

Regression Multicollinearity: Sometimes it's helpful

Robert M Lynch and Brian Kim
Monfort College of Business
University of Northern Colorado
Greeley, CO 80639
robert.lynch@unco.edu

Many introductory textbooks include multiple regression topics with discussions about model building, forecasting and variable screening methods. Frequently the case is made that good predictors have the following characteristics:

(1) The independent variables X should be highly correlated with the dependent variable Y and uncorrelated with other independent variables.(1,2)

This is intuitively reasonable and usually provides a set of independent variables that may lead to a satisfactory regression model. At the same time, writers often suggest that:

(2) Independent variables that are highly correlated with each other demonstrate a condition referred to as multicollinearity and once one of the collinear variables is entered into the model, the entry of the second will demonstrate non-significant results and little, if any increase, in R^2 .
(1,2)

In this note, the authors provide an illustration suggesting this is not always the case and demonstrate that multicollinear independent variables taken together can improve forecasting and can reduce residual error substantially.

Illustration

Three variables, a Y , X_1 and X_2 were generated using Excel's random number generator. Two hundred fifty observations drawn from normal distributions for each of the three variables were generated. The tables below show the means, standard deviations, and correlations for the three variables.

Table 1
Means, standard deviations and correlations

	Y	X_1	X_2
<i>Means</i>	97.4410	208.4928	101.0905
<i>Standard Deviation</i>	20.2311	28.7478	20.8551
<hr/>			
<i>Correlations</i>	Y	X_1	X_2
Y	1.0000		
X_1	0.6503	1.0000	

X_2	-0.0872	0.6821	1.0000
-------	---------	--------	--------

From Table 1, one will note that variables Y and X_1 have a modest to high correlation while Y with X_2 has a low to near-zero correlation. By (1) above, this would suggest X_1 may be a viable predictor while X_2 may not. Similarly, the correlation between X_1 and X_2 is modest to high, demonstrating some multicollinearity as noted in (2) above. An initial review might suggest that X_1 is a good predictor of Y , X_2 is not and some multicollinearity is present.

Table 2 presents the multiple regression analysis for the three variables. From the table, R^2 approaches 1.00 and both variables are significant.

Table 2
Multiple regression analysis

<i>Regression Statistics</i>					
R^2	0.9497				
Standard Error	4.5546				
Observations	250				
<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	96,790.94	48,395.47	2,332.93	0.00
Residual	247	5,123.89	20.74		
Total	249	101,914.83			
	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	
Intercept	0.0260	2.1134	0.0123	0.9902	
X_1	0.9341	0.0137	68.0330	0.0000	
X_2	-0.9629	0.0189	-50.8758	0.0000	

From the correlations, the R^2 for the simple model Y with X_1 is $(.6503)^2 = .4229$ yet when X_2 is introduced into the model R^2 increases to .9497, an increase of .5268 in R^2 while the simple correlation between Y and X_2 is only -.0872. We have here a case where the collinearity among the independent variables serves to improve the overall R^2 and a largely uncorrelated variable with Y becomes an important contributor.

Discussion

Though the characteristics for good predictors described in (1) and (2) above are good, they are not complete. There are cases where a variable may be a good predictor yet be uncorrelated with the dependent variable. Moreover, some forms of collinearity may lead to substantial improvement in prediction when collinear variables are included in the model. Interestingly, both forward and backward elimination approaches to model building would have included X_2 in the final model.

When the writers discuss this with students, the question of what the data look like and how were the variables created arises. The authors simply created two variables drawn from normal populations with equal means and standard deviations. Variable Y was the first variate created, X_1 was the sum of the two variates created, and X_2 was the second variate created. A small random error was then added to X_1 to reduce R^2 below 1.00. Thus X_1 is related to Y and X_2 , while Y and X_2 were uncorrelated

This suggests that an independent variable which measures, say 2 components and the dependent variable measures one of the components while another independent variable measures the other, similar regression results can occur.

A simple illustration might be suggested by an instrument (eg, a test) that measures two uncorrelated dimensions, perhaps *spatial relations* and *verbal reasoning*. An investigator builds a model where *verbal reasoning* serves as the dependent variable, and *spatial relations* and a *total score* (*spatial relations* + *verbal reasoning*) serve as independent variables. The authors are not suggesting this is a good model building approach because *verbal reasoning* is present in both the dependent and independent variables but merely suggest it will yield the results noted above. This can also occur when one divides a variable by another and includes both in an analysis. For example, suppose one measures *total personal income* and *population*. Frequently a variable, *per capita income* is created. If *personal income* is unrelated to *population*, the *per capita income* measure will be correlated with both the *total personal income* and *population* variables. If *total personal income* was to serve as the dependent variable and *population* and *per capita income* were included as independent variables, the same behavior as noted above would surface.

References

1. D.A. Lind, W.G. Marchal and S.A. Wathen, *Statistical Techniques in Business and Economics*, 12th Edition. New York: McGraw Hill, 2005.
2. Fred Kerlinger, *Foundations of Behavioral Research*, 2nd Edition. New York: Holt, Rinehart and Winston. 1973.