Received: 12 September 2016 | Revised: 24 October 2017 | Accepted: 19 January 2018 | Published on: 7 May 2018

DOI: 10.1002/smj.2783

WILEY STRATEGIC MANAGEMENT JOURNAL

RESEARCH ARTICLE

Multicollinearity: How common factors cause Type 1 errors in multivariate regression

Arturs Kalnins

Department of Management and Organizations, Tippie College of Business, University of Iowa, Iowa City, Iowa

Correspondence
Arturs Kalnins, Tippie College of Business, S388 Pappajohn Business Building, University of Iowa, Iowa City, IA 52242.
Email: arturs-kalnins@uiowa.edu

Research Summary: In multivariate regression analyses of correlated variables, we sometimes observe pairs of estimated beta coefficients large in absolute magnitude and opposite in sign. *T*-statistics are also large, suggesting meaningful findings. I found 64 recently published *Strategic Management Journal* articles with results exhibiting these characteristics. In this article, I demonstrate that such results may be Type 1 errors (false positives): If regressors are correlated via an unobservable common factor, estimated beta coefficients will misleadingly tend toward infinite magnitudes in opposite directions, even if the variables' real effects are small and of the same sign. Diagnostics such as Variance Inflation Factors (VIF) will misleadingly validate Type 1 errors as legitimate results. After establishing general results via mathematical analysis and simulation, I provide guidelines for detection and mitigation.

Managerial Summary: This article demonstrates mathematically how regression analyses with correlated independent variables may generate beta coefficients of opposite sign to the variables' true effects. To assess the likelihood of this possibility, I propose that: if (a) absolute correlation of two independent variables is about ±0.3 or more (smaller correlations may be problematic for large data sets), (b) the two variables have beta coefficients of opposite sign, if correlated positively, and of the same sign, if correlated negatively, and (c) the bivariate correlation of one independent variable with the dependent variable is of the opposite sign from the beta coefficient, then the beta might be a false positive. To facilitate such analysis, authors should provide complete correlation tables, including dependent variables, interaction terms, and quadratic terms.

KEYWORDS
analytic model, econometrics, multicollinearity, multivariate regression, research methods

1 | INTRODUCTION

When using multivariate ordinary least squares (OLS) regression, or its maximum-likelihood counterparts such as logit, probit, or hazard rate models, to analyze business performance and survival duration based on archival data sources, I have often observed unrealistically large estimated coefficients and large *t*-statistics when independent variables are correlated. The unrealistic coefficients come in pairs—not only are they large in absolute value, but also they are of opposite sign and one of the signs may not align with prior expectations. And, the more highly correlated the pair of independent variables, the larger the absolute values of their coefficients. While such estimated coefficients could plausibly be surprising results worthy of publication, my concern has been that they may be nothing more than Type 1 errors, that is, false positives where no legitimate result exists. Whenever I have come across such results in my own work, I have assumed they are problematic and have not pursued publication.

Yet I regularly read published research where results fit this description, and where these results are considered legitimate by their authors. I provide two examples, not because I believe they are poor papers—in fact, in many ways they are both quite carefully done—but because their empirical data structures illustrate forms of multicollinearity that are canonical to strategic management research. Indeed, I do not believe that the authors of these articles, or those of the many Strategic Management Journal studies reviewed later in this article, have engaged in any inappropriate research practices: they have by and large followed commonly accepted principles and practices in their research design and interpretation. However, I do believe that the commonly accepted principles and practices can be improved upon to better detect and mitigate the likelihood of Type 1 errors due to multicollinearity. Articulating necessary improvements is one major aim of this article.

First, Yli-Renko and Janakiraman (2008) analyzed number of new products developed by firms: R&D spending and firm size are included as control variables and are correlated 0.61. In the regressions presented, R&D spending has a positive coefficient and firm size has a negative coefficient. However, the bivariate correlation of R&D spending with new products developed is negative and the bivariate correlation of firm size with new products developed is positive. Multicollinearity is a possible culprit of creating Type 1 errors for both of these variables, that is, flipping the signs of both bivariate correlations and inflating the magnitude of the effects such that seemingly meaningful regression coefficients appear to emerge from the analysis. While these variables are only controls in this study, this example demonstrates the possibly deleterious effects of multicollinearity on archetypal strategic management research variables that are typically correlated in practice.

Second, Pfarrer, Pollock and Rindova (2010) reported that firms' high reputations are associated with lower likelihoods of positive earnings surprises. But the same regression contains the control variable "prior high reputation," correlated 0.72 with [current] high reputation, and that control variable is associated with a greater likelihood of positive earnings surprises. It seems puzzling to me that a variable measured at one point in time would have a meaningful effect opposite to that of the same variable measured at previous points in time—indeed, the bivariate correlations of both

reputation variables with positive earnings surprises are negative and of similar magnitude. It is possible that multicollinearity created a Type 1 error by flipping the sign of one variable's bivariate correlation with the dependent variable, and inflating the absolute magnitude of both variables' effects. The authors address the multicollinearity issue by stating "All variance inflation factors were below five, with an average of 2.4. Thus, multicollinearity is not a concern" (Pfarrer, Pollock and Rindova, 2010: 1140).

The possibility that results involving correlated variables may, in fact, be Type 1 errors is problematic for the academic field of strategic management. Our field is empirically driven, and we are careful to consider previous results when motivating new research questions, extending the work of others, crafting hypotheses, and assessing the likely validity of empirical results. On the one hand, researchers may spend too much effort generating new theory and extending the theory of others in directions that assume the Type 1 errors provide a meaningful base of knowledge. On the other hand, any new results opposing the Type 1 errors may be unfairly discounted. In addition, the strategy field makes extensive use of control variables to isolate the effects of the variables that are of interest. Control variables can only have a beneficial effect in this regard if they are correlated with the hypothesized variables. Yet, due to multicollinearity, the addition of control variables might generate Type 1 errors among variables of interest rather than isolate their true effects. For these reasons, an analysis of the possible distortion of regression coefficients due to multicollinearity and the possible Type 1 errors that result is a worthwhile methodological effort to improve validity of future findings relevant to strategic management.

When I have consulted methods texts and journal articles, I have found allusions to multicollinearity generating results that are "nonsense" (Frisch, 1934, p. 6) and multicollinearity causing wrong signs for regression coefficients (Gujarati, 2003; Kennedy, 2003), though the cause of these wrong signs is not made clear. Simulations have provided examples where beta coefficients become large and opposite in sign (e.g., Day et al., 2004; Yoo et al., 2014). But I also found highly cited references that dismissed concerns about multicollinearity. The typical argument is that, if the standard OLS assumptions hold—in particular, that dependent variable y is a perfectly specified linear function of only observable independent variables plus an exogenous error term—multicollinearity does not bias coefficient estimates, but merely inflates their standard errors (e.g., Goldberger, 1989, 1991). This point has been repeated in popular statistical methods textbooks (e.g., Gujarati, 2003; Judge, Hill, Griffiths, Lutkepohl, & Lee, 1982; Kennedy, 2003), and is proved in Appendix S1. If I accept this textbook perspective as valid for the commonplace multivariate regressions that strategy researches estimate from archival data, a conclusion is that multicollinearity may lead to Type 2 errors (false negatives), but should not ever lead to Type 1 errors (false positives). Due to this perspective, "do nothing" has become a popular refrain among scholars who have found unusual results among highly correlated variables (see, e.g., Gujarati, 2003; Kennedy, 2003).

Empirically, many authors have used VIF statistics to dismiss multicollinearity concerns. More than 58,000 manuscripts on Google Scholar reference the VIF and multicollinearity, including more than 12,000 in 2017. VIFs are the diagonal elements of the inverse of the moments matrix $M = X'X$ of independent variables. Rules of thumb exist that if the maximum VIF is less than 10, 8, or 5, then multicollinearity problems are not likely to exist. Alternatively, Belsley, Kuh, and Welsch (1980) suggested the Condition Index, the square root of the ratio of the largest eigenvalue to the smallest eigenvalue of the matrix M. Over 1,750 manuscripts on Google Scholar reference multicollinearity and the Condition Index. Belsley et al. (1980) referred to a Condition Index of 30 or above as "high," but also cautioned that condition numbers as low as 8 or 10 may signal multicollinearity problems. Yet, other texts cite simple "rules of thumb" that multicollinearity problems are not likely

to exist if bivariate correlations are below cut-off values, typically 0.7 or 0.8 (e.g., Judge et al., 1982).

Because of my inability to reconcile the large and oddly signed coefficients of collinear variables in published studies and in my own data with the perspectives that (a) multicollinearity only inflates standard errors, and (b) a low VIF statistic or Condition Index allows a dismissal of multicollinearity-related concerns, I initiated this research project to understand the mathematical relationship between multicollinearity and Type 1 errors in OLS regression. The project has three main objectives. First, to establish whether multicollinearity is a prevalent phenomenon in strategic management research, using as primary evidence an analysis of all articles published in the *Strategic Management Journal* between January 2013 and December 2017. Second, to assess analytically and via simulation the possibility that, when variables are collinear, their estimated beta coefficients become biased and misleading. Third, to provide guidelines for the detection of, and approaches for the mitigation of, concerns about multicollinearity.

In the analysis section, I show that, if the OLS assumption of fully observable independent variables is violated, then multicollinearity may lead to biased results and Type 1 errors. Specifically, if regressors are correlated via an unobservable common factor, that is, a latent variable common to multiple observable variables, their estimated beta coefficients will often tend toward infinite magnitudes in opposite directions as their correlation increases. This result remains true even if their real effects are small and of the same sign. Standard errors will indeed grow in magnitude, but often not enough to eliminate large, false *t*-statistics that are the basis of Type 1 errors. Large data sets only exacerbate this problem because they will generate large *t*-statistics. Finally, I demonstrate that the VIF statistic, the Condition Index, and bivariate rules of thumb are not valid detection approaches if common-factor multicollinearity is present in a data set. These diagnostics will often misleadingly validate Type 1 errors.

2 | PREVALENCE OF MULTICOLLINEARITY, COMMON FACTORS AND POSSIBLE TYPE 1 ERRORS

While researchers would prefer to observe in isolation the variables that are causing an effect of theoretical interest, social scientists have long been aware that this is often not possible. In 1920, Karl Pearson succinctly identified the issue that underlies the problems of multicollinearity in archival empirical research. He stated, "For us the unobservable variables may be supposed to be uncorrelated causes, and to be connected by unknown functional relations with the [observable] correlated variables" (Pearson, 1920, p. 27). In the case analyzed here, the uncorrelated causes are (a) the unobservable common factor, and (b) the substantive but idiosyncratic factors associated with some variables. The common factor might be substantive (e.g., firm resources, reputation) as envisioned by Pearson, or it might be a simple measurement error that is common to multiple variables. The unknown functional relations involve the relative amounts of the unobservable common factor, idiosyncratic factors, and noise that appear in the observable variables.

Applying this discussion to strategic management, the archival data sources used in strategic management research often do not consist of those variables we wish to test. Godfrey and Hill (1995) argued that strategy constructs such as opportunism, utility, and resource stocks are necessarily unobservable variables that can be empirically analyzed only through proxy variables, which are related via Pearson's unknown functional relations. Even if the proxies are measured perfectly, the

presence of unobservable constructs, such as resource stocks, violates the key OLS assumption for unbiasedness. The unknown functional relations between the proxies and the unobservable will create the correlations between proxies necessary for Type 1 errors.

In addition, the use of proxy variables exacerbates the potential for measurement error. Boyd, Gove, and Hitt (2005) discussed the effect of proxy-related measurement error on statistical power—it decreases power and may create Type 2 errors. I demonstrate that, when multiple variables in an OLS regression are subject to the same source of measurement error, Type 1 errors may result as well. In sum, I believe that data structures with correlations due to unobservable common factors are more common in strategic management research than are cases where regression equations can be perfectly specified. These common factors represent a form of mis-specification that is likely pervasive in any empirically based field of social science and can lead to Type 1 errors if not properly addressed. It may well be as problematic as endogeneity but has received far less attention to date.

3 | ARTICLES PUBLISHED IN THE *STRATEGIC MANAGEMENT JOURNAL* FROM 2013 TO THE PRESENT

I examined all 618 articles published by the *Strategic Management Journal* between January 2013 and December 2017. I evaluated whether common-factor multicollinearity could be driving some of the reported results.[1] My evaluation criteria were: First, two hypothesized variables, or one hypothesized variable and a control, were correlated about ±0.3 or more. I show in the following that correlations of ±0.3 are sufficiently large to generate Type 1 errors. Second, the two variables had to have beta coefficients of the opposite sign, if correlated positively, and of the same sign, if correlated negatively. Third, the authors stated that at least one of the two beta coefficients played a role in supporting a hypothesis. Fourth, each hypothesis-supporting beta coefficient had to differ in one of two ways from its variable's underlying bivariate correlation with the dependent variable: The bivariate correlation could be (a) the opposite sign from the beta, or (b) designated as not statistically different from zero.

I was not able to assess multicollinearity between interaction terms, or between an interaction term and control variables because of the curious tradition in strategy and management research of not providing descriptive statistics or correlations for interaction terms. The same problem holds for quadratic terms. Interaction and quadratic variables should be susceptible to the same problems that I describe later. Given the prevalence of testing interaction and quadratic terms in strategic management research, the real extent of the multicollinearity problem might be greater than I am able to estimate.

Based on these four criteria, I identified 64 articles (10.3% of all 618) across 65 *SMJ* issues that contained at least one result reported as a meaningful, theoretically-based finding, but that may be a Type 1 error. In the regression results in 41 of the 64 articles, at least one hypothesized coefficient affected by multicollinearity flips sign from that of its bivariate correlation with the dependent variable. In the remaining 23, a bivariate correlation designated as not statistically different from zero turns into a hypothesis-supporting beta coefficient.

I am not arguing that these results are necessarily Type 1 errors because they meet the four criteria; the results may accurately reflect the underlying relationships they are testing. I am pointing out that the structural features of these regressions are consistent with an enhanced possibility of Type 1 errors. And, as I stated earlier, I do not believe the authors of these articles engaged in any sort of poor research practice. I lay out tests below that authors of articles such

[1]I thank an anonymous reviewer for suggesting this analysis.

as these could employ in the future to mitigate the possibility that their results are interpreted as Type 1 errors.

Appendix S2 lists the 64 articles. The first two columns list the authors, volume, and page number. The third column of the table lists the two independent variables that are correlated. At least one is designated as a hypothesized variable. The fourth column displays the correlation between the two independent variables; this is the cause of the multicollinearity problem. The fifth column contains the bivariate correlation between the dependent variable and each of the two independent variables. The sixth column contains the beta coefficients from a regression for the two independent variables. In every case, the variable(s) associated with a hypothesis are deemed to be statistically significant by the authors. The seventh column highlights whether a variable's beta sign has flipped relative to the bivariate correlation or simply gained statistical significance. If a dependent variable was already statistically significantly related to an independent variable in a bivariate correlation, then even if multicollinearity is present, it is not likely the cause of the relationship.

Plausible common factors are evident among some of the pairs of variables in Appendix S2. Some articles contain positively correlated variables that represent multiple "types" of the same common factor. One article contains two types of conflict (cognitive and affective) in a single regression, while other articles contain multiple types of exploration/exploitation, knowledge, governance, technological patents, FDI host countries, and complexity. Other studies in Appendix S2 have correlated independent variables where the common factor may be firm performance or size. These include combinations of profits and revenues, ROE and ROA, target productivity and sales, financial resources and marketing budget, and performance and pay dispersion. In the mitigation section that follows, I examine in detail the implications of profits and revenues as independent variables in one study listed in the appendix.

Regarding diagnostics, in 31 of the 64 articles with possible Type 1 errors, authors state that they can dismiss multicollinearity concerns via low values of VIF statistics. I demonstrate in the following that the VIF often will not detect problems related to common-factor multicollinearity. Eight articles claim to address multicollinearity concerns by mean-centering variables, which Echambadi and Hess (2007) demonstrated to be a harmless but ineffective solution.

4 | ANALYTIC AND SIMULATION MODELS: EFFECTS OF MULTICOLLINEARITY

4.1 | Analytic model of the effects of multicollinearity due to an unobservable common factor

I analyze the effects of multicollinearity that arise from the presence of an unobservable factor that is common to two observable variables. This is an important and realistic alternative to the case of a perfectly specified regression because, as I previously argued, non-experimental variables may often be the sum of a common factor and an idiosyncratic term, the former of which may represent a common source of measurement error or a meaningful but unobservable causal variable. The common factor structure is consistent with that described by Pearson (1920) as most representative of an empiricist's reality. Common factors were also proposed by Frisch (1934) as the basis for the canonical form of multicollinearity. Further, the unobservable common factor is a necessary part of the form of multicollinearity implied by textbooks that use a Venn diagram chart (e.g., Gujarati, 2003; Kennedy, 2003) to illustrate the concept.

As per Sastry (1970), the formula for each regression coefficient based on covariances is:

$$\beta_i = \frac{|M_i|}{|M|}, \tag{1}$$

where M is the moments matrix $X'X$ (the covariance matrix $X'X$/n may also be used) of all the independent variables in the regression. $|M|$ is the determinant of M. The matrix M_i is matrix M with Column i switched out in favor of a column vector that consists of all moments or covariances of dependent variable y with the independent variables. Row i from M is left intact in matrix M_i.

If there are only two scalar variables in X, or only two that are correlated with each other (I designate them as x_1 and x_2), Equation 1 for estimated coefficient β_1 simplifies to:

$$\beta_1 = \frac{var(x_2)\,cov(x_1 y) - cov(x_1 x_2)\,cov(x_2 y)}{var(x_1)\,var(x_2) - (cov(x_1 x_2))^2}.$$

If the x_1 and x_2 variables are bivariate standard normal, then on average, $var(x_1) = var(x_2) = 1$, while $cov(x_1 x_2) = corr(x_1 x_2)$. For this average case, I can rewrite the equation as:

$$\beta_1 = \frac{cov(x_1 y) - corr(x_1 x_2)\,cov(x_2 y)}{1 - (corr(x_1 x_2))^2} \tag{2}$$

The same equalities hold for β_2 by substituting x_2 for x_1 and vice versa.

Now, consider a specification underlying the bivariate distribution of x_1 and x_2. The observable regressor variables x_1 and x_2 are correlated via an unobservable common factor x_n. The variables x_1 and x_2 also have idiosyncratic components, x_{i1} and x_{i2}, respectively, that are independent of each other and of common factor x_n. Pearson's (1920) functional relations are represented by a and c.

$$x_1 = a x_n + c x_{i1};$$

$$x_2 = a x_n + c x_{i2}$$

Without loss of generality, I can scale x_{i1}, x_{i2}, and x_n such that they are standard normal. I then monotonically transform x_1 and x_2 in the following fashion.

$$x_1 = \sqrt{\frac{a}{a+c}} x_n + \sqrt{\frac{c}{a+c}} x_{i1};$$

$$x_2 = \sqrt{\frac{a}{a+c}} x_n + \sqrt{\frac{c}{a+c}} x_{i2}$$

After this transformation, x_1 and x_2 will be standard normally distributed for any values of a and c, and x_1 and x_2 will have a bivariate correlation equal to $\frac{a}{a+c}$, which I replace with θ. If the variables are negatively correlated, the signs of one variable's values can be flipped so that the correlation is positive and so that the square root is a real value. The sign of that variable's coefficient, as derived below, will then need to be flipped but all other conclusions remain the same. Thus:

$$x_1 = \sqrt{\theta} x_n + \sqrt{1-\theta} x_{i1};$$

$$x_2 = \sqrt{\theta} x_n + \sqrt{1-\theta} x_{i2}$$

I define the dependent variable y as a linear function of the unobservable variables, that is, the common factor and the idiosyncratic terms, and a standard normal error term e. Note that in the definition below there is no correlation between the error term and the unobservables. Therefore, this mis-specification is not an endogeneity problem:

$$y = \gamma x_n + \delta_1 x_{i1} + \delta_2 x_{i2} + e.$$

Based on this structure of x_1, x_2, and y, it can be shown that:

$$cov(x_1 y) = \gamma\sqrt{\theta} + \delta_1\sqrt{1-\theta};$$

$$cov(x_2 y) = \gamma\sqrt{\theta} + \delta_2\sqrt{1-\theta}.$$

Inserting these values into Equation 2 yields:

$$\beta_1 = \frac{\gamma\sqrt{\theta} + \delta_1\sqrt{1-\theta} - \theta\left(\gamma\sqrt{\theta} + \delta_2\sqrt{1-\theta}\right)}{1-\theta^2};$$

$$\beta_1 = \frac{\gamma\sqrt{\theta}(1-\theta)}{1-\theta^2} + \frac{(\delta_1 - \theta\delta_2)(\sqrt{1-\theta})}{1-\theta^2};$$

$$\beta_1 = \frac{\gamma\sqrt{\theta}}{1+\theta} + \frac{(\delta_1 - \theta\delta_2)(\sqrt{1-\theta})}{(1+\theta)(1-\theta)};$$

$$\beta_1 = \frac{\gamma\sqrt{\theta}}{1+\theta} + \frac{(\delta_1 - \theta\delta_2)}{(1+\theta)\sqrt{1-\theta}}. \quad (3)$$

Replace θ with 1 − ε:

$$\beta_1 = \frac{\gamma\sqrt{1-\varepsilon}}{2-\varepsilon} + \frac{(\delta_1 - \delta_2 + \delta_2\varepsilon)}{(2-\varepsilon)\sqrt{\varepsilon}}.$$

As ε approaches 0 and gets arbitrarily close, this can be rewritten as:

$$\lim_{\varepsilon\to 0}\beta_1 = \frac{\gamma}{2} + \frac{(\delta_1 - \delta_2)}{2\sqrt{0}} = \pm\infty.$$

I note that I only need to examine a one-sided limit because it is impossible to approach a correlation of 1 from above. Thus, I conclude that:

$$\lim_{\varepsilon\to 0}\beta_1 = +\infty \; if \; \delta_1 - \delta_2 > 0;$$

$$= -\infty \; if \; \delta_1 - \delta_2 < 0;$$

$$= \frac{\gamma}{2} \; if \; \delta_1 - \delta_2 = 0.$$

The results for β_2 can be obtained by exchanging the two inequality signs.

4.2 | The standard error

I now consider the standard error, whose general formula can be written:

$$SE(\beta_i) = \sqrt{M_{ii}^{-1}\frac{RSS(\beta)/n}{n-k-1}}, \quad (4)$$

where RSS is the residual sum of squares, n is the number of observations, and k is the number of variables. M_{ii}^{-1} is the iith element (the ith diagonal element) of the inverse of the covariance matrix M. M_{ii}^{-1} also represents the VIF statistic that we discuss in the following section. The general formula for the RSS is:

$$RSS(\beta)=n\left(var(y)-\sum_{i=1}^{N}\beta_i^2\,var(x_i)-\sum_{i\neq j}2\beta_i\beta_j\,cov\left(x_ix_j\right)\right). \tag{5}$$

Given the equation $y = \gamma x_n + \delta_1 x_{i1} + \delta_2 x_{i2} + e$, and given the assumption that the unobservable components of y, x_{i1}, x_{i2}, and x_n are uncorrelated and distributed standard normally, $var(y)=\gamma^2+\delta_1^2+\delta_2^2+1$. Then, by substituting in values for β_1 and β_2 in terms of δ_1, δ_2, and θ from Equation (3), and after manipulation, RSS can be rewritten for the two-variable case as:

$$RSS(\beta)=n\frac{(1-\theta)\gamma^2-2\sqrt{\theta}\sqrt{1-\theta}\gamma(\delta_1+\delta_2)+\theta\delta_1^2+\theta\delta_2^2+2\theta\delta_1\delta_2+1+\theta}{1+\theta} \tag{6}$$

Given the fact that e, x_1 and x_2 are assumed to be standard normally distributed, x_1 and x_2 have a correlation of θ, and $k = 2$ because there are two variables; I can rewrite Equation (4) as:

$$SE(\beta_1)=\sqrt{\frac{(1-\theta)\gamma^2-2\sqrt{\theta}\sqrt{1-\theta}\gamma(\delta_1+\delta_2)+\theta\delta_1^2+\theta\delta_2^2+2\theta\delta_1\delta_2+1+\theta}{(n-3)\left(1-\theta^2\right)(1+\theta)}}..$$

I can replace θ with $1 - \varepsilon$. As ε approaches 0 and gets arbitrarily close, SE can be rewritten as:

$$SE(\beta_1)=\sqrt{\frac{\delta_1^2+\delta_2^2+2\delta_1\delta_2+2}{4\varepsilon(n-3)}},$$

and thus, $\lim_{\varepsilon\to 0} SE(\beta_1)=\infty$. However, this finding does not imply that the t-statistic will approach 0. The t-statistic at this point where ε gets arbitrarily close to 0 is determined by:

$$t=\lim_{\varepsilon\to 0}\frac{\beta_1}{SE(\beta_1)}=\frac{(\delta_1-\delta_2)}{2\sqrt{\varepsilon}}\frac{2\sqrt{\varepsilon(n-k-1)}}{\sqrt{\delta_1^2+\delta_2^2+2\delta_1\delta_2+2}}=\frac{\sqrt{n-3}\,(\delta_1-\delta_2)}{\sqrt{(\delta_1+\delta_2)^2+2}}.$$

To achieve a t-statistic of at least 2 (the t-statistic associated with statistical significance at $p < 0.05$ is $t = 1.96$), for example, I would need to ensure that:

$$t=\frac{\sqrt{n-3}\,(\delta_1-\delta_2)}{\sqrt{(\delta_1+\delta_2)^2+2}}>2,\ or$$

$$n>\frac{4(\delta_1+\delta_2)^2+8}{(\delta_1-\delta_2)^2}+3.$$

This does not seem like a difficult hurdle in empirical studies. Given that $\sigma_\varepsilon = 1$, for a large value of $\delta_1 = 0.5$, and $\delta_2 = 0$, n would have to equal only 39 to achieve a t-statistic of 2 in the limit for both β_1 and β_2. For even a small value of $\delta_1 = 0.1$ and $\delta_2 = 0$, relative to $\sigma_\varepsilon = 1$, n would have to equal ~800 to achieve a t-statistic of 2 in the limit that is feasible for sample sizes used in many current empirical studies. The larger the sample size, the more likely a Type 1 error.

I conclude that, under the conditions specified, the coefficient estimates β_1 and β_2 may be very different than the actual effects δ_1, δ_2, and γ of x_{i1}, x_{i2}, and x_n, respectively, on y.

> **Conclusion 1:** *If the dependent variable y is not a perfectly specified linear function of the observable independent variables, but rather the two independent variables share an unobservable common factor, and the dependent variable y is a linear function of at least one idiosyncratic component of one of the independent variables:*

a. Estimated β coefficients of the two independent variables will only equal their true idiosyncratic component values when their correlation is zero. Anytime the variables' idiosyncratic effects are unequal, the two estimated β coefficients will always approach positive and negative infinity as their correlation approaches one, regardless of the true values.
b. If both true effects are of the same sign but of different magnitudes, there will be a correlation above which one of the estimated β effects will exhibit a sign opposite to that of its true value. This is not a meaningful value.
c. As the correlation of the two independent variables approaches one, the estimated β coefficients may exhibit large but false t*-statistics, possibly causing at least one Type 1 error.*
d. Large sample sizes do not mitigate the problem of large, false t*-statistics. In fact, they make it worse by increasing the size of the* t*-statistics.*

4.3 | The variance inflation factor and the condition index for the two-variable case

I consider a numerical example where actual effects are not large relative to $\sigma_\varepsilon = 1$: $\delta_1 = 0.5$, $\delta_2 = 0.1$, and $\gamma = 0$, and correlation $\theta = 0.3$ are inserted into Equation (3), yielding:

$$\beta_1 = \frac{(0.5 - 0.3^{*}0.1)}{(1+0.3)\sqrt{1-0.3}} = 0.43;$$

$$\beta_2 = \frac{(0.1 - 0.3^{*}0.5)}{(1+0.3)\sqrt{1-0.3}} = -0.046.$$

While β_1 is not far from its real value, β_2 has the opposite sign of δ_2, due to multicollinearity. Based on Equation (4), the standard error will be:

$$SE(\beta_1) = SE(\beta_2) = \sqrt{\frac{0.3(0.5+0.1)^2 + 1 + 0.3}{(n-3)(1-0.09)(1+0.3)}}.$$

The value of β_2 is small, but with a large n of 2,200, its t-statistic would equal approximately 2 based on the previous standard error. The effect of β_2 is a Type 1 error.

I now consider the VIF, Condition Index, and bivariate rules of thumb for this two-variable example. Both VIFs, the diagonal elements M_{ii}^{-1} of the inverse of correlation matrix $M = X'X/n$, will equal $1/(1 - \theta^2) = 1.1$. Despite a Type 1 error, the VIF is very low—just 0.1 away from the minimum possible VIF value of 1. The typical rules of thumb state that the VIF should be less than 5, 8, or 10 to ensure no problems of multicollinearity. The VIF for this example hits these critical values at approximately $\theta = 0.89$, 0.93, and 0.95, respectively, suggesting inappropriately that correlations below that level are not problematic. If the correlation of the variables in the example was really 0.895, the estimate of the β_2 coefficient would be −0.57, almost six times greater than its real value of $\delta_2 = 0.1$ and of the opposite sign. The value of −0.57 would have a t-statistic > 2 with a sample size as low as 65.

The Condition Index is the square root of the ratio of the largest eigenvalue to the smallest eigenvalue of $M = X'X$. In the previous example, the larger eigenvalue is $1 + \theta$ and the smaller is $1 - \theta$. Thus, the Condition Index is $\sqrt{\frac{1+\theta}{1-\theta}}$, equal to 1.36 if $\theta = 0.3$. The Condition Index reaches

Belsley et al.'s (1980) most conservative critical value of 8 only at approximately $\theta = 0.97$, leading to the inappropriate conclusion that correlations below that high level are not problematic. I have shown with this numerical example that a Condition Index can be as low as 1.36 for a coefficient that is a substantively meaningless Type 1 error.

Finally, bivariate "rules of thumb" exist that state bivariate correlations are likely problematic only if they exceed 0.7 or 0.8 (e.g., Judge et al., 1982). The previous example, where $\theta = 0.3$ and where multicollinearity problems are present, demonstrates why these rules are inadequately conservative, at least for the case of common-factor multicollinearity.

Conclusion 2: *Type 1 errors may occur in regressions where diagnostic values such as the (a) VIF, (b) Condition Index, and (c) bivariate rules of thumb are far below even their most conservative cut-off values, falsely suggesting the absence of a multicollinearity problem.*

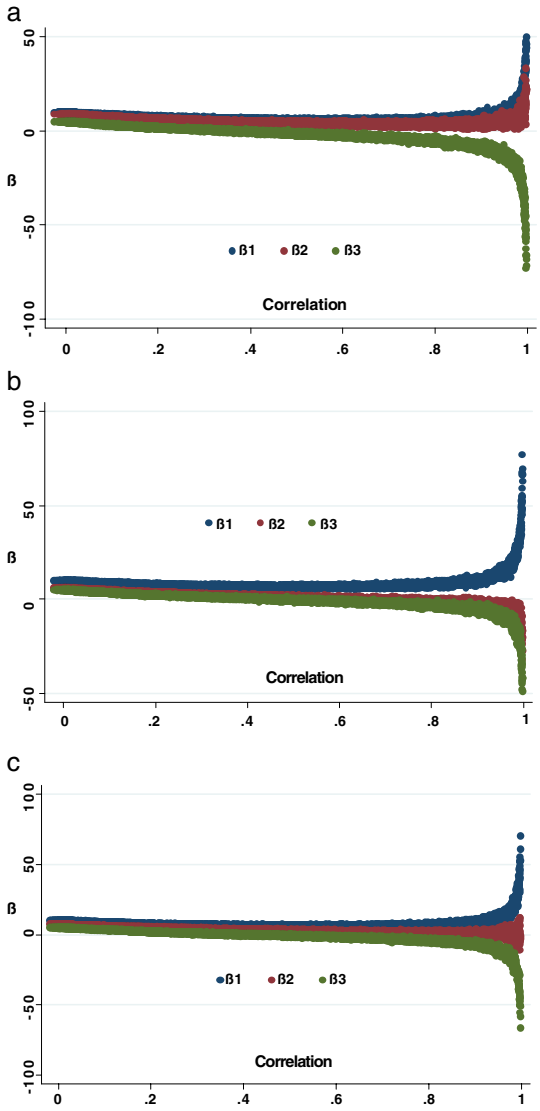

FIGURE 1 (a) Simulated three-variable case ($\delta_1 = 10$, $\delta_2 = 9$, $\delta_3 = 5$). β_2 approaches $+\infty$. (b) Simulated three-variable case ($\delta_1 = 10$, $\delta_2 = 6$, $\delta_3 = 5$). Now β_2 approaches $-\infty$. (c) Simulated three-variable case ($\delta_1 = 10$, $\delta_2 = 7.5$, $\delta_3 = 5$). Now β_2 approaches either $+$ or $-\infty$ based on the realizations of the random component ε. But the mean of β_2 values approach 0

4.4 | Simulations with three variables of theoretical interest

In this section, I present simulation results, shown in Figure 1a–c, to demonstrate how the addition of a third variable of theoretical interest will affect the regression results. I keep the same structure from the two-variable case: All three observed independent variables consist of a standard normally distributed unobservable common factor x_n and unobservable individual components x_{i1} through x_{i3}. Specifically,

$$x_k = \sqrt{\theta}x_n + \sqrt{1-\theta}x_{ik},\ where\, k = 1,2,3.$$

The observed variables x_k are thus distributed standard normal, and the bivariate correlations between all pairs of the x_k will necessarily be θ. The STATA simulation code that generates the results found in Figure 1 and in the following text is presented in full in Appendix S3.

I define the first dependent variable as $y = 10x_{i1} + 9x_{i2} + 5x_{i3} + \varepsilon$, the second as $y = 10x_{i1} + 7.5x_{i2} + 5x_{i3} + \varepsilon$, and the third as $y = 10x_{i1} + 6x_{i2} + 5x_{i3} + \varepsilon$, where ε is distributed standard normal. I chose these values to demonstrate the variation in the effects of multicollinearity on the variable in the middle in terms of effects on *y*. Just like in the two-variable case, the variable with the largest real idiosyncratic effect (x_1 with a value of $\delta_1 = 10$) will always have an estimated β_1 that will approach $+\infty$ as θ approaches 1, and the variable with the smallest (x_3 with a value of $\delta_3 = 5$) will always have an estimated β_3 that will approach $-\infty$. I have discovered, through many simulations, shown and not shown, that the estimated value β_2 in the middle will approach $+\infty$ as θ approaches 1 when $\delta_2 > (\delta_1 + \delta_3)/2$. β_2 will approach $-\infty$ when $\delta_2 < (\delta_1 + \delta_3)/2$.

In Figure 1a, when $\delta_2 = 9 > (\delta_1 + \delta_3)/2$, the estimated β_2 will approach $+\infty$. What happens to the estimated β_3 in this case is of interest. Even though the real relationship between x_3 and *y* is positive ($\delta_3 = 5$), when the correlation θ is 0.4, the estimated β_3 has flipped signs (is negative) 85% of the time (85 times out of 100 trials). When $\theta = 0.5$, the estimated β_3 is negative 98% of the time and 81% of the time its *t*-statistic < -2. This provides a clear example of the possibility of Type 1 errors and flipped coefficient signs even at moderate correlations among variables of theoretical interest. Further, the maximum VIF statistics and Condition Indices among all the regressions where $\theta = 0.5$ are 1.69 and 2.21, respectively. Based on these low values, researchers might incorrectly conclude there is no multicollinearity problem.

In Figure 1b, I observe that when $\delta_2 = 6 < (\delta_1 + \delta_3)/2$, the estimated β_2 will approach $-\infty$, the opposite of the case of Figure 1a, just because of a shift in the value of δ_2 from 9 to 6.

In Figure 1c, when $\delta_2 = 7.5 = (\delta_1 + \delta_3)/2$, the estimated β_2 can approach either + or $-\infty$, based on the values of ε realized in each regression. Regarding β_2, the median value approaches 0, not 7.5, as the correlation θ approaches 1, increasing the likelihood of a Type 2 error. Regarding β_3, much like the case shown in Figure 1a, when the correlation θ is 0.4, the estimated β_3 has flipped signs (is negative) 63% of the time. When $\theta = 0.5$, the estimated β_3 is negative 98% of the time and 57% of the time its *t*-statistic < -2. The maximum VIF statistics and condition indices among all the regressions where $\theta = 0.5$ are 1.68 and 2.20, respectively. This simulation demonstrates that common-factor multicollinearity can cause both a possible Type 2 error (x_2) as well as a Type 1 error (x_3) within the same regression.

> **Conclusion 3:** *In a regression with three independent variables that share a common factor, regardless of the size or signs of the variables' true effects:*
>
> *a. The estimated β coefficient of the variable with the largest real effect on* y *will always approach positive infinity,*
> *b. The one with the smallest real effect will approach negative infinity,*

TABLE 1 Average OLS results of 10,000 simulated regressions in each row with 1,000 observations each

	Mean effect of controls on y is 0. τ = N(0,1)						Mean effect of controls on y is 1. τ = N(1,1)					
	t-statistic > +2		t-statistic < −2		Diagnostics (max)		t-statistic > +2		t-statistic < −2		Diagnostics (max)	
Controls Count	Count	Avg. β	Count	Avg. β	VIF	CI	Count	Avg. β	Count	Avg. β	VIF	CI
0: x_1 only	251	0.402	253	−0.399	1.000	1.111	204	0.512	221	−0.528	1.000	1.102
1	1251	0.522	1287	−0.511	1.327	1.726	190	0.581	2780	−0.677	1.361	1.769
2	1540	0.538	1555	−0.541	1.725	2.291	121	0.601	4450	−0.748	1.750	2.299
3	1670	0.541	1646	−0.548	1.956	2.601	70	0.596	5533	−0.793	1.939	2.599
4	1628	0.547	1607	−0.549	2.076	2.836	33	0.590	6307	−0.822	2.035	2.848
5	1627	0.539	1616	−0.540	2.132	2.993	23	0.526	6882	−0.842	2.168	3.091
6	1597	0.529	1617	−0.532	2.171	3.190	11	0.507	7306	−0.859	2.189	3.321
7	1586	0.517	1586	−0.527	2.285	3.452	5	0.538	7659	−0.868	2.297	3.477
8	1537	0.507	1543	−0.511	2.296	3.637	7	0.475	7952	−0.875	2.326	3.709
9	1532	0.495	1535	−0.499	2.347	3.822	1	0.475	8178	−0.879	2.371	3.949
10	1484	0.485	1508	−0.490	2.357	3.903	1	0.415	8472	−0.878	2.387	3.985
11	1480	0.476	1489	−0.478	2.423	4.144	1	0.423	8725	−0.878	2.425	4.095
12	1456	0.468	1488	−0.464	2.440	4.252	1	0.433	8888	−0.880	2.438	4.243
13	1490	0.452	1474	−0.456	2.474	4.389	2	0.420	9043	−0.881	2.524	4.472
14	1486	0.443	1504	−0.440	2.520	4.417	1	0.401	9203	−0.882	2.536	4.538
15	1476	0.432	1496	−0.430	2.533	4.524	0		9352	−0.881	2.548	4.669
16	1480	0.420	1523	−0.418	2.549	4.669	0		9478	−0.882	2.564	4.788
17	1526	0.405	1548	−0.404	2.551	4.844	0		9572	−0.882	2.582	4.858
18	1513	0.396	1587	−0.391	2.552	4.993	0		9681	−0.882	2.617	5.028
19	1551	0.382	1626	−0.377	2.566	5.157	0		9763	−0.881	2.732	5.195
20	1581	0.371	1636	−0.368	2.583	5.240	0		9831	−0.883	2.737	5.392

Note. $y = \Sigma\, \tau_k x_{ik} + 3\varepsilon$, k = 2...21; Variable x_1 has no real direct relationship with y. All variables $x_2 - x_{21}$ have bivariate correlations θ = U(0, 0.6), due to a common factor.

c. The one in the middle can approach positive or negative infinity based on whether its effect is larger or smaller, respectively, than the mean effect of the other two.
d. Out of the three variables, there can be two Type 1 errors.
e. VIF statistics and Condition Indexes may falsely validate these Type 1 errors.

4.5 | Simulations with one variable of theoretical interest plus control variables

Table 1 demonstrates how the addition of control variables can distort the effect of a variable of theoretical interest. In these regressions, the observed variable x_1 is the variable of theoretical interest and observed variables, x_2 through x_{21} are controls. Using the same structure as for the 2-variable and 3-variable regressions, all 21 observed independent variables consist of an unobservable factor x_n common to all and unobservable individual components x_{i1} through x_{i21}. Because x_2 through x_{21} are control variables I allow the amount of the common factor in each control variable to vary randomly:

$$x_k = \sqrt{\theta_k} x_n + \sqrt{1-\theta_k} x_{ik}, \text{ where } k = 1-21.$$

The common factor x_n and individual components x_{ik}, are distributed standard normal. The θ_k are distributed uniformly between 0 and 0.6, apart from θ_1 which is fixed at 0.3. These parameter values provide a realistic case where fully exogenous control variables vary in their relevance to the variable of theoretical interest. The STATA simulation code that generates the results found in Table 1 is presented in full in Appendix S4. I define the dependent variable as:

$$y = 3\varepsilon + \sum_{k=2}^{21} \tau_k x_{ik}.$$

The ε and all the τ_k are distributed standard normal. The magnitude 3ε was chosen so that the coefficients of determination R^2 would approximate those found in typical strategic management studies. The observed variable x_1 of theoretical interest has no direct effect on y, therefore $\tau_1 = 0$ and the summation begins with $k = 2$. In an unbiased regression, the estimated β of x_1 should equal 0. Indeed, because all the τ_k have means of 0, β_1 does have a mean of 0. However, because of the random nature of the τ_k, the β values will vary. I observe how many times Type 1 errors may occur out of 10,000 simulations total for each row of Table 1. I define as cases where β_1 is positive or negative with t-statistics > 2 or < -2. In the first row of Table 1, where there are no control variables, I observe that in 251 cases, β_1 has a t-statistic > 2, and in 253 cases, the t-statistic < -2, for a total of 504 (5.04%) Type 1 errors. This is the exact number to be expected purely from random chance, as per any standard textbook's t-statistic table.

In the next row, I add only one control variable x_2 whose average effect τ_2 is 0, and whose correlation θ with x_1 is on average only 0.3. Even though β_1 still has a mean of 0, the number of cases where the β_1 t-statistic > 2 jumps to 1,251 and the number where the t-statistic < -2 jumps to 1,287, for a total of 2,538 (25.38%) Type 1 errors. Further, the average absolute value of β_1 jumps about 25% to 0.52. The inclusion of a control variable x_2 does not help to isolate the true value of β_1: Because of the common-factor multicollinearity, adding a control variable raises the likelihood of a Type 1 error from 5 to 25%.

Additional control variables, even with mean effects of 0, amplify the Type 1 error problem. The maximum of Type 1 errors occurs with three control variables: the number of cases where the β_1 t-statistic > 2 hits its maximum of 1,670 and the number where the t-statistic < -2 hits its maximum of 1,646, for a total of 3,316 (33.16%) Type 1 errors. Adding additional controls makes the Type 1 error percentage for β_1 slightly lower, but eventually it rises again.

Finally, in the fifth and sixth columns of Table 1, I observe that the maximum, across the 10,000 maximum VIF statistics for these regressions, is always below 3 and the Condition Index is always below 6. Based on these diagnostic results being below any published cut-off value, a researcher might incorrectly conclude there is no multicollinearity problem.

I conducted a second set of 10,000 simulations for each row in the second panel of Table 1 with the τ_k are still distributed as normal but now with a mean of 1. The observed variable x_1 of theoretical interest still has no direct effect on y, so in an unbiased regression, the β of x_1 should still equal 0. In the first row of Table 1, with no control variables, I observe in Columns 7–10 that in 204 cases, β_1 has a t-statistic > 2, and in 221 cases, the t-statistic < -2, for a total of 425 (4.25%) possible Type 1 errors, slightly less than expectation from random chance.

In the second row, I add one control variable x_2 whose average effect τ_2 is 1 and whose correlation θ with x_1 is on average only 0.3. The number of cases where the t-statistic > 2 goes down slightly to 190, but the number where the t-statistic < -2 jumps to 2,780, for a total of 2,970 (29.70%) Type 1 errors. Further, the absolute value of β_1 in this case jumps about 15% relative to the no-control case. The inclusion of a control variable adds bias rather than helps isolate the true value of the coefficient of x_1. Because of common-factor multicollinearity, adding a control variable raises the likelihood of a Type 1 error from 5% to almost 30%.

Adding additional control variables, all with mean effects τ of 1, make the Type 1 error problem worse. The positive real effects of the controls push β_1 further and further into negative territory as the number of controls increases. By the time I have added 20 control variables, a Type 1 error is virtually assured: β_1 exhibits an average value of -0.88, and 9,381 times out of 10,000, a t-statistic < -2. Unlike the case in Columns 1–6, the negative relationship between the number of controls and possible Type 1 errors is monotonically increasing.

> **Conclusion 4:** *If control variables are added to a regression, and these controls share a common factor with a variable of theoretical interest, they may increase the likelihood of a Type 1 error of the theoretically hypothesized variable even at moderate correlations such as 0.3 between controls and hypothesized variables.*

5 | DETECTION OF PROBLEMATIC MULTICOLLINEARITY AND TYPE 1 ERRORS

5.1 | Detection of possible Type 1 errors in published studies

Previously, I analyzed five years of *SMJ* articles and I presented criteria for detecting possible Type 1 errors in published studies. These same criteria can be used by researchers to detect Type 1 errors among results that appear to support hypotheses. A Type 1 error may be present if three criteria all hold. First, two variables, at least one of which appears to support a hypothesis, are correlated about ± 0.3 or more, but possibly even lower with large sample sizes. Second, the two variables have beta coefficients of opposite sign, if correlated positively, or of the same sign, if correlated negatively. Third, a hypothesized variable's bivariate correlation with the dependent variable is (a) the opposite sign from the beta, or (b) designated as not statistically different from zero. If these criteria are met in a research project, the authors can still possibly mitigate the multicollinearity concerns by presenting multiple specifications or by combining the collinear variables. Both approaches are described in detail later. If multicollinearity concerns cannot be mitigated, authors should consider abandoning the project.

These criteria for detection may be all that can be employed by a reader without access to the original data. Of the 2013–2017 *SMJ* articles I reviewed, 39 were regression-based pieces that did not contain correlation tables; Type 1 errors might exist among these, but we cannot tell. For this reason, I believe it is imperative for empirical work to report full bivariate correlation tables, including dependent variables, interaction terms, and polynomial terms. The practice of excluding interaction terms and quadratic terms from correlation tables has no valid mathematical basis and should be abandoned.

Further, given the frequent use of fixed effects in strategic management research, regression tables should include columns with only the variables of theoretical interest and the fixed effects. Controls can then be added in subsequent specifications. These additional columns will help readers assess the effects of the fixed effects alone on the variables of interest, relative to their bivariate correlations with the dependent variable, and then the effects of the controls relative to the fixed effects-only specification. If, for example, the effect of a variable of theoretical interest exhibits a sign flip when regressed only on fixed effects, then at least we would know that a possibly collinear control variable is not causing the flip. This might be a legitimate difference between within-effects and between-effects. On the other hand, if fixed effects do not cause a sign flip, but subsequent control variables do so, we could then examine the correlation of the residuals of the variable of theoretical interest and the control variable after regressing out the fixed effects. The bivariate correlation of residuals might suggest a multicollinearity problem even if the raw bivariate correlations do not appear to do so.

5.2 | Some ability to detect possible Type 1 errors via the joint *F*-test

Many textbooks suggest multicollinearity can lead to large standard errors, and therefore, to small *t*-statistics and Type 2 errors, but that joint *F*-statistics for the collinear variables will remain meaningful (e.g., Gujarati, 2003; Kennedy, 2003). The conclusion of these texts is that researchers can gain some insight about the impact of multicollinearity in a regression by conducting an F-test. Consider again the regression with two variables that share an unobservable common factor. The general form of the *F*-statistic is:

$$F_{(2,n-3)}=\frac{ESS/2}{RSS/(n-3)}.$$

The equation for the Residual Sum of Squares (RSS) was derived as Equation (6). The Estimated Sum of Squares (ESS) for a regression that includes both observable variables x_1 and x_2 can be derived similarly, as:

$$ESS=n\frac{2\theta\gamma^2+2\sqrt{\theta}\sqrt{1-\theta}\gamma(\delta_1+\delta_2)+\delta_1^2+\delta_2^2-2\theta\delta_1\delta_2}{1+\theta}.$$

Replace θ with 1 − ε. As ε gets arbitrarily close to 0, ESS and RSS can be rewritten as:

$$\lim_{\varepsilon\to 0}ESS=n\frac{2\gamma^2+\delta_1^2+\delta_2^2-2\delta_1\delta_2}{2}=n\frac{2\gamma^2+(\delta_1-\delta_2)^2}{2};$$

$$\lim_{\varepsilon\to 0}RSS=n\frac{(\delta_1+\delta_2)^2+2}{2};$$

$$\lim_{\varepsilon\to 0}F_{(2,n-3)}=\frac{\left(2\gamma^2+(\delta_1-\delta_2)^2\right)/2}{\left((\delta_1+\delta_2)^2+2\right)/n-3}.$$

Using the same numerical examples with which I assessed the n required to generate a Type 1 error based on the t-statistic, for a large value of $\delta_1 = 0.5$ and $\delta_2 = 0$, relative to $\sigma_\varepsilon = 1$, n would have to equal 60 to achieve the 95th percentile of the F-distribution. Yet, this number is larger than 39, which I calculated as the minimum n required to achieve a t-statistic based Type 1 error. For a small value of $\delta_1 = 0.1$ and $\delta_2 = 0$, n would have to equal 1,210 to achieve that threshold. This number is larger than 800 needed to achieve a t-statistic based Type1 error.

I conclude that as correlation $\theta \rightarrow 1$ in the unobservable common factor case, the t-tests for the coefficients β_1 and β_2, and the joint F-test for the two variables x_1 and x_2 will often exhibit large values even if the effects of only one of δ_1 and δ_2 is non-zero and the sample sizes n are typical of empirical studies. Thus, a large joint F-statistic should not be interpreted as evidence that a multicollinearity problem does not exist, in other words, that the estimated effects of both independent variables x_1 and x_2 are real and unbiased, as is sometimes argued in empirical work.

However, the discrepancy between the n required to achieve a t-statistic-based and joint F-statistic-based Type 1 error shows that there will be a range of n within which high t-statistics, but a weak joint F-statistic can be the sign of a multicollinearity-based problem. In this case, researchers should be concerned about the possibility of t-statistic based Type 1 errors.

> **Conclusion 5:** *If two independent variables are correlated moderately, such as ±0.3 or more, they may have large but false* t*-statistics and their signs may go in opposite directions:*
>
> *a. Even with small effects, a Type 1 error may emerge with* n > *1200.*
> b. *A joint* F*-test with a high value, and thus, a low* p*-value is not a sign that multicollinearity problems are absent.*
> c. *A joint* F*-test with a low value may signify that the large* t*-statistics may be Type 1 errors due to multicollinearity.*

6 | MITIGATION OF MULTICOLLINEARITY CONCERNS AND TYPE 1 ERRORS

If, based on the detection strategies previously discussed, multicollinearity does appear to be causing a Type 1 error, what can be done? I first discuss strategies that appear valid and helpful to mitigate likelihood of Type 1 error due to common-factor multicollinearity, and follow with an overview of strategies suggested in the literature that do not appear valuable.

6.1 | Appropriate mitigation

6.1.1 | Presentation of multiple specifications with and without each collinear variable

To mitigate multicollinearity concerns, authors should show separate regressions with only one, then only the other, and then both collinear variables of interest. If the signs are consistent and magnitudes are roughly consistent in all specifications, then it is unlikely that multicollinearity is distorting results. If the coefficients switch signs or vastly change magnitudes due to the addition of a correlated variable I do not believe that a hypothesis can be conclusively supported. I recognize that this conservative practice will cause some Type 2 errors, but I believe that is a small price to pay relative to the likelihood of a false positives and Type 1 errors. Note that this is a more comprehensive approach than the suggestion of many methods texts to merely "drop" one of the correlated

variables (e.g., Gujarati, 2003; Kennedy, 2003; Maddala, 2001). I believe the reader must be able to compare specifications with and without each "dropped" variable.

To present an example, I forensically constructed results from a recently published *Strategic Management Journal* article (one of those listed in Appendix S2) that were not actually presented in that research, based on the correlation matrix and descriptive statistics. The coefficients from an OLS regression equation can be perfectly reconstructed from a covariance matrix that includes all variables in the regression. I return to Equation 1:

$$\beta_i = \frac{|M_i|}{|M|}.$$

In the case of the *SMJ* article, the *M* matrix consists of Rows and Columns 6–17 of the published correlation matrix. The matrix M_i is matrix *M* with Column i switched out in favor of Column 1, Rows 6–17. Row i from *M* is left intact.

The standard error for each β_i derived from Equation 1 can be calculated based on the correlation matrix using Equations (4) and (5). In lieu of Equation (6), the following form for RSS, based on Equation (5), allows straightforward computation for large covariance matrices:

$$RSS = n\left(var(y) - \beta^T M \beta\right).$$

Table 2 presents the forensic results. The *SMJ* article only presents the tenth and final column. Based on the results in that column, the authors conclude that their hypothesis regarding profit is supported: Profit appears to have a negative effect on the dependent variable, and the *t*-statistics are great enough for the authors to judge the hypothesis as being supported. The authors cite the low VIF statistic as a reason for dismissing multicollinearity concerns, but as I previously discussed, that statistic is misleading in the presence of an unobserved common factor.

The authors did not provide the results from the first eight columns from Table 2. I constructed these columns, along with an equivalent of their presented result in the ninth column, using the correlation table, descriptive statistics, and the strategy previously described. In the first eight columns, I replaced the full matrices *M* and M_i with appropriate submatrices.

Columns 1–4 in Table 2 demonstrate that the *Profit* variable has trivially small coefficients and *t*-statistics, and flips sign twice as I add additional control variables. When the *Revenue* control variable is added in Column 9, which is correlated 0.75 with Profit, the *Profit* variable's coefficient grows almost tenfold in magnitude relative to that shown in Column 4. Further, only when the Revenue variable is added does the Profit variable's *t*-statistic grow to a level typically considered to be meaningful. Similarly, Columns 5–8 demonstrate that the Revenue control variable flips sign from negative to positive as I add additional control variables. In Column 9, the Revenue variable's coefficient grows almost fourfold in magnitude, relative to that shown in Column 8. Only when the Profit variable is added, does the Revenue control variable's *t*-statistic grow to a level typically considered to be meaningful.

My analysis suggests that a possible unobserved common factor combined with only a small real effect of revenues alone might have generated these results, and might have created a Type 1 error in the form of a false negative relationship of profit with the dependent variable.

In this example, the multicollinearity concerns could not be mitigated with this specification. Had the authors showed meaningful coefficients and *t*-statistics at least in Columns 4 and 8, in addition to Column 9, they would have had a credible argument that multicollinearity concerns had been mitigated and that their results were not Type 1 errors. I encourage authors to present, at the very

TABLE 2 Forensically constructed results from recent *SMJ* article: Revenue and Profit Correlated 0.75

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)[a]	(10)
×1		0.0860	0.1354	0.2921		0.0989	0.1427	0.2793	0.4412	0.41
		(0.1668)	(0.1740)	(0.1800)		(0.1657)	(0.1731)	(0.1722)	(0.1766)	
×2		0.0051	0.2136	0.3003		0.0425	0.2270	0.2584	0.4303	0.39
		(0.1955)	(0.2127)	(0.2118)		(0.1857)	(0.2080)	(0.2000)	(0.2023)	
×3		0.0996	0.1011	0.1542		0.1396	0.0956	0.0842	0.1129	0.12
		(0.1634)	(0.1672)	(0.1592)		(0.1701)	(0.1695)	(0.1571)	(0.1479)	
×4		−0.3372	−0.2814	−0.0332		−0.2530	−0.3135	−0.0982	−0.1076	−0.10
		(0.1932)	(0.2004)	(0.2514)		(0.2272)	(0.2332)	(0.2495)	(0.2340)	
×5			−0.0632	−0.0696			−0.0647	−0.1383	−0.1940	−0.18
			(0.1768)	(0.1904)			(0.1769)	(0.1930)	(0.1826)	
×6			0.1648	0.0664			0.1983	0.1016	0.3605	0.33
			(0.1727)	(0.1680)			(0.1693)	(0.1682)	(0.1948)	
×7			−0.4500	−0.4552			−0.4710	−0.5584	−0.6961	−0.66
			(0.2220)	(0.2113)			(0.2379)	(0.2229)	(0.2176)	
×8			0.4146	0.3881			0.4225	0.4266	0.4832	0.48
			(0.1588)	(0.1513)			(0.1606)	(0.1520)	(0.1447)	
×9				0.3924				0.3588	0.5793	0.56
				(0.1892)				(0.1740)	(0.1900)	
×10				−0.2003				−0.3242	−0.5589	−0.53
				(0.2164)				(0.2381)	(0.2461)	
Revenue					−0.2300	−0.1440	0.0914	0.2779	1.0529	0.96
					(0.1539)	(0.2411)	(0.2528)	(0.2636)	(0.4220)	
					[−1.494]	[−0.597]	[0.361]	[1.054]	[2.495]	
Profit	−0.1000	0.0214	0.0643	−0.0808					−0.7528	−0.68
(H: -)	(0.1573)	(0.2002)	(0.2091)	(0.2108)					(0.3323)	
	[−0.635]	[0.107]	[0.307]	[−0.383]					[−2.265]	

Note. Standard errors in parentheses; *t*-statistics for Profit and Revenue variables in brackets. Standard errors not provided by authors, as per Column 10. Coefficient and *t*-statistic of Profit variable increases when Revenue control variable is added.

[a] Columns 9 and 10 should match exactly. They are close but do not match perfectly, possibly because of rounding error from the two-decimal correlation table.

TABLE 3 Forensically constructed results from recent *SMJ* article: Revenue and Profits combined as one variable

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
×1		0.0933	0.1380	0.2855	0.2698		0.1622	0.2771	0.3463	0.3425
		(0.1692)	(0.1760)	(0.1790)	(0.1797)		(0.1635)	(0.1680)	(0.1707)	(0.1756)
×2		0.0301	0.2177	0.3128	0.2651		0.0946	0.3614	0.3901	0.3813
		(0.1963)	(0.2130)	(0.2055)	(0.2112)		(0.1779)	(0.1987)	(0.1955)	(0.2078)
×3		0.1155	0.0956	0.1135	0.1226		0.1163	0.1677	0.1528	0.1551
		(0.1702)	(0.1724)	(0.1628)	(0.1631)		(0.1537)	(0.1479)	(0.1452)	(0.1485)
×4		−0.3085	−0.2995	−0.2053	−0.0587		−0.4604	−0.4787	−0.3920	−0.3585
		(0.2134)	(0.2181)	(0.2101)	(0.2573)		(0.1904)	(0.1975)	(0.2016)	(0.3090)
×5			−0.0645	0.0017	−0.0917			0.0532	0.0705	0.0544
			(0.1797)	(0.1721)	(0.1965)			(0.1686)	(0.1655)	(0.2015)
×6			0.1780	0.0495	0.0534			0.2374	0.1339	0.1312
			(0.1662)	(0.1673)	(0.1674)			(0.1521)	(0.1637)	(0.1675)
×7			−0.4627	−0.4973	−0.4828			−0.5120	−0.5394	−0.5326
			(0.2333)	(0.2206)	(0.2212)			(0.2024)	(0.1990)	(0.2077)
×8			0.4186	0.3773	0.3978			0.4274	0.3998	0.4013
			(0.1618)	(0.1538)	(0.1552)			(0.1465)	(0.1446)	(0.1473)
×9				0.3970	0.3554				0.2726	0.2726
				(0.1803)	(0.1852)				(0.1788)	(0.1818)
×10					−0.2261					−0.0342
					(0.2289)					(0.2360)
"Resources"	−0.0943	−0.0287	0.0445	−0.0045	0.0316					
Revenue+Profit	(0.0843)	(0.1239)	(0.1258)	(0.1207)	(0.1262)					
	[−1.119]	[−0.232]	[0.354]	[−0.038]	[0.250]					
"Profitability"						−0.1260	−0.2113	−0.2600	−0.2031	−0.1951
Profit/Revenue						(0.1025)	(0.1065)	(0.1001)	(0.1049)	(0.1199)
						[−1.229]	[−1.984]	[−2.597]	[−1.935]	[−1.628]

Note. Standard errors in parentheses; *t*-statistics for "Resources" and "Profitability" variables in brackets. In Columns 1–5, Revenue and Profits are added. In Columns 6–10, Profit is divided by Revenue (STATA's corr2data is used).

least, the equivalents of Columns 4, 8, and 9 when analyzing correlated variables within the same set of regressions.

6.1.2 | Variable combination strategies: Summation

Theory may exist such that collinear variables may be combined in a logical fashion. For example, Gujarati (2003: 363) states "even if we cannot estimate one or more regression coefficients with greater precision, a linear combination of them can be estimated relatively efficiently." The simplest such combination would be the sum. To continue the profit and revenue example using the published *SMJ* article, suppose the researcher's primary interest is the effect of resource availability on a dependent variable, rather than separate effects of profit or revenues. Both profit and revenues are associated with the construct of resource availability, a point made by the authors of the *SMJ* article. If a common-factor multicollinearity problem is detected when profit and revenues are both included in a regression, combining the two variables will eliminate distortion of the beta coefficients toward the infinite, if the sum is not highly correlated with other variables in the regression.

Because the Profit and Revenue variables are standardized in the published article, their sum will not weight the effect of one variable more highly than the other. In Table 3, I created a variable that is the sum of the Profit and Revenue variables. It is, of course, the task of the authors to interpret the sum of variables such as profit and revenue in a sound and meaningful way. I do not attempt that here. I wish only to test for the possible presence of an effect of the sum. Columns 1–5 demonstrate that the sum has very little effect on the dependent variable. Thus, the authors' suggestion that resource availability may be driving their results does not appear to be supported by the evidence. In this case, summation did not mitigate multicollinearity concerns. Beyond the case of the sum, the researcher may consider principal component techniques to determine appropriate linear combinations of collinear variables that can be estimated (e.g., Gujarati, 2003; Kennedy, 2003). But, again, the focus must be on combinations that are theoretically meaningful and reasonably interpretable.

6.1.3 | Variable combination strategies: Division

Alternatively, there might be a theoretical basis to include the quotient of the Profit and Revenue variables. Like summation, the quotient eliminates the issue of the correlation between the two variables (e.g., Maddala, 2001). Perhaps Revenue was included as a control variable in the published regression because it serves as a proxy for the size of the firm; this factor is likely also common to the Profit variable. The Profit/Revenue quotient can then become a meaningful proxy for "Profitability," where the effect of a possible common factor such as size is reduced, if valid theory exists to support its use. In Table 3, Columns 6–10, I included a simulated quotient variable created by using the corr2data function in Stata. This variable could not be derived precisely from the correlation table as could the sum because the correlations of a product or quotient do not have a fixed relationship with the correlations of the component variables. Thus, the results presented here will not be the same as those obtained using the actual data.

The Profit/Revenue quotient is not highly correlated with any of the control variables: The highest correlation is .26. Likely thanks to the low correlations, the Profit/Revenue coefficient remains relatively stable in magnitude in Columns 6–10 regardless of the choice of control variables in each regression; this is a good sign. Further, the *t*-statistics are sufficiently large that the quotient variable might be considered to have a meaningful effect on the dependent variable.

I argue that the quotient approach may have successfully mitigated multicollinearity concerns in this study, both between variables of interest such as Profit and Revenue, and between the quotient and control variables as well. The relatively stable coefficients across

multiple specifications are evidence of a real, robust effect. I do not argue that the quotient is the preferred independent variable for this *SMJ* study; that would require theory about a construct such as "Profitability," and the authors do not provide that. Nor does it imply that the authors' hypothesis is necessarily supported. There is a danger that a form of an independent variable, such as the sum or a quotient, may be chosen for a study because of its *t*-statistic (Goldfarb & King, 2016). To avoid such a perception, authors should be able to verbally justify a specification based on theory such that a rationale exists for its use other than large *t*-statistics.

6.2 | Inappropriate mitigation

6.2.1 | Discredited mitigation approaches

I have demonstrated that low values of VIF statistics or Condition Indexes do not mitigate multicollinearity concerns. Three additional popular mitigation approaches for multicollinearity have been convincingly discredited in recent publications. These include mean-centering the independent variables (Echambadi & Hess, 2007), residualizing one independent variable by regressing it on the other ones (Kennedy, 2003), and orthogonalization of the variables (Mitchell, 1991). I next discuss additional approaches that also appear to be of little value for the case of common-factor multicollinearity.

6.2.2 | Penalized regression approaches (Ridge regression)

Ridge regression (Hoerl & Kennard, 1970) "penalizes" estimated coefficients by pushing them toward zero. The underlying logic is that multicollinearity has made the coefficients too large and that they need to be reduced in size in a systematic fashion. Controversy exists whether Ridge regression, in fact, accomplishes the reduction in a systematic fashion or whether the reduction is simply arbitrary (Kennedy, 2003, p. 216, provides a summary of this debate).

Moving beyond that debate, I have demonstrated that common-factor multicollinearity can often result in signs of coefficients being flipped, with high *t*-statistics in the wrong direction. Ridge regression cannot mitigate the problem because it can only reduce the size of the beta coefficients. It cannot flip them back to the correct sign with a meaningful *t*-statistic. I conclude that Ridge regression often cannot mitigate concerns about common-factor multicollinearity.

6.2.3 | Collecting additional data

I have shown that common-factor multicollinearity and the Type 1 errors that may result are not small data set problems. I demonstrated analytically that estimated beta coefficients approach positive and negative infinity even with an infinite data population. An implication of this result is that collecting more data of the same type as already in a regression is not useful, in contrast to the arguments made in many methods textbooks (e.g., Judge et al., 1982; Maddala, 2001). My result also suggests that replication by others may not cull the Type 1 errors: If the same unobservable common factor exists in samples collected by different researchers, even at different time periods or different industrial or geographic settings, the same false positives are likely to reappear. In the era of the "replication crisis" this conclusion is particularly salient: even careful and rigorous replications may not be enough to validate the presence of a real relationship between two variables in the social sciences.

6.2.4 | Adding control variables

I have shown that exogenous control variables correlated via a common factor with a variable of theoretical interest may cause a Type 1 error regarding the latter, rather than isolating its true effect. Adding controls may make the problem worse. In general, controls should be used to demonstrate that an observed effect does not go away; controls should not be used to make an effect appear. If an effect exists only with a certain set of controls, it is likely a Type 1 error.

6.2.5 | Do nothing

Finally, I believe I have shown that "do nothing" is not a valid option regarding common-factor multicollinearity. The "do nothing" argument is based on the perspective that multicollinearity leads only to Type 2 errors, but not Type 1 errors.

7 | CONCLUSION

I have presented evidence that multicollinearity is a concern in a nontrivial proportion of strategic management research that employs multivariate regression due to the frequent presence of unobservable common factors and proxy variables. Whenever multiple proxies are included for the same unobservable construct, multicollinearity exists as does a common-factor problem. Ignorance of this problem has likely led to Type 1 errors in published research. Although this problem is distinct from that of endogeneity–error terms remain fully exogenous here–I argue that the common-factor multicollinearity problem should be afforded equal attention.

I have shown that econometricians' (e.g., Goldberger, 1989, 1991) perspective that multicollinearity causes only Type 2 errors is not valid in the presence of an unobservable common factor. Estimated beta coefficients of the correlated variables become misleading, tending toward infinite magnitudes in opposite directions even if their real effects are small and of the same sign. Further, these coefficients may be associated with large and false *t*-statistics. I conclude that multicollinearity causes Type 1 errors as well as Type 2 errors.

I have gone beyond OLS regression and conducted extensive simulations for forms of regression that require maximum likelihood estimation such as logit, probit, tobit, and exponential hazard models. In all these cases, all the conclusions of this article hold across a broad range of simulation values. High correlations due to an unobservable common factor may lead to Type 1 errors using these forms of regression just as they do for OLS regression. To evaluate this claim by using the STATA code I have provided in Appendix S4, replace the "regress" command with "probit" or "logit" for example and create a binary dependent variable that is based on the current continuous variable y.

Finally, I have illustrated how researchers can detect and mitigate common-factor multicollinearity concerns by (a) providing full bivariate correlation tables in all regression-based empirical studies that include all interaction terms and quadratic terms, and (b) providing separate specifications that include each collinear variable alone followed by a specification with both. If all the specifications provide consistent results, the concern has been mitigated. Should (a) and (b) suggest that a multicollinearity problem remains, researchers can (c) combine the collinear variables in a fashion—creating a sum or a quotient for example—that makes the most sense theoretically.

ACKNOWLEDGEMENTS

I would like to thank Arturs Kalnins (Sr.), Myles Shaver, Helene Shapiro, and seminar participants at the University of Maryland for helpful comments on this article, along with *SMJ* Editor Sendil

Ethiraj and two anonymous reviewers for their substantial efforts. I would also like to thank Nicole McQuiddy-Davis for expert research assistance.

REFERENCES

Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: Wiley.

Boyd, B. K., Gove, S., & Hitt, M. A. (2005). Construct measurement in strategic management research: Illusion or reality? *Strategic Management Journal*, *26*(3), 239–257.

Day, N. E., Wong, M. Y., Bingham, S., Khaw, K. T., Luben, R., Michels, K. B., … Wareham, N. J. (2004). Correlated measurement error—Implications for nutritional epidemiology. *International Journal of Epidemiology*, *33*(6), 1373–1381.

Echambadi, R., & Hess, J. D. (2007). Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*, *26*(3), 438–445.

Frisch, R. (1934). *Statistical confluence analysis by means of complete regression systems*. Oslo: Universitets Økonomiske Instituut.

Godfrey, P. C., & Hill, C. W. (1995). The problem of unobservables in strategic management research. *Strategic Management Journal*, *16*(7), 519–533.

Goldberger, A. S. (1989). The ET interview. *Econometric Theory*, *5*(1), 133–160.

Goldberger, A. S. (1991). *A course in econometrics*. Cambridge, MA: Harvard University Press.

Goldfarb, B., & King, A. A. (2016). Scientific apophenia in strategic management research: Significance tests & mistaken inference. *Strategic Management Journal*, *37*(1), 167–176.

Gujarati, D. N. (2003). *Basic econometrics* (4th ed.). New York: McGraw-Hill.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Judge, G. G., Hill, R. C., Griffiths, W., Lutkepohl, H., & Lee, T. C. (1982). *Introduction to the theory and practice of econometrics*. New York: Wiley.

Kennedy, P. (2003). *A guide to econometrics* (5th ed.). Cambridge, MA: MIT Press.

Maddala, G. S. (2001). *Introduction to econometrics* (3rd Edition). Chichester, U.K.: John Willey and Sons Inc.

Mitchell, D. W. (1991). Invariance of results under a common orthogonalization. *Journal of Economics and Business*, *43*(2), 193–196.

Pearson, K. (1920). Notes on the history of correlation. *Biometrika*, *13*(1), 25–45.

Pfarrer, M. D., Pollock, T. G., & Rindova, V. P. (2010). A tale of two assets: The effects of firm reputation and celebrity on earnings surprises and investors' reactions. *Academy of Management Journal*, *53*(5), 1131–1152.

Sastry, M. V. R. (1970). Some limits in the theory of multicollinearity. *American Statistician*, *24*(1), 39–40.

Yli-Renko, H., & Janakiraman, R. (2008). How customer portfolio affects new product development in technology-based entrepreneurial firms. *Journal of Marketing*, *72*(5), 131–148.

Yoo, W., Mayberry, R., Bae, S., Singh, K., He, Q. P., & Lillard, J. W., Jr. (2014). A study of effects of multicollinearity in the multivariable analysis. *International Journal of Applied Science and Technology*, *4*(5), 9–19.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Kalnins A. Multicollinearity: How common factors cause Type 1 errors in multivariate regression. *Strat Mgmt J*. 2018;39:2362–2385. https://doi.org/10.1002/smj.2783