IMPACT OF SAMPLE SIZE ON MULTICOLLINEARITY WITH HIGH DIMENSIONAL DATA IN LOGISTIC REGRESSION ANALYSIS

Gerald Ohene Agyekum[1*], Samuel Akwasi Adarkwa[1], and Richard Yaw Kusi[1]

[1]*Department of Statistical Sciences, Kumasi Technical University.*

[*]*Corresponding author: Gerald Ohene Agyekum,geraldagyekum45@gmail.com*

Abstract

In fields like epidemiology and biological sciences, logistic regression is essential for predicting or categorizing binary outcomes. However, multicollinearity, where predictor variables are highly correlated, can impact the model and lead to erroneous conclusions about each predictor's effect. While maximum likelihood estimation is commonly used to obtain model parameters, it can be problematic with small sample sizes. There is littleresearch on how sample size affects multicollinearity in high-dimensional binary logistic regression. To address this, researchers often suggest using methods like variable droppingor principal component analysis. This study aimed to evaluate the feasibility of using PCA to manage multicollinearity in logistic regression with large column vectors and examine how sample size affects multicollinearity with samples of size 100, 200, 500, 1000, 1500 and 2000. Results indicate that standard errors (SEs) and Variance Inflation Factors (VIFs) decrease with larger sample sizes and increase as sample sizes decrease, even with no correlation between predictors. Suggesting that sample size plays a crucial role in multicollinearity. The study recommends a sample size of at least 500 to avoid issues with multicollinearity in logistic regression. If obtaining a sample of this size is not possible, using Principal Component Analysis (PCA) is a useful alternative.

Keywords: logistic regression, multicollinearity, high dimensional data, principal component analysis

1.0 INTRODUCTION

In contrast to linear regression, which is utilized for forecasting continuous results, logistic regression aims to classify or forecast discrete or categorical outcomes, for instance, the possi- bility that a patient will succeed in cognitive rehabilitation (Maroof, 2012). This is particularly important in epidemiology and biological sciences, where it is necessary to predict or categorize a binary outcome or the likelihood of an event happening. Logistic regression is utilized to evaluate the probability of factors, such as high blood pressure, age, and cholesterol, influenc- ing the possibility of suffering from heart disease. However, the logistic regression model can sometimes be impacted by multicollinearity, which violates its underlying assumptions.

A scenario known as multicollinearity occurs when there is a stronger correlation between two or more explanatory variables in the model (Aidoo et al., 2021). Multicollinearity, however, has the potential to bias or skew the estimated model parameter as well as increase the estimated standard errors related to the model parameter estimates (Lavery et al., 2019; Zahid & Ramzan, 2012; Sinan & Alkan, 2015). Multicollinearity allegedly undermines the statistical significance of independent variables, which makes it difficult to understand model parameters (Mackinnon & Puterman, 1989; Allen, 1997). Multicollinearity does not influence the model's goodness of fit, but it does lead to the incorrect conclusion when the goal is to forecast the effect of each predictor variable. This has an impact on how coefficients of predictor variables should be interpreted as a measure of how much the outcome variable will change with a unit change in the predictor while holding other variables constant. Due to the current limitation, the overlapping of variables can lead to an unrealistic contribution from each explanatory variable. The maximum likelihood estimation approach is typically used to obtain the parameters of the logistic model. When a large sample size is involved, the maximum likelihood estimators are known to be objective and consistent (Murphy, Rossini, & van der Vaart, 1997). Small samples are, however, frequently used in the medical sciences.

The effect of sample size on multicollinearity has been the subject of numerous studies, but little is known about binary logistic regression, particularly when dealing with high dimensional column vectors. Because of this, it might be difficult to determine at what sample size a researcher should be concerned about the problem of multicollinearity.

Numerous solutions to multicollinearity problems have been suggested by experts over the years (Aidoo et al., 2021). The drop variable approach, which entails deleting highly correlated variables from the model, is one of the most popular and basic methods for handling multicollinearity issues (Chen, 2012). Principal component analysis (PCA), which identifies orthogonal directions of maximum variance in the original data set and projects it onto a low dimensional subspace composed of the highest variance components without losing much information, is another method for dealing with the multicollinearity problem.

1.1 Study Objectives

The aim of the study is to investigate how sample size affects multicollinearity when performing logistic regression analysis. In addition, the study aims to ascertain the feasibility of PCA in handling multicollinearity problems in logistic regression and the extent to which the method can yield estimates with less variance Inflation Factor (VIF).

2.0 METHODS AND MATERIALS

2.1 *The Logistic Regression Model*

For the analysis of data with categorical dependent variables, the multivariate statistical tool known as logistic regression (LR) is frequently utilized. It is among the top resources for creating and applying binary linear models for classification.

To use the logistic regression model on the data, certain assumptions must be made, aswith any other method. Contrary to presuming a linear relationship between the predictor and outcome variables, logistic regression assumes a relationship between the logit of the independent and dependent values. Additionally, the outcome variable needs to be categorical, and its categories must be both mutually exclusive and exhaustive as well (Hosmer, Lemeshow, & Sturdivant, 1989).

Let $x_1, x_2, x_3, x_4, \ldots x_p$ Let us denote a collection of continuous predictor variables without any error. We will now consider n observations of these variables, which will be represented in the matrix format; $\chi = (x_{ij})_{n\times p}$.

let $Y = (y_1, y_2, y_3, y_4, \ldots, y_n)'$ be a random sample outcome variable Y associated with the observations in χ that is, $y_i \in \{0, 1\}$ $i = 1, 2, 3, 4, \ldots n$. Then the LR model is given by;

$$y_i = \pi_i + \varepsilon_i, \qquad i = 1, 2, 3, 4, \ldots p, \qquad (1)$$

Where:

y_i is the outcome

ε_i is the error term

π represents the expected value of Y given X = x1, ..., *xp, and it is expressed in the model as follows:*

$$\pi_i = P\{Y = 1|x_{i1}, \ldots, x_{ip}\} = \frac{e^{\beta_\circ+\beta_1 X_1+\ldots+\beta_p X_p}}{1+e^{\beta_\circ+\beta_1 X_1+\ldots+\beta_p X_p}} \qquad (2)$$

Where: $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \ldots, \beta_p$ are the model parameters.

Since the above logistic model is non-linear, the logit transformation would be employed to make it linear. This is given by:

$$ln\left[\frac{p(x)}{1-p(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \qquad (3)$$

The linear summation can be obtained from the logit of p(x). The ratio of p(x) divided by 1-p(x), whose log value yields the logit, expresses the likelihood that a COVID-19 patient would pass away, with x denoting the predictor variables.

2.2 Techniques for Detecting Multicollinearity

2.2.1 Tolerance

In order to identify the presence of multicollinearity, one may employ the utilization of tolerance, which measures the proportion of the variation in a specific independent variable that remains unaccounted for by the remaining independent variables. The specific explanatory variables which have a certain level of tolerance is given by:

$$Tolerance = 1 - R^2 \qquad (4)$$

If the tolerance value approaches 1, it suggests minimal multicollinearity, whereas if it approaches 0, there is a greater chance of multicollinearity.

2.2.2 VIF

The Variance Inflation Factor (VIF) is an indicator that reveals the degree by which multicollinearity inflates the estimated coefficient's variance. A Variance Inflation Factor value which is ≥ 5 indicates the presence of multicollinearity (Shrestha, 2020). A VIF value is given by;

$$VIF_k = \frac{1}{1-R_k^2} \qquad (5)$$

Where;

R_k^2 is the coefficient of determination for regressing the predictor variable on the rest of the variables.

2.2.3 Correlation Coefficient

The Pearson correlation coefficient (also known as the "correlation coefficient") is a most frequently employed indication of a linear relationship among two normally distributed variables. The Pearson coefficient is commonly calculated using a Least-Squares fit, with a value of 1 denoting a perfect positive link, a value of -1 denoting a perfect negative link, and a value of 0 indicating the lack of a relationship between variables. The usual rule is that multicollinearity exists when the correlation between variables exceeds 0.7 or 0.9.

2.3 Principal Component Analysis

There is always the danger of multicollinearity when dealing with high-dimensional data. Karl Pearson pioneered principal component analysis in the early 20th century, and Harold Hotellingexpanded on it in 1933. It is a multivariate method used to describe the variability of a set of variables in terms of a smaller number of uncorrelated linear spans, such as variables with thehighest variance, known as principal components.

PCA produces a feature subspace that maximises variation along the axes, we must first standardise the original data set ($mean$ = 0, $variance$ = 1). The main goal here is to convertthe data set X of D dimensions into a new sample set Z of smaller dimension P ($P < D$), here$Z$ is the PC of X.

The principal components can be obtained as follows:

1. Organize the data set. With X having a set of n vectors where X_i element is an instanceof our data set.
2. Find the mean of variables using the equation;

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (6)$$

3. Calculate the variance;

$$s^2 = \frac{1}{n-1}[x_i - \bar{x}]^2$$

4. Calculate the covariance;

$$S = \frac{1}{n-1}[x_i - \bar{x}_i][x_j - \bar{x}_j] \qquad (7)$$

5. Calculate the eigenvalues and eigenvectors of the covariance matrix. Determine the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors indicate the orientations of the new feature space, while the eigenvalues indicate their respective magnitudes.

   Let us say A is a $d \times d$ matrix then a non-zero x in $\Re^d$ is the eigenvector of A. If Ax is also a scalar multiple of x that is: $Ax = \lambda x$ for some scalar λ. The scalar λ is termed as an eigenvalue of A and λ is corresponded by x. Since the eigenvector corresponding to the eigenvalue is non-zero of the matrix A, they satisfy the equation:

$$(\lambda I - A)x = 0 \qquad (8)$$

By using this definition, we can refer to the set E as encompassing all vectors x that fulfill the equation stated below, representing the associated eigen space.

$$E\{x : (A - \lambda I)x\} = 0 \qquad (9)$$

6. Once the eigen space is obtained from the covariance matrix, the next step is to order the eigenvectors with regard to their eigenvalues from highest to lowest ($\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \lambda_4, \ldots, \geq \lambda_n$). The proportion of total variation accounted for by the i^{th} PC is given by:

$$\frac{\lambda_i}{\sum_{i=1}^{n} \lambda_i}, i = 1,2, \ldots, n \qquad (10)$$

   by doing so, we eliminate components that are less significant and retain PCs that provide good approximations of the original data.

2.4 Statistical Software

The R software version R.4.1.3 was used for fitting the LR models at 0.05 significance level andsimulation of data.

3.0 Simulation Setting

In the simulation, 1000 samples (replications) of sizes n= 100, 200, 500, 1000, 1500, and 2000 were generated based on model (2) with 20 covariates. In order to study the effect of sample size on multicollinearity, two different logit models were created during the data generation stage. These two models were simulated in two different ways as:

3.1 Version 1

1. Generate 20 independent normal predictor variables $X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9, X_{10}, \ldots\ldots, X_{20}$ from a multivariate normal distribution with mean vector $\mu' = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, \ldots\ldots\ldots\ldots, 0)$ and covariance matrix:

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}$$

2. Use the information in **1** to generate an outcome variable *y* using the logit function:

$$\mu = \frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\ldots+\beta_pX_p}}{1 + e^{\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\ldots+\beta_pX_p}} \qquad (11)$$

with model parameters $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \ldots\beta_p$

3. Generate the observed response variable *y* such that:

$$y = \begin{cases} 0 \text{ if } \mu < 0.5 \\ 1 \text{ if } \mu \geq 0.5 \end{cases} \qquad (12)$$

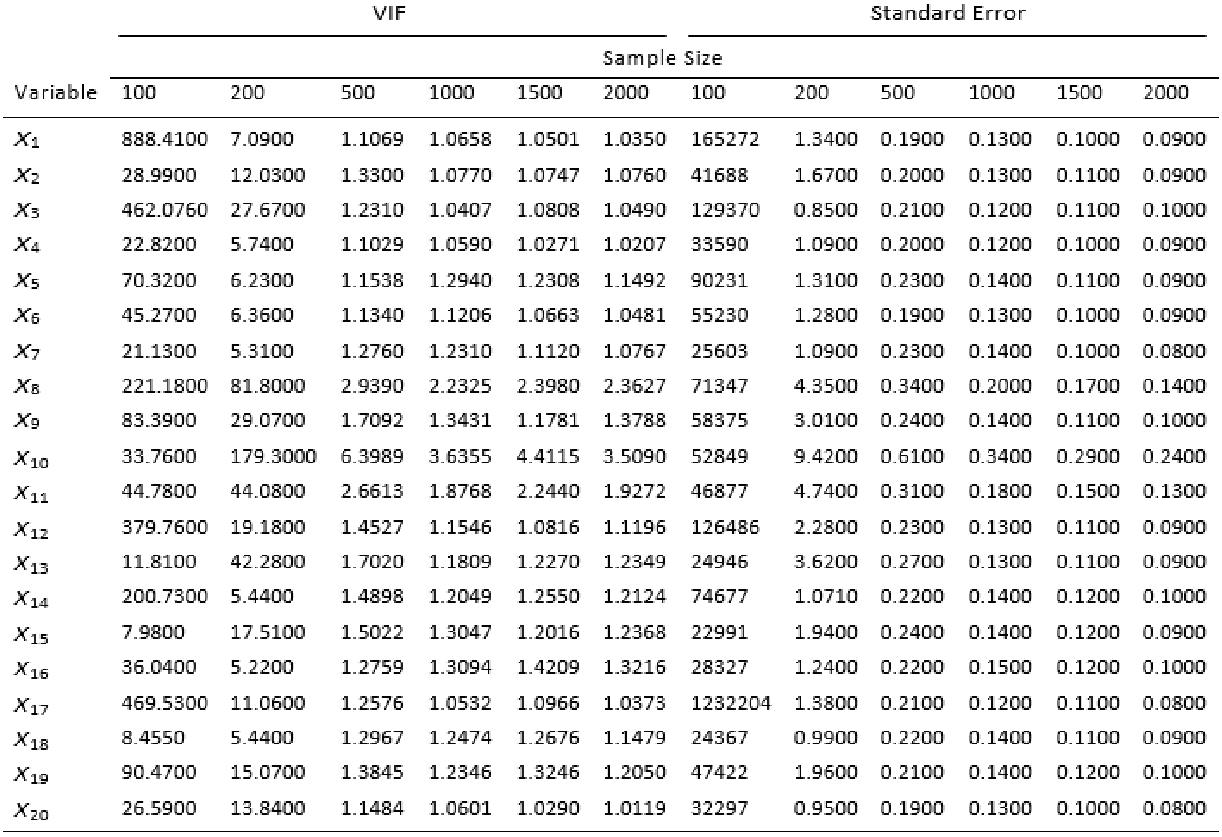
4. Fit a binary logistic model $y \sim 0.2 + 0.12X_1 + 0.4X_2 + 0.3X_3 + 0.03X_4 + 0.6X_5 + 0.43X_6 + 0.4X_7 + 2.1X_8 + 0.72X_9 + 4.3X_{10} + 1.9X_{11} + 0.6X_{12} + 0.8X_{13} + 0.7X_{14} + 0.7X_{15} + 0.8X_{16} + 0.2X_{17} + 0.6X_{18} + 0.9X_{19} + 0.11X_{20}$
5. Perform the steps numbered 1 to 4 in a repetitive manner, iterating 1000 times for various samples, and determine the suitable parameters through computation.

3.2 Version 2

1. To address the issue of high multicollinearity in a large data set X containing 100 observations and 20 variables, principal component analysis was utilized. The data was sourced from a multivariate normal distribution with a mean of zero and a variance of one for comparative analysis.
2. High correlation is fixed at 0.80 to induce problem with multicollinearity deliberately.
3. All other procedures remain the same as outlined in the first simulation setting.

4.0 Results and Discussion

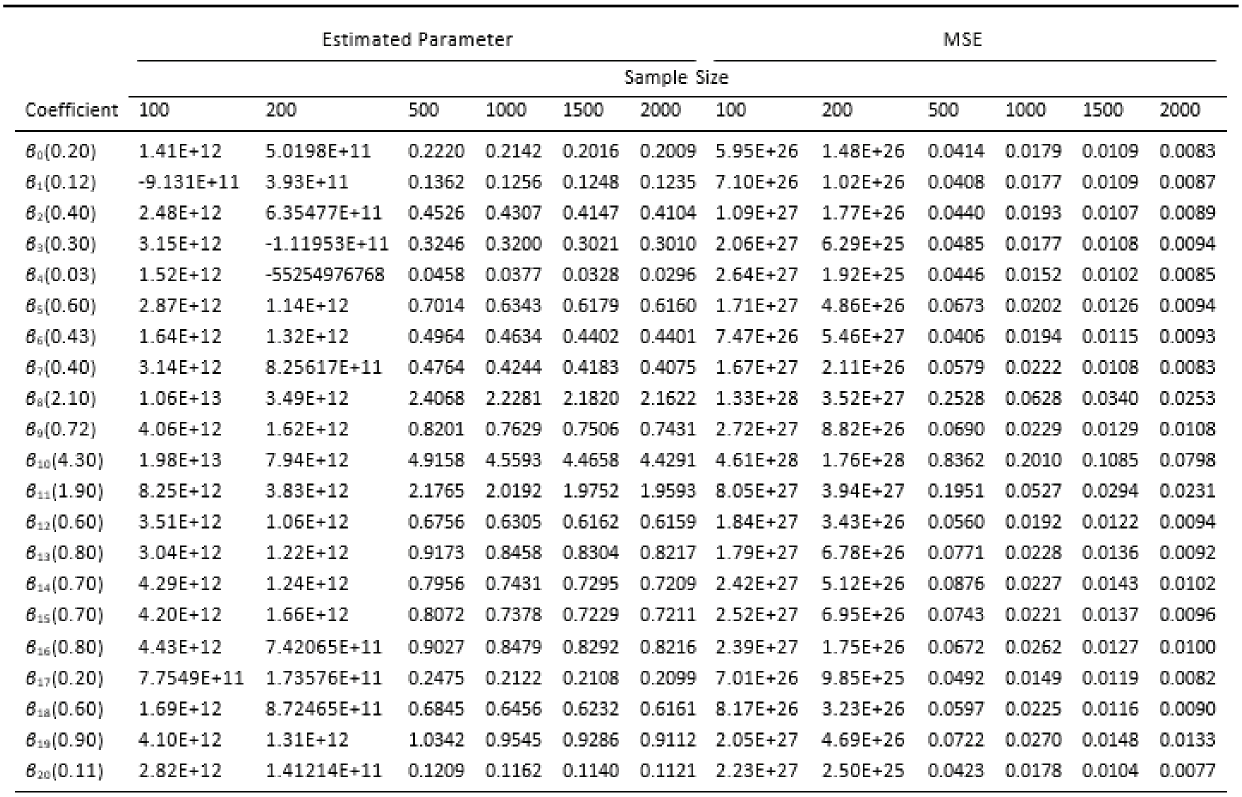
Table 1: VIF and standard errors for binary logistic model variables with 20 predictor variables.



	VIF						Standard Error					
	Sample Size											
Variable	100	200	500	1000	1500	2000	100	200	500	1000	1500	2000
X_1	888.4100	7.0900	1.1069	1.0658	1.0501	1.0350	165272	1.3400	0.1900	0.1300	0.1000	0.0900
X_2	28.9900	12.0300	1.3300	1.0770	1.0747	1.0760	41688	1.6700	0.2000	0.1300	0.1100	0.0900
X_3	462.0760	27.6700	1.2310	1.0407	1.0808	1.0490	129370	0.8500	0.2100	0.1200	0.1100	0.1000
X_4	22.8200	5.7400	1.1029	1.0590	1.0271	1.0207	33590	1.0900	0.2000	0.1200	0.1000	0.0900
X_5	70.3200	6.2300	1.1538	1.2940	1.2308	1.1492	90231	1.3100	0.2300	0.1400	0.1100	0.0900
X_6	45.2700	6.3600	1.1340	1.1206	1.0663	1.0481	55230	1.2800	0.1900	0.1300	0.1000	0.0900
X_7	21.1300	5.3100	1.2760	1.2310	1.1120	1.0767	25603	1.0900	0.2300	0.1400	0.1000	0.0800
X_8	221.1800	81.8000	2.9390	2.2325	2.3980	2.3627	71347	4.3500	0.3400	0.2000	0.1700	0.1400
X_9	83.3900	29.0700	1.7092	1.3431	1.1781	1.3788	58375	3.0100	0.2400	0.1400	0.1100	0.1000
X_{10}	33.7600	179.3000	6.3989	3.6355	4.4115	3.5090	52849	9.4200	0.6100	0.3400	0.2900	0.2400
X_{11}	44.7800	44.0800	2.6613	1.8768	2.2440	1.9272	46877	4.7400	0.3100	0.1800	0.1500	0.1300
X_{12}	379.7600	19.1800	1.4527	1.1546	1.0816	1.1196	126486	2.2800	0.2300	0.1300	0.1100	0.0900
X_{13}	11.8100	42.2800	1.7020	1.1809	1.2270	1.2349	24946	3.6200	0.2700	0.1300	0.1100	0.0900
X_{14}	200.7300	5.4400	1.4898	1.2049	1.2550	1.2124	74677	1.0710	0.2200	0.1400	0.1200	0.1000
X_{15}	7.9800	17.5100	1.5022	1.3047	1.2016	1.2368	22991	1.9400	0.2400	0.1400	0.1200	0.0900
X_{16}	36.0400	5.2200	1.2759	1.3094	1.4209	1.3216	28327	1.2400	0.2200	0.1500	0.1200	0.1000
X_{17}	469.5300	11.0600	1.2576	1.0532	1.0966	1.0373	1232204	1.3800	0.2100	0.1200	0.1100	0.0800
X_{18}	8.4550	5.4400	1.2967	1.2474	1.2676	1.1479	24367	0.9900	0.2200	0.1400	0.1100	0.0900
X_{19}	90.4700	15.0700	1.3845	1.2346	1.3246	1.2050	47422	1.9600	0.2100	0.1400	0.1200	0.1000
X_{20}	26.5900	13.8400	1.1484	1.0601	1.0290	1.0119	32297	0.9500	0.1900	0.1300	0.1000	0.0800

The study examined the effect of varying sample sizes on multicollinearity in logistic regression with high dimensional data, using the first simulation setting. In Table 1, VIF values and standard errors were presented for different sample sizes. Despite setting the correlation between predictor variables to zero, multicollinearity was still present and more severe in smaller sample sizes, especially those below 500. However, increasing the sample size resulted in a decrease in the severity of multicollinearity. The standard errors of predictor variables also increased with higher VIF values and were more prominent in smaller sample sizes. Nonetheless, the standard errors decreased significantly as sample size increased. Overall, the findings demonstrate that sample size can induce multicollinearity in logistic regression with high dimensional data, indicating that multicollinearity is particularly sensitive to changes in sample size.

Table 2: Mean of the predicted parameter and the average of their square error for the binary logistic model containing 20 variables, with varying sample sizes.

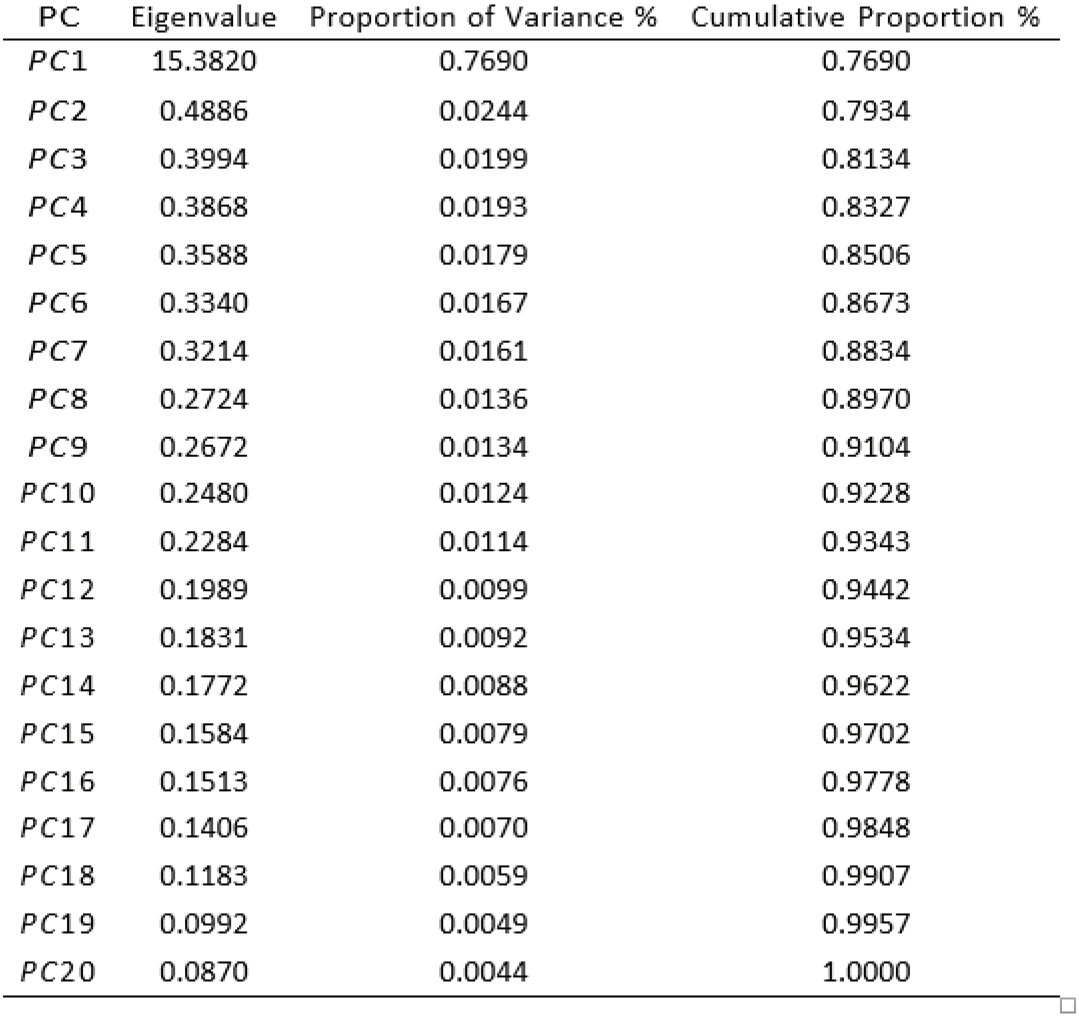


	Estimated Parameter						MSE					
	Sample Size											
Coefficient	100	200	500	1000	1500	2000	100	200	500	1000	1500	2000
β_0(0.20)	1.41E+12	5.0198E+11	0.2220	0.2142	0.2016	0.2009	5.95E+26	1.48E+26	0.0414	0.0179	0.0109	0.0083
β_1(0.12)	-9.131E+11	3.93E+11	0.1362	0.1256	0.1248	0.1235	7.10E+26	1.02E+26	0.0408	0.0177	0.0109	0.0087
β_2(0.40)	2.48E+12	6.35477E+11	0.4526	0.4307	0.4147	0.4104	1.09E+27	1.77E+26	0.0440	0.0193	0.0107	0.0089
β_3(0.30)	3.15E+12	-1.11953E+11	0.3246	0.3200	0.3021	0.3010	2.06E+27	6.29E+25	0.0485	0.0177	0.0108	0.0094
β_4(0.03)	1.52E+12	-55254976768	0.0458	0.0377	0.0328	0.0296	2.64E+27	1.92E+25	0.0446	0.0152	0.0102	0.0085
β_5(0.60)	2.87E+12	1.14E+12	0.7014	0.6343	0.6179	0.6160	1.71E+27	4.86E+26	0.0673	0.0202	0.0126	0.0094
β_6(0.43)	1.64E+12	1.32E+12	0.4964	0.4634	0.4402	0.4401	7.47E+26	5.46E+27	0.0406	0.0194	0.0115	0.0093
β_7(0.40)	3.14E+12	8.25617E+11	0.4764	0.4244	0.4183	0.4075	1.67E+27	2.11E+26	0.0579	0.0222	0.0108	0.0083
β_8(2.10)	1.06E+13	3.49E+12	2.4068	2.2281	2.1820	2.1622	1.33E+28	3.52E+27	0.2528	0.0628	0.0340	0.0253
β_9(0.72)	4.06E+12	1.62E+12	0.8201	0.7629	0.7506	0.7431	2.72E+27	8.82E+26	0.0690	0.0229	0.0129	0.0108
β_{10}(4.30)	1.98E+13	7.94E+12	4.9158	4.5593	4.4658	4.4291	4.61E+28	1.76E+28	0.8362	0.2010	0.1085	0.0798
β_{11}(1.90)	8.25E+12	3.83E+12	2.1765	2.0192	1.9752	1.9593	8.05E+27	3.94E+27	0.1951	0.0527	0.0294	0.0231
β_{12}(0.60)	3.51E+12	1.06E+12	0.6756	0.6305	0.6162	0.6159	1.84E+27	3.43E+26	0.0560	0.0192	0.0122	0.0094
β_{13}(0.80)	3.04E+12	1.22E+12	0.9173	0.8458	0.8304	0.8217	1.79E+27	6.78E+26	0.0771	0.0228	0.0136	0.0092
β_{14}(0.70)	4.29E+12	1.24E+12	0.7956	0.7431	0.7295	0.7209	2.42E+27	5.12E+26	0.0876	0.0227	0.0143	0.0102
β_{15}(0.70)	4.20E+12	1.66E+12	0.8072	0.7378	0.7229	0.7211	2.52E+27	6.95E+26	0.0743	0.0221	0.0137	0.0096
β_{16}(0.80)	4.43E+12	7.42065E+11	0.9027	0.8479	0.8292	0.8216	2.39E+27	1.75E+26	0.0672	0.0262	0.0127	0.0100
β_{17}(0.20)	7.7549E+11	1.73576E+11	0.2475	0.2122	0.2108	0.2099	7.01E+26	9.85E+25	0.0492	0.0149	0.0119	0.0082
β_{18}(0.60)	1.69E+12	8.72465E+11	0.6845	0.6456	0.6232	0.6161	8.17E+26	3.23E+26	0.0597	0.0225	0.0116	0.0090
β_{19}(0.90)	4.10E+12	1.31E+12	1.0342	0.9545	0.9286	0.9112	2.05E+27	4.69E+26	0.0722	0.0270	0.0148	0.0133
β_{20}(0.11)	2.82E+12	1.41214E+11	0.1209	0.1162	0.1140	0.1121	2.23E+27	2.50E+25	0.0423	0.0178	0.0104	0.0077

Table 2 displays the estimated parameters and their associated Mean Squared Error (MSE), including measurement bias. The findings indicate that larger sample sizes, particularly 500 or greater, result in estimated parameters that asymptotically approach their true values on average. This discovery aligns with the conclusions drawn by Bujang et al. (2018), who recommended utilizing a minimum sample size of 500. This sample size is crucial for generating statistical values that accurately reflect the parameters within the targeted population when using logistic regression. For variables with high VIF values, there is significant variability between the estimated and true values, particularly for sample sizes of 100 and 200. This demonstrates that sample size is a factor in accounting for biases in model parameters when multicollinearity is present.

In an attempt to deal with the issue of multicollinearity between predictor variables in the logistic regression model with a sample of size 100, principal component analysis which projects the original data unto a subspace of uncorrelated principal components was employed using the second simulated data where the correlation between predictor variables was set at 0.8.

Table 3: Total variance explained



PC	Eigenvalue	Proportion of Variance %	Cumulative Proportion %
*PC*1	15.3820	0.7690	0.7690
*PC*2	0.4886	0.0244	0.7934
*PC*3	0.3994	0.0199	0.8134
*PC*4	0.3868	0.0193	0.8327
*PC*5	0.3588	0.0179	0.8506
*PC*6	0.3340	0.0167	0.8673
*PC*7	0.3214	0.0161	0.8834
*PC*8	0.2724	0.0136	0.8970
*PC*9	0.2672	0.0134	0.9104
*PC*10	0.2480	0.0124	0.9228
*PC*11	0.2284	0.0114	0.9343
*PC*12	0.1989	0.0099	0.9442
*PC*13	0.1831	0.0092	0.9534
*PC*14	0.1772	0.0088	0.9622
*PC*15	0.1584	0.0079	0.9702
*PC*16	0.1513	0.0076	0.9778
*PC*17	0.1406	0.0070	0.9848
*PC*18	0.1183	0.0059	0.9907
*PC*19	0.0992	0.0049	0.9957
*PC*20	0.0870	0.0044	1.0000

Table 3 presents the principal components together with eigenvalues as well as their total variance explained. The first PC with an eigenvalue of 15.382 has the largest variance that accounts for about 76.9% of the total variation. However, the remaining 19 PCs have eigenvalues less than 1 which indicates that each of the 19 PCs the PC explains less than a single original variable. As a rule of thumb, the first PC collectively accounts for about 76.9% of the variability of the original data set losing only 23.1% of the information. Therefore, only 1 PC is extracted from the twenty PCs without much information loss. This implies that the original information was reduced from a 20-dimension data set which was high into 1-dimensional data while at the same time maximizing the variability of the original data set. The remaining PCs are considered insignificant hence, they are omitted from the analysis.

Based on the screen plot in Figure 1, PC1 (76.9%) captures the most variation, whereas the remaining PCs explain less variations in the original data.

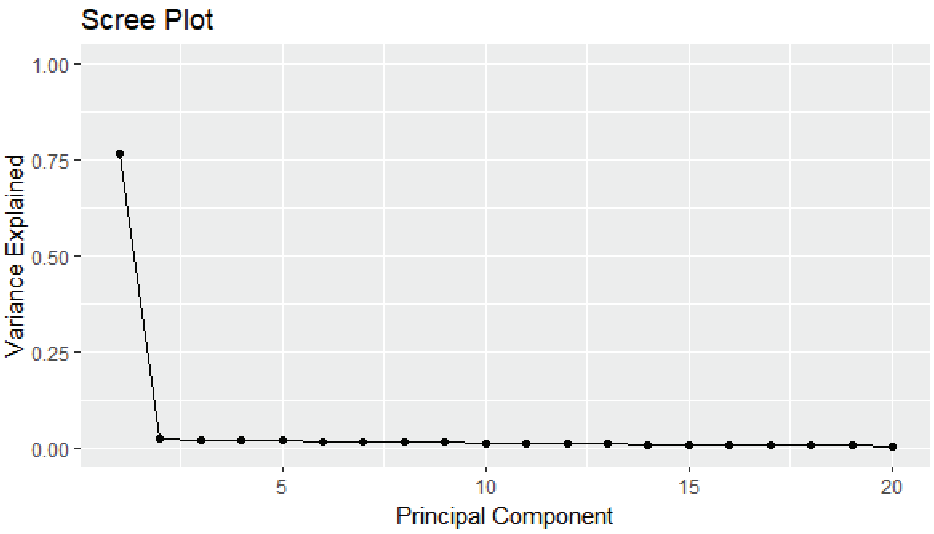


Figure 1: Scree plot for the 20 principal components

Table 4: Comparison of VIF values and standard errors of traditional logistic regression model to principal component logistic regression.

	LR With all Predictor variables			LR with 1 PC	
Variable	Standard Error	VIF	PC	Standard Error	VIF
X_1	1.44E+05	82.3200	PC1	0.5951	1.3099
X_2	1.74E+05	109.9100			
X_3	1.44E+05	37.3900			
X_4	8.14E+04	19.1800			
X_5	7.83E+04	10.9700			
X_6	3.02E+05	214.7900			
X_7	1.52E+05	84.6400			
X_8	8.44E+04	12.9700			
X_9	8.14E+04	15.6500			
X_{10}	1.42E+05	37.0600			
X_{11}	2.04E+05	62.4600			
X_{12}	1.45E+05	48.1200			
X_{13}	7.89E+04	16.1400			
X_{14}	9.70E+04	25.2100			
X_{15}	1.56E+05	75.9200			
X_{16}	6.73E+04	17.3100			
X_{17}	1.50E+05	67.4600			
X_{18}	9.80E+04	47.3900			
X_{19}	9.91E+04	16.0000			
X_{20}	1.79E+05	74.1400			

Table 4 revealed that the traditional logistic regression model had very large standard errors and VIF values for its predictors. In contrast, using principal component analysis (PCA) to reduce the predictors to only one component that explains 76.9% of the variance in the data resulted in a significantly lower VIF value of less than 2. This suggests that when faced with high numbers of predictor variables and low sample sizes, PCA is a more effective method for addressing multicollinearity in logistic regression than the traditional method of regressing on all predictors.

5.0 CONCLUSION

The purpose of this study was to examine how sample size affects multicollinearity in logistic regression analysis of high dimensional data. The study's significance lies in its ability to informresearchers about the impact of different sample sizes on the incidence of multicollinearity andhow they can address it. The study successfully achieved its goals, revealing that when the correlation between predictor variables is set to zero, standard errors and VIF decrease with increasing sample sizes and increase with decreasing sample sizes, indicating that sample size plays a crucial role in the problem of multicollinearity.

Drawing implications from the findings, it can be deduced that a minimum sample size of 500 emerges as an optimal benchmark to circumvent multicollinearity challenges when dealing with high-dimensional data in logistic regression analyses. Should the feasibility of collecting such a sizable sample be constrained, the study suggests an alternative approach: the utilization of principal component analysis (PCA). This technique can effectively ameliorate multicollinearity by transforming original variables into orthogonal components, thereby enhancing the statistical robustness of the analysis.

It is recommended that, forthcoming research endeavours should delve into the nuances of applying PCA within scenarios characterized by limited data availability. Furthermore, investigating the impact of diverse manifestations of multicollinearity on various modeling techniques holds the promise of yielding nuanced insights into optimizing both sample sizes and overall model performance. Additionally, exploring the generalizability of our findings across diverse regression models and disparate fields of study will undoubtedly contribute to a richer understanding of the intricate interplay between sample size and multicollinearity.

REFERENCES

Aidoo, E. N., Appiah, S. K., & Boateng, A. (2021). Brief research report: A monte carlo simulation study of small sample bias in ordered logit model under multicollinearity. *TheJournal of Experimental Education*, *89*(4), 742–750.

Allen, M. P. (1997). The problem of multicollinearity. *Understanding regression analysis*, 176–180.

Bujang, M. A., Sa'at, N., Bakar, T. M. I. T. A., & Joo, L. C. (2018). Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *The Malaysian journal of medical sciences: MJMS*, *25*(4), 122.

Chen, G. J. (2012). A simple way to deal with multicollinearity. *Journal of Applied Statistics*,

39 (9), 1893–1909.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (1989). The multiple logistic regression model. *Applied logistic regression*, *1*, 25–37.

Lavery, M. R., Acharya, P., Sivo, S. A., & Xu, L. (2019). Number of predictors and multi- collinearity: What are their effects on error and bias in regression? *Communications in Statistics-Simulation and Computation*, *48* (1), 27–38.

Mackinnon, M. J., & Puterman, M. L. (1989). Collinearity in generalized linear models.

Communications in statistics-theory and methods, *18* (9), 3463–3472.

Murphy, S., Rossini, A., & van der Vaart, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *Journal of the American Statistical Association*, *92* (439), 968–976.

Shrestha, N. (2020). Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, *8* (2), 39–42.

Sinan, A., & Alkan, B. B. (2015). A useful approach to identify the multicollinearity in the presence of outliers. *Journal of Applied Statistics*, *42* (5), 986–993.

Zahid, F. M., & Ramzan, S. (2012). Ordinal ridge regression with categorical predictors.

Journal of Applied Statistics, *39* (1), 161–171.