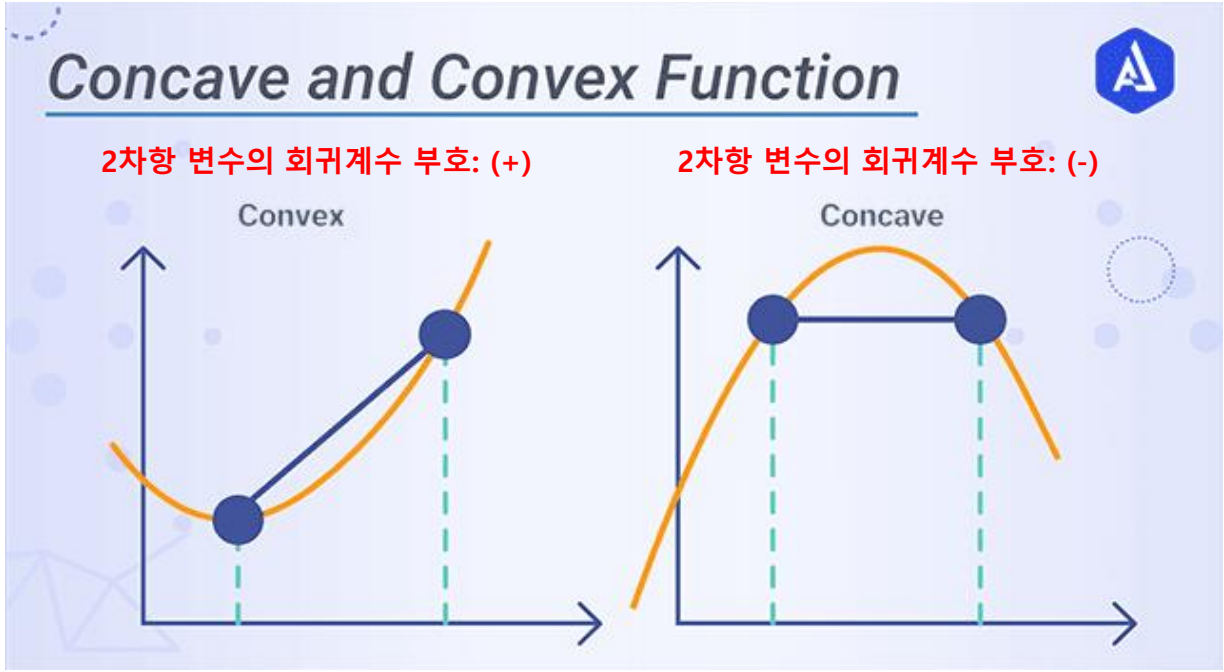
회귀분석: 전제조건과 최소자승법

숙명여자대학교 경영학부 오중산

다중회귀분석

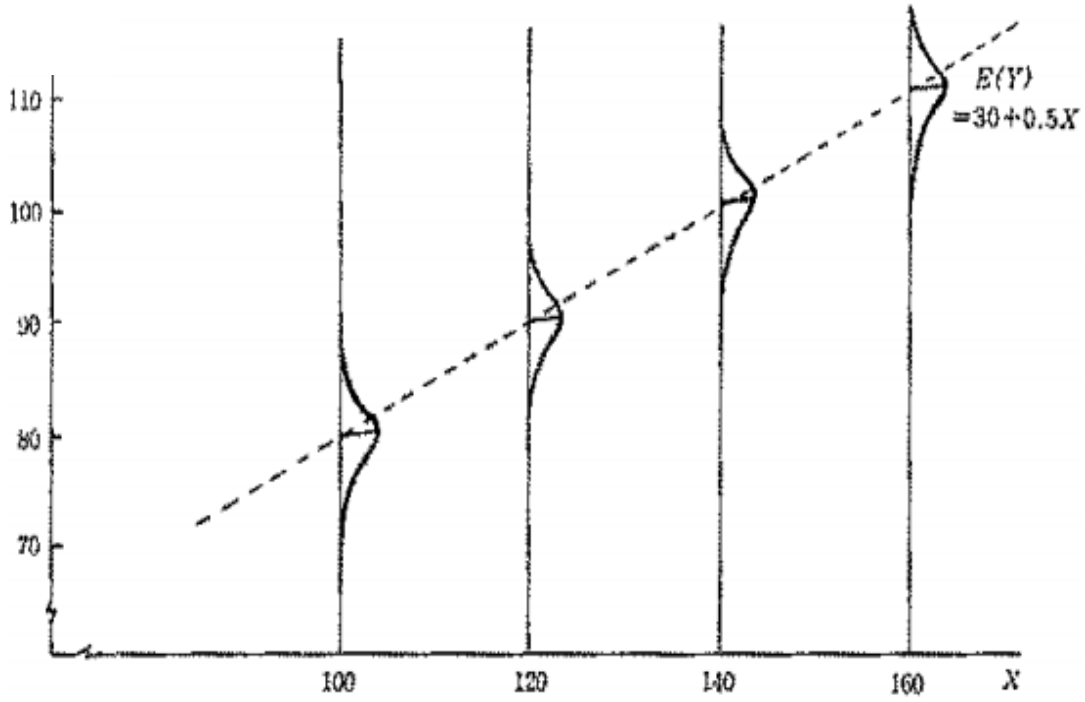
- 다중회귀분석의 네 가지 전제조건
  - ◆ 조건1: 선형성(linearity)
    - ▪ DV와 IV 간의 관계는 선형이어야 함
      - ❖ IV 차수가 1이고, +/-로 연결되어야 함
    - ▪ 반드시 지켜야 할 전제조건은 아니며, 차수가 2인 경우 회귀계수 부호 해석에 유의해야 함



Source: akira.ai

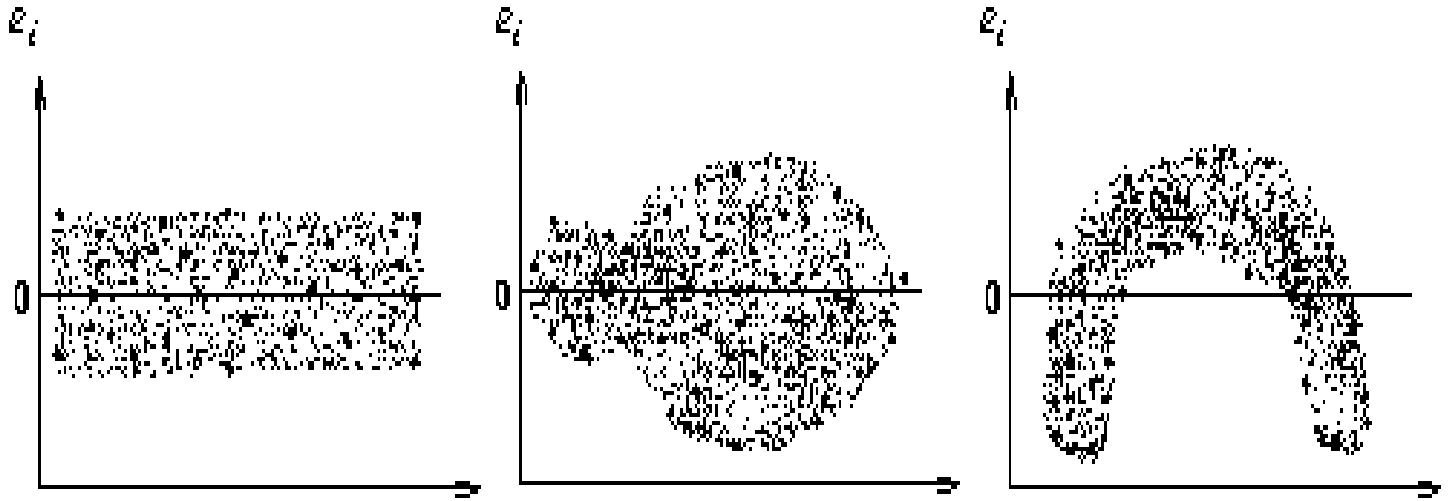
다중회귀분석

- 다중회귀분석의 네 가지 전제조건
  - ◆ 조건2: 정규성(normality)
    - $Y_i \sim N(\hat{Y}_i, \sigma^2)$
      - ❖ 특정 IV에 대해, $E[Y_i] = \alpha + \beta_j X_{ji} \approx a + b_j X_{ji} = \hat{Y}_i$
      - ❖ 기준을 완화해 특정 IV가 아닌, 전체 사례를 대상으로 할 수 있음
    - $e_i \sim N(0, \sigma^2)$라는 조건으로 대체할 수 있음



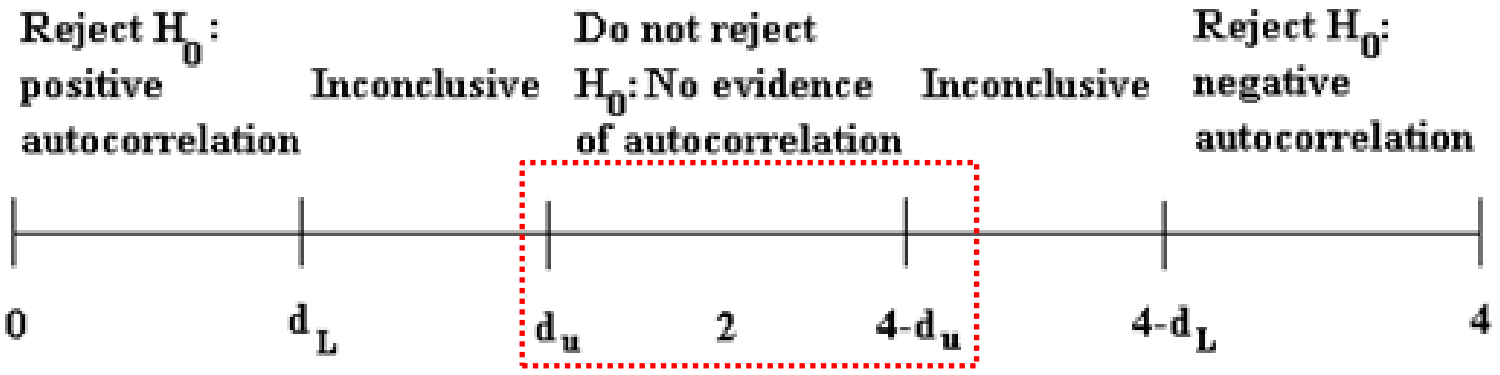
다중회귀분석

- 다중회귀분석의 네 가지 전제조건
  - 조건3: 등분산성(homoscedasticity)
    - 모든 IV에 대해 Y_i 혹은 e_i의 σ^2는 일정함
    - 기준을 완화하여 임의 추출한 두 집단에 대해 분산이 같아야 함



다중회귀분석

- 다중회귀분석의 네 가지 전제조건
  - 조건4: 독립성(independence): Y_i는 Y_p와 독립 ($i \neq p$)
    - 서로 다른 사례 간에 DV 측정에 영향을 주지 않음
    - e_i는 e_p와 서로 독립이라는 조건으로 대체 가능
      - 독립성은 오차의 자기상관(autocorrelation)이 없음을 의미
    - 오차의 자기상관은 Durbin-Watson (DW) 통계량을 통해 확인
      - Lower critical value (d_L)와 Upper critical value (d_u) 구하기



다중회귀분석

- 세 가지 변동
  - SST (sum of squares total) = $\sum(Y_i - \bar{Y})^2$
    - DV의 분산(특성) 전체를 의미
    - $\overline{Y}$는 DV 표본 평균값
  - SSR (sum of squares regression) = $\sum(\hat{Y}_i - \bar{Y})^2$
    - DV의 분산(특성) 중 IV로 설명되는 부분
    - SSR이 클수록 설명력이 좋은 회귀식이라고 할 수 있음
  - SST = SSR + SSE(= $\sum(Y_i - \hat{Y}_i)^2$)
    - SSE(sum of squares residuals)는 DV의 분산(특성) 중 IV로 설명되지 못하는 부분
    - SSE가 작을수록 좋은 회귀식이라고 할 수 있음

다중회귀분석

- 최소자승법(ordinary least squares: OLS)
  - ◆SSE가 최소(혹은 SSR이 최대)가 되도록 α와 β의 추정치인 a와 b를 추정
    - min SSE = min $\{\Sigma\ (\varepsilon_i\ (= Y_i - \hat{Y}_i))^2\}$
    - DV의 분산(특성) 중 IV로 설명되지 못하는 부분을 최소화
    - SSE를 a와 b로 각각 편미분한 후, 연립방정식을 풀면 a와 b를 구할 수 있음

다중회귀분석

- 결정계수(coefficient of determination)
  - R^2 = SSR / SST
    - R^2는 DV의 분산(특성) 중에서 회귀식(혹은 IV)으로 설명 가능한 비율
      - R^2는 0에서 1사이의 값을 가지며, 최저 기준은 없음
      - R^2 크기보다 IV와 DV의 인과관계에 대한 통계적 유의성($\beta_j \neq 0$)이 더 중요함
    - OLS는 SSE를 최소화하고 SSR을 최대화하므로, R^2를 최대화하는 방법
      - SSE가 작을수록(SSR이 클수록) R^2는 1에 가까워 모형적합도(Goodness of Fit: GoF)가 높음

다중회귀분석

- R^2의 한계
  - IV 개수(k)가 증가하면 R^2가 커지는 경향이 있음
    - DV의 분산에 대한 IV의 설명력(혹은 GoF)이 높아지지 못하거나, 설명력이 높아진 정도가 미미한데 수치상으로만 R^2가 증가할 수 있음
  - IV를 늘리거나 추가하는 것은 '모형의 간명성' 원칙에 위배될 수 있음
    - 모형이 복잡해 지더라도 기존 IV로 설명하지 못한 DV의 분산에 대해 새로운 IV가 설명할 수 있다면 IV 추가 가능

다중회귀분석

- R^2의 한계
  - ◆ R^2_{adj}: 수정(Adjusted) R^2 $\quad R^2_{adj} = R^2 - \frac{SSE/df_e}{SST/df_t} = R^2 - \frac{n-1}{n-(k+1)}(1-R^2)$
    - $R^2_{adj} \leq R^2$
      - ❖ k가 증가하면 R^2는 증가하지만, R^2_{adj}는 커질 수도 있고 작아질 수도 있음
      - ❖ 어떤 의미에서 R^2_{adj}가 R^2 보다 더 정확한 GoF 지표이므로, R^2 뿐만 아니라, R^2_{adj}를 함께 고려해야 함
    - 새로운 IV를 추가하려며 아래 두 가지 조건을 모두 만족해야 함
      - ❖ 조건1: 새로운 R^2_{adj}가 기존 R^2_{adj} 보다 증가해야 함
      - ❖ 조건2: R^2의 증가된 정도($\Delta R^2 \sim F$)가 통계적으로 유의해야 함