김은영

제5장 회귀모형의 선택

5.2 변수선택의 기준

5.2.1 결정계수 R_p^2

- $R_p^2 = \dfrac{SSR_p}{SST} = 1 - \dfrac{SSE_p}{SST}$　　여기서, SSE_p: 현재모형 하에서의 잔차제곱합
- R_p^2을 최대화하는 모형을 선택한다.
- 단점: p가 증가하면 R_p^2도 증가한다. 즉, 최대모형일 때 R_p^2가 최대가 된다.
  모든 설명변수가 전부 고려되는 최대모형이 항상 바람직한 것은 아니다.

5.2.2 수정된 결정계수 R_{ap}^2와 s_p^2

- $R_{ap}^2 = 1 - \dfrac{SSE_p/(n-p)}{SST/(n-1)} = 1 - \dfrac{s_p^2}{SST/(n-1)}$

5.2.3 Mallows' C_p

- 현재모형: $y = X\beta + \epsilon$
- 실제모형: $y = \mu + \epsilon$　　여기서, μ: 미지의 모수

- Mallows' C_p의 목적

: 실제모형과 현재모형 간의 평균오차제곱(Mean Squared Error; MSE)을 최소화

- $\hat{y}$: 현재모형 하에서의 적합치
- i-번째 적합치의 평균오차제곱(MSE): 분산과 편의(bias)의 제곱의 합으로 표현

$$E(\hat{y}_i - \mu_i)^2 = E\left[\left\{\hat{y}_i - E(\hat{y}_i)\right\} + \left\{E(\hat{y}_i) - \mu_i\right\}\right]^2 = Var(\hat{y}_i) + \left\{E(\hat{y}_i) - \mu_i\right\}^2$$

■ Γ_p에 대한 정의

$$\Gamma_p = \frac{MSE}{\sigma^2} = \sum_{i=1}^{n} \left[Var(\hat{y}_i) + \{E(\hat{y}_i) - \mu_i\}^2 \right] / \sigma^2$$

$$= \frac{1}{\sigma^2} \left[tr\{Cov(\hat{y})\} + \{E(\hat{y}) - \mu\}^t \{E(\hat{y}) - \mu\} \right]$$

$$= \frac{1}{\sigma^2} \left[p\sigma^2 + \mu^t (I-H)\mu \right]$$

$$= p + \frac{1}{\sigma^2} \mu^t (I-H)\mu$$

여기서, $\sum_{i=1}^{n} Var(\hat{y}_i) = tr\left[Cov(\hat{y})\right] = tr\left[Cov(Hy)\right] = tr\left[H \bullet Cov(y) \bullet H\right] = tr(H\sigma^2)$

$$= tr\left[X(X^tX)^{-1}X^t\sigma^2\right] = tr\left[(X^tX)^{-1}X^tX\sigma^2\right] = tr(I_p \bullet \sigma^2) = p\sigma^2$$

$$\sum_{i=1}^{n} \{E(\hat{y}_i) - \mu_i\}^2 = \{E(\hat{y}) - \mu\}^t \{E(\hat{y}) - \mu\} = \{E(Hy) - \mu\}^t \{E(Hy) - \mu\}$$

$$= \{-(I-H)\mu\}^t \{-(I-H)\mu\} = \mu^t (I-H)\mu$$

여기서, H, $(I-H)$: 멱등행렬

- Γ_p를 최소화하면 적합지의 분산과 편의를 동시에 작게 해주는 것
- Γ_p는 미지의 모수 σ^2, μ 등을 포함 → Γ_p에 대한 적절한 추정치가 필요하다.

■ Γ_p의 추정치를 구하기 위해

먼저 현재모형 하에서 오차제곱합 SSE_p의 기댓값을 구해보자.

$$E(SSE_p) = E(e^te) = E\left[\{(I-H)y\}^t \{(I-H)y\}\right] = E\left[y^t (I-H) y\right]$$

$$= \mu^t (I-H)\mu + tr\left[(I-H) \bullet I\sigma^2\right] \quad \text{여기서, } E(y^tAy) = \mu^tA\mu + tr(A\Sigma)$$

$$= \mu^t (I-H)\mu + (n-p)\sigma^2$$

$$\rightarrow \frac{E(SSE_p)}{\sigma^2} = \frac{\mu^t (I-H)\mu}{\sigma^2} + (n-p)$$

$$cf)\ \Gamma_p = p + \frac{1}{\sigma^2} \mu^t (I-H)\mu$$

$$\rightarrow \frac{E(SSE_p)}{\sigma^2} - (n-2p) = \frac{1}{\sigma^2} \{\mu^t (I-H)\mu + (n-p)\sigma^2 - (n-2p)\sigma^2\} = \frac{\mu^t (I-H)\mu}{\sigma^2} + p = \Gamma_p$$

$$\Rightarrow \frac{E(SSE_p)}{\sigma^2} - (n-2p) = \Gamma_p$$

■ Mallows' C_p

• $E(SSE_p)$와 σ^2 대신 SSE_p와 s^2를 사용해 Γ_p의 추정치로 변수선택의 기준으로 사용

• C_p를 최소화하는 모형을 선택

$$C_p = \frac{SSE_p}{s^2} - (n-2p)$$

여기서, SSE_p : 현재모형에 기반하여 계산, s^2 : 최대모형에 기반하여 계산

※ C_p의 첫 번째 항: 모형의 적합도. 설명변수가 많아질수록 작아진다.

두 번째 항: 설명변수가 많아질수록 커진다. 벌칙항.

⇒ Mallows' C_p: 적합도와 사용된 설명변수의 개수에 대한 적절한 타협

5.2.4 Allen's $PRESS_p$ (*PRediction Error Sum of Squares*)

■ R_p^2, R_{ap}^2, C_p : 관측된 자료들에 대한 현재모형의 적합도가 얼마나 좋은지

■ 정확도가 높은 예측이 회귀분석의 목적인 경우
적합도보다 예측도가 높은 모형을 선택해야 할 필요가 있다. → $PRESS_p$

■ Allen's $PRESS_p$

$$PRESS_p = \sum_{i=1}^{n} \left(y_i - \hat{y}_{i(i)}\right)^2$$

여기서, $\hat{y}_{i(i)}$: i-번째 관측치를 제외시킨 후 $(n-1)$개의 관측치로 회귀모형을 구한 다음, x_i에서 예측한 y값

$y_i - \hat{y}_{i(i)}$: 예측오차(prediction error)

■ $PRESS_p$는 n개의 자료를 나누어서 $(n-1)$개는 추정에 이용하고 나머지 한 개는 예측의 정확도 계산에 사용하는 것이다. *cf) n-fold cross validation*

■ $PRESS_p$의 값을 최소화하는 모형을 최적모형으로 선택하면 된다.

$$\hat{y}_{i(i)} = x_i^t \hat{\beta}_{(i)} = x_i^t \left[\hat{\beta} - \frac{(X^tX)^{-1} x_i e_i}{1-h_{ii}}\right] = \hat{y}_i - \frac{x_i^t (X^tX)^{-1} x_i e_i}{1-h_{ii}} = \hat{y}_i - \frac{h_{ii} e_i}{1-h_{ii}}$$

$$PRESS_p = \sum_{i=1}^{n}\left(y_i - \hat{y}_{i(i)}\right)^2 = \sum_{i=1}^{n}\left(y_i - \hat{y}_i + \frac{h_{ii}e_i}{1-h_{ii}}\right)^2 = \sum_{i=1}^{n}\left(e_i + \frac{h_{ii}e_i}{1-h_{ii}}\right)^2 = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$