Korean J Fam Med. 2015;36:42-43
ISSN: 2005-6443 (Print) · ISSN: 2092-6715 (Online)

http://dx.doi.org/10.4082/kjfm.2015.36.1.42

Comments on Statistical Issues in January 2015

Commentary

Yong Gyu Park

Department of Biostatistics, The Catholic University of Korea College of Medicine, Seoul, Korea

In this section, we address the problem of multicollinearity in multiple regression analysisthat appeared in the article titled, "Correlation between frailty and cognitive function in non-demented community dwelling older Koreans," published in November 2014 by Kim et al.[1]

REVIEWER'S COMMENT: DID YOU CHECK THE MULTICOLLINEARITY?

This is one of the most frequent comments made about articles usingmultiple regression analysis. Multicollinearity indicates that independent (explanatory) variables are not mutually independent, but have some linearly correlated relationship. When we performthe multiple regression analysis,it is usually assumed that independent variables are deemed as given (fixed) values and only the dependent (response) variable is regarded as a random variable. This is why we adopt the term multicollinearity, instead of saying correlation,which denotes an association occurring among random variables. Depending on the situation, the term multicollinearitymay representa relationship of perfect linear combinations between independent variables. However,we confine its meaning tothe case of near-linear dependence.

It is foredoomed thatsome degree of associationwill exist among the independent variables in a multiple regression analysis. The very reason we perform a multiple regression analysis is to controlfor the interdependency among independent variables. If there is no associationamong the independent variables, we do not have to conduct the multiple regression analysis. The results from the multiple regression analysis will be the sameas those from simple regression analyses using independent variables one by one.

We do not have to worry about the multicollinearity problem when the degree of association between independent variables is not too high. However, in the case of an extremely high degree of association, some regression coefficients or their standard errors cannot be correctly calculated (estimated); that is, the phenomenaare such that no regression coefficients can be obtained, or extremely large standard errors in the analysis results might occur. In these cases, we say, "We could not obtain proper estimates from the multiple regression model due to the multicollinearity (near-linear dependency)."

Hence, the reviewer's comment, "Did you check the multicollinearity?" precisely implies that, "You might present the wrong analysis resultsdue to the multicollinearity. Did you check for this?" However, most authors provide tablesthat contain the final results of a multiple regression analysis. Therefore, if reviewers look at these tables carefully, they can find the problems caused by multicollinearityby themselves. Thus, if the problem is detected, it is appropriateeither to point out concretelythat the estimation of the regression model is wrong due to multicollinearity, or not to bring the issue up at all. If someone asksresearchersloosely whether they have checked for it, it could be said that he or she does not understand the concept of multicollinearity.

In performing multiple regression analysis, the most popular measure used to check for multicollinearity is the variance inflation factor (VIF). The VIF of independent variable (x_j) is defined as follows: $VIF_j = (1 - R_j^2)^{-1}$, where R_j^2 is the coefficient of determination obtained when xj is regressed on all the remaining independent variables. If x_j is nearly orthogonal to the remaining independent variables, R_j^2 is small and VIF_j is close to unity, while if x_j is nearly dependent on some subset of the remaining independent variables, R_j^2 is near unity and VIF_j is large. Practical experience indicates that if any of the VIFs exceeds 5 or 10, it is a sure sign that the associated regression coefficients are poorly estimated because of multicollinearity.[2)]

CONFLICT OF INTEREST

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

REFERENCES

1. Kim S, Park JL, Hwang HS, Kim YP. Correlation between frailty and cognitive function in non-demented community dwelling older Koreans. Korean J Fam Med 2014;35:309-20.
2. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. Hoboken (NJ): John Wiley & Sons, Inc.; 2006.