

# A Standardization Technique to Reduce the Problem of Multicollinearity in Polynomial Regression Analysis

Doo-Sub Kim

*Hanyang University, Department of Sociology*

*Sung Dong-Ku*

*Seoul, 133-791 KOREA*

*duskim@email.hanyang.ac.kr*

This paper attempts to explain how the problem of multicollinearity can be reduced in polynomial regression analysis. A simple standardization technique is illustrated to deal with curvilinearity and multicollinearity problems.

Suppose we have a cubic relationship to solve, and that the independent variable  $X$  has large values.

$$(1) \quad Y = c_0 + c_1X + c_2X^2 + c_3X^3$$

Then, it is very likely that the problem of multicollinearity arises since the independent variables  $X$ ,  $X^2$ , and  $X^3$  in equation (1) are collinear with each other. In case of severe multicollinearity, this regression equation cannot be solved since the rank of the matrix is less than its order. If the tolerance is very small, but still larger than the specified tolerance level, unstable *beta* weights (standardized regression coefficients) and calculations can be expected. In this case, multicollinearity can also be detected by size of the standard errors of *beta*.

The problem of multicollinearity can be removed or reduced substantially by standardizing the linear, quadratic, and cubic terms in the polynomial regression equation. First, it is suggested that the independent variable is transformed in such a way that the resulting mean is zero and the resulting standard deviation is one. For example, the independent variable  $X$  can be standardized by using the following linear transformation:

$$(2) \quad \text{Standardized Variable}(Z) = (X - \bar{X})/S_x$$

Then, new quadratic and cubic terms are created by taking squared and cubed values of this standardized variable.

$$(3) \quad \text{Squared Variable}(Z^2) = ((X - \bar{X})/S_x)^2$$

$$(4) \quad \text{Cubed Variable}(Z^3) = ((X - \bar{X})/S_x)^3$$

Then, a polynomial regression equation (1) becomes

$$(5) \quad Y = d_0 + d_1Z + d_2Z^2 + d_3Z^3.$$

The rationale is that, as illustrated in equation (2), no information is lost as a result of this standardization. The original metric can always be recovered when the mean and the standard deviation of the original variable are given. One of the advantages of this standardization is that multicollinearity among the linear, quadratic, and cubic terms is substantially reduced, while the correlation coefficients with other variables are not

affected by this transformation. After the above standardization, under the standard normal distribution, 99.73 percent of the sample falls in the range of  $Z$  value between  $-3$  and  $3$ , regardless of the value range of the original variable. Therefore, the quadratic and cubic terms of the standardized variable vary between  $0$  and  $9$ , and between  $-27$  and  $27$ , respectively. As a result, multicollinearity, as well as correlation coefficients among the linear, quadratic, and cubic terms of the standardized variable, declines substantially.

Theoretically, the equations (1) and (5) should provide the same fit and result in the same value of  $R^2$ . However, severe multicollinearity is likely to exist among  $X$ ,  $X^2$ , and  $X^3$  in equation (1). If so, one or more independent variables may be dropped from the stepwise regression procedures mechanically since their coefficients are not significantly different from zero. But the true situation may not be that the variable has no effect but simply that the set of sample data has not enabled us to pick it up.

To justify the above standardization, Kim(1987) tried to examine a cubic relationship between family income and fertility, by applying actual survey data to the regression models (1) and (5). It turned out that the empirical results from the these two regression equations are completely different. It was found that the quadratic term  $X^2$  in equation (1) was dropped since the specified tolerance level was not met by this variable. It was also indicated that the problem of multicollinearity in equation (1) resulted in a very large beta value for the cubic term.

Severe multicollinearity also results in reduced precision of estimation so that it becomes very difficult to disentangle the relative effect of each independent variable on the dependent variable. In addition, estimates of regression coefficients may become very sensitive to a particular set of sample data, and the sampling variances of the coefficients may be very large(Johnston, 1972: 160).

Another advantage of the above transformation into the standardized form is that calculations become more accurate. Without the above standardization, we may lose accuracy because of rounding errors in the course of calculating the variance or covariance. This is especially true when a variable with large values, such as income, is included as an independent variable in the regression equation, involving many variables and many cases.

For more discussion on the problems of multicollinearity and advantages of the standardization in this paper, see Kim(1987, 1993).

## REFERENCES

- Johnston, J. (1972). *Econometric Methods*. Second Edition. McGraw-Hill. New York.
- Kim, Doo-Sub. (1987). *Socioeconomic Status, Inequality and Fertility*. The Population and Development Studies Center, Seoul National University. Seoul, Korea.
- Kim, Doo-Sub. (1993). *Regression Analysis for the Social Sciences* (in Korean). Bupmunsa. Seoul, Korea.

Note: The author acknowledges his indebtedness to Dr. James Sakoda for his guidance and criticisms in the course of developing the main idea of this paper.