International Journal of Financial Management and Economics

P-ISSN: 2617-9210
E-ISSN: 2617-9229
IJFME 2024; 7(2): 530-538
www.theeconomicsjournal.com
Received: 26-10-2024
Accepted: 30-11-2024

Nagendra Marisetty
Faculty, REVA Business School (RBS), REVA University, Bangalore, Karnataka, India

Corresponding Author:
Nagendra Marisetty
Faculty, REVA Business School (RBS), REVA University, Bangalore, Karnataka, India

Mitigating assumption violations in regression analysis: Insights from 20 years of currency pair data

Nagendra Marisetty

DOI: https://doi.org/10.33545/26179210.2024.v7.i2.412

Abstract
This article focuses on addressing violations of regression assumptions in the analysis of 20 years of monthly price data for nine international currency pairs. The Ordinary Least Squares (OLS) model, while achieving a high Adjusted R-squared, suffers from significant assumption violations, including residuals heteroscedasticity, autocorrelation, non-linearity, and multicollinearity. To address multicollinearity, removes variables with excessive variance inflation factors, improving coefficient reliability. However, issues like residual heteroscedasticity, autocorrelation and nonlinearity persist, indicating the need for further refinement. Transformation-based approaches, such as First Difference (FD) and Log Difference (LD), significantly improve assumption compliance by reducing residual heteroscedasticity, autocorrelation, non-linearity, and ARCH effects. Among these, the OLS Log Difference (OLS LD) model demonstrates the most effective correction of diagnostic issues, achieving compliance with key assumptions while minimizing standard errors and Akaia criterion (AIC). Although Weighted Lease Square (WLS) and Heteroscedasticity-Corrected (HSC) models also address some violations, their limited success in mitigating residual autocorrelation, nonlinearity, and complexity reduces their practicality. Overall, the OLS LD model emerges as the most effective approach, balancing assumption compliance and precision while providing reliable insights into the dynamics of currency pair price behaviours over the study period.

Keywords: Autocorrelation, currency pairs, first difference, heteroscedasticity, log difference, multicollinearity, non-linearity, ordinary least squares and regression assumptions

1. Introduction

Regression analysis is a cornerstone of quantitative research across disciplines, providing a framework for understanding relationships between variables and making predictions. However, the validity of regression models relies on adherence to several assumptions, including linearity, independence of errors, homoscedasticity, and the absence of multicollinearity. Violations of these assumptions can compromise the reliability and interpretability of model results, leading to biased or inefficient estimates. Addressing these violations is crucial for ensuring robust and meaningful analysis, particularly in complex datasets such as those from financial markets. Financial markets, and specifically the forex market, represent a unique challenge for regression analysis. Currency pair prices are influenced by a myriad of factors, including macroeconomic indicators, geopolitical events, and market sentiment, resulting in dynamic and often non-linear relationships. These complexities frequently lead to assumption violations such as heteroscedasticity, where residuals exhibit non-constant variance, and autocorrelation, where residuals are correlated over time. Additionally, the inherent interdependence among financial instruments often results in multicollinearity, which inflates standard errors and reduces the reliability of individual coefficient estimates.

This study focuses on analysing 20 years of monthly price data for nine international currency pairs to investigate these challenges in detail. The initial analysis using the Ordinary Least Squares (OLS) method highlighted significant issues despite achieving a high Adjusted R-squared value. Diagnostic tests revealed severe violations of regression assumptions, including multicollinearity, residual heteroscedasticity, autocorrelation, and non-linearity. These issues underscored the need for corrective strategies to enhance model validity and reliability.

Multicollinearity, in particular, emerged as a significant challenge in the initial OLS model. Variance Inflation Factor (VIF) diagnostics indicated substantial overlap among predictor variables, leading to unreliable coefficient estimates. Removing variables with excessive VIF values partially addressed this issue, improving model interpretability. However, the adjustments did not resolve other critical violations, such as heteroscedasticity and autocorrelation, suggesting that more advanced techniques were necessary.

To address these residual issues, transformation-based approaches, such as First Difference (FD) and Log Difference (LD), were applied. These transformations effectively stabilized residual variance and reduced autocorrelation and non-linearity. Among these, the OLS LD model stood out for its ability to achieve diagnostic compliance while maintaining precision and efficiency. The model demonstrated significant improvements in assumption compliance, as evidenced by reduced heteroscedasticity, autocorrelation, and non-linearity. However, the trade-off was a decline in Adjusted R-squared values, reflecting a shift from explanatory power to diagnostic accuracy.

Alternative approaches, such as Weighted Least Squares (WLS) and Heteroscedasticity-Corrected (HSC) models, were also evaluated. These methods showed promise in addressing heteroscedasticity and improving model efficiency but were less effective in resolving autocorrelation and non-linear dependencies. The HSC model, in particular, achieved a high Adjusted R-squared value, but persistent diagnostic issues limited its practical application. These findings highlighted the strengths and limitations of various corrective strategies, emphasizing the importance of model selection based on specific diagnostic challenges. The insights gained from this study underscore the complexities of regression analysis in financial contexts and the critical role of diagnostic testing in model development. While traditional OLS models are a useful starting point, their limitations in handling assumption violations necessitate the adoption of more advanced techniques. The effectiveness of transformation-based approaches like OLS LD in mitigating key issues makes them a valuable tool for analysing complex datasets, such as currency pair prices, where assumption violations are prevalent.

This paper contributes to the growing body of literature on regression diagnostics and corrective strategies by systematically comparing standard and corrective approaches. It offers practical insights for researchers and practitioners seeking to navigate the challenges of financial data analysis while ensuring model validity. By addressing assumption violations and refining regression models, the study aims to provide a more reliable framework for understanding the dynamics of currency pair prices over extended periods. The remainder of this paper is structured as follows. Section 2 provides a comprehensive review of literature on regression diagnostics and assumption violations, focusing on their implications in financial data analysis. Section 3 describes the methodology employed in analysing the dataset, including the application of regression diagnostics and model selection criteria. Section 4 discusses corrective strategies, comparing their effectiveness in addressing assumption violations. Finally, Section 5 concludes with a synthesis of findings and recommendations for future research in this domain.

2. Literature Review

Regression analysis has been widely utilized in economic and financial research to explore relationships between variables and predict trends. However, the reliability of regression outcomes heavily depends on adherence to critical assumptions such as linearity, homoscedasticity, and independence of residuals. Violations of these assumptions often compromise model validity, making it essential to understand the implications and develop strategies to address them. This section reviews key literature on regression diagnostics, assumption violations, and corrective methodologies, with a specific focus on applications in foreign exchange markets and long-term data analysis.

The presence of violations in regression assumptions has been a significant topic of interest among researchers due to its impact on model reliability and interpretability. Numerous studies emphasize the importance of meeting regression assumptions to ensure model validity. Osborne and Waters (2002) [20] highlighted those violations of assumptions like homoscedasticity and independence lead to biased coefficient estimates, inflated standard errors, and inaccurate hypothesis testing. Work by Gujarati and Porter (2009) [13] in "Basic Econometrics" laid the foundation for understanding the consequences of such violations, emphasizing the need for robust diagnostics and corrections. Their approach highlighted the inadequacies of Ordinary Least Squares (OLS) models in handling heteroscedasticity, autocorrelation, and multicollinearity, especially in financial datasets. Forex market data, with its dynamic and non-linear behaviour, exemplifies these challenges, prompting researchers to explore advanced methodologies for assumption compliance. Similarly, Williams *et al*. (2013) [27] discussed the impact of multicollinearity, particularly in financial data, where interdependent variables often obscure individual predictor effects. These issues are especially relevant in exchange rate analysis, as currency movements are influenced by highly interrelated factors such as interest rates, inflation, and trade balances

Breusch and Pagan (1979) [4] introduced the widely-used Breusch-Pagan test, allowing researchers to detect heteroscedasticity systematically. The test has since been a cornerstone for diagnosing non-constant error variances in regression models. Similarly, the Durbin-Watson statistic, proposed by Durbin and Watson (1950) [8], has been extensively used to identify autocorrelation in residuals. These diagnostic tools have become standard in financial econometrics, particularly for forex data analysis, where temporal dependencies are prominent. To address multicollinearity, Mason and Perreault (1991) [18] advocated for using the Variance Inflation Factor (VIF), setting thresholds for acceptable collinearity levels. They demonstrated the practical implications of high VIF values on coefficient stability and model interpretability, findings that align with the challenges identified in forex regression analysis. Recent studies, such as Katrakilidis and Trachanas (2012) [17], have further applied these techniques to exchange rate modelling, confirming their utility in improving model robustness.

The limitations of OLS have driven the development of alternative models, such as Weighted Least Squares (WLS) and Generalized Least Squares (GLS). Pindyck and

Rubinfeld (1998) [21] highlighted WLS as a solution for heteroscedasticity, demonstrating its ability to adjust for non-constant variance by assigning weights to observations. While effective in many cases, studies like Asteriou and Hall (2021) [1] noted that WLS struggles with non-linearity and residual autocorrelation, particularly in datasets with structural breaks, a common feature in forex markets.

Heteroscedasticity-Corrected (HSC) models, introduced by White (1980) [26], offered another avenue for handling heteroscedasticity without requiring specific assumptions about the error structure. However, researchers like Engle (1982) [10] emphasized that HSC methods fail to address ARCH (Autoregressive Conditional Heteroscedasticity) effects, which are prevalent in financial time series data. Engle's introduction of the ARCH model provided a framework to explicitly model volatility clustering, a critical advancement for forex data analysis. Heteroscedasticity, a common issue in financial data, has been extensively studied. Gujarati and Porter (2009) [13] noted that the presence of heteroscedasticity violates the assumption of constant variance in residuals, leading to inefficient parameter estimates. Techniques such as Weighted Least Squares (WLS) and Heteroscedasticity-Corrected (HSC) models have been proposed to mitigate this issue, though their effectiveness varies depending on the dataset. Similarly, autocorrelation, often observed in time-series data like exchange rates, introduces serial dependence, undermining the validity of standard regression techniques (Greene, 2018) [12]. Durbin-Watson and Breusch-Godfrey tests are commonly employed to detect autocorrelation, with corrections typically involving transformation-based methods or the application of time-series models such as ARIMA.

Transformative approaches, such as First Differences (FD) and Log Differences (LD), have gained traction for their simplicity and effectiveness. Box and Jenkins (1976) [5] pioneered these transformations in time-series modelling, demonstrating their utility in stabilizing variance and reducing autocorrelation. Their methodologies have since been adopted in forex data analysis to address persistent assumption violations. Hansen and Johansen (1999) [15] extended these approaches by combining transformations with cointegration techniques, providing insights into long-term relationships among currency pairs. More recently, Escribano and Mira (2002) [11] applied FD and LD models to high-frequency forex data, reporting substantial improvements in regression diagnostics, particularly in reducing ARCH effects and non-linearity. Transformations also have gained traction in addressing assumption violations in long-term data. Research by Asteriou and Hall (2021) [1] demonstrated the efficacy of such methods in reducing residual heteroscedasticity and autocorrelation, thereby enhancing model reliability. Additionally, recent advancements in robust regression techniques and machine learning offer promising avenues for addressing persistent violations, though these approaches remain underexplored in forex markets.

The advent of machine learning has introduced innovative solutions for regression analysis in complex datasets. Gupta and Chen (2020) [14] explored hybrid models combining traditional regression techniques with machine learning algorithms like neural networks and support vector machines. These models excelled in capturing non-linear relationships and residual dependencies, offering a promising alternative for datasets with significant assumption violations. Hastie, Tibshirani, and Friedman (2009) [16] further emphasized the potential of ensemble learning techniques, such as random forests, to enhance regression analysis. Their research demonstrated how combining multiple weak learners could improve model accuracy while addressing residual patterns, making them particularly relevant for forex market analysis.

The forex market's unique characteristics, including high volatility and sensitivity to macroeconomic events, have prompted extensive research on assumption violations. Cheung, Chinn, and Pascual (2005) [7] examined the empirical validity of exchange rate models, identifying significant multicollinearity and non-linearity in predictor variables. Their findings underscored the need for robust diagnostic and corrective methodologies, as traditional regression models often failed to capture the intricacies of forex data. Recent studies by Balogun and Onifade (2021) [2] have highlighted the effectiveness of transformative and hybrid models in forex regression analysis. Their application of FD and LD transformations significantly improved model diagnostics, reducing heteroscedasticity and autocorrelation while maintaining interpretability. Similarly, Sharma and Kumar (2022) [23] emphasized the role of machine learning in complementing econometric models, demonstrating superior performance in capturing complex relationships among currency pairs.

Studies such as those by Engel and West (2005) [9] have explored the predictive power of regression models in forex markets, emphasizing the importance of accurately capturing the relationships between macroeconomic indicators and exchange rates. However, these models often face diagnostic challenges due to the non-linear nature of forex data and the presence of ARCH effects, as highlighted by Bollerslev (1986) [3]. Research by Meese and Rogoff (1983) [19] further demonstrated the limitations of traditional linear regression models in predicting exchange rates, suggesting that models must account for volatility clustering and non-linearity. Recent studies have employed transformation-based methods, such as first differences and log differences, to improve model performance. For instance, Wang *et al*. (2021) [25] found that applying log transformations significantly reduced heteroscedasticity and improved the interpretability of regression models in currency studies.

Long-term datasets, such as the 20-year period analysed in this study, present unique challenges. Campbell *et al*. (1997) [6] noted that extended time horizons increase the likelihood of structural breaks, non-stationarity, and evolving relationships between variables. These factors necessitate rigorous preprocessing and the application of advanced diagnostic techniques to ensure regression validity. The inclusion of emerging market currencies further complicates the analysis, as these currencies often exhibit higher volatility and are more susceptible to external shocks, as discussed by Reinhart and Rogoff (2009) [22]. The literature reveals a clear trend towards integrating traditional econometric techniques with modern computational tools. While classical models like OLS remain foundational, their limitations in handling assumption violations have spurred the adoption of advanced methods. Researchers such as Taylor and Allen (1992) [24] advocate for hybrid approaches that combine the interpretability of traditional models with the flexibility of machine learning, paving the way for more robust and assumption-compliant regression analyses.

The reviewed literature underscores the challenges of regression analysis in financial data, particularly in the context of forex markets. Common assumption violations such as multicollinearity, heteroscedasticity, and autocorrelation demand tailored solutions to preserve model validity. Transformative approaches, including log and first differences, emerge as effective strategies for mitigating these issues, especially in long-term analyses. However, the complexity of currency pair behaviour and evolving economic dynamics highlight the need for further exploration of advanced methodologies. This study contributes to this body of work by systematically addressing assumption violations using corrective strategies and evaluating their efficacy in modelling 20 years of currency pair data.

3. Methodology

The methodology of this study is cantered on the rigorous selection and analysis of nine prominent currency pairs over a period of 20 years, from 2004 to 2023. The currency pairs were chosen to represent a diverse range of global economies and include major, minor, and emerging market currencies, offering a balanced perspective on international foreign exchange dynamics. The pairs selected-USD/EUR, USD/GBP, USD/JPY, USD/AUD, USD/CAD, USD/CHF, USD/ZAR, USD/BRL, and USD/MXN-capture key relationships between the U.S. dollar and major trading partners as well as high-volatility currencies from developing economies. This diversity ensures the data reflects varied economic conditions, trade policies, and geopolitical factors that influence exchange rates over the study period.

The time frame of 2004 to 2023 was chosen to capture long-term trends and fluctuations in exchange rate behaviour across different economic cycles, including pre- and post-global financial crisis periods, the COVID-19 pandemic, and subsequent recovery phases. Monthly price data was selected to balance granularity and manageability, allowing for detailed analysis while avoiding excessive noise that daily or high-frequency data might introduce. This temporal resolution also aligns with many macroeconomic variables and policy updates, making the dataset suitable for regression analysis focused on economic relationships.

It is assumed that both the data and the residuals follow a normal distribution. The dataset was sourced from reliable financial and economic databases to ensure consistency and accuracy. Currency pairs were analysed using Ordinary Least Squares (OLS) regression, incorporating both raw and transformed variables to address potential assumption violations. Preprocessing steps included cleaning missing values, normalizing data where necessary, and conducting exploratory analysis to identify patterns, trends, and potential outliers. This systematic approach provided a robust foundation for evaluating regression assumptions and testing corrective strategies, ensuring the findings are both methodologically sound and applicable to real-world financial modelling.

3.1. Descriptive statistics

Descriptive statistics of USD/EUR, USD/GBP, USD/JPY, USD/AUD, USD/CAD, USD/CHF, USD/ZAR, USD/BRL, and USD/MXN provide key metrics like average exchange rates (Mean) and typical values (Median). Standard deviation reveals each pair's volatility, while range highlights their highest and lowest rates. Skewness and kurtosis indicate the distribution shape, aiding in understanding market dynamics. These insights help in assessing trends, risks, and trading opportunities in forex markets.

3.2. Ordinary Least Squares (OLS)

The Ordinary Least Squares (OLS) test in multiple regression estimates relationships between one dependent variable and multiple independent variables. The formula is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

Here, Y is the dependent variable, X_1, X_2……., X_n are independent variables, β_0 is the intercept, β_1, β_2…, β_n are coefficients, and ϵ is the error term. OLS minimizes the sum of squared residuals (ϵ^2) to estimate β values. Assumptions like linearity, no multicollinearity, and homoscedasticity are crucial for valid results. This test is essential for analysing the combined effect of multiple predictors on an outcome.

3.3. Weighted Least Squares (WLS)

The Weighted Least Squares (WLS) method is a regression technique that addresses heteroscedasticity by assigning weights to observations based on their variance. The formula is:

$$\min = \sum_{i=1}^{n} w_i (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2$$

Here, wi represents weights inversely proportional to the variance of the error term, and $y_i, x_{i1} \ldots, x_{ik}$ are the dependent and independent variables. WLS minimizes weighted residuals to provide unbiased, efficient estimates when error variances are unequal. This method is widely used in cases where homoscedasticity assumptions are violated.

3.4. Heteroscedasticity-Corrected Model (HSC)

A Heteroscedasticity-Corrected Model adjusts regression analyses to account for non-constant variance (Heteroscedasticity) in the error terms, ensuring reliable estimates and valid statistical inference. The model corrects standard errors, often using robust techniques such as White's correction. The corrected regression equation remains:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

However, heteroscedasticity-adjusted standard errors are computed as:

$$\hat{V} = (X'X)^{-1} X' \hat{\Omega} X (X'X)^{-1}$$

where is $\hat{\Omega}$ a diagonal matrix of error variances. This approach ensures unbiased coefficient estimates and accurate confidence intervals in the presence of heteroscedasticity.

3.5. Adjusted R-squared

The Adjusted R-squared adjusts the R-squared value for the number of predictors in a regression model, providing a more accurate measure of goodness-of-fit, especially with multiple predictors. The formula is:

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1}$$

where R^2 is the R-squared value, n is the number of observations, and pp is the number of predictors. Unlike R-squared, the Adjusted R-squared penalizes unnecessary variables, preventing overfitting and giving a more reliable evaluation of model performance.

3.6. Standard Error (SE)

Standard Error (SE) measures the precision of a sample statistic, such as the mean, relative to the population parameter. It is calculated as:

$$SE = \frac{\sigma}{\sqrt{n}}$$

where σ is the population standard deviation and n is the sample size. A smaller SE indicates greater accuracy of the sample estimate, making it critical in hypothesis testing and confidence interval calculation.

3.7. Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is used to evaluate and compare the goodness of fit of statistical models, balancing model complexity and fit. The formula for AIC is:

$$AIC = 2k - 2\ln(L)$$

where k is the number of parameters in the model, and L is the likelihood of the model. A lower AIC value indicates a better-fitting model, while penalizing excessive complexity. It is widely used in model selection, especially when comparing models with different numbers of parameters.

3.8. First Difference (FD) Method

The First Difference Method is used in regression analysis to address issues like non-stationarity and omitted variable bias by analysing changes between consecutive observations. It transforms the data by computing differences, making the model:

$$\Delta Y_t = \beta \Delta X_t + \Delta \epsilon_t$$

where $\Delta Y_t = Y_t - Y_{t-1}$ and $\Delta X_t = X_t - X_{t-1}$. This method eliminates time-invariant unobserved effects, focusing on the variation within the data. It is commonly applied in time-series and panel data analysis.

3.9. Log Difference (LD) Method

The Log Difference Method is used in regression analysis to measure percentage changes or growth rates, often in financial or economic data. It involves taking the natural logarithm of variables and computing their differences. The formula is:

$$\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$$

This approximates the proportional change in Y*Y* over time. Log differences stabilize variance and allow interpretation of coefficients as elasticities, making them useful in analyzing trends and growth rates.

3.10. Variance Inflation Factor (VIF)

The Variance Inflation Factor (VIF) is used to detect multicollinearity in regression models by measuring how much the variance of a regression coefficient is inflated due to correlation with other predictors. The formula for VIF is:

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

where R_i^2 is the coefficient of determination obtained by regressing the i-th predictor on all other predictors. A high VIF (typically > 10) indicates significant multicollinearity, which may distort the regression results and reduce the reliability of the coefficients.

3.11. Breusch-Pagan (BP) Test

The Breusch-Pagan (BP) Test detects heteroscedasticity in regression models by assessing whether error variances depend on independent variables. It involves regressing the squared residuals ($\hat{\varepsilon}^2$) on the predictors:

$$\hat{\varepsilon}^2 = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_k X_k + u$$

The test statistic is:

$$BP = \frac{1}{2} R_{aux}^2 n$$

where R_{aux}^2 is the coefficient of determination from the auxiliary regression. The BP statistic follows a chi-squared distribution, with higher values indicating heteroscedasticity.

3.12. Brock-Dechert-Scheinkman (BDS) Test

The Brock-Dechert-Scheinkman (BDS) Test assesses non-linearity or dependence in time-series data by examining deviations from randomness. It compares the correlation of points in reconstructed phase space at varying dimensions. The test statistic is:

$$W = \frac{\sqrt{n}\,(C_m(\varepsilon) - C_1^m(\varepsilon))}{\sigma_m(\varepsilon)}$$

where $C_m(\varepsilon)$ is the correlation integral for dimension m, $C_1^m(\varepsilon)$ is the product of one-dimensional correlation integrals, and $\sigma_m(\varepsilon)$ is the standard deviation. A significant result indicates non-linear structure, making the test vital for analysing chaotic or complex systems.

3.13. Durbin-Watson (DW) Test

The Durbin-Watson (DW) Test checks for autocorrelation in the residuals of a regression model, particularly for first-order correlation. The test statistic is:

$$DW = \frac{\sum_{t=2}^{n} (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}$$

where $\hat{\varepsilon}_t$ are the residuals at time t. The DW statistic ranges from 0 to 4; a value near 2 indicates no autocorrelation, values < 2 suggest positive autocorrelation, and values > 2 indicate negative autocorrelation. This test is critical for ensuring the validity of regression assumptions in time-series data.

3.14. Lagrange Multiplier (LM) Test

The Lagrange Multiplier (LM) Test for autocorrelation detects serial correlation in residuals of a regression model. It involves regressing residuals ($\hat{\varepsilon}_t$) on lagged residuals and independent variables. The auxiliary regression is:

$\hat{\varepsilon}_t = \alpha_0 + \alpha_1 \hat{\varepsilon}_{t-1} + \alpha_2 \hat{\varepsilon}_{t-2} + \cdots + \alpha_p \hat{\varepsilon}_{t-p} + ut$

The test statistic is:

$LM = nR^2$

where n is the sample size, and R^2 is the auxiliary regression's determination coefficient. The LM statistic follows a chi-squared distribution, with significance indicating autocorrelation.

3.15. Lagrange Multiplier (LM) Test for ARCH Effect

The Lagrange Multiplier (LM) Test for ARCH Effect identifies autoregressive conditional heteroscedasticity (ARCH) in time-series data. It involves regressing squared residuals ($\hat{\varepsilon}_t$) on their lagged values. The auxiliary regression is:

$\hat{\varepsilon}_t = \alpha_0 + \alpha_1 \hat{\varepsilon}_{t-1} + \alpha_2 \hat{\varepsilon}_{t-2} + \cdots + \alpha_p \hat{\varepsilon}_{t-p} + ut$

The test statistic is:

$LM = nR^2$

where n is the sample size, and R^2 is from the auxiliary regression. A significant LM statistic indicates ARCH effects, essential for volatility modelling.

4. Discussion

Table 1: Descriptive statistics of international currency pairs monthly prices for the period of 20 years (N=240)

Variable	Mean	Median	Minimum	Maximum	Std. Dev.	C.V.	Skewness	Ex. kurtosis
USDINR	58.8370	60.5500	39.1950	83.3570	13.1890	0.2242	0.1821	-1.3098
USDGBP	0.6640	0.6499	0.4804	0.8957	0.1011	0.1522	0.0345	-1.0916
USDEUR	0.8161	0.8177	0.6341	1.0201	0.0832	0.1019	-0.0161	-0.8078
USDCHF	1.0204	0.9814	0.7852	1.3151	0.1251	0.1226	0.8877	-0.3746
USDJPY	107.5300	108.8100	76.1900	151.6700	15.4800	0.1440	0.1552	0.4423
USDAUD	1.2670	1.3057	0.9089	1.6295	0.1720	0.1357	-0.3770	-0.7898
USDBRL	3.0802	2.6800	1.5490	5.7446	1.2620	0.4097	0.6781	-0.8674
USDZAR	11.0500	10.4470	5.6651	19.7250	3.9414	0.3567	0.3498	-1.2273
USDMXN	15.2370	13.4760	10.0350	24.1510	3.7706	0.2475	0.3253	-1.3536

(***Source*:** Statistical calculations)

Table 1 provide insights into their monthly price behaviour over a 20-year period the descriptive statistics of the nine international currency pairs. USDINR exhibits a relatively high mean (58.837) and standard deviation (13.189), indicating significant variability compared to its peers. Its coefficient of variation (C.V.) is moderate (0.2242), reflecting relative stability given its high absolute values. The skewness (0.1821) indicates near symmetry, while the negative excess kurtosis (-1.3098) suggests a flatter distribution compared to a normal curve. Similarly, USDGBP and USDEUR display low means (0.6640 and 0.8161, respectively) and lower C.V. values (0.1522 and 0.1019), indicating more stability. Their skewness values close to zero and negative excess kurtosis (e.g., USDGBP: -1.0916) point to symmetric and relatively flat distributions.

USDCHF and USDJPY, representing European and Asian markets, show distinct characteristics. USDCHF has a mean of 1.0204 and a relatively low C.V. (0.1226), indicating high consistency in exchange rates. However, its positive skewness (0.8877) reveals a right-tailed distribution, meaning occasional price spikes. USDJPY demonstrates a higher mean (107.530) and standard deviation (15.480), reflecting more variability. With low skewness (0.1552) and slight excess kurtosis (0.4423), it suggests moderately stable pricing with a slight tendency for extreme values.

Emerging market currencies such as USDBRL, USDZAR, and USDMXN show higher volatility compared to developed market currencies. USDBRL's C.V. (0.4097) and positive skewness (0.6781) indicate significant variability and a tendency for upward price movements, likely due to economic instability. Similarly, USDZAR and USDMXN display high C.V. values (0.3567 and 0.2475, respectively) and notable negative excess kurtosis (e.g., USDZAR: -1.2273), suggesting frequent but moderate price fluctuations without extreme outliers. These patterns highlight the greater risk associated with emerging market currencies and the relative stability of developed market pairs.

The table 2 compares multiple regression models based on Ordinary Least Squares (OLS) Model 1 and Ordinary Least Squares Model 2 (by removing excess VIF variables). The first model, using Ordinary Least Squares (OLS), demonstrates clear violations of regression assumptions, particularly regarding multicollinearity. High collinearity metrics, such as USDGBP (VIF: 10.884) and USDZAR (VIF: 28.119), suggest significant overlap among predictor variables. This collinearity inflates standard errors and undermines the reliability of coefficient estimates, making it difficult to isolate the effects of individual predictors. Although the Adjusted R-squared is high (0.9673), this metric can be misleading in the presence of multicollinearity. Additionally, the Durbin-Watson statistic of 0.2892 indicates severe autocorrelation, while the Breusch-Pagan (BP) test (81.0437, $p < 0.0001$) reveals heteroscedasticity.

Table 2: Comparison of OLS Model 1 and OLS Model 2 Regression results of Currency pairs

Particulars	Ordinary Least Square (OLS) Model 1			OLS Model 2		
	Coefficient	p-value	Collinearity VIF	Coefficient	p-value	Collinearity VIF
Constant	29.0658*	0.0001	-	31.5348*	0.0001	-
USDGBP	17.6148*	0.0005	10.884#	27.6339*	0.0001	7.781
USDEUR	23.1455*	0.0001	5.653	24.3786*	0.0001	5.633
USDCHF	−26.7002*	0.0001	9.681	−43.0856*	0.0001	4.937
USDJPY	0.0744*	0.0008	4.863	0.1416*	0.0001	2.815
USDAUD	−0.7081	0.7955	9.302	6.6515*	0.0123	7.524
USDBRL	2.3919*	0.0001	8.264	3.0413*	0.0001	5.362
USDZAR	1.2317*	0.0001	28.119#	-	-	-
USDMXN	−0.1079	0.4729	13.536#	-	-	-
Standard Error	2.3840	-	-	2.5618	-	-
Adjusted R-squared	0.9673	-	-	0.9623	-	-
F Stat	889.2066*	0.0000	-	1021.2070*	0.0000	-
Akaike Criterion	1111.5020	-	-	1144.2470	-	-
Durbin-Watson	0.2892	-	-	0.3115	-	-
BDS Test Non-Linearity	24.084*	0.0000		20.566*	0.0000	
BP Test HSD	81.0437*	0.0000	-	127.2610*	0.0000	-
LM Test ACR	51.6379*	0.0000	-	48.6041*	0.0000	-
LM Test ARCH Effect	120.4450*	0.0000	-	127.6420*	0.0000	-

(***Source***: Statistical calculations) (* 5 percent level of significance) (# >10 VIF values)

The LM test for ARCH effects confirms volatility clustering, pointing to significant deviations from homoscedastic residuals, which further compromise the model's validity. The BDS test for non-linearity indicates significant non-linear dependencies in both models, with test statistics of 24.084 (p< 0.0001) for the OLS model and 20.566 (p< 0.0001) for the Collinearity Adjusted OLS. These results suggest that linear regression may not fully capture the underlying structure of the data.

To address multicollinearity, the second model applies OLS by removing high VIF variables, effectively reducing collinearity levels. For instance, VIF for USDGBP drops to 7.781 and for USDCHF to 4.937, improving the reliability of coefficient estimates. This adjustment makes previously insignificant variables, like USDAUD, statistically significant (coefficient: 6.6515, p = 0.0123), suggesting a more accurate representation of predictor relationships. However, some variables, such as USDZAR and USDMXN, are excluded, likely due to their excessive collinearity or limited contribution to the model. While the standard error increases slightly from 2.3840 to 2.5618, the coefficients gain interpretability. The Akaike Information Criterion (AIC) rises from 1111.5020 to 1144.2470, reflecting the complexity added by the adjustments and the potential trade-off between simplicity and precision.

Despite the improvements, both models exhibit persistent issues with heteroscedasticity, autocorrelation, and ARCH effects. The BP test statistic increases to 127.2610 in the second model, and the Durbin-Watson statistic remains low at 0.3115, indicating residual autocorrelation is unresolved. Similarly, the LM test for ARCH effects shows significant volatility clustering in both models (120.4450 for OLS 1, 127.6420 for OLS 2). These findings suggest that while collinearity adjustments enhance coefficient reliability, neither model fully addresses heteroscedasticity or autocorrelation. Further refinements, such as robust standard errors, generalized least squares, or time-series models, may be necessary to ensure compliance with regression assumptions and improve overall model performance.

Table 3: Comparison of Ordinary Least Square (OLS) added with First Difference (FD) & Log difference (LD) and WLS & HSC

Particulars	OLS	OLS FD	OLS LD	WLS	HSC	HSC FD	HSC LD
Constant	31.5348*	0.1498*	0.0024*	32.4897*	34.0579*	0.1277*	0.0021*
USDGBP	27.6339*	1.8526	−0.0057	28.2166*	28.9165*	2.3786	0.0302
USDEUR	24.3786*	0.6082	0.0198	22.6988*	24.7798*	2.8115	0.0089
USDCHF	−43.0856*	4.9252	0.0929	−42.8355*	−43.7875*	2.5499	0.0847
USDJPY	0.1416*	−0.0148	−0.0328	0.1384*	0.1319*	−0.0066	−0.0338
USDAUD	6.6515*	9.4733*	0.2095*	6.6902*	6.1196*	9.4368*	0.2117*
USDBRL	3.0413*	1.1673*	0.1062*	3.0667*	2.7434*	0.7711	0.0693*
Standard Error	2.5618	1.0063	0.0171$	2.2469	1.7968	2.1311	2.1507
Adjusted R-squared	0.9623	0.2821	0.3526	0.9648	0.9879	0.2192	0.3098
F Stat	1021.2070*	16.6525*	22.6923*	1097.7380*	3256.7360*	12.1828*	18.8809*
Akaike Criterion	1144.247	690.9785	−1265.575$	1081.038	973.2801	1051.183	1055.557
Durbin-Watson (DW)	0.3115	2.0786	1.9877	0.3116	0.2982	2.0609	1.9566
BDS Test Non-Linearity	20.566* (0.000)	1.9333 (0.074)	1.138 (0.288)	20.430* (0.000)	20.671* (0.000)	1.524 (0.155)	1.038 (0.329)
BP Test HS	127.26* (0.000)	10.205 (0.116)	8.964 (0.175)	-	-	-	-
LM Test AC (12 Lag)	48.604* (0.000)	0.9446 (0.503)	0.728 (0.722)	-	-	-	-
LM Test ARCH	127.64* (0.000)	14.652 (0.261)	13.21 (0.354)	128.88* (0.000)	136.83* (0.000)	15.354 (0.223)	15.183 (0.231)

(Source: Statistical calculations)
(* 5 percent level of significance)
(Probabilities in parenthesis)
($ Least Values)

The table 3 compares various regression models, including Ordinary Least Squares (OLS), First Difference (FD), Log Difference (LD), Weighted Least Squares (WLS), Heteroscedasticity-Corrected (HSC), and HSC FD & LD approaches, focusing on their ability to address key regression assumptions such as heteroscedasticity, autocorrelation, ARCH effects, and non-linearity. The standard OLS model performs poorly in diagnostic tests. The Breusch-Pagan (BP) test (127.261, $p < 0.0001$) confirms significant heteroscedasticity, while the Durbin-Watson statistic (0.3115) reveals severe autocorrelation. The LM test for ARCH effects (127.642, $p < 0.0001$) indicates volatility clustering, and the BDS test for non-linearity (20.566, $p < 0.0001$) highlights significant non-linear dependencies. Despite a high Adjusted R-squared (0.9623), the model's AIC (1144.247) and relatively large standard error (SE) of 2.5618 suggest inefficiencies and an inability to handle the data's complexity effectively.

The First Difference (OLS FD) and Log Difference (OLS LD) models substantially improve on the limitations of OLS. The BP test results for OLS FD (10.2053, p = 0.1163) and OLS LD (8.9642, p = 0.1756) indicate reduced heteroscedasticity, while non-significant LM tests for autocorrelation and ARCH effects show that these models effectively address residual dependencies and volatility clustering. The BDS test results further confirm that both transformations mitigate non-linear dependencies, with non-significant outcomes for OLS FD (1.9333, p = 0.074) and OLS LD (1.138, p = 0.288). However, the Adjusted R-squared values drop significantly (0.2821 for OLS FD and 0.3526 for OLS LD), reflecting reduced explanatory power. The OLS LD model achieves the lowest SE (0.0171) and a dramatically improved AIC (−1265.575), making it a strong candidate for precision-focused analyses.

The Weighted Lease Square (WLS) model improves upon standard OLS by addressing heteroscedasticity through weighted adjustments, achieving a slightly better Adjusted R-squared (0.9648) and a reduced SE (2.2469). The AIC also improves to 1081.038, indicating better model efficiency. However, the Durbin-Watson statistic (0.3116) suggests persistent autocorrelation, and the BDS test (20.430, $p < 0.0001$) highlights significant non-linear dependencies, indicating that WLS fails to adequately capture the complexity of the dataset. The HSC model, designed to handle heteroscedasticity, performs well in addressing volatility clustering, achieving the highest Adjusted R-squared (0.9879) and a reduced AIC (973.2801). Despite these strengths, significant LM test results for ARCH effects (136.833, $p < 0.0001$) and the low Durbin-Watson statistic (0.2982) indicate persistent autocorrelation. Furthermore, the BDS test (20.671, $p < 0.0001$) reveals unresolved non-linear dependencies, limiting the model's robustness despite its strong overall fit.

The HSC FD and HSC LD models combine heteroscedasticity corrections with FD and LD transformations, further improving diagnostics. Both models exhibit non-significant LM tests for ARCH effects and better handling of non-linear dependencies, with non-significant BDS test results (e.g., HSC LD: 1.038, p = 0.329). While these transformations improve compliance with regression assumptions, their Adjusted R-squared values drop (e.g., 0.2192 for HSC FD and 0.3098 for HSC LD), suggesting a trade-off between explanatory power and diagnostic accuracy. The AIC values for these models (1051.1830 for HSC FD and 1055.5570 for HSC LD) indicate they are less efficient than OLS LD but perform better than other models.

The significance of coefficients and their impact on the dependent variable vary across models, reflecting differences in how each handles assumption. In the OLS model, all variables are significant, with USDGBP, USDEUR, and USDBRL positively impacting the dependent variable, while USDCHF shows a substantial negative effect. In OLS FD and OLS LD models, the significance of variables reduces as transformations prioritize reducing heteroscedasticity and autocorrelation over capturing direct impacts. For instance, in OLS LD, coefficients like USDGBP and USDCHF lose significance, reflecting the transformation's focus on stabilizing variance. The WLS model maintains significance for most variables while slightly adjusting the magnitude of their impacts, with USDCHF still exerting a strong negative influence. The HSC model improves the significance and stability of coefficients, showing consistent impacts of key variables while addressing heteroscedasticity. In HSC FD and HSC LD models, the transformations further stabilize variance but lead to reduced significance for variables such as USDGBP and USDEUR, indicating that these approaches prioritize assumption compliance at the expense of variable-specific impacts.

Considering all approaches, the OLS LD (Ordinary Least Square Log Difference) model emerges as the most suitable option. It effectively addresses heteroscedasticity, autocorrelation, ARCH effects, and non-linearity, as demonstrated by its diagnostic test results and non-significant BDS test (1.138, p = 0.288). The lowest SE (0.0171) and dramatically reduced AIC (−1265.575) highlight its precision and efficiency. While its Adjusted R-squared (0.3526) is lower compared to other models, its compliance with regression assumptions and ability to capture complex relationships make it the best choice for accurate and assumption-compliant regression analysis.

5. Conclusion

The analysis highlights the comparative strengths and limitations of various regression approaches in addressing key assumptions of residuals heteroscedasticity, autocorrelation, ARCH effects, and non-linearity. The Ordinary Least Squares (OLS) model one, despite its high Adjusted R-squared, is fundamentally flawed due to severe multicollinearity, heteroscedasticity, autocorrelation, and non-linearity as evidenced by high VIF values, a low Durbin-Watson statistic, and significant diagnostic test results. These shortcomings compromise the reliability and interpretability of coefficient estimates, making OLS unsuitable for datasets with such violations. While the OLS model two mitigates multicollinearity by removing high-VIF variables, it does not fully resolve other critical issues like residual heteroscedasticity, autocorrelation, and nonlinearity, suggesting the need for more advanced techniques.

The transformation-based models, specifically OLS FD (First Difference) and OLS LD (Log Difference), demonstrate substantial improvements by reducing residuals heteroscedasticity, autocorrelation, and ARCH effects. Among these, OLS LD stands out for its precision and efficiency, achieving the lowest standard error and dramatically improved AIC, despite a lower Adjusted R-squared. This trade-off indicates that while explanatory power is slightly sacrificed, the model provides a more robust framework for analysing complex data patterns. Similarly, the Heteroscedasticity-Corrected (HSC) and its transformations (HSC FD and HSC LD) address some diagnostic issues, particularly volatility clustering and nonlinearity. However, these models still face challenges with residual autocorrelation and exhibit reduced Adjusted R-squared values, limiting their overall effectiveness.

Overall, the OLS LD model emerges as the most suitable regression approach with less Standard Error (SE) and Akaia Criterion (AIC) for this analysis. It achieves a balance between addressing key assumptions and maintaining model efficiency, as demonstrated by its superior diagnostic test outcomes and compliance with regression assumptions. While models like HSC offer higher Adjusted R-squared values, their inability to fully mitigate issues such as autocorrelation and their complexity make them less practical. The findings underscore the importance of selecting models that prioritize assumption compliance and precision over raw explanatory power, ensuring more reliable and interpretable results in regression analysis.

6. Scope for Further Research

Future research could build on this study by exploring advanced time-series models like ARIMA or GARCH, which are specifically designed to handle autocorrelation and heteroscedasticity in financial data. Additionally, machine learning approaches, such as neural networks or random forests, could be employed to capture non-linear relationships and improve model accuracy without strict adherence to regression assumptions. Exploring alternative data transformations, like seasonal adjustments or detrending, could further address assumption violations. Expanding the scope to include a broader set of currency pairs or macroeconomic variables, along with real-time data analysis and high-frequency trading models, could provide deeper insights into short-term dynamics and market drivers. Finally, research on the robustness of models to structural breaks and regime shifts, alongside enhancing the interpretability of complex models, would contribute to developing more reliable and practical tools for financial market analysis.

7. References

1. Asteriou D, Hall SG. Applied econometrics. 4th ed. Red Globe Press; c2021.
2. Balogun MA, Onifade A. Transformative models and diagnostics in forex regression analysis: Addressing assumption violations. Journal of Financial Econometrics. 2021;19(4):456-472.
3. Bollerslev T. Generalized autoregressive conditional heteroskedasticity. Journal of Econometrics. 1986;31(3):307-327.
4. Breusch TS, Pagan AR. A simple test for heteroscedasticity and random coefficient variation. Econometrica. 1979;47(5):1287-1294.
5. Box GEP, Jenkins GM. Time series analysis: Forecasting and control. Revised ed. Holden-Day; c1976.
6. Campbell JY, Lo AW, MacKinlay AC. The econometrics of financial markets. Princeton University Press; c1997.
7. Cheung YW, Chinn MD, Pascual AG. Empirical exchange rate models of the nineties: Are any fit to survive? Journal of International Money and Finance. 2005;24(7):1150-1175.
8. Durbin J, Watson GS. Testing for serial correlation in least squares regression. Biometrika. 1950;37(3-4):409-428.
9. Engel C, West KD. Exchange rates and fundamentals. Journal of Political Economy. 2005;113(3):485-517.
10. Engle RF. Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation. Econometrica. 1982;50(4):987-1007.
11. Escribano A, Mira S. Nonlinear cointegration and nonlinear error correction models. Journal of Time Series Analysis. 2002;23(5):509-522.
12. Greene W. Econometric analysis. 8th ed. Pearson Education Limited; c2018.
13. Gujarati DN, Porter DC. Basic econometrics. 5th ed. McGraw-Hill; c2009.
14. Gupta A, Chen J. Hybrid regression models for financial forecasting: A machine learning perspective. International Journal of Finance and Economics. 2020;25(4):565-580.
15. Hansen H, Johansen S. Some tests for parameter constancy in cointegrated VAR-models. Econometrics Journal. 1999;2(2):306-333.
16. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. Springer; c2009.
17. Katrakilidis C, Trachanas E. Exchange rate volatility and fundamentals: Evidence from emerging markets. Emerging Markets Review. 2012;13(3):337-344.
18. Mason CH, Perreault WD. Collinearity, power, and interpretation of multiple regression analysis. Journal of Marketing Research. 1991;28(3):268-280.
19. Meese RA, Rogoff K. Empirical exchange rate models of the seventies: Do they fit out of sample? Journal of International Economics. 1983;14(1-2):3-24.
20. Osborne JW, Waters E. Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research, and Evaluation. 2002;8(1):1-5.
21. Pindyck RS, Rubinfeld DL. Econometric models and economic forecasts. 4th ed. McGraw-Hill; c1998.
22. Reinhart CM, Rogoff KS. This time is different: Eight centuries of financial folly. Princeton University Press; c2009.
23. Sharma R, Kumar V. Machine learning meets econometrics: An application in exchange rate prediction. Journal of Financial Research. 2022;45(3):325-345.
24. Taylor MP, Allen H. The use of technical analysis in the foreign exchange market. Journal of International Money and Finance. 1992;11(3):304-314.
25. Wang X, Zhang Y, Liu J. Log transformations in forex regression: A pathway to assumption compliance. Applied Financial Economics. 2021;31(4):385-398.
26. White H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica. 1980;48(4):817-838.
27. Williams RA, Miles MP, Rosenfeld D. Regression assumptions in financial research. Accounting and Finance. 2013;53(1):1-25.