The Effect of Multicollinearity on Prediction in Regression Models

Daniel J. Mundfrom **Michelle DePoy Smith** **Lisa W. Kay**
Eastern Kentucky University

It has long been known and there is ample literature in support of the notion that the presence of multicollinearity in a dataset can, and often will, have detrimental effects on one's ability to determine which of the model predictors are actually responsible for, or contributing to, the variation in the measured/observed response (Montgomery, Peck, & Vining, 2001; Pedhazur, 1982). There also exist some indications that the presence of multicollinearity in the data does not, or at least may not, impact one's ability to accurately estimate or predict the value of the response variable for any specific set of measurements/observations on the predictors (Kutner, Nachtsheim & Neter, 2004; Weiss, 2012). This idea, although seemingly logical on the face of it, is not widely present in regression textbooks, nor is there an abundance of research literature that supports it. The purpose of this study was to examine this relationship, or lack thereof, in a variety of situations that vary in the number of predictors, the strength of the association between the predictors and the response, the size of the sample, and the level of the multicollinearity among the predictors.

Virtually every statistics textbook that includes chapters on multiple regression at least touches on the concept of multicollinearity and the problems that it can cause in arriving at an acceptable model. The focus of these discussions is almost unilaterally restricted to the determination of which independent variables are needed/appropriate in an optimal model and which are unnecessary because of their inter-connectedness to other independent variables in the model (Adeboye, Fagoyinbo, & Olatayo, 2014). Various procedures or "rules" are presented to aid the researcher in deciding which variables to keep and which ones can be discarded, always, and usually, in the context of arriving at a reduced model that will still adequately predict/explain the desired response with each independent variable making its own unique, "sizeable" contribution to that prediction or explanation (Montgomery, Peck, & Vining, 2001; Willis, & Perlack, 1978).

Some textbooks differentiate between an effect that multicollinearity may have on the ability to determine an optimal set of predictors and an effect it may have on predicting or estimating the value of the response variable. When this distinction is addressed, the typical statement is somewhere along the lines of such an effect on the prediction of the response is negligible or non-existent (Frost, 2013; Kutner, Nachtsheim & Neter, 2004; Weiss, 2012). It is rare to see any justification or empirical evidence in support of such claims.

Whereas it is not suggested here that these assertions either are, or may not be, true, it seems prudent to see if such claims can be supported by data or, if not, under what conditions, multicollinearity may have some effect on the ability to accurately predict or estimate the value of the response. This examination is not exhaustive of all possible regression scenarios involving multiple predictors at various levels of multicollinearity. Rather, it is a first step in an exploration of whether or not a potential effect of multicollinearity on prediction is something about which researchers and data analysts need to be concerned.

Methods

Two different regression models were investigated in this study. The first model was a two-variable model in which a single variable, X_2, which was collinear with the existing variable, X_1, in a simple linear regression model, was added to the model to create a model in which both variables were relatively highly correlated with the response variable, Y, and also moderately to highly correlated with each other. These two models are respectively: $Y_1 = b_0 + b_1X_1 + e$ and $Y_2 = b_0 + b_1X_1 + b_2X_2 + e$.

The second model was a three-variable model in which a single variable, X_3, which was collinear with both of the existing variables, X_1 and X_2, was added to the model to create a model in which all three variables were relatively highly correlated with the response variable, Y; X_1 was moderately correlated with both X_2 and X_3; and the correlation between X_2 and X_3 was varied from being relatively uncorrelated with each other to being very highly correlated with each other. These two models are respectively: $Y_1 = b_0 + b_1X_1 + b_2X_2 + e$ and $Y_2 = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$.

In the two-variable model, correlations between Y and X_1 were varied across the values 0.8, 0.85, and 0.9; correlations between Y and X_2 were varied across the values 0.7, 0.75, 0.8, 0.85, and 0.9; correlations between X_1 and X_2 were varied across the values 0.3, 0.5, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. The cases in which the values of the correlation between X_1 and X_2 were set at 0.3 and 0.5 were used as baseline conditions, in which the two independent variables were not collinear in an effort to better understand the effect of introducing an additional independent variable into a model which was collinear with the previous independent variable.

In the three-variable model, correlations between Y and X_1 were varied across the values 0.8 and 0.9; correlations between Y and X_2 were varied across the values 0.7, 0.75, and 0.8; correlations between Y and X_3 were varied across the values 0.7 and 0.75; correlations between X_1 and X_2 and between X_1 and X_3 were fixed at 0.5; and correlations between X_2 and X_3 were varied across the values 0.3, 0.5, 0.7, 0.75, 0.8, 0.85, 0.9, and 0.95. Again, the cases in which the values of the correlation between X_1 and X_3 were set at 0.3 and 0.5 were used as baseline conditions, in which the three independent variables were not collinear for the same reason as stated above with the two-variable model.

Sample sizes were set at 20, 50, and 100 in all the scenarios investigated for both the two-variable models and the three-variable models. Although it typically is probably not the case that a collinear variable is treated as being added to a model that already contains one or two independent variables; in order to control the conditions of this study, that method is what was employed. In conjunction with that, in order to see the effect of the additional collinear variable, the correlation between the independent variable(s) and Y had to be greater than or equal to the correlation between the collinear variable and Y in order for the correlation coefficient between the two predicted values of Y to be comparable. It may seem that these conditions are limiting in terms of the generalizability of the findings, but it is merely an artifact of creating specific scenarios for comparison purposes.

Data

Initially, data were generated from a Multivariate Normal Distribution (MVN) with mean = 0, variance of Y = 25, variance of X_1 = 9, variance of X_2 = 4, variance of X_3 = 16, and covariances determined by the given correlations. Next, data were generated from a multivariate *t* distribution for two variables for 3 and 5 degrees of freedom, and for three variables for 3 and 5 degrees of freedom, thereby creating distributions with somewhat heavier tails. Further, data were generated from a multivariate uniform distribution (marginals were standard uniform) for two variables and three variables; hence, the distribution had even thicker tails than the multivariate *t* distributions. For all the combinations of conditions described above in each of the three sample sizes previously mentioned and for each of the three distributions, 2000 replications were simulated using R (Mundfrom et al, 2011).

Results

For the two-variable model and for each of the combinations of conditions, we used R to generate a matrix of results containing the original values of Y, X_1, and X_2, the predicted values of Y using the SLR model and two-variable MLR model denoted by $\hat{y}_1$ and $\hat{y}_2$, respectively, the endpoints of a confidence interval based on Y_1, and the endpoints of a confidence interval based on Y_2. Next, we computed the confidence interval widths based on Y_1 and Y_2. Finally, we computed the ratio of the mean difference in the confidence interval widths to the standard deviation of the Y values. The correlations between Y with $\hat{y}_1$ and $\hat{y}_2$, between $\hat{y}_1$ and $\hat{y}_2$, and between all the variables in the model were also computed. The results in the following tables are selected representative results for a variety of treatment conditions.

Summary Statistics of Simulations

For the simulations computed we found minimal differences (maximum difference being 0.094) between the sample correlations and model correlations. In addition, we computed the mean, median, minimum, and maximum of the ratios of the mean difference in the confidence interval widths to the standard deviation of the y values and the correlations between $\hat{y}_1$ and $\hat{y}_2$.

Tables 3-6 display the mean, minimum, and maximum sample correlations for each of the model conditions investigated herein that were obtained in the replications simulated in this study.

Table 1. Two-Variable Model Results (Average of 2000 Simulations).

$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1,$ $\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$. $\rho_{Y,\ X_1} = 0.8, \rho_{Y,\ X_2} = 0.75$

		Multivariate Normal		Multivariate *t df*=5		Multivariate *t df*=3		Multivariate Uniform	
$\boldsymbol{\rho_{X_1,\ X_2}}$	$\boldsymbol{n}$	$\boldsymbol{r_{\hat{y}_1, \hat{y}_2}}$	$\mathrm{M_DCIW_{SD}}$	$\boldsymbol{r_{\hat{y}_1, \hat{y}_2}}$	$\mathrm{M_DCIW_{SD}}$	$\boldsymbol{r_{\hat{y}_1, \hat{y}_2}}$	$\mathrm{M_DCIW_{SD}}$	$\boldsymbol{r_{\hat{y}_1, \hat{y}_2}}$	$\mathrm{M_DCIW_{SD}}$
0.70	20	0.934	−0.081	0.924	−0.063	0.914	−0.045	0.948	−0.105
0.70	50	0.944	−0.046	0.938	−0.036	0.930	−0.026	0.949	−0.052
0.70	100	0.946	−0.031	0.943	−0.025	0.938	−0.019	0.949	−0.034
0.75	20	0.945	−0.105	0.936	−0.087	0.929	−0.071	0.959	−0.128
0.75	50	0.957	−0.062	0.951	−0.053	0.941	−0.040	0.962	−0.068
0.75	100	0.959	−0.043	0.957	−0.036	0.949	−0.029	0.962	−0.045
0.80	20	0.957	−0.130	0.950	−0.116	0.938	−0.096	0.969	−0.150
0.80	50	0.970	−0.078	0.965	−0.067	0.954	−0.054	0.974	−0.083
0.80	100	0.972	−0.054	0.969	−0.047	0.961	−0.038	0.975	−0.056
0.85	20	0.970	−0.156	0.958	−0.132	0.943	−0.108	0.977	−0.167
0.85	50	0.981	−0.090	0.975	−0.079	0.964	−0.066	0.985	−0.096
0.85	100	0.984	−0.062	0.981	−0.056	0.972	−0.045	0.986	−0.065
0.90	20	0.977	−0.172	0.966	−0.151	0.956	−0.129	0.981	−0.178
0.90	50	0.990	−0.100	0.985	−0.089	0.973	−0.075	0.992	−0.104
0.90	100	0.993	−0.070	0.990	−0.063	0.982	−0.053	0.995	−0.071
0.95	20	0.982	−0.178	0.968	−0.154	0.953	−0.132	0.978	−0.179
0.95	50	0.993	−0.104	0.988	−0.092	0.975	−0.078	0.991	−0.106
0.95	100	0.996	−0.072	0.993	−0.065	0.983	−0.054	0.996	−0.074

Note. The same statistics were calculated for the three-variable model, where in these cases, the predicted values of Y_1 and Y_2, are the predicted values from the MLR model with two independent variables and the MLR model with three independent variables respectively. $\mathrm{M_DCIW_{SD}}$ is the Mean Difference in Confidence Interval Widths divided by the Standard Deviation of *y*.
$\mathrm{M_DCIW_{SD}}$ = Mean[CIW($\hat{y}_1$) - CIW($\hat{y}_2$)]/SD(y).

Conclusions

It does not appear that the effect of multicollinearity on the value of the predicted response is as simple as the textbooks convey. Clearly, including a collinear variable will decrease the degrees of freedom for the squared error term by one while not significantly reducing the error. The average differences in confidence interval widths obtained in the simulations were greater than the difference we would obtain solely as the result of the loss of one degree of freedom for the squared error term. Hence, this seems to quantify that the impact of multicollinearity on the confidence interval widths is more than just negligible or non-existent. Normality tests did not suggest rejection of the null hypothesis of normality for many scenarios with the multivariate *t* and multivariate uniform data; this was especially true for *t* with *df* = 5. In many cases, level of normality does not appear to make much of a difference in confidence interval width for predictions based on the data.

From these data it does appear that multicollinearity does have an effect on that prediction in at least some, if not most, of the scenarios studied. The basic struggle we faced is how to best quantify that effect. We are not sure that we have adequately conquered that struggle.

Three outcomes, however, are quite clear from our data:

One, the size of the sample has an effect, with larger samples appearing to mitigate, to some extent at least, the effect of the multicollinearity of the predicted value of Y. Specifically, it appears to affect confidence interval width for smaller sample sizes more than it does for larger sample sizes, and it makes a bigger difference in confidence interval width for the model with two predictor variables than for the model with three predictor variables.

Two, there is an effect of the mean difference in the width of the confidence intervals based on the predicted values of Y_1 and Y_2, with the wider interval being associated with the “collinear” model.

Lastly, the presence of multicollinearity in the data appears to have a larger effect with fewer variables in the model. Specifically, the width of the confidence intervals for the mean difference between the predicted values for Y_1 and Y_2 were wider in the two-variable model than in the three-variable model.

Table 2. Three-Variable Model Results (Average of 2000 Simulations).

$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, $\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$

$\rho_{Y,\,X_1} = 0.8$, $\rho_{Y,\,X_2} = 0.75$, $\rho_{Y,\,X_3} = 0.7$, $\rho_{X_1,\,X_2} = 0.5$, $\rho_{X_1,\,X_3} = 0.5$

		Multivariate Normal		Multivariate *t df*=5		Multivariate *t df*=3		Multivariate Uniform	
$\boldsymbol{\rho_{X_1,\,X_2}}$	$\boldsymbol{n}$	$\boldsymbol{r_{\hat{y}_1,\hat{y}_2}}$	$\text{M}_\text{D}\text{CIW}_\text{SD}$	$\boldsymbol{r_{\hat{y}_1,\hat{y}_2}}$	$\text{M}_\text{D}\text{CIW}_\text{SD}$	$\boldsymbol{r_{\hat{y}_1,\hat{y}_2}}$	$\text{M}_\text{D}\text{CIW}_\text{SD}$	$\boldsymbol{r_{\hat{y}_1,\hat{y}_2}}$	$\text{M}_\text{D}\text{CIW}_\text{SD}$
0.70	20	0.983	−0.081	0.976	−0.051	0.982	−0.070	0.952	−0.049
0.70	50	0.988	−0.047	0.983	−0.032	0.986	−0.040	0.953	−0.014
0.70	100	0.989	−0.032	0.986	−0.022	0.988	−0.028	0.954	−0.006
0.75	20	0.986	−0.095	0.981	−0.066	0.983	−0.080	0.955	−0.051
0.75	50	0.991	−0.054	0.987	−0.039	0.989	−0.047	0.955	−0.017
0.75	100	0.992	−0.037	0.988	−0.027	0.991	−0.033	0.955	−0.008
0.80	20	0.990	−0.107	0.984	−0.076	0.988	−0.091	0.956	−0.058
0.80	50	0.994	−0.061	0.989	−0.045	0.993	−0.054	0.956	−0.019
0.80	100	0.995	−0.042	0.992	−0.031	0.994	−0.037	0.957	−0.012
0.85	20	0.992	−0.114	0.986	−0.083	0.989	−0.097	0.961	−0.068
0.85	50	0.996	−0.066	0.991	−0.048	0.994	−0.058	0.961	−0.027
0.85	100	0.998	−0.045	0.994	−0.034	0.997	−0.040	0.960	−0.017
0.90	20	0.993	−0.117	0.985	−0.084	0.989	−0.101	0.966	−0.082
0.90	50	0.997	−0.067	0.992	−0.049	0.995	−0.060	0.965	−0.036
0.90	100	0.999	−0.047	0.995	−0.036	0.998	−0.042	0.966	−0.024
0.95	20	0.989	−0.102	0.981	−0.071	0.985	−0.085	0.971	−0.096
0.95	50	0.993	−0.058	0.988	−0.043	0.991	−0.051	0.971	−0.047
0.95	100	0.994	−0.040	0.991	−0.030	0.993	−0.036	0.971	−0.032

Table 3. Generated Sample Statistics the Ratios of the Mean Difference in the Confidence Interval Widths to the Standard Deviation of the *y* Values for the Two-Variable Model

Model	**Count**	**Mean**	**Median**	**Maximum**	**Minimum**
Multivariate Normal	213	-0.055	-0.058	0.219	-0.178
Multivariate *t df* = 5	213	-0.045	-0.049	0.221	-0.157
Multivariate *t df* = 3	213	-0.034	-0.039	0.226	-0.136
Multivariate Uniform	216	-0.074	-0.069	0.015	-0.188

Table 4. Generated Sample Statistics the Correlation $\hat{y}_1$ and $\hat{y}_2$ for the Two-Variable Model

Model	**Count**	**Mean**	**Median**	**Maximum**	**Minimum**
Multivariate Normal	213	0.970	0.976	0.999	0.908
Multivariate *t df* = 5	213	0.966	0.971	0.997	0.899
Multivariate *t df* = 3	213	0.959	0.963	0.993	0.890
Multivariate Uniform	216	0.976	0.982	0.998	0.922

Table 5. Generated Sample Statistics the Ratios of the Mean Difference in the Confidence Interval Widths to the Standard Deviation of the *y* Values for the Three-Variable Model

Model	**Count**	**Mean**	**Median**	**Maximum**	**Minimum**
Multivariate Normal	87	-0.044	-0.041	0.062	-0.117
Multivariate *t df* = 5	87	-0.029	-0.030	0.065	-0.084
Multivariate *t df* = 3	87	-0.037	-0.036	0.063	-0.101
Multivariate Uniform	234	-0.021	-0.016	0.020	-0.094

Table 6. Generated Sample Statistics the Correlation $\hat{y}_1$ and $\hat{y}_2$ for the Three-Variable Model

Model	Count	Mean	Median	Maximum	Minimum
Multivariate Normal	87	0.991	0.993	1.000	0.967
Multivariate *t df* = 5	87	0.987	0.989	0.999	0.965
Multivariate *t df* = 3	87	0.989	0.992	1.000	0.964
Multivariate Uniform	234	0.966	0.965	0.990	0.944

References

Adeboye, N. O., Fagoyinbo, I. S., & Olatayo, T. O. (2014). Estimation of the effect of multicollinearity on the standard error of regression coefficients. *IOSR Journal of Mathematics*, *10*(4), 16-20.

Frost. J. (2013, May 2). What are the effects of multicollinearity and when can I ignore them?

[Blog post]. Retrieved from http://blog.mimitab.com/blog/adventures-in-statistics/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them

Kutner, M. H., Nachtsheim, C. J., & Neter, J. 2004. *Applied linear regression models* (4th ed.). New York: McGraw-Hill/Irwin Series.

Montgomery, D. C., Peck, E. A., & Vining, C. G. (2001). *Introduction to linear regression analysis* (3rd ed.). New York: Wiley.

Mundfrom, D., Schaffer, J., Shaw, D., Preecha, C., Ussawarujikulchai, A, Supawan, P., & Kim, M. (2011). Number of replications required in Monte Carlo simulation studies: A synthesis of four studies. *Journal of Modern Applied Statistical Methods*, *10*(1), 19-28.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd ed.). New York: Holt, Rinehart, and Winston.

Weiss, N. A. (2012). *Introductory statistics* (10th ed.). Boston: Pearson.

Willis, C. E., & Perlack, R. D. (1978). Multicollinearity: Effects, symptoms, and remedies. *Journal of the Northeastern Agricultural Economics Council*, *7*(1), 55-61.

Send correspondence to:

Daniel J. Mundfrom
Eastern Kentucky University
Email: daniel.mundfrom@eku.edu