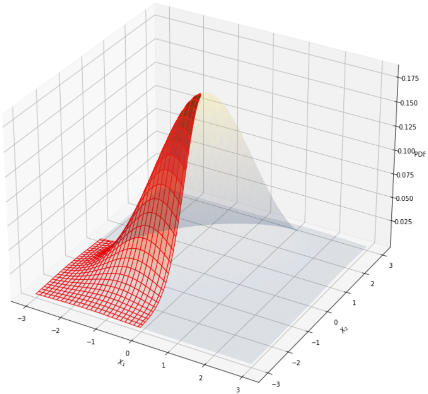
- 오차의 차이 $(\varepsilon_{ij} - \varepsilon_{ik})$는 정규분포이므로, 선택확률은 다음과 같은 $(J-1)$ 차원 누적정규(cumulative normal) 분포로 계산된다.
- 만약에 1, 2, 3의 선택 대안이 있다면, $P(y_i = 1)$은 다음과 같다.

$$P(y_i = 1) = \int_{-\infty}^{V_{i1}-V_{i2}} \int_{-\infty}^{V_{i1}-V_{i3}} \phi_2(\varepsilon_{i1} - \varepsilon_{i2}, \varepsilon_{i1} - \varepsilon_{i3})\, d(\varepsilon_{i1} - \varepsilon_{i3}) d(\varepsilon_{i1} - \varepsilon_{i2})$$

- 이때 ϕ_2는 이변량 정규분포(bivariate normal)임을 의미하고, 두 변수의 결합분포(joint distribution)의 확률밀도함수(PDF)이다.

다항 프로빗 회귀모형

- 이변량 정규분포 ϕ_2는 가령 이렇게 생겼다. 적분으로 계산된 확률은 이러한 2차원 공간 안의 면적으로 색칠된 영역이다(Why?).



다항 프로빗 회귀모형

- 다항 프로빗 모형의 우도 함수는 다음과 같다.

$$\mathcal{L} = \prod_{i=1}^{N} P(y_i = j_i \mid X_i)$$

- 이때 각 $P(y_i = j)$는 다변량 정규분포의 적분으로 계산되어야 하나, 일반적으로 3 차원 이상일 경우, 닫힌 해를 구하기 어렵다.
- 하는 수 없이 수치 해석 또는 시뮬레이션(e.g., maximum simulated likelihood 또는 Markov Chain Monte Carlo)을 사용하게 된다. 이때 다항 프로빗 모형이 수렴(convergence)한다는 보장은 아무래도 상대적으로 낮다(Why?).
- 같은 표본 크기에 대해 선택 대안이 많아지면, 계산 비용은 기하급수적으로 높아진다 (J차원 다변량 정규분포의 적분이 필요하기 때문이다).

다항 프로빗 회귀모형

식별성 문제도 남아있다.

- 다항 프로빗 모형의 식별성(identification)은 두 가지 측면이 모두 만족되어야 한다.
- 첫째, 위치 식별성(location identification), 즉 하나의 효용을 기준으로 삼고(e.g., $U_{iJ} = 0$), 나머지는 상대 효용으로 표현한다.
- 둘째, Σ와 관련해서도 자유롭게 추정될 수 있도록 내버려 둘 수 없다.
- 따라서 스케일 식별 제약(scale identifcation)을 위해 Σ의 대각 전체 또는 일부 값을 고정해야 한다(e.g., $\mathrm{Var}(\epsilon_{ij} - \epsilon_{ik}) = 1$).
- 결론적으로 $J - 1$개의 상대 효용만 추정되고, Σ도 일부만 자유롭게 움직일 수 있다.

다항 프로빗 회귀모형

다항 프로빗 모형에서의 한계효과는 닫힌 해로 찾을 수 없다.

- 다항 프로빗의 확률모형은 다음과 같았다.

$$P(Y_i = j) = \int_{\mathcal{R}_j} \phi_{J-1}(\varepsilon) d\varepsilon$$

- 그러므로 이를 x_{ir}에 대해 미분하면 다음과 같다.

$$\frac{\partial P(Y = j)}{\partial x_{ir}} = \int_{\mathcal{R}_j} \frac{\partial}{\partial x_{ir}} \phi_{J-1}(\varepsilon) d\varepsilon$$

- 여기서는 닫힌 해가 없고 수치적 방법이 필요하다. 주로 시뮬레이션에 기반하여 AME의 근사값을 찾게 된다.

다항 프로빗 회귀모형

결론을 내려보자.

- 다항 로짓 회귀모형은 계산 과정이 명쾌하지만, IIA 문제에 취약하다.
- Hausman 검정을 수행하고, IIA 문제가 심각하지 않음을 주장한다.
- IIA 문제를 피할 수 없어 보인다면, 다항 프로빗 회귀모형을 사용한다. 이때 다항 로짓 회귀모형의 추정 결과와 크게 다르지 않은지 확인해보자(Why?).
- 다항 프로빗 회귀모형은 유연한 공분산 구조의 설정이 가능하지만 계산 비용이 몹시 크다.
- 다행스럽게도 대다수의 리뷰어는 수학에 어둡기 때문에 이 문제로 태클을 걸 것 같지는 않다.