제 9 장 Analysis of collinear data

◆ Multicollinearity

▷ model : $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p + \boldsymbol{\epsilon}$$

- Ideal case : $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$: orthogonal
- multicollinearity : $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$: near linearly dependent

[Note] It is not a modeling error, but a condition of deficient data...

▷ Questions to be answered...

- How does multicollinearity affect statistical inference and forecasting ?
- How can multicollinearity be detected ?
- What can be done to resolve the difficulties associated with multicollinearity ?

[Note] If multicollinearity is a potential problem,
the three issues must be treated simultaneously by necessity...

▷ Effect on inference

◎ **Example** (EEO data on Table 9.1-9.2)

[Note] The context of the example is borrowed from research on equal opportunity.
"concerning the lack of availability of equal educational opportunities for individuals
by reason of race, color, religion or national origin in public educational institutions..."

- Variables : Y : ACHV : achievement
  X_1: FAM : home environment
  X_2: PEER : peer group in the school
  X_3: SCHOOL : school facilities
- Model : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \quad (n = 70)$
- Result : Table 9.3 on p.237 : all variables are not significant
  Fig 9.1 on p.238 : OK
  Fig 9.2 on p.239 : high correlation between predictors
  $F \sim F(3, 66)$, $F = 5.72\,(p - value = 0.0015)$ (중회귀모형으로 얻어볼 것)

⇒ a typical situation : reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
not reject $H_0 : \beta_j = 0$, for $j = 1, 2, 3$

deficiency in data : (Fig 9.2 : no information (+,-), (-,+) part)

⇒ ideal case : Table 9.4 on p.240

▷ Effects on forecasting

◎ **Example** (Import data(French Economy data) : Table 9.5 : 1949-1966, $n = 18$)

- Variables : Y : IMPORT : imports

  X_1: DOPROD : domestic production

  X_2: STOCK : stock formation

  X_3: CONSUM : domestic consumption

- Model : $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Result : Table 9.6 on p.242 : reject $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

  not reject $H_0 : \beta_j = 0$, for $j = 1, 2, 3$

  Fig 9.3 on p.242 : autocorrelation ? Fact : European common market began at 1960..

  ⇒ Need to concentrate on 1949-1959

- Result : Table 9.7, Fig 9.4 : better

  However, DOPROD : $\widehat{\beta}_1$: negative, not significant, contrary to prior expectation

  $Corr(X_1, X_3) = 0.997$, $\widehat{X}_3 = 6.259 + 0.686 X_1$

- Forecasting(we must be confident that the character and strength of the overall relationship will hold into future periods.)

$$\hat{Y} = -10.10 - 0.051X_1 + 0.587X_2 + 0.287X_3$$

[Note] The forecast will be accurate as long as
the future values of X_1, X_2 ,X_3 have the relationship that $x_3 \approx 0.7x_1 + 6$.

eg. Year 1960, $x_1 \rightarrow x_1 + 10$ X_2, X_3 : unchanged

$$\widehat{Y(1960)} = \widehat{Y(1959)} - (0.051)(10) = \widehat{Y(1959)} - 0.51$$

$\Rightarrow$ not appropriate, because not hold $x_3 \approx 0.7x_1 + 6$

If $x_1 \rightarrow x_1 + 10$, then $x_3 \rightarrow x_3 + (2/3)(10)$ should be true... then

$$\widehat{Y(1960)} = \widehat{Y(1959)} - (0.51)(10) + (0.287)(6.67) = \widehat{Y(1959)} + 1.5$$

$\Rightarrow$ probably a better forecast...

[Note] The case where X_1 increases alone corresponding to a change in the basic structure of the data...
the forecast cannot be expected to produce meaningful forecasts...

◆ Detection of multicollinearity

▷ Indication of multicollinearity (p.245)

- Large changes in the estimated coefficients when a variable is added or deleted.
  (See Table 9.8 on p.246)
- Large changes in the coefficients when a data point is altered or dropped
- The algebraic signs of the estimated coefficients do not conform to prior expectation.
  (See Table 9.7 : $\widehat{\beta}_1$: negative, not significant)
- Coefficient of variables that are expected to be important have large s.e.(small t-values)

eg. Advertising expenditure data (Table 9.9), $n = 22$

A_t : advertising expenditure

P_t : promotion expenditure

E_t : sales expense

S_t : aggregate sales of a firm in period t.

- model : $S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \epsilon_t$
- result : Table 9.10, 9.11, Fig 9.5, 9.6 : do not suggest any problem
  ⇒ To see the changes when we delete A_t from the model
  See file hadi ch9r.txt :

```
m2= lm(S_t~P_t +E_t +lagA_t + lagP_t)
Coefficients:
             Estimate   Std. Error   t value    Pr(>|t|)
(Intercept)   10.5094     2.4576      4.276    0.000510 ***
P_t            3.7018     0.7571      4.889    0.000138 ***
E_t           22.7942     2.1804     10.454    8.04e-09 ***
lagA_t        -0.7692     0.8746     -0.880    0.391388
lagP_t        -0.9687     0.7423     -1.305    0.209273

 Multiple R-squared: 0.9077
```

coef of P_t : 8.372 -> 3.7018

A_{t-1}, P_{t-1} : change the signs

BUT $R^2 = 0.917 \rightarrow R^2 = 0.9097$ does not change much..

The regression of A_t on P_t, A_{t-1}, P_{t-1}

$$\widehat{A_t} = 4.63 - 0.87P_t - 0.86A_{t-1} - 0.95P_t$$

Approximate rule imposed on the budget :

$$A_t + P_t + A_{t-1} + P_{t-1} \approx 5$$

[Note] 여러 변수들이 복합적으로 선형의 관계가 있을 때 multicoll이 쉽게 진단되지 않을 수도 있다.

▷ Variance Inflation Factor (VIF)

$$VIF_j = \frac{1}{1-R_j^2}, \qquad R_j^2 \ : \ R^2 \text{ for the model } \quad X_j = \beta_0 + \sum_{i \neq j} \beta_i X_i + \epsilon$$

- If $R_j^2 \approx 1$, $\Rightarrow$ $VIF_j \approx \infty$
- If $R_j^2 \approx 0$, $\Rightarrow$ $VIF_j \approx 1$
- If $R_j^2 > 10$, $\Rightarrow$ severe multicollinearity

  See Table 9.12

◆ Principal component approach (통계수학 5장 부분 읽어올 것)

Use unit s.d. version of standardization ($\tilde{Z}$: unit length version)

$$\tilde{\boldsymbol{y}} = \tilde{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}$$

$$\tilde{Z}'\tilde{Z} = Corr(X) = Corr(\tilde{X}) = Corr(\tilde{Z}), \qquad \tilde{X}'\tilde{X} = (n-1)\tilde{Z}'\tilde{Z}$$

$\Rightarrow$ λ_j: j-th largest eigen value of $Corr(X)$, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

$\boldsymbol{v}_j$: orthonormal eigen vector associated with λ_j

eg. Import Data : Table 9.14 on p.254

$$\lambda_1 = 1.999, \quad \lambda_2 = 0.998, \quad \lambda_3 = 0.003$$

$$\boldsymbol{v}_1 = \begin{pmatrix} 0.706 \\ 0.044 \\ 0.707 \end{pmatrix}, \quad \boldsymbol{v}_2 = \begin{pmatrix} -0.036 \\ 0.999 \\ -0.026 \end{pmatrix}, \quad \boldsymbol{v}_3 = \begin{pmatrix} -0.707 \\ -0.007 \\ 0.707 \end{pmatrix}$$

$\Lambda = diag(\lambda_1, \ldots, \lambda_p)$: diagonal matrix

$$V = (\boldsymbol{v}_1, \ \ldots, \boldsymbol{v}_p), \quad V'V = VV' = I \qquad \Rightarrow \quad V'\tilde{Z}'\tilde{Z}V = \Lambda \qquad \Rightarrow \quad \tilde{Z}'\tilde{Z} = V\Lambda V'$$

$\Rightarrow$ $\tilde{\boldsymbol{y}} = \tilde{X}VV'\boldsymbol{\theta} + \boldsymbol{\epsilon} = C\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, $\quad C = \tilde{X}V, \quad \boldsymbol{\alpha} = V'\boldsymbol{\theta}$

$\Rightarrow$ $\boldsymbol{c}_j = \tilde{X}\boldsymbol{v}_j = \tilde{\boldsymbol{x}}_1 v_{1j} + \cdots + \tilde{\boldsymbol{x}}_p v_{pj}$: j=th principal component

$\Rightarrow$ 변수의 형태로 표현하면 $\quad C_1 = 0.706\tilde{X}_1 + 0.044\tilde{X}_2 + 0.707\tilde{X}_3, \quad C_2 = -0.036\tilde{X}_1 + 0.999\tilde{X}_2 + 0.026\tilde{X}_3,$

$$C_3 = 0.707\tilde{X}_1 + 0.007\tilde{X}_2 + 0.707\tilde{X}_3$$

$\Rightarrow$ $C'C = V'\widetilde{X}'\widetilde{X}V = (n-1)V'\widetilde{Z}'\widetilde{Z}V = (n-1)\Lambda$

$\Rightarrow$ C_j : mean is 0 [centered되어 있으므로 모든 변수의 평균이 0이다.]

$\Rightarrow$ $\frac{1}{n-1}C'C = \Lambda$: Variance-Covariance matrix of C

$\Rightarrow$ $\lambda_3 = 0.003$의 의미는 C_3의 분산이 0에 가까우므로 상수에 가깝다.

즉, roughly... $\widetilde{X}_1 + \widetilde{X}_3 \approx const$ (나중에 자세히...) $\Rightarrow$ multicollinearity

eg. EEO data : $\lambda_1 = 2.952$, $\lambda_2 = 0.040$, $\lambda_3 = 0.008$

advertizing data : $\lambda_1 = 1.701$, $\lambda_2 = 0.288$, $\lambda_3 = 1.145$, $\lambda_4 = 0.859$, $\lambda_5 = 0.007$

▷ Condition number (A measure of the overall multicollinearity)

$$\kappa = \sqrt{\frac{\lambda_1}{\lambda_p}} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}}$$

If $\kappa > 15$ $\Rightarrow$ multicollinearity exists

eg. EEO : 19.20

Import : 25.81

Advertizing : 15.59

If $\kappa > 30$ $\Rightarrow$ corrective action should always be taken.

▷ Multicollinearity sources

(1) Import data

$$\lambda_3 = 0.003 \quad \Rightarrow \quad \frac{1}{(n-1)}\boldsymbol{c}_3{}'\boldsymbol{c}_3 = Var(C_3) = 0.003$$

$$\Rightarrow \quad C_3 =- 0.707\widetilde{X}_1 - 0.007\widetilde{X}_2 + 0.707\widetilde{X}_3 \approx 0$$

$$\Rightarrow \quad 0.007\widetilde{X}_2 \approx 0 \quad \Rightarrow \quad \widetilde{X}_1 \approx \widetilde{X}_3 \quad \Rightarrow \quad \frac{X_1 - \overline{x_1}}{s_1} \approx \frac{X_3 - \overline{x_3}}{s_3}$$

$$\Rightarrow \quad s_1 = 29.9995, \quad s_3 = 20.6344 \quad \Rightarrow \quad X_3 \propto \frac{2}{3}X_1 \quad \text{(same as before)}$$

(2) Advertizing data

$$\lambda_5 = 0.007\,, \quad \boldsymbol{v}_5 = (0.514,\ 0.489,\ -0.010,\ 0.428,\ 0.559)'$$

$$\Rightarrow \quad C_5 = 0.514\widetilde{X}_1 + 0.489\widetilde{X}_2 - 0.010\widetilde{X}_3 + 0.428\widetilde{X}_4 + 0.559\widetilde{X}_5 \approx 0$$

$$\Rightarrow \quad 0.010\widetilde{X}_3 \approx 0 \quad \Rightarrow \quad \widetilde{X}_1 \approx - 0.951\widetilde{X}_2 - 0.83\widetilde{X}_4 - 1.087\widetilde{X}_5$$

$$\Rightarrow \quad s_1 = 0.4347, \quad s_2 = 0.4647, \quad s_4 = 0.4099, \quad s_5 = 0.4879$$

$$\Rightarrow (*) \quad X_1 + X_2 + X_4 + X_5 \approx const \quad \text{(same as before)}$$

the next smallest eigen value : $\lambda_4 = 0.859$

⇒ (*) is the only source of multicollinearity