김은영

제4장 회귀진단

4.1 서론

회귀진단(regression diagnostics) [모형검토(model checking)]
: 적합된 회귀모형의 전반적인 검토

(1) 선형성(linearity)

- 고전적 선형회귀모형들의 가정: $E(Y) = \sum_{j=0}^{p-1} \beta_j X_j$
- 실제 자료들: 선형성의 가정을 만족하지 못하는 경우가 흔히 있다.
- 대안: 비선형모형, 일반화 선형모형 또는 비모수 회귀모형 등

(2) 오차항의 분포

- $\epsilon_i \sim i.i.d. N(0, \sigma^2)$
- 독립성 가정이 옳지 못하고 잔차들 간 자기상관이 있을 것 같다 → 더빈-왓슨 검정
- 등분산 가정이 의심된다 → 변수 변환, 가중 최소제곱법
- 정규분포 가정이 옳지 못한 것 같다 → 변수 변환, 로버스트 회귀 등

(3) 다중공선성(multicollinearity)

■ 설명변수들 간에 높은 상관관계가 있을 경우 β의 추정에 필요한 행렬 $X^t X$가 거의 비정칙 행렬에 가깝게 되고 이러한 상황에서 구한 최소제곱 추정치 $\hat{\beta}$은 매우 불안정한 값이 된다.
- 대안: 능형회귀 등

(4) 영향력 관측치(influential observations)

■ 최소제곱법에서 구한 $\hat{\beta}$, s^2 등은 흔히 한 개 또는 소수개의 관측치에 의해 큰 영향을 받는 경우가 있다. 이 경우 추정치에 큰 영향을 미치는 관측치를 '영향력 관측치'라 부르며, 이러한 관측치의 탐색으로 보다 안정된 추정치를 구할 수 있다.

4.2 잔차

회귀모형: $y = X\beta + \epsilon$

잔차: $e = y - \hat{y} = y - Hy = (I-H)y$ 　　　여기서, $\hat{y} = X\hat{\beta} = X(X^tX)^{-1}X^ty = Hy$

※ 오차항의 독립성을 가정하고 구한 잔차들은 더 이상 독립이 아니고 서로 상관관계를 가지고 있다.

$$E(e) = \left[I - X(X^tX)^{-1}X^t\right]X\beta = X\beta - X(X^tX)^{-1}X^tX\beta = X\beta - X\beta = 0$$

$$Cov(e) = Cov[(I-H)y] = (I-H)Cov(y)(I-H) = (I-H)\sigma^2$$

여기서, $I-H = \begin{bmatrix} 1-h_{11} & -h_{12} & \cdots & -h_{1n} \\ & 1-h_{22} & \cdots & -h_{2n} \\ & & \ddots & \\ & & & 1-h_{nn} \end{bmatrix}$: 멱등행렬

$H = (h_{ij})$ $(i = 1, \cdots, n; j = 1, \cdots, n)$ 여기서, h_{ij}: H 행렬의 (i, j)번째 원소

$\Rightarrow$ $Var(e_i) = (1-h_{ii})\sigma^2$

$Cov(e_i, e_j) = -h_{ij}\sigma^2$ 　　$(i \neq j)$

$$X_{(n\times p)} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & \cdots & X_{n,p-1} \end{bmatrix} = \begin{bmatrix} X_1^t \\ \vdots \\ X_n^t \end{bmatrix}$$

$\rightarrow$ $X_i = \begin{bmatrix} 1 & X_{i1} & \cdots & X_{i,p-1} \end{bmatrix}^{-1}$: $p-$벡터

$$\Rightarrow h_{ij} = (0 \quad 0 \quad 0 \quad 1 \quad 0 \quad \cdots \quad 0)X(X^tX)^{-1}X^t \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} = x_i^t(X^tX)^{-1}x_j$$

※ 잔차는 척도무관(scale-invariant)한 값이 아니다. 변수 단위에 무관한 값으로 변환시킬 필요가 있다. 척도무관 변환의 대표적인 것이 ‘표준화’인데 여기서는 두 가지 형태의 ‘표준화 잔차’를 소개한다.

4.2.1 내 표준화 잔차 (Internally Studentized Residual)

$$r_i = \frac{e_i - E(e_i)}{S.E.(e_i)} = \frac{e_i}{s\sqrt{1-h_{ii}}}$$

4.2.2 외 표준화 잔차 (Externally Studentized Residual)

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$$ 여기서, $s_{(i)}^2 = s^2 \bullet \frac{n-p-r_i^2}{n-p-1}$

※ r_i와 r_i^*의 관계: $r_i^* = r_i\sqrt{\frac{n-p-1}{n-p-r_i^2}}$