Multicollinearity in Statistical Modeling: A Review

September 2026

Joost de Winter

Faculty of Mechanical Engineering, Department of Cognitive Robotics, Delft University of Technology, Delft, The Netherlands

Corresponding author: j.c.f.dewinter@tudelft.nl

Abstract

The presence of strong linear dependencies among predictor variables, a condition known as multicollinearity, is a common concern in quantitative research across diverse fields such as econometrics, psychology, ecology, and medicine. Its main effect, the inflation of the variance of regression coefficients, is well documented, but views on its interpretation, diagnosis, and remediation differ widely. This review examines the conceptualization of multicollinearity, from a simple data deficiency to an amplifier of bias from model misspecification. Next, this review documents an array of diagnostic tools, from the widely used Variance Inflation Factor (VIF) and condition number to more specialized techniques for detecting collinearity-influential observations. This review also assesses the impact of multicollinearity in hierarchical and geographically weighted models, and identifies its interaction with measurement error and outliers. Finally, the review discusses strategies to address its consequences, including model respecification, biased estimation techniques such as ridge and principal component regression, and modern machine learning approaches. In conclusion, the appropriate response to multicollinearity depends on the research goal, the underlying data-generating process, and the analytical context.

Keywords: multicollinearity; regression analysis; variance inflation factor; ridge regression; collinearity diagnostics; biased estimation

1. Introduction

Multiple regression analysis is a cornerstone of quantitative research. It is a powerful tool for modeling the relationships between a dependent variable and a set of independent, or predictor, variables. A key requirement for estimating a multiple regression model is that the predictors are not perfectly linearly dependent. Even when this requirement is met, strong linear dependencies among predictors can leave little independent variation to estimate their separate effects, a situation commonly referred to as multicollinearity.

The term multicollinearity was introduced by Ragnar Frisch in 1934 to describe a perfect or near-perfect linear relationship among a set of variables [1], as cited in [2]. It is important to distinguish between its two forms. Perfect multicollinearity, where one predictor is an exact linear combination of others, is a mathematical issue that makes unique estimation of the affected coefficients impossible. A far more common and insidious problem is *near-multicollinearity*, which occurs when predictors are nearly

linearly dependent. This condition results in an ill-conditioned design matrix (i.e., the matrix containing the predictor variables), which, while not preventing estimation, can compromise the reliability and interpretation of statistical results [3,4,5]. See Textbox 1 for an illustration.

Textbox 1: Illustration of multicollinearity

Figure 1 illustrates a least-squares linear regression with two predictors (x_1 and x_2). In the left panel, these two predictors are uncorrelated ($r_{12,p} = 0.00$), and in the right panel they are strongly related ($r_{12,p} = 0.90$). The subscript p refers to the population value, that is, the value obtained if the sample size were infinite; symbols without this subscript refer to sample estimates.

All variables have a normal distribution with a mean of 0 and a standard deviation of 1. In both panels, the first predictor variable correlates 0.50 with the outcome y ($r_{y1,p} = 0.50$), and the second predictor variable correlates 0.40 with it ($r_{y2,p} = 0.40$). Using a random number generator, 10,000 individuals were drawn from a population.

In the left panel, where the two predictor variables are uncorrelated, both regression coefficients are positive. More specifically, $\beta_1 \approx 0.497$ and $\beta_2 \approx 0.404$, close to the population correlation coefficients of $r_{y1,p} = \beta_{1,p} = 0.50$ and $r_{y2,p} = \beta_{2,p} = 0.40$.

In the right panel, where multicollinearity exists ($r_{12,p} = 0.90$), $\beta_1 \approx 0.754$ and $\beta_2 \approx -0.281$. Because the sample size is large, at $n = 10{,}000$, the regression coefficients are again close to their population values ($\beta_{1,p} \approx 0.737$ and $\beta_{2,p} \approx -0.263$). Because x_1 and x_2 correlate strongly, the plane fitted through this point cloud has slopes that no longer resemble the zero-order correlations of 0.50 and 0.40. The sign reversal is a property of the population: it is the partial relationship obtained when the outcome is regressed simultaneously on two strongly correlated predictors, and an infinitely large sample would yield the same negative coefficient. The finite-sample consequence of multicollinearity (i.e., inflated sampling variability of the estimated coefficients) is illustrated in Textbox 2. For reference, the population variance inflation factor (VIF) is 1.00 in the left panel and 5.26 in the right panel.

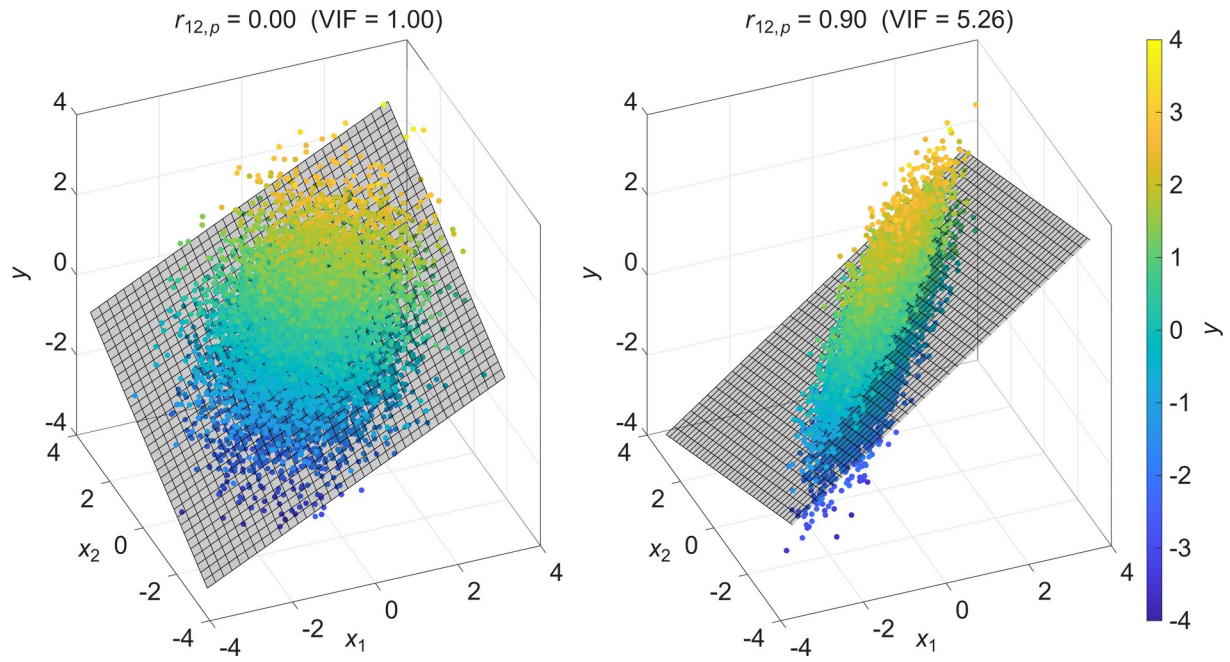

Figure 1. Simulation result for 10,000 random individuals, where the predictor variables x_1 and x_2 are uncorrelated in the population (**left panel)** or strongly correlated in the population (**right panel**). A translucent least-squares fitted plane is depicted, with slopes equal to the regression coefficients β_1 and β_2. The 10,000 markers are color-coded according to their *y*-value. In both panels, $r_{y1,p}$ = 0.50 and $r_{y2,p}$ = 0.40.

The consequences of multicollinearity are solidly established in the statistical literature. Multicollinearity leaves the coefficient estimates of an otherwise correctly specified model unbiased and increases their variances and standard errors [6,7]. This leads to unstable estimates that are sensitive to minor changes in the data, wide confidence intervals, and low statistical power for individual predictors, often resulting in a highly significant overall model fit with no significant individual coefficients [3,4,5,8].

Despite this general understanding, an inspection of the literature reveals a rich and often contentious debate regarding the proper way to conceptualize, diagnose, and address multicollinearity. Some scholars frame multicollinearity as a data deficiency problem, akin to a small sample size [9]. Others argue that multicollinearity is a feature of the data that ordinary least squares (OLS) handles correctly, with the resulting uncertainty honestly reflecting the limitations of the data [7,10]. A third perspective holds that severe multicollinearity amplifies the bias from model misspecification. A fourth perspective uses multicollinearity as a clue for theory building.

This paper reviews these perspectives based on methodological and applied studies. This review first explores the ways multicollinearity has been interpreted and its consequences for statistical modeling. Then, it presents an overview of the diagnostic tools that have been developed, from simple heuristics to advanced statistical tests and visualization techniques. Following this, proposed remedies are reviewed, including data-level solutions, model respecification, and the use of alternative estimation techniques such as

biased regression and machine learning. Finally, this paper examines specific contexts where multicollinearity presents unique challenges, such as in moderated regression, structural equation modeling, and spatial analysis. While earlier work has reviewed and simulated collinearity methods in ecology [11] or corrected common misconceptions in international business research [12], the present review offers broader disciplinary coverage, integrates more recent methodological developments, and provides new Monte Carlo simulations that illustrate key concepts. The present study is a narrative review: no formal database search protocol, inclusion criteria, or cutoff dates were applied. Instead, sources were selected for their conceptual or historical significance, methodological contribution, and disciplinary breadth. Accordingly, the coverage of individual estimators and application areas is selective.

2. Conceptualizing the Problem

The appropriate response to multicollinearity is contingent on how the phenomenon is conceptualized. The literature reveals a range of interpretations, from a simple data problem to a signal of theoretical or methodological problems.

The most traditional view, articulated, for example, by Belsley [9], is that multicollinearity is a data problem. This frames multicollinearity as a deficiency in one's data, i.e., a condition of micronumerosity (an approximate synonym for small sample size; [13]) where the data lack sufficient independent variation to permit the disentangling of individual predictor effects [14]. This view is often summarized with the quote from Blanchard [15] (p. 449) that "multicollinearity is God's will, not a problem with OLS or statistical techniques in general" [16,17].

A stronger version of this perspective argues that multicollinearity is not a problem at all. Morrissey and Ruxton [7] contend that the increased standard errors produced by OLS are an honest and correct reflection of the uncertainty in the data. In this view, multicollinearity is "part of the information that multiple regression uses to infer direct effects" [7] (p. 6), and attempts to resolve it are misguided interventions that often introduce bias. Vanhove [10] echoes this, using the analogy that multicollinearity is a problem for analysis "in the same way that Belgium's lack of mountains is detrimental to the country's chances of hosting the Winter Olympics: It is an unfortunate fact of life, but not something that has to be solved" (p. 3). This school of thought attributes the perceived issues to faulty shortcuts, such as equating non-significance with a zero effect [10].

A third framework holds that multicollinearity amplifies the bias from model misspecification [14]. Kalnins [18] argues that when multicollinearity arises from an unobservable common factor shared by two predictors, the standard OLS model can be viewed as misspecified. In that situation, the misspecification biases the coefficients, and the multicollinearity amplifies this bias, which can lead to spurious findings. Under a correctly specified model that satisfies the standard exogeneity assumptions, by contrast, multicollinearity does not introduce bias. This perspective shifts the viewpoint from data deficiency to theoretical deficiency. Similarly, Schisterman et al. [19] argue that the impact of collinearity cannot be understood without first defining the causal structure of the variables with a directed acyclic graph (DAG). In their view, too, multicollinearity amplifies the bias that results from model misspecification.

Finally, some researchers have framed multicollinearity as a diagnostic tool for theory building. Schwarz et al. [20] propose that high inter-correlation among a set of theoretically related first-order constructs in structural equation modeling (SEM) should be interpreted as an empirical signal of a higher-order structure. This view transforms the statistical issue into a clue that can guide the researcher toward a more accurate and parsimonious theoretical model.

3. Statistical Consequences

Regardless of the conceptual framework, the statistical consequences of multicollinearity on OLS are well established. The primary effect is the inflation of the variance of the estimated regression coefficients [21]. For any non-intercept coefficient β_j, its variance is proportional to the term $1 / (1 - R^2_j)$, where R^2_j is the coefficient of determination from regressing predictor x_j on all other predictors. As R^2_j approaches 1, this term, known as the Variance Inflation Factor (VIF; [22], citing [23]), approaches infinity, leading to several adverse effects.

3.1. Estimation and Inference

The most cited consequences relate to the stability and interpretation of parameter estimates. Coefficients become highly sensitive to small changes in the data, a phenomenon known as “bouncing betas” [24] (p. 36). This instability reflects inflated sampling variance, resulting in large standard errors and wide confidence intervals [3,21]. The effect on precision can be quantified directly: multicollinearity inflates the length of a coefficient’s confidence interval by a factor of the square root of its VIF [25]. This can cause estimated coefficients to have signs opposite to what theory or simple correlations would suggest [5,26]. Textbox 2 provides an illustration of these effects.

Textbox 2: A visual demonstration of how multicollinearity amplifies sampling error

A Monte Carlo simulation was used to demonstrate how correlation between two predictors inflates the sampling variability of their regression coefficients while everything else about the population is held constant. Data were generated from a population in which the two predictors, x_1 and x_2, are drawn from a bivariate normal distribution with unit variances and have identical true (unstandardized) regression coefficients ($\beta_{1,p} = \beta_{2,p} = 0.40$); for each observation, the outcome equals this weighted sum of the predictors plus an independent Gaussian error term with a constant *SD* of 1.0. For each condition, 10,000 samples of $n = 50$ were drawn, and the coefficients β_1 and β_2 were estimated using OLS.

The results show that when the predictors are uncorrelated in the population, the estimates scatter around the true values with essentially independent errors: the correlation between β_1 and β_2 across repeated samples is $r \approx 0.00$, and each coefficient has a sampling standard deviation of approximately 0.15 (left panel). When $r_{12,p} = 0.80$, the estimates remain unbiased, with means still equal to about 0.40, but the sampling standard deviation of each coefficient rises to approximately 0.25 (right panel). This inflation by a factor of about 1.7 closely matches the theoretical factor given by the square root of the VIF, $\sqrt{(1/(1 - 0.80^2))} \approx 1.67$. In addition, the estimates become strongly negatively correlated across samples ($r \approx -0.80$): because the data cannot fully distinguish the two correlated predictors, a random overestimation of one coefficient is compensated for by an underestimation of the other.

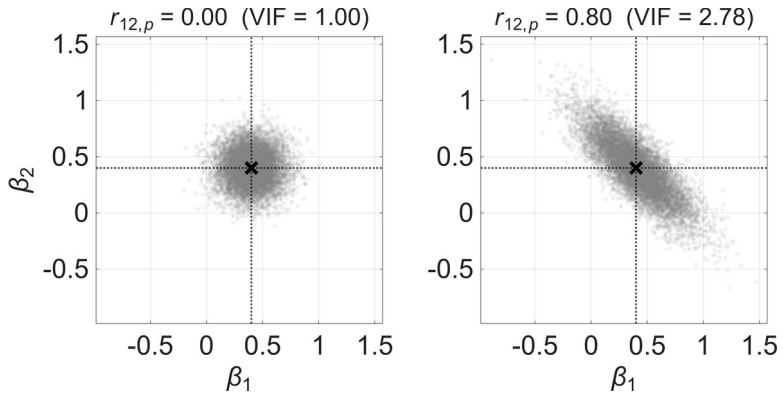

Figure 2. 10,000 estimated coefficient pairs (β_1, β_2) from samples of n = 50 drawn from a population with identical true coefficients ($\beta_{1,p} = \beta_{2,p} = 0.40$; marked by the cross) and constant residual variance, where the predictors are uncorrelated ($r_{12,p}$ = 0.00; **left panel**) or highly correlated ($r_{12,p}$ = 0.80; **right panel**). Collinearity inflates the sampling variability of each coefficient ($SD \approx 0.25$ vs. 0.15) and induces a strong negative correlation between the estimates ($r \approx -0.80$ vs. 0.00). The corresponding population VIFs are 1.00 and 2.78.

The inflated standard errors reduce the power of significance tests. This increases the probability of Type II errors, where a researcher fails to detect a true effect [27]. This often leads to the classic paradox of a highly significant overall model fit (based on the *F*-test) with no statistically significant individual predictors [5]. Kalnins [18] shows that when two predictors share an unobserved common factor, i.e., the model is misspecified, OLS coefficients can be biased away from zero; in such settings, multicollinearity amplifies this bias, inflating *t*-statistics and raising Type I error rates. Because the bias persists as the sample grows, larger samples produce even larger spurious *t*-statistics.

3.2. Prediction

An important distinction is often made between a model's explanatory and predictive goals. For prediction, multicollinearity does not necessarily reduce average point accuracy, provided the new data preserve the same dependency among predictors as the training data. The explanation is that when two predictors (e.g., a person's age and years of experience) are correlated, their coefficients become individually unreliable. The model might put too much weight on experience and, to compensate, put too little weight on age. When predicting an outcome for a new person whose age and experience fit the typical pattern (known as the collinear ridge), the errors in the two coefficients cancel each other out, and the prediction remains reasonably accurate (e.g., [7,28]).

This stability can be misleading, however, as predictions at combinations of predictors that depart from the observed ridge (i.e., outside the data region) become highly uncertain and often poor [21,29]. A simulation study by Mundfrom et al. [30] found that adding a collinear predictor widened the confidence intervals of the predicted values, more so at higher predictor correlations and in smaller samples, although the widening was small (at most about 0.19 standard deviations of the outcome).

3.3. Interaction with Other Data Problems

The effects of multicollinearity can be compounded by other common data issues. The combination of multicollinearity and measurement error is particularly pernicious. Measurement error in a predictor attenuates its own coefficient, and this underestimated effect can be misattributed to a correlated predictor, which leads to large biases [6]. Zidek et al. [31] termed this a “transfer of apparent causality” (p. 448), where a non-causal surrogate variable can become spuriously significant. Furthermore, measurement error can mask the presence of multicollinearity in standard diagnostics such as the aforementioned VIF (e.g., reduced reliability leading to reduced VIF; see Case 1 in [32]).

Outliers also interact with multicollinearity. Jurczyk [33] demonstrated both analytically and with simulations that robust regression methods designed to handle outliers, such as Least Trimmed Squares (LTS), can fail more often as multicollinearity increases. The collinearity creates flat regions in the loss function, which allows the estimator to fit an outlier without a large penalty, thereby masking it. To address outliers and multicollinearity together, researchers have proposed combined solutions, such as the Robust Dawoud–Kibria (RDK) estimator, which integrates a robust M-estimator (to downweight outliers) into the structure of a biased (shrinkage) estimator (to combat collinearity; [34]).

4. Diagnostic Tools

The literature contains an array of tools for diagnosing multicollinearity, ranging from simple heuristics to complex statistical tests and visualizations. The choice and interpretation of these diagnostics have also been a subject of debate.

4.1. Simple Heuristics and Correlation-Based Measures

The most basic diagnostic is the examination of the pairwise correlation matrix of the predictor variables. While insufficient for detecting complex dependencies involving multiple variables [35], it represents a common first step. Reported rules of thumb for problematic correlations vary, with thresholds of $|r| > 0.7$ [11], $> 0.7–0.8$ [5], or $> 0.8–0.9$ [4,8] frequently cited. A more sophisticated rule, attributed to Klein [36] (p. 101), suggests that multicollinearity is harmful when a pairwise correlation between two predictors is as large as or larger than the overall multiple correlation coefficient (R) of the model [37].

4.2. Variance Inflation Factor and Tolerance

The most widely used diagnostic is the VIF, which quantifies the inflation in the variance of a coefficient due to its linear relationship with other predictors [23]. Its reciprocal, tolerance, is also commonly reported [38].

While no universal threshold exists, a VIF greater than 10 is the most frequently cited guideline for severe multicollinearity [5,23]. More conservative thresholds of VIF > 5 [39] (p. 203), VIF > 4 [5], or even VIF > 2.5 [38] have also been proposed. However, several studies caution against the rigid application of these rules. Lavery et al. [40] showed that problematic effects can occur with VIFs well below 10, while Vatcheva et al. [41] found that a VIF less than 5 does not always indicate low multicollinearity. Conversely, O’Brien [42] argued that VIF rules of thumb are inappropriate because they ignore other factors that strongly influence the variance of regression coefficients, such as a large sample size or a high proportion of explained variance in the dependent variable.

4.3. Eigenvalue-Based Diagnostics

A more comprehensive approach involves the analysis of the singular values of the column-scaled predictor matrix, a method pioneered by Belsley et al. [3].

The condition number (also known as the maximum condition index), defined as the ratio of the largest to the smallest singular value, provides a single summary measure of the matrix's ill-conditioning. Belsley et al. [3] indicated that condition indexes around 10 signal weak dependencies, while values around 30 imply moderate to strong dependencies that can be troublesome. Therefore, a threshold of 30 is typically used to flag dependencies requiring further analysis. In practice, condition indexes are typically computed using singular value decomposition (SVD) for numerical stability; the singular values equal the square roots of the eigenvalues of the cross-product matrix.

To identify the specific variables involved in a near-dependency, Belsley et al. [3] introduced variance-decomposition proportions (VDPs). The diagnostic is the co-occurrence of a high condition index with high VDPs (e.g., > 0.5) for two or more variables, which pinpoints the source and degrading effect of the multicollinearity (see also [5]). This diagnostic is widely available in statistical software; however, implementations differ in their default data preprocessing. Some packages compute these indexes using the correlation matrix (standardized data), while others follow Belsley's recommendation to use the uncentered, column-scaled matrix. The latter approach allows for the detection of multicollinearity involving the intercept, which is masked when data are centered.

4.4. Advanced and Specialized Diagnostics

The literature contains various extensions and novel diagnostic tools designed for specific contexts or to overcome the limitations of standard measures. For assessing collinearity among subsets of coefficients (e.g., for a set of dummy variables), Fox and Monette [22] developed the generalized VIF (GVIF), which measures the inflation of the joint confidence region for a set of parameters. It can be calculated as GVIF = (det R_{11} × det R_{22}) / det R, where R is the correlation matrix of all non-intercept regressors, partitioned so that R_{11} corresponds to the submatrix for the coefficient set of interest, and R_{22} to the submatrix for the remaining regressors. To compare blocks of different sizes on a standard-error scale, a dimension-adjusted form, $\mathrm{GVIF}^{1/(2df)}$, is often used, where df is the number of coefficients in the subset [22].

To distinguish the mere presence of collinearity from its damaging effects, Belsley [43] proposed a non-central-F signal-to-noise test that, when combined with a high condition index (≥ 30), indicates harmful collinearity. More recently, Mohammadi [44] introduced a test of harmful multicollinearity that compares the p-values obtained from OLS and generalized ridge regression to determine whether nonsignificant coefficients are an artifact of collinearity. Chennamaneni et al. [45] proposed the C^2 metric, a diagnostic tool for moderated regressions that assesses the impact of collinearity on statistical inference. It categorizes collinearity based on the potential for an insignificant coefficient to become significant if collinearity were reduced, using the classifications harmful but salvageable collinearity, harmful but hard-to-salvage collinearity, and irrelevant collinearity.

Researchers have developed specialized diagnostics to address challenges beyond standard multicollinearity detection. Hadi [46] introduced the concept of “collinearity-influential points” (p. 145) to identify individual observations that disproportionately create or mask near-linear dependencies. His proposed δ_i influence measure approximates the relative change in the condition index upon deleting the *i*-th observation. Because outliers can mask the true correlational structure of data, Sinan and Alkan [47] proposed a robust procedure. In this method, diagnostics are calculated from a correlation matrix estimated using the Minimum Covariance Determinant (MCD) method, which is resistant to outliers. For Geographically Weighted Regression (GWR), where collinearity can vary spatially, Wheeler [48] and Wheeler and Tiefelsdorf [49] adapted standard tools to create local VIFs and local condition numbers, calculated from the spatially weighted data matrix at each location, allowing researchers to map where estimates may be unstable. To make dense diagnostic tables more intuitive, Friendly and Kwan [50] proposed novel visualizations such as tableplots and a collinearity biplot. Unlike standard biplots, the collinearity biplot is based on the smallest eigenvalues, directly visualizing the near-linear dependencies among predictors.

5. Remedies

The strategies for managing multicollinearity are as varied and debated as its diagnostics. They range from data collection and model respecification to the use of alternative estimation techniques.

5.1. Data-Level and Design Solutions

The most widely recommended, though often impractical, solution is to collect more or better data [3,21,51]. Additional data can break the dependency patterns observed in the initial sample. However, when new observations replicate the same dependency structure, the standard errors decrease, which may yield statistically significant coefficients, while the collinearity diagnostics remain unchanged [52]. A more proactive approach is prevention through study design. In experimental settings, researchers can manipulate variables to ensure they are orthogonal (e.g., with balanced factorial designs [45]). Tests of interactions between correlated continuous predictors should include quadratic terms to avoid spurious interactions, even though this adds collinearity [53]. For observational studies involving spatial data, Heikkila [54] showed that multicollinearity among distance variables can be minimized by carefully selecting the geographic range from which data are drawn.

5.2. Model Respecification

A common approach is to alter the set of predictor variables. The simplest method is to drop one or more of the highly correlated variables. However, this is widely criticized as it can introduce omitted-variable bias, a more severe problem than the precision loss from multicollinearity [6,7,12]. Furthermore, such an action fundamentally alters the model being tested. As O’Brien [42] notes, dropping a variable means the remaining coefficients no longer represent the same partial relationships, which means “the theory being tested by the model has changed” (p. 683). Stepwise regression, an automated form of variable deletion, is particularly discouraged in the presence of multicollinearity as its selection criteria become unstable and unreliable [55].

Correlated variables that measure a similar underlying construct can also be combined into a single index or factor score, often guided by Principal Component Analysis (PCA) or factor analysis [11,27,56]. Finally, in SEM, high correlation among first-order constructs can be addressed by specifying a higher-order factor model, which is both more parsimonious and theoretically richer [20].

5.3. Biased and Regularized Estimation

A major class of solutions involves using estimators that introduce a small amount of bias in exchange for a substantial reduction in variance, leading to a lower overall Mean Squared Error (MSE).

Proposed by Hoerl and Kennard [57], ridge regression (RR) adds a small constant, k, to the diagonal of the cross-product matrix, $\beta_{RR} = (X'X + k \cdot I)^{-1} \cdot X'Y$, which stabilizes the inverse and shrinks the coefficients. In this formulation, the predictors are centered (and typically scaled), so that the intercept is not penalized. A vast literature exists on methods for selecting the optimal k, with recent work proposing estimators that directly incorporate the condition index [58].

Principal Component Regression (PCR) involves performing PCA on the predictors, deleting the components associated with the smallest eigenvalues (i.e., directions of near-dependency or low information), and regressing the dependent variable on the remaining orthogonal components [59]. Park [60] showed that, over a region of interest, the linear restrictions that minimize the integrated MSE of predicted responses produce the PCR estimator. PCR has also been extended to semiparametric models; for example, Aydın et al. [61] introduced a PCR estimator and a ridge-PCR variant that use kernel partial residuals to estimate both the parametric and non-parametric components.

Partial Least Squares (PLS) builds orthogonal components from the covariance between the predictors and the dependent variable; in the example of Wold et al. [62], it predicted as well as ridge regression and better than PCR with two components (for a tutorial, see [63]).

Lasso (Least Absolute Shrinkage and Selection Operator) is a regularization technique that uses an L1 penalty, calculated as the sum of the absolute values of the model coefficients, which can shrink some coefficients to exactly zero. This property allows it to perform variable selection and regularization simultaneously [64]. However, its variable selection can be inconsistent when predictors are highly correlated. To address this limitation, the Elastic Net was developed as a hybrid of Ridge and Lasso, proving particularly effective in such cases [65] (for an overview, see [66]). A large body of research has proposed further biased estimators designed to offer more flexibility and efficiency than standard ridge regression, including one-parameter options such as the Liu estimator [67] and the Kibria–Lukman estimator [68], and two-parameter options such as the Liu-type [69] and Dawoud–Kibria [70] estimators, with recent proposals incorporating adjustments such as logarithmic transformations and customized penalization to improve estimation efficiency [71].

Textbox 3 compares OLS, ridge regression, and Lasso under severe multicollinearity, for coefficient estimation and for prediction.

Textbox 3: A simulation of remedies for multicollinearity

As in Textbox 2, data were generated from a population with fixed unstandardized coefficients and an explicit residual term. The 20 predictors were drawn from a multivariate normal distribution in which each predictor has a mean of 0 and an *SD* of 1, arranged in two 10-variable blocks with a within-block population correlation of 0.90; predictors from different blocks were uncorrelated. For each observation, the outcome equaled a weighted sum of the predictors plus an independent Gaussian error term with an *SD* of 4.0; the predictors thus accounted for about 45% of the outcome variance (population $R^2 \approx 0.45$). Only five predictors in the first block had nonzero weights (unstandardized population coefficients: $\beta_{1,p} = 1.0$, $\beta_{3,p} = 0.8$, $\beta_{4,p} = 0.6$, $\beta_{7,p} = 0.9$, $\beta_{9,p} = 0.5$); the second block had all-zero coefficients. Without the residual term, OLS would recover all 20 coefficients exactly in any sample with $n > 20$; it is the residual noise, amplified by the strong predictor correlations, that makes the coefficients difficult to estimate.

In each Monte Carlo replication, (1) OLS as a baseline, (2) Ridge, with the penalty that minimized the 10-fold cross-validation error, and (3) Lasso, with the penalty selected by 10-fold cross-validation and the 1-SE rule (200 λ values; $\lambda\text{min}/\lambda\text{max} = 10^{-6}$), were fitted to a fresh training sample of $n = 80$. All three fitted models were then evaluated on an independent test sample of 10,000 new observations drawn from the same population. This complete procedure was repeated 1,000 times. To avoid leakage, both *x* and *y* were standardized using training-sample statistics only, and prediction errors were converted back to the original *y*-units. Figure 3 shows a representative coefficient realization from these 1,000 replications and the Monte Carlo summary.

Left panel (coefficients, standardized scale): Because the models were fitted to z-scored training data, the plotted estimates, as well as the true coefficients shown for comparison, are expressed on the standardized scale; for example, the population coefficient $\beta_{1,p} = 1.0$ corresponds to a standardized value of approximately 0.19. To avoid highlighting an arbitrary draw, the displayed realization was selected automatically as one with the median Lasso model size (3 predictors) and predictive performance close to the Monte Carlo averages. In this representative realization, the OLS estimates deviate markedly from the true coefficients, whereas Ridge strongly shrinks the estimates. The Monte Carlo results confirm that this is systematic: the mean coefficient MSE was 0.085 for OLS, compared with 0.007 for Ridge and 0.006 for Lasso. Setting all coefficients to zero gives an even lower coefficient MSE (about 0.005): the gain of Ridge and Lasso over OLS comes from shrinkage, and neither method reliably identified which predictors in the first block carry the effect. In this realization, Lasso selected three predictors (x_1, x_2, and x_8), with standardized coefficients of approximately 0.26, 0.09, and 0.05, respectively; only x_1 has a nonzero true coefficient. This illustrates a characteristic difficulty of variable selection under strong multicollinearity: when many predictors carry similar information, Lasso can select correlated proxies in place of the variables that generated the outcome.

Right panel (predictive performance across replications): The bars show the mean independent-test RMSE across the 1,000 replications, and the error bars show the empirical 2.5th–97.5th percentile intervals of the replication-level RMSEs. Mean test RMSE was 4.66 for OLS, 4.13 for Ridge, and 4.37 for Lasso. Because the residual *SD* was 4.0, no model can attain a mean test RMSE below this value; Ridge approached this floor, whereas OLS exceeded it markedly due to estimation error. This excess reflects the 20 coefficients estimated from 80 noisy observations: in a repetition of the simulation with uncorrelated predictors, OLS attained the same mean test RMSE, as expected from Section 3.2. The correlations mainly affect the coefficients: the sampling variance of the OLS coefficients (in original units) is about nine times larger than with uncorrelated predictors, in line with the within-block VIF of 9.0. Ridge attained the lowest test RMSE in 91.6% of the replications and Lasso in the remaining 8.4%; OLS never did. Most of the difference between Ridge and Lasso reflects the 1-SE rule, which favors sparse models at some cost in prediction accuracy: with the penalty at the minimum cross-validation error, Lasso reached a mean test RMSE of about 4.16, close to that of Ridge.

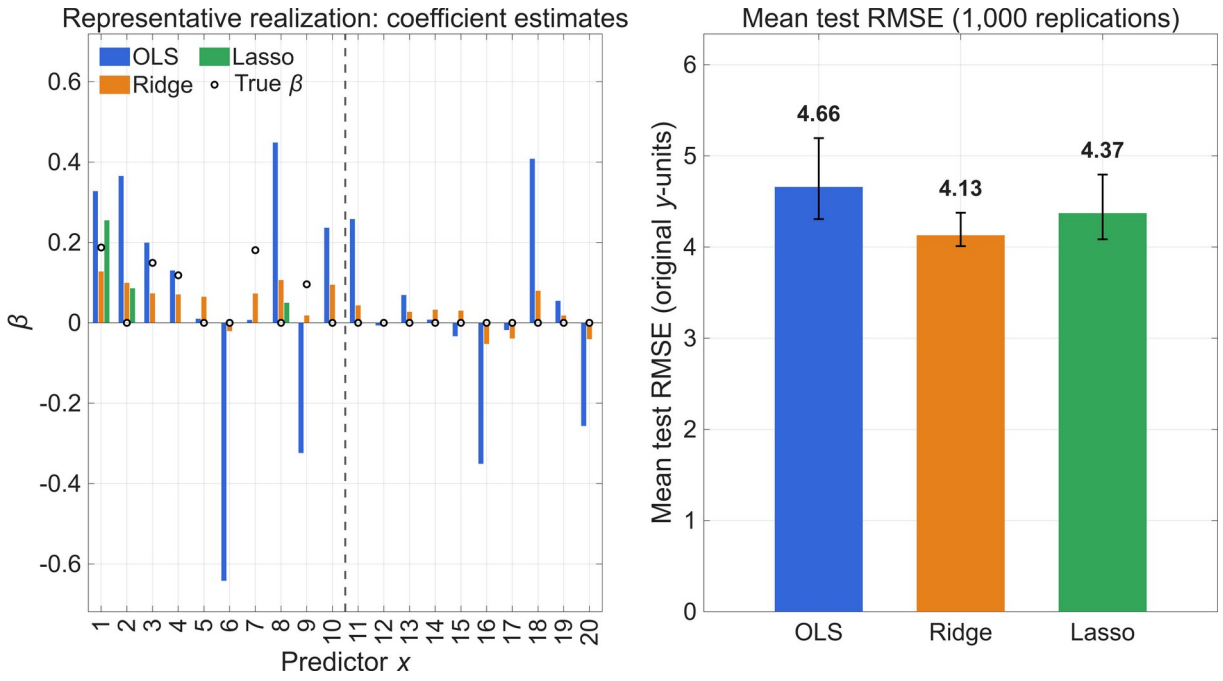

Figure 3. Simulation of OLS, Ridge, and Lasso under severe multicollinearity. In each replication, the models are fitted to n = 80 observations with 20 predictors arranged in two blocks (x_1–x_{10} and x_{11}–x_{20}, separated by the vertical dashed line); predictors correlate at 0.90 within each block in the population and are uncorrelated across blocks. **Left panel:** Standardized coefficient estimates for a representative replication selected from the 1,000 Monte Carlo runs to have the median Lasso model size and near-average predictive performance; the true coefficients are shown as open circles. In this realization, Lasso retained x_1, x_2, and x_8. **Right panel:** Mean independent-test RMSE across the 1,000 replications. Each fitted model was evaluated on a new test sample of 10,000 observations, and the error bars show the empirical 2.5th–97.5th percentile intervals.

5.4. Bayesian and Machine Learning Approaches

More recent approaches have adopted different modeling paradigms. By incorporating prior information, Bayesian methods can stabilize estimates. Adepoju and Ojo [72] demonstrated that a Bayesian model with an informative Normal-Gamma prior produced more precise estimates with narrower credible intervals than OLS. Curtis and Ghosh [16] proposed a hierarchical model that uses a Dirichlet process prior to simultaneously perform variable selection and predictor clustering. The model does this by assigning regression coefficients to groups, where all predictors in a group share an identical coefficient value. Variable selection is accomplished by allowing one of these groups to have a coefficient of zero, effectively removing those predictors from the model.

Non-parametric methods can be less sensitive to multicollinearity for prediction, although interpretation and variable-importance measures can still be affected. In the case studies of De Veaux and Ungar [73], neural networks, owing to their redundant and overparameterized architecture, were relatively insensitive to correlated inputs in terms of predictive performance; this finding may not generalize to other architectures and settings. Genetic Programming can automatically perform predictor selection as part of its

model evolution process [74]. However, these methods often come at the cost of interpretability.

6. Special Topics and Debates

6.1. Moderated and Polynomial Regression

The inclusion of interaction or polynomial terms creates “structural” or “nonessential” multicollinearity, as terms such as $x \cdot z$ and x^2 are typically strongly correlated with x and z, especially when the predictors have nonzero means. Such correlation is not a mathematical necessity, however: depending on the means and the shape of the joint predictor distribution, product terms can be weakly correlated, or even uncorrelated, with their constituent variables [75]. A long-standing debate has focused on the practice of mean-centering variables to mitigate this. Echambadi and Hess [76] argued that centering is an illusory fix that does not alter the underlying model-level collinearity. In contrast, Iacobucci et al. [77] proposed a distinction between “micro” multicollinearity (correlations involving the interaction term) and “macro” multicollinearity (overall model properties). They argued that centering helps alleviate the former, which aids in interpreting main effects. McClelland et al. [78] countered that multicollinearity is a red herring in moderated multiple regression, as the significance test of the interaction term is unaffected by centering. There is broad agreement that centering is a useful tool for improving the interpretability of the lower-order coefficients (after centering, these represent simple slopes at the sample means; before centering, they represent slopes at zero, a point that may lie far outside the data) and that it does not change the statistical test of the highest-order term [75]. Whether centering also “alleviates” collinearity is largely a matter of definition: the reparameterization leaves the fitted model and every estimand unchanged (see Textbox 4), while individual collinearity diagnostics can decrease or increase depending on the data [75].

Textbox 4: Mean-centering in moderated regression

This simulation shows what mean-centering changes in a moderated regression and what it leaves unchanged. For n = 800 observations, the predictors x and z were drawn from a bivariate normal distribution with means of 2 and −1, *SD*s of 1, and a population correlation of 0.50. The outcome was generated as $y = x + z + 1.5x \cdot z + \varepsilon$, where ε was normally distributed with an *SD* of 1.0. The same model was fitted twice with OLS: once with the raw predictors (uncentered) and once with x and z centered at their sample means before the product term was formed (centered).

Both fits produced identical fitted values and R^2. Centering reduced the VIFs of x, z, and $x \cdot z$ from 2.20, 5.19, and 3.85 to 1.35, 1.35, and 1.00 (Figure 4, right panel), and it reduced the condition number of the column-scaled design matrix from 10.27 to 1.75. The coefficients of the lower-order terms changed because they refer to different quantities in the two fits. In the uncentered fit, the coefficient of x is the slope of x at z = 0; in the centered fit, it is the slope of x at the sample mean of z (−0.96). When the same quantity is computed from both fits, the estimates and standard errors are identical (Figure 4, left panel): the slope of x at z = 0 was 1.02 (SE = 0.05), the slope at the mean of z was −0.41 (SE = 0.04), and the interaction coefficient was 1.50 (SE = 0.03). Centering thus changes the meaning of the lower-order coefficients and the values of the collinearity diagnostics, and it leaves the information in the data unchanged.

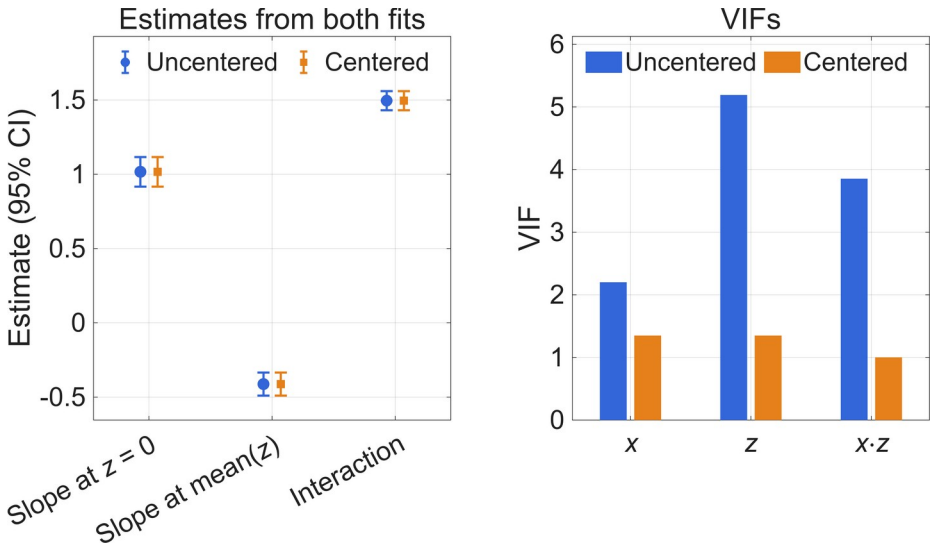

Figure 4. Uncentered and mean-centered fits of the same moderated regression model (n = 800). **Left panel:** Estimates and 95% confidence intervals of three quantities computed from each fit: the slope of x at z = 0, the slope of x at the sample mean of z, and the interaction coefficient. **Right panel:** VIFs of x, z, and $x \cdot z$ in each fit.

6.2. Structural Equation Models (SEM) and Multilevel Models (MLM)

Multicollinearity also presents unique challenges in more complex models. Grewal et al. [79] showed that in SEM, correcting for measurement error often increases the correlations among latent exogenous constructs, which can exacerbate multicollinearity and lead to high Type II error rates. Kock and Lynn [80] introduced the concept of lateral collinearity (predictor-criterion redundancy) and proposed a “full collinearity test” using VIFs on all latent variables to detect it. Shieh and Fouladi [81] found that Level 1 multicollinearity in MLM leaves fixed-effect point estimates unbiased and causes a substantial downward bias in Level 2 variance components. Yu et al. [82] found that multicollinearity at Level 2 is more detrimental than at Level 1 and, using REML estimation, that variance components remained unbiased; they proposed a “top-down” diagnostic method using the condition number and variance decomposition proportions. The challenge of multicollinearity and the use of biased estimation also extends to other complex frameworks, such as models for zero-inflated count data. For example, Ali and Lukman [83] developed a ridge penalized Zero-Inflated Probit Bell (ZIPBell) model and showed with simulations that ridge regularization stabilizes parameter estimates and reduces mean squared error under severe multicollinearity.

6.3. Geographically Weighted Regression (GWR)

In spatial models, such as GWR and spatial autoregressive models, multicollinearity can be a particularly challenging phenomenon. Wheeler and Tiefelsdorf [49] demonstrated that GWR can induce strong negative correlation among local coefficients even when global variables are uncorrelated, an artifact of the local smoothing process. Fotheringham and Oshan [84] later argued that the severity of this issue was overestimated, showing through simulation that GWR is robust to multicollinearity except in extreme cases, and that standard significance testing is a more effective guard against

misinterpretation than local diagnostics. To address multicollinearity in spatial autoregressive models directly, Chavez-Chong et al. [85] proposed a ridge regularization procedure that estimates both the regression coefficients and the spatial dependence parameter simultaneously, offering a method to stabilize estimates in an explanatory modeling context.

7. Discussion and Conclusion

The extensive literature on multicollinearity reveals a problem that is both simple in its origin and complex in its consequences. While early work focused on detection and the development of biased estimators to improve the precision of OLS, the modern literature has become more multifaceted. There is a growing consensus that multicollinearity is a data condition, compatible with the regression assumptions, whose implications must be judged in the context of the research question. The simulations in the textboxes illustrate this: the sign of a coefficient reversed in the population (Textbox 1), standard errors grew with the square root of the VIF (Textbox 2), regularization improved prediction without identifying which correlated predictors carry the effect (Textbox 3), and centering changed the meaning of the lower-order coefficients while leaving the fitted model unchanged (Textbox 4).

The choice of diagnostic tools has expanded substantially, from the ubiquitous VIF to specialized measures for assessing influential points, distinguishing harmful from benign collinearity, and visualizing complex dependency structures. Similarly, the array of remedial strategies has grown to include sophisticated regularization techniques, robust estimators, and machine learning models. However, a recurring theme is the danger of applying these remedies without a clear theoretical justification. Practices such as stepwise variable deletion are now widely discouraged, as the omitted-variable bias they can introduce is considered a more severe error than the imprecision caused by multicollinearity [12,86].

In conclusion, the literature suggests that there is no one-size-fits-all solution. The most effective approach involves a deep understanding of the data-generating process, a clear articulation of the model's purpose, whether for explanation or prediction, and the judicious application of a suite of diagnostic tools. As many authors have concluded, the best remedies are often fundamental research practices: careful, theory-driven model specification and, whenever possible, the collection of more and better data [8,87].

8. Acknowledgments

Gemini 2.5 Pro, GPT-5, and Claude Opus (versions 4.8 and 5.5) were used during the preparation of this manuscript to summarize and cross-check the cited papers, polish the language, and develop and check the simulation code. All material was critically evaluated and revised by the author.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Frisch, R. *Statistical Confluence Analysis by Means of Complete Regression Systems*; University Institute for Economics: Oslo, Norway, 1934; Volume 5.
2. Blalock, H.M., Jr. Correlated independent variables: The problem of multicollinearity. *Social Forces* **1963**, *42*, 233–237. https://doi.org/10.2307/2575696
3. Belsley, D.A.; Kuh, E.; Welsch, R.E. Detecting and assessing collinearity. In *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*; Belsley, D.A., Kuh, E., Welsch, R.E.; John Wiley & Sons: Hoboken, NJ, 1980; pp. 85–191. https://doi.org/10.1002/0471725153
4. Farrar, D.E.; Glauber, R.R. Multicollinearity in regression analysis: The problem revisited. *The Review of Economics and Statistics* **1967**, *49*, 92–107. https://doi.org/10.2307/1937887
5. Slinker, B.K.; Glantz, S.A. Multiple regression for physiological data analysis: The problem of multicollinearity. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology* **1985**, *249*, R1–R12. https://doi.org/10.1152/ajpregu.1985.249.1.R1
6. Freckleton, R.P. Dealing with collinearity in behavioural and ecological data: Model averaging and the problems of measurement error. *Behavioral Ecology and Sociobiology* **2011**, *65*, 91–101. https://doi.org/10.1007/s00265-010-1045-6
7. Morrissey, M.B.; Ruxton, G.D. Multiple regression is not multiple regressions: The meaning of multiple regression and the non-problem of collinearity. *Philosophy, Theory, and Practice in Biology* **2018**, *10*, Article 3. https://doi.org/10.3998/ptpbio.16039257.0010.003
8. Franke, G.R. Multicollinearity. In *Wiley International Encyclopedia of Marketing*; Sheth, J.N., Malhotra, N.K., Eds.; John Wiley & Sons, 2010. https://doi.org/10.1002/9781444316568.wiem02066
9. Belsley, D.A. Harmful collinearity and short data. In *Conditioning Diagnostics: Collinearity and Weak Data in Regression*; Belsley, D.A.; John Wiley & Sons, 1991; pp. 205–244.
10. Vanhove, J. Collinearity isn't a disease that needs curing. *Meta-Psychology* **2021**, *5*, Article MP.2020.2548. https://doi.org/10.15626/MP.2021.2548
11. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; García Marquéz, J.R.; Gruber, B.; Lafourcade, B.; Leitão, P.J.; Münkemüller, T.; McClean, C.; Osborne, P.E.; Reineking, B.; Schröder, B.; Skidmore, A.K.; Zurell, D.; Lautenbach, S. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x
12. Lindner, T.; Puck, J.; Verbeke, A. Misconceptions about multicollinearity in international business research: Identification, consequences, and remedies. *Journal of International Business Studies* **2020**, *51*, 283–298. https://doi.org/10.1057/s41267-019-00257-1
13. Goldberger, A.S. *A Course in Econometrics*; Harvard University Press: Cambridge, MA, 1991.
14. Winship, C.; Western, B. Multicollinearity and model misspecification. *Sociological Science* **2016**, *3*, 627–649. https://doi.org/10.15195/v3.a27

15. Blanchard, O.J. [Vector Autoregressions and Reality]: Comment. *Journal of Business & Economic Statistics* **1987**, *5*, 449–451. https://doi.org/10.2307/1391994
16. Curtis, S.M.; Ghosh, S.K. A Bayesian approach to multicollinearity and the simultaneous selection and clustering of predictors in linear regression. *Journal of Statistical Theory and Practice* **2011**, *5*, 715–735. https://doi.org/10.1080/15598608.2011.10483741
17. Hill, R.C.; Adkins, L.C. Collinearity. In *A Companion to Theoretical Econometrics*; Baltagi, B.H., Ed.; Blackwell Publishing Ltd, 2003; pp. 256–278.
18. Kalnins, A. Multicollinearity: How common factors cause Type 1 errors in multivariate regression. *Strategic Management Journal* **2018**, *39*, 2362–2385. https://doi.org/10.1002/smj.2783
19. Schisterman, E.F.; Perkins, N.J.; Mumford, S.L.; Ahrens, K.A.; Mitchell, E.M. Collinearity and causal diagrams: A lesson on the importance of model specification. *Epidemiology* **2017**, *28*, 47–53. https://doi.org/10.1097/EDE.0000000000000554
20. Schwarz, C.; Schwarz, A.; Black, W.C. Examining the impact of multicollinearity in discovering higher-order factor models. *Communications of the Association for Information Systems* **2014**, *34*, Article 62. https://doi.org/10.17705/1CAIS.03463
21. Gunst, R.F.; Webster, J.T. Regression analysis and problems of multicollinearity. *Communications in Statistics* **1975**, *4*, 277–292. https://doi.org/10.1080/03610927308827246
22. Fox, J.; Monette, G. Generalized collinearity diagnostics. *Journal of the American Statistical Association* **1992**, *87*, 178–183. https://doi.org/10.1080/01621459.1992.10475190
23. Marquardt, D.W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. *Technometrics* **1970**, *12*, 591–612. https://doi.org/10.1080/00401706.1970.10488699
24. Smith, K.W.; Sasaki, M.S. Decreasing multicollinearity: A method for models with multiplicative functions. *Sociological Methods & Research* **1979**, *8*, 35–56. https://doi.org/10.1177/004912417900800102
25. Montgomery, D.C.; Peck, E.A. *Introduction to Linear Regression Analysis*; John Wiley & Sons: New York, NY, 1982.
26. Kidwell, J.S.; Brown, L.H. Ridge regression as a technique for analyzing models with multicollinearity. *Journal of Marriage and the Family* **1982**, *44*, 287–299. https://doi.org/10.2307/351539
27. Mason, C.H.; Perreault, W.D., Jr. Collinearity, power, and interpretation of multiple regression analysis. *Journal of Marketing Research* **1991**, *28*, 268–280. https://doi.org/10.1177/002224379102800302
28. Kroll, C.N.; Song, P. Impact of multicollinearity on small sample hydrologic regression models. *Water Resources Research* **2013**, *49*, 3756–3769. https://doi.org/10.1002/wrcr.20315
29. Askin, R.G. Multicollinearity in regression: Review and examples. *Journal of Forecasting* **1982**, *1*, 281–292. https://doi.org/10.1002/for.3980010307

30. Mundfrom, D.J.; DePoy Smith, M.; Kay, L.W. The effect of multicollinearity on prediction in regression models. *General Linear Model Journal* **2018**, *44*, 24–28. https://doi.org/10.31523/glmj.044001.003
31. Zidek, J.V.; Wong, H.; Le, N.D.; Burnett, R. Causality, measurement error and multicollinearity in epidemiology. *Environmetrics* **1996**, *7*, 441–451. https://doi.org/10.1002/(SICI)1099-095X(199607)7:4%3C441::AID-ENV226%3E3.0.CO;2-V
32. Gokmen, S.; Dagalp, R.; Kilickaplan, S. Multicollinearity in measurement error models. *Communications in Statistics - Theory and Methods* **2022**, *51*, 474–485. https://doi.org/10.1080/03610926.2020.1750654
33. Jurczyk, T. Outlier detection under multicollinearity. *Journal of Statistical Computation and Simulation* **2012**, *82*, 261–278. https://doi.org/10.1080/00949655.2011.638634
34. Dawoud, I.; Abonazel, M.R. Robust Dawoud–Kibria estimator for handling multicollinearity and outliers in the linear regression model. *Journal of Statistical Computation and Simulation* **2021**, *91*, 3678–3692. https://doi.org/10.1080/00949655.2021.1945063
35. Mansfield, E.R.; Helms, B.P. Detecting multicollinearity. *The American Statistician* **1982**, *36*, 158–160. https://doi.org/10.1080/00031305.1982.10482818
36. Klein, L.R. *An Introduction to Econometrics*; Prentice-Hall: Englewood Cliffs, NJ, 1962.
37. Willis, C.E.; Perlack, R.D. Multicollinearity: Effects, symptoms, and remedies. *Journal of the Northeastern Agricultural Economics Council* **1978**, *7*, 55–61. https://doi.org/10.1017/S0163548400001989
38. Midi, H.; Sarkar, S.K.; Rana, S. Collinearity diagnostics of binary logistic regression model. *Journal of Interdisciplinary Mathematics* **2010**, *13*, 253–267. https://doi.org/10.1080/09720502.2010.10700699
39. Sheather, S. *A Modern Approach to Regression with R*; Springer: New York, NY, 2009. https://doi.org/10.1007/978-0-387-09608-7
40. Lavery, M.R.; Acharya, P.; Sivo, S.A.; Xu, L. Number of predictors and multicollinearity: What are their effects on error and bias in regression? *Communications in Statistics - Simulation and Computation* **2019**, *48*, 27–38. https://doi.org/10.1080/03610918.2017.1371750
41. Vatcheva, K.P.; Lee, M.; McCormick, J.B.; Rahbar, M.H. Multicollinearity in regression analyses conducted in epidemiologic studies. *Epidemiology (Sunnyvale)* **2016**, *6*, Article 1000227. https://doi.org/10.4172/2161-1165.1000227
42. O'Brien, R.M. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* **2007**, *41*, 673–690. https://doi.org/10.1007/s11135-006-9018-6
43. Belsley, D.A. Assessing the presence of harmful collinearity and other forms of weak data through a test for signal-to-noise. *Journal of Econometrics* **1982**, *20*, 211–253. https://doi.org/10.1016/0304-4076(82)90020-3
44. Mohammadi, S. A test of harmful multicollinearity: A generalized ridge regression approach. *Communications in Statistics - Theory and Methods* **2022**, *51*, 724–743. https://doi.org/10.1080/03610926.2020.1754855
45. Chennamaneni, P.R.; Echambadi, R.; Hess, J.D.; Syam, N. Diagnosing harmful collinearity in moderated regressions: A roadmap. *International Journal of*

Research in Marketing **2016**, *33*, 172–182. https://doi.org/10.1016/j.ijresmar.2015.08.004
46. Hadi, A.S. Diagnosing collinearity-influential observations. *Computational Statistics & Data Analysis* **1988**, *7*, 143–159. https://doi.org/10.1016/0167-9473(88)90089-8
47. Sinan, A.; Alkan, B.B. A useful approach to identify the multicollinearity in the presence of outliers. *Journal of Applied Statistics* **2015**, *42*, 986–993. https://doi.org/10.1080/02664763.2014.993369
48. Wheeler, D.C. Diagnostic tools and a remedial method for collinearity in geographically weighted regression. *Environment and Planning A: Economy and Space* **2007**, *39*, 2464–2481. https://doi.org/10.1068/a38325
49. Wheeler, D.; Tiefelsdorf, M. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems* **2005**, *7*, 161–187. https://doi.org/10.1007/s10109-005-0155-6
50. Friendly, M.; Kwan, E. Where's Waldo? Visualizing collinearity diagnostics. *The American Statistician* **2009**, *63*, 56–65. https://doi.org/10.1198/tast.2009.0012
51. Johnston, J. *Econometric Methods*; McGraw-Hill, 1963.
52. Salmerón-Gómez, R.; García-García, C.B.; Rodríguez-Sánchez, A. Enlarging of the sample to address multicollinearity. *Computational Economics* **2026**, *67*, 1877–1899. https://doi.org/10.1007/s10614-025-10920-5
53. Cortina, J.M. Interaction, nonlinearity, and multicollinearity: Implications for multiple regression. *Journal of Management* **1993**, *19*, 915–922. https://doi.org/10.1177/014920639301900411
54. Heikkila, E. Multicollinearity in regression models with multiple distance measures. *Journal of Regional Science* **1988**, *28*, 345–362. https://doi.org/10.1111/j.1467-9787.1988.tb01087.x
55. Xi, W.-F.; Jiang, Q.-W.; Yang, A.-M. Using stepwise regression to address multicollinearity is not appropriate. *International Journal of Surgery* **2024**, *110*, 3122–3123. https://doi.org/10.1097/JS9.0000000000001208
56. Morrow-Howell, N. The *M* word: Multicollinearity in multiple regression. *Social Work Research* **1994**, *18*, 247–251. https://doi.org/10.1093/swr/18.4.247
57. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. https://doi.org/10.2307/1267351
58. Dar, I.S.; Chand, S.; Shabbir, M.; Kibria, B.M.G. Condition-index based new ridge regression estimator for linear regression model with multicollinearity. *Kuwait Journal of Science* **2023**, *50*, 91–96. https://doi.org/10.1016/j.kjs.2023.02.013
59. Morzuch, B.J.; Ruark, G.A. Principal components regression to mitigate the effects of multicollinearity. *Forest Science* **1991**, *37*, 191–199. https://doi.org/10.1093/forestscience/37.1.191
60. Park, S.H. Collinearity and optimal restrictions on regression parameters for estimating responses. *Technometrics* **1981**, *23*, 289–295. https://doi.org/10.2307/1267793
61. Aydın, D.; Yılmaz, E.; Chamidah, N.; Mohammad, S. Principal components regression based on kernel smoothing for semiparametric models with

multicollinearity. *Journal of Statistical Computation and Simulation* **2025**, *95*, 2088–2116. https://doi.org/10.1080/00949655.2025.2479658
62. Wold, S.; Ruhe, A.; Wold, H.; Dunn, W.J., III. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing* **1984**, *5*, 735–743. https://doi.org/10.1137/0905052
63. Geladi, P.; Kowalski, B.R. Partial least-squares regression: A tutorial. *Analytica Chimica Acta* **1986**, *185*, 1–17. https://doi.org/10.1016/0003-2670(86)80028-9
64. Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **1996**, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x
65. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **2005**, *67*, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x
66. Tomaschek, F.; Hendrix, P.; Baayen, R.H. Strategies for addressing collinearity in multivariate linguistic data. *Journal of Phonetics* **2018**, *71*, 249–267. https://doi.org/10.1016/j.wocn.2018.09.004
67. Liu, K. A new class of biased estimate in linear regression. *Communications in Statistics - Theory and Methods* **1993**, *22*, 393–402. https://doi.org/10.1080/03610929308831027
68. Kibria, B.M.G.; Lukman, A.F. A new ridge-type estimator for the linear regression model: Simulations and applications. *Scientifica* **2020**, Article 9758378. https://doi.org/10.1155/2020/9758378
69. Liu, K. Using Liu-type estimator to combat collinearity. *Communications in Statistics - Theory and Methods* **2003**, *32*, 1009–1020. https://doi.org/10.1081/STA-120019959
70. Dawoud, I.; Kibria, B.M.G. A new biased estimator to combat the multicollinearity of the Gaussian linear regression model. *Stats* **2020**, *3*, 526–541. https://doi.org/10.3390/stats3040033
71. Alharthi, M.F.; Akhtar, N. Newly improved two-parameter ridge estimators: A better approach for mitigating multicollinearity in regression analysis. *Axioms* **2025**, *14*, Article 186. https://doi.org/10.3390/axioms14030186
72. Adepoju, A.A.; Ojo, O.O. Bayesian method for solving the problem of multicollinearity in regression. *Afrika Statistika* **2018**, *13*, 1823–1834. https://doi.org/10.16929/as/1823.135
73. De Veaux, R.D.; Ungar, L.H. Multicollinearity: A tale of two nonparametric regressions. In *Selecting Models from Data: Artificial Intelligence and Statistics IV*; Cheeseman, P., Oldford, R.W., Eds.; Springer: New York, NY, 1994; pp. 393–402. https://doi.org/10.1007/978-1-4612-2660-4_40
74. Garg, A.; Tai, K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *International Journal of Modelling, Identification and Control* **2013**, *18*, 295–312. https://doi.org/10.1504/IJMIC.2013.053535
75. Shieh, G. Clarifying the role of mean centring in multicollinearity of interaction effects. *British Journal of Mathematical and Statistical Psychology* **2011**, *64*, 462–477. https://doi.org/10.1111/j.2044-8317.2010.02002.x

76. Echambadi, R.; Hess, J.D. Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science* **2007**, *26*, 438–445. https://doi.org/10.1287/mksc.1060.0263
77. Iacobucci, D.; Schneider, M.J.; Popovich, D.L.; Bakamitsos, G.A. Mean centering helps alleviate “micro” but not “macro” multicollinearity. *Behavior Research Methods* **2016**, *48*, 1308–1317. https://doi.org/10.3758/s13428-015-0624-x
78. McClelland, G.H.; Irwin, J.R.; Disatnik, D.; Sivan, L. Multicollinearity is a red herring in the search for moderator variables: A guide to interpreting moderated multiple regression models and a critique of Iacobucci, Schneider, Popovich, and Bakamitsos (2016). *Behavior Research Methods* **2017**, *49*, 394–402. https://doi.org/10.3758/s13428-016-0785-2
79. Grewal, R.; Cote, J.A.; Baumgartner, H. Multicollinearity and measurement error in structural equation models: Implications for theory testing. *Marketing Science* **2004**, *23*, 519–529. https://doi.org/10.1287/mksc.1040.0070
80. Kock, N.; Lynn, G.S. Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems* **2012**, *13*, 546–580. https://doi.org/10.17705/1jais.00302
81. Shieh, Y.-Y.; Fouladi, R.T. The effect of multicollinearity on multilevel modeling parameter estimates and standard errors. *Educational and Psychological Measurement* **2003**, *63*, 951–985. https://doi.org/10.1177/0013164403258402
82. Yu, H.; Jiang, S.; Land, K.C. Multicollinearity in hierarchical linear models. *Social Science Research* **2015**, *53*, 118–136. https://doi.org/10.1016/j.ssresearch.2015.04.008
83. Ali, E.; Lukman, A.F. Ridge-penalized Zero-Inflated Probit Bell model for multicollinearity in count data. *Journal of Applied Statistics* **2026**, *53*, 633–658. https://doi.org/10.1080/02664763.2025.2530551
84. Fotheringham, A.S.; Oshan, T.M. Geographically weighted regression and multicollinearity: Dispelling the myth. *Journal of Geographical Systems* **2016**, *18*, 303–329. https://doi.org/10.1007/s10109-016-0239-5
85. Chavez-Chong, C.O.; Hardouin, C.; Fermin, A.-K. Ridge regularization for spatial autoregressive models with multicollinearity issues. *AStA Advances in Statistical Analysis* **2025**, *109*, 25–52. https://doi.org/10.1007/s10182-024-00496-0
86. Arceneaux, K.; Huber, G.A. What to do (and not do) with multicollinearity in state politics research. *State Politics & Policy Quarterly* **2007**, *7*, 81–101. https://doi.org/10.1177/153244000700700105
87. York, R. Residualization is not the answer: Rethinking how to address multicollinearity. *Social Science Research* **2012**, *41*, 1379–1386. https://doi.org/10.1016/j.ssresearch.2012.05.014