KJA
Korean Journal of Anesthesiology

Statistical Round
pISSN 2005-6419 • eISSN 2005-7563

# Multicollinearity and misleading statistical results

Jong Hae Kim

*Department of Anesthesiology and Pain Medicine, School of Medicine, Daegu Catholic University, Daegu, Korea*

Multicollinearity represents a high degree of linear intercorrelation between explanatory variables in a multiple regression model and leads to incorrect results of regression analyses. Diagnostic tools of multicollinearity include the variance inflation factor (VIF), condition index and condition number, and variance decomposition proportion (VDP). The multicollinearity can be expressed by the coefficient of determination ($R_h^2$) of a multiple regression model with one explanatory variable ($X_h$) as the model's response variable and the others ($X_i$ [$i \neq h$]) as its explanatory variables. The variance ($\sigma_h^2$) of the regression coefficients constituting the final regression model are proportional to the VIF ($\frac{1}{1-R_h^2}$). Hence, an increase in $R_h^2$ (strong multicollinearity) increases $\sigma_h^2$. The larger $\sigma_h^2$ produces unreliable probability values and confidence intervals of the regression coefficients. The square root of the ratio of the maximum eigenvalue to each eigenvalue from the correlation matrix of standardized explanatory variables is referred to as the condition index. The condition number is the maximum condition index. Multicollinearity is present when the VIF is higher than 5 to 10 or the condition indices are higher than 10 to 30. However, they cannot indicate multicollinear explanatory variables. VDPs obtained from the eigenvectors can identify the multicollinear variables by showing the extent of the inflation of $\sigma_h^2$ according to each condition index. When two or more VDPs, which correspond to a common condition index higher than 10 to 30, are higher than 0.8 to 0.9, their associated explanatory variables are multicollinear. Excluding multicollinear explanatory variables leads to statistically stable multiple regression models.

**Keywords:** Biomedical research; Biostatistics; Multivariable analysis; Regression; Statistical bias; Statistical data analysis.

## Introduction

A prospective randomized controlled trial assesses the effects of a single explanatory variable on its primary outcome variable and the unknown effects of the other explanatory variables are minimized by randomization [1]. However, in observational or retrospective studies, randomization is not performed before the collection of data and there exist the confounding effects of explanatory variables other than the one of interest. To control the confounding effects on a single response variable, multivariable regression analyses are used. However, most of the time, explanatory variables are intercorrelated and produce significant effects on one another. This relationship between explanatory variables compromises the results of multivariable regression analyses. The intercorrelation between explanatory variables is termed as "multicollinearity." In this review, the definition of multicollinearity, measures to detect it, and its effects on the results of multiple linear regression analyses will be discussed. In the appendix following the main text, the concepts of multicollinearity and measures for its detection are described with as much detail as possible along with mathematical equations to aid readers who are unfamiliar with statistical mathematics.

Corresponding author: Jong Hae Kim, M.D.
Department of Anesthesiology and Pain Medicine, School of Medicine, Daegu Catholic University, 33 Duryugongwon-ro 17-gil, Namgu, Daegu 42472, Korea
Tel: +82-53-650-4979, Fax: +82-53-650-4517
Email: usmed12@gmail.com
ORCID: https://orcid.org/0000-0003-1222-0054

Received: March 3, 2019.
Revised: May 17, 2019.
Accepted: July 8, 2019.

Korean J Anesthesiol 2019 December 72(6): 558-569
https://doi.org/10.4097/kja.19087

## Multicollinearity

Exact collinearity is a perfect linear relationship between two explanatory variables $X_1$ and $X_2$. In other words, exact collinearity occurs if one variable determines the other variable (e.g., $X_1 = 100 - 2X_2$). If such relationship exists between more than two explanatory variables (e.g., $X_1 = 100 - 2X_2 + 3X_3$), the relationship is defined as multicollinearity. Under multicollinearity, more than one explanatory variable is determined by the others. However, collinearity or multicollinearity do not need to be exact to determine their presence. A strong relationship is enough to have significant collinearity or multicollinearity. A coefficient of determination is the proportion of the variance in a response variable predicted by the regression model built upon the explanatory variable (s). However, the coefficient of determination ($R^2$) from a multiple linear regression model whose response and explanatory variables are one explanatory variable and the rest, respectively, can also be used to measure the extent of multicollinearity between explanatory variables. $R^2 = 0$ represents the absence of multicollinearity between explanatory variables, while $R^2 = 1$ represents the presence of exact multicollinearity between them. The removal of one or more explanatory variables from variables with exact multicollinearity does not cause loss of information from a multiple linear regression model.

## Variance Inflation Factor

The variance of regression coefficients is proportional to

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

which is called the variance inflation factor. Considering the range of $R^2$ ($0 \leq R^2 \leq 1$), $R^2 = 0$ (complete absence of multicollinearity) minimizes the variance of the regression coefficient of interest, while $R^2 = 1$ (exact multicollinearity) makes this variance infinite (Fig. 1). The reciprocal of the variance inflation factor ($1 - R^2$) is known as the tolerance. If the variance inflation factor and tolerance are greater than 5 to 10 and lower than 0.1 to 0.2, respectively ($R^2$ = 0.8 to 0.9), multicollinearity exists. Although the variance inflation factor helps to determine the presence of multicollinearity, it cannot detect the explanatory variables causing the multicollinearity.

As previously mentioned, strong multicollinearity increases the variance of a regression coefficient. The increase in the variance also increases the standard error of the regression coefficient (because the standard error is the square root of the variance). The increase in the standard error leads to a wide 95% confidence interval of the regression coefficient. The inflated variance also results in a reduction in the t-statistic to determine whether the regression coefficient is 0. With a low t-statistic value, the regression coefficient becomes insignificant. The wide confidence interval and insignificant regression coefficient make the final predictive regression model unreliable.

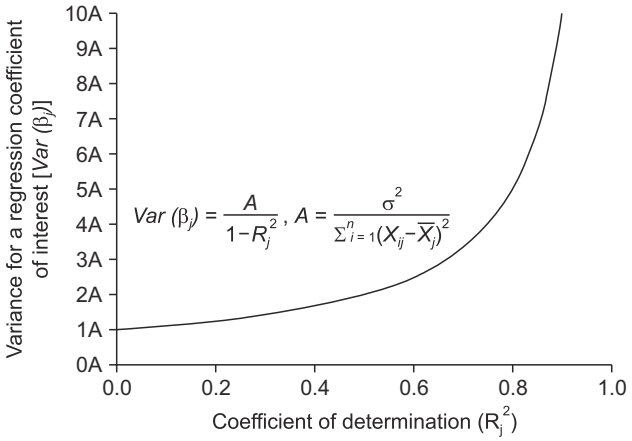


**Fig. 1.** The effects of the coefficient of determination ($R_j^2$) from a regression model [$X_{ij} = \gamma_0 + \sum_{l=1}^{k} \gamma_l X_{il} + \epsilon_i$ ($i = 1, 2, \ldots, n$; $l = 1, 2, \ldots, k$; $l \neq j$)] on the variance of a regression coefficient of interest [$Var(\beta_j)$]. The presence of multicollinearity (an increase in $R_j^2$) inflates $Var(\beta_j)$. $X_{ij}$: $j$th explanatory variable of a regression model [$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + , \ldots , + \beta_k X_{ik} + \varepsilon_i$ ($i = 1,2, \ldots , n$)].

## Condition Number and Condition Index

The eigenvalues (λ) obtained from the calculation using a matrix composed of standardized explanatory variables can be used to diagnose multicollinearity. The total number and sum of the eigenvalues are equal to the number of explanatory variables. The average of the eigenvalues is 1. Because the total sum of the eigenvalues is constant, the presence of their high maximum value indicates that the other eigenvalues are low relative to the maximum ($\lambda_{max}$). Eigenvalues close to 0 indicate the presence of multicollinearity, in which explanatory variables are highly intercorrelated and even small changes in the data lead to large changes in regression coefficient estimates. The square root of the ratio between the maximum and each eigenvalue ($\lambda_1, \lambda_2, \ldots, \lambda_k$) is referred to as the condition index:

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

The largest condition index is called the condition number. A condition number between 10 and 30 indicates the presence of multicollinearity and when a value is larger than 30, the multicollinearity is regarded as strong.

## Variance Decomposition Proportion

Eigenvectors derived from eigenvalues are used to calculate the variance decomposition proportions, which represent the extent of variance inflation by multicollinearity and enable the determination of the variables involved in the multicollinearity. Each explanatory variable has variance decomposition proportions corresponding to each condition index. The total sum of the variance decomposition proportions for one explanatory variable is 1. If two or more variance decomposition proportions corresponding to condition indices higher than 10 to 30 exceed 80% to 90%, it is determined that multicollinearity is present between the explanatory variables corresponding to the exceeding variance decomposition proportions.

## Strategies to Deal with Multicollinearity

Erroneous recording or coding of data may inadvertently cause multicollinearity. For example, the unintentional duplicative inclusion of the same variable in a regression analysis yields a multicollinear regression model. Therefore, preventing human errors in data handling is very important. Theoretically, increas-

**Table 1.** Raw Data from Reference [4]

| Serial number | PVV/GW (cm/s/100 g) | PSV/GW (cm/s/100 g) | EDV/GW (cm/s/100 g) | HVV/GW (cm/s/100 g) | GW/SLV (%) | GRWR (%) | Regeneration rate (%) |
|---|---|---|---|---|---|---|---|
| 1 | 16.36 | 8.9 | 3.47 | 6.02 | 57.42 | 1.11 | 158.76 |
| 2 | 26.68 | 21.22 | 3.53 | 12.07 | 61.38 | 1.36 | 197.19 |
| 3 | 12.49 | 16.62 | 2 | 8.88 | 67.42 | 1.47 | 144.73 |
| 4 | 8.45 | 22.86 | 6.71 | 7.46 | 69.94 | 1.31 | 140.06 |
| 5 | 10.19 | 14.23 | 4.75 | 2.06 | 65.68 | 1.25 | 129.71 |
| 6 | 19.53 | 17.35 | 1.95 | 7.54 | 59.63 | 1.14 | 162.59 |
| 7 | 20.65 | 10.48 | 2.21 | 4.88 | 59.42 | 1.07 | 178.48 |
| 8 | 22.96 | 14.23 | 4.25 | 3.69 | 75.08 | 1.73 | 120.9 |
| 9 | 21.22 | 21.64 | 4.1 | 11.94 | 43.42 | 0.87 | 191.24 |
| 10 | 8.11 | 3.16 | 0.78 | 8.82 | 75.12 | 1.47 | 150.03 |
| 11 | 24.74 | 7.84 | 1.68 | 3.68 | 57.65 | 1.08 | 173.44 |
| 12 | 11.38 | 15.71 | 3.56 | 7.2 | 39.93 | 0.74 | 211.98 |
| 13 | 15.82 | 15.04 | 2.4 | 9.89 | 51.27 | 1.02 | 193.49 |
| 14 | 8.36 | 9.01 | 2.01 | 3.4 | 50.52 | 0.94 | 164.04 |
| 15 | 12.04 | 9.72 | 2.27 | 6.03 | 51.6 | 1.05 | 156.97 |
| 16 | 10.97 | 4.58 | 1.73 | 5.55 | 56.63 | 1.03 | 208.36 |
| 17 | 7.97 | 9.33 | 0.57 | 4.17 | 79.09 | 1.61 | 154.62 |
| 18 | 7.46 | 6.11 | 1.73 | 2.99 | 57.2 | 1.07 | 137.38 |
| 19 | 29.09 | 15.71 | 3.41 | 9.35 | 56.44 | 1.1 | 180.15 |
| 20 | 10.3 | 8.54 | 2.32 | 10.78 | 60.43 | 1.17 | 228.47 |
| 21 | 7.82 | 4.41 | 1.07 | 4.19 | 59.52 | 1 | 153.62 |
| 22 | 14.71 | 6.29 | 1.77 | 6.16 | 65.05 | 1.3 | 121.31 |
| 23 | 8.54 | 6.73 | 1.27 | 5.52 | 65.65 | 1.17 | 157.37 |
| 24 | 23.05 | 11.34 | 5.39 | 3 | 33.57 | 0.63 | 211.27 |
| 25 | 13.12 | 5.86 | 1.89 | 10.92 | 52.93 | 0.9 | 178.16 |
| 26 | 7.41 | 9.11 | 2.05 | 5.5 | 53.72 | 0.91 | 174.89 |
| 27 | 14.59 | 5.59 | 1.26 | 3.75 | 58.62 | 1.14 | 142.98 |
| 28 | 8.52 | 6.52 | 1 | 6.92 | 56.61 | 1.11 | 165.59 |
| 29 | 18.97 | 6.35 | 2.94 | 5.61 | 56.41 | 1.07 | 141.54 |
| 30 | 35.41 | 36.36 | 14.23 | 15 | 41.52 | 0.89 | 238.22 |
| 31 | 4.55 | 1.27 | 3.13 | 2.83 | 70.91 | 1.27 | 138.42 |
| 32 | 22.59 | 28.7 | 10.51 | 10.35 | 32.74 | 0.66 | 247.45 |
| 33 | 9.21 | 4.55 | 1.19 | 7.92 | 72.2 | 1.34 | 140.27 |
| 34 | 18.32 | 11.61 | 2.91 | 8.07 | 52.23 | 1.02 | 216.06 |
| 35 | 5.69 | 6.88 | 1.18 | 2.78 | 72.12 | 1.39 | 144.18 |
| 36 | 11.21 | 11.92 | 3.31 | 10.29 | 60.65 | 1.69 | 156.22 |

PVV/GW: peak portal venous flow velocity per 100 g of the initial graft weight, PSV/GW: peak systolic velocity of the hepatic artery per 100 g of the initial graft weight, EDV/GW: end diastolic velocity of the hepatic artery per 100 g of the initial graft weight, HVV/GW: peak hepatic venous flow velocity per 100 g of the initial graft weight, GW/SLV: graft-to-standard liver volume ratio, GRWR: graft-to-recipient weight ratio.

ing the sample size reduces the standard errors of regression coefficients and hence, decreases the degree of multicollinearity [2]. However, standard errors are not always reduced under strong multicollinearity. Old explanatory variables can be replaced with newly collected ones, which predict a response variable more accurately. However, the inclusion of new cases or explanatory variables to an already completed study requires significant additional time and cost or is simply technically impossible.

Combining multicollinear variables into one can be another option. Variables that belong to a common category are usually multicollinear. Hence, combining each variable into a higher hierarchical variable can reduce the multicollinearity. In addition, one of the variables in nearly exact collinearity can be presented by an equation with the other variable. The inclusion of the equation in the multiple regression model removes one of the collinear variables. Principal component analysis or factor analysis can also generate a single variable that combines multicollinear variables. However, with this procedure, it is not possible to assess the effect of individual multicollinear variables.

Finally, the multicollinear variables identified by the variance decomposition proportions can be discarded from the regression model, making it more statistically stable. However, the principled exclusion of multicollinear variables alone does not guarantee the remaining of the relevant variables, whose effects on the response variable should be investigated in the multivariable regression analysis. The exclusion of relevant variables produces biased regression coefficients, leading to issues more serious than multicollinearity. Ridge regression is an alternative modality to include all the multicollinear variables in a regression model [3].

## Numerical Example

In this section, multicollinearity is assessed from variance inflation factors, condition numbers, condition indices, and variance decomposition proportions, using data (Table 1) from a previously published paper [4]. The response variable considered, is the liver regeneration rate two weeks after living donor liver transplantation (LDLT). Four of the six explanatory variables considered, are hepatic hemodynamic parameters measured one day after LDLT, which include the peak portal venous flow velocity (PVV), the peak systolic velocity (PSV) and the end diastolic velocity (EDV) of the hepatic artery, and the peak hepatic venous flow velocity (HVV). These parameters are standardized by dividing them by 100 g of the initial graft weight (GW). The other explanatory variables considered are the graft-to-recipient weight ratio (GRWR) and the GW to standard liver volume ratio (GW/SLV).

Because the shear stress generated by the inflow from the portal vein and hepatic artery into a partial liver graft serves as a driving force for liver regeneration [5], it is assumed that the standardized PPV, PSV, and EDV (PVV/GW, PSV/GW, and EDV/GW, respectively) are positively correlated with the liver regeneration rate. A positive correlation between the standardized HVV (HVV/GW) and liver regeneration rate is also expected because the inflow constitutes the outflow through the hepatic vein from a liver graft. In addition, the smaller is a liver graft, the higher is the shear stress. Hence, the graft weight relative to the recipient weight and standard liver volume is expected to be negatively correlated with the liver regeneration rate. The expected univariate correlations between each explanatory variable and liver regeneration rate were found in the above-cited paper [4]. Significant correlations between the hepatic hemo-

**Table 2.** Correlation Matrix between Explanatory Variables

| | | PVV/GW (cm/s/100 g) | PSV/GW (cm/s/100 g) | EDV/GW (cm/s/100 g) | HVV/GW (cm/s/100 g) | GW/SLV (%) | GRWR (%) |
|---|---|---|---|---|---|---|---|
| PVV/GW | Pearson's correlation coefficient | 1 | 0.649† | 0.591† | 0.456† | −0.459† | −0.262 |
| | Two-tailed P value | | < 0.001 | < 0.001 | 0.005 | 0.005 | 0.122 |
| PSV/GW | Pearson's correlation coefficient | | 1 | 0.841† | 0.610† | −0.442† | −0.217 |
| | Two-tailed P value | | | < 0.001 | < 0.001 | 0.007 | 0.203 |
| EDV/GW | Pearson's correlation coefficient | | | 1 | 0.450† | −0.504† | −0.330* |
| | Two-tailed P value | | | | 0.006 | 0.002 | 0.049 |
| HVV/GW | Pearson's correlation coefficient | | | | 1 | −0.310 | −0.109 |
| | Two-tailed P value | | | | | 0.066 | 0.528 |
| GW/SLV | Pearson's correlation coefficient | | | | | 1 | 0.886† |
| | Two-tailed P value | | | | | | < 0.001 |
| GRWR | Pearson's correlation coefficient | | | | | | 1 |
| | Two-tailed P value | | | | | | |

PVV/GW: peak portal venous flow velocity per 100 g of the initial graft weight, PSV/GW: peak systolic velocity of the hepatic artery per 100 g of the initial graft weight, EDV/GW: end diastolic velocity of the hepatic artery per 100 g of the initial graft weight, HVV/GW: peak hepatic venous flow velocity per 100 g of the initial graft weight, GW/SLV: graft-to-standard liver volume ratio, GRWR: graft-to-recipient weight ratio. *P < 0.05, †P < 0.01.

dynamic parameters (PVV/GW, PSV/GW, EDV/GW, and HVV/GW) and between the relative graft weights (GRWR and GW/SLV) are anticipated because they share common characteristics. Therefore, the use of all of these explanatory variables for multiple linear regression analysis might lead to multicollinearity.

As expected, the correlation matrix of the explanatory variables shows a significant correlation between the variables (Table 2). Based on the regression coefficients calculated from the multiple linear regression analysis using the six explanatory variables (Table 3A), we obtain the following regression model.

*Liver regeneration rate*

$$= 232.797 + 0.221 \times PPV/GW - 0.050 \times PSV/GW + 0.690 \times EDV/GW + 4.083 \times HVV/GW - 0.905 \times GW/SLV - 37.594 \times GRWR$$

$$(R^2 = 0.682, P < 0.001)$$

While an increase in the PSV/GW leads to an increase in the liver regeneration rate, according to the results of the simple linear regression analysis [4] a unit increase in the PSV/GW reduces, albeit insignificantly, the regeneration rate by 0.05%. In addition, although the effect of a unit change in the GRWR on the regeneration rate is the strongest (37.954% decrease in the regeneration rate per unit increase), its regression coefficient is not statistically significant due to its inflated variance, which leads to a high standard error of 32.665 and a wide 95% confidence interval between –104.4 and 29.213. These unreliable results are produced by multicollinearity presented by the high variance inflation factors of the regression coefficients for GW/SLV, GRWR, and PSV/GW (more than or very close to 5), which are 7.384, 6.011, and 4.948, respectively. They are indicated with asterisks in Table 3A. The three condition indices of more than 10 (daggers in Table 3B) also indicate that there are three linear dependencies that arise from multicollinearity. However, they cannot identify the explanatory variables with multicollinearity.

The variance decomposition proportions exceeding 0.8 are

**Table 3A**. Regression Coefficients of Multiple Linear Regression Model for Six Explanatory Variables

| | Unstandardized coefficients | Standard error | Standardized coefficients | t-statistic | P | 95% Confidence interval for the unstandardized coefficients | | Collinearity statistics | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower bound | Upper bound | Tolerance | Variance inflation factor |
| Intercept | 232.797 | 29.542 | | 7.880 | < 0.001 | 172.376 | 293.217 | | |
| PVV/GW | 0.221 | 0.642 | 0.050 | 0.344 | 0.733 | −1.093 | 1.534 | 0.525 | 1.905 |
| PSV/GW | −0.050* | 1.026 | −0.011 | −0.048 | 0.962* | −2.148 | 2.049 | 0.202 | 4.948* |
| EDV/GW | 0.690 | 2.506 | 0.056 | 0.275 | 0.785 | −4.435 | 5.816 | 0.261 | 3.834 |
| HVV/GW | 4.083 | 1.411 | 0.396 | 2.893 | 0.007 | 1.197 | 6.970 | 0.585 | 1.709 |
| GW/SLV | −0.905 | 0.845 | −0.305 | −1.071 | 0.293 | −2.633 | 0.823 | 0.135 | 7.387* |
| GRWR | −37.594* | 32.665* | −0.295 | −1.151 | 0.259* | −104.400* | 29.213* | 0.166 | 6.011* |

*Refer to the main text for details. PVV/GW: peak portal venous flow velocity per 100 g of the initial graft weight, PSV/GW: peak systolic velocity of the hepatic artery per 100 g of the initial graft weight, EDV/GW: end diastolic velocity of the hepatic artery per 100 g of the initial graft weight, HVV/GW: peak hepatic venous flow velocity per 100 g of the initial graft weight, GW/SLV: graft-to-standard liver volume ratio, GRWR: graft-to-recipient weight ratio.

**Table 3B**. Collinearity Diagnostics

| Eigenvalue | Condition Index | Variance decomposition proportions | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Intercept | PVV/GW | PSV/GW | EDV/GW | HVV/GW | GW/SLV | GRWR |
| 6.164 | 1.000 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 0.555 | 3.332 | 0.00 | 0.00 | 0.02 | 0.07 | 0.00 | 0.00 | 0.00 |
| 0.119 | 7.209 | 0.00 | 0.09 | 0.00 | 0.25 | 0.45 | 0.00 | 0.00 |
| 0.099 | 7.883 | 0.00 | 0.74 | 0.02 | 0.02 | 0.24 | 0.00 | 0.00 |
| 0.043 | 11.938† | 0.01 | 0.01 | 0.87* | 0.55 | 0.22 | 0.00 | 0.00 |
| 0.017 | 18.975† | 0.47 | 0.09 | 0.08 | 0.11 | 0.04 | 0.00 | 0.15 |
| 0.003 | 47.323† | 0.51 | 0.06 | 0.02 | 0.00 | 0.04 | 0.99* | 0.84* |

*Refer to the main text for details, †Refer to the main text for details. PVV/GW: peak portal venous flow velocity per 100 g of the initial graft weight, PSV/GW: peak systolic velocity of the hepatic artery per 100 g of the initial graft weight, EDV/GW: end diastolic velocity of the hepatic artery per 100 g of the initial graft weight, HVV/GW: peak hepatic venous flow velocity per 100 g of the initial graft weight, GW/SLV: graft-to-standard liver volume ratio, GRWR: graft-to-recipient weight ratio.

**Table 4A**. Regression Coefficients of Multiple Linear Regression Model for Four Explanatory Variables Following the Exclusion of Two Variables

| | Unstandardized coefficients | Standard error | Standardized coefficients | t-statistic | P | 95% Confidence interval for the unstandardized coefficients | | Collinearity statistics | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Lower bound | Upper bound | Tolerance | Variance inflation factor |
| Intercept | 209.393 | 19.653 | | 10.655 | < 0.001 | 169.311 | 249.476 | | |
| PVV/GW | 0.392 | 0.593 | 0.088 | 0.661 | 0.514 | −0.817 | 1.601 | 0.599 | 1.670 |
| EDV/GW | 1.006 | 1.664 | 0.082 | 0.605 | 0.550 | −2.388 | 4.401 | 0.575 | 1.738 |
| HVV/GW | 4.410 | 1.239 | 0.428 | 3.559 | 0.001 | 1.882 | 6.937 | 0.738 | 1.355 |
| GRWR | −68.832 | 14.014* | −0.541 | −4.912 | < 0.001* | −97.413* | −40.251* | 0.879 | 1.137* |

*Refer to the main text for details. PVV/GW: peak portal venous flow velocity per 100 g of the initial graft weight, EDV/GW: end diastolic velocity of the hepatic artery per 100 g of the initial graft weight, HVV/GW: peak hepatic venous flow velocity per 100 g of the initial graft weight, GRWR: graft-to-recipient weight ratio.

**Table 4B**. Collinearity Diagnostics

| Eigenvalue | Condition Index | Variance decomposition proportions | | | | |
|---|---|---|---|---|---|---|
| | | Intercept | PVV/GW | EDV/GW | HVV/GW | GRWR |
| 4.409 | 1.000 | 0.00 | 0.01 | 0.01 | 0.01 | 0.00 |
| 0.369 | 3.456 | 0.01 | 0.01 | 0.39 | 0.00 | 0.03 |
| 0.107 | 6.410 | 0.02 | 0.07 | 0.38 | 0.69 | 0.05 |
| 0.096 | 6.766 | 0.00 | 0.85 | 0.18 | 0.30 | 0.00 |
| 0.018 | 15.697† | 0.97 | 0.06 | 0.03 | 0.01 | 0.91* |

*Refer to the main text for details, †Refer to the main text for details. PVV/GW: peak portal venous flow velocity per 100 g of the initial graft weight, EDV/GW: end diastolic velocity of the hepatic artery per 100 g of the initial graft weight, HVV/GW: peak hepatic venous flow velocity per 100 g of the initial graft weight, GRWR: graft-to-recipient weight ratio.

indicated with asterisks in Table 3B. Those corresponding to the highest condition index (condition number), i.e., 0.99 and 0.84, indicate that the most dominant linear dependence of the regression model is explained by 99% and 84% of the variance inflation in the regression coefficients of GW/SLV and GRWR. The strong linear dependence between the two explanatory variables is also supported by the highest Pearson's correlation coefficient ($R$ = 0.886) between them (Table 2). However, the second strongest correlation between PSV/GW and EDV/GW ($R$ = 0.841), which can be found in Table 2, does not seem to cause multicollinearity in the multiple linear regression model. Although the variance inflation factor of the PSV/GW (4.948) is lower than 5, it is very close to it. In addition, only one of their variance decomposition proportions corresponding to the condition index of 11.938 is over 0.8. However, if the cut-off value of the variance decomposition proportion for the diagnosis of multicollinearity is set to 0.3 according to the work of Liao et al. [6], the two explanatory variables are multicollinear. Therefore, excluding the GW/SLV from the regression model is justified, but whether the PSV/GW is removed from the regression model is not clear.

The exclusion of GW/SLV and PSV/GW produced a stable regression model (Table 4A) which is

*Liver regeneration rate*

$$= 209.393 + 0.392 \times PVV/GW + 1.006 \times EDV/GW + 4.410 \times HVV/GW - 68.832 \times GRWR$$

$$(R^2 = 0.669, P < 0.001)$$

All the variance inflation factors became less than 2. Particularly, the variance inflation factor of the regression coefficient for the GRWR was reduced from 6.011 to 1.137 with a decrease in its standard error from 32.665 to 14.014 and narrowing of its 95% confidence interval from (−104.4, 29.213) to (−97.413, −40.251). In accordance with the above changes, the probability value for the regression coefficient became less than 0.05 (from 0.259 to < 0.001). Although there is still a condition number of more than 10, only one variance decomposition proportion of more than 0.9 is present (Table 4B). It needs to be noted that the intercept term is not important for this analysis. The small change in the coefficient of determination ($R^2$) from 0.682 to 0.669 indicates a negligible loss of information.

## Conclusions

Multicollinearity distorts the results obtained from multiple

linear regression analysis. The inflation of the variances of the regression coefficients due to multicollinearity makes the coefficients statistically insignificant and widens their confidence intervals. Multicollinearity is determined to be present if the variance inflation factor and condition number are more than 5 to 10 and 10 to 30, respectively. However, they cannot detect which explanatory variables are multicollinear. To identify the variables with multicollinearity, the variance decomposition proportion is used. If the variance decomposition proportions of more than 0.8 to 0.9 correspond to the condition indices of more than 10 to 30, the explanatory variables, which are associated with the variance decomposition proportions corresponding to common condition indices, are multicollinear. In conclusion, the diagnosis of multicollinearity and exclusion of multicollinear explanatory variables enable the formulation of a reliable multiple linear regression model.

## Conflicts of Interest

No potential conflict of interest relevant to this article was reported.

## References


1. Kim JH, Kim TK, In J, Lee DK, Lee S, Kang H. Assessment of risk of bias in quasi-randomized controlled trials and randomized controlled trials reported in the Korean Journal of Anesthesiology between 2010 and 2016. Korean J Anesthesiol 2017; 70: 511-9.
2. Vatcheva KP, Lee M, McCormick JB, Rahbar MH. Multicollinearity in regression analyses conducted in epidemiologic studies. Epidemiology (Sunnyvale) 2016; 6: 227.
3. McDonald GC. Ridge regression. Wiley Interdiscip Rev Comput Stat 2009; 1: 93-100.
4. Byun SH, Yang HS, Kim JH. Liver graft hyperperfusion in the early postoperative period promotes hepatic regeneration 2 weeks after living donor liver transplantation: a prospective observational cohort study. Medicine (Baltimore) 2016; 95: e5404.
5. Sato Y, Tsukada K, Hatakeyama K. Role of shear stress and immune responses in liver regeneration after a partial hepatectomy. Surg Today 1999; 29: 1-9.
6. Liao D, Valliant R. Condition indexes and variance decompositions for diagnosing collinearity in linear model analysis of survey data. Surv Method 2012; 38: 189-202.

# 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 ($X_1$, $X_2$, … , $X_k$, $k > 2$, *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:

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

## Variance Inflation Factor

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + , \ldots , + \beta_k X_{ik} + \varepsilon_i \; (i = 1,2, \ldots , n) \quad \varepsilon_i \sim N(0, \sigma^2),$$

where $\beta_j (j = 0,1,2, \ldots , k)$ and $\varepsilon_i$ are the regression coefficients and error, respectively. Each error ($\varepsilon_1$, $\varepsilon_2$, … , $\varepsilon_n$) is stochastically independent and is normally distributed with a mean of 0 and a variance of $\sigma^2$. The variance of $\beta_j$ [$Var(\beta_j)$] 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 [$X_{ij} = \gamma_0 + \sum_{l=1}^{k} \gamma_l X_{il} + \epsilon_i \; (i = 1, 2, \ldots , n; \; l = 1,2, \ldots , k; \; l \neq j)$] with the response variable of $X_{ij}$, the explanatory variables of $X_{il}$, the regression coefficients of $\gamma_0$ and $\gamma_l$, and the error of $\epsilon_i$. Assuming that $\sum_{i=1}^{n} (X_{ij} - \overline{X}_j)^2$ and $\sigma^2$ 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 ($\beta_j$), 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 $Y_i=\beta_0+\beta_1X_{i1}+\beta_2X_{i2}+,\ldots,+\beta_kX_{ik}+\varepsilon_i$ ($i=1,2,\ldots,n$) 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:

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

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 ($\lambda_1$, $\lambda_2$) 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 ($\lambda_1, \lambda_2, \dots, \lambda_k$) of the correlation matrix can be calculated.

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

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

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

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

By solving the $k^{th}$ degree polynomial equation of the variable $\lambda$, we can obtain $k$ eigenvalues ($\lambda_1, \lambda_2, \dots, \lambda_k$). The number of eigenvalues ($\lambda_1, \lambda_2, \dots, \lambda_k$) 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 ($\lambda_1, \lambda_2, \dots, \lambda_k$) is termed "condition index" and is expressed as

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

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

## Variance Decomposition Proportion

Eigenvectors are calculated from their corresponding eigenvalues. The relationship between two eigenvalues ($\lambda_1, \lambda_2$) 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 ($v_1, v_2, \dots, v_k$) consisting of $k$ elements in one column, which correspond to $k$ eigenvalues ($\lambda_1, \lambda_2, \dots, \lambda_k$).

The eigenvector corresponding to the eigenvalue $\lambda_s$ ($s = 1, 2, \dots, k$) 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 $\beta_j$ ($j = 1, 2, \ldots, k$), which are defined as

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

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