

**DEVELOPMENT OF SOME INFLUENTIAL MEASURES
IN ONE PARAMETER NEW RIDGE-TYPE ESTIMATORS
OF LINEAR REGRESSION MODELS**

OLUWAROHUNBI Grace Olufunke
MATRIC NO: PSC/F/22/23/0033
B.Sc. (ILORIN), M.Sc. (OGUN)

A THESIS SUBMITTED TO THE DEPARTMENT OF STATISTICS

**IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF
THE DEGREE OF DOCTOR OF PHILOSOPHY IN STATISTICS, OLABISI
ONABANJO UNIVERSITY, AGO-IWOYE OGUN STATE, NIGERIA**

SEPTEMBER, 2026

ABSTRACT

In linear regression analysis, influential points are observations that exert disproportionate impact on regression models. If not properly identified and addressed, they can distort parameter estimates, fitted values, and statistical inferences. Although, Ordinary Least Squares (OLS) estimator is widely used for estimating parameter coefficients in regression analysis, it performs poorly when multicollinearity and outliers are present. Many existing biased estimators are developed to mitigate only multicollinearity, while robust regressions are used to reduce the effect of outliers on regression estimates. In practice, these two problems often occur together. Similarly, traditional influential diagnostic measures developed under the OLS framework, such as Cook's Distance (CD) and Difference in Fits (DFFITS), may not perform well when multicollinearity and outliers exist simultaneously. Hence, this study developed influential diagnostic measures within One-parameter new ridge-type estimator's framework, to improve detection of influential points when multicollinearity and outliers co-exist in a dataset.

The Kibria–Lukman (KL) estimator was combined with six robust regression techniques, namely: Modified Maximum Likelihood (MM), Least Absolute Deviation (LAD), Least Median of Squares (LMS), Least Quartile of Squares (LQS), Least Trimmed Squares (LTS), and S estimators to produce a class of Robust Kibria–Lukman (RKL) estimators. Based on these estimators, generalized forms of Cook's Distance, DFFITS, Difference in Betas (DFBETAS), Covariance Ratio's, Hadi's, Pena's, and Atkinson's statistics were derived, along with their approximate case-deletion expressions for Cook's Distance and DFFITS. Their performance were evaluated through a Monte Carlo simulation study, replicated 1,000 times, with sample sizes from 10 to 500, correlation levels from 0.80 to 0.99, outlier magnitudes from 0 to 10, and contamination rates of 10% and 20%.

The findings showed that the Covariance Ratio based on the Robust Kibria–Lukman_M estimator (CVR_KL_M) was the overall best-performing diagnostic measure, with the highest total frequency of 100% detection (2,622). It was followed by CVR_KL_LAD (2,603) and CVR_KL_MM (2,325). CVR_KL_M also showed stable performance across different sample sizes, with detection frequencies of 314 at $n=10$, 330 at $n=20$, 30, 50, 100 and 250 and 329 at $n=40$ and 500. In real life data applications, the proposed measures identified observations 2, 16, 1, 12, and 9 as influential in the economic dataset, while several existing methods failed to detect observations 2, 12, and 9. In the Hussein dataset, the proposed methods consistently identified observations 15, 20, 16, 12, and 4 as influential.

The study concluded that the proposed Robust Kibria–Lukman influential measures were more effective, stable, and reliable than the existing diagnostic measures for detecting influential points in a dataset affected by both multicollinearity and outliers. Particularly, CVR_KL_M emerged as the most efficient diagnostic measure. The study therefore provides a statistically sound and practically useful approach for identifying influential points in linear regression models under these combined conditions. It is therefore recommended that applied researchers and data analysts should use the proposed Robust Kibria–Lukman-based diagnostic measures, especially CVR_KL_M, when analysing regression data affected by both multicollinearity and outliers.

Keywords: Influential measures, Influential points, Multicollinearity, Outliers, Regression analysis

Word count: 496 words

CERTIFICATION

This is to certify this research work presented in this thesis was carried out by OLUWAROHUNBI Grace Olufunke with Matric No: PSC/F/22/23/0033, under our supervision in Department of Statistics, Olabisi Onabanjo University, Ago-Iwoye, Nigeria.

Prof. Timothy. O. Olatayo

Main Supervisor

.....

Signature/ Date

Dr. Abass. I. Taiwo

Co-supervisor

.....

Signature / Date

Dr. Abass. I. Taiwo

Head of Department

.....

Signature/ Date

DEDICATION

This thesis is dedicated to Almighty God and to my husband who is my benefactor.

ACKNOWLEDGEMENTS

To the Immortal, Invisible and only wise God be all the glory, honour and praise for making my dream of earning a Ph.D degree in Statistics to become a reality. Indeed God shows mercy to whom He chose to show mercy. He came through for me at every difficult turning point of this pursuit. Hence, I say Hallelujah.

My profound gratitude goes to my main supervisor, an aptitude scholar, Prof. Timothy. O. Olatayo for making available his wealth of experience. His guidance, push and support saw this research to its successful completion. Sir, I am greatly indebted to you. Similarly, I am grateful to Dr. Abass. I. Taiwo, the co-supervisor of this thesis and the Head of Department of Statistics. Your availability, listening ears and useful criticism is highly appreciated.

My sincere thanks to Dr. Biodun. T. Efuwape and Dr Emmanuel F. Ologunleko whose wealth of experience help to shape this thesis. Your labour of love would not go unrewarded. I thank profusely Dr Adedayo F. Adedotun, the Departmental Post Graduate Coordinator who ensured that the Post Graduate format was meticulously followed in fulfillment of the school requirement to earn a Ph.D. God bless you sir.

Additionally, I thank Mr. Ridwan. O. Raheem, Mr. Hammed .B. Buari and Mr. Adekanmi. A. Oyewole who were always available during seminars and gave their useful suggestions to better this work. Sirs, your labour will not be forgotten in a hurry. Also, I am grateful to Mr. L. O. Idris for his readiness to provide every needed Administrative assistance in Department of statistics.

I express my sincere gratitude goes to Prof Samuel Olutunde Odetunde, Dean, Faculty of Science who stirred me on. Your support is very well appreciated, sir. I say thank you to

Dr. T. J. Adejumo who made available many materials used for the research. God bless you, sir.

I profusely thank Dr. (Mrs.) J. O. Oyebanji, my sister-in-law, and her husband, Prof. W. O. Oyebanji, your belief in me kept me on my toes. Thank you sir and ma for cheering me on. I thank Prof. Teju Jolayemi, who encouraged me to earn more degrees in the field of Statistics. Sir, your encouragement and the solid background given to me in my undergraduate kept me going. I celebrate you sir.

To the Women Organization Secretary, Mrs. Taiwo Muyiwa, Dr. (Mrs.) Wuraola Ogunleye, Mrs. Fatima Adewole and Mr. Ibrahim Rmanla. You facilitated the work in your different capacities and your commitment to duty is highly commendable. May you really be blessed. Thank you Mr. Gideon Dauda, the Bishop's driver who drove me to school tirelessly.

I say a "big thank you" to my husband, The Rt. Revd Michael. A Oluwarohunbi, my backbone, you envisioned and nurtured this dream to reality. You provided necessary logistics to facilitate this work. Undoubtedly, you are my treasure and the elder brother I never had. Surely, you will reap the fruit of your labour of love. Again, I say thank you, my love.

To my wonderful children, Oladipo, Temiloluwa and Modupe, your encouragement made it happen. May you succeed in all your endeavours. Amen.

TABLE OF CONTENTS

Content	Page
Title Page	i
Abstract	ii
Certification	iv
Dedication	v
Acknowledgements	vi
Table of Contents	viii
List of Tables	xi
List of Figures	xiii
Abbreviations	xiv
 CHAPTER ONE: INTRODUCTION	
1.1 Background of the Study	1
1.2 Statement of the Problem	10
1.3 Aim and Objectives of the study	11
1.4 Significance of the study	11
1.5 Scope of the Study	12
 CHAPTER TWO: LITERATURE REVIEW	
2.1 Concept of Outliers and its Detection	13
2.2 Robust Regression Techniques to Handle Outlying Cases	17
2.3 Review of Some One-Parameter Estimators	20
2.4 Review of some Two-parameter Estimators	24
2.5 Review of Some Existing Influential Diagnostic Measures	25

CHAPTER THREE: MATERIALS AND METHOD

Content	Page
3.1 Materials	32
3.2 The Development of Influential Measures in Robust Regression RKL Estimators	34
3.2.1 The Canonical form of Generalized RKL Estimators	35
3.2.2 Determination of Mean Squares Error (MSE) for (RKL) Estimators	36
3.2.3 Generalized Version of Robust Cook's D in Robust Kibria-Lukman (RKL) Estimator	39
3.2.4 Generalized Version of Robust DFFITs for the RKL Estimator	40
3.2.4.1 Derivation of Approximate Case Deletion formula for RKL Estimator	42
3.2.5 Generalized Version of Robust Covratio Measure in RKL Estimator	44
3.2.6 Generalized Version of Robust Hadi's Measure for the RKL Estimator	45
3.2.7 Generalized Version of Robust Pena's Statistic in RKL Estimator	46
3.2.8 Generalized Version of Robust Atkinson Diagnostic in RKL Estimator	47
3.3 Simulation procedures to generate exogenous variables, response variable and estimation of Mean Squares Error.	48
3.3.1 Parameters for Simulation	49
3.3.2 Algorithm used to generate the Explanatory variables, Dependant variable, Error terms, and Mean Square Error	50
3.4 Determination of Influential point and Percentage of influential point detected	51
3.4.1 Criterion for Investigation of the Proposed Influential Diagnostic Tools	51

CHAPTER FOUR: RESULTS AND DISCUSSION

Content	Page
4.1 Investigation of Proposed Influential Measures in the presence of multicollinearity and outliers	52
4.2 Investigation of the Influential measures without Robust Regression	53
4.3 Discussion	111

CHAPTER FIVE: SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Summary	114
5.2 Conclusion	115
5.3 Recommendations	116

References

Appendices

LIST OF TABLES

Table	Title	Page
3.1:	Summary of Mathematics of Generalised version of some proposed Influential Measures in RKL estimator	46
4.1:	Frequency of the influential measures proposed based on OLS estimator that detected all Influential points (100%), when counted over all the specifications	53
4.2a:	Frequency of the influential measures proposed based on Robust-M estimator that detected all Influential points (100%),when counted over all the specifications	57
4.2b:	Frequency of the influential measures proposed based on Robust-MM estimator that detected all Influential points (100%), when counted over all the specifications	59
4.2c:	Frequency of the influential measures proposed based on Robust-LAD estimator that detected all Influential points (100%), when counted over all the specifications	61
4.2d:	Frequency of the influential measures proposed based on Robust-LMS estimator that detected all Influential points (100%),when counted over all the specifications	63
4.2e:	Frequency of the influential measures proposed based on Robust-LQS estimator that detected all Influential points (100%), when counted over all the specifications	65
4.2f:	Frequency of the influential measures proposed based on Robust-LTS estimator that detected all Influential points (100%), when counted over all the specifications	67
4.2g:	Frequency of the influential measures proposed based on Robust-S estimator that detected all Influential points (100%), when counted over all the specifications	69
4.3a:	Summary of best influential measure at each sample size when counted over all the specifications for Robust-M estimator	85
4.3b:	Summary of best influential measure at each sample size when counted over all the specifications for Robust-MM estimator	87
4.3c:	Summary of best influential measure at each sample size when counted over all the specifications for Robust-LAD estimator	89
4.3d:	Summary of best influential measure at each sample size when counted over all the specifications for Robust-LMS estimator	91
4.3e:	Summary of best influential measure at each sample size when counted over all the specifications for Robust-LQS estimator	93
4.3f:	Summary of best influential measure at each sample size when counted	95

	over all the specifications for Robust-LTS estimator	
4.3g:	Summary of best influential measure at each sample size when counted over all the specifications for Robust-S estimator	97
4.4:	Overall best Influential Measures when counted over all the specifications and sample sizes at each sample size for each robust regression estimators	99
4.5:	Best Influential Measures when counted over all the specifications at each sample sizes and across all robust regressions	103
4.6:	Five most influential observations based on some already existing influential Diagnostic Tools when compared with the newly Developed ones using Economic data set.	106
4.7:	Five most influential points detected by already existing influential measures when compared with the newly proposed ones, using Hussein's data	109

LIST OF FIGURES

Figure	Title	Page
4.1:	The graphical representation of best influential measures proposed based on OLS that detected all Influential points (100%), when counted over all specifications	55
4.2a:	The graphical representation of best influential measures proposed based on Robust-M estimator that detected all Influential points (100%), when counted over all specifications.	71
4.2b:	The graphical representation of best influential measures proposed based on Robust-MM estimator that detected all Influential points (100%), when counted over all specifications	73
4.2c:	The graphical representation of best influential measures proposed based on Robust-LAD estimator that detected all Influential points (100%), when counted over all specifications.	75
4.2d:	The graphical representation of best influential measures proposed based on Robust-LMS estimator that detected all Influential points (100%), when counted over all specifications.	77
4.2e:	The graphical representation of best influential measures proposed based on Robust-LQS estimator that detected all Influential points (100%), when counted over all specifications.	79
4.2f:	The graphical representation of best influential measures proposed based on Robust-LTS estimator that detected all influential points (100%), when counted over all specifications.	81
4.2g:	The graphical representation of best influential measures proposed based on Robust-S estimator that detected all Influential points (100%), when counted over all specifications.	83
4.3:	Bar chart of overall best Influential Measures when counted over all the specifications, for all robust regressions.	101

ABBREVIATIONS

Abbreviations	Full Meaning
ATK	Atkinson statistic
ATK_KL	Atkinson in Kibria-Lukman Estimator
ATK_KL_LAD	Atkinson in Robust-LAD Kibria-Lukman
ATK_KL_LMS	Atkinson in Robust-LMS Kibria-Lukman
ATK_KL_LQS	Atkinson in Robust-LQS Kibria-Lukman
ATK_KL_LTS	Atkinson in Robust-LTS Kibria-Lukman
ATK_KL_M	Atkinson in Robust-M Kibria-Lukman
ATK_KL_MM	Atkinson in Robust-MM Kibria-Lukman
ATK_KL_S	Atkinson in Robust-S Kibria-Lukman
ATK_LAD	Atkinson in Robust-LAD
ATK_LMS	Atkinson in Robust-LMS
ATK_LQS	Atkinson in Robust-LQS
ATK_LTS	Atkinson in Robust-LTS
ATK_M	Atkinson in Robust-M
ATK_MM	Atkinson in Robust-MM
ATK_S	Atkinson in Robust-S
CKD	Cook's Distance
CKD_KL	Cook's D in Kibria-Lukman Estimator
CKD_KL_LAD	Cook's D in Robust-LAD Kibria-Lukman
CKD_KL_LMS	Cook's D in Robust-LMS Kibria-Lukman
CKD_KL_LQS	Cook's D in Robust-LQS Kibria-Lukman
CKD_KL_LTS	Cook's D in Robust-LTS Kibria-Lukman

CKD_KL_M	Cook's D in Robust-M Kibria-Lukman
CKD_KL_MM	Cook's D in Robust-MM Kibria-Lukman
CKD_KL_S	Cook's D in Robust-S Kibria-Lukman
CKD_LAD	Cook's D in Robust-LAD
CKD_LMS	Cook's D in Robust-LMS
CKD_LQS	Cook's D in Robust-LQS
CKD_LTS	Cook's D in Robust-LTS
CKD_M	Cook's D in Robust-M
CKD_MM	Cook's D in Robust-MM
CKD_S	Cook's D in Robust-S
COVRATIO	Covariance Ratio
CVR	COVRATIO measure
CVR_KL	COVRATIO in Kibria-Lukman Estimator
CVR_KL_LAD	COVRATIO in Robust-LAD Kibria-Lukman
CVR_KL_LMS	COVRATIO in Robust-LMS Kibria-Lukman
CVR_KL_LQS	COVRATIO in Robust-LQS Kibria-Lukman
CVR_KL_LTS	COVRATIO in Robust-LTS Kibria-Lukman
CVR_KL_M	COVRATIO in Robust-M Kibria-Lukman
CVR_KL_MM	COVRATIO in Robust-MM Kibria-Lukman
CVR_KL_S	COVRATIO in Robust-S Kibria-Lukman
CVR_LAD	COVRATIO in Robust-LAD
CVR_LMS	COVRATIO in Robust-LMS
CVR_LQS	COVRATIO in Robust-LQS
CVR_LTS	COVRATIO in Robust-LTS
CVR_M	COVRATIO in Robust-M

CVR_MM	COVRATIO in Robust-MM
CVR_S	COVRATIO in Robust-S
DFFITS	Difference in fits
DFT	DFFITS
DFT_KL	DFFITS in Kibria-Lukman Estimator
DFT_KL_LAD	DFFITS in Robust-LAD Kibria-Lukman
DFT_KL_LMS	DFFITS in Robust-LMS Kibria-Lukman
DFT_KL_LQS	DFFITS in Robust-LQS Kibria-Lukman
DFT_KL_LTS	DFFITS in Robust-LTS Kibria-Lukman
DFT_KL_M	DFFITS in Robust-M Kibria-Lukman
DFT_KL_MM	DFFITS in Robust-MM Kibria-Lukman
DFT_KL_S	DFFITS in Robust-S Kibria-Lukman
DFT_LAD	DFFITS in Robust-LAD
DFT_LMS	DFFITS in Robust-LMS
DFT_LQS	DFFITS in Robust-LQS
DFT_LTS	DFFITS in Robust-LTS
DFT_M	DFFITS in Robust-M
DFT_MM	DFFITS in Robust-MM
DFT_S	DFFITS in Robust-S
HAD	Hadi's measure
HAD_KL	Hadi's measure in Kibria-Lukman Estimator
HAD_KL_LAD	Hadi's measure in Robust-LAD Kibria-Lukman
HAD_KL_LMS	Hadi's measure in Robust-LMS Kibria-Lukman
HAD_KL_LQS	Hadi's measure in Robust-LQS Kibria-Lukman
HAD_KL_LTS	Hadi's measure in Robust-LTS Kibria-Lukman

HAD_KL_M	Hadi's measure in Robust-M Kibria-Lukman
HAD_KL_MM	Hadi's measure in Robust-MM Kibria-Lukman
HAD_KL_S	Hadi's measure in Robust-S Kibria-Lukman
HAD_LAD	Hadi's measure in Robust-LAD
HAD_LMS	Hadi's measure in Robust-LMS
HAD_LQS	Hadi's measure in Robust-LQS
HAD_LTS	Hadi's measure in Robust-LTS
HAD_M	Hadi's measure in Robust-M
HAD_MM	Hadi's measure in Robust-MM
HAD_S	Hadi's measure in Robust-S
KL	Kibria-Lukman estimator
KL_LAD	LAD-estimator+Kibria-Lukman estimator
KL_LMS	LMS-estimator+Kibria-Lukman estimator
KL_LQS	LQS-estimator+Kibria-Lukman estimator
KL_LTS	LTS-estimator+Kibria-Lukman estimator
KL_M	M-estimator+Kibria-Lukman estimator
KL_MM	MM- estimator+Kibria-Lukman estimator
KL_S	S-estimator+Kibria-Lukman estimator
LAD	Least Absolute Deviation estimator
LE	LIU estimator
LMS	Least Median of Squares estimator
LQS	Least Quartile of Squares estimator
LTS	Least Trimmed Square estimator
M	M-estimator
MM	MM-estimator

MSE	Mean Square Error
OLS	Ordinary Least Squares
PEN	Pena's Statistics
PEN_KL	Pena's Statistics in Kibria-Lukman Estimator
PEN_KL_LAD	Pena's Statistics in Robust-LAD Kibria-Lukman
PEN_KL_LMS	Pena's Statistics in Robust-LMS Kibria-Lukman
PEN_KL_LQS	Pena's Statistics in Robust-LQS Kibria-Lukman
PEN_KL_LTS	Pena's Statistics in Robust-LTS Kibria-Lukman
PEN_KL_M	Pena's Statistics in Robust-M Kibria-Lukman
PEN_KL_MM	Pena's Statistics in Robust-MM Kibria-Lukman
PEN_KL_S	Pena's Statistics in Robust-S Kibria-Lukman
PEN_LAD	Pena's Statistics in Robust-LAD
PEN_LMS	Pena's Statistics in Robust-LMS
PEN_LQS	Pena's Statistics in Robust-LQS
PEN_LTS	Pena's Statistics in Robust-LTS
PEN_M	Pena's Statistics in Robust-M
PEN_MM	Pena's Statistics in Robust-MM
PEN_S	Pena's Statistics in Robust-S
RE	Ridge Estimator
S	S-estimator

CHAPTER ONE

INTRODUCTION

1.1 Background to the study

An influential point is any observation that has a strong impact on regression analysis such that its removal can have great effect on regression coefficient, prediction or the overall fitness of a model (Montgomery *et al.*, 2021). They are described as data points that meaningfully affect the regression model particularly when along with high leverage they have high residual (Fox. 2016). They are capable of controlling or disproportionate regression solution since they occupy extreme position in a data space (Draper and Smith, 1998). In other words, influential points are data points that can significantly affect various aspects in regression analysis, such as predicted responses, parameter coefficients estimates, and results of testing hypothesis (Cook and Weisberg, 1982). Regression analysis is widely used to examine relationships between dependent and independent variables across various fields (Montgomery *et al.*, 2021). However, influential points can substantially affect regression results, including the coefficient of determination, potentially leading to misleading conclusions. Therefore, detecting influential points is essential for reliable regression modelling (Belsley *et al.*, 1980). OLS estimator is a frequently used approach for parameter estimation because of its simplicity and efficiency under assumption like homoscedasticity, absence of multicollinearity, no autocorrelation and normality of errors (Seber and Lee, 2021). Nonetheless, some of these dataset frequently violate real-life dataset. High correlations between independent variables and they can inflate the variances of OLS estimators, resulting in an unstable and unreliable coefficient of parameters (Kutner et al., 2004).

Hoerl and Kennard (1970) developed the ridge regression estimator, which improved the OLS technique by adding a penalty term that shrinks coefficient estimates and reduces

variance in order to mitigate the impacts of high correlation in a regression model. Despite its advantages, ridge regression estimator remains sensitive to outliers. Conversely, robust regression techniques, which are designed to minimize the influence of extreme values and provide solutions for handling outliers, often fail to address multicollinearity effectively (Huber, 1981).

Outliers, on the other hand, are observations which deviate significantly from the trend of all other observations. They are capable of exerting disproportionate influence on Regression models and can lead to biased estimates and poor predictive performance.

Robust estimators are used to deal with Outliers so as to ensure that Coefficient estimates of a Regression model are unaffected by a small change in the dataset. Some examples of such Robust Estimators are Maximum likelihood-type, M (Huber, 1964), MM (Yohai, 1987), Least Trimmed Squares -LTS (Rousseeuw and Driessen, 1998) and Scaled - S (Rousseeuw and Yohai, 1984), Least Winsorized Squares -LWS (Rousseeuw and Leroy, 1987), Least Median of Squares -LMS (Rousseeuw, 1984), Least Quantile of Squares - LQS (Koenker and Bassett, 1984) and Least Absolute Deviations - LAD (Boscovich, 1757).

A one-parameter ridge-type estimator integrates the strengths of ridge regression and robust regression, providing a more reliable method for addressing both multicollinearity and outliers (Maronna *et al.*, 2006). When a dataset contains influential points, key considerations need to be done such as validating data accuracy and comparing regression models constructed with and without the influential point (Atkinson, 1985). . Thus, if the models generate inconsistent results, it is advisable to proceed with care. Calculating the regression equation with and without the outlier is one method for determining the influence of an outlier (Cook, 1977). Therefore, detecting influential observations is a

component of assessment that should be performed on any dataset, to determine the validity of such dataset; failure to examine this in regression analysis affect parameter estimation by increasing the variance of the parameters, thereby compromising the reliability and precision of estimated coefficients (Weisberg, 2005).

Cook's Distance (Cook, 1977) “,Difference in fits” -DFFITS (Welsch and Kuh, (1977), Studentized deleted residual (Welsch and Kuh, 1977), “Difference in betas’- DFBETAS (Belsley, Kuh & Welsch, 1980), COVRATIO - CVR (Belsley et al,1980), Atkinson Statistics -ATK (G Atkinson,1998), Hadi’s Measure - HAD (Ali Hadi, 1992), Pena’s statistic (PEN), proposed by Daniel Peña (2005), is one of the traditional diagnostic measures used to detect influential points in classical regression models.

While influential measures are essential tools in regression analysis, used to assess the validity of regression model, identify influential data points, improve model accuracy and detect hidden patterns or anomalies in a Dataset. Sometimes, it leads to detecting violations of assumptions such as homoscedasticity, normality, and linearity. Regression analysis is a widely used statistical method for analysing data , particularly for economic data. It is used to study explain the connection between endogenous variable and exogenous variable(s) - regressors. According to Gujarati (2003), Regression analysis explores connection between an endogenous variable and regressors.

When residual sum of squares is minimized, parameters in a regression analysis are estimated using the OLS estimator or Linear Least Squares approach. This method fits a line between the dataset to minimize the sum of the squares error between observed and fitted value. It is efficient when the following assumptions are meant:

- (i) Predictive variables must be measured without error.
- (ii) All the regressors are assumed to be stochastic.

- (iii) The variable e_i must be a real random variable.
- (iv) Every disturbance term e_i is independent of the explanatory variables i.e

$$E(e_i x_i) = 0.$$

The random error term e_i has zero mean value for each x_i symbolically $E(e_i) = 0$

- (v) The variance of each e_i is the same for all the x_i values, i.e $V(e_i) = \sigma^2$ constant.
- (vi) The values of each e_i should be normally distributed i.e $e_i \sim N(0, \sigma^2)$.
- (vii) The residuals are uncorrelated, i.e $Cov(e_i e_j) = 0$, for all $i \neq j$.
- (viii) Regression models should be correctly specified

Severe multicollinearity and outliers can compromise the reliability of OLS estimates. In classical regression models, multicollinearity happen in varying degrees, classified as perfect, high, and moderate correlation. It is possible, but not typical, for economic data to be completely free of multicollinearity problems. Perfect multicollinearity, on the other hand, is far less likely. When the explanatory variables have perfect linear relationship, then, multicollinearity is considered perfect. In this situation, Lukman (2017) confirmed that matrix $(X^T X)$ does not have full rank hence, its determinant as well as inverse do not exist. Consequently, OLS regression coefficients cannot be estimated and their standard errors are infinite.

In addition, Gujarati (1995) stated that in case of near perfect or high multicollinearity, the following consequences may likely occur:

- (i) OLS estimator may produce large variance and covariance despite the fact that it is Best Linear Unbiased Estimator (BLUE).
- (ii) Acceptance of null hypothesis as a result of wide range of confidence interval.
- (iii) Individual regression coefficients may be statistically insignificant despite a high and significant overall goodness of fit.

(iv) The OLS estimators can be sensitive to small changes in a dataset.

Furthermore, Lukman (2017) suggested that as the severity of multicollinearity increases, the determinant tends toward zero and its inverse increases significantly, resulting in round-off mistakes during computations of regression estimates. Hence, the regression coefficients may not be accurate, thus, rendering them imprecise due to huge variances.

Many literature have proposed several methods for detecting multicollinearity in linear regression. Some are:

- (i) when the coefficient of determination (R^2) is high but there are few significant t-ratios, the F-test will be rejected in some cases, indicating that the coefficient of parameters are simultaneously different from zero, especially when (R^2) is as high as 0.8, but individual t-test would show that few of coefficient of parameters are statistically insignificant. Gujarati (1995).
- (ii) High zero-order correlation between two regressors ($r > 0.8$). On this issue, Gujarati (1995) suggested that large zero-order correlations could indicate collinearity. It is not required for them to be high to have collinearity in any given circumstance. He further stated that it is sufficient but not a necessary for it to be said that multicollinearity exist because it might exist even when the zero-order correlations between two regressors are quite low (Gujarati, 1995; Lukman, 2017).
- (iii) Using Variance Inflation Factor (VIF); The equation is given as the inverse of 'one minor R_j^2 , the coefficient of determination obtained from regressing X_j on other regressors. If the Variance Inflation Factor (VIF) exceeds 10, that variable is said to be correlated (Gujarati, 1995; Lukman, 2017).
- (iv) Another methods of detecting if multicollinearity exist in the dataset, is the use of Condition Number (CN) and Condition index (CI). CN is defined as the square root

of the ratio of maximum eigenvalue to minimum eigenvalue. While CI is square root of the ratio of maximum eigenvalue of the j th eigenvalue.

A CN or CI between 10 and 30, signifies there exist moderate multicollinearity and when greater than 30, it indicates the existence of severe multicollinearity in a dataset (Gujarati, 1995 & Lukman, 2017).

On the other hand the occurrence of extreme value in a dataset could be the result of a potential violation in the assumed theory of a regression model and it calls for further investigation by the researcher. Also, the pathological appearance of outliers in a certain form of datasets, may indicate that the mechanism for generating such occurrence might differ from the one that generated other occurrences. Thus, it is important to examine why the anomalies occur in a dataset. Outliers could emerge due to a variety of factors. Thus, Anscombe (1960) categorized outliers into two primary types: those that arise from data errors and those that result from the natural variability inherent in the dataset. Similarly, Osborne and Amy (2004) identified several potential sources of outliers, including measurement variability, errors in measurement, mistakes in data collection and entry, flawed theoretical assumptions, deliberate misreporting, issues with standardization, and incorrect distributional assumptions. They emphasized that outliers can adversely affect statistical inference by reducing power, accuracy, and increasing error rates. Consequently, strategies such as data transformation, re-coding, or trimming can help mitigate these negative effects.

Sunitha *et al.*, (2013) further explored the causes of outliers, highlighting additional factors such as mechanical malfunctions, shifts in system behaviour, fraudulent activities, human errors, instrument failures, and natural deviations within populations. Detecting outliers is essential to prevent their escalation and mitigate their potential impact on statistical analyses. Furthermore, identifying outliers aids in recognizing errors and minimizing their

contaminating effects, thereby improving data integrity and ensuring more reliable analytical outcomes.

Some outliers arise due to rare or extreme real-world events. Natural disasters, economic crashes, or medical anomalies can create outliers in datasets. For instance, stock market data during financial crises exhibit extreme fluctuations that deviate significantly from normal trends (Taleb, 2007). These types of outliers may carry important information and should be analyzed in context rather than dismissed as errors. So understanding the characteristics of data distributions is crucial when determining if an observation is truly an outlier or part of a long-tailed distribution.

Several methods have been developed for detecting outliers in datasets. One widely used approach is the graphical method introduced by Box and Cox (Box and Cox, 1964) which includes tools such as histograms, run sequence plots, and box plots. These visual techniques help identify potential outliers, examine the normality of data distribution, and determine whether a dataset contains single or multiple outliers. Grubbs' test identifies one outlier in approximately normally distributed univariate data (Grubbs, 1969). When multiple outliers are suspected, the Tietjen-Moore test (Tietjen & Moore, 1972) can be used, though it requires having a prior knowledge of the number of extreme values available in a given dataset. The Generalized Extreme Studentized Deviate (ESD) test and Studentized residuals, proposed by Rosner (1983), is used for the detection of one or more outliers in univariate datasets and assist in evaluating assumptions in multiple regression models.

Stevens (2006) introduced Studentized deleted residuals to assess the significance of individual observations. The study also noted that the Extreme Studentized Deviate test is particularly effective for datasets with more than 10 observations from a normal

distribution, whereas the Dixon test is better suited for smaller samples (fewer than 10 observations) and does not require strong distributional assumptions.

Additionally, Yinaze *et al.*, (2011) proposed a modified box plot procedure for detecting outliers in location-scale data. They compared this method with the traditional box plot, which visualizes characteristics such as location, spread, and skewness, and found that the modified approach is more effective for datasets with symmetric distributions.

Recent research has increasingly focused on enhancing outlier detection by integrating machine learning with robust statistical techniques. Advances in deep learning have led to the development of autoencoder-based anomaly detection models, which effectively identify complex outliers in large datasets (Zhao *et al.*, 2020). Additionally, hybrid approaches that combine statistical and computational methods, such as isolation forests and robust clustering algorithms, have significantly improved the precision of outlier detection (Liu *et al.*, 2022). Another notable trend is the application of explainable AI (XAI) techniques to interpret outliers in critical fields such as healthcare and finance (Molnar, 2023). Gaining insights into the underlying causes of detected anomalies enhances decision-making and minimizes the risk of false positives in anomaly detection systems. Outliers can arise from multiple sources, including measurement and data entry errors, sampling biases, natural variability, data processing mistakes, fraudulent activities, violations of statistical assumptions, extreme events, and intrinsic characteristics of data distributions. Effective detection and management of outliers ditto influential points contribute to maintaining data integrity and improving the reliability of statistical analyses. It also enhance more accurate and robust conclusions in research and applied analytics.

Consequently, many researchers have integrated some estimators to find influential locations in place of the aforementioned statistics. As a result, several versions of Cook's Distance, DFFITS and Pena's statistic for One and Two-parameter have been proposed for

detecting influential points in classical models. It is pertinent to state that some of these One and Two-Parameter estimators were not used with certain robust regression techniques to detect influential points in dataset. Therefore, the focus of this research explored developing influential diagnostic measures within this framework, which are critical for assessing the performance and reliability of the estimator.

1.2 Statement of the Problem

The reliability of classical regression rely not only on appropriateness of the estimation technique but also on its ability to identify observations that can exert disproportionate effects on estimated regression coefficients, fitted values and statistical inferences. Two major challenges that undermine the reliability of regression estimates are multicollinearity and outliers. Multicollinearity increases estimator variance, destabilizes regression coefficients, and reduces the reliability of statistical inferences and predictions. Similarly, outliers distort parameter estimates and compromise predictive performance of regression models. The simultaneous occurrence of these two problems further complicates regression analysis, as a value that is unusual in the regressor space or response variable may exert substantial influence on the estimated regression coefficients.

Several estimation techniques have been developed to address these challenges independently. Biased estimators, particularly ridge-type estimators, have been widely deployed to mitigate the adverse effects of high correlation, by including a biasing parameter to downsize the variance of parameter estimates. However, conventional ridge-type estimators remain sensitive to outliers, which can substantially distort their estimates. Conversely, robust regression techniques, such as ‘M’, ‘MM’, ‘Least Absolute Deviation’ (LAD), ‘Least Median of Squares’ (LMS), ‘Least Trimmed Squares’ (LTS), ‘Least Quartile of Squares’ (LQS) and ‘S’ estimators are designed to downsize the effects of outliers but may not adequately address the instability associated with severe

multicollinearity. Consequently, the application of either conventional biased estimators or robust regression techniques in isolation may be inadequate when multicollinearity and outliers occur simultaneously. Although some researchers have explored hybrid estimation approaches that combine the advantages of biased and robust estimators, limitation exist in the development of appropriate diagnostic measures within these combined estimation frameworks, hence it remained an important methodological consideration.

Furthermore, conventional influential diagnostic measures, such as ‘Cook’s Distance’, ‘Difference in fits’ (DFFITS), ‘Difference in betas’ (DFBETAS), ‘Covariance ratio’ (COVRATIO), Hadi’s measure, Pena’s statistic and Atkinson’s statistic, were primarily developed within ordinary least squares framework. Their effectiveness may be compromised when the underlying regression model is affected by multicollinearity and outliers simultaneously. Although several researchers like Adejumo et. at (2024) and Adejumo et. at (2025) have extended influential diagnostic measures to One-parameter and Two-parameter ridge-type estimators, including Kibria–Lukman estimator, much of the existing work has focused on addressing multicollinearity, with comparatively limited attention given to the development and evaluation of influential diagnostic measures that integrate robust regression techniques with one-parameter new ridge-type estimators. This, creates a methodological gap in identification of influential observations when the combined effects of multicollinearity and outliers are present. Additionally, non availability of suitable diagnostic measures within such hybrid estimation frameworks may limit the ability of researchers to distinguish genuinely influential observations from those whose apparent influence is attributable to the instability of parameter estimates or the presence of extreme values.

The study is motivated by the need to improve influential-point detection when multicollinearity and outliers coexist. The limitations of existing diagnostic measures under these combined conditions, together with the relatively limited development of influential diagnostics for robust one-parameter new ridge-type estimators, provide a compelling basis for further methodological investigation. Consequently, this study provides a more reliable diagnostic framework for identifying influential cases, thereby improving the stability of parameter estimates, the validity of statistical inference and the overall reliability of regression models in the presence of multicollinearity and outliers simultaneously.

1.3 Aim and Objectives of the study

The aim of this study is to propose influential diagnostic measures for a One-parameter new ridge-type estimators in linear regression models that effectively detect influential points in the presence of both multicollinearity and outlier. Specific objectives of the study are as follows to:

- i. develop robust versions of One-parameter new ridge-type estimators
- ii. estimate some diagnostic measures based on the newly developed one-parameter new Ridge-type estimators.
- iii. Obtain Approximate case deletion formulas for Cook's Distance and DFFITS in robust One-parameter new ridge-type estimator.
- iv. assess performance of these measures through simulation studies and real-life datasets.
- v. compare the results with existing methods and make necessary recommendation

1.4 Significance of the study

The importance of detecting influential points when developing a linear regression model cannot be overemphasised because failure to do so can invalidate the conclusions reached. When extreme values are excluded from a regression analysis, the issue of key points emerge. While, multicollinearity is yet another challenge in regression analysis, and there is little doubt that the twin problems may coexist. Notably, several researchers had proposed, estimators that can mitigate problems of outliers and multicollinearity, concurrently. However, this study is significant for several reasons; it developed influential diagnostic measures in one-parameter Robust Regression Kibria-Lukman estimators and ensured that the one-parameter new ridge-type estimator can be reliably used in situations where both multicollinearities, as well as when outliers are present. Also, since accurate diagnostics allow researchers and practitioners to identify and mitigate the impact of influential data points can lead to more robust and trustworthy models, the study will advance the field of robust regression by integrating influential diagnostics into the new ridge-type estimation framework, thereby, providing a novel approach to addressing common regression challenges.

1.5 Scope of the study

The scope of this study includes theoretical development of influential diagnostic measures for Robust One-Parameter New Ridge-type estimator and applying these measures in various scenarios involving multicollinearity and outliers. A comparative analysis of the new influential diagnostics were done with existing diagnostic measures in bias regression models using simulated and real-life datasets, to validate effectiveness of the proposed influential measures.

CHAPTER TWO

LITERATURE REVIEW

2.1 Concept of Outliers and its Detection

Several diagnostic techniques had been developed to detect observations that have influence on regression models. Cook's Distance (Cook, 1977), DFFITs (Welsch and Kuh, 1977), Welsch's Distance (Welsch, 1982), and DFBETAS are examples of diagnostic tools, which assess the influence of individual observation on fitted values and regression coefficients. Meloun and Militky (2001) examined influential-point diagnostics based on the OLS estimator, while Leiva et al. (2006) developed influence measures based on Log-Birnbaum-Saunders regression with censored data. Jahufer and Jianbao (2008) proposed a deletion-based modified ridge regression technique for detecting influential observations. Recognizing that multiple leverage observations can contribute to multicollinearity, Bagheri and Midi (2009) proposed robust estimation techniques and reported DRGP-MM estimator as the most efficient among estimators considered.

Bagheri *et al.* (2010) used Monte Carlo experiment to evaluate influence of multicollinearity - important points on dataset with multicollinearity. Jahufer (2013) suggested an acceptable approximate case deletion formula for detecting location of influential observations using Liu estimator. In order to find influential points and robust diagnostic tools in regression analysis, Ayinde et al. (2015) compared several robust estimators and influential-point diagnostics such as Cook's Distance, Welsch-Kuh's Distance, DFFITS, DFBETAS, OLS, MM, LTS, and S estimators, using real-life datasets and RMSE as the performance criterion. Asar and Erisoglu (2016) proposed a diagnostic tool based on Two-parameter ridge regression, developing generalized versions of DFFITS and Cook's Distance along with approximate case-deletion formulas. Emami and Emami (2016) introduced a diagnostic measure for assessing influence of individual observations

on fitted ridge regression models. Despite the fact that several diagnostic measures are available to discover influential points in regression analysis, particularly when contamination is caused either by multicollinearity or outliers, the problems of extreme observation and multicollinearity in regression models may be unavoidable. Lukman and Ayinde (2016) proposed an approximate case-deletion formula based on Two-parameter Liu-Ridge estimator, alongside conventional Cook's Distance and DFFITS. Penã et al. (2016) developed generalized standardized corrected pearson residuals (GSCPR) for detecting multiple leverage points in generalized linear models, while Lukman et al. (2016) applied modified Peña's statistics to biased estimators. Kashif et al. (2018) developed a Peña-based influential measure for One-parameter Liu estimator, and Alkasadi et al. (2019) proposed a DFFIT-based approach for detecting outliers in multiple circular regression models. Hussein (2021) developed Cook's Distance and DFFIT measures, including approximate case-deletion formulas, for Liu-type estimator under extreme multicollinearity.

Adejumo et al. (2024) developed Cook's Distance, DFFITS, Peña's, Atkinson's, Hadi's, and COVRATIO diagnostics for new Two-parameter (NTP) estimator. Their simulation results indicated that CVRNTP outperform other influential measures when multicollinearity and outliers exist in a dataset. Similarly, Adejumo et al. (2025) developed corresponding influential diagnostic measures for Kibria–Lukman estimator, with CVR-KL demonstrating the strongest performance among measures considered when multicollinearity and outliers are present in a dataset.

Barnett and Lewis (1994) define extreme points as values that differ from other data in a data collection. While Leverage points are explanatory variables that appear to be outliers with little or no impact on the estimate of the coefficient of a regression model. Any outlier with significant impact on fitted regression line is said to be influential. Every Influential

point is an Outlier but not all Outliers are Influential points. Outliers can be classified into three types: Influential points, high-leverage points or combinations of the two oddities. Outliers in data can be identified through a variety of techniques that includes, graphical method (Box and Cox, 1964), Dixon test (1960) and Grubbs' test (1969). These methods are limited to detection of outlier in univariate dataset. Tietjen-Moore test (1972), Studentised residual (Velleman and Welsch, 1981), generalized extreme studentized deviate test (Rosner, 1983), and deleted studentized residual (Steven, 2006) are used to detect multiple outliers. They are all assumed to follow approximately normal distribution.

Some graphical tools useful in checking the normality assumption and identifying potential outliers include Histogram, run sequence plot, and box plot. In addition to showing potential outliers, several of these methods help to determine if there is need to check for single outlier or multiple outliers in a dataset.

In regression analysis, the generalized ESD test and Studentized residuals Rosner (1983) were proposed to ascertain the extreme values in a univariate dataset which follows approximately Normal distribution and to examine various assumptions of the multiple regression model. Also, Studentized deleted residuals by Steven (2006) is meant to test the significance of an observation in the dataset. In addition, he claimed Extreme Studentized Deviate test is an excellent test for datasets that have more than 10 observations from a dataset that is normally distributed. He further affirmed that, for a dataset of less than 10 observations, the Dixon test should be used and as such, it does not require distribution assumption.

Yinaze, *et al.* (2011) discovered a new box plot procedure for outlier detection in location-scale data. They compared the traditional box plot. Rahman *et al.* (2012) developed a new

approach for finding extreme values in multiple regression models and to verify whether outlying cases come from the same or different population via mathematical calculation.

David (2012) claimed that if outliers can radically alter the outcome of analysis and can also violate OLS assumption, his research suggest dropping such outlier may be appropriate. However the study stressed that dropping outliers can introduce bias into the researcher's results. Hence, it concluded that dropping outliers is justified only if the dataset is bad for some reasons

2.2 Robust Regression Techniques to Handle Outlying Cases

Robust regression methods were developed to curtail influence of outliers on regression coefficients. Robust regression methods - 'M', 'S' and 'MM' estimators minimize a modified loss function, which down weights the influence of extreme values.

OLS estimators tend to misbehave, especially when an iota of odd observation appears in a set of data, causing parameters predicted to grow or drop. Robust regression estimators have been shown to reduce these anomalies; hence, to analyze, compare, and contrast performance of several robust regression methods, Fitrianto and Xin (2022) conducted research in this area. The robust regression methods considered were 'M' 'MM', 'LTS', and 'S' estimators. The research, found M robust regression to have outperformed other methods, when a dataset is contaminated in the y-direction using the Root Means Squared Error (RMSE) criterion. Their research also focused on the detection of leverage and influential points, using deleted studentized residual and Cook's Distance diagnostic measures. They eventually concluded that LTS estimator introduced and S estimator outperformed M, MM, and OLS estimators when contamination occurs in the x-direction. Others estimators considered were LMS estimator and Least Winsorized Squares (LWS) estimator, to minimize effects of outliers on regression models. Similarly, Alguraibawi *et*

al. (2015) conducted a robust diagnostic analysis to assess impact of leverage points on parameter estimates in regression analysis.

❖ **M-Estimator**

Huber (1964) proposed the M estimator as a robust approach for handling outliers. It replaced OLS squared-error loss+ with a robust objective function applied to standardized residuals. The method uses all observations without requiring a specific data distribution.

❖ **MM-Estimator**

Yohai (1987) introduced MM-estimator. It uses regression approaches to reduce the effect of extreme observation in dataset. To obtain the final estimates, this technique employed many M-estimation techniques. The sum of an objective function is applied to a scale of the residual obtained from the residual of *S*-estimator (Susanti *et al.*, 2014). The primary goal of the MM estimator is to generate robust parameter estimates that are highly efficient.

❖ **Least Median of Squares Estimator (LMS)**

Rousseeuw (1984) introduced LMS estimator as an alternative to OLS, minimizing the median (50th percentile) of squared residuals rather than their mean.

❖ **Least Trimmed Squares (LTS) Estimator**

Rousseeuw (1984) proposed LTS estimator as a more efficient robust alternative to LMS. It minimizes sum of smallest ordered squared residuals over trimmed subset of observations (Rousseeuw and Leroy, 1987).

❖ **S-Estimator**

Rousseeuw and Yohai (1984) developed the S-estimator as a high-breakdown robust estimator that minimizes the scale of regression residuals. It can also provide initial estimates for M-estimators, improving their efficiency.

❖ **Least Absolute Deviation (LAD)**

Boscovich first introduced LAD in 1757 (Birkes and Dodge, 1993). It is also known as Least Absolute Error (LAE) estimator. In LAD, coefficient of parameters are estimated such that total absolute deviation of error terms are minimized. It is like the Ordinary Least Square estimator except that it considers only the absolute values of the residuals. It aims at finding a function that approximate a dataset by minimizing sum of absolute deviations between the observed and fitted values. This approach is highly useful when the residual distribution is not normal. Furthermore, LAD is robust when anomalies occur in the y-direction (Ahmed and Maha, 2016).

❖ **Least Quantile of Squares (LQS) Estimator**

Rousseeuw (1984) developed LQS estimator. It is an extension of Least Median of Square. It is based on selecting smallest quantile and the selected quantile's sum of square of error are minimized. Furthermore, LQS was designed such that a proportion of observations is q , while the remaining $1-q$ observations are considered extreme.

2.3 **Review of Some One-Parameter Estimators**

To address high correlation among regressors, Hoerl and Kennard (1970) proposed ridge regression estimator as a one-parameter alternative to OLS. Ridge parameter was introduced into $(X^T X)$ matrix to lessen the impact of multicollinearity of which OLS estimator has a high variance α . Valley estimator was proposed by Leamer, (1977) for problems in which predictor variables are almost orthogonal- the predictor variables are not too correlated. Valley estimator introduce a biased parameter to reduce the variance of

parameter estimates ,so as to produce a more stable and reliable parameter coefficients. In other words, the valley estimator implore the benefit of trading bias for reduced variance

To address challenges of multicollinearity, Liu (1993) proposed Liu estimator by combining the properties of the Stein and ridge estimators, using a biased parameter to shrink coefficients and reduce variance. Subsequently, Kibria and Lukman (2020) developed another biased estimator to further mitigate effects of multicollinearity on regression models. In addition, Abdul-Wahab *et al.* (2005) used Principal Component Analysis (PCA) technique in multiple regression analysis to model troposphere ozone and to identify major factors impacting ozone levels. Similarly, Çamdevyren *et al.* (2005) analysed water quality using PCA in multiple linear regression. Sousa *et al.* (2007) proposed a PCR model for predicting Ozone concentrations that used PCs as input. They compared the model to two additional models: a feed forward artificial neural network and multiple linear regression.

Alabi *et al.* (2008) explored the impact of strong correlation on Ordinary Least Squares estimator by considering a regression model with two regressors that reflect multicollinearity. They discovered that the power rates of the constant value are the same as those of the postulated values, and the error rates fall asymptotically at all levels of multicollinearity. Rajab *et al.* (2013) improved the prediction of ozone levels by combining a multiple regression model with PCA approach. All of these studies showed that incorporating principal components (PCs) as independent variables in regression model improved prediction using the model and it reduced the model complexity by mitigating multicollinearity.

When there is multicollinearity among the regressors, Alabi *et al.*, (2014) investigated what sample size would be required if a traditional estimator like OLS is to be utilized. A regression constant and two exogenous variables with two regression coefficients were

used in the investigation under various multicollinearity levels (0, 0.25, 0.5, 0.7, 0.75, 0.8, 0.9, and 0.99). They replicated the Monte Carlo experiments 1000 times using various sample sizes, they came to the conclusion that the sample size required to have a reliable estimates depend largely on the degree to which hypothesized values deviated from the observed values, regardless of the severity of correlation among regressors.

Adeboye *et al.*, (2014) examined several tests for detecting multicollinearity and fitted an OLS model to Nigeria's GDP dataset so as to determine the effect of multicollinearity on variance of fitted regression model. Ultimately, they concluded that multicollinearity had a significant effect on the model by increasing variance of parameter coefficients thereby causing the estimate of the coefficient of parameters to become insignificant.

Donugade (2015) proposed an Adjusted Ridge Estimator (AR) for Linear Regression models and compared it with some existing Ridge and Kibria's Estimators with the view to deal with severe multicollinearity and as such improve estimation accuracy. The new estimator (AR) did not require the estimation of the optimal biased parameter (k) as it is the case in the Ridge and Kibria's estimators. The study then compared performance of the new estimator (AR) with some already existing estimators using stimulated and real-life dataset. The evaluation was done with four (4) different magnitude of multicollinearity (0.6, 0.99, 0.999, 0.9999) and the result showed that AR had a smaller MSE. Consequently, the study concluded that the new estimator (AR) was a more robust and efficient method for estimating parameter coefficients of regression model as it outperformed some already existing Ridge and Kibria estimators.

Dawoud and Kibria (2020) developed a biased estimator for Gaussian regression models by introducing a biasing parameter, reporting improved MSE and coefficient stability compared with OLS and some existing estimators. Owolabi *et al.* (2022) proposed a novel Two-Parameter estimator, while Jegede *et al.* (2022) developed a robust Jackknife Kibria–

Lukman M estimator to address the effects of multicollinearity and outliers. Arumairajan and Kayathiri (2002) proposed a stochastic restricted Two-parameter estimator, whereas **Abidoeye *et al.*** (2022) developed an unbiased Modified Two-Parameter estimator. **Majeed *et al.*** (2022) introduced a robust Kibria–Lukman estimator for y -direction outliers, and **Idowu *et al.*** (2023) proposed a Two-Parameter Kibria–Lukman estimator for multicollinearity.

Oyeleke *et al.* (2024) compared combinations of One and Two-parameter ridge-type and robust regression estimators for simultaneously addressing multicollinearity and outliers, with MRT–MM, MRT–M, and MRT–LTS showing competitive performance based on MSE. Extending the research of **Majeed *et al.*** (2022), **Adejumo *et al.*** (2024) developed a robust KL-M estimator for handling effects of outliers in X -direction.

2.4 Review of some Two-parameter Estimators

Several researchers have proposed Two-Parameter estimators to address multicollinearity and improve the performance of regression models. **Idowu *et al.*** (2023) proposed a Two-Parameter estimator for Gamma regression models to mitigate the effects of correlated regressors and reported superior performance compared with MLE, GRE, GLE, and GLTE based on MSE. Similarly, **Oladapo *et al.*** (2024) developed new Biased Two-parameter (NBTP) estimator for Logistic regression models under extreme multicollinearity. The proposed estimator incorporated an additional biasing parameter and demonstrated improved performance and greater parameter stability compared with OLS, LLE, LRE, LKLE, and LMRTE.

Since multicollinearity and outliers frequently occur simultaneously in real-life datasets, hybrid robust ridge-type estimators have also been developed to address their combined effects. Midi and Zahari (2007) combined MM estimator with ridge regression to develop robust-MM (R_MM) estimator, while Lukman *et al.* (2020) proposed a Two-Parameter

ridge-type modified M-estimator to mitigate effects of multicollinearity and outliers simultaneously. These developments highlight importance of integrating robust estimation with shrinkage techniques for improving regression estimation in the presence of multiple data anomalies.

2.5 Review of Some Existing Influential Diagnostic Measures

Influential diagnostic measures are tools used to identify data points that have a disproportionate effect on regression models. Many diagnostic tools have been developed to identify and assess the influence of individual extreme data value using classical linear model. Some of the commonly used diagnostic measures are Cook's Distance, DFFITS and DFBETAS. These diagnostics help in evaluating influence of individual observation on parameter estimates and overall goodness of fit of a regression model. The challenge in hybrid models is to adapt these diagnostics to address the effects of multicollinearity and outliers in regression analysis. Another diagnostic measure is leverage, which identifies data points with extreme predictor values relative to other observations (Chatterjee and Hadi, 1988). High leverage points have the potential to distort the regression line, thus affecting the estimated coefficients. Leverage is often used alongside Cook's Distance to identify influential observations.

Studentized residuals also serve as an important diagnostic tool for detecting outliers. These residuals are adjusted for leverage, offering a more accurate measure of each data point's influence on the regression model (Barnett & Lewis, 1994). The Generalized Extreme Studentized Deviate (ESD) test (Rosner, 1983), extended the concept of studentized residuals, making it particularly useful for identifying multiple outliers in a univariate data.

The presence of outliers particularly, influential observations can significantly distort the inference of hypothesis testing. Osborne and Overbay (2004) highlighted that summary statistics such as mean, standard deviation and variance are highly sensitive to extreme values. The value of mean, in particular, shifts toward outliers, which can lead to a misleading representation of the data's central tendency. Similarly, variance and standard deviation can be inflated by outliers, thus, making the data appear more dispersed than it truly is (Barnett & Lewis, 1994). Rousseeuw and Leroy (2003) emphasized that robust regression estimators like M-estimators and LAD (least absolute deviation) regression, offer greater resistance to the influence of outliers. These estimators reduce the effect of extreme values on the parameter estimates, with M-estimators using a weighting function to diminish the influence of outliers fitted regression model.

In relation to multicollinearity, influential observations can exacerbate the effects of highly correlated predictors, complicating the acquisition of reliable coefficient estimates (Belsley, Kuh, & Welsch, 1980). In such cases, techniques like ridge regression or other forms of regularization are employed to mitigate multicollinearity and also reduce the influence of outliers. Recent research have focused on enhancing the detection and management of influential observations. One notable development is the introduction of machine learning techniques for identifying outliers and influential points in complex datasets. Zhao *et al.*, (2020) proposed auto-encoder-based anomaly detection models, which are effective for identifying outliers in large, high-dimensional datasets. These models are particularly useful in big data contexts, where traditional diagnostic measures may be computationally prohibitive.

Furthermore, hybrid methods combining statistical techniques with machine learning have shown promise in improving the accuracy of outlier detection. Liu *et al.*, (2022)

demonstrated that combining isolation forests with robust clustering algorithms can enhance the robustness of regression models by more accurately identifying and managing influential points. The presence of influential points often necessitates appropriate measures, which could involve removing, transforming, or down-weighting these observations. Cook (1977) suggested deleting influential points to assess the robustness of regression models. However, this approach may not always be feasible, particularly in small datasets where every observation holds significant value. In such cases, transformations, such as log transformations, or the use of robust regression techniques are more appropriate (Huber, 1981).

In practice, the way influential observations are handled should consider the context and domain of the analysis. For instance, in financial modeling, outliers may represent rare but impactful events, such as stock market crashes or financial crisis (Cont, 2001). Removing these outliers could distort the risk assessment. Similarly, in medical research, outliers may represent rare but important conditions, such as unique disease presentations, which are crucial for understanding disease progression (Zimek *et al.*, 2012). Therefore, it is vital to adopt a context-sensitive approach when dealing with influential data points to preserve the integrity of the analysis. ‘Cook’s Distance’ (Cook 1977), is an influential-point diagnostic that measures effect of individual observation on fitted values and estimated regression coefficients. A large D_i indicates that deleting the observation substantially changes the fitted model, suggesting that the observation is highly influential.

Welsch and Kuh (1977) proposed Difference in fits (DFFITs) as a diagnostic measure that evaluates influence of individual observation on its own fitted value. In other words, it measures the difference between the fitted value for the i th observation of a full dataset and its fitted value when i th observation is deleted from the dataset, divided by standard deviation of fitted model at that point. Specifically, it measures influence observation i has

on its fitted value $\hat{Y}$. Thus, DFFITS is the standardized change of fitted value when the i th observation is deleted. Jahufer (2013) suggest that observations for which $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ for large datasets and those for which absolute value of $DFFITS_i$ is greater than one, for small or medium size datasets, warrants attention.

As an alternative influential measure, Atkinson (1985) modified Cook's Distance by replacing the variance (s^2) of full data with the variance ($s_{(-i)}^2$) after the i th observation is deleted. Any observation that its estimate of *Atkinson statistics* > 1 is said to be influential. Belsley *et al.*, (1980) proposed another influential measure as an alternative to Cook's D and DFFITS for detecting influential points on regression model (LRM) based on determinant ratio. They defined COVRATIO statistic as ratio of determinant of covariance matrix for the estimated parameter coefficients excluding subset D_m to the determinant of covariance matrix for estimated parameter coefficients for full dataset. An observation is considered influential, if $|COVRATIO_i - 1| \geq \frac{3p}{n}$

Another version of COVRATIO statistic was proposed by Abuzaid *et al.* (2011) as an extension to COVRATIO statistic proposed (Belsley *et al.*, 1980). The study explored the approach of detecting influential point in circular regression model. Through simulation study, they derived a cut-off point for the statistic and found out that its cut-off points are dependent on sample size. They affirmed that the proposed statistic is capable of detecting high influential point when sample size is large and parameters are concentrated. They noted that the higher the concentrated parameter, the higher the influential detection by the statistic. The cut-off point for COVRATIO statistic is: $\left| COVRATIO_{(-i)} - 1 \right| \geq \frac{16}{n}$

In order to detect influential points in regression model using data obtained from a survey, Li and Valliant (2011) came up with a new method called forward search method. The method was applied to some types of complex survey data using diagnostics based on linear regression. In their study they incorporated some sample weights, including the effect of stratification. They defined forward search method as method which involves tracking number of different statistics whose values can be affected by influential observations. They claimed that any little change in ED_i indicates that an influential point has penetrated which can be as a result of observation(s) entered has a heavy weight, residual or leverage.

Hadi (1992) proposed a new influential diagnostic measure based on the fact that outliers can be outrageous in the dependent variable or explanatory variables. The cut-off point for Hadi's statistic is $H_i^2 = \text{mean}(H_i^2) + c\sqrt{\text{var}(H_i^2)}$ where c is a constant chosen appropriately i.e 2 or 3 as suggested by El-Fallah and El-Salam (2013). Hence, any observation whose estimated influential measure statistics is greater than the cut off point, is considered to be an influential point.

To further detect influential points in dataset, Pena (2005) proposed a new influential diagnostic measure different from conventional Cook's Distance and DFFITS. His study showed how each data-point affected other set of observations. Also, He was able to show that the new diagnostic measure was asymptotically normal distribution and it is capable of detecting groups of outliers that could not be detected by Cook's Distance diagnostic measure. He further revealed that the statistic tends to be normally distributed if samples and regressors are numerous. He affirmed that the cut-off point of the diagnostic tools is considered influential, if the value of S_i is greater than $(S_i - E(S_i))/SD(S_i)$ because both

mean and standard deviation of S_i are usually affected by influential point. He then claimed that a point is influential if $|S_i - \text{median}(S_i)| \geq 4.5 \text{MAD}(S_i)$.

Since, implication of influential points on the regression analysis is always challenging, therefore their location and accessibility of their impact on the regression line is necessary most especially in the presence of multicollinearity. In this regard, Emami and Emami (2016) developed a new influential diagnostic tool for one parameter estimator that can mitigate collinearity problem. Their study demonstrated that employing ridge regression to address multicollinearity can substantially alter the influence of certain observations. This finding was supported by simulation studies and an application to a real-life dataset.

Detection of influential points become more critical especially when regressors are highly correlated. Therefore, this made Kashif *et al.*, (2019) to propose Pena's statistic version for detecting influential point in presence of multicollinearity, by considering modified ridge regression estimator (MMRE). They were able to show that when robust estimator that can mitigate problem of multicollinearity in regression analysis, the influence of some observations could be changed significantly. With aid of simulation study, empirical results were obtained and were compared with results obtained with real life data.

An approximate deletion formula was proposed by Emami and Emami (2017) in order to detect influential observations in modified Liu-type estimator (MLE). In their study, influential diagnostic measure in MLE for Cook's D and DFFITS versions were defined.

CHAPTER THREE

MATERIALS AND METHODS

This chapter discussed materials used in this research and the methodology of influential measures that were developed. The simulation procedures used to evaluate performance of the developed diagnostic measures were also discussed extensively.

3.1 Materials

In this study, One-parameter new ridge-type estimator was used (Kibria and Lukman, 2020). It is an alternative estimator to already existing estimators that can mitigate problems arising from multicollinearity in a dataset. Its mathematical definition is as follows:

$$\hat{\beta}_{KL} = (X^T X + kI_p)^{-1} (X^T X - kI_p) \hat{\beta}_{OLS} \quad (3.0)$$

where $k = \frac{\hat{\sigma}^2}{2\hat{\alpha}_i^2 + (\hat{\sigma}^2 / \lambda_i)}$, $\hat{\beta}_{OLS}$ is the OLS estimator and I_n is an n identity matrix such

that σ^2 is MSE and $\hat{\alpha}_i$ is an unbiased estimator of $\hat{\beta}_{OLS}$ and λ_i is ith Eigen value $X^T X$.

Study has shown that the presence of extreme values in real-life datasets may be inevitable and it has been noted to affect KL-estimator (Majeed *et al.*, 2022). Also, in many real-life data they occur simultaneously along with multicollinearity. Hence, several influential measures developed in some one or two-parameter ordinary KL- estimators may not be too efficient in identifying influential points, since the estimator does not handle outliers. It is to this end, that some Robust Regression Estimators that can take care of outlier were used so that the twin problem can be mitigated. The Robust Regression Estimators used are given as follows:

I. M Estimator

$$\hat{\beta}_M = \min \sum_{i=1}^n \rho \left(\frac{y_i - x_{ij} \hat{\beta}}{s} \right). \quad (3.1)$$

where ρ is a Tukey's bi-weight function, y_i is dependent variable, x_{ij} is independent variable,

$\hat{\beta}_j$ is OLS estimator and s is a scalar expressed as:

$$s = \frac{\text{median}|e_i - \text{median}(e_i)|}{0.6745}.$$

II. MM Estimator

$$\hat{\beta}_{MM} = \sum \rho \left(\frac{y_i - \sum_{j=0}^k x_{ij} \hat{\beta}_j}{s_{MM}} \right) x_{ij}. \quad (3.2)$$

where s_{MM} is a standard deviation obtained from the residual of S -estimator and ρ is a Tukey's bi-weight function, y_i is dependent variable, x_{ij} is independent variable and $\hat{\beta}_j$ is OLS estimator.

III. LTS Estimator

$$\hat{\beta}_{LTS} = \min \left(\sum_{i=1}^h (e_i^2)_{i=n} \right). \quad (3.3)$$

where $(e^2)_{1:n} \leq (e^2)_{2:n} \leq \dots \leq (e^2)_{n:n}$ are ordered squared errors arrange from least to the highest number and $h = n(1 - j) + 1$ (Rousseeuw and Leory, 1987) such that j is the trimmed percentage.

IV. LMS Estimator

$$\hat{\beta}_{LMS} = \min_{\beta} [\text{med}(e_i^2)] \quad (3.4)$$

where, e_i is the residual from the OLS.

V. S Estimator

$$\hat{\beta}_S = \min \hat{\sigma}_S (e_1, \dots, e_n). \quad (3.5)$$

where, $\hat{\sigma}_S = \sqrt{\frac{1}{nk} \sum_{i=1}^n w_i e_i^2}$ where $K = 0.199$ and initial estimate

$$\hat{\sigma}_S = \frac{\text{median}|e_i - \text{median}(e_i)|}{0.6745}$$

VI. LAD Estimator

$$\hat{\beta}_{LAD} = \min \sum_{i=1}^n |y_i - x_i \hat{\beta}| \quad (3.6)$$

VII. LQS Estimator

$$\hat{\beta}_{LQS} \in \arg \min_{\beta} |r_{(q)}| \quad (3.7)$$

where $r_{(q)}$ is residual corresponding to qth ordered absolute residual: $|r_{(1)}| \leq |r_{(2)}| \leq \dots \leq |r_{(n)}|$.

3.2 The Development of Influential Measures in Robust Regression Kibria-Lukman (RKL) Estimator.

This subsection discusses the methodology of the influential measures that were developed based on Robust Regression Kibria-Lukman Estimator (RKL) in achievement of objectives one and two of this study. The method employed in this research was that Kibria-Lukman estimator was combined with each of MM, LTS, LMS, S, LAD and LQS estimator and each respective hat matrices of fitted values were used to replace OLS hat matrix fitted values of the following conventional diagnostic tools: Cook's Distance (CKD), DFFITS (DFT), COVRATIO (CVR), Hadi's measure (HAD), Pena's statistic (PEN) and Atkinson measure (ATK). Majeed *et al.* (2022) and Adejumo *et al.* (2024) expressed M version of Kibria-Lukman estimator as:

$$\hat{\beta}_M^{KL} = (X^T X + k I_p)^{-1} (X^T X - k I_p) \hat{\beta}_M. \quad (3.8)$$

Where $X^T X$ is the information matrix, k is biasing parameter for KL – estimator and $\hat{\beta}_M$ is parameter estimator from M estimator as defined in equation (3.1) and p is number of independent variable.

From which other robust versions of KL-estimator were developed in this study under some conditions by integrating the aforementioned robust regression estimators (MM, LTS, LMS, LAD, LQS and S) expressed as equations (3.1) to (3.7) into KL- estimator defined in equation (3.0). The following equations were obtained:

$$\hat{\beta}_{MM}^{KL} = (X^T X + k_{MM} I_p)^{-1} (X^T X - k_{MM} I_p) \hat{\beta}_{MM}. \quad (3.8a)$$

$$\hat{\beta}_{LTS}^{KL} = (X^T X + k_{LTS} I_p)^{-1} (X^T X - k_{LTS} I_p) \hat{\beta}_{LTS}. \quad (3.8b)$$

$$\hat{\beta}_{LMS}^{KL} = (X^T X + k_{LMS} I_p)^{-1} (X^T X - k_{LMS} I_p) \hat{\beta}_{LMS}. \quad (3.8c)$$

$$\hat{\beta}_{LAD}^{KL} = (X^T X + k_{LAD} I_p)^{-1} (X^T X - k_{LAD} I_p) \hat{\beta}_{LAD}. \quad (3.8d)$$

$$\hat{\beta}_{LQS}^{KL} = (X^T X + k_{LQS} I_p)^{-1} (X^T X - k_{LQS} I_p) \hat{\beta}_{LQS}. \quad (3.8e)$$

$$\hat{\beta}_S^{KL} = (X^T X + k_S I_p)^{-1} (X^T X - k_S I_p) \hat{\beta}_S. \quad (3.8f)$$

Where $X^T X$ is the design matrix, k_{MM} , k_{LTS} , k_{LMS} , k_{LAD} , k_{LQS} and k_S are the biasing parameters for each of robust version of KL estimator stated in equations (3.8a) – (3.8f). Likewise.

$\hat{\beta}_{MM}$, $\hat{\beta}_{LTS}$, $\hat{\beta}_{LMS}$, $\hat{\beta}_{LAD}$, $\hat{\beta}_{LQS}$ and $\hat{\beta}_S$ are the each of robust regression estimator itemized in equations (3.2) – (3.7)

Therefore, the generalized version of RKL-estimator is defined as:

$$\hat{\beta}_{RKL} = (X^T X + k_R I_p)^{-1} (X^T X - k_R I_p) \hat{\beta}_R. \quad (3.9)$$

where $k_R = \frac{\hat{\sigma}_{Robust}^2}{2\alpha_{iRobust}^2 + \frac{\hat{\sigma}_{Robust}^2}{\lambda_i}}$ is biased parameter for RKL- estimator, λ_i is ith eigenvalue

of $X^T X$, $\hat{\alpha}_{iRobust} = Q^T \hat{\beta}_R$, $\sigma_{Robust}^2 = \frac{(e_{iRobust})^T (e_{iRobust})}{n-p}$ is MSE of each of the robust regression estimator with full data such that $\hat{\beta}_R$ is each robust regression estimator defined in equations (3.2) – (3.7). Such that $e_{iRobust} = Y_i - \hat{Y}_i^{Robust}$ and $\hat{\beta}_R^{KL}$ is robust KL-estimator based on M, MM, LTS, LMS, S, LAD and LQS estimators.

3.2.1 The Canonical form of the Generalized RKL Estimator

Consider the general form of classical linear regression model:

$$y = X\hat{\beta} + U \quad (3.10)$$

Where $\hat{\beta}$ is the OLS estimates

The canonical form of equation (3.10) is expressed as:

$$y = F\hat{\theta} + U \quad (3.11)$$

Where $F = XT$ and $\theta = T^T \beta$, where β is the OLS estimator, T is the orthogonal matrix such that $F^T F = T^T X^T X T = \lambda = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$ where $\lambda_1, \lambda_2, \dots, \lambda_p > 0$ are ordered eigenvalues of $X^T X$.

Let $\hat{\theta}_R$ be any robust regression estimates of equations $\sum_{i=1}^n \mathcal{G}\left(\frac{e_i}{t}\right) = 0$ and

$$\sum_{i=1}^n \mathcal{G}\left(\frac{e_i}{t}\right) s_i = 0, \quad e_i = y_i - s_i^T \hat{\theta}_R \quad \text{where } t \text{ is a scale parameter and } \mathcal{G} \text{ is a selected useful}$$

function (Hampel *et al.*, 1986) and Huber (1981).

Canonical form of some existing estimators are as follows:

$$\hat{\theta} = \Gamma^{-1} F^T y, \quad (3.12)$$

where $\Gamma = F^T F$

$$\hat{\theta} = (\Gamma + kI_p)^{-1} F^T y, \quad (3.13)$$

$$\hat{\theta}_R(k) = (I_p + kF^{-1})\hat{\theta}_R, \quad (3.14)$$

$$\hat{\theta}_R(d) = (\Gamma + I_p)^{-1}(\Gamma + dI_p)\hat{\theta}_R, \quad (3.15)$$

$$\hat{\theta}_R(KL) = (\Gamma + kI_p)^{-1}(\Gamma - kI_p)\hat{\theta}_R, \quad (3.16)$$

$$\hat{\theta}_R(TP) = (\Gamma + kI_p)^{-1}(\Gamma + kdI_p)\hat{\theta}_R, \quad (3.17)$$

$$\hat{\theta}_R(DK) = (\Gamma + k(1+d)I_p)^{-1}(\Gamma - k(1+d)I_p)\hat{\theta}_R, \quad (3.18)$$

Hence, the canonical form of generalized RKL estimates of θ is defined as:

$$\hat{\theta}_R^{KL} = (\Gamma + k_R I_p)^{-1} (\Gamma - k_R I_p) \hat{\theta}_R. \quad (3.19)$$

where k_R is the generalized robust biasing parameter for generalized robust KL estimator

and it is each of robust biasing parameter for each of robust KL estimator.

3.2.2 Determination of Mean Squares Error (MSE) for (RKL) Estimator

Recall that, the MSE of OLS estimates $\hat{\theta}$ can be expressed as:

$$\begin{aligned} MSE(\hat{\theta}) &= E(\hat{\theta} - \theta)^T (\hat{\theta} - \theta), \\ &= tr(cov(\hat{\theta})) + bias(\hat{\theta})^T bias(\hat{\theta}). \end{aligned} \quad (3.20)$$

Hence, for practical purpose (3.20) can be written as:

$$MSE(\hat{\theta}) = \sum_{i=1}^p \sigma^2 / \lambda_i. \quad (3.21)$$

The MSE of robust regression can be expressed as:

$$MSE(\hat{\theta}_R) = \sum_{i=1}^p \Omega_{ii}. \quad (3.22)$$

where Ω_{ii} is diagonal element for $\text{cov}(\hat{\theta}_R)$ equivalent to Ω which is finite and p is the number of independent variables.

Hoerl and Kennard (1970) provided the MSE for Ridge estimator $\hat{\theta}(k)$ as follows:

$$MSE(\hat{\theta}(k)) = \sigma^2 \sum_{i=1}^p \frac{\lambda_i}{(\lambda_i + k)^2} + \sum_{i=1}^p \frac{k^2 \theta_i^2}{(\lambda_i + k)^2} \quad (3.23)$$

In (1993), Liu estimated the MSE for the Liu estimator $\hat{\theta}_d$ as follows:

$$MSE(\hat{\theta}_d) = \sigma^2 \sum_{i=1}^p \frac{(\lambda_i + d)^2}{(\lambda_i + 1)^2 \lambda_i} + (d - 1)^2 \sum_{i=1}^p \frac{\theta_i^2}{(\lambda_i + 1)^2} \quad (3.24)$$

Where d is biasing parameter of Liu estimator

MSE for robust ridge estimator $\hat{\theta}_R(k)$ is defined as:

$$MSE(\hat{\theta}_R(k)) = \sum_{i=1}^p \frac{\lambda_i^2 \Omega_{ii} + k^2 \theta_i^2}{(\lambda_i + k)^2} \quad (3.25)$$

Also, MSE for robust Liu estimator $\hat{\theta}_R(d)$ denoted as $MSE(\hat{\theta}_R(d))$ is given as:

$$MSE(\hat{\theta}_R(d)) = \sum_{i=1}^p \frac{(\lambda_i + d)^2 \Omega_{ii} + (1 - d)^2 \theta_i^2}{(\lambda_i + 1)^2} \quad (3.26)$$

MSE of robust Dawoud-Kibria estimator was expressed as:

$$MSE((DK)) = \sum_{i=1}^p \frac{(\mu_i - k(1 - d))^2 \Omega_{ii}}{(\mu_i + k(1 + d))^2} + \sum_{i=1}^p \frac{4k^2 (1 - d)^2 \alpha_i^2}{(\mu_i + k(1 + d))^2} \quad (3.27)$$

Therefore, following the submission of Lukman *et al.*, (2020), the MSE of generalized RKL can be expressed as:

$$MSE(\hat{\theta}_R(KL)) = \sum_{i=1}^p \frac{(\lambda_i - k)^2 \Omega_{ii}}{\lambda_i (\lambda_i + k)^2} + \sum_{i=1}^p \frac{4k^2 \theta_i^2}{(\lambda_i + k)^2} \quad (3.28)$$

In order to ascertain the effectiveness of the newly developed RKL estimator over others, the following conditions were upheld:

- (i) The function $\mathcal{J}$ should be non - decreasing and skew- symmetric function

- (ii) The residuals should be symmetric; they must have mean = zero and finite variance.
- (iii) The diagonal element for $\text{cov}(\hat{\theta}_R)$ which is Ω_{ii} is finite.

The fitted values for the generalized RKL estimator can be derived as follows.

$$\hat{Y}_i^{RKL} = X\hat{\beta}_R^{KL}. \quad (3.29)$$

$$\hat{Y}_i^{RKL} = X(X^T X + k_R I_p)^{-1}(X^T X - k_R I_p)\hat{\beta}_R. \quad (3.30)$$

$$\hat{Y}_i^{RKL} = X(X^T X + k_R I_p)^{-1}(X^T X - k_R I_p)(X^T X)^{-1}X^T Y. \quad (3.31)$$

$$\hat{Y}_i^{RKL} = H^{RKL} Y. \quad (3.32)$$

$$\text{where } H^{RKL} = X(X^T X + k_R I_p)^{-1}(X^T X - k_R I_p)(X^T X)^{-1}X^T,$$

This is equivalent to hat matrix of the OLS when $k_R = 0$. Since H^{RKL} is not an idempotent, therefore, not a projection matrix. Hence, its i th diagonal element of H^{RKL} is written as:

$$h_{ii}^{RKL} = x_i(X^T X + k_R I_p)^{-1}(X^T X - k_R I_p)(X^T X)^{-1}x_i^T. \quad (3.33)$$

where x_i denotes i th row of $X^T X$

Likewise, i th fitted value written in terms of the element of H^{RKL} is as follows:

$$\hat{Y}_i^{RKL} = \sum_{i=1}^n h_{ii}^{RKL} Y_i. \quad (3.34)$$

Therefore, equation (3.35) is the partial derivative of $\hat{Y}_i^{RKL}$ with respect to Y_i ;

$$\frac{\partial \hat{Y}_i^{RKL}}{\partial Y_i} = h_{ii}^{RKL}. \quad (3.35)$$

The error of i th RKL is defined as:

$$e_i^{RKL} = Y_i - \hat{Y}_i^{RKL}. \quad (3.36)$$

3.2.3 Generalized Version of Robust Cook's D in Robust Kibria-Lukman (RKL)

Estimator

Cook's Distance is defined as

$$D_i = \frac{(\hat{\beta}_i - \beta)' (X'X)(\hat{\beta}_i - \beta)}{P\sigma^2} \quad (3.37)$$

where $\hat{\beta}_i$ and $\hat{\beta}$ give estimates based on all observations (n) and compares them with the estimates obtained after deleting the (i)th observation. An equivalent algebraic expression is given as:

$$D_i = \frac{r_i^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right) \quad (3.38)$$

where h_{ii} is ith diagonal elements of hat matrix and r_i is ith studentised residual which is given as:

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}}.$$

$$\text{such that, } e_i = y_i - \hat{y} \text{ and } \hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - p}.$$

The influential points for which $D_i > 1$, warrants attention.

The generalized robust Cook's Distance based on Robust Kibria-Lukman estimator is obtained by replacing the conventional estimator and hat-matrix elements in Cook's Distance (Cook, 1977) with their robust Kibria-Lukman counterparts. Using the Sherman-Morrison-Woodbury theorem (Rao, 1973), the corresponding Robust Kibria-Lukman Cook's Distance measure is derived.

$$D_i^{RKL} = \frac{(r_{iRKL})^2}{(p+1)} \left(\frac{h_{ii}^{RKL}}{1 - h_{ii}^{RKL}} \right) \quad (3.39)$$

$$\text{where } r_{iRKL} = \frac{e_i^{RKL}}{\sqrt{MSE_{Robust}(1 - h_{ii}^{RKL})}}$$

such that $MSE_{Robust} = \frac{(e_i^{Robust})^T e_i^{Robust}}{n-p}$, $e_i^{RKL} = \sum_{i=1}^n (Y_i - \hat{Y}^{RKL})^2$ and $e_i^{Robust} = Y_i - \hat{Y}_i^{Robust}$.

h_{ii}^{RKL} is the diagonal element of an hat matrix H^{RKL} for Robust Regression KL estimator when data is full. The cut-off point $D_i^{RKL} > 1$ implies any observation that its value of $D_i^{RKL} > 1$ is influential.

3.2.4 Generalized Robust Version of DFFITS for the RKL Estimator

DFFITS is defined as:

$$DFFITS_i = \frac{\hat{Y} - \hat{Y}_{(-i)}}{\sqrt{S_{(-1)}^2 h_{ii}}} \quad (3.40)$$

where $\hat{Y}$ is the predicted value for all observation, $\hat{Y}_{(-i)}$ is the remaining observation obtained when i th case is omitted from a dataset, $S_{(-1)}^2$ is the estimated mean square error MSE of $\hat{Y}_{(-1)}$. Also, it could be expressed as:

$$DFFITS_i = \frac{x_i [\hat{\beta} - \hat{\beta}_{(i)}]}{S(x_i \hat{\beta})}. \quad (3.41)$$

where $S(x_i \hat{\beta})$ is estimate of standard error of fitted values, x_i is i th row of X matrix, $\hat{\beta}_{(i)}$ is the Ordinary Least Squares estimator of β , when the i th observation is omitted].

Algebraically, it could further be expressed as:

$$(DFFITS)_i = \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2} \frac{e_i}{\sqrt{\hat{\sigma}^2 (1-h_{ii})}} = \left(\frac{h_{ii}}{1-h_{ii}} \right)^{1/2} t_i. \quad (3.42)$$

where $\hat{\sigma}^2$ is estimate of σ , h_{ii} is the diagonal elements of hat matrix and

$t_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}$ is studentized residual (also called external studentized residual) Lukman

and Ayinde (2020).

The generalized version of the Difference in fits (DFFITS) for robust KL estimator is hereby obtained, following the conventional diagnostic for (DFFITS) proposed based on OLS:

$$DFFITS_i^{RKL} = \frac{|\hat{Y}_i^{RKL} - \hat{Y}_{(-i)}^{RKL}|}{S_{(-i)RKL} \sqrt{h_{ii}^{RKL}}}. \quad (3.43)$$

where $\hat{Y}_i^{RKL}$ is fitted model of the robust KL estimator when the data is complete and $\hat{Y}_{(-i)}^{RKL}$ is the fitted model for the robust KL estimator, when i th observation is deleted, $S_{(-i)RKL}$ is robust KL standard error estimate, when i th observation is deleted and h_{ii}^{RKL} is the diagonal elements of hat matrix $X(X^T X + k_R I_p)^{-1}(X^T X - k_R I_p)(X^T X)^{-1} X^T$.

The observation for which $|DFFITS_i^{RKL}| > 2 \sqrt{\frac{p}{n}}$ for large data, such observation can be considered to be influential.

3.2. 4.1 Derivation of Approximate Case- Deletion formulas for RKL estimator

In line with of objective three this subsection discusses the derivation of approximate Case -Deletion formulas for Cook's D and DFFITS in RKL framework;

If i th row is deleted from $\hat{\beta}_{RKL}$ in (3.9), then:

$$\hat{\beta}_{(i)RKL} = (G_{(i)} + kI_p)^{-1}(F_{(i)} + kI_p)\hat{\beta}_{(i)R} \quad (3.44)$$

where $G_{(i)} = X_{(i)}^T X_{(i)}$ and $X_{(i)}$ is the matrix X without the i th row. $X_{(i)}$ is scaled so that $G_{(i)}$ is in correlation form.

Hence, $\hat{\beta}_{(i)RKL}$ is obtained as follows by imposing Sherman-Morrison-Woodbury (SMW)

theorem:

$$\hat{\beta}_{(i)RKL} = (G - x_i^T x_i + kI_p)^{-1} (G - x_i^T x_i + kI_p) \hat{\beta}_{(i)R} \quad (3.45)$$

$$\hat{\beta}_{(i)RKL} = (F_k - x_i^T x_i)^{-1} (G - x_i^T x_i + kI_p) \hat{\beta}_{(i)R} \quad (3.46)$$

where, $F_k = G + kI_p$, such that $G = X^T X$ Therefore,

$$\hat{\beta}_{(i)RKL} = \left(F_k - \frac{F_k^{-1} x_i^T x_i F_k^{-1}}{1 - x_i F_k^{-1} x_i^T} \right)^{-1} (G - kI_p - x_i^T x_i) \hat{\beta}_{(i)R} \quad (3.47)$$

where, $\hat{\beta}_{(i)RKL} = (F_k - x_i^T x_i)^{-1} = \left(F_k^{-1} + \frac{F_k^{-1} x_i^T x_i F_k^{-1}}{1 - x_i F_k^{-1} x_i^T} \right)$ such that, $L_{ii} = x_i F_k^{-1} x_i^T$

By expanding (3.47), it leads to (3.48)

$$\begin{aligned} \hat{\beta}_{(i)RKL} &= F_k^{-1} (F_k - kI_p) \hat{\beta}_{(i)R} - \left(\frac{F_k^{-1} x_i^T x_i F_k^{-1}}{1 - L_{ii}} \right) \hat{\beta}_{(i)R} + \\ &\left(\frac{F_k^{-1} x_i^T x_i F_k^{-1}}{1 - L_{ii}} \right) (G - kI_p) \hat{\beta}_{(i)R} \end{aligned} \quad (3.48)$$

$$\hat{\beta}_{(i)RKL} = \hat{\beta}_{RKL} + \frac{F_k^{-1} x_i^T}{1 - L_{ii}} [x_i F_k^{-1} (G - kI_p) \hat{\beta}_{(i)R} - x_i^T F_k^{-1} x_i^T x_i \hat{\beta}_{(i)R} - (1 - L_{ii}) \hat{\beta}_{(i)R}]$$

$$\hat{\beta}_{(i)RKL} = \hat{\beta}_{RKL} + \frac{F_k^{-1} x_i^T}{1 - L_{ii}} [\hat{y}_{iRKL} - L_{ii} y_i - (1 - L_{ii}) y_i] \quad (3.49)$$

where, $\hat{y}_{iRKL} = x_i F_k^{-1} (G - kI_p) \hat{\beta}_{(i)R} = x_i F_k^{-1} (G - kI_p) G^{-1} X^T y_i$ and

$$\hat{\beta}_{RKL} = F_k^{-1} (G - kI_p) \hat{\beta}_{(i)R}$$

$$\hat{\beta}_{(i)RKL} = \hat{\beta}_{RKL} + \frac{F_k^{-1} x_i^T}{1 - L_{ii}} (\hat{y}_{iRKL} - y_i) \quad (3.50)$$

$$\hat{\beta}_{(i)RKL} = \hat{\beta}_{RKL} - \frac{F_k^{-1} x_i^T}{1 - L_{ii}} (y_i - \hat{y}_{iRKL}) \quad (3.51)$$

$$\hat{\beta}_{(i)RKL} = \hat{\beta}_{RKL} - \frac{F_k^{-1} x_i^T e_i^{RKL}}{1 - L_{ii}}. \quad (3.52)$$

where, $e_i^{RKL} = (y_i - \hat{y}_{iRKL}) = (1 - h_{ii}^{RKL})y_i$

$$\text{Therefore, } \hat{\beta}_{RKL} - \hat{\beta}_{(i)RKL} \cong \frac{e_i^{RKL} F_k^{-1} x_i^T}{1 - L_{ii}} \quad (3.53)$$

Theorem I: Approximate case-deletion formula of Cook's Distance based on RKL-estimator

Approximate case- deletion formula for D_{iRKL} is written as:

$$D_{iRKL} = \frac{1}{pS^2} \left[\frac{e_i^{RKL} F_k^{-1} x_i^T}{1 - L_{ii}} \right]^T G \left[\frac{e_i^{RKL} F_k^{-1} x_i^T}{1 - L_{ii}} \right] \quad (3.54)$$

$$= \frac{1}{pS^2} \frac{x_i^T}{(1 - L_{ii})^2} F_k^{-1} (e_i^{RKL})^2 G F_k^{-1} x_i^T \quad (3.55)$$

where $F_k^{-1} = (G + kI_p)^{-1}$

Therefore,

$$\begin{aligned} D_{iRKL} &= D_{iRKL}^* \frac{1}{pS^2} \left[\frac{e_i^{RKL}}{1 - L_{ii}} \right]^2 x_i^T (G + kI_p)^{-1} (G + kI_p)^{-1} x_i^T \\ &= \frac{1}{pS^2} \left[\frac{e_i^{RKL}}{1 - L_{ii}} \right]^2 \sum_{j=1}^n L_{ij}^2 \end{aligned} \quad (3.56)$$

Also, approximate case -deletion formula for D_{iRKL} can be obtained as follows:

$$D_{iRKL}^{a*} = \frac{1}{pS^2} \left[\frac{e_i^{RKL}}{1 - L_{ii}} \right]^2 x_i^T (G + kI_p)^{-1} (G + kI_p)^{-1} G x_i^T (G + kI_p) (G + kI_p)^{-1} x_i^T \quad (3.57)$$

$$D_i^{RKL} = \frac{1}{pS^2} \left[\frac{e_i^{RKL}}{1 - L_{ii}} \right]^2 \sum_{j=1}^n (h_{ij}^{RKL})^2 \quad (3.58)$$

Theorem II: The approximate case-deletion formula for DFFITS based on RKL estimator

Approximate case-deletion formula for DFFITS based on RKL estimator is expressed as:

$$DFFITS_i^{RKL} = \left[\frac{L_{ii}}{1 - L_{ii}} \right] \frac{e_i^{RKL}}{S(x_i \hat{\beta}_{RKL})} \quad (3.59)$$

$$DFFITS_i^{RKL} = \left[\frac{L_{ii}}{1 - L_{ii}} \right] \frac{e_i^{RKL}}{S_{(i)} \sqrt{\sum_{j=1}^n (h_{ij}^{RKL})^2}} \quad (3.60)$$

where $S_{(i)} = \sqrt{\frac{(n-p)S^2 - \frac{e_i^2}{1-h_{ii}}}{n-p}}$ such that $L_{ii} = x_i F_k^{-1} x_i^T$ and $F_k^{-1} = (G + kI_p)^{-1}$

3.2.5 Generalized version of Robust Covratio Measure in RKL Estimator

The statistic for Covratio influential Measure is defined as:

$$COVRATIO_i = \left[\frac{\det(S_i^2 (X_i^T X_i)^{-1})}{\det(S^2 (X^T X)^{-1})} \right] \quad (3.61)$$

The equation 3.61 is further defined as:

$$COVRATIO_i = \frac{\left[\frac{n-p-r_i^2}{n-p-1} \right]}{1-h_{ii}} \quad (3.62)$$

Where $r_i = \frac{e_i}{\hat{\sigma} \sqrt{1-h_{ii}}}$, p is number of independent variables.

COVRATIO measure can alternatively be written as:

$$COVRATIO_i = \frac{(S_{(-i)}^2)^{(p+1)}}{MSE^{(p+1)}} \left(\frac{1}{1-h_{ii}} \right) \quad (3.63)$$

where $MSE^{(p+1)} = \frac{SSE}{n-p+1}$, such that $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and $S_{(-i)}^2$ is the variance when ith

observation is omitted, $(S_{(-i)}^2)^{(p+1)}$ is the variance when ith observation is omitted with

constant. When $|COVRATIO_i - 1| \geq \frac{3p}{n}$, the observation is considered Influential

Following the procedures of COVRATIO measure proposed by Belsley *et al.*, (1980) based on OLS, generalized version of COVRATIO measure based on robust KL estimator is hereby defined as:

$$CVR_i^{RKL} = \frac{(S_{(-i)RKL}^2)^{(p+1)}}{MSE_R} \left(\frac{1}{1 - h_{ii}^{RKL}} \right) \quad (3.64)$$

where $MSE_R = \frac{(e_i^R)^T e_i^R}{n-p}$, such that $e_i^R = Y_i - \hat{Y}_i^R$ and $S_{(-i)RKL}^2$ is the variance for the robust

KL estimator when ith observation is deleted.

Any observation for which $|CVR_i^{RKL} - 1| \geq \frac{3p}{n}$ is influential.

3.2.6 Generalized version of Robust Hadi's Measure for the RKL Estimator

The statistic for Hadi's Influential Measure based on OLS is expressed as:

$$H_i^2 = \frac{h_{ii}}{1 - h_{ii}} + \left(\frac{p}{1 - h_{ii}} \right) \left(\frac{d_i^2}{1 - d_i^2} \right). \quad (3.65)$$

where, $d_i = \frac{e_i}{\sqrt{SSE}}$, p represent the number of explanatory variables, while h_{ii} is its hat

matrix.

The cut-off point for Hadi's statistic in equation (3.65), is given as:

$H_i^2 = mean(H_i^2) + c\sqrt{var(H_i^2)}$, where 'c' is a constant chosen appropriately which is 2 or

3 according to El-Fallah and El-Salam (2013).

The generalized Hadi's measure for the Robust Kibria–Lukman estimator is obtained by replacing the conventional OLS hat-matrix diagonal elements and residuals with their Robust KL counterparts. The resulting diagnostic measure is expressed as:

$$H_{iRKL}^2 = \frac{h_{ii}^{RKL}}{1 - h_{ii}^{RKL}} + \left(\frac{p}{1 - h_{ii}^{RKL}} \right) \left(\frac{d_{iRKL}^2}{1 - d_{iRKL}^2} \right). \quad (3.66)$$

where, $d_{iRKL} = \frac{e_i^{RKL}}{\sqrt{SSE_{iRKL}}}$, such that $e_i^{RKL} = Y_i - \hat{Y}_i^{RKL}$ $SSE_{iRKL} = \sum_{i=1}^n (Y_i - \hat{Y}_i^{RKL})^2$ and p is the number of explanatory variables and h_{ii}^{RKL} is the diagonal elements of it hat matrix for robust KL estimator.

The point for which its $H_{iRKL}^2 = \text{mean}(H_{iRKL}^2) + c\sqrt{\text{var}(H_{iRKL}^2)}$ is an influential point, where c is a constant chosen appropriately to equal 2.

3.2.7 Generalized version of Robust Pena's Statistic in RKL Estimator

The Pena's measure based on OLS for ith case is defined as:

$$S_i = \frac{1}{p\hat{\sigma}^2 h_{ii}} \sum_{j=1}^n \frac{h_{ji}^2 e_j^2}{(1 - h_{jj})^2}. \quad (3.67)$$

where h_{ii} is ith diagonal element and h_{ji} is jth element of the hat matrix H .

The generalized version for Pena's measure based on robust KL estimator is obtained by replacing diagonal elements of hat matrix of fitted values and residual used in conventional Pena's statistic based on ordinary least squares estimator with diagonal element of its hat matrix denoted by h_{ii}^{RKL} , e_i^{RKL} obtained from robust KL estimator.

Hence, following Emami and Emami (2016) the robust influential measure can be expressed as:

$$S_{iRKL} = \frac{1}{p\hat{\sigma}_R^2 h_{ii}^{RKL}} \sum_{j=1}^n \frac{(h_{ji}^{RKL})^2 e_{jRKL}^2}{(1 - h_{ji}^{RKL})^2}. \quad (3.68)$$

where $e_{jiRKL} = Y_i - \hat{Y}_i^{RKL}$, $\sigma_R^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i^R)^2}{n - p}$, h_{ii}^{RKL} is i th diagonal element of hat matrix H^{RKL} . The cut-off point for the influential measure is given as: $|S_{iRKL} - \text{median}(S_{iRKL})| \geq 4.5 \text{MAD}(S_{iRKL})$.

3.2.8 Generalized version of Robust Atkinson Diagnostic in RKL Estimator

Atkinson (1985) defined the **Atkinson diagnostic** statistic as:

$$A_i = \frac{(\hat{\beta} - \hat{\beta}_{(-i)})^T X^T X (\hat{\beta} - \hat{\beta}_{(-i)}) (n - p - 1)}{(p + 1) s_{(-i)}^2}. \quad (3.69)$$

where $\hat{\beta}$ is OLS estimates of parameter coefficients for full data, $\hat{\beta}_{(-i)}$ is OLS parameter estimates when i th observation is deleted. He further expressed the statistic as follows:

$$A_i = \frac{r_i^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right) \quad (3.70)$$

such that $r_{(-i)} = \frac{e_{(-i)}}{\sqrt{\hat{\sigma}_{(-i)}^2 (1 - h_{ii})}}$, $e_{(-i)} = y_i - \hat{y}_{(-i)}$ and $\sigma_{(-i)}^2 = \frac{\sum_{i=1}^n e_{(-i)}^2}{n - p}$.

any observation that its $A_i > 1$ is influential.

Furthermore, generalized version of the Atkinson diagnostic measure based on robust KL estimator can be obtained using the conventional Atkinson diagnostic based on ordinary least squares and is defined as:

$$ATK_i^{RKL} = \frac{r_{(-i)RKL}^2}{p} \left(\frac{h_{ii}}{1 - h_{ii}} \right) \quad (3.71)$$

such that $r_{(-i)RKL} = \frac{e_{(-i)}^{RKL}}{\sqrt{\hat{\sigma}_{(-i)RKL}^2 (1 - h_{ii}^{RKL})}}$, $e_{(-i)}^{RKL} = Y_i - \hat{Y}_{(-i)}^{RKL}$ and $\sigma_{(-i)RKL}^2 = \frac{(e_{(-i)}^{RKL})^T (e_{(-i)}^{RKL})}{n - p}$.

Any observation that its $ATK_i^{RKL} > 1$ is influential.

Table 3.1: Summary of Mathematics of Generalised version of some proposed Influential Measures in RKL estimator

S/N	Estimator	Influential Measure	Mathematics	Cut off point
	RKL	Cook's D version (CKD)	$D_i^{RKL} = \frac{(r_{iRKL})^2}{p} \left(\frac{h_{ii}^{RKL}}{1 - h_{ii}^{RKL}} \right)$ $r_{iRKL} = \frac{e_i^{RKL}}{\sqrt{MSE_{RKL}} (1 - h_{ii}^{RKL})}$ where,	$D_i^{RKL} > 1$
		DFFITs version (DFT)	$DFFITs_i^{RKL} = \frac{ \hat{Y}_i^{RKL} - \hat{Y}_{(-i)}^{RKL} }{S_{(-i)RKL} \sqrt{h_{ii}^{RKL}}}$	$DFFITs_i^{RKL} > 2\sqrt{\frac{p}{n}}$
		Hadi's Measure Version (HAD)	$H_{iRKL}^2 = \frac{h_{ii}^{RKL}}{1 - h_{ii}^{RKL}} + \left(\frac{p}{1 - h_{ii}^{RKL}} \right) \left(\frac{d_{iRKL}^2}{1 - d_{iRKL}^2} \right)$	$H_{iRKL}^2 = \text{mean}(H_{iRKL}^2) + c\sqrt{\text{var}(H_{iRKL}^2)}$
		Pena's Statistic Version (PEN)	$S_{iRKL} = \frac{1}{p\hat{\sigma}_{Robust}^2 h_{ii}^{RKL}} \sum_{j=1}^n \frac{(h_{ji}^{RKL})^2 e_{jRKL}^2}{(1 - h_{ji}^{RKL})^2}$	$ S_{iRKL} - \text{median}(S_{iRKL}) \geq 4.5MAD(S_{iRKL})$
		COVRATIO Measure (CVR)	$COVRATIO_i^{RKL} = \frac{(S_{(-i)RKL}^2)^{(p+1)}}{MSE_{Robust}} \left(\frac{1}{1 - h_{ii}^{RKL}} \right)$	$ COVRATIO_i^{RKL} - 1 \geq \frac{3p}{n}$
		Atkinson Diagnostic (ATK)	$ATK_i^{RKL} = \frac{r_{(-i)RKL}^2}{p} \left(\frac{h_{ii}^{RKL}}{1 - h_{ii}^{RKL}} \right).$	$ATK_i^{RKL} > 1$

The Table 3.1 shows the generalised version of all influential measures (Cook's Distance, DFFITS, COVRATIO, Pena's statistics, Hadi's measure and Atkinson Statistics) based on the robust Regression Kibria- Lukeman that were obtained in this study. For ease of reference, the table also shows the mathematics and cut off points for the diagnostic.

3.3 Simulation procedure to generate exogenous variables, response variables and estimated Mean Squares Error.

The simulation studies of this research was conducted using R-statistical codes. All the exogenous variables were generated using equation given in equation (3.72) used by Lukman *et al.* (2020).

$$x_{ij} = (1 - \rho^2)^{\frac{1}{2}} z_{ij} + \rho z_{ip+1}, i = 1, 2, \dots, n, j = 1, 2, \dots, p. \quad (3.72)$$

Where

z_{ij} are independent standard normal pseudo-random numbers,

ρ^2 means the correlation between any two exogenous variables.

To ensure different degree of multicollinearity among explanatory variables, the following five (5) different levels of correlations were considered: 0.8, 0.85, 0.9, 0.95 and 0.99. The numbers of exogenous variables $p = 3$. This variables were expressed in a standardized form. Also, the response variables was generated using the following equations:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_{ip} + e_i, i = 1, 2, 3, \dots, p. \quad (3.73)$$

Where e_i = error term independently and identically normally distributed with mean zero (0) ,

variance σ^2 that is $e_i \sim iidN(0, \sigma^2)$. Zero intercept was assumed for the model in (3.73)

and the values of β was chosen to satisfy the constraints $\beta^T \beta = 1$ suggested by Lukman *et al.*, (2020). The studies were replicated 1000 times for each sample size $n = 10, 20, 30, 40, 50, 100, 250$ and 500 , with standard deviation (1, 5, and 10) used in a research conducted by Kaciranlar and Sakallioğlu (2001), Chang and Yang (2012), Lukamn *et al.*,

(2020) and Ayinde, *et al.*, (2020). Furthermore, for each replication, MSE for each estimator was obtained as

$$MSE(\beta^*) = \frac{1}{1000} \sum_{i=1}^{1000} (\beta^* - \beta)^2 \quad (3.74)$$

Where

β^* is estimator of both existing and proposed estimator (both ordinary and robust estimators). In the same vain, individual estimated coefficient parameters (both ordinary and robust estimators) considered in the above research were used.

3.3.1 Parameters for Simulation

The simulation study of this research was conducted using R-programming codes with the following parameters.

- (i) Sample size (n) = (10, 20, 30, 40, 50, 100, 250 and 500)
- (ii) Equation to generate exogenous variables $x_{ij} = (1 - \rho^2)^{\frac{1}{2}} z_{ij} + \rho z_{ip+1}$, $i = 1, 2, \dots, n$,
 $j = 1, 2, \dots, p$
- (iii) Magnitude of outliers in the x-direction (Mx) = (0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10)
- (iv) percentage of outliers in the x-direction (px) = (10% and 20%)
- (v) Point of outliers; $X(i)_{\text{outlier}} = Mx * \text{Max}(X_i) + X_i$
- (vi) $\beta_i = (0, 1)$
- (vii) $e_i \sim iidN(0, \sigma^2)$
- (viii) Replication (Rep) = 1000
- (xi) Levels of multicollinearity (0.8, 0.85, 0.9, 0.95 and 0.99)
- (xii) $MSE(\beta^*) = \frac{1}{1000} \sum_{i=1}^{1000} (\beta^* - \beta)^2$

3.3.2 The Algorithm for generating explanatory variables, Response variable, Error terms, and Mean Squares Error

- (i) Choose sample size n to work with..
- (ii) Generate exogenous variables using the equation: $x_{ij} = (1 - \rho^2)^{\frac{1}{2}} z_{ij} + \rho z_{ip+1}$, $i = 1, 2, \dots, n, j = 1, 2, \dots, p$
- (iii) Choose a percentage of the dataset to be replaced with outliers.
- (iv) Choose a particular magnitude of outlier to invoke into the data to be randomly generated in step 2
- (v) Randomly select observations to make up the percentage of the generated dataset to be replaced with outliers.
- (vi) Invoke outliers into dataset using; $X(i)_{\text{outlier}} = Mx * \text{Max}(X_i) + X_i$ where Mx is the magnitude of outliers in x-direction
- (vii) Replace outliers in original data generated in (iii)
- (viii) Generate response variable using:
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots + \beta_p x_{ip} + e_i, i = 1, 2, 3, \dots, p.$$
- (ix) Obtain MSE of the estimators in step (viii)
- (x) Compute the estimated MSE for each replicate and for each of the estimators by dividing the result in step (ix) by the number of replications.
- (xi) Choose another magnitude of outliers to invoke into the data and repeat step (iv) to (xi) until the magnitudes of outliers are exhausted.
- (xii) Repeat step (ii) to (xii) until the sample sizes are exhausted.

3.4 Determination of Influential point and Percentage of influential point detected

A number of significant influential measures covered in sections 3.2.3 through 3.2.8 were obtained and compared with their different cut-off point. Similarly, the percentage (%) of

influential point was calculated, that is the total number of influential points found were divided by the total number of outliers imposed in section 3.3.2 step (iv).

3.4.1 Criterion for Investigation of the Proposed Influential Measures

The following criteria were used for the selection of the best influential measures:

- i. **Hundred percent detection:** The best influential measure was defined as the diagnostic tool or influential measure with the highest count at 100% detection.
- iii. **Highest frequency:** The highest frequency was used to identify the overall best influential measure. When the specifications, which include percentage of outliers (P_x), magnitude of outliers (M_x), error variances, different levels of multicollinearity (R), and sample size (n), are totaled, the influential measure with the highest frequency is deemed to be the best overall.

CHAPTER FOUR

RESULTS AND DISCUSSION

This Chapter discuss results obtained from the new influential measures based on robust Regression- KL Estimators when applied to both simulation study and real- life data. Their performances were also compared to existing tools.

4.1 Investigation of the Developed Influential Measures in the presence of multicollinearity and outliers

The results obtained from the simulation studies were presented in Tables 4.1, 4.2a to 4.2g and graphically illustrated in Figures 4.1, 4.2a to 4.2g respectively. They displayed the frequencies and ranks of the best consistent influential measures that detected correctly, all influential points (100%), when counted over different percentage of outliers, magnitude of outliers, levels of multicollinearity, different variance and sample sizes.

4.1.1 Investigation of the Influential measures without Robust Regression

The result of simulation study for the influential measures based on OLS that correctly detected all influential points (100%), at different sample sizes (10, 20, 30, 40, 50, 100, 250 and 500) such as error variances (1, 5 and 10), multicollinearity levels (0, 0.8, 0.9, 0.95 and 0.99), percentage of outliers (10% and 20%), and magnitude of outliers (0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10).

Table 4.1: Frequency of the influential measures proposed based on OLS estimator that correctly detect Influential points (100%) when counted over all the specifications

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_OLS	9	0	9	237	330	330	329	329	1573	1
HAD_OLS	126	94	110	0	47	0	0	0	377	2
PEN_OLS	134	0	0	0	0	0	0	0	134	3
ATK_OLS	29	0	0	0	0	0	0	0	29	4

Source: Author's analysis, 2025

The table 4.1 shows Covratio influential measure proposed based on OLS (CVR_OLS) has a total count of 1573 which is far from Hadi Measure proposed based on OLS (HAD_OLS) which has a total count of 377. It presents the performance of influential measures based on OLS that correctly detected Influential points (100%) at different sample sizes when counted over all specifications considered in the study which includes; sample sizes (10, 20, 30, 40, 50, 100, 250 and 500), error variances (1, 5 and 10), multicollinearity levels (0, 0.8, 0.9, 0.95 and 0.99), percentage of outliers (10% and 20%), and magnitude of outliers (0, 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10). COVRATIO proposed based on OLS (CVR_OLS) which has the highest counts indicates that it can detect influential points better than others considered in the study when data is contaminated with multicollinearity and outliers without being curtailed.

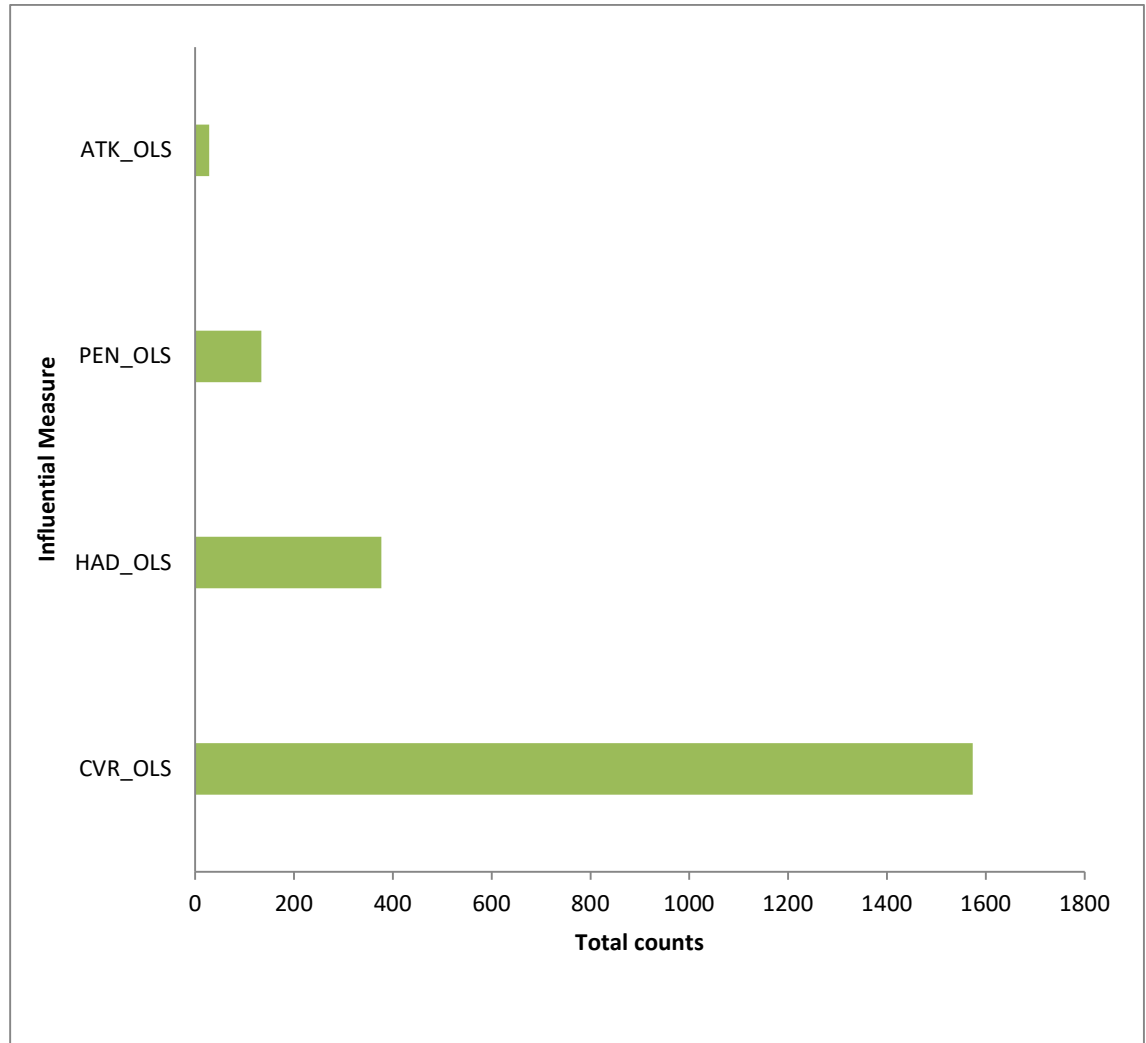


Figure 4.1: The graphical representation of best influential measures proposed based on OLS that detected all influential points (100%), when counted over all specifications.

From the graph shown in Fig 4.1 CVR_OLS has the highest bar followed by HAD_OLS while ATK has the lowest Bar when counted over of all the specification. In other words, the Bar Chart shows the performance of the Influential measure without Robust Regression and affirms that CVR_OLS outperform all the other diagnostic measures when multicollinearity and outliers are not mitigated.

Table 4.2a: Frequency of the influential measures proposed based on Robust-M estimator which detected all Influential points (100%), when counted over all the specifications

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_KL_M	314	330	330	329	330	330	330	329	2622	1
DFT_KL_M	42	103	149	214	227	243	276	278	1532	2
CVR_M	0	0	0	177	330	330	330	330	1497	3
HAD_M	118	89	108	0	47	0	0	0	362	4
HAD_RDG_M	110	94	101	0	45	0	0	0	350	5
HAD_KL_M	114	102	106	0	8	0	0	1	331	6
ATK_M	25	0	0	0	0	0	0	0	25	7

Source: Author's analysis, 2025

Table 4.2a shows Covratio influential measure proposed based on Kibra-Lukman _M (CVR KL_M) has a total count of 2622 followed by DFFIT influential measure proposed based on Kibra-Lukman _M (DFT KL_M) which has a total count of 1532. As sample size increases the frequency of 100% detection of influential point increases for some Kibra-Lukman _M diagnostics. These measures perform better as sample size increases. However, HAD_M, HAD_RDG_M, HAD_KL_M and ATK_M were not able to detect influential point as sample size increased to 100, while CVR_M performed woefully at small sample size.

Table 4.2b: Frequency of the influential measures proposed based on Robust-MM estimator that detected all Influential points (100%), when counted over all the specifications

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_KL_MM	108	267	311	324	327	330	330	328	2325	1
DFT_KL_MM	42	103	157	214	232	235	276	276	1535	2
CVR_MM	0	0	0	0	198	276	330	328	1132	3
HAD_MM	124	86	100	0	47	0	0	0	357	4
HAD_KL_MM	118	103	106	0	0	0	0	0	327	5
PEN_MM	128	0	0	0	0	0	0	0	128	6
PEN_KL_MM	10	0	0	0	0	0	0	0	10	7

Source: Author's analysis, 2025

Table 4.2b shows Covratio influential measure proposed based on Kibra-Lukman_MM (CVR_KL_MM) has a total count of 2325 followed by DFFIT influential measure proposed based on Kibra-Lukman_MM (DFT_KL_MM) which has a total count of 1535 and Pena's Measure (PEN_KL_MM) has a total count of 10. The table shows as sample size increases, the frequency of 100% detection of influential point increases, for some Kibra-Lukman_MM diagnostics. These measures perform better as sample size increases. However, HAD_MM, HAD_RDG_MM, HAD_KL_MM and ATK_MM were not able to detect influential point as sample size increased to 100, while CVR_MM performed woefully, at small sample sizes

Table 4.2c: Frequency of influential measures proposed based on Robust-LAD estimator that detected all Influential points (100%), when counted over all the specifications

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_KL_LAD	300	326	330	330	330	330	330	327	2603	1
DFT_KL_LAD	43	102	149	214	227	244	275	276	1530	2
CVR_LAD	0	0	0	21	330	330	330	327	1338	3
HAD_LAD	119	0	107	0	47	0	0	0	273	4
HAD_RGD_LAD	107	0	102	0	46	0	0	0	255	5
HAD_KL_LAD	112	0	107	0	8	0	0	0	227	6
ATK_KL_LAD	8	0	0	0	0	0	0	0	8	7

Source: Author's analysis, 2025

Table 4.2c shows Covratio influential measure proposed based on Kibra-Lukman_LAD (CVR_KL_LAD) has a total count of 2603 to rank 1st followed by DFFIT influential measure proposed based on Kibra-Lukman_LAD (DFT_KL_LAD) which has a total count of 1530 and Atkinson's Measure (ATK_KL_LAD) has a total count of 8 to rank 7th. The table shows that some influential measure proposed base on Kibra-Lukman_LAD perform better as sample size increases. However, HAD_LAD, HAD_RDG_LAD and HAD_KL_LAD were not able to detect influential point as sample size increased to 100, while CVR_LAD performed woefully at small sample sizes. ATK_LAD perform poorly at all sample sizes as it was only able to detect 100 % influential point 8 times at the sample size of 10

Table 4.2d: Frequency of the influential measures proposed based on Robust-LMS estimator that detected all Influential points (100%), when counted over all the specifications

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_KL_LMS	18	157	236	294	313	330	329	330	2007	1
DFT_KL_LMS	43	103	151	215	232	235	274	278	1531	2
CVR_LMS	0	0	0	0	0	0	228	330	558	3
HAD_LMS	132	85	109	0	47	0	0	0	373	4
HAD_KL_LMS	119	105	108	0	0	0	0	0	332	5
PEN_LMS	135	0	0	0	0	0	0	0	135	6
PEN_KL_LMS	9	0	0	0	0	0	0	0	9	7

Source: Author's analysis, 2025

Table 4.2d shows Covratio influential measure proposed based on Kibra-Lukman _LMS (CVR_KL_LMS) has a total count of 2007 to rank 1st followed by DFFIT influential measure proposed based on Kibra-Lukman _LMS (DFT_KL_LMS) which has a total count of 1531 to rank 2nd and Pena's Measure (PEN_KL_LMS) has a total count of 9 to rank 7th. The table further shows that some influential measure proposed base on Kibra-Lukman _LMS perform better as sample size increases. However, HAD_LMS, HAD_RDG_LMS and HAD_KL_LMS were not able to detect influential point as sample size increased to 100, while CVR_LMS performed woefully at small sample sizes, PEN_KL_LMS performed well at sample size 10. PEN_LMS perform poorly at all sample sizes as it was only able to detect 100 % influential point 9 times at the sample size of 10

Table 4.2e: Frequency of the influential measures proposed based on Robust-Least Quantile Squares (LQS) estimator that detected all Influential points (100%), when counted over all the specifications

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_KL_LQS	56	178	245	300	313	330	330	329	2081	1
DFT_KL_LQS	42	102	151	215	232	236	275	278	1531	2
CVR_LQS	0	0	0	0	0	0	330	329	659	3
HAD_LQS	128	88	108	0	47	0	0	0	371	4
HAD_KL_LQS	119	100	109	0	0	0	0	0	328	5
PEN_LQS	134	0	0	0	0	0	0	0	134	6
CKD_LQS	18	0	0	0	0	0	0	0	18	7
PEN_KL_LQS	17	0	0	0	0	0	0	0	17	8

Source: Author's analysis, 2025

Table 4.2e shows Covratio influential measure proposed based on Kibra-Lukman_LQS (CVR_KL_LQS) has a total count of 2081 to rank 1st followed by DFFIT influential measure proposed based on Kibra-Lukman_LQS (DFT_KL_LQS) which has a total count of 1531 to rank 2nd and Pena's Measure (PEN_KL_LQS) has a total count of 8 to rank 7th. The table further shows that some influential measure proposed base on Kibra-Lukman_LQS perform better as sample size increases. However, HAD_LQS, HAD_RDG_LQS and HAD_KL_LQS were not able to detect influential point as sample size increased to 100, while CVR_LQS performed woefully at small sample sizes, CKD_LQS performed well at sample size 10. In addition, PEN_KL_LQS and PEN_LQS perform poorly at all sample sizes as it was only able to detect 100 % influential point 18 and 17 times respectively at the sample size of 10

Table 4.2f: Frequency of the influential measures proposed based on Robust-Least Trimmed Squares (LTS) estimator that detected all Influential points (100%), when counted over all the specifications

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_KL_LTS	57	176	251	292	316	330	330	330	2082	1
DFT_KL_LTS	42	102	151	215	231	235	276	278	1530	2
CVR_LTS	0	0	0	0	0	0	234	330	564	3
HAD_KL_LTS	121	103	110	0	0	0	0	0	334	4
HAD_LTS	133	85	108	0	47	0	0	0	373	5
PEN_LTS	134	0	0	0	0	0	0	0	134	6
ATK_LTS	29	0	0	0	0	0	0	0	29	7
PEN_KL_LTS	13	0	0	0	0	0	0	0	13	8

Source: Author's analysis, 2025

Table 4.2f shows Covratio Influential measure proposed based on Kibra-Lukman _LTS(CVR KL_LTS) has a total count of 2082 followed by DFFIT influential measure proposed based on Kibra-Lukman _LTS (DFT KL_LTS) which has a total count of 1530 and Pena's Measure (PEN_KL_LTS) has a total count of 8. The table further shows that some influential measure proposed base on Kibra-Lukman _LTS perform better as sample size increases. However, HAD_LTS, HAD_RDG_LTS and HAD_KL_LTS were not able to detect influential point as sample size increased to 100, while CVR_LTS performed woefully at small sample sizes, PEN_LTS performed well at sample size 10. In addition, ATK_LTS and PEN_KL_LTS perform poorly at all sample sizes as it was only able to detect 100 % influential point 29 and 13 times respectively at the sample size of 10.

Table 4.2g: Frequency of the influential measures proposed based on Robust-S estimator that detected all Influential points (100%), when counted over all the specifications

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_KL_S	102	212	280	316	327	330	330	327	2224	1
DFT_KL_S	42	99	151	214	232	235	271	231	1475	2
HAD_S	126	72	106	0	47	0	0	10	361	3
HAD_KL_S	119	92	110	0	0	0	0	0	321	4
CVR_S	0	0	0	0	0	252	23	15	290	5
PEN_S	126	0	0	0	0	0	0	0	126	6
ATK_S	28	0	0	0	0	0	0	0	28	7

Source: Author's analysis, 2025

Table 4.2g shows Covratio Influential measure proposed based on Kibra-Lukman _S (CVR KL_S) has a total count of 2224 followed by DFFITS influential measure proposed based on Kibra-Lukman _S (DFT KL_S) which has a total count of 1475 and Atkinson's Measure based on S (ATK__S) has a total count of 28. The table further shows that some influential measure proposed base on Kibra-Lukman _S perform better as sample size increases. However, HAD_S and HAD_KL_S were not able to detect influential point as sample size increased to 100, while CVR_S performed woefully at small sample sizes, PEN_S performed well at sample size 10 by detecting influential observations, 126 times. In addition, ATK_S performed poorly at all sample sizes as it was only able to detect 100 % influential point 28 times at the sample size of 10.

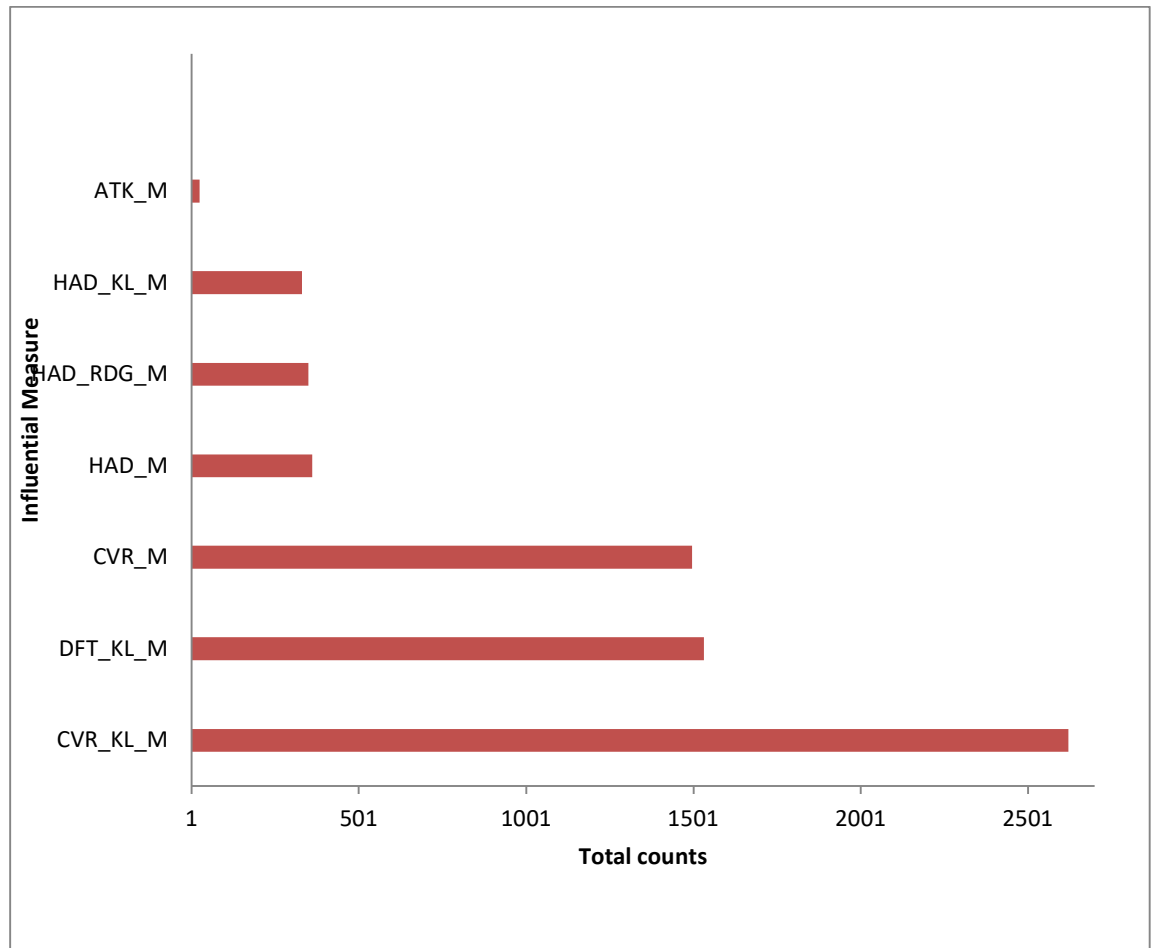


Figure 4.2a: The graphical representation of best influential measures proposed based on Robust-M estimator that detected all influential points (100%), when counted over all specifications.

Fig 4.2a shows that CVR_KL_M has the highest bar followed by DFT_KL_M while ATK has the smallest Bar when counted over of all the specification. In other words the Bar Chart shows the performance of the Influential measure developed based on Kibra-Lukman_M . The Bar representing DFT_KL_M is slightly higher than that of CVR_M showing that either DFT_KL_M or CVR_M could be used when dealing with medium sample sizes.

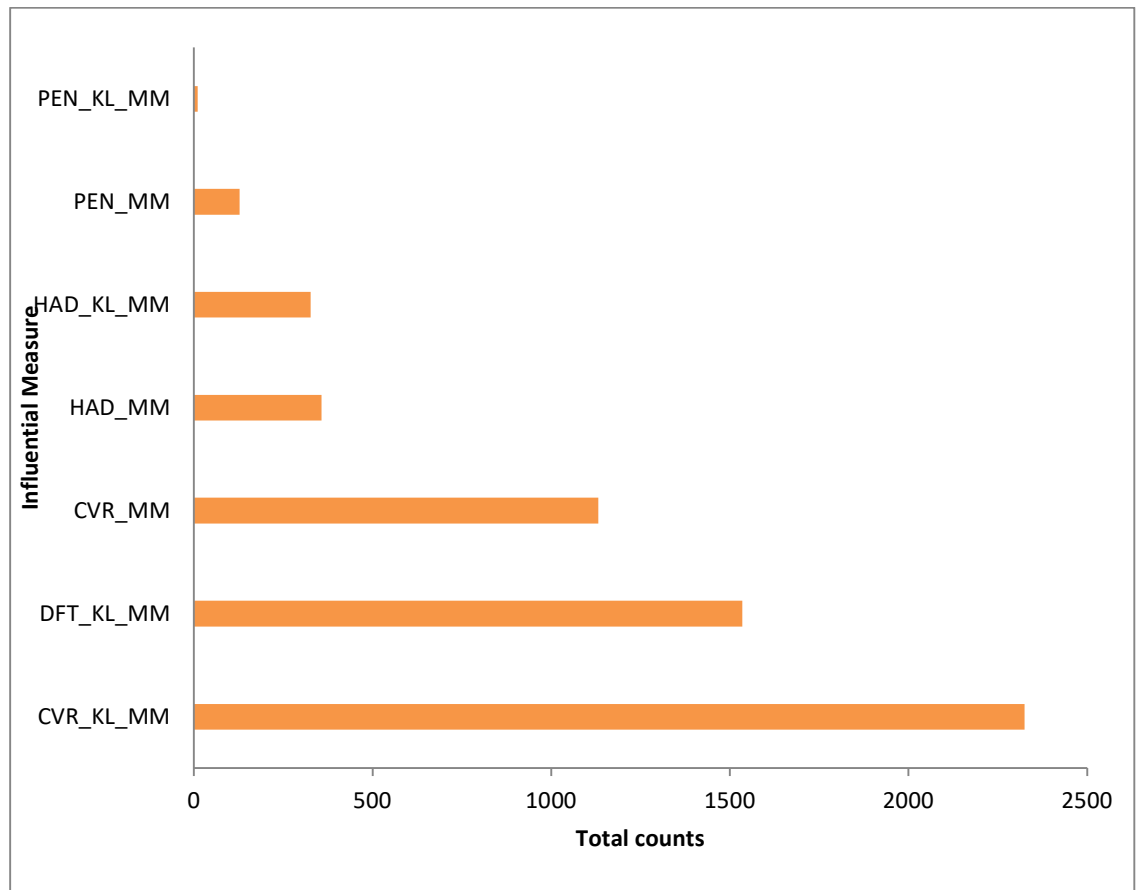


Figure 4.2b: The graphical representation of best influential measures proposed based on Robust-MM estimator that detected all influential points (100%) when counted over all specifications.

Fig 4.2b shows that CVR_KL_MM has the highest bar followed by DFT_KL_MM while PEN_KI_MM and PEN_MM have the smallest Bar when counted over of all the specification. The Bar representing DFT_KL_MM is slightly higher than that of CVR_MM showing that either DFT_KL_M or CVR_M could be used when dealing with medium sample sizes.

In other words the Bar Chart shows the performance of the Influential measure developed based on Kibra-Lukman_MM , at different sample sizes.

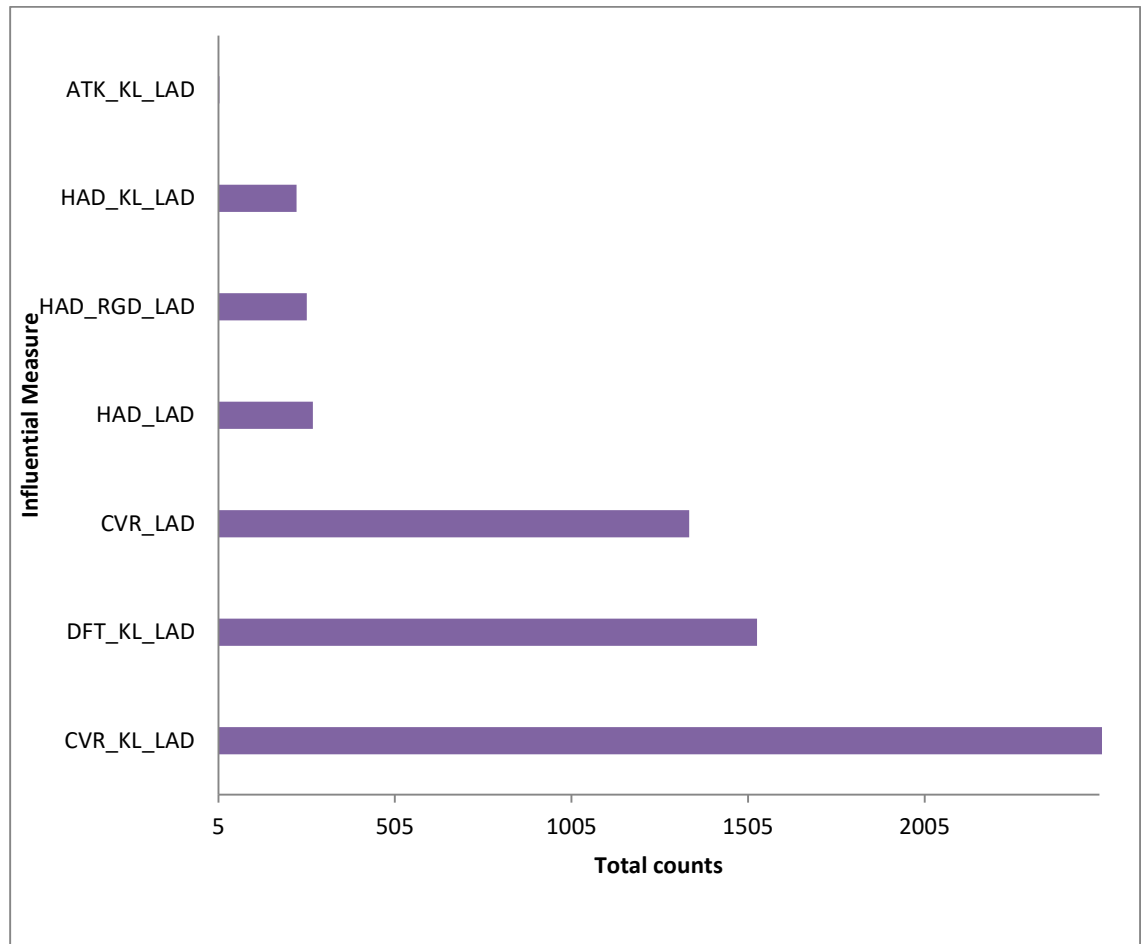


Figure 4.2c: The graphical representation of best influential measures proposed based on Robust-LAD estimator that detected all influential points (100%) when counted over all specifications.

Fig 4.2c shows that CVR_KL_LAD has the highest bar followed by DFT_KL_LAD while ATK_LAD has the smallest Bar when counted over of all the specification. The Fig shows that the Bars for HAD_LAD, HAD_KL_LAD and HAD_RGD_LAD have about the same height. Hence, for medium sample sizes and of the three diagnostics can be used. In other words the Bar Chart shows the performance of the Influential measure developed based on Kibra-Lukman_LAD at different sample sizes.

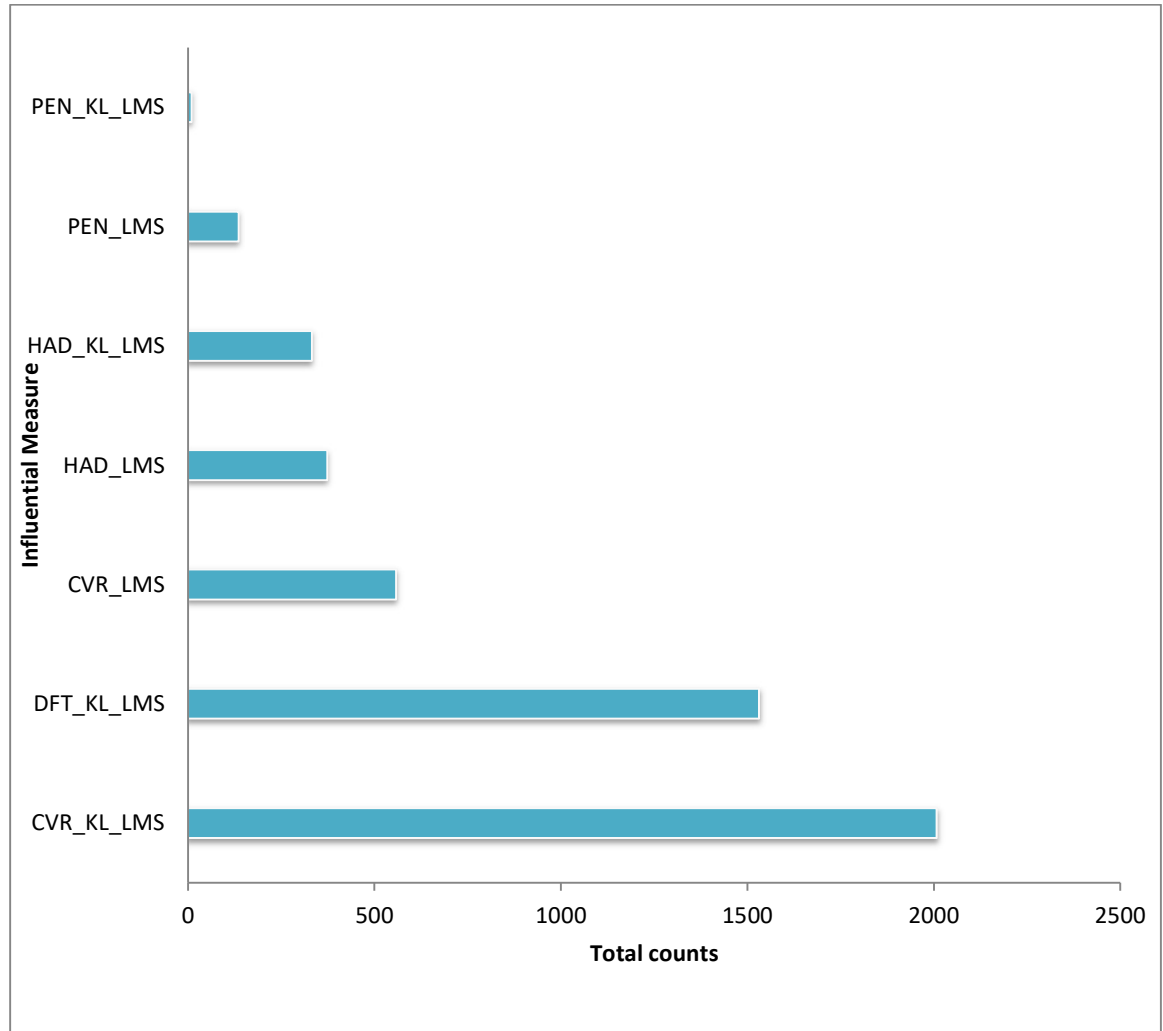


Figure 4.2d: The graphical representation of best influential measures proposed based on Robust-LMS estimator that detect influential point (100%), when counted over all specifications.

Fig 4.2d which is Bar Chart showing the performance of the Influential measure developed based on Kibra-Lukman _LMS, It indicates that CVR_KL_LMS has the highest bar followed by DFT_KL_LMS while PEN_KL_LMS has the smallest Bar when counted over of all the specification. Also, it shows that influential measures developed based Kibra-Lukman _LMS perform distinctly at different sample sizes.

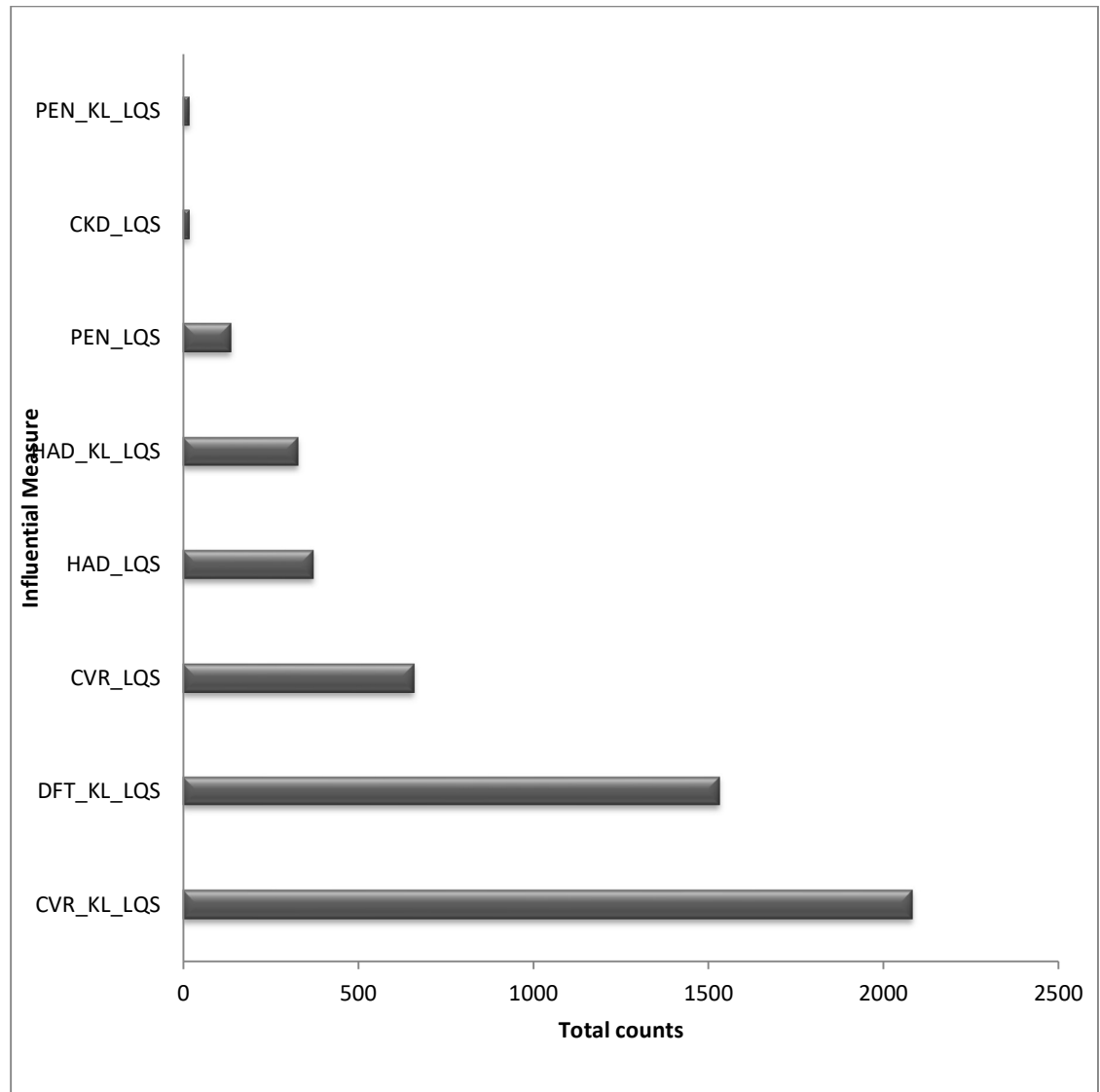


Figure 4.2e: The graphical representation of best influential measures proposed based on Robust-LQS estimator which detect influential points (100%), when counted over all specifications.

Fig 4.2e is a Bar Chart showing the performance of Influential measure developed based on Kibra-Lukman _LQS, indicates that CVR_KL_LQS has the highest bar followed by DFT_KL_LQS while PEN_KL_LQS has the smallest Bar when counted over of all the specification. Also, it shows that influential measures developed based Kibra-Lukman _LMS perform distinctly at different sample sizes with CKD-LQS and PEN_KL_LQS performing very poorly.

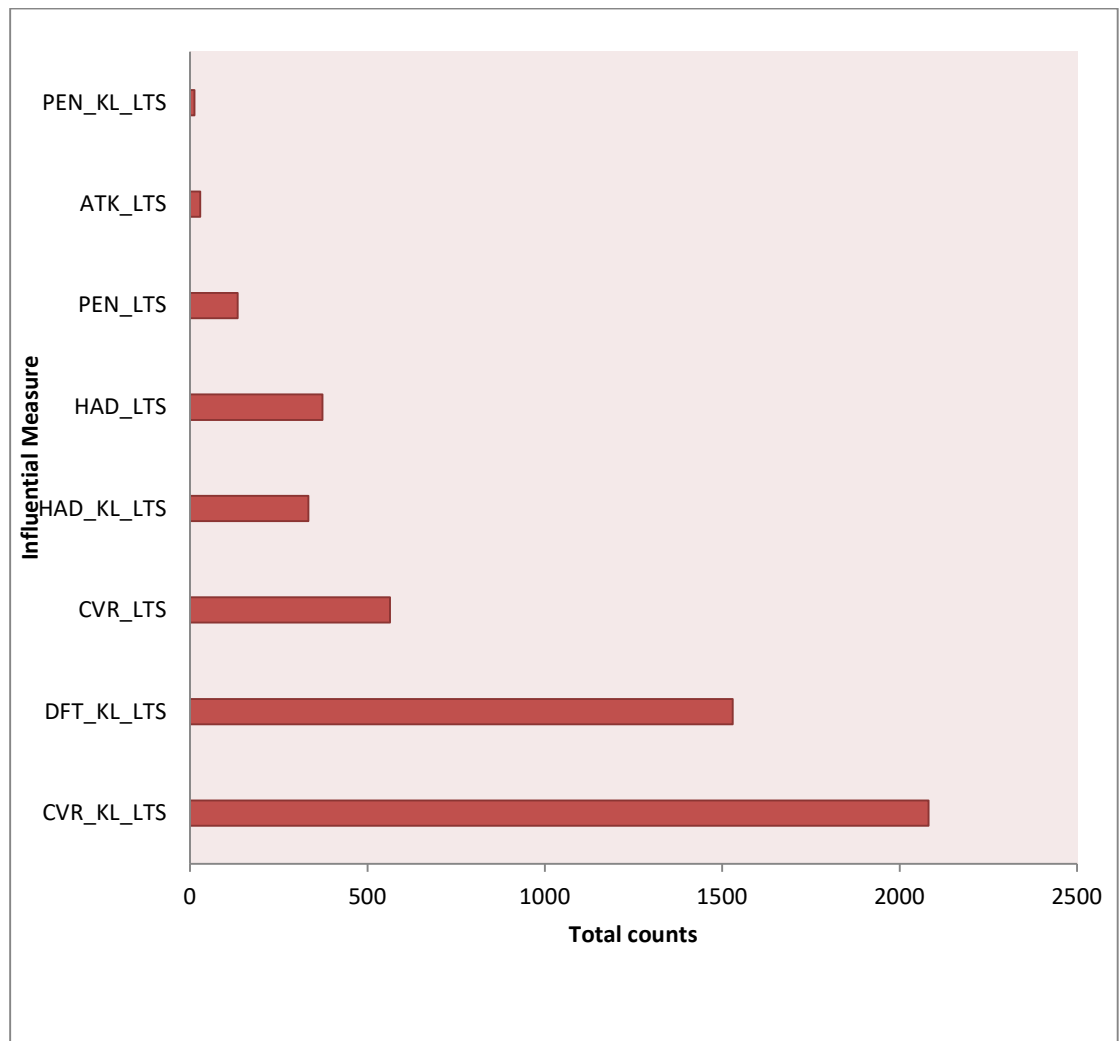


Figure 4.2f: The graphical representation of best influential measures proposed based on Robust-LTS estimator that detect influential points (100%), when counted over all specifications.

Fig 4.2f is the Bar Chart showing the performance of the Influential measure developed based on Kibra-Lukman _LTS. It indicates that CVR_KL_LTS has the highest bar followed by DFT_KL_LTS while PEN_KL_LTS has the smallest Bar when counted over of all the specification. This reveals that CVR_KL_LTS perform best across all scenario and PEN_KL_LTS perform worst in detecting observation that has great impart on a regression model among all the influential measures developed bases on Kibra-Lukman _LTS.

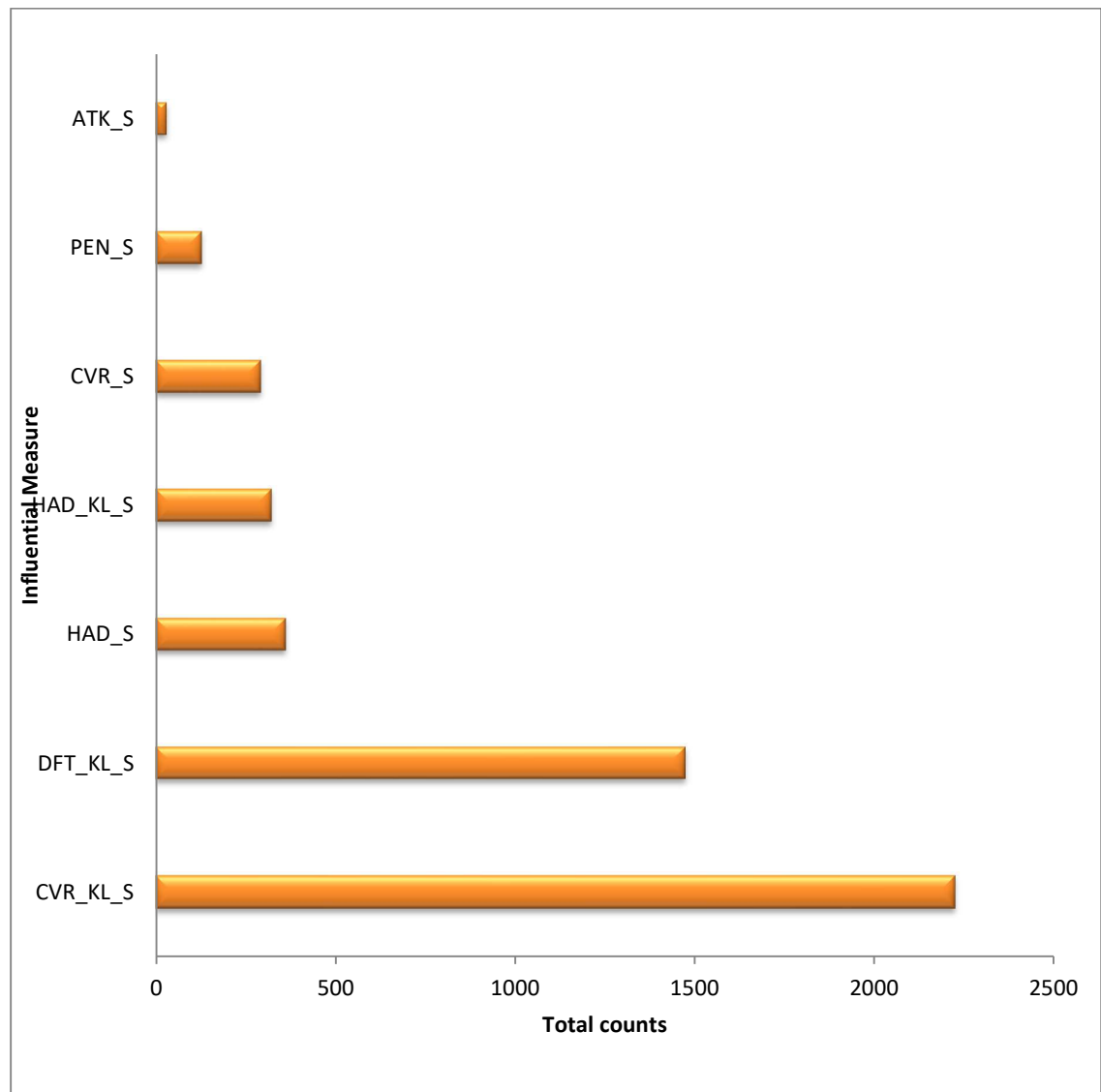


Figure 4.2g: The graphical representation of the best influential measure proposed based on Robust-S estimator which detect influential points (100%) when counted using all specifications.

Fig 4.2g is Bar Chart showing performance of Influential measure developed based on Kibra-Lukman _S, it indicates that CVR_KL_S has the highest bar followed by DFT_KL_S while PEN_KL_S has the smallest Bar when counted over of all the specification. This reveals that CVR_KL_LTS perform best across all scenario and ATK_S perform worst in detecting Influential point in dataset of which dual contaminants such as multicollinearity and outliers are dealt with simultaneously .

Table 4.3a: Summary of best influential measure at each sample size when counted over all the specifications for Robust-M estimator

Sample size (n)	IFM	Frequency
10	CVR_KL_M	314
20	CVR_KL_M	330
30	CVR_KL_M	330
40	CVR_KL_M	329
50	CVR_KL_M and CVR_M	330
100	CVR_KL_M and CVR_M	330
250	CVR_KL_M and CVR_M	330
500	CVR_M	330
Overall best	CVR_KL_M	2622

Source: Author's analysis, 2025

Table 4.3a shows CVR_KL_M had the highest frequency at each sample size when counted over all the specification and the highest total frequency of 2622 . Therefore, CVR_KL_M is adjudged the best performed influential measure developed based on Kibra-Lukman_M.

Table 4.3b: Summary of best influential measure at each sample size when counted over all the specifications for Robust-MM estimator

Sample size (n)	IFM	Frequency
10	PEN_MM	128
20	CVR_KL_MM	267
30	CVR_KL_MM	311
40	CVR_KL_MM	324
50	CVR_KL_MM	327
100	CVR_KL_MM	330
250	CVR_KL_MM and CVR_MM	330
500	CVR_KL_MM and CVR_MM	328
Overall best	CVR_KL_MM	2325

Source: Author's analysis, 2025

Table 4.3b shows that CVR_KL_MM had highest frequency at different sample size when counted over all the specification and the highest total frequency of 2325. While CVR_KL_MM and CVR_MM performed best at large sample sizes 250 and 500. Therefore, CVR_KL_MM is adjudged the best performed influential measure developed based on on Kibra-Lukman_MM.

Table 4.3c: Summary of best influential measure at each sample size when counted over all the specifications for Robust-LAD estimator

Sample size (n)	IFM	Frequency
10	CVR_KL_LAD	300
20	CVR_KL_LAD	326
30	CVR_KL_LAD	330
40	CVR_KL_LAD	330
50	CVR_KL_LAD and CVR_LAD	330
100	CVR_KL_LAD and CVR_LAD	330
250	CVR_KL_LAD and CVR_LAD	330
500	CVR_KL_LAD and CVR_LAD	327
Overall best	CVR_KL_LAD	2603

Source: Author's Compilation, 2025

Table 4.2c shows that CVR_KL_LAD had highest frequency at different sample size when counted over all the specification and the highest total frequency of 2603. When sample size is large CVR_KL_LAD and CVR_LAD. Therefore, CVR_KL_LAD is adjudged the best performed influential measure developed based on Kibra-Lukman_LAD.

Table 4.3d: Summary of best influential measure at each sample size when counted over all the specifications for Robust-LMS estimator

Sample size (n)	IFM	Frequency
10	PEN_LMS	135
20	CVR_KL_LMS	157
30	CVR_KL_LMS	236
40	CVR_KL_LMS	294
50	CVR_KL_LMS	313
100	CVR_KL_LMS	330
250	CVR_KL_LMS	329
	CVR_KL_LMS and CVR_LMS	
500		330
Overall best	CVR_KL_LMS	2007

Source: Author's Compilation, 2025

Table 4.3d shows that CVR_KL_LMS had highest frequency at different sample size when counted over all the specification and the highest total frequency of 2007. At sample size of 500 CVR_KL_LMS and CVR_LMS performed identically. Therefore, CVR_KL_LMS is adjudged the best performed influential measure and PEN_LMS performed worst among all the influential measures developed based on Kibra-Lukman_LMS.

Table 4.3e: Summary of best influential measure at each sample size when counted over all the specifications for Robust-LQS estimator

Sample size (n)	IFM	Frequency
10	PEN_LQS	134
20	CVR_KL_LQS	178
30	CVR_KL_LQS	245
40	CVR_KL_LQS	300
50	CVR_KL_LQS	313
100	CVR_KL_LQS	330
	CVR_KL_LQS and CVR_LQS	330
250		
	CVR_KL_LQS and CVR_LQS	329
500		
Overall best	CVR_KL_LQS	2081

Source: Author's Compilation, 2025

Table 4.3e shows that CVR_KL_LQS had the highest frequency at each sample size when counted over all the specification and the highest total frequency of 2081. While CVR_KL_LQS and CVR_LQS performed best at large sample sizes 250 and 500. Therefore, CVR_KL_LQS is adjudged the best performed and PEN_LQS the worst influential measure developed based on Kibra-Lukman_LQS .

Table 4.3f: Summary of best influential measure at each sample size when counted over all the specifications for Robust-LTS estimator

Sample size (n)	IFM	Frequency
	PEN_LTS	134
10		
20	CVR_KL_LTS	176
30	DFT_KL_LTS	251
40	CVR_KL_LTS	292
50	CVR_KL_LTS	316
100	CVR_KL_LTS	330
250	CVR_KL_LTS	330
	CVR_KL_LTS and CVR_LTS	330
500		
Overall best	CVR_KL_LTS	2025

Source: Author's Compilation, 2025

Table 4.3f shows that CVR_KL_LTS had the highest frequency at sample size 20 and above when counted over all the specification and the highest total frequency of 2025 . CVR_KL_LTS and CVR_LTS performed well at large sample sizes such as 250 -500. Therefore, CVR_KL_LTS is adjudged the best performed influential measure developed based on Kibra-Lukman_LTS.

Table 4.3g: Summary of best influential measure at each sample size when counted over all the specifications for Robust-S estimator

Sample size (n)	IFM	Frequency
10	HAD_S and PEN_S	126
20	CVR_KL_S	212
30	CVR_KL_S	280
40	CVR_KL_S	316
50	CVR_KL_S	327
100	CVR_KL_S	330
250	CVR_KL_S	330
500	CVR_KL_S	327
Overall best	CVR_KL_S	2224

Source: Author's Compilation, 2025

Table 4.3g shows that CVR_KL_S had the highest frequency at sample size 20 and above when counted over all the specification and the highest total frequency of 2224. CVR_KL_S performed well at all sample (low, medium and large) sizes. Therefore, CVR_KL_S is adjudged the best performed influential measure developed based on Kibra-Lukman_S.

Table 4.4: Overall best Influential Measures when counted over all the specifications and different sample sizes for each robust regression estimators

IFM	Sample size (n)								Total	Rank
	10	20	30	40	50	100	250	500		
CVR_KL_M	314	330	330	329	330	330	330	329	2622	1
CVR_KL_LAD	300	326	330	330	330	330	330	327	2603	2
CVR_KL_MM	108	267	311	324	327	330	330	328	2325	3
CVR_KL_S	102	212	280	316	327	330	330	327	2224	4
CVR_KL_LTS	57	176	251	292	316	330	330	330	2082	5
CVR_KL_LQS	56	178	245	300	313	330	330	329	2081	6
CVR_KL_LMS	18	157	236	294	313	330	329	330	2007	7

Source: Author's Compilation, (2025)

Table 4.4 shows CVR_KL_M had the highest total frequency of 2622 at all sample size when counted over all the specification. Therefore, CVR_KL_M ranked 1st (The best performed) followed by CVR_KL_LAD and CVR_KL_MM ranking 2nd and 3rd respectively. CVR_KL_LTS and CVR_KL_LQS perform almost identically to rank 5th and 6th respectively. Consequently, CVR_KL_M performed excellently well among influential measure developed based on Kibra-Lukman_ Robust Regression.

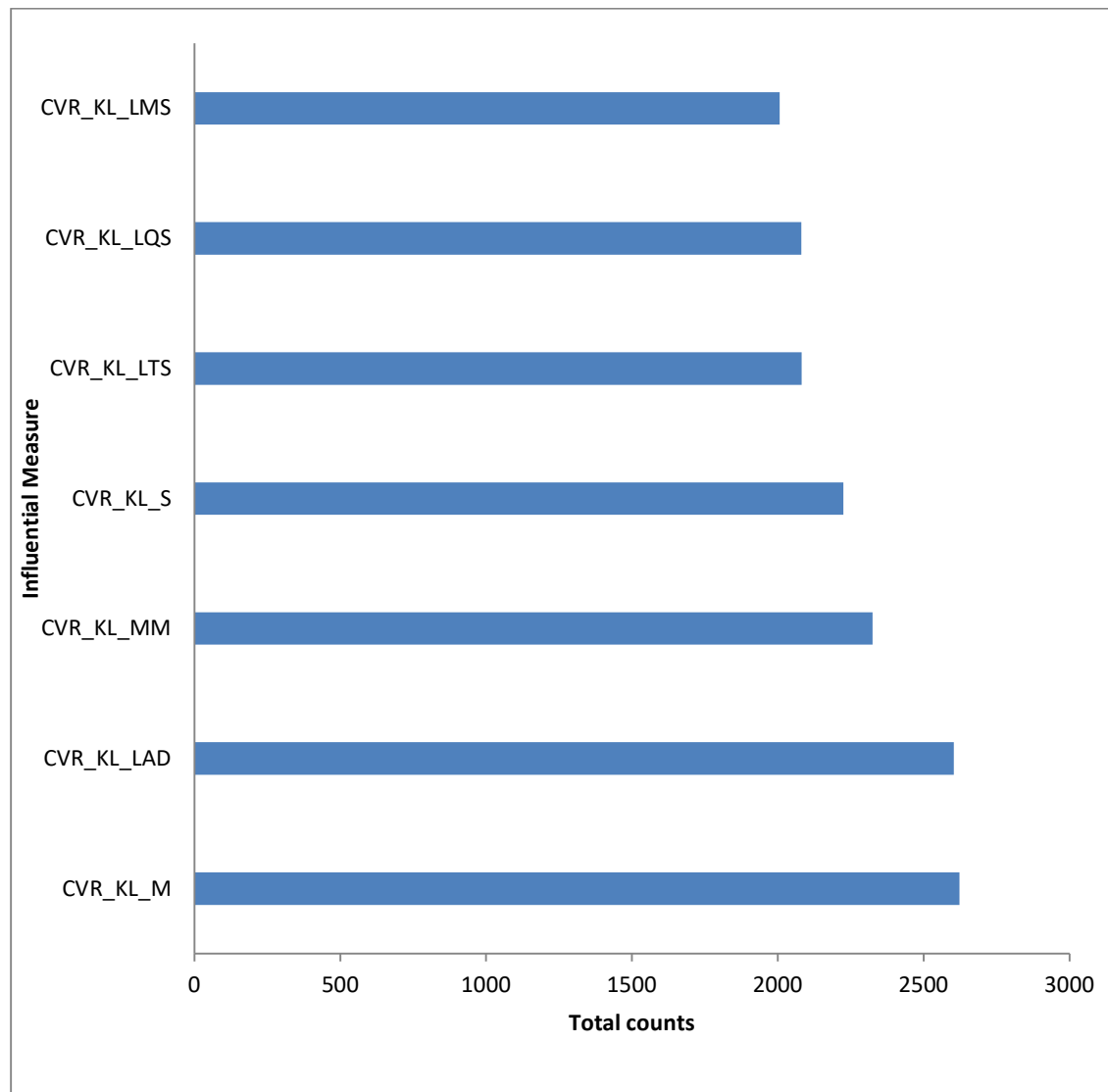


Figure 4.3: Bar chart of overall best Influential Measures when counted over all the specifications for all robust regressions.

Fig 4.3 buttress the fact that CVR_KL_M which has the highest Bar performed very well above all the other overall best influential Measures like CVR_KL_LAD and CVR_KL_MM. Therefore, CVR_KL_M is adjudged the best performing influential measure capable of detecting influential point when multicollinearity and outliers are mitigated.

Table 4.5: Best Influential Measures when counted over all the specifications at each sample sizes and across all robust regressions

Category	Sample size (n)	Influential Measures	Frequency
Small sample size	10	CVR_KL_M	314
	20	CVR_KL_M	330
	30	CVR_KL_M and CVR_KL_LAD	330
Moderate sample size	40	CVR_KL_LAD	330
	50	CVR_KL_M, CVR_M, CVR_KL_LAD and CVR_LAD	330
	100	CVR_KL_M, CVR_M, CVR_KL_MM, CVR_KL_LAD, CVR_LAD, CVR_KL_LMS, CVR_KL_LQS, CVR_KL_LTS and CVR_KL_S	330
High Sample size	250	CVR_KL_M, CVR_M, CVR_KL_MM, CVR_MM, CVR_KL_LAD, CVR_LAD, CVR_KL_LQS, CVR_LQS, CVR_KL_LTS and CVR_KL_S	330
	500	CVR_M, CVR_KL_LMS, CVR_LMS, CVR_KL_LTS, and CVR_LTS	330
Overall best influential measure		CVR_KL_M	2622

Source : Author's compilation, (2025)

Table 4.5 presents different influential measures developed in this study which performed well at different sample size when counted over all the specification. CVR_KL_M, CVR_M, CVR_KL_MM, CVR_MM, CVR_KL_LAD, CVR_LAD, CVR_KL_LQS, CVR_LQS, CVR_KL_LTS and CVR_KL_S performs well at large sample size of 250 and CVR_M, CVR_KL_LMS, CVR_LMS, CVR_KL_LTS, and CVR_LTS perform excellently when sample size increase to 500. In view of the foregoing, CVR_KL_M which had the highest frequency count (2622) was adjudged the overall best performed influential measure developed based on Kibra-Lukman_ Robust Regression.

4.2 Application to Real-life Dataset

Real-life dataset were used to illustrate the performance of influential measures estimated in this research, its results were compared with existing diagnostics and shown in Tables 4.3 and 4.4.

4.2.1 Application to Economic Dataset

The first dataset was an economic dataset that had been used by several authors such as Jahufer and Jianbao (2009), Jahufer (2013), Yasin and Murat (2016), Ullah *et al.* (2013), Kashif *et al.* (2019), Lukman and Ayinde (2018) and Longley (1967). The regression model was given as follows:

$$y_i = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 + \gamma_4 x_4 + \gamma_5 x_5 + \gamma_6 x_6 + e_i \quad (4.1)$$

Where

y = total derived employment;

x_1 = gross national product implicit price deflator;

x_2 = gross national product;

x_3 is unemployment;

x_4 is size of armed forces;

x_5 is non-institutional population 14 years of age and over;

x_6 is time.

Walker and Birch, (1988) reported that scaled Condition Number (CN) of the dataset is 43.275. Likewise, Liu (1993) claimed that the dataset exhibited severe multicollinearity because the values of Variance Inflation Factor (VIF), was estimated to be 128.29, 103.43, and 70.87. Similarly, Midi *et al.* (2010) supported the claim of Liu (1993) and spotted the presence of outliers in the dataset.

Table 4.6: Five most Influential observations detected based on some already existing Influential measures when compared with the newly proposed ones, using Economic dataset.

		Existing Method	
Author(s)	Year	Influential points Detected	Biased Method
Cook	1977	5, 16, 4, 10, 15	Cook's Distance based on OLS
Jahufer	2013	15, 4, 1, 6, 16	Cook's Distance and DFFITs with Liu
Ullah <i>et al.</i> ,	2013	16, 10, 4, 6, 1	Influential points with Liu estimator when d = 0.1, 0.5 and 0.9
		16, 10, 4, 6, 5	
		16, 5, 4, 10, 15	
Yasin and Murat	2016	16, 10, 6, 1 and 4	Influential points with Two-parameter Ridge
Lukman and Ayinde	2018	16, 10, 4, 5 and 15	DFFITs with Two-parameter Ridge-Liu
Kashif <i>et al.</i> ,	2018	3, 10, 11, 4 and 5	Pena's Statistic for the Liu regression for d = 0.1, 0.5, 0.9 and 1. 0
		10, 3, 4, 15 and 16	
		10, 4, 15, 5 and 16	
		15, 10, 5, 4 and 16	
Kashif <i>et al.</i> ,	2019	16, 15, 5, 4 and 10	Modified Pena's Statistic for the Modified Ridge Regression for K = 0.0002
Proposed Method			
CVR_KL_M	Proposed	2, 16, 1, 12 and 9	Covratio in Robust-M Kibria-Lukman
CVR_KL_LAD	Proposed	2, 16, 1, 12 and 9	Covratio in Robust-LAD Kibria-Lukman
CVR_KL_MM	Proposed	2, 16, 1, 12 and 9	Covratio in Robust-MM Kibria-Lukman
CVR_KL_S	Proposed	2, 16, 1, 12 and 9	Covratio in Robust-S Kibria-Lukman
CVR_KL_LTS	Proposed	2, 16, 1, 12 and 9	Covratio in Robust-LTS Kibria-Lukman
CVR_KL_LQS	Proposed	2, 16, 1, 12 and 9	Covratio in Robust-LQS Kibria-Lukman
CVR_KL_LMS	Proposed	2, 16, 1, 12 and 9	Covratio in Robust-LMS Kibria-Lukman

Source: Author's Compilation, (2025)

Table 4.6 indicate that the proposed Influential measures detected observations 16 and 1 as in some of the already existing ones compared, but not in the same order. Where, cases 2, 12, and 9 were not detected by the pre-existing influential measures; this merely indicates that more influential points were detected by the newly suggested diagnostic tools than the pre-existing ones. Hence, they perform better than the existing measures when detecting influence point in the presence of multicollinearity and outliers in a dataset.

4.2.2 Application to Manufacturing Dataset

The manufacturing data that Hussein and Abdalla (2012) adopted is what was used in this investigation. The regression model for the dataset is shown below:

$$y = \gamma_1 x_1 + \gamma_2 x_2 + \gamma_3 x_3 \quad (4.2)$$

where

y = product value in the manufacturing sector;

x_1 = value of imported intermediate;

x_2 = represents imported capital commodities;

x_3 = indicates value of imported raw materials.

Hussein and Abdalla (2012) stated that multicollinearity existed in the dataset because the values of Variance Inflation Factor (VIF) was estimated to be (128.29, 103.43, and 70.87). Likewise, Lukman, *et al.* (2015) supported this claim of Hussein and Abdalla (2012) and also spotted the presence of outliers in the dataset.

The value of parameter k used in this study was same as the one used by Yasin and Murat (2016) and Kashif *et al.* (2019), estimated to be 0.0012.

Table 4.7: Five Most influential points detected by other existing influential measures compared with the newly proposed ones using Hussein's dataset

Existing Method			
Author(s)	Year	Influential points in order	Method
Cook	1977	15, 16, 31, 30 and 20	Cooks D in OLS
Ayinde <i>et al.</i> ,	2015	30, 31, 16, 21 and 14	Cooks D in M-estimator
Ayinde <i>et al.</i> ,	2015	15, 30, 31, 14 and 12	Cooks D in MM-estimator
Ayinde <i>et al.</i> ,	2015	15, 16, 31, 30 and 20	DFFITs based on OLS
Ayinde <i>et al.</i> ,	2015	31, 15, 30, 14 and 12	DFFITs in MM-estimator
Proposed Method			
CVR_KL_M	Proposed	15, 20, 16, 12 and 4	Covratio in Robust-M Kibria-Lukman
CVR_KL_LAD	Proposed	15, 20, 16, 12 and 4	Covratio in Robust-LAD Kibria-Lukman
CVR_KL_MM	Proposed	15, 20, 16, 12 and 4	Covratio in Robust-MM Kibria-Lukman
CVR_KL_S	Proposed	15, 20, 16, 12 and 4	Covratio in Robust-S Kibria-Lukman
CVR_KL_LTS	Proposed	15, 20, 16, 12 and 4	Covratio in Robust-LTS Kibria-Lukman
CVR_KL_LQS	Proposed	15, 20, 16, 12 and 4	Covratio in Robust-LQS Kibria-Lukman
CVR_KL_LMS	Proposed	15, 20, 16, 12 and 4	Covratio in Robust-LMS Kibria-Lukman

Source: Author's compilation, 2025

Table 4.4 shows that all the developed Influential measures identified all the observations that the existing ones identified even more but at different order. While cases 4 was not part of the most influential points identified by already existing influential measures, this simply means that the new proposed diagnostic tools is more robust to influential points detection than some of the already existing ones when outliers and multicollinearity problems are being curtailed in regression analysis.

4.3 Discussion

The findings of this study gives important insights about the behaviour of Influential diagnostic measures when multicollinearity and outliers which occur simultaneously in dataset are mitigated. The primary objective of this study was to develop robust versions of the one-parameter new ridge-type estimator that can detect influential points more effectively than existing methods when multicollinearity and outliers are not curtailed. The results of both simulation studies and real-life dataset applications showed that the developed Robust Regression Kibria–Lukman (RKL) estimators performed very well in detecting influential points, thereby fulfilling the aim of the study.

This study successfully developed the robust versions of the KL estimator, thus objective one was meant. The combination of the KL estimator with robust regression techniques (M, MM, LTS, LMS, LAD, S and LQS), resulted in the development of six new estimators capable of simultaneously mitigating multicollinearity and resisting the negative effects of outliers. This outcome is in consonance with the literature, particularly the works of Bagheri and Midi (2009) and Ayinde *et al.*, (2015), who emphasized the necessity and benefits of employing robust estimation in contaminated datasets. The study which enhance estimation stability supports earlier arguments that biased estimators alone are insufficient when a dataset contain the dual anomalies of outliers and multicollinearity.

The second objective, which centred on deriving theoretical diagnostic measures based on the robust KL framework, was achieved. Generalized versions of Cook's Distance, DFFITS, COVRATIO, Atkinson, Hadi's and Pena's diagnostic measures were successfully derived for the RKL estimator. These successes align with existing research of scholars like Asar and Erisoglu (2016) and Emami and Emami (2016) who developed specialized diagnostic measures for ridge-type and Liu-type estimators. Therefore, this study

effectively extends the diagnostic framework by contributing a new family of robust ridge-type estimators, consequently, filling an identified gap in previous studies.

In addition, Approximate case deletion formula for Cook's D and DFFITS were obtained in line with objective three. This accomplishment is similar to the works of Jahufer (2013) and Emami and Emami (2017), who created deletion frameworks for Liu-type estimators. These newly developed tools significantly, strengthens the practical utility of influential measures in detecting influential points. It also substantially improves efficient computation of influential diagnostics without fitting models repeated.

As required by objective four, simulation study was used to determine the performance of the new measures. The result offered strong empirical evidence in support of the effectiveness of the developed measures. The performance of diagnostic tools in influential point detection followed predictable patterns supported by statistical theory such as the need to mitigate multicollinearity and outliers in a dataset especially when they occur simultaneously. The results shown by Tables 4.1- 4.2g and graphically illustrated by Figures 4.1 and 4.2a - 4.2g respectively reveals that as the magnitude of outliers increased, the proportion of detected influential points also increased, a result that aligns with the classical understanding of leverage and residual amplification discussed by Belsley et al (1980). Similarly, when sample size increased, percentage detection of thr newly developed diagnostic measures also increased. Conversely, as error term increased, the rate of detection of the measures, reduces. These patterns buttress the submission in prior studies that increase error variance affects model stability and efficiency.

Furthermore, the simulation study results shows the Covratio measure developed under the Robust KL_M framework (CVR_KL_M) outperformed the other measures in this study, as it consistently produced the highest frequency of 100% influential point detection

across all specifications used. Its effectiveness in detecting influential observations in the presence of this dual anomalies affirms the superiority of combining robustness with ridge-type regularization. Even, when OLS-based Covratio (CVR_OLS) performed relatively well without integrating robust regression, its performance reduced when multicollinearity and outliers were curtailed. This validates the assertion of Barnett and Lewis (1994) on vulnerability of OLS diagnostics in contaminated data. Although, the literature also identified CVR as a strong diagnostic tool, this study enhances this understanding by its evaluation outcomes that shows that CVR in a robust KL estimators significantly outperforms the traditional counterpart.

The final objective which focused on comparing the developed measures with existing ones, further demonstrates the effectiveness of the RKL-based measures, particularly CVR_KL_M in detecting influential observations perform much more than the existing diagnostics. This finding aligns with earlier empirical results from Emami and Emami (2016) and Kashif *et al.*, (2019), who established that ridge-type and Liu-type modifications alter the influence structure of data and may outperform classical diagnostics. Thus, the study agrees with literature that neither OLS estimators nor traditional diagnostics adequately address the joint effects of outliers and multicollinearity. Hence, the importance of developing influential measures using estimators such as RKL estimators to effectively handle the problem arising from the presence of multicollinearity and outliers simultaneously in a dataset, cannot be overemphasized. Notably, many real - life are infested by these two anomalies and if parameter estimates are to be stable and predictions reliable, then detecting influential point with estimators such as RKL estimator are not negotiable.

In view of the foregoing, the results of this research advances regression analysis by providing comprehensive evidence that robust one-parameter ridge-type estimators especially, the RKL variants are better equipped to manage the twin challenges of multicollinearity and outliers when they co-exist. It also confirms that integrating robustness and shrinkage within a unified framework produces a more reliable and resilient diagnostic tool.

CHAPTER FIVE

SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 SUMMARY

Detecting influential points in any dataset is a necessary component of assessment that should be done in regression analysis, to ascertain and validate integrity of data. Failure of which has undue influence on the regression model. It can impact the parameter estimation by increasing the variance and coefficient of parameters. By so doing the reliability of estimated coefficients and precision of hypothesis decisions are compromised (Weisberg, 2005).

Consequent upon the foregoing, the study developed influential diagnostic measures that could detect Influential points reliably when multicollinearity and outliers occur simultaneously in a dataset—a situation that challenges statistical modelling in real-life data. The study integrated one-parameter Kibria–Lukman estimator into some robust regression techniques namely: MM, LTS, LMS, LAD, S and LQS methods. These resulted in development of a six Robust Regression _kibria–Lukman (RKL) estimators, from which generalized versions of Cook’s Distance, DFFITs, COVRATIO, Atkinson’s statistic, Hadi’s measure and Pena’s statistic were derived. Afterwhich, approximate case-deletion formulas for Cook’s Distance and DFFITS under RKL estimators were obtained. Additionally, a detailed simulation study was carried out using different specification such as sample size, different levels of multicollinearity, vary degrees of error variances, different proportions and magnitudes of outliers in order to evaluate the performances of the newly developed diagnostic measures. These measures were also applied to Real-life

data and the outcomes were compared with already existing ones using 100% of influential points detection and the frequency as criteria.

Across all scenarios deployed in the simulation study, $Covratio_KL_M$ (CVR_KL_M) was found to consistently outperformed all other measures developed based on Robust_KL estimators. When applied to real-life economic and manufacturing datasets, the newly developed measures performed excellently well even more than existing measures supporting the fact that the developed influential measures on one-parameter new ridge-type estimators was statistically beneficial as it enhances assessing influence more efficiently without repeatedly re-estimating the model. Particularly, CVR_KL_M identified influential observations much more than the already existing measures.

5.2 CONCLUSION

The newly developed influential measures based on Robust Kibria-Lukman (RKL) estimator provide a reliable framework for identifying influential observations when the dual problems of multicollinearity and outlier are mitigated in a dataset. In other words, this research propose the use of the developed influential measures when datasets are heavily contaminated with extreme values and high correlation among predictor . This is premised upon the result obtained from simulation studies and application to real-life dataset which show the newly developed diagnostic measures are very efficient and effective in detecting influential observations even much more than some existing measures. While traditional diagnostics such as Cook's Distance and COVATIO based on OLS became less sensitive when the twin anomalies are present in datasets.

Additionally, the study projects the benefits of integrating robustness with ridge-type shrinkage to produce influential diagnostics that are considerably more accurate, reliable

and dependable among a wide spectrum of data variability. Similarly, the research puts CVR_KL_M as the most effective measure based on robust Kibria-Lukman in detecting influential points even under severe contamination and high multicollinearity. Therefore, the study affirms the importance of designing diagnostic tools that can accommodate multiple forms of data irregularities simultaneously rather than repetitive model fitting.

This study has made important contribution to regression modelling as well as diagnostic analysis. It introduces a unified hybrid approach that integrates robustness and biased shrinkage in a single estimator, to effectively handle the dual problem of outliers and multicollinearity that conventional estimators struggle with. Most significantly, it develops set of generalized influential diagnostic measures specifically tailored to the robust Kibria-Lukman estimator. Prior to this work, no diagnostic measure has been developed based on robust KL estimator to detect influence observations.

Also, approximate case-deletion formulas for Cook's Distance and DFFITS based on robust KL estimators were derived to further enhance the practical utility of the new diagnostic measures by reducing computational burden as well as simplifying implementation.

The comprehensive simulation study and real-life applications validate CVR_KL_M as the most effective diagnostic measure for identifying influential points at all sample sizes. Thus, provides practitioners and researchers with a highly reliable diagnostic tool for detecting influential observations in complex datasets.

Furthermore, the study empirically assessed the performance of the influential measures developed under Robust KL estimators to have outperformed those derived from ridge

regression, Liu-type estimators and other single-parameter approaches. Thereby placing the RKL framework as a strong foundation for future research in regression analysis.

5.3 RECOMMENDATIONS

Based on the specifications considered in the study and the results obtained from this research, the following are recommended:

- i. When datasets contain or are suspected to contain both outliers and multicollinearity, Analysts should adopt the Covratio_Kibria Lukman_M (CVR_KL_M) measure to spot influential points given its strong and consistent performance across all scenarios.
- ii. Sample size should be considered when choosing diagnostic measure. For smaller datasets, CVR_KL_M and Covratio_Kibria Lukman_LAD (CVR_KL_LAD) variations should be used. For moderate sample sizes, CVR_KL_LAD, CVR_KL_M, Covratio_M (CVR_M) and Covratio_LAD (CVR_LAD) should be used and for large datasets, Covratio_Kibria Lukman_M (CVR_KL_M), Covratio_Kibria Lukman_MM (CVR_KL_MM), Covratio_Kibria Lukman_LAD (CVR_KL_LAD), Covratio_Kibria Lukman_LSM (CVR_KL_LMS), Covratio_Kibria Lukman_LQS (CVR_KL_LQS), Covratio_Kibria Lukman_LTS (CVR_KL_LTS), Covratio_Kibria Lukman_S (CVR_KL_S), (CVR_M), Covratio_MM (CVR_MM), (CVR_LAD), Covratio_LQS (CVR_LQS), Covratio_LMS (CVR_LMS) and Covratio_LTS (CVR_LTS) can be used interchangeably depending on the estimation approach.
- iii. At any sample size, COVRATIO based on robust Kibria-Lukman_M estimator (CVR_KL_M) can be used, as it shows high accuracy and reliability.
- iv. When datasets exhibit signs of contamination and high correlation among predictors, researchers should proceed with caution. It should be noted that traditional OLS

diagnostic tools tend to underestimate the influence of extreme values especially, when multicollinearity and outliers coexist.

- v. Given effectiveness of the newly developed influential measures as shown by results obtained in the study, these diagnostic tools should be incorporated into analytical practices across of fields of study for which real-life data are collected such as economics, engineering, environmental science, health research and manufacturing.

REFERENCES

- Abdul-Wahab, S. A., Bakheit, C. S., & Al-Alawi, S. M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environmental Modelling & Software*, 20(10), 1263–1271.
- Abidoeye, A. O., Ajayi, I. M., Lukman, A. F., & Ogunjobi, J. O. (2022). Unbiased modified two-parameter estimator for the linear regression model. *Journal of Scientific Research*, 14(3), 785–795.
- Abuzaid, A., Mohamad, I., Hussin, A. G. & Rambli, A. (2011). COVRATIO statistic for simple circular regression model. *Chiang Mai Journal of Science*. 38(3), 321–330.
- Adeboye, N. O., Fagoyinbo, I. S., & Olatayo, T. O. (2014). Estimation of the effect of multicollinearity on the standard error for regression coefficients. *IOSR Journal of Mathematics*, 10(4, Version 1), 16–20. <https://doi.org/10.9790/5728-10411620>
- Adejumo, T. J., Amin, M., Ayinde, K., & Adewuyi, E. T. (2026). Influence diagnostics in the linear model with Kibria–Lukman estimator. *Communications in Statistics—Simulation and Computation*, 55(8), 3179–3196. <https://doi.org/10.1080/03610918.2025.2484628>
- Ahmad, S., & Aslam, M. (2022). Another proposal about the new two-parameter estimator for linear regression model with correlated regressors. *Communications in Statistics—Simulation and Computation*, 51(6), 3054–3072. <https://doi.org/10.1080/03610918.2019.1705975>
- Ahmed, M. G. & Maha, E. Q. (2016). Regression Estimation in the presence of outliers: A comparative study. *International Journal of Probability & Statistics*, 5(3).65-72. <https://doi.org/10.5923/j.ijps.20160503.01>
- Alabi, O. O., Ayinde, K. & Olatayo, T.O. (2008). Effect of Multicollinearity on Power Rates of the Ordinary Least Squares Estimators. *Journal of Mathematics & Statistics*, 4(2), 75- 80.
- Alabi, O.O., Olatayo, T.O. & Afolabi, F. R. (2014). Empirical Determination of the Tolerable sample size for OLS Estimator in the presence of Multicollinearity. *Applied Mathematics*, 5, 1870- 1877. <https://doi.org/10.4236/am.2014.513180>
- Alguraibawi, M., Midi, H., & Imon, A. H. M. R. (2015). A new robust diagnostic plot for classifying good and bad high leverage points in a multiple linear regression model. *Mathematical Problems in Engineering*, Article 279472. <https://doi.org/10.1155/2015/279472>
- Alkasadi, N. A., Ibrahim, S., Abuzaid, A. H. M., Yusoff, M. I., Hamid, H., Leow, W. Z., & Abd Razak, A. (2019). Outlier detection in multiple circular regression model using

- DFFITc Statistic. *Sains Malaysiana*, 48(7), 1557–1563.
<https://doi.org/10.17576/jsm-2019-4807-25>
- Anscombe, F. J. (1960). Rejection of outliers. *Technometrics*, 2(2), 123–147.
<https://doi.org/10.1080/00401706.1960.10489888>
- Arumairajan, S. & Kayathiri, S. (2002). A New Stochastic Restricted Two-Parameter Estimator in Multiple Linear Regression Model. *Vavuniya Journal of Science*.1, 38–47
- Asar, Y., & Erişoğlu, M. (2016). Influence diagnostics in two-parameter ridge regression. *Journal of Data Science*, 14(1), 33–52.
[https://doi.org/10.6339/JDS.201601_14\(1\).0003](https://doi.org/10.6339/JDS.201601_14(1).0003)
- Atkinson, A. C. (1985). Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis (Oxford Statistical Science Series, 1). *Oxford University Press*.
- Atkinson, A. C., & Cheng, T. C. (2000). On linear robust regression with incomplete data. *Computational Statistics & Data Analysis*, 33(4), 361–380.
[https://doi.org/10.1016/S0167-9473\(99\)00061-4](https://doi.org/10.1016/S0167-9473(99)00061-4)
- Ayinde, K., Lukman, A. F., & Arowolo, O. T. (2015). Robust regression diagnostics of influential observations in linear regression model. *Open Journal of Statistics*, 5, 273–283. <https://doi.org/10.4236/ojs.2015.54029>
- Ayinde, K., Lukman, A. F., Alabi, O. O., & Bello, H. A. (2020). A new approach of principal component regression estimator with applications to collinear data. *International Journal of Engineering Research and Technology*, 13(7), 1616–1622.
<https://doi.org/10.37624/IJERT/13.7.2020.1616-1622>
- Bagheri, A., & Midi, H. (2009). Robust estimations as a remedy for multicollinearity caused by multiple high leverage points. *Journal of Mathematics and Statistics*, 5(4), 311–321. <https://doi.org/10.3844/jmssp.2009.311.321>
- Bagheri, A., Midi, H., & Imon, A. H. M. R. (2010). The effect of collinearity–influential observations on collinear data sets: A Monte Carlo simulation study. *Journal of Applied Sciences*, 10(18), 2086–2093.
- Barnett, V. & Lewis, T. (1994). Outliers in Statistical Data (3rd Edition), *John Wiley & Sons*.
- Belsley, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics: Identifying influential data and sources of collinearity. *John Wiley & Sons*.
<https://doi.org/10.1002/0471725153>
- Birkes, D. & Dodge, Y. D. (1993). Alternative Methods of Regression, *John Wiley & Sons*, New York.

- Boscovich, R.J. (1757). De Litteraria Expeditione per Pontificiam ditionem, et Synopsis amplioris Operis. Bononiensi Scientiarum et Artium. *Instituto Atque Academia Commentarii*, 4, 353–396.
- Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–243. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>
- Çamdevýren, H., Demýr, N., Kanik, A., & Keskýn, S. (2005). Use of principal component scores in multiple linear regression models for prediction of chlorophyll-a in reservoirs. *Ecological Modelling*, 181(4), 581–589. <https://doi.org/10.1016/j.ecolmodel.2004.06.043>
- Chang, X., & Yang, H. (2012). Combining two-parameter and principal component regression estimators. *Statistical Papers*, 53(3), 549–562. <https://doi.org/10.1007/s00362-011-0364-7>
- Chatterjee, S., & Hadi, A. S. (1988). Sensitivity analysis in linear regression. *John Wiley & Sons*.
- Cont, R. (2001). Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2), 223–236. <https://doi.org/10.1080/713665670>
- Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19(1), 15–18.
- Cook, R. D., & Weisberg, S. (1982). Residuals & Influence in Regression. *Chapman & Hall*.
- Dawoud, I., & Kibria, B. M. G. (2020). A new biased estimator to combat the multicollinearity of the Gaussian linear regression model. *Stats*, 3(4), 526–541. <https://doi.org/10.3390/stats3040033>
- Dixon, W. J. (1960). Simplified Estimation from Censored Normal Samples. *The Annals of Mathematical Statistics*. 31(2), 385–391.
- Donugade, M. D. (2015). Ridge regression & its applications. *International Journal of Statistics & Applications*, 5(2), 45–50.
- Dovoedo, Y. H. (2011). Contributions to outlier detection methods: Some theory and applications [Doctoral dissertation, The University of Alabama]. The University of Alabama Libraries. <https://ir.ua.edu/items/a78a814a-37dd-48ec-8e38-fd588a593a59>
- Draper, N. R., & Smith, H. (1998). Applied regression analysis (3rd ed.). *John Wiley & Sons*.

- El-Fallah, M. & El-Salam, A. (2013). Alternative Outlier Detection Procedures in Linear Regression Analysis: A Comparative Study. *International Journal of Mathematics & Statistics Studies*, 2(1), 25-33.
- Emami, H., & Emami, M. (2016). Influence diagnostics in modified Liu-type estimator. *Calcutta Statistical Association Bulletin*, 68(1-2), 82-91. <https://doi.org/10.1177/0008068316668426>
- Emami, H., & Emami, M. (2016). New influence diagnostics in ridge regression. *Journal of Applied Statistics*, 43(3), 476-489. <https://doi.org/10.1080/02664763.2015.1070804>
- Fitrianto, A., & Xin, S. H. (2022). Comparisons between robust regression approaches in the presence of outliers and high leverage points. *Barekeng: Jurnal Ilmu Matematika dan Terapan*, 16(1), 243-252. <https://doi.org/10.30598/barekengvol16iss1pp243-252>
- Fox, J. (2016). Applied regression analysis & generalized linear models (3rd ed.). Sage Publications.
- Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, 11(1), 1-21. <https://doi.org/10.1080/00401706.1969.10490657>
- Gujarati, D. N. (1995). Basic econometrics (3rd ed.). McGraw-Hill.
- Gujarati, D. N. (2003). Basic Econometrics, New Delhi; Tata, McGraw-Hill. New York.
- Hadi, A. S. (1992). A new measure of overall potential influence in linear regression. *Computational Statistics & Data Analysis*, 14(1), 1-27. [https://doi.org/10.1016/0167-9473\(92\)90078-T](https://doi.org/10.1016/0167-9473(92)90078-T)
- Hoerl, A. E. & Kennard, R. W. & Baldwin, K. F. (1975). Ridge Regression. Some simulation. *Communications in Statistics*, 4(2), 105-123.
- Hoerl, A. E. & Kennard, R. W. (1970). Ridge Regression. Biased Estimation for Non-orthogonal Problems. *Technometrics*, 12(1), 55-67.
- Huber, P. J. (1964). Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35, 73-101.
- Huber, P. J. (1981). Robust statistics. John Wiley & Sons.
- Hussein, E. (2021). Leverage & Influential observations on the Liu-type estimator in the linear regression model with severe multicollinearity. *Heliyon*. <https://doi.org/10.1016/j.heliyon.2021.e07792>.
- Hussein, Y. A. & Abdalla, A. A. (2012). Generalized Two stages Ridge Regression Estimator for Multicollinearity & Autocorrelated errors. *Canadian Journal on Science & Engineering Mathematics*, 3(3), 79-85.

- Idowu, J. I., Fashoranbaku, A. S., & Ayinde, K. (2023). A two-parameter estimator for correlated regressors in Gamma regression model. *Science World Journal*, 18(4). <https://doi.org/10.4314/swj.v18i4.8>
- Idowu, J. I., Owolabi, A. T., Oladapo, O. J., Ayinde, K., Oshuporu, O. A., & Alao, A. N. (2023). Mitigating multicollinearity in linear regression model with two-parameter Kibria–Lukman estimators. *WSEAS Transactions on Systems and Control*, 18, 612–635. <https://doi.org/10.37394/23203.2023.18.63>
- Jahufer, A. (2013). Detecting global influential observations in Liu regression model. *Open Journal of Statistics*, 3(1), 5–11. <https://doi.org/10.4236/ojs.2013.31002>
- Jahufer, A., & Chen, J. (2009). Assessing global influential observations in modified ridge regression. *Statistics & Probability Letters*, 79(4), 513–518. <https://doi.org/10.1016/j.spl.2008.09.019>
- Jegede, S. L., Lukman, A. F., Ayinde, K. & Odeniyi, K. A. (2022). Jackknife Kibria Lukman M-Estimator, Simulation & Application. *Journal of the Nigerian Society of Physical Sciences*, 4, 251-264. <https://doi.org/10.46481/jnsps.2022.664>
- Kaciranlar, S. & Sakallioglu. (2001). Combining the Liu estimator & the principal component regression estimator. *Communication Statistics Theory and Methods* 30, 2699-2705.
- Kashif, M., Amanullah, M., & Aslam, M. (2018). Pena’s statistic for the Liu regression. *Journal of Statistical Computation and Simulation*, 88(13), 2473–2488. <https://doi.org/10.1080/00949655.2018.1468444>
- Kashif, M., Ullah, M. A., & Aslam, M. (2019). Influential diagnostics with Pena’s statistic for the modified ridge regression. *Communications in Statistics—Simulation and Computation*. <https://doi.org/10.1080/03610918.2019.1634204>
- Kibria, B. M. G., & Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model: Simulations and applications. *Scientifica*, Article 9758378. <https://doi.org/10.1155/2020/9758378>
- Koenker, R., & Bassett, G. (1984). Robust tests for heteroscedasticity based on regression quantiles. *Econometrica*, 52(5), 1343–1350.
- Kutner, M. H., Nachtsheim, C. J., & Neter, J. (2004). Applied linear regression models (4th ed.). *McGraw-Hill*.
- Leamer, E. E. (1977). Valley regression: Biased estimation for orthogonal problems (*UCLA Economics Working Paper No. 102*). University of California, Los Angeles.
- Leiva, V., Barros, M., Paula, G. A., & Galea, M. (2007). Influence diagnostics in log-Birnbaum–Saunders regression models with censored data. *Computational*

Statistics & Data Analysis, 51(12), 5694–5707.
<https://doi.org/10.1016/j.csda.2006.09.020>

- Li, J. & Valliant, R. (2011). Detecting group of influential observations in linear regression using Survey data—Adapting the forward search method. *Pakistan Journal of Statistics*. 27(4), 507–528.
- Liu, K. (1993). A new class of biased estimate in linear regression. *Journal of Communications in Statistics. Theory & Methods*. 22,2, 393–402.
<https://doi.org/10.1080/03610929308831027>
- Liu, K., Zhang, Y., & Wang, L. (2022). A new ridge-type estimator for handling multicollinearity in linear regression models. *Communications in Statistics—Simulation & Computation*, 51(4), 1892–1908.
- Longley, J. W. (1967). An appraisal of least squares programs for electronic computer from the point of view of the use. *Journal of American Statistical Association*. 62, 819–841.
- Lukman, A. F. (2017). Classification Based Ridge Parameter Estimation Techniques for Solving the Multicollinearity Problem. Unpublished PhD thesis. Ladoke Akintola University of Technology, Ogbomoso.
- Lukman, A. F., & Ayinde, K. (2018). Detecting influential observations in two-parameter Liu–Ridge estimator. *Journal of Data Science*, 16(2), 207–218.
[https://doi.org/10.6339/JDS.201804_16\(2\).0001](https://doi.org/10.6339/JDS.201804_16(2).0001)
- Lukman, A. F., Ayinde, K., Kibria, B. M. G., & Jegede, S. L. (2020). Two-parameter modified ridge-type M-estimator for linear regression model. *The Scientific World Journal*, 2020, Article 3192852. <https://doi.org/10.1155/2020/3192852>
- Lukman, A. F., Ayinde, K., Oluyemi, O. A., Akanbi, O. B., & Onate, C. A. (2018). Classification-based ridge estimation techniques of Alkhamisi methods. *Journal of Probability and Statistical Science*, 16(2), 165–181.
- Lukman, A. F., Osowole, O. I., & Ayinde, K. (2015). Two stage robust ridge method in a linear regression model. *Journal of Modern Applied Statistical Methods*, 14(2), 53–67. <https://doi.org/10.22237/jmasm/1446350820>
- Majeed, A., Ahmad, S., Aslam, M., & Kashif, M. (2022). A robust Kibria–Lukman estimator for linear regression model to combat multicollinearity and outliers. *Concurrency and Computation: Practice and Experience*.
<https://doi.org/10.1002/cpe.7533>
- Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). Robust statistics: Theory & methods. *John Wiley & Sons*.

- McCallum, B. T. (1970). Artificial Orthogonalization in Regression Analysis. *Review of Economic Statistics*, 52, 110-113.
- Meloun, M., & Militký, J. (2001). Detection of single influential points in OLS regression model building. *Analytica Chimica Acta*, 439(2), 169–191. [https://doi.org/10.1016/S0003-2670\(01\)01040-6](https://doi.org/10.1016/S0003-2670(01)01040-6)
- Midi, H. & Zahari, M. (2007), A simulation study on Ridge Regression Estimators in the presence of outliers & multicollinearity. *Journal of Teknologi*, 47 (1). 59–74.
- Molnar, C. (2023). Interpretable machine learning (2nd ed.). *Leanpub*.
- Montgomery, D. C., & Peck, E. A. (1992). Introduction to Linear Regression Analysis. New York, *John Wiley & Sons*
- Montgomery, D. C., Peck, E. A., & Vinning, G. G. (2021). Introduction to Linear Regression Analysis (6th ed.). *John Wiley & Sons*.
- Oladapo, O. J., Idowu, J. I., Owolabi, A. T., Ayinde, K., & Adejumo, T. J. (2024). A new ridge-type estimator in the logistic regression model with correlated regressors. *WSEAS Transactions on Business and Economics*, 21, 2528–2541. <https://doi.org/10.37394/23207.2024.21.208>
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6), 1–8
- Osborne, J. W., & Waters, E. (2004). Identifying & addressing outliers. *Journal of Experimental Education*, 73(1), 101–118.
- Owolabi, A. T., Ayinde, K. & Alabi, O. O. (2022). A new ridge-type estimator for the linear regression model with correlated regressors. *Concurrency and Computation: Practice and Experience*. <https://doi.org/10.1002/cpe.6933>
- Oyeleke K. T, Olatayo, T. O & Efuwape B. T. (2024). Handling Multicollinearity and Outliers: A Comparative Study of Some One and Two–Parameter Estimators Using Real-Life Data. *International Journal of Development Mathematics* 1(4) 177–190.
- Pena, D. (2005). New Statistic for influence in linear regression. *Technometrics*, 47(1), 1-12. <https://doi.org/10.1198/004017004000000662>
- Peng, L. Y., Midi, H., Rana, S., & Fitrianto, A. (2016). Identification of multiple outliers in a generalized linear model with continuous variables. *Mathematical Problems in Engineering*, 2016, Article 5840523. <https://doi.org/10.1155/2016/5840523>
- Rahman, S. M. A. K., Sathik, M. M., & Kannan, K. S. (2012). Multiple linear regression models in outlier detection. *International Journal of Research in Computer Science*, 2(2), 23–28.

- Rajab, J. M., MatJafri, M. & Lim, H. (2013). Combining multiple regression & principal component analysis for accurate predictions for column ozone in Peninsular Malaysia, *Atmospheric Environment*, 71, 36–43.
- Rao, C. R. (1973). Linear Statistical Inference & its Applications (2nd ed.), *John Wiley & Sons*.
- Rosner, B. (1983). Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2), 165-172.
- Rousseeuw, P. J. & Yohai (1984). Robust regression by means of S-estimator. In W. H. J. Franke & D. Martin (Eds.); Robust & Nonlinear Time Series Analysis, 256 – 272. *Springer- verlag*.
- Rousseeuw, P. J. (1984). Least Median of Squares Regression. *Journal of the American Statistical Association*, 79, 871-880.
- Rousseeuw, P. J., & Leroy, A. M. (1987). Robust regression and outlier detection. *John Wiley & Sons*. <https://doi.org/10.1002/0471725382>
- Rousseeuw, P. J., & Van Driessen, K. (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery*, 12(1), 29–45. <https://doi.org/10.1007/s10618-005-0024-4>
- Rousseeuw, P. J., & Yohai, V. J. (1984). Robust regression by means of S-estimators. In J. Franke, W. Härdle, & D. Martin (Eds.), Robust and nonlinear time series analysis (Lecture Notes in Statistics, Vol. 26, pp. 256–272). Springer. https://doi.org/10.1007/978-1-4615-7821-5_15
- Sakallıoğlu, S., & Kaçiranlar, S. (2008). A new biased estimator based on ridge estimation. *Statistical Papers*, 49, 669–689.
- Seber, G. A. F., & Lee, A. J. (2021). Linear Regression Analysis (2nd ed.). Wiley.
- Sousa, S. I. V., Martins, F. G., Alvim-Ferraz, M. C. M., Pereira, M. C. (2007). Multiple linear regression & artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modeling and Software*, 22, 97–103.
- Susanti, Y., Pratiwi, H., Sulistijowati H., S., & Liana, T. (2014). M estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics*, 91(3), 349–360. <https://doi.org/10.12732/ijpam.v91i3.7>
- Taleb, N. N. (2007). The black swan: The impact of the highly improbable. Random House.
- Tibshirani, R. (1996). Regression Shrinkage & Selection via the Lasso. *Journal of the Royal Statistical Society, Series B (Methodological)*, 58(1), 267-288.

- Tietjen, G. L., & Moore, R. H. (1972). Some Grubbs-type statistics for the detection of several outliers. *Technometrics*, 14(3), 583–597. <https://doi.org/10.1080/00401706.1972.10488948>
- Ullah, M. A., Pasha, G. R. & Aslam, M. (2013). Assessing Influence on the Liu Estimators in Linear Regression Models. *Communications in Statistics–Theory & Methods*, 42(17), 3100–3116.
- Uzuke, C. A., & Ezeilo, I. C. (2021). On identifying influential observations in the presence of multicollinearity. *Open Journal of Statistics*, 11, 290–302. <https://doi.org/10.4236/ojs.2021.112016>
- Velleman, P. F. & Welsch, R. E. (1981). Efficient Computing of Regression Diagnostics. *American Statistician*. 35(4), 234-242.
- Walfish, S. (2006, November 2). A review of statistical outlier methods. *Pharmaceutical Technology*, 30(11). <https://www.pharmtech.com/view/review-statistical-outlier-methods>
- Walker, E. & Birch, J. B. (1988). Influence Measures in Ridge Regression. *Technometrics*, 30(2), 221–227.
- Weisberg, S. (2005). Applied linear regression (3rd ed.). John Wiley & Sons.
- Welsch, R. E. & Kuh, E. (1977). Linear regression diagnostics. (Technical Report 923-77). Sloan school of management, Massachusetts Institute of Technology.
- Welsch, R. E. (1982). Influence function & regression diagnostics. Modern Data Analysis. New York. Academic Press.
- Yohai, V.J. (1987). High Breakdown-point & High Efficiency Robust Estimates for Regression. *The Annals of Statistics*. 15(20), 642-656.
- Yuliana, & Susanti, Y. (2008). Estimasi M dan sifat-sifatnya pada regresi linear robust. *Jurnal Math-Info*, 1(11), 8–16.
- Zhao, Y., Nasrullah, Z., & Li, Z. (2019). PyOD: A Python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, 20(96), 1–7. <https://www.jmlr.org/papers/v20/19-011.html>
- Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis & Data Mining*, 5(5), 363–387

APPENDICES

<https://1drv.ms/w/c/9FD5FA8A4AAFE3C0/IQDYTjplcLucTqunVRLKvfUuAVX8G3M2uVHxoHtvKebRwf0?e=iqxdaL>