Փոփոխականների ընտրություն

Մալոուզի հայտանիշ

Ենթադրենք ունենք գծային ռեգրեսիայի

$$Y = X\beta^* + \xi$$

մոդելը և գիտենք, որ X-ը ոչ պատահական է, իսկ ξ-ն պատահական՝ 0 միջինով և $\mathbb{E}[\xi_j^2] = \sigma^2, \forall j$: Ենթադրենք նաև, որ ξ-ի կոորդինատները չեն կոռելացված են, այսինքն ξ-ի կովարիացիոն մատրիցն է $Var[\xi] = \sigma^2 \cdot I_n$.

Դիցուք ունենք $\{1,\dots,p\}$ տարբեր ենթաբազմություններ՝ $S_1, S_2, \dots, S_m$: Կամայական $S \in \{S_1,\dots,S_m\}$-ի համար կարող ենք հաշվել Ն.փ.ք.-ն հաշվի առնելով միայն այն բացատրական փոփոխականները, որոնց ինդեքսը պատկանում է S-ին՝

$$\hat{\beta}^S = (X_S^T X_S)^{-1} X_S^T Y$$

<u>Հարց</u> Ինչպե՞ս է արդյոք $\hat{\beta}^{S_1},\dots,\hat{\beta}^{S_m}$ գնահատականներից ընտրել լավագույնը:

Իհարկե, հարցին պատասխանելու համար անհրաժեշտ է նշել թե ինչ ենք հասկանում լավագույն ասելով: Այստեղ նպատակահարմար է լավագույն համարել այն գնահատականը, որի ռիսկը նվազագույնն է: $\hat{\beta}^S$-ի ռիսկը կսահմանենք այսպես

$$R(S) = \mathbb{E}\left[\|X_S\hat{\beta}^S - X\beta^*\|^2\right]:$$

Մալոուի հայտանիշը հիմնված է հետևյալ թեորեմի վրա:

<u>Թեորեմ</u> Կամայական $S \subset \{1,\dots,p\}$ -ի համար,

$$\hat{R}(S) = \|Y - X_S \hat{\beta}^S\|^2 + 2\sigma^2 |S| - n\sigma^2$$

պատահական մեծությունը (կամ վիճակագրությունը) $R(S)$-ի անշեղ գնահատական է:
(Այստեղ $|S|$-ով ենք նշանակել S-ի տարրերի քանակը:)

<u>Ապացույց</u>

Գիտենք, որ $X_S \hat{\beta}^S = P_S Y$, որտեղ P_S-ը պրոյեկցիոն մատրիցն է X_S-ի սյուներով ծնված տարածության վրա:
Հետևաբար, մի կողմից

$$\begin{aligned}
R(S) &= \mathbb{E}\left[\|P_S Y - X\beta^*\|_2^2\right] = \mathbb{E}\left[\|(P_S - I)X\beta^* + P_S\xi\|^2\right] \\
&= \mathbb{E}\left[\|(P_S - I)X\beta^*\|^2 + 2\langle (P_S - I)X\beta^*, P_S\xi\rangle + \|P_S\xi\|^2\right] \\
&= \|(P_S - I)X\beta^*\|^2 + \mathbb{E}\left[\|P_S\xi\|^2\right] \\
&= \|(P_S - I)X\beta^*\|^2 + \sigma^2 \cdot |S|.
\end{aligned}$$

Մյուս կողմից,

$$\begin{aligned}
\mathbb{E}\left[\hat{R}(S)\right] &= \mathbb{E}\left[\|Y - P_S Y\|^2\right] + 2\sigma^2|S| - n\sigma^2 \\
&= \mathbb{E}\left[\|(I - P_S)X\beta^* + (I - P_S)\xi\|^2\right] + 2\sigma^2|S| - n\sigma^2 \\
&= \|(I - P_S)X\beta^*\|^2 + \mathbb{E}\left[\|(I - P_S)\xi\|^2\right] + 2\sigma^2|S| - n\sigma^2 \\
&= \|(I - P_S)X\beta^*\|^2 + (n - |S|)\sigma^2 + 2\sigma^2|S| - n\sigma^2 \\
&= R(S) \qquad \blacksquare
\end{aligned}$$