Assessing the Impact of Multicollinearity in Various Regression Scenarios

Lisa W. Kay[1], Daniel J. Mundfrom[1], Michelle L. DePoy Smith[1]
[1]Eastern Kentucky University, 521 Lancaster Avenue, Richmond, KY 40475

Abstract
Although it is well known that multicollinearity can impede one's ability to evaluate model predictors (Montgomery, Peck, & Vining, 2001; Pedhazur, 1982), it has been suggested that the presence of multicollinearity may not affect the accuracy of a prediction of the response variable, given a set of observations taken on the predictor variables (Kutner, Nachtsheim & Neter, 2004; Weiss, 2012). In a previous study, the authors explored a model's ability to make predictions under different scenarios by varying the number of predictors, the strength of the association between the predictor variables and the response variable, the sample size, and the level of multicollinearity (Mundfrom, Smith & Kay, 2016). Simulations in the study indicated that confidence intervals are wider in the presence of multicollinearity. Furthermore, differences in confidence interval width appeared to depend on degree of taintedness in multivariate normal data, sample size, and number of predictors in the model. The purpose of the present study was to determine which scenarios result in more appreciable effects and to examine alternative ways of measuring the impact of multicollinearity on predictions.

Key Words: multicollinearity, regression, model, predictors

1. Introduction

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). Although it is well known that multicollinearity can impede one's ability to evaluate model predictors (Montgomery, Peck, & Vining, 2001; Pedhazur, 1982), it has been suggested that the presence of multicollinearity may not affect the accuracy of a prediction of the response variable, given a set of observations taken on the predictor variables (Kutner, Nachtsheim, & Neter, 2004; Weiss, 2012). In a previous study, the authors explored a model's ability to make predictions under different scenarios by varying the number of predictors, the strength of the association between the predictor variables and the response variable, the sample size, the level of multicollinearity, and departures from normality (i.e., taintedness) (Mundfrom, Smith, & Kay, 2016). Simulations in the study indicated that confidence intervals are wider in the presence of multicollinearity. Furthermore, differences in confidence interval width appeared to depend on degree of taintedness in multivariate normal data, sample size, and number of predictors in the model. The present study extends

the authors' previous work to include data from multivariate *t* and multivariate uniform distributions to further examine the effect of multicollinearity with non-normal data.

2. Methods

As in the authors' previous work (Mundfrom, Smith, & Kay, 2016), 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, $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.

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, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ and $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

Simulations were conducted with different combinations of the following: number of quantitative predictor variables (2, 3), correlation between each predictor variable and response variable (0.7, 0.75, 0.8, 0.85, 0.9), and correlation between two predictor variables (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 for the two-variable model 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. Similarly, the cases in which the values of the correlation between X_1 and X_3 for the three-variable model were set at 0.3 and 0.5 were used as baseline conditions. 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 is probably not typically 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 a result of creating specific scenarios for comparison purposes.

3. Data

Initially, data were generated from a multivariate normal distribution with $\mu_Y = \mu_{X_1} = \mu_{X_2} = \mu_{X_3} = 0$, $\sigma_Y^2 = 25$, $\sigma_{X_1}^2 = 9$, $\sigma_{X_2}^2 = 4$, $\sigma_{X_3}^2 = 16$, and covariances determined by the given correlations. Then, 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, Schaffer, Shaw, Preecha, Ussawarujikulchai, Supawan, & Kim, (2011).

4. 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_1 and Y_2, the predicted values from the SLR model and the two-variable MLR model respectively, the correlation between the predicted values for the two models, the endpoints of a confidence interval based on Y_1, the endpoints of a confidence interval based on Y_2, and the ratio of the mean difference in the confidence interval widths to $SD(y)$. The results in the following tables are selected representative results for a variety of treatment conditions.

Table 1. Two-Variable Model Results

$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, \quad \hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$
$\rho_{Y,\,X_1} = 0.8, \rho_{Y,\,X_2} = 0.75$
Average of 2000 Simulations

		Multivariate Normal Data		Multivariate *t* Data *df* = 5		Multivariate *t* Data *df* = 3		Multivariate Uniform Data	
$\rho_{X_1,\,X_2}$	n	$r_{\hat{y}_1,\,\hat{y}_2}$	Mean [CIW($\hat{y}_1$) − CIW($\hat{y}_2$)] /SD(y)	$r_{\hat{y}_1,\,\hat{y}_2}$	Mean [CIW($\hat{y}_1$) − CIW($\hat{y}_2$)] /SD(y)	$r_{\hat{y}_1,\,\hat{y}_2}$	Mean [CIW($\hat{y}_1$) − CIW($\hat{y}_2$)] /SD(y)	$r_{\hat{y}_1,\,\hat{y}_2}$	Mean [CIW($\hat{y}_1$) − CIW($\hat{y}_2$)] /SD(y)
0.7	20	0.934	−0.081	0.924	−0.063	0.914	−0.045	0.948	−0.105
0.7	50	0.944	−0.046	0.938	−0.036	0.930	−0.026	0.949	−0.052
0.7	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
0.8	20	0.957	−0.130	0.950	−0.116	0.938	−0.096	0.969	−0.150
0.8	50	0.970	−0.078	0.965	−0.067	0.954	−0.054	0.974	−0.083
0.8	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
0.9	20	0.977	−0.172	0.966	−0.151	0.956	−0.129	0.981	−0.178
0.9	50	0.990	−0.100	0.985	−0.089	0.973	−0.075	0.992	−0.104
0.9	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

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.

Table 2. Three-Variable Model Results

$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, $\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$ $\rho_{Y,\,X_1} = 0.8$, $\rho_{Y,\,X_2} = 0.75$, $\rho_{Y,\,X_3} = 0.7$, $\rho_{X_1,\,X_2} = 0.5$, $\rho_{X_1,\,X_3} = 0.5$ Average of 2000 Simulations									
		Multivariate Normal Data		**Multivariate *t* Data *df* = 5**		**Multivariate *t* Data *df* = 3**		**Multivariate Uniform Data**	
$\rho_{X_2,\,X_3}$	n	$r_{\hat{y}_1,\,\hat{y}_2}$	**Mean [CIW($\hat{y}_1$) − CIW($\hat{y}_2$)] /SD(y)**	$r_{\hat{y}_1,\,\hat{y}_2}$	**Mean [CIW($\hat{y}_1$) − CIW($\hat{y}_2$)] /SD(y)**	$r_{\hat{y}_1,\,\hat{y}_2}$	**Mean [CIW($\hat{y}_1$) − CIW($\hat{y}_2$)] /SD(y)**	$r_{\hat{y}_1,\,\hat{y}_2}$	**Mean [CIW($\hat{y}_1$) − CIW($\hat{y}_2$)] /SD(y)**
0.7	20	0.983	−0.081	0.976	−0.051	0.982	−0.070	0.952	−0.049
0.7	50	0.988	−0.047	0.983	−0.032	0.986	−0.040	0.953	−0.014
0.7	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
0.8	20	0.990	−0.107	0.984	−0.076	0.988	−0.091	0.956	−0.058
0.8	50	0.994	−0.061	0.989	−0.045	0.993	−0.054	0.956	−0.019
0.8	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
0.9	20	0.993	−0.117	0.985	−0.084	0.989	−0.101	0.966	−0.082
0.9	50	0.997	−0.067	0.992	−0.049	0.995	−0.060	0.965	−0.036
0.9	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

5. Conclusions

Results of the current study seem consistent with the results of the authors' previous study (Mundfrom, Smith, & Kay, 2016). Multicollinearity has an effect on prediction in at least some scenarios. It is not clear how best to quantify the effect of multicollinearity on prediction, but several ways seem to be informative. Simulations indicate that confidence intervals are wider in the presence of multicollinearity. The $r_{\hat{y}_1, \hat{y}_2}$ values are very high; they indicate stronger correlations than for the authors' previous study that employed "tainted" multivariate normal data. 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. Multicollinearity 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.

References

Adeboye, N. O., Fagoyinbo, I. S., & Olatayo, T. O. (2014). Estimation of the effect of multicollinearity on the standard error for regression coefficients. *IOSR Journal of Mathematics*, *10*(4): 16−20. http://www.iosrjournals.org/iosr-jm/papers/Vol10-issue4/Version-1/D010411620.pdf

Dr. John Ruscio Professional + Personal Pages. (n.d.). Programs. Retrieved from <http://ruscio.pages.tcnj.edu/quantitative-methods-program-code/>

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

Montgomery, D. C., Peck, E. A., & Vining, G. 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.

Mundfrom, D. J., Smith, M. L. D., & Kay, L. W. (2016). The effect of multicollinearity on prediction in various regression scenarios. *JSM 2016 Proceedings of the Section of Statistical Programmers and Analysts*, 3282−3287.

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

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