



The Impact of Bootstrap on PCR and PLS Predictive Accuracy under Multicollinearity and Small Sample Sizes: A Simulation Study

¹ Nawal Mohammed Othman Mustafa, ²Ali Meelad Abd Allateef
nawal.mustafa@uob.edu.ly¹, ali.abdallateef@uob.edu.ly²

^{1,2} Faculty of Arts and Sciences – Al Marj, University of Benghazi – Libya

Received: 29/6/2026 Revised: 20/7/2026 Accepted: 28/7/2026 Publication: 5/8/2026

Abstract

In this paper, we study how resampling affects the predictive accuracy of principal component regression (PCR) and partial least squares (PLS) models in the presence of small samples and multicollinearity. Monte Carlo simulations were carried out with 500 iterations by drawing 500 bootstrap samples per iteration. For the analysis, three different sample sizes ($n = 30, 50, 100$), two values for the predictive variables ($p = 5, 10$) and three degrees of multicollinearity ($r = 0.80, 0.90, 0.99$) were considered while comparing the following techniques: standard PCR, bootstrap-PCR, standard PLS, and bootstrap-PLS. The performance of these techniques was further assessed based on the average mean squared error (AMSE) and mean variance (AVAR). According to the simulation results, it can be summarized that resample testing is advantageous both for PCR and PLS in statistical estimators. It is important to mention that the improvement is more remarkable as multicollinearity becomes stronger and the sample sizes (n) get smaller. The Bootstrap-PLS approach had the best performance over all other estimators.

Keywords:

Bootstrap resampling; principal component regression (PCR); partial least squares regression (PLS); multicollinearity; small sample size.

المخلص

في هذا البحث ندرس تأثير تقنية بوتستراب على أداء كل من انحدار المكونات الرئيسية (PCR) وانحدار المربعات الصغرى الجزئية (PLS) في وجود مشكلة التعدد الخطي وصغر حجم العينة، تم اعتماد المحاكاة في المقارنة، والتحقق من أداء تم استخدام عدة مقاييس من أبرزها متوسط مربع الخطأ (MSE) ومتوسط التباين (AVAR). ولقد توصلنا من خلال نتائج أن تطبيق بوتستراب على كل من نموذجي PCR و PLS يؤدي إلى تحسين شامل في جودة التقدير، ومن ملاحظة نتائج تبين إن انحدار المربعات الصغرى الجزئية المحسّن تفوق على جميع النماذج الأخرى.

1-Introduction

Multiple linear regression is also an extremely popular method of analysis for understanding the relationship among variables and predicting their future values. However, this approach is not free of certain limitations due to statistical issues that result in loss of precision and instability in the regression coefficients. The problem of (near-)perfect multicollinearity among predictive variables is particularly salient, because the phenomenon results in inflated variance and instability of coefficient estimates. These issues are magnified with small samples and hence they are extremely sensitive to whatever influence that data sources are under. This is difficult for much field research because large samples are logistically difficult to access (just as they are in medical research, for example). As a solution to the problem, researchers have resorted to dimensionality reduction techniques such as principal component regression (PCR) and partial least squares regression (PLS). Nevertheless, small sample sizes have detrimental effects on these tools and there is a pressing need for methods that can enhance their performances. Bootstrap is an important technique in this area and has been shown to improve the estimation accuracy and reduce the variability of predictions. Hence, we propose to do an all-inclusive comparison of the four models: PCR, Bootstrap-PCR, PLS and Bootstrap-PLS via simulated data in this work

1.1 Multiple Linear Regression Models

A multiple regression model is a model that relates a dependent variable (also called a response variable) to two or more independent variables (also called predictor variables). The general multiple linear regression model is as follows:

$$Y_i = B_0 + B_1X_{i1} + B_2X_{i2} + \dots + B_{p-1}X_{ip-1} + \varepsilon$$

where:

Y_i : is the value of the response variable in the i th trial.

$X_{i1}, X_{i2}, \dots, X_{ip-1}$: are the values of the $p - 1$ predictor variables in the i th trial.

$B_0, B_1, \dots, B_{p-1}$ are the regression coefficients.

ε_i is the error term in the i th trial.

The model is called a multiple linear regression model because it is a linear function of the regression coefficients $B_0, B_1, \dots, B_{p-1}$. The model is called a first-order model with $p - 1$ predictor variables.

1.2Multicollinearity

Multicollinearity refers to a situation where a few or all the predictor variables are markedly correlated with each other. Under such a situation, regression coefficients cannot be estimated with precision and the standard deviations of the estimated regression coefficients are

very high. This means that the estimation of each regression coefficient is not exact and the possible influence of predictor variables can hardly be disentangled.

1.3 Principal Component Regression (PCR).

Principal Component Regression "An alternative method for dealing with multicollinearity is principal component regression (PCR). Instead of using the original regressor variables $x_1, x_2, \dots, x_k$ in the regression model, we can use the principal components $z_1, z_2, \dots, z_k$. The principal components are orthogonal linear combinations of the original regressors. In principal component regression, we fit a standard linear regression model using a subset of m principal components ($m < k$) as the regressors:

$$Y = \alpha_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \dots + \alpha_m Z_m + \varepsilon$$

Because the principal components are uncorrelated, the least squares estimator of the regression coefficients α has desirable properties, and the severe instability caused by multicollinearity among the original regressors is avoided. Typically, components associated with the smallest eigenvalues are omitted, retaining only those m components that account for the vast majority of the total variance in the regressor variables."

1.4 Partial Least Squares Regression (PLS)

In Partial Least Squares (PLS) regression, the explanatory variables are converted into a new set of finite, uncorrelated components. Then, regression analysis is applied to these components rather than to the original explanatory variables. This is the case, for example, when the explanatory variables are strongly correlated or when there are more explanatory variables than observations, in which case Ordinary Least Squares (OLS) regression may deliver coefficient estimates whose standard errors are extremely large, or it may not even be possible to perform the calculations. In contrast to multiple regression, the PLS algorithm does not presume the stability of the explanatory variables and, therefore, is capable of handling errors due to the measurement of explanatory variables; hence, the method is more robust against measurement errors and uncertainties (Wold, Sjöström, & Eriksson, 2001).

Mathematical formulation

The PLS model is a joint analysis of the matrix of predictors and the vector of responses. The model can be represented mathematically as:

$$y = xb + \varepsilon$$

y : – is the response (dependent) variable vector $n \times 1$

T : – is the matrix of latent score components extracted from the predictor matrix $n \times k$

q : – is the vector of regression coefficients (loadings) associated with the latent components $k \times 1$

E : – is the vector of random errors

n denotes the sample size

k is the number of retained latent components, where $K \leq \min(n, p)$

Where N denotes the sample size, and A is the number of retained latent components. The latent score matrix is obtained as

$$T = XW$$

The end regression model can be expressed in terms of the original independent variables as.:

$$\hat{B}_{PLS} = W(P^T W)^{-1} Q^T$$

1.5 Bootstrap Resampling Framework

Bootstrap is a flexible, non-parametric resampling method—that is, it makes no prior assumptions about the form of the probability distribution—which considers the observed data as representative of the population and samples from it with replacement to estimate the sampling distribution of a statistic and evaluate properties of estimators, such as variance. Proposed by Efron in 1979 to determine the precision of statistical estimates, it does not rely on the assumptions that are required in parametric procedures by virtue of being easily used in contexts where distributional assumptions are not met. (Efron, 1979; Efron & Tibshirani, 1994)

Mathematical formulation

Let the original dataset be denoted by

$$D = \{X_1, X_2, \dots, X_n\}$$

where n represents the sample size. A bootstrap sample is generated by randomly drawing n observations with replacement from the original dataset:-

$$D^{*b} = \{X_1^{*b}, X_2^{*b}, \dots, X_n^{*b}\}, b = 1, 2, \dots, B$$

where B denotes the number of

bootstrap replications. Each bootstrap sample is used to calculate a bootstrap estimate of the statistic of interest:-

$$\hat{\theta}^{*b} = s(D^{*b})^n$$

where $s(\cdot)$ represents the statistical estimator and $\hat{\theta}^{*b}$ is the bootstrap estimate obtained from the b -th bootstrap sample..

The bootstrap estimate of the standard error is calculated as

$$SE_{BOOT}(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{b=1}^B (\hat{\theta}^{(b)} - \bar{\theta}^*)^2}$$

Where

$$\bar{\theta}^* = \frac{1}{B} \sum_{b=1}^B \hat{\theta}^{*(b)}$$

Similarly, the bootstrap estimate of bias is

$$\hat{Bias}_{boot} = \bar{\theta}^* - \hat{\theta}$$

where $\hat{\theta}$ is the estimate obtained from the original sample

2. The Simulation Study

A comparative study of the performance of the PCR and PLS with and without Bootstrap will be carried out based on the use of functions and predefined formulas in the R programming language, using two numerical summary measures: **Average Mean Squared Error (AMSE)** and **Average Variance (AVAR)**. Simulations were based on 500 Monte Carlo iterations, and 500 Bootstrapped resampling for each iteration, based on three sample sizes ($n = 30, 50$ and 100), two number of independent variables ($p = 5$ and 10) and three degrees of multicollinearity ($r = 0.80, 0.90$ and 0.99).

Simulation Design

The simulation study employed a full factorial design consisting of the factors shown in Table

Table 1.1. Experimental factors and levels used in the simulation study

Factor	Levels	Values
<i>Sample size (n)</i>	3	30, 50, 100
<i>Number of predictors (p)</i>	2	5, 10
<i>Multicollinearity level</i>	3	Low ($r \approx 0.80$), Moderate ($r \approx 0.90$), Severe ($r \approx 0.99$)
Regression method	4	PCR, Bootstrap-PCR, PLS, Bootstrap-PLS
<i>Monte Carlo replications</i>	—	500 per cell
<i>Bootstrap replications (B)</i>	—	500 per model fit

2.1 Numerical Summary for the simulation study

For the Bootstrap-based methods (Bootstrap-PCR and Bootstrap-PLS), 500 bootstrap samples were generated in each Monte Carlo replication. Each simulation scenario was repeated 500 Monte Carlo times. Results of simulations are presented in Tables 1.2-1.5 include the results of the accuracy of four regression methods: Principal Component Regression (PCR), Bootstrap Principal Component Regression (Bootstrap-PCR), Partial Least Squares Regression (PLS) and Bootstrap Partial Least Squares Regression (Bootstrap-PLS). The models in question have been assessed based on various numbers of predictors and various levels of multicollinearity, as well as sample sizes.

Table 1.2. Average Mean Squared Error (AMSE) of PCR, Bootstrap-PCR, PLS, and Bootstrap-PLS under different levels of multicollinearity and sample sizes ($p = 5$)

		Multicollinearity Level	n	PCR	Bootstrap-PCR	PLS	Bootstrap- PLS
AMSE	P=5	Moderate ($\rho=0.80$)	30	0.8959408	0.8818335	0.8101803	0.8108287
			50	0.9435222	0.9331429	0.8848689	0.8846948
			100	0.9687197	0.9642374	0.9395041	0.9394758
		High ($\rho=0.90$)	30	0.9112342	0.8969275	0.828536	0.8286953
			50	0.93351	0.9236335	0.8758259	0.8757862
			100	0.9868447	0.9818481	0.9566216	0.9565861
		Severe ($\rho=0.99$)	30	0.9060134	0.8889636	0.8147344	0.8151918
			50	0.9411486	0.9318893	0.8849842	0.8846637
			100	0.9688093	0.9640804	0.9402516	0.9401965

Table 1.3. Average Mean Squared Error (AMSE) of PCR, Bootstrap-PCR, PLS, and Bootstrap-PLS under different levels of multicollinearity and sample sizes ($p = 10$)

		Multicollinearity Level	n	PCR	Bootstrap-PCR	PLS	Bootstrap- PLS
AMSE	P=10	Moderate ($\rho=0.80$)	30	0.8946216	0.8762591	0.6932614	0.701744
			50	0.9371442	0.925106	0.7991141	0.8015106
			100	0.968134	0.961752	0.8946973	0.8950404
		High ($\rho=0.90$)	30	0.9083339	0.8890727	0.7076107	0.7149333
			50	0.9444655	0.9310001	0.8040574	0.8065864
			100	0.9635943	0.9569235	0.889772	0.8900989
		Severe ($\rho=0.99$)	30	0.8970919	0.8768668	0.6983781	0.7055237
			50	0.9491937	0.9362173	0.8126783	0.8147673
			100	0.9748259	0.9688860	0.9019582	0.9022267

Table 1.4. Average Variance (AVAR) of PCR, Bootstrap-PCR, PLS, and Bootstrap-PLS under different levels of multicollinearity and sample sizes ($p = 5$)

		Multicollinearity Level	n	PCR	Bootstrap-PCR	PLS	Bootstrap- PLS
AVAR	P=5	Moderate ($\rho=0.80$)	30	0.02492456	0.01383198	0.1237079	0.09314081
			50	0.01548736	0.008770635	0.07855816	0.06646589
			100	0.00913793	0.004804649	0.04076203	0.03720083
		High ($\rho=0.90$)	30	0.05259922	0.02813033	0.2445741	0.1879858
			50	0.03086371	0.0164854	0.1615954	0.1340281
			100	0.01633641	0.008762683	0.08257327	0.07448459
		Severe ($\rho=0.99$)	30	0.4779454	0.2454562	2.556712	1.919775
			50	0.3528653	0.1743716	1.607657	1.358372
			100	0.1843895	0.09569094	0.7979605	0.7255784

Table 1.5 Average Variance (AVAR) of PCR, Bootstrap-PCR, PLS, and Bootstrap-PLS under different levels of multicollinearity and sample sizes ($p = 10$)

		Multicollinearity Level	n	PCR	Bootstrap-PCR	PLS	Bootstrap- PLS
AVAR	P=10	Moderate ($\rho=0.80$)	30	0.0100120	0.004259263	0.103865	0.06076604
			50	0.0065794	0.002740651	0.0746665	0.05139885
			100	0.0035665	0.001422103	0.0401541	0.03241277
		High ($\rho=0.90$)	30	0.0185644	0.006924237	0.2029628	0.1209407
			50	0.0118640	0.00452473	0.1502737	0.1030647
			100	0.0064773	0.002378838	0.0808118	0.06496109
		Severe ($\rho=0.99$)	30	0.1778455	0.06609346	2.049032	1.210759
			50	0.1227966	0.04422563	1.423772	0.9754039
			100	0.0753325	0.024498970	0.8165654	0.65904020

Analysis of Simulation Results

results of the Simulations

Tables 1.2 and 1.3 report the results for PCR, Bootstrap-PCR, PLS, and Bootstrap-PLS with two different numbers of predictors ($p = 5$ and $p = 10$), three degrees of multicollinearity ($\rho = 0.80, 0.90, 0.99$), and three sample sizes ($n = 30, 50, 100$). are as follows:

- Effect of Bootstrap Resampling

1-For $p=5$ and $p=10$, Bootstrap-PCR also outperformed the traditional PCR estimator for the majority of simulation settings in terms of the AMSE. This suggests that bootstrap resampling enhances the PCR predictive quality by having a smaller prediction error.

2-The differences between the standard PLS and Bootstrap-PLS estimators for PLS were small and negligible for all the simulation settings. This implies that the effect of bootstrap resampling on the predictive performance of PLS is very limited.

Comparison of PCR and PLS

For all levels of multicollinearity and sample sizes PLS and Bootstrap-PLS outperformed PCR and Bootstrap-PCR in terms of obtaining lower AMSE values. This shows that for both $p = 5$ and $p = 10$, PLS has higher predictive accuracy than PCR.

• Influence of Sample Size

For both values of p , the AMSE values became more consistent with increasing sample size ($n = 30, 50, 100$). Predictive accuracy also depends on sampling size and the degree of multicollinearity, multiple relationships were found in this study.

• Impact of Multicollinearity

There was no clear pattern in AMSE when the correlation coefficient was increased from $\rho = 0.80$ to $\rho = 0.99$. Increased multicollinearity also led to higher AMSE in some scenarios, but only minor differences in others. Hence, the influence of multicollinearity appears to rely on the number of predictors, sample size, and estimation method.

• Impact of the Number of Predictors

A comparison between $p = 5$ and $p = 10$ in Tables 1.2 and 1.3 indicates nearly the same results. PLS and Boot-PLS were again the best performing methods in both situations, and again Bootstrap-PCR was clearly superior to the standard PCR estimator. Hence, the dominance of PLS and the advantage of bootstrapping for PCR remained unaffected by the number of predictors.

results of the Simulation for Tables 1.4 and 1.5: AVAR (Average Variance).

Tables 1.4 and 1.5 displays the numbers for the Average Variance (AVAR) for the four methods under comparison, PCR, Bootstrap-PCR, PLS, and Bootstrap-PLS under three multicollinearity levels ($\rho = 0.80, 0.90$, and 0.99), three sample sizes ($n = 30, 50$, and 100), and two numbers of predictor variables ($p = 5$, and $p = 10$). As AVAR represents the variability of the estimated model, it is more desirable

to have smaller values for AVAR which implies higher stability of estimation and statistical efficiency.

For $p = 5$ (Table 1.4), at the moderate multicollinearity level ($\rho = 0.80$), for all procedures the AVAR values were relatively small. However, Bootstrap-PCR was always superior in terms of variance to PCR. For instance, when $n = 30$, the AVAR value decreased by around 44% from 0.0249 for PCR to 0.0138 after applying bootstrap. Again, Bootstrap-PLS reduced the variance from 0.1237 to 0.0931, which is close to a 25% decrement. For all methods, the AVAR values gradually decreased as the sample size grew from 30 to 100, suggesting that estimators become more stable with increasing sample size.

When multicollinearity rose to $\rho = 0.90$, the variance of all estimates increased substantially. Nevertheless, bootstrap was also very successful in bounding the variability of the estimator. Bootstrap-PCR decreased AVAR from 0.0526 to 0.0281 for $n = 30$, whereas Bootstrap-PLS decreased it from 0.2446 to 0.1880. The variance reduction also became more evident than under moderate multicollinearity, exhibiting the importance of bootstrap increasingly as correlation among predictor gets higher.

When severe multicollinearity ($\rho = 0.99$) was present, the conventional estimators became highly unstable. For instance, when $n = 30$, PCR had an AVAR of 0.4779 and Bootstrap-PCR devised it to 0.2455, which is a decrease of almost 49%. In a similar manner, PLS yielded an AVAR of 2.5567, whereas Bootstrap-PLS reduced it to 1.9198, which is about 25% reduction. While the variance was still relatively large due to the extremely high correlation between the predictors, bootstrap greatly enhanced the estimator stability in this challenging scenario.

A similar trend can be seen for $p = 10$ (Table 1.5). For all levels of sample size and multicollinearity, Bootstrap-PCR had the lowest AVAR among the PCR-based methods, and Bootstrap-PLS performed best among the PLS-model. At $\rho = 0.80$, $n = 30$, Bootstrap-PCR decreased the variance by about 57% and Bootstrap-PLS obtained a reduction of approximately 42% over the conventional PLS. These

gains continued as the sample size increased, all the methods showing decreasing AVARs.

The impact of bootstrap was more pronounced under $\rho = 0.90$. E.g., Bootstrap-PCR reduces AVAR from 0.0186 to 0.0069 at $n = 30$, which corresponds to a reduction more than 60%, while Bootstrap-PLS reduces AVAR from 0.2030 to 0.1209, which corresponds to a reduction close to 40%. These results show that bootstrap is more advantageous for higher multicollinearity.

that even under very high multicollinearity, bootstrap can be still very effective 2.0490 to 1.2108, that is a reduction of about 41%. These suggest extraordinary gains. E.g., Bootstrap-PCR reduced AVAR from 0.1778 to 0.0661 for $n = 30$, that is a reduction of about 63%, while Bootstrap-PLS reduced AVAR from too highly correlated. Bootstrap, however, continued to yield When using the hardest case ($\rho = 0.99$), the variance of the estimators also increased for all methods, as a result of the predictors being.

precision estimates. Second, a rise in multicollinearity from $\rho = 0.80$ to $\rho = 0.99$ drastically increases estimator variance, indicating the detrimental impact strong predictor dependence has on estimation The two aspects of the simulation design are then associated with two notable findings. First, the effect of larger sample sizes was to consistently decrease AVAR for every method, at every level of multicollinearity and for every number of predictors, thus confirming that larger samples lead to more stable parameter.

Overall Discussion

Overall, the simulation study suggests that Bootstrap improves the performance of both principal component regression (PCR) and partial least squares regression (PLS) for different degree of co linearity and sample size. The Bootstrap largely decrease the AMSE variability of PCR, therefore the results of Bootstrap-PCR were in general better than those of the standard PCR. Improvements for PLS were smaller, which is not surprising given the good performance of the standard PLS model. Bootstrap, nonetheless, decreased the estimates variances for the two techniques in the majority of the situations, and the greatest gains were attained at very high co linearity levels coupled with small sample sizes. The same is true for the predictive power of the two methods, which was also enhanced by

the method in the majority of the cases. In summary, the Bootstrap-PLS approach is superior to the simple PLS approach while the Bootstrap-PCR method yields slightly better than the classical PCR method.

References

1. Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26.
2. Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC Press.
3. Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press.
4. Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson.
5. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.
6. Jolliffe, I. T. (2002). *Principal component analysis* (2nd ed.). Springer.
7. Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models* (5th ed.). McGraw-Hill Irwin.
8. Krichene, E., Hmadi, M. S. A., & Al-Gajamiya, S. K. (2026). A Fair Comparative Framework for Time-Series Forecasting Using ARIMAX, XGBoost, and LSTM: Evidence from Libya. *Al-Farooq Journal of Sciences*, 2(3), 69-85.
9. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). Wiley.
10. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2021). *Introduction to linear regression analysis* (6th ed.). Wiley.
11. Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58(2), 109–130.. [https://doi.org/10.1016/S0169-7439\(01\)00155-1](https://doi.org/10.1016/S0169-7439(01)00155-1).