Appendix

This appendix shows the mathematical description of the definition of multicollinearity and its diagnostics, which was not presented in the main text.

Multicollinearity

If two explanatory variables X_1 and X_2 have a linear relationship, as follows,

$$c_1X_1 + c_2X_2 = c_0$$
$$\Leftrightarrow X_1 = c_0 - \frac{c_2}{c_1}X_2$$
$$\Leftrightarrow X_2 = c_0 - \frac{c_1}{c_2}X_1,$$

where c_0, c_1, and c_2 are arbitrary constants, the relationship is called exact collinearity. If the relationship between more than two explanatory variables ($X_1, X_2, \dots, X_k$, $k > 2$, *k is a natural number*) is or approximates

$$c_1X_1 + c_2X_2 + \cdots + c_kX_k = c_0,$$

where c_k ($k > 2$, *k is a natural number*) is an arbitrary constant, multicollinearity occurs. Under multicollinearity, more than one explanatory variable X_h is determined by the other explanatory variables as follows:

$$X_h \cong \left(c_0 - \sum_{j \neq h} c_j X_j\right) / \ c_h \ (j = 1,2,\dots,k) \ j \neq h$$

Variance Inflation Factor

A multiple linear regression model with n sample observations of k explanatory variables ($X_1, X_2, \dots, X_k$) and a response variable (Y) is given by

$$Y_i = \beta_0 + \beta_1X_{i1} + \beta_2X_{i2} + , \dots , + \beta_kX_{ik} + \varepsilon_i \ (i = 1,2, \dots , n) \quad \varepsilon_i \sim N(0, \sigma^2),$$

where $\beta_j (j = 0,1,2, \dots , k)$ and ε_i are the regression coefficients and error, respectively. Each error ($\varepsilon_1, \varepsilon_2, \dots , \varepsilon_n$) is stochastically independent and is normally distributed with a mean of 0 and a variance of σ^2. The variance of β_j [$Var(\beta_j)$] is

$$Var(\beta_j) = \sigma^2 \left(\frac{1}{1 - R_j^2}\right)\left(\frac{1}{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2}\right)$$

where $\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2 = (X_{1j} - \overline{X}_j)^2 + (X_{2j} - \overline{X}_j)^2 + \cdots + (X_{nj} - \overline{X}_j)^2$ is the sum of squares of the difference between each value of X_{ij} and the mean of X_{ij} ($\overline{X}_j$) and R_j^2 is the coefficient of determination from the regression model [$X_{ij} = \gamma_0 + \sum_{l=1}^{k} \gamma_l X_{il} + \epsilon_i$ ($i = 1, 2, \dots , n$; $l = 1,2, \dots , k$; $l \neq j$)] with the response variable of X_{ij}, the explanatory variables of X_{il}, the regression coefficients of γ_0 and γ_l, and the error of ϵ_i. Assuming that $\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2$ and σ^2 are constant, $Var(\beta_j)$ is solely dependent on $\frac{1}{1 - R_j^2}$ and an increase in R_j^2 leads to an increase in $Var(\beta_j)$ and vice versa. Because $0 \leq R_j^2 \leq 1$, $R_j^2 = 0$ minimizes $Var(\beta_j)$ while $R_j^2 \approx 1$ makes $Var(\beta_j)$ infinite (Fig. 1). This means that the complete absence of multicollinearity ($R_j^2 = 0$) between explanatory variables minimizes the variance of the regression coefficient for an explanatory variable of interest, whereas exact multicollinearity ($R_j^2 = 1$) between them inflates the variance infinitely. Because of its significant effects on the variance of a regression coefficient, the term

$$\frac{1}{1-R_j^2}$$

is called the variance inflation factor; its reciprocal is known as the tolerance.

The variance inflated by strong multicollinearity increases the standard error of the regression coefficient $\left(\sqrt{Var(\beta_j)}\right)$ and widens the 95% confidence interval of a regression coefficient (β_j), which is

$$\beta_j \pm t_{(n-k-1;\,0.025)}\left(\sqrt{Var(\beta_j)}\right),$$

where $t_{(n-k-1;\,0.025)}$ is the critical t-statistic at 2.5% (= $\frac{100-95}{2}$%) level under the degree of freedom $n-k-1$. The increase in the variance also results in a reduction in t-statistic

$$T=\frac{\beta_j-0}{\sqrt{Var(\beta_j)}}$$

for the hypothesis test (H_0: $\beta_j = 0$ *versus* H_1: $\beta_j \neq 0$), which produces an insignificant result.

Condition Number and Condition Index

Each explanatory variable (X_{ij}) from a multiple linear regression $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + , \dots , + \beta_k X_{ik} + \varepsilon_i$ (i = 1,2, … , n) can be standardized by dividing the difference between each of its values (X_{ij}) and their mean ($\overline{X}_j$) by the square root of the sum of squares of all the differences:

$$Z_{ij}=\frac{X_{ij}-\overline{X}_j}{\sqrt{\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)^2}}\,(j=1,2,\dots,k)$$

Then, we obtain an $n \times k$ matrix (Z) of the standardized explanatory variables:

$$Z=\begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1k} \\ Z_{21} & Z_{22} & \cdots & Z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nk} \end{pmatrix}$$

By transposing Z, so that the rows become columns and vice versa, we obtain the $k \times n$ transposed matrix (Z^T):

$$Z^T=\begin{pmatrix} Z_{11} & Z_{21} & \cdots & Z_{n1} \\ Z_{12} & Z_{22} & \cdots & Z_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1k} & Z_{2k} & \cdots & Z_{nk} \end{pmatrix}$$

The multiplication of Z^T by Z produces a $k \times k$ square matrix. As shown below, the multiplications of each element from the a^{th} row of Z^T and the b^{th} column of Z yield the element from the b^{th} column of the a^{th} row in $Z^T \times Z$:

$$Z^T \times Z=\begin{pmatrix} Z_{11} & Z_{21} & \cdots & Z_{n1} \\ Z_{12} & Z_{22} & \cdots & Z_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1k} & Z_{2k} & \cdots & Z_{nk} \end{pmatrix}\times\begin{pmatrix} Z_{11} & Z_{12} & \cdots & Z_{1k} \\ Z_{21} & Z_{22} & \cdots & Z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} Z_{11}Z_{11}+Z_{21}Z_{21}+\cdots+Z_{n1}Z_{n1} & Z_{11}Z_{12}+Z_{21}Z_{22}+\cdots+Z_{n1}Z_{n2} & \cdots & Z_{11}Z_{1k}+Z_{21}Z_{2k}+\cdots+Z_{n1}Z_{1k} \\ Z_{12}Z_{11}+Z_{22}Z_{21}+\cdots+Z_{n2}Z_{n1} & Z_{12}Z_{12}+Z_{22}Z_{22}+\cdots+Z_{n2}Z_{n2} & \cdots & Z_{12}Z_{1k}+Z_{22}Z_{2k}+\cdots+Z_{n2}Z_{nk} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1k}Z_{11}+Z_{2k}Z_{21}+\cdots+Z_{nk}Z_{n1} & Z_{1k}Z_{12}+Z_{2k}Z_{22}+\cdots+Z_{nk}Z_{n2} & \cdots & Z_{1k}Z_{1k}+Z_{2k}Z_{2k}+\cdots+Z_{nk}Z_{nk} \end{pmatrix}$$

Each element of the square matrix is equivalent to a correlation coefficient (r) of two explanatory variables (X_{ih} and X_{ij}).

$$Z_{1h}Z_{1j}+Z_{2h}Z_{2j}+\cdots+Z_{nh}Z_{nj}$$

$$= \frac{X_{1h}-\overline{X}_h}{\sqrt{\sum_{i=1}^n (X_{ih}-\overline{X}_h)^2}} \frac{X_{1j}-\overline{X}_j}{\sqrt{\sum_{i=1}^n (X_{ij}-\overline{X}_j)^2}} + \frac{X_{2h}-\overline{X}_h}{\sqrt{\sum_{i=1}^n (X_{ih}-\overline{X}_h)^2}} \frac{X_{2j}-\overline{X}_j}{\sqrt{\sum_{i=1}^n (X_{ij}-\overline{X}_j)^2}} + \cdots + \frac{X_{nh}-\overline{X}_h}{\sqrt{\sum_{i=1}^n (X_{ih}-\overline{X}_h)^2}} \frac{X_{nj}-\overline{X}_j}{\sqrt{\sum_{i=1}^n (X_{ij}-\overline{X}_j)^2}} = r_{hj}$$

Therefore, the matrix Z^TZ can be expressed as follows:

$$Z^TZ = \begin{pmatrix} r_{11} & r_{11} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{2k} & \cdots & r_{kk} \end{pmatrix}$$

To calculate the eigenvalues of a square matrix, its determinant needs to be known. The determinant of a 2 × 2 matrix is

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

The determinant of a 3 × 3 matrix is

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$

$$= a(ei - fh) - b(di - fg) + c(dh - eg)$$

Using the above equations for the determinant of a square matrix, the eigenvalues (λ_1, λ_2) of the 2 × 2 correlation matrix can be obtained:

$$\left| \begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} - \lambda \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right| = 0$$

$$\begin{vmatrix} r_{11}-\lambda & r_{12} \\ r_{21} & r_{22}-\lambda \end{vmatrix} = 0$$

$$(r_{11}-\lambda)(r_{22}-\lambda) - r_{12}r_{21} = 0$$

$$\lambda^2 - (r_{11}+r_{22})\lambda + r_{11}r_{22} - r_{12}r_{21} = 0$$

$$\lambda = \frac{(r_{11}+r_{22}) \pm \sqrt{(r_{11}+r_{22})^2 - 4(r_{11}r_{22} - r_{12}r_{21})}}{2} \because ax^2+bx+c=0 \Leftrightarrow x$$

$$= \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

If generalized, the eigenvalues ($\lambda_1, \lambda_2, \ldots, \lambda_k$) of the correlation matrix can be calculated.

$$\left|\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{2k} & \cdots & r_{kk} \end{pmatrix} - \lambda \times \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}\right| = 0$$

$$\begin{vmatrix} r_{11}-\lambda & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22}-\lambda & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{2k} & \cdots & r_{kk}-\lambda \end{vmatrix} = 0$$

$$(r_{11}-\lambda)\begin{vmatrix} r_{22}-\lambda & \cdots & r_{2k} \\ \vdots & \ddots & \vdots \\ r_{2k} & \cdots & r_{kk}-\lambda \end{vmatrix} + r_{21}\begin{vmatrix} r_{21} & r_{23} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{3k} & \cdots & r_{kk}-\lambda \end{vmatrix} + \cdots + r_{1k}\begin{vmatrix} r_{12} & \cdots & r_{2(k-1)} \\ \vdots & \ddots & \vdots \\ r_{1k} & \cdots & r_{k(k-1)} \end{vmatrix} = 0$$

$$(\lambda-\lambda_1)(\lambda-\lambda_2)\ldots(\lambda-\lambda_k) = 0$$

By solving the k^{th} degree polynomial equation of the variable λ, we can obtain k eigenvalues ($\lambda_1, \lambda_2, \ldots, \lambda_k$). The number of eigenvalues ($\lambda_1, \lambda_2, \ldots, \lambda_k$) from the $k \times k$ matrix is k and their mean and total sum are 1 and k, respectively.

The square root of the ratio between the maximum and each eigenvalue ($\lambda_1, \lambda_2, \ldots, \lambda_k$) is termed "condition index" and is expressed as

$$\kappa_s = \sqrt{\frac{\lambda_{max}}{\lambda_s}}(s = 1, 2, \ldots, k)$$

The largest condition index is called the "condition number."

Variance Decomposition Proportion

Eigenvectors are calculated from their corresponding eigenvalues. The relationship between two eigenvalues (λ_1, λ_2) and their eigenvectors (v_1, v_2) is as follows:

$$\left[\begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix} - \lambda_s \times \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right] \times \begin{pmatrix} v_{1s} \\ v_{2s} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} (s = 1, 2)$$

By solving the above equation, the ratio (R_s) between the two elements ($v_{1s} = R_s \times v_{2s}$) is obtained. As long as the ratio is maintained, the values of the two elements can be chosen arbitrarily. Then, two eigenvectors can be obtained.

$$v_1 = \begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix}, v_2 = \begin{pmatrix} v_{12} \\ v_{22} \end{pmatrix}$$

With

$$\left[\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ r_{21} & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{2k} & \cdots & r_{kk} \end{pmatrix} - \lambda_s \times \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}\right] \times \begin{pmatrix} v_{1s} \\ v_{2s} \\ \vdots \\ v_{ks} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

we have k eigenvectors ($v_1, v_2, \ldots, v_k$) consisting of k elements in one column, which correspond to k eigenvalues ($\lambda_1, \lambda_2, \ldots, \lambda_k$).

The eigenvector corresponding to the eigenvalue λ_s ($s = 1, 2, \ldots, k$) are expressed as

$$v_s = \begin{pmatrix} v_{1s} \\ v_{2s} \\ \vdots \\ v_{ks} \end{pmatrix}$$

There are k variance decomposition proportions for the regression coefficient β_j ($j = 1, 2, \dots, k$), which are defined as

$$\pi_{js} = \frac{\frac{v_{js}^2}{\lambda_s}}{\frac{v_{j1}^2}{\lambda_1} + \frac{v_{j2}^2}{\lambda_2} + \cdots \frac{v_{jk}^2}{\lambda_k}} = \frac{\frac{v_{js}^2}{\lambda_s}}{\sum_{s=1}^{k} \frac{v_{js}^2}{\lambda_s}} \ (s = 1, 2, \dots, k)$$

The total sum of the variance decomposition proportions for β_j ($\pi_{j1} + \pi_{j2} + \cdots + \pi_{jk} = \sum_{s=1}^{k} \pi_{js}$) is 1.