ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 6 No. 1 (2026) 374-384
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Comparing Multiple Linear and Robust Regression Models in Assessing Economic Indicators' Impact on GDP

Tan Shun Hong[1], Azme Khamis[1]*

[1] *Department of Mathematics and Statistics, Faculty of Applied Science and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, MALAYSIA*

*Corresponding Author: azme@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2026.06.01.044

Article Info

Received: 4 January 2026
Accepted: 20 January 2026
Available online: 9 July 2026

Keywords

Gross Domestic Product, Economic Indicators, Multiple Linear Regression, Robust Multiple Linear Regression, MM-Estimation

Abstract

Macroeconomic data for 142 countries in 2022 and 2023 were obtained from the World Bank, International Monetary Fund, and Trading Economics, covering interest rate, inflation rate, unemployment rate, public debt, trade balance, gross fixed capital formation, fiscal balance, and current account. Correlation analysis reveals that most indicators exhibit weak linear relationships with GDP, while trade balance shows the strongest and statistically significant negative correlation with GDP (2022: $r = -0.314$, $p < 0.001$; 2023: $r = -0.301$, $p < 0.001$). Stepwise MLR identifies unemployment rate, public debt, and current account as significant predictors of GDP in both years. In the MLR models, unemployment rate consistently shows a negative effect on GDP (2022: $\beta = -0.0768$; 2023: $\beta = -0.0776$), while public debt and current account exhibit positive effects. Model evaluation indicates that both MLR and RMLR provide comparable predictive accuracy. The MLR model achieves slightly lower prediction errors, with MAE values of 1.1458 (2022) and 1.3923 (2023), and RMSE values of 1.8092 (2022) and 1.7877 (2023). The corresponding R^2 values for the MLR model are 20.09% in 2022 and 22.03% in 2023, marginally higher than those of the RMLR model. These results suggest that, after logarithmic transformation and assumption checking, the MLR model is sufficient for GDP estimation, while RMLR serves as a valuable robustness check that confirms the stability of the findings. Future research may extend this study by incorporating additional years of data, exploring nonlinear or panel regression frameworks, and examining regional or income-group heterogeneity to further enhance the explanatory power of GDP models.

1. Introduction

Gross Domestic Product (GDP) is the primary global economic measure. As a major macroeconomic tool for economic assessment, GDP provides a recognized framework for tracking national economic conditions by evaluating total output [1]. The determinants of GDP remain a fundamental issue in economic research and policy evaluation. However, many studies tend to focus only on a single country or region, limiting the general applicability of their results. Differences in the methods used to measure GDP growth rates across countries reduce the effectiveness of comparing GDP growth, especially when there are significant differences in variables such as inflation, consumer price index, and national accounting systems. In addition to the issue of indicator correlation, the characteristics of global economic data introduce additional complexity, as these data often exhibit extreme values, outliers, and structural heterogeneity [2]. These limitations suggest the need to consider alternative estimation methods that are less sensitive to outliers.

Previous studies have demonstrated that Ordinary Least Squares (OLS) or Multiple Linear Regression (MLR) models are highly sensitive to outliers and leverage points, often resulting in unstable coefficient estimates and inflated error measures. The presence of extreme observations leads to instability in OLS intercepts and increased standard errors, whereas robust regression methods, particularly MM-estimation, produced stable results with lower mean squared errors [3]. Similarly, M-estimation could handle minor deviations, it struggled with high-leverage points, whereas MM-estimation consistently provided the most reliable estimates under various contamination scenarios [4]. Simulation studies that MM-estimators outperform traditional OLS when regression assumptions are violated, due to their high breakdown point and efficiency across different error distributions [5]. In an economic context, robust regression techniques improved model stability and reduced estimation errors when analysing GDP-related data containing outliers [6]. Overall, empirical evidence supports the use of robust multiple linear regression methods, particularly MM-estimation, as a complementary approach to MLR when data irregularities or assumption violations are present.

The objectives of this study are to identify the relationship between selected economic indicators and GDP using Pearson Correlation Coefficient, to develop a Multiple Linear Regression (MLR) model using stepwise regression and a Robust Multiple Linear Regression (RMLR) model using MM estimation with selected economic indicators as independent variables for estimating GDP, and to compare the performance of the MLR and RMLR models based on Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and coefficient of determination (R^2) to identify the more suitable model for GDP estimation. This paper is structured as follows. Section 1 introduces the background, problem statement, and research objectives. Section 2 describes the methodology used in this study. Section 3 presents the results obtained from the regression models. Section 4 discusses the findings, and the final section concludes the study.

2. Methodology

In this section, the methodology used to conduct the study is discussed. First part talking about the data collection. Part two is show the strength and direction between two quantitative variables. Part three and part four is explain about model MLR, RMLR and their assumption checking. Part four talking about model evaluation about MAE and RMSE.

2.1 Data Collection

The data constitutes the macroeconomic indicators of the year 2022 and year 2023 which are GDP, Policy Interest Rate, Inflation Rate, Unemployment Rate, Public Debt, Trade Balance, Gross Fixed Capital Formation (GFCF), Fiscal Balance and Current Account. These indicators were chosen to explore recent performances of the global economy and its interdependencies and show in Table 1. Three primary sources were used to obtain data, including the World Bank DataBank, the International Monetary Fund (IMF) DataMapper and Trading Economics.

Table 1 The unit of variable and description of variables

Variables	Unit	Description
GDP (dependent variable)	USD Billion	Total value of GDP
Interest Rate	Percentage	The interest rate set by the central bank
Inflation Rate	Annual Variation in percentage	Measures the year-over-year percentage change of CPI in percentage
Unemployment rate	Percentage	The percentage of the labor force
Public Debt	Percentage of GDP	The proportion of government debt compared to GDP value
Trade Balance	USD Billion	The difference between a country's exports and imports of goods and services
GFCF	Annual Variation in percentage	Measures the year-over-year percentage change in investments made in fixed assets
Fiscal Balance	Percentage of GDP	The proportion of the government's fiscal position compared to GDP value
Current Acccount	Percentage of GDP	The proportion of the net trade in goods and services, income, and transfers compared to GDP value

2.2 Pearson Correlation Coefficient

The correlation analysis is a statistic used to measure the strength and direction of the linear relationship between two quantitative variables. Pearson correlation coefficient is the most used statistic to measure the degree of the relationship between linearly related variables [7].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

where r is Pearson correlation coefficient between x and y, n is total sample size, x_i is value of x for i-th observation, y_i is value of y for i-th observation, $\bar{x}$ is the mean value of x for all observation, $\bar{y}$ is mean value of y for all observation.

2.3 Multiple Linear Regression (MLR)

MLR is a method of statistical modelling that describes the relationship between two or more independent variables and one dependent variable. The aim is to know the overall influence of the independent variables on the dependent variable and to estimate the value of the dependent variable given the values of the independent variables [8].

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1} + \varepsilon_i \tag{2}$$

where i is 1, 2, 3, …, n denotes the observations, $x_{i1}, \ldots, x_{i,p-1}$ is the independent variables for observation i, y_i is the dependent variable for observation i, β_0 is the y-intercept of the regression when all independent variables equal 0, β_k is partial regression coefficients or slopes, k is 1, 2, …, $p-1$ and ε_i is random error term, independently and identically distributed.

2.3.1 Assumption of MLR

The assumptions must be followed to ensure that the results are valid and reliable in MLR. Firstly, there is normality of residuals can be evaluated Anderson-Darling test. Secondly, linear relationship between the dependent variable and each independent variable or called linearity. Thirdly, homoscedasticity must be present that the residuals should have constant variance across all levels of the independent variables by Breusch-Pagan test. Lastly, there is no multicollinearity check by Variance Inflation Factor (VIF) [9][10].

2.3.2 Anderson-Darling Test

Normality test is a very important step in many statistical analyses and applications. The invalid conclusions and unreliable results can violate if the distribution is normally distributed [11].

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)\left[\ln\left(F\left(Y_i\right)\right) + \ln\left(1 - F\left(Y_{n-i+1}\right)\right)\right] \tag{3}$$

where n is the sample size, Y_i is ordered data points and $F(Y_i)$ is the cumulative distribution function of the hypothesized distribution evaluated at Y_i.

2.3.3 Box-Cox Transformation

This method is a transformation of non-normal dependent variables into normal shape distribution and stabilize variance [12].

$$y(\lambda) = \begin{cases} y^{\lambda} - 1, \ \lambda \neq 0 \\ \log(y), \ \lambda = 0 \end{cases} \tag{4}$$

where λ is transformation parameter. All values of λ are considered and the optimal value from the data used. The optimal value is the best approximation of a normal distribution curve.

2.3.3 Breusch-Pagan Test

The Breusch-Pagan test is a statistical test performed to identify the existence of heteroscedasticity in a MLR model [13].

$$\hat{\varepsilon}_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \ldots \gamma_k x_{ik} + u_i \tag{5}$$

where $x_{i1}, \ldots, x_{ik}$ are the independent variables for observation i, $\hat{\varepsilon}_i^2$ is the dependent variable for observation i, γ_0 is the intercept of the auxiliary regression, $\gamma_1, \ldots, \gamma_k$ is partial regression coefficients in the auxiliary regression, k is the number of independent and u_i is the error term in auxiliary regression. The Breusch–Pagan test (BP) statistic is computed as the product of the sample size n, and the coefficient of determination, $\hat{R}^2$obtained from the auxiliary regression. The null hypothesis is rejected if the test statistic exceeds the critical value from the chi-square distribution$\chi^2_{k,a}$, indicating heteroscedasticity is present.

2.3.4 Variance Inflation Factor (VIF)

One way to detect multicollinearity is using VIF, which quantifies how much the variance of a coefficient is increased due to multicollinearity. VIF above 5 show that multicollinearity might exist, and further investigation is required [14].

$$VIF_i = \frac{1}{1 - R_i^2} \tag{6}$$

where $\boldsymbol{R_i^2}$ is the unadjusted for R^2 for i-th independent variables.

2.3.5 Stepwise Regression

Stepwise regression repeatedly chooses the independent variables are necessary for building a prediction of model. It simplifies large sets of data by systematically including or removing variables according to their importance for the model which makes it helpful for spotting the most important independent and keeping the main information. The process repeats itself until adding more variables does not result in meaningful improvements and the model contains just the most important features [15].

2.4 Robust Multiple Regression (RMLR)

RMLR model addresses extreme values or high leveraged observations in economics, which tend to be valid and informative data rather than errors [16]. Conventional methods for multivariate analysis are usually affected by extreme values because apply least squares to find the best fit of curve to minimizing the sum of square error. The purpose of a robust method is to minimize or eliminate the impact of unusual data so the average of most of the data is more decisive. Moreover, strong methods are value-additional since they enable analysts to identify and manage extreme values or high-leverage observations by combining exploratory methods and standard diagnostic measures to verify model stability and reliability [17].

2.4.1 MM-Estimation

MM-estimation is constructed using a three-step procedure. First, an initial regression estimate $\hat{\beta}_0$is obtained using an S-estimator. Next, the scale of the residuals, $\hat{\sigma}$, is estimated using an M-estimator of scale based on the initial regression estimate $\hat{\beta}_0$ [5].

$$\widehat{\beta} = \underset{\beta}{argmin}\ (r_1(\beta), \ldots, r_n(\beta)) \tag{10}$$

where $\hat{\beta}$ is the estimated of the regression coefficients, $\hat{\sigma}$ is robust error scale estimate and $r_i(\beta) = y_i - x_i^T\beta$ is the residual of observation i. Then, using $\hat{\beta}_0$as the starting value and the fixed scale $\hat{\sigma}$, an M-estimator of the final regression coefficients $\hat{\beta}$is computed.

$$\frac{1}{n}\sum_{i=1}^{n} \varrho\left(\frac{r_i(\widehat{\beta})}{\widehat{\sigma}}\right) = \delta \tag{11}$$

where ρ is a bounded loss function, $\hat{\sigma}$ is robust error scale estimate and δ is a constant, the expected value of $\rho(r)$ under standard normal distribution. Finally, the MM-estimation of the regression coefficients is obtained by minimizing.

$$L(\beta) = \sum_{i=1}^{n} \varrho\left(\frac{r_i(\widehat{\beta})}{\widehat{\sigma}}\right) \tag{12}$$

where $L(\beta)$ is the total loss over all residuals after scaling and applying ρ, β is vector of regression coefficients, ρ is a bounded loss function, $r_i(\beta)$ is β is the residual of observation i, and $\hat{\sigma}$ is robust error scale estimate.

2.5 Assumption of RMLR

In real-world data, a few observations tend to deviated from the fitted regression line. The primary goal of regression analysis is to understand the relationship between the dependent variable and the independent variables that make up the majority of the data, rather than focusing excessively on these isolated observations. However, these extreme values give some warrant special attention and they can make significantly impact on the model estimations and statistical conclusion [4].

2.5.1 Hat Matrix Leverage Value

Leverage values are derived from the hat matrix, and it quantify how much each observation's predictor values can influence the regression fit [8].

$$\hat{h}_{ii} = X\left(X^T X\right)^{-1} X^T \tag{7}$$

where $\hat{h}_{ii}$ is the leverage value of i-th observation, X is the design matrix of regression model and X^T is transpose of the design matrix. The leverage threshold defined as two times the ratio of $(p+1)$to the sample size n, where $p + 1$ represents the total number of independent variables in the model including the intercept. Any observation with a leverage value $\hat{h}_{ii}$ exceeds than this threshold is considered a high-leverage observation.

2.5.2 Studentized Residuals

Studentized residuals are typical procedures used to identify unusually large residuals in the dependent variable to identify extreme values that is greater than absolute value of 3 [8].

$$z_i = \frac{e_i}{\hat{\sigma}\sqrt{1-\hat{h}_{ii}}} \tag{8}$$

where e_i *is* the difference between actual value and predicted value on i-th observation, $\hat{\sigma}$ is standard error standard deviation and $\hat{h}_{ii}$ is the leverage value of i-th observation.

2.5.3 Cook's Distance

Cook's distance is used to assess the impact of observations resulting from its combined effects on studentized residual and leverage value [8].

$$D_i = \frac{e_i}{\hat{\sigma}\sqrt{1-\hat{h}_{ii}}} \cdot \frac{\hat{h}_{ii}}{\left(1-\hat{h}_{ii}\right)^2} \tag{9}$$

where e_i is the difference between actual value and predicted value on i-th observation, $\hat{\sigma}$ is residual standard deviation and $\hat{h}_{ii}$ is the leverage value of i-th observation. An observation considered highly influential when its Cook's distance value, $D_i$exceeds the threshold value defined as 4 divided by sample size n.

2.6 Model Evaluation

The study applies popular regression evaluation methods to measure the model of MLR and RMLR. The MAE and RMSE are examples of these, which examine things such as the average magnitude of prediction mistakes and how well the variables related.

2.6.1 Mean Absolute Error (MAE)

MAE is appropriate if outliers have been affected by incorrect data. As a result, MAE tends not to downgrade the performance of outliers, which is why the performance of the model remains constant and bound. However, if the test set includes a high number of outliers, the performance of the model will not be very good [18].

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_t - \hat{y}_t\right| \tag{13}$$

where n is number of observations, y_t is actual value for i-th observation and $\hat{y}_t$ is predicted value by regression model for i-th observation.

2.6.2 Root Mean Square Error (RMSE)

RMSE is calculated by taking the square root of Mean Square Error. RMSE functions similarly to the standard deviation of residuals, indicating the degree of dispersion of the residuals [19].

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_t - \hat{y}_t\right)^2} \tag{14}$$

where n is number of observations, y_t is actual value for i-th observation and $\hat{y}_t$ is predicted value by regression model for i-th observation.

2.6.3 Coefficient of determination (R^2)

The coefficient of determination measures the part of the dependent variable's changes that the independent variables can explain [18].

$$R^2 = 1 - \frac{SSE}{SST} \tag{14}$$

where SSE is the sum of squared errors and SST is the total sum of squares

3. Results and Discussions

3.1 Correlation Analysis

Table 2 show the result of Pearson Correlation coefficient between GDP and selected economic indicator.

Table 2 Pearson Correlation coefficient between GDP and selected economic indicator

Dependent Variable	Independent Variables	Pearson Correlation coefficient 2022	*p*-value	Pearson Correlation coefficient 2023	*p*-value
GDP	Interest Rate	-0.053	0.530	-0.051	0.550
	Inflation Rate	-0.073	0.389	-0.060	0.479
	Unemployment rate	-0.098	0.244	-0.093	0.269
	Public Debt	0.201	0.016	0.223	0.008
	Trade Balance	-0.314	0.000	-0.301	0.000
	GFCF	-0.001	0.993	-0.063	0.457
	Fiscal Balance	-0.056	0.509	-0.069	0.415
	Current Account	0.063	0.459	0.047	0.578

The Pearson correlation analysis for 2022 and 2023 indicates that most selected economic indicators do not exhibit statistically significant linear relationships with GDP at the 5% significance level. Interest Rate, Inflation Rate, Unemployment Rate, GFCF, Fiscal Balance, and Current Account show weak and not statistically significant correlations with GDP in both years. In contrast, Public Debt demonstrates a weak but statistically significant positive correlation with GDP, while Trade Balance exhibits a moderate and statistically significant negative correlation across both years. The consistency of these findings suggests that among the selected indicators, Trade Balance has the strongest linear association with GDP.

3.2 MLR Model

3.2.1 Normality Test

Table 3 shows the results of the normality test of GDP by using the Anderson-Darling test.

Table 3 Normality test of GDP in 2022 and 2023

Anderson-Darling test	GDP 2022	*ln* GDP 2022	GDP 2023	*ln* GDP 2023
p-value	<0.005	0.076	<0.005	0.121

The original data for GDP values in 2022 and 2023 are not normally distributed. Hence, the Box–Cox transformation is applied, and the natural logarithm (*ln*) transformation identified and provided an approximately normal distribution. This transformation satisfies the normality assumption required for regression analysis for both the 2022 and 2023 datasets.

3.2.2 Linearity Test

The figure 1 and 2 show the linearity test using the residual plot versus each independent variable in 2022 and 2023. The residual plots 2022 and 2023 show that the residuals are randomly distributed around the zero line across all independent variables, indicating that the linear specification of the model is appropriate. No clear systematic patterns or non-linear relationships are observed. The residual plots support the adequacy of the MLR model, as the key assumptions of linearity are satisfied.

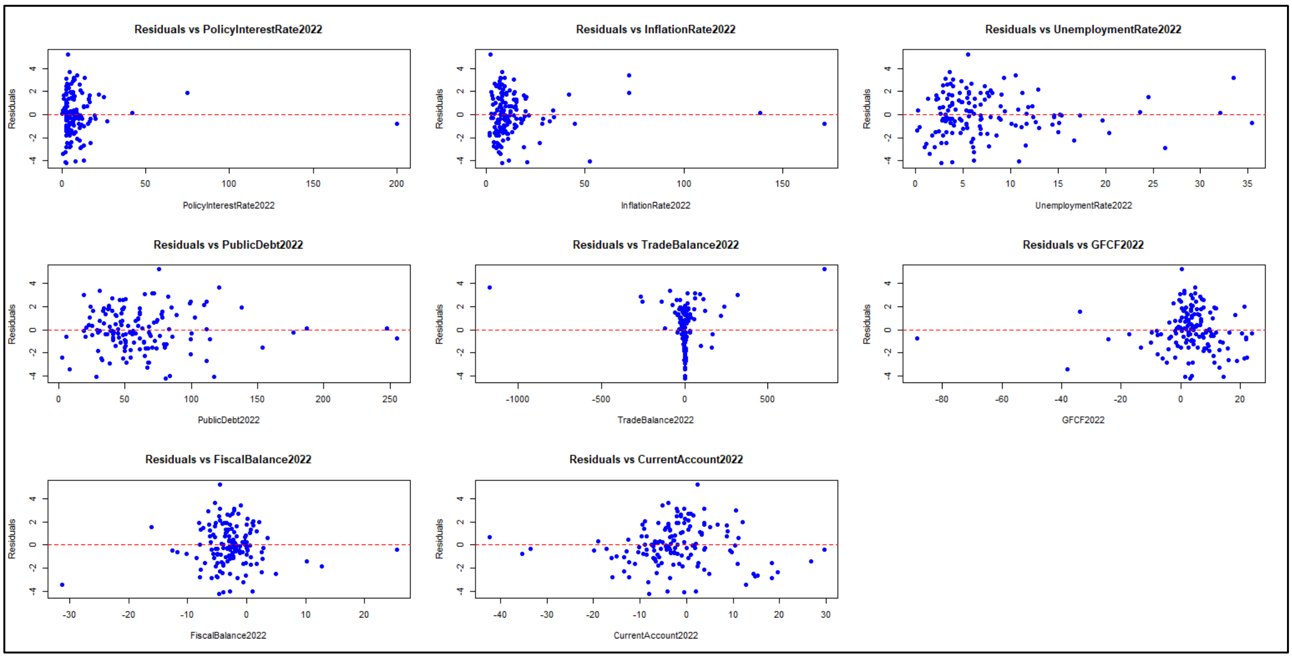

Fig. 1 Residual plot vs each independent variable 2022

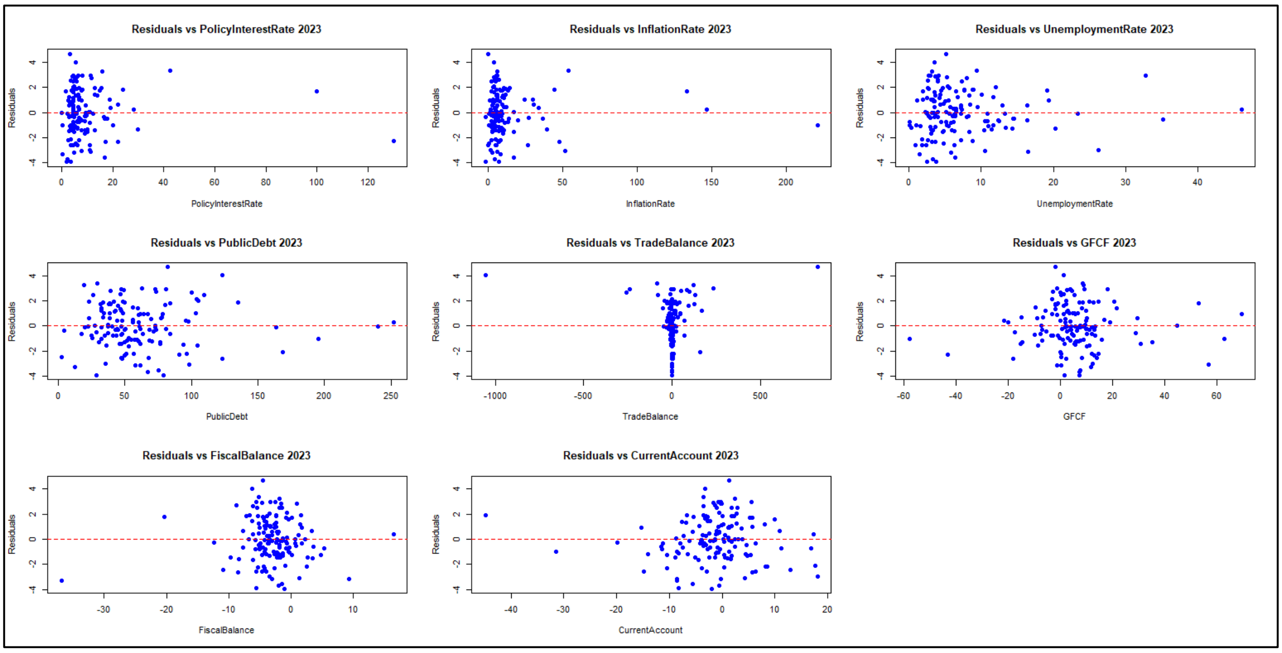

Fig. 2 Residual plot vs each independent variable 2023

3.2.3 Breusch-Pagan Test

Before the Breusch-Pagan test, the auxiliary regression of squared residuals produced coefficient of determination as 0.1066 in 2022 and 0.0518 in 2023 with a total sample size of 142 observations. For the chi-square critical value at the 5% significance level with 8 degrees of freedom is 15.507. Table 4 shows the result of Breusch-Pagan test of 2022 and 2023.

Table 4 Breusch-Pagan test of 2022 and 2023

Breusch-Pagan test	2022	2023
Test-statistic value	15.1372	7.3556

For both 2022 and 2023, the Breusch-Pagan test statistics of 15.1372 and 7.3556 respectively do not exceed the Chi-Square critical value of 15.507, indicating that the null hypothesis of homoscedasticity cannot be rejected and the models satisfy the assumption of constant variance.

3.2.4 Multicollinearity Check

The multicollinearity check using the Variance Inflation Factor (VIF) to ensure that independent variables are not highly correlated. Table 5 shows the VIF results of ln GDP between the independent variables in 2022 and 2023.

Table 5 VIF result in 2022 and 2023

VIF result	Interest Rate	Inflation Rate	Unemployment rate	Public Debt	Trade Balance	GFCF	Fiscal Balance	Current Account
2022	1.25	2.06	1.09	1.33	1.08	1.62	1.27	1.22
2023	1.25	1.87	1.17	1.31	1.05	1.22	1.07	1.21

The results of the VIF show that all independent variables have values well below the common threshold of 4. This suggests that there is no significant multicollinearity among the variables in 2022 and 2023.

3.2.5 Stepwise Regression

The stepwise regression results for the MLR model in 2022 and 2023 identify Current Account, Public Debt, and Unemployment Rate as the significant economic indicators associated with GDP. Table 6 presents the inferences of regression coefficients of the final MLR models after the stepwise selection procedure. The t-test results indicate that all selected variables are statistically significant at the 5% significance level in both years. In addition, the Variance Inflation Factor (VIF) values are close to 1, suggesting that multicollinearity is not a concern.

Table 6 Inferences of regression coefficient MLR after stepwise selection in 2022 and 2023

Year	2022				2023			
	Coefficient	t-value	*p*-value	VIF	Coefficient	t-value	*p*-value	VIF
Constant	4.275	12.97	0.000		4.111	13.29	0.000	
Unemployment Rate	-0.0768	-3.12	0.002	1.02	-0.0776	-3.27	0.001	1.06
Public Debt	0.0151	3.63	0.000	1.07	0.01808	4.31	0.000	1.07
Current Account	0.0712	4.44	0.000	1.05	0.0858	4.30	0.000	1.02

From an economic perspective, the Unemployment Rate exhibits a negative relationship with GDP (2022: -0.0768, t = -3.12, p = 0.002; 2023: -0.0776, t = -3.27, p = 0.001), consistent with economic theory, as higher unemployment reflects underutilization of labour resources and leads to lower economic output. Public Debt shows a positive association with GDP (2022: 0.0151, t = 3.63, p < 0.001; 2023: 0.0181, t = 4.31, p < 0.001), suggesting that government borrowing may contribute to economic activity through expansionary fiscal policies, although the magnitude of the effect remains modest. The positive coefficient of the Current Account (2022: 0.0712, t = 4.44, *p* < 0.001; 2023: 0.0858, t = 4.30, p < 0.001) indicates that improvements in external balance are associated with higher GDP, potentially due to increased net exports or improved international competitiveness. These findings support the study's objective of identifying key economic indicators that influence GDP using the MLR model.

3.3 RMLR model

3.3.1 Extreme values and Influential points

Table 7 shows the result of extreme values and influential points from stepwise regression of MLR models in 2022 and 2023. These results are test by studentized residual, hat matrix leverage value and Cook's Distance.

Table 7 Result of extreme values and influential points

Test	2022		2023	
	Case number	Proportion of sample	Case number	Proportion of sample
Studentized residual	No observations exceeded the threshold	0%	No observations exceeded the threshold	0%

Hat matrix leverage values	16, 36, 51, 65, 70, 72, 87, 97, 107, 115, 118, 121, 123 and 135	9.86%	16, 36, 51, 65, 70, 72, 115, 118, 121 and 123	7.04%
Cook's Distance	36, 72, 118, 122, 129, 137	4.23%	36, 70, 72, 115, 118, 122 and 137	4.93%

In both years, no observations exceeded the ±3 threshold for studentized residuals, indicating the absence of extreme outliers in the response variable. High-leverage observations accounted for approximately 9.86% of the sample in 2022 and 7.04% in 2023, reflecting the presence of several observations with unusual predictor combinations, which is expected in cross-country economic data due to structural heterogeneity. Influential points based on Cook's distance represented 4.23% of observations in 2022 and 4.93% in 2023, suggesting that a small proportion of cases may moderately affect the estimated coefficients. Overall, fewer than 10% of observations were identified as extreme or influential across all diagnostic measures in both years, indicating that the datasets are not dominated by outliers or highly influential points. The identified observations likely represent valid economic conditions rather than data errors, supporting the robustness of the MLR stepwise models.

3.3.2 MM-Estimation

Table 8 shows the result of inferences of regression coefficient RMLR model 2022 and 2023 by using MM-estimation including the coefficients, *t*-value and *p*-value. The results show that Unemployment Rate, Public Debt, Trade Balance, and Current Account in 2022 and 2023 have stronger effects in MM-estimation with very small p-values.

Table 8 Inferences of regression coefficient RMLR model 2022 and 2023

Term	2022			2023		
	Coefficient	*t*-value	*p*-value	Coefficient	*t*-value	*p*-value
Intercept	4.4042	12.1136	0.0000	4.2463	11.5583	0.0000
Policy Interest Rate	-0.0060	-0.6197	0.5365	0.0039	0.3161	0.7524
Inflation Rate	0.0041	0.3670	0.7142	-0.0013	-0.1500	0.8810
Unemployment Rate	-0.0804	-3.0834	0.0025	-0.0696	-2.6825	0.0082
Public Debt	0.0130	2.7222	0.0074	0.0144	2.9853	0.0034
Trade Balance	-0.0038	-3.1482	0.0020	-0.0037	-2.8521	0.0050
GFCF	0.0124	0.7332	0.4648	-0.0123	-1.0715	0.2859
Fiscal Balance	0.0084	0.2352	0.8144	-0.0158	-0.4628	0.6443
Current Account	0.0754	4.2502	0.0000	0.0903	3.9795	0.0001

The remaining variables such as Policy Interest Rate, Inflation Rate, GFCF and Fiscal Balance demonstrate limited influence in the robust regression. This reinforces that only certain macroeconomic indicators play a meaningful role in explaining GDP variations under the robust regression framework. Economically, the results mirror the MLR model. The negative association of the Unemployment Rate indicates that higher unemployment reduces GDP, while Public Debt contributes positively to GDP. Trade Balance shows a negative effect, highlighting the importance of net exports for economic performance, and improvements in the Current Account are associated with higher GDP. These relationships are stable across both years, demonstrating that the RMLR model using MM-estimation effectively identifies key indicators while reducing the influence of outliers.

3.4 Model Evaluation

Table 9 shows the result of MAE, RMSE and Coefficient of Determination for MLR and RMLR model in 2022 and 2023.

Table 9 Result of MAE and RMSE Coefficient of Determination for MLR and RMLR model in 2022 and 2023

Model	MAE	RMSE	Coefficient of Determination, R^2

MLR Model 2022	1.1458	1.8092	20.09
RMLR Model 2022	1.3913	1.8195	19.17
MLR Model 2023	1.3923	1.7877	22.03
RMLR Model 2023	1.3931	1.7932	19.27

The results indicate that the MLR and RMLR models exhibit very similar predictive accuracy in both 2022 and 2023. In 2022, the MLR model achieved lower MAE (1.1458) and RMSE (1.8092) compared to the RMLR model. In 2023, the prediction errors of both models were very close, with the MLR model again showing slightly lower values. The coefficient of determination (R^2), which measures the proportion of variance in GDP explained by the model, was also comparable between the two models. In 2022, the MLR model achieved an R^2 of 20.09%, while the RMLR model recorded 19.17%. A similar pattern was observed in 2023, where the MLR model attained an R^2 of 22.03%, compared to 19.27% for the RMLR model. These relatively small differences suggest that both models capture GDP variability to a comparable extent.

The slightly better performance of the MLR model, despite the presence of some high-leverage and influential observations, can be explained by two factors. First, the logarithmic transformation of GDP helps stabilize variance and reduces the influence of atypical observations, thereby diminishing the potential advantage of robust regression. Second, the MLR model was constructed using stepwise selection, which removes non-significant predictors and indirectly limits the impact of leverage points. In contrast, the RMLR model was estimated using all selected predictors without stepwise selection. Consequently, the minor differences in predictive errors and R^2 are consistent with the diagnostic results, and the MLR model remains appropriate for GDP estimation under these conditions.

4. Conclusion

The Pearson correlation analysis for 2022 and 2023 indicates that most selected economic indicators do not exhibit statistically significant linear relationships with GDP at the 5% significance level. Interest Rate, Inflation Rate, Unemployment Rate, Gross Fixed Capital Formation (GFCF), Fiscal Balance, and Current Account display weak and non-significant correlations with GDP in both years. In contrast, Public Debt shows a weak but statistically significant positive correlation with GDP, while Trade Balance demonstrates a moderate and statistically significant negative correlation across both years.

The MLR models for 2022 and 2023, developed using stepwise regression, satisfied key regression assumptions, including linearity, homoscedasticity, and low multicollinearity, following the natural logarithm transformation of GDP. Diagnostic tests based on studentized residuals, leverage values, and Cook's Distance indicate that extreme and highly influential observations are limited and do not dominate the datasets, supporting the stability and reliability of the estimated coefficients.

Robust multiple linear regression (RMLR) models based on MM estimation were developed in parallel as a robustness check and produced results consistent with the MLR estimates. A comparison of predictive performance for both years shows that the MLR model consistently achieved equal or slightly lower MAE and RMSE than the RMLR model, indicating marginally better predictive accuracy. The coefficient of determination (R^2) was also comparable between models, with MLR explaining 20.09% and 22.03% of GDP variation in 2022 and 2023, respectively, while RMLR explained 19.17% and 19.27%.

The slightly better performance of the MLR model can be attributed to the logarithmic transformation of GDP, which stabilizes variance and reduces the influence of atypical observations, as well as the use of stepwise selection, which removes non-significant predictors and indirectly mitigates the impact of leverage points. Overall, the selected economic indicators are suitable for GDP estimation, the MLR model is the most appropriate predictive approach for the given dataset, and the RMLR results confirm the robustness and stability of the regression analysis.

Acknowledgement

The authors would thank the Faculty of Applied Sciences and Technology, Universiti Tun Hussein Onn Malaysia, for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design****: Tan Shun Hong, Azme Khamis;* ***solve the governing equation****: Tan Shun Hong, Azme Khamis;* ***data collection****: Tan Shun Hong;* ***analysis and interpretation of results****: Tan Shun Hong, Azme Khamis;* ***draft manuscript preparation****: Tan Shun Hong, Azme Khamis. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Stobierski, T. (2021). What is GDP & why is it important? Harvard Business School Online. https://online.hbs.edu/blog/post/why-is-gdp-important/

[2] Hosseiny, A. (2015). Violation of invariance of measurement for GDP growth rate and its consequences. arXiv. https://doi.org/10.48550/arXiv.1507.04848

[3] Jana, P., Rosadi, D., & Supandi, E. (2023). Comparison of robust estimation on multiple regression model. BAREKENG: Jurnal Ilmu Matematika dan Terapan, 17(2), 979–988. https://doi.org/10.30598/barekengvol17iss2pp0979-0988

[4] Fitrianto, A., & Xin, S. (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/barekengvol16iss1pp241-250

[5] Yu, C., & Yao, W. (2017). Robust linear regression: A review and comparison. Communications in Statistics – Simulation and Computation, 46(8), 6261–6282. https://doi.org/10.1080/03610918.2016.1202271

[6] Khan, D. M., Yaqoob, A., Zubair, S., Khan, M. A., Ahmad, Z., & Alamri, O. A. (2021). Applications of robust regression techniques: An econometric approach. Mathematical Problems in Engineering, 2021, 6525079. https://doi.org/10.1155/2021/6525079

[7] Li, G., Zhang, A., Zhang, Q., Wu, D., & Zhan, C. (2022). Pearson correlation coefficient-based performance enhancement of broad learning system for stock price prediction. IEEE Transactions on Circuits and Systems II: Express Briefs, 69(5), 2413–2417. https://doi.org/10.1109/TCSII.2022.3160266

[8] Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2013). Applied linear statistical models (5th ed.). McGraw-Hill Education (India) Private Limited.

[9] Alita, D., Putra, A. D., & Darwis, D. (2021). Analysis of classic assumption test and multiple linear regression coefficient test for employee structural office recommendation. Indonesian Journal of Computing and Cybernetics Systems, 15(3), 295–306. https://doi.org/10.22146/ijccs.65586

[10] Tranmer, M., Murphy, J., Elliot, M., & Pampaka, M. (2020). Multiple linear regression (2nd ed.; CMIST Working Paper 2020-01). Cathie Marsh Institute, University of Manchester. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/2020-1-multiple-linear-regression.pdf

[11] Demir, S. (2022). Comparison of normality tests in terms of sample sizes under different skewness and kurtosis coefficients. International Journal of Assessment Tools in Education, 9(2), 397–409. https://doi.org/10.21449/ijate.1101295

[12] Plummer, A. (2025). Box Cox transformation and target variable: Explained. Built In. https://builtin.com/data-science/box-cox-transformation-target-variable

[13] Akewugberu, H. O., Umar, S. M., Musa, U. M., Ishaaq, O. O., Ibrahim, A., & Osi, A. A. (2024). Breusch–Pagan test: A comprehensive evaluation of its performance under different heteroscedastic structures. FUDMA Journal of Sciences, 8(6), Article 2826. https://fjs.fudutsinma.edu.ng/index.php/fjs/article/view/2826

[14] Marcoulides, K. M., & Raykov, T. (2019). Evaluation of variance inflation factors in regression models using latent variable modeling methods. Educational and Psychological Measurement, 79(5), 874–882. https://doi.org/10.1177/0013164418817803

[15] Manishankar, S., Kumaraperumal, R., Ragunath, K., & Kannan, B. (2021). Selection of environmental covariates using stepwise regression. Pharma Innovation, 10(5), 380–385. https://www.researchgate.net/publication/356253306_Selection_of_environmental_covariates_using_stepwise_regression

[16] Scott, D. W., & Wang, Z. (2021). Robust multiple regression. Entropy, 23(1), 88. https://doi.org/10.3390/e23010088

[17] Møller, S. F., von Frese, J., & Bro, R. (2005). Robust methods for multivariate data analysis. Journal of Chemometrics, 19(10), 549–563. https://doi.org/10.1002/cem.962

[18] Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Computer Science, 7, e623. https://doi.org/10.7717/peerj-cs.623

[19] Tatachar, A. V. (2021). Comparative assessment of regression models based on model evaluation metrics. International Research Journal of Engineering and Technology (IRJET), 8(9), Article V8I9127. https://www.irjet.net/archives/V8/i9/IRJET-V8I9127.pdf