제 10 장 Biased estimation of regression coefficients

◆ Principal components regression

◎ **Example** (Import data(French Economy data) : Table 9.5 : 1949-1959, $n=11$)

• Variables : Y : IMPORT, X_1: DOPROD, X_2: STOCK, X_3: CONSUM

▷ Method :

• Original Model : $Y=\beta_0+\beta_1 X_1+\beta_2 X_2+\beta_3 X_3+\epsilon$

• standarize variables : $\widetilde{Y}=\dfrac{Y-\overline{y}}{s_y}$, $\widetilde{X}_j=\dfrac{X_j-\overline{x_j}}{s_j}$, $j=1, \ldots, 3$ $\Rightarrow$ $\widetilde{Y}=\theta_1\widetilde{X}_1+\theta_2\widetilde{X}_2+\theta_3\widetilde{X}_3+\epsilon'$

◦ result : Table 10.7 on p.273 : $\hat{\theta}=(-0.339, 0.213, 1.303)'$ $\Rightarrow$ $\widehat{\beta}_j=\widehat{\theta}_j\dfrac{s_y}{s_1}$, $\widehat{\beta}_0=\overline{y}-\sum_j\widehat{\beta}_j\overline{x_j}$

• Use principal components

$$\tilde{\boldsymbol{y}}=\widetilde{X}VV'\boldsymbol{\theta}+\boldsymbol{\epsilon}=C\boldsymbol{\alpha}+\boldsymbol{\epsilon}, \quad C=\widetilde{X}V, \quad \boldsymbol{\alpha}=V'\boldsymbol{\theta} \quad \Rightarrow \quad \widetilde{Y}=\alpha_1 C_1+\alpha_2 C_2+\alpha_3 C_3+\epsilon'$$

◦ result : Table 10.8 on p.274 : $\hat{\boldsymbol{\alpha}}=(0.690, 0.191, 1.16)'$ $\Rightarrow$ $\hat{\boldsymbol{\theta}}=V\hat{\boldsymbol{\alpha}}=(-0.339, 0.213, 1.303)'$

[Note] $Var(\hat{\boldsymbol{\alpha}})=(C'C)^{-1}\sigma^2=(n-1)^{-1}\Lambda^{-1}\sigma^2$ $\Rightarrow$ $Var(\widehat{\alpha}_j)=\dfrac{\sigma^2}{(n-1)\lambda_j}\propto\dfrac{1}{\lambda_j}$ (책과 차이가 나는 이유를 설명)

$$Var(\hat{\boldsymbol{\theta}})=(n-1)^{-1}(V\Lambda^{-1}V')\sigma^2 \quad \Rightarrow \quad Var(\widehat{\theta}_j)=\frac{1}{(n-1)}\left(\frac{v_{j1}^2}{\lambda_1}+\frac{v_{j2}^2}{\lambda_2}+\frac{v_{j3}^2}{\lambda_3}\right)\sigma^2$$

- Remove dependence among the predictors

Consider two models :

① with one principal component :

$$\tilde{Y} = \alpha_1 C_1 + \epsilon', \qquad \widehat{\alpha_1} = 0.690 \quad \Rightarrow \quad \widehat{\boldsymbol{\alpha}_1} = (\widehat{\alpha_1}, 0, 0)'$$

$$\Rightarrow \quad \widehat{\boldsymbol{\theta}}_1 = V\widehat{\boldsymbol{\alpha}_1} = \widehat{\alpha_1}\boldsymbol{v}_1 = (0.487, 0.030, 0.487)'$$

$$\Rightarrow \quad Var(\widehat{\theta_{1,j}}) = \frac{v_{j1}^2}{(n-1)\lambda_1}\sigma^2 \qquad \Rightarrow \quad \sum_{j-1}^{p} Var(\widehat{\theta_{1,j}}) = \frac{1}{(n-1)\lambda_1}\sigma^2 \sum_{j=1}^{p} v_{j1}^2 = \frac{1}{(n-1)\lambda_1}\sigma^2$$

② with two principal components

$$\tilde{Y} = \alpha_1 C_1 + \alpha_2 C_2 + \epsilon', \qquad \widehat{\alpha_1} = 0.690, \quad \widehat{\alpha_2} = 0.191 \quad \Rightarrow \quad \widehat{\boldsymbol{\alpha}_2} = (\widehat{\alpha_1}, \widehat{\alpha_2}, 0)'$$

$$\Rightarrow \quad \widehat{\boldsymbol{\theta}}_2 = V\widehat{\boldsymbol{\alpha}_2} = \widehat{\alpha_1}\boldsymbol{v}_1 + \widehat{\alpha_2}\boldsymbol{v}_2 = (0.480, 0.221, 0.483)'$$

$$\Rightarrow \quad Var(\widehat{\theta_{2,j}}) = \frac{1}{(n-1)}\left(\frac{v_{j1}^2}{\lambda_1} + \frac{v_{j2}^2}{\lambda_2}\right)\sigma^2 \qquad \Rightarrow \quad \sum_{j=1}^{p} Var(\widehat{\theta_{2,j}}) = \frac{1}{(n-1)}\left(\frac{1}{\lambda_1} + \frac{1}{\lambda_2}\right)\sigma^2$$

$\Rightarrow$ result : C_1 is combination of X_1, X_3 $\Rightarrow$ model ① determined the coefficients of X_1, X_3…

C_2 represents X_2 $\Rightarrow$ model ② determined the coefficients of X_2…

$\Rightarrow$ model ② is plausible representation of the IMPORT relationship… (see Table 10.9 on p.276)

[reg213f]3. 다음에 주어진 정보를 이용하여 다음 질문에 답하라.

종속변수 y와 세 개의 독립변수 x_1, x_2, x_3 값의 평균과 표준편차는 다음과 같다. $(n=16)$

$\overline{y}=31.03,\quad \overline{x_1}=13.97,\quad \overline{x_2}=5.64,\quad \overline{x_3}=2.03,\quad s_y=19.4025,\quad s_1=2.2877,\quad s_2=0.4900,\quad s_3=0.5035$

독립변수들의 상관계수 행렬에 대하여 eigen value와 eigen vector는 다음과 같다.

* **eigen values** (λ) : 2.6505, 0.3432, 0.006223,   * **eigen vectors** (V) : $\begin{pmatrix} 0.6013 & -0.3339 & 0.7259 \\ -0.5339 & -0.8438 & 0.05420 \\ -0.5944 & 0.4202 & 0.6857 \end{pmatrix}$

(A) $y=\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\epsilon$, $\boldsymbol{y}=X\boldsymbol{\beta}+\boldsymbol{\epsilon}$

(B) $\tilde{y}=\theta_1\widetilde{x_1}+\theta_2\widetilde{x_2}+\theta_3\widetilde{x_3}+\epsilon'$, $\tilde{\boldsymbol{y}}=\widetilde{X}\boldsymbol{\theta}+\boldsymbol{\epsilon}'$ (변수를 평균은 0 표준편차는 1이 되도록 표준화 한 것)

(C) $\tilde{y}=\alpha_1 c_1+\alpha_2 c_2+\alpha_3 c_3+\epsilon'$, $\tilde{\boldsymbol{y}}=\widetilde{X}VV'\boldsymbol{\theta}+\boldsymbol{\epsilon}'=C\boldsymbol{\alpha}+\boldsymbol{\epsilon}'$ (principal component 회귀모델)

(1) $x_{i2}=5.8$일 때 $\widetilde{x_{i2}}$를 구하라.

(2) $\widetilde{x_j}$들로부터 첫 번째 principal component c_1을 구하는 식을 쓰라.

(3) 모델 (A)에 다중공선성의 문제가 있는지 판단하고, 그 근거를 제시하라.

(4) 가장 심각하게 다중공선성의 문제를 야기하는 독립변수들($\widetilde{x_j}$)간의 관계를 기술하라.

(5) 모델(B)를 적합시켜서 $\hat{\boldsymbol{\theta}}=(-1.0546,\ -0.2188,\ 0.08130)'$의 결과를 얻었다.

처음 2개의 주성분c_1, c_2를 이용한 주성분 회귀분석을 시행한다고 할 때 $\widehat{\boldsymbol{\theta}_{PC}}$를 구하라.

(6) 위의 (5)에서 구한 $\widehat{\boldsymbol{\theta}_{PC}}$를 이용하여 모델(A)의 β_0, β_1의 추정치를 구하라.

(7) $\sum_{j=1}^{3} Var(\hat{\theta}_j)$가 $\sum_{j=1}^{3} Var(\widehat{\theta_{PC,j}})$에 비해 몇 % 크겠는가?

▷ A caution on PC regression

◎ **Example** (Hald data : Table 10.10 on p.278)

4 predictors, with 4 PC's ⇒ result Table10.12 on p.279

with 3 PC's ⇒ result Table10.13

See Fig10.3 on p.280