http://dx.doi.org/10.4082/kjfm.2015.36.6.357 • *Korean J Fam Med 2015;36:357-358*

KJFM
Korean Journal of Family Medicine

eISSN: 2092-6715

■ Commentary

Comments on Statistical Issues in November 2015

Kyung Do Han, Yong Gyu Park

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

In this section, we explain the definition and solution to avoid the multi-collinearity in multivariate analysis, which appeared in the articles titled, “Time to first cigarette and hypertension in Korean male smokers” and “Barrier factors to the completion of diabetes education in Korean diabetic adult patients: Korea National Health and Nutrition Examination Surveys 2007–2012”, published in September 2015 by Lee et al.[1] and by Kim et al.,[2] respectively.

WHAT IS AND HOW TO CHECK THE MULTI-COLLINEARITY?

Multi-collinearity indicates that independent (explanatory) variables are not mutually independent, but have some linearly correlated relationship in multiple (logistic) regression analysis. It is foredoomed that some degree of association will exist among the independent variables in the multivariate analysis. However, when the degree of association between independent variables is extremely high, some coefficients or their standard errors cannot be correctly calculated (estimated); that is, the phenomena are such that no 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 multivariate model due to the multi-collinearity (near-linear dependency).”[3]

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 x_j 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.[4]

HOW TO AVOID THE MULTI-COLLINEARITY?

The simplest and most intuitive method to avoid the multi-collinearity in analysis is using only independent variables with low correlation to each other. Firstly, these independent variables could be chosen by subjective method. For instance, when a researcher has to choose between Body Mass Index (BMI) and body weight, and his/her initial intention focused on BMI, then the independent variable should be the former, regardless of the variable which has higher correlation with the dependent variable.

Secondly, from the statistical point of view, a researcher can select the variable having the highest correlation with the dependent variable. The simplest way is to compare the values of correlation between the competing independent variables with a dependent variable. Also, the easiest method for selecting variables without multicollinearity, is applying a stepwise method in the variable selection for multivariate statistical analysis programs. However, if a researcher only based on statistical methods for variable selection, its final result would be far from the original intention of the researcher, or clinically unexplainable. On the other hand, ridge regression can handle analysis for the independent variables with multi-collinearity. However, ridge regression is generally not used because the estimated coefficients are biased and its method is not easy to understand.

CONFLICT OF INTEREST

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

REFERENCES

1. Lee S, Jang M, Noh HM, Oh HY, Song HJ, Park KH, et al. Time to first cigarette and hypertension in Korean male smokers. Korean J Fam Med 2015;36:221-6.
2. Kim HT, Lee K, Jung SY, Oh SM, Jeong SM, Choi YJ. Barrier factors to the completion of diabetes education in Korean diabetic adult patients: Korea National Health and Nutrition Examination Surveys 2007-2012. Korean J Fam Med 2015;36:203-9.
3. Park YG. Comments on statistical issues in January 2015. Korean J Fam Med 2015;36:42-3.
4. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. Hoboken (NJ): John Wiley & Sons Inc.; 2006.

http://dx.doi.org/10.4082/kjfm.2015.36.6.357