Journal of Modern Applied Statistical Methods

Volume 15 | Issue 1 Article 18

5-2016

The Goldilocks Dilemma: Impacts of Multicollinearity -- A Comparison of Simple Linear Regression, Multiple Regression, and Ordered Variable Regression Models

Grayson L. Baird
University of Wyoming, grayson_baird@brown.edu

Stephen L. Bieber
University of Wyoming

The Goldilocks Dilemma: Impacts of Multicollinearity -- A Comparison of Simple Linear Regression, Multiple Regression, and Ordered Variable Regression Models

Cover Page Footnote
We would like to thank Jared Studyvin for his assistance throughout the development of this project and for his insightful review of the final manuscript.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 332-357.

ISSN 1538 − 9472

The Goldilocks Dilemma: Impacts of Multicollinearity—A Comparison of Simple Linear Regression, Multiple Regression, and Ordered Variable Regression Models

Grayson L. Baird
University of Wyoming
Laramie, Wyoming

Stephen L. Bieber
University of Wyoming
Laramie, Wyoming

A common consideration concerning the application of multiple linear regression is the lack of independence among predictors (multicollinearity). The main purpose of this study is to introduce an alternative method of regression originally outlined by Woolf (1951) that eliminates the relatedness between the predictors in a multiple predictor setting.

Keywords: multicollinearity, collinearity, multiple linear regression, ordered variable regression, OVR

Introduction

Social and behavioral scientists often use multiple linear regression (MLR) to answer research questions that involve multiple predictor variables in both experimental and observational research settings. These scientists must consider a host of issues when applying MLR, such as the appropriateness of measurement, sampling, design, and model assumptions. This paper will focus on one commonly encountered problem in the application of MLR: the situation in which the predictor variables are related to one another, a condition generally referred to as multicollinearity.

Although multicollinearity can be defined as a condition in which the predictor variables are correlated with each other to some degree, the literature provides several alternative names and definitions. For instance, Darlington (1968) referred to the relatedness between predictors as intercorrelation. Kutner,

Grayson Baird is a research statistician in the Lifespan Biostatistics Core, Rhode Island Hospital. Email him at: grayson_baird@brown.edu. Dr. Bieber is a professor in the Department of Statistics.

Nachtsheim, Neter, and Li (2004) also referred to the relationship between predictors as intercorrelation, but go on to note that extreme intercorrelation is often referred to as multicollinearity. Similarly, both Cohen, Cohen, West, and Aiken (2003), and Nie, Hull, Jenkins, Steinbrenner, and Bent (1975) directly referred to multicollinearity as the situation where two or more predictors are highly intercorrelated with each other. Although Gordon (1968) noted others have used the term multicollinearity, he refers to high correlation among predictors as redundancy, though many texts reserve the term redundancy to denote the squared value of the intercorrelation (see Cohen et al., 2003). Weisberg (2005) refined the concept of multicollinearity by distinguishing different levels of collinearity, where some relatedness between predictors is referred to as approximate collinearity, strong relatedness is referred to as collinearity, and perfect relatedness is referred to as exact collinearity.

Throughout the references above, the terms such as high, extreme, some, and strong are not numerically specified. However, Tabachnick and Fidell (2007) do specifically define thresholds for multicollinearity, indicating that clear multicollinearity exists when predictors correlate above .90, where correlations above .70 may also be suggestive of multicollinearity. For most social and behavioral science researchers, these values are so unattainably high that they could leave the impression that multicollinearity never needs to be considered nor viewed as problematic within their data.

Gordon (1968) noted that "statistics texts focus upon conditions of extremely high correlation because it is at that point that the resulting problems become most nearly statistical ones."(p. 596). Alternatively, Cohen et al. (2003), Kutner et al. (2007), and Weisberg (2005) referred to multicollinearity as a problematic condition, where Nie et al. (1975) refer to multicollinearity as a condition that can cause problems. All the aforementioned authors go on to discuss various problems of multicollinearity in application.

Gordon (1968) observed that discussions of multicollinearity in general are brief in statistical texts and Weisberg (2005) observed that [multi]collinearity itself has no precise definition. Therefore, it appears that multicollinearity does not have a unified definition or meaning and in fact can denote a variety of different concepts in the literature. Given the imprecise nature of the term multicollinearity, the correlation between predictors in this paper will be referred to as simply relatedness. It is hoped that this neutral term expresses the idea of correlation between predictors, without implying unintended connotations such as strength or threshold of correlation, being problematic or not, etc.

Therefore, the intent of this study aims not at defining multicollinearity, but rather discussing and demonstrating the impacts of related predictors on the MLR model and statistics, for any value of relatedness greater than zero. An alternative to MLR, called ordered variable regression (OVR), will be presented in this paper, which resolves the issue of related predictors entirely by creating and using predictors that are perfectly unrelated.

Relationship between Predictor Variables

The predictor variables in the multiple linear regression (MLR) model can be either independent of each other ($r_{12} = 0$) or correlated to each other ($r_{12} \neq 0$) [for simplicity and without loss of generalizability, only two predictors, X_1 and X_2, will be considered throughout this paper]. If two predictors are related to each other, then their redundancy (see Cohen et. al. 2003) can be expressed as r_{12}^2 (i.e., the squared value of their correlation; shared variance).

Figures 1 and 2 will be used extensively throughout this article to present the numerous and varied impacts of the relationship between the predictor variables on the response variable (Y). In these Venn diagrams, the area within any circle is equal to 1 (the total variance of any variable = 1.00), thus the partitions of these circles represent proportions of variance (see Kerlinger & Pedhazur, 1973).

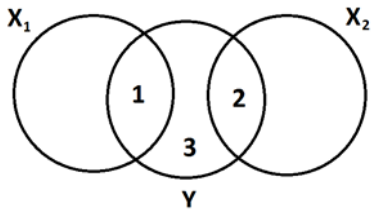

Figure 1. Predictors are unrelated

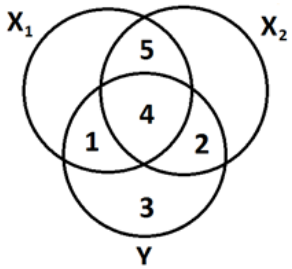

Figure 2. Predictors are related

Figure 1 illustrates the situation when the two predictors are independent (i.e., the circles representing the predictors do not intersect). In this situation, regions 1 and 2 represent the proportions of the response variable Y accounted for by the predictors X_1 and X_2 respectively; specifically the size of region 1 is r_{YX1}^2 and the size of region 2 is r_{YX2}^2. Region 3 represents that part of Y that can't be predicted by X_1 or X_2, which will be referred to as the error.

Conversely, Figure 2 illustrates the situation when the two predictors are related to each other (i.e., the circles representing the predictors do intersect). Regions 1 and 2, as they did in Figure 1, represent the unique contributions to the response variable from X_1 and X_2, respectively. Unlike Figure 1, Figure 2 contains two additional regions, 4 and 5, which reflect the redundancy $\left(r_{12}^2\right)$ between the two predictors. The impact of this relationship between the variables X_1 and X_2 complicates the prediction of the response variable by adding a new piece, region 4 (the shared influence of both predictors on *Y*) to the circle representing *Y*. In this figure, region 3 is that part of *Y* which can't be predicted by X_1 and/or X_2; once again the error.

Implications and Impacts When the Predictors Are Not Related

Simple Linear Regression

The simple linear regression (SLR) model in which only X_1 is used to predict *Y* can be expressed as

$$\hat{Y} = b_1 X_1 \tag{1}$$

where b_1 is the least squares estimate of the slope associated with X_1 and is the answer to the research question, "how is X_1 predictive of *Y*?" [without loss of generality, it is possible to consider all of the regression models presented in this article from the perspective in which X_1, X_2, and *Y* are standardized. As a consequence and for convenience, the intercept is always 0.] Similarly,

$$\hat{Y} = b_2 X_2 \tag{2}$$

where b_2 is the least squares estimate of the slope associated with X_2 and is the answer to the research question, how is "X_2 predictive of *Y*?" Because the variables have all been standardized,

$$b_1 = r_{YX1} \tag{3}$$

and

$$b_2 = r_{YX\,2} \tag{4}$$

There are two related questions to "how is X_i predictive of Y?" These are "is the relationship between X_i and Y statistically significant," and "how much of Y is predicted by X_i?" [throughout this paper, the subscript i will be used to designate either X_1 ($i = 1$), X_2 ($i = 2$), or a result derived from X_i]. The first of these two questions is answered by converting the slope into the t-distribution using

$$t_i\left(n-p-1\right) = \frac{b_i}{SE\left(b_i\right)} = \frac{b_i}{\left[\dfrac{\sigma_i^2}{\left(n-1\right)s^2\left(X_i\right)}\right]^{1/2}} \tag{5}$$

where n is the sample size, p is the number of predictors (i.e., $p = 1$), SE(b_i) is the standard error of the slope (b_i), and σ_i^2 is error variance of Y [which is associated with X_i and is estimated using the mean squared error (MSE$_i$)], and $s^2(X_i)$ is the variance of the predictor X_i.

The second of these questions is answered using the coefficient of determination or R^2, which represents the proportion of variance in Y accounted for (explained, predicted) by either

$$R_i^2 = b_i^2 = r_{YXi}^2 \tag{6}$$

where b_i^2 equals the size of region i. The coefficient of determination can also be calculated through the use of the sums of squares presented in the analysis of variance table. Although unnecessary in this section, it is presented for consistency with subsequent sections of this article. Within the context of SLR, the sums of squares can be partitioned as follows

$$SS_{Total} = SS_{SLR\ Model(i)} + SS_{Error(i)} \tag{7}$$

where SS_{Total} is the total variation found in Y (associated with regions 1, 2, and 3 in Figure 1; as the circles of Figures 1 and 2 have been standardized to variance = 1, the sums of squares are associated with (represented by) the regions in concept, but not equal to them in size.), $SS_{SLR\ Model(i)}$ is the variation in Y associated with predictor variable X_i and $SS_{Error(i)}$ is the variation in Y not associated with the predictor variable X_i. Hence, when $i = 1$ (predicting Y from X_1),

$SS_{SLR\ Model(1)}$ is represented by region 1, and $SS_{Error(1)}$ by regions 2 and 3. Conversely, when $i = 2$ (predicting Y from X_2), $SS_{SLR\ Model(2)}$ is represented by region 2, and $SS_{Error(2)}$ by regions 1 and 3. From this context, the coefficient of determination for SLR models is

$$R_i^2 = \frac{SS_{SLR\ Model(i)}}{SS_{Total}} = \frac{SS_{SLR\ Model(i)}}{SS_{SLR\ Model(i)} + SS_{Error(i)}} \tag{8}$$

The values for R_i^2 as determined by Equations 6 and 8 are identical. Lastly, the significance of R_i^2 can be found by using the Omnibus F-statistic (abbreviated throughout the paper as F),

$$F_i\left(p, n-p-1\right) = \frac{SS_{SLR\ Model(i)} \big/ p}{SS_{Error(i)} \big/ \left(n-p-1\right)} = \frac{SS_{SLR\ Model(i)} \big/ p}{MSE_i} \tag{9}$$

where the mean squared error (MSE_i) is the estimate of the error variance associated with predictor X_i, which was identified as σ_i^2 in Equation 5.

Multiple Linear Regression

Based on the foundational elements for the simple linear regression (SLR) model above, it is possible to develop the multiple linear regression (MLR) model, in which Y is predicted jointly by both X_1 and X_2. In a parallel form to the preceding section it is possible to start with the fundamental research question, which is "how are X_1 and X_2 jointly predictive of Y?" The answer to this question is found in the MLR model

$$\hat{Y} = c_1 X_1 + c_2 X_2 \tag{10}$$

where c_1 and c_2 are the least squares estimates of MLR parameters.

Although Equation 10 is considered to be the answer to the question posed, it rests heavily upon how the word jointly is interpreted (this distinction will be considered at length in next section considering the implications and impacts when the predictors are related). In its standard application, MLR produces an additive model (no interaction terms) and thus defines jointly as independent of one another. As a consequence, the coefficient c_1 is actually the answer to the

question, "how is X_1 predictive of Y independent of X_2," and c_2 is the answer to the question, "how is X_2 predictive of Y independent of X_1?" From this perspective, it can be seen that the coefficients from the MLR answer a similar, yet very distinct question from the context of SLR.

At this point, the most important and logical question is "what is the relationship between c_1 and b_1, and between c_2 and b_2?" Within the context of standardized variables, the MLR coefficients, c_1 and c_2, can be linked with the bi-variate correlations as follows from Darlington (1968).

$$c_1 = \frac{r_{YX1} - r_{YX2} r_{12}}{1 - r_{12}^2} \quad \text{and} \quad c_2 = \frac{r_{YX2} - r_{YX1} r_{12}}{1 - r_{12}^2} \tag{11}$$

The relationship of Equation 11 with the part correlations (McNemar, 1962), which are also called the semi-partial correlations (Nunnally, 1967), will be discussed at length in the consideration of Equation 27. At this point, the relationship is inconsequential, because $r_{12} = 0$ and as a result Equation 11 reduces to

$$c_i = r_{YXi} = b_i \tag{12}$$

Thus, if the two predictor variables are not related, then the MLR, c_1 and c_2, are identical to their SLR counterparts, b_1 and b_2. In addition, the italicized portion of the MLR questions above (independent of) can be deleted and also simplify to their SLR counterparts.

The test of the significance of the regression coefficients c_1 and c_2 is once again found through the *t*-statistics, which in the context for MLR is

$$t_i\left(n - p - 1\right) = \frac{c_i}{SE\left(c_i\right)} = \frac{c_i}{\left[\left(\frac{\sigma^2}{\left(n-1\right)s^2\left(X_i\right)}\right)\left(\frac{1}{1-r_{12}^2}\right)\right]^{1/2}} \tag{13}$$

where p is the number of predictors (i.e., $p = 2$). Because $r_{12} = 0$, Equation 13 reduces to

$$t_i\left(n-p-1\right)=\frac{c_i}{SE\left(c_i\right)}=\frac{c_i}{\left[\dfrac{\sigma^2}{\left(n-1\right)s^2\left(X_i\right)}\right]^{1/2}} \tag{14}$$

Although Equation 14 is similar in appearance to Equation 5, they are not identical. The standard errors of the regression coefficients [SE(c_i)] are smaller than the corresponding standard errors [SE(b_i)], because the size of the MSE (σ^2) from the MLR model has been reduced to region 3 only (hence, $\sigma^2 < \sigma_1^2$ and $\sigma^2 < \sigma_2^2$). Therefore, the value of the *t*-statistics from the MLR model will be larger than in the SLR models; however, they will not necessarily result in smaller *p*-values given that the degrees of freedom have been reduced by one.

As in the previous section, the MLR answer to the question, "how much of Y is predicted by X_1 and X_2," is found using the coefficient of determination. As presented in Darlington (1968), the coefficient of determination within the context of two predictor variables is

$$R^2 = c_1^2 + c_2^2 + 2c_1c_2r_{12} \tag{15}$$

using $r_{12} = 0$ and the result of Equation 12

$$R^2 = b_1^2 + b_2^2 \tag{16}$$

Thus, the coefficient of determination from the multiple regression reduces to the sum of the coefficients of determination from the two separate simple regressions, see Equation 6.

From the context of the partitioning of the sums of squares,

$$SS_{Total} = SS_{MLR\ Model} + SS_{Error} \tag{17}$$

where

$$SS_{MLR\ Model} = SS\left(X_1 | X_2\right) + SS\left(X_2 | X_1\right) \tag{18}$$

Specifically, $SS(X_1 | X_2)$ reflects the amount of variation in Y associated with the first predictor independent of any association with the second predictor

(represented by region 1 in Figure 1) and $SS(X_2|X_1)$ corresponds with the amount of variation in Y associated with X_2 independent of any association with X_1 (region 2). Given that the predictors are not related, then it can be logically deduced from the results above that $SS(X_1|X_2) = SS_{SLR\ Model(1)}$, $SS(X_2|X_1) = SS_{SLR\ Model(2)}$, and

$$SS_{SLR\ Model} = SS_{SLR\ Model(1)} + SS_{SLR\ Model(2)} \tag{19}$$

Hence, the amount of variation in Y accounted for jointly by X_1 and X_2 is simply the sum of their variation from the simple regressions. The simultaneous use of both predictors results in a single model reflecting both predictive regions (1 and 2), while reducing the error to its appropriate minimum (region 3 only). Thus the coefficient of determination becomes

$$\begin{aligned} R^2 &= \frac{SS_{MLR\ Model}}{SS_{Total}} = \frac{SS_{MLR\ Model}}{SS_{MLR\ Model} + SS_{Error}} \\ &= \frac{SS_{SLR\ Model(1)} + SS_{SLR\ Model(2)}}{SS_{SLR\ Model(1)} + SS_{SLR\ Model(2)} + SS_{Error}} \end{aligned} \tag{20}$$

or the sum of the two coefficients of determination presented in Equation 8.

As with the individual tests of the coefficients, presented in Equation 14, the Omnibus F-statistic is not a simple extension from the SLR results, due to the reduction in the error term and degrees of freedom. The Omnibus F-statistic for the multiple regression is

$$F\left(p, n-p-1\right) = \frac{SS_{MLR\ Model}/p}{SS_{Error}/\left(n-p-1\right)} = \frac{\left(SS_{SLR\ Model(1)} + SS_{SLR\ Model(2)}\right)/p}{MSE} \tag{21}$$

Numerical Example

To illustrate the points made above when considering the SLR and MLR models, and their corresponding results, a numerical example is presented in Table 1 for data in which $r_{12} = .000$. Due to round off errors associated with the standardization of any data set, the actual value of the relatedness of X_1 and X_2 will not be perfectly zero. For these data the relatedness of X_1 and X_2 is 6.00E-18.

The results can be found by using the regression routine in most statistical computer packages. The exception is that a general linear model routine needs to be performed in order to obtain the sums of squares breakdown information specific to each predictor [$SS(X_1|X_2)$ and $SS(X_2|X_1)$] in the MLR context.

In summary, when the predictors are not related, the coefficients produced by the MLR model are identical to the coefficients produced by the SLR models. As a consequence, the R^2 and model sum of squares for the MLR model are the additive composites of the R^2 and model sum of squares produced by the SLR models. Thus, the data of Table 1 confirms the derived results presented in Equations 12, 16, 19, and 20. These results will hold for any data set in which the predictors are unrelated.

Implications and Impacts When the Predictors Are Related

Simple Linear Regression

This section is essentially the duplicate of the simple regression section when the predictors are not related. The primary difference is found in region 4 of Figure 2. What is the impact of this difference on the results presented previously?

The questions of "how is X_i predictive of Y" remain the same and b_i (the estimates of the slopes) are still the answers. However, b_1 is now associated with regions 1 and 4, and b_2 is now associated with regions 2 and 4. Similarly, all of the results presented in Equations 3 through 9 remain the same, but are expanded to include region 4. Hence, any discussion of X_1 now includes both regions 1 and 4, and any discussion of X_2 now includes regions 2 and 4.

It is important to note that even though all of the results are identical, regardless of whether the predictors are related or not, the answers to the fundamental questions, "how is X_i predictive of Y," are now more complex. The first predictor is no longer solely predictive of Y (represented by region 1), but this prediction is now supplemented by a shared element associated with the second predictor (region 4). The same situation exists when the focus of the SLR is the second predictor. As a consequence, although the fundamental regression questions remain simple, the answers aren't. Unfortunately these two aspects of the predictor variables are fused together in the answers b_i and can't be separated within the context of SLR.

Table 1. Comparison of Simple and Multiple Regression when the predictors are not related

	SLR(X_1)	SLR(X_2)	MLR	Comments
Coefficient				
X_1	b_1 = .467		c_1 = .467	$b_1 = c_1 = r_{YX1}$, Equations 3, 11
X_2		b_2 = .312	c_2 = .312	$b_2 = c_2 = r_{YX2}$, Equations 4, 11
Sums of Squares				
X_1	16.115		16.115	MLR result = SLR result for X_1, Discussion for Equation 19
X_2		7.222	7.222	MLR result = SLR result for X_2, Discussion for Equation 19
Model	16.115	7.222	23.337	MLR result = sum of the SLR results for X_1 and X_2, Equation 19
Error	57.884	66.778	50.663	
Total	74.000	74.000	74.000	
R^2 Estimates				
R^2	.218	.097	.315	MLR result = sum of the SLR results for X_1 and X_2, Equation 16
R^2 from *SS*			.315	$SS_{MLR\ Model}$ / SS_{Total} = 23.337 / 74.000 = .315, Equation 20

*Note: r_{12} = .000, r_{YX1} = .467, r_{YX2} = .312, n = 75

It is important to note that even though all of the results are identical, regardless of whether the predictors are related or not, the answers to the fundamental questions, "how is X_i predictive of Y," are now more complex. The first predictor is no longer solely predictive of Y (represented by region 1), but this prediction is now supplemented by a shared element associated with the second predictor (region 4). The same situation exists when the focus of the SLR is the second predictor. As a consequence, although the fundamental regression questions remain simple, the answers aren't. Unfortunately these two aspects of the predictor variables are fused together in the answers b_i and can't be separated within the context of SLR.

Looking at Figure 2 it can be seen that SS_{Total} is now represented by regions 1, 2, 3, and 4. As a result, the SLR for X_1 produces

$$SS_{SLR\ Model(1)} = SS\left(X_1 \middle| X_2\right) + SS_{Shared} \tag{22}$$

which corresponds with regions 1 and 4, and an $SS_{Error(1)}$ corresponding to regions 2 and 3. [SS_{Shared} will be defined later in Equations 41 and 42.] Similarly, the result of the SLR for X_2 produces

$$SS_{SLR\ Model(2)} = SS\left(X_2 \middle| X_1\right) + SS_{Shared} \tag{23}$$

which corresponds with regions 2 and 4, and an $SS_{Error(2)}$ corresponding to regions 1 and 3.

Multiple Linear Regression

The previous section with MLR when the predictors were not related began with the logical research question, "how are X_1 and X_2 jointly related to Y?" However, because the two predictor variables are now related, the definition of the word jointly is much more complicated than in this previous section. In fact, there are now at least three distinct definitions of this word, which each lead to decidedly different conclusions in regard to the regression coefficients, coefficients of determination, sums of squares, and statistical tests.

Definition 1. Jointly is viewed as the composite of the influence of X_1 to Y and the influence of X_2 to Y. This definition reflects jointly as the sum of the two separate SLR questions, "how is X_1 predictive of Y" and "how is X_2 predictive of Y." The answer to this question is

$$\hat{Y} = b_1 X_1 + b_2 X_2 \tag{24}$$

Using Equation 10 and the result of Equation 12, it was found that when $r_{12} = 0$ it is possible for MLR to generate the model presented in Equation 24. However, when $r_{12} \neq 0$, Equation 24 can't be estimated by any single regression model, because b_1 and b_2 must be estimated separately. Thus, Equation 24 should only be considered as a conceptual combination of the two predictors.

From the previous section, the coefficients of determination for these two simple regressions are R_1^2 for X_1 (regions 1 and 4 in Figure 2) and R_2^2 for X_2 (regions 2 and 4 in Figure 2). As a result, if the two were added together to provide a combined estimate, then

$$R^2 = R_1^2 + R_2^2 = b_1^2 + b_2^2 = \text{Region 1 + Region 4 + Region 2 + Region 4} \tag{25}$$

Thus, the combined estimate presented in Equation 25 would double count region 4 and artificially inflate the jointly determined R^2 by the size of region 4. This was not the case for Equation 16, because region 4 didn't exist. Hence, the use of this definition to determine the joint R^2 is accurate only when the predictors aren't related.

In practice, this first definition of jointly would result in answering the multiple regression question from the context of performing two simple regressions and combining their results at the level of discussion rather than at the level of a predictive model. Although the multiple application of SLR in the presence of multiple predictors may be found in the literature, their results should be viewed with considerable caution. As pointed out in the section above, their answers are not as simple as their questions imply (they can't be interpreted independently), and the R^2 from their conceptual combination (jointly determined influence) will increasingly be over estimated as $|r_{12}|$ increases (increasing the size of region 4).

Definition 2. Jointly is viewed as the composite of the influence of X_1 to Y independent of X_2 and the influence of X_2 to Y independent of X_1. In this context the word jointly reflects a simultaneous relationship and leads directly to the traditional MLR model

$$\hat{Y} = c_1 X_1 + c_2 X_2 \tag{26}$$

In appearance this is exactly Equation 10. However, is it? As presented in Equation 11, duplicated here, it is known that

$$c_1 = \frac{r_{YX1} - r_{YX2}r_{12}}{1 - r_{12}^2} \quad \text{and} \quad c_2 = \frac{r_{YX2} - r_{YX1}r_{12}}{1 - r_{12}^2} \tag{27}$$

To begin, given that $r_{12} \neq 0$, Equation 27 doesn't simplify as Equation 11 did, and the MLR coefficients (c_i) won't equal their SLR counterparts (b_i). A close inspection of Equation 27 reveals that the MLR coefficients are functions of the part correlations (McNemar, 1962) [although it is common to speak about the multiple regression coefficients as addressing the question of the relationship between a predictor and dependent variable partialling out the influence of other predictors, this process as actually accomplished through the part correlations, not the partial correlations. Symbolically, $r_{YX1.X2}$ refers to the partial correlation and $r_{Y(X1.X2)}$ refers to the part correlation.] The part correlation of X_1 with Y removing the influence of X_2 from Y only (directly represented by region 1 in Figure 2) is

$$r_{Y(X1.X2)} = \frac{r_{YX1} - r_{YX2}r_{12}}{\sqrt{1 - r_{12}^2}} \tag{28}$$

and of X_2 with Y removing the influence of X_1 from Y only (represented by region 2) is

$$r_{Y(X2.X1)} = \frac{r_{YX2} - r_{YX1}r_{12}}{\sqrt{1 - r_{12}^2}} \tag{29}$$

As a consequence, substituting Equations 28 and 29 into Equation 27, the coefficients from the MLR model are

$$c_1 = \frac{r_{Y(X1.X2)}}{\sqrt{1 - r_{12}^2}} \quad \text{and} \quad c_2 = \frac{r_{Y(X2.X1)}}{\sqrt{1 - r_{12}^2}} \tag{30}$$

Thus, although Equation 26 looks very similar to Equation 10, it is dramatically different. This is the first impact of the relatedness of the predictors; the MLR regression coefficients are no longer equal to their SLR counterparts. In MLR the coefficients, through their association with the process of part correlation, have

had the shared influence (represented by region 4) removed in comparison to the coefficients from SLR. This is the direct result of the additive nature of the MLR model presented in Equation 26.

What additional impact does this second definition of jointly have on the other results in the multiple predictor setting? Within the context of MLR, the test of the significance of the regression coefficients c_1 and c_2 is found through the *t*-statistic [Equation 13 is duplicated below]

$$t_i\left(n-p-1\right)=\frac{c_i}{SE\left(c_i\right)}-\frac{c_i}{\left[\left(\frac{\sigma^2}{\left(n-1\right)s^2\left(X_i\right)}\right)\left(\frac{1}{1-r_{12}^2}\right)\right]^{1/2}} \tag{31}$$

Unlike Equation 13, Equation 31 doesn't reduce to Equation 14 because $r_{12}\neq 0$. As a note, $1\Big/\left(1-r_{12}^2\right)$ of Equation 31 is commonly referred to as the Variance Inflation Factor (VIF). Hence, the impact of the relatedness between the two predictors is the inflation of SE (because the VIF must be greater than 1), which results in a decrease in the magnitude (and thus significance) of the *t*-statistic. The second impact of the relatedness of the predictors is that their independent contributions to predicting *Y* are less statistically significant.

What is the impact on the coefficient of determination? From Figure 2, it can be seen that R^2 should be the combined influence from X_1 and X_2 independently (region 1 and 2), and the shared influence of X_1 and X_2 (region 4), such that

$$R^2 = \text{ region 1 + region 2 + region 4} \tag{32}$$

When $r_{12} = 0$, it is easy to relate the regions of Figure 1 with the components of the R^2; as found in Equation 6. However, now that $r_{12}\neq 0$, how do the results from the MLR model correspond with the components of R^2? Using Equations 28 and 29 along with the research question posed by the definition of jointly as simultaneously, it can be seen that the MLR model, found in Equation 26, produces

$$\text{region 1}=r_{Y(X1.X2)}^2 \text{ and } \text{region 2}=r_{Y(X2.X1)}^2 \tag{33}$$

such that

$$R^2_{MLR\ Model} = \text{region } 1 + \text{region } 2 = r^2_{Y(X1.X2)} + r^2_{Y(X2.X1)} \tag{34}$$

and

$$R^2 = R^2_{MLR\ Model} + R^2_{Shared} \tag{35}$$

where R^2_{Shared} equals the size of region 4. Substituting the results of Equation 30 into Equation 33 produces the association between the regions of Figure 2 and the MLR coefficients as

$$\text{region } 1 = \left(1 - r^2_{12}\right)c^2_1 \quad \text{and} \quad \text{region } 2 = \left(1 - r^2_{12}\right)c^2_2 \tag{36}$$

Recalling that R^2 for the MLR equals Equation 15, the size of region 4 can be established in terms of the part correlations as

$$\text{region } 4 = R^2_{Shared} = \frac{r_{12}}{1 - r^2_{12}}\left[r_{12}r^2_{Y(X1.X2)} + r_{12}r^2_{Y(X2.X1)} + 2r_{Y(X1.X2)}r_{Y(X2.X1)} \right] \tag{37}$$

and in terms of the MLR coefficients as

$$\text{region } 4 = R^2_{Shared} = r_{12}\left[r_{12}c^2_1 + r_{12}c^2_2 + 2c_1c_2 \right] \tag{38}$$

It can be seen from this discussion that R^2 is actually a combination of two separate and independent pieces; that piece associated with the model ($R^2_{MLR\ Model}$; regions 1 and 2) and that piece associated with the shared influence (R^2_{Shared} ; region 4). The third impact of the relatedness of the predictors is that the R^2 is unequal to $R^2_{MLR\ Model}$, begin inflated by the size of region 4, unless $r_{12} = 0$.

These results for the R^2 can also be illustrated by examining the sum of squares. The determination of the sums of squares using this second definition of jointly is often referred to as Type III sums of squares, which is presented in Equation 40.

$$SS_{Total} = SS_{MLR\ Model} + SS_{Shared} + SS_{Error} \tag{39}$$

where

$$SS_{MLR\ Model} = SS\left(X_1 | X_2\right) + SS\left(X_2 | X_1\right) \tag{40}$$

and

$$SS_{Shared} = SS\left(X_1, X_2\right) = SS_{SLR\ Model(1)} - SS\left(X_1 | X_2\right) \tag{41}$$

from Equation 22 and

$$SS_{Shared} = SS\left(X_1, X_2\right) = SS_{SLR\ Model(2)} - SS\left(X_2 | X_1\right) \tag{42}$$

from Equation 23. As expressed in Equation 18, $SS(X_1|X_2)$ reflects the amount of variation in Y associated with the first predictor independent of any association with the second predictor (region 1) and $SS(X_2|X_1)$ corresponds with the amount of variation in Y associated with X_2 independent of any association with X_1 (region 2). SS_{Shared} reflects the joint influence of X_1 and X_2 (represented by region 4), and SS_{Error} now correctly corresponds with region 3 only.

In many textbooks and statistical programs, it appears that the SS_{Model} is not calculated directly, but rather determined indirectly through the simple subtraction whereby $SS_{Model} = SS_{Total} - SS_{Error}$. This calculation works perfectly when $r_{12} = 0$, but when $r_{12} \neq 0$ it mistakenly includes the SS_{Shared} in the SS_{Model} and inflates the sums of squares associated with the model, such that

$$SS_{Model} = SS_{Total} - SS_{Error} = SS_{MLR\ Model} + SS_{Shared} \tag{43}$$

This is perhaps best explained and illustrated by Woolf (1951, see p. 113). Therefore, the R^2 can be calculated using the sums of squares as

$$R^2_{MLR\ Model} = \frac{SS_{MLR\ Model}}{SS_{Total}} = \frac{SS\left(X_1 | X_2\right) + SS\left(X_2 | X_1\right)}{SS_{Total}} \tag{44}$$

and as

$$R^2 = \frac{SS_{Total} - SS_{Error}}{SS_{Total}} = \frac{SS_{MLR\ Model} + SS_{Shared}}{SS_{Total}} \tag{45}$$

It can be seen in Equation 45 that the R^2 calculated by the simple subtraction method is once again inflated by SS_{Shared} (region 4) in comparison to the $R^2_{MLR\ Model}$, as presented in Equation 44.

The last impact of the relatedness of the predictors on the MLR results is seen in the determination of the Omnibus *F*-statistic

$$F\left(p, n-p-1\right) = \frac{SS_{Model}/p}{SS_{Error}/\left(n-p-1\right)} = \frac{\left(SS_{MLR\ Model} + SS_{Shared}\right)/p}{MSE} \tag{46}$$

whose value can be partitioned such that the components of *F* are equal to the sum of

$$F_{MLR\ Model} = \frac{SS_{MLR\ Model}/p}{SS_{Error}/\left(n-p-1\right)} = \frac{\left[SS\left(X_1 \middle| X_2\right) + SS\left(X_2 \middle| X_1\right)\right]/p}{MSE} \tag{47}$$

and

$$F_{Shared} = \frac{SS_{Shared}/p}{SS_{Error}/\left(n-p-1\right)} = \frac{SS_{Shared}/p}{MSE} \tag{48}$$

As with R^2, the use of SS_{Model} results in the inflation of the *F* by SS_{Shared} (region 4).

In summary, the MLR coefficients are the direct answers to the research questions posed at the beginning of this section (Definition 2 of the word jointly) and the *t*-statistics provide the appropriate significance tests of these relationships. However, both the coefficient of determination and the Omnibus *F*-statistic are inflated in relation to the MLR model by a function of the amount of shared variance (region 4). Hence, the MLR model (c_1 and c_2) is not consistent with the commonly reported summary statistics (R^2 and *F*). These results will be demonstrated in the numerical example section below.

Definition 3. Jointly is viewed as the composite of the influence of X_1 to *Y* (from Definition 1) and the influence of X_2 to *Y* independent of X_1 (Definition 2). In this context, the word jointly affects an ordered relationship (note either X_1 or X_2 can be represented in the first question, with the other predictor in the second. For convenience only, X_1 will be used in the first question and X_2 in the second). Together this ordered relationship could be represented in the model

$$\hat{Y} = b_1 X_1 + c_2 X_2 \tag{49}$$

where b_1 comes from Equation 24 and c_2 comes from Equation 26. This model will be referred to here as Ordered Variable Regression (OVR). [Unlike Definition 1 that was only a conceptual combination of the two predictors, Definition 3 actually leads to a determinable model, which will be presented later in this section.]

Another way of viewing these two influences is from the context of stepwise regression, in which b_1 is the answer to the question, "what does X_1 contribute to Y," and c_2 is the answer to the question, "what does X_2 contribute to Y beyond what is already being contributed by X_1?"

The significance of these two regression coefficients have already been presented in Equation 14 and Equation 31, respectively. The determination of the sums of squares using this third definition of jointly is often referred to as Type I sums of squares, which is presented in Equation 51.

$$SS_{Total} = SS_{OVR\ Model} + SS_{Error} \tag{50}$$

Specifically,

$$SS_{OVR\ Model} = SS_1 + SS_2 \tag{51}$$

where

$$SS_1 = SS\left(X_1 \middle| X_2\right) + SS_{Shared} \tag{52}$$

is consistent with Equation 22, which corresponds with regions 1 and 4, and

$$SS_2 = SS\left(X_2 \middle| X_1\right) \tag{53}$$

corresponds with region 2, hence

$$SS_{OVR\ Model} = SS\left(X_1 \middle| X_2\right) + SS\left(X_2 \middle| X_1\right) + SS\left(X_1, X_2\right) = SS_{MLR\ Model} + SS_{Shared} \tag{54}$$

corresponds with regions 1, 2, and 4. The OVR model (Definition 3) now contains region 4, where the MLR model (Definition 2) did not. Now, SS_{Total} - SS_{Error} does

equal SS_{Model}. The R^2 determined from the OVR model does actually include the shared variation and does equal Equation 45. Thus, whereas Equation 45 is inflated for the determination of R^2 when associated with the MLR model, it is now correct for the OVR model. Likewise, the F determined from Equation 46 is now appropriate for the OVR model by the result of Equation 54. As a consequence, the regression model and these summary statistics are now in agreement, which was not the case for the MLR model.

The ordered variable regression (OVR) can easily be performed within any statistical package using the following steps. [Although only presented for two predictors, the steps can easily be expanded to include any number of predictors. In addition, alternative orderings can easily be proposed, considered, and compared using the same method.] First, determine the order for considering the predictors. This is perhaps the hardest step, but most researchers have little or no trouble placing their predictors in some order based on logic, theory, convenience, and/or cost considerations. As a consequence, the research questions answered by the OVR model are arguably more consistent with real questions than those actually answered by the MLR model. For illustration, let X_1 be the predictor of primary interest. Second, obtain the residuals (X_{2res}) from the regression in which X_1 is the independent variable and X_2 is the dependent variable. The correlation between X_1 and these residuals will be zero. Thus the entire earlier section when the predictors are not related of this article applies. Third, perform the regression in which X_1 and X_{2res} are the predictors of the response variable Y. The result of this regression will be the OVR model expressed in Equation 49. Which will produce

$$R^2_{OVR\ Model} = b_1^2 + r^2_{Y(X2.X1)} = \text{Region 1 + Region 4 + Region 2} = R^2 \qquad (55)$$

the R^2 value indicated in Equation 32 because $r^2_{Y(X2.X1)}$ is region 2 (from Equation 33) and

$$b_1^2 = r^2_{Y(X1.X2)} + \frac{r_{12}}{1-r_{12}^2}\left[r_{12}r^2_{Y(X1.X2)} + r_{12}r^2_{Y(X2.X1)} + 2r_{Y(X1.X2)}r_{Y(X2.X1)} \right] \qquad (56)$$

is region 1 + region 4 from Equations 33 and 37.

Numerical Example

To illustrate the points made in the two sections above for the SLR, MLR, and OVR models, and their corresponding results, a numerical example is presented in Table 2 for data in which $r_{12} = .469$. These results can be found by using the regression routine in any of the major statistical computer packages. The exception is that a general linear model routine needs to be run in order to obtain the sums of squares breakdown information specific to each predictor in the MLR and OVR contexts. Due to round off errors in the computation of X_{2res}, the actual correlation of X_1 and X_{2res} is -5.2E-16 instead of perfect zero.

At this point, it may seem that the OVR model is nothing more than hierarchical multiple regression analysis (HMR) or forward step regression using type I sum of squares. It is true that OVR and HMR share a common approach in that predictors are entered into the model sequentially and the additive contribution of each predictor can be reflected in the type I sum of squares. However, OVR differs from HMR in that the additive contribution of each predictor is reflected in both the type I sum of squares and the model coefficients. This is illustrated in Table 3. Of course, the OVR produces the same model as the HMR when predictors are not related.

It should be noted the concept of [what is referred to in this paper as] OVR was proposed by Woolf (1951) as a second method of calculating multiple linear regression. The novelty presented here is in the application of OVR as a method of regression modeling when faced with multicollinearity; guided by theory, OVR can be used to incrementally model the natural relatedness between predictors. As a consequence, OVR not only provides an alternative method of dealing with multicollinearity in a regression context, but more importantly, it allows the evaluation of research questions that assume or hypothesize hierarchical relatedness among predictors.

In summary, when the predictors are related, the coefficients of the SLR, MLR, HMR, and OVR models are not equal, but differ from one another in a predictable manner based on the amount of the relatedness between the two predictors. The data confirmed that the overall summary and test statistics (R^2 and F) associated with MLR are all inflated in relation to the model by the inclusion of the shared variance; as indicated in Equations 35, 45, and 46. In contrast, the data showed that these statistics are consistent with the OVR model which does included the shared variance. The implications and impacts of the results presented in Tables 2 and 3 will hold for any value of the relatedness between predictors that is different from zero.

Table 2. Comparison of Simple, Multiple, and Ordered Regression when the predictors are related

	SLR(X_1)	SLR(X_2)	MLR	OVR	Comments
Coefficient					
X_1	b_1 = .505		c_1= .389	.505	$b_1 = r_{YX1}$, $b_1 \neq c_1$, OVR slope = b_1, Equations 3, 27, 30, 49
X_2		b_1 = .429	c_2 =.246	.246	$b_2 = r_{YX2}$, $b_2 \neq c_2$, OVR slope = c_2, Equations 4, 27, 30, 49
Sums of Squares					
X_1	18.854		8.753	18.854	Equations 23, 52
X_2		13.592	3.491	3.491	Equations 24, 52
Model	18.854	13.592	12.244	22.345	MLR (Equation 40), OVR (Equation 54)
Error	55.146	60.408	51.655	51.655	
Shared			10.101		Equations 41, 42
Total	74.000	74.000	74.000	74.000	MLR (Equation 39), OVR (Equation 50)
					For the *SS* to be additive, MLR must add in SS_{Shared}
R^2 Estimates					
X_1	.255		.118	.255	MLR (Equation 33), OVR (Equation 6)
X_2		.184	.047	.047	MLR, OVR (Equation 33)
R^2_{Model}			.137	.302	MLR (Equation 34), OVR (Equations 35, 55)
R^2_{Shared}			.165		MLR (Equations 37, 38), OVR is included in the model
R^2			.302		For the R^2 to be additive, MLR must add in R^2_{Shared}
F Statistics					
F_{Model}			8.533	15.573	MLR (Equation 47), OVR (Equation 46)
F_{Shared}			7.040		MLR (Equation 48)
F			15.573	15.573	Overall *F* value, MLR (Equation 46) must include F_{Shared}

*Note: r_{12} = .469, r_{YX1} = .505, r_{YX2} = .429, $r_{Y(X1.X2)}$ = .344, $r_{Y(X2.X1)}$ = .217, n = 75

Table 3. Comparison of Hierarchical Multiple Regression, Forward Step Regression, and Ordered Variable Regression when predictors are related

	HMR	FSR	OVR	Comments
Coefficient				
X_1	.389	.389	.505	Note that the coefficients produced by HMR and FSR are identical to MLR from Table 3, but not OVR.
X_2	.246	.246	.246	
Sums of Squares (Type I)				
X_1	18.854	18.854	18.854	
X_2	3.491	3.491	3.491	
Model	22.345	22.345	22.345	Note that the type I sum of squares matches across all models
Error	51.655	51.655	51.655	
Shared				
Total	74.000	74.000	74.000	
R^2 Estimates				
X_1	.2548	.2548	.2548	
X_2	.0472	.0472	.0472	
R^2_{Model}	.302	.302	.302	

*Note: HMR and FSR models were run using SAS Software 9.3, using PROC REG, GLM and STEPWISE (SAS Institute Inc., Cary, NC)

Conclusion

Although there is no agreed upon definition of multicollinearity in the literature, the impacts of multicollinearity (or interrelatedness of the predictors) are straightforward, as presented in both of the implications and impacts sections of this article; regardless of the size of the relatedness. Specifically, when the predictors are interrelated, the model coefficients for the SLR models, the MLR model, and the OVR model are all different. What is more, the shared contribution resulting from the interrelatedness in MLR is included in the overall R^2 and F, but not in the model coefficients nor in the MLR model itself. However, this is not a problem for the OVR model as the same shared contribution is included in the R^2 and F as well as the model coefficients (and thus the OVR model).

Although rare, when no interrelatedness exists between the predictors, the SLR, MLR, and OVR coefficients and R^2 values are all consistent with each other. In addition, the MLR and OVR model coefficients, R^2 values, and F test statistic are all identical. In short, when interrelatedness does not exist between the predictors, all three definitions of joint contribution and their corresponding models are identical. This is summarized in Table 4.

Table 4. Summary of Simple, Multiple, and Ordered Regression when the predictors are related (regions of Figure 2)

	SLR Models		MLR Model	OVR Model	Comments
	X_1	X_2			
Coefficients					Shared contribution is not included in MLR but is included in OVR
X_1	1,4		1	1,4	
X_2		2,4	2	2	
R^2 Estimates					
R^2_{Model}	1,4	2,4	1,2	1,2,4	Shared contribution is included in R^2 and F for MLR, although MLR does not contain shared contribution. This is not a problem for the OVR
R^2	1,4	2,4	1,2,4	1,2,4	
F Statistics					
F_{Model}	1,4	2,4	1,2	1,2,4	
F	1,4	2,4	1,2,4	1,2,4	

*Note. The shaded area indicates problems (the impacts) associated with the application of MLR.

Multicollinearity defined as the simple relatedness between predictors ($r_{12} \neq 0$) is a universal condition that exists within real data unless the predictors have been experimentally designed to be independent of each other. Consequently, the use of MLR will result in the impacts of multicollinearity as presented in this paper to an increasing degree as $|r_{12}|$ increases. Multicollinearity defined as a problematic condition that exists once $|r_{12}|$ increases beyond some threshold level, still results in the impacts presented in this paper. This second definition of multicollinearity is plagued by the need to ascertain a logical, reasonable, and appropriate threshold value. Although this is probably the more common of the two definitions, it presents the researcher with the hope of zero impact when in truth some degree of impact actually does exist (albeit smaller than the threshold amount). In either case, MLR results in a model that doesn't include the relatedness between the predictors.

OVR is presented as a method of modeling data when relatedness exists between predictors, a common issue in applied research. However, the behavior

and generalizability of OVR with regard to other common applied issues, such as small sample size and departures of model assumptions, needs to be examined. Therefore, an essential next step in the research is to use Monte Carlo simulations to evaluate statistical power (of the corresponding *F* and *t* tests) and robustness of estimation and efficiency of OVR under conditions where asymptotic behavior often breaks down.

When faced with a regression problem with multiple related predictors, a researcher is confronted with the Goldilocks dilemma (see Nestrick, 1962). It is possible to address the problem from the perspective of the multiple application of simple regression (the papa bear solution which over includes the shared variance, Equation 25), from the perspective of multiple regression (the mama bear solution which doesn't include the shared variance, Equation 34) and from the perspective of order variable regression (the baby bear solution which appropriately considers the shared variance, Equation 55).

Acknowledgements

We would like to thank Jared Studyvin for his assistance throughout the development of this project and for his insightful review of the final manuscript.

References

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Darlington, R. B. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin, 69*(3), 161-182. doi:10.1037/h0025471

Gordon, R. A. (1968). Issues in multiple regression. *American Journal of Sociology, 73*(5), 592–616.

Kerlinger, F. N. & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston, Inc.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). Homewood, IL: McGraw-Hill/Irwin.

McNemar, Q. (1962). *Psychological statistics*. (3rd ed.). New York: Wiley.

Nestrick, N. (Ed.). (1962). *The three bears and Goldilocks*. New York: Platt & Munk.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. (1975). *Statistical package for the social sciences (SPSS)* (2nd ed.). New York: McGraw-Hill.

Nunnally, J. C. (1967). *Psychometric theory*. New York: McGraw Hill.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

Weisberg, S. (2005). *Applied Linear Regression* (3rd ed.). New York: Wiley.

Woolf, B. (1951). Computation and interpretation of multiple regressions. *Journal of the Royal Statistical Society. Series B (Methodological), 13*(1), 100-119.