Contemporary Mathematics
https://ojs.wiserpub.com/index.php/CM/

Research Article

Navigating Multicollinearity in Linear Regression Models: Implications for Big Data Analysis

Salomi du Plessis[1], Mohammad Arashi[1,2], Salomon Millard[1*], Gaonyalelwe Maribe[1]

[1] Department of Statistics, Faculty of Natural and Agricultural Science, University of Pretoria, Pretoria 0002, South Africa
[2] Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, P. O. Box 1159, Mashhad 91775, Iran
E-mail: sollie.millard@up.ac.za

Received: 18 August 2025; **Revised:** 29 September 2025; **Accepted:** 21 October 2025

Abstract: The consequences of multicollinearity in regression analysis involving small, moderate, or high-dimensional datasets are well-established, and many notable solutions exist. However, the consequences of multicollinearity when considering big data, specifically data with a large number of observations, are not well established. In this paper, we determine the impact of multicollinearity on the linear regression model when applied to big data by numerically evaluating the bias, variance, and signs of the estimated regression coefficients. An extensive simulation study shows that multicollinearity does not substantially alter the statistical measures under consideration. Our analysis is also applied to a real-world dataset for method demonstration.

***Keywords*:** big data, cross-validation, multicollinearity, multiple linear regression, sufficient statistics

MSC: 62J05, 62H20, 62F10, 65C60

1. Introduction

Since the development of computer technology that can collect, collate, and store enormous amounts of data, from a volume, velocity, and variety perspective, big data has emerged as a discipline in the new era of science [1]. Big data received significant attention, such that the American government provided $200 million toward big data research programs after the relevance of big data was debated at the World Economic Forum in 2012. Soon after, the term was incorporated into the Oxford English Dictionary [2].

Although the high variety and velocity aspects of big data require novel statistical methodology, the data considered in this paper speak to the high volume aspect of big data. Access to large datasets enables us to better understand the relationship between a response variable and a set of predictor variables, but we often face major issues during the analysis of such data. In the context of regression analysis, multicollinearity is perceived as one of these issues. The authors of [3] give a detailed summary of statistical techniques to mitigate the effects of multicollinearity in big data. Furthermore, the authors of [4, 5] proposed efficient methodologies for obtaining shrinkage estimators to be used in the presence of multicollinearity in big data. Multicollinearity refers to the existence of near-perfect linear relationships between the predictor variables in a regression problem. That is, some predictor variables in the regression model can be linearly predicted from other predictor variables in the regression model with a substantial degree of accuracy [6].

Copyright ©2026 Salomon Millard, et al.
DOI: https://doi.org/10.37256/cm.7320268305
This is an open-access article distributed under a CC BY license
(Creative Commons Attribution 4.0 International License)
https://creativecommons.org/licenses/by/4.0/

Consider the multiple linear regression model given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is the response variable, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ is the matrix of predictor variables with $\mathbf{x}_i \in \mathbb{R}^p$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)^\top$ is the vector of regression coefficients, and $\mathbf{e} = (e_1, \ldots, e_n)^\top$ is the error vector. One of the main assumptions of the multiple linear regression model is that there are no near-perfect linear relationships between the predictor variables, and violating this assumption has severe consequences in small or moderate-sized datasets. The two major consequences of multicollinearity in the linear regression model are difficulties in interpreting the regression coefficients as well as unstable regression coefficients [7]. Regression coefficients are interpreted as an estimate of the effect of a one-unit change in a predictor variable whilst keeping the remaining predictor variables constant. If two or more predictor variables have a near-perfect linear relationship, the changes in one of the predictor variables are not independent of changes in the predictor variables with which it is linearly related, and hence, the regression coefficient is an imprecise estimate of the effect of independent changes in the predictor variable. Furthermore, the standard errors of the regression coefficients will be inflated in the presence of multicollinearity. Multicollinearity is a phenomenon of samples, and thus the observed correlation structure can differ severely if multiple samples are collected [8]. As a result, the estimated regression coefficients may change erratically in response to small changes in the data, even resulting in changes in the signs of the estimated regression coefficients [9].

The consequences of multicollinearity in linear regression models have been thoroughly studied for small or moderate-sized datasets, and many remedies, such as removing highly correlated predictors, data transformations, and regularization techniques, exist. Interested readers may refer to the studies of Arashi et al. [10] on the theoretical development of shrinkage learners in the seemingly unrelated semiparametric model, Roozbeh [11] on optimal ridge estimation in restricted logistic regression models, Roozbeh et al. [12] considering a robust approach to the ridge estimator, mitigating the effect of outliers, and Chiang et al. [13] focussing on exact ridge regression for big data.

However, in an era where large data sets with many features that naturally exhibit multicollinearity are common, there is a need to establish whether multicollinearity is as problematic as in small or moderate-sized datasets and hence whether it is still necessary to consider the traditional solutions to multicollinearity. The goal of this paper therefore is to determine whether the problems that arise due to multicollinearity in small or moderate-sized datasets are also prevalent when we consider big data, where big data refers to data with a large number of observations. We will consider the bias, variance, and signs of the estimated regression coefficients of the multiple linear regression model in an extensive simulation study. To the best of our knowledge, no literature with the same goal is available. This paper is structured as follows: Section 2 contains some computational considerations for implementing the simulation and application, section 3 provides the details and results of the simulation study, section 4 provides the results of a real-world application, and section 5 concludes the paper.

2. Computational considerations

Despite the ease with which the linear regression model can be estimated for small or moderate-sized datasets, we encounter challenges in the big data domain. Some of these challenges are that the high volume of data frequently exceeds the storage capacity of a single computer and that the time required to obtain results becomes infeasible due to the computational burden of such a high volume of data. To address these big data challenges, the authors of [5] proposed an algorithm that uses a series of sufficient statistics to obtain the ordinary least squares estimates of the multiple linear regression model and perform model validation by means of K-fold cross-validation. The sufficient statistics array given in (2) is used to obtain the estimator of the regression coefficients and the covariance matrix of the regression coefficients in the multiple linear regression model.

$$\mathbf{A} := \sum_{i=1}^{n} \mathbf{A}_i = \begin{bmatrix} S_{yy} & \mathbf{S}_{xy}^{\top} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} y_i \mathbf{x}_i^{\top} \\ \sum_{i=1}^{n} \mathbf{x}_i y_i & \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^{\top} \end{bmatrix}. \tag{2}$$

The sufficient statistic array can be efficiently updated at the row or batch level. Using the sufficient statistic array, the ordinary least squares estimator of the regression coefficients for the linear regression model given in (1) is

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}$$

whereas the covariance matrix of the regression coefficients is given by

$$V(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2 \mathbf{S}_{xx}^{-1}$$

with $\hat{\sigma}^2 = \left(S_{yy} - \mathbf{S}_{xy}^{\top} \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy}\right) / (n-p)$. This methodology also allows for performing K-fold cross-validation with reduced computational complexity compared to traditional methods [5]. Suppose we partition the dataset into K blocks containing n_k observations. The sufficient statistics array of each block is then given by $\sum_{i=1}^{n_k} \mathbf{A}_i$ such that the sufficient statistics array of the complete dataset is

$$\mathbf{A} := \sum_{k=1}^{K} \mathbf{A}_{(k)} := \sum_{k=1}^{K} \sum_{i=1}^{n_k} \mathbf{A}_i.$$

The sufficient statistics array for any training dataset consists of $(K-1)$ of the K blocks. We will elaborate on the use of K-fold cross-validation in section 4.

3. Simulation study

In this section, we will determine whether the consequences of multicollinearity that arise when using small or moderate-sized datasets remain problematic when the multiple linear regression model is applied to highly correlated big data. We will consider the effect of multicollinearity on the bias, variance, and signs of the estimated regression coefficients.

For the sake of completeness, consider again the multiple linear regression model given by $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ with $\mathbf{y}$ the response variable, $\mathbf{X}$ the matrix of predictor variables, $\boldsymbol{\beta}$ the vector of regression coefficients, and $\mathbf{e}$ the error vector. The predictor variables were generated using

$$x_{ij} = \left(1 - \gamma^2\right)^{1/2} z_{ij} + \gamma z_{i(p+1)},$$

where $i \in (1, \ldots, n)$, $j \in (1, \ldots, p)$ and z_{ij} independent standard normal pseudo-random numbers. The correlation between any two predictor variables is given by γ^2. The simulation study considers $p = 100$, $\gamma \in (0.9,\ 0.95,\ 0.99,\ 0.999)$ and $n \in (1{,}000,\ 10{,}000,\ 100{,}000,\ 1{,}000{,}000,\ 10{,}000{,}000)$. Furthermore, we consider some plausible symmetric, asymmetric, and heavy-tailed distributions that one may face in real data analysis. These include regression coefficients simulated from

the Gumbel, uniform, autoregressive, and a mixture of Gaussian distributions. An illustration of the distribution of the 4 sets of regression coefficients is provided in Figure 1.

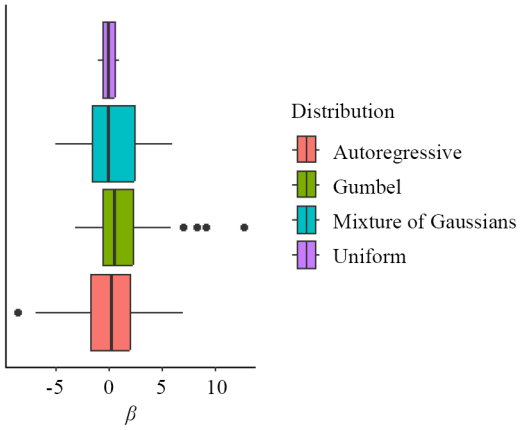


Figure 1. Distribution of the 4 sets of regression coefficients included in the simulation

The simulation was repeated 1,000 times for every $(n,\ \gamma,\ \beta)$ combination. The average condition index of $\mathbf{X}^\top\mathbf{X}$ for every combination of $(n,\ \gamma)$ is given in Table 1. As a general rule, a condition index between 10 and 30 indicates moderate multicollinearity, whereas a condition index greater than 30 indicates strong multicollinearity [14].

Table 1. Average condition index of $\mathbf{X}^\top\mathbf{X}$ for every $(n,\ \gamma)$ combination

Sample size	$\gamma = 0.9$	$\gamma = 0.95$	$\gamma = 0.99$	$\gamma = 0.999$
1,000	29.847	43.898	101.150	322.122
10,000	22.885	33.686	77.702	247.316
100,000	21.323	31.401	72.401	230.448
1,000,000	20.873	30.737	70.868	225.614
10,000,000	20.735	30.533	70.400	224.123

For each $(n,\ \gamma,\ \beta)$ combination, the bias, variance, and sign of β_j are recorded at every iteration. We will report on scaled versions of bias and variance in order to give these values relative to the value of β_j. The scaled bias and variance of β_j are calculated using $bias_s\left(\hat{\beta}_j\right) = \dfrac{|bias(\hat{\beta}_j)|}{|\beta_j|}$ and $var_s\left(\hat{\beta}_j\right) = \dfrac{var(\hat{\beta}_j)}{|\beta_j|}$, respectively. Hence, the scaled bias vector is given by $bias_s\left(\hat{\boldsymbol{\beta}}\right) = \left(bias_s\left(\hat{\beta}_0\right),\ bias_s\left(\hat{\beta}_1\right),\ \ldots,\ bias_s\left(\hat{\beta}_p\right)\right)$ whereas the scaled variance vector is given by $var_s\left(\hat{\boldsymbol{\beta}}\right) = \left(var_s\left(\hat{\beta}_0\right),\ var_s\left(\hat{\beta}_1\right),\ \ldots,\ var_s\left(\hat{\beta}_p\right)\right)$. Tables 2 and 3 provide the minimum, maximum, and median of the scaled bias and variance vectors for each combination $(n,\ \gamma,\ \beta)$.

3.1 *Bias evaluation*

As with small or moderate-sized datasets, we expect the regression coefficient to be unbiased. The results on the scaled biases given in Table 2 confirm that the scaled biases are close to zero and decrease as the sample size increases. In certain instances, the maximum of the scaled biases may appear large. This occurs when the true value of β_j is very close to zero, and as a result, the denominator of the scaled bias is close to zero. However, these multiples of values that are very close to zero are also very small.

Table 2. Scaled bias of the estimated regression coefficients for the various (n, γ, β) scenarios

Distribution of β	Statistic	$n = 1,000$	$n = 10,000$	$n = 100,000$	$n = 1,000,000$	$n = 10,000,000$
			$\gamma = 0.9$			
Gumbel	Min	0.0000271	0.0000082	0.0000002	0.0000013	0.0000003
	Median	0.0011412	0.0003140	0.0000968	0.0000390	0.0000113
	Max	0.0581371	0.0138212	0.0018336	0.0017207	0.0003976
Mixture of Gaussians	Min	0.0000106	0.0000055	0.0000009	0.0000001	0.0000001
	Median	0.0012029	0.0002903	0.0001138	0.0000392	0.0000081
	Max	0.2205165	0.0568773	0.0131850	0.0064995	0.0013526
Uniform	Min	0.0001070	0.0000189	0.0000164	0.0000118	0.0000001
	Median	0.0079542	0.0023969	0.0006617	0.0002464	0.0000666
	Max	1.3139785	0.2192294	0.0658390	0.0238750	0.0128006
Autoregressive	Min	0.0000132	0.0000030	0.0000018	0.0000000	0.0000001
	Median	0.0008353	0.0001834	0.0000665	0.0000240	0.0000094
	Max	0.0281529	0.0150646	0.0087760	0.0043744	0.0008047
			$\gamma = 0.95$			
Gumbel	Min	0.0000079	0.0000008	0.0000013	0.0000008	0.0000001
	Median	0.0015956	0.0004192	0.0001695	0.0000536	0.0000113
	Max	0.0329302	0.0143048	0.0033073	0.0018097	0.0006204
Mixture of Gaussians	Min	0.0000181	0.0000233	0.0000021	0.0000014	0.0000014
	Median	0.0013839	0.0004809	0.0001493	0.0000433	0.0000136
	Max	0.1578881	0.1279908	0.0289075	0.0042428	0.0049535
Uniform	Min	0.0000591	0.0000614	0.0000618	0.0000019	0.0000023
	Median	0.0107481	0.0028807	0.0011103	0.0003627	0.0001061
	Max	0.9414992	0.1662014	0.0306548	0.0579284	0.0057851
Autoregressive	Min	0.0000894	0.0000127	0.0000002	0.0000003	0.0000001
	Median	0.0018491	0.0003410	0.0001435	0.0000311	0.0000101
	Max	0.0831200	0.0093326	0.0067265	0.0075282	0.0024724
			$\gamma = 0.99$			
Gumbel	Min	0.0000369	0.0000029	0.0000052	0.0000022	0.0000002
	Median	0.0036256	0.0011230	0.0003267	0.0000965	0.0000331
	Max	0.1406935	0.0623205	0.0085779	0.0023633	0.0011666
Mixture of Gaussians	Min	0.0000171	0.0000058	0.0000162	0.0000009	0.0000001
	Median	0.0033637	0.0012162	0.0003125	0.0001140	0.0000300
	Max	0.6639392	0.3920458	0.0603905	0.0250671	0.0046959
Uniform	Min	0.0005965	0.0004114	0.0000220	0.0000198	0.0000009
	Median	0.0192090	0.0077398	0.0020234	0.0007101	0.0002169
	Max	4.8319850	1.6475852	0.3267947	0.1033133	0.0373694
Autoregressive	Min	0.0000050	0.0000091	0.0000096	0.0000002	0.0000002
	Median	0.0026418	0.0007858	0.0002350	0.0000812	0.0000370
	Max	0.3875684	0.0317697	0.0071282	0.0068368	0.0026723
			$\gamma = 0.999$			
Gumbel	Min	0.0002609	0.0000163	0.0000075	0.0000062	0.0000001
	Median	0.0136633	0.0023543	0.0015197	0.0003597	0.0001006
	Max	0.4747464	0.1197787	0.0696467	0.0052329	0.0025570
Mixture of Gaussians	Min	0.0000149	0.0000642	0.0000049	0.0000009	0.0000004
	Median	0.0095786	0.0030228	0.0009610	0.0003225	0.0001018
	Max	2.7361390	0.6543861	0.1978810	0.1145064	0.0155506
Uniform	Min	0.0003523	0.0001212	0.0002090	0.0000547	0.0000001
	Median	0.0769363	0.0269822	0.0074643	0.0025183	0.0007022
	Max	14.1286932	1.6787861	1.7381428	0.6960135	0.0485963
Autoregressive	Min	0.0002164	0.0000497	0.0000028	0.0000011	0.0000004
	Median	0.0094705	0.0026498	0.0006719	0.0003131	0.0000798
	Max	0.2238803	0.0966304	0.0215575	0.0114785	0.0016246

3.2 *Variance evaluation*

When considering the minimum, median, and maximum of the scaled variance vectors in Table 3, higher levels of multicollinearity result in estimated regression coefficients with larger variances. However, the scaled variances become extremely small when we start to consider $1,000,000$ or more observations. So small that the variances, although inflated, do not lead to unstable regression coefficients.

Table 3. Scaled variance of the estimated regression coefficients for the various $(n,\ \gamma,\ \beta)$ scenarios

Distribution of β	Statistic	$n = 1,000$	$n = 10,000$	$n = 100,000$	$n = 1,000,000$	$n = 10,000,000$
			$\gamma = 0.9$			
Gumbel	Min	0.0051742	0.0004676	0.0000460	0.0000045	0.0000005
	Median	0.0058208	0.0005269	0.0000525	0.0000052	0.0000005
	Max	0.0063673	0.0005700	0.0000584	0.0000058	0.0000006
Mixture of Gaussians	Min	0.0051335	0.0004705	0.0000455	0.0000046	0.0000005
	Median	0.0057875	0.0005272	0.0000522	0.0000052	0.0000005
	Max	0.0064695	0.0005863	0.0000575	0.0000059	0.0000006
Uniform	Min	0.0050850	0.0004559	0.0000469	0.0000047	0.0000005
	Median	0.0058503	0.0005287	0.0000525	0.0000052	0.0000005
	Max	0.0064818	0.0005713	0.0000577	0.0000057	0.0000006
Autoregressive	Min	0.0053126	0.0004666	0.0000462	0.0000047	0.0000005
	Median	0.0058040	0.0005286	0.0000525	0.0000052	0.0000005
	Max	0.0063494	0.0005823	0.0000581	0.0000059	0.0000006
			$\gamma = 0.95$			
Gumbel	Min	0.0102474	0.0008908	0.0000923	0.0000092	0.0000009
	Median	0.0111626	0.0010200	0.0001029	0.0000102	0.0000010
	Max	0.0125349	0.0011574	0.0001153	0.0000113	0.0000012
Mixture of Gaussians	Min	0.0101502	0.0008888	0.0000901	0.0000090	0.0000009
	Median	0.0112926	0.0010109	0.0001014	0.0000102	0.0000010
	Max	0.0126913	0.0011744	0.0001136	0.0000111	0.0000011
Uniform	Min	0.0103272	0.0009293	0.0000911	0.0000089	0.0000009
	Median	0.0112907	0.0010278	0.0001012	0.0000103	0.0000010
	Max	0.0130797	0.0011571	0.0001114	0.0000119	0.0000011
Autoregressive	Min	0.0102301	0.0009158	0.0000913	0.0000089	0.0000009
	Median	0.0112115	0.0010322	0.0001019	0.0000101	0.0000010
	Max	0.0128931	0.0011948	0.0001164	0.0000118	0.0000011
			$\gamma = 0.99$			
Gumbel	Min	0.0488907	0.0045185	0.0004493	0.0000435	0.0000045
	Median	0.0554186	0.0050382	0.0004988	0.0000494	0.0000050
	Max	0.0622955	0.0056381	0.0005481	0.0000558	0.0000055
Mixture of Gaussians	Min	0.0497655	0.0042898	0.0004401	0.0000435	0.0000045
	Median	0.0546515	0.0050351	0.0004966	0.0000497	0.0000050
	Max	0.0605330	0.0056473	0.0005689	0.0000550	0.0000056
Uniform	Min	0.0478949	0.0045236	0.0004382	0.0000432	0.0000042
	Median	0.0549684	0.0049948	0.0004984	0.0000495	0.0000050
	Max	0.0628675	0.0056345	0.0005546	0.0000560	0.0000054
Autoregressive	Min	0.0493462	0.0045342	0.0004241	0.0000439	0.0000044
	Median	0.0556207	0.0050204	0.0005003	0.0000497	0.0000050
	Max	0.0620436	0.0058613	0.0005481	0.0000555	0.0000055
			$\gamma = 0.999$			
Gumbel	Min	0.5006129	0.0439428	0.0042373	0.0004432	0.0000434
	Median	0.5576276	0.0496719	0.0049172	0.0004934	0.0000498
	Max	0.6140826	0.0542087	0.0054517	0.0005591	0.0000577
Mixture of Gaussians	Min	0.4756353	0.0449244	0.0044084	0.0004334	0.0000439
	Median	0.5502615	0.0495016	0.0049450	0.0004932	0.0000495
	Max	0.6044623	0.0541005	0.0055234	0.0005399	0.0000539
Uniform	Min	0.4894209	0.0440284	0.0044447	0.0004140	0.0000438
	Median	0.5584280	0.0498676	0.0049218	0.0004955	0.0000493
	Max	0.6177261	0.0565172	0.0055751	0.0005627	0.0000541
Autoregressive	Min	0.4905192	0.0441435	0.0045184	0.0004325	0.0000445
	Median	0.5527114	0.0503074	0.0049754	0.0005007	0.0000489
	Max	0.6452749	0.0576867	0.0056224	0.0005758	0.0000552

3.3 *Sign evaluation*

Lastly, we want to determine whether the signs of the regression coefficients change throughout each iteration of a simulation for a $(n,\ \gamma,\ \beta)$ combination. The number of times that each regression coefficient is estimated as a positive value is calculated to determine whether multicollinearity in large datasets results in regression coefficients with volatile sign changes. To visualize these results, the regression coefficients of each $\boldsymbol{\beta}$ scenario were divided into 10 groups, each containing approximately the same number of regression coefficients. The number of times that the regression coefficients within a group are estimated as positive values is then provided in a stacked barplot where the various colors represent the level of multicollinearity. The black dot at the top of each bar represents the total number of positive regression coefficients that would have been obtained if the signs of the regression coefficients were estimated correctly at every iteration of the simulation study. These results are given in Figures 2 and 3. When the sample size is $1,000$ the regression coefficients with negative values are often estimated as positive values. Similarly, those regression coefficients with positive values are often estimated as negative values. As the sample size increases, we see fewer regression coefficients that are estimated with incorrect signs. When the sample contains $10,000,000$ observations, we witness only a few sign changes at the highest level of multicollinearity and at intervals for which β_j is very close to zero, in these scenarios between -0.1 and 0.1. We need to consider whether a regression coefficient that is almost 0 has practical significance. If it does not, then estimating those regression coefficients with an incorrect sign on a few occasions is not problematic and does not justify considering methods, such as shrinkage estimators, for addressing multicollinearity, which comes at a cost, especially in the big data domain.

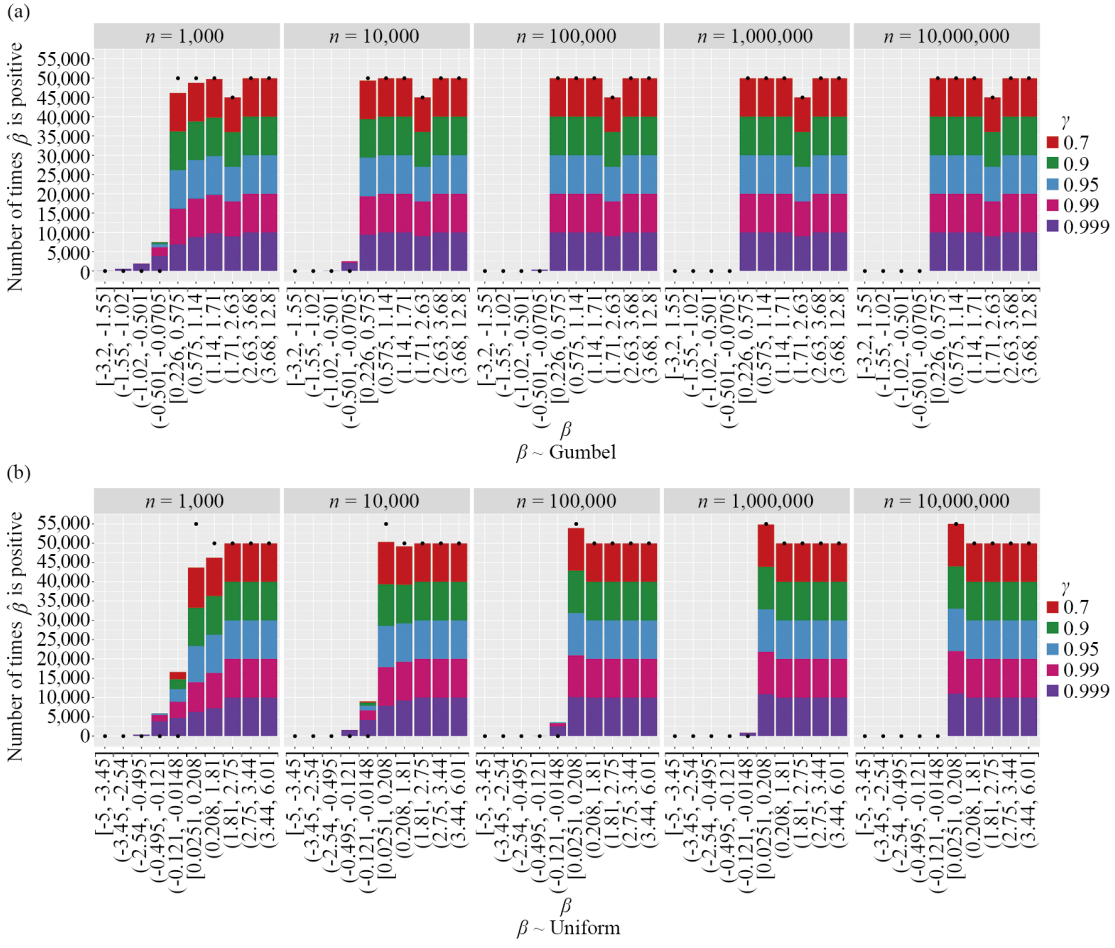


Figure 2. Number of times that the regression coefficients are estimated as a positive value for the various $(n,\ \gamma,\ \beta)$ scenarios

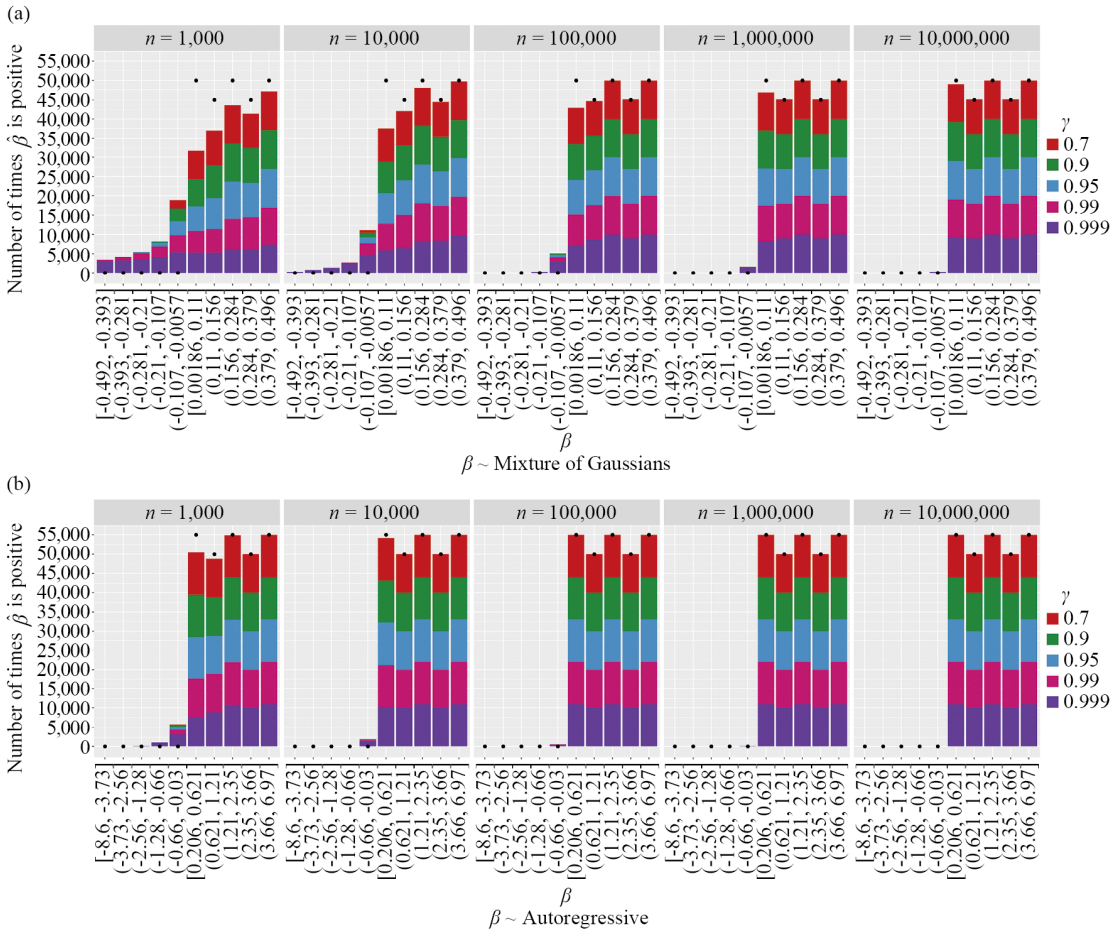


Figure 3. Number of times that the regression coefficients are estimated as a positive value for the various $(n,\ \gamma,\ \beta)$ scenarios

3.4 *Simulation results summary*

Considering the specific simulation setup and specifically bias, variance and sign sensitivity for the different sample sizes, it is apparent that the impact of multicollinearity becomes negligible when the sample size is $1,000,000$ or larger.

4. Application

The dataset used in this application was obtained from the Bureau of Transportation Statistics, which provides information on U.S. transportation systems. The response variable is the arrival delay in minutes. Variables such as the origin and destination airports, the time at which flights are scheduled to arrive and depart, the departure delay, the time and distance from the origin to the destination airport, and the date and time of these flights for major air carriers were used as the predictor variables. We've selected air traffic data from 2010 to 2020, resulting in $\pm\ 6,000,000$ observations and 90 predictor variables. The condition index of $1,640,313$ indicates an extremely high level of multicollinearity as we would expect due to the nature of correlated features such as the scheduled departure and arrival time as well as the time and distance between the origin and departure airports. The application aims to determine the impact of multicollinearity on the variance and signs of the estimated regression coefficients using subsets of the complete data. To capture the variability in samples obtained from a population for which the predictor variables exhibit high multicollinearity, random samples of size $n = 4,000,000$ were drawn from the complete dataset. Thereafter, the Ordinary Least Squares (OLS) estimates of the

model given in (1) were calculated for each sample. This process was repeated 1,000 times. Since the OLS estimates are unbiased in the presence of multicollinearity and we do not have $\boldsymbol{\beta}$, Figure 4 provides $E(\hat{\boldsymbol{\beta}}_j)$ for each regression coefficient.

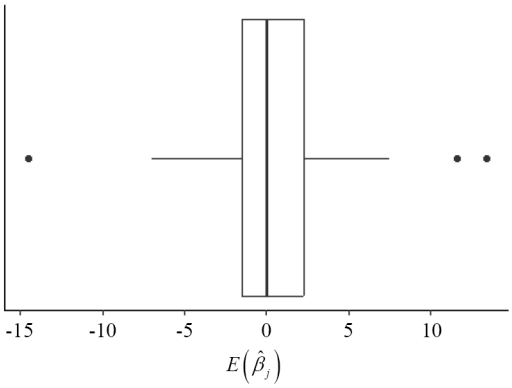


Figure 4. Distribution of $E(\hat{\boldsymbol{\beta}})$

4.1 *Variance evaluation*

To determine the impact of multicollinearity on the magnitude of the variances of the estimated regression coefficients, the scaled variance of each β_j was calculated as $var_s\left(\hat{\beta}_j\right) = \dfrac{var(\hat{\beta}_j)}{|E(\hat{\beta}_j)|}$ such that the scaled variance vector is $var_s\left(\hat{\boldsymbol{\beta}}\right) = \left(var_s\left(\hat{\beta}_0\right), var_s\left(\hat{\beta}_1\right), \ldots, var_s\left(\hat{\beta}_p\right)\right)$. The variance of each regression coefficient is less than 0.5% of its expected value, with the minimum, median and maximum of $var_s(\hat{\boldsymbol{\beta}})$ respectively given by 8×10^{-8}, 2×10^{-4} and 4×10^{-3}.

4.2 *Sign evaluation*

We also consider the number of times that each regression coefficient is estimated as a positive value to determine whether multicollinearity results in regression coefficients with volatile sign changes. Only three of the OLS estimates, with expected values close to zero, changed signs throughout the 1,000 iterations. The expected values of these three regression coefficients, as well as the number of times that they were estimated as positive values, are given in Table 4.

Table 4. OLS estimates with changing signs

$E(\hat{\beta}_j)$	Number of times that $\hat{\beta}_j$ is positive
−0.033	8
−0.019	21
0.037	984

Suppose we want confirmation that the OLS estimates of a different real-world dataset that exhibits high multicollinearity will not be unstable. The K-fold cross-validation as described in section 2 could be utilized to obtain the sampling distribution of the regression coefficients without substantially increasing the computational burden.

5. Discussion

It is widely accepted that the estimation of regression coefficients can be plagued by significant difficulties when multicollinearity is present in small or moderate-sized datasets, resulting in inaccurate estimation. However, research on the impact of multicollinearity when considering large volumes of data is limited. As such, we deemed it necessary to determine the practical significance of multicollinearity in the context of big data analytics and provide insights into how multicollinearity should be addressed within this context.

The purpose of this research was to determine the impact of multicollinearity on the estimated regression coefficients of the linear regression model when applied to large volumes of data. This was achieved by computationally evaluating the bias, variance, and signs of the estimated regression coefficients. The results of the simulation and application showed that the estimated regression coefficients are unbiased, have small variances, and have consistent signs. Therefore, the regression coefficients are stable and as a result we have shown through a numerical approach that the presence of multicollinearity does not have a significant impact on the estimates of the linear regression model when considering big data that exhibits high multicollinearity. As such, within the context of the scenarios considered in this paper, specifically the sample size, number of predictor variables and levels of multicollinearity, it is not necessary to utilise the computationally expensive estimation strategies that were developed to deal with multicollinearity in small or moderate-sized datasets.

Acknowledgments

This work was based upon research supported in part by the National Research Foundation (NRF) of South Africa (SA), grant RA211204653274, nr 151035. The opinions expressed and conclusions arrived at are those of the authors and are not necessarily to be attributed to the NRF. Mohammad Arashi's work is based on research supported in part by the London Mathematical Society (LMS), grant reference: MA-2425-29.

Conflict of interest

The authors declare no conflict of interests.

References

[1] Wang C, Chen MH, Schifano E, Wu J, Yan J. Statistical methods and computing for big data. *Statistics and Its Interface*. 2016; 9(4): 399–414. Available from: https://doi.org/10.4310/SII.2016.v9.n4.a1.

[2] Diebold FX. What's the big idea? "Big Data" and its origins. *Significance*. 2021; 18(1): 36–37. Available from: https://doi.org/10.1111/1740-9713.01490.

[3] Chan JYL, Leow SMH, Bea KT, Cheng WK, Phoong SW, Hong ZW, et al. Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*. 2022; 10(8): 1283. Available from: https://doi.org/10.3390/math10081283.

[4] Zhang T, Yang B. An exact approach to ridge regression for big data. *Computational Statistics*. 2017; 32(3): 909–928. Available from: https://doi.org/10.1007/s00180-017-0731-5.

[5] Du Plessis S, Arashi M, Maribe G, Millard SM. Efficient estimation and validation of shrinkage estimators in big data analytics. *Mathematics*. 2023; 11(22): 4632. Available from: https://doi.org/10.3390/math11224632.

[6] Saleh AME, Arashi M, Kibria BG. *Theory of Ridge Regression Estimation with Applications*. Hoboken, NJ, USA: John Wiley & Sons; 2019.

[7] Alin A. Multicollinearity. *WIREs Computational Statistics*. 2010; 2(3): 370–374. Available from: https://doi.org/10.1002/wics.84.

[8] Coetsee J, Arashi M, Bekker A, Millard SM. Preliminary testing of the Cobb-Douglas production function and related inferential issues. *Communications in Statistics-Simulation and Computation*. 2017; 46(6): 469–483. Available from: https://doi.org/10.1080/03610918.2014.968724.

[9] Goldstein R. Conditioning diagnostics: Collinearity and weak data in regression. *Technometrics*. 1993; 35(1): 85–86. Available from: https://doi.org/10.1080/00401706.1993.10484997.

[10] Arashi M, Roozbeh M, Amini M. Theoretical development of shrinkage learners in the seemingly unrelated semiparametric model. *Journal of Statistical Research*. 2025; 59(1): 131–143. Available from: https://doi.org/10.3329/jsr.v59i1.83690.

[11] Roozbeh M. Optimal ridge estimation in the restricted logistic semiparametric regression models using generalized cross-validation. *Journal of Applied Statistics*. 2025: 1–20. Available from: https://doi.org/10.1080/02664763.2025.2541252.

[12] Roozbeh M, Maanavi M, Mohamed NA. A robust counterpart approach for the ridge estimator to tackle outlier effect in restricted multicollinear regression models. *Journal of Statistical Computation and Simulation*. 2023; 94(2): 279–296. Available from: https://doi.org/10.1080/00949655.2023.2243361.

[13] Chiang W, Liu X, Zhang T, Yang B. A study of exact ridge regression for big data. In: *2018 IEEE International Conference on Big Data (Big Data)*. Seattle, WA, USA: IEEE; 2018. p.3821–3830.

[14] Kim JH. Multicollinearity and misleading statistical results. *Korean Journal of Anesthesiology*. 2019; 72(6): 558–569. Available from: https://doi.org/10.4097/kja.19087.