제 5 장

일원 배치법과 분산과 불완전 계수 행렬

5.1 서론

선형모형에서 설계행렬 (design matrix) X가 완전계수 (full rank) 행렬일 때 회귀계수의 추정치는 최소제곱법에서 구해진 정규방정식의 유일한 해로 구해진다.

$$(\mathrm{X}^t\mathrm{X})\boldsymbol{\beta} = \mathrm{X}^t\mathrm{y} \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathrm{X}^t\mathrm{X})^{-1}\mathrm{X}^t\mathrm{y}$$

그러나 여러 가지 실험이나 자료의 형태에서 설계행렬 X의 계수가 완전하지 않을 때 (less than full rank) 가 있으며

$$rank(\mathrm{X}) = r < p = \text{ number of columns in X}$$

이러한 경우에는 정규방정식에서 유일한 해가 존재하지 않는다. 이 장에서는 이러한 경우의 해결 방법을 알아 보고 일원 배치법에 어떻게 적용되는 자를 알아본다.

5.2 불완전 계수 행렬의 계수 추정

설계행렬 X의 계수가 완전하지 않을 때 회귀 계수를 추정하기 위한 방법으로서 다음과 같은 세 가지 방법이 있다.

1. 모수의 재조정 (reparameterization)

   X의 계수가 완전하지 않을 때 설계행렬의 열을 다시 구성하여 계수를 완전하게 하는 방법이 있다. 즉 $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2)$ 으로 표시하고 X_1 을 $n \times r$ $(r < p)$ 라고 하며 어떤 행렬 F가 존재하여 $\mathrm{X}_2 = \mathrm{X}_1\mathrm{F}$ 의 관계를 가진다고 가정하자. 이러한 관계는 X_2 의 열들이 X_1 의 열들의 선형결합으로 표현될 수 있다는 것을 의미한다. 이러한 경우에 선형모형은 다음과 같이 표현될 수 있다.

$$\mathrm{y} = \mathrm{X}\boldsymbol{\beta} + \mathrm{e} = \mathrm{X}_1(\mathrm{I}, \mathrm{F})\boldsymbol{\beta} + \mathrm{e} = \mathrm{X}_1\boldsymbol{\alpha} + \mathrm{e}$$

여기서 새롭게 조정된 계수 $\boldsymbol{\alpha}$와 처음의 계수 $\boldsymbol{\beta}$는 다음과 같은 관계가 있다.

$$\boldsymbol{\alpha} = (\mathrm{I}, \mathrm{F})\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$$

따라서 새롭게 구성된 선형모형 $\mathrm{y} = \mathrm{X}_1\boldsymbol{\alpha} + \mathrm{e}$에서 새로운 계수의 추정치는 $\hat{\boldsymbol{\alpha}} = (\mathrm{X}_1^t\mathrm{X}_1)^{-1}\mathrm{X}_1^t\mathrm{y}$ 이다.

2. 부가 조건의 이용 회귀계수에 부가 조건(side condition)을 주면 유일한 계수의 추정치를 구할 수 있다. 즉 $(p-r)\times p$ 행렬 H를 고려하고 $\mathrm{H}\boldsymbol{\beta} = 0$이라는 부가조건을 가정하자. 즉 모든 $\boldsymbol{\eta} = R(\mathrm{X})$에 대하여 $\boldsymbol{\eta} = \mathrm{X}\boldsymbol{\beta}$와 $\mathrm{H}\boldsymbol{\beta} = 0$를 만족하는 $\boldsymbol{\beta}$는 유일하게 존재한다.

$$\begin{pmatrix} \boldsymbol{\eta} \\ 0 \end{pmatrix} = \begin{pmatrix} \mathrm{X} \\ \mathrm{H} \end{pmatrix} \boldsymbol{\beta} \equiv \mathrm{G}\boldsymbol{\beta}, \quad \text{with} \quad \mathrm{G} = \begin{pmatrix} \mathrm{X} \\ \mathrm{H} \end{pmatrix}$$

이러한 부가 조건 $\mathrm{H}\boldsymbol{\beta} = 0$과 정규방정식 $(\mathrm{X}^t\mathrm{X})\boldsymbol{\beta} = \mathrm{X}^t\mathrm{y}$를 동시에 만족하는 유일한 해를 구하고 이를 최소제곱추정량으로 한다. 이러한 부가 조건을 주는 방법은 분산분석을 이용하는 여러 가지 선형 모형 (예: 일원 배치법)에 자주 사용된다.

3. 일반화 역함수의 이용

X의 계수가 완전하지 않을 때 일반화 역행렬 (generalized inverse matrix)를 이용하면 회귀계수의 추정치를 구할 수 있다.

여기서 $m \times n$ 행렬 A의 일반화 역행렬 A^-는 다음을 만족하는 행렬이다.

$$\mathrm{A} = \mathrm{A}\mathrm{A}^-\mathrm{A}$$

일반화 역행렬은 일반적으로 유일하지 않다. A가 정방행렬이고 정칙행렬일 때 유일하게 존재하며 $\mathrm{A}^- = \mathrm{A}^{-1}$이다. 정규방정싱의 좌변에 $\mathrm{X}^t\mathrm{X}(\mathrm{X}^t\mathrm{X})^-$를 곱하면

$$\mathrm{X}^t\mathrm{X}(\mathrm{X}^t\mathrm{X})^-\mathrm{X}\mathrm{y} = \mathrm{X}^t\mathrm{X}(\mathrm{X}^t\mathrm{X})^-\mathrm{X}^t\mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{X}^t\mathrm{X}\hat{\boldsymbol{\beta}} = \mathrm{X}^t\mathrm{y}$$

이므로 $\hat{\boldsymbol{\beta}} = (\mathrm{X}^t\mathrm{X})^-\mathrm{X}^t\mathrm{y}$는 정규방정식의 해가 된다. 앞에서 언급하였듯이 일반화 역함수를 이용한 계수의 추정량은 유일하지 않다. 그러나 반응변수의 추정량 $\hat{\mathrm{y}} = \mathrm{X}\hat{\boldsymbol{\beta}}$는 추정된 계수에 관계없이 유일하다.

5.3 추정가능한 계수의 함수

선형모형의 설계행렬이 불완전하면 반응변수의 기대값 $\boldsymbol{\beta} = E(\mathrm{y})$의 유일한 추정치 $\mathrm{X}\hat{\boldsymbol{\beta}}$가 존재하지 않는다. 이러한 사실은 계수 $\boldsymbol{\beta}$의 모든 면을 추정할 수 없다는 것을 의미한다.

회귀계수의 선형결합 $\psi = \mathrm{c}^t\boldsymbol{\beta}$를 고려하고 만약에 이에 대한 불편 선형추정량이 존재한다면 $\psi = \mathrm{c}^t\boldsymbol{\beta}$를 추정가능(estimable)하다고 말한다. 즉, y의 선형 결합 $\mathrm{a}^t\mathrm{y}$가 있어서 $E(\mathrm{a}^t\mathrm{y}) = \psi$이면 ψ를 추정가능하다고 한다.

추정가능한 계수의 함에 대한 정의를 보면 다음과 같은 관계를 쉽게 알 수 있다.

$$E(\mathrm{a}^t\mathrm{y}) = \mathrm{a}^t E(\mathrm{y}) = \mathrm{a}^t\mathrm{X}\boldsymbol{\beta} = \mathrm{c}^t\boldsymbol{\beta}$$

따라서 선형결합 $\psi = \mathrm{c}^t\boldsymbol{\beta}$이 추정가능하려면 $\mathrm{c}^t = \mathrm{a}^t\mathrm{X}$가 성립함을 알 수 있다. 따라서 $\mathrm{c}^t = \mathrm{a}^t\mathrm{X}$ 이면 $\psi = \mathrm{c}^t\boldsymbol{\beta}$는 추정가능하다.

5.3.1 일원배치법

이 절에서는 두 개 이사의 집단의 평균을 비교할 수 있는 통계적 방법인 일원배치법(one-way classification or one-way model)에 대한 추론을 고려해보자.

서로 다른 모집단의 갯수를 I라고 하고 각 모집단의 평균의 차이에 대하여 관심이 있다. 따라서 각각의 모집단에서 표본을 추출하는데 그 크기를 $J_1,\ J_2, \ldots, J_I$라고 하자. i번째 머집단에서 추출한 j번째 관측값을 y_{ij}라고 한다면 다음과 같은 일원배치법 모형을 생각할 수 있다.

$$y_{ij} = \mu + \alpha_i + e_{ij}, \quad i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J_i \tag{5.1}$$

예를 들어 서로다른 보집단은 I개의 서로 다른 처리(treatments)를 임의로 적용한 개체들의 집단이며 μ는 모든 개체들의 공통적인 모평균으러 생각할 수 있고 α_i는 i번째 처리의 효과로 생각할 수 있다. e_{ij}는 처리로서 설명할 수 없는 오차를 나타내면 이들은 서로 독립이고 평균이 0이고 분산이 σ^2인 정규분포를 따른다고 가정하자. ($e_{ij} \sim N(0, \sigma^2)$)

일원배치 모형 (5.1)를 선형모형 $\mathrm{y} = \mathrm{X}\boldsymbol{\beta} + \mathrm{e}$로 표시하면 각 행렬은 다음과 같이 표시 된다.

$$\begin{bmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1J_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2J_2} \\ \vdots \\ y_{I1} \\ y_{I2} \\ \vdots \\ y_{IJ_I} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & . & . & 0 \\ 1 & 1 & 0 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & 0 & 1 & . & . & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & . & . & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \\ 1 & 0 & 0 & . & . & 1 \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & . & . & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_I \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1J_1} \\ e_{21} \\ e_{22} \\ \vdots \\ e_{2J_2} \\ \vdots \\ e_{I1} \\ e_{I2} \\ \vdots \\ e_{IJ_I} \end{bmatrix}$$

위의 선형모형에서 일반적인 최소제곱법에 의한 회귀계수 $\boldsymbol{\beta}$를 구할 수 있는 정규방정식 $(\mathrm{X}^t\mathrm{X})\hat{\boldsymbol{\beta}} = \mathrm{X}^t\mathrm{y}$을 구하면 다음과 같다.

$$\begin{pmatrix} n & J_1 & J_2 & \cdot & \cdot & J_I \\ J_1 & J_1 & 0 & \cdot & \cdot & 0 \\ J_2 & 0 & J_2 & \cdot & \cdot & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ J_I & 0 & 0 & \cdot & \cdot & J_I \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot\alpha_I \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^I J_i \bar{y}_{i.} \\ J_1 \bar{y}_{1.} \\ J_2 \bar{y}_{2.} \\ . \\ . \\ J_I \bar{y}_{I.} \end{pmatrix}$$

위의 계획행렬 X 또는 정규방정식의 $\mathrm{X}^t\mathrm{X}$에서 행렬이 full rank가 아님을 알수있다.

$$rank(\mathrm{X}) = rank(\mathrm{X}^t\mathrm{X}) = I$$

정규방정식은 다음과 같이 나타낼 수 있다.

$$n\hat{\mu} + \sum_{i=1}^I J_i \hat{\alpha}_i = \sum_{i=1}^I J_i \bar{y}_{i.}, \quad J_i\hat{\mu} + J_i\hat{\alpha}_i = J_i\bar{y}_{i.}, \quad i = 1, 2, \ldots, I$$

위 정규 방정식에서 α_i의 최소제곱추정량은 언제나 $\hat{\alpha}_i = \bar{y}_{i.} - \hat{\mu}$의 형태로 나타나며 어떤 특정한 일반화 역행렬를 이용하면 추정량을 구할 수 있다 ($\hat{\boldsymbol{\beta}} = (\mathrm{X}^t\mathrm{X})^-\mathrm{X}^t\mathrm{y}$). 또한 앞절에서 본 것처럼 부가조건 $\mathrm{H}\boldsymbol{\beta} = 0$을 주면 유일한 계수의 추정치를 구할 수 있다.

몇 개의 보편적인 부가조건과 대응하는 계수의 추정치를 알아보자.

1. $\sum_{i=1}^{I} \hat{\alpha}_i = 0$ (sum-to-zero condition, `R package`의 default condition).

   Sum-to-zero 조건에서는 계수의 추정치가 다음과 같이 주어진다.

$$\hat{\mu} = \sum_i \bar{y}_{i.}/I, \quad \hat{\alpha}_i = \bar{y}_{i.} - \hat{\mu}, \quad i = 1, 2, \ldots, I$$

2. $\hat{\alpha}_I = 0$ (set-to-zero condition, `SAS`의 default condition).

   Sum-to-zero 조건에서는 계수의 추정치가 다음과 같이 주어진다.

$$\hat{\mu} = \bar{y}_{I.}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{I.}, \quad i = 1, 2, \ldots, I-1, \quad \hat{\alpha}_I = 0$$

3. $\sum_{i=1}^{I} J_i \hat{\alpha}_i = 0$

   위의 조건에서는 계수의 추정치가 다음과 같이 주어진다.

$$\hat{\mu} = \sum_i J_i \bar{y}_{i.}/n = \bar{y}_{..}, \quad \hat{\alpha}_i = \bar{y}_{i.} - \hat{\mu}, \quad i = 1, 2, \ldots, I$$

여기서 주목할 사항은 어떠한 부가 조건이나 일반화 역행렬에 대해서도 예측치 $\hat{y}_{ij}$는 변하지 않는다.

$$\hat{y}_{ij} = \hat{\mu} + \hat{\alpha}_i \equiv \bar{y}_{i.}$$

따라서 $SSE = \sum_i \sum_j (y_{ij} - \hat{y}_{ij})^2$ 도 변하지 않는다.

일원배치 모형 (5.1)에서 추정가능한 모수의 형태를 보면

$$\begin{aligned}
\psi &= \mathrm{a}^t \mathrm{X} \boldsymbol{\beta} \\
&= \sum_{i=1}^{I} \sum_{j=1}^{J_i} a_{ij}(\mu + \alpha_i) \\
&= \sum_{i=1}^{I} c_i(\mu + \alpha_i) \quad (c_i = \sum_{j=1}^{J_i} a_{ij}) \\
&= c_0\mu + c_1\alpha_1 + c_2\alpha_2 + \cdots + c_I\alpha_I \quad (c_0 = \sum_{i=1}^{I} c_i)
\end{aligned}$$

따라서 $\mu_i \equiv \mu + \alpha_i$ 라고 하면 추정가능한 함수의 형태는 $\psi = \sum_{i=1}^{I} c_i(\mu + \alpha_i) = \sum_{i=1}^{I} c_i\mu_i$ 이며 이는 $\hat{\psi} = \sum_{i=1}^{I} c_i\bar{y}_{i.}$ 형태로 추정된다. 예를 들어 $\psi = \alpha_1 - \alpha_2$ 이면 $c_0 = 0, \ c_1 = 1, \ c_2 = -1, \ , c_3 = 0, \ldots, c_I = 0$ 으로 추정가능한 함수의 조건을 만족하며 이는 $\hat{\psi} = \bar{y}_{1.} - \bar{y}_{2.}$ 으로 추정된다. 반면 $\psi = \alpha_1 + \alpha_2$ 이면 $c_1 = 1, \ c_2 = 1, \ , c_3 = 0, \ldots, c_I = 0$ 이너마 하며 따라서 $c_0 = 2 \neq 0$ 으로 추정가능한 함수의 조건을 만족하지 않는다.

위에서 각 수준에 대한 모수 α_i 의 선형 결합 $\psi = \sum_{i=1}^{I} c_i\alpha_i$ 는 $\sum_{i=1}^{I} c_i = 0$ 을 만족할 때 추정가능한 함수이다. 이러한 조건을 만족하는 함수를 대비함수 (contrast) 라고 한다. 일원배치에서 각

수준의 차이 $\psi = \alpha_i - \alpha_j$ 는 대비함수이며 각 수분의 차이에 대한 추론에 중요한 모수이다. 대비함수는 $\hat{\psi} = \sum_{i=1}^{I} c_i \bar{y}_{i.}$ 으로 추정되며 그 분산은 $\sigma^2 \sum_{i=1}^{I} c_i^2/J_i$ 이다. 따라서 $100(1-\alpha)\%$ 신뢰구간은 $\hat{\psi} \pm t(n-I, \alpha/2) S(\sum_{i=1}^{I} c_i^2/J_i)^2$ 으로 주어진다.

만약에 일원배치 모형 (5.1)를 모수의 재조정(reparameterization)을 이용하면 다음과 같은 full rank를 만족하는 새로운 모형을 만들 수 있으며 모든 추정가능한 계수의 추정은 모형 (5.1)과 동일하다.

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, 2, \ldots, I, \ j = 1, 2, \ldots, J_i$$