회귀분석: 가설검정과 모형의 간명성

숙명여자대학교 경영학부 오중산

다중회귀분석에서 가설검정

- 다중회귀분석에서의 전체 가설
  - 귀무가설
    - 어떤 독립변수도 종속변수를 설명할 수 없음
    - X_{ji}와 Y_i사이에 통계적으로 유의한(significant) 인과관계가 전혀 존재하지 않음
    - H_0: 모든 $\beta_j = 0$ ($j = 1, 2, 3, \ldots, k$)
  - 대립가설
    - 적어도 하나의 독립변수는 종속변수를 설명할 수 있음
    - X_{ji}와 Y_i사이에 유의한 (양의 혹은 음의) 인과관계가 적어도 하나는 존재함
    - H_a: 적어도 하나의 $\beta_j \neq 0 (> 0$ 혹은 $< 0)$ ($j = 1, 2, 3, \ldots, k$)

다중회귀분석에서 가설검정

- 다중회귀분석에서의 개별 가설
  - 독립변수 개수(k) 만큼의 가설쌍이 존재
    - $H_0: \beta_j = 0$ ($j = 1, 2, 3, \ldots, k$)
    - $H_a: \beta_j \neq 0 (> 0$ 혹은 $< 0)$ ($j = 1, 2, 3, \ldots, k$)
- 가설검정(hypothesis test) 이란?
  - 추정치(b_j)를 이용하여 모수(β_j)에 대해 의사 결정
  - 전체 가설은 F통계량(=MSR / MSE = (SSR / k) / (SSE / (n-k-1)))을 통해 가설검정
  - 개별 가설은 k개 가설쌍에 대해 t통계량을 이용하여 가설검정

다중회귀분석에서 가설검정

- 가설검정(hypothesis test) 이란?
  - ◆ 양/단측 *t*-검정
    - 회귀계수 추정치(b_j)에 대해 *t*-통계량에 따른 *p*-value를 산출하여 유의수준과 비교
      - ❖ 양측검정: $2p\text{-value} \leq \alpha$ 이면, b_j가 통계적으로 유의하므로 H_a 채택
      - ❖ 단측검정: $p\text{-value} \leq \alpha$ 이면, b_j가 통계적으로 유의하므로 H_a 채택
    - H_a 채택은 X_{ji}가 Y_i에 유의한 영향을 미친다는 의미로, 두 변수 간에 양의/음의 인과관계 성립

$$t_j = \frac{b_j - 0}{b_j\text{의 불편 표준오차}} = \frac{b_j}{\sqrt{\hat{\sigma}^2 / \sum (X_i - \bar{X})^2}} = \frac{b_j}{\hat{\sigma} \div \sqrt{\sum X_i^2 - n \cdot (\overline{X})^2}}$$

$\hat{\sigma}$은 DV(혹은 오차) 표준편차의 불편추정량 $= \sqrt{\frac{\sum \varepsilon_i^2}{df=(n-2)}}$

다중회귀분석에서 가설검정

- IV의 중요도: 여러 개 IV 중에서 어떤 IV가 가장 중요한가?
  - “어떤 IV가 DV에 가장 많은 영향을 미치는가?” 혹은 “어떤 IV가 DV를 가장 잘 설명하는가”라는 질문과 동일함
  - IV에 대한 척도 내용/단위 혹은 분포가 다르므로, 이 질문에 대해 단순히 b_j 절대값 크기를 비교하여 답을 할 수 없음
    - DV: 체중 / IV: 신장(*cm*), 일별 섭취량(kcal), 일별 운동시간(분), 음주여부(Y/N) 등….

다중회귀분석에서 가설검정

- IV의 중요도: 여러 개 IV 중에서 어떤 IV가 가장 중요한가?
  - ◆ 여러 IV의 척도 내용/단위 혹은 분포가 다른 상황에서 IV의 '상대적 중요성' 확인 방법
    - IV 간에 '표준화 회귀계수 추정치'의 크기를 비교해야 함
      - ❖ 표준화 회귀계수 추정치란, DV와 IV를 z-통계량으로 표준화한 후 추정된 회귀계수 추정치
    - 유의하게 추정된 표준화 회귀계수 추정치에 한해, 절대값 크기 순서대로 IV의 중요성이 높음

IV 추가의 정당성 확인

- IV의 개수(k)와 모형의 간명성(parsimony)
  - 회귀식에서 n은 적절히 많되 k는 적을수록, 즉 '간명'할수록 바람직함
  - 모형의 간명성을 강조하면 R^2가 낮게 산출되거나, DV에 유의한 영향을 미치는 IV가 부족하거나, DV에 대한 설명을 1~2개 IV에 지나치게 의존할 수 있음
  - 반면, IV를 늘리면 모형 간명성이 떨어지고, 실제로는 GoF 혹은 DV에 대한 설명력이 개선되는 것이 아니라, R^2 만 증가하는 문제가 발생할 수 있음
  - 결국 적정 수의 k를 정하는 것이 요구되며, 이를 위해 k를 최소화한 기본 모형(base model)에서 IV를 순차적으로 늘리는 접근 방식이 필요함

IV 추가의 정당성 확인

- IV 추가의 정당성 확인 방법
  - base model과 new model
    - base model은 기존에 알려진 IV만으로 구성하고, new model은 base model에 새로운 IV를 추가
    - base model의 IV를 통제변수(control variable: CV)라고 하며, DV에 대한 CV의 설명력을 통제한 후, new model을 추정함으로써 DV에 대한 새로운 IV의 추가 설명력이 유의한지 확인
  - 1) R^2_{adj} 증가 2) ΔR^2이 (F-분포) 통계적으로 유의하면 new model 선택
    - 두 조건을 만족할 때 base model에 비해 new model의 설명력이 좋다고 함

$$F = \frac{\Delta R^2 / \Delta df}{MSE_{new}}$$

      - DV에 대해 CV로 설명하지 못하는 특성을 새로운 IV가 설명할 수 있음
      - 회귀식의 간명성은 다소 떨어지나, 이러한 간명성 훼손을 만회할 정도로 GoF가 개선됨