Multiple Regression

회귀직선의 유의성 검정

회귀직선의 유의성 검정 (F-test)

- 가설 : $H_0 : \beta_1 = \cdots = \beta_p = 0 \ vs. \ H_1$: not H_0
- 검정통계량

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-(p+1)))} \sim_{H_0} F(p, n-p-1)$$

- 검정통계량의 관측값 : f
- 유의수준 α에서의 기각역 : $f \geq F_\alpha(p, n-p-1)$
- 유의확률 $= P(F \geq f)$

회귀직선의 유의성 검정

■ 회귀직선의 유의성 검정을 위한 분산분석표

요인	제곱합	자유도(df)	평균제곱(MS)	f	유의확률
회귀	SSR	p	$MSR = \dfrac{SSR}{p}$	$f = \dfrac{MSR}{MSE}$	$P(F \geq f)$
잔차	SSE	$n-(p+1)$	$MSE = \dfrac{SSE}{n-p-1}$		
계	SST	$n-1$			

회귀계수에 대한 추론

- $\beta_i, \quad i = 0, 1, \ldots, p$ 에 대한 추론
  - $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} Y$
  - $\mathrm{Var}(\hat{\beta}_i) = c_{ii}\sigma^2 (i = 0, 1, 2, \ldots, p)$

    c_{ii} : $(i+1)$th diagonal elements of $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}$
  - $s.e.(\hat{\beta}_i) = \sqrt{c_{ii} MSE}$
  - $\hat{\beta}_i$ 의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{\beta}_i \pm t_{\alpha/2}(n-p-1)\sqrt{c_{ii} MSE}$$

회귀계수에 대한 추론

- $\beta_i,\ \ i = 0, 1, \ldots, p$ 에 대한 추론
  - 가설검정 : $H_0 : \beta_i = \beta_i^0$
  - 검정통계량 : $T = \dfrac{\hat{\beta}_i - \beta_i^0}{\hat{\sigma}\sqrt{c_{ii}}} \sim_{H_0} t(n-p-1)$, 관측값 : t_0

대립가설	유의확률	유의수준 α 기각역
$H_1 : \beta_i > \beta_i^0$	$P(T \geq t_0)$	$t_0 \geq t_\alpha(n-p-1)$
$H_1 : \beta_i < \beta_i^0$	$P(T \leq t_0)$	$t_0 \geq t_\alpha(n-p-1)$
$H_1 : \beta_i \neq \beta_i^0$	$P(\|T\| \geq \|t_0\|)$	$\|t_0\| \geq t_{\alpha/2}(n-p-1)$

회귀계수에 대한 추론

$\boldsymbol{\beta}$ 의 선형함수

$$\mathbf{q}^\top \boldsymbol{\beta} = (q_0, q_1, \ldots, q_k) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} = q_0\beta_0 + q_1\beta_1 + \cdots + q_k\beta_k$$

- $\mathrm{Var}(\mathbf{q}^\top \hat{\boldsymbol{\beta}}) = \mathbf{q}^\top \mathrm{Var}(\hat{\boldsymbol{\beta}})\mathbf{q} = \mathbf{q}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\mathbf{q}\sigma^2$
- $s.e.(\mathbf{q}^\top \hat{\boldsymbol{\beta}}) = \sqrt{\mathbf{q}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\mathbf{q}\sigma^2}$
- $\mathbf{q}^\top \boldsymbol{\beta}$의 $100(1-\alpha)\%$ 신뢰구간

$$\mathbf{q}^\top \hat{\boldsymbol{\beta}} \pm t_{\alpha/2}(n-p-1)\sqrt{\mathbf{q}^\top (X^\top X)^{-1}\mathbf{q} \cdot MSE}$$

회귀계수에 대한 추론

$\boldsymbol{\beta}$ 의 선형함수

- 가설검정 : $H_0 : \mathbf{q}^\top \boldsymbol{\beta} = c$
- 검정통계량 : $T = \dfrac{\mathbf{q}^\top \boldsymbol{\beta} - c}{\hat{\sigma}\sqrt{\mathbf{q}^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\mathbf{q}}} \sim_{H_0} t(n-p-1)$
- 관측값 : t_0

대립가설	유의확률	유의수준 α 기각역
$H_1 : \beta_i > \beta_i^0$	$P(T \geq t_0)$	$t_0 \geq t_\alpha(n-p-1)$
$H_1 : \beta_i < \beta_i^0$	$P(T \leq t_0)$	$t_0 \geq t_\alpha(n-p-1)$
$H_1 : \beta_i \neq \beta_i^0$	$P(\|T\| \geq \|t_0\|)$	$\|t_0\| \geq t_{\alpha/2}(n-p-1)$

평균반응예측

- $\boldsymbol{X} = \boldsymbol{x}_0 = (x_{01}, \ldots, x_{0p})^\top$가 주어졌을 때 평균반응의 예측
  - 평균반응 : $\mu_0 = E(y|\boldsymbol{x}_0) = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} = \boldsymbol{x}_0^\top \boldsymbol{\beta}$
  - 평균반응 추정량 : $\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p} = \boldsymbol{x}_0^\top \hat{\boldsymbol{\beta}}$
  - $\mathrm{Var}(\hat{\mu}_0) = \boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0 \sigma^2$
  - $\widehat{s.e.(\hat{\mu}_0)} = \sqrt{\boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0 MSE}$
  - $\hat{\mu}_0$의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{\mu}_0 \pm t_{\alpha/2}(n-p-1) s.e.(\hat{\mu}_0)$$

예측

- $\boldsymbol{X} = \boldsymbol{x}_0$가 주어졌을 때 $y = y_0$ 예측
  - $y_0 = \beta_0 + \beta_1 x_{01} + \cdots + \beta_p x_{0p} + \epsilon_0$
  - 예측값 : $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \cdots + \hat{\beta}_p x_{0p}$
  - $\mathrm{Var}(\hat{y}_0) = \left(1 + \boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0\right) \sigma^2$
  - $\widehat{s.e.(\hat{y}_0)} = \sqrt{\left(1 + \boldsymbol{x}_0^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{x}_0\right) MSE}$
  - $\hat{y}_0$의 $100(1-\alpha)\%$ 신뢰구간

$$\hat{\mu}_0 \pm t_{\alpha/2}(n-p-1) s.e.(\hat{y}_0)$$

Partial F test

- Reduced model(RM) vs. Full model(FM)

$$\text{FM} \;:\; y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \cdots + \beta_p x_{ip} + \epsilon_i$$

$$\text{RM} \;:\; y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \epsilon_i$$

- extra sum of squares : $SSR_{FM} - SSR_{RM}$
  - SSR_{RM} : $\beta_1, \ldots, \beta_q$ 에 의한 회귀제곱합
  - SSR_{FM} : $\beta_1, \ldots, \beta_q, \ldots, \beta_p$ 에 의한 회귀제곱합
  - $SSR_{FM} - SSR_{RM}$: FM에서 $\beta_1, \ldots, \beta_q$ 이 주어졌을 때, $\beta_{q+1}, \ldots, \beta_p$ 에 의한 회귀제곱합

Partial F test

- 가설 : $H_0 : \beta_{q+1} = \cdots = \beta_p = 0 \ vs. \ H_1$: not H_0
- 검정통계량

$$F = \frac{(SSR_{FM} - SSR_{RM})/(p-q)}{SSE_{FM}/(n-p-1)} \sim_{H_0} F(p-q, n-p-1)$$

- 검정통계량의 관측값 : f
- 유의수준 α에서의 기각역 : $f \geq F_\alpha(p-q, n-p-1)$
- 유의확률 $= P(F \geq f)$

Testing the general linear hypothesis

- General Linear Hypothesis

$$H_0 : H\boldsymbol{\beta} = 0 \;\; vs. \;\; H_1 : \text{not } H_0$$

- $FM : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
- $H_0 : \beta_1 = \beta_3, \quad i.e. \; H = (0, 1, 0, -1)$

$$\begin{aligned} RM : y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_1 x_3 + \epsilon \\ &= \beta_0 + \beta_1 (x_1 + x_3) + \beta_2 x_2 + \epsilon \\ &= \gamma_0 + \gamma_1 z_1 + \gamma_2 x_2 + \epsilon \end{aligned}$$

Testing the general linear hypothesis

- General Linear Hypothesis

$$H_0 : H\boldsymbol{\beta} = 0 \;\; vs. \;\; H_1 : \text{not } H_0$$

- $FM : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$
- $H_0 : \beta_1 = \beta_3, \;\; \beta_2 = 0 \;\; i.e. \; H = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$

$$\begin{aligned} RM : y &= \beta_0 + \beta_1 x_1 + \beta_1 x_3 + \epsilon \\ &= \beta_0 + \beta_1 (x_1 + x_3) + \epsilon \\ &= \gamma_0 + \gamma_1 z_1 + \epsilon \end{aligned}$$

Testing the general linear hypothesis

General Linear Hypothesis

$$H_0 : H\boldsymbol{\beta} = 0 \;\; vs. \;\; H_1 : \text{not } H_0$$

- $H : r \times (p+1)$ matrix, $rank(H) = r$
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$
- 검정통계량

$$F = \frac{(SSR_{FM} - SSR_{RM})/(r)}{SSE_{FM}/(n-p-1)} \sim_{H_0} F(r, n-p-1)$$

Example

- 예제 7.2
  - 다음의 데이터

y	x_1	x_2	x_3
8	2	1	4
10	-1	2	1
9	1	-3	4
6	2	1	2
12	1	4	6

에 절편 없는 중회귀모형, $y_j = \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \epsilon_j$ 이 적절하다고 판단되어 이를 적합시켰다. $H_0 : \beta_1 = \beta_2 + 4$에 대한 검정을 수행하시오.

Example

- 예제 7.2

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} = \begin{pmatrix} 11 & 3 & 21 \\ 2 & 31 & 20 \\ 21 & 20 & 73 \end{pmatrix}^{-1} = \begin{pmatrix} 0.2145 & 0.0231 & -0.0680 \\ 0.0231 & 0.0417 & -0.0181 \\ -0.0680 & -0.0181 & 0.0382 \end{pmatrix}$$

$$\boldsymbol{y}^\top \boldsymbol{y} = 425, \qquad X^\top \boldsymbol{y} = \begin{pmatrix} 39 \\ 55 \\ 162 \end{pmatrix}$$

Example

- 예제 7.2

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} X^\top \boldsymbol{y} = \begin{pmatrix} -1.39 \\ 0.27 \\ 2.54 \end{pmatrix}$$

Table: 분산분석표

요인	제곱합	자유도
회귀	SSR=372.9	3
잔차	SSE=52.1	2
계	SST=425.0	5