Chapter 9 Multicollinearity

남준우·허인 (2018), 제8장

Gujarati/Porter (2018), Chapter 10

(1) Nature of Multicollinearity

(2) Problems of Multicollineartiy

(3) Detection of High Multicollinearity

(4) Remedy of High Multicollinearity

► Related to ass. (d) No exact linear relationship among X variables.

(1) Nature of Multicollinearity

① Perfect Collinearity(완전공선성)

• There exists an exact linear relationship among all or several X's.

► There exists constants $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k) \neq 0$ such that $\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_k X_k = 0 \quad (X_1 = 1)$.

► $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_k) \neq 0 \Leftrightarrow$ Not all of them are zero simultaneously.

► $X_k = -\dfrac{1}{\lambda_k}\left(\lambda_1 X_1 + \lambda_2 X_2 + \cdots + \lambda_{k-1} X_{k-1}\right) = a_1 X_1 + \cdots + a_{k-1} X_{k-1}$.

(Examples)

(a) X_{2i} :annual income, X_{3i} :monthly income, $X_{2i} = 12 X_{3i}$ for all i.

(b) X_t : export, M_t :*import*, B_t : *trade balance*, $B_t = X_t - M_t$ for all t.

② <u>High Multicollinearity</u>: Highly correlated but not perfectly.

► $X_{ki} = a_1 X_{1i} + \cdots + a_{k-1} X_{k-1\,i} + v_i$, v_i: random error.

(Example)

X_2	X_3	X_3*
10	50	52
15	75	75
18	90	97
24	120	129
30	150	152

$\gamma_{23} = 1, \quad \gamma_{23*} = 0.98$

(Examples)

① GDP vs. GNI

② 경상수지 vs. 무역수지

③ 근로소득 vs. 종합소득

(Notes)

(a) If two variables move together in systematic way (GDP and GNP), we can not identify each variable's effect.

(b) If two variables are highly correlated, then 'other variables are constant' are not well satisfied.

(2) Problems of Multicollineartiy

① Perfect collinearity makes the coefficient indeterminate.

(Example) $y_i = b_1 + b_2 X_i + b_3 Z_i + e_i$

$$b_2 = \frac{S_{Xy} S_Z{}^2 - S_{Zy} S_{XZ}}{S_X{}^2 S_Z{}^2 - S_{XZ}{}^2}.$$ If $X_i = \lambda Z_i$, then $b_2 = \frac{0}{0}$.

(Example) $y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \varepsilon_i$

If $X_i = \lambda Z_i$, then

$$\begin{aligned} y_i &= \beta_1 + \beta_2 X_i + \beta_3 \lambda X_i + \varepsilon_i \\ &= \beta_1 + (\beta_2 + \beta_3 \lambda) X_i + \varepsilon_i \end{aligned}.$$

So, LS of y on $(1.X)$ yields $(\hat{\beta}_1, \widehat{\beta_2 + \lambda \beta_3})$ not $(\hat{\beta}_1, \hat{\beta}_2)$.

② <u>Perfect collinearity makes the variance infinite</u>.

(Example) $y_i = b_1 + b_2 X_i + b_3 Z_i + e_i$

Since $V(b_2) = \dfrac{\sigma^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2 (1 - \gamma_{XZ}{}^2)}$, if $X_i = \lambda Z_i$, $(\gamma_{XZ}{}^2 = 1)$, $V(b_2) = \infty$.

③ <u>High multicollinearity</u> ⇒ variance(or standard error) of the coefficient ↑

► Estimated coefficients are unstable to small changes in sample.

• High multicollinearity ⇒ t-ratio ↓, C.I. wider.

⇒ difficult to reject $H_0 : \beta_j = 0$.

► In fact, X_j is an important variable($\beta_j \neq 0$),

but high multicollinearity makes it as unimportant($\beta_j = 0$).

④ <u>High R^2, F-statistics but low t-ratio's</u>.

► High multicollinearity do not provide enough information to estimate separate effects.

⑤ <u>Accurate forecasting</u>.

(3) Detection of High Multicollinearity

► The problem is the degree of multicollinearity not its existence.

① High R^2 but few significant t-ratio's.

② Pairwise correlation coefficients among X's.

③ R^2 from auxiliary regression: LS of X_j on other X's.

► $X_{ji} = a_1 + a_2 X_{2i} + \cdots + a_{k-1} X_{ki} + v_i, \quad \text{get} \quad R^2$.

(4) Remedy of High Multicollinearity

① Drop one of the variables.

② Transformation of the data.

③ Ridge regression

④ Principal component