Journal of Modern Applied Statistical Methods

Volume 18 | Issue 1 Article 11

3-25-2020

Sampling the Porridge: A Comparison of Ordered Variable Regression with F and R^2 and Multiple Linear Regression with Corrected F and R^2 in the Presence of Multicollinearity

Grayson L. Baird
Brown University, grayson_baird@brown.edu

Stephen L. Bieber
University of Wyoming

Recommended Citation

Baird, G. L., & Bieber, S. L. (2019). Sampling the porridge: A comparison of ordered variable regression with F and R^2 and multiple linear regression with corrected F and R^2 in the presence of multicollinearity. Journal of Modern Applied Statistical Methods, 18(1), eP2522. doi: 10.22237/jmasm/1556668860

Journal of Modern Applied Statistical Methods
May 2019, Vol. 18, No. 1, eP2522.
doi: 10.22237/jmasm/1556668860

בס"ד

ISSN 1538 − 9472

Sampling the Porridge: A Comparison of Ordered Variable Regression with *F* and *R*² and Multiple Linear Regression with Corrected *F* and *R*² in the Presence of Multicollinearity

Grayson L. Baird
Brown University
Providence, RI

Stephen L. Bieber
University of Wyoming
Laramie, WY

Differences between the multiple linear regression model with Corrected R^2 and Corrected F and the ordered variable regression model with R^2 and F when intercorrelation is present are illustrated with simulated and real-world data.

Keywords: Multicollinearity, collinearity, intercorrelation, ordered variable regression, multiple linear regression, MLR, OVR

Introduction

Recent work by Baird and Bieber (2016) provided a framework whereby the correlation occurring between two or more predictors and a mutually dependent variable, referred to here as intercorrelation, can either be included in the regression model or removed completely. The model including intercorrelation was originally established by Woolf (1951) as a second method of regression, referred to here as ordered variable regression (OVR), and was demonstrated in the context of multicollinearity by Baird and Bieber. In its simplest form, the OVR model is fit by regressing X_2 on X_1, and the residuals derived from this fit result in a new predictor, $X_{2\text{ resid}}$, which is now orthogonal with X_1:

$$X_2 = b_1 X_1 + b_0, \tag{1}$$

doi: 10.22237/jmasm/1556668860 | Accepted: February 19, 2018; Published: March 25, 2020.
Correspondence: Grayson L. Baird, grayson_baird@brown.edu

which yields $X_{2\ \text{resid}}$. Then the OVR model is fit by regressing Y on X_1 and $X_{2\ \text{resid}}$:

$$Y = c_1 X_1 + c_2 X_{2\ \text{resid}} + c_0 \,. \tag{2}$$

The OVR model is evaluated for overall fit and statistical significance using R^2 and F, both of which include the redundancy resulting from the intercorrelation between the predictors and are derived using the Type I sums of squares. The model removing intercorrelation refers to traditional multiple linear regression (MLR), though instead of using R^2 and F to assess fit and statistical significance, Baird and Bieber (2016) provide the Corrected R^2 and a Corrected F that do not include the intercorrelation between predictors and are derived using the Type III sums of squares. For clarity, an abbreviated review of the distinction between OVR and MLR, along with traditional R^2 and F and the Corrected R^2 and F, is provided here; a full discussion can be found in Baird and Bieber.

When two or more predictors correlate with each other and a dependent variable in a regression context, a certain amount of redundancy is introduced; this redundancy will be illustrated using Venn diagrams. The left side of Figure 1 illustrates the situation where Areas 1 and 2 represent the unique and independent contributions on Y from predictors X_1 and X_2. The right side of Figure 1 illustrates the situation where Areas 1 and 2 also represent the unique contributions on Y from predictors X_1 and X_2, but the two predictors also share contribution, redundancy, represented by Area 4.

When no intercorrelation exists between two predictors (i.e., $r_{12} = 0.00$), the MLR and OVR model coefficients are identical in value and can both be represented with the left side of Figure 1. Likewise, the t values corresponding with said coefficients are also identical between the MLR and OVR models, as are the F and R^2, and the sums of squares, from which the F and R^2 values are derived. Thus, Areas 1 and 2 in the left side of Figure 1 represent both MLR and OVR model coefficients and corresponding t values; F and R^2 reflect Areas 1 and 2's additive composite, for both the MLR and OVR models.

However, when intercorrelation is present between predictors, redundancy is removed from the MLR coefficients, represented by Area 4 in the right side of Figure 1, leaving the non-redundant contributions, as represented by Areas 1 and 2 of the right side of Figure 1. This is evidenced in equations (3) and (4), showing ($\Sigma\, x_1 x_2$) being removed:

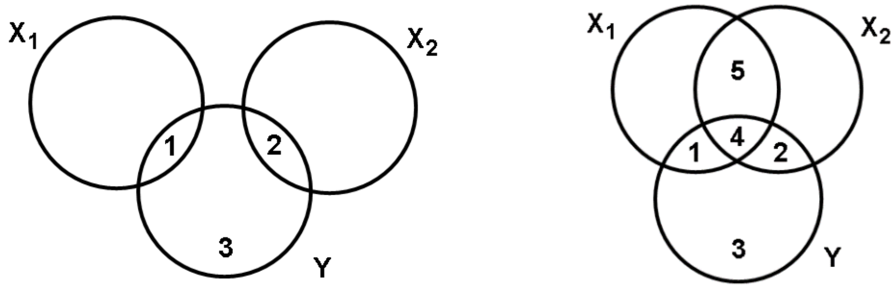

Figure 1. Predictors are unrelated (left); Predictors are related (right)

$$b_1 = \frac{\left(\sum x_2^2\right)\left(\sum x_1 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_2 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)^2}, \tag{3}$$

$$b_2 = \frac{\left(\sum x_1^2\right)\left(\sum x_2 y\right) - \left(\sum x_1 x_2\right)\left(\sum x_1 y\right)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - \left(\sum x_1 x_2\right)^2}. \tag{4}$$

This removal is also evidenced in the Type III sums of squares of X_1 and X_2:

$$\mathrm{SS}_{X_1} = \mathrm{SS}\left(X_1 \mid X_2\right), \tag{5}$$

$$\mathrm{SS}_{X_2} = \mathrm{SS}\left(X_2 \mid X_1\right). \tag{6}$$

Conversely, when any value of intercorrelation is present between predictors, the first coefficient of OVR retains the redundancy, as represented by Areas 1 and 4 in the right side of Figure 1, while the redundancy is removed from the second coefficient, as represented by Area 2 of the right side of Figure 1. This is evidenced by equations (7) and (8), which draw from equations (1) and (2), but replace X_2 with $X_{2\ \mathrm{resid}}$. Because redundancy is retained in the first OVR predictor and removed from the second, the two OVR predictors, X_1 and $X_{2\ \mathrm{resid}}$, are orthogonal (i.e., $\Sigma\ x_1 x_{2\ \mathrm{resid}} = 0$), thus equations (3) and (4) reduce to equations (7) and (8).

$$c_1 = \frac{\left(\sum x_1 y\right)}{\left(\sum x_1^2\right)}, \tag{7}$$

$$c_2 = \frac{\left(\sum x_{2\,\text{resid}} y\right)}{\left(\sum x_{2\,\text{resid}}^2\right)}. \tag{8}$$

This behavior is also evidenced by the Type I sums of squares for X_1 and X_2, where redundancy remains in the first predictor and is removed from the second:

$$\text{SS}_{X_1} = \text{SS}\left(X_1\right), \tag{9}$$

$$\text{SS}_{X_2} = \text{SS}\left(X_2 \mid X_1\right). \tag{10}$$

Because X_1 and $X_{2\,\text{resid}}$ are orthogonal [see equations (1) & (2)], it follows that the Type I sums of squares for X_1 and $X_{2\,\text{resid}}$ are identical in value to the Type I sums of squares for X_1 and X_2.

$$\text{SS}_{X_1} = \text{SS}\left(X_1\right), \tag{11}$$

$$\text{SS}_{X_{2\,\text{resid}}} = \text{SS}\left(X_{2\,\text{resid}} \mid X_1\right) = \text{SS}\left(X_2 \mid X_1\right). \tag{12}$$

Confusion arises when the model, the model fit, and inference of the model do not correspond with each other. As seen in equations (13) and (14), F and R^2 are calculated using the Type I sums of squares and thus contain the redundancy introduced by intercorrelation. Because intercorrelation is removed from the MLR coefficients [see equations (3) & (4)], F and R^2 provide inflated estimates of statistical significance and fit for the MLR model (see Baird & Bieber, 2016; also see Woolf, 1951). However, because redundancy is included in one or more OVR coefficients [see equations (1) & (2)], the F and R^2 reflect the OVR model:

$$R^2 = \frac{\left[\text{SS}\left(X_1\right) + \text{SS}\left(X_2 \mid X_1\right)\right]}{\text{SS}\left(\text{Total}\right)}, \tag{13}$$

$$F = \frac{\left[\text{SS}\left(X_1\right) + \text{SS}\left(X_2 \mid X_1\right)\right]}{\text{SS}\left(\text{Error}\right)} \cdot \frac{\left(N - P - 1\right)}{P} \tag{14}$$

given

$$\sum\left(Y_i - \bar{Y}\right)^2 = \mathrm{SS}\left(\mathrm{Total}\right), \tag{15}$$

$$\sum\left(Y_i - \hat{Y}_i\right)^2 = \mathrm{SS}\left(\mathrm{Error}\right), \tag{16}$$

where P is the number of parameters for sample size N.

Fortunately, Corrected F and Corrected R^2, which are calculated using the Type III sums of squares, cannot contain redundancy resulting from intercorrelation and provide appropriate values of statistical significance and fit for the MLR model when intercorrelation is present (Baird & Bieber, 2016).

$$R^2_{\mathrm{Corrected}} = \frac{\left[\mathrm{SS}\left(X_1 \mid X_2\right) + \mathrm{SS}\left(X_2 \mid X_1\right)\right]}{\mathrm{SS}\left(\mathrm{Total}\right)}, \tag{17}$$

$$F_{\mathrm{Corrected}} = \frac{\left[\mathrm{SS}\left(X_1 \mid X_2\right) + \mathrm{SS}\left(X_2 \mid X_1\right)\right]}{\mathrm{SS}\left(\mathrm{Error}\right)} \cdot \frac{\left(N - P - 1\right)}{P}, \tag{18}$$

given

$$\sum\left(Y_i - \bar{Y}\right)^2 = \mathrm{SS}\left(\mathrm{Total}\right), \tag{19}$$

$$\sum\left(Y_i - \hat{Y}_i\right)^2 = \mathrm{SS}\left(\mathrm{Error}\right). \tag{20}$$

As with the Corrected F and Corrected R^2, intercorrelation is removed from the t values used to evaluate the individual MLR coefficients, via the unstandardized coefficients [see equations (3) & (4)] and their standard errors [equation (22)], where b_k is an unstandardized MLR coefficient, SE_{b_k} is its standard error, σ_Y is the standard deviation of Y, and σ_{X_k} is the standard deviation for k predictors, with N sample size and P number of predictors; R^2_{12} is the redundancy term.

$$t_{b_k} = \frac{b_k}{SE_{b_k}} \tag{21}$$

where

$$SE_{b_k} = \frac{\sigma_Y}{\sigma_{X_k}} \cdot \left(\frac{1 - R^2}{\left(1 - R_{12}^2\right) \cdot \left(N - P - 1\right)} \right)^{1/2} . \tag{22}$$

Because the intercorrelation occurring between predictors is retained in the first OVR predictor and removed from the second [see equations (1) & (2)], the OVR predictors are orthogonal; thus, no redundancy is removed from their standard errors [see equation (24)], akin to the corresponding unstandardized OVR coefficients [equations (7) & (8)], where c_k is an unstandardized OVR coefficient and SE_{c_k} is its standard error.

$$t_{c_k} = \frac{c_k}{SE_{c_k}} \tag{23}$$

where

$$SE_{c_k} = \frac{\sigma_Y}{\sigma_{X_k}} \cdot \left(\frac{1 - R^2}{N - P - 1} \right)^{1/2} \tag{24}$$

Although Baird and Bieber (2016) provide a framework wherein the values of the model, model fit, and statistical significance are consistent with each other, a closer review of this framework reveals an inconsistency in the MLR model when intercorrelation is present. As can be seen in examples contained in Baird and Bieber (2016, Table 1, p. 342), when there is no intercorrelation between the predictors, the squared values of the MLR standardized coefficients, added together, equal the R^2 value:

$$b_1^2 + b_2^2 = 0.467^2 + 0.312^2 = 0.315, \quad R^2 = 0.315 . \tag{25}$$

Similarly, when intercorrelation is present, the standardized coefficients of the OVR model (located at the bottom of Table 2 of Baird & Bieber, 2016, p. 353), squared and added together, equal the R^2. However, the squared standardized coefficients of the MLR model add to neither the Corrected R^2 nor the R^2.

$$\text{OVR:} \quad c_1^2 + c_2^2 = 0.505^2 + 0.217^2 = 0.302, \quad R^2 = 0.302\,, \tag{26}$$

$$\text{MLR:} \quad b_1^2 + b_2^2 = 0.389^2 + 0.246^2 = 0.212, \quad R^2_{\text{Corrected}} = 0.137 \tag{27}$$

This result indicates a discrepancy between the standardized coefficient values and the fit of the model. However, this discrepancy only arises when intercorrelation is present, and only for the MLR model.

The primary purpose of the present study is to empirically demonstrate the differences between the MLR model (with Corrected R^2 and Corrected F) and the OVR model (with R^2 and F) as outlined in Baird and Bieber (2016) using both simulation and real-world data. The simulation study is provided to illustrate the concepts outlined by Baird and Bieber in a controlled but artificial fashion. The real-world data study is provided to demonstrate these concepts with real data from applied settings.

A secondary aim of this study is to examine the differences between the MLR and OVR models not previously outlined by Baird and Bieber (2016); namely, the relationship between the standardized and unstandardized coefficients, their corresponding t values, with model statistical significance and fit. The simulation results will be used to identify the source of the aforementioned discrepancy between the standardized MLR coefficients and Corrected R^2 values and with it, a possible solution. Then, the simulation and real-world results will be used to confirm that R^2 and F reflect the OVR model by deriving them, respectively, from the standardized OVR coefficients and their corresponding t values; likewise, the results will be used to confirm that Corrected R^2 and Corrected F reflect the MLR model by deriving them, respectively, from the standardized MLR coefficients and their corresponding t values.

Methods

Simulation Study

Design. Simulations were designed to examine OVR and MLR models under increasing values of intercorrelation and sample size. As can be seen in Table 1, each row in the table references a population with a particular degree of correlation (and, equivalently, covariance) between two predictor variables (i.e., ρ_{12} = 0.00, 0.10, 0.20, 0. 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99). No intercorrelation between predictors is included as a control comparison. Four sample sizes of interest are referenced within each row, representing the samples drawn from each population. In all, there are 11 populations and 44 samples.

Populations. . Populations for all 11 values of intercorrelation were generated using the covariance matrix found in equation (28). The covariance parameter between predictors, σ_{12}, was incrementally increased in order to increase the value of the intercorrelation (i.e., 0.00, 0.1245, 0.246, 0.368, 0.49, 0.6115, 0.733, 0.855, 0.977, 1.099, 1.219). The remaining variance and covariance parameters (σ_{11}, σ_{22}, σ_{13}, σ_{23}, σ_{33}) were held constant (i.e., 1.22, 1.22, 6099, 6099, 97710314) for all populations.

$$\mathbf{Cov} = \begin{bmatrix} 1.22 & & \\ \sigma_{12} & 1.22 & \\ 6099 & 6099 & 97710314 \end{bmatrix}, \tag{28}$$

where **Cov** = [X_1 X_2 X_3] and Mean = (3, 3, 21343). The resulting unstandardized [equation (29)] and standardized [equation (30)] population models for both MLR and OVR when σ_{12} = 0:

$$Y = B_0 + 5000X_1 + 5000X_2, \tag{29}$$

$$Y = 0.56X_1 + 0.56X_2. \tag{30}$$

Table 1. Study design

σ_{12}	ρ_{12}	Samples
0.000	0.00	*n* = 20, 30, 50, 100
0.125	0.10	*n* = 20, 30, 50, 100
0.246	0.20	*n* = 20, 30, 50, 100
0.368	0.30	*n* = 20, 30, 50, 100
0.490	0.40	*n* = 20, 30, 50, 100
0.612	0.50	*n* = 20, 30, 50, 100
0.733	0.60	*n* = 20, 30, 50, 100
0.855	0.70	*n* = 20, 30, 50, 100
0.977	0.80	*n* = 20, 30, 50, 100
1.099	0.90	*n* = 20, 30, 50, 100
1.219	0.99	*n* = 20, 30, 50, 100

Note, the relationship between each predictor and *Y* is identical, along with the variance of said predictors. Though perhaps rare in application, this scenario was designed so that the respective changes in the predictors may be evaluated relative to each other, for each population. In total, 11 populations were created, each with $N = 1000000$.

Samples. Samples were drawn from each aforementioned population for four sample sizes ($n = 20, 30, 50, 100$). Evidence from a recent meta-analysis by Mundform et al. (2011) suggested the optimal number of sample replicates for Monte Carlo simulations to produce stable results for the purpose of evaluation to be around 8000. Therefore, 8000 replicates were used for each of the 44 samples. Sampling with replacement was used in order to optimize the sampling design (Sawilowsky, 2003).

All simulations were conducted using SAS Software 9.4 (SAS Inc., Cary, NC). In an effort to reduce simulation error, populations from a multivariate normal distribution were generated using PROC IML with the RANDNORMAL function, which uses the Mersenne-Twister pseudorandom number generator (Matsumoto & Nishimura, 1998). In an effort to control for all aspects of the populations, the random seed used for simulating each population was held constant across all populations (i.e., so that any differences observed in the population could not be attributed to varying seeds). Samples were drawn using PROC SURVEYSELECT with unstructured random sampling. To reduce systematic simulation error, a different seed was used for each sample (i.e., to emulate random sampling). Code for populations and samples is provided in Appendix A.

Real Data Study

Dataset. Data were selected from a published, real-world, and publicly accessible dataset via Kuiper (2008a). The dataset example, by Kuiper (2008b), examined vehicle Price using three MLR models, namely Mileage and Liter size (Model 1), Mileage and number of Cylinders (Model 2), Mileage, Liter size, and number of Cylinders (Model 3). Although Liter size and Cylinder number both significantly predicted price for Models 1 and 2, when placed into the same model together (Model 3), Liter size was no longer statistically significant. Kuiper (2008b) concluded that the effect of multicollinearity was demonstrated by Liter no longer remaining a significant predictor in the full model, and that the source of the multicollinearity was the correlation between Liter and Cylinder, ($r = 0.96$), as they both reflect aspects of engine size.

Design. This dataset was selected to demonstrate the differences in how intercorrelation is modeled between MLR and OVR and how intercorrelation influences each model's respective fit and statistical significance in a real-world case of multicollinearity. In order to examine the three-predictor model, the OVR model was fit in the following way: regress X_2 on X_1, and the residuals derived from this fit result in a new predictor, $X_{2\ \mathrm{resid}}$:

$$X_2 = b_1 X_1 + b_0, \tag{31}$$

which yields $X_{2\ \mathrm{resid}}$. Next, regress X_3 on X_1 and $X_{2\ \mathrm{resid}}$, and the residuals derived from this fit result in a new predictor, $X_{3\ \mathrm{resid}}$:

$$X_3 = b_1 X_1 + c_2 X_{2\ \mathrm{resid}} + b_0, \tag{32}$$

which yields $X_{3\ \mathrm{resid}}$. Finally, regress Y on X_1, $X_{2\ \mathrm{resid}}$, and $X_{3\ \mathrm{resid}}$, resulting in the final OVR model:

$$Y = c_1 X_1 + c_2 X_{2\ \mathrm{resid}} + c_3 X_{3\ \mathrm{resid}} + c_0 \tag{33}$$

Statistical Methods

Unless stated otherwise, all analyses were conducted using SAS Software 9.4 (SAS Inc., Cary, NC). The following parameters were evaluated: unstandardized and standardized coefficients, sums of squares, R^2; t values and F values were also evaluated. Code for the MLR and OVR models is provided in Appendix A. PROC MEANS was used to summarize the replicate results, where the mean and 95% confidence intervals were calculated. Figures were provided for interpretation using PROC SGPANNEL. Areas under the curve were estimated using trapezoidal numerical integration with the Pracma package (Borchers, 2015) using R (R Foundation for Statistical Computing, Vienna, Austria).

Results

Simulation Study

Unstandardized Coefficients. As illustrated in Figure 2, when no intercorrelation exists between predictors, the value of the coefficients for the MLR and OVR are identical. However, as intercorrelation between the two predictors increases, the unstandardized MLR and OVR coefficients are affected differently. Specifically, as intercorrelation approaches a value of 0.99, both unstandardized MLR coefficients reduce simultaneously and equally in value. Note that as the intercorrelation approaches .90, the value of both unstandardized MLR coefficients is almost half of their value when intercorrelation was zero. This behavior is consistent with equations (3) and (4), revealing that as intercorrelation increases, the removal of intercorrelation from the MLR coefficients will reduce the value of these coefficients. It is important to note that when intercorrelation reaches a value of 0.99, both unstandardized MLR coefficients diverge in value, as one gets larger and the other smaller in value, revealing the instability of coefficient values at perfect or near-perfect intercorrelation (see Cohen et al., 2003).

Conversely, as intercorrelation approaches a value of .99, the first unstandardized OVR coefficient remains unchanged in value, while the second unstandardized OVR coefficient reduces in value in a fashion identical to both unstandardized MLR coefficients. Thus, as intercorrelation approaches .90, the second OVR coefficient decreases to half of its original value and the first OVR coefficient retains its original value. This behavior is consistent with equations (7)

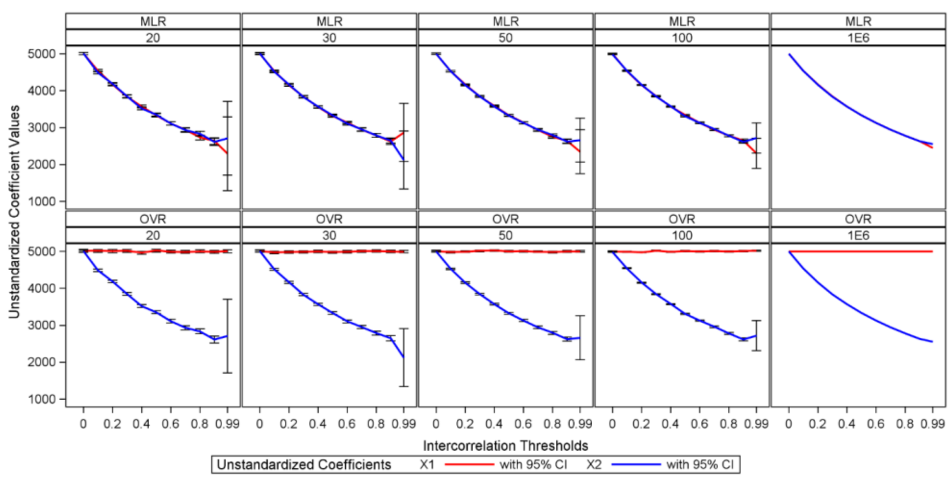

Figure 2. Unstandardized coefficients; mean values of unstandardized coefficients with 95% confidence intervals are provide across increasing sample sizes (*n* = 20, 30, 50, 100) and thresholds of intercorrelation (r_{12} = 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99) for both X_1 (red) and X_2 (blue); coefficient values are presented by model: MLR (top) and OVR (bottom); for reference, population parameter values are provided at the far right and are denoted as having a size of one million

and (8), revealing no intercorrelation term in the first OVR coefficient, thus leaving it unaffected from changes in intercorrelation, while all intercorrelation is removed in the second OVR predictor, thus reducing the second coefficient's value as intercorrelation increases. It should be clarified that intercorrelation was not removed from the second OVR coefficient, unlike the MLR coefficients [see equations (3) & (4)], but instead was removed from the second predictor [see equations (1) and (2)], which is reflected by the second OVR coefficient.

The means of the unstandardized MLR and OVR coefficient estimates approximate their population parameters increasingly well and their interval estimates reduced in value as sample size increased. However, when intercorrelation attains a value of .99, both unstandardized MLR coefficient estimate values diverge from their respective population parameter value and their interval estimate values inflate relative to all other intercorrelation conditions. In contrast, this behavior holds true only for the second unstandardized OVR coefficient, which mirrors both unstandardized MLR coefficients, while the first unstandardized OVR coefficient remains unchanged in value and variation. This behavior reveals the inflation of variation of coefficients at perfect or near-perfect intercorrelation (See Cohen et al., 2003).

Standardized Coefficients. As illustrated in Figure 3, the behavior of the standardized coefficients is similar to that of the unstandardized coefficients: when no intercorrelation exists between predictors, all standardized MLR and OVR coefficients are identical in value. As intercorrelation increases to a value of .90, both MLR coefficients reduce simultaneously and equally in value, reducing to almost half of their value relative to when intercorrelation was zero. Conversely, as intercorrelation approaches a value of .99, the first standardized OVR coefficient remains unchanged in value, while the second approaches zero.

As with the unstandardized coefficients, the means of the standardized MLR and OVR coefficient estimates more accurately approximate their respective population parameter values as sample size increases and their interval estimates reduced in value. Likewise, as intercorrelation attains a value of .99, both standardized MLR coefficient estimates diverge from their population parameter value and their interval values inflate relative to all other intercorrelation conditions. However, neither the first nor second standardized OVR coefficients deviate in value nor do their respective interval estimates widen in value when intercorrelation attains a value of .99.

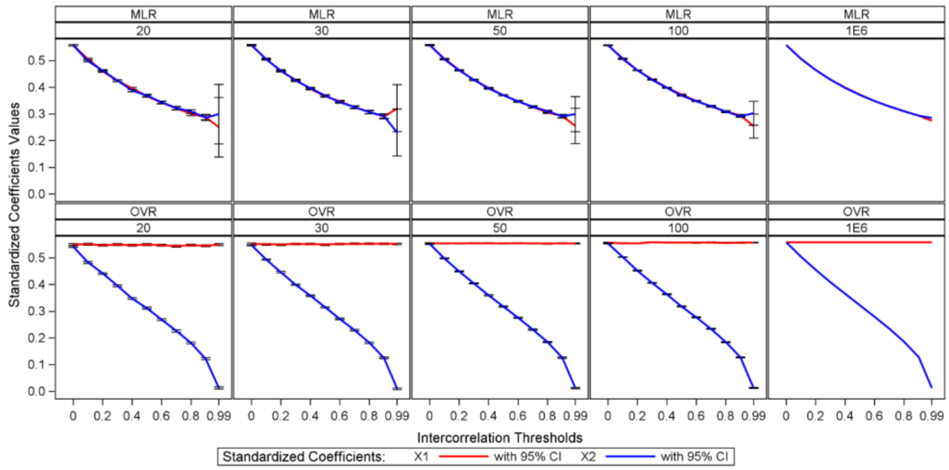

Figure 3. Standardized coefficients; mean values of standardized coefficients with 95% confidence intervals are provide across increasing sample sizes (n = 20, 30, 50, 100) and thresholds of intercorrelation (r_{12} = 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99) for both X_1 (red) and X_2 (blue); coefficient values are presented by model: MLR (top) and OVR (bottom); for reference, population parameter values are provided at the far right and are denoted as having a size of one million

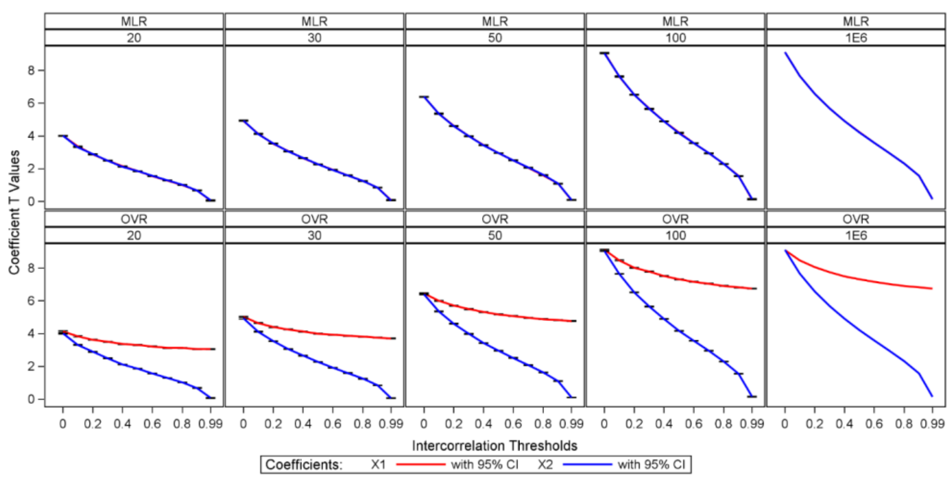

Figure 4. t values of coefficients; mean values of t with 95% confidence intervals are provide across increasing sample sizes (n = 20, 30, 50, 100) and thresholds of intercorrelation (r_{12} = 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99) for both X_1 (red) and X_2 (blue); coefficient values are presented by model: MLR (top) and OVR (bottom); note that the population of t values are not to scale nor do they exist in reality; they are provided for comparison of behavior, not value

t *Values.* As illustrated in Figure 4, the *t* values mirror the behavior of their corresponding coefficient values: when intercorrelation is zero, the *t* values are identical in value between MLR and OVR. As intercorrelation approaches .99, the respective *t* values corresponding with the MLR coefficients approach a value of zero. This is consistent with equations (3)-(4) and (21)-(22): as intercorrelation increases, the unstandardized coefficients reduce in size while their standard errors increase in size. Conversely, the *t* values corresponding with the first OVR coefficient remain roughly unaffected by intercorrelation; this is consistent with equations (7) and (23)-(24), which reveal the intercorrelation term for the first OVR coefficient and its standard error is zero. The *t* values corresponding with the second OVR coefficient approach zero as intercorrelation increases. This behavior is consistent with equations (1)-(2), revealing that intercorrelation is removed from the second predictor; this removal is reflected by its coefficient [equation (8)] and standard error [equation (24)], which will therefore reduce the *t* value to zero as intercorrelation increases.

As sample size increases, the values of *t* increase, with one exception: as intercorrelation increases to a value of .99, the values of *t* corresponding with the first coefficient of the OVR model decrease slightly. Also, the size of the interval estimates decrease for *t*, as sample size increases.

Sums of Squares. As illustrated in Figures 5 and 6, the sums of squares values corresponding with both the MLR and OVR predictors are identical when the value of intercorrelation is zero. As intercorrelation approaches a value of .99, the sums of squares values corresponding with the MLR predictors approach a value of zero in unison. This is consistent with equations (5) and (6), revealing that as intercorrelation increases between predictors, the removal of redundancy will reduce the value of the Type III sums of squares corresponding with each predictor, where complete redundancy will result in a value of zero. Conversely, as intercorrelation approaches a value of .99, the sums of squares corresponding with the first OVR predictor does not change in value while the value of the sums of squares for the second OVR predictor approaches zero in an identical fashion to the sums of squares corresponding with the MLR predictors. This is consistent with equations (9)-(12), which reveal that no redundancy is removed from the Type I sums of squares for the first predictor, while the redundancy between the first and second predictors is removed from the second predictor's sums of squares.

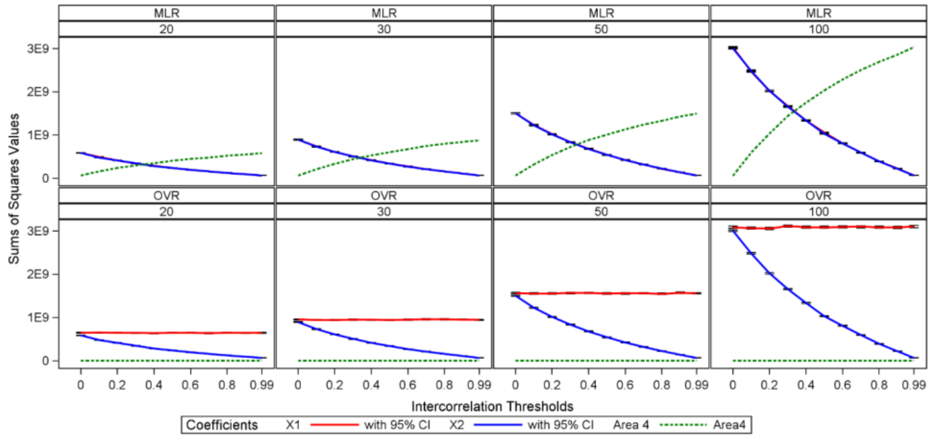

Figure 5. Predictor sums of squares for samples; mean values of X_1 (red), X_2 (blue), and Area 4 (green) with 95% confidence intervals are provide across increasing sample sizes (n = 20, 30, 50, 100) and thresholds of intercorrelation (r_{12} = 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99) for both; values are presented by model: MLR (top) and OVR (bottom); Note that Area 4 is part of X_1 (red) for OVR but is shown as a green line for comparison with MLR

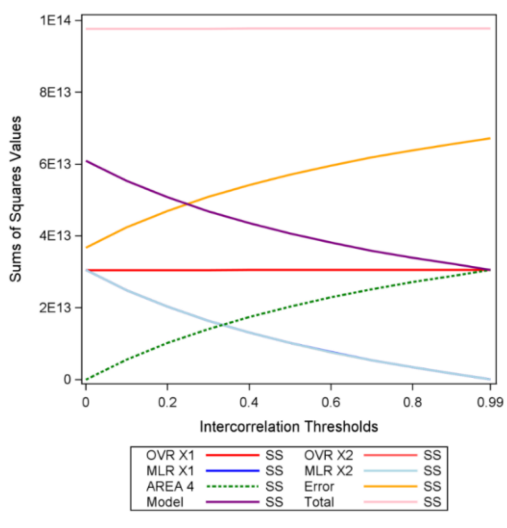

Figure 6. Sums of squares for populations; population values of sums of squares are provided across thresholds of intercorrelation (r_{12} = 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99) for X_1 (dark) and X_2 (light) along with Area 4 (green), Error (orange), Model Total (purple), and Total (pink) for both the MLR (blue) and OVR (red) models; Note that OVR X_2 and MLR X_1 are hidden behind MLR X_2

Table 2. Area under the curve of the population sums of squares (Figure 6) for X_1, X_2, Area 4, Error, and Total using trapezoidal numerical integration

Component	OVR Type I	MLR Type III
X_1	0.30895	0.12082
X_2	0.12061	0.12061
Area 4	Part of X_1	0.18805
Error	0.56058	0.56058
Total	0.99014	0.99006

The total and error sums of squares are identical for the Type I and III sums of squares [see equations (15)-(16), (19)-(20)]. This is evident in Figure 6, which shows the values of the total sums of squares and the error sums of squares are identical for both the MLR and OVR models. As intercorrelation approaches a value of .99, the value of the total sums of squares remains unchanged, while the values of the error sums of squares increase. It is important to note that although the total and the error sums of squares are identical for both models, only the sums of squares of the OVR model add to the total sums of squares when the value of intercorrelation is not zero. Specifically, when the value of intercorrelation is above zero, the sums of squares for the MLR predictors do not add to the total sums of squares, as evidenced by Figure 6 and in Table 2, which presents the value of the area under the curve for each component of the sums of squares.

The discrepancy between the total sums of squares and the sums of squares of the MLR predictors is illustrated in Figures 5 and 6 as "Area 4" (see the right side of Figure 1). When the value of intercorrelation is zero, the value of Area 4 is also zero. As intercorrelation approaches a value of .99, the value of Area 4 approaches the value of the sums of squares of the first predictor of the OVR model (and also the value of all predictors when intercorrelation is zero). This reveals Area 4 to be the mathematical complement of the sums of squares reduction in the MLR predictors, as also evidenced by the areas under the curve for the entire range of intercorrelation presented in Table 2. The empirical discrepancy observed here between the total sums of squares and the sums of squares of the MLR predictors, or Area 4, confirms the deficit observed by Woolf (1951).

R^2 and Corrected R^2 Values. As illustrated in Figure 7, when the value of intercorrelation is zero, R^2 and Corrected R^2 are identical in value. However, as the value of intercorrelation approaches .99, R^2 reduces to half of its original value when no intercorrelation was present, while Corrected R^2 approaches a value of zero; this behavior is consistent with equations (13) and (17), respectively. For the

OVR model, when intercorrelation nears .99, sums of squares for the first predictor retains its original value while the second approaches zero—thus the combined sums of squares for both predictors when intercorrelation approaches one is half of their combined value relative to when intercorrelation is zero—the exact value of R^2. For the MLR model, both predictor sums of squares approach a value of zero in unison as the value of intercorrelation approaches .99—the exact value of Corrected R^2.

The mean values of R^2 and Corrected R^2 more accurately approximate their respective population parameters as sample size increases as well as their interval estimates reduce in value. Note, a clear inflation in value of both R^2 and the Corrected R^2 exists due to no adjustment factor being used, though this inflation diminishes with increasing sample size.

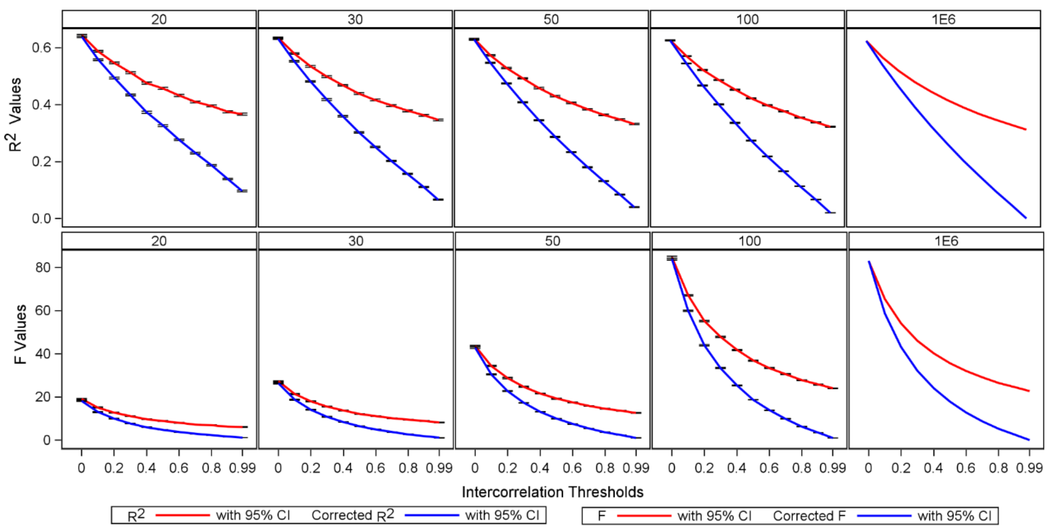

Figure 7. R^2 and Corrected R^2, F and Corrected F; mean values of R^2 (Red) and Corrected R^2 (Blue) and F (Red) and Corrected F (Blue) with 95% confidence intervals are provide across increasing sample sizes (n = 20, 30, 50, 100) and thresholds of intercorrelation (r_{12} = 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99) for both; for reference, population parameter values are provided at the far right and are denoted as having a size of one million; note that the population of F and Corrected F are not to scale nor do they exist in reality; they are provided for comparison of behavior, not value

F *and Corrected* F *Values.* As illustrated in Figure 7, when the value of the intercorrelation is zero, *F* and Corrected *F* are identical in value. However, as the value of intercorrelation increases to a value of .99, the value of Corrected *F* approaches a value of zero, while *F* nears half of its value relative to when the value of intercorrelation was zero; this behavior is consistent with equations (14) and (18), respectively. For the OVR model, when intercorrelation nears .99, the sums of squares for the first predictor retains its original value while the second approaches zero—thus the combined sums of squares for both predictors when intercorrelation approaches one would be half of their combined value relative to when intercorrelation was zero—the approximate value of *F*. For the MLR model, both predictor sums of squares approach a value of zero in unison as the value of intercorrelation approaches .99—the exact value of Corrected *F*. In addition, the value of *F* and Corrected *F* increased with the increase in sample size. Note that the population for *F* is not to scale (nor does it exist in reality) and is provided for comparison in behavior only, not value.

Real Data

As illustrated in Table 3, the results demonstrating multicollinearity from Kuiper (2008b) are replicated using three MLR models of vehicle Price: Mileage and Cylinder (MLR Model 1), Mileage and Liter (MLR Model 2), Mileage, Liter, Cylinder (MLR Model 3). Evidence of multicollinearity was confirmed when comparing the three models, where Liter and Cylinder both significantly predict Price for Model 1 and Model 2, respectively, but when placed into the same model together (MLR Model 3), Liter no longer remains a statistically significant predictor. For comparison, two OVR models, along with Corrected F and Corrected R^2 values for the MLR models, are also included. For brevity, both OVR models assume Mileage to be the most important predictor, thus it is the first predictor in both models. Liter precedes Cylinder in OVR Model 1; thus, Liter is assumed to be more important than Cylinder; Cylinder precedes Liter in OVR Model 2; thus, Cylinder is assumed to be more important than Liter.

The results from the real-data study complement those found in simulation study. Although the full MLR model (i.e., Model 3) and the two OVR models have identical predictors, each model produces a different set of coefficient values, save for Mileage, which is not highly correlated with any other predictor. Specifically, the unstandardized coefficients of the MLR Model 3 for Cylinder and Liter correspond in value with the value of the last unstandardized coefficient of OVR model 1 and 2 (Cylinder and Liter, respectively). This result demonstrates that the

MLR unstandardized coefficient values for Cylinder and Liter represent their respective, non-redundant contribution after removing any overlapping contribution from each other and Mileage.

These results also reveal the last predictor in each OVR model produces an unstandardized coefficient value that represents the unique, non-redundant contribution after removing any overlapping contribution from only preceding predictors; at the same time, coefficient values preceding the last predictor are unaffected by the subsequent predictors. Thus, the unstandardized coefficient value of Liter for OVR Model 1 is unaffected by Cylinder's presence and is therefore the same value as Liter's coefficient value in the MLR model that does not include Cylinder (MLR Model 2). Likewise, the unstandardized coefficient value of Cylinder for OVR Model 2 is unaffected by Liter's presence, which is therefore the same value as Cylinder's value in the MLR Model 1 that does not include Liter.

The results also demonstrate the differences between R^2 and F and Corrected R^2 and Corrected F when intercorrelation is present, as seen in the simulation results. As can be seen in Table 3, the inflation of R^2 and F are large, relative to the value of Corrected R^2 and Corrected F, when Liter and Cylinder are included together in MLR Model 3. The values of Corrected R^2 and Corrected F can be confirmed using the Type III sums of squares. Note that when intercorrelation is not high, the inflation of R^2 and F relative to Corrected R^2 and Corrected F is minimal, such as is the case with MLR models 1 and 2, with only Mileage and Cylinder or Mileage and Liter, respectively.

A general observation of the real-data results, aside from comparisons with the simulation study results, demonstrates how OVR is an alternative modeling approach to MLR. Suppose a researcher would like to use all three predictors in the model, but the predictors have a particular order of importance. For instance, let us assume that Mileage is the most important aspect of Price among the three predictors. Next, Kuiper (2008b) noted that Liter size is a more precise measure of the engine than number of Cylinders. Therefore, let us consider Liter a more important predictor of Price than Cylinder, thereby making Cylinder the least important predictor. Found in Table 3 is the resulting model, OVR Model 1, along with the MLR Model 3 for comparison.

Table 3. Real-data example comparing MLR and OVR models

		MLR model 1	**MLR model 2**	**MLR model 3**	**OVR model 1**		**OVR model 2**
Unstandardized coefficient	Intercept	3145.8 (1325.9)	9426.6 (1095.1)	4707.6 (1602.9)	24765 (741.99)	Intercept	24765 (741.98)
	X_1 (mileage)	-0.15243 (0.03)	-0.16003 (0.04)	-0.15443 (0.03)	-0.173 (0.04)	X_1 (mileage)	-0.173 (0.04)
	X_2 (liter)		4968.29 (258.8)	1545.3 (893.4)	4968.29 (256.4)	X_2 (cylinder)	4027.67 (204.4)
	X_3 (cylinder)	4027.67 (204.6)		2847.9 (712.0)	2847.93 (712.0)	X_3 (liter)	1545.25 (893.4)
Standardized coefficient	X_1 (mileage)	-0.12639	-0.13269	-0.12805	-0.14305	X_1 (mileage)	-0.14305
	X_2 (liter)		0.55567	0.17283	0.55558	X_2 (cylinder)	0.56512
	X_3 (cylinder)	0.56536		0.39976	0.11468	X_3 (liter)	0.04959
Type I sums of squares	X_1 (mileage)	1605590375	1605590375	1605590375	1605590375	X_1 (mileage)	1605590375
	X_2 (liter)		24218240323	24218240323	24218240323	X_2 (cylinder)	25057212321
	X_3 (cylinder)	25057212321		1031948046	1031948046	X_3 (liter)	192976048
	Error	51798580168	52637552166	51605604120	51605604120		51605604120
	Total	78461382864	78461382864	78461382864	78461382864		78461382864
Type III sums of squares	X_1 (mileage)	1252374754	1381011542	1283996660	1605590375	X_1 (mileage)	1605590375
	X_2 (liter)		24218240323	192976048	24218240323	X_2 (cylinder)	25057212321
	X_3 (cylinder)	25057212321		1031948046	1031948046	X_3 (liter)	192976048
	Error	51798580168	52637552166	51605604120	51605604120		51605604120
	Total	78461382864	78461382864	78461382864	78461382864		78461382864
t value, *p*-value	X_1 (mileage)	-4.40, <.0001	-4.58, <.0001	-4.46, <.0001	-4.99, .0001	X_1 (mileage)	-4.99, <.0001
	X_2 (liter)		19.20, <.0001	1.73, 0.0841	19.38, <.0001	X_2 (cylinder)	19.71, <.0001
	X_3 (cylinder)	19.68, <.0001		4.00, <.0001	4.00, <.0001	X_3 (liter)	1.73, 0.0841
Fit & test statistics	*F* value, *p*-value	206.15, <.0001	196.48, <.0001	138.77, <.0001	138.77, <.0001		138.77, <.0001
	Corrected *F*, *p*-value	203.17	194.53	12.96	138.77, <.0001		138.77, <.0001
	R^2	0.3398	0.3291	0.3423	0.3423		0.3423
	Corrected R^2	0.3353	0.3263	0.032	0.3423		0.3423

Note: Estimate (Standard Error); N = 804; Also note: The Type III sums of squares for the MLR models do not add to the total sums of squares (i.e., Woolf's deficit)

Although the order of the predictors in the MLR Model 3 is identical to that of the OVR Model 1, the coefficient values are different in both value and interpretation. Specifically, for the MLR model, with every one-unit increase in Mileage, the Price of the vehicle decreases $0.15, after removing the contribution of both Liter and Cylinder; for everyone one-unit increase in Liter, the Price of the vehicle increases $1,545.30, after removing the contribution of Mileage and Cylinder; for every one-unit increase in Cylinder, the Price of the vehicle increases $2,847.90, after removing the contribution of Mileage and Liter. Note, Liter is no longer a statistically significant contributor to the model and thus could be "dropped". For the OVR model, with every one-unit increase in Mileage, the Price of the vehicle decreased $0.17; for every one-unit increase in Liter size, the Price of the vehicle increases $4,968.29, after removing the contribution of Mileage; for every one-unit increase in Cylinder, the least important predictor, the Price of the vehicle increases $2,847.93, after removing the contribution of Mileage and then Liter size.

The MLR model removes the redundant contribution from all predictors, simultaneously, while the OVR model removes the redundant contribution, sequentially. The differences in the Liter and Cylinder values, between the MLR and OVR models, reveal the magnitude of the intercorrelation between these predictors, with the MLR model removing all intercorrelation, as reflected by the Corrected R^2 and Corrected F while the OVR retains the intercorrelation, as evidenced by the R^2 and F. In doing so, the OVR Model 1 addresses a possible theoretical or pragmatic need to model the predictors in a specific order and the resulting coefficient values reflect this order. What's more, this chosen order allowed all predictors to be statistically significant in the OVR Model 1 but not the MLR model.

Standardized Coefficients and Corrected Standardized MLR Coefficients with Model Fit

The simulation results illustrate an inconsistency in the value of the standardized coefficients. Specifically, although both unstandardized coefficients of the MLR model are identical in value to the second unstandardized OVR coefficient (see Figure 2), this is not true for the standardized coefficients when intercorrelation is present (see Figure 3). As can be seen in equation (34), the value of the standardized coefficients is the product of the unstandardized coefficient and the ratio of the standard deviation of a given predictor and the standard deviation of the dependent variable,

$$\beta_k = b_k \cdot \frac{\sigma_{X_k}}{\sigma_Y} \tag{34}$$

where β_k is the standardized coefficient, b_k is the unstandardized coefficient, σ_{X_k} is the standard deviation, all of the k predictor, and σ_Y is the standard deviation of Y. Given that the values between the unstandardized MLR coefficients and the second unstandardized OVR coefficient are identical (see Figure 2), and because the standard deviation of the dependent variable, Y, is constant, the observed difference between the standardized MLR coefficients and the second standardized OVR coefficient, as seen in Figure 3, must be attributed to the difference in value of the standard deviation of the predictors. Note, the discrepancy between the standardized MLR coefficients and the second OVR coefficient occurs only when intercorrelation is present.

The reason for this discrepancy is clear: when intercorrelation is present, the unstandardized MLR coefficients and second OVR coefficient reduce proportionally and identically with the increase in intercorrelation. Although the unstandardized MLR coefficients reduce in value as intercorrelation increases, the standard deviation of each MLR predictor remains unchanged in value. However, this is not the case with the second standardized OVR coefficient—both the second unstandardized OVR coefficient and the second OVR predictor reduce in value, together, as intercorrelation increases. Because the second OVR predictor is the residual of the first predictor regressed on the second [see equations (1) & (2)], the second OVR predictor and its standard deviation change in concert with the second unstandardized OVR coefficient.

The discrepancy between the standardized MLR coefficients and the second OVR coefficient has a more general impact on interpretation. As demonstrated graphically with Figures 3 and 7, when no intercorrelation is present, the standardized coefficient values for both the MLR and OVR models reflect their individual contributions to the model — this can be verified by squaring the coefficient values and the adding them together, which results in the value of the R^2 (and equivalently, Corrected R^2 value). This remains true for the standardized OVR coefficients when intercorrelation is present — with each increase intercorrelation, the standardized coefficients of the OVR model, squared and added together, equals the value of R^2. Thus, when intercorrelation is near complete, the first OVR coefficient represents almost everything and the second almost nothing.

Conversely, with each increase in intercorrelation, the standardized coefficients of the MLR model, squared and added together, do not equal the value of Corrected R^2 (nor R^2). Because the value of the standard deviation of the MLR predictors is constant, the resulting standardized MLR coefficients represent not their contribution to the model, but rather how much they contribute, relative to themselves without intercorrelation. This can be seen when intercorrelation is near complete: the standardized MLR coefficients added together equal their respective values when no intercorrelation was present (see Figure 3). Fortunately, a solution exists which resolves the discrepancy between the standardized MLR coefficients and the second standardized OVR coefficient. More importantly, this solution also enables the standardized MLR coefficients to reflect their contribution to the model without changing the value of their corresponding unstandardized MLR coefficients.

Because the unstandardized MLR coefficients change in value when intercorrelation is present while the standard deviations of the MLR predictors remain constant, the resulting standardized coefficients reflect only a partial removal of redundancy from the model. Thus, to reflect the full removal of redundancy from the standardized MLR coefficients, the standard deviation of the predictors also needs to reduce in value. To calculate the (Corrected) standard deviation, fit an individual OVR model for each predictor originally in the MLR model, entering each of these predictors last, allowing all the other predictors to proceed it. For the two-predictor case, simply regress X_1 on X_2, and the residuals derived from this fit results in a new predictor, $X_{1\ \text{resid}}$:

$$X_1 = c_2 X_2 + c_0\,, \tag{35}$$

which yields $X_{1\ \text{resid}}$. Likewise, regress X_2 on X_1, and the residuals derived from this fit results in a new predictor, $X_{2\ \text{resid}}$:

$$X_2 = c_1 X_1 + c_0\,, \tag{36}$$

which yields $X_{2\ \text{resid}}$. Next, calculate the standard deviation of each new (residual) predictor from each OVR model:

$$\sigma\left(X_{1\ \text{resid}}\right) = \sigma_{X_{1\ \text{resid}}} \quad \text{and} \quad \sigma\left(X_{2\ \text{resid}}\right) = \sigma_{X_{2\ \text{resid}}}\,. \tag{37}$$

Finally, calculate the (Corrected) standardized MLR coefficients using the (Corrected) standard deviation values derived from the new (residual) predictors:

$$\beta_{1(\text{Corrected})} = b_1 \cdot \frac{\sigma_{X_{1\,\text{resid}}}}{\sigma_Y}, \tag{38}$$

$$\beta_{2(\text{Corrected})} = b_2 \cdot \frac{\sigma_{X_{2\,\text{resid}}}}{\sigma_Y}. \tag{39}$$

Note, the only difference in equations (38) and (39) from equation (34) is the use of the respective standard deviations of each new (residual) predictor (i.e., the Corrected standard deviation).

The Corrected standardized MLR coefficients are demonstrated empirically using the simulation study results, illustrated in Figure 8. As can be seen comparing Figure 8 with Figure 7, the squared Corrected standardized MLR coefficients for each level of intercorrelation now add to the corresponding Corrected R^2, just as the squared standardized OVR coefficients for each level of intercorrelation add to R^2. Now the MLR standardized coefficients reflect their individual contribution to the model; they also are identical in value and behavior to the second standardized OVR coefficient. Note, at near perfect intercorrelation, the variances of the estimates are no longer inflated nor are they unstable.

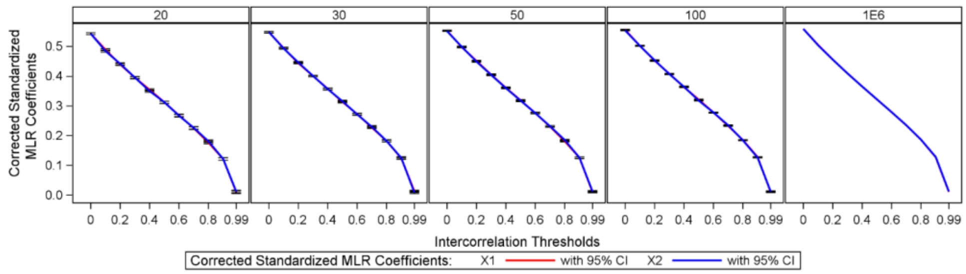

Figure 8. Corrected standardized MLR coefficients values; mean values of Corrected standardized MLR coefficients with 95% confidence intervals are provided across increasing sample sizes (n = 20, 30, 50, 100) and thresholds of intercorrelation (r_{12} = 0.00, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, 0.99) for both X_1 (red) and X_2 (blue); for reference, population parameter values are provided at the far right and are denoted as having a size of one million

Using *t* Values to Confirm Model Statistical Significance and Model Fit

The calculation of the *t* values for the MLR and OVR coefficients is identical, though as seen in equations (21)-(22), because the OVR predictors are orthogonal, the *t* values reduce to equations (23)-(24) given that there is no redundancy term. It should be noted that the *t* values for the MLR coefficients can also be calculated without a redundancy term if using the Corrected standard deviations, as evidenced by equation (40). The former is the traditional formula for *t* that explicitly removes the intercorrelation. The latter is the same equation, but instead of removing the intercorrelation, the Corrected standard deviation of the MLR predictor is used [see equation (38)], which does not include any redundancy to remove.

$$t_{b_k} = \frac{b_k}{\frac{\sigma_Y}{\sigma_{X_k}} \cdot \left(\frac{1-R^2}{\left(1-R_{12}^2\right) \cdot \left(N-P-1\right)} \right)^{1/2}} = \frac{b_k}{\frac{\sigma_Y}{\sigma_{X_{k\ \text{resid}}}} \cdot \left(\frac{1-R^2}{\left(N-P-1\right)} \right)^{1/2}} \tag{40}$$

The simulation results will now be used to illustrate how *F* and Corrected *F*, R^2 and Corrected R^2 can be derived using the *t* values. As can be seen in Figures 4 and 7 of the simulation study results, the *t* values corresponding with the MLR coefficients for each level of intercorrelation, squared, added together and divided by the number of predictors (i.e., $P = 2$), equal the value of the corresponding Corrected *F* values.

$$F_{\text{Corrected}} = \frac{t_{b_1}^2 + t_{b_2}^2}{P} \tag{41}$$

As seen in Figures 4 and 7, the *t* values corresponding with the OVR coefficients for each level of intercorrelation, squared, added together and divided by the number of predictors (i.e., $P = 2$), equal the value of the corresponding *F* values.

$$F = \frac{t_{c_1}^2 + t_{c_2}^2}{P} \tag{42}$$

The Corrected R^2 can also be derived from the MLR *t* values, using Corrected *F* values. For each level of intercorrelation, the *t* values, squared, added together, and divided by the number of predictors (i.e., $P = 2$), equals the same value as the Corrected F value, which is then used to derive the Corrected R^2:

$$R^2_{\text{Corrected}} = \frac{F_{\text{Corrected}}}{\left(F + \frac{N-P-1}{P}\right)}, \tag{43}$$

where N is the sample size for P number of predictors. Likewise, the t values corresponding with unstandardized coefficients of OVR model, squared, added together, and divided by the number of predictors can be used to derive the F, which is then used to derive the R^2:

$$R^2 = \frac{F}{\left(F + \frac{N-P-1}{P}\right)}. \tag{44}$$

Of course, the Corrected R^2 and R^2 can be used to derive the Corrected F and F:

$$F_{\text{Corrected}} = \frac{R^2_{\text{Corrected}}}{\left(1-R^2\right)} \cdot \frac{N-P-1}{P}, \tag{45}$$

$$F = \frac{R^2}{\left(1-R^2\right)} \cdot \frac{N-P-1}{P}, \tag{46}$$

where N is the sample size for P number of predictors.

Empirical Demonstration: Using Standardized Coefficients and *t* Values to Derive Model Fit and Statistical Significance

Tables 4 and 5 empirically demonstrate the two-predictor situation for the MLR and OVR models using the simulation results, where the relationships between the unstandardized coefficients, standardized coefficients, Corrected standard deviations, t values, F and Corrected F values, R^2 and Corrected R^2 values, and Type I and Type III sums of squares are provided, for each increase in intercorrelation. For reference, the standard deviation of the predictors and the original standardized MLR coefficients are provided. For precision, only the population values are provided.

Table 4. Simulation study (population) MLR model results

r_{12}		B	σ	Corr σ	β	Corr β	t values	TYPE III X SS	Error SS Total SS	F values	Corr F values	$\Sigma(t^2)/2$	R^2 values	Corr R^2 values	Corr $\Sigma(\beta^2)$
0.00	X_1	5003.59	1.10	1.10	0.56	0.56	911.69	3.05E+13	3.67E+13	8.30E+05	8.31E+05	8.31E+05	0.62	0.62	0.62
	X_2	5000.77	1.10	1.10	0.56	0.56	911.38	3.05E+13	9.77E+13						
0.10	X_1	4540.82	1.10	1.10	0.51	0.50	766.34	2.49E+13	4.24E+13	6.53E+05	5.87E+05	5.87E+05	0.57	0.51	0.51
	X_2	4537.98	1.10	1.10	0.51	0.50	765.94	2.49E+13	9.77E+13						
0.20	X_1	4164.91	1.10	1.08	0.47	0.46	657.50	2.03E+13	4.70E+13	5.40E+05	4.32E+05	4.32E+05	0.52	0.42	0.42
	X_2	4162.14	1.10	1.08	0.46	0.46	657.05	2.03E+13	9.77E+13						
0.30	X_1	3845.27	1.10	1.05	0.43	0.41	567.87	1.64E+13	5.09E+13	4.61E+05	3.22E+05	3.22E+05	0.48	0.34	0.34
	X_2	3842.65	1.10	1.05	0.43	0.41	567.41	1.64E+13	9.77E+13						
0.40	X_1	3571.17	1.10	1.01	0.40	0.37	490.76	1.31E+13	5.42E+13	4.02E+05	2.41E+05	2.41E+05	0.45	0.27	0.27
	X_2	3568.78	1.10	1.01	0.40	0.37	490.32	1.30E+13	9.77E+13						
0.50	X_1	3334.4	1.10	0.96	0.37	0.32	421.92	1.02E+13	5.71E+13	3.56E+05	1.78E+05	1.78E+05	0.42	0.21	0.21
	X_2	3332.35	1.10	0.96	0.37	0.32	421.52	1.01E+13	9.77E+13						
0.60	X_1	3126.98	1.10	0.88	0.35	0.28	357.72	7.63E+12	5.96E+13	3.20E+05	1.28E+05	1.28E+05	0.39	0.16	0.16
	X_2	3125.43	1.10	0.88	0.35	0.28	357.40	7.62E+12	9.77E+13						
0.70	X_1	2942.97	1.10	0.79	0.33	0.23	294.96	5.38E+12	6.19E+13	2.90E+05	8.69E+04	8.69E+04	0.37	0.11	0.11
	X_2	2942.18	1.10	0.79	0.33	0.23	294.74	5.37E+12	9.78E+13						
0.80	X_1	2779.06	1.10	0.66	0.31	0.19	230.18	3.38E+12	6.39E+13	2.66E+05	5.30E+04	5.30E+04	0.35	0.07	0.07
	X_2	2779.57	1.10	0.66	0.31	0.19	230.11	3.38E+12	9.78E+13						
0.90	X_1	2631.49	1.10	0.48	0.29	0.13	155.84	1.59E+12	6.56E+13	2.45E+05	2.43E+04	2.43E+04	0.33	0.03	0.03
	X_2	2634.92	1.10	0.48	0.29	0.13	155.98	1.60E+12	9.78E+13						
0.99	X_1	2453.35	1.10	0.04	0.27	0.01	13.38	1.20E+10	6.72E+13	2.27E+05	1.87E+02	1.87E+02	0.31	0.00	0.00
	X_2	2553.88	1.10	0.04	0.29	0.01	13.93	1.30E+10	9.78E+13						

Note: Corr denotes "Corrected"; N = 1000000; STD(Y): 9887.0; R^2 and F are provided for comparison only, they do not correspond with MLR when intercorrelation is present; when r_{12} = 0.99, $\beta_1 + \beta_2$ = .56, as indicated on page 24

Table 5. Simulation study (population) OVR model results

r_{12}		B	σ	β	t values	TYPE III X SS	Error SS Total SS	F values	$\Sigma(t^2)/2$	R^2 values	$\Sigma(\beta^2)$
0.00	X_1	4997.85	1.10	0.56	910.64	3.05E+13	3.67E+13	8.30E+05	8.30E+05	0.62	0.62
	X_2	5000.77	1.10	0.56	911.38	3.05E+13	9.77E+13				
0.10	X_1	4998.74	1.10	0.56	847.95	3.05E+13	4.24E+13	6.53E+05	6.53E+05	0.57	0.57
	X_2	4537.98	1.10	0.50	765.94	2.49E+13	9.77E+13				
0.20	X_1	4999.49	1.10	0.56	805.61	3.05E+13	4.70E+13	5.40E+05	5.40E+05	0.52	0.52
	X_2	4162.14	1.08	0.46	657.05	2.03E+13	9.77E+13				
0.30	X_1	5000.16	1.10	0.56	774.22	3.05E+13	5.09E+13	4.61E+05	4.61E+05	0.48	0.48
	X_2	3842.65	1.05	0.41	567.41	1.64E+13	9.77E+13				
0.40	X_1	5000.79	1.10	0.56	750.06	3.05E+13	5.42E+13	4.02E+05	4.02E+05	0.45	0.45
	X_2	3568.78	1.01	0.37	490.32	1.30E+13	9.77E+13				
0.50	X_1	5001.37	1.10	0.56	730.94	3.05E+13	5.71E+13	3.56E+05	3.56E+05	0.42	0.42
	X_2	3332.35	0.96	0.32	421.52	1.01E+13	9.77E+13				
0.60	X_1	5001.93	1.10	0.56	715.37	3.05E+13	5.96E+13	3.20E+05	3.20E+05	0.39	0.39
	X_2	3125.43	0.88	0.28	357.40	7.62E+12	9.77E+13				
0.70	X_1	5002.49	1.10	0.56	702.37	3.05E+13	6.19E+13	2.90E+05	2.90E+05	0.37	0.37
	X_2	2942.18	0.79	0.23	294.74	5.37E+12	9.78E+13				
0.80	X_1	5003.08	1.10	0.56	691.42	3.05E+13	6.39E+13	2.66E+05	2.66E+05	0.35	0.35
	X_2	2779.57	0.66	0.19	230.11	3.38E+12	9.78E+13				
0.90	X_1	5003.76	1.10	0.56	682.06	3.05E+13	6.56E+13	2.45E+05	2.45E+05	0.33	0.33
	X_2	2634.92	0.48	0.13	155.98	1.60E+12	9.78E+13				
0.99	X_1	5005.03	1.10	0.56	674.18	3.05E+13	6.72E+13	2.27E+05	2.27E+05	0.31	0.31
	X_2	2553.88	0.04	0.01	13.93	1.30E+10	9.78E+13				

Note: N = 1000000; STD(Y): 9887.0

Table 6. Confirmation using real-world data

		OVR A	OVR B	OVR C	MLR	
Unstandardized coefficients	Intercept	6185.75	24765	24765	4707.6	
	Mileage	**X_3: -0.154**	-0.173	-0.173	**-0.1544**	
	Liter	4990.39	**X_3: 1545.2**	4968.29	**1545.3**	
	Cylinder	2976.36	4027.67	**X_3: 2847.9**	**2847.9**	
					Original	Corrected
Standardized coefficients	Mileage	-0.1279	-0.14305	-0.14305	-0.12805	**-0.12792**
	Liter	0.5582	0.04959	0.55558	0.17283	**0.04959**
	Cylinder	0.11995	0.56512	0.11468	0.39976	**0.11468**
Standard deviations	Price				9884.85	9884.85
	Mileage	8188.2			8196.32	8188.2
	Liter		0.3172448		1.10556	0.31725
	Cylinder			0.3980532	1.38753	0.39805
Type I sums of squares	Mileage	**1283996660**	1605590375	1605590375	1605590375	
	Liter	24442819155	**192976048**	24218240323	24218240323	
	Cylinder	1128962928	25057212321	**1031948046**	1031948046	
	Error	51605604120	51605604120	51605604120	51605604120	
	Total	78461382864	78461382864	78461382864	78461382864	
Type III sums of squares	Mileage	**1283996660**	1605590375	1605590375	**1283996660**	
	Liter	24442819155	**192976048**	24218240323	**192976048**	
	Cylinder	1128962928	25057212321	**1031948046**	**1031948046**	
	Error	51605604120	51605604120	51605604120	51605604120	
	Total	78461382864	78461382864	78461382864	78461382864	
t value, *p*-value	Mileage	**-4.46, <.0001**	-4.99, <.0001	-4.99, .0001	**-4.46, <.0001**	
	Liter	19.47, <.0001	**1.73, 0.0841**	19.38, <.0001	**1.73, 0.0841**	
	Cylinder	4.18, 0.0841	19.71, <.0001	**4.00, <.0001**	**4.00, <.0001**	
Fit & test statistics	*F* value, *p*-value	**138.77, <.0001**	**138.77, <.0001**	**138.77, <.0001**	138.77, <.0001	
	Corrected *F*, *p*-value	138.77, <.0001	138.77, <.0001	138.77, <.0001	**12.96, <.0001**	**12.96**
	R^2	**0.3423**	**0.3423**	**0.3423**	0.342	
	Corrected R^2	0.3423	0.3423	0.3423	**0.032**	**0.032**

Note: *N* = 804; bolded numbers reflect order being last; the Type III sums of squares for the MLR model do not add to the total sums of squares (i.e., Woolf's deficit)

For each level of intercorrelation, the Corrected standardized MLR coefficients values, squared and added together, equal the value of Corrected R^2 while the standardized OVR coefficients values, squared and added together, equal the value of R^2. Note, the standardized MLR coefficients, squared and added together, never equal the value of Corrected R^2 nor R^2, except when intercorrelation is zero. Likewise, the t values corresponding with the MLR coefficients, squared, added together, and divided by the number of predictors (i.e., 2), equal Corrected F while the t values corresponding with the OVR coefficients, squared, added together, and divided by the number of predictors (i.e., 2), equal F. To verify these results, Corrected R^2 and Corrected F can be derived from the Type III sums of squares for the MLR model while R^2 and F can be derived from the Type I sums of squares for the OVR model, as outlined in Baird and Bieber (2016).

The process of calculating the Corrected standardized MLR coefficients will now be demonstrated for the three-predictor situation using the real-world data. As can be seen in Table 6, three OVR models are fitted, where Mileage, Liter, and Cylinder are placed last in each model, respectively OVR A, OVR B, and OVR C. Second, the standard deviation is calculated for each of the these resulting (residual) predictors. Third, these (Corrected) standard deviations are used to calculate the (Corrected) standardized MLR coefficients representing Mileage, Liter, and Cylinder (last column). Note that the Corrected standardized MLR coefficients, squared and combined, add to the Corrected R^2.

For reference, the standardized OVR coefficients, squared and combined, add to R^2 and the t values, corresponding with the MLR and OVR coefficients, squared, added together, and divided by three, add Corrected F and F, respectively. It is important to point out that the Corrected standardized coefficients of the MLR are now identical in value to each last standardized coefficient of each OVR model, whereas the corresponding unstandardized coefficients of the MLR model were always identical in value to the last unstandardized coefficients in each OVR model.

Discussion

This study was designed to empirically demonstrate the effects of intercorrelation between the MLR and OVR models, Type I and III sum of squares, and values of model statistical significance and model fit, as previously outlined by Baird and Bieber (2016). This demonstration was achieved using two separate sources of empirical evidence: simulated and real-world data. The simulation study was engineered to illustrate the differences between the two models and their corresponding model fits and test statistics in a controlled fashion: the last row, last column, and diagonal elements where identical for each increase in value of the off-diagonal elements in the first row and column of the population covariances. Therefore, predictors X_1 and X_2 serve as a reference for each other as intercorrelation increases. The predictors also serve as a reference for each other and for themselves between models, given that the same population covariances were used to evaluate both the MLR and OVR models.

Although the simulation study reflects an artificial situation in application, the pure effects of intercorrelation between the MLR and OVR models and their corresponding fit and test statistics from this design may be demonstrated clearly. On the other hand, the real-world dataset was chosen to illustrate the differences between the two models and their corresponding fit and test statistics in an applied fashion: the source of the data were collected to address certain questions of an applied nature occurring in the auto industry, where more than two predictors are of interest and multicollinearity is considered a problem, outside of the context of the current study.

A second aim of this study was to use the simulation results to identify the source of the discrepancy that occurs between the standardized coefficients of the MLR model and model fit when intercorrelation is present. In revealing the source of this discrepancy, a solution was also provided using the simulation dataset and confirmed using the real-world dataset. For clarity, study aims 1 and 2 will now be considered together, in concert, as they constitute an internally consistent framework.

The results from the simulation study confirm the framework provided in Baird and Bieber (2016). Although the MLR and OVR model coefficients, their corresponding t values, and values of model fit and statistical significance are identical when no intercorrelation is present, sharp differences exist when intercorrelation is present. As intercorrelation approaches a value of .99, the value of both unstandardized MLR coefficients identically reduce to almost half of their value relative to when no intercorrelation was present. Thus, the individual contribution of each predictor becomes indistinguishable from the other—reflecting one single predictor, not two separate predictors—both canceling each other out as neither can take "credit." As the MLR coefficient values become indistinguishable from each other, their unique, non-redundant contributions diminish to zero, as evidenced by the Type III sums of squares of the predictors, Corrected standardized MLR coefficients, and *t* values.

Conversely, as intercorrelation approaches a value of .99, the value of the first unstandardized OVR coefficient retains its original value, while only the value of the second OVR coefficient reduces to half of its original value in a fashion identical with both unstandardized MLR coefficients. Here, the contribution of the second predictor diminishes proportionally to the increase in redundancy while the contribution of the first predictor, which is unchanged, assumes all "credit". This is evidenced by the first standardized OVR coefficient retaining its original value, while the second OVR coefficient approaches a value of zero. Likewise, the *t* value corresponding with the first OVR predictor retains its original value (approximately), while the *t* value corresponding with the second OVR predictor approaches a value of zero. These results are mirrored by the sums of squares, where the sums of squares of the first predictor retains its original value meanwhile the second approaches a value of zero in the same fashion as the Type III sums of squares for both MLR predictors.

The findings from the simulation study also reveal differences between traditional and Corrected *F* and R^2 and how these test statistics and model fits are consistent with the results from the OVR and MLR models, respectively. As anticipated mathematically in Baird and Bieber (2016) and demonstrated empirically here, as the value of intercorrelation approaches .99, Corrected R^2 and Corrected *F* approach a value of zero in a manner proportional to and coterminous with the values of the Type III sums of squares for each MLR predictor, from which both were derived. Evidence that Corrected R^2 and Corrected *F* reflect the MLR model was confirmed using the Corrected standardized coefficients and their corresponding *t* values.

As anticipated in Baird and Bieber (2016) and demonstrated empirically here, as the value of intercorrelation increases to .99, R^2 and *F* approach a value of half of their original value when intercorrelation was zero in a manner proportional to and coterminous with the values of the combined Type I sums of squares for each OVR predictor, from which both were derived. Evidence that traditional R^2 and *F* reflect the OVR model was further demonstrated using the standardized coefficients and their corresponding *t* values.

The simulation study also demonstrates the differences between the MLR and OVR models regarding sample size and intercorrelation. As anticipated, when sample size increased, the means of the coefficients (standardized and unstandardized) approached their corresponding population parameters and the intervals estimates reduced in size, relative to the corresponding means and interval estimates with smaller sample sizes. As expected, the values of *t* and *F* also increased in value relative to corresponding values with smaller sample sizes. Also anticipated, the values of both R^2 and Corrected R^2 where higher than their corresponding population parameters, but reduced as sample size increased, thereby revealing the inflation that takes place when not adjusting for the number of parameters in the model, especially with smaller sample sizes (Miles, 2014). The adjustment factor was not used here so the relationship between R^2 and Corrected R with the standardized coefficients and sums of squares could be clearly elucidated. It should be noted that this adjustment, originally devised for R^2, would also adjust for inflation of Corrected R^2 equally well:

$$R^2_{\text{adj}} = 1 - \left[\frac{\left(1 - R^2\right)\left(N - 1\right)}{\left(N - P - 1\right)} \right], \tag{47}$$

$$\text{Corrected } R^2_{\text{adj}} = 1 - \left[\frac{\left(1 - R^2_{\text{Corrected}}\right)\left(N - 1\right)}{\left(N - P - 1\right)} \right]. \tag{48}$$

The aforementioned changes due to increasing sample size did not appear to influence the differences between the MLR and OVR models (and their respective test statistics and fits) due to increasing intercorrelation—that is, the difference as intercorrelation increased did not seem to be modified as sample size increased, with one exception: near-perfect intercorrelation. Specifically, when intercorrelation was at the .99 level, the mean of the standardized and unstandardized coefficients of the MLR model deviated greatly from their respective population parameter values and their confidence intervals inflated in size, as anticipated (see Cohen et al., 2003). However, this behavior also held true only for the second unstandardized coefficient of the OVR model. The standardized OVR coefficients and the Corrected standardized MLR coefficients did not deviated from their population parameter, nor did their confidence intervals inflate in size.

Although the simulation study results confirm the framework provided by Baird and Bieber (2016), this framework was also applied to real-world data, using more than two predictors. The real-world data results confirmed the findings from the simulation results, along with demonstrating the framework in applied settings. As noted by Kuiper (2008b) using this dataset, evidence of multicollinearity was found when comparing the three models, where Liter and Cylinder both significantly predict Price for MLR Model 1 and MLR Model 2, respectively, but when placed into the same model together (MLR Model 3), Liter no longer remained a significant predictor. However, OVR provided an alternative modeling approach, where, in the case of these data, designating Liter before Cylinder allowed all three predictors to be statistically significant in the model. Also demonstrated was how a specific order of predictor importance could be modeled and interpreted using OVR, relative to MLR.

The real-world data were used to confirm, in an applied setting, how R^2 and F values are larger than Corrected R^2 and Corrected F values. This inflation is also reflected between the t values of the OVR and the t values MLR models, as well as the standardized coefficients of the OVR model and the Corrected standardized coefficients of the MLR model. For greater context, it should be noted that this inflation, identified in the simulation results as "Area 4", is the original discrepancy observed by Woolf (1951). More importantly, the results from the real-world study demonstrate the unity of how R^2 and F can be derived from the standardized OVR coefficients, t values, and Type I sums of squares, and how Corrected R^2 and Corrected F can be derived from the Corrected standardized MLR coefficients, t values, and Type III sums of squares, from published, non-engineered data. This unity is especially relevant considering that the F and Corrected F, R^2 and Corrected R^2 can all be derived from the t values alone, even though the t values and the unstandardized coefficients with which they correspond have never been modified from their original value.

There are limitations of the present study. The results address correlation that is positive and linear only. This design was used because the literature referencing multicollinearity are usually in the context of correlation that is linear and most often positive (cf. Mela & Kopalle, 2002). Thus, subsequent empirical studies looking at different types of correlation, such as suppressor effects (see Cohen et al., 2003), also need be explored. The results reflect the effects of increasing intercorrelation for MLR and OVR using a specified and constant value of correlation between the predictors and the dependent variable (i.e., Area 4). Because Area 4 is not a function of the correlation between predictors, but rather the simultaneous correlation between predictors and each predictor's respective correlation with the dependent variable, caution must be used to not reduce the findings observed here to correlation between predictors only.

This study empirically confirms and advances the framework proposed by Baird and Bieber (2016), as an extension of Woolf (1951), and demonstrates the internal consistency of this framework, showing how the coefficient values, t values, and sums of squares values can all be used to derive identical values of fit and statistical significance for their respective models. An essential next step of evaluating this framework is to consider when and why the OVR model would be used in place of the MLR model, or vice versa, in applied settings, along with demonstrating how this framework compares to competing approaches of dealing with multicollinearity, such as ridge regression, principal component regression, hierarchical regression, and stepwise regression.

References

Baird, G. L., & Bieber, S. L. (2016). The goldilocks dilemma: Impacts of multicollinearity -- A comparison of simple linear regression, multiple regression, and ordered variable regression models. *Journal of Modern Applied Statistical Methods, 15*(1), 332-357. doi: 10.22237/jmasm/1462076220

Borchers, H. (2015). Pracma: Practical numerical math functions [R package, version 1.3]. Retrieved from https://cran.r-project.org/package=pracma

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge.

Kuiper, S. (2008a). 2005 car data [Data set]. Retrieved from https://amstat.tandfonline.com/doi/suppl/10.1080/10691898.2008.11889579

Kuiper, S. (2008b). Introduction to multiple regression: How much is your car worth? *Journal of Statistics Education, 16*(3). doi: 10.1080/10691898.2008.11889579

Matsumoto, M., & Nishimura, T. (1998). Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation, 8*(1), 3-30. doi: 10.1145/272991.272995

Mela, C. F., & Kopalle, P. K. (2002). The impact of collinearity on regression analysis: The asymmetric effect of negative and positive correlations. *Applied Economics, 34*(6), 667-677. doi: 10.1080/00036840110058482

Miles, J. (2014). *R* squared, Adjusted *R* squared. In N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri, & J. L. Teugels (Eds.), *Wiley StatsRef: Statistics reference online, 2014.* doi: 10.1002/9781118445112.stat06627

Mundform, D. J., Schaffer, J., Kim, M.-J., Shaw, D., Thongteeraparp, A., & Supawan, P. (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. doi: 10.22237/jmasm/1304222580

Sawilowsky, S. S. (2003). You think you've got trivials? *Journal of Modern Applied Statistical Methods, 2*(1), 218-225. doi: 10.22237/jmasm/1051748460

Woolf, B. (1951). Computation and interpretation of multiple regressions. *Journal of the Royal Statistical Society: Series B (Methodological), 13*(1), 100-119. doi: 10.1111/j.2517-6161.1951.tb00074.x

Appendix A

I. Code for Populations

```
/*creating population for [.00]*/
proc iml;
N = 1000000;     /* population size   */

/* specify population mean and covariance */
Mean = {3, 3, 21343};
Cov =  {1.22 .0000 6099,
         .0000 1.22 6099,
          6099 6099 97710314};

call randseed(121982);  /*seed remains constant for populations */
X = Randnormal(N, Mean, Cov);
create population_0 from X[c={"X1" "X2" "Y"}];
append from X;
close population_0;
quit;
```

II. Code for Samples

```
proc surveyselect data=Population_0 out=sample0a seed=14159 method=urs
sampsize=20 rep=8000 OUTHITS;
run;
```

III. Code for MLR

```
PROC REG Data= sample0a outest=mlrs0a tableout alpha=0.05 noprint
RIDGE=0;
MODEL Y=X1 X2 /rsquare MSE OUTSTB OUTVIF;
by Replicate;
RUN;
```

IV. Code for OVR

```
*Step 1, make new predictor 2;
PROC REG Data= sample0a noprint;
MODEL X2= X1;
```

```
OUTPUT OUT= res0a  residual=yresid; by Replicate; RUN; quit;

*Step 2, OVR Model;
PROC REG Data=res0a  outest=ovrs0a tableout alpha=0.05 noprint RIDGE=0;
MODEL Y=X1 yresid /rsquare MSE OUTSTB OUTVIF;
by Replicate;
RUN;
```