⟨**권이태**⟩

Problem. 1. *동일한 환자로부터 약물에 대한 반응을 약물의 용량을 늘려가면서 세 번에 걸쳐 측정하는 실험을 10명에게 수행하였다. 그 결과가 아래와 같다.*

환자(i)	약물반응(Y)	약물용량(X_1)	성별(X_2)	나이(X_3)
1	70	10	남	59
1	75	20	남	59
1	86	30	남	59
2	60	10	여	55
⋮	⋮	⋮	⋮	⋮
10	80	30	남	61

이 자료의 분석을 위해 아래의 회귀모형을 제안하였다.

$$y_{ij} = \beta_0 + x_{1ij}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + \epsilon_{ij}$$

이때 y_{ij}는 i번째 환자의 j번째 관측값, x_{1ij}는 해당 실험에서의 약물용량, x_{2i}는 남자면 1, 여자면 0인 더미변수, x_{3i}는 나이를 의미한다.

(a) 위 모형을 $Y = X\beta + \epsilon$ 형태로 표현하라. 단, Y는 30차원 벡터이다.

(b) 환자가 가지는 3개의 약물반응값 $(y_{i1}, y_{i2}, y_{i3})^T$은 x_i들에 조건부로 3×3 공분산행렬 Ω를 가진다고 하자. 전체 환자가 Ω를 공유한다고 할 때, 30×30 공분산 행렬 $\mathrm{Var}(Y|X) = \Sigma$를 Ω로 표현하여라.

(c) 이 자료에 대해 오차항의 독립성 가정이 만족되는지 논의하고, $\hat{\beta} = (X^T X)^{-1} X^T y$가 BLUE인지 대답하여라. 만약 그렇지 않다면, 어떠한 추정량이 BLUE인가? X와 y, Σ로써 표현하여라. 또한 해당 추정량의 편의와 분산 역시 구하여라.

(d) 실제로는 $\mathrm{Var}(Y) = \Sigma$를 몰라 (c)에서 구한 추정량을 사용하지 못할 수 있다. 이때 어떠한 방법을 사용할 수 있는가?

Problem. 2. *(a) i번째 실험 참여자의 소득수준이 x_{i1}, 교육기간이 x_{i2}, 성별이 남자인지 여부가 x_{i3}이라고 한다. 이들의 대출승인 여부가 y_i로 주어진다고 하자. y_i가 평균이*

$$\mu_i = \mathbb{E}[y_i|x_{i1}, x_{i2}, x_{i3}] = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})}$$

인 베르누이 확률변수를 따른다고 할 때, $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}$를 μ_i에 대한 식으로 써라.

(b) 성별이 남자인지 여부에 따른 대출승인여부의 오즈비를 $\beta_0, \beta_1, \beta_2, \beta_3$의 함수로써 제시하여라.

(c) 위의 모형을 $\beta_0, \beta_1, \beta_2, \beta_3$*에 대한 최대가능도법을 통해 풀고자 한다. 가능도함수*

$$L(\beta_0, \beta_1, \beta_2, \beta_3 | y, x_1, x_2, x_3)$$

를 구하여라. 표본의 개수는 n*이라 가정한다.*

(d) 좋은 조건 하에서

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}$$

의 최대가능도추정량 $\hat{\beta}$*는 강오목인 로그가능도함수* $l(\beta)$*에 대한 가능도방정식*

$$\nabla_\beta l(\beta) = \mathbf{0}$$

의 근이 됨이 알려져 있다. μ_i*의 추정량* $\hat{\mu}_i$*를* $\hat{\beta}$*와* x_{i1}, x_{i2}, x_{i3}*로 표현하고,*

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{\mu}_i, \quad \sum_{i=1}^{n} x_{ij} y_i = \sum_{i=1}^{n} x_{ij} \hat{\mu}_i \ (j = 1, 2, 3)$$

가 성립함을 확인하여라.

Problem. 3. *한국이는 동급생 n명에게 설문조사를 통해 주변국 p개에 대한 호감도 점수를 얻었다. 한국이는 정치 성향, 문화적 영향력 등의 m개의 요인이 존재하여 이들이 호감도 점수를 결정할 것이라 믿어 아래의 요인분석 모형을 제시하였다(단, $m \ll p$).*

$$X = \mu + LF + \epsilon$$

이때 X는 p개 국가에 대한 호감도 점수, μ는 평균, L은 factor loading($p \times m$차원), F는 factor($m \times 1$차원), ϵ은 오차항이다. 이때

$$\mathbb{E}[F] = 0,\ Cov(F) = I_m, \quad \mathbb{E}[\epsilon] = 0,\ Cov(\epsilon) = \Psi = \mathrm{diag}(\psi_i), \quad Cov(\epsilon, F) = 0$$

을 가정한다.

(a) (μ, L, F, ϵ)*과 차원은 같고 값은 다른* $(\mu', L', F', \epsilon')$*이 존재하여,*

$$X = \mu + LF + \epsilon = \mu' + L'F' + \epsilon'$$

이 성립하고 주어진 가정들을 모두 만족한다. 그러한 $(\mu', L', F', \epsilon')$*를 하나만 찾아라.*

(Hint: 직교행렬 A에 대하여 $L' = LA, F' = A^T F$를 고려하라.)

(b) 인자분석의 식별불가능성 문제를 해결하는 방법에는 L에 제약조건을 주는 방법과 해석이 쉬운 요인회전 A를 찾는 방법이 있다. 각각이 어떻게 적용되는지 위 문제의 상황을 상정하여 설명하시오. 제약조건을 주는 방법의 경우 제약의 개수와 종류를, 요인회전 A를 찾는 방법의 경우 어떤 기준이 사용될 수 있는지 식으로 예를 드시오.

Problem. 4. *선형회귀모형 $y = X\beta + \epsilon$을 추정하고자 한다. 관측값은 n개, 설명변수는 p개 존재한다. $k \times p$차원 제약행렬 T와 $k \times 1$차원 상수행렬 c가 있어 $T\beta = c$임이 이미 알려져 있다고 하자. T의 행들은 선형독립이라고 한다. 이때 β의 최소제곱추정량, 즉*

$$\begin{cases} \text{minimize} & (y - X\beta)^T(y - X\beta) \\ \text{subject to} & T\beta = c \end{cases}$$

의 해 $\tilde{\beta}$를 찾고, 그 편의와 분산을 구하여라. $\epsilon \sim N(0, \sigma^2 I_n)$이라 하자. 이 추정량은 OLS 추정량과 비교하여 어떻게 다른가?

(Hint: $\tilde{\beta} = \hat{\beta}_{OLS} + (X^TX)^{-1}T^T[T(X^TX)^{-1}T^T]^{-1}(c - T\hat{\beta}_{OLS})$)