7강. 변수선택

◈ 담당교수 : 김성수 교수

■ 주요용어

용어	해설
다중공선성	두 설명변수 X_1과 X_2가 임의의 상수 c_0, c_1, c_2에 대하여 다음과 같은 선형관계가 성립하면 $c_1X_1 + c_2X_2 = c_0$ 두 변수 사이에 완벽한 공선성(exact collinearity)이 있다고 한다. 설명변수의 수가 2가 넘는 경우, 즉 $k>2$일 때, 설명변수 $X_1, \cdots, X_k$들이 임의의 상수 $c_0, c_1, \cdots, c_k$에 대하여 다음과 같은 관계가 성립하거나 또는 근사적으로 성립할 때 $c_1X_1 + c_2X_2 + \cdots + c_kX_k = c_0$ 이 설명변수들 사이에 다중공선성(multicollinearity)이 존재한다고 한다.
분산팽창인자 (variance inflation factor)	설명변수 X_h와 나머지 설명변수들간의 다중표본상관계수의 제곱을 R_h^2라고 할 때, $1/(1-R_j^2)$를 j 번째 분산팽창인자라고 한다. 일반적으로 k개의 VIF_j중 가장 큰 값이 5~10을 넘으면 다중공선성이 있다고 판정한다.
수정결정계수	수정결정계수(adjusted coefficient of determination), $\overline{R}_p^2$는 변수선택기준으로 널리 쓰이는 통계량으로, 결정계수를 보완한 통계량이다. 식은 다음과 같다. $\overline{R}_p^2 = 1 - \frac{SSE_p/(n-p-1)}{SST/(n-1)} = 1 - \left(\frac{n-1}{n-k-1}\right)(1-R_p^2)$
Mallows의 Cp통계량	변수선택기준으로 널리 쓰이는 통계량으로, 식은 다음과 같다. $C_p = \frac{SSE_p}{\hat{\sigma}^2} - (n-2p-2)$ C_p를 기준으로 k개의 변수 중에서 p개의 변수를 선택할 때 $C_p \approx p+1$이면서, 작은 값을 가지는 모형을 선택한다.

■ 연습문제

1. 설명변수 X_h와 나머지 설명변수들간의 다중표본상관계수의 제곱을 R_h^2라고 할 때 j번째 분산팽창인자를 구하는 식은 ?

정답 및 해설 : $1/(1-R_j^2)$

2. Mallows의 C_p 통계량을 구하는 식은 ?

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

정답 및 해설 :

3. 모든 가능한 회귀를 적합하고자 한다. 명령 (a) 는?

```
> install.packages("leaps")
> library(leaps)
> all.lm = ( a )(Y ~. , data=hald)
> (rs=summary(all.lm))
```

정답 및 해설 : (a) = regsubsets

■ 참고사이트

- 강명욱,김영일,안철환,이용구, 『회귀분석』, 율곡출판사, 1996.
- 박성현, 『회귀분석』 (제3판), 민영사, 2007.
- 이태림, 김성수, 성내경, 『통계패키지』, 방송대출판부, 2009.
- Faraway, J.J. (2002), Practical Regression and Anova Using R, (www.google.com에서 검색 후, pdf 파일로 다운받을 수 있음)
- Peter Dalgaard (2005), Introductory Statistics with R, Springer, (www.google.com에서 검색 후, pdf 파일로 다운받을 수 있음)