AJO-DO

Multicollinearity

Michail Tsagris[a] and Nikolaos Pandis[b]
Rethymno, Greece, and Bern, Switzerland

When fitting a multiple linear regression model,[1,2] we must safeguard against the problem of multicollinearity. Multicollinearity is a phenomenon that occurs when 2 or more independent variables are highly (but not perfectly) correlated.[2] Multicollinearity can either inflate (or deflate) the standard errors of the coefficients, and as a result, the coefficients can, falsely, become nonsignificant (or significant). Another effect of multicollinearity is that of a sign change of the coefficient in which a negative effect can become positive and vice versa.

We can imagine a scenario in which we would like to predict the final after treatment overjet by fitting a regression model using several independent variables which can be highly correlated, such as initial overjet, ANB angle, Wits appraisal, and so on. To make the example more tangible and for simplicity, suppose there is 1 response variable (y) which is the final overjet, and 2 independent variables, x_1 and x_2, for which n measurements are available. We can further assume that x_1 is the patient's initial overjet and x_2 is another variable that is highly correlated with the initial overjet and that both x_1 and x_2 are continuous variables. The final overjet is related to the independent variables via a linear regression model:

final overjet $= \alpha + \beta_{1*}$ initial overjet $+ \beta_2 x_2 + e$, where e denotes the random error component.

In reality, the variable x_2 is a constructed variable, and we generated several versions of this variable with increasing correlation with the variable x_1. We fitted the same linear regression model above using the different x_2 variables with the increasing correlations

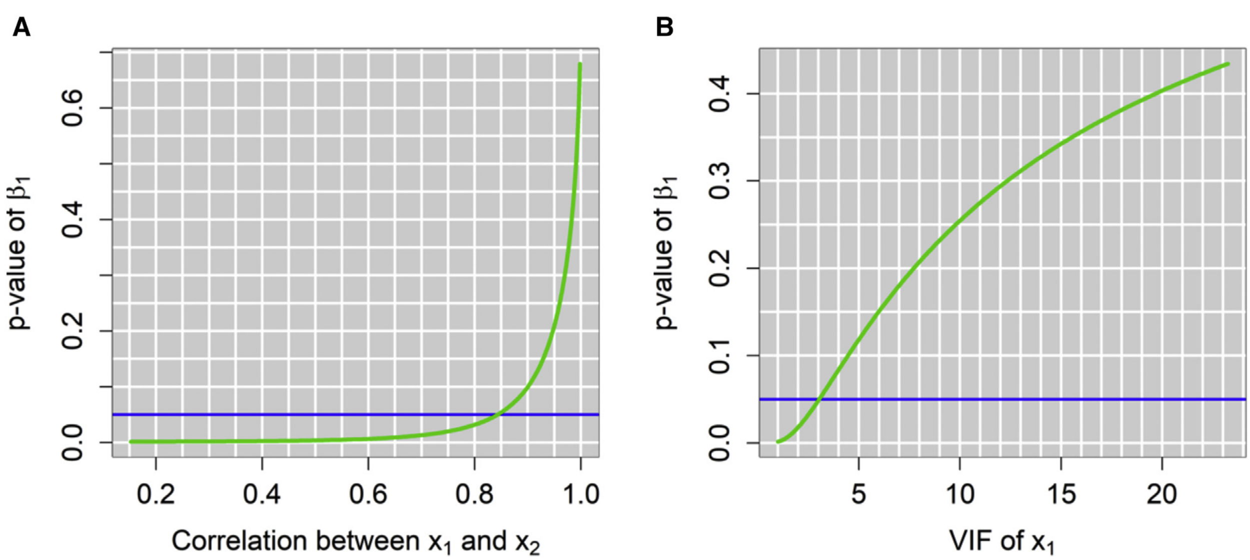

Fig. The *P* value of β_1 as a function of the correlation r between x_1 and x_2 **(A)** and as a function of VIF **(B)**. The *blue line* signifies the *P* value equal to 0.05.

[a]University of Crete, Rethymnon, Greece.
[b]University of Bern, Bern, Switzerland.
Address correspondence to: Nikolaos Pandis, Department of Orthodontics and Dentofacial Orthopedics, University of Bern, Freiburgstrasse 7, CH-3010 Bern, Switzerland; e-mail, npandis@yahoo.com.

Am J Orthod Dentofacial Orthop 2021;159:695-6
0889-5406/$36.00

https://doi.org/10.1016/j.ajodo.2021.02.005

with variable x_1. We plotted the *P* values of the coefficient β_1 from the different models vs the correlation between x_1 and x_2, and we can see an inversely proportional relationship between the *P* value and the correlation between x_1 and x_2. Figure, *A* visualizes this relationship and shows that as the correlation between x_1 and x_2 increases, so does the *P* value. Therefore, a significant and expected effect of the initial overjet on the final overjet disappears because of the presence of the variable x_2, which is highly correlated with x_1 (initial overjet). In this scenario, the variable x_2 should not be included in the model. In addition, it is interesting to note that the relationship in Figure, *A* is clearly nonlinear. This means that for every increase in the correlation by 1 factor, the *P* value increases by an increasingly larger factor. For small to moderate values, the increase in the *P* value is small, but when the correlation becomes larger than 0.84, the *P* value of β_1 exceeds the cutoff point of 0.05.

A second example is a regression model in which the distance walked by cardiovascular patients is predicted by the heart rate (HR) during the recuperation phase at minute 1 and minute 2 after the exercise.

M_1: distance = 492.9012 + 0.4761HR recuperation 2

M_2: distance = 469.171 − 32.687HR recuperation 2 + 32.617HR recuperation 1

The *P* value of HR recuperation 2 in M_1 is equal to 0.156, indicating that this variable is nonsignificant at the 0.05 significance level, whereas the *P* value of the same variable in M_2 is far smaller than 0.001 and hence this variable is highly statistically significant. The explanation of this phenomenon is the high correlation observed between recuperation 2 and recuperation 1, which is equal to 0.996.

To check the effect of multicollinearity, the variance inflation factor (VIF) must be computed for each variable. The minimum value of VIF equals 1 in the case of independent variables, whereas VIF increases with increasing correlations among the independent variables in Figure, *B*. A rule of thumb is that if the VIF for an independent variable is greater than 5 or 10, the multicollinearity of this variable is suspiciously high.

REFERENCES

1. Schmidt AF, Finan C. Linear regression and the normality assumption. J Clin Epidemiol 2018;98:146-51.
2. Wooldridge JM. Introductory Econometrics: A Modern Approach. 5th ed. Boston: Cengage Learning; 2012.