로짓회귀분석:
로짓회귀분석 소개 및 추정치 전환

숙명여자대학교 경영학부 오중산

로짓회귀분석 소개

- 로짓회귀분석(Logistic regression)이란?
  - 회귀분석의 한 형태로서, DV를 범주형 척도인 이진수로 측정
    - 로짓회귀분석에서 표준화 회귀계수 추정치는 추정되지 않음
  - DV는 사건(event) 발생여부 혹은 분류와 관련됨
    - 발생여부의 경우, 측정값 0은 “No”(발생하지 않음)을, 1은 “Yes”(발생함)를 의미
    - 분류의 경우, 측정값 0은 해당 사례가 첫 번째 집단에 속함을, 1은 두 번째 집단에 속함을 의미

로짓회귀분석 소개

- 로짓회귀분석의 목적
  - 목적1: 인과관계 규명
    - 기존 사례를 통해 추정된 로짓회귀식은 DV와 IV간의 인과관계를 규명함
  - 목적2: 새로운 사례 분류 예측
    - 추정된 로짓회귀식(예측모형)을 활용해 새로운 사례 분류(사건 발생 여부) 예측
    - 로짓회귀분석 역시 ML의 지도학습 유형에 해당됨

로짓회귀분석 소개

- 로짓회귀분석 예시
  - 예시1: 기업부도 예측모형
    - 신용평가기관에서 부채비율/영업이익률/상환 연체율/매출증가율과 같은 IV를 통해 기업의 부도에 영향을 미치는 요인을 파악하거나, 부도 여부 예측
  - 예시2: 고객이탈 예측모형
    - 신용카드회사에서 고객의 행동 특성 관련 IV(예: 사용액, 사용빈도)를 측정하여 이탈에 영향을 미치는 요인을 파악하거나, 이탈할지 여부를 예측
  - 예시3: 직원 이직 예측모형
    - 기업에서 어떤 직원의 행동, 성과, 인구통계학적 속성 등과 관련된 IV를 측정하여 이직에 영향을 미치는 요인을 파악하거나, 이직을 예측

로짓회귀분석에서 DV 추정치의 전환

- DV 측정치 vs 추정치
  - 측정치(실제값)는 0/1이지만, 추정치(logit value)는 -∞ ~ +∞ 구간의 실수(real number)
    - DV 추정치가 -∞에 가까울수록 사건은 발생하지 않고, +∞에 가까울수록 사건은 발생
  - 그렇다면, DV 추정치를 측정치(실제값)처럼 0이나 1로 변환할 수 없을까?
    - 1단계: DV 추정치를 0~1 사이의 확률값으로 전환
    - 2단계: 전환된 확률값을 기준값(0.5)에 따라 0이나 1로 변환

로짓회귀분석에서 DV 추정치의 전환

- 1단계: DV 추정치를 0~1 사이의 확률값(Pr)으로 전환
  - ◆ 추정할 로짓회귀식 $\hat{Y}_i = a + b_j X_{ji}$ ($j = 1, 2, 3, \ldots, k$)

$$= ln\{\Pr \div (1 - \Pr)\} = ln(\text{Odds})$$

  - ◆ $e^{\hat{Y}i} = \Pr \div (1 - \Pr) = \text{Odds}$ (자연로그의 밑인 무리수 $e = \lim_{x \to \infty} \left(1 + \frac{1}{x}\right)^x = 2.71828\ldots$)
    - ▪ $\Pr = \frac{e^{\hat{Y}i}}{1 + e^{\hat{Y}i}}$ 추정치가 $-\infty$이면 Pr은 0, 추정치가 $+\infty$이면 Pr은 1, 추정치가 0이면 Pr은 0.5

로짓회귀분석에서 DV 추정치의 전환

- 2단계: 전환된 확률값을 기준값(0.5)에 따라 0이나 1로 변환
  - ◆ 확률값이 1에 가까울수록, 사건 발생 가능성이 높음
  - ◆ 확률값이 기준값인 0.5 이상이면 DV를 1로, 작으면 0으로 변환
    - ▪ 추정치 → 확률 → 0/1 분류(예측값) 과정을 거침

Logit value ($\hat{Y}_i$)	Probability (Pr $= \frac{e^{\hat{Y}_i}}{1+e^{\hat{Y}_i}}$)	Odds ($\frac{Pr}{1-Pr}$)	0/1 분류(예측값)
-∞	**0.00**	**0.00**	**0**
-2.20	**0.10**	**0.11**	
-0.85	**0.30**	**0.43**	
0.00	**0.50**	**1.00**	**1**
0.85	**0.70**	**2.33**	
+∞	**1.00**	+∞	

로짓회귀분석:
로짓회귀식 추정 및 모형적합도

숙명여자대학교 경영학부 오중산

로짓회귀식 추정

- 로짓회귀계수 추정치 추정
  - Wald statistics
    - 로짓회귀분석에서는로짓회귀계수 추정치의 Wald-통계량에 대한 유의성을 통해 가설검정
    - Wald-통계량은 자유도가 1인 χ^2 분포를 따르며, 도출된 p-value와 α를 비교하면 됨(단측검정)
      - 회귀계수 추정치가 유의한 양수이면, 해당 IV가 사건 발생 확률을 높임(예측값 1)
      - 회귀계수 추정치가 유의한 음수이면, 해당 IV가 사건 발생 확률을 낮춤(예측값 0)

로짓회귀식 추정

- Hit ratio (or percent concordance)
  - ◆ DV 측정치(실제값)와 예측값 간의 일치도 관련 지표
    - Hit ratio는 0~1 사이 값을 가지며, 0은 0으로, 1은 1로 예측(분류)한 빈도가 많을수록 1에 수렴
    - 아래 교차표에서 hit ratio가 높으면 모형의 (예측)정확도가 높다고 말할 수 있음

구분		예측값		Hit ratio
		0	1	
측정치	0	41	14	0.745
	1	17	28	0.622
합계		58	42	**0.690**

모형적합도

- 로짓회귀식의 모형적합도
  - Hosmer and Lemeshow test
    - χ^2-test를 통해 통계적으로 유의하지 않으면 모형이 적합함
  - -2LL (-2 log likelihood value)
    - 최소값은 0이며, 작을수록 모형적합도(GoF)가 좋음
    - IV 추가의 정당성 확인에 활용됨
      - IV를 추가하면 -2LL값은 감소함
      - χ^2-test를 통해 Δ-2LL이 유의하면 IV 추가 가능하며, 해당 로짓회귀식의 모형적합도가 더 좋음

모형적합도

- 로짓회귀식의 모형적합도
  - ◆ *P*seudo R^2
    - ▪ 다중회귀분석에서의 R^2와 비슷한 지표
    - ▪ $R^2_{\text{Logit}} = \{-2LL_{\text{null}} - (-2LL_{\text{model}})\} \div -2LL_{\text{null}}$
      - ❖ null은 IV 없이 상수만으로 구성된 아무 의미가 없는 로짓회귀식이며, model은 null에 IV 추가됨
    - ▪ 두 가지 종류의 *P*seudo R^2
      - ❖ Cox & Snell R^2 와 Nagelkerke R^2
      - ❖ 두 값 모두 1에 가까울수록 모형적합도가 좋으며, 기준은 없음