EJE | EUROPEAN JOURNAL OF ECONOMICS

ISSN: 2669-2384
VOLUME 5, ISSUE 2

Unveiling the Pitfalls of Endogeneity in Forecasting: An Insight from Monte Carlo Simulation

Hifsa Mobeen

Lecturer, Department of Economics, National University of Modern Languages, Mirpur AJK Campus, Pakistan

Hifsa.mobeen@numl.edu.pk

https://orcid.org/0000-0001-7370-033X

ABSTRACT

This paper investigates how endogeneity impairs the forecasting performance of cointegration-adjusted regression models. We employ a Monte Carlo simulation framework to generate data under three scenarios. In this regard, we have estimated three model variants: a correctly specified benchmark model (M_1), a model with the issue of endogeneity (M_2), and a endogeneity corrected model (M_3). Endogeneity is addressed using simulated instrumental variables that satisfy the relevance, exogeneity and exclusion restriction conditions. Forecast accuracy is evaluated across three models using various error metrics. The results reveal that models subject to endogeneity yield systematically poorer forecasts, while endogeneity correction through valid instruments significantly improves predictive performance of the model. The paper emphasizes that failure to address endogeneity in cointegration-adjusted regression models undermines structural validity, leading to biased forecasts and potentially flawed economic or financial decisions. Accurate model specification is thus essential for credible and policy-relevant forecasting.

Keywords: Endogeneity, Simulation, Cointegration Adjusted Regression Models, Forecasting

Cite this article as: Mobeen, H. (2025). Unveiling the Pitfalls of Endogeneity in Forecasting: An Insight from Monte Carlo Simulation. *European Journal of Economics*, 5(2), 78-105. https://doi.org/10.33422/eje.v5i2.1221

1. Introduction

Forecasting is not a precise science and can be quite wrong; yet, it is a necessary endeavor because the alternative is to do nothing and hope for the best. Cointegration in time series econometrics has several implications for empirical macroeconometrics, time series econometrics, and financial econometrics (Engle and Granger, 1987). Forecasts for certain economic or financial indicators are utilized as input in the decision-making process for a wide range of operations since the consequence of a decision made today depends on something that will happen in the future. This assumption holds for activities necessitating forecasting for various financial and economic data.

Cointegration analysis is a useful tool for identifying meaningful linkages that go beyond simple correlation, providing insights into the dynamics of economic systems. When these insights are appropriately utilized, they serve as the foundation for solid forecasting models, policy assessments, and informed decision-making processes.

The assumptions used to develop a model reflect the quality of econometric practice. Three conventional assumptions are commonly practiced regarding any linear econometric model.

i. Greene (2012) interpreted error term ε_t as the net effect of omitted regressors on regressand y_t.

ii. $E(y_t/x_t)$ is finite.

iii. $E(\varepsilon_t/x_t) = 0$.

Econometricians and statisticians interpret assumption (iii) differently. Greene (2012) demonstrated that x_t is determined outside of the system i.e. x_t is exogenous. Engle et al. (1983) posit to treat x_t exogenous corresponding to the purpose of the inference we desire to conduct. Friedman and Schwartz (1982) suggest that x_t can be regarded as exogenous in some cases and endogenous in some other cases. However, Pratt and Schlaifer (1988) strongly claimed that it is meaningless to accept the orthogonality of x_t and ε_t, if ε_t hold relevant omitted regressors. Similarly, it is futile to accept the orthogonality of x_t and ε_t in a model having more than two variables i.e. a multivariate model.

In this regard, two thoughtful propositions have been forwarded by Pratt and Schlaifer (1984) and Pratt and Schlaifer (1988). According to the first proposition, the assumption that covariates are orthogonal to omitted variables that are present in error terms is meaningless. The second proposition states that even if we incorporate all the relevant variables in the model, the coefficient of covariates does not provide the true effect of covariates on the outcome variable. This is due to the reason that there always exist indirect effects of included covariates on outcome variables that cannot be ignored. Because the total effect of the covariate on the outcome variable comprises direct and indirect effects. These two assertions jointly suggest that the included covariates cannot be uncorrelated with the error term in the regression model. In the presence of omitted variable, the linear combination of two or more *I(1)* variables will no longer be stationary (Pratt & Schlaifer, 1984; Pratt & Schlaifer, 1988). Therefore, there won't exist cointegration among underlying non-stationary variables (Swamy & Muehlen, 2020). This provokes an econometric issue called endogeneity (Pedroni, 2007; Herzer & Vollmer, 2012; Swamy et al., 2017).

Now from these propositions, two points originate. First, we don't know by any means whether these omitted variables that are present in the error term are stationary or not. If in case these omitted variables are non-stationary then the stationary linear combination of x_t and y_t stated by Pedroni (2019) and Greene (2012) is questionable. Second, if these omitted variables are correlated with x_t then there arises a problem of endogeneity. Here, even though if we consider omitted variables stationary due to correlation with x_t the estimated cointegration relationships will be spurious. If these omitted variables are non-stationary and also correlate with the included covariate x_t then the problem doubles. On one side non-stationary linear combination and on the other side issue of endogeneity. In the presence of endogeneity, there won't exist cointegration among variables in reality. This endogeneity issue renders a spurious cointegrating relationship (Swamy & Muehlen, 2020). So, we have to consider these econometric problems in regression model before estimating the cointegration relationship and (iii) becomes void, if the error term and coefficients of a regression model are not unique.

The violation of exogeneity assumption and non-uniqueness of coefficients of a regression model due to omitted variable is illustrated by model adopted from Clarke (2005) and Hunushek and Jackson (1977). Suppose the true multivariate regression model is;

$$Y_t = \alpha_0 + \sum_{i=1}^{k} \beta_{it} X_{it} + \varepsilon_t \qquad \text{Where, } i = 1, \dots, k \tag{1}$$

Corresponding underspecified model, where we intentionally omit X_{dt} is;

$$Y_t = \alpha_0 + \sum_{i=2}^{k-m} \beta_{it} X_{it} + \varepsilon_t \qquad \text{Where, } i = 1, \dots, k-m \tag{2}$$

Writing equation (2) in matrix form;

$$\boldsymbol{y} = \boldsymbol{N\gamma} + \boldsymbol{X\beta} + \boldsymbol{\mu} \tag{3}$$

Equation (3) $\boldsymbol{y}$ represents $(T \times 1)$ vector of observations on regressand Y_t. $\boldsymbol{N}$ is a matrix holding endogenous regressors X_{k-m} and $\boldsymbol{\gamma}$ denote the vector of their corresponding coefficients. $\boldsymbol{X}$ is a matrix representing exogenous regressors and $\boldsymbol{\beta}$ is a vector of their corresponding coefficients. Whereas, $\boldsymbol{\mu}$ is the vector of error terms.

In conventional practice, a well-admitted assumption is that; error terms hold relevant omitted regressors that do not result in omitted variable bias in the model. Hence, the coefficients of included regressors in a model and error term are unique. However, coefficients of included regressors in a model and error term can be declared unique only if they remain unchanged by adding and subtracting the product of the omitted variable coefficient on the right-hand side of the model.

Let the t^{th} element of $\boldsymbol{y}$ is;

$$\boldsymbol{y_t} = \boldsymbol{\gamma' n_t} + \boldsymbol{\beta' x_t} + \boldsymbol{\mu_t} \tag{4}$$

Where $\boldsymbol{y_t} = (y_1, y_2, \dots, y_M)'$, $\boldsymbol{n_t}$ is transpose of t^{th} row $(n_1, n_2, \dots, n_M)$ of $\boldsymbol{N}$ and $\boldsymbol{\gamma'}$ is equal to $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \dots, \gamma_M)'$. $\boldsymbol{x_t}$ is transpose of t^{th} row $(x_1, x_2, \dots, x_k)$ of $\boldsymbol{X}$ and $\boldsymbol{\beta'}$ is the transpose of column vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_k)'$. Now suppose if $\boldsymbol{w_t}$ is the matrix of the omitted regressors and $\boldsymbol{\omega}$ is the vector of their corresponding coefficients. Then $\boldsymbol{\mu_t}$ can be written as $\boldsymbol{\mu_t} = \boldsymbol{\omega w_t}$ and equation (4) can be written as;

$$\boldsymbol{y_t} = \boldsymbol{\gamma' n_t} + \boldsymbol{\beta' x_t} + \boldsymbol{\omega' w_t} \tag{5}$$

Where, $\boldsymbol{w_t} = (w_1, w_2, \dots, w_L)'$ is column vector of observation on omitted regressors that error term holds and $\boldsymbol{\omega'} = (\omega_1, \omega_2, \dots, \omega_L)$ is a row vector of their corresponding coefficients. Pratt and Schlaifer (1984), Swamy et al. (2017), and Swamy and Muehlen (2020) claimed that relevant omitted regressors that constitute error terms are not unique. Hence, if a relevant variable is omitted from the model then the error term of the corresponding model cannot be unique and the coefficients of included regressors are inconsistent.

To prove the non-uniqueness of the coefficient and error term of a model, equation (5) can be written as;

$$y_t = \sum_{h=1}^{M} \gamma_h y_{th} + \sum_{k=1}^{K} \beta_k x_{tk} + \sum_{l=1}^{L} \omega_l w_{tl} \tag{6}$$

Suppose, one of the values of the subscript k is k' and λ is one of the values the subscript l can assume. Then $\omega_\lambda x_{tk'}$ is the corresponding product of one element from $\boldsymbol{\omega}$ and one element from x_t.

Adding and subtracting $\omega_\lambda x_{tk'}$ on R.H.S of equation (6) as;

$$y_t = \sum_{h=1}^{M} \gamma_h y_{th} + \sum_{\substack{k=1 \\ k \neq k'}}^{K} \beta_k x_{tk} + (\beta_{k'} + \omega_\lambda) x_{tk'} + \sum_{\substack{l=1 \\ l \neq \lambda}}^{L} \omega_l w_{tl} + \omega_\lambda (w_{t\lambda} - x_{tk'}) \tag{7}$$

From equation (6) to equation (7) it can be observed that coefficients β_k of equation (6) change from $\beta_{k'}$ to $(\beta_{k'} + \omega_\lambda)$ and the sum of omitted regressors $\sum_{l=1}^{L} \omega_l w_{tl}$ that made up the error term change from $\omega_\lambda w_{tk}$ to $\omega_\lambda (w_{t\lambda} - x_{tk'})$. As the coefficient β and omitted regressors w_t are unknown, it cannot be proved that the values $(\beta_{k'} + \omega_\lambda)$ and $(w_{t\lambda} - x_{tk'})$ are invalid. Hence, it can be authentically stated that the coefficient $\beta_{k'}$ and the term $\omega_\lambda w_{t\lambda}$ of $\sum_{l=1}^{L} \omega_l w_{tl}$ assuming two distinct expressions through equation (6) to equation (7) are not unique. It means that the exogeneity assumption $E(\varepsilon_t / x_t) = 0$ does not hold in practice.

After recognizing these drawbacks and their corresponding consequences, a struggle was created to use various methods, models and techniques that can cover these flaws to some extent. In the context of cointegration analysis, the endogeneity issue in econometric models has long been recognized in previous studies. However, its consequences for spurious cointegration relationships and corresponding forecasts are predominantly ignored. Methods such as Fully Modified Ordinary Least Square (FM-OLS) were introduced in early works by Phillips (1986) and Phillips & Hansen (1990) to correct for endogeneity in cointegrated systems. Phillips (1988) developed Dynamic Ordinary Least Square (DOLS) to address this issue by incorporating leads and lags of regressors in the regression model. Stock & Watson (1993) and Hansen (1992) acknowledged that traditional OLS estimates may be biased when endogenous regressors are present in the model. Therefore, the authors implemented the Instrumental Variables (IV) technique to address endogeneity. Saikkonen (1991) and Gregory & Hansen (1996) used structural breaks in the regression model to mitigate endogeneity and improve the precision of cointegration tests.

Nevertheless, these initial studies predominantly concentrated on enhancing estimation efficiency and did not explicitly investigate the direct impact of endogeneity on the forecasting outcomes, despite the fact that these studies have addressed the issue of endogeneity. Yet a single study empirically addressed the impact of endogeneity issue in the cointegration-adjusted regression model on the spurious cointegration relationships. Swamy and Muehlen (2020) recommended that to establish the true long-run relationships among non-stationary variables through the JJ test, we have to mend the issue of endogeneity in the underlying regression model. These authors proposed a Time-Varying Coefficient (TVC) model to solve the issue of endogeneity and to avoid the consequences of spurious cointegration relationships before applying the JJ cointegration test. However, this research paper fails to address the influence of endogeneity on forecasting outcomes.

Unlike prior studies in literature, which have mostly focused on improving estimates from their methods and models, this study systematically examines the impact of endogeneity on forecasting accuracy of the cointegration-adjusted regression models. By using the instrumental variable (IV) technique before performing the Johansen Juselius (JJ) test using cointegration-adjusted regression model, this study shows how controlling for endogeneity improves the reliability of corresponding forecasts. This contribution in the literature constitutes a substantial step forward in long-run econometric modeling.

In this paper we are going to investigate what will be the forecasting performance of the true model (M_1), the model with the issue of endogeneity (M_2) and the endogeneity corrected model (M_3) in the presence of cointegration among regressors. In this regard, we formulated two research questions and endeavored to answer them through a simulation experiment. First, does addressing the issue of endogeneity in M_3 aid in forecasting accuracy? Granger (1986), Engle and Yoo (1987) and Kuo (2016) argued that the longer the forecasting horizon is, the larger the percentage reduction in error metrics becomes. Therefore, we have analyzed the second research question, what is the behavior of forecast accuracy over various forecast horizons? Third, what is the behavior of forecast accuracy over various sample sizes? These two research questions are addressed by calculating and comparing the percentage reduction in M_3 relative to M_2 for three forecast horizons, 5 periods ahead, 10 periods ahead and 20 periods ahead, in three scenarios. Fourth, does the increase in the number of cointegration relationships enhance the forecast accuracy? This research question is analyzed by comparing forecast error metrics over the $r = 1$, $r = 2$ and $r = 3$.

2. Methodology

The study aims to investigate the impact of endogeneity on forecasting accuracy of cointegration-adjusted regression models. A simulation experiment is conducted based on three scenarios to evaluate the pitfalls of endogeneity. Details are discussed here as under.

2.1. Simulation Experiment

The simulation experiment comprises three models and three scenarios. The three models are the true model (M_1), model with the issue of endogeneity (M_2) and the endogeneity corrected model (M_3). These three models are simulated based on following three scenarios:

1. **Scenario-I ($r = 1$)**: $k - 1$ cointegration relationships.
2. **Scenario-II ($r = 2$)**: $k - 2$ cointegration relationships.
3. **Scenario-III ($r = 3$)**: $k - 3$ cointegration relationships

Figure 1 briefly presents the layout of the simulation experiment.

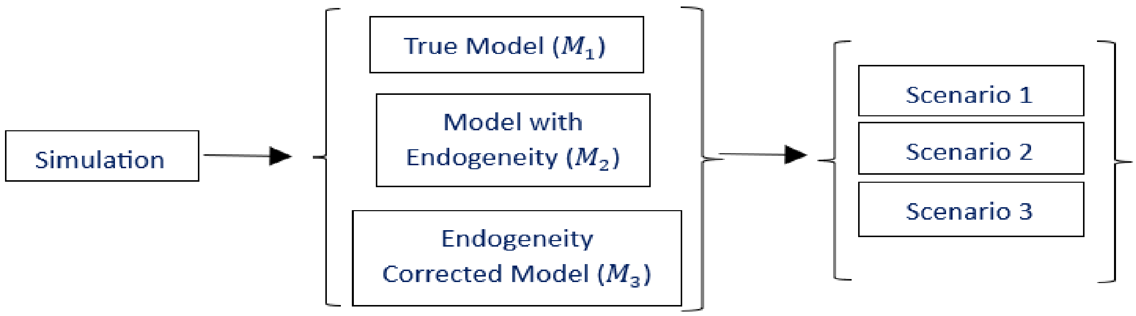


Figure 1. Simulation Experiment Flow Diagram
Source: Author's work.

To proceed toward the construction of M_1, M_2 and M_3 following pre-requisites need to be met.

i. Regressors must be correlated with each other.
ii. The regressors should be *I(1)*.
iii. There must exist cointegration between *I(1)* regressors.

2.1.1. Construction of True Model M_1

Using multivariate linear regression model we generated Data Generating Process (DGP) following Swamy and Muehlen (2020). Following Dolado et al. (2001) and Semadeni et al. (2013). we generated dependent variable as a linear combination of correlated as well as cointegrated random walk processes as;

$$Y_t = \alpha_0 + \sum_{i=1}^{k} \beta_{it} X_{it} + \varepsilon_t \qquad \text{Where, } \varepsilon_t \sim i.i.d.\ N(0,\ 1)\ and\ \ i = 1, \ldots k \qquad (8)$$

In equation (8), α_0 is intercept and β_{it} is coefficient of $i = 1, \ldots k$ regressors. Following Kremers and Erisson (1992) and Semadeni et al. (2013), the values of β_{it} coefficient is taken as $\{0 < \beta_{it} < 1\}$. The k regressors (X_{it}) in equation (8) are correlated and cointegrated Random Walk (RW) processes. We introduced the correlation among the regressors with the help of variance-covariance matrix Σ using the *mvrnorm* function in R. Cointegration is introduced by introducing common stochastic trends among k regressors following Dolado et al. (2001); Charemza and Deadman (1997); Pfaff (2008); and Ripley (2009).

$$X_{it} = \varphi_i \sum_i^k X_{it-1} + \delta_i c.t + e_{it} \quad \text{where } i = 1, \dots, k \quad e_{it} \sim MVN(0, \Sigma) \tag{9}$$

In equation (9) the $c.t$ refers to common stochastic trends specifically referring to non-stationary components that X_{it} time series share. The DGP of $c.t$ is $\sum_i^t \eta_i$ where $\sum_i^t \eta_i \sim N(0,1)$ is generated following Dolado et al. (2001) and Pfaff (2008). $e_{it} \sim MVN(0, \Sigma)$ is the multivariate normally distributed error term. Theoretically, in equation (8), when we generate $\varepsilon_t \sim N(0, 1)$ and X_{it} as correlated random walk processes with common stochastic trends $c.t$, the regressors X_{it} do not correlate with the error term ε_t of the model. This means in our true model given in equation (8) we don't have the issue of endogeneity.

The reliability of a simulation experiment depends on the reproducibility of results using different sample sizes. Three sample sizes $n = 202$, $n = 502$, and $n = 1002$ are generated. A lag order of 2 is selected for validation and generalization. The experiment is iterated 1000 times in all scenarios.

2.1.2. Construction of Model with the Issue of Endogeneity M_2

Following Greene (2003); Wooldridge, (2006); Pacini (2019) and Hirukawa et al. (2023), the model with the potential problem of endogeneity is constructed by intentionally omitting a variable. Let the M_2 with an omitted variable (X_{dt}) is:

$$Y_t = \alpha_0 + \sum_{i=1}^{k-m} \beta_{it} X_{it} + v_t \quad \text{Where, } v_t = \varepsilon_t + X_{dt} \text{ and } i = 1, \dots k - m \tag{10.a}$$

And,

$$X_{it} = \varphi_i \sum_i^{k-m} X_{it-1} + \delta_i c.t + e_t \quad \text{where } i = 1, \dots, k - m, \quad e_t \sim MVN(0, \Sigma) \tag{10.b}$$

In equation (10.a) and (10.b) $k - m$ denote the number of endogenous regressors (X_{mt}), due to omitting an important regressor (X_{dt}). According to Greene (2003), in M_2, X_{mt} regressors will be endogenous when $cor(X_{mt}, v_t) \neq 0$ if;

i. X_{mt} correlate with X_{dt} (omitted variable) and

ii. $\beta_d \neq 0$

i.e. $(cor(X_{mt}, v_t) = cor(X_{mt}, \beta_d X_{dt} + \varepsilon_t) = cor(X_{mt}, X_{dt})\beta_d + cor(X_{mt}, \varepsilon_t))$

and $cor(X_{mt}, \varepsilon_t) = 0$.

For X_{mt} to be endogenous, conditions (i) and (ii) are satisfied. We verified the presence of the endogeneity problem by applying the Durbin Wu Hausman (1978) test for endogeneity. Results of the Durbin Wu Hausman test is provided in the result and discussion section.

2.1.3. Construction of Endogeneity Corrected Model M_3

To address the endogeneity issue, constructed in M_2 we used the instrumental variable technique (Reiersøl, 1945; Sargan, 1958; Stock and Watson, 2003; Wooldridge, 2006). Following Semadeni et al. (2013) and Gui et al. (2023), we have constructed IV as a function of the endogenous regressor. The DGP of IV is as follows;

$$Z_{it} = \alpha_0 X_{mt} + e_t \quad \text{where } e_t \sim N(0, 1) \text{ and } i = 1, \dots, k, t = 1,2, \dots, n \tag{11}$$

In equation (11), α_0 is the coefficient i.e. used to correlate the IV i.e. Z_{it} with endogenous regressor X_{mt}. Following Semadeni et al. (2013), we vary the strength of the instrument Z_{it} by choosing the value of α_0 between 0.7 and 0.9. The correlation of IV with error is generated as equal to zero by adding $e_t \sim N(0, 1)$, in the DGP of IV in equation (11).

After successfully constructing a valid and relevant IV the instrumental variable Z_{it} is used to determine the estimated form of endogenous independent regressor X_{mt}[1]. In the next step, we used the predicted value of $\hat{X}_{mt}$ as regressor and proceed according to JJ cointegration approach. Therefore, the endogeneity corrected model (M_3) is as follows;

$$Y_t = \alpha_0 + \sum_{i=1}^{k-m} \beta_{it} \hat{X}_{it} + v_t \qquad \text{Where, } v_t \sim i.i.d.\ N(0,\ 1)\ and\ \ i = 1, \ldots k - m \quad (12)$$

2.2. Testing Forecast Accuracy

To assess the forecasting performance of cointegration-adjusted regression models three error metrics have been used i.e. Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percent Error (MAPE).

$$RMSE = \sqrt{\frac{1}{h}\sum_{t=1}^{h}(Y_t - \hat{Y}_t)^2} \quad (13)$$

$$MAE = \frac{1}{h}\sum_{t=1}^{h}\left|Y_t - \hat{Y}_t\right| \quad (14)$$

$$MAPE = \frac{1}{h}\sum_{t=1}^{h}\left|\frac{Y_t - \hat{Y}_t}{Y_t}\right| \quad (15)$$

In the above equation, Y_t represent actual value while $\hat{Y}_t$ represent the forecasted value. h indicate the forecast horizon.

3. Simulation Results and Discussion

3.1. Data and Data Description

The simulated data sets consist of three sample sizes $n = 200, 500$ and 1000 respectively. We simulated three regressors X_{1t}, X_{2t} and X_{3t} and one regressand Y_t for each n. In each data set, we constructed $r = 1, r = 2$ and $r = 3$ among the regressors, respectively. Cointegrated series generated in a simulation experiment are shown in Figure 2. The Augmented Dickey and Fuller (1979) unit root test is used to check for unit roots in simulated data series. The results provided in Table 1 shows that the ADF test fails to reject the null hypothesis at the level. However, it rejects the null hypothesis at the first difference, indicating a stationary data series. Statistical significance of unit root is determined based on finite critical values.

[1] The intuition behind this approach is that regressing X_{mt} on Z_{it} partials out any common variance between X_{mt} and Z_{it}, so the predicted value $\hat{X}_{mt}$ does not share any variance that is related to the error term.

Table 1: Results of Unit Root Test

		$r = 1$		$r = 2$		$r = 3$		
N	**Variables**	**Level**	**1st Diff.**	**Level**	**1st Diff.**	**Level**	**1st Diff.**	**Order of Integ.**
200	X_{1t}	-2.003	-7.173***	-1.614	-6.846***	-2.096	-8.418***	$I(1)$
	X_{2t}	-2.168	-8.335***	-2.980	-7.822***	-2.490	-8.762***	$I(1)$
	X_{3t}	-1.619	-5.481***	-1.850	-7.123***	-2.882	-9.174***	$I(1)$
	Y_t	-1.988	-7.929***	-1.790	-7.573***	-1.994	-8.473***	$I(1)$
500	X_{1t}	-2.403	-10.284***	-2.111	-10.282***	-2.851	-9.594***	$I(1)$
	X_{2t}	-2.796	-10.442***	-1.335	-10.094***	-2.794	-9.788***	$I(1)$
	X_{3t}	-0.548	-10.066***	-2.202	-10.144***	-2.679	-10.208***	$I(1)$
	Y_t	-2.331	-11.707***	-1.554	-12.933***	-2.736	-9.321***	$I(1)$
1000	X_{1t}	-2.216	-11.319***	-1.925	-11.029***	-2.192	-11.429***	$I(1)$
	X_{2t}	-1.909	-11.752***	-1.659	-12.072***	-2.134	-11.544***	$I(1)$
	X_{3t}	-2.098	-10.258***	-1.538	-11.420***	-2.147	-11.716***	$I(1)$
	Y_t	-1.881	-12.162***	-1.508	-11.575***	-2.199	-10.836***	$I(1)$

*Source: Author's calculations. Note: -3.431482, -2.861925, and -2.567018 are critical values at 1%, 5%, and 10% level of significance, respectively. Whereas, ***, **, and * indicate significance at 1%, 5%, and 10% level of significance, respectively.*

3.2. Results of Correlation Matrix

Correlations among the regressors is generated using the MASS package in R. The results in Table 2 reveal that the correlation between simulated regressors and error term is almost near zero, indicating that the regressors are exogenous. The study also found no issues of severe multicollinearity among the simulated regressors.

Table 2: Correlation Matrix

Variables	X_{1t}	X_{2t}	X_{3t}	**Error Term**
X_{1t}	1			
X_{2t}	0.35	1		
X_{3t}	0.37	0.44	1	
Error Term	-0.034	0.031	-0.056	1

Source: Author's calculations.

3.3. Results of the Hausman Test for Endogeneity

After constructing and testing the correlation among regressors and the error term we induced the endogeneity by omitting a regressor from the model. The Hausman (1978) test for endogeneity is applied to test the evidence of endogeneity[2]. The null hypothesis of the Hausman test states no endogeneity. The endogeneity of more than one regressor is detected by analyzing the F-statistics of the structural equation. The results of the test for endogeneity in case of scenario I show that when omitting a regressor, it makes X_{1t} and X_{3t} endogenous. It is confirmed by the significance of v_hat_X_{2t} and v_hat_X_{3t} and F-statistics for all sample sizes, reported in Table 3. The endogeneity of X_{1t} and X_{3t} after omitting a relevant variable is appearing due to the correlation among regressors by construction. The results of Hausman Test for Endogeneity for scenario II and III are reported in Table 4 and 5.

[2] Hausman (1978) test works by computing the reduced form for potential endogenous regressors, including all exogenous variables and additional Instrumental Variables (IVs). If the coefficient of residuals is statistically significant, it indicates the potential endogenous variable is truly endogenous.

Table 3: Results of the Hausman Test for Endogeneity in M_2: Scenario I $(r = 1)$

	n=200		n=500		n=1000	
Omitting X_1						
Variables	**Coeff.**	**P-value**	**Coeff.**	**P-value**	**Coeff.**	**P-value**
X_2	-0.117	0.055*	-0.215	0.040**	-0.121	0.027**
X_3	0.544	0.000***	0.368	0.000***	0.506	0.000***
$v\_hat\_X_{2t}$	0.306	0.000***	0.509	0.000***	0.797	0.000***
$v\_hat\_X_{3t}$	-0.473	0.009***	-0.362	0.000***	-0.646	0.000***
***F*-stats**	269.19	0.000***	1396.35	0.000***	5227.21	0.000***
Omitting X_2						
X_1	0.206	0.023**	-0.272	0.009***	0.531	0.002***
X_3	0.524	0.049**	0.572	0.031**	0.218	0.032**
$v\_hat\_X_{1t}$	-0.621	0.000***	0.853	0.000***	-0.353	0.000***
$v\_hat\_X_{3t}$	-0.425	0.040**	0.377	0.000***	-0.326	0.000***
***F*-stats**	826.49	0.000***	2283.0	0.000***	5248.0	0.000***
Omitting X_3						
X_1	0.474	0.026**	0.577	0.013**	0.657	0.003***
X_2	0.438	0.070*	0.233	0.035**	0.588	0.014**
$v\_hat\_X_{1t}$	-0.121	0.008***	0.164	0.000***	-0.644	0.000***
$v\_hat\_X_{2t}$	-0.139	0.025**	0.265	0.045**	-0.289	0.001***
***F*-stats**	1592.33	0.000***	9797.65	0.000***	36403.5	0.000***

Source: Author's calculations.
H_0: *No. endogeneity.*

Table 4:Results of the Hausman Test for Endogeneity in M_2: Scenario II $(r = 2)$

	n=200		n=500		n=1000	
Omitting X_1						
Variables	**Coeff.**	**P-value**	**Coeff.**	**P-value**	**Coeff.**	**P-value**
X_2	-0.341	0.011**	0.216	0.033**	-0.489	0.024***
X_3	0.663	0.004***	0.628	0.011**	0.582	0.000***
$v\_hat\_X_{2t}$	0.546	0.000***	0.726	0.000***	0.272	0.000***
$v\_hat\_X_{3t}$	-0.156	0.041**	-0.126	0.0177**	-0.189	0.000***
***F*-stats**	142.167	0.000***	548.61	0.000***	2873.07	0.000***
Omitting X_2						
X_1	0.374	0.026**	0.662	0.0157**	0.419	0.007***
X_3	0.254	0.048**	0.605	0.025**	0.307	0.022**
$v\_hat\_X_{1t}$	0.218	0.000***	0.177	0.000***	0.184	0.000***
$v\_hat\_X_{3t}$	-0.151	0.008***	0.271	0.003***	0.129	0.000***
***F*-stats**	379.93	0.000***	1865.0	0.000***	5703.2	0.000***
Omitting X_3						
X_1	0.714	0.000***	0.703	0.000***	0.699	0.000***
X_2	0.442	0.014**	0.424	0.059*	0.316	0.023**
$v\_hat\_X_{1t}$	-0.546	0.028**	-0.404	0.090*	-0.385	0.007***
$v\_hat\_X_{2t}$	-0.142	0.045**	-0.125	0.037**	-0.166	0.022**
***F*-stats**	733.85	0.000***	4383.78	0.000***	18051.7	0.000***

Source: Author's calculations.
H_0: *No. endogeneity.*

Table 5: Results of the Hausman Test for Endogeneity in M_2: Scenario III $(r = 3)$

	n=200		n=500		n=1000	
Omitting X_1						
Variables	**Coeff.**	**P-value**	**Coeff.**	**P-value**	**Coeff.**	**P-value**
X_2	-0.620	0.040**	0.761	0.030**	0.563	0.013**
X_3	0.709	0.007***	0.618	0.000***	0.578	0.000***
$v\_hat\_X_{2t}$	0.711	0.000***	-0.471	0.000***	-0.278	0.000***
$v\_hat\_X_{3t}$	-0.234	0.017**	-0.118	0.021**	0.288	0.000***
***F*-stats**	276.29	0.000***	1494.11	0.000***	5289.2	0.000***

Omitting X_2						
X_1	0.813	0.000***	0.783	0.000***	0.773	0.000***
X_3	0.285	0.018**	0.257	0.000***	0.247	0.000***
$v\_hat\_X_{1t}$	-0.204	0.065*	-0.683	0.005***	-0.560	0.017**
$v\_hat\_X_{3t}$	-0.336	0.037**	-0.303	0.023**	-0.185	0.010***
***F*-stats**	1067.02	0.000***	3650.24	0.000***	5468.21	0.000***
Omitting X_3						
X_1	0.714	0.000***	0.704	0.000***	0.699	0.000***
X_2	0.334	0.015**	0.337	0.056*	0.291	0.021**
$v\_hat\_X_{1t}$	-0.154	0.054*	-0.404	0.079*	-0.381	0.009***
$v\_hat\_X_{2t}$	-0.346	0.035**	-0.380	0.040**	0.382	0.024**
***F*-stats**	742.76	0.000***	4301.03	0.000***	17302.8	0.000***

Source: Author's calculations.

H_0: *No. endogeneity.*

3.4. Results of Instrumental Variable Validity

Instrumental Variable (IV) is a variable used to isolate variation in an endogenous variable that is uncorrelated with the error term. Its validity is determined by three characteristics: relevance, exogeneity, and exclusion restriction[3]. The results of IV relevance and exogeneity are presented in Tables 6 to 8 for scenario I, scenario II, and scenario III respectively. Relevance is determined by the instrument's substantial correlation with the endogenous explanatory variable, which is examined using the F-statistic from the first-stage regression of 2SLS. Exogeneity is determined by the instrumental variable's being uncorrelated with the error term. The results indicate that all IVs (z1, z2, and z3) have zero correlation with the error term v_t of the M_2, confirming their exogeneity. Exclusion restriction is justified intuitively by the fact that Z_{it} is generated as a function of endogenous regressor X_{mt}, ensuring that it is uncorrelated with all factors directly affecting Y_t.

Table 6: Results of Instruments Relevance and Exogeneity: Scenario I (r=1)

	n=200	n=500	n=1000	n=200	n=500	n=1000
Omitting X_1						
IV	**F-stats**	**F-stats**	**F-stats**	**Cor. with v_t**	**Cor. with v_t**	**Cor. with v_t**
z2	100.69	190.66	533.16	0.000	0.000	0.000
z3	60.07	236.53	491.32	0.000	0.000	0.000
Omitting X_2						
z1	105.68	451.91	1297.80	0.000	0.000	0.000
z3	47.40	213.85	595.77	0.000	0.000	0.000
Omitting X_3						
z1	102.21	436.54	1210.55	0.000	0.000	0.000
z2	45.31	190.81	585.56	0.000	0.000	0.000

Source: Author's calculations.

Table 7: Results of Instruments Relevance and Exogeneity: Scenario II ($r = 2$)

	n=200	n=500	n=1000	n=200	n=500	n=1000
Omitting X_1						
IV	**F-stats**	**F-stats**	**F-stats**	**Cor. with v_t**	**Cor. with v_t**	**Cor. with v_t**
z2	29.34	138.58	385.02	0.000	0.000	0.000
z3	33.12	205.25	791.78	0.000	0.000	0.000
Omitting X_2						
z1	120.11	599.99	1855.22	0.000	0.000	0.000

[3] Exclusion restriction means the instrumental variable cannot directly affect the dependent variable except through its endogenous explanatory variable.

z3	22.91	111.80	320.55	0.000	0.000	0.000
Omitting X_3						
z1	219.38	646.99	1755.79	0.000	0.000	0.000
z2	45.20	141.71	333.04	0.000	0.000	0.000

Source: Author's calculations.

Table 8: Results of Instruments Relevance and Exogeneity: Scenario III $(r = 3)$

	n=200	**n=500**	**n=1000**	**n=200**	**n=500**	**n=1000**
Omitting X_1						
IV	**F-stats**	**F-stats**	**F-stats**	**Cor. with v_t**	**Cor. with v_t**	**Cor. with v_t**
z2	36.17	122.07	885.07	0.000	0.000	0.000
z3	20.94	123.14	501.60	0.000	0.000	0.000
Omitting X_2						
z1	72.87	420.87	1601.28	0.000	0.000	0.000
z3	28.79	168.93	641.84	0.000	0.000	0.000
Omitting X_3						
z1	117.70	522.18	1907.83	0.000	0.000	0.000
z2	42.04	227.56	930.93	0.000	0.000	0.000

Source: Author's calculations.

3.5. Results of Forecast Evaluation of Cointegration-Adjusted Regression Models with and without the Issue of Endogeneity

This section evaluates the forecasting accuracy of three models M_1 (true model), M_2 (model with the issue of endogeneity), and M_3 (endogeneity corrected model) in the presence of cointegration among regressors. The detailed results are provided in Table 9 to Table 17. The first research question asks whether addressing endogeneity in M_3 improves its forecast accuracy relative to M_2. The results clearly support that M_3 consistently outperforms M_2 across all scenarios ($r = 1, r = 2, r = 3$), forecast horizons (5, 10, and 20 periods), and sample sizes (200, 500, 1000), with substantial reductions in RMSE, MAE, MAPE, and Theil statistics. For example, in scenario I with a sample size of 200 and a 5-period horizon, M_3 achieves a 70.84% lower RMSE compared to M_2. These improvements confirm that correcting for endogeneity through IV techniques significantly enhances forecasting accuracy in cointegration-adjusted regression models.

The second research question explores how forecast accuracy changes over various forecast horizons. The analysis reveals that the improvement in M_3 accuracy relative to M_2 is most prominent at shorter horizons and declines as the horizon extends. In scenario I, the RMSE reduction for M_3 relative to M_2 drops from 70.84% at a 5-period forecast to 49.7% at 20 periods. This pattern is consistent across all scenarios, suggesting that the benefits of endogeneity correction are strongest in short-term forecasts, likely due to growing uncertainty in long-term predictions. These findings align with the views of Clements and Hendry (1995) but contrast with studies like Engle and Yoo (1987) and Granger (1986), who advocate improved long-run forecasts in cointegration-adjusted models.

The third research question examines how forecast accuracy varies with changes in sample size and the number of cointegration relationships. Results indicate that increasing the sample size does not significantly improve the relative forecast performance of M_3, and in some cases, the reduction in forecast errors is marginal or mixed. Likewise, increasing the number of cointegrating relationships from $r = 1$ to $r = 3$ does not enhance M_3 advantage over M_2; in fact, the improvement diminishes in later scenarios. For example, the RMSE reduction in scenario III is consistently lower than in scenarios I and II.

From the above discussion, we can conclude that addressing the issue of endogeneity in cointegration-adjusted M_3 helps in improving the forecasting accuracy of M_3 relative to M_2 in the shorter forecast horizon, in relatively smaller sample size and in $r = 1$. However, an increase in forecast horizons, extending sample size and increase in number of cointegration relationship does not play a role in this regard.

In Figure 3 to Figure 11 graphical depiction of the forecast evaluation of M_1, M_2 and M_3 for each scenario is presented. The graphical depiction also supports addressing the issue of endogeneity in M_3 in the presence of cointegration among regressors helps in improving the forecasts' accuracy

Table 9: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario I $(r = 1)$ *Omitting* X_1

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
5 Periods	M_1	1.99	1.70	103.80	0.46	1.90	1.64	135.32	0.34	1.92	1.66	66.68	0.28
	M_2	7.75	5.90	261.68	0.85	6.50	5.69	248.08	0.69	6.47	5.67	260.77	0.60
	M_3	**2.26**	**1.97**	**108.30**	**0.49**	**2.30**	**2.00**	**88.62**	**0.38**	**2.24**	**1.94**	**94.77**	**0.26**
% FE Reduction in M_3 Relative to M_2		70.84	66.61	58.61	42.35	64.62	64.85	64.28	44.93	65.38	65.78	63.66	56.67
10 Periods	M_1	2.20	1.85	189.23	0.50	2.19	1.84	122.96	0.37	2.18	1.82	79.66	0.29
	M_2	7.06	6.04	412.61	0.76	6.88	5.87	351.04	0.58	6.84	5.84	197.43	0.50
	M_3	**2.69**	**2.30**	**240.85**	**0.54**	**2.73**	**2.33**	**234.14**	**0.39**	**2.62**	**2.23**	**85.28**	**0.30**
% FE Reduction in M_3 Relative to M_2		61.92	61.92	41.61	28.95	60.32	60.31	33.3	32.76	61.69	61.13	56.81	40.0
20 Periods	M_1	2.58	2.16	194.58	0.55	2.60	2.17	140.19	0.40	2.61	2.19	130.32	0.33
	M_2	7.57	6.41	354.66	0.78	7.47	6.31	277.43	0.71	7.28	6.15	218.18	0.59
	M_3	**3.81**	**2.27**	**248.81**	**0.58**	**3.15**	**2.67**	**134.22**	**0.43**	**3.28**	**2.80**	**87.85**	**0.37**
% FE Reduction in M_3 Relative to M_2		49.7	64.6	29.9	25.6	57.8	57.7	51.7	39.4	54.9	54.3	59.6	37.3

Source: Author's calculations.

Table 10: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario I $(r = 1)$ *Omitting* X_2

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
5 Periods	M_1	1.99	1.70	103.80	0.46	1.90	1.64	135.32	0.34	1.92	1.66	66.68	0.28
	M_2	7.69	6.86	361.03	0.83	7.57	6.75	393.99	0.70	7.41	6.61	254.37	0.68
	M_3	**2.53**	**2.20**	**118.85**	**0.52**	**2.55**	**2.23**	**103.64**	**0.42**	**2.50**	**2.20**	**87.05**	**0.34**
% FE Reduction in M_3 Relative to M_2		67.1	67.9	67.1	37.3	66.3	67.0	73.7	40.0	66.3	66.7	65.8	50.0
10 Periods	M_1	2.20	1.85	189.23	0.50	2.19	1.84	122.96	0.37	2.18	1.82	79.66	0.29
	M_2	7.05	6.02	339.71	0.74	5.94	5.92	357.01	0.73	6.87	6.87	187.30	0.67
	M_3	**2.99**	**2.55**	**209.55**	**0.56**	**3.07**	**2.63**	**149.27**	**0.51**	**3.05**	**2.60**	**98.93**	**0.38**
% FE Reduction in M_3 Relative to M_2		57.6	57.6	38.3	24.3	48.3	55.6	58.2	30.1	55.6	62.2	47.2	43.3
20 Periods	M_1	2.58	2.16	194.58	0.55	2.60	2.17	140.19	0.40	2.61	2.19	130.32	0.33
	M_2	7.52	6.37	438.43	0.77	7.27	6.16	368.73	0.72	7.24	6.13	282.55	0.71
	M_3	**3.70**	**3.17**	**208.85**	**0.62**	**3.80**	**3.26**	**179.65**	**0.49**	**3.93**	**3.37**	**151.09**	**0.42**

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
% FE Reduction in M_3 Relative to M_2		50.8	50.2	52.4	19.5	47.7	47.1	51.3	31.9	45.7	45.0	46.5	40.8

Source: Author's calculations.

Table 11: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario I ($r = 1$) Omitting X_3

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
5 Periods	M_1	1.99	1.70	103.80	0.46	1.90	1.64	135.32	0.34	1.92	1.66	66.68	0.28
	M_2	7.85	6.03	354.87	0.75	7.60	5.77	341.63	0.60	7.40	5.63	244.10	0.59
	M_3	**2.43**	**2.11**	**118.91**	**0.50**	**2.45**	**2.14**	**168.58**	**0.40**	**2.50**	**2.18**	**116.01**	**0.34**
% FE Reduction in M_3 Relative to M_2		69.0	65.0	66.5	33.3	67.8	62.9	50.6	33.3	66.2	61.3	52.5	42.4
10 Periods	M_1	2.20	1.85	189.23	0.50	2.19	1.84	122.96	0.37	2.18	1.82	79.66	0.29
	M_2	7.05	6.01	397.78	0.85	6.87	5.88	269.97	0.68	6.81	5.80	199.14	0.56
	M_3	**2.89**	**2.47**	**203.49**	**0.59**	**3.02**	**2.58**	**162.90**	**0.42**	**2.97**	**2.54**	**92.12**	**0.36**
% FE Reduction in M_3 Relative to M_2		59.0	58.9	48.8	30.6	56.0	56.1	39.6	38.2	56.4	56.2	53.7	35.7
20 Periods	M_1	2.58	2.16	194.58	0.55	2.60	2.17	140.19	0.40	2.61	2.19	130.32	0.33
	M_2	7.54	6.38	326.54	0.76	7.27	6.15	254.45	0.63	7.20	6.08	321.64	0.46
	M_3	**3.52**	**2.99**	**222.82**	**0.60**	**3.63**	**3.09**	**158.78**	**0.48**	**3.75**	**3.21**	**205.07**	**0.36**
% FE Reduction in M_3 Relative to M_2		53.3	53.1	31.7	21.1	50.1	49.8	37.6	23.8	47.9	47.2	36.3	21.7

Source: Author's calculations.

Table 12: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario II ($r = 2$) Omitting X_1

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
5 Periods	M_1	1.98	1.71	114.60	0.48	1.98	1.69	102.28	0.35	1.91	1.64	62.28	0.24
	M_2	6.17	5.25	494.39	0.77	6.13	5.22	205.99	0.62	6.11	5.22	237.05	0.52
	M_3	**2.08**	**1.80**	**127.30**	**0.50**	**2.11**	**1.83**	**105.65**	**0.37**	**2.20**	**1.91**	**85.84**	**0.31**
% FE Reduction in M_3 Relative to M_2		66.29	65.71	74.23	35.06	65.57	64.94	48.73	40.32	63.99	63.41	63.78	40.38
10 Periods	M_1	2.21	1.85	141.11	0.46	2.21	1.86	122.34	0.36	2.24	1.88	56.44	0.28
	M_2	6.44	5.28	400.57	0.74	6.43	5.30	275.32	0.61	6.48	5.32	193.61	0.51

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
	$\boldsymbol{M_3}$	**2.33**	**1.95**	**229.23**	**0.49**	**2.43**	**2.06**	**160.02**	**0.39**	**2.51**	**2.13**	**124.90**	**0.30**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		63.77	63.07	42.77	33.78	62.20	61.13	41.89	36.07	61.26	59.96	35.50	41.18
20 Periods	$\boldsymbol{M_1}$	2.54	2.13	303.39	0.54	2.60	2.17	117.12	0.40	2.64	2.22	103.65	0.35
	$\boldsymbol{M_2}$	6.59	5.33	605.17	0.74	6.66	5.41	297.55	0.62	6.60	5.36	276.42	0.52
	$\boldsymbol{M_3}$	**2.71**	**2.28**	**186.66**	**0.55**	**2.78**	**2.34**	**137.37**	**0.43**	**2.91**	**2.47**	**116.93**	**0.34**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		58.87	57.22	69.17	25.68	58.23	56.75	53.84	30.65	55.91	53.92	57.70	34.62

Source: Author's calculations.

Table 13: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario II ($r = 2$) *Omitting* X_2

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
5 Periods	$\boldsymbol{M_1}$	1.98	1.71	114.60	0.48	1.98	1.69	102.28	0.35	1.91	1.64	62.28	0.24
	$\boldsymbol{M_2}$	6.26	5.38	279.93	0.74	6.21	5.27	192.75	0.62	6.07	5.16	367.17	0.50
	$\boldsymbol{M_3}$	**2.08**	**1.80**	182.55	**0.49**	**2.16**	**1.87**	**75.27**	**0.40**	**2.12**	**1.82**	**77.19**	**0.27**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		66.77	66.54	34.79	33.78	65.37	64.54	60.94	35.48	65.07	64.73	78.98	46.0
10 Periods	$\boldsymbol{M_1}$	2.21	1.85	141.11	0.46	2.21	1.86	122.34	0.36	2.24	1.88	56.44	0.28
	$\boldsymbol{M_2}$	6.31	5.16	334.23	0.75	6.12	5.07	235.39	0.61	6.48	5.30	168.06	0.51
	$\boldsymbol{M_3}$	**2.41**	**2.04**	**188.15**	**0.51**	**2.47**	**2.11**	**149.20**	**0.40**	**2.46**	**2.10**	**74.27**	**0.32**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		61.85	60.58	43.68	32.00	59.69	58.38	36.60	34.43	62.04	60.38	55.76	37.25
20 Periods	$\boldsymbol{M_1}$	2.54	2.13	303.39	0.54	2.60	2.17	117.12	0.40	2.64	2.22	103.65	0.35
	$\boldsymbol{M_2}$	6.51	5.31	426.49	0.72	6.57	5.29	266.78	0.61	6.81	5.51	204.97	0.53
	$\boldsymbol{M_3}$	**2.75**	**2.32**	**324.62**	**0.55**	**2.84**	**2.41**	**135.69**	**0.45**	**3.12**	**2.67**	**120.29**	**0.39**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		57.77	56.31	23.87	23.61	56.73	54.43	49.15	26.20	54.17	51.57	41.31	26.42

Source: Author's calculations.

Table 14: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario II ($r = 2$) *Omitting* X_3

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
	$\boldsymbol{M_1}$	1.98	1.71	114.60	0.48	1.98	1.69	102.28	0.35	1.91	1.64	62.28	0.24

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
5 Periods	$\boldsymbol{M_2}$	5.48	4.67	331.65	0.79	5.22	4.45	260.31	0.59	5.28	4.49	138.59	0.50
	$\boldsymbol{M_3}$	**2.09**	**1.80**	**121.47**	**0.48**	**2.18**	**1.88**	**93.20**	**0.38**	**2.13**	**1.83**	**73.39**	**0.30**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		61.88	61.45	63.33	39.24	58.30	57.76	64.24	35.44	59.55	59.16	47.03	40.00
10 Periods	$\boldsymbol{M_1}$	2.21	1.85	141.11	0.46	2.21	1.86	122.34	0.36	2.24	1.88	56.44	0.28
	$\boldsymbol{M_2}$	5.56	4.55	253.35	0.74	5.53	4.55	165.25	0.58	5.46	4.39	306.91	0.46
	$\boldsymbol{M_3}$	**2.34**	**1.97**	**146.30**	**0.51**	**2.40**	**2.03**	**138.76**	**0.37**	**2.46**	**2.08**	**64.26**	**0.31**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		57.90	56.74	42.23	31.08	56.60	55.37	16.00	36.21	54.98	52.64	79.05	32.61
20 Periods	$\boldsymbol{M_1}$	2.54	2.13	303.39	0.54	2.60	2.17	117.12	0.40	2.64	2.22	103.65	0.35
	$\boldsymbol{M_2}$	5.82	4.74	484.77	0.75	5.59	4.54	353.88	0.58	5.61	4.58	220.14	0.48
	$\boldsymbol{M_3}$	**2.68**	**2.25**	**332.15**	**0.53**	**2.79**	**2.35**	**136.53**	**0.41**	**2.83**	**2.39**	**125.06**	**0.37**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		54.02	52.55	31.52	29.33	50.15	48.23	61.43	29.31	49.56	47.85	43.18	22.92

Source: Author's calculations.

Table 15: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario III ($r = 3$) Omitting X_1

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil	RMSE	MAE	MAPE	Theil
5 Periods	$\boldsymbol{M_1}$	2.25	1.95	202.90	0.43	2.15	1.85	52.10	0.27	2.13	1.83	54.88	0.22
	$\boldsymbol{M_2}$	5.73	4.84	319.00	0.70	6.42	5.44	247.52	0.63	5.57	4.73	115.93	0.42
	$\boldsymbol{M_3}$	**2.94**	**2.42**	**210.05**	**0.44**	**2.18**	**1.88**	**68.71**	**0.30**	**2.22**	**1.92**	**69.19**	**0.23**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		**48.69**	**50.0**	**34.19**	**37.14**	**66.03**	**65.44**	**72.23**	**52.38**	**60.14**	**59.41**	**40.33**	**45.24**
10 Periods	$\boldsymbol{M_1}$	2.53	2.13	200.38	0.43	2.60	2.18	237.59	0.31	2.56	2.16	106.21	0.27
	$\boldsymbol{M_2}$	6.04	4.97	349.73	0.68	5.61	4.61	320.34	0.50	5.42	4.45	312.81	0.39
	$\boldsymbol{M_3}$	**3.07**	**2.46**	**220.94**	**0.47**	**2.86**	**2.53**	**266.32**	**0.33**	**2.58**	**2.16**	**132.56**	**0.29**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		**49.17**	**50.50**	**36.82**	**30.88**	**48.85**	**42.12**	**16.87**	**34.00**	**52.40**	**51.46**	**57.61**	**25.64**
20 Periods	$\boldsymbol{M_1}$	3.20	2.71	167.72	0.53	3.20	2.71	122.72	0.38	3.05	2.57	64.68	0.31
	$\boldsymbol{M_2}$	6.17	5.03	342.45	0.67	6.07	4.94	318.58	0.57	6.35	5.18	122.69	0.44
	$\boldsymbol{M_3}$	**3.23**	**2.72**	**203.26**	**0.56**	**3.27**	**2.87**	**145.73**	**0.38**	**3.21**	**2.70**	**66.44**	**0.35**
% FE Reduction in $\boldsymbol{M_3}$ Relative to $\boldsymbol{M_2}$		**47.65**	**45.92**	**40.64**	**16.42**	**46.13**	**41.91**	**54.26**	**33.33**	**49.45**	**47.88**	**45.86**	**20.45**

Source: Author's calculations.

Table 16: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario III ($r = 3$) *Omitting* X_2

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	**MAE**	**MAPE**	**Theil**	**RMSE**	**MAE**	**MAPE**	**Theil**	**RMSE**	**MAE**	**MAPE**	**Theil**
5 Periods	M_1	2.25	1.95	202.90	0.43	2.15	1.85	52.10	0.27	2.13	1.83	54.88	0.22
	M_2	5.85	4.96	379.46	0.72	5.24	4.46	261.67	0.49	5.17	4.38	209.76	0.40
	M_3	**2.25**	**2.20**	**225.65**	**0.48**	**2.16**	**1.88**	**75.25**	**0.28**	**2.20**	**1.91**	**62.27**	**0.23**
% FE Reduction in M_3 Relative to M_2		61.54	55.65	40.54	33.33	58.78	57.62	71.26	42.86	57.45	56.39	70.31	42.50
10 Periods	M_1	2.53	2.13	200.38	0.43	2.60	2.18	237.59	0.31	2.56	2.16	106.21	0.27
	M_2	6.03	4.98	347.62	0.69	5.65	4.65	481.20	0.48	5.53	4.55	228.76	0.39
	M_3	**2.61**	**2.21**	**225.83**	**0.44**	**2.76**	**2.36**	**276.52**	**0.31**	**2.59**	**2.18**	**133.64**	**0.31**
% FE Reduction in M_3 Relative to M_2		56.72	55.62	35.06	36.23	51.15	49.25	42.53	35.42	53.16	52.09	41.57	20.51
20 Periods	M_1	3.20	2.71	167.72	0.53	3.20	2.71	222.72	0.38	3.05	2.57	64.68	0.31
	M_2	6.25	5.11	342.10	0.69	7.36	5.60	524.03	0.63	6.31	5.11	145.72	0.43
	M_3	**3.25**	**2.78**	**181.25**	**0.57**	**3.48**	**2.99**	**245.04**	**0.46**	**3.30**	**2.65**	**78.95**	**0.32**
% FE Reduction in M_3 Relative to M_2		48.00	45.60	47.03	17.39	52.72	46.61	53.23	26.98	47.70	48.14	45.82	25.58

Source: Author's calculations.

Table 17: Forecast Evaluation of Cointegration Adjusted-Regression Models based on Scenario III ($r = 3$) *Omitting* X_3

Forecast Horizon	Models	n = 200				n = 500				n = 1000			
		RMSE	**MAE**	**MAPE**	**Theil**	**RMSE**	**MAE**	**MAPE**	**Theil**	**RMSE**	**MAE**	**MAPE**	**Theil**
5 Periods	M_1	2.25	1.95	202.90	0.43	2.15	1.85	52.10	0.27	2.13	1.83	54.88	0.22
	M_2	5.85	4.97	359.06	0.71	5.71	4.86	148.54	0.55	5.50	4.67	137.13	0.41
	M_3	**2.34**	**2.03**	**127.21**	**0.43**	**2.28**	**1.97**	**90.72**	**0.30**	**2.30**	**2.00**	**80.92**	**0.24**
% FE Reduction in M_3 Relative to M_2		59.91	59.16	64.57	39.44	60.07	59.48	38.92	45.45	58.18	57.19	41.0	41.46
10 Periods	M_1	2.53	2.13	200.38	0.43	2.60	2.18	237.59	0.31	2.56	2.16	106.21	0.27
	M_2	6.09	5.03	341.56	0.70	6.04	4.99	355.05	0.53	5.84	4.80	230.76	0.43
	M_3	**2.61**	**2.21**	**235.41**	**0.45**	**2.70**	**2.28**	**254.37**	**0.34**	**2.72**	**2.32**	**153.65**	**0.27**
% FE Reduction in M_3 Relative to M_2		57.14	56.06	31.08	35.71	55.30	54.31	28.37	35.85	53.41	51.67	33.42	37.21
20 Periods	M_1	3.20	2.71	167.72	0.53	3.20	2.71	222.72	0.38	3.05	2.57	64.68	0.31
	M_2	6.32	5.17	349.35	0.89	6.73	5.48	412.48	0.58	6.26	5.10	181.54	0.45
	M_3	**3.24**	**2.74**	**189.14**	**0.59**	**3.20**	**2.77**	**305.17**	**0.39**	**3.24**	**2.75**	**106.36**	**0.33**
% FE Reduction in M_3 Relative to M_2		48.73	47.0	45.86	33.71	52.45	49.45	26.01	32.76	48.24	46.08	41.41	26.67

Source: Author's calculations.

The study presents graphical depictions of forecast evaluations of M_1, M_2, and M_3 for each scenario, highlighting the impact of addressing endogeneity in M_3 with cointegration among regressors on forecast accuracy. Results show that M_3 error metrics are closer to the true regression model M_1, while M_2 error metrics are higher. The impact of addressing endogeneity is more pronounced in shorter forecast horizons and smaller sample sizes. The study also reveals that increasing the number of cointegration relationships does not improve the forecasting performance of M_3 relative to M_2, regardless of the sample size or forecast horizon.

4. Conclusion

This chapter explores how endogeneity affects the forecasting performance of cointegration-adjusted regression models across various scenarios involving different cointegration relationships $(r = 1,\ r = 2, r = 3)$, sample sizes $(n = 200, 500, 1000)$, and forecast horizons (5, 10, and 20 periods). The evaluation, based on RMSE, MAE, MAPE, and Theil statistics, addresses four key research questions. The findings for the first question indicate that correcting for endogeneity using instrumental variables significantly improves the forecast accuracy of M_3 relative to M_2, confirming the importance of addressing omitted variable bias in cointegrated settings. Regarding the second question, results reveal that the forecasting advantage of M_3 diminishes as the forecast horizon increases, suggesting that the gains from endogeneity correction are more impactful in the short term due to rising uncertainty over time.

For the third research question, the study finds that increasing the number of cointegrating relationships does not enhance the forecast performance of M_3 relative to M_2. These results hold consistently across different sample sizes and forecast horizons. Ultimately, the chapter provides a practical framework for researchers aiming to improve forecasting in models with cointegrated variables and potential endogeneity issues. While correcting for endogeneity improves predictive accuracy, the study also highlights the complex nature of forecasting, emphasizing that gains are conditional on specific data structures and model specifications similar to the challenge of predicting weather, where some uncertainty always remains.

Statements and Declarations

1. Author Contributions

The author solely contributed to the study conception and design. The author performed material preparation, data generation and statistical analysis. The manuscript is written by the author and checked by Dr. Saud Ahmed Khan and Dr. Hafsa Hina as peer reviewers. The author read and approved the final manuscript.

2. Funding

The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

3. Data Availability Statement

The simulation codes and data that support the findings of this paper are obtainable from the author upon reasonable request.

References

Charemza, W. W. (1997). New Directions in Econometric Practice.

Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science*, *22*(4), 341-352. https://doi.org/10.1080/07388940500339183

Clements, M. P., & Hendry, D. F. (1995). Forecasting in cointegrated systems. *Journal of Applied Econometrics*, *10*(2), 127-146. https://doi.org/10.1002/jae.3950100204

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427-431. https://doi.org/10.1080/01621459.1979.10482531

Dolado, J. J., Gonzalo, J., & Marmol, F. (2001). Cointegration. *A Companion to Theoretical Econometrics*, 634-654. https://doi.org/10.1002/9780470996249.ch31

Engle, R. F., & Granger, C. W. (1987). Co-integration and error correction: representation, estimation, and testing. Econometrica: *Journal of the Econometric Society*, 251-276. https://doi.org/10.2307/1913236

Engle, R. F., Hendry, D. F., & Richard, J.-F. (1983). Exogeneity. Econometrica: *Journal of the Econometric Society*, 277-304. https://doi.org/10.2307/1911990

Engle, R. F., & Yoo, B. S. (1987). Forecasting and testing in co-integrated systems. *Journal of Econometrics*, 35(1), 143-159. https://doi.org/10.1016/0304-4076(87)90085-6

Friedman, M., & Schwartz, A. J. (1982). Monetary Trends in the United States and United Kingdom: Their Relation to Income, Prices and Interest Rates, 1867-1975. https://doi.org/10.7208/chicago/9780226264257.001.0001

Granger, C. J. (1986). Developments in the study of cointegrated economic variables. Oxford Bulletin of Economics and statistics, 48(3), 213-228. https://doi.org/10.1111/j.1468-0084.1986.mp48003002.x

Greene, W. H. (2003). Econometric analysis. *Pretence Hall.*

Greene, W. H. (2012). Econometric analysis, 71e. Stern School of Business, New York University.

Gregory, A. W., & Hansen, B. E. (1996). Residual-based tests for cointegration in models with regime shifts. *Journal of econometrics*, *70*(1), 99-126. https://doi.org/10.1016/0304-4076(69)41685-7

Gui, R., Meierer, M., Schilter, P., & Algesheimer, R. (2023). REndo: internal instrumental variables to address endogeneity. *Journal of Statistical Software*, 107, 1-43. https://doi.org/10.18637/jss.v107.i03

Hansen, B. E. (1992). Efficient estimation and testing of cointegrating vectors in the presence of deterministic trends. *Journal of econometrics*, *53*(1-3), 87-121. https://doi.org/10.1016/0304-4076(92)90081-2

Hausman J.A. (1978). Specification Tests in Econometrics. *Econometrica*, *46*(6), 1251-1271.

Herzer, D., & Vollmer, S. (2012). Inequality and growth: evidence from panel cointegration. *The Journal of Economic Inequality*, *10*(4), 489-503. https://doi.org/10.2307/1913827

Hirukawa, M., Murtazashvili, I., & Prokhorov, A. (2023). Yet another look at the omitted variable bias. *Econometric Reviews*, *42*(1), 1-27. https://doi.org/10.1080/07474938.2022.2157965

Hunushek, E., & Jackson, J. (1977). Statistical Methods for the Social Sciences: Orlando, Florida: Academic Press. https://doi.org/10.1016/B978-0-08-091857-0.50006-6

Kremers, J. J., Ericsson, N. R., & Dolado, J. J. (1992). The power of cointegration tests. *Oxford Bulletin of Economics and Statistics*, *54*(3), 325-348. https://doi.org/10.1111/j.1468-0084.1992.tb00005.x

Kuo, C.-Y. (2016). Does the vector error correction model perform better than others in forecasting stock price? An application of residual income valuation theory. *Economic Modelling*, *52*, 772-789. https://doi.org/10.1016/j.econmod.2015.10.016

Pacini, D. (2019). Two-sample least squares projection. *Econometric Reviews*, *38*(1), 95-123. https://doi.org/10.1080/07474938.2016.1222068

Pedroni, P. (2007). Social capital, barriers to production and capital shares: implications for the importance of parameter heterogeneity from a nonstationary panel approach. *Journal of Applied Econometrics*, *22*(2), 429-451. https://doi.org/10.1002/jae.948

Pedroni, P. (2019). Panel cointegration techniques and open challenges. *Panel data econometrics*, 251-287. https://doi.org/10.1016/B978-0-12-814367-4.00010-1

Pfaff, B. (2008). *Analysis of integrated and cointegrated time series with R*. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-75967-8

Phillips, P. C. (1986). Understanding spurious regressions in econometrics. *Journal of econometrics*, *33*(3), 311-340. https://doi.org/10.1007/978-0-387-75967-8

Phillips, P. C. (1988b). Regression theory for near-integrated time series. *Econometrica: Journal of the Econometric Society*, 1021-1043. https://doi.org/10.2307/1911357

Phillips, P. C., & Hansen, B. E. (1990). Statistical inference in instrumental variables regression with I (1) processes. *The Review of Economic Studies*, *57*(1), 99-125. https://doi.org/10.1002/jae.948

Pratt, J. W., & Schlaifer, R. (1984). On the nature and discovery of structure. *Journal of the American Statistical Association*, *79*(385), 9-21. https://doi.org/10.1080/01621459.1984.10477054

Pratt, J. W., & Schlaifer, R. (1988). On the interpretation and observation of laws. *Journal of Econometrics*, *39*(1-2), 23-52. https://doi.org/10.1016/0304-4076(88)90039-5

Reiersøl, O. (1945). Confluence analysis by means of instrumental sets of variables. Doctoral Dissertation.

Ripley, B. D. (2009). *Stochastic simulation*. John Wiley & Sons.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica: Journal of the Econometric Society*, 393-415. https://doi.org/10.2307/1907619

Saikkonen, P. (1991). Asymptotically efficient estimation of cointegration regressions. *Econometric theory*, *7*(1), 1-21. https://doi.org/10.1017/S0266466600004217

Semadeni, M., Withers, M. C., & Trevis Certo, S. (2013). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal*, *35*(7), 1070–1079. https://doi.org/10.1002/smj.2136

Stock, J. H., & Watson, M. W. (1993). A simple estimator of cointegrating vectors in higher order integrated systems. *Econometrica: Journal of the Econometric Society*, 783-820. https://doi.org/10.2307/2951763

Stock, J., & Watson, M. W. (2003). Introduction to Econometrics. https://doi.org/10.5871/bacad/9780197262733.003.0001

Swamy, P., Mehta, J. S., & Chang, I.-L. (2017). Endogeneity, time-varying coefficients, and incorrect vs. correct ways of specifying the error terms of econometric models. *Econometrics*, 5(1), 8. https://doi.org/10.3390/econometrics5010008

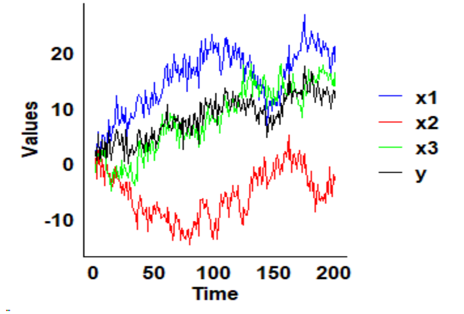
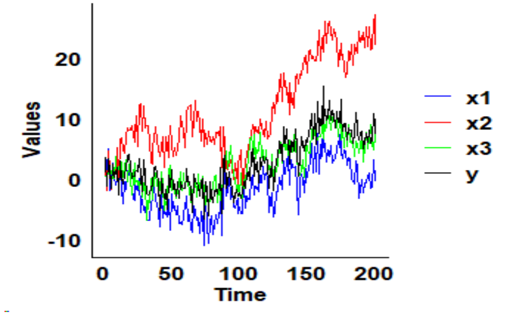
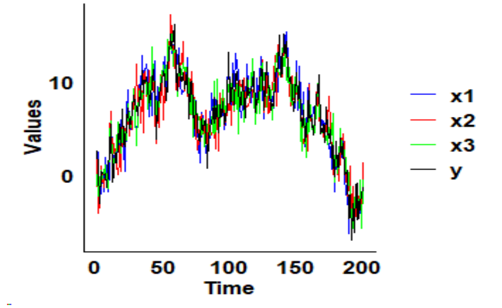
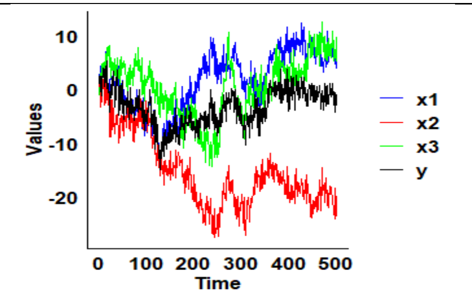
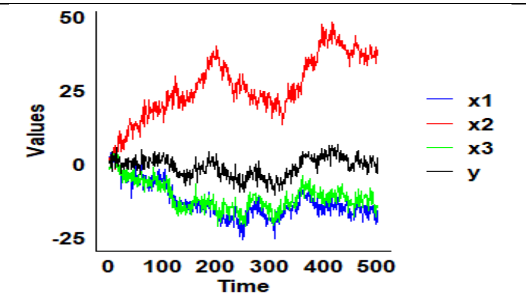
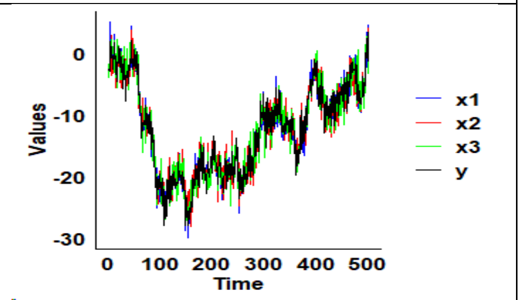
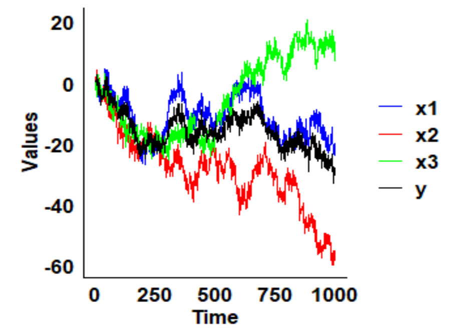
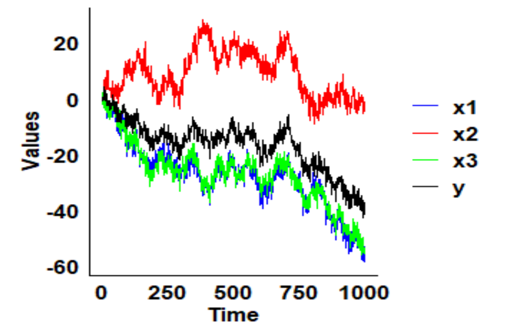
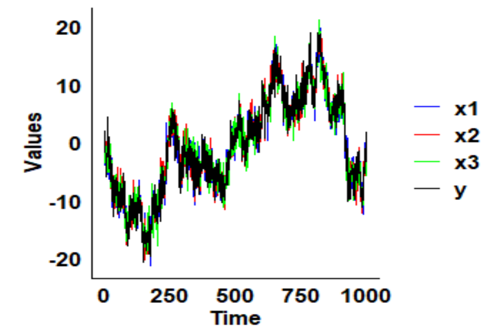
Swamy, P., & von zur Muehlen, P. (2020). Cointegration: Its fatal flaw and a proposed solution. *Sustainable Futures*, *2*, 100038. https://doi.org/10.1016/j.sftr.2020.100038

Wooldridge, J. (2006). Introductory Econometrics: A Modern Approach, *4*, 310-311.

Appendix

Unveiling the Pitfalls of Endogeneity in Forecasting: An Insight from Monte Carlo Simulation

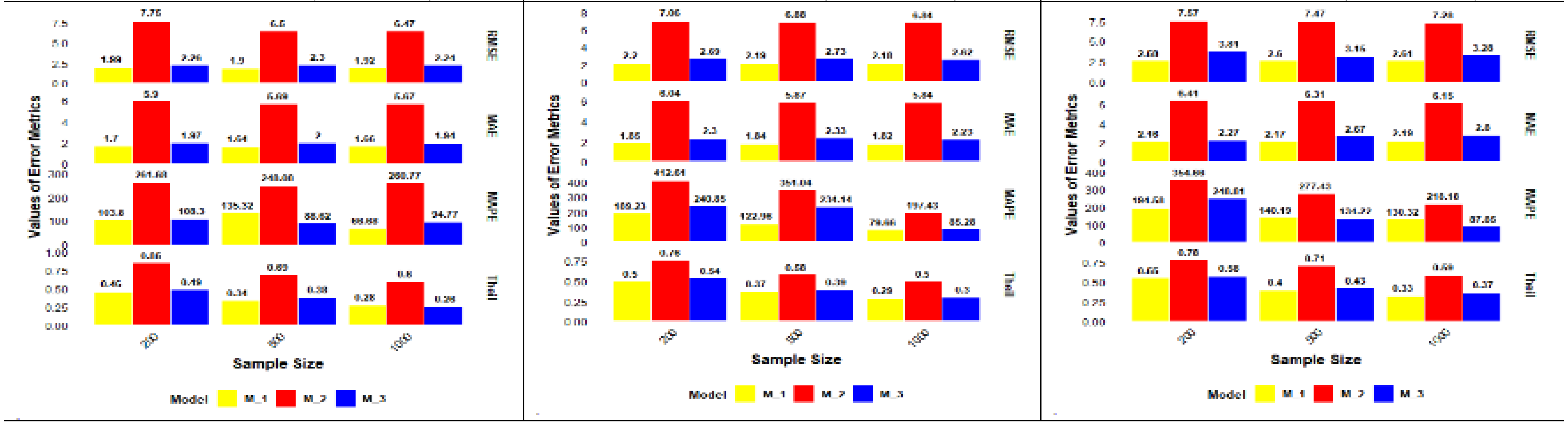
Figure 2. Graph of Cointegrated Series

<table>
<tr><th>N</th><th>$r = 1$</th><th>$r = 2$</th><th>$r = 3$</th></tr>
<tr><td>200</td>
<td>
</td>
<td>
</td>
<td>
</td></tr>
<tr><td>500</td>
<td>
</td>
<td>
</td>
<td>
</td></tr>
<tr><td>1000</td>
<td>
</td>
<td>
</td>
<td>
</td></tr>
</table>

Source: Author's work

Figure 3. Graphical Depiction of Forecast Evaluation, Scenario I $(r = 1)$ *omitting* X_1

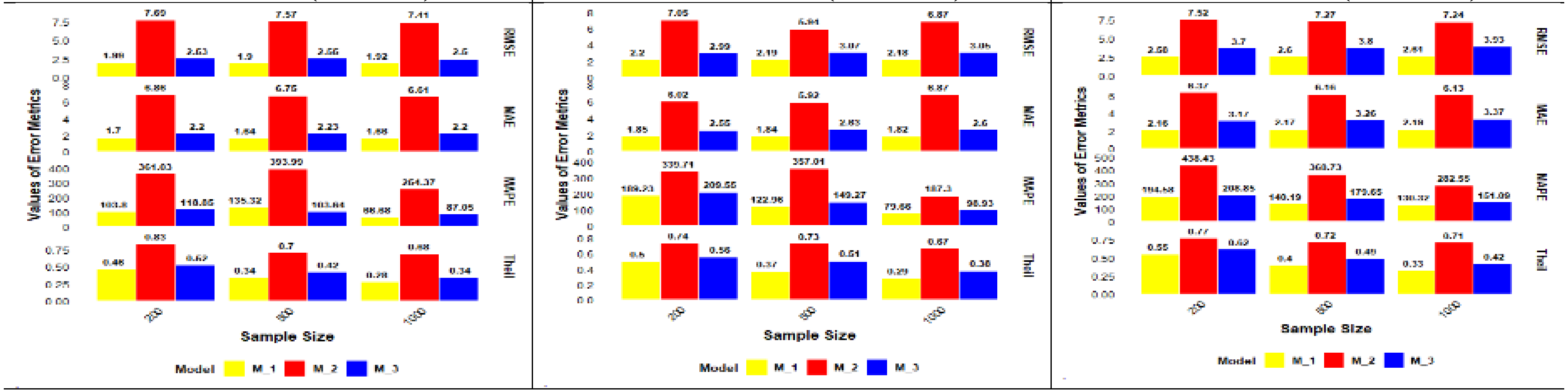
Forecast Horizon (5 Periods) | **Forecast Horizon (10 Periods)** | **Forecast Horizon (20 Periods)**



Source: Author's calculations.

Figure 4. Graphical Depiction of Forecast Evaluation, Scenario I $(r = 1)$ *omitting* X_2

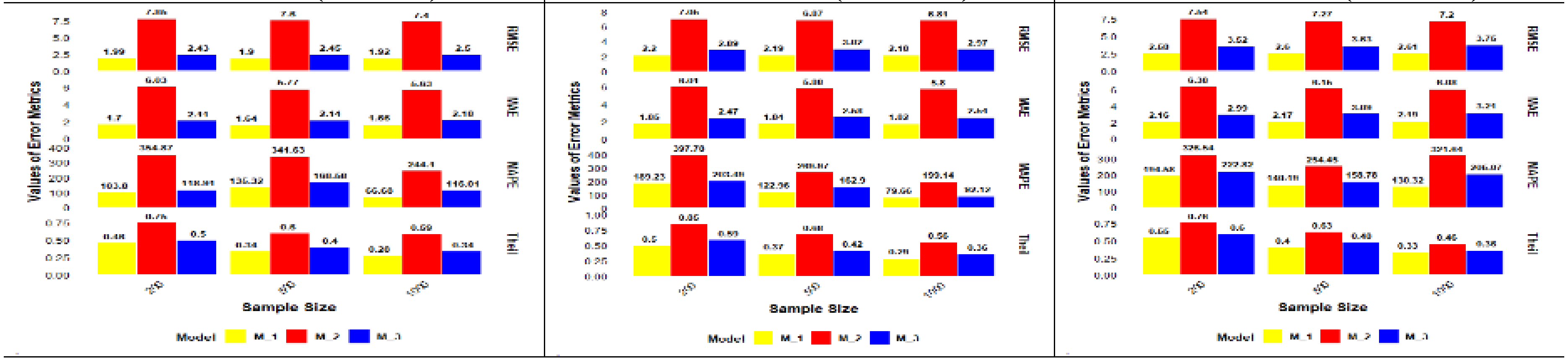
Forecast Horizon (5 Periods) | **Forecast Horizon (10 Periods)** | **Forecast Horizon (20 Periods)**



Source: Author's calculations.

Figure 5. Graphical Depiction of Forecast Evaluation, Scenario I ($r = 1$) *omitting* X_3

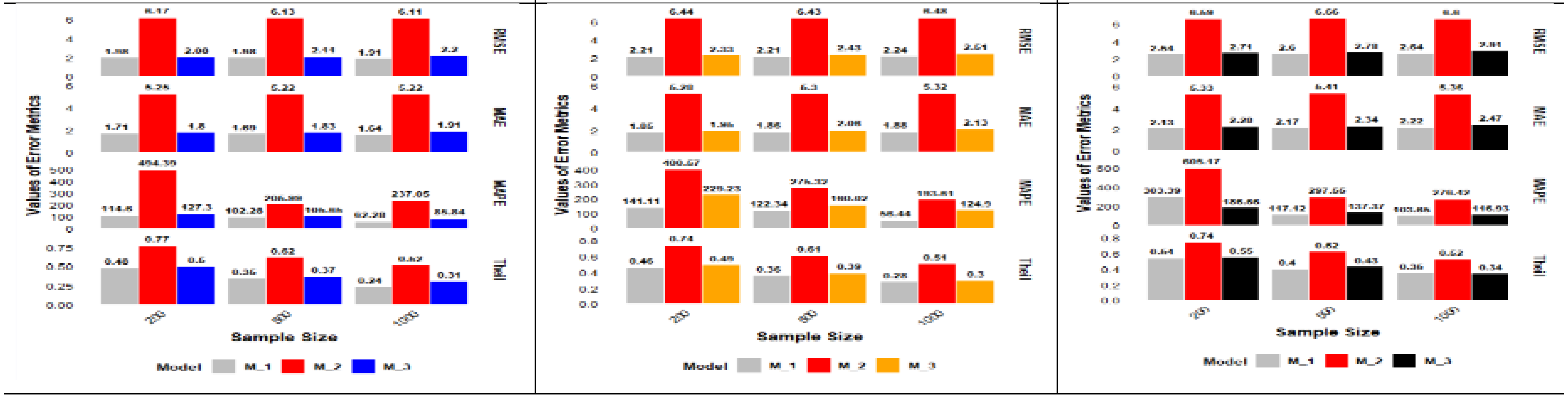
Forecast Horizon (5 Periods)	Forecast Horizon (10 Periods)	Forecast Horizon (20 Periods)



Source: Author's calculations.

Figure 6. Graphical Depiction of Forecast Evaluation, scenario II ($r = 2$) *omitting* X_1

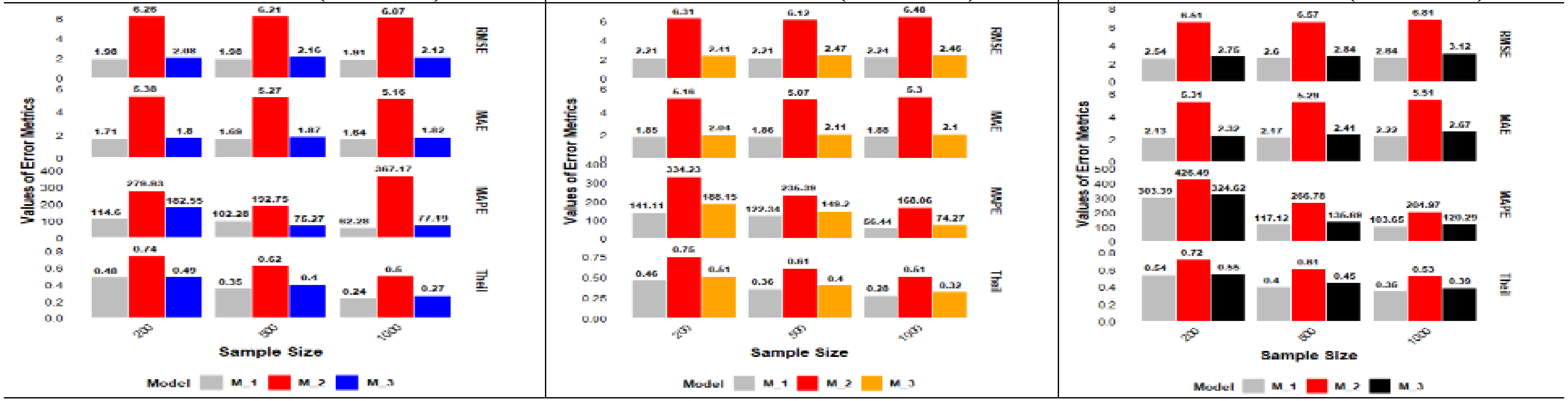
Forecast Horizon (5 Periods)	Forecast Horizon (10 Periods)	Forecast Horizon (20 Periods)



Source: Author's calculations.

Figure 7. Graphical Depiction of Forecast Evaluation, scenario II $(r = 2)$ *omitting* X_2

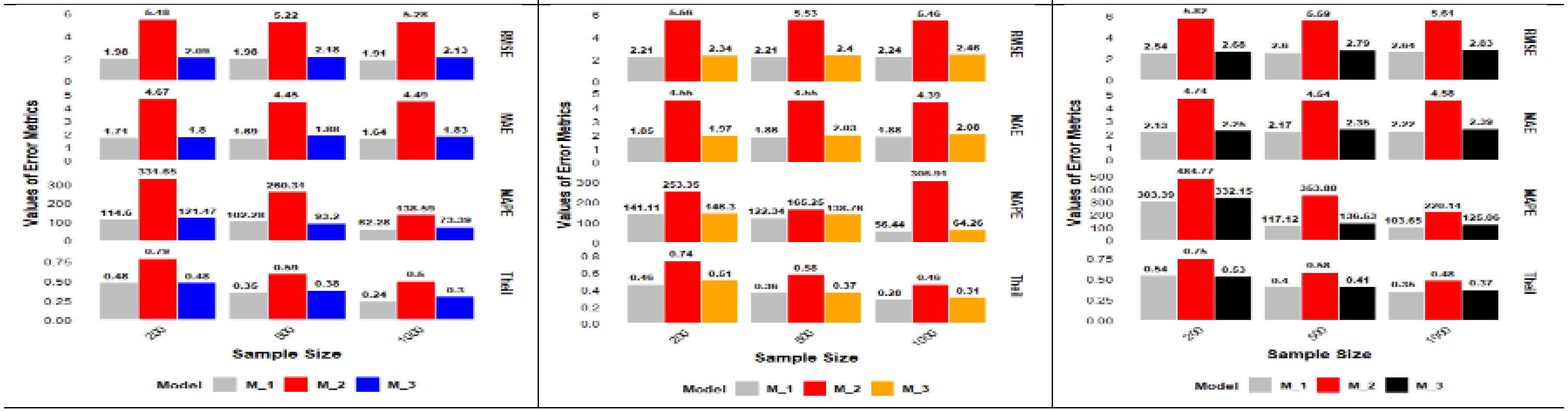
Forecast Horizon (5 Periods)	Forecast Horizon (10 Periods)	Forecast Horizon (20 Periods)



Source: Author's calculations.

Figure 8. *Graphical Depiction of Forecast Evaluation, scenario II* $(r = 2)$ *omitting* X_3

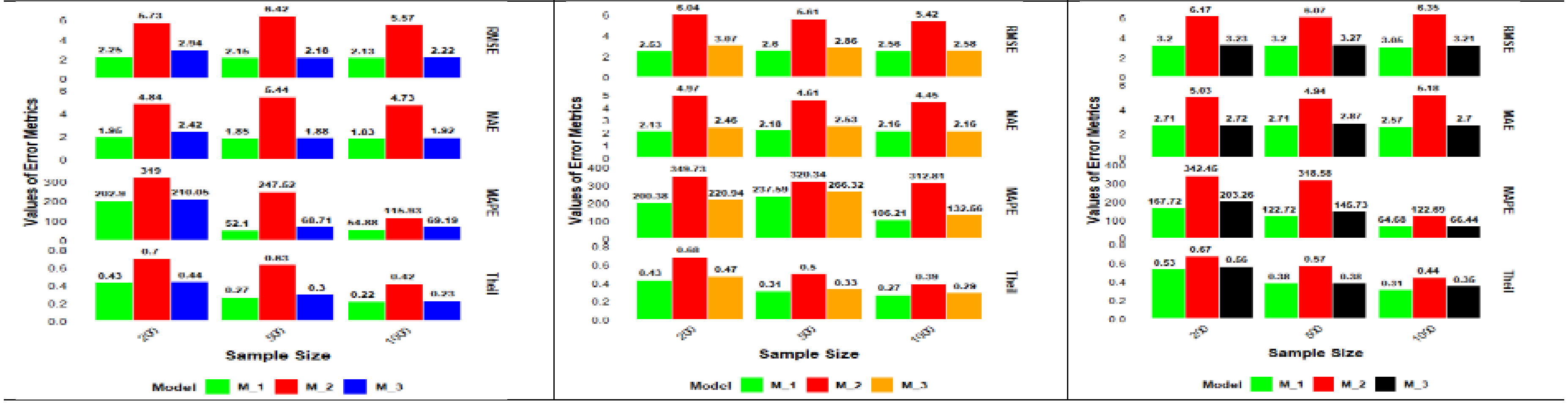
Forecast Horizon (5 Periods)	Forecast Horizon (10 Periods)	Forecast Horizon (20 Periods)



Source: Author's calculations.

Figure 9. Graphical Depiction of Forecast Evaluation, Scenario III ($r = 3$) omitting X_1

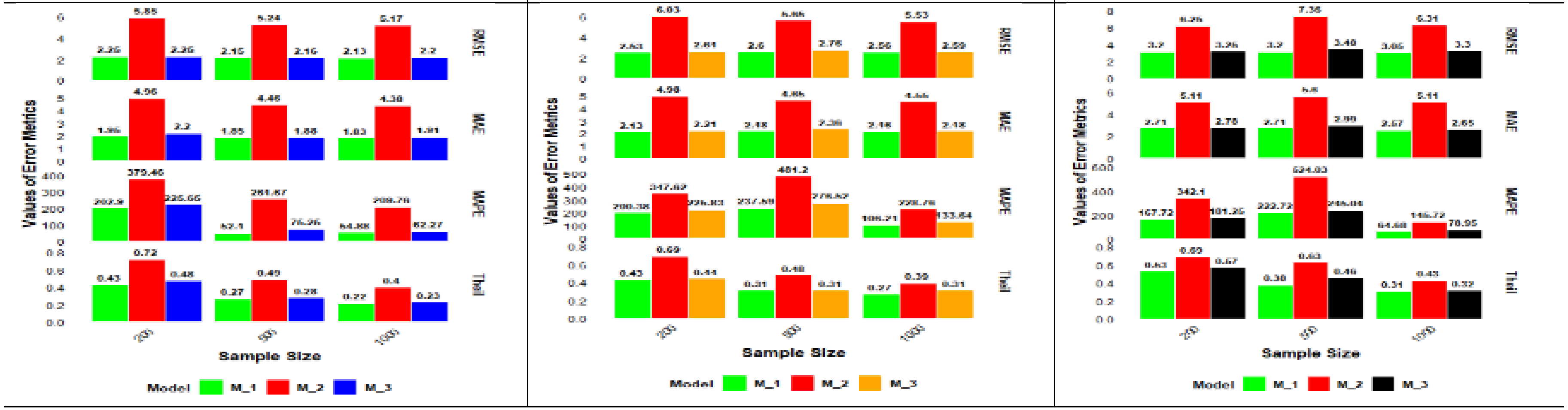
Forecast Horizon (5 Periods)	Forecast Horizon (10 Periods)	Forecast Horizon (20 Periods)



Source: Author's calculations.

Figure 10. Graphical Depiction of Forecast Evaluation, Scenario III ($r = 3$) omitting X_2

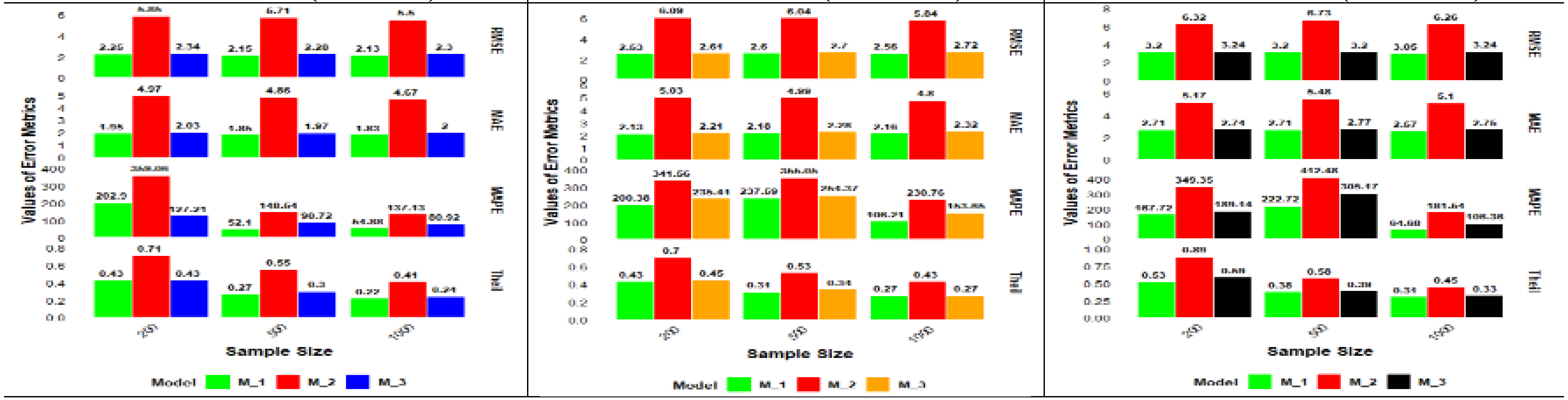
Forecast Horizon (5 Periods)	Forecast Horizon (10 Periods)	Forecast Horizon (20 Periods)



Source: Author's calculations.

Figure 11. Graphical Depiction of Forecast Evaluation, Scenario III ($r = 3$) omitting X_3

Forecast Horizon (5 Periods)	Forecast Horizon (10 Periods)	Forecast Horizon (20 Periods)



Source: Author's calculations.