5강. 중회귀모형 (2)

◈ 담당교수 : 김성수 교수

■ 주요용어

용어	해설
중회귀모형	독립변수의 수가 k개인 중회귀모형은 일반적으로 다음과 같이 표현된다. $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i$ $\beta_0, \beta_1, \cdots, \beta_k$: 회귀계수 $\epsilon_i \sim N(0, \sigma^2)$ 이고 서로 독립 $i = 1, 2, \cdots, n$ 여기서 $\beta_0, \beta_1, \cdots, \beta_k$ 는 모집단의 회귀계수이고, ϵ_i 는 반응변수 Y_i를 측정할 때 발생하는 오차이다.
햇행렬	중회귀모형에서 추정값과 잔차벡터는 $\hat{Y} = Xb = X(X'X)^{-1}X'Y = HY$ $e = Y - \hat{Y} = Y - Xb = Y - X(X'X)^{-1}X'Y = (I - X(X'X)^{-1}X')Y = (I - H)Y$ 이 된다. 여기서 $H = X(X'X)^{-1}X'$ 을 나타내며, 이를 햇행렬(hat matrix) 이라 한다.
잔차	중회귀모형을 데이터에 적합시켜 얻은 추정된 회귀식의 값 $\hat{Y}_i$과 관찰값 Y_i의 차이를 잔차, $e_i = Y_i - \hat{Y}_i$ 라고 하며, 잔차벡터는 다음과 같이 간단히 표시된다. $e = Y - \hat{Y} = Y - Xb = Y - X(X'X)^{-1}X'Y = (I - X(X'X)^{-1}X')Y = (I - H)Y$
총제곱합	총변동은 총제곱합 (total sum of squares, SST)으로, $SST = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n(\bar{Y})^2 = Y'Y - n(\bar{Y})^2$ 이를 행렬로 표현하면 다음과 같다. $SST = Y'Y - n(\bar{Y})^2 = Y'(I - \frac{J}{n})Y$ 여기서 J 행렬은 모든 요소가 1인 $n \times n$ 행렬을 의미한다. SST에 대한 자유도는 $n-1$이다.
잔차제곱합	잔차제곱합(sum of squares due to residual errors, SSE)은 다음과 같다. $SSE = \sum (Y_i - \hat{Y}_i)^2 = \sum e_i^2 = e'e = [(I - H)Y]'[(I - H)Y] = Y'(I - H)'(I - H)Y = Y'(I - H)Y$ SSE에 대한 자유도는 $n-k-1$이다.

회귀제곱합	회귀제곱합 (sum of squares due to regression, SSR)은 $SSR = \sum(\hat{Y}_i - \bar{Y})^2 = \sum \hat{Y}_i^2 - n(\bar{Y})^2 = \hat{Y}'\hat{Y} - n(\bar{Y})^2$ 이고, $\hat{Y} = Xb$로부터 $SSR = b'X'Y - n(\bar{Y})^2 = Y'HY - Y'(\frac{J}{n})Y = Y'(H - \frac{J}{n})Y$ 이다. SSR의 자유도는 독립변수의 수와 같은 k가 된다.
결정계수	결정계(Coefficient of determination)수는 회귀방정식에 의하여 설명되는 변동 SSR이 총변동 SST에 비하여 어느 정도인가를 나타내 주는 값으로 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$ 이다.

■ 연습문제

1. $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$을 적합시켰을 때, $x_1 = 10,\ x_2 = 10$에서 E(y)를 95% 신뢰구간으로 추정하고자 한다.
R함수 (a) 는?

```
> pred.x = data.frame(X1=10, X2=10)
 > pc = ( a )(market2.lm, int="c", newdata=pred.x)
 > pc

       fit       lwr       upr
1 20.70503 19.95796  21.45209
```

```

```

정답 및 해설 : predict

2. 다음은 추가제곱합을 구하기 위한 R 분석 결과이다.

유의수준 0.05에서 검정할 때, 변수 X_2는 (X_1, X_3, X_4) 모형에 유의한 변수인지 검정하시오.

```
> h3.lm = lm(Y ~ X1 + X3 + X4, data=health)
> h4.lm = lm(Y ~ X1 + X2 + X3 + X4, data=health)
```

```
> anova(h3.lm, h4.lm)
Analysis of Variance Table

Model 1: Y ~ X1 + X3 + X4
Model 2: Y ~ X1 + X2 + X3 + X4
  Res.Df   RSS Df Sum of Sq    F Pr(>F)
1     26 20856
2     25 20551  1    304.62 0.3706 0.5482
```

정답 및 해설 : p-값 = 0.5482 > 0.05 이므로 귀무가설을 받아들인다. 즉, 변수 X_2는 유의하지 않다.

3. 추가변수그림을 그리고자 한다. R 패키지 car에서 이용하는 함수 (b)는 ?

```
> library(car)
> h4.lm = lm(Y ~ X1 + X2 + X3 + X4, data=health)
> ( b )(h4.lm)
```

정답 및 해설 : avPlots

■ 참고사이트

- 강명욱,김영일,안철환,이용구, 『회귀분석』, 율곡출판사, 1996.
- 박성현, 『회귀분석』 (제3판), 민영사, 2007.
- 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)
- R 사이트 바로가기
  https://www.r-project.org/
- R Studio 사이트 바로가기
  https://www.rstudio.com/