제5장 회귀모형의 선택

5.3 변수선택의 방법

■ 변수선택의 과정에서 어떤 방법으로 계산할 것인가가 중요한 문제다.

- 모든 가능한 회귀 (all possible regression)
- 앞으로부터의 단계적 회귀 (forward stepwise regression)
- 앞으로부터의 선택 (forward selection)
- 뒤로부터의 제거 (backward elimination)

5.3.1 모든 가능한 회귀 (all possible regression)

■ 최대모형이 $(k-1)$개의 변수로 구성되어 있을 때 가능한 현재모형의 가짓수

$$\binom{k-1}{0}+\binom{k-1}{1}+\cdots+\binom{k-1}{k-1}=2^{k-1}$$

■ 모든 가능한 회귀: 2^{k-1}개의 모든 현재모형에 대해 어떤 변수선택의 기준(F-값을 최대화)을 적용시켜 최적모형을 구하는 것

■ 장점: 모든 가능한 경우에 대해 확인절차를 거치므로 가장 확실한 방법이 될 수 있다.

■ 단점: 계산량이 매우 많다.

5.3.2 앞으로부터의 단계적 회귀 (forward stepwise regression)

[단계 1]

현재모형이 한 개의 설명변수를 가지고 있다면(단순선형회귀모형) 기울기가 0인지에 대해 **부분 F-검정**을 실시한다.

$$F_i^* = \frac{MSR_i}{s_i^2} \qquad (i=1,\ \cdots,\ k-1)$$

- $\max\limits_{1\le i\le(k-1)} F_i^* > F_\alpha(1, n-2)$ → 해당 설명변수(X_i)를 추가, [단계 2]로 이동
- $\max\limits_{1\le i\le(k-1)} F_i^* < F_\alpha(1, n-2)$ → 여기서 중지. 영모형 채택

[단계 2]

[단계 1]에서 X_7이 추가되었다고 가정하자.

이제 X_7에 또 다른 설명변수가 추가된, 두 개의 설명변수에 대한 현재모형에 대해 **부분 F-검정**을 실시한다.

$$F_i^* = \frac{MSR(X_i | X_7)}{s^2(X_7,\ X_i)} = \left[\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right]^2 \qquad (i \neq 7)$$

- $\max_{i \neq 7} F_i^* > F_\alpha(1,\ n-3)$ → 해당 설명변수(X_i)를 추가, [단계 3]으로 이동
- $\max_{i \neq 7} F_i^* < F_\alpha(1,\ n-3)$ → 여기서 중지. 최종선택된 모형은 $y_i = \beta_0 + \beta_1 X_{i7} + \epsilon$

[단계 3]

[단계 2]에서 X_3가 추가되었다고 가정하자.

현재모형에는 X_7과 X_3가 있다.

[단계 3]에서는 이들 두 변수에 대해 **부분 F-검정**을 실시한다.

이미 [단계 2]에서 X_3에 대한 **부분 F-검정**이 실시되었으므로 [단계 3]에서는 X_7에 대한 **부분 F-검정**만 실시한다.

$$F_7^* = \frac{MSR(X_7 | X_3)}{s^2(X_3,\ X_7)}$$

- $F_7^* > F_\alpha(1,\ n-3)$ → X_3과 X_7은 모형에 남게 됨. [단계 4]로 이동
- $F_7^* < F_\alpha(1,\ n-3)$ → X_7을 제거. [단계 4]로 이동

[단계 4]

[단계 3]에서 X_7이 남아있게 된다고 하자.

이제 X_3과 X_7이 아닌 변수들에 대해 [단계 2]와 [단계 3]을 반복 적용한다.

$$F_i^* = \frac{MSR(X_i | X_3,\ X_7)}{s^2(X_i,\ X_3,\ X_7)} \qquad (i \neq 3,\ 7)$$

[예제 5.2]

5.3.3 앞으로부터의 선택 (forward selection)

- 이 방법은 **앞으로부터의 단계적 회귀**를 단순화시킨 것
- 한 번 선택된 변수는 제거되지 않는 것이 차이점이다.

5.3.4 뒤로부터의 제거 (backward elimination)

- 최대모형에서 출발한다.
- 가장 작은 F^* 값을 가지는 것이 기준치(F_α)보다 작으면 제거해 나간다.

$$F_i^* = \frac{MSR_i}{s_i^2} \qquad (i = 1,\ \cdots,\ k-1)$$

- $\min\limits_{1 \le i \le (k-1)} F_i^* < F_\alpha(1,\ n-2)$ → 해당 설명변수(X_i)를 제거
- $\min\limits_{1 \le i \le (k-1)} F_i^* > F_\alpha(1,\ n-2)$ → 여기서 중지