The Effect of Multicollinearity on Prediction in Regression Models

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

Abstract
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.

Key Words: multicollinearity, regression analysis, simulations

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, N. O., Fagoyinbo, I. S., & Olatayo, T. O., 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, C. E. & Perlack, R. D., 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 (Kutner, Nachtsheim & Neter, 2004; Weiss, 2012; Frost, 2013). 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.

2. 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: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ 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 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$.

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.

3. 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. Then this distribution was "tainted" by initially replacing 5% of randomly selected observations with a new observation formed by adding to it a randomly generated quantity from a normal distribution with mean = 0 and standard deviation = 8, thereby creating a distribution with somewhat heavier tails. A third distribution was also created in a similar manner with 10% of the original observations being replaced with a "tainted" value from a normal distribution with mean = 0 and standard deviation = 8 resulting in a distribution which in turn had somewhat heavier tails than the first "tainted" distribution.

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 endpoints of a confidence interval based on Y_1, and the endpoints of a confidence interval based on Y_2. The results in the following tables are selected representative results for a variety of treatment conditions. More complete results will be included in the final version.

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 SD(y) = 4.964		**Multivariate Normal Data 5% Tainted Data SD(y)= 5.261**		**Multivariate Normal Data 10% Tainted Data SD(y)= 5.545**	
$\rho(x_1,x_2)$	n	$r(\hat{y}_1, \hat{y}_2)$	Mean[$CIW(\hat{y}_1)$ - $CIW(\hat{y}_2)$]	$r(\hat{y}_1, \hat{y}_2)$	Mean[$CIW(\hat{y}_1)$ - $CIW(\hat{y}_2)$]	$r(\hat{y}_1, \hat{y}_2)$	Mean[$CIW(\hat{y}_1)$ - $CIW(\hat{y}_2)$]
0.3	20	0.826	1.673	0.806	0.427	0.803	-0.241
0.3	50	0.829	1.012	0.823	0.127	0.821	-0.232
0.3	100	0.830	0.717	0.833	0.048	0.837	-0.187
0.5	20	0.881	0.351	0.843	-0.130	0.821	-0.514
0.5	50	0.887	0.231	0.863	-0.125	0.846	-0.332
0.5	100	0.891	0.152	0.876	-0.112	0.864	-0.253
0.7	20	0.934	-0.373	0.881	-0.480	0.846	-0.683
0.7	50	0.943	-0.226	0.900	-0.297	0.882	-0.425
0.7	100	0.947	-0.153	0.912	-0.206	0.895	-0.294
0.75	20	0.948	-0.513	0.886	-0.522	0.855	-0.724
0.75	50	0.956	-0.304	0.906	-0.316	0.889	-0.440
0.75	100	0.959	-0.212	0.921	-0.223	0.901	-0.303
0.8	20	0.959	-0.637	0.894	-0.595	0.860	-0.732
0.8	50	0.970	-0.381	0.917	-0.342	0.892	-0.440
0.8	100	0.973	-0.268	0.930	-0.239	0.910	-0.312
0.85	20	0.970	-0.735	0.904	-0.641	0.859	-0.762
0.85	50	0.980	-0.443	0.928	-0.357	0.897	-0.434
0.85	100	0.984	-0.310	0.938	-0.239	0.914	-0.304
0.9	20	0.980	-0.835	0.913	-0.689	0.860	-0.755
0.9	50	0.990	-0.495	0.933	-0.357	0.901	-0.439
0.9	100	0.993	-0.344	0.945	-0.240	0.921	-0.307
0.95	20	0.983	-0.865	0.917	-0.696	0.874	-0.765
0.95	50	0.993	-0.513	0.935	-0.346	0.910	-0.447
0.95	100	0.996	-0.356	0.949	-0.230	0.924	-0.304

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 SD(y) = 4.966		Multivariate Normal Data 5% Tainted Data SD(y)= 5.265		Multivariate Normal Data 10% Tainted Data SD(y)= 5.549	
$\rho(x_2,x_3)$	n	$r(\hat{y}_1, \hat{y}_2)$	Mean[$CIW(\hat{y}_1)$ - $CIW(\hat{y}_2)$]	$r(\hat{y}_1, \hat{y}_2)$	Mean[$CIW(\hat{y}_1)$ - $CIW(\hat{y}_2)$]	$r(\hat{y}_1, \hat{y}_2)$	Mean[$CIW(\hat{y}_1)$ - $CIW(\hat{y}_2)$]
0.3	20	0.939	0.691	0.910	0.113	0.881	-0.167
0.3	50	0.941	0.411	0.909	0.032	0.876	-0.135
0.3	100	0.941	0.291	0.903	0.012	0.875	-0.118
0.5	20	0.965	-0.003	0.928	-0.204	0.901	-0.408
0.5	50	0.969	-0.004	0.930	-0.134	0.897	-0.254
0.5	100	0.969	0.003	0.928	-0.095	0.892	-0.169
0.7	20	0.984	-0.396	0.946	-0.445	0.916	-0.556
0.7	50	0.987	-0.226	0.946	-0.237	0.909	-0.304
0.7	100	0.989	-0.158	0.943	-0.157	0.905	-0.204
0.75	20	0.987	-0.460	0.952	-0.491	0.915	-0.585
0.75	50	0.991	-0.268	0.949	-0.257	0.912	-0.313
0.75	100	0.992	-0.188	0.949	-0.174	0.910	-0.212
0.8	20	0.990	-0.503	0.951	-0.497	0.919	-0.581
0.8	50	0.994	-0.300	0.955	-0.276	0.917	-0.322
0.8	100	0.995	-0.209	0.950	-0.176	0.913	-0.213
0.85	20	0.992	-0.552	0.956	-0.543	0.917	-0.591
0.85	50	0.997	-0.327	0.958	-0.287	0.919	-0.327
0.85	100	0.998	-0.226	0.953	-0.179	0.917	-0.217
0.9	20	0.993	-0.573	0.957	-0.530	0.916	-0.586
0.9	50	0.997	-0.334	0.958	-0.280	0.918	-0.317
0.9	100	0.999	-0.232	0.957	-0.179	0.919	-0.218
0.95	20	0.989	-0.500	0.960	-0.536	0.921	-0.579
0.95	50	0.993	-0.291	0.957	-0.255	0.920	-0.318
0.95	100	0.994	-0.201	0.961	-0.175	0.919	-0.208

5. 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. This loss of one degree of freedom for the error term is likely to have a larger effect, if any, with the smaller sample sizes.

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.

Four 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. Two, the non-normality of the "tainted" distributions also showed the collinearity having a larger effect on the predictions with smaller values for the correlation between the two predicted values. Third, 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. And, four, 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.

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. http://www.iosrjournals.org/iosr-jm/papers/Vol10-issue4/Version-1/D011620.pdf

Frost. J. (2013, May 2). What are the effects of multicollinearity and when can I ignore them? [Web log 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 Edition. New York: McGraw-Hill/Irwin Series.

Montgomery, D. C., Peck, E. A., & Vining, C. G. (2001). Introduction to Linear Regression Analysis, 3rd Edition. 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 Edition. Holt, Rinehart, and Winston.

Weiss, N. A. (2012). Introductory Statistics, 10th Edition. 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. http://purl.umn.edu/159045