Journal of Engineering and Applied Sciences
MEDWELL PUBLICATIONS

Dealing with Multicollinearity in Regression Analysis: A Case in Psychology

[1]Solly Matshonisa Seeletse and [2]Motlalepula Grace Phalwane
[1]*Department of Statistics and Operations Research, Sefako Makgatho Health Sciences University, 1 Molotlegi Street, Ga-Rankuwa, Gauteng Province, South Africa*
[2]*Department of Community Dentistry, Sefako Makgatho Health Sciences University, 1 Molotlegi Street, Ga-Rankuwa, Gauteng Province, South Africa*

Key words: Principal components, ridge regression, stepwise regression, multicollinearity, exploratory

Corresponding Author:
Solly Matshonisa Seeletse
Department of Statistics and Operations Research, Sefako Makgatho Health Sciences University, 1 Molotlegi Street, Ga-Rankuwa, Gauteng Province, South Africa

Page No.: 2693-2703
Volume: 15, Issue 13, 2020
ISSN: 1816-949x
Journal of Engineering and Applied Sciences

Abstract: In regression analysis, the main interest is to predict the response variable using the exploratory variables by estimating parameters of the linear model. However, in reality, the exploratory variables may share similar characteristics. This interdependency between the exploratory variables is called multicollinearity and causes parameter estimation in regression analysis to be unreliable. Different approaches to address the multicollinearity problem in regression modelling include variable selection, principal component regression and ridge regression. In this study, the performances of these techniques in handling multicollinearity in simulated data are compared. Out of the four regression models compared, principal regression model produced the best model to explain the variability and its parameter estimates were precise and addressing multicollinearity.

INTRODUCTION

Regression analysis is a statistical tool to investigate relationships between variables[1]. According to Chatterjee *et al*.[2], regression analysis ascertains the quantitative effect of variables on other variables and assesses the statistical significance of the estimated relationships. Regression analysis with one independent variable is the simple linear regression where only one dependent variable is regressed on one independent variable. Multiple linear regression is when a dependent variable is regressed on >1 independent variable.

A major issue possible in multiple regression analysis is the intercorrelation between independent variables, termed multicollinearity[3]. Dormann *et al*.[4] state that collinearity describes a situation where some predictor variables in a statistical model are linearly related. Multicollinearity is as a problem to every investigator when their main focus is to predict the outcome of a dependent variable from a set of independent variables in multiple linear regression[5].

Multicollinearity problems cause instability problems in regression analysis[6, 7]. Fekedulegn *et al*.[8] explain the disadvantages of using Ordinary Least Square (OLS) regression for estimating the regression parameters when multicollinearity exists between independent variables. OLS produces regression coefficients that could violate the practical situation. Coefficients could fluctuate in sign and magnitude due to small changes in the dependent or independent variables. They can also inflate insignificant standard errors.

There are several methods to detect the presence of multicollinearity. Dormann *et al*.[4] demonstrate that multicollinearity cannot be solved.

Aim and objectives: The aim of this study is to describe multicollinearity and its impact on multiple linear regression analysis. A real-life scenario in Psychology is used for illustrations. The level of multicollinearity between the independent variables in the data is then controlled through a predefined correlation structure.

Methods to address the multicollinearity in the data include stepwise regression, Principal Component Regression (PCR) and ridge regression. These models are assessed through model fit measures and the interpretability of the results.

Multicollinearity problem: Gujarati[9] illuminates that regression analysis estimates dependency between the dependent and the independent variables the parameters. It is not involved in estimating the interdependency between independent variables which is multicollinearity. Farrar and Glauber[10] studied multicollinearity in regression analysis and considered the proper treatment of multicollinearity, its detection or diagnosis and how to correct it with possible additional information.

According to Naes and Mevik[5], a central issue with multicollinearity in data analysis occurs when estimating the parameters of a regression model. When conducting regression analysis, the model fit is assessed using measures such as the p-value and the (R^2) value. R^2 indicates the proportion of variation in the dependent variable accounted for by the independent variables in the model[2]. Whenever the model gives a high R^2 and an overall p-value below 0.05, the model is considered to be a good fit. However, if the multicollinearity in the data is severe, both R^2 and p-value can mislead the "proper specification and effective estimation of the type of structural relationship commonly sought through the use regression[10].

Dormann *et al*.[4] state that multicollinearity can lead to inflated variance of regression parameters as well as incorrect identification of important predictors in the regression analysis. Mela and Kopalle[11] highlight other problems of multicollinearity. They state that various econometric references such as Belsly *et al*.[12], Greene[13] and Kmenta[14] indicate that collinearity increases estimates of parameter variance, yields high R^2 in the face of low parameter significance and results in parameters with incorrect signs and implausible magnitudes.

Multicollinearity diagnostics: Lafi and Kaneene[15] state that in data analysis, the first step is to run diagnostic tests to investigate the existence of multicollinearity. There are different ways to detect multicollinearity such as the correlation matrix, scatterplots, tolerance, Variance Inflation Factor (VIF) and Condition Indices (CI). According to Farrar and Glauber[10], the simpler and easiest way of detecting multicollinearity is in the main diagonal elements of the inverted correlation matrix of the predictor variables. Liu *et al*.[16] states that when the correlation coefficient between two independent variables is large, there is an indication of possible multicollinearity.

Tolerance and VIF are commonly used to detect multicollinearity. Tolerance is the complement of R_i^2, the squared multiple correlation of the ith variable with other independent variables[17]. It is interpreted as the proportion of variance in the ith independent variable that is not related to the other independent variables in the model. VIF is the reciprocal of the tolerance. The relationship between tolerance and VIF is inversely proportional, therefore, variables with low tolerance tend to have large VIF. This would suggest that those variables are collinear. Chatterjee *et al*.[2] suggest that a VIF value >10 indicates the existence of multicollinearity.

CI values are the square roots of ratios of the largest eigenvalue to each successive eigenvalue. A CI value >15 indicates a possible problem and an index >30 suggest a serious problem with multicollinearity[16]. Chatterjee *et al*.[2] state that this criterion is based on empirical observation rather than on pure theory.

Dealing with multicollinearity: Multicollinearity is a difficult problem to solve in regression analysis. In some cases it cannot be entirely eliminated[18]. Dormann *et al*.[4] argue that the problem of multicollinearity cannot be solved and that statistical methods cannot separate collinear variables. Despite this there are different methods to address multicollinearity in regression modelling.

One way to deal with the multicollinearity is through stepwise regression where predictors are added to or removed from the model sequentially[19]. Stepwise regression is used when there is evidence of multicollinearity by sequentially adding or removing some of the regressors into the model according to some criteria such as the F-test of the significance of the independent variables[20]. Based on this test, only variables that are significant are included in the model and any variable that become insignificant at subsequent steps are removed from the model. Although, stepwise regression removes the multicollinearity in the model, this approach ignores the unique contribution of the excluded variables in the regression model which could lead to a loss of power[21].

Hotelling[22] developed an iterative procedure for calculating eigenvalues and eigenvectors of any symmetric matrix, called Principal Component Analysis (PCA). He defined PCA as a method of transforming the original independent variables into new uncorrelated variables. The main objective of PCA is data reduction. In PCA, the original correlated predictor variables are replaced by their uncorrelated principal components in a regression analysis, thereby addressing the problem of multicollinearity and making the regression model more stable[23-25].

Lafi and Kaneene[15] also describe how PCA can be used to correct for multicollinearity. Through a PCA, linear combinations of predictors are created that are uncorrelated with each other and explain as much of the variance in the dataset as possible. Calmes[26] and Kornblut and Wilson[27] used Principal Component Regression (PCR) to determine whether test scores and other economic and education-related variables are good indicators of economic performance. PCR delivered useful results in both cases.

The issue of which components to choose for a PCR has been debated. Mansfield *et al*.[28] suggested that the predictive power in the regression are minimised when components with small variance are deleted. The criteria of deletion of principal components in regression are based on the magnitudes of the eigenvalues of the predictor variables or statistical tests of the significance of the components[29]. Kendall[25] , Massy[30], Jeffers[31] and Hawkins[32] recommend deleting components with small variances. Hocking[33] used a different approach for choosing the principal components. Instead of ignoring components with low variance, he defined a rule for retaining principal components in regression.

Ridge regression has been extensively reviewed in the literature of applied statistics as a method for dealing with multicollinearity. The purpose of ridge regression is to reduce the high variances of the estimated coefficients at the expense of incurring some bias[34].

Niemela-Nyrhimnen and Leskinen[34] illustrate the use of ridge regression in mitigating the effects of multicollinearity in structural equation modelling. They used these two methods with slightly differing ridge estimation procedures. The methods produced the same point estimates of path coefficients. However, one method had smaller standard errors of parameter estimates and larger squared multiple correlations than the other one.

Mahajan *et al*.[35] studied the application of ridge regression in the presence of multicollinearity when analysing parameter estimation in marketing models. They compared OLS estimates and ridge regression estimates and found that OLS estimates are unbiased with large variance and ridge estimates are biased but with smaller variance. According to Mahajan *et al*.[35] ridge regression is a method that could overcome multicollinearity and produce stable estimates that are closer to the true values of the coefficients the analyst is trying to develop.

MATERIALS AND METHODS

The research project consists of three sections; data simulation, regression models and evaluation. A sample of n = 300 respondents are simulated in order to create a high collinear independent variables. The regression model is formulated using the psychometric measures and the three methods (PCR, ridge regression and stepwise regression) are applied to overcome the problem of multicollinearity. The evaluation of the severity of multicollinearity using VIF and CI techniques is performed. The study also gives an overview on the assumptions on residuals for a proper data analysis report.

Data simulation: To assess methods of dealing with multicollinearity data are simulated according to a pre-specified correlation structure such that there is a strong relation between the dependent and all independent variables but a subset of the independent variables are highly correlated. For the purpose of this research a dataset is simulated that reflect a real-world scenario in Psychology related to academic achievement and cognitive measures.

Gasic-Pavisic *et al*.[36] state that Locus Cf control (LOC) is a cognitive component of self-concept. They define LOC as "the extent to which an individual believes he or she is at the mercy of external forces (external LOC) that is the extent to which one is responsible for events that occur in one's life and the extent to which one can control the effect of ones actions (internal LOC). Self-esteem is seen as an evaluative component of self-concept[36] and reflects a person's positive evaluation of self. According to Gasic-Pavisic *et al*.[36] there is a strong relationship between LOC and self-esteem.

In their research Ahmad *et al*.[37] show that self-esteem and academic achievement are strongly related. Bar-Tal and Bar-Zohar[38] also suggest that internal LOC is related to academic performance.

For this analysis a sample of n = 300 respondents are simulated to reflect the dependence and interdependence relationships between these four psychometric measures. To achieve this the following multivariate random normal variables are simulated using the mvrnorm function of the MASS library in R[39] with a mean vector of 0 and a correlation matrix such that all three independent

variables are strongly correlated with academic achievement and that LOC and self-esteem are highly correlated:

Where:
- Y = Standardised test scores of academic achievement of secondary school learners where high values indicate high achievement
- X_1 = LOC in standardised form where high values indicate internal LOC
- X_2 = Self-esteem in standardised form where high values indicate high self-esteem
- X_3 = Intellect in standardised form where high values indicate high non-verbal reasoning

Regression models

OLS regression: An initial OLS regression is fitted to illustrate and assess the extent of the multicollinearity in the data using the regression procedure in SPSS. According to Chatterjee *et al.*[2] the linear relationship between Y and X_1, X_2 and X_3 is formulated as follows:

$$Y = \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\varepsilon \quad (1)$$

OLS regression estimates the β_2 values by minimising the sum of squares of the errors. This yields the estimated regression model[2]:

$$\hat{Y} = \hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\hat{\beta}_3X_3 \quad (2)$$

Where:
- $\hat{\beta}_0$ = Unbiased estimated intercept or constant
- $\hat{\beta}_1$ = Unbiased parameter estimates for variable

In matrix notation the regression model and least squares estimates of the regression parameters are given by Eq. 3 and 4, respectively:

$$y_{OLS} = X\beta+\varepsilon \quad (3)$$

$$\hat{\beta}_{OLS} = \left(X^TX\right)^{-1}X^Ty \quad (4)$$

Models addressing multicollinearity: For the purpose of this research three different methods are applied to overcome the problem of multicollinearity in regression analysis, namely stepwise regression, PCR and ridge regression. SPSS is used to generate the output for the stepwise regression, PCA and PCR. The linear ridge function in the ridge library of R is used for the ridge regression analysis.

Stepwise regression: As stepwise regression is a series of OLS regression models where independent variables are added and removed according to some specific criteria, it follows that the regression model and parameter estimates are essentially the same as in Eq. 3 and 4 with the only difference the size of the input data matrix. The SPSS procedure produces models for each step of the stepwise process. All of these models are evaluated and the optimal model selected.

PCR: Johnson and Wichern[40] state that PCA is concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. The general purposes of PCA are data reduction and interpretation. It is a data reduction technique in such a way that the original correlated exploratory variables are transformed to a new set of variables the principal components which are ordered, so that, the first few retain most of the variation present in all of the original variables.

PCR incorporates the principal components as the new predictor variables in regression. PCA uses an orthogonal transformation to convert data from X onto an orthogonal basis. It converts X using a linear combination of its columns into principal components or loadings P and scores S given by:

$$X = SP^T \quad (5)$$

P contains the coefficients of the linear combinations of the original variables and S gives the coordinates of X in the new orthogonal basis or the principal component space. The first principal component is such that it has the highest variance from the data and it keeps decreasing with subsequent principal components. The PCR has a least squares solution similar to that in Eq. 3 and 4. However, in terms of the principal components they are:

$$\hat{\beta}_{PCR} = \left(S^TS\right)^{-1}S^Ty \quad (6)$$

Ridge regression: In ridge regression, the degree of bias is added to the regression estimates to reduce the standard errors yielding the more reliable estimates. The parameters are estimated by adding a small value k to the diagonal elements of the correlation matrix where k is a positive quantity:

$$\hat{\beta}_{PCR} = \left(XX'+kI\right)^{-1}X'y \quad (7)$$

A key obstacle in using ridge regression is to choose an appropriate value of k[41]. In this analysis 21 different k values are used ranging from 0.05-1.00 with an increment of 0.05.

Evaluation: Various measures are used to assess the quality of the regression analyses, specifically assumptions regarding the residuals as well as the model fit.

Level of multicollinearity: To assess the existence and extent of multicollinearity in the data, both VIF and CI are evaluated. The VIF value forms part of the SPSS regression output. For each independent variable it is calculated using equation[2]:

$$VIF_j = \frac{1}{1-R_j^2}\ j = 1, 2, ..., p \quad (8)$$

Where:
R_j^2 = The proportion of variation in variabl
X_j = Explained by all other variables

X_i, i = 1,..., p, i ≠ j in a regression of X_j on X_i

CI values are calculated as a function of the eigenvalues of the correlation matrix as in the PCA procedure using equation[2]:

$$CI_i = \sqrt{\frac{\lambda_1}{\lambda_i}},\ j = 1, 2,..., p \quad (9)$$

Where:
λ_1 = The maximum eigenvalue of the correlation matrix
λ_i = The minimum eigenvalue of the correlation matrix of size i

Assumptions: According to Chatterjee *et al*.[2] the residuals are assumed to be independently and identically distributed normal random variables. The Shapiro-Wilk test[41] tests whether the residuals follow a normal distribution. The hypothesis:

- H_0: the residuals are normally distributed
- H_1: the residual are not normally distributed

If the null hypothesis is rejected, the residuals can be considered to be non-normal. If the null hypothesis is not rejected, then the assumption of normality is probably valid. In addition to the Shapiro-Wilk test, the Quantile-Quantile (QQ) plot of the theoretical quantiles from the normal distribution vs. the sample quantiles are used to visually assess normality. The histogram of the residuals is also often used. All these tests are in the stats library of R[43].

The Durbin-Watson statistic tests the autocorrelation in regression analysis[2]. It mainly tests whether the residuals are independently distributed. The statistic is defined as:

$$d = \frac{\sum_{t=1}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \quad (10)$$

If the value of d is near 2, there is strong evidence that the residuals are not autocorrelated. A graphical representation of the residual values vs. predicted values determines whether any trend or pattern exists that may indicate a model misfit.

Model fit: The model fit of a regression analysis is assessed using the overall model p-value from the ANOVA table, the R^2 value, the adjusted R^2 value, the Mean Square Error (MSE) and the parameter estimates. R^2 can be interpreted as the proportion of variation in the dependent variable that is accounted for by the independent variables in the regression model[2]. It is given by:

$$d = 1 - \frac{SS_E}{SS_T} \quad (11)$$

SS_T and SS_E the total sum of squared deviation and sum of squared residuals, respectively, given by:

$$SS_T = \sum(y_i - \bar{y})^2 \quad (12)$$

$$SS_E = \sum(y_i - \hat{y}_i)^2 \quad (13)$$

According to Chatterjee *et al*.[2], the adjusted R^2 is used to compare models with differing number of predictor variables. For n observations and p variables, it is defined as:

$$R_a^2 = 1 - \frac{SS_E/(n-p-1)}{SS_T/(n-p)} \quad (14)$$

The MSE shows the average over all n residual values and is given by:

$$MSE = \frac{1}{n} SS_E \quad (15)$$

A lower MSE value a better fit. The estimates of the regression parameters are also assessed in terms of their significance and interpretability. Since, the data are simulated such that all three independent variables are positively correlated with the dependent variables, the linearity of the predictors should be correctly reflected in the good model.

RESULTS AND DISCUSSION

This study presents the analysis produced by the data simulated. It includes the results of the correlation matrix to evidence multicollinearity in the data.

Data structure: Multivariate random normal data were simulated to signify the interrelationship between four psychological measurements: academic achievement (Y), LOC (X_1), self-esteem (X_2) and intellect (X_3). All variables were simulated with mean zero and variance one. The tables follow below for illustrating the results. Table 1 shows the summary statistics. The means and standard deviations are nearly zero and one respectively. Table 2 gives the relationship between the academic achievement and cognitive measures. A strong linear relationship exists between the dependent variable (Y) and independent variables (X_i). The predictors LOC and self-esteem have a strong interrelation, r = 0.9742 which indicates that multicollinearity exists in the data.

Ordinary least squares regression: The baseline (Eq. 2) is first modelled in order to have an initial inspection of the regression analysis and assess the impact of multicollinearity. Diagnostic check for outlier is vital for accurate analysis. Table 3 shows that the assumption of normality is violated, since, the Shapiro-Wilk test is >0.05. This is due to the fact that data is highly correlated. The standard error of the estimate is large. To check for possible outliers the default SPSS is ±3 with one extreme observation if using ±2 it gets 16 outliers. Thus, the removal of outlier is not considered because there is a possibility that this points are important as part of the data.

Table 3 shows ANOVA output. ANOVA determines if the regression is significant. Since, p<0.05, the regression model is significant. Table 3 shows a good model with and R^2 = 0.9651. The Durbin-Watson tests if observations are independent. Its value of 1.9030 is close to 2. This indicates that the residuals are not correlated after the model is fitted. Table 4 shows that the regression equation is:

$$\text{Academic achievement} = 0.0032 + 0.9294\,\text{LOC} - \text{self esteem} + 0.5861\,\text{intellect}$$

The equation implies that academic achievement realization requires an increase in LOC by 0.9294, a decrease in self-esteem by 0.2646 and an increase in intellect by 0.5861. This opposes known realities where academic achievement requires a high LOC, Self-esteem and Intellect. The discrepancy in the parameter estimates is caused by the multicollinearity effect. This is indicated by the VIF which shows a severe multicollinearity problem. The correlation between LOC and intellect is positive but in the regression model the self-esteem coefficient is negative. VIFs of LOC and self-esteem exceed 10 which signals severe multicollinearity.

Table 1: Summary statistics

Variables	N	Minimum	Maximum	Mean	SD
Y	300	-2.7981	3.8893	0.0149	1.0432
X_1	300	-3.0539	2.6993	0.0127	1.0334
X_2	300	-3.2324	2.7838	0.0089	1.0263
X_3	300	-2.3913	3.0730	0.0037	0.9736

Table 2: Correlation matrix

Variables	Y	X_1	X_2	X_3
Y	1.0000	0.8126	0.7336	0.7459
X_1	0.8126	1.0000	0.9742	0.2663
X_2	0.7336	0.9742	1.0000	0.1775
X_3	0.7459	0.2663	0.1775	1.0000

Table 3: OLS Model summary

Assumptions	Values
Shapiro-Wilk test	0.5787
Durban-Watson statistic	1.9030
ANOVA p-value	0.0000
R^2	0.9651
Adjusted R^2	0.9647
Standard error of the estimate	0.1959
Mean square error	0.0379

Table 4: OLS parameter estimates

Variables	B	SE	Beta	p-values	VIF
Constant	0.0032	0.0113	-	0.7753	-
LOC	0.9294	0.0535	0.9206	0.0000	23.8
Self-esteem	-0.2646	0.0528	-0.2604	0.0000	22.9
Intellect	0.5861	0.0130	0.5469	0.0000	1.3

Table 5 shows the necessary diagnostics to detect the effect of multicollinearity. All condition index values are below 15. However, the variables LOC and Self-esteem have large variance proportion. Thus, collinearity exist in spite the lower condition index value.

CI and VIF values surprising give slightly different results. However, since, the CI value is <15, the severity of multicollinearity is moderate. Because of the worth of LOC and Self-esteem if they are removed, the purpose of the regression analysis would not be fulfilled.

The residuals against the predicted values are not entirely randomly dispersed, violating the assumption of homoscedasticity. This means that the variance between the residual is not constant. This causes the cases with larger disturbance "noise" to have more "pull" than other observations.

Residual analysis: Plots for the task of residual analysis are presented next and then discussed thereafter.

Figure 1 shows in a histogram that the residuals are symmetrically distributed. Figure 2 displays a plot of residuals versus predicted values. Homoscedasticity is evidenced. Therefore, the residuals are not randomly distributed. Figure 3 is a normal QQ plot showing that the residuals are not all aligned in the x-y axis symmetry. This signals the violation of the normality assumption. Therefore, the residuals are not normally distributed.

Table 5: OLS condition indices

			Variance proportions			
Dim	Eig	CI	(Constant)	LOC	Self-esteem	Intellect
1	2.0669	1.0000	0.0001	0.0095	0.0095	0.0308
2	0.9998	1.4378	0.9996	0.0000	0.0000	0.0001
3	0.9118	1.5056	0.0000	0.0011	0.0029	0.8007
4	0.0216	9.7934	0.0003	0.9894	0.9875	0.1685

Dim = Dimension; Eig = Eigenvalue; CI-Condition Index, Con = Constant

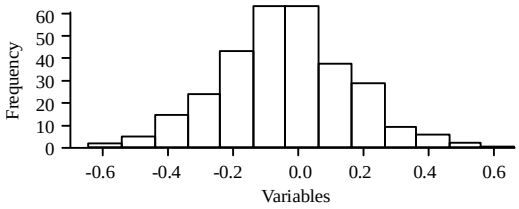

Fig. 1: Histogram of residuals

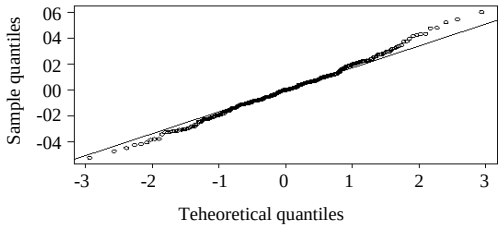

Fig. 2: Residuals vs. predicted values

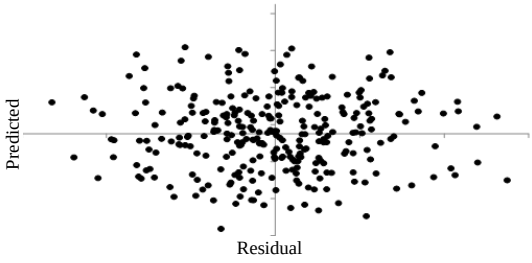

Fig. 3: Normal QQ plot

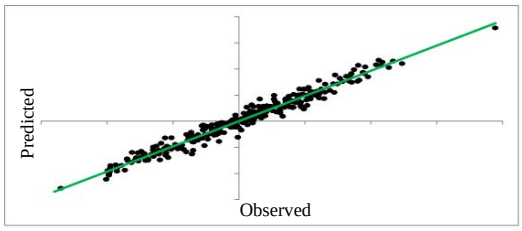

Fig. 4: Linear relationship between dependent and independent variables

Figure 4 is a regression line showing a strong linear relationship between the dependent and independent variables.

Stepwise regression: The tables below present the results of stepwise regression. Table 6 summaries forward stepwise regression of regression models. The R^2 from model 1-3 increases from 66-97%. However, the R^2 change decreases as each variable is added per model due to multicollinearity that causes parameter estimates to be unstable. Also, when a variable is added to a model, it decreases the values of standard errors and MSEs. This isolates model 2 as the best regression model.

Table 7 also shows that model 2 is the best. The R^2 value increases when Intellect variable is added with p = 0.05 and the parameter estimate signs are in the same direction. Also, the VIF drastically decreases which signals the multicollinearity problem. There is a major change in the residual QQ plot. It displays a normally distributed pattern amongst the residuals. The results on both (Fig. 3 and 4) are unchanged probably because of outlier influences.

Principal component regression: The tables below present the results. Table 9 is the PCA output whereby three components were extracted. Component 1 has the largest variance of 68.9% followed by Components 2 and 3 with variances of 30.4 and 0.7%, respectively. Components 1 and 2 effectively explain most of the variation of 99.3%. The dimensionality of the data can be reduced with little loss of information, thereby eliminating multicollinearity.

Table 10 shows that Component 1 constitutes LOC and self-esteem variables and Component 2 consists of the intellect variable only. Component 1 is associated with internal motivation and Component 2 has to do with the IQ of the respondent.

Table 11 is an output of the regression using the principal components 'component 1' and 'component 2' as the input variables. This model shows that it is statistically significant with $p<0.05$. The assumption of normality is violated since the Shapiro-Wilk test is not less 0.05. Hence, the assumption of independence is not violated. This is an indication that correlation does not exist between the components. The PCR model is a good model with $R^2 = 95\%$.

Table 12 shows the sign of direction on parameter estimates as the same and with no multicollinearity. The p-value of the constant is non-significant with value of 0.266. This indicates that the model will be without the intercept that is. Table 13 shows the PCR

Table 6: Stepwise regression model summary

	Models		
Variables	1	2	3
Assumptions			
Shapiro-Wilk test p-value	0.8084	0.6444	0.5787
Durban-Watson statistic	2.0423	1.9082	1.9030
Model			
Variables in model	LOC	LOC+Intellect	LOC+Intellect+ Self-esteem
ANOVA p-value	0.0000	0.0000	0.0000
R^2	0.6604	0.9621	0.9651
R^2 change	0.6604	0.3017	0.0030
Adjusted R^2	0.6592	0.9619	0.9647
Standard error of the estimate	0.6090	0.2037	0.1959
Mean square error	0.3684	0.0411	0.0379

Table 7: Stepwise regression parameter estimates

Variables	B	SE	Beta	p-values	VIF
Constant	0.0041	0.0118		0.7260	
LOC	0.6672	0.0118	0.6609	0.0000	1.0763
Intellect	0.6107	0.0126	0.5699	0.0000	1.0763

Table 8: Stepwise regression condition indices

			Variance proportions		
Dimension	Eigen value	Condition index	(Constant)	LOC	Intellect
1	1.2668	1.0000	0.0014	0.3663	0.3657
2	0.9996	1.1257	0.9979	0.0002	0.0020
3	0.7336	1.3141	0.0007	0.6335	0.6323

Table 9: PCA total variance explained

	Initial eigen values		
Components	Total	Variance (%)	Cum (%)
1	2.0667	68.89	68.89
2	0.9118	30.39	99.28
3	0.0216	0.72	100.00

Table 10: Component matrix

Variables	Component 1	Component 2	Component 3
LOC	0.9839	-0.1450	-0.1047
Self-esteem	0.9662	-0.2366	0.1024
Intellect	0.4064	0.9137	0.0099

Table 11: PCR Model summary

Models	Values
Assumptions	
Shapiro-Wilk test p-value	0.5846
Durban-Watson statistic	1.9000
Models	
ANOVA p-value	0.0000
R^2	0.9513
Adjusted R^2	0.9509
Standard error of the estimate	0.2311
Mean square error	0.0529

condition indices to be <15. This shows that multicollinearity is minimal at a value of 1 in 3 dimensional surfaces.

Ridge regression: The ridge regression was done for 21 different ridge parameters, ranging from 0-1 with an increment of 0.05. The ridge trace shows the regression coefficients or parameter estimates for all 3 independent

Table 12: PCR parameter estimates

Variables	B	SE	Beta	p-values	VIF
Constant	0.0149	0.0133		0.267	
Internal motivation	0.9144	0.0134	0.8765	0.000	1.00
Intellect	0.4463	0.0134	0.4278	0.000	1.00

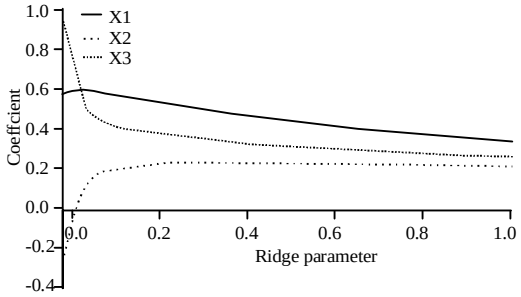

Fig. 5: Ridge trace

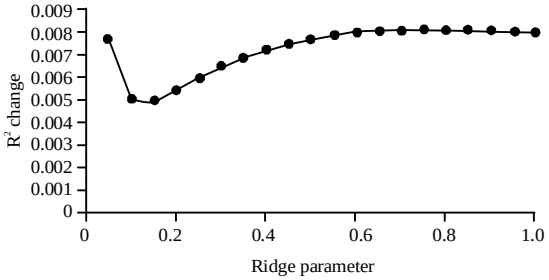

Fig. 6: Ridge trace

variables across all ridge parameters. These assist to determine the place where the lines stabilize. The figures below are used for this purpose.

It seems from (Fig. 5-8) that as though at the ridge parameter of 0.4, the lines stabilize. Figure 5-8 show the change in the estimates and the change in R^2 between ridge parameters. From these graphs it looks like 0.4 is good enough.

Figure 6 shows that the ridge parameter controls which regression coefficient is estimated the least. At ridge parameter = 0.4, the regression coefficient of variable intellect is least efficient in the estimation.

Table 13: PCR condition indices

			Variance proportions		
Dimension	Eigen value	Condition index	(Constant)	Internal motivation	Intellect
1	1.0000	1.0000	0.0000	1.0000	0.0000
2	1.0000	1.0000	1.0000	0.0000	0.0000
3	1.0000	1.0000	0.0000	0.0000	1.0000

Table 14: Ridge regression model summary

Models	Values
Assumptions	
Shapiro-Wilk test p-value	0.9919
Durban-Watson statistic	1.9981
Models	
ANOVA p-value	0.0000
R^2	0.9154
Adjusted R^2	0.9146
Mean square error	0.4912

Table 15: Ridge regression parameter estimates

Variables	B	SE	p-values
Constant	0.0067	-	-
LOC	0.3347	0.1467	0.0000
Self-esteem	0.2415	0.1504	0.0000
Intellect	0.4710	0.2176	0.0000

Table 16: Model comparison

Models	OLS	Step reg	PCR	Ridge
R^2	0.9651	0.9621	0.9513	0.9154
MSE	0.0379	0.0411	0.0529	0.4912
Constant	0.0032	0.0041	0.0149	0.0067
LOC	0.9294	0.6672	0.9144	0.3347
Self-esteem	-0.2646	NA		0.2415
Intellect	0.5861	0.6107	0.4463	0.4710

OLS = Ordinary Least Squares; Step Reg = Stepwise Regression; PCR = Principal Components Regression; Ridge = Ridge regression

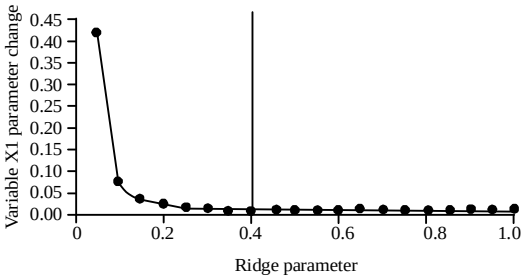

Fig. 7: Ridge trace

Figure 7 indicates the LOC parameter change from 0.41 to 0 when ridge parameter = 0.4. Figure 8 shows that self-esteem parameter changes from -0.41-0 when ridge parameter = 0.4. In Fig. 9, intellect parameter change from -0.01-0.015 when ridge parameter = 0.4. In, Fig. 10 of residual vs. predicted, a random spread is not indicated. This could be due to the resulting penalization of the regression.

Table 14 demonstrates that ridge regression model is statistically significant with p-<0.05. The assumption of normality is violated, since, the Shapiro-Wilk test is not less 0.05. Hence, the assumption of independence is not

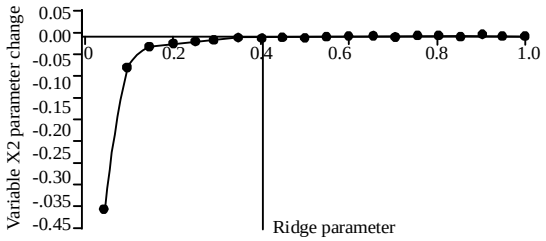

Fig. 8: Ridge trace

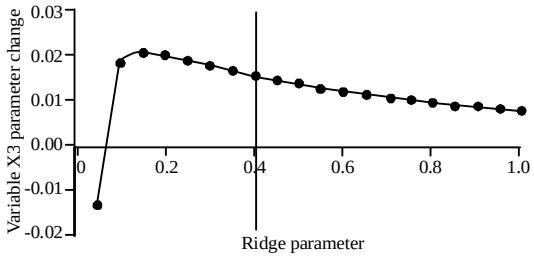

Fig. 9: Ridge trace

violated. This is an indication that correlation does not exist between the components. The PCR Model is a good model with R^2 = 92%.

Table 15 shows the sign of direction on parameter estimates as the same and with no multicollinearity. The p-value of the constant is significant with value of 0.0067. This shows that the model will have a nonzero intercept.

Model comparison: The OLS Model has the highest R^2 = 0.97% and least MSE = 0.0379 compared to the other three models. This identifies OLS as the best model. However, OLS showed shortcomings such as parameter estimate instability and collinearity.

Stepwise regression model also showed suitability with R^2 = 96% and MSE = 0.041. Thus, it has the smallest variance. Its main problem though is that it excludes one of the variables (Self-esteem) that is vital for inferences on psychometrical measures. PCR Model produced R^2 = 95% and MSE = 0.052. The parameter estimate had the same sign of direction. This meant that the variables are independent.

Lastly, ridge regression has a good R^2 = 0.91 but its MSE = 0.491 is the highest amongst the four models. This model showed to be having stable parameter estimates,

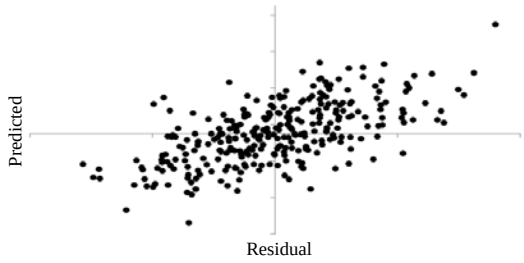

Fig. 10: Residual vs. predicted for ridge regression model

the assumptions of normality and random distribution hold. However, since, its residuals showed a linear shape, it suggests that a transformation should be performed. From the four models, PCR is superior to predict the academic achievement of school learners.

CONCLUSION

Multicollinearity is a common problem in research and statistical analyses. It is particularly significant when the aim is to predict a dependent variable among a set of independent that share some characteristics. The OLS Model produced in this research paper, this model has the highest R^2 among the other models but with imprecise regression coefficients.

The advantage of stepwise regression model is that it can produce as many different models using variable selection. Three psychometric measures were examined but Model 2 excluded self-esteem. This expresses that academic achievement can be obtained regardless of whether a learner has self-esteem or not. This model is thus biased towards learners who have low self-esteem. Ridge regression displayed the least parameter estimates in a model in order to eliminate those that are correlated and it can also handle data with outliers. But for this data ridge regression performed the least in this data analysis and it did not produce the desired results.

The PCR firstly removes collinearity factor in that data and then fit a regression model using uncorrelated variables 'components'. This model can explain 95 % of variability.

RECOMMENDATIONS

It is recommended that diagnostic analysis should be performed firstly in order recognise any existence of multicollinearity. Also, other models available in modern literature such as partial least squares and factor analysis using different rotation should be tried out in order to remedy the effects of multicollinearity.

ACKNOWLEDGEMENTS

The Department of Actuarial Science and Mathematical Statistics of the University of the Witwatersrand, Johannesburg provided knowledge and research training to the first author in the methods used in the sudy during a formal study towards the BSc Honours degree. That knowledge has proved handy during this study.

REFERENCES

01. Alan, O.S., 1993. An introduction to regression analysis. Master Thesis, University of Chicago Law School, Chicago, Illinois.
02. Chaterjee, S., A.S. Hadi and B. Price, 2000. Regression Analysis by Examples. 3rd Edn. Wiley VCH, New York.
03. Bowerman, B.L. and R.T. O'Connell, 1990. Linear Statistical Models: An Applied Approach. 2nd Edn., Duxbury Press, Grove, California, USA., ISBN:9780534229856, Pages: 1024.
04. Dormann, C.F., J. Elith, S. Bacher, C. Buchmann and G. Carl *et al*., 2013. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36: 27-46.
05. Naes, T. and B.H. Mevik, 2001. Understanding the collinearity problem in regression and discriminant analysis. J. Chemometrics, 15: 413-426.
06. Weisberg, S., 1985. Applied Regression Analysis. 2nd Edn., Wiley, New York, USA., Pages: 324.
07. Martens, M. and T. Naes, 1989. Multivariate Calibration. J. Wiley and Sons, Ltd., Chichester.
08. Fekedulegn, D., M.P.M. Siurtain and J.J. Colbert, 1999. Parameter estimation of nonlinear models in forestry. Silva Fennica, 33: 327-336.
09. Gujarati, D.N., 1999. Study guide for Essentials of Econometrics by Gujarati. 2nd Edn., McGraw-Hill, New York, USA., ISBN: 9780075619352, Pages: 534.
10. Farrar, D.E. and R.R. Glauber, 1967. Multicollinearity in regression analysis: The problem revisited. Rev. Econ. Stat., 49: 92-107.
11. Mela, C.F. and P.K. Kopalle, 2002. The impact of collinearity on regression analysis: The asymmetric effect of negative and positive correlations. Applied Econom., 34: 667-677.
12. Belsley, D.A., E. Kuh and R.E. Welsch, 1980. Regression Diagnostics: Identifying Influential Data and Sources of Colinearity. John Willey and Sons Inc., New York.
13. Greene, W.H., 1990. Econometric Analysis. 2nd Edn., Macmillan Publishing Company, New York, USA., ISBN: 9780023463907, Pages: 783.
14. Kmenta, J., 1986. Elements of Econometrics. 2nd Edn., Macmillan Publishers, New York, pp: 55-115.
15. Lafi, S.Q. and J.B. Kaneene, 1992. An explanation of the use of principal-components analysis to detect and correct for multicollinearity. Preventive Vet. Med., 13: 261-275.

16. Liu, R.X., J. Kuang, Q. Gong and X.L. Hou, 2003. Principal component regression analysis with SPSS. Comput. Methods Programs Biomed., 71: 141-147.
17. O'brien, R.M., 2007. A caution regarding rules of thumb for variance inflation factors. Qual. Quantity, 41: 673-690.
18. Tsutsumi, M., E. Shimizu and Y. Matsuba, 1997. A comparative study on counter-measures for multicollinearity in regression analysis. J. East. Asia Soc. Transp. Stud., 2: 1891-1904.
19. Paul, R.K., 2006. Multicollinearity: Causes, Effects and Remedies. IASRI, New Delhi, India,.
20. Luetjohann, H., 1968. The stepwise regression algorithm seen from the statistician's point of view. Research Memoranda No. 11, Institutional Repository at HIS, Austria, Europe. https://irihs.ihs.ac.at/id/eprint/11/1/fo11.pdf.
21. Graham, M.H., 2003. Confronting multicollinearity in ecological multiple regression. Ecology, 84: 2809-2815.
22. Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. J. Educ. Psychol., 24: 498-520.
23. Hotelling, H., 1957. The relations of the newer multivariate statistical methods to factor analysis. Br. J. Stat. Psychol., 10: 69-79.
    Jackson, J.E., 1991. A User's Guide to Principal Components. Wiley, New York.
24. Jolliffe, I.T., 2002. Principal Component Analysis. 2nd Edn., Springer-Verlag, New York, USA.
25. Kendall, M.G., 1957. A Course in Multivariate Analysis. 5th Edn., Griffin, London, UK., Pages: 185.
26. Calmes, J., 2010. Obama calls for new Sputnik Moment. New York Times, New York, USA. https://thecaucus.blogs.nytimes.com/2010/12/06/obama-calls-for-new-sputnik-moment/
27. Kornblut, A.E. and S. Wilson, 2011. State of the Union 2011: Win the future,'Obama says. The Washington Post, Washington DC., USA.
28. Mansfield, E.R., J.T. Webster and R.F. Gunst, 1977. An analytic variable selection technique for principal component regression. J. Royal Stat. Soc., 26: 34-40.
29. Gunst, R.F. and R.L. Mason, 1980. Regression Analysis and its Application: A Data-Oriented Approach. Marcel Dekker, New York, USA., ISBN: 9780824769932, Pages: 524.
30. Massy, W.F., 1965. Principal components regression in exploratory statistical research. J. Am. Stat. Assoc., 60: 234-256.
31. Jeffer, J.N.R., 1967. Two case studies in the application of principal component analysis. J. R. Statist. Soc. Ser. C (Applied Statist.), 16: 225-236.
32. Hawkins, D.M., 1973. On the investigation of alternative regressions by principal component analysis. J. R. Stat. Soc., 22: 275-286.
33. Hocking, R.R., 1976. The analysis and selection of variables in linear re-gression. Biometrics, 32: 1-49.
34. Niemela-Nyrhinen, J. and E. Leskinen, 2014. Multicollinearity in marketing models: Notes on the application of ridge trace estimation in structural equation modelling. Electron. J. Bus. Res. Methods, 12: 3-15.
35. Mahajan, V., A.K. Jain and M. Bergier, 1977. Parameter estimation in marketing models in the presence of multicollinearity: An application of ridge regression. J. Marketing Res., 14: 586-591.
36. Gasic-Pavisic, S., S. Joksimovic and D. Janjetovic, 2006. General self-esteem and locus of control of young sportsmen. Proc. Inst. Educ. Res., 38: 385-400.
37. Ahmad, I., A. Zeb, S. Ullah and A. Ali, 2012. Relationship between Self-Esteem and academic achievements of students: A case of government secondary schools in district Swabi, KPK, Pakistan. Int. J. Social Sci. Educ., 3: 361-369.
38. Bar-Tal, D. and Y. Bar-Zohar, 1977. The relationship between perception of locus of control and academic achievement: Review and some educational implications. Contemp. Educ. Psychol., 2: 181-199.
39. Venables, W.N. and B.D. Ripley, 2002. Modern Applied Statistics with S. 4th Edn., Springer, Berlin, Germany, ISBN:9780387954578, Pages: 495.
40. Johnson, R.A. and D.W. Wichern, 2007. Applied Multivariate Statistical Analysis. 6th Edn., Pearson Prentice Hall, New York, ISBN-13: 978-0131877153, pp: 800.
41. Hoerl, A.E. and R.W. Kennard, 1970. Ridge regression: Biased estimation for non-orthogonal problems. Technometrica, 12: 55-88.
42. Shapiro, S.S. and M.B. Wilk, 1965. An analysis of variance test for normality (Complete samples). Biometrika, 52: 591-611.
43. R Core Team, 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.