회귀분석: 더미변수와 다중공선성

숙명여자대학교 경영학부 오중산

더미 변수(dummy variable)

- 더미변수란?
  - ◆ 다른 IV와 달리 문자/범주형 척도로 측정되며, 범주가 두 개로 구분되는 IV
    - ▪ 범주는 0(No) 혹은1(Yes)과 같은 이진값
    - ▪ 예) 휴학 여부, 운전면허취득 여부, 흡연 여부
  - ◆ 통제변수와 더미변수
    - ▪ 더미변수를 CV로 사용하는 경우가 많음
      - ❖ 예: 매출액(DV)에 대해 R&D 역량이나 마케팅 역량과 같은 IV의 설명력을 확인하기 전에, 기업규모 (대기업 여부) / 주식시장 상장여부 / 제조업 여부와 같은 더미변수를 CV로 활용한 base model 구성

더미 변수(dummy variable)

- 집단이 여러 개로 구분될 때 더미변수 개수
  - 더미변수 개수 = 집단 개수(t)-1
    - 집단은 서로 배타적이고 포괄적인(mutually exclusive and exhausted) 조건을 만족해야 함
    - Reference 집단
      - T-1개 더미변수 측정값이 모두 NO(0)로 표시되는 집단
      - 다른 집단은 reference 집단 대비 DV가 증가/감소한다고 설명
  - 예시: 학년(t = 4)에서 4학년을 reference로 선정
    - dV_{1i}: 1학년 여부, dV_{2i}: 2학년 여부, dV_{3i}: 3학년 여부
    - 1학년(1, 0, 0) / 2학년(0, 1, 0) / 3학년(0, 0, 1) / 4학년(0, 0, 0)
    - 어떤 더미변수 b_j가 양(음)으로 유의하면, 4학년 대비 해당 학년일 때 DV가 증가(감소)함

다중공선성(multicollinearity)

- 다중공선성이란?
  - IV간에 상관관계가 높으면, 다중공선성으로 인해 다음과 같은 문제가 발생함
    - 유의해야 할 회귀계수 추정치가 유의하지 않게 추정됨
    - 회귀계수 추정치 부호가 반대로 유의하게 추정됨
  - 다중공선성 방지나 모형 간명성을 위해 의미가 유사하거나 중복되는 IV는 제거하는 것이 바람직함

다중공선성(multicollinearity)

- 다중공선성 검사
  - 공차한계(tolerance)
    - 어떤 IV에 대해 다른 IVs가 설명하지 못하는 정도(= 1 - R^{2*})로 0 ~ 1 사이의 값을 가짐
      - R^{2*} 는 특정 IV를 DV로, 다른 IVs를 IV로 하는 회귀식을 추정했을 때의 R^2
    - 어떤 IV의 공차한계가 낮을수록, 다른 IVs와 관련성이 높아서 다중공선성을 일으킬 수 있음
      - IV의 공차한계가 0.19 미만이면, 해당 IV는 다중공선성을 일으킨다고 판단
      - 공차한계 역수인 분산팽창요인(variance inflation factor: VIF)이 5.3보다 큰 IV 역시 다중공선성을 일으킬 수 있음

다중공선성(multicollinearity)

- 다중공선성에 대한 판단과 해소 방안
  - 다중공선성 해소 방안
    - STEP1: VIF가 5.3 이상(혹은 공차한계가 0.19 미만)인 IV 파악
    - STEP2: STEP1에서 파악된 IVs 중에서 제거할 IV 순서 결정
      - 내용적 중요도가 떨어지거나, 다른 IV로 대체 가능한 IV를 우선적으로 제거
    - STEP3: 제거 우선순위 1위 IV를 제거한 후 회귀식 재추정
    - STEP4: 만약 STEP3에서 모든 IV의 VIF가 5.3 미만이 아니면 STEP2로 돌아가 반복