2.1.4 구간추정과 가설검정

2.1.3절의 성질 ①에 의해

$$Z = \frac{\hat{\beta}_i - \beta_i}{\sigma_\varepsilon \sqrt{c_{ii}}} \sim N(0, 1)$$

이고, 성질 ⑤에 의해

$$\chi^2 = \frac{(n-p-1)s^2}{\sigma_\varepsilon^2} \sim \chi^2(n-p-1)$$

이다. 또한 두 통계량 Z와 χ^2는 서로 독립이므로 다음의 통계량

$$T = \frac{Z}{\sqrt{\dfrac{\chi^2}{(n-p-1)}}} = \frac{\hat{\beta}_i - \beta_i}{s\sqrt{c_{ii}}} \tag{2-9}$$

는 자유도 $n-p-1$을 갖는 t분포를 따른다. 모수 β_i에 대한 신뢰구간 추정과 가설검정은 (2-9)의 T통계량에 기초한다.

(1) 신뢰구간 추정

모수 β_i에 대한 $100(1-\alpha)\%$ 신뢰구간은 다음과 같다.

$$\beta_i \pm t_{\frac{\alpha}{2}}(n-p-1)\, s\sqrt{c_{ii}}$$

단, $t_{\frac{\alpha}{2}}(n-p-1)$은 자유도 $n-p-1$을 갖는 t분포의 $100\left(1-\frac{\alpha}{2}\right)$ 백분위수이며 이 값은 부록 B의 표에 수록되어 있다.

(2) 개별 모수의 유의성검정

귀무가설 "$H_0: \beta_i = 0$" 에 대한 대립가설 "$H_1: \beta_i \neq 0$" 의 검정통계량은

$$T = \frac{\hat{\beta}_i}{s\sqrt{c_{ii}}}$$

이다. 표본으로부터 계산된 T통계량의 값이 t라면 유의수준 α의 검정 규칙은 아래와 같다. 이때 H_0를 기각하는 경우는 모수 β_i가 통계적으로 유의하여 독립변수 X_i가 모형에 필요함을 의미한다.

(i) $|t| > t_{\frac{\alpha}{2}}(n-p-1)$이면 H_0를 기각한다.
(ii) $|t| \leq t_{\frac{\alpha}{2}}(n-p-1)$이면 H_0를 기각하지 못한다.

대부분의 통계패키지는 **유의확률**(significant probability) 혹은 **$p-$값**($p-$value)을 출력해 준다. 예를 들어 개별 모수의 검정에서 $p-$값은

$$p-\text{값} = P(T > |t|)$$

으로 계산된다. 따라서 통계패키지에서 출력되는 $p-$값을 이용한 유의수준 α에서의 검정 규칙은 다음과 같다.

(i) $p-$값$<\alpha$이면 H_0를 기각한다.
(ii) $p-$값$\geq\alpha$이면 H_0를 기각하지 못한다.

(3) 회귀모형의 유의성검정

귀무가설 "$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$"의 의미는 p개의 모든 독립변수들이 종속변수를 설명하는 데 아무런 영향을 주지 못하므로 결국 회귀모형이 유의하지 않음을 의미한다. 이에 대한 대립가설은 "H_1: 적어도 하나의 $\beta_i \neq 0$"이다. 이러한 회귀모형의 유의성검정은 2.1.3절의 성질 ④에서 설명한 분산분석표에 근거하여 수행된다. MSR과 MSE가 서로 독립인 χ^2-분포를 따르므로 F가 $F(p, n-p-1)$분포를 따름을 이용하여 검정통계량 $F = \frac{\text{MSR}}{\text{MSE}}$의 관측값이 f라면 유의수준 α에서의 검정 규칙은 다음과 같다.

(i) $f > F_{\alpha}(p, n-p-1)$이면 H_0를 기각한다.

(ii) $f \leq F_{\alpha}(p, n-p-1)$이면 H_0를 기각하지 못한다.

단, $F_{\alpha}(p, n-p-1)$은 자유도 p와 $n-p-1$을 갖는 F분포의 $100(1-\alpha)$ 백분위수로서 이 값은 부록 B의 표에 수록되어 있다.

(4) 모수집합의 유의성검정

모형에 두 개 이상의 설명변수들이 추가될 때 추가된 변수들의 종속변수에 대한 설명력이 통계적으로 유의한지를 검정할 필요가 있다. 다음의 두 모형을 생각해 보자.

$$(\text{축소모형}) : Z_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_q X_{tq} + \varepsilon_t$$
$$(\text{완전모형}) : Z_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \cdots + \beta_q X_{tq} + \beta_{q+1} X_{t,q+1} + \cdots + \beta_p X_{tp} + \varepsilon_t$$

완전모형은 축소모형을 포함하며, 축소모형에 없는 $p-q$개의 변수들 $X_{q+1}, X_{q+2}, \cdots, X_p$를 더 포함한다. 이처럼 $p-q$개의 변수들을 추가함으로써 생기는 회귀제곱합의 증가분이 통계적으로 유의하다면 축소모형에 이러한 변수들을 추가할 수 있을 것이다.

먼저 축소모형에 대한 모수를 $\boldsymbol{\beta}(1) = (\beta_0(1), \beta_1(1), \cdots, \beta_q(1))'$이라 놓으면 $\boldsymbol{\beta}(1)$의 LSE는

$$\hat{\boldsymbol{\beta}}(1) = (\hat{\beta}_0(1), \hat{\beta}_1(1), \cdots, \hat{\beta}_q(1))' = (\boldsymbol{X}'_{(1)} \boldsymbol{X}_{(1)})^{-1} \boldsymbol{X}'_{(1)} \boldsymbol{Z}$$

이다. 여기서 $\boldsymbol{X}_{(1)}$은 축소모형에 대한 디자인행렬(design matrix)이다. 축소모형의 회귀제곱합은

$$\text{SSR}(X_1, X_2, \cdots, X_q) = \hat{\boldsymbol{\beta}}(1)' \boldsymbol{X}'_{(1)} \boldsymbol{Z} - n\overline{Z}^2$$

이고, 여분의 변수 $X_{q+1}, X_{q+2}, \cdots, X_p$를 추가함으로써 생기는 추가제곱합(extra sum of squares)은

$$
\begin{aligned}
&\mathrm{SSR}(X_{q+1}, X_{q+2}, \cdots, X_p \mid X_1, X_2, \cdots, X_q) \\
&\quad = \mathrm{SSR}(X_1, X_2, \cdots, X_p) - \mathrm{SSR}(X_1, X_2, \cdots, X_q) \\
&\quad = \hat{\boldsymbol{\beta}} X'Z - \hat{\boldsymbol{\beta}}(1)'X'_{(1)}Z
\end{aligned}
$$

이다. 단, $\hat{\boldsymbol{\beta}}$는 완전모형에 대한 모수 $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$의 LSE이다. 이제 귀무가설 "$H_0 : \beta_{q+1} = \beta_{q+2} = \cdots = \beta_p = 0$"에 대한 대립가설 "$H_1$: 적어도 하나의 $\beta_i \neq 0$ $(i = q+1, q+2, \cdots, p)$"의 검정을 위한 검정통계량은

$$
F = \frac{\mathrm{SSR}(X_{q+1}, X_{q+2}, \cdots, X_p \mid X_1, X_2, \cdots, X_q)/(p-q)}{\mathrm{SSE}/(n-p-1)}
$$

이고, F통계량의 관측값이 f라면 유의수준 α에서의 검정 규칙은 다음과 같다.

(i) $f > F_\alpha(p-q, n-p-1)$이면 H_0를 기각한다.
(ii) $f \leq F_\alpha(p-q, n-p-1)$이면 H_0를 기각하지 못한다.

2.1.5 잔차분석

잔차분석(residual analysis)은 잔차들을 분석하여 잠정적으로 설정된 회귀모형이 과연 자료를 잘 설명하는지를 판단하는 분석 단계이다. 잔차는 오차의 추정값으로 생각할 수 있으므로 회귀모형 설정 시 오차항에 대한 가정에 잔차가 잘 부합되는지를 검토한다. 오차항에 대한 가정이 만족되지 않아 회귀모형에 수용되지 못하고 잔차에 남겨둔 정보가 있다면 이를 밝혀 잠정모형을 갱신한다. 반대로 오차항에 대한 가정이 위배되지 않는다면 잠정모형을 최종 예측모형으로 사용할 수 있다.

중회귀모형 $\boldsymbol{Z} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$에서 오차항 $\boldsymbol{\varepsilon}$은 $N_n(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I})$분포를 따르고 각 ε_i들은 상관관계가 없어 서로 독립이나 잔차벡터 $\boldsymbol{e} = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = (e_1, e_2, \cdots e_n)'$의 공분산행렬 $\mathrm{Cov}(\boldsymbol{e})$는 대각행렬이 아니므로 e_i들 간에는 상관관계가 존재하며, 그것은 단지 디자인행렬 $\boldsymbol{X}$에만 의존함을 알 수 있다. 그러나 e_i들 간의 상관관계는 추정해야 할 모수